Sozialwissenschaftliche Fakultät

Methodenmodul B.MZS11 (Statistik I)

Skript zu den Lerneinheiten

Lerneinheit 1: Wozu Statistik?

Empirische Sozialwissenschaften befassen sich mit:

- Regelmäßigkeiten (aber auch Abweichungen von Regelmäßigkeiten) in Vorstellungen, Verhalten und Interaktionen von Menschen, deren Ursachen und Konsequenzen.

Um festzustellen, ob bei einem Phänomen eine Regelmäßigkeit auftritt, müssen viele Beobachtungen vorliegen.

Beispiel: Es wird vermutet, dass sich die Wahlbeteiligung von Männern und Frauen systematisch unterscheidet. Es mag sein, dass mein Nachbar zur Linken nie wählt, meine Nachbarin zur Rechten dagegen an allen Wahlen teilnimmt. Aber die Gefahr eines falschen Induktionsschlusses ist hoch, wenn ich daraus schließe, dass Männer sich grundsätzlich seltener als Frauen an Wahlen beteiligen.

Aussagekräftigere Hinweise geben möglicherweise die Antworten von insgesamt 3234 im März bis Juli 1998 befragten Männern und Frauen, über die in einer als repräsentativ betrachteten Stichprobe der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) 1998 Antworten auf die Frage zur Wahlbeteiligung vorlagen:

{ (ja, männlich), (ja, weiblich), (weiß nicht, männlich), (ja, männlich), (weiß nicht, weiblich,) (nein, männlich), (nein, männlich), (ja, männlich), (ja, weiblich), (ja, weiblich), (nein, weiblich), (ja, männlich), (ja, weiblich), (ja, männlich), (ja, männlich), (ja, männlich), (ja, weiblich), (ja, weiblich), ... **}**

Warum Statistik?

Ohne statistische Aufbereitung der Antworten lässt sich nicht erkennen, ob es Unterschiede zwischen den Antworten von Männern und Frauen gibt.

Eine mögliche Aufbereitung ist die Zusammenstellung aller Antworten in einer Tabelle:

Beabsichtigte Wahlbeteiligung	Geschlecht Mann	Geschlecht Frau	Geschlecht Mann	Geschlecht Frau	Geschlecht Mann	Geschlecht Frau	Geschlecht Mann	Geschlecht Frau
- ja	998	1090	66.6%	62.8%	71.1%	65.0%	79.5%	72.2%	94.0%	90.6%
- nein	64	113	4.2%	6.5%	4.6%	6.7%	5.1%	7.5%	6.0%	9.4%
- weiß nicht	194	307	13.0%	17.7%	13.8%	18.3%	15.4%	20.3%		
- keine Angabe	148	168	9.9%	9.7%	10.5%	10.0%				
- nicht wahlberecht.	94	58	6.3%	3.3%						
Total	1498	1736	100.0% (1498)	100.0% (1736)	100.0% (1404)	100.0% (1678)	100.0% (1256)	100.0% (1510)	100.0% (1062)	100.0% (1203)
Prozentsatzdifferenz zwischen Männern und Frauen bei Wahlbeteiligung = ja:			3.8 Punkte		6.1 Punkte		7.3 Punkte		3.4 Punkte	
			Prozentuierung aller Antwortmöglichkeiten		Prozentuierung nur wahlberecht. Befragte		Prozentuierung ohne Verweigerungn		Prozentuierung nur ja/nein	

Um zu prüfen, ob Frauen mehr oder weniger oft wählen als Männer, werden die Angaben nach Geschlecht getrennt prozentuiert.
Problematisch erscheint allerdings, dass je nach Prozentuierungsbasis die Unterschiede zwischen Männern und Frauen von 3.4 bis 7.3 Prozentpunkten variieren.

Warum Statistik?

Hinzu kommt möglicherweise das Problem ungleicher Auswahlchancen:

- Die Allbus-Stichprobe ist so konstruiert, dass ca. 1/3 der Befragten aus den neuen Bundesländern kommen; tatsächlich beträgt der Bevölkerungsanteil in den neuen Ländern aber nur ca. 16%.
- In der Allbus-Stichprobe wird jeweils 1 Person aus einem Haushalt befragt; je mehr Personen in einem Haushalt leben, desto geringere ist daher die Chance, für die Stichprobe ausgewählt zu werden.

Über ***Gewichtungen*** können die ungleichen Auswahlchancen ausgeglichen werden:

Beabsichtigte Wahlbeteiligung	Geschlecht Mann	Frau	Geschlecht Mann	Frau
- ja	998	1090	1010	1098
- nein	64	113	61	110
- weiß nicht	194	307	184	281
- keine Angabe	148	168	142	167
- nicht wahlberecht.	94	58	110	65
Total	1498	1736	1514	1721

Prozentuierung u. Differenz

	ungewichtet Mann	Frau	gewichtet Mann	Frau	
	94.0	90.6	94.3	90.9	
	3.4		3.4		Punkte

Im Beispiel ergeben ungewichtete und gewichtete Daten die gleichen Prozentsatzdifferenzen, wenn nur „ja“ vs. „nein“ betrachtet wird.

ungewichtete Daten

gewichte Daten

Gewichtungsvariablen:
- Region: alte/neue Länder
- Haushaltsgröße

Warum Statistik?

Schließlich stellt sich die Frage, ob die verwendeten Daten für die Fragestellung überhaupt aussagekräftig sind?

Es ergeben sich jedenfalls wiederum jeweils andere Werte, wenn nicht nach der beabsichtigten Wahlbeteiligung, sondern nach der tatsächlichen Wahlbeteiligung bei der letzten Wahl gefragt wird. Noch andere Werte ergeben sich, wenn die repräsentative Wahlstatistik herangezogen wird, die an einer großen Stichprobe die tatsächliche Wahlbeteiligung erfasst.

Beabsichtigte Wahlbeteiligung	Wahlabsicht BTW 1998 Mann	Frau	Rückerinnerung BTW 1994 Mann	Frau	BTW 2002 Mann	Frau
- ja	94.3%	90.9%	91.8%	91.8%	79.9%	79.4%
- nein	5.7%	9.1%	8.2%	8.2%	21.1%	21.6%
Prozentsatzdifferenz	3.4 Punkte		0.0 Punkte		0.5 Punkte	
	(1071)	(1208)	(1336)	(1575)		

gewichtete Daten

gewichtete Daten

Wahlbeteiligung bei BTW 2002 nach repräsentativer Wahlstatistik

Tatsächliche Beteiligung Bundestagswahl 1998 82.3%

Tatsächliche Beteiligung Bundestagswahl 1994 79.1%

Warum Statistik?

Das Beispiel zeigt:
In Abhängigkeit von den herangezogenen Daten und der Art der statistischen Aufbereitung (im Beispiel Gewichtung und Prozentuierungsbasis) können sich unterschiedliche Schlussfolgerungen ergeben, die sich auch widersprechen können.
Bei sich widersprechenden Befunden kann jedoch nur einer empirisch zutreffend sein: Die Wahlbeteiligung von Männern und Frauen ist entweder verschieden oder gleich!

Daraus folgt für die empirische Forschung:
Erst das Zusammenspiel von statistischen Kenntnissen und inhaltlicher Erfahrung ermöglicht es, Ergebnisse empirischer Untersuchungen angemessen zu beurteilen:

- Die Statistik hilft, die Datenmengen so zu strukturieren, dass überhaupt Aussagen möglich sind.
- Die inhaltliche Erfahrung hilft bei der Beurteilung der Frage, ob die Daten und deren Aufbereitung für die jeweilige Fragestellung relevant sind.

Statistikkenntnisse sind also eine notwendige (aber nicht hinreichende) Bedingung, um empirische Regelmäßigkeiten aufzudecken.

In der Statistik-Ausbildung werden diese notwendigen Kenntnisse vermittelt.
Beispielhafte Aufgaben helfen, nicht nur den Lehrstoff zu verfestigen, sondern auch, die Rolle inhaltlicher Gesichtspunkte einzuschätzen.
Umfassende statistische Kompetenz ergibt erst die Anwendung in der Fachwissenschaft.

Gegenstand der Statistik

In der Statistik geht es um die
***mathematische Modellierung von Verteilungen*.**

Was bedeutet das?

a) ***Verteilung***: (Betrachtung von) Eigenschaften einer Menge von Einheiten
Beispiele:
- *Gemeinsamkeiten und Unterschiede beim Einkommen von Haushalten*
- *Bewertungen von Parteien in einem Bundesland*
- *Häufigkeiten, Art und Intensität von Konflikten zwischen Partnern*
- *Konsum alkoholischer Getränke einer Person in einem Zeitraum*

b) ***Modellierung***: **Abstraktion von realen Einheiten**
durch Konzentration auf relevante und Ignorierung irrelevanter Aspekte
⇒ Informationsverdichtung u. Informationsreduktion

Beispiel für ein Modell: Straßenkarte als Modell einer Landschaft
Beispiele für statistische Modelle:

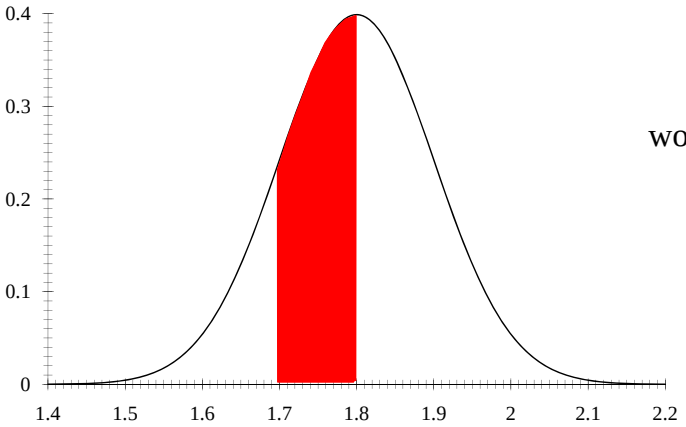
- *Beschreibung von Verteilungen durch typische Werte:*
  z.B. das durchschnittliche Jahreseinkommen eines Haushalts
- *Beschreibung der Unterschiedlichkeit der Einheiten:*
  z.B. der Abstand zwischen den mittleren Einkommen der 25% einkommensschwächsten und den 25% einkommensreichsten Haushalten
- *Beschreibung von Zusammenhängen:*
  z.B: Pro zusätzlichem Jahr an Bildung erhöht sich das Einkommen im Mittel um 250.--€/Monat

Was ist Statistik ?

c) ***Mathematische Modellierung***:

Modellformulierung in „Sprache“ der Mathematik (Symbole u. Formeln)

Beispiel: *Verteilung der Körpergröße (X) von Erwachsenen in einer Population. Empirische Verteilungen lassen sich oft durch spezifische mathematische Formeln beschreiben. So ist die Körpergröße in einer Gruppe in der Regel annähernd* ***normalverteilt*** *und folgt damit folgender Formel:*



$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma_X^2}} \cdot e^{-\frac{1}{2} \cdot \frac{(x-\mu_X)^2}{\sigma_X^2}}$$

wobei: μ_X =: durchschnittliche Körpergröße (Erwartungswert)

σ_X^2 =: Ausmaß der Unterschiedlichkeit der Körpergrößen (Varianz)

Im Beispiel: $\mu_X = 1.80$ *und* $\sigma^2_X = 0.1$

Die Fläche unter der Kurve gibt den Anteil der Verteilung an:
Im Beispiel hat ein Anteil von 34.% eine Körpergröße zwischen 1.70 und 1.80 Meter.

Klassische Einteilung der Statistik

B.MZS.11	Deskriptive Statistik	Induktive Statistik / Inferenzstatistik
Univariate Verteilungen	Verteilungsparameter (Quantile, Lagemaße, Streuungsmaße)	Wahrscheinlichkeitstheorie, Schätzen und Testen
Bivariate Verteilungen	Beschreibung und Prüfung von bivariaten Zuammenhängen	
Multivariate Verteilungen	Drittvariablenkontrolle Konditionale u. Partielle Effekte Prüfung der Angemessenheit statistischer Modelle	

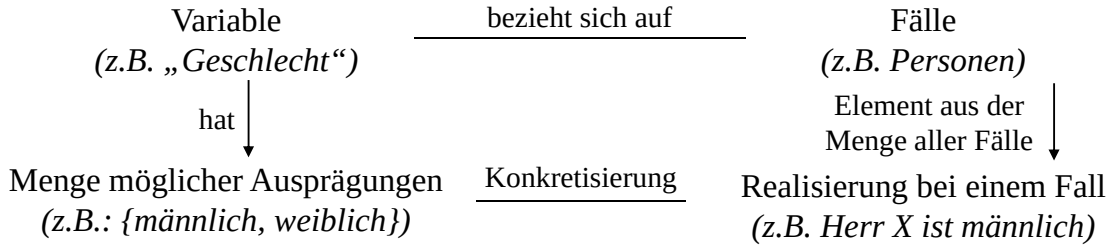
Lerneinheit 2:
Von der Datenmatrix zur univariaten Häufigkeitstabelle

Empirische Verteilungen beziehen sich auf Gemeinsamkeiten und Unterschiede zwischen interessierenden Eigenschaften einer Menge empirischer Objekte.

- Die ***empirischen Objekte*** (z.B. Menschen, Institutionen, Gesellschaften aber auch Ereignisse wie Scheidungen) werden in der statistischen Analyse meist als ***Untersuchungseinheiten*** oder ***Fälle*** bezeichnet.
- Die ***Gesamtmenge*** aller interessierenden Fälle bildete die ***Grundgesamtheit*** oder ***Population***. Wird nur eine Teilmenge einer Population empirisch betrachtet, spricht man von einer ***Stichprobe*** *(engl:* ***Sample****)*.
- Die interessierenden ***Eigenschaften*** oder ***Merkmale*** der Objekte einer Population werden in der Statistik als ***Variablen*** bezeichnet.
- Die ***möglichen Auftretensformen*** einer Variablen (z.B: Eigenschaft ist vorhanden oder nicht vorhanden oder Eigenschaft ist in bestimmten Ausmaß vorhanden) sind die ***Ausprägungen*** oder ***Werte*** *(engl:* ***values*** *oder codes)* einer Variable; die ***Menge aller möglichen Ausprägungen*** bilden den ***Wertebereich*** einer Variablen.
- Die tatsächlich ***vorkommende Ausprägung*** einer Variable bei einem Fall zu einem Zeitpunkt wird als ***Realisierung*** oder ***Realisation*** *(engl:* ***realisation****)* bezeichnet.

Variablen, Ausprägungen und Realisierungen

Es ist wichtig, zwischen Variablen und deren Ausprägungen sowie zwischen Fällen und deren Realisierungen zu unterscheiden:



Wenn in einer Menge **alle Fälle** bei einer Variable die ***gleiche Ausprägung*** haben, reduziert sich die Variable zu einer ***Konstanten***.

Variablen lassen sich entsprechend den Eigenschaften ihrer Ausprägungen unterscheiden in:

- diskrete vs. kontinuierliche Variablen:
  Die Ausprägungen ***kontinuierlicher*** oder ***stetiger Variablen*** lassen sich nur als ***reelle Zahlen*** darstellen während Ausprägungen ***diskreter Variablen*** durch ***ganze Zahlen*** kodiert werden können.
  Obwohl empirische Variablen konzeptionell kontinuierlich sein können (z.B. die Körpergröße von Personen), sind sie aufgrund begrenzter Messgenauigkeiten praktisch nur diskret messbar.

Variablen, Ausprägungen und Realisierungen

- Diskrete Variablen mit wenigen möglichen Ausprägungen werden auch als ***kategoriale Variablen*** bezeichnet und ihre Werte entsprechend als ***Kategorien***.
  Hat eine kategorieale Variable nur ***zwei Ausprägungen*** ist diese Variable ***dichotom***, hat sie ***drei Ausprägungen*** ist sie ***trichotom***. Diskrete Ausprägungen mit mehr als drei Ausprägungen sind ***polytom***.
- Entsprechend dem ***Skalen***- oder ***Messniveau*** von Variablen wird unterschieden zwischen:
  - ***nominalskalierten (nominalen) Variablen***, bei denen Realisationen nur hinsichtlich der Gleichheit oder Verschiedenheit von Ausprägungen verglichen werden können,
  - ***ordinalskalierten (ordinalen) Variablen***, bei denen Realisierungen zusätzlich hinsichtlich eines mehr oder weniger an der betrachteten Eigenschaft vergleichen werden können
  - und ***metrischen Variablen***, bei denen zusätzlich Vergleiche hinsichtlich ihres Unterschiedlichkeit möglich sind. Gibt es einen Bezugspunkt, so dass Ausprägungsverhältnisse informativ sind (z.B: Haushalt A hat doppelt so viel Einkommen pro Monat wie Haushalt B), dann hat eine metrische Variable ***Ratio***- oder ***Proportionalskalenniveau***, ansonsten ***Intervallskalenniveau***.
- Bei empirischen Variablen kann es vorkommen, dass nicht bei allen Fällen die tatsächliche Realisierung beobachtet wurde. Um Fälle mit empirisch beobachteten und empirisch nicht beobachteten Ausprägungen bei einer interessierenden Variable unterscheiden zu können, wírd der Wertebereich um spezielle Ausprägungen für ***fehlende*** oder ***ungültige Messungen*** *(engl:* ***missing values****)* ergänzt.

Ungültige Werte

Vor allem in der ***Umfrageforschung*** *(engl:* ***survey research****)* macht es Sinn, zwischen unterschiedlichen Arten von ungültigen Werten zu unterscheiden, so nach Ausfällen, weil eine Realisierung nicht vorkommen kann (z.B: Anzahl der Arbeitsstunden bei nicht beschäftigten Personen), nach Antwortverweigerungen und nach Fehlern bei der Erhebung.
Dabei haben sich Konventionen für die Kodierung ungültiger Fälle eingespielt, die möglichst eingehalten werden sollten:

	Endziffer	einstellige Variablen	zweistellige Variablen	dreistellige Variablen
Verweigerung	7	7	97	997
weiß nicht	8	8	98	998
keine Angabe	9	9	99	999
trifft nicht zu	0	0	0	0

Die Datenmatrix

Ausgangspunkt der statistischen Datenanayse ist eine Datenmatrix, die die in der Regel als Zahlenwerte kodierten (gemessenen) Realisierungen aller beobachteten Variablen einer Stichprobe oder Population enthält.

Als Beispiel wird von einer kleinen Umfrage ausgegangen, die neben Fragen zur Demokratiezufriedenheit und Beeinflussbarkeit der Politik, das Geburtsjahr und als Beobachtungsgröße das Geschlecht der Befragten enthält.

Beispielfragebogen

FRAGE	ANTWORT	Kode
1. Sind Sie mit der Art und Weise, wie die Demokratie in der Bundesrepublik funktioniert, alles in allem gesehen ...	... sehr zufrieden,..... ... eher zufrieden,..... ... eher unzufrieden,..... ... oder völlig unzufrieden?..... *weiß nicht* [1] *keine Angabe*	*4* *3* *2* *1* *8* *9*
2. Nun einige Aussagen, über die man verschiedener Ansicht sein kann. Sagen Sie mir bitte jeweils, ob Sie der Aussage eher zustimmen oder eher nicht zustimmen. a) Leute wie ich haben so oder so keinen Einfluss darauf, was die Regierung tut b) Die Parteien wollen nur die Stimmen der Wähler, ihre Ansichten interessieren sie nicht	stimme eher zu / stimme eher nicht zu / *weiß nicht* / *keine Angabe* *1* *2* *8* *9* *1* *2* *8* *9*	
ohne Abfrage eintragen! *Das Interview wurde geführt mit...*	*einem Mann.....* *einer Frau.....*	*1* *2*
4. Zum Schluss noch eine Frage zur Statistik. Sagen Sie mir bitte, in welchem Jahr Sie geboren sind.	*Geburtsjahr vierstellig eintragen!* 1943 *keine Angabe*	*9999*

[1] Kursiver gedruckter Text ist für den Interviewer bestimmt und wird nicht vorgelesen.

Die beobachteten oder erfragten Variablen werden über ein Erhebungsinstrument erfasst. Zusätzlich werden meist weitere Variablen, etwa zum Erhebungstag und Erhebungsort und bei Befragungen Informationen über den Interviewer bzw. die Interviewerin aufgenommen

Den einzelnen Fällen sollten unbedingt eindeutige Indexnummern zugewiesen werden, die etwa nach der Reihenfolge der Beobachtung oder dem Eingang der Daten gebildet werden. Diese Nummern heißen ***Fallnummern*** (oder Identifikationsnummern).

Datenmatrix

Beispiel einer Datenmatrix:

	Merkmale der Untersuchungseinheiten (Variablen)					
	Fall-nummer ID	Antwort Frage 1 F1	Antwort Frage 2a F2A	Antwort Frage 2b F2B	Ge-schlecht F3	Geburts-jahr F4
Fälle	1	3	2	2	1	1943
	2	2	8	1	2	1960
	3	4	1	2	2	1957
	4	9	8	1	1	1939
	5	2	2	1	2	9999
	6	8	8	1	1	1956
	7	4	1	2	2	1970
	8	1	1	2	1	1920
	9	3	3	1	2	1956
	10	4	2	2	2	1966

In einer Datenmatrix sind die Informationen i.a. so angeordnet, dass jede Zeile die gesamten verfügbaren Informationen (Realisierungen aller Variablen) bei einem Fall enthält, und dass jede Spalte alle Realisierungen einer Variablen über alle Fälle enthält.

Information über den ersten Fall (Realisierungen aller Variablen bei Fall 1)

Informationen über die erste Variable (Realisierungen aller Fälle bei Variable F1)

Univariate Verteilungen

Univariate Verteilungen beziehen sich auf eine einzige Variable. In einer Datenmatrix gibt jede Spalte der Matrix die univariate Verteilung einer Variable über die Menge der beobachteten Fälle wieder.

Fall-nummer	Antwort Frage 1 F1
1	3
2	2
3	4
4	9
5	2
6	8
7	4
8	1
9	3
10	4

Um die Realisierung eines Falles zu identifizieren, werden meist Fallnummern als Indizes verwendet.
Wenn - wie im Beispiel - F1 eine Variable bezeichnet, dann ist $F1_3$ die Realisierung dieser Variable beim dritten Fall der in der Datenmatrix aufgenommenen Fälle, im Beispiel also der Wert „4“.
Zur Unterscheidung zwischen Variablen und Ausprägungen, werden Variablen oft durch große, Ausprägungen durch kleine Buchstaben gekennzeichnet, also X für die Variable und x_3 für die Realisierung des dritten Falles.

Um Variablen zu bezeichnen, werden oft Variablennamen mit wenigen Zeichen und Buchstaben verwendet, im Beispiel etwa F1 für die erste Variable des Beispielfragebogens, also die Frage nach der Demokratiezufriedenheit.
In Formeln werden metrische Variablen oft durch Buchstaben vom Ende des Alphabets (also W, X, Y, Z) und nominalskalierte Variablen oft durch Buchstaben vom Anfang des Alphabets (A, B, C) symbolisiert.

Univariate Verteilungen

Fall-nummer	Antwort Frage 1 F1
1	3
2	2
3	4
4	9
5	2
6	8
7	4
8	1
9	3
10	4

Fall-nummer	geord-nete Falln. (i)	geordn. Antwort F1
8	1	1
2	2	2
5	3	2
1	4	3
9	5	3
3	6	4
7	7	4
10	8	4
6	9	8
4	10	9

Rang-nummer (k)	Rang-nummer (k)
1	1
2	2.5
2	2.5
3	4.5
3	4.5
4	7
4	7
4	7
5	9
6	10

Für spezielle Berechnungen kann es notwendig sein, anstelle der ursprünglichen Reihenfolge der Realisationen die Fälle aufsteigend zu sortieren.
Um die Positionen geordneter Realisierungsreihen von den ursprünglichen Positionen zu unterscheiden, werden die Indizes dann meist in Klammern dargestellt. $F1_{(4)}$ ist also die Realisation des viertkleinsten Falles von F1 in der Datenmatrix, im Beispiel der Fall mit der Fallnummer 1 mit der Ausprägung 3.

Anstelle der beobachteten Ausprägungen werden bisweilen auch Rangnummern entsprechend der ordinalen Anordnung der Ausprägungen vergeben.

Treten bei verschiedenen Fällen die gleichen Rangnummern auf (sog. ***Bindungen***, *engl:* ***ties***), können auch mittlere Rangnummern vergeben werden.

Häufigkeitstabellen

Wenn eine univariate Verteilung betrachtet wird und die Ausprägungen in einer Spalte der Datenmatrix wiederholt vorkommen, ist es sinnvoller, anstelle der Betrachtung einer Spalte der Datenmatrix die Verteilung in einer ***univariaten Häufigkeitstabelle*** darzustellen.

Als Beispiel zeigt die folgende Tabelle die Verteilung der Antworten auf die Frage nach der Demokratizufriedenheit (F1):

F1: Demokratiezufriedenheit			Anteile		kumulierte
Ausprägung	Code	Häufigkeit	insgesamt	nur gültige	Anteile
völlig unzufrieden	1	1	0.100	0.125	0.125
eher unzufrieden	2	2	0.200	0.250	0.375
eher zufrieden	3	2	0.200	0.250	0.625
sehr zufrieden	4	3	0.300	0.375	1.000
weiß nicht	8	1	0.100	--	
keine Angabe	9	1	0.100	--	
Summe		10	1.000	1.000	
(gültige Fälle: 8; ungültige Fälle 2)					

In einer Häufigkeitstabelle werden die Häufikgeiten aller (in der Datenmatrix vorkommenden) Ausprägungen einer Variable nach ihren Werten (Codes) aufsteigend sortiert zusammengefasst.

Häufigkeitstabellen

F1: Demokratiezufriedenheit			Anteile		kumulierte
Ausprägung	Code	Häufigkeit	insgesamt	nur gültige	Anteile
völlig unzufrieden	1	1	0.100	0.125	0.125
eher unzufrieden	2	2	0.200	0.250	0.375
eher zufrieden	3	2	0.200	0.250	0.625
sehr zufrieden	4	3	0.300	0.375	1.000
weiß nicht	8	1	0.100	--	
keine Angabe	9	1	0.100	--	
Summe		10	1.000	1.000	
(gültige Fälle: 8; ungültige Fälle 2)					

Die Tabelle enthält neben den sowohl sprachlich (1. Spalte) als auch numerisch (2. Spalte) wiedergebenen Ausprägungen die ***absoluten Häufigkeiten*** mit der jede Ausprägung im Datensatz vorkommt.

Im Beispiel kommt die 1. Ausprägung ("völlig unzufrieden", Code „1") mit der absoluten Häufigkeit 1 vor, die zweite Ausprägung („eher unzufrieden") mit der Häufigkeit 2, die dritte Ausprägung („eher zufrieden") mit der Häufigkeit 2, die 4. Ausprägung ("sehr zufrieden") mit der absoluten Häufigkeit 3, die ungültige Ausprägung "weiß nicht" mit der absoluten Häufigkeit 1 und die ungültige Ausprägung „keine Angabe" mit der Häufigkeit 1.

Häufigkeitstabellen

F1: Demokratiezufriedenheit			Anteile		kumulierte
Ausprägung	Code	Häufigkeit	insgesamt	nur gültige	Anteile
völlig unzufrieden	1	1	0.100	0.125	0.125
eher unzufrieden	2	2	0.200	0.250	0.375
eher zufrieden	3	2	0.200	0.250	0.625
sehr zufrieden	4	3	0.300	0.375	1.000
weiß nicht	8	1	0.100	--	
keine Angabe	9	1	0.100	--	
Summe		10	1.000	1.000	
(gültige Fälle: 8; ungültige Fälle 2)					

Aus der Tabelle ist auch ersichtlich, dass es neben den vier gültigen Ausprägungen zwei Ausprägungen gibt, die als ungültig deklariert sind.
Ob eine Ausprägung als „ungültig“ bewertet wird, hängt von der jeweiligen Fragestellung ab.

Dies Festlegung ungültiger Werte hat Auswirkungen auf die Berechnung der ***Anteile*** (***relativen Häufigkeiten***), die sich aus der Divison der absoluten Häufigkeiten durch die Gesamtzahl berechnen.

> *Anteile können sich auf die gesamte Fallzahl (4. Spalte) oder nur auf die Zahl der Fälle mit gültigen Antworten (5. Spalte) beziehen.*

Häufigkeitstabellen

F1: Demokratiezufriedenheit			Anteile		kumulierte
Ausprägung	Code	Häufigkeit	insgesamt	nur gültige	Anteile
völlig unzufrieden	1	1	0.100	0.125	0.125
eher unzufrieden	2	2	0.200	0.250	0.375
eher zufrieden	3	2	0.200	0.250	0.625
sehr zufrieden	4	3	0.300	0.375	1.000
weiß nicht	8	1	0.100	--	
keine Angabe	9	1	0.100	--	
Summe		10	1.000	1.000	
(gültige Fälle: 8; ungültige Fälle 2)					

In der *letzten Spalte* werden die relativen Häufigkeiten der gültigen Fälle aufsummiert.

> *Die Zahl 0.375 in der Zeile mit dem Code 2 „eher unzufrieden“ ist also die Summe der Anteile, die diesen oder einen kleineren Wert aufweisen, hier also die Summe der völlig unzufriedenen (Anteil = 0.125) plus der eher unzufriedenen (Anteil = 0.250) Personen: 0.375 = 0.125 + 0.250.*

Kumulierte Anteile machen nur ab ordinalem Messniveau Sinn.
Da ungültige Werte sich nicht in die Rangordnung der übrigen Ausprägungen einordnen lassen, werden sie bei der Kumulierung nicht berücksichtigt.

Konventionen

Zur Darstellung in statistischen Formeln gibt es eine Reihe von Konventionen, mit denen Variablen, Ausprägungen und Realisierungen, gemessene Werte und Transformationen gekennzeichnet werden.

Variable	X, Y, Z, V2, V_2
Ausprägung	x, y, z, v_2, v_2
Anzahl der Fälle	n
Realisation des i-ten Falles (i=1,2,...,n) der Variablen X	x_i
Realisation des i-ten sortierten Falles	$x_{(i)}$
Mittelwert der k-ten Gruppe bei gruppierten Daten	$m_{(k)}$
Ausprägung k (k=1,2,...,K) der Variablen X	x_k, $x_{(k)}$
Anzahl der Fälle mit der Ausprägung x_k	n_x, n_k, $n_{(k)}$
Anteil der Fälle mit der Ausprägung x_k	p_x, p_k, $p_{(k)}$
Prozent der Fälle mit der Ausprägung x_k	$p_k\% = p_k \cdot 100$

Anwendungsbeispiele:
Berechnungsformel für Anteile: $p_{(k)} = \frac{n_{(k)}}{n}$ bzw. $p_x = \frac{n_x}{n}$

Berechnungsformel für kumulierte Anteile: $cp_{(k)} = p\left(X \le x_{(k)}\right) = \sum_{j=1}^{k} p_{(j)} = \frac{\sum_{j=1}^{k} n_{(j)}}{n}$

Häufigkeitstabellen: Berechnung von Anteilen

		$n_{(k)}$	$p_{(k)}$	$p_{(k)}$	$cp_{(k)} = \sum p_{(k)}$
F1			Anteile		kumulierte
Ausprägung	Code	Häufigkeit	insgesamt	nur gültige	Anteile
völlig unzufrieden	1	1	0.100	0.125	0.125
eher unzufrieden	2	2	0.200	0.250	0.375
eher zufrieden	3	2	0.200	0.250	0.625
sehr zufrieden	4	3	0.300	0.375	1.000
weiß nicht	8	1	0.100	--	
keine Angabe	9	1	0.100	--	
Summe		n = 10	1.000	1.000	
(gültige Fälle: 8; ungültige Fälle 2)					

$p_{(k)} = \frac{n_{(k)}}{n}$ bzw. $p_x = \frac{n_x}{n}$

Bei der Indizierung wird manchmal auch die zugeordnete Ausprägung als Indexwert verwendet.

$p_x = \frac{n_x}{n}$

$p_{(1)} = n_{(1)} / n = 1/10 = 0.1$
$p_{(2)} = n_{(2)} / n = 2/10 = 0.2$
$p_{(3)} = n_{(3)} / n = 2/10 = 0.2$
$p_{(4)} = n_{(4)} / n = 3/10 = 0.3$
$p_{(5)} = n_{(5)} / n = 1/10 = 0.1$
$p_{(6)} = n_{(6)} / n = 1/10 = 0.1$

alternativ:
$= p_8 = n_8 / n$
$= p_9 = n_9 / n$

Häufigkeitstabellen: Interpretation

F003: Demokratiezufriedenheit			Anteile		kumulierte
Ausprägung	Code	Häufigkeit	insgesamt	nur gültige	Anteile
sehr unzufrieden	1	94	0.056	0.057	0.057
ziemlich unzufrieden	2	169	0.102	0.102	0.159
etwas unzufrieden	3	283	0.170	0.171	0.330
etwas zufrieden	4	396	0.238	0.239	0.568
ziemlich zufrieden	5	580	0.349	0.350	0.919
sehr zufrieden	6	135	0.081	0.081	1.000
weiß nicht	8	6	0.004	--	
keine Angabe	9	0	0.000	--	
Summe		1663	1.000	1.000	

(BTW 05 Nachwahl nur alte Länder; gültige Fälle: 1657)

Vor der Interpretation einer empirischen Häufigkeitstabelle sollte zunächst versucht werden die Datenqualität zu beurteilen. Hinweise hierzu geben insbesondere:

- Angaben über die Datenquelle

  Ist die Datenquelle nicht angegeben, ist in jedem Fall Vorsicht angeraten.
  Im Beispiel kommen die Daten aus einer repräsentativen persönlichen Umfrage, die im Anschluss an die Bundestagswahl 2005 durch professionelle Interviewer des Umfrageinstituts Infratest durchgeführt wurden. Aufgeführt sind nur Angaben der Befragten aus den alten Bundesländern.

Häufigkeitstabellen: Interpretation

F003: Demokratiezufriedenheit			Anteile		kumulierte
Ausprägung	Code	Häufigkeit	insgesamt	nur gültige	Anteile
sehr unzufrieden	1	94	0.056	0.057	0.057
ziemlich unzufrieden	2	169	0.102	0.102	0.159
etwas unzufrieden	3	283	0.170	0.171	0.330
etwas zufrieden	4	396	0.238	0.239	0.568
ziemlich zufrieden	5	580	0.349	0.350	0.919
sehr zufrieden	6	135	0.081	0.081	1.000
weiß nicht	8	6	0.004	--	
keine Angabe	9	0	0.000	--	
Summe		1663	1.000	1.000	

(BTW 05 Nachwahl nur alte Länder; gültige Fälle: 1657)

Vor der Interpretation einer empirischen Häufigkeitstabelle sollte zunächst versucht werden die Datenqualität zu beurteilen. Hinweise hierzu geben insbesondere:

- Hinweise über ungültige Fälle

  Auf die Frage nach dem Funktionieren der Demokratie in der BRD haben nur 0.4% der Befragten aus den alten Ländern eine ausweichende Antwort („weiß nicht“) gegeben.

Häufigkeitstabellen: Interpretation

F003: Demokratiezufriedenheit			Anteile		kumulierte
Ausprägung	Code	Häufigkeit	insgesamt	nur gültige	Anteile
sehr unzufrieden	1	94	0.056	0.057	0.057
ziemlich unzufrieden	2	169	0.102	0.102	0.159
etwas unzufrieden	3	283	0.170	0.171	0.330
etwas zufrieden	4	396	0.238	0.239	0.568
ziemlich zufrieden	5	580	0.349	0.350	0.919
sehr zufrieden	6	135	0.081	0.081	1.000
weiß nicht	8	6	0.004	--	
keine Angabe	9	0	0.000	--	
Summe		1663	1.000	1.000	

(BTW 05 Nachwahl nur alte Länder; gültige Fälle: 1657)

Hinweise zur Datenqualität:

- Fallzahl
  Bei Stichproben, die nur auf wenigen Fällen (n < 100 oder gar n < 50) beruhen, ist damit zu rechnen, dass die Häufigkeitsverteilung über unterschiedliche Stichproben deutlich variieren. Bei Prozentuierungen bzw. Anteilen sind daher Nachkommastellen eher nicht zu interpretieren, da selbst bei „guten“ Auswahlverfahren mit Stichprobenschwankungen von ±5% (bzw. ±0.05) zu rechnen ist.
  Im Beispiel ist die Fallzahl der sehr hoch, was auf vermutlich genaue Schätzungen hinweist.

Häufigkeitstabellen: Interpretation

F003: Demokratiezufriedenheit			Anteile		kumulierte
Ausprägung	Code	Häufigkeit	insgesamt	nur gültige	Anteile
sehr unzufrieden	1	94	0.056	0.057	0.057
ziemlich unzufrieden	2	169	0.102	0.102	0.159
etwas unzufrieden	3	283	0.170	0.171	0.330
etwas zufrieden	4	396	0.238	0.239	0.568
ziemlich zufrieden	5	580	0.349	0.350	0.919
sehr zufrieden	6	135	0.081	0.081	1.000
weiß nicht	8	6	0.004	--	
keine Angabe	9	0	0.000	--	
Summe		1663	1.000	1.000	

(BTW 05 Nachwahl nur alte Länder; gültige Fälle: 1657)

Hinweise zur Datenqualität:

- bei Umfragen Frageformulierung und Antwortvorgaben:
  Die Frageformulierung ist nicht explizit sichtbar!
  Bei den Antwortvorgaben stellt sich die Frage, ob die Befragten zwischen „ziemlich“ und „sehr“ unterscheiden können.

Häufigkeitstabellen: Interpretation

F003: Demokratiezufriedenheit Ausprägung	Code	Häufigkeit	Anteile insgesamt	Anteile nur gültige	kumulierte Anteile
sehr unzufrieden	1	94	0.056	0.057	0.057
ziemlich unzufrieden	2	169	0.102	0.102	0.159
etwas unzufrieden	3	283	0.170	0.171	0.330
etwas zufrieden	4	396	0.238	0.239	0.568
ziemlich zufrieden	5	580	0.349	0.350	0.919
sehr zufrieden	6	135	0.081	0.081	1.000
weiß nicht	8	6	0.004	--	
keine Angabe	9	0	0.000	--	
Summe		1663	1.000	1.000	

(BTW 05 Nachwahl nur alte Länder; gültige Fälle: 1657)

Interpretation:
Interpretiert werden in erster Linie die relativen Häufigkeiten und ab ordinalem Messniveau auch die kumulierten Anteile.

> *Der Anteil der mit dem Funktionieren der Demokratie sehr unzufriedenen Personen ist mit knapp 6% recht gering. Insgesamt sind aber doch etwa 1/3 (33%) aller Befragten in den alten Bundesländern zumindest etwas unzufrieden. Auf der anderen Seite sind aber über 40% (100% – 56.8%) ziemlich oder sehr zufrieden.*

Häufigkeitstabellen bei gruppierten Daten

Wenn eine Variable sehr viele Ausprägungen hat, was insbeondere bei stetigen Variablen der Fall ist, werden oft vor der Darstellung der Verteilung in einer Häufigkeitstabelle aus Gründen der Übersichtlichkeit Ausprägungen zu Klassen (Gruppen) zusammengefasst.

Messtheoretisch gesehen ist jede ***Klassenbildung*** eine unzulässige Transformation, da verschiedene Ausprägungen einer Variablen zu einem Wert zusammengefasst werden.
Inhaltlich bedeutet die Zusammenfassung von Ausprägungen einer Variablen zu Klassen stets einen ***Informationsverlust***.

Bei der Zusammenfassung von Ausprägungen zu Klassen sollten folgende Regeln berücksichtigt werden:

1. Die ***Klassengrenzen*** dürfen sich ***nicht überschneiden***, d.h. jede Ausprägung darf nur einer einzigen Klasse zugeordnet werden.
2. Die Klassen sollen ***lückenlos aufeinander folgen***, d.h. innerhalb des Wertebreichs soll jede Zahl einer Klasse zugeordnet werden können (→ ***exakte Klassengrenzen***),
3. Die ***Klassenbreiten*** sollen ***möglichst jeweils gleich*** sein.
   (Ausnahmen: ungleiche Klassenbreite bei der ersten oder letzten Klasse, wenn diese sonst sehr gering besetzt wären).

Bisweilen werden Klassen aber auch bewust so gebildet, dass sie in etwa gleich stark besetzt sind. Als Folge sind die Klassenbreiten dann i.a. unterschiedlich breit.)

Häufigkeitstabellen bei gruppierten Daten

$u_{(k)}$ $o_{(k)}$	$m_{(k)}$	$n_{(k)}$		$p_{(k)}$	$cp_{(k)}$
Ausprägung in Jahren exakte Klassengenzen)	Code = Klassenmitte	Häufigkeit	Prozente	Gültige Prozente	Kumulierte Prozente
17.5 bis <29.5	23.5	507	14.8	14.9	14.9
29.5 bis < 44.5	37.0	943	27.6	27.6	42.5
44.5 bis <59.5	52.0	904	26.4	26.5	69.0
59.5 bis <74.5	67.0	812	23.7	23.8	92.8
74.5 bis <94.5	84.5	246	7.2	7.2	100.0
keine Angabe	--	10	0.3	Missing	
Total		3422	100.0	100.0	

Gültige Fälle: 3412 Fehlende Fälle: 10

(Quelle: Allbus 2006, gewichtet nach Ost-West)

Als Wert (Code) für die Ausprägungen der Klassen Variablen wird die Klassenmitte $m_{(k)}$ einer Klasse berechnet, das ist der Durchschnittswert aus Ober- und Untergrenze einer Klasse:

$$m_{(k)} = \frac{u_{(k)} + o_{(k)}}{2}$$

$$m_{(1)} = (17.5 + 29.5)/2 = 23.5$$

$$\dots$$

$$m_{(5)} = (74.5 + 93.5)/2 = 84.5$$

Bei exakten Klassengrenzen ist in Berechnungen der Wert der Klassenobergrenze mit dem Wert der Klassenuntergrenze der nächsten Klasse gleichgesetzt: $u_{(k)} = o_{(k-1)}$.

Lerneinheit 3: Verteilungsfunktionen und Quantile

Mathematisch gesehen sind die in einer Häufigkeitstabelle aufgeführten Werte Abbildungen oder Funktionen, wobei die Argumente dieser Funktionen die Ausprägungen (Werte) der betrachteten Variable sind und die Funktionswerte die absoluten, relativen oder kumulierten Häufigkeiten.

Im folgenden Beispiel wird die Verteilung der Antworten auf die Frage nach der allgemeinen Wirtschaftslage aus dem Allbus 2006 betrachtet.

Bewertung der allgemeinen Wirtschaftslage

Ausprägung	Code	Häufigkeit	Prozente	Gültige Prozente	Kumulierte Prozente
sehr gut	1	35	1.0	1.0	1.0
gut	2	434	12.7	12.7	13.7
teils/teils	3	1619	47.3	47.5	61.2
schlecht	4	1100	32.1	32.3	93.5
sehr schlecht	5	222	6.5	6.5	100.0
weiß nicht	8	11	0.3	Missing	
keine Angabe	9	1	0.0	Missing	
Total		3422	100.0	100.0	

Gültige Fälle: 3410 Fehlende Fälle: 12

(Daten: ALLBUS 2006, gewichtet nach Ost-West)

Verteilungsfunktion

Bewertung der allgemeinen Wirtschaftslage

Ausprägung	Code	Häufigkeit	Prozente	Gültige Prozente	$\hat{F}(X = x_k)$ Kumulierte Prozente
sehr gut	1	35	1.0	1.0	1.0
gut	2	434	12.7	12.7	13.7
teils/teils	3	1619	47.3	47.5	61.2
schlecht	4	1100	32.1	32.3	93.5
sehr schlecht	5	222	6.5	6.5	100.0
weiß nicht	8	11	0.3	Missing	
keine Angabe	9	1	0.0	Missing	
Total		3422	100.0	100.0	

Gültige Fälle: 3410 Fehlende Fälle: 12

(Daten: ALLBUS 2006, gewichtet nach Ost-West)

Von besonderen Interesse für Statistiker ist die sogenannte ***Verteilungsfunktion*** F(x), die angibt, wie hoch der Anteil der Realisierungen einer Verteilung ist, die kleiner oder gleich dem Argumentwert x sind.

In der Häufigkeitstabelle gibt die letzte Spalte mit den kumulierten Prozentwerten oder Anteilen die Werte der Verteilungsfunktion für die Ausprägungen der tabellierten Variable wieder. Da i.a. davon ausgegangen wird, dass beobachteten Daten auf Stichproben beruhen, die die Verteilungsfunktion in der Population schätzen, spricht man hier von der ***empirischen Verteilungsfunktion*** $\hat{F}(X)$, die zur Unterscheidung der Verteilungsfunktion in der Population durch ein Dach (engl: hat, „^“) auf dem „F“ gekennzeichnet ist.

Berechnung der empirischen Verteilungsfunktion

Die ***empirische Verteilungsfunktion*** lässt sich berechnen, indem für eine beliebige Zahl X=x ausgerechnet wird, wieviele Fälle der Verteilung kleiner oder gleich diesem Wert sind. Dazu müssen alle Fälle zunächst der Größe nach sortiert werden.

Soll für die Beispieldatenmatrix (aus L02) die empirische Verteilungsfunktion des Geburtsjahrs berechnet werden, sind zunächst die Werte der letzte Spalte der Datenmatrix der Größe nach zu ordnen:

Fall-nummer ID	Antwort Frage 1 F1	Antwort Frage 2a F2A	Antwort Frage 2b F2B	Ge-schlecht F3	Geburts-jahr F4
1	3	2	2	1	1943
2	2	missing	1	2	1960
3	4	1	2	2	1957
4	missing	missing	1	1	1939
5	2	2	1	2	missing
6	missing	missing	1	1	1956
7	4	1	2	2	1970
8	1	1	2	1	1920
9	3	3	1	2	1956
10	4	2	2	2	1966

geordn. Geburts-jahr
1920
1939
1943
1956
1956
1957
1960
1966
1970
missing

geordn. G.jahr X $x_{(i)}$	$\sum_{i=1}^{n} \frac{i}{n}$	$\hat{F}(X)$
1920	1/9	0.111
1939	2/9	0.222
1943	3/9	0.333
1956	4/9	--
1956	5/9	0.556
1957	6/9	0.667
1960	7/9	0.778
1966	8/9	0.889
1970	9/9	1.000
missing		

Die Verteilungsfunktion berechnet sich dann nach: $\hat{F}\left(X = x_{(i)}\right) = p\left(X \le x_{(i)}\right) = \sum_{i=1}^{n} \frac{i}{n}$

Haben Realisierungen die gleichen Ausprägungen (im Beispiel die Fälle mit den Geburtsjahren 1956), dann wird in der Verteilungsfunktion nur der letzte Wert berücksichtig!

Berechnung der empirischen Verteilungsfunktion

Bewertung der allgemeinen Wirtschaftslage

Ausprägung	Code	Häufigkeit	Anteile	Gültige Anteile	$\hat{F}(X = x_k)$ Kumulierte Anteile
sehr gut	1	35	0.010	0.010	0.010
gut	2	434	0.127	0.127	0.137
teils/teils	3	1619	0.473	0.475	0.612
schlecht	4	1100	0.321	0.323	0.935
sehr schlecht	5	222	0.065	0.065	1.000
weiß nicht	8	11	0.003	Missing	
keine Angabe	9	1	0.000	Missing	
Total		3422	1.000	1.000	

Gültige Fälle: 3410 Fehlende Fälle: 12

(Daten: ALLBUS 2006, gewichtet nach Ost-West)

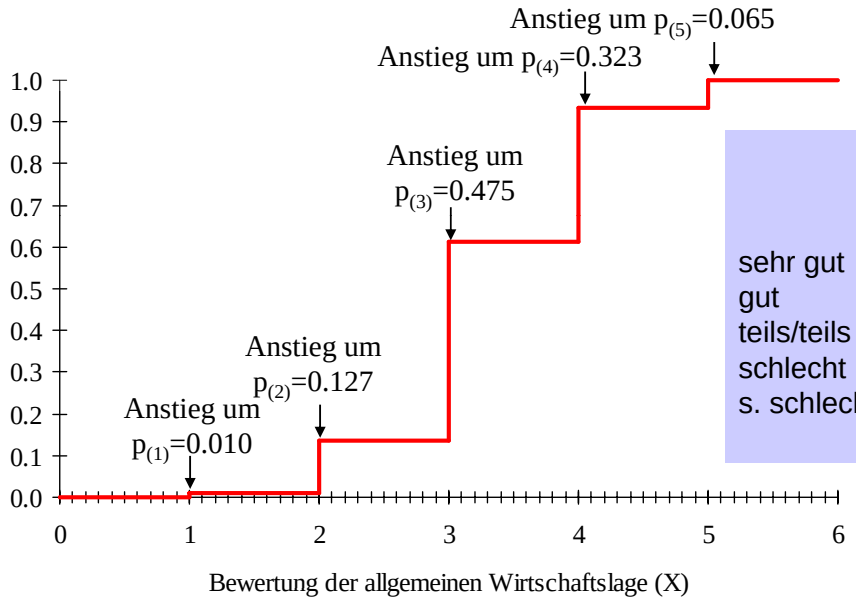
Liegt eine Verteilung als (ungruppierte) Häufigkeitstabelle vor, dann enthält die letzte Spalte mit den kumierten Anteilen die Funktionswerte der Verteilungsfunktion für die in der Tabelle aufgeführten Ausprägungen.

Formal berechnet sich hier die Funktion nach: $\hat{F}\left(X = x_k\right) = p\left(X \le x_k\right) = \sum_{j=1}^{k} \frac{n_{(j)}}{n} = \sum_{j=1}^{k} p_{(j)} = cp_{(j)}$

Verteilungsfunktionen lassen sich grafisch darstellen, wobei entlang der X-Achse die Ausprägungen der Variablen und entlang der Y-Achse die Werte der Verteilungsfunktion aufgetragen werden.

Grafische Darstellung der Verteilungsfunktion

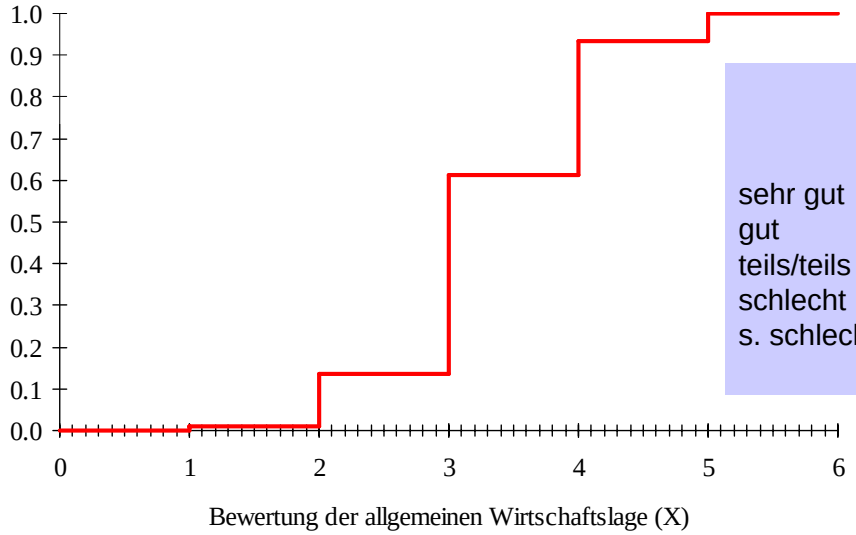
Es ergibt sich dann eine treppenförmige Gestalt *(eng: step function)*:



	X	Gültige Anteile	Kumulierte Anteile
	0	0.000	0.000
sehr gut	1	0.010	0.010
gut	2	0.127	0.137
teils/teils	3	0.475	0.612
schlecht	4	0.323	0.935
s. schlecht	5	0.065	1.000
	6	0.000	1.000

Bei jeder Ausprägung der Variablen steigt die Funktion um die relative Häufigkeit dieser Ausprägung an.

Quantile



	X	Gültige Anteile	Kumulierte Anteile
	0	0.000	0.000
sehr gut	1	0.010	0.010
gut	2	0.127	0.137
teils/teils	3	0.475	0.612
schlecht	4	0.323	0.935
s. schlecht	5	0.065	1.000
	6	0.000	1.000

Betrachtet man anstelle der Verteilungsfunktion ***umgekehr***t für welchen Wert der Variable gilt, dass ein vorgegebener Anteil aller Realisierungen kleiner oder gleich diesem Wert sind, ergeben sich die Quantilwerte.
Mathematisch berechnen sich die ***Quantilwerte*** daher aus der ***Inversen der Verteilungsfunktion***, d.h. die Rolle von X (Argumentwerte) und Y (Funktionswerte) werden vertauscht:

$$Q(p) = \hat{F}^{-1}(X)$$

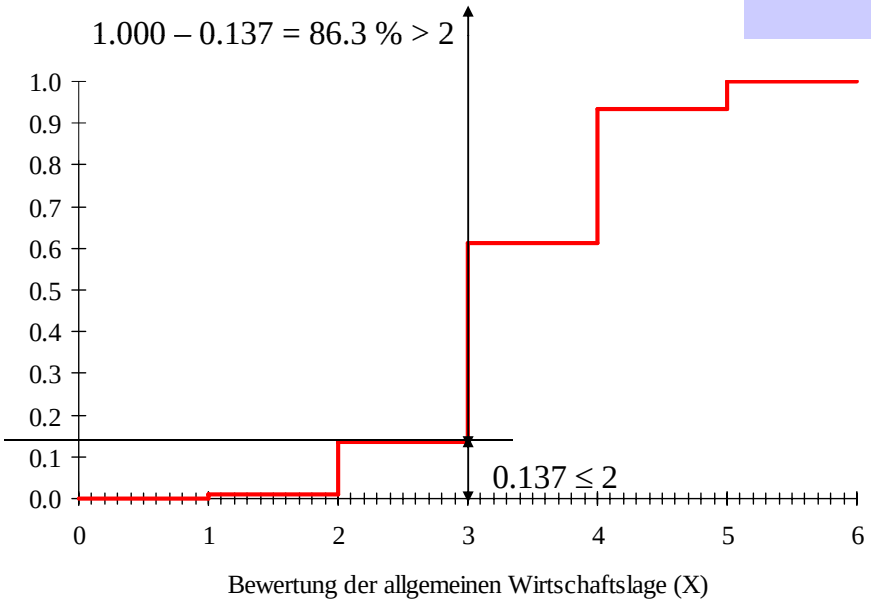
Quantile

Quantile teilen somit eine Verteilung in zwei Teilmengen auf:

$0.137 \leq 2$

$1.000 - 0.137 = 0.863\% > 2$

	X	Gültige Anteile	Kumulierte Anteile
	0	0.000	0.000
sehr gut	1	0.010	0.010
gut	2	0.127	0.137
teils/teils	3	0.475	0.612
schlecht	4	0.323	0.935
s. schlecht	5	0.065	1.000
	6	0.000	1.000



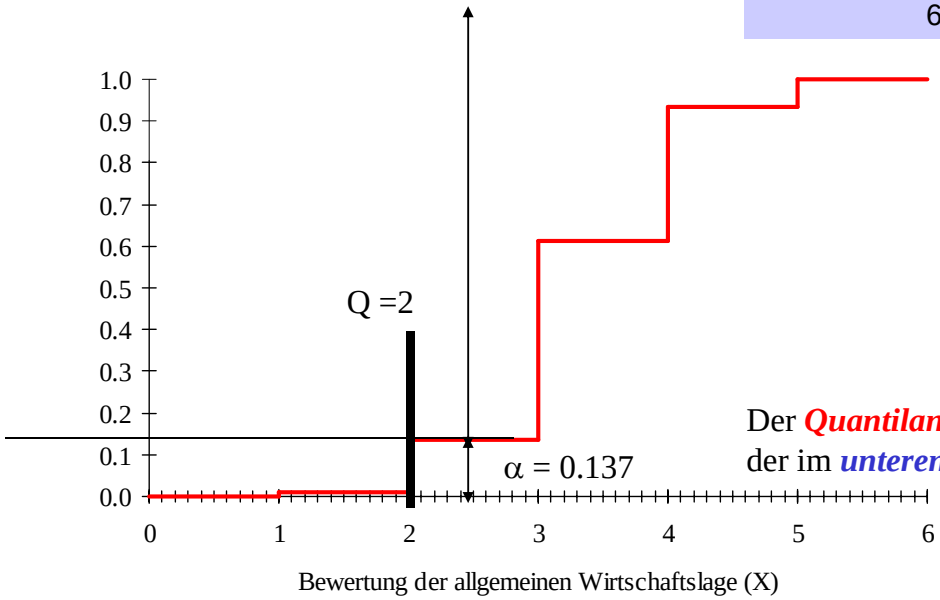
Quantile

Der ***Quantilwert Q*** gibt die Trennstelle an, an der die Teilung erfolgt

$\alpha = 0.137$

$Q = 2$

	X	Gültige Anteile	Kumulierte Anteile
	0	0.000	0.000
sehr gut	1	0.010	0.010
gut	2	0.127	0.137
teils/teils	3	0.475	0.612
schlecht	4	0.323	0.935
s. schlecht	5	0.065	1.000
	6	0.000	1.000

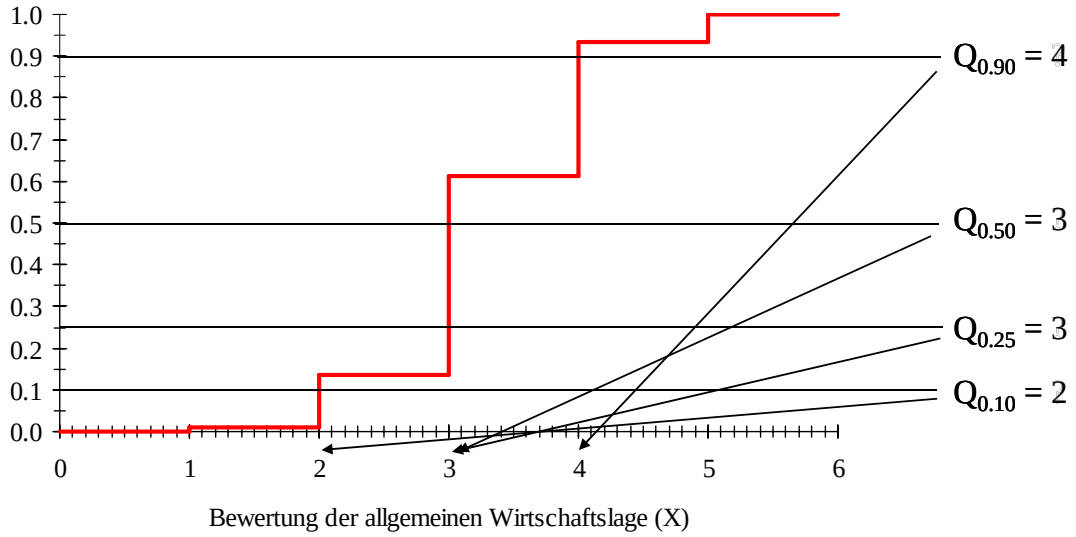


Der ***Quantilanteil α*** gibt den Anteil an, der im ***unteren Teilbereich*** liegt.

Quantile

Da die Verteilungsfunktion treppenförmig und nicht stetig ist, ist die Umkehrfunktion nicht ganz eindeutig.
Der ***empirische Quantilwert*** Q_α ist definiert als der ***kleinste Wert*** für den gilt, dass ***mindestens*** ein ***Anteil*** α von allen Realisierungen ***kleiner oder gleich*** diesem Wert ist.

	X	Gültige Anteile	Kumulierte Anteile
	0	0.000	0.000
sehr gut	1	0.010	0.010
gut	2	0.127	0.137
teils/teils	3	0.475	0.612
schlecht	4	0.323	0.935
s. schlecht	5	0.065	1.000
	6	0.000	1.000



Quantile: Berechnung aus Häufigkeitstabellen ungruppierter Daten

Wenn eine Häufigkeitstabelle ungruppierter Daten vorliegt, können die Quantilwerte direkt aus der Häufigkeitstabelle abgelesen werden:
Der Quantilwert ist die Ausprägung, bei der in der Spalte mit den kumulierten Anteilen bzw. kumulierten Prozentwerten erstmals der Quantilanteil erreicht oder überschritten wird:

	X	Gültige Anteile	Kumulierte Anteile
sehr gut	1	0.010	0.010
gut	2	0.127	0.137
teils/teils	3	0.475	0.612
schlecht	4	0.323	0.935
s. schlecht	5	0.065	1.000

$Q_{10\%} = Q_{0.10} = ?$

$0.010 < 0.10 \Rightarrow Q_{10\%} > 1$

$0.137 > 0.10 \Rightarrow Q_{10\%} \leq 2$

„2“ ist die kleinste Ausprägung, für die gilt, mindestens 10% aller Fälle sind $\leq 2 \Rightarrow Q_{0.1} = 2$.

In der Spalte mit den kumulierten Anteilen oder Prozenten sind nur die Quantilanteile der Ausprägungen einer Variable aufgeführt, aus denen dieWerte der übrigen Quantilanteile folgen:

	X	Gültige Anteile	Kumulierte Anteile
sehr gut	1	0.010	0.010
gut	2	0.127	0.137
teils/teils	3	0.475	0.612
schlecht	4	0.323	0.935
s. schlecht	5	0.065	1.000

$Q_{\alpha=0.0\%}$ bis $Q_{\alpha=1.0\%}$ = 1 (sehr gut)
$Q_{\alpha>1.0\%}$ bis $Q_{\alpha=13.7\%}$ = 2 (gut)
$Q_{\alpha>13.7\%}$ bis $Q_{\alpha=61.2\%}$ = 3 (teils/teils)
$Q_{\alpha>61.2\%}$ bis $Q_{\alpha=95.5\%}$ = 4 (schlecht)
$Q_{\alpha>95.5\%}$ bis $Q_{\alpha=100\%}$ = 5 (sehr schlecht)

Quantile: Berechnung aus geordneten Messwerten

Wie die Verteilungsfunktion können auch Quantilwerte direkt aus den Messwerten berechnet werden, wenn die Messwertreihe vorher der Größe nach sortiert sind:

Fall-nummer ID	Antwort Frage 1 F1	Antwort Frage 2a F2A	Antwort Frage 2b F2B	Ge-schlecht F3	Geburts-jahr F4
1	3	2	2	1	1943
2	2	missing	1	2	1960
3	4	1	2	2	1957
4	missing	missing	1	1	1939
5	2	2	1	2	missing
6	missing	missing	1	1	1956
7	4	1	2	2	1970
8	1	1	2	1	1920
9	3	3	1	2	1956
10	4	2	2	2	1966

F4 geord-net	Posi-tion (j)
1920	1
1939	2
1943	3
1956	4
1956	**5**
1957	6
1960	7
1966	8
1970	9
missing	missing

Beispiel: $Q_{50\%}$ = ?
n = 9 gültige Werte
Schritt 1: $i = 9 \times 0.5 = 4.5$
Schritt 2: 4.5 ist keine ganze Zahl, daher aufrunden: $4.5 \Rightarrow j = 5$
Schritt 3: $Q_{50\%}$ ist dann $x_{(5)}$
$x_{(5)} = 1956$
Das 50%-Quantil der Geburtsjahre ist das Jahr 1956

Die Berechnung erfolgt in drei Schritten:
Schritt 1: Multiplikation des Quantilanteils mit der Fallzahl: $i = n \cdot \alpha$
Schritt 2: Falls i keine ganze Zahl ist, sondern Nachkommastellen hat, Aufrunden zur nächsten ganzen Zahl j, anderenfalls: $j = i$.
Schritt 3: Der Quantilwert Q_α ist der Wert der Variablen auf dem j-ten Position: $x_{(j)}$.

Quantile: Berechnung aus geordneten Messwerten

Beispiel, bei dem nicht aufgerundet werden muss.
Schritt 1: Multiplikation des Quantilanteils mit der Fallzahl: $i = n \cdot \alpha$
Schritt 2: Falls i keine ganze Zahl ist, sondern Nachkommastellen hat, Aufrunden zur nächsten ganzen Zahl j, anderenfalls: $j = i$.
Schritt 3: Der Quantilwert Q_α ist der Wert der Variablen auf der j-ten Position: $x_{(j)}$.

X
4
1
3
5
5
2
6
6
1
6

X geord-net	Posi-tion (j)
1	1
1	2
2	3
3	4
4	5
5	6
5	7
6	8
6	9
6	10

Beispiel: $Q_{60\%}$ = ?
n = 10 gültige Werte
Schritt 1: $i = 10 \times 0.6 = 6$
Schritt 2: 6 ist ganze Zahl, daher nicht aufrunden:
$\Rightarrow i = j = 6$
Schritt 3: $Q_{60\%}$ ist dann $x_{(6)}$: $x_{(6)} = 5$
Das 60%-Quantil von X ist der Wert 5
(„5“ ist der kleinste Wert für den gilt, dass mindestens 60% aller Realisierungen kleiner oder gleich diesem Wert sind).

Quantilberechnung bei gruppierten Daten metrischer Variablen

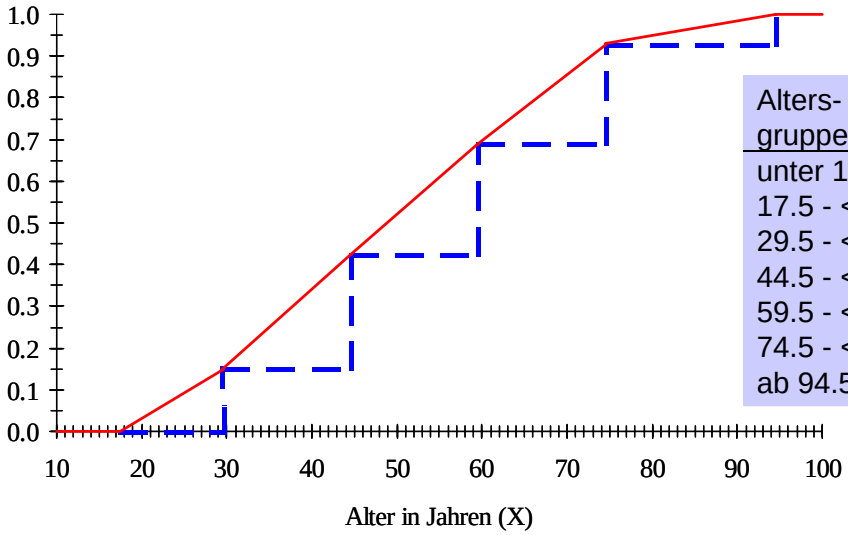
Bei gruppierten Daten metrischer Variablen wird eine andere Vorgehensweise zur Berechnung von Quantilen eingesetzt, die versucht, den Informationsverlust durch die Gruppierung (Klassenbildung) zu kompensieren.

$u_{(k)}$ $o_{(k)}$		$m_{(k)}$	$n_{(k)}$		$p_{(k)}$	$cp_{(k)}$
Ausprägung in Jahren (exakte Klassengenzen)		Code = Klassenmitte	Häufigkeit	Prozente	Gültige Prozente	Kumulierte Prozente
17.5 bis <29.5		23.5	507	14.8	14.9	14.9
29.5 bis < 44.5		37.0	943	27.6	27.6	42.5
44.5 bis <59.5		52.0	904	26.4	26.5	69.0
59.5 bis <74.5		67.0	812	23.7	23.8	92.8
74.5 bis <94.5		84.5	246	7.2	7.2	100.0
keine Angabe		--	10	0.3	Missing	
Total			3422	100.0	100.0	
Gültige Fälle: 3412	Fehlende Fälle: 10					
(Quelle: Allbus 2006, gewichtet nach Ost-West)						

Als Beispiel wird die Altersverteilung der Befragten aus dem Allbus 2006 betrachtet, die in der Häufigkeitstabelle in 5 Altersgruppen zusammengefasst ist.
Von der empirischen Verteilungsfunktion sind nur die Funktionswerte an den Klassengrenzen bekannt.

So ist keine befragte Person jünger als 17.5 Jahre, 14.9% sind jünger als 29.5 Jahre, 42.5% sind jünger als 44.5 Jahre, 69% jünger als 59.5 Jahr, 92.8 jünger als 74.5 und 100% sind jünger als 94.5 Jahre.

Summenkurve



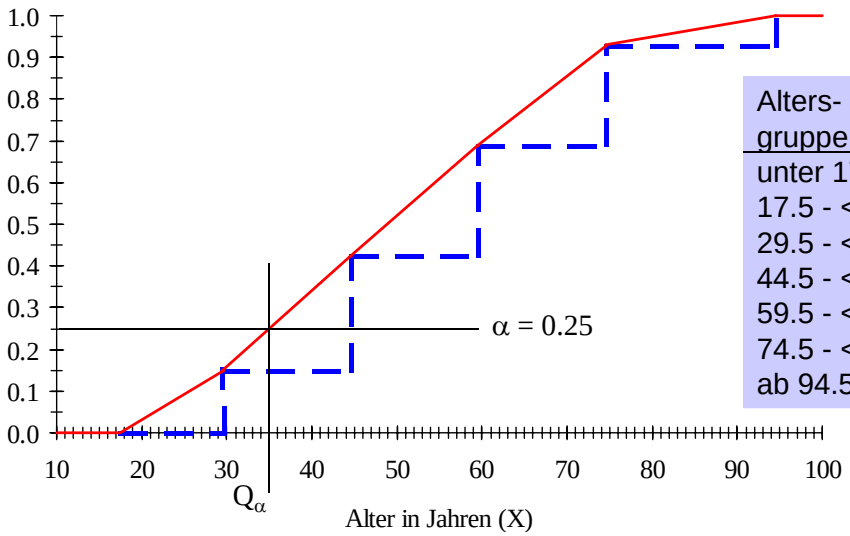
Alters-gruppe	Klasse k	Gültige Anteile	Kumulierte Anteile
unter 17.5	0	0.000	0.000
17.5 - <29.5	1	0.149	0.149
29.5 - <44.5	2	0.276	0.425
44.5 - <59.5	3	0.265	0.690
59.5 - <74.5	4	0.328	0.928
74.5 - <94.5	5	0.072	1.000
ab 94.5	6	0.000	1.000

Aus den gruppierten Daten ist nicht ersichtlich, wie die Verteilung innerhalb der Altersgruppen verläuft. Für die Berechnung der Quantile bei solchen gruppierten Daten wird **unterstellt**, dass sich alle Fälle innerhalb einer Klasse ***gleichmäßig über die gesamte Klassenbreite verteilen***.

Es wird also z.B. unterstellt, dass sich die 507 Fälle (14.9%) in der untersten Altersgruppe gleichmäßig auf den Bereich von 17.5 bis 29.5 Jahre verteilen.

Unter dieser Annahme lässt sich die Verteilungsfunktion grafisch durch Linienabschnitte annähern, die jeweils die kumulierten Anteile an den Intervallgrenzen verbinden. Die resultierende Kurve wird als ***Summenkurve*** bezeichnet.

Quantilberechnung bei gruppierten Daten metrischer Variablen

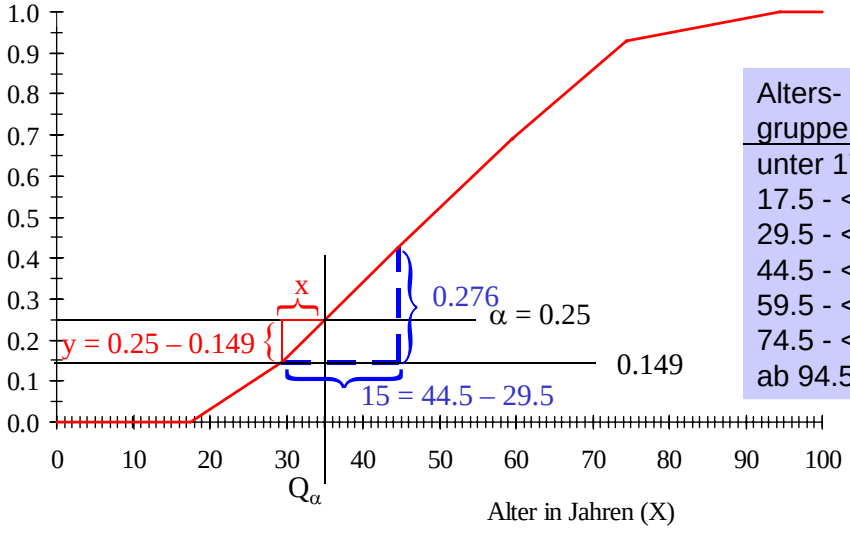


Alters-gruppe	Klasse k	Gültige Anteile	Kumulierte Anteile
unter 17.5	0	0.000	0.000
17.5 - <29.5	1	0.149	0.149
29.5 - <44.5	2	0.276	0.425
44.5 - <59.5	3	0.265	0.690
59.5 - <74.5	4	0.328	0.928
74.5 - <94.5	5	0.072	1.000
ab 94.5	6	0.000	1.000

Der Quantilwert Q_α wird dann als der Wert an der X-Achse bestimmt, an der eine horizontale Gerade auf der Höhe α der senkrechten Y-Achse die Summenkurve schneidet.

So ergibt sich z.B. für das 25%-Quantil der Altersverteilung im Allbus 2006 die Zahl 34.99. Dieser Wert ist der Schnittpunkt einer senkrechten Gerade mit der X-Achse an der Stelle, an der die horizontale Linie in der Höhe 0.25 der Y-Achse die Summenkurve schneidet.

Quantilberechnung bei gruppierten Daten metrischer Variablen



Alters-gruppe	Klasse k	Gültige Anteile	Kumulierte Anteile
unter 17.5	0	0.000	0.000
17.5 - <29.5	1	0.149	0.149
29.5 - <44.5	2	0.276	0.425
44.5 - <59.5	3	0.265	0.690
59.5 - <74.5	4	0.328	0.928
74.5 - <94.5	5	0.072	1.000
ab 94.5	6	0.000	1.000

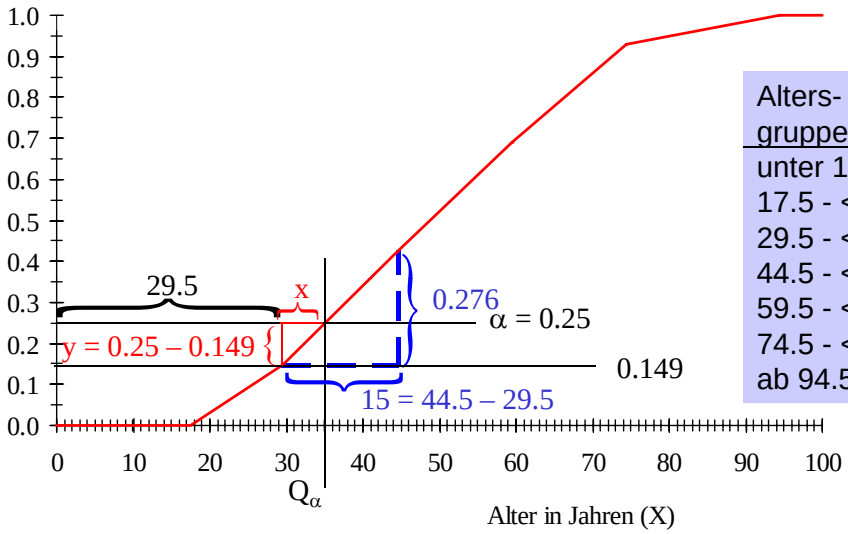
Rechnerisch lässt sich der Quantilwert durch ***lineare Interpolation*** bestimmen:

Die Summenkurve steigt in jedem durch die Klassengrenzen festgelegten Intervall gleichmäßig an, wobei sich die Steigung als Quotient aus der relativen Häufigkeit in der Klasse geteilt durch die Intervallbreite ergibt, in der zweiten Klasse also 0.276/(44.5–29.5) = 0.276 / 15 beträgt.

Der Quotient y/x mit y = 0.25 – 0.149 ist gleich dieser Steigung :

$$\frac{y}{x} = \frac{0.25 - 0.149}{x} = \frac{0.276}{15} = \frac{0.276}{44.5 - 29.5}$$

Quantilberechnung bei gruppierten Daten metrischer Variablen



Alters-gruppe	Klasse k	Gültige Anteile	Kumulierte Anteile
unter 17.5	0	0.000	0.000
17.5 - <29.5	1	0.149	0.149
29.5 - <44.5	2	0.276	0.425
44.5 - <59.5	3	0.265	0.690
59.5 - <74.5	4	0.328	0.928
74.5 - <94.5	5	0.072	1.000
ab 94.5	6	0.000	1.000

Auflösen nach x: $\frac{y}{x} = \frac{0.25 - 0.149}{x} = \frac{0.276}{44.5 - 29.5} \Rightarrow x = \frac{0.25 - 0.149}{0.276 / (44.5 - 29.5)} = 5.489$

und addieren des Wertes der Untergrenze des Intervalls (hier: 29.5) ergibt den gesuchten Quantilwert: $Q_{0.25} = 29.5 + 5.489 = 34.99$.

Quantilberechnung bei gruppierten Daten metrischer Variablen

Alters-gruppe	Klasse k	Gültige Anteile	Kumulierte Anteile
unter 17.5	0	0.000	0.000
17.5 - <29.5	1	0.149	0.149
29.5 - <44.5	2	0.276	0.425
44.5 - <59.5	3	0.265	0.690
59.5 - <74.5	4	0.328	0.928
74.5 - <94.5	5	0.072	1.000
ab 94.5	6	0.000	1.000

Werden anstelle der Zahlen des Beispiels die Symbole der Häufigkeitstabelle eingesetzt, ergibt sich die allgemeine Formel zur Berechnung von Quantilen innerhalb der k-ten Klasse bei Vorliegen einer Häufigkeitstabelle gruppierter metrischer Daten durch Interpolation:

$$Q_\alpha = o_{(k-1)} + \frac{\alpha - cp_{(k-1)}}{p_{(k)} / \left(o_{(k)} - o_{(k-1)}\right)} = o_{(k-1)} + \frac{\alpha - cp_{(k-1)}}{p_{(k)}} \cdot \left(o_{(k)} - o_{(k-1)}\right)$$

Nach dieser Formel beträgt das 25%-Quantil, dass in der 2. Klasse liegt:

$$Q_{25\%} = 29.5 + \frac{0.25 - 0.149}{0.276} \cdot (44.5 - 29.5) = 34.99$$

Der tatsächlich zutreffende Wert des 25%-Quantils beträgt für die ungruppierte Altersverteilung 36 Jahre.

Die Abweichung zwischen dem über die Summenkurve interpolierten Quantilwert und dem tatsächlichen Wert ist um so größer, je stärker die tatsächliche Verteilung innerhalb der Klasse, in der das Quantil liegt, von einer Gleichverteilung abweicht.

Bedeutung von Quantilen

Quantilanteile und Quantilwerte sind deswegen von großem Interesse, weil sie Informationen über eine Verteilung geben:

- So besagt z.B. das 50%-Quantil, bei welchem Wert in etwa die „Mitte“ einer Verteilung liegt,
- Die Differenzen des 5%- und des 95%-Quantils geben entsprechend an, in welchen Grenzen die mittleren 90% aller Fälle liegen.
- Die Gesamtheit aller Quantile enthält alle Informationen über eine Verteilung.

Voraussetzung für die Berechnung von Quantilen ist allerdings ***mindestens ordinales***, besser metrisches ***Skalenniveau***.
Bei ordinalen Skalenniveau sind Quantilwerte Ausprägungen von Rangplätzen (Kategorien).

Besondere Namen:

- Das 25%-, das 50-% und das 75%-Quantil werden auch als ***Quartile*** bezeichnet, weil sie die Verteilung in vier gleich stark besetzte Klassen aufteilen;
- entsprechend werden das 10%-, 20%-, 30%-, ..., 90%-Quantil als ***Dezentile*** bezeichnet, weil sie die Verteilung in 10 gleich stark besetzte Klassen aufteilen;
- das 1%-, 2%-, ..., 98%-, 99%-Quantil werden analog als ***Perzentile*** bezeichnet.

Anwendungsbeispiel:

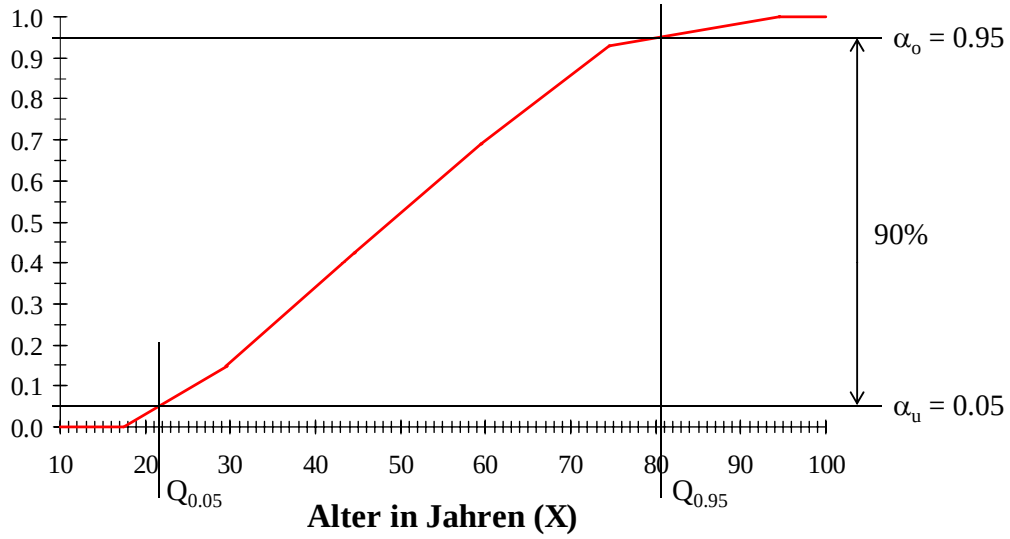
Fragestellung: In welchen Bereich um das 50%-Quantil liegen 90% aller Fälle der Altersverteilung im Allbus 2006?

Ausprägung in Jahren exakte Klassengenzen)	Code = Klassenmitte	Häufigkeit	Prozente	Gültige Prozente	Kumulierte Prozente
17.5 bis <29.5	23.5	507	14.8	14.9	14.9
29.5 bis < 44.5	37.0	943	27.6	27.6	42.5
44.5 bis <59.5	52.0	904	26.4	26.5	69.0
59.5 bis <74.5	67.0	812	23.7	23.8	92.8
74.5 bis <94.5	84.5	246	7.2	7.2	100.0
keine Angabe	--	10	0.3	Missing	
Total		3422	100.0	100.0	

Gültige Fälle: 3412 Fehlende Fälle: 10
(Quelle: Allbus 2006, gewichtet nach Ost-West)

Das 50%-Quantil teilt die Verteilung in eine obere und eine untere Hälfte.
Wenn 90% um das 50%-Quantil verteilt sind, liegen jeweils 45% unterhalb und oberhalb dieses Werts.
Der gesuchte Bereich wird daher durch das 5%-Quantil (5% = 50% – 45%) und durch das 95%-Quantil (95% = 50% + 45%) begrenzt.
Bei gruppierten Daten können die Intervallgrenzen grafisch über die Summenkurve oder rechnerisch durch Interpolation bestimmt werden.

Anwendung von Quantilen bei gruppierten Daten



Aus der Summenkurve geht hervor, dass 90% aller Befragten der Stichprobe des Allbus 2006 zwischen etwa 22 und 81 Jahren alt sind.

Quantilberechnung bei gruppierten Daten: Interpolation innerhalb der Quantilklasse

	$u_{(k)}$ $o_{(k)}$ Ausprägung in Jahren exakte Klassengenzen)	$m_{(k)}$ Code = Klassenmitte	$n_{(k)}$ Häufigkeit	Prozente	$p_{(k)}$ Gültige Prozente	$cp_{(k)}$ Kumulierte Prozente
k=1	17.5 bis <29.5	23.5	507	14.8	14.9	14.9
k=2	29.5 bis < 44.5	37.0	943	27.6	27.6	42.5
k=3	44.5 bis <59.5	52.0	904	26.4	26.5	69.0
k=4	59.5 bis <74.5	67.0	812	23.7	23.8	92.8
k=5	74.5 bis <94.5	84.5	246	7.2	7.2	100.0
	keine Angabe	--	10	0.3	Missing	
	Total		3422	100.0	100.0	

Gültige Fälle: 3412 Fehlende Fälle: 10

(Quelle: Allbus 2006, gewichtet nach Ost-West)

$$Q_\alpha = o_{(k-1)} + \frac{\alpha - cp_{(k-1)}}{p_{(k)}} \cdot \left(o_{(k)} - o_{(k-1)}\right)$$

Da $cp_{(1)}$ = 14.9% > 5% liegt das 5%-Quantil in der ersten Klasse: k = 1.

$$Q_{0.05} = o_{(1-1)} + \frac{0.05 - cp_{(1-1)}}{p_{(1)}} \cdot \left(o_{(1)} - o_{(1-1)}\right) = 17.5 + \frac{0.05 - 0}{.149} \cdot \left(29.5 - 17.5\right) = 21.53$$

Die Untergrenze des gesuchten Intervalls ist daher 21.53.

Quantilberechnung bei gruppierten Daten: Interpolation innerhalb der Quantilklasse

	$u_{(k)}$ $o_{(k)}$	$m_{(k)}$	$n_{(k)}$		$p_{(k)}$	$cp_{(k)}$
	Ausprägung in Jahren exakte Klassengenzen)	Code = Klassenmitte	Häufigkeit	Prozente	Gültige Prozente	Kumulierte Prozente
k=1	17.5 bis <29.5	23.5	507	14.8	14.9	14.9
k=2	29.5 bis < 44.5	37.0	943	27.6	27.6	42.5
k=3	44.5 bis <59.5	52.0	904	26.4	26.5	69.0
k=4	59.5 bis <74.5	67.0	812	23.7	23.8	92.8
k=5	74.5 bis <94.5	84.5	246	7.2	7.2	100.0
	keine Angabe	--	10	0.3	Missing	
	Total		3422	100.0	100.0	

Gültige Fälle: 3412 Fehlende Fälle: 10

(Quelle: Allbus 2006, gewichtet nach Ost-West)

$$Q_\alpha = o_{(k-1)} + \frac{\alpha - cp_{(k-1)}}{p_{(k)}} \cdot \left(o_{(k)} - o_{(k-1)}\right)$$

Da $cp_{(4)} = 92.8\% < 95\%$ liegt das 95%-Quantil in der letzten Klasse: k = 5.

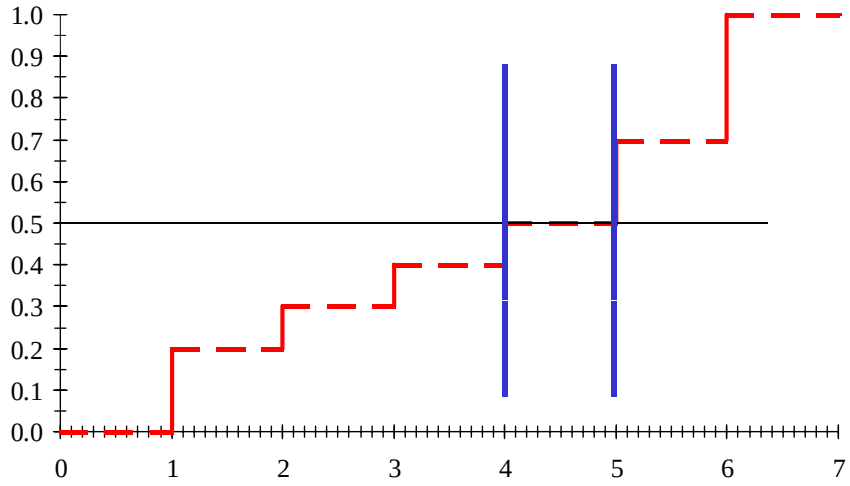
$$Q_{0.95} = o_{(5-1)} + \frac{0.95 - cp_{(5-1)}}{p_{(5)}} \cdot \left(o_{(5)} - o_{(5-1)}\right) = 74.5 + \frac{0.95 - 0.928}{.072} \cdot \left(94.5 - 74.5\right) = 80.61$$

Die Obergrenze des gesuchten Intervalls ist daher 80.61.
90% der Befragten sind daher zwischen 21.53 und 80.61 Jahre alt.

Uneindeutigkeit bei der Berechnung von Quantilen

Die vorgestellte Berechnungsweise bei ungruppierten Daten erfolgt nach der Methode der ***empirischen Quantile***. Andere Methoden können zu abweichenden Quantilwerten führen.
Ursache ist die erwähnte Unstetigkeit der empirischen Verteilungsfunktion.

X	Position (i)
1	1
1	2
2	3
3	4
4	5
5	6
5	7
6	8
6	9
6	10



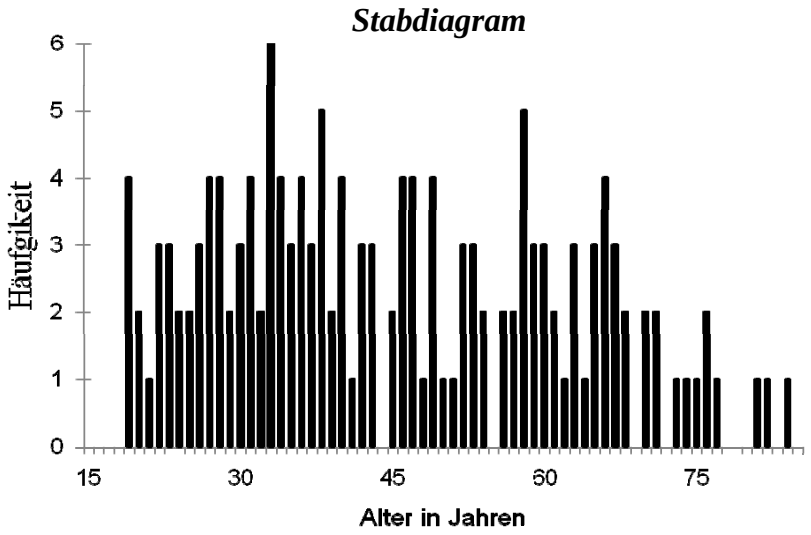
So ist bei den links wiedergegebenen n=10 Fällen das 50%-Quantil $Q_{0.50} = 4$.
Da es keine Realisierung zwischen 4 und 5 gibt, verläuft die empirische Verteilunsgfunktion in diesem Bereich waagerecht. Daher gilt für alle Zahlen zwischen 4 und <5, dass 50% der Realisierungen kleiner oder gleich dieser Zahl sind.
Anstelle des kleinsten Wertes wird daher nach anderen Berechnungsmethoden der Mittelwert zwischen 4 und 5 berechnet oder wie bei gruppierten Daten linear interpoliert.

Lerneinheit 4: Grafische Darstellung univariater Verteilungen

Darstellungen metrischer Verteilungen

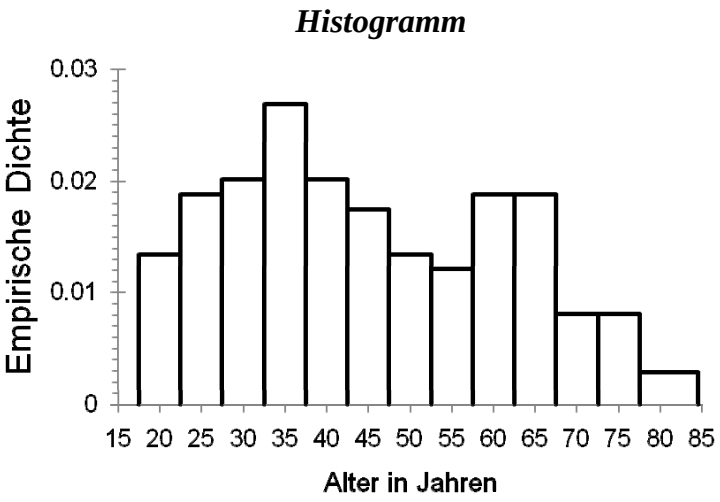
Grafische Darstellungen können einen Eindruck von der Form einer Verteilung vermitteln.

Das Beispiel zeigt die Altersverteilung einer zufälligen Auswahl von 149 Befragten aus der Stichprobe des Allbus 1996.



In ***Stabdiagrammen*** werden die absoluten oder relativen Häufigkeiten der Ausprägungen als senkrechte Linien symbolisiert. Dies ergibt einen schnellen Überblick über die Form der Verteilung einer metrischen Variable.

Darstellungen metrischer Variablen



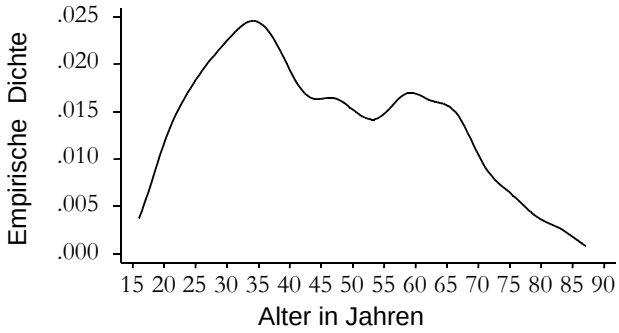
Die Balkenhöhe ist dann die sogenannte ***empirischen Dichte*** der Verteilungsfunktion:

empirische Dichte: $\hat{f}_{(k)} = \frac{p_{(k)}}{(o_{(k)} - u_{(k)})}$

In ***Histogrammen*** wird die Häufigkeitsverteilung durch einander berührende Balken dargestellt. Histogramme sind besonders für die Darstellung der Verteilung bei ***gruppierten Daten*** und ***stetigen Variablen*** sinnvoll, da sie das ***Prinzip der Flächentreue*** berücksichtigen:
Die **Fläche** eines Balkens entspricht der **relativen Häufigkeit** $p_{(k)}$ in dem durch die Balkenbreite definierten Intervall mit den Intervallgrenzen $u_{(k)}$ und $o_{(k)}$.
Da die Fläche eines Rechtecks gleich der Breite mal der Höhe ist, ergibt sich die Balkenhöhe als Quotient aus der relativen Häufigkeit $p_{(k)}$ in dem jeweiligen Intervall geteilt durch die Intervallbreite $(o_{(k)} - u_{(k)})$.

Darstellungen metrischer Variablen

Kern-Dichte-Schätzer



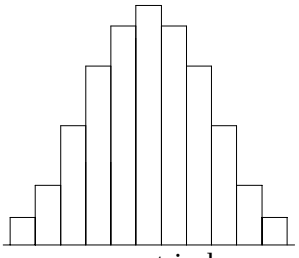
In Abhängigkeit von der verwendeten Formel und der Länge des berücksichtigten Abstands um den jeweiligen Wert, für den die emprische Dichte geschätzt wird, sind die resultierenden Kurvenverläufe glätter oder zerklüfteter.

Die Form eines Histogramms hängt allerdings nicht nur von der Verteilung, sondern auch von den Intervallbreiten und der gewählten Untergrenze für das erste (ganz links angeordnete) Intervall ab.

Um dieses Problem zu umgehen, sind insbesondere für stetigen Variablen sogenannte ***Kern-Dichte-Schätzer*** entwickelt worden.
Diese berechnen die empirische Dichte einer Verteilung nicht nur für ein Intervall, sondern für jede Ausprägung der Variable, wobei jeweils alle Realisierungen in einem vorgegebenen Abstand berücksichtigt werden und der Einfluss eines Wertes auf die berechnete Dichte mit steigendem Abstand sinkt.
Werden die Dichten der Punkte verbunden, ergibt sich eine Kurve, die die Form einer (stetigen) Verteilung besser wiedergibt, als die Balken eines Histogramms.

Verteilungsformen

Mit Hilfe von Kern-Dichte-Schätzern bzw. Histogrammen lassen sich Verteilungen nach kennzeichnenden Charakteristika, wie Schiefe, U-Förmigkeit etc. beschreiben.

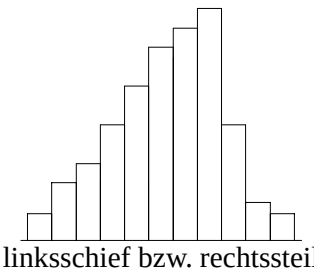


symmetrisch

Eine Verteilung ist ***symmetrisch***, wenn der Abstand eines beliebigen Quantils mit dem Quantilanteil α vom 50%-Quantil der Verteilung gleich dem Abstand des Quantils mit dem Quantilanteil $1-\alpha$ vom 50%-Quantil ist:

$$|Q_{\alpha}-Q_{0.5}| = |Q_{1-\alpha}-Q_{0.5}|$$

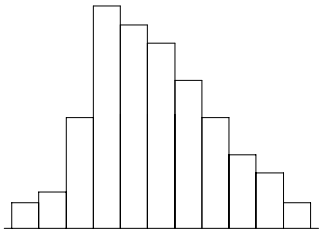
z.B. $|Q_{0.1}-Q_{0.5}| = |Q_{0.9}-Q_{0.5}|$



linksschief bzw. rechtssteil

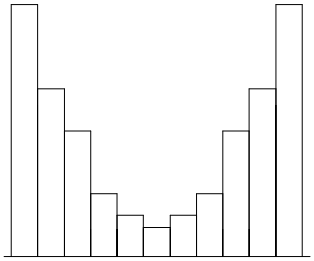
Das Gegenteil von symmetrischen Verteilungen sind ***schiefe*** Verteilungen.
Eine Verteilung ist ***linksschief*** oder ***rechtssteil***, wenn sie von links langsamer ansteigt als sie nach rechts abfällt.

Verteilungsformen



rechtsschief bzw. linkssteil

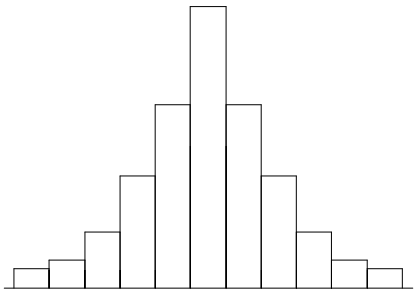
Dagegen ist eine Verteilung ***rechtsschief*** oder ***linkssteil***, wenn sie nach rechts langsamer abfällt als sie von links ansteigt.



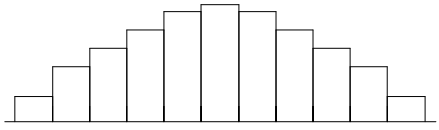
uförmig bimodal, symmetrisch

Hat eine Verteilung nur einen Gipfel ist sie eingipflig oder ***unimodal***. Hat sie zwei Gipfel ist sie dagegen ***bimodel***, bei drei Gipfeln ***trimodal***.
Mehrgipflige (***multimodale***) Verteilungen können ein Hinweis darauf sein, dass sich die betrachtete Population aus unterscheidbaren Teilpopulationen zusammensetzt.

Verteilungsformen



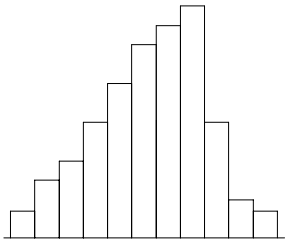
unimodal, steil ansteigend, symmetrisch



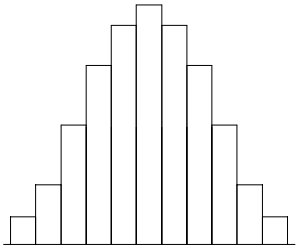
unimodal, flach ansteigend symmetrisch

Gelegentlich wird auch die ***Steilheit*** oder ***Wölbung*** *(engl: **curtosis** oder **excess**)* einer Verteilung betrachtet.
Eine Verteilung verläuft steiler als eine zweite Verteilung mit gleichem Wertebereich, wenn sie an den Rändern flacher und in der Mitte stärker ansteigt.

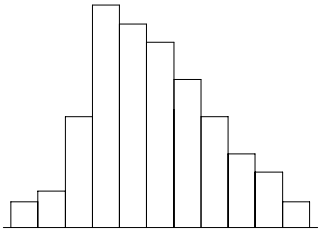
Verteilungsformen: Überblick



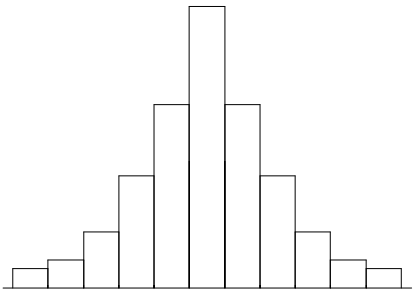
unimodal,
linksschief bzw. rechtssteil



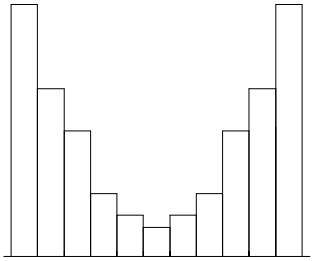
unimodal,
symmetrisch



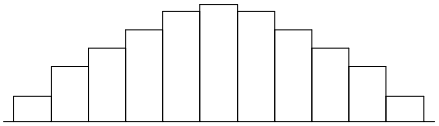
unimodal,
rechtsschief bzw. linkssteil



unimodal, steil ansteigend,
symmetrisch



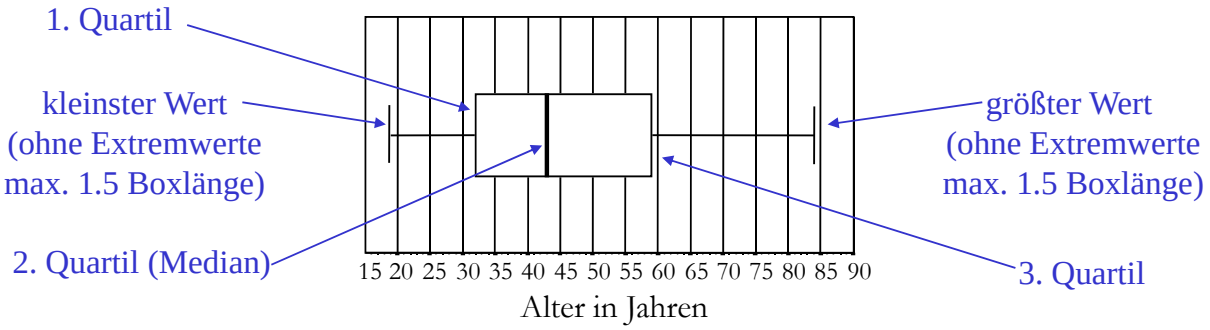
uförmig bimodal,
symmetrisch



unimodal, flach ansteigend
symmetrisch

Darstellungen metrischer Variablen

Box-Plot

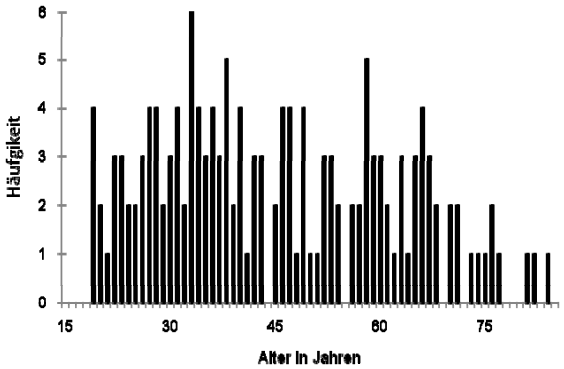


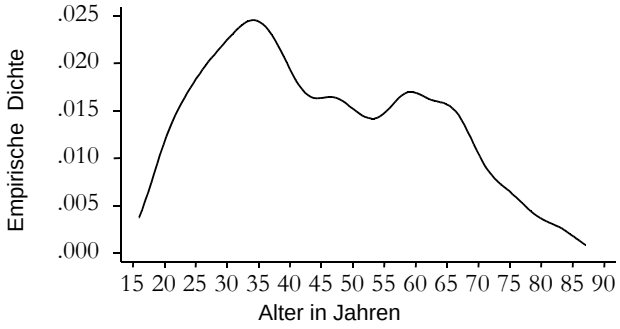
Box-Plots konzentrieren sich auf wenige Merkmale einer Verteilung:

- die „Box“ gibt die Lage der mittleren 50% aller Realisierungen einer Verteilung an;
- ein Strich in der Box kennzeichnet den Median, der die Verteilung in zwei gleich stark besetzte Hälften teilt;
- Linien links und rechts von der Box zeigen (mit Ausnahme möglicher extremer Ausreißerwerte) den Wertebereich an;
- gibt es Extremwerte, die mehr als 1.5 mal weiter vom oberen oder unteren Ende der Box entfernt sind, als die Box selbst lang ist, werden diese durch zusätzliche Punkte oder Sternchen außerhalb der Linien gekennzeichnet.

Bei der Berechnung der Quartile und des Medians werden i.a. etwas andere Rechenformeln verwendet als bei empirischen Quantilen.

Darstellungen metrischer Variablen: Interpretation





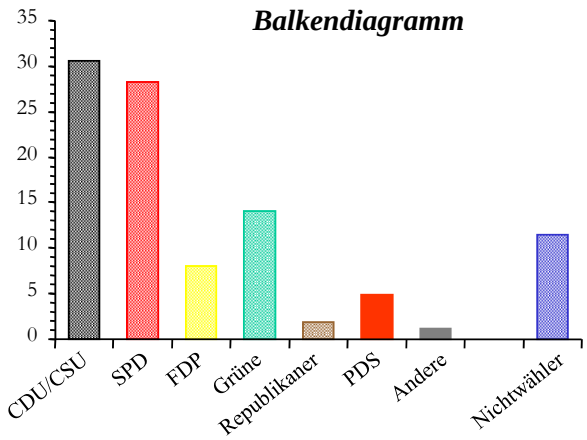
Vor allem die grafische Darstellung der Verteilung mittels Kern-Dichte-Schätzung weist darauf hin, dass die Verteilung linkssteil bzw. rechtsschief ist, zudem bimodal ist, wobei der zweite Gipfel deutlich geringer ist als der erste Gipfel. Die Rechtsschiefe lässt sich auch aus dem Box-Plot der Verteilung ablesen, die Mehrgipfligkeit dagegen nicht.

Für die inhaltliche Interpretation muss der Erhebungszeitraum 1996 berücksichtigt werden. Der erste Gipfel bei Mitte 30 bezieht sich auf Befragte, die in den 60er Jahren des 20. Jhds. geboren sind. Der zweite Gipfel mit Befragten der Altersgruppe zwischen 55 und 65 bezieht sich auf Befragte aus der ersten Hälfte der 30er Jahren.

Grafische Darstellung nominalskalierter Variablen

Bei nominalskalierten Variablen ist zu beachten, dass durch die grafische Anordnung nicht der Eindruck entsteht, dass die Zahlenwerte (Codes) der Ausprägungen inhaltliche Bedeutung haben.

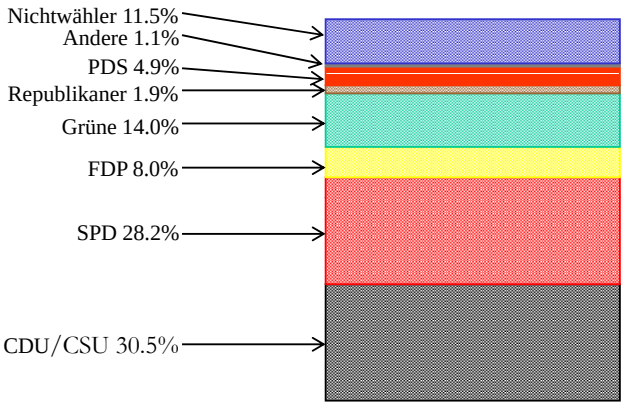
Das Beispiel zeigt die Antwortverteilung der Wahlabsichtsfrage aus dem Allbus 1996



Balkendiagramme entsprechen Stabdiagrammen.
Für jede Ausprägung wird ein Balken gezeichnet, dessen Länge der absoluten oder relativen Besetzungshäufigkeit entspricht.
Um den Eindruck einer ordinalen oder stetigen Variable zu vermeiden, dürfen sich die Balken nicht berühren.

Darstellungen nominalskalierter Variablen

Säulendiagramm

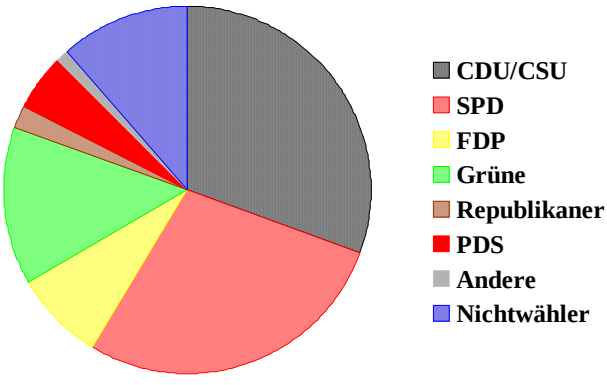


In ***Säulendiagrammen*** wird ein Balken in Teilabschnitte eingeteilt, wobei jeder Abschnitt für eine Ausprägung steht.
Die Abschnittsbreite entspricht der (relativen) Häufigkeit dieser Ausprägung.
Säulendiagramme eignen sich gut beim Vergleich von Verteilungen in Subgruppen.

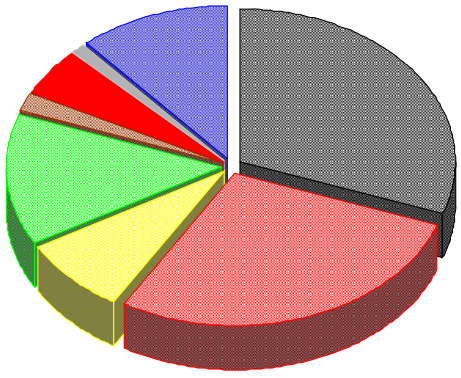
Die Anordnung der Teilabschnitte ist bei nominalen Skalenniveau irrelevant und kann daher nach pragmatischen Gesichtspunkten erfolgen.

Darstellungen nominalskalierter Variablen

Kreisdiagram

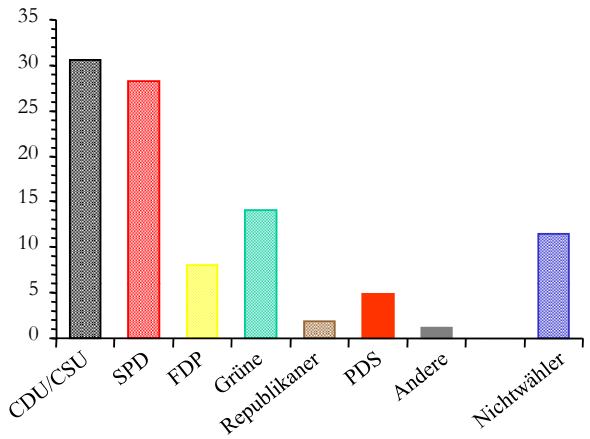


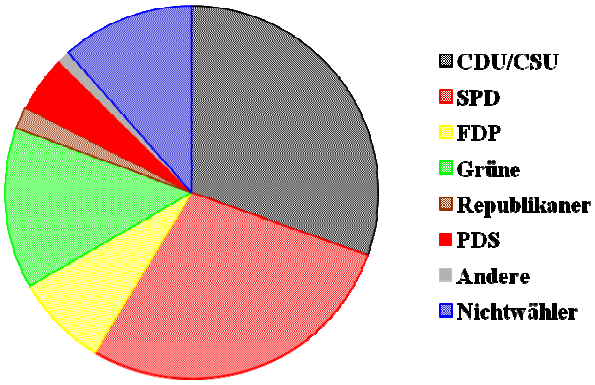
Tortendiagram



In ***Kreisdiagrammen*** und ***Tortendiagrammen*** wird ein Kreis bzw. ein Zylinder in Segmente zerteilt, die für die Ausprägungen stehen.
Die relative Häufigkeit einer Ausprägung wird durch den Umfang des zugehörigen Segments, d.h. seinem Winkelanteil an den insgesamt 360° des Kreisumfangs bestimmt.
Durch unterschiedliche große Kreise bzw. Torten können bei Vergleichen von Verteilungen in Subgruppen unterschiedliche Gruppengrößen berücksichtigt werden.
Ein Nachteil von Kreis- oder Tortendiagrammen ist, dass es oft nicht einfach ist, die relativen Größenverhältnisse der Ausprägungen über die Segmentumfänge abzuschätzen.

Darstellungen nominalskalierter Variablen: Interpretation





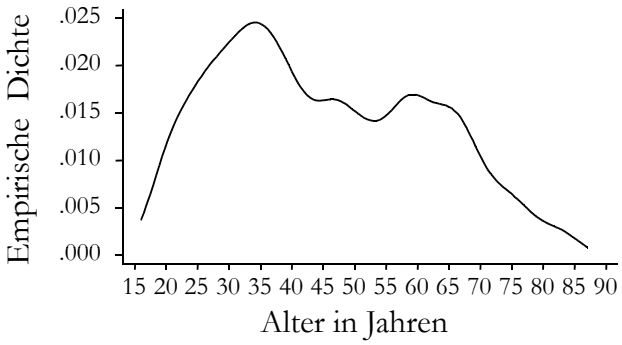
Aus der grafischen Darstellung wird deutlich, dass die Mehrheit der Befragten 1996 eine der beiden „großen“ Parteien CDU/CSU bzw. SPD zu wählen beabsichtigten, wenn am nächsten Sonntag (nach der Befragung) eine Bundestagswahl wäre.
Da es sich um ungewichtete Datenh handelt, der Anteil der Befragten aus den neuen Bundesländern höher ist als der tatsächliche Anteil und sich das Wahlverhalten in den neuen und alten Bundesländern unterscheidet, geben die Grafiken nicht die tatsächliche Verteilung der Wahlabsichten wieder, wie sie sich im Sommer 1996 für die Bundesrepublik insgesamt darstellte.

Probleme der grafischen Darstellungen ordinaler Variablen

Für ordinale Verteilungen haben sich keine speziellen Darstellungsweisen durchgesetzt. Da Box-Plots vor allem Quantile darstellen, die ab ordinalen Messniveau berechenbar sind, liegt es nahe, bei ordinalen Variablen Box-Plots zu verwenden. Mehrgipfligkeit lässt sich dann allerdings nicht erkennen. In der Praxis werden daher auch Stabdiagramme und Histogramme für die Darstellung ordinaler Variablen verwendet. Dabei werden anstelle der metrischen Werte Rangplätze dargestellt.

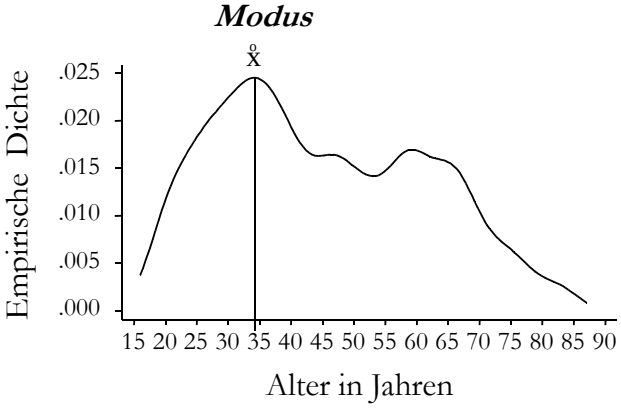
Aufgrund der Hierarchie der Messniveaus können auch Balken, Säulen- oder Kreisdiagramme zur Darstellung ordinaler Verteilungen verwendet werden, wobei es sich anbietet, bei Balken- und Säulendiagramme die Balken bzw. Säulenabschnitte entsprechend der Rangfolge der Ausprägungen anzuordnen. Tatsächlich dürften diese beiden Darstellungsformen am geeignetsten sein, da sie einerseits die Form der Verteilung wiedergeben, um Unterschied zu Histogrammen oder gar Darstellungen von Kern-Dichte-Schätzungen aber nicht so leicht suggerieren, dass die Abstände zwischen den Ausprägungen sich im Sinnen von Zahlenabständen interpretieren lassen.

Lerneinheit 5: Lagemaße



Grafische Darstellungen geben eine gute Übersicht über eine Verteilung, setzen aber ein Darstellungsmedium voraus. Rein sprachlich lassen sie sich nur schwer beschreiben. Vor allem, wenn es darum geht, (viele) Verteilungen zu vergleichen, stoßen grafische Darstellungen auch schnell an ihre Grenzen.
Anstelle alle Realisierungen zu betrachten, benötigt man daher in der Statistik oft eine oder wenige ***Kenngrößen,*** die sogenannten ***Verteilungsparameter***, die charakteristisch für die ganze Verteilung sind.
Ein Parameter, der gewissermaßen repräsentativ für eine Verteilung sein soll, wird auch als ***typischer Wert*** bezeichnet. Da ein typischer Wert bei metrischen Verteilungen den Ort oder die Lage der Verteilung auf der Achse der Zahlen angibt, spricht man auch von einem ***Lagemaß***.

Der Modus oder Modalwert

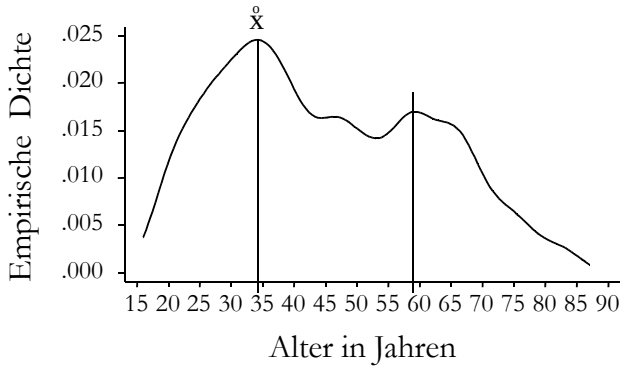


Es liegt nahe, als charakteristischen Wert einer Verteilung einfach die ***Ausprägung*** zu benennen, ***die am häufigsten realisiert*** wird. Dieser Wert wird als ***Modus*** oder ***Modalwert*** (engl: *mode*) einer Verteilung bezeichnet. In Formeln wird der Modus oft durch einen kleinen Kreis über dem Symbol für die betrachtete Variable dargestellt:

$$\overset{o}{x} = \left\{ x_{(k)} \middle| n_{(k)} > n_{(j)} \text{ für alle } j \neq k \right\}$$

In der abgebildeten Altersverteilung liegt der Modus bei etwa 34 Jahren: dieses Alter kommt in der Verteilung am häufigsten vor.

Modus bei mehrgipfligen Verteilungen



Sinnvoll ist die Wahl des Modus als charakteristischen Wert einer Verteilung nur, wenn es tatsächlich nur eine Ausprägung gibt, die am häufigsten vorkommt. Bei bi- oder multimodalen (mehrgipfligen) Verteilungen muss daher ein Gipfel besonders herausragen.

> *Die in der Abbildung wiedergegebene Altersverteilung ist zwar bimodal, da es neben dem Maximum von 34 Jahren ein zweites bei 59 Jahren gibt. Das zweite Maximum hat allerdings eine deutlich geringere empirische Dichte.*

Hinweis:
In Statistikprogrammen wird auch dann ein Modalwert berechnet, wenn mehrere Maxima mit gleichen Häufigkeiten auftreten. Der ausgewiesene Wert ist dann entweder das erste oder das letzte Maximum der Verteilung.

Modus in Häufigkeitstabellen

Bewertung der allgemeinen Wirtschaftslage

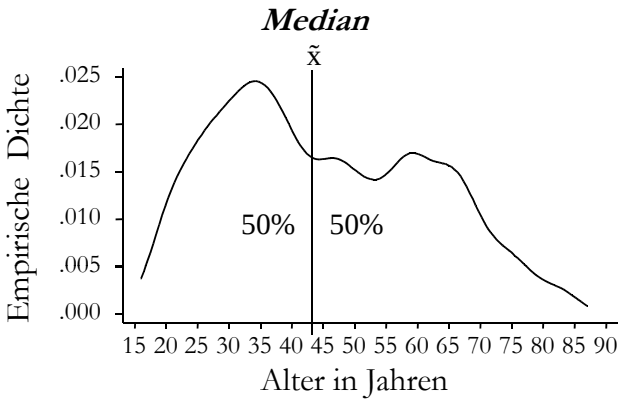
Ausprägung	Code	Häufigkeit	Prozente	Gültige Prozente	Kumulierte Prozente
sehr gut	1	35	1.0	1.0	1.0
gut	2	434	12.7	12.7	13.7
teils/teils	3	1619	47.3	47.5	61.2
schlecht	4	1100	32.1	32.3	93.5
sehr schlecht	5	222	6.5	6.5	100.0
weiß nicht	8	11	0.3	Missing	
keine Angabe	9	1	0.0	Missing	
Total		3422	100.0	100.0	

Gültige Fälle: 3410 Fehlende Fälle: 12
(Daten: ALLBUS 2006, gewichtet nach Ost-West)

Bei der Häufigkeitsverteilung der Beurteilung der allgemeinen Wirtschaftsslage im Allbus 2006 ist der Modus der Wert 3, d.h. die Ausprägung „teils/teils“.
Von den Befragten wird diese Kategorie mit 1619 Realisierungen am häufigsten gewählt.

Bei gruppierten Daten wird als Modus die Klassenmitte der Klasse gewählt, die die größe Besetzung aufweist. Dies macht offensichtlich nur Sinn, wenn die Klassenbildung nicht so gewählt ist, dass die Klassen gleiche Häufigkeiten aufweisen.
In der Regel ist der Modus bei gruppierten Daten nicht informativ!

Der Median



Ein anderer charakteristischer Wert einer Verteilung ist der ***Median***, das ist der Wert, der eine Verteilung in ***zwei gleich stark besetzte Hälften*** zerteilt. Für den Median gilt also, dass jeweils (mindetens) die Hälfte der Realisierungen sowohl kleiner oder gleich als auch größer oder gleich diesem Wert sind.
Der Medien wird durch eine Tilde (~) über dem Symbol für eine Ausprägung der betrachteten Variable gekennzeichnet.
Der Median von X ist also $\tilde{x}$.

Die Definition des Medians weist auf Ähnlichkeiten mit dem 50%-Quantil hin. Tatsächlich sind der Median und das empirische 50%-Quantil bei vielen Verteilungen identisch.

Berechnung des Medians bei gerader Fallzahl

Bei geraden Fallzahlen kann es allerdings zu Abweichungen kommen, da bei der Berechnung des Medians eine andere Berechnungsmethode verwendet wird.

X
4
1
3
5
5
2
6
6
1
6

X geordnet	Position (i)
1	1
1	2
2	3
3	4
4	**5**
5	**6**
5	7
6	8
6	9
6	10

Für die Berechnung müssen zunächst alle gültigen Realisierungen der Größe nach angeordnet werden.

Bei einer ***geraden Fallzahl*** ist der ***Median*** dann definiert als der ***Mittelwert*** der beiden Fälle auf den ***Rangplätzen (n/2)*** und ***(n/2+1)***:

$$\tilde{x} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Bei dieser Berechnungsweise sind dann jeweils mindestens n/2 Realisierungen kleiner oder gleich dem Median und n/2 Realisierungen größer oder gleich dem Median.

Im Beispiel der Verteilung mit 10 Fällen ist der Median der Mittelwert aus den beiden Realisierungen an der 5. (= 10/2) und an der 6. (= 10/2 +1) Position nach der aufsteigenden Anordnung aller Realisierungen, hier also der Mittelwert von 4 (= $x_{(5)}$) und 5 (= $x_{(6)}$). Im Beispiel ist der Median daher 4.5 (= (4+5)/2) und damit größer als das empirische 50%-Quantil, das den Wert 4 aufweist.
Es sind 5 (= 10/2) Realisierungen kleiner/gleich und 5 Realisierungen größer/gleich 4.5.

Berechnung des Medians bei ungerader Fallzahl

Bei ungerader Fallzahl ist der Median stets gleich dem empirischen 50%-Quantil.

Als Beispiel wird der Median für die n=9 gültigen Fälle des Geburtsjahres aus der Beispieldatenmatrix (in L02) berechnet.

Geburtsjahr F4
1943
1960
1957
1939
missing
1956
1970
1920
1956
1966

F4 geordnet	Position (i)
1920	1
1939	2
1943	3
1956	4
1956	**5**
1957	6
1960	7
1966	8
1970	9
missing	missing

Die Realisierungen der Verteilung müssen wiederum zunächst der Größe nach geordnet werden.

Bei einer ***ungeraden Fallzahl*** ist der ***Median*** dann die Realisierung mit dem ***Rangplatz (n+1)/2***:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

Dann sind nämlich jeweils mindestens (n–1)/2 Realisierungen kleiner oder gleich dem Median und (n–1)/2 Realisierungen größer oder gleich dem Median.

Im Beispiel der 9 gültigen Fälle ist der Median gleich dem Wert der Realisierung des Falles, der nach der Anordnung der Größe nach an der 5. (= (9+1)/2) Position steht, hier also das Jahr 1956. Vier (= (9-1)/2) Fälle der Verteilung sind dann kleiner oder gleich, d.h. früher oder im gleichem Jahr geboren, und vier Fälle sind größer oder gleich, d.h. im gleichem Jahr oder später geboren.

Berechnung des Medians bei Häufigkeitstabellen ungruppierter Daten

Bewertung der allgemeinen Wirtschaftslage

Ausprägung	Code	Häufigkeit	Prozente	Kumulierte Prozente
sehr gut	1	35	1.0	1.0
gut	2	434	12.7	13.7
teils/teils	3	1619	47.5	61.2
schlecht	4	1100	32.3	93.5
sehr schlecht	5	222	6.5	100.0
Total		3410	100.0	

(Daten: ALLBUS 2006, gewichtet nach Ost-West)

Kumulierte Häufigkeiten
35
469
2088
3188
3410

Wenn eine Verteilung als Häufigkeitstabelle vorliegt, kann der Median in der Regel direkt aus der Häufigkeitstabelle abgelesen werden: Es ist die Ausprägung, bei der die kumulierten Anteile den Wert 0.5 bzw. 50% erstmals überschreiten.

Bei den Allbus-Daten 2006 zur Bewertung der allgemeinen Wirtschaftslage ist dies der Wert 3 bzw. die Kategorie „teils/teils“, bei der die kumulierten Prozente erstmals >50% sind.

Der Median ist der Mittelwert der Fälle mit den Rangplätzen 1705 (= 3410/2) und 1706. Beide Realisierungen weisen die dritte Ausprägung der Variablen auf. Daher ist jeweils die Hälfte der insgesamt 3410 gültigen Fälle kleiner gleich und gleichzeitig größer oder gleich diesem Wert.

Berechnung des Medians bei Häufigkeitstabellen ungruppierter Daten

Wirtschaftslage in BRD Ausprägung	Code	Häufigkeit	Kumulierte Prozente	Kumulierte Häufigkeiten
sehr gut	1	35	1.0	35
gut	2	434	13.7	469
teils/teils	3	1236	**50.0**	1705 = 3410/2
schlecht	4	1483	93.5	3188
sehr schlecht	5	222	100.0	3410
Total		3410		

(Quelle: fiktive Daten in Anlehnung an Allbus 2006)

Wenn allerdings bei einer Ausprägung die kumulierte relative Häufigkeit exakt (d.h. ohne Rundungsfehler) den Wert 50% erreicht, was nur bei gerader Fallzahl möglich ist, dann ist der Median gemäß seiner Definition der Mittelwert aus dieser und der nächsten Ausprägung.

Im Beispiel mit fiktiven Daten weist der Median somit den Wert 3.5 = (3 + 4) /2 auf. Da eine Mittelwertberechnung nur bei metrischen Variablen zulässig ist, ist es hier sinnvoller, zu sagen: Der Median liegt genau zwischen der 3. („teils/teils“) und der 4. („schlecht“) Kategorie der Variablen.

Berechnung des Medians bei Häufigkeitstabellen gruppierter Daten

	$u_{(k)}$ $o_{(k)}$ Ausprägung in Jahren exakte Klassengenzen)	$m_{(k)}$ Code = Klassenmitte	$n_{(k)}$ Häufigkeit	Prozente	$p_{(k)}$ Gültige Prozente	$cp_{(k)}$ Kumulierte Prozente
k=1	17.5 bis <29.5	23.5	507	14.8	14.9	14.9
k=2	29.5 bis < 44.5	37.0	943	27.6	27.6	42.5
k=3	44.5 bis <59.5	52.0	904	26.4	26.5	69.0
k=4	59.5 bis <74.5	67.0	812	23.7	23.8	92.8
k=5	74.5 bis <94.5	84.5	246	7.2	7.2	100.0
	keine Angabe	--	10	0.3	Missing	
	Total		3421	100.0	100.0	

Gültige Fälle: 3411 Fehlende Fälle: 10

(Quelle: Allbus 2006, gewichtet nach Ost-West)

Bei gruppierten Daten ist der Median das über die Summenfunktion interpolierte 50%-Quantil:

$$\tilde{x} = o_{(k-1)} + \frac{0.5 - cp_{(k-1)}}{p_{(k)}} \cdot \left(o_{(k)} - o_{(k-1)}\right) = 44.5 + \frac{0.5 - 0.425}{26.5} \cdot (59.5 - 44.5) = 48.75$$

In der gruppierten Altersverteilung der Befragten aus dem Allbus 2006 liegt der Median in der 3. Klasse der Altersgruppe von 44.5 bis unter 59.5 Jahren.
Die lineare Interpolation ergibt für den Median (das 50%-Quantil) den Wert 48.75 Jahre. Dieser Wert liegt sehr dicht bei dem auf der Basis der ungruppierten Daten berechneten Median, der einen Wert von 49 Jahren aufweist.

Minimierungseigenschaft des Medians

Wirtschaftslage in BRD			Kumulierte
Ausprägung	Code	Häufigkeit	Prozente
sehr gut	1	35	1.0
gut	2	434	13.7
teils/teils	3	1236	**50.0**
schlecht	4	1483	93.5
sehr schlecht	5	222	100.0
Total		3410	

(Quelle: fiktive Daten in Anlehnung an Allbus 2006)

X	$n_{(k)}\cdot\lvert x-3.5\rvert$	$n_{(k)}\cdot\lvert x-2.5\rvert$	$n_{(k)}\cdot\lvert x-3\rvert$	$n_{(k)}\cdot\lvert x-4\rvert$
1	87.5	52.5	70	105
2	651.0	217.0	434	868
3	618.0	618.0	0	1236
4	741.5	2224.5	1483	0
5	333.0	555.0	444	222
$\sum$	2431.0	3667.0	2431	2431

Der ***Median*** weist die Eigenschaft auf, dass die ***Summe der absoluten***, d.h. vorzeichenbereinigten, ***Differenzen*** aller Realisierungen vom Median ***minimal*** ist:

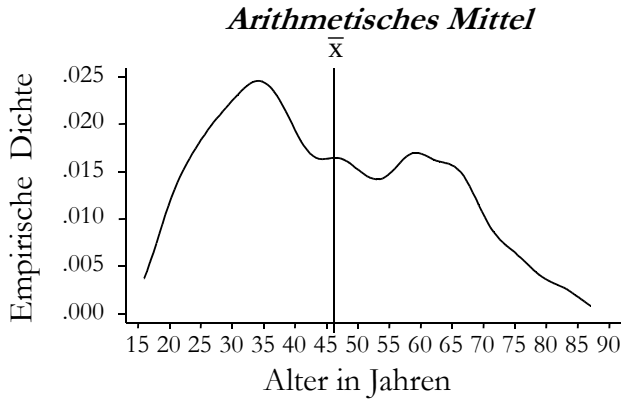
$$\sum_{i=1}^{n}\left|x_i-\tilde{x}\right| \le \sum_{i=1}^{n}\left|x_i-a\right| \quad \text{für alle Werte a}$$

Im Beispiel beträgt die Summe der absoluten Differenzen der Realisierungen vom Median 2431. Die Summe der Abweichungen vom Wert 2.5 ist dagegen mit 3367 deutlich größer.

Allerdings ist diese Eigenschaft bei gerader Fallzahl nicht immer eindeutig. Sie gilt dann für alle Werte zwischen den Ausprägungen $x_{(n/2)}$ bis $x_{(n/2+1)}$.

So beträgt im Beispiel die Summe der Abweichungen aller Realisierungen von $x_{(3405)} = 3$ und $x_{(3406)} = 4$ wie beim Median 2431.

Das arithmetisches Mittel



Der bei metrischen Daten am häufigsten verwendete typische Wert einer Verteilung ist das ***arithmetische Mittel*** *(engl:* ***mean****)*, das auch als ***Mittelwert*** oder ***Durchschnitt*** bezeichnet wird.

Der Mittelwert einer Verteilung berechnet sich aus derr Summe aller Realisierungen (mit gültigen Werten) geteilt durch die Anzahl dieser Realisierungen:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Berechnung des arithmetischen Mittels für eine Variable der Datematrix

Für die Berechnung des Mittelwerts einer Variable (Spalte) der Datenmatrix müssen alle gültigen Realisierungen in der Spalte aufsummiert und anschließend durch die Zahl der Fälle mit gültigen Werten dividiert werden.

Geburts-jahr F4	Alter (X)
1943	65
1960	48
1957	51
1939	69
missing	missing
1956	52
1970	38
1920	88
1956	52
1966	42
Summe	505
$\frac{\text{Summe}}{9}$	56.1

Als Beispiel soll aus dem Geburtsjahr des Beispielfragebogens aus L02 das durchschnitt liche Alter der Befragten berechnet werden. Das Alter ergibt sich, wenn für jede gültige Realisierung vom Erhebungszeitpunkt (z.B. 2008) das Geburtsjahr abgezogen wird.

Für das Beispiel ergibt sich:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{65+48+51+69+52+38+88+52+42}{9} = \frac{505}{9} = 56.1$$

Das durchschnittliche Alter beträgt im Beispiel 56.1 Jahre.

Im Unterschied zur Berechnung des Medians müssen die Realisierungen bei der Berechnung des Mittelwerts nicht der Größe nach angeordnet sein.

Berechnung des arithmetischen Mittels in einer Häufigkeitstabelle

	Alter (X)
	65
	48
	51
	69
	missing
	52
	38
	88
	52
	42
Summe	505
$\frac{\text{Summe}}{9}$	56.1

Alter in Jahren	Häufigkeit	Anteile	gültige Anteile	kumulierte Anteile
38	1	0.100	0.111	0.111
42	1	0.100	0.111	0.222
48	1	0.100	0.111	0.333
51	1	0.100	0.111	0.444
52	2	0.200	0.222	0.667
65	1	0.100	0.111	0.778
69	1	0.100	0.111	0.889
88	1	0.100	0.111	1.000
999	1	0.100	--	
Total	10	1.000	1.000	

Liegen die Realisierungen in einer Häufigkeitstabelle vor, ist zu beachten, wie oft eine Ausprägung vorkommt:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{\overbrace{38}^{1\times} + \overbrace{42}^{1\times} + \overbrace{48}^{1\times} + \overbrace{51}^{1\times} + \overbrace{52+52}^{2\times} + \overbrace{65}^{1\times} + \overbrace{69}^{1\times} + \overbrace{88}^{1\times}}{9} = 56.1$$

$$= \frac{1}{9} \cdot (1\times 38 + 1\times 42 + 1\times 48 + 1\times 51 + 2\times 52 + 1\times 65 + 1\times 69 + 1\times 88)$$

$$= \frac{1}{n} \sum_{k=1}^{K} n_k \cdot x_k \text{ mit } n = \sum_{k=1}^{K} n_k$$

Berechnung des arithmetischen Mittels in einer Häufigkeitstabelle

	Alter (X)
	65
	48
	51
	69
	missing
	52
	38
	88
	52
	42
Summe	505
Summe / 9	56.1

Alter in Jahren	Häufigkeit	Anteile	gültige Anteile	kumulierte Anteile
38	1	0.100	0.111	0.111
42	1	0.100	0.111	0.222
48	1	0.100	0.111	0.333
51	1	0.100	0.111	0.444
52	2	0.200	0.222	0.667
65	1	0.100	0.111	0.778
69	1	0.100	0.111	0.889
88	1	0.100	0.111	1.000
–9	1	0.100	--	
Total	10	1.000	1.000	

In Häufigkeitstabellen berechnet sich der Mittelwert also als mit den Auftretenshäufigkeiten ***gewichtetes Mittel*** der gültigen Ausprägungen. Anstelle der Gewichtung mit den absoluten Häufigkeiten kann auch über die relativen Häufigkeiten gewichtet werden:

$$\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{K} n_k \cdot x_k = \frac{n}{n}\cdot\sum_{k=1}^{K}\frac{n_k \cdot x_k}{n} = \sum_{k=1}^{K} p_k \cdot x_k$$

$$= \frac{1}{9}\cdot 38 + \frac{1}{9}\cdot 42 + \frac{1}{9}\cdot 48 + \frac{1}{9}\cdot 51 + \frac{2}{9}\cdot 52 + \frac{1}{9}\cdot 65 + \frac{1}{9}\cdot 69 + \frac{1}{9}\cdot 88 = 56.1$$

Aufgrund größerer Rundungsfehler ist die Verwendung relativer Häufigkeiten oft ungenauer.

Berechnung des arithmetischen Mittels in gruppierten Häufgkeitstabellen

	u_k – o_k Ausprägung in Jahren exakte Klassengenzen)	m_k Code = Klassenmitte	n_k Häufigkeit	Prozente	p_k Gültige Prozente	cp_k Kumulierte Prozente
k=1	17.5 bis <29.5	23.5	507	14.8	14.9	14.9
k=2	29.5 bis < 44.5	37.0	943	27.6	27.6	42.5
k=3	44.5 bis <59.5	52.0	904	26.4	26.5	69.0
k=4	59.5 bis <74.5	67.0	812	23.7	23.8	92.8
k=5	74.5 bis <94.5	84.5	246	7.2	7.2	100.0
	keine Angabe	--	10	0.3	Missing	
	Total		3422	100.0	100.0	

Gültige Fälle: 3411 Fehlende Fälle: 10

(Quelle: Allbus 2006, gewichtet nach Ost-West)

n_k	m_k	$n_k \cdot m_k$
507	23.5	11914.5
943	37.0	34891.0
904	52.0	47008.0
812	67.0	54404.0
246	84.5	20786.0
3411		169004.5
Summe/3411		49.5

p_k	m_k	$n_k \cdot m_k$
.149	23.5	3.5015
.276	37.0	10.2120
.265	52.0	13.7800
.238	67.0	15.9460
.072	84.5	6.0840
3411		49.5

Bei gruppierten Daten werden statt der Ausprägungen die Mittelwerte jeder Klasse herangezogen:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{K} n_k \cdot m_k = \sum_{k=1}^{K} p_k \cdot m_k$$

Anstelle der absoluten Häufigkeiten kann wiederum mit den relativen Häufigkeiten gewichtet werden.

Berechnung des arithmetischen Mittels in gruppierten Häufgkeitstabellen

Klassenmitten

n_k	m_k	$n_k \cdot m_k$
507	23.5	11914.5
943	37.0	34891.0
904	52.0	47008.0
812	67.0	54404.0
246	84.5	20786.0
3411		169004.5
Summe/3411		49.5

Klassenmittelwerte

n_k	m_k	$n_k \cdot m_k$
507	23.19	11757.33
943	37.81	35654.83
904	51.88	46899.52
812	66.82	54257.84
246	79.98	19675.08
3411		168244.60
Summe/3411		49.3

Der auf der Basis der gruppierten Daten berechnete Mittelwert liegt sehr dicht beim tatsächlichen Mittelwert, der 49.3 Jahre beträgt.

Abweichungen zwischen dem über gruppierten Daten berechneten Mittelwert und dem Mittelwert der ungruppierten Verteilung sind Folge davon, dass die Klassenmitten von den Mittelwerten der Realisierungen in den Alterklassen abweichen können.

Im Beispiel gibt es vor allem bei den stark besetzten Altersklassen nur relativ geringe Abweichungen zwischen der jeweiligen Klassenmitte und dem Klassenmittelwert.
Werden die Klassenmitten durch die korrekten Klassenmittelwerte ersetzt, ergibt sich der korrekte Altersmittelwert von 49.3 Jahren in der Allbus-Stichprobe.

Minimierungseigenschaften des arithmetischen Mittels

Während der Median der Wert ist, der die absoluten Abweichungen aller Realisierungen von diesem Wert minimiert, minimiert das arithmetische Mittel die quadrierten Abweichungen:

- **Die Summe der quadrierten Abweichungen vom Mittelwert ist minimal.**

Als Konsequenz aus dieser Eigenschaft folgt:

- **Die Summe der Abweichungen vom Mittelwert ist null.**

Die folgenden Beispieldaten verdeutlichen diese Eigenschaft:

	X	X–3	$(X-3)^2$	X–2.9	$(X-2.9)^2$	X–3.1	$(X-3.1)^2$
	1	–2	4	–1.9	3.61	–2.1	4.41
	2	–1	1	–0.9	0.81	–1.1	1.21
	3	0	0	0.1	0.01	–0.1	0.01
	4	1	1	1.1	1.21	0.9	0.81
	5	2	4	2.1	4.41	1.9	3.61
$\sum$	15	0	10	0.5	10.05	–0.5	10.05
$1/n \cdot \sum$	3						

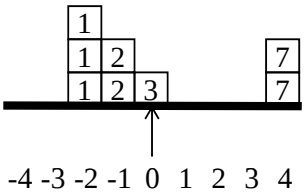
$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \qquad \sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - a)^2 \quad \text{für beliebige } a \neq \bar{x}$$

Anders als beim Median gilt die Minimierungseigenschaft des Mittelwertes auch bei geraden Fallzahlen nur für den Mittelwert selbst.

Das arithmetische Mittel als Schwerpunkt einer Verteilung

Die Minimierungseigenschaft des Mittelwerts führt dazu, dass der Mittelwert auch der ***Schwerpunkt*** der betrachteten Verteilung ist.

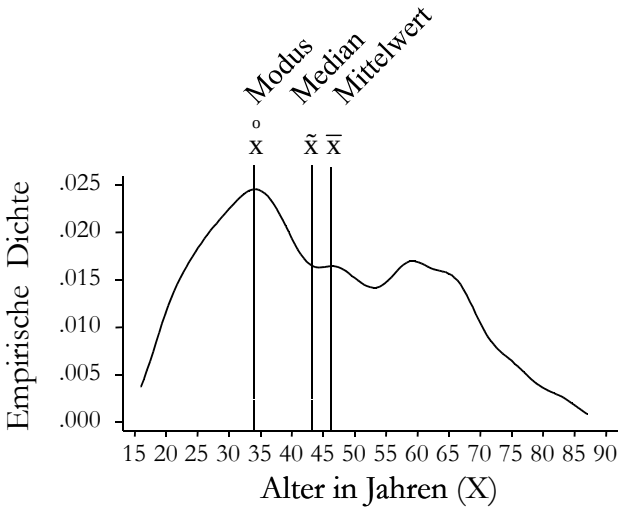
X	n_k	p_k	cp_k	$x_k \cdot p_k$	$n_k \cdot (x_k - 3)$
1	3	0.375	0.375	0.375	–6
2	2	0.250	0.625	0.500	–2
3	1	0.125	0.750	0.375	0
7	2	0.250	1.000	1.750	8
∑	8	1.000		3.000	0



Im Beispiel wird eine Verteilung von 8 Fällen betrachtet, deren Mittelwert 3.0 ist.

Werden alle Realisierungen durch jeweils ein gleich großes Gewicht symbolisiert und die Gewichte so auf eine Wippe gelegt, dass Gewichte mit gleichen Ausprägungen übereinander liegen und der Abstand zwischen den Gewichten unterschiedlicher Ausprägungen dem Abstand der Werte dieser Ausprägunge entspricht, so ist die Wippe genau dann waagerecht, wenn ihr Gelenk, an dem sie sich drehen kann, der Position des arithmetischen Mittels der Verteilung entspricht.

Auswahl eines geeingeten Lagemaßes



In der Praxis stellt sich die Frage, welches Lagemaß zur Charakterisierung einer Verteilung herangezogen werden sollte. Als Entscheidungshilfe werden meist drei Kriterien berücksichtigt:

- das Skalenniveau der betrachteten Variable,
- die Robustheit des Lagemaßes gegenüber Ausreißern
- und die Informationshaltigkeit des Lagemaßes.

Robustheit gegenüber Ausreißern

Ausreißerwerte sind Ausprägungen an den Rändern der Verteilung, die einerseits eine geringe Auftretenshäufigkeit haben und andererseits einen großen Abstand zu den meisten anderen Werten aufweisen.
Robustheit gegenüber Ausreißern bezieht sich nun darauf, wie empfindlich ein Parameter auf die Realisierung von Ausreißerwerten reagiert. Vergleich man darauf die drei Lagemaße Modus, Median und arithmetrisches Mittel so gilt:

- Der Median ist sehr unempfindlich gegenüber Ausreißerwerten.
  Da der Wert des Medians nur vom mittleren Wert oder den beiden mittleren Werten einer Verteilung nach der Anordnung der Größe nach abhängt, spielen die Werte an den Rändern der Verteilung überhaupt keine Rolle.
- Der Modus ist ebenfalls in der Regel robust gegenüber Ausreißerwerten, jedenfalls dann, wenn die Häufigkeit des Modalwertes deutlich höher ist als die aller anderern Ausprägungen.
- Das arithmetische Mittel reagiert dagegen empfindlich gegenüber Ausreißerwerten.
  Da der Mittelwert die quadrierten Abstände zu sich minimiert, bedeutet das, dass Werte an den Rändern der Verteilung, die sehr weit von den übrigen Werten entfernt sind, den Mittelwert relativ stärker beeinflussen und in ihre Richtung „ziehen“.
  Dies ergibt sich auch aus der Schwerpunkteigenschaft des Mittelwerts. Auf einer Wippe kann ein relativ kleines Gewicht, das weiter vom Drehpunkt entfernt ist, größere Gewichte, die dichter am Dehpunkt sind, ausgleichen.

Vorausgesetztes Messniveau

Die drei Lagemaße unterscheiden sich auch bei dem Messniveau, dass sie voraussetzen:

- Der Modus betrachtet ausschließlich die Auftretenshäufigkeit einer Ausprägung. Er kann daher bereits ab Nominalskalenniveau verwendet werden, da nach jeder bei Nominalskalenniveau zulässigen Transformation der Modus der transformierten Verteilung gleich dem transformierten Modalwert ist.
- in den Median fließen relative Größenpositionen ein. Er ist daher wie alle Quantile erst ab Ordinalskalenniveau zulässig. Da der Median bei gerader Fallzahl der Mittelwert aus den Realisierungen $x_{(n/2)}$ und $x_{(n/2+1)}$ ist, wird hier streng genommen sogar metrisches Messniveau vorausgesetzt, falls die beiden Realisierungen unterschiedliche Ausprägungen aufweisen. Allerdings kann in dieser Situation auch festgehalten werden, dass der Median genau zwischen den Ausprägungen des n/2-ten und des (n/2+1)-ten Falles liegt. Alternativ kann anstelle des Medians auch das empirische 50%-Quantil als Lagemaß verwendet werden, dass grundsätzlich ab Ordinalskalenniveau berechnenbar ist.
- In die Berechnung des Mittelwertes gehen (implizit) die Abstände zwischen den Realisierungen ein. Mittelwerte unterstellen daher metrisches Messniveau.
  Wenn die Ausprägungen der Variablen als Rangplätze interpretiert werden und die unbekannten tatsächlichen Abstände zwischen den Ausprägungen in etwa der Differenz der Rangnummen entsprechen, dann ist der Mittelwert relativ robust gegenüber monotonen Transformationen, ändert also nicht seine relative Position in einer Verteilung. In der Praxis wird der Mittelwert daher auch oft bei stenggenommen nur ordinalskaliertenen Variablen berechnet.

Informationshaltigkeit

Der Informationsgehalt der Lagemaße steht in einem umgekehrten Verhältnis zum vorausgesetzten Skalenniveau:

- Das arithmetischen Mittel hat den höchsten Informationsgehalt, da in dessen Berechnung alle Realisierungen einfließen.
- Der Informationsgehalt des Median ist geringer. Allerdings wird zumindest die Größenordnung der Ausprägungen berücksichtigt.
- Am geringsten ist der Informationsgehalt des Modalwerts.

Kenngröße	Skalenniveau	Robustheit	Informationsgehalt
Modus	ab Nominalskala	bedingt	gering
Median	ab Ordinalskala	hoch	mäßig
Mittelwert	nur metrisch	gering	hoch

Aus den Eigenschaften der Lagemaße folgt, dass bei nominalskalierten Variablen nur der Modus als Lagemaß in Frage kommt.
Bei eindeutig ordinalen Variablen sollte der Median verwendet werden.
Bei metrischen Daten wird meist der Mittelwert herangezogen. Wenn allerdings die Gefahr besteht, dass Werte an den Rändern der Verteilung den Mittelwert sehr stark beeinflussen , dann kann auch bei metrischen Daten der Median geeigneter sein.

So sind Einkommensverteilungen oft extrem schief verteilt mit einer relativ großen Menge von Geringverdienern und nur relativ wenigen Spitzenverdienern. Letztere beeinflussen den Mittelwert allerdings relativ stärker, so dass das Durchschnittsseinkommen einen überhöhten Eindruck von der „typischnen“ Finanzkraft der Untersuchungseinheiten geben kann.

Spezifische Mittelwerte

Geburtsjahr F4	Alter 2008-F4 (X)	Alter geordn. (X)
1943	65	~~-7991~~
1960	48	38
1957	51	42
1939	69	48
9999	-7991	51
1956	52	52
1970	38	52
1920	88	65
1956	52	69
1966	42	88
Summe	-7486	505
Summe / 9	-748.6	56.1

Die fehlende Robustheit gegenüber Ausreißern zeigt sich an den Beispieldaten, wenn bei der Berechnung des Alters das ungültige Geburtsjahr 9999 fälschlich als gültig betrachtet wird und in die Berechnung des Alters einbezogen wird.

Das Durchschnittsalter erreicht dann anstelle des tatsächlichen Wertes von 56.1 Jahren bei den 9 gültigen Fällen den unmöglichen Wert -748.6 Jahre auf der Basis aller 10 Fälle.

Dem gegenüber ändert sich der Median durch die fehlerhafte Einbeziehung des ungültigen Wertes kaum.

Bei der Berücksichtigung aller 10 Fälle beträgt das mittlere Alter gemessen über den Median 51.5 Jahre und bei ausschließlicher Einbeziehung der 9 gültigen Fälle 52 Jahre.

Getrimmte Mittel

Eine Möglichkeit den Mittelwert unempfindlicher gegenüber Ausreißerwerte zu machen, besteht darin, jeweils einen vorgegebenen Prozentsatz der kleinsten und größten Realisationen nicht zu berücksichtigen, den Mittelwert also nur auf der Basis der Werte in der Mitte der Verteilung zu berechnen. Wenn jeweils die x% kleinsten und die x% größten Werte ausgeschlossen werden, ergibt sich das ***x%-getrimmte Mittel***.

Getrimmtes Mittel

Position (i)	Alter geordn. (X)
1	-7991
2	38
3	42
4	48
5	51
6	52
7	52
8	65
9	69
10	88
Summe	-7486
Summe / 9	-748.6

Position (i)	Alter geordn. (X)
--	~~-7991~~
1	38
2	42
3	48
4	51
5	52
6	52
7	65
8	69
--	~~88~~
Summe	417
Summe / 8	52.1

Da die Beispielverteilung nur auf 10 Realisierungen beruht, ergibt sich das 10%-getrimmte Mittel, wenn jeweils der kleinste und der größte Fall von der Berechnung ausgeschlossen werden.

Wird der (ungültige) kleinste Wert -7991 und der (gültige) größte Wert 88 aus der Berechnung ausgeschlossen, beträgt der Mittelwert der verbleibenden 8 Fälle 52.1 Jahre.

Sinnvoller als das Trimmen zum Ausschluss ungültiger Fälle ist offensichtlich, die ungültigen Fälle tatsächlich auch als ungültig zu deklarieren und von vornherein nicht in die Analyse einzubeziehen.

Getrimmte Mittel können als Kompromiss betrachtet werden, um einerseits den Informationsgehalt von Mittelwerten und andererseits die Robustheit des Medians zu erhalten.
Nachteilig ist hierbei, dass wie bei der Berechnung des Medians die Verteilung zunächst der Größe nach angeordnet sein muss. Außerdem gelten die Minimierungseigenschaften des Mittelwerts nicht mehr für die gesamte Verteilung, sondern nur für die nach dem Trimmen verbleibenden Fälle.

Getrimmtes arithmetisches Mittel

Notwendig ist die Berechnung von getrimmten Mitteln bei gruppierten Daten, bei denen die Unter- und/oder Obergrenze der ersten bzw. letzten Klasse nicht bekannt ist.

Dies ist oft bei Einkommensverteilungen der Fall, bei der die oberste Klasse nach oben offen ist wie bei der folgenden Wiedergabe einer fiktiven Einkommenverteilung:

	u_k – o_k Einkommensklasse	n_k Häufigkeit	p_k Anteile	cp_k kum. Anteile	m_k Klassenmitte
k=1	0 € bis < 500 €	150	0.150	0.150	250
k=2	500 € bis < 1500 €	200	0.200	0.350	1000
k=3	1500 € bis < 5000 €	300	0.300	0.650	3250
k=4	5000 € bis < 10000 €	200	0.200	0.850	7500
k=5	10000 € bis < 25000 €	100	0.100	0.950	17500
k=6	25000 € und mehr	50	0.050	1.000	?
	Summe	1000	1.000		

(Quelle: fktive Daten)

Da in der abgebildeten Häufigkeitstabelle die Obergrenze der höchsten Einkommensklasse unbekannt ist, lässt sich deren Klassenmitte nicht berechnen und damit auch kein Mittelwert für gruppierte Daten.

Eine mögliche Lösung ist, die oberste Klasse mit n_6=50 Fällen auszulassen und entsprechend das 5%/getrimmte Mittel zu berechnen.

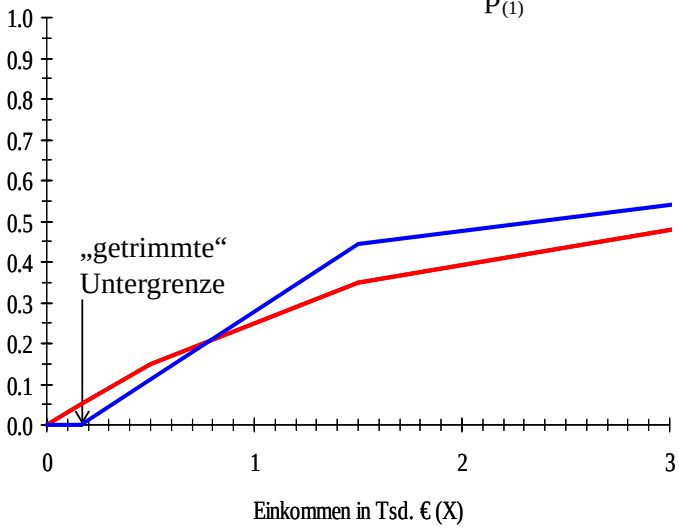
Dabei wird wie bei der Berechnung von Quantilen über die Summenkurve davon ausgegangen, dass sich die Fälle in einer Klasse gleichmäßig über die gesamte Klassenbreite verteilen.

Getrimmtes arithmetisches Mittel

Die Untergrenze des Intervalls verschiebt sich dann entsprechend um den Anteil der ausgelassenen Fälle in dieser Klasse

Wenn α_t bzw. n_t den Anteil bzw. die Häufigkeit der Fälle bezeichnet, um die die erste Klasse getrimmt wird, berechnet sich die durch die Trimmung verschobene Untergrenze u_t nach:

$$u_t = u_{(1)} + \frac{\alpha_t}{p_{(1)}} \cdot \left(o_{(1)} - u_{(1)}\right) = u_{(1)} + \frac{n_t}{n_{(1)}} \cdot \left(o_{(1)} - u_{(1)}\right)$$



Da im Beispiel um den Anteil der obersten Einkommensklasse getrimmt wird, beträgt $n_t = n_K = 50$ bzw. $\alpha_t = p_K = 0.05$. Die getrimmte Untergenze der ersten Klasse berechnet sich daher nach:

$$u_t = 0 + \frac{50}{150} \cdot \left(500 - 0\right) = 167$$

$$= 0 + \frac{0.05}{0.15} \cdot \left(500 - 0\right) = 167$$

Durch die Trimmung reduziert sich die Fallzahl, was Auswirkungen auf den Verlauf der Summenkurve hat.

Getrimmtes arithmetisches Mittel

Einkommensklasse	Häufigkeit	Anteile	kum. Anteile
0 € bis < 500 €	150	0.150	0.150
500 € bis < 1500 €	200	0.200	0.350
1500 € bis < 5000 €	300	0.300	0.650
5000 € bis < 10000 €	200	0.200	0.850
10000 € bis < 25000 €	100	0.100	0.950
25000 € und mehr	50	0.050	1.000
Summe	1000	1.000	

(Quelle: fktive Daten)

$$u_t = 0 + \frac{50}{150} \cdot \left(500 - 0\right) = 167$$

$$= 0 + \frac{0.05}{0.15} \cdot \left(500 - 0\right) = 167$$

⇓

Einkommensklasse	Häufigkeit	Anteile	kum. Anteile
0 € bis < 167 €	~~50~~		
167 € bis < 500	100	0.111	0.111
500 € bis < 1500 €	200	0.333	0.444
1500 € bis < 5000 €	300	0.333	0.667
5000 € bis < 10000 €	200	0.222	0.889
10000 € bis < 25000 €	100	0.111	1.000
25000 € und mehr	~~50~~		
Summe	900	1.000	

(Quelle: fktive Daten)

m_k

Klassenmitte
333.5
1000.0
3250.0
7500.0
17500.0

Getrimmtes arithmetisches Mittel

Einkommensklasse	Häufigkeit	Anteile	kum. Anteile	Klassenmitte	$m_k \cdot p_k$
0 € bis < 167 €	~~50~~				
167 € bis < 500	100	0.111	0.111	333.5	37.037
500 € bis < 1500 €	200	0.333	0.444	1000.0	222.222
1500 € bis < 5000 €	300	0.333	0.667	3250.0	1083.333
5000 € bis < 10000 €	200	0.222	0.889	7500.0	1666.667
10000 € bis < 25000 €	100	0.111	1.000	17500.0	1944.444
25000 € und mehr	~~50~~				
Summe	900	1.000			4953.703

(Quelle: fktive Daten)

Das 5%-getrimmte arithmetische Mittel der Einkommensverteilung beträgt 4953.7 €. Werden also nur die mittleren 90% der Realisierungen betrachtet, beträgt das mittlere Einkommen knapp 5000 €.

Ganz analog zum Verschieben der Untergrenze des ersten Intervalls kann auch die Obergrenze des letzten Intervalls um n_t Fälle bzw. um den Anteil α_t nach unten verschoben werden. Die getrimmte Obergrenze berechnet sich dann nach:

$$o_t = o_{(K)} - \frac{n_t}{n_{(K)}} \cdot \left(o_{(K)} - u_{(K)}\right) = o_{(K)} - \frac{\alpha_t}{p_{(K)}} \cdot \left(o_{(K)} - u_{(K)}\right)$$

Wenn für das Datenbeispiel das 10%-getrimmte Mittel berechnet werden sollte, müsste die Untergrenze der 1. Klasse um 100 Fälle bzw. 0.1 nach oben verschobem werden. Zusätzlich würde die 6. (zweitletzte) Klasse um 50 Fälle oder 0.05 nach unten verschoben werden.

Geometrisches Mittel

Das arithmetische Mittel wird bisweilen verwendet, um den durchschnittlichen Anstieg oder Rückgang von Veränderungen einer Verteilung zu berechnen.

Angenommen, das Vermögen einer Person erhöhe sich im 1. Jahr um 200 €, im 2. Jahr reduziert es sich um 50 € und im 3. Jahr erhöht es sich um 200 €. Dann beträgt der durchschnittliche Anstieg des Vermögens pro Jahr (200–50+300)/3 = 150 €.

Liegen allerdings anstelle additiver Veränderungen ***Veränderungsraten*** vor, dann führt die Berechnung des arithmetischen Mittels zu falschen Ergebnissen.

Wenn das Ausgangsvermögen 1000 € beträgt und es im ersten Jahr um 20% ansteigt, im zweiten Jahr um 5% sinkt und im 3. Jahr um 30% steigt, dann ist die durchschnittliche Veränderungsrate nicht (20%–5%+30%)/3 = 15%, sondern höher wie die folgende Berechnung zeigt:

Ausgangssumme:		1000.00 €	
Anstieg im 1. Jahr um 20% von 1000.00 €		+200.00 €	
Vermögen nach 1 Jahr:	=	1200.00 €	
Reduktion im 2. Jahr um 5% von 1200.00 €		–60.00 €	
Vermögen nach 2 Jahren	=	1140.00 €	
Anstieg im 3. Jahr um 30% von 1140.00 €		+342.00 €	
Vermögen nach 3 Jahren	=	1482.00 €	entspricht 48.2% von 1000 €

48.2% in drei Jahren ergeben 16.067% pro Jahr. Allerdings ist der durchschnittliche „Zuwachs“ pro Jahr sogar geringer als 15%. Ursache hierfür ist, dass Veränderungsraten - wie beim Zinseszins - multiplikativ und nicht additiv wirken.

Geometrisches Mittel

Die korrekte Veränderungsrate ergibt sich, wenn anstelle des arithmetischen Mittels das ***geometrische Mittel*** über die Veränderungsraten berechnet wird.

Das geometrische Mittel von n Faktoren ist die n-te Wurzel aus dem Produkt dieser Faktoren:

$$\bar{x}_{geom.} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Im Beispiel müssen zunächst aus den prozentualen Veränderungen Veränderungsfaktoren berechnet werden:

Ein Antieg um 20% entspricht dem Faktor 1.2 (=1+0.2), eine Reduktion um 5% dem Faktor 0.95 (=1–0.05) und ein Anstieg um 30% dem Faktor 1.3 (=1+0.3). Das geometrische Mittel berechnet sich dann nach:

$$\bar{x}_{geom.} = \sqrt[3]{1.20 \cdot 0.95 \cdot 1.30} = \sqrt[3]{1.482} = 1.140117 \text{ oder } 14.0117\%$$

Die durchschnittliche Veränderung beträgt also 14.0117%. Steigt das Vermögen in jedem Jahr um diesen Wert, ergibt sich nach 3 Jahren ein Anstieg um 48.2%:

Ausgangssumme:		1000.000 €	
1. Jahr: +14.0117% von 1000.000 €		+140.117 €	
Vermögen nach 1 Jahr:	=	1140.117 €	
2. Jahr: +14.0117% von 1140.117 €		+159.870 €	
Vermögen nach 2 Jahren	=	1299.867 €	
3. Jahr: + 14.0117% con 1299.867 €		+182.133 €	
Vermögen nach 3 Jahren	=	1482.000 €	entspricht 48.2% von 1000 €

Geometrisches Mittel

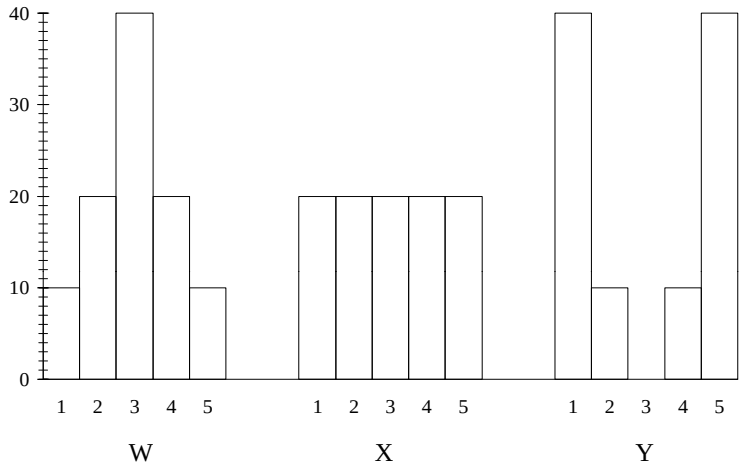
Da durch Logarithieren Produkte zu Summen und Potenzen zu Produkten werden, lässt sich das geometrische Mittel über den Antilogarithmus des arithmetisches Mittels der logarithmierten Veränderungsfaktoren berechnen:

$$\bar{x}_{geom.} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \exp\left(\frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i)\right)$$

Für das Beispiel ergibt sich:

$$\frac{\ln(1.2) + \ln(0.95) + \ln(1.3)}{3} = \frac{0.1823 + (-0.0513) + 0.2624}{3} = 0.1311 \Rightarrow e^{0.1311} = 1.140117$$

Lerneinheit 6: Streuungsmaße und weitere Kenngrößen



Neben einem typischen Wert, der eine Verteilung repräsentieren kann, ist auch von Bedeutung, wie repräsentativ dieser Wert ist, ob also eher mit großen oder mit kleinen Abweichungen zu rechnen ist.

So gilt für die drei über Histogramme abgebildeten Verteilungen von W, X und Y, dass sowohl der Median wie der Mittelwert jeweils der Wert 3.0 ist, obwohl sich die Verteilungen sehr deutlich unterscheiden. Nur bei W stimmen die beiden Lagemaße auch gleichzeitig mit dem Modalwert überein; bei der u-förmigen Verteilung Y sind Mittelwert und Median dagegen Werte, die überhaupt nicht realisiert werden!

Streuungsmaße für metrische Verteilungen

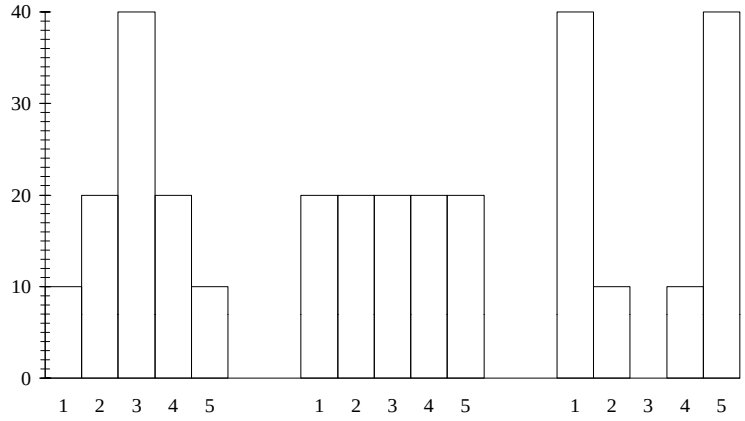
	W			X			Y		
Wert	n_k	p_k	cp_k	n_k	p_k	cp_k	n_k	p_k	cp_k
1	10	0.1	0.1	20	0.2	0.2	40	0.4	0.4
2	20	0.2	0.3	20	0.2	0.4	10	0.1	0.5
3	40	0.4	0.7	20	0.2	0.6	0	0.0	0.5
4	20	0.2	0.9	20	0.2	0.8	10	0.1	0.6
5	10	0.1	1.0	20	0.2	1.0	40	0.4	1.0
∑	100	1.0		100	1.0		100	1.0	

Ergänzend zu einem typischen Wert werden daher Kennwerte (Verteilungsparameter) benötigt, um das Ausmaß der Unterschiedlichkeit der Realisierungen einer Verteilung zu erfassen. Solche Maße für die Unterschiedlichkeit oder Streuung einer Verteilung werden als ***Streuungsmaße*** bezeichnet.

Ein einfaches und naheliegendes Maß zur Erfassung der Unterschiedlichkeit der Realisierungen ist die ***Spannweite*** *(engl.* ***Range****)*, die als Abstand (Differenz) zwischen größter und kleinster Realisierung einer Verteilung definiert ist:

$$R = x_{(n)} - x_{(1)}$$

Spannweite

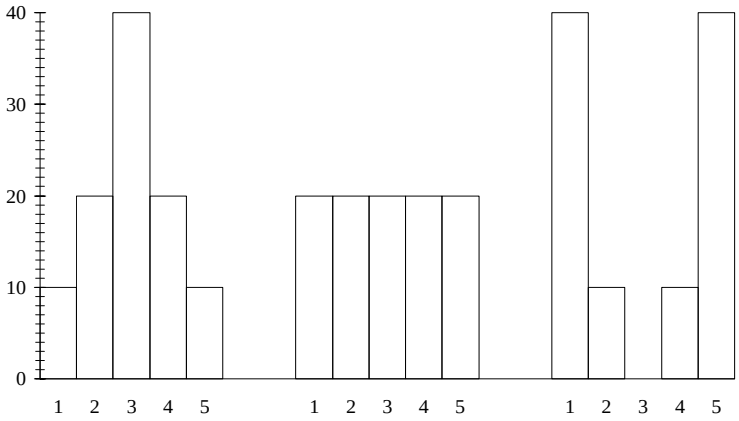


$R_W = 5 - 1 = 4$ $R_X = 5 - 1 = 4$ $R_Y = 5 - 1 = 4$

Die drei Verteilungen W, X und Y weisen trotz unterschiedlicher Verteilungen alle die gleiche Spannweite von 4 auf.

Tatsächlich ist die Spannweite nicht sehr aussagekräftig, da sie keinerlei Informationen darüber berücksichtigt, wie die Realisierungen innerhalb des Wertebereichs einer Verteilung streuen. Darüber hinaus ist die Spannweite bei Stichprobendaten sehr empfindlich gegenüber Ausreißerwerten.

(Mittlerer) Quartilabstand



$IQR_W=2$ $IQR_X=2$ $IQR_Y=4$

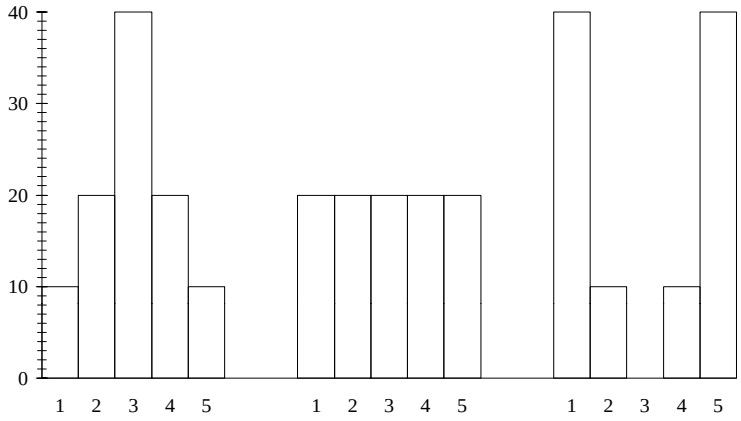
Anstelle des kleinsten und größten Wertes können auch die Abstände von Quantilen betrachtet werden. So wird recht häufig der ***Quartilabstand*** *(engl.* ***interquartil range****)* berechnet, das ist die Differenz des dritten vom ersten Quartil:

$$IQR = Q_{0.75} - Q_{0.25}$$

	W			X			Y		
Wert	n_k	p_k	cp_k	n_k	p_k	cp_k	n_k	p_k	cp_k
1	10	0.1	0.1	20	0.2	0.2	40	0.4	0.4
2	20	0.2	0.3	20	0.2	0.4	10	0.1	0.5
3	40	0.4	0.7	20	0.2	0.6	0	0.0	0.5
4	20	0.2	0.9	20	0.2	0.8	10	0.1	0.6
5	10	0.1	1.0	20	0.2	1.0	40	0.4	1.0
$\sum$	100	1.0		100	1.0		100	1.0	

Da bei den beiden Verteilungen W und X das 25%-Quantil 2 und das 75%-Quantil 4 sind, ist bei beiden Verteilung der Quartilabstand 2.
Die u-förmige Verteilung weist dagegen einen doppelt so großen Quartilabstand von 4 auf.

Mittlerer Quartilabstand



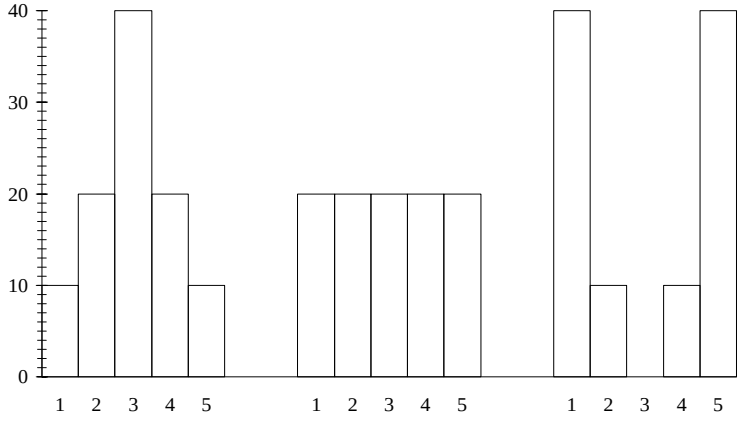
$mIQR_W=1$ $mIQR_X=1$ $mIQR_Y=2$

Der Quartilanstand wird in Box-Plots zur Festlegung der Boxlänge herangezogen wird, wobei allerding die Quartilwerte nach einer Formel berechnet werden, die von den empirischen Quantilwerten $Q_{25\%}$ und $Q_{75\%}$ abweicht.

Anstelle des Quartilabstands wird bisweilen auch der ***mittlere Quartilabstand*** berechnet, der die Hälfte des Quartilabstands ist:

$$mIQR = \frac{Q_{0.75} - Q_{0.25}}{2}$$

Durchschnittliche absolute Abweichung



$AD_W=0.8$ $AD_X=1.2$ $AD_Y=1.8$

Ein Maß, dass alle Realisationen einer Verteilung berücksichtigt und mit zunehmender Unterschiedlichkeit größere Werte aufweist, ist die ***durchschnittliche absolute Abweichung*** *(engl.* ***absolute deviation****)*, also der Mittelwert der vorbezeichenbereinigten Differenzen aller Realisierungen vom Mittelwert:

$$AD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Durchschnittliche absolute Abweichung

W	n_k	p_k	cp_k	$p_k \cdot w_k$	$p_k \cdot \lvert w_k - 3 \rvert$
1	10	0.1	0.1	0.1	0.2
2	20	0.2	0.3	0.4	0.2
3	40	0.4	0.7	1.2	0.0
4	20	0.2	0.9	0.8	0.2
5	10	0.1	1.0	0.5	0.2
∑	100	1.0		3.0	0.8

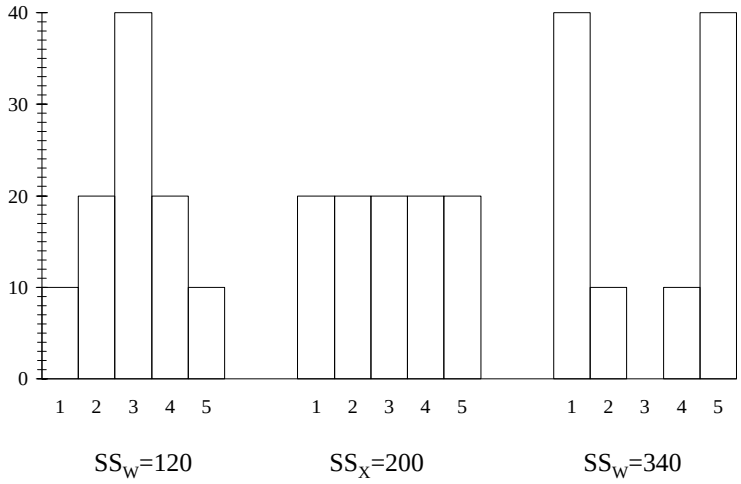
X	n_k	p_k	cp_k	$p_k \cdot w_k$	$p_k \cdot \lvert w_k - 3 \rvert$
1	20	0.2	0.2	0.2	0.4
2	20	0.2	0.4	0.4	0.2
3	20	0.2	0.6	0.6	0.0
4	20	0.2	0.8	0.8	0.2
5	20	0.2	1.0	1.0	0.4
∑	100	1.0		3.0	1.2

Y	n_k	p_k	cp_k	$p_k \cdot w_k$	$p_k \cdot \lvert w_k - 3 \rvert$
1	40	0.4	0.4	0.4	0.8
2	10	0.1	0.5	0.2	0.1
3	0	0.0	0.5	0.0	0.0
4	10	0.1	0.6	0.4	0.1
5	40	0.4	1.0	2.0	0.8
∑	100	1.0		3.0	1.8

Die durchschnittlichen Abweichungen unterscheiden sich bei den drei Verteilungen deutlich: Am geringsten ist der Wert von AD mit 0.8 bei der unimodalen Verteilung, größer bei der Gleichverteilung und am größten bei der u-förmigen Verteilung.

Bei symmetrischen Verteilungen, bei denen arithmetisches Mittel und Median zusammenfallen, ist die durchschnittliche absolute Abweichung gleichzeitig ein ***definiertes Minimum***, da es dann das durchschnittliche Minimum der absoluten Abweichungen ist.
Weichen Median und Mittelwert voneinander ab, ist die Minimaleigenschaft allerdings nicht garantiert. Daher wird anstelle des Maßes bisweilen auch die durchschnittliche absolute Abweichung vom Median als alternatives Streuungsmaß berechnet.

Variation



SS_W=120 SS_X=200 SS_W=340

Für den Mittelwert gilt, dass die Summe der quadrierten Abweichungen vom Mittelwert ein absoluter Minimalwert ist. Der resultierende Minimalwert wird ***Variation*** oder ***mittelwertbereinigte Quadratsumme*** *(engl:* ***sum of squares****, abgekürzt:* ***SS_X****)* genannt und ist Ausgangsgröße für die in der Statistik am häufigsten verwendeten Streuungsmaße:

$$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Variation

W	n_k	p_k	cp_k	$p_k \cdot w_k$	$n_k \cdot (w_k - 3)^2$
1	10	0.1	0.1	0.1	40
2	20	0.2	0.3	0.4	20
3	40	0.4	0.7	1.2	0
4	20	0.2	0.9	0.8	20
5	10	0.1	1.0	0.5	40
Σ	100	1.0		3.0	120

X	n_k	p_k	cp_k	$p_k \cdot w_k$	$n_k \cdot (w_k - 3)^2$
1	20	0.2	0.2	0.2	80
2	20	0.2	0.4	0.4	20
3	20	0.2	0.6	0.6	0
4	20	0.2	0.8	0.8	20
5	20	0.2	1.0	1.0	80
Σ	100	1.0		3.0	200

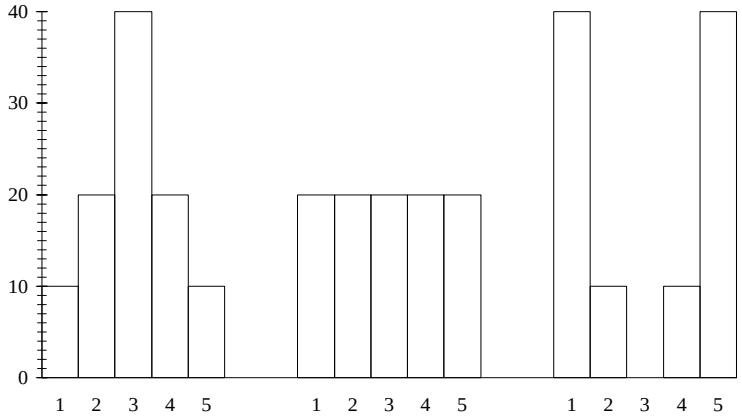
Y	n_k	p_k	cp_k	$p_k \cdot w_k$	$n_k \cdot (w_k - 3)^2$
1	40	0.4	0.4	0.4	160
2	10	0.1	0.5	0.2	10
3	0	0.0	0.5	0.0	0
4	10	0.1	0.6	0.4	10
5	40	0.4	1.0	2.0	160
Σ	100	1.0		3.0	340

Bei der Berechnung der Variation ist es nicht unbedingt nötig, für jede Ausprägung zunächst die Differenz zum Mittelwert zu berechnen und diese dann zu quadrieren. Umformungen zeigen, dass die Variation auch als Differenz zwischen der Summe der quadrierten Realisierungen und dem Produkt aus Fallzahl und Quadrat des Mittelwerts berechnet werden kann:

$$SS_X = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 + \bar{x}^2 - 2 \cdot x_i \cdot \bar{x}) = \sum_{i=1}^{n} x_i^2 + n \cdot \bar{x}^2 - 2 \cdot \bar{x} \cdot \sum_{i=1}^{n} x_i$$

$$= \sum_{i=1}^{n} x_i^2 + n \cdot \bar{x}^2 - 2 \cdot \bar{x} \cdot (n \cdot \bar{x}) = \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2$$

Stichprobenvarianz



$SS_W = 120$; $s_W^2 = 1.2$ $SS_X = 200$; $s_X^2 = 2.0$ $SS_Y = 340$; $s_Y^2 = 3.4$

Wird die Variation durch die Fallzahl geteilt, ergibt sich die ***(Stichproben-) Varianz***, das ist die durchschnittliche quadrierte Abweichung vom Mittelwert:

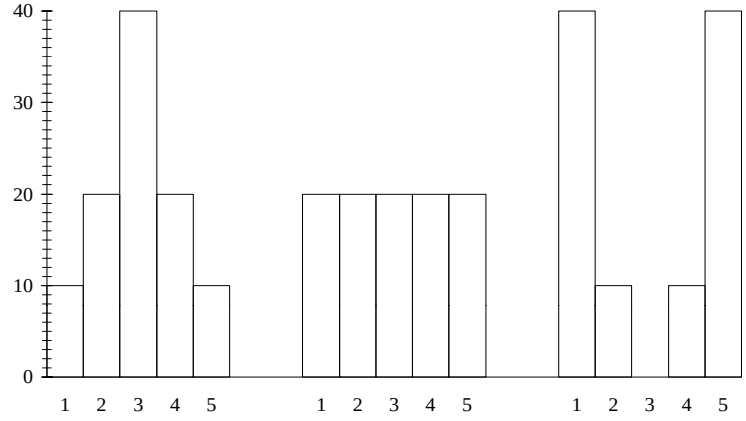
$$s_X^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{SS_X}{n}$$

Hinweis:

In Statistikprogrammen, Taschenrechnern und manchen Statistikbüchern wird bei der Berechnung der Varianz die Variation i.a. nicht durch die Fallzahl n, sondern durch die ***Zahl der Freiheitsgrade*** *n – 1 geteilt:* $\hat{\sigma}_X^2 = SS_X / (n-1)$.

Dieser Quotient ist eine Schätzung der Populationsvarianz auf der Basis von Stichprobendaten.

Standardabweichung

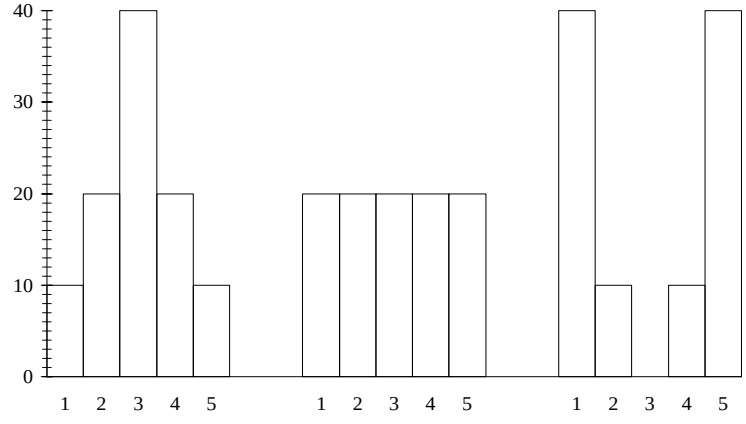


$s_W^2 = 1.2$; $s_W = 1.095$ $s_X^2 = 2.0$; $s_x = 1.414$ $s_Y^2 = 3.4$; $s_Y = 1.844$

Da die Einheit der Varianz das Quadrat der Einheit der betrachteten Verteilung ist, wird meistens die ***Standardabweichung*** *(engl:* ***standard deviation***) als Maß für die Streuung verwendet, die die positive Quadratwurzel aus der Varianz ist:

$$s_X = \left|\sqrt{s_X^2}\right| = \left|\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right| = \left|\sqrt{\frac{SS_X}{n}}\right|$$

Variantionskoeffizient



$s_W = 1.095$; $V_W = 0.365$ $s_x = 1.414$; $V_x = 0.471$ $s_Y = 1.844$; $V_Y = 0.615$

Sind die Ausprägungen einer Variable große Zahlen, dann ist oft auch die Standardabweichung hoch, sind die Ausprägungen kleinere Zahlen, gilt dies oft auch für die Standardabweichung.

Der ***Variationskoeffizient*** berücksichtigt dies, da er als Quotient der Standardabweichung geteilt durch das arithmetisches Mittel definiert ist:

$$V_X = \frac{s_x}{\bar{x}} = \frac{\sqrt{s_X^2}}{\bar{x}} = \frac{\sqrt{SS_X}}{\sqrt{n}\cdot\bar{x}}$$

Der Variationsindex ist eine einheitslose Größe und wird oft in Prozent angeben.

Variantionskoeffizient

Inaltlich sinnvoll ist der Variationskoeffizient nur bei Rationalskalenniveau und positivem Mittelwert. Letzteres ist der Fall, wenn es keine negativen Ausprägungen geben kann.

Eine mögliche Verallgemeinerung besteht daher darin, statt durch den Mittelwert durch die Abweichung des Mittelwerts von der kleinsten Realisation zu teilen:

$$V_X^* = \frac{s_x}{\bar{x} - x_{(1)}} = \frac{\sqrt{s_X^2}}{\bar{x} - x_{(1)}} = \frac{\sqrt{SS_X}}{\sqrt{n} \cdot \left(\bar{x} - x_{(1)}\right)}$$

Für die drei Verteilungen ergeben sich Werte von 0.548 für W, 0.707 für X und 0.922 für Y.

Wenn durch die Differenz zwischen Mittelwert und kleinstem Wert geteilt wird, ist es nicht sinnvoll, den resultierenden Wert als Prozentwert darzustellen.

Welches Streuungsmaßes sollte verwendet werden?

So wie der Mittelwert aufgrund seines Informationsgehalts bei metrischen Variablen das bevorzugte Lagemaß ist, ist es bei den Streuungsmaßen die Varianz, bzw. Standardabweichung. Der Variationskoeffizient wird in der Ökonometrie sehr viel genutzt, da wichtige ökonomische Größen Ratioskalen mit nichtnegativen Wertebereichen sind.
In der explorativen Datenanalyse wird wegen der stärkeren Unempfindlichkeit gegenüber Ausreißern auch häufiger die absolute Abweichung und der Quartilabstand herangezogen.

Berechnung von Variation und Standardabweichung für eine Variable der Datematrix

Die Berechnung der Variation und abgeleiteter Streuungsmaße für eine Variable der Datenmatrix ist am einfachsten, wenn neben der Spalte der Datenmatrix, die die Realisierungen der Variable enthält, eine zweite Spalte mit den quadrierten Realisierungen generiert wird und jeweils die Summe der gültigen Spaltenwerte berechnet wird.

Fallnr. IS	Alter (X)	Alter^2 (X^2)
1943	65	4225
1960	48	2304
1957	51	2601
1939	69	4761
missing	missing	missing
1956	52	2704
1970	38	1444
1920	88	7744
1956	52	2704
1966	42	1764
Summe	505	30251
$\frac{\text{Summe}}{n_{valid} = 9}$	56.111	3361.222

Die Variation berechnet sich aus diesen Summen nach:

$$SS_X = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Für die Beispieldaten ergibt sich:

$$SS_X = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{9} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2$$

$$= 30251 - \frac{505^2}{9} = 30251 - 9 \cdot 56.111^2 = 58587.111$$

Berechnung von Variation und Standardabweichung für eine Variable der Datematrix

Für die Varianz, die Standardabweichung und den Variationskoeffizienten folgt dann:

Fallnr. IS	Alter (X)	Alter² (X²)
1943	65	4225
1960	48	2304
1957	51	2601
1939	69	4761
missing	missing	missing
1956	52	2704
1970	38	1444
1920	88	7744
1956	52	2704
1966	42	1764
Summe	505	30251
$\frac{\text{Summe}}{n_{valid}=9}$	56.111	3361.222

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 = \frac{\sum_{i=1}^{n}x_i^2}{n} - \left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2$$

Für die Beispieldaten ergibt sich:

$$s_X^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}x_i^2 - \bar{x}^2 = \frac{\sum_{i=1}^{n}x_i^2}{n} - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n^2} = \frac{SS_X}{n}$$

$$= 3361.222 - 56.111^2 = \frac{30251}{9} - \frac{505^2}{9^2} = \frac{58587.111}{9} = 212.765$$

Dann folgt für die Standardabweichung:

$$s_X = \sqrt{s_X^2} \qquad \textit{im Beispiel}: s_X = \sqrt{212.7654321} = 14.586$$

Der Variationskoeffizienten brechnet sich schließlich nach:

$$V_X = s_X / \bar{x}$$

$$\textit{im Beispiel}: \; V_X = 14.58648 / 46.11111 = 0.2600 = 26.0\%$$

Rechenschema für Häufigkeitstabellen

Die Berechnung der Variation und abgeleiteter Größen gilt auch für Berechnungen auf der Basis von Häufigkeitstabellen, wenn jeweils anstelle von Summen und Mittelwerten von (quadrierten) Realisierungen mit den über die (relativen) Häufgkeiten gewichteten Summen bzw. Mitteln der Ausprägungen und deren Quadrate gerechnet wird.

$$SS_X = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{k=1}^{K}n_k\cdot\left(x_i - \bar{x}\right)^2 = \sum_{k=1}^{K}n_k\cdot x_i^2 - n\cdot\bar{x}^2 = \sum_{k=1}^{K}n_k\cdot x_i^2 - \frac{\left(\sum_{k=1}^{K}n_k\cdot x_k\right)^2}{n}$$

$$s_X^2 = \frac{1}{n}\cdot\sum_{k=1}^{K}n_k\cdot x_k^2 - \bar{x}^2 = \frac{\sum_{k=1}^{K}n_k\cdot x_k^2}{n} - \frac{\left(\sum_{k=1}^{K}n_k\cdot x_k\right)^2}{n^2} = \frac{SS_X}{n}$$

Die kann am Beispiel der univariaten Verteilung W demonstriert werden:

W	n_k	p_k	cp_k	$n_k \cdot w_k$	$n_k \cdot (w_k)^2$
1	10	0.1	0.1	10	10
2	20	0.2	0.3	40	80
3	40	0.4	0.7	120	360
4	20	0.2	0.9	80	320
5	10	0.1	1.0	50	250
Σ	100	1.0		300	1020

$$SS_W = 1020 - \frac{300^2}{100} = 120$$

$$s_W^2 = \frac{1020}{100} - \frac{300^2}{100^2} = \frac{120}{100} = 1.2$$

Rechenschema für Häufigkeitstabellen

Bei Gewichtung mit relativen Häufigkeiten gilt:

$$s_X^2 = \sum_{k=1}^{K} p_k \cdot (x_k - \bar{x}) = \sum_{k=1}^{K} p_k \cdot x_k^2 - \bar{x}^2 = \sum_{k=1}^{K} p_k \cdot x_k^2 - \left(\sum_{k=1}^{K} p_k \cdot x_k \right)^2 \quad \text{und} \quad SS_X = n \cdot s_X^2$$

W	n_k	p_k	cp_k	$p_k \cdot w_k$	$p_k \cdot (w_k)^2$
1	10	0.1	0.1	0.10	0.10
2	20	0.2	0.3	0.40	0.80
3	40	0.4	0.7	1.20	3.60
4	20	0.2	0.9	0 80	3.20
5	10	0.1	1.0	0.50	2.50
∑	100	1.0		3.00	10.20

$$s_W^2 = 10.20 - 3^2 = 1.2$$

$$SS_W = 100 \cdot 1.2 = 120$$

Standardabweichung und Variationskoeffizient werden anschließend aus der Varianz oder der Variation berechnet:

im Beispiel: $s_X = \sqrt{1.2} = 1.095$ *und :* $V_X = \frac{\sqrt{1.2}}{3} = 0.365 = 36.5\%$

Bei gruppierten Daten werden in den Schemata statt der Ausprägungen x_k die Klassenmitten m_k eingesetzt. Da dabei die Streuung der Realisierungen innerhalb der Klassen nicht berücksichtigt wird, führt die Berechnung der Variation auf der Basis gruppierter Werte tendenziell zur Unterschätzung der tatsächlichen Variation der Realisierungen der Verteilung.

Tschebyscheffsche Ungleichung

Die Bedeutung der Varianz bzw. Standardabweichung als Streuungsmaß liegt auch darin, das gezeigt werden kann, das bei allen metrischen Verteilungen das Intervall von ±k Standardabweichungen um den Mittelwert mindestens $1 - 1/k^2$ aller Realisierungen einer Verteilung enthält bzw. umgekehrt maximal $1/k^2$ aller Realisierungen außerhalb dieses Intervalls liegen.

Diese Eigenschaft ist als ***Tschebyscheffsche Ungleichung*** bekannt:

$$p(\bar{x} - k \cdot s_x \le X \le \bar{x} + k \cdot s_x) \ge 1 - \frac{1}{k^2}$$

Für die 9 gültigen Realisierungen der Altersvereilung aus dem Beispielfragebogen ergab sich ein Mittelwert von 56.11 Jahren und eine Standardabweichung von 14.586 Jahren. Aus der Tschebyscheffschen Ungeichung folgt dann:

k=1.5: Mindestens $1–1/1.5^2$ = 55.5% aller Fälle sind zwischen 34 (≈56.11 – 1.5·14.586) und 78 Jahre alt;

k=2.0: Mindestens $1-1/2^2$ = 75.0% aller Fälle sind zwischen 27 (≈56.11 – 2·14.586) und 85 Jahre alt.

Tatsächlich liegen sowohl 88.9% aller Realisierungen im ersten wie im zweiten Intervall.

Nur bei sehr extremen Verteilungen wird die Tschebyscheffsche Ungleichung exakt erfüllt. In der Regel liegen deutlich mehr als $1/1–k^2$ Fälle im Intervall ±k Standardabweichungen um den Mittelwert. Interessant ist gleichwohl, dass allein auf der Basis von Mittelwert und Varianz für beliebige Verteilungen Aussagen über Anteilsintervalle möglich sind.

Streuungsmaße für Verteilungen nominalskalierter Variablen

Alle vorgestellten Maße gehen streng genommen von metrischen Daten aus, da Abstandsinformationen für ihre Berechnung verwendet werden. Bei nominalskalierten Variablen ist daher keines dieser Maße anwendbar.

Als Substitut für die Spannweite kann bei Nominalskalenniveau die Anzahl der Ausprägungen als Streuungsmaß verwendet werden, die aber wie die Spannweite nicht sehr informativ ist. Informationshaltiger sind Maße, die die Häufigkeiten der Ausprägungen berücksichtigen. In der Literatur findet sich als ein Streuungsmaß für nominalskalierte Variablen die Differenz der quadrierten und aufsummierten relativen Häufigkeiten von 1:

$$1 - \sum_{k=1}^{K} p_k^2$$

Nur bei einer Konstante ist der Wert null. Der Wert ist um so größer, aber stets kleiner 1, je mehr Ausprägungen eine Verteilung hat und je gleichmäßiger die Realisierungen über die Ausprägungen streuen. Das jeweils bei Gleichverteilung erreichte Maximum beträgt 1 – 1/K. Der ***Index qualitativer Variation*** ist die auf den Wertebereich 0 bis 1 standardisierte Form dieses Maßes:

$$IQV = \frac{K}{K-1} \cdot \left(1 - \sum_{k=1}^{K} p_k^2\right)$$

Streuungsmaße für Verteilungen nominalskalierter Variablen

Ein weiteres Streuungsmaß für nominalskalierte Variablen ist die ***Devianz***. Es lässt sich zeigen, dass die Devianz proportional mit dem ***Informationsgehalt*** einer Verteilung ansteigt. Der Informationsgehalt wird dabei an der durchschnittlichen Anzahl der zu beantwortenden binären richtig/falsch-Fragen gemessen, die notwendig sind, um die Ausprägung einer zufällig herausgegriffenen Realisierung zu ermitteln.

Für die mathematische Statistik ist die Devianz vor allem aufgrund ihres Beziehung zur sogenannten Likelihood-Funktion für Zufallsstichproben nominalskalierter Variablen von Bedeutung.

Die ***absolute Devianz*** D_X einer Variablen X ist ein Analogon zur Variation bei metrischen Verteilungen. Sie berechnet sich nach:

$$D_X = -2\sum_{k=1}^{K} n_k \cdot \ln\left(\frac{n_k}{n}\right) = -2\sum_{k=1}^{K} n_k \cdot \ln(p_k)$$

Bei der Berechnung der ***relativen Devianz*** d_X erfolgt die Gewichtung der Logarithmen über die relativen Häufigkeiten:

$$d_X = -2\sum_{k=1}^{K} p_k \cdot \ln(p_k) = \frac{D_x}{n}$$

Auch für die Devianz gilt, dass der Maximalwert, der bei K Ausprägungen bei der relativen Devianz –2·ln(K) beträgt, jeweils bei einer Gleichverteilung erreicht wird und der Minimalwert null, wenn nur eine einzige Ausprägung (mit allen Fällen) besetzt wird.

Streuungsmaße für Verteilungen nominalskalierter Variablen

Als Beispiel soll die ***Devianz*** der (gültigen Ausprägungen) der Konfession der Befragten aus dem Allbus 2006 berechnet werden:

Konfession	Häudigkeit	Prozent	gültige Prozent	$-2\cdot n_k\cdot\ln(p_k)$	$-2\cdot p_k\cdot\ln(p_k)$
evang. ohne Freikirchen	1169	34.2	34.3	2501.718	0.73404
evang. Freikirche	89	2.6	2.6	649.639	0.18978
Römisch-katholisch	1042	30.5	30.6	2467.811	0.72471
andere christl. Religion	76	2.2	2.2	580.140	0.16794
nicht-christliche Religion	138	4.0	4.1	881.595	0.26192
ohne Religionszugehör.	890	26.0	26.2	2384.151	0.70185
verweigert	10	0.3	--		
keine Angabe	8	0.2	--		
Total:	3422	100.0	100.0	9465.054	2.78024

Gültige Fälle 3404 Fehlende Fälle: 18
(Allbuss 2006 Ost-West-gewichtet)

$$D_X = -2\sum_{k=1}^{K} n_k \cdot \ln(p_k) = 9465.054 \; ; \; d_X = -2\sum_{k=1}^{K} p_k \cdot \ln(p_k) = 2.780 = \frac{D_x}{n} = \frac{9465.054}{3304}$$

Der ***Index qualitativer Variation*** berechnet sich bei dieser Verteilung nach $(1 - .343^2 - .026^2 - .305^2 - .022^2 - .041^2 - .262^2) \cdot 6/(6-1) = 0.861$.

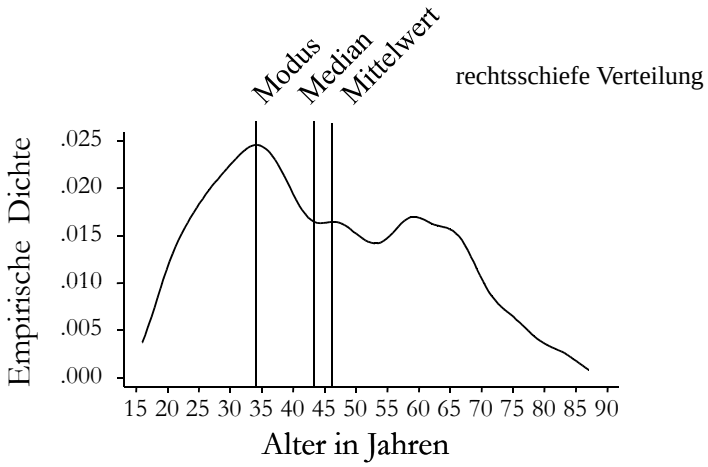
Streuung von ordinalen Variablen

Für ordinale Variablen finden sich in der Literatur bislang keine speziellen Streuungsmaße. Bisweilen wird der Quartilabstand verwendet. Da er aber durch die Berechnung über eine Differenz metrische Informationen nutzt, kann er ähnlich wie der Median bei gleicher Fallzahl nur im Sinne einer Aussage über Ränge oder Ausprägungsanzahlen genutzt werden. Ein Quartilabstand von 2 weist also darauf hin, dass die drei (=2+1) - bezogen auf den Median - mittleren Kategorien mindestens 50% aller Fälle enthalten.

Wenn die Ordinalität einer Variable ignoriert wird und wie bei einer metrischen Variablen der Mittelwert als Lagemaß verwendet wird, liegt es nahe, auch die Varianz bzw. Standardabweichung als Streuungsmaß zu verwenden.

Da die Messnievaus hierarchisch geordnet sind, kann stets auf Kenngrößen für ein niedrigeres Messniveau zurückgegriffen werden.
In diesem Sinne kann etwa die Devianz oder der Index qualitativer Variation auch als Streuungsmaß für Verteilungen ordinaler Variablen berücksichtigt werden. Allerdings geht dabei möglicherweise relevante Information verloren. So ist es bereits bei Ordinalskalenniveau sinnvoll, von u-förmigen Verteilungen zu sprechen und diesen eine höhere Streuung zuzuweisen als Gleichverteilungen. Auf Nominalskalenniveau ist die Anordnung dagegen irrelevant, weswegen u-förmige Verteilungen wie alle Verteilungen, bei denen mindestens eine Ausprägung stärker besetzt ist als andere auf einer Nominalskala eine geringere Streuung aufweisen als Gleichverteilungen.

Weitere Verteilungkenngrößen



Neben der Streuung ist oft auch von Interesse, ob eine Verteilung (annähernd) symmetrisch oder schief verteilt ist.

Hinweise auf die ***Schiefe*** (engl.: *skewness*) einer Verteilung gibt der Vergleich von Modus, Median und Mittelwert:

- Für unimodale, symmetrische Verteilungen gilt: Modus = Median = Mittelwert,
  bei mehrgipfligen, symmetrischen Verteilungen gilt: Median = Mittelwert;
- bei einer eindeutig rechtsschiefen Verteilung gilt: Modus < Median < Mittelwert;
- bei einer eindeutig linksschiefen Verteilung gilt: Modus > Median > Mittelwert.

Momente

Wie bei der Streuung sind auch zur Erfassung der Schiefe Kennwerte vorgeschlagen worden. Ausgangspunkt solcher Kennwerte sind die ***Momente*** einer Verteilung, das sind die Mittelwerte von Potenzen der Realisierungen. Ähnlich wie die Gesamtheit aller Quantile eine Verteilung charakterisieren, gilt dies auch für die Menge aller Momente.

Dabei wird zwischen Rohmomenten oder Momenten um den Usprung und zentralen Momenten unterschieden.
Das ***k-te (Roh-) Moment*** ist der Durchschnittswert über alle mit k potenzierten Realisierungen einer Verteilung:

$$\text{k-tes Rohmoment } m'_k = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^k$$

Werden vor der Potenzierung die Differenzen vom ersten Moment berechnet, ergeben sich die ***zentralen Momente***:

$$\text{k-tes zentrales Moment } m_k = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - m'_1 \right)^k$$

Schiefekoeffizient

Aus der Definition der Momente folgt, dass das erste (Roh-) Moment einer Verteilung ihr Mittelwert ist.

Die Varianz einer Verteilung ist die Differenz des quadrierten ersten Rohmoments vom zweiten Rohmoment und gleichzeitig das zweite zentrale Moment.
Da die Summe und damit auch der Mittelwert aller Abweichungen vom Mittelwert Null sind, folgt, dass das erste zentrale Moment einer Verteilung immer Null ist.

Als Kenngröße der Schiefe wird im ***Schiefekoeffizient*** einer Verteilung das dritte Moments um den Erwartungswert einer Verteilung durch die dritte Potenz der Standardabweichung dividiert:

$$\text{Schiefekoeff.} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{s_X^3} = \frac{m_3}{\left(\sqrt{m_2}\right)^3}$$

Bei symmetrischen Verteilungen ist der Schiefekoeffizient Null.
Ist eine Verteilung eher rechtsschief, dann ist der Schiefekoeffizient positiv, da die Werte am rechten Rand weiter vom Mittelwert entfernt sind als Werte am linken Rand und durch Potenzierung mit 3 Realisierungen am rechten Rand große positive und am linken Rand große negative Beiträge zum Schiefekoeffizienten bilden.
Umgekehrt ist der Schiefekoeffizient bei eher linksschiefen Verteilungen negativ.

Steilheit oder Kurtosis

In analoger Weise wie bei der Schiefe ist zur Erfassung der ***Wölbung*** oder ***Steilheit*** die Kurtosis als Quotient des vierten Moments um den Erwartungswert geteilt durch die vierte Potenz der Standardabweichung (bzw. das Quadrat der Varianz) definiert, wobei von diesem Wert i.a. noch die Zahl 3 abgezogen wird:

$$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{\left(s_X^2\right)^2} - 3 = \frac{m_4}{\left(m_2\right)^2} - 3$$

Die Potenzierung mit 4 führt dazu, dass der Wert der Kurtosis um so höher wird, je mehr Realisierungen es gibt, die weit vom Erwartungswert entfernt sind. je höher der Wert ist, desto flacher oder gar u-förmig ist eine Verteilung. Umgekehr ist bei unimodalen Verteilungen der Wert um so kleiner, je steiler die Verteilungen in der Nähe des Mittelwerts ansteigt.
Die abzuziehende Zahl 3 führt dazu, dass die Kurtosis einer sogeannten ***Normalverteilung*** genau null ist. Negative Werte weisen also auf eine Verteilung hin, die steiler ist als eine Normalverteilung, während positive Werte auf eine weniger steil ansteigende Verteilung hinweisen.

Da die Berechnung von Momenten metrisches Messniveau voraussetzt, beziehen sich auch die Maße zur Erfassung der Schiefe und der Steilheit auf metrische Variablen.

Lerneinheit 7:
Lineartranformationen und Zusamenfassungen von Subgruppen

In vielen Situationen interessieren anstelle der ursprünglichen Variablen Transformationen dieser Variablen.

Ein Beispiel ist die Analyse der Altersverteilung aus dem Beispielfragebogen, eine Variable, die im ursprünglichen Fragebogen nicht enthalten war, sondern aus dem erfragten Geburtsjahr berechnnet wurde.
Das Alter berechnet sich nach der Formel:

Alter = aktuelles Jahr – Geburtsjahr.

Viele solcher Transformationen lassen sich durch mathematische Gleichungen darstellen.
Die allgemeine Form einer solchen Gleichung lautet:

$$Y = g(X)$$

wobei X die ursprüngliche Variable und Y die transformierte Variable bezeichnet und g(...) eine beliebige mathematische Funktion ist.

Bei der Berechnung des Alters aus dem Geburtsjahr hat g(...) folgende Form:

$$g(x) =: a - 1 \cdot x$$

Im Beispiel wurde Mittelwert und Varianz des Alters aus den transformierten Geburtsjahren berechnet. Es stellt sich die Frage, ob es auch möglich ist, Mittelwert und Varianz des Alters bei Kenntnis nur von Mittelwert und Varianz der ursprünglichen Variable Geburtsjahr zu berechnen.

Lineartransformationen

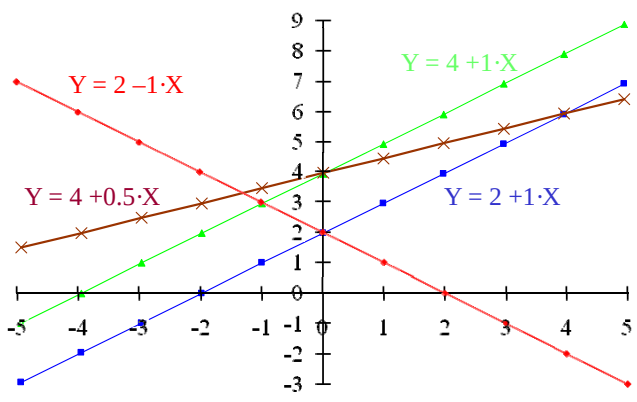
Wenn die Gleichung Y=g(X) eine lineare Gleichung ist, ist es tatsächlich sehr einfach, aus dem Mittelwert und der Varianz bzw. Variation oder Standardabweichung der Ursprungsvariable X die entsprechende Kennwerte der transformierten Variable Y zu berechnen. Man bezeichnet eine solche Transformation auch als ***Lineartransformation***.

In Lineartransformationen wird Y aus X erzeugt durch eine lineare Gleichung der Form:

$$Y = a + b \cdot X$$

In der Gleichung stehen „a“ und „b“ für zwei Zahlen, die ***Koeffizienten*** der linearen Gleichung.

Die Berechnung des Alters aus dem Geburtsjahr ist eine Linearfunktion, bei der „a“ das betrachtete Jahr und „b“ die Zahl –1 ist.



Lineare Gleichungen lassen sich in einem Koordinatensystem als Graden einzeichnen.
Die ***Konstante a*** gibt dabei den Wert von Y an, wenn X=0. Grafisch ist das der Schnittpunkt der Geraden mit der senkrechten Y-Achse.
Das ***Gewicht b*** gibt die Steigung der Geraden an. Inmmer, wenn der Wert von X um +1 Einheit ansteigt, verändert sich der Wert von Y um b Einheiten.

Mittelwert und Variation von Lineartransformationen

Bei der Anwendung der Lineartransformation wird jede Realisation durch die lineare Gleichung transformiert:

$$y_i = a + b \cdot x_i \quad \text{für } i = 1,2,...,n$$

Für den Mittelwert von Y gilt dann:

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = \frac{1}{n} \cdot \sum_{i=1}^{n} (a + b \cdot x_i) = \frac{1}{n} \cdot \sum_{i=1}^{n} a + \frac{1}{n} \cdot \sum_{i=1}^{n} b \cdot x_i = \frac{1}{n} \cdot n \cdot a + \frac{b}{n} \sum_{i=1}^{n} x_i = a + b \cdot \bar{x}$$

Der Mittelwert der linear transformierten Vaiable Y ergibt sich also durch Anwendung der Lineartransformation auf den Mittelwert von X.

Für die Variation von Y gilt dann:

$$SS_Y = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \left((a + b \cdot x_i) - (a + b \cdot \bar{x}) \right)^2 = \sum_{i=1}^{n} \left(b \cdot (x_i - \bar{x}) \right)^2 = \sum_{i=1}^{n} b^2 \cdot (x_i - \bar{x})^2 = b^2 \cdot SS_X$$

Die Variation der linear transformierten Variable Y ist also gleich dem Quadrat des Gewichtes b mal der Variation von X.
Der Wert der Konstanten a ist somit für die Variation von Y irrelevant.

Da sich die Varianz und die Standardabweichung aus der Variation berechnen lassen, folgt weiter:

$$s_Y^2 = \frac{SS_Y}{n} = \frac{b^2 \cdot SS_X}{n} = b^2 \cdot s_X^2 \text{ und } s_Y = \left| \sqrt{s_Y^2} \right| = \left| \sqrt{b^2 \cdot s_X^2} \right| = |b| \cdot s_X$$

Auswirkungen von Lineartransformationen

Als Beispiel soll Mittelwert, Variation und Varianz des Alters aus den enstprechenden Kennwerten des Geburtsjahrs berechnet werden.

	Geburts-jahr (X)	Geburts-jahr² (X²)
	1943	3775249
	1960	3851600
	1957	3829894
	1939	3759721
	missing	missing
	1956	3825936
	1970	3880900
	1920	3686400
	1956	3825936
	1966	3865156
Summe	17567	34290747
Summe / 9	1951.889	3910083

Lineartransformation
$Y = 2008 + (-1) \cdot X$
$\Rightarrow$

	Alter (Y)	Alter² (Y²)
	65	4225
	48	2304
	51	2601
	69	4761
	missing	missing
	52	2704
	38	1444
	88	7744
	52	2704
	42	1764
Summe	505	30251
Summe / 9	56.111	3361.222

$$\bar{y} = a + b \cdot \bar{x} = 2008 + (-1) \cdot 1951.889 = 56.111$$

$$SS_Y = b^2 \cdot SS_X = (-1)^2 \cdot SS_X = 1914.8889$$

$$s_Y^2 = b^2 \cdot s_X^2 = (-1)^2 \cdot 212.7654$$

$$s_Y = |b| \cdot s_X = |-1| \cdot 14.586$$

$\bar{x} = 17567/9 = 1951.889$
$SS_X = 34290747 - 17567^2/9 = 1914.8889$
$s_X^2 = 1914.8889/9 = 212.7654$
$s_X = 14.586$

$\bar{y} = 505/9 = 56.111$
$SS_Y = 30251 - 505^2/9 = 1914.8889$
$s_Y^2 = 1914.8889/9 = 212.7654$
$s_Y = 14.586$

Zentrierung und Normierung

Bei Kenntnis von Mittelwert und Varianz einer Verteilung ist es möglich, mit Hilfe einer Lineartransformation diese Verteilung so zu ändern, dass die resultierende Verteilungen einen vorgegebenen Mittelwert und/oder eine vorgegebene Varianz aufweist.
Genutzt wird dies in der Statistik u.a., um Variablen zu zentrieren, zu normieren oder zu standardisieren.

Zentrierung ist die Verschiebung des ursprünglichen Wertebereichs, so dass die ***zentrierte Variable*** einen Mittelwert von Null aufweist. Man spricht in diesem Zusammenhang auch davon, dass die Variable ***mittelwertfrei*** oder ***mittelwertbereinigt*** ist. Möglich ist dies durch eine Lineartransformation mit einem Gewicht von 1 und einer Konstanten, die das Negative des Mittelwerts der Ausgangsvariable ist:

$$Y = a + b \cdot X \text{ mit } a = -\bar{x} \text{ und } b = 1:$$

$$Y = -\bar{x} + 1 \cdot X = X - \bar{x} \Rightarrow \bar{y} = 0 \; ; \; SS_Y = SS_X \; ; \; s_Y^2 = s_X^2 \; ; \; s_Y = s_x$$

Dem gegenüber bedeutet ***Normierung*** eine Stauchung oder Streckung des ursprünglichen Wertebereichs. so dass die ***normierte Variable*** eine ***Varianz*** und eine Standardabweichung von ***1*** hat. Die Variation der normierten Variable ist dann gleich der Fallzahl. Erreicht wird dies durch eine Lineartransformation, bei der das Gewicht der Kehrwert aus der Standardabweichung der Ursprungsvariable ist:

$$Y = a + b \cdot X \text{ mit } a = 0 \text{ und } b = \frac{1}{s_X}:$$

$$Y = 0 + \frac{1}{s_X} \cdot X = \frac{X}{s_X} \Rightarrow \bar{y} = \frac{\bar{x}}{s_X} \; ; \; SS_Y = n \; ; \; s_Y^2 = 1 \; ; \; s_Y = 1$$

Normierung und Standardisierung

Die Division durch die Standardabweichung normiert die Streuung, verändert aber gleichzeitig den Mittelwert der Ausgangsvariable in Richtung 0. Soll der ursprüngliche Mittelwert unverändert bleiben kann dies durch geeignete Wahl der Konstante garantiert werden:

$$Y = a + b \cdot X \text{ mit } a = \left(1 - \frac{1}{s_X}\right) \cdot \bar{x} \text{ und } b = \frac{1}{s_X}:$$

$$Y = \left(1 - \frac{1}{s_X}\right) \cdot \bar{x} + \frac{1}{s_X} \cdot X = \frac{X - \bar{x}}{s_X} + \bar{x} \Rightarrow \bar{y} = \bar{x} \; ; \; SS_Y = n \; ; \; s_Y^2 = 1 \; ; \; s_Y = 1$$

Von ***Standardisierung*** spricht man schließlich, wenn gleichzeitig zentriert und normiert wird. Eine ***standardisierte Variable*** hat also einen ***Mittelwert*** von ***0*** und eine ***Varianz*** von ***1***. Erreicht wird dies durch eine lineare Transformation, bei der zunächst der Mittelwert abgezogen und anschließend durch die Standardabweichung geteilt wird. Da die standardisierten Realisierungen bisweilen auch als ***Z-Werte*** bezeichnet werden, wird die standardisierende Transformation auch als ***Z-Transformation*** bezeichnet. In diesem Sinne wird in der Formel Z statt Y für die transformierte Variable verwendet:

$$Z = a + b \cdot X \text{ mit } a = \frac{-\bar{x}}{s_X} \text{ und } b = \frac{1}{s_X}:$$

$$Z = \frac{-\bar{x}}{s_X} + \frac{1}{s_X} \cdot X = \frac{X - \bar{x}}{s_X} \Rightarrow \bar{z} = 0 \; ; \; SS_Z = n \; ; \; s_Z^2 = 1 \; ; \; s_Z = 1$$

Zentrierung, Normierung und Standardisierung

Als Beispiel sollen die 9 gültigen Fälle der Altersverteilung zentriert, normiert und standardisiert werden.

	Ausgangsvariable		→ zentriertes Alter: $Y = X - 56.111$		→ normiertes Alter: $Y = X / 14.586$		→ standardisiertes Alter: $Z = (X - 56.111)/14.586$	
	Alter (X)	Alter² (X^2)	Y	Y^2	Y	Y^2	Z	Z^2
	65	4225	8.889	79.012	4.456	19.859	0.609	0.371
	48	2304	–8.111	65.790	3.291	10.830	–0.556	0.309
	51	2601	–5.111	26.123	3.497	12.226	–0.350	0.123
	69	4761	12.889	166.123	4.731	22.378	0.884	0.781
	missing	missing	missing	missing	missing	missing	missing	missing
	52	2704	–4.111	16.901	3.565	12.710	–0.282	0.080
	38	1444	–18.111	328.012	2.605	6.787	–1.242	1.543
	88	7744	31.889	1016.901	6.033	36.399	2.186	4.779
	52	2704	–4.111	16.901	3.565	12.710	–0.282	0.080
	42	1764	–14.111	199.123	2.879	8.291	–0.967	0.935
Summe	505	30251	0.001	1914.886	34.622	142.190	0.000	9.001
Summe / 9	56.111	3361.222	0.000	212.765	3.847	212.765	0.000	1.000

$\bar{x} = 56.111$; $s_X = 14.586$

$\bar{y} = 0$; $SS_Y = 1914.886$; $s_Y^2 = 212.765$; $s_Y = 14.586$

$\bar{y} = 3.847$; $SS_Y = 9$; $s_Y^2 = 1$; $s_Y = 1$

$\bar{x} = 0$; $SS_Y = 9.00$; $s_Y^2 = 1$; $s_Y = 1$

Mittelwerte und Varianzen bei Zusammenfassungen von Subgruppen

In vielen Anwendungen werden Verteilungen betrachtet, bei denen sich die Population bzw. Gesamtstichprobe aus Subgruppen (Teilpopulationen oder Teilstichproben) zusammensetzt.

So besteht etwa die Stichprobe des Allbus aus zwei getrennt erhobenen Teilstichproben in den alten und in den neuen Bundesländern.

Ähnlich wie bei Linearkombinationen ist es auch bei Zusammenfassungen möglich, aus den Mittelwerten und Varianzen der Teilgruppen den Mittelwert und die Varianz der Gesamtheit aller Realisierungen aus allen Gruppen zu berechnen.

Als Beispiel wird eine Variable Y betrachtet, für die in einer Teilgruppe A n_A=6 Realisierungen und in einer Teilgruppe B n_B=4 Realisierungen vorliegen:

	Y in Gruppe A	Y^2 in Gruppe A
	3	9
	4	16
	1	1
	4	16
	3	9
	3	9
Summe	18	60
Summe / 6	3.0	10.0

$n_A = 6$

$\bar{y}_A = \frac{18}{6} = 3.0$

$SS_A = 60 - \frac{18^2}{6} = 6.0$

$s_A^2 = \frac{60}{6} - \frac{18^2}{6^2} = 1.0$

	Y in Gruppe B	Y^2 in Gruppe B
	1	1
	7	49
	2	4
	6	36
Summe	16	90
Summe / 4	4.0	22.5

$n_B = 4$

$\bar{y}_B = \frac{16}{4} = 4.0$

$SS_B = 90 - \frac{16^2}{4} = 26.0$

$s_B^2 = \frac{90}{4} - \frac{16^2}{4^2} = 6.5$

Mittelwerte und Varianzen bei Zusammenfassungen von Subgruppen

Der Gesamtmittelwert ist die Summe der Realisierungen in allen Gruppen geteilt durch die Gesamtfallzahl:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\left(\sum_{i=1}^{n_A} y_{i\in A} + \sum_{i=n_A+1}^{n_A+n_B} y_{i\in B}\right) = \frac{1}{n}\cdot\left(n_A\cdot\bar{y}_A + n_B\cdot\bar{y}_B\right) = \frac{n_A}{n}\cdot\bar{y}_A + \frac{n_B}{n}\cdot\bar{y}_B$$

Bei gegebenen Mittelwerten in den Gruppen ist der Gesamtmittelwert das mit der relativen Gruppenfallzahlen gewichtete Mittel der Gruppenmittelwerte.

Bezogen auf die Beispieldaten ergibt sich somit für den Gesamtmittelwert:

$$\bar{y} = \frac{6}{10}\cdot 3 + \frac{4}{10}\cdot 4 = 3.4$$

Die Gesamtvariation ist die Summe der quadrierten Abweichungen vom Gesamtmittelwert:

$$\begin{aligned} SS_Y &= \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 = \sum_{i=1}^{n_A}\left(y_{i\in A}-\bar{y}\right)^2 + \sum_{i=n_A+1}^{n_A+n_B}\left(y_{i\in B}-\bar{y}\right)^2 \\ &= \sum_{i=1}^{n_A}\left(\left(y_{i\in A}-\bar{y}_A\right)+\left(\bar{y}_A-\bar{y}\right)\right)^2 + \sum_{i=n_A+1}^{n_A+n_B}\left(\left(y_{i\in B}-\bar{y}_B\right)+\left(\bar{y}_B-\bar{y}\right)\right)^2 \\ &= \sum_{i=1}^{n_A}\left(y_{i\in A}-\bar{y}_A\right)^2 + \underbrace{\sum_{i=1}^{n_A}\left(\bar{y}_A-\bar{y}\right)^2}_{=n_A\cdot\left(\bar{y}_A-\bar{y}\right)^2} + \sum_{i=n_A+1}^{n_A+n_B}\left(y_{i\in B}-\bar{y}_B\right)^2 + \underbrace{\sum_{i=n_A+1}^{n_A+n_B}\left(\bar{y}_B-\bar{y}\right)^2}_{=n_B\cdot\left(\bar{y}_B-\bar{y}\right)^2} \end{aligned}$$

Mittelwerte und Varianzen bei Zusammenfassungen von Subgruppen

$$\begin{aligned} SS_Y &= \sum_{i=1}^{n_A}\left(y_{i\in A}-\bar{y}_A\right)^2 + \underbrace{\sum_{i=1}^{n_A}\left(\bar{y}_A-\bar{y}\right)^2}_{=n_A\cdot\left(\bar{y}_A-\bar{y}\right)^2} + \sum_{i=1}^{n_A}\left(y_{i\in A}-\bar{y}_A\right)^2 + \underbrace{\sum_{i=n_A+1}^{n_A+n_B}\left(\bar{y}_B-\bar{y}\right)^2}_{=n_B\cdot\left(\bar{y}_B-\bar{y}\right)^2} \\ &= \underbrace{\sum_{i=1}^{n_A}\left(y_{i\in A}-\bar{y}_A\right)^2 + \sum_{i=n_A+1}^{n_A+n_B}\left(y_{i\in B}-\bar{y}_B\right)^2}_{SS_W} + \underbrace{\left(n_A\cdot\left(\bar{y}_A-\bar{y}\right)^2 + n_B\cdot\left(\bar{y}_B-\bar{y}\right)^2\right)}_{SS_B} = SS_W + SS_B \end{aligned}$$

Die Gesamtvariation setzt sich also zusammen aus der Summe der Binnenvariationen innerhalb der einzelnen Gruppen (SS_W nach der engl. Bezeichnung „*variation within*“) plus der Variation der Gruppenmittelwerte um den Gesamtmittelwert (SS_B nach der engl. Bezeichnung „*variation between*“).

Dass diese Aufteilung der Gesamtvariation möglich ist, liegt daran, dass bei der Auflösung des Quadrates in den Formeln der beiden Teilvariationen jeweils ein Summand wegfällt, da für Variationen generell gilt:

$$\begin{aligned} &\sum_{i=1}^{n}\left(\left(x_i-\bar{x}\right)+\left(\bar{x}-a\right)\right)^2 = \sum_{i=1}^{n}\left(\left(x_i-\bar{x}\right)^2+\left(\bar{x}-a\right)^2+2\cdot\left(x_i-\bar{x}\right)\cdot\left(\bar{x}-a\right)\right) \\ &= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 + \sum_{i=1}^{n}\left(\bar{x}-a\right)^2 + 2\cdot\left(\bar{x}-a\right)\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right) = SS_X + n\cdot\left(\bar{x}-a\right)^2 + 2\cdot\left(\bar{x}-a\right)\cdot 0 \end{aligned}$$

Mittelwerte und Varianzen bei Zusammenfassungen von Subgruppen

Da die Varianz die Variation geteilt durch die Fallzahl ist, folgt schließlich:

$$\frac{SS_Y}{n} = \frac{\sum_{i=1}^{n_A}\left(y_{i\in A} - \bar{y}_A\right)^2}{n} + \frac{\sum_{i=n_B+1}^{n_A+n_B}\left(y_{i\in B} - \bar{y}_B\right)^2}{n} + \left(\frac{n_A\cdot\left(\bar{y}_A - \bar{y}\right)^2}{n} + \frac{n_B\cdot\left(\bar{y}_B - \bar{y}\right)^2}{n}\right)$$

$$= \frac{SS_A}{n} + \frac{SS_B}{n} + \left(\frac{n_A\cdot\left(\bar{y}_A - \bar{y}\right)^2}{n} + \frac{n_B\cdot\left(\bar{y}_B - \bar{y}\right)^2}{n}\right)$$

$$= \frac{n_A\cdot s_A^2}{n} + \frac{n_B\cdot s_B^2}{n} + \left(\frac{n_A\cdot\left(\bar{y}_A - \bar{y}\right)^2}{n} + \frac{n_B\cdot\left(\bar{y}_B - \bar{y}\right)^2}{n}\right)$$

$$= \left(\frac{n_A}{n}\cdot s_A^2 + \frac{n_B}{n}\cdot s_B^2\right) + \left(\frac{n_A}{n}\cdot\left(\bar{y}_A - \bar{y}\right)^2 + \frac{n_B}{n}\cdot\left(\bar{y}_B - \bar{y}\right)^2\right) = s_W^2 + s_B^2 = s_Y^2$$

Die Gesamtvarianz ist also die Summe aus der mit den relativen Gruppenfallzahlen gewichteten durchschnittlichen Varianz in den Gruppen plus der – ebenfalls mit den relativen Gruppenfallzahlen gewichteten – durchschnittlichen quadrierten Abweichung der Gruppenmittelwerte vom Gesamtmittelwert. Analog zur Variation wird der Summand als Binnenvarianz s^2_W (engl.: *variance within* oder *intraclass-variance*) und der zweite Summand als Varianz der Gruppenmittelwerte s^2_B (engl.: *variance beetween* oder *interclass-variance*) bezeichnet.

Mittelwerte und Varianzen bei Zusammenfassungen von Subgruppen

Angewendet auf die Beispieldaten ergibt sich:

$n_A = 6$ $\qquad$ $n_B = 4$

$\bar{y}_A = \frac{18}{6} = 3.0$ $\qquad$ $\bar{y}_B = \frac{16}{4} = 4.0$

$SS_A = 60 - \frac{18^2}{6} = 6.0$ $\qquad$ $SS_B = 90 - \frac{16^2}{4} = 26.0$

$s_A^2 = \frac{60}{6} - \frac{18^2}{6^2} = 1.0$ $\qquad$ $s_B^2 = \frac{90}{4} - \frac{16^2}{4^2} = 6.5$

$$\bar{y} = \frac{6}{10}\cdot 3 + \frac{4}{10}\cdot 4 = 3.4$$

$$s_Y^2 = \left(\frac{6}{10}\cdot 1 + \frac{4}{10}\cdot 6.5\right) + \left(\frac{6}{10}\cdot\left(3-3.4\right)^2 + \frac{4}{10}\cdot\left(4-3.4\right)^2\right)$$

$$= 3.2 + 0.24 = 3.44$$

	Y	Y^2
	3	9
	4	16
	1	1
	4	16
	3	9
	3	9
	1	1
	7	49
	2	4
	6	36
Summe	34	150
$\frac{\text{Summe}}{10}$	3.4	15.0

Die gleichen Resultate ergeben sich, wenn Mittelwert und Varianz nach Zusammenfassen der beiden Gruppen über die Gesamtmenge der 10 Realisierungen berechnet wird:

$$\bar{y} = \frac{34}{10} = 3.4$$

$$s_Y^2 = \frac{150}{10} - \frac{34^2}{10^2} = 3.44$$

Mittelwerte und Varianzen bei Zusammenfassungen von Subgruppen

Die Resultate können auf mehr als zwei Gruppen verallgemeinert werden.
Wenn es k = 1, 2, ..., K Gruppen mit insgesamt $n = n_1, n_2, ..., n_K$ Realisierungen in den Gruppen gibt, berechnen sich Mittelwert und Varianz der Zusammenfassung der Gruppen nach:

$$\bar{y} = \sum_{k=1}^{K} \frac{n_k}{n} \cdot \bar{y}_k \text{ und } s_Y^2 = \sum_{k=1}^{K} \frac{n_k}{n} \cdot s_k^2 + \sum_{k=1}^{K} \frac{n_k}{n} \cdot \left(\bar{y}_k - \bar{y}\right)^2 = s_w^2 + s_B^2$$

Da n_k/n gleich dem relativen Anteil p_k der Realisierungen in Teilgruppe an der Gesamtfallzahl ist, ist es für die Berechnung des Gesamtmittelwertes und der Gesamtvarianz hinreichend, die relativen Gruppengrößen zu kennen.

Aus der Formel ist sichtbar,

- dass die Gesamtvarianz nur dann genau gleich dem gewichteten Durchschnitt der Varianzen in den Teilgruppen ist, wenn alle Mittelwerte gleich groß und damit gleich dem Gesamtmittelwert sind, da dann der zweite Summand in der Formel für die Varianz null wird.
- Wenn die Unterschiede zwischen den Subgruppenmittelwerten relativ zu den Standardabweichungen gering sind, dann ist auch die Differenz zwischen dem gewichteten Durchschnitt der Varianzen in den Teilgruppen von der Gesamtvarianz gering.
  Daher wird bisweilen als Annäherung nur dieser Teil der Varianz berücksichtigt:

$$s_{pooled}^2 = \sum_{k=1}^{K} \frac{n_k}{n} \cdot s_k^2 \approx s_Y^2 \text{ wenn } \bar{y}_k - \bar{y} \approx 0 \text{ für alle k}$$

Mittelwerte und Varianzen bei Zusammenfassungen von Subgruppen

$$\bar{y} = \sum_{k=1}^{K} \frac{n_k}{n} \cdot \bar{y}_k \text{ und } s_Y^2 = \sum_{k=1}^{K} \frac{n_k}{n} \cdot s_k^2 + \sum_{k=1}^{K} \frac{n_k}{n} \cdot \left(\bar{y}_k - \bar{y}\right)^2 = s_w^2 + s_B^2$$

- Da bei der Berechnung der Varianz aus gruppierten Häufigkeitstabellen nur der rechte Summand s^2_B der Varianz zwischen den Klassen berücksichtigt wird, unterschätzt die so berechnete Varianz gruppierter Daten in der Regel die Gesamtvarianz.
  Die Nichtberücksichtigung der Varianz in den Gruppen s^2_W wird teilweise kompensiert (und evtl. sogar überkompensiert), da bei der Berechnung der Varianz aus gruppierten Häufigkeitstabellen nicht die tatsächlichen Klassenmittelwerte, sondern die Klassenmitten in die Berechnung der Varianz s^2_B zwischen den Klassen eingehen.

Gewichtungen

Bei Stichproben entsprechen die relativen Fallzahlen in den Teilgruppen oft nicht den relativen Anteilen der Teilpopulationen.

*So setzt sich die Allbus-Stichprobe 2006 aus 2299 (= 67.2%) Befragten aus den alten Ländern (Westen) und aus 1122 (= 32.8%) Befragten aus den neuen Ländern (Osten) zusammen. Dies ist eine Folge des Stichprobenplans, der einen überproportional hohen Anteil (**Oversampling**) aus den neuen Bundesländer vorsieht.*

Als Rechenbeispiel wird im folgenden angenommen, dass im obigen Beispiel für die Zusammenfassung von 2 Teilgruppen die Teilpopulation A π_A = 80% und die Teilpopulation B π_B = 20% der Fälle umfasst. Die Stichprobenanteile waren p_A = 60% und p_B = 40%, sodass die zweite Gruppe bezogen auf die Gesamtpopulation überrepräsentiert ist.

Um Aussagen über die Gesamtpopulation zu treffen, müssen dann bei der Zusammenfassung der Gruppen nicht die relativen Gruppengrößen, sondern die relativen Populationsanteile verwendet werden:

$$n_A = 6\ ;\ \pi_A = 0.8\ ;\ \bar{y}_A = \frac{18}{6} = 3.0 \qquad n_B = 4\ ;\ \pi_B = 0.2\ ;\ \bar{y}_B = \frac{16}{4} = 4.0$$

$$SS_A = 60 - \frac{18^2}{6} = 6.0\ ;\ s_A^2 = \frac{60}{6} - \frac{18^2}{6^2} = 1.0 \qquad SS_B = 90 - \frac{16^2}{4} = 26.0\ ;\ s_B^2 = \frac{90}{4} - \frac{16^2}{4^2} = 6.5$$

$$\bar{y} = 0.8 \cdot 3 + 0.2 \cdot 4 = 3.2$$

$$s_Y^2 = \left(0.8 \cdot 1 + 0.2 \cdot 6.5\right) + \left(0.8 \cdot \left(3 - 3.2\right)^2 + 0.2 \cdot \left(4 - 3.2\right)^2\right) = 2.1 + 0.16 = 2.26$$

Gewichtungen

Sollen bei disproportionalen Fallzahlen in den Teilgruppen Mittelwert und Varianz für die Gesamtheit nicht getrennt für die Gruppen berechnet und erst dann für die Gesamtheit zusammengefasst werden, müssen die Daten gewichtet werden.
Die Gewichte ergeben sich als Quotienten aus den Populationsanteilen geteilt durch die Gruppenfallzahlen.

Im Beispiel ergeben sich für die Realisierungen die folgenden Gewichte:

	Gruppe A	*Gruppe B*
Populationsanteil der Gruppe:	$\pi_A = 0.8$	$\pi_B = 0.2$
Relative Fallzahl in Gruppe:	$p_A = 6/10 = 0.6$	$p_B = 4/10 = 0.4.$
Gewichte:	$\pi_A / p_A = 0.8/0.6 = 1.333$	$\pi_A / p_A = 0.4/0.2 = 0.500.$

Formal bedeutet Gewichtung, dass jeder Fall als eine eigene Subgruppe betrachtet wird, weshalb die Varianz innerhalb der Gruppen s^2_B null ist. Wenn W die Gewichtungsvariable ist, berechnen sich dann Mittelwert, Variation und Varianz der gewichteten Daten nach:

$$\bar{y} = \frac{1}{N} \cdot \sum_{i=1}^{n} w_i \cdot y_i \text{ mit } N = \sum_{i=1}^{n} w_i \qquad SS_Y = \sum_{i=1}^{n} w_i \cdot \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} w_i \cdot y_i^2 - \frac{\left(\sum_{i=1}^{n} w_i \cdot y_i\right)^2}{\sum_{i=1}^{n} w_i}$$

$$s_Y^2 = \frac{1}{N} \cdot \sum_{i=1}^{n} w_i \cdot \left(y_i - \bar{y}\right)^2 = \frac{\sum_{i=1}^{n} w_i \cdot y_i^2}{\sum_{i=1}^{n} w_i} - \left(\frac{\sum_{i=1}^{n} w_i \cdot y_i}{\sum_{i=1}^{n} w_i}\right)^2$$

Gewichtungen

Für die Berechnung werden also die Werte bzw. quadrierten Werte jeder Realisierung vor dem Aufsummieren jeweils mit ihrem individuellen Gewicht multipliziert:

Y	Gewicht W	W·Y	W·Y²
3	1.333	4.000	12.000
4	1.333	5.333	21.333
1	1.333	1.333	1.333
4	1.333	5.333	21.333
3	1.333	4.000	12.000
3	1.333	4.000	12.000
1	0.500	0.500	0.500
7	0.500	3.500	24.500
2	0.500	1.000	2.000
6	0.500	3.000	18.000
Summe	10.00	32.00	125.00
$\frac{\text{Summe}}{\sum w_i}$	1.00	3.200	12.500

$$\bar{y} = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot y_i = \frac{32}{10} = 3.2$$

$$SS_Y = \left(\sum_{i=1}^{n} w_i \cdot y_i^2 \right) - \frac{\left(\sum_{i=1}^{n} w_i \cdot y_i \right)^2}{\sum_{i=1}^{n} w_i} = 125 - \frac{32^2}{10} = 22.6$$

$$s_Y^2 = \frac{\left(\sum_{i=1}^{n} w_i \cdot y_i^2 \right)}{\sum_{i=1}^{n} w_i} - \left(\frac{\sum_{i=1}^{n} w_i \cdot y_i}{\sum_{i=1}^{n} w_i} \right)^2 = \frac{125}{10} - \frac{32^2}{10^2} = 2.26$$

Das Beispiel zeigt, dass das Rechnen mit gewichteten Daten bei Mittelwerten, Variationen und Varianzen zu den gleichen Ergebnissen führt wie die getrennte Berechnung in den Teilgruppen und eine anschließende Zusammenfassung entsprechend den Populationsanteilen.

Gewichtungen

Berechnet man mit den gewichteten Daten jeweils für die Gruppen getrennt, ergeben sich die gleichen Ergebnisse wie bei ungewichteter Berechnung innerhalb jeder Gruppe:

Y	Gewicht W	W·Y	W·Y²
3	1.333	4.000	12.000
4	1.333	5.333	21.333
1	1.333	1.333	1.333
4	1.333	5.333	21.333
3	1.333	4.000	12.000
3	1.333	4.000	12.000
1	0.500	0.500	0.500
7	0.500	3.500	24.500
2	0.500	1.000	2.000
6	0.500	3.000	18.000
Summe	10.00	32.00	125.00
$\frac{\text{Summe}}{\sum w_i}$	1.00	3.200	12.500

$$\sum_{i=1}^{6} w_i = 8 \; ; \; \sum_{i=1}^{6} w_i \cdot y_i = 24 \; ; \; \sum_{i=1}^{6} w_i \cdot y_i^2 = 80$$

$$\Rightarrow \bar{y}_{\text{gewichtet } \in A} = \frac{24}{8} = 3 \; ; \; s^2_{\text{gewichtet } \in B} = \frac{80}{8} - 9 = 1$$

Y in Gruppe A	Y² in Gruppe A
3	9
4	16
1	1
4	16
3	9
3	9
∑=18	∑=60

$$\sum_{i=1}^{4} w_i = 2 \; ; \; \sum_{i=1}^{6} w_i \cdot y_i = 8 \; ; \; \sum_{i=1}^{6} w_i \cdot y_i^2 = 45$$

$$\Rightarrow \bar{y}_{\text{gewichtet } \in B} = \frac{8}{2} = 4 \; ; \; s^2_{\text{gewichtet } \in B} = \frac{45}{2} - 16 = 6.5$$

Y in Gruppe B	Y² in Gruppe B
1	1
7	49
2	4
6	36
∑=16	∑=90

$$n_A = 6 \; ; \; \pi_A = 0.8 \; ; \; \bar{y}_A = \frac{18}{6} = 3.0$$

$$SS_A = 60 - \frac{18^2}{6} = 6.0 \; ; \; s_A^2 = \frac{60}{6} - \frac{18^2}{6^2} = 1.0$$

$$n_B = 4 \; ; \; \pi_B = 0.2 \; ; \; \bar{y}_B = \frac{16}{4} = 4.0$$

$$SS_B = 90 - \frac{16^2}{4} = 26.0 \; ; \; s_B^2 = \frac{90}{4} - \frac{16^2}{4^2} = 6.5$$

Gewichtungen

Soll mit gewichteten Daten gerechnet werden, werden auch in Häufigkeitstabellen für die aboluten und relativen Häufigkeiten Gewichte benutzt:

Y	Gewicht W
3	1.333
4	1.333
1	1.333
4	1.333
3	1.333
3	1.333
1	0.500
7	0.500
2	0.500
6	0.500

Y gewichtet			kumul.
Wert	Häufigkeit	Anteil	Anteil
1	1.833	0.183	0.183
2	0.500	0.050	0.233
3	4.000	0.400	0.633
4	2.667	0.267	0.900
6	0.500	0.050	0.950
7	0.500	0.050	1.000
Total	10.000	1.000	

Y gewichtet u. gerundet			kumul.
Wert	Häufigkeit	Anteil	Anteil
1	2	0.167	0.167
2	1	0.083	0.250
3	4	0.333	0.583
4	3	0.250	0.833
6	1	0.167	0.917
7	1	0.167	1.000
Total	12	1.000	

Wenn bei den Häufigkeiten zu ganzen Zahlen gerundet wird, kann es bei den Fallzahlen zu Abweichungen kommen, die sich auch auf die Quantilberechnungen auswirken.

So enthält der Allbus 2006 3421 Fälle. Werden Häufigkeitstabellen auf der Basis von Ost-West-gewichteten Daten berechnet, weisen die resultierenden Tabellen bisweilen 3422 Fälle auf.

Lerneinheit 8: Elementare Wahrscheinlichkeitstheorie

Die ***Wahrscheinlichkeitstheorie*** beschäftigt sich mit Regelmäßigkeiten im Auftreten von unsicheren Ereignissen, also Ereignissen, die zwar möglich sind, aber nicht mit Sicherheit eintreten.

Ein Beispiel aus der Sozialforschung ist etwa die Frage, ob der Mittelwert von Stichprobendaten nahe beim Mittelwert in der Population liegt, aus der die Population kommt. Ist dies der Fall, kann der Stichprobenmittelwert als akzeptable Schätzung des Populationsmittelwerts herangezogen werden, was immer dann sinnvoll ist, wenn die exakte Bestimmung eines Populationsmittelwerts zu aufwendig oder sogar unmöglich ist.

Da der Induktionsschluss von einer Stichprobe auf eine Grundgesamtheit das unvermeidbare Risiko eines Fehlschlusses beinhaltet, wäre es wünschenswert, wenn es möglich wäre, dieses Risiko abzuschätzen und möglichst klein zu halten.

Mit Hilfe der Wahrscheinlichkeitstheorie kann gezeigt, dass dies möglich ist, wenn die Stichprobenziehung als ein sogenanntes Zufallsexperiment aufgefasst werden kann.

Die Vorstellung eines ***Zufallsexperiments*** ist der Ausgangspunkt der Wahrscheinlichkeitstheorie, wobei ein Zufallsexperiment als eine wiederholbare Versuchsanordnung aufgefasst wird, bei der nicht vorhersehbar ist, welches Resultat sich einstellen wird.

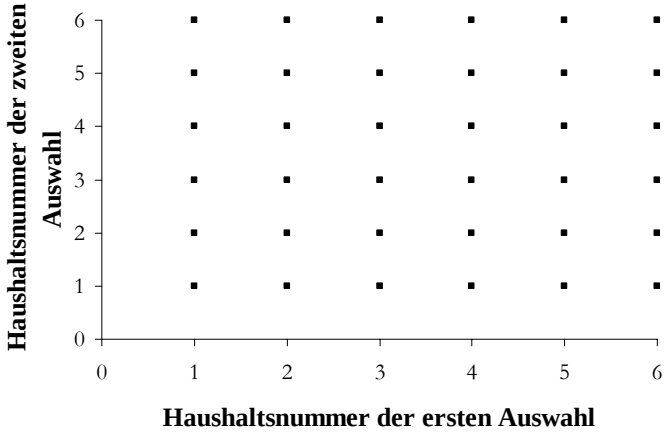
Historisch gesehen ist die Wahrscheinlichkeitstheorie u.a. im Kontext von Glücksspielen entstanden, die geradezu paradigmatisch für Zufallsexperimente sind. Ziel war es, die Wahrscheinlichkeit von Gewinnaussichten zu berechnen.

Zufallsexperiment

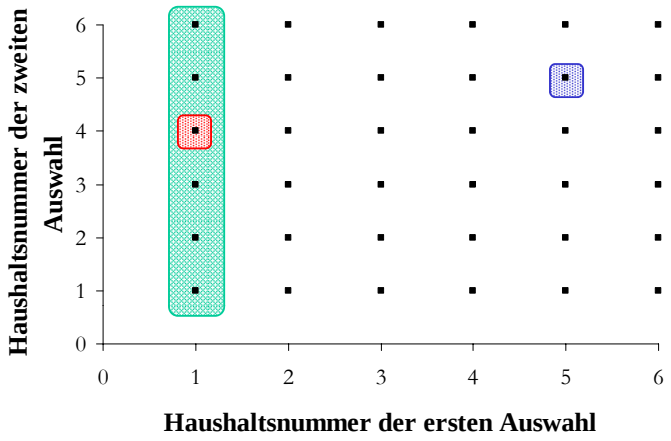
Ein ***Zufallsexperiment*** als definiert als eine Situation, die drei Eigenschaften aufweist:

1. Das Experiment ist (zumindest theoretisch) unter völlig identischen Bedingungen ***beliebig oft wiederholbar***.
2. Als Ergebnis des Experiments ***tritt*** genau ***ein Ereignis aus*** einer klar definierten ***Menge von möglichen Ereignissen*** auf.
3. ***Vor*** der ***Durchführung*** des Experiments ist ***unbekannt***, ***welches Ereignis*** auftreten wird.

Ein Beispiel für ein Zufallsexperiment kann das Werfen zweier Würfel sein, oder die zweimalige zufällige Auswahl eines Haushalts aus einer Menge von 6 Haushalten. Die Ergebnissen beider Zufallsexperimente lassen sich als Punkte in ein Koordinantensystem eintragen.



Zufallsexperimente und Wahrscheinlichkeit

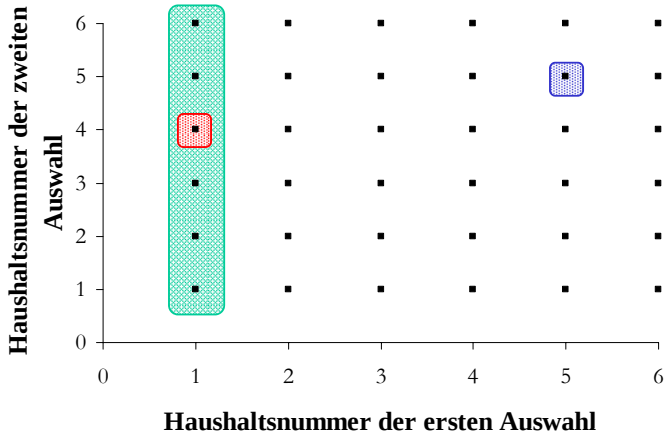


Die - im Beispiel 36 - möglichen Ergebnisse des Werfens zweier Würfel oder der wiederholten zufälligen Auswahl eines Haushalts sind im Koordinatensystem als Punkte mit den Positionen (x,y) ein getragen, wobei x die waagerechte und y die senkrechte Position angibt.

Wenn im Beispiel der X-Wert die Nummer des zuerst ausgewählten Haushalts angibt und der Y-Wert die Nummer des zweiten ausgewählten Haushalts, dann gilt:

- *Der Punkt (1,4) steht für das Ereignis, zunächst Haushalts Nr. 1 und dann Haushalts Nr. 4 auszuwählen; der Punkt (5,5) für das Ereingis, zweimal Haushalt 5 auszuwählen.*
- *Die grün gekennzeichnete Fläche mit den Punkten (1,1), (1,2), ..., (1,6) steht dann für das Ereignis, zuerst Haushalt Nr. 1 auszuwählen und dann irgendeinen der 6 Haushalte.*

Universum

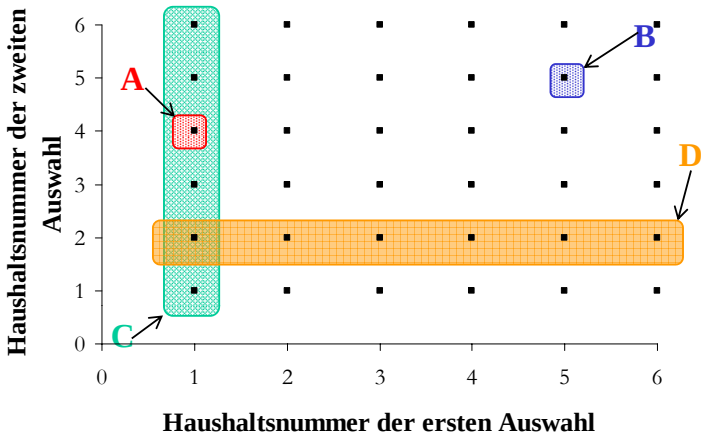


Die Gesamtmenge aller möglichen Egebnisse des Zufallsexperiments wird als ***Ereignisraum*** oder ***Universum*** des Zufallsexperiments bezeichnet und mathematisch durch den großen griechischen Buchstaben **Ω** (Omega) symbolisiert.

Alle denkbaren Ereignisse des Zufallsexperiments sind dann mengentheoretisch Teilmengen des Universums.

Durch diesen Bezug auf Ereignisse und Mengen von Ereignissen ist es möglich, auch Wahrscheinlichkeiten mengentheoretisch zu beschreiben und damit Wahrscheinlichkeitsaussagen über komplexe Ereignisse zu machen, die sich mengentheoretisch aus anderen Mengen zusammensetzen.

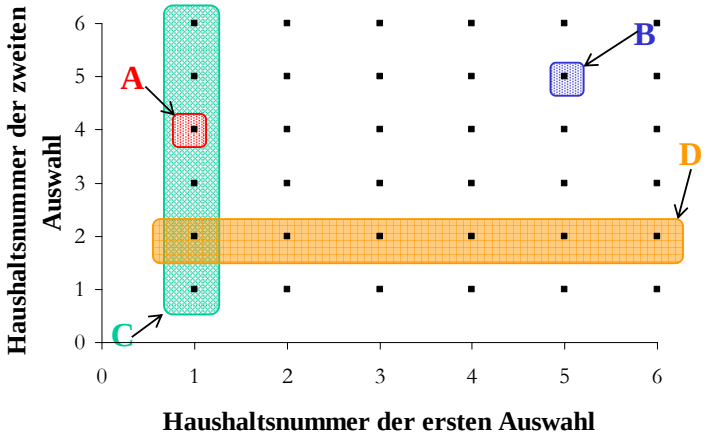
Zufallsexperimente und Wahrscheinlichkeit



Mengen und damit auch die Ereignisse eines Zufallsexperiments werden oft durch große Buchstaben bezeichnet:

- *A kann z.B. das Ereignis bezeichnen, zunächst Haushalt 1 und dannn Haushalt 4 auszuwählen,*
- *B das Ereignis, zweimal Haushalt Nr. 5 auszuwählen,*
- *C das (komplexe) Ereignis, zuerst Haushalt Nr. 1 auszuwählen,*
- *und D das Ereignis, zunächst irgendeinen Haushalt und dann Haushalt Nr. 2 zu befragen.*

Disjunkte Ereignisse



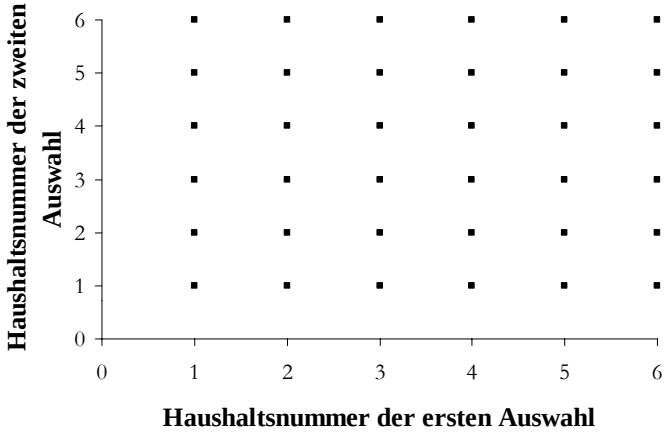
Zwei beliebige Ereignisse A und B heißen ***disjunkt***, wenn sie nicht gleichzeitig auftreten können.

- Die ***Schnittmenge*** **A∩B** der beiden Ereignisse, das ist ihr ***gemeinsames Auftreten*** (Ereignis A und Ereignis B), ist dann die sogenannte ***leere Menge*** {}.

  *In der Abbildung sind die Ereignisse A und B disjunkt: Es ist **unmöglich**, zunächst Haushalt 1 und dann Haushalt 4 auszuwählen und gleichzeitig in beiden Befragungen Haushalt 5 auszuwählen.*

  *Die Ereignisse C und D sind dagegen nicht disjunkt: Es ist **möglich**, zunächst Haushalt 1 (Ereignis C) und dann Haushalt 2 (Ereignis D) auszuwählen.*

Elementarereignisse

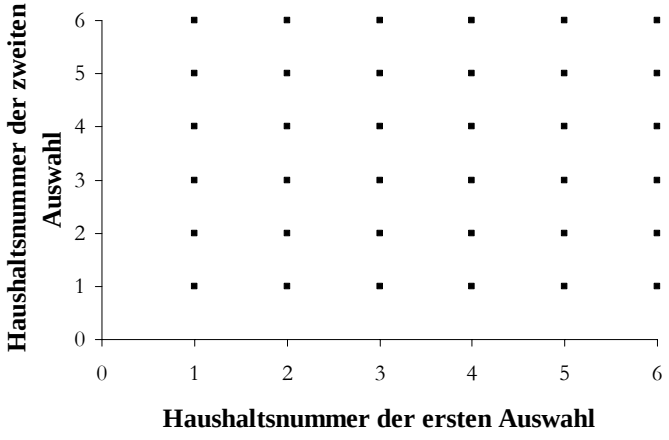


Eine ***exhaustive Zerlegung*** ist die vollständige Aufteilung eines Ereignisraums Ω in disjunkte Teilmengen.

Die Ereignisse einer exhaustiven Zerlegung heißen ***Elementarereignisse***, wenn diese Ereignisse nicht weiter in Teilereignisse zerlegt werden können. Elementarereignisse sind also die kleinstmöglichen Ergebnisse eines Zufallsexperiments.

In der Abbildung ist jeder Punkt ein Elementarereignis. Die insgesamt 36 Punkte ergeben eine exhaustive Zerlegung des Ereignisraums.

Apriori-Wahrscheinlichkeit

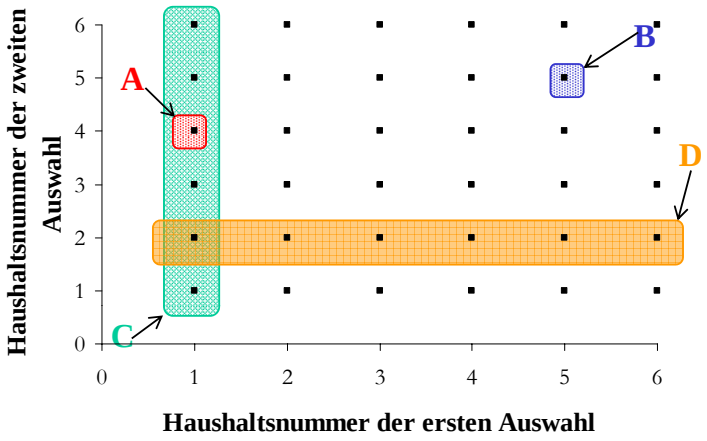


Es ist oft zu erwarten, dass jedes Elementarereignis mit gleicher Wahrscheinlichkeit auftritt. Wenn der Ereignisraum eines Zufallsexperiment zu n Elementarereignissen führt, dann hat somit jedes Elementarereignisses eine Warscheinlichkeit von 1/n.

Die einem Ereigniss ausschließlich aufgrund theoretischer Übelegungen zugesprochene Auftretenswahrscheinlichkeit wird als ***Apriori-Wahrscheinlichkeit*** bezeichnet.

Im Beispiel ist so die Apriori-Wahrscheinlichkeit jedes der 36 Elementarereignisse genau 1/36.

Berechnung von Wahrscheinlichkeiten



Nach der Zuordnung von Apriori-Wahrscheinlichkeiten zu den Elementarereignissen lassen sich mit mengentheoretischen Überlegungen die Wahrscheinlichkeiten aller Ereignisse eines Zufallsexperiments berechnen.

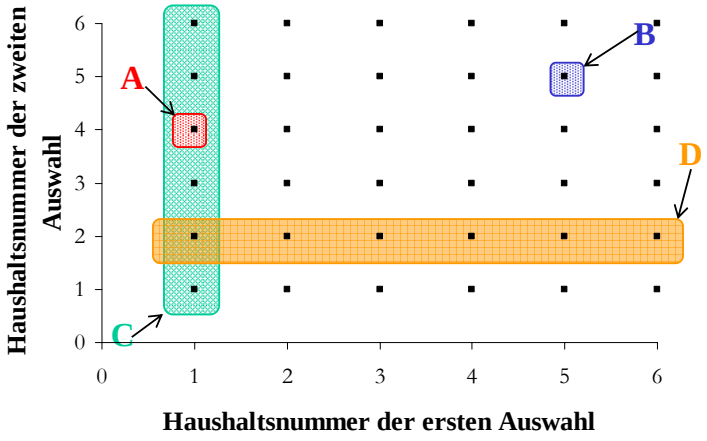
Entsprechend dieser Vorstellung beträgt die Wahrscheinlichkeit des Eignisses A wie auch des Ereignisses B jeweils 1/36: Pr(A) = Pr(B) = 1/36.

Die Wahrscheinlichkeit des Ereignisses C beträgt dann Pr(C)= 6/36 = 1/6.

Die gleiche Wahrscheinlichkeit von 1/6 hat auch das Ereignisses D: Pr(D) = 1/6.

Die Realisierungswahrscheinlichkeit eines Ereignisses X wird in diesem Skript durch Pr(X) symbolisiert („Pr“ als Abkürzung für die englische Bezeichnung „Probability“).

Wahrscheinlichkeit von Zusammenfassungen von Ereignissen



Es ist möglich, disjunkte wie nicht disjunkte Ereignisse zu einem komplexen Ereignis zusammenzufassen:

*Das komplexe Ereignis A **oder** B fasst die beiden disjunkten Ereignisse A und B zusammen,*

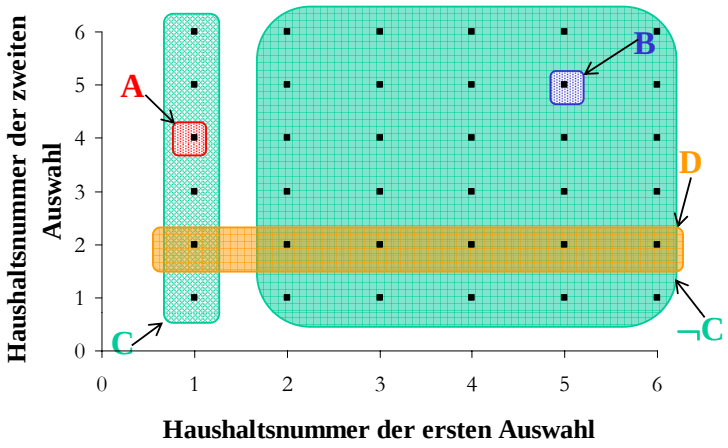
*das komplexe Ereignis C **oder** D die beiden nicht disjunkten Ereignisse C und D.*

Formal werden solche ***Zusammenfassungen*** oder Vereinigungen von Mengen durch das Symbol $\cup$ dargestellt: Die ***Vereinigungsmenge*** von A und B ist $A \cup B$.

Die Wahrscheinlichkeit der Vereinigungsmenge $A \cup B$ beträgt $Pr(A \cup B) = 2/36 = 1/18$,

die Wahrscheinlichkeit der Vereinigungsmenge $C \cup D$ beträgt $Pr(C \cup D) = 11/36$.

Komplemenäre Ereignisse



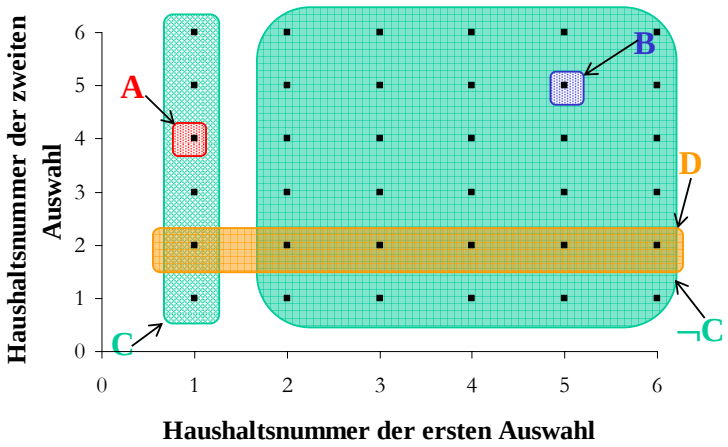
Zwei disjunkte Ereignisse heißen ***komplementär***, wenn ihre Vereinigungsmenge den gesamten Ereignisraum Ω umfasst.

Das zu einem Ereignis ***komplementäre Ereignis*** wird oft durch das Symbol ¬ („nicht“) dargestellt, da das Komplementärereignis das „Gegenteil“ eines Ereignisse ist.Das Ereignis **¬C** ist das Komplentärereignis zum Ereignis C.

Im Beispiel ist ¬C das Ereignis, in der ersten Befragung nicht Haushalt 1 zu befragen.
Die Wahrscheinlichkeit von ¬C beträgt 30/36 = 5/6.
Diese Wahrscheinlichkeit ergibt sich auch nach:

$$Pr(\neg C) = Pr(\Omega) - Pr(C) = 1 - 1/6 = 5/6.$$

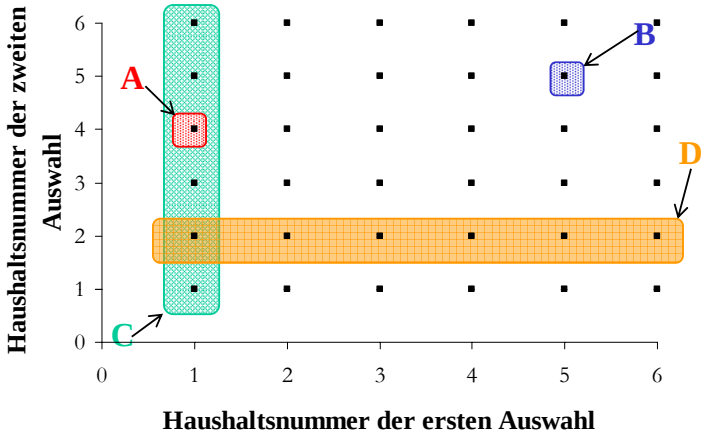
Klassischer Wahrscheinlichkeitsbegriff



Ähnlich zur Apriori-Wahrscheinlichkeit wurde ursprünglich die Wahrscheinlichkeit eines Ereignisses als Zahl der günstigen Möglichkeiten durch die Zahl der Möglichkeiten insgesamt berechnet. Diese Definition wird als ***klassische Wahrscheinlichkeit*** bezeichnet.

So beträgt die klassische Wahrscheinlichkeit, mit zwei Würfeln insgesamt eine Augenzahl von 18 (=2·6) zu erreichen, gleich 1/36, weil dieses Ereignis nur 1 mal unter 36 Möglichkeiten auftritt.

Axiomatische Wahrscheinlichkeitstheorie



Mit Hilfe der Mengentheorie ist es möglich, sämtliche Aussagen der Wahrscheinlichkeitstheorie von Ereignissen auf nur 3 Axiome der ***axiomatischen Wahrscheinlichkeitstheorie*** zurückzuführen:

A1: Die Wahrscheinlichkeit jedes beliebigen Ereignisses A ist eine reelle Zahl zwischen Null und Eins: $0 \leq Pr(A) \leq 1$

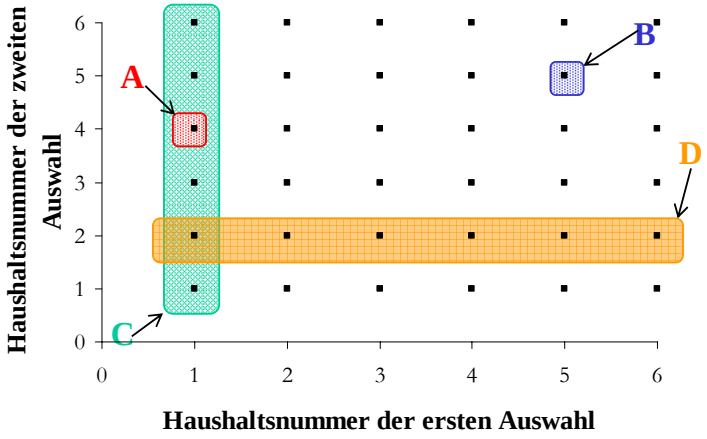
A2: Irgendein Ereignis des Ereignisraums (Universums) Ω muss auftreten. Die Wahrscheinlichkeit des Universums ist daher das sichere Ereignis mit der Wahrscheinlichkeit 1:

$$Pr(\Omega) = 1$$

A3: Die Wahrscheinlichkeit der Vereinigungsmenge zweier disjunkter Ereignisse A oder B ist die Summe der Wahrscheinlichkeit von A und der Wahrscheinlichkeit von B:

$$Pr(A \cup B) = Pr(A) + Pr(B) \text{ wenn } A \cap B = \{\}$$

Axiomatische Wahrscheinlichkeitstheorie



Im Beispiel:

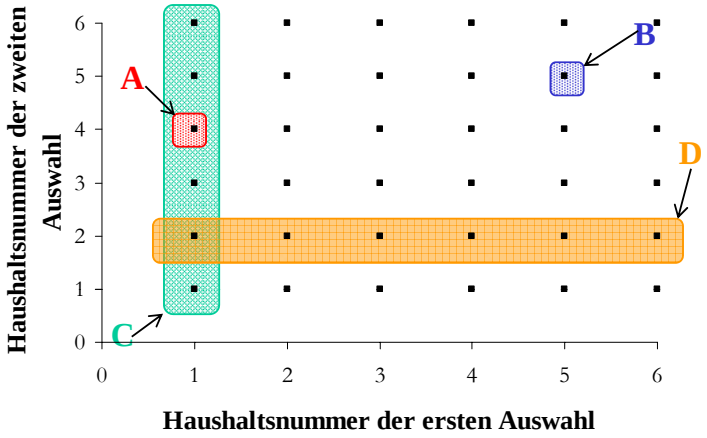
Pr (A) = Pr(B) = 1/36 ; Pr(D) = 6/36.

Dann folgt: *Pr(A∪B) = 1/36 + 1/36 = 2/36*

Pr(B) + Pr(C) = 1/36 + 6/36 = 7/36

Die Wahrscheinlichkeit, zweimal Haushalt 5 auszuwählen <u>oder</u> zunächst Haushalt 1, beträgt 7/36.

Additionstheorem



Aus den drei Axiomen der Wahrscheinlichkeitstheorie folgt für die Wahrscheinlichkeit der Vereinigungsmenge zweier beliebiger (disjunkter wie nicht disjunkter) Ereignisse A und B:

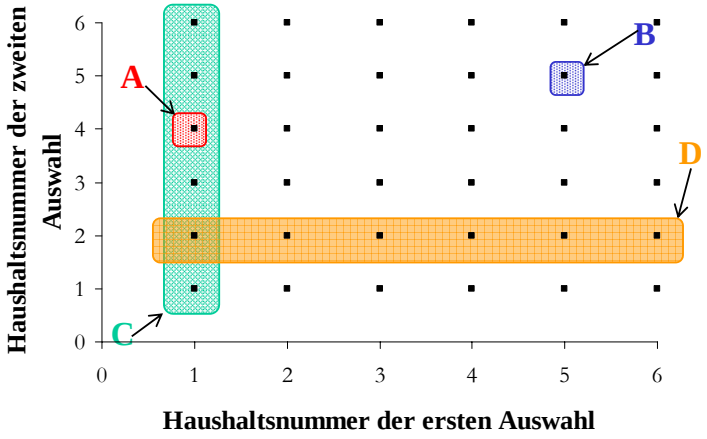
$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B).$$

Dieser Satz wird als ***Additionstheorem*** der Wahrscheinlichkeitstheorie bezeichnet.

Im Beispiel:

$$Pr(C \cup D) = Pr(C) + Pr(D) - PR(C \cap D) = 6/36 + 6/36 - 1/36 = 11/36$$
$$Pr(A \cup C) = Pr(A) + Pr(C) - PR(A \cap C) = 1/36 + 6/36 - 1/36 = 6/36$$
$$Pr(B \cup D) = Pr(B) + Pr(D) - PR(B \cap D) = 1/36 + 6/36 - 0/36 = 7/36$$

Wahrscheinlichkeit bedingter Ereignisse



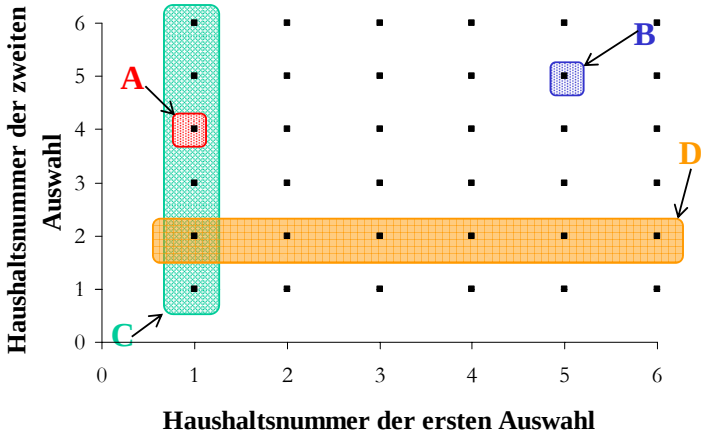
Oft ist man an der Wahrscheinlichkeit des Auftretens eines Ereignisses A unter der Bedingung interessiert, dass ein zweites Ereignis B auftritt. Das Ereignis B wird dann als ***bedingendes Ereignis*** bezeichnet, das Ereignis A als ***bedingtes Ereignis***.

Da das Auftreten des bedingendes Ereignisses B vorausgesetzt wird, reduziert sich der mögliche Ereignisraum für das bedingte Ereigniss A auf das Auftreten des Ereignises B. Die Wahrscheinlichkeit des bedingenden Ereignisses engt gewissermaßen den ursprünglichen Ereignisraum Ω auf das Ereignis B ein.

Die ***bedingte Wahrscheinlichkeit*** des Ereignisses ***A gegeben B*** ist daher die ***Wahrscheinlichkeit***, dass ***A und B gemeinsam*** auftreten, ***geteilt durch*** die ***Wahrscheinlichkeit***, dass ***B*** auftritt:

$$Pr(A|B) = Pr(A \cap B) / Pr(B).$$

Wahrscheinlichkeit bedingter Ereignisse



Wenn im Beispiel C das bedingende Ereignis ist, ergeben sich so folgende bedingte Wahrscheinlichkeiten für A, B und C:

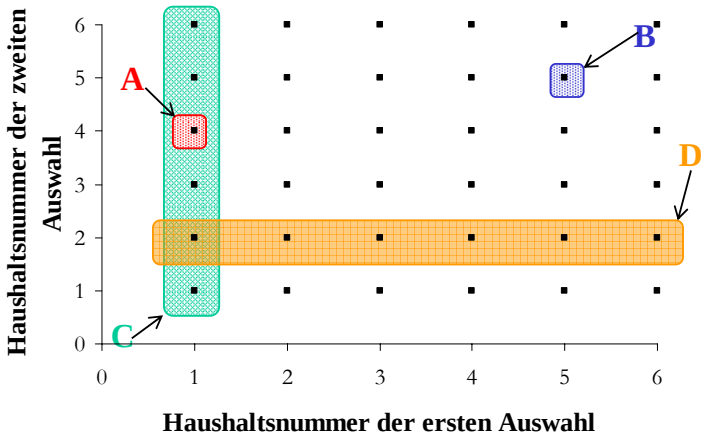
$Pr(A|C) = Pr(A \cap C) / Pr(C) = 1/36 / 6/36 = 1/6;$

$Pr(D|C) = Pr(D \cap C) / Pr(C) = 1/36 / 6/36 = 1/6;$

$Pr(B|C) = Pr(B \cap C) / Pr(C) = 0/36 / 6/36 = 0.$

Da durch die Wahrscheinlichkeit des bedingenden Ereignis geteilt wird, ist die bedingte Wahrscheinlichkeit nur für bedingende Ereignisse sinnvoll, die möglich sind, also eine Auftretenswahrscheinlichkeit größer Null haben.
Sind bedingendes und bedingtes Ereignis disjunkt, ist die bedingte Wahrscheinlichkeit Null.

Wahrscheinlichkeit bedingter Ereignisse



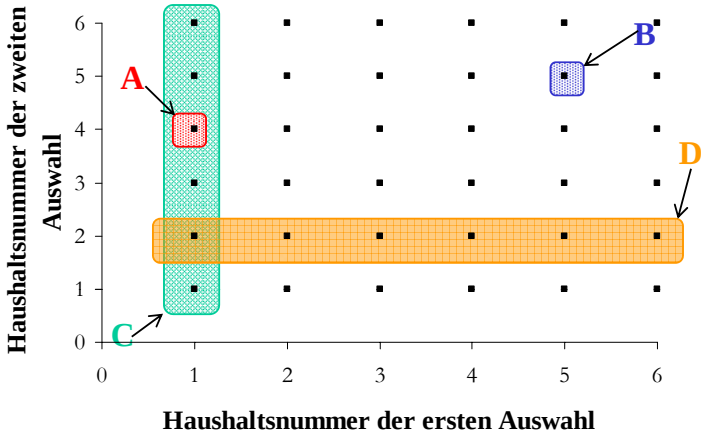
Bedingte Wahrscheinlichkeiten bilden die Grundlage der statistischen Zusammenhangsanalyse. Zu beachten ist, dass es sich zunächst um rein formale Aussagen handelt, ohne einen zeitlichen Bezug, wie er z.B. bei kausalen Beziehungen vorausgesetzt wird. Es ist daher auch möglich, die bedingte Wahrscheinlichkeit eines Ereignisses zu berechnen unter der Bedingung, dass ein zukünftiges Ereignis eintreten wird.

Im Beispiel:

$Pr(C|D) = Pr(C \cap D) / Pr(D) = 1/36 / 6/36 = 1/6.$

Die Wahrscheinlichkeit zunächst Haushalt 1 auszuwählen, wenn in der zweiten Auswahl Haushalt 2 erreicht werden wird, beträgt 1/6.

Statistische Unabhängigkeit

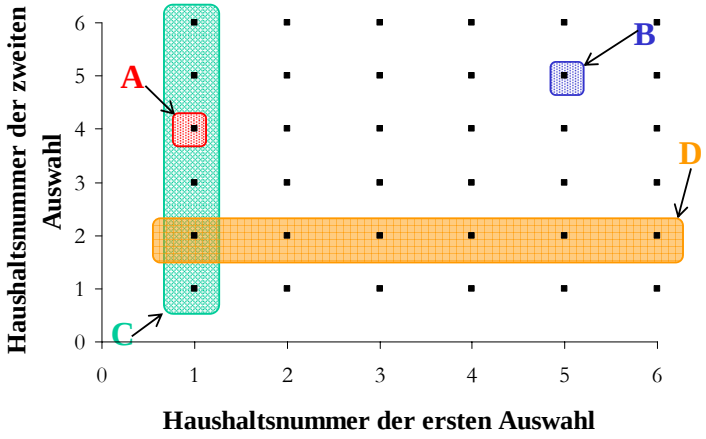


Über die bedingte Wahrscheinlichkeit wird die ***statistische Unabhängigkeit*** definiert:
Zwei Ereignisse A und B sind genau dann statistisch unabhängig voneinander, wenn die bedingte Wahrscheinlichkeit von A gegeben B gleich der (unbedingten) Wahrscheinlichkeit von A ist, bzw. wenn die bedingte Wahrscheinlichkeit von B gegeben A gleich der (unbedingten) Wahrscheinlichkeit von B ist:

$$Pr(A|B) = Pr(A) \text{ und } Pr(B|A) = Pr(B).$$

Im Beispiel:
Da Pr(D|C) = 1/6 gleich Pr(D) = 1/6, sind C und D statistisch unabhängig voneinander.

Multiplikationstheorem



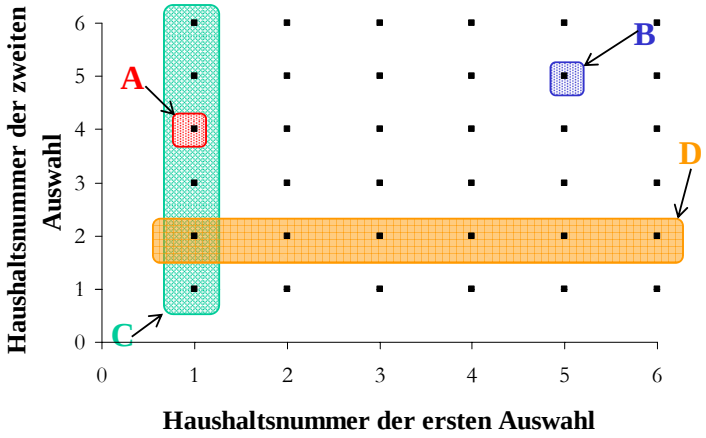
Aus der Umformung der Wahrscheinlichkeit eines bedingten Ereignisses folgt, dass die ***Wahrscheinlichkeit des gleichzeitigen Auftretens*** zweier Ereignisse gleich dem ***Produkt der bedingten Wahrscheinlichkeit des einen Ereignisse mal der unbedingten Wahrscheinlichkeit des bedingenden Ereignisses*** ist:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A).$$

Dieser Zusammenhang ist als ***Multiplikationstheorem*** bekannt.

Im Beispiel:
$Pr(A \cap C) = Pr(A|C) \cdot Pr(C) = 1/6 \cdot 1/6 = 1/36;$
$Pr(B \cap C) = Pr(B|C) \cdot Pr(C) = 0/6 \cdot 1/6 = 0.$

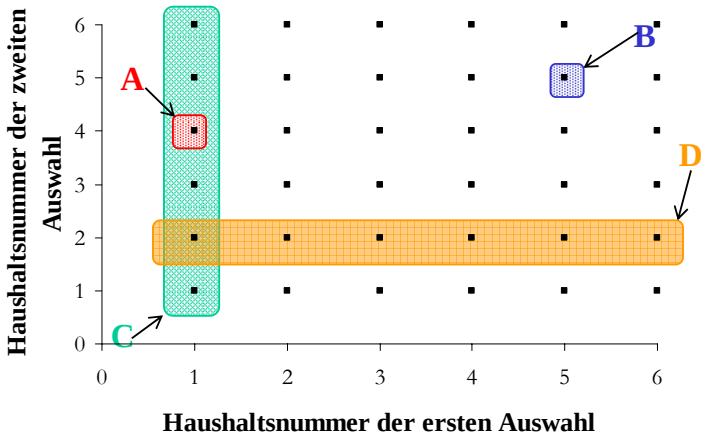
Multiplikationstheorem



Bei statistischer Unabhängigkeit ist das gemeinsame (gleichzeitige) Auftreten zweier Ereignisse gleich dem Produkt der beiden Auftretenswahrscheinlichkeiten.

$Pr(C \cap D) = Pr(C|D) \cdot Pr(D) = Pr(D|C) \cdot Pr(C) = Pr(C) \cdot Pr(D) = 1/6 \cdot 1/6 = 1/36$
Also sind C und D statistisch unabhängig:

Theorem von Bayes



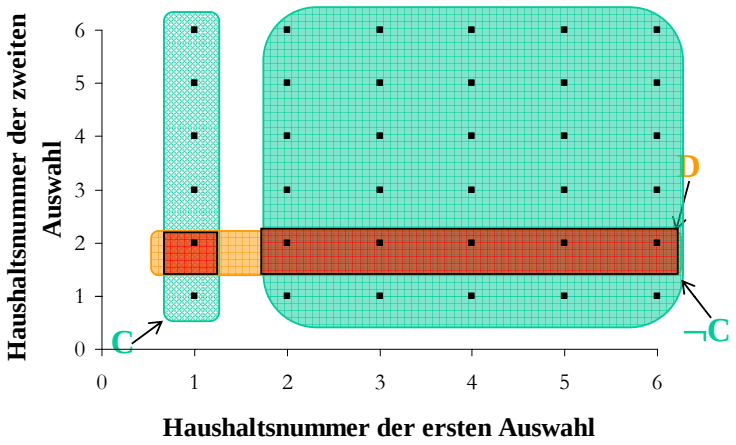
Mit Hilfe der unbedingten Wahrscheinlichkeiten lassen sich bedingten Wahrscheinlichkeiten für den Fall berechnen, dass bedingtes und bedingendes Ereignis ausgetauscht werden:

Schritt1:

Da ein Ereignis A und sein Komplementärereignis ¬A eine exhaustive Zerlegung des Universums bilden, ist die Wahrscheinlichkeit eines beliebigen Ereignisses B gleich der Summe der Wahrscheinlichkeiten des gleichzeitigen Auftretens von A und B sowie der von ¬A und B:

$$\Pr(B) = \Pr(A \cap B) + \Pr(\neg A \cap B)$$
$$= \Pr(B|A) \cdot \Pr(A) + \Pr(B|\neg A) \cdot \Pr(\neg A)$$

Theorem von Bayes



Im Beispiel:

$Pr(D) = Pr(C \cap D) + Pr(\neg C \cap D) = 1/36 + 5/36 = 6/36$

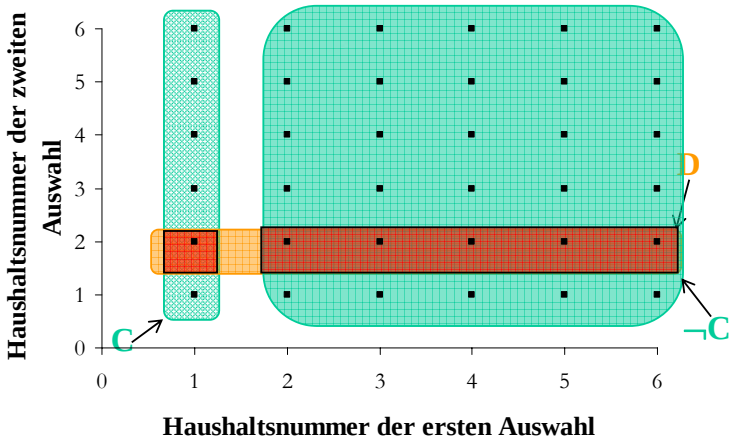
$= Pr(D|C) \cdot Pr(C) + Pr(D|\neg C) \cdot Pr(\neg C) = 1/6 \cdot 6/36 + 5/30 \cdot 30/36$

Schritt 2:

Die bedingte Wahrscheinlichkeit von A gegeben B ist dann eine Funktion der bedingten Wahrscheinlichkeiten von B gegeben A und von B gegeben ¬A:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)} = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B|A) \cdot \Pr(A) + \Pr(B|\neg A) \cdot \Pr(\neg A)}$$

Theorem von Bayes



Diese Beziehung ist als ***Satz von Bayes*** oder ***Bayessches Theorem*** bekannt.

Im Beispiel:

$$\Pr(C|D) = \frac{\Pr(D|C) \cdot \Pr(C)}{\Pr(D)} = \frac{1/6 \cdot 1/6}{1/6}$$

$$= \frac{\Pr(D|C) \cdot \Pr(C)}{\Pr(D|C) \cdot \Pr(C) + \Pr(D|\neg C) \cdot \Pr(\neg C)} = \frac{1/6 \cdot 1/6}{1/6 \cdot 1/6 + 5/30 \cdot 30/36}$$

Anwendung des Theorem von Bayes: Kumulierung von Wissen

Der Satz von Bayes ist die Grundlage der ***Bayesschen Statistik***, in der u.a. versucht wird, mit Hilfe von Daten Wissen zu kumulieren.

Ausgangspunkt ist die ***subjektive Wahrscheinlichkeit*** über ein Ereignis A, das ist die Sicherheit, mit der eine Ausage für wahr gehalten wird.

Beispiel:
Die Aussage „50% halten die Wirtschaftslage für gut“ sei durch A symbolisiert. Ein Forscher könnte etwa vermuten, dass diese Aussage mit einer ***subjektiven Apriori-Wahrscheinlichkeit*** *Pr(A) = 0.75 wahr ist.*
In einer Stichprobe von 100 Personen zeigt sich, dass nur 40% der Bevökerung die Wirtschaftslage für gut halten. Dies sei das ***empirische Datum B****. Die Wahrscheinlichkeit dieses tatsächlich aufgetretetenen (beobachteten) Datums (Ereignis B) ist eins: P(B)=1*

Die Wahrscheinlichkeit, dass von 100 Personen maximal 40% die Wirtschaftslage für gut halten, wenn es tatsächlich 50% in der Population sind, beträgt in einer einfachen Zufallsauswahl Pr(B|A) = 0.025.
Aus dem Satz von Bayes folgt dann:

$$\Pr(A|B) = \left(\Pr(B|A)\cdot\Pr(A)\right)/\Pr(B) = (0.025\cdot 0.75)/1 = 0.01875$$

Angesichts der Daten sinkt daher die subjektive Wahrscheinlichkeit von 0.75 auf nur noch 0.01875. Dies ist die sogenante ***Aposteriori-Wahrscheinlichkeit****, die Vorwissen bzw. Vorvermutungen mit empirischen Erkenntnissen rational kombiniert.*

Anwendung des Theorem von Bayes: Vermeidung von Fehlurteilen

Der Satz von Bayes kann auch helfen, Fehlschlüsse zu vermeiden.

Beispiel:
Mit Hilfe eines Tests wird mit Sicherheit, d.h. Wahrscheinlichkeit von 1 entdeckt, ob ein Vogel an der gefährlichen Form der Vogelgrippe gestorben ist;
mit einer Fehlerwahrscheinlichkeit von 1% (=0.01) wird bei einem toten Vogel fälschlicherweise Vogelgrippe diagnostiziert, obwohl sie nicht vorliegt.
In einer Region haben 0.5% (=0.005) der Vögel Vogelgrippe.
Bei einem toten Vogel zeigt der Test Vogelgrippe an. Wie wahrscheinlich ist es, dass der Vogel tatsächlich an der Vogelgrippe gestorben ist?

Intuitiv möchte man meinen, dass die gesuchte Wahrscheinlichkeit 99% beträgt, da der Test nur 1% Fehler macht. Tatsächlich ergibt sich eine Wahrscheinlichkeit von mehr als 33%!

A ist das Ereignis „Vogel hat Vogelgrippe“, B das Ereignis „Test zeigt Vogelgrippe an“.
Dann ist Pr(B|A) = 1, Pr(B|¬A) = 0.01, Pr(A) = 0.005 und Pr(¬A) = 0.995.

Die gesuchte Wahrscheinlichkeit ist dann die bedingte Wahrscheinlichkeit, dass ein Vogel Vogelgrippe hat, wenn der Test dies anzeigt: Pr(A|B).

Mit Hilfe des Satzes von Bayes errechnet sich diese Wahrscheinlichkeit als:

$$\Pr(A|B) = \frac{\Pr(B|A)\cdot\Pr(A)}{\Pr(B|A)\cdot\Pr(A) + \Pr(B|\neg A)\cdot\Pr(\neg A)} = \frac{1\cdot 0.005}{1\cdot 0.005 + 0.01\cdot 0.995} = 0.334$$

Lerneinheit 9:
Stichprobenziehung als Zufallsexperiment

Eine wichtige Anwendung der Wahrscheinlichkeitstheorie in den Sozialwissenschaften besteht bei der Beurteilung von Stichprobendaten, die Aussagen über Populationseigenschaften ermöglichen sollen.
Ausgangspunkt ist eine ***Population*** von insgesamt ***N Elementen***. Aus dieser Population wird eine ***Stichprobe*** von ***n*** ausgewählt. Die Stichprobenziehung ist ein Zufallsexperiment, wenn die Auswahl der Stichprobenelemente aus der Population nach einem Zufallsprinzip erfolgt.

So kann die Auswahl mittels einer Urne erfolgen. Dabei wird wie bei einer Lotterie für jedes der N Elemente eine numerierte Kugel mit der Fallnummer des Elements in eine Urne gelegt, die gut durchmischt wird. Nacheinander werden dann n Kugeln gezogen. Die Nummern auf den gezogenen Kugeln bestimmen die ausgewählten Elemente, die in die Stichprobe aufgenommen werden.

In der Realität erfolgt die Ziehung von Stichproben heutzutage mit Computerprogrammen, die z.B. mit Hilfe eines Zufallszahlengenerators zufällig Personen aus der Einwohnermeldeamtsdatei einer Stadt auswählen. Die Programme können eine Urnenwahl so simulieren, dass die Auswahl der physikalischen Auswahl mittels einer echten Urne entspricht.

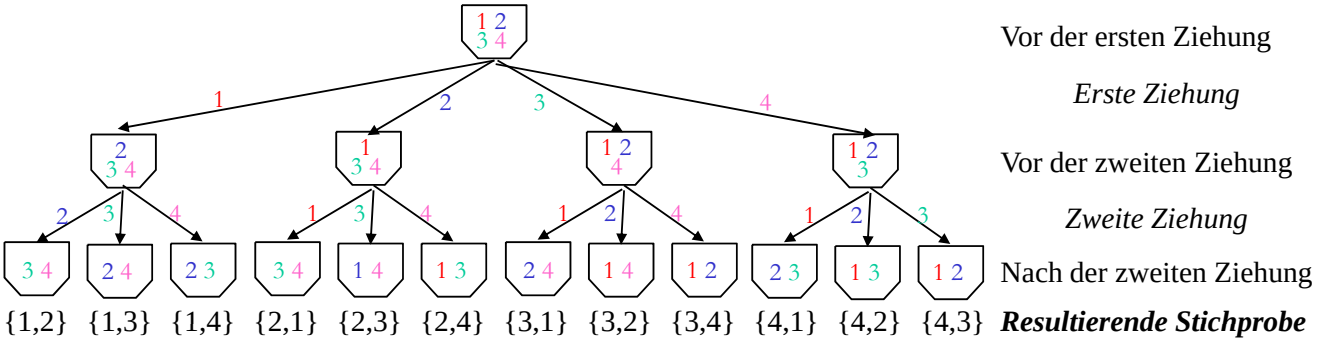
Wenn ein Element nur einmal ausgewählt werden kann, also nach der Auswahl nicht mehr in die Urne zurückgelegt wird, spricht man von einem ***Urnenmodell ohne Zurücklegen***.

Urnenmodell ohne Zurücklegen

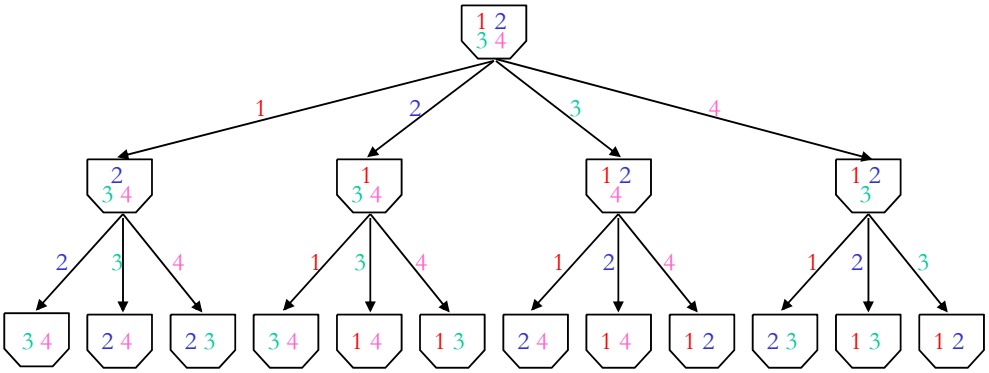
Vor der ersten Ziehung sind dann N Kugeln in der Urne. Es gibt somit auch N mögliche Resultate. Nach der ersten Ziehung sind nur noch N–1 Kugeln in der Urne, so dass für die zweite Ziehung N–1 Möglichkeiten bestehen, eine Kugel auszuwählen.
Nach der zweiten Ziehung sind noch N–2 Kugeln in der Urne, so dass es in der dritten Ziehung N–2 Möglichkeiten gibt.
Nach jeder Ziehung reduziert sich also die Zahl der Kugeln in der Urne um 1. Vor der n-ten Ziehung sind somit (N–n+1) Kugeln in der Urne, nach der n-ten Ziehung (N–n) Kugeln.

Grafisch lässt sich das gesamt Vorgehen mit Hilfe eines ***Ereignisbaums*** darstellen.

Um übersichtlich zu bleiben, wird als Beispiel die Auswahl von n=2 Elementen (Fällen) aus N=4 Elementen einer Population dargestellt.



Urnenmodell ohne Zurücklegen



Vor der ersten Ziehung

Erste Ziehung

Vor der zweiten Ziehung

Zweite Ziehung

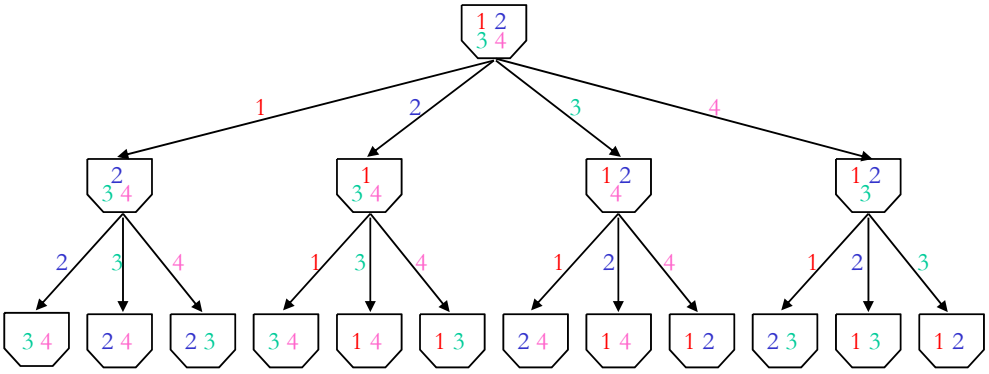
Nach der zweiten Ziehung

{1,2} {1,3} {1,4} {2,1} {2,3} {2,4} {3,1} {3,2} {3,4} {4,1} {4,2} {4,3} ***Resultierende Stichprobe***

Insgesamt gibt es im Beispiel 12 = 4 · 3 mögliche Ergebnisse des Zufallsexperiments „Zufälliges Ziehen von n=2 Elemenen aus N=4 Elementen“.

Geht man davon aus, dass bei jedem Ziehungsschritt die gleiche Auswahlwahrscheinlichkeit für eine der Kugeln in der Urne vorliegt, dann beträgt die Wahrscheinlichkeit für jedes Ergebnis vor der ersten Ziehung 1/N, im Beispiel 1/4 und vor der zweiten Ziehung 1/(N–1), im Beispiel 1/3, wobei die Auswahrscheinlichkeit des zweiten Ziehungsschritts eine bedingte Wahrscheinlichkeit ist, gegeben den ersten Ziehungsschritt.
Nach dem Multiplikationstheorem der Wahrscheinlichkeitstheorie beträgt dann die Wahrscheinlichkeit jeder Stichprobe 1/N ·1/(N–1) · ... · 1/(N–n+1), im Beispiel 1/4 ·1/(4–2+1) = 1/12.

Einfache Zufallsauswahl ohne Zurücklegen



Vor der ersten Ziehung

Erste Ziehung

Vor der zweiten Ziehung

Zweite Ziehung

Nach der zweiten Ziehung

{1,2} {1,3} {1,4} {2,1} {2,3} {2,4} {3,1} {3,2} {3,4} {4,1} {4,2} {4,3} ***Resultierende Stichprobe***

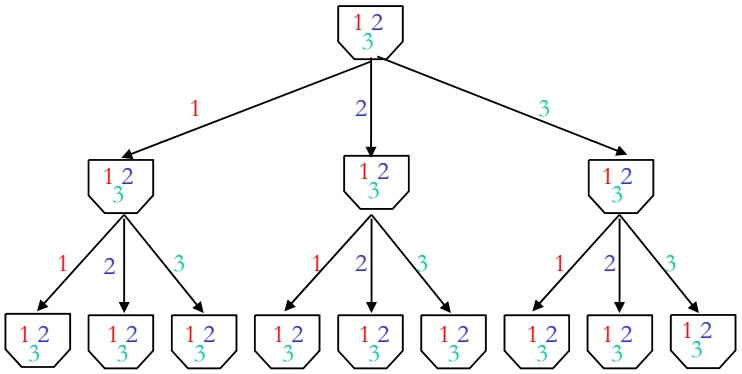
In einer Stichprobe kommt eine Nummer genau einmal vor, über alle 12 Stichproben kommt jede Nummer sechsmal vor. Die Wahrscheinlichkeit, dass ein beliebiges Element ausgewählt wird, beträgt also für jede Nummer 6/12 bzw. 0.5.
Es gibt jeweils zwei Stichproben mit gleichen Fällen, z.B. {1,2} und {2,1}.

Eine Zufalls- oder Wahrscheinlichkeitsauswahl heißt ***einfache Zufallsauswahl***, wenn jedes Element mit gleicher Wahrscheinlichkeit und auch ***jede mögliche Stichprobe gleicher Fallzahl*** und gleichen Anzahlen wiederholt gezogener Elemente mit jeweils ***gleicher Wahrscheinlichkeit*** ausgewählt wird.
Da jedes Element nur einmal ausgewählt werden kann, handelt es sich um eine ***einfache Zufallsauswahl ohne Zurücklegen***.

Einfache Zufallsauswahl mit Zurücklegen

Bei einer ***einfachen Zufallsauswahl mit Zurücklegen*** kann jede Nummer mehrfach ausgewählt werden, da die entsprechende Kugel nach der Ziehung wieder in die Urne zurückgelegt wird.

Das folgende Beispiel zeigt den Ereignisbaum einer einfachen Zufallsauswahl mit Zurücklegen von n=2 Elementen aus N=3 Elementen.



Vor der ersten Ziehung

Erste Ziehung

Vor der zweiten Ziehung

Zweite Ziehung

Nach der zweiten Ziehung

Resultierende Stichprobe

Vor jeder Ziehung beträgt die Auswahlwahrscheinlichkeit jeder Nummer 1/N, im Beispiel 1/3. Die einzelnen Ziehungen sind statistisch unabhängig voneinander. Die Auswahlwahrscheinlichkeit jeder Stichprobe beträgt daher $(1/N)^n$, im Beispiel 1/9 (= $(1/3)^2$ = 1/3· 1/3).

Im Beispiel wird jedes Element insgesamt sechsmal in fünf Stichproben ausgewählt. Die Wahrscheinlichkeit ein beliebiges Element genau einmal auszuwählen, beträgt 4/9, die Wahrscheinlichkeit ein beliebiges Element zweimal auszuwählen, 1/9.

Kombinatorik: Permutationen, Variationen, Kombinationen

Mit Hilfe der ***Kombinatorik*** lassen sich für beliebige Populations- und Stichprobengrößen Formeln herleiten, mit denen sich die Zahl der Stichproben berechnen lassen.

Bei einer einfachen Zufallsauswahl ohne Zurücklegen ergibt sich die Zahl der Stichproben durch die Anzahl der Möglichkeiten n Elementen aus N Elementen auszuwählen. Für den ersten Auswahlschritt stehen N Elemente zu Verfügung. Nach jedem Auswahlschritt reduziert sich die Zahl der verbleibenden Elememnte um 1, so dass sich für die folgenden Schritte auch die Zahl der Auswahlmöglichkeiten jeweils um 1 reduziert. Die Zahl der Möglichkeiten, n Elemente aus N ohne Zurücklegen auszuwählen beträgt daher:

$$\underbrace{N\cdot(N-1)\cdot(N-2)\cdot\ldots\cdot(N-n+2)\cdot(N-n+1)}_{\text{Produkt aus n Faktoren}}=\prod_{i=1}^{n}(N-i+1)$$

Die Formel gibt also die Zahl der möglichen Anordnungen an, wenn n Elemente aus N bei Berücksichtigung der Reihenfolge ausgewählt werden. In der Kombinatorik bezeichnet man diese Zahl der Anordnungen als ***Variationen***, die durch das Symbol $_N\mathbf{V_n}$ abgekürzt wird. Somit gilt:

$$_N V_n=\prod_{i=1}^{n}(N-i+1)=N\cdot(N-1)\cdot(N-2)\cdot\ldots(N-n+2)\cdot(N-n+1)$$

Im Beispiel der einfachen Zufallsauswahl von n=2 Elementen aus einer Population des Umfangs N=4 ergaben sich daher $_4V_2$ = 4·3 = 12 unterscheidbare Stichproben.

Permutationen

Werden der Reihe nach alle N Elemente der Population ausgewählt, berechnet sich die Zahl der unterscheidbaren Anordnungen, also der möglichen Reihenfolgen der Auswahl der Elemente nach:

$$\underbrace{N \cdot (N-1) \cdot (N-2) \cdot \ldots 3 \cdot 2 \cdot 1}_{\text{Produkt aus N Faktoren}} = \prod_{i=1}^{N} i = N!$$

Das Ausrufungszeichen hinter der Zahl N steht für das Fakultätszeichen. Das Produkt aller Zahlen von 1 bis N bezeichnet man in der Mathematik auch als „***Fakultät von N***“ oder „***N-Fakultät***“. Die Fakultät von 1 ist 1, die von 2 ist 2 (= 2·1), die von 3 ist 6 (= 3·2·1) usw.. Definitorisch ist zudem auch die Fakultät von 0 auf den Wert 1 festgesetzt: 0! = 1! = 1.

In der Kombinatorik bezeichnet man die Zahl der möglichen Anordnungen von N Elementen als ***Permutationen*** und verwendet hierfür das Symbol $\mathbf{P_n}$.

Bei n=4 Elementen gibt es also P_4= 4! = 4·3·2·1 = 24 Möglichkeiten, diese Zahlen anzuordnen. Bei n=5 sind es schon P_5= 5! = 5·4·3·2·1 = 120 verschiedene Anordnungen. Die Zahl der Permuationen nimmt also sehr schnell zu und übersteigt bereits bei n= 10 mit 10! = 3 628 800 die Millionengrenze.

Variationen und Kombinationen ohne Wiederholung

Die Zahl der ***Variationen*** $\mathbf{{}_N V_n}$, also der möglichen Anordnungen von n Elementen aus N Elementen, lässt sich auch als Quotient der Permutationen von N geteilt durch die Zahl der Permutationen von N–n darstellen:

$${}_N V_n = \frac{N \cdot (N-1) \cdot (N-2) \cdot \ldots (N-n+1) \cdot (N-n) \cdot (N-n-1) \cdot \ldots 2 \cdot 1}{(N-n) \cdot (N-n-1) \cdot \ldots 2 \cdot 1} = \frac{N!}{(N-n)!} = \frac{P_N}{P_{N-n}}$$

Bei Permutationen und Variationen wird die Zahl möglicher Anordnungen betrachtet. Die Reihenfolge der Auswahl der Elemente ist daher von Bedeutung.
Man kann sich aber auch die Frage stellen, wie viele Möglichkeiten es gibt, n Elemente aus N Elementen auszuwählen, wenn die Reihenfolge keine Rolle spielen soll.

Wenn die Reihenfolge keine Rolle spielt, werden also z.B. die Anordnungen (1,2,3) (1,3,2), (2,1,3), (2,3,1), (3,1,2) und (3,2,1) nicht unterschieden.

Die Möglichkeiten, n Elemente aus N Elementen ohne Berücksichtigung der Reihenfolge auszuwählen, werden in der Kombinatorik als ***Kombinationen*** bezeichnet. Bei n ausgewählten Elementen gibt es n! Anordnungen, die in der Zahl der Kombinationen nicht unterschieden werden. Die Zahl der Kombinationen $\mathbf{{}_N K_n}$ ergibt sich daher dadurch, dass die Zahl der Variationen durch die Anzahl der Permutationen der ausgewählten Elemente geteilt wird:

$${}_N K_n = \frac{{}_N V_n}{P_n} = \frac{P_N}{P_{N-n} \cdot P_n} = \frac{N!}{(N-n)! \cdot n!} = \frac{N \cdot (N-1) \cdot \ldots \cdot (N-n+1)}{n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1} = \binom{N}{n}$$

Binomialkoeffizient

Der in der Formel der Kombinationen ganz rechts stehende Ausdruck heißt ***Binomialkoeffizient***. Der Binomialkoeffizient von a und b wird als „a über b“ ausgesprochen, wobei a für die die obere und b für die untere Zahl steht und die obere Zahl nicht kleiner als die untere sein darf. Der Binomialkoeffizient „ a über b“ berechnet sich dann nach:

$$\binom{a}{b} = \frac{a!}{b!\cdot(a-b)!} = \frac{a\cdot(a-1)\cdot(a-2)\cdot\ldots 2\cdot 1}{b\cdot(b-1)\cdot(b-2)\cdot\ldots 2\cdot 1\cdot(a-b)\cdot(a-b-1)\cdot\ldots\cdot 2\cdot 1}$$

$$= \frac{a\cdot(a-1)\cdot(a-2)\cdot\ldots\cdot(a-b)\cdot(a-b-1)\cdot\ldots\cdot 2\cdot 1}{b\cdot(b-1)\cdot(b-2)\cdot\ldots 2\cdot 1\cdot(a-b)\cdot(a-b-1)\cdot\ldots\cdot 2\cdot 1} = \prod_{i=1}^{b}\frac{(a-i+1)}{i}$$

Der Binomialkoeffizient „N über n“ gibt also die Anzahl der Kombination an, die sich ergeben, wenn man n Elemente aus N ohne Berücksichtigung der Reihenfolge auswählt. Aufgrund der Symmetrie in der Formel ergibt sich die gleiche Anzahl von Kombinationen, wenn man N–n Elemente aus N Elementen auswählt.
Da die Summe aus n und N–n wieder N ergibt, ist die Zahl der Kombinationen von n aus N (oder von N–n aus N) auch gleichzeitig die Zahl der Möglichkeiten, eine Menge von N Elementen in zwei Teilmengen von n und von N–n Elementen zu zerlegen.

Variationen und Kombinationen mit Wiederholung

Im Urnenmodell mit Wiederholungen ist es möglich, dass die gleichen Elemente mehrfach ausgewählt werden. Wenn n aus N Elementen mit Zurücklegen ausgewählt werden, reduziert sich daher nicht die Zahl der auszuwählenden Elemente nach der Auswahl eines Elements. Die Anzahl der möglichen ***Variationen*** (Anordnungen) von n Elementen aus N Elementen ***mit Wiederholung*** ${}_{N}V^{*}_{n}$ berechnet sich daher als n-te Potenz von N:

$${}_{N}V^{*}_{n} = N^{n} = \underbrace{N\cdot N\cdot\ldots\cdot N}_{\text{n mal}} = \prod_{i=1}^{n} N$$

Komplizierter ist die Zahl der Kombinationen mit Wiederholung zu berechnen. Wenn n=2 Elemente aus N Elementen ausgewählt werden, gibt es N^2 Variationen. Von diesen Variationen haben N zweimal das gleiche Element, z.B. (1,1) oder (2,2) und $N^2–N = N\cdot(N–1)$ haben zwei unterschiedliche Elemente, z.B. (1,2) oder (2,1), die nicht unterschieden werden. Es gibt insgesamt „N+1 über 2“ Möglichkeiten, n=2 aus N Elemementen ohne Berücksichtigung der Anordnung auszuwählen.

Bei einer einfachen Zufallsauswahl von n=2 aus N=3 Elementen {a,b,c} mit Zurücklegen gibt es „4 über 2“ =(4! / (2!·(4-2)!) = 6 unterscheidbare Stichproben ohne Berücksichtigung der Ziehungsreihenfolge, nämlich {a,a}, {b,b}, {c,c}, {a,b}, {a,c}, und {b,c}.

Noch komplexer wird es, wenn die Zahl der ausgewählten Elemente weiter ansteigt.

Bei einer einfachen Zufallsauswahl von n=3 aus N=3 Elementen {a,b,c} mit Zurücklegen gibt es 10 unterscheidbare Stichproben ohne Berücksichtigung der Reihenfolge:

Variationen und Kombinationen mit Wiederholung

{a,a,a}, {b,b,b}, {c,c,c}, {a,a,b}, {a,a,c}, {b,b,a}, {b,b,c}, {c,c,a}, {c,c,b} und {a,b,c}.
Werden n=4 aus N=3 Elemente mit Zurücklegen ausgewählt, gibt es 15 unterscheidbare Stichproben: {a,a,a,a}, {b,b,b,b}, {c,c,c,c}, {a,a,a,b}, {a,a,a,c}, {b,b,b,a}, {b,b,b,c}, {c,c,c,a}, {c,c,c,b}, {a,a,b,b}, {a,a,c,c}, {b,b,c,c}, {a,a,b,c}, {b,b,a,c} und {c,c,a,b}.

Generell berechnet sich die Zahl der ***Kombinationen mit Wiederholungen*** ${}_N\mathrm{K}^*_n$, also die Zahl der Möglichkeiten, n Elemente aus N Elementen ohne Berücksichtigung der Reihenfolge auszuwählen, wenn jeder der N Elemente beliebig oft in der Stichprobe der n Elemente vorkommen kann nach „N+n–1 über n“:

$$ {}_N\mathrm{K}^*_n = \binom{N+n-1}{n} = \frac{(N+n-1)!}{n!\cdot(N-1)!} = \frac{(N+n-1)\cdot(N+n-2)\cdot\ldots\cdot N}{n\cdot(n-1)\cdot\ldots\cdot 2\cdot 1} $$

Die Anwendung der Formel lässt sich an den vorgestellten Beispielen demonstrieren.
Bei einer Auswahl von n=2 aus N=3 Elementen mit Zurücklegen, ist N+n–1 = 2+3–1 = 4. Entsprechend gibt es „4 über 2“ = 4!/(2!·2!) = 6 unterscheidbare Stichproben, wenn die Reihenfolge keine Rolle spielt.
Bei n=3 aus N=3 ist N+n–1 = 3+3–1 = 5. Es gibt also „5 über 3“ = 5!/(3!·2!) = 10 unterscheidbare Stichproben.
Bei n=4 aus N=3 ist N+n–1 = 3+4–1 = 6. Es gibt also „6 über 4“ = 6!/(4!·2!) = 15 unterscheidbare Stichproben.

Realisierungswahrscheinlichkeit von Stichproben bei einfachen Zufallsauswahlen

Die Berechnung der Zahl der unterschiedlichen Stichproben ist nur ein Zwischenschritt, um die Realisierungswahrscheinlichkeiten von Stichproben zu berechnen.

Das Kennzeichen einer einfachen Zufallsauswahl ist, dass in einer Stichprobe jeweils alle Elemente und auch alle Teilmengen mit jeweils gleicher Anzahl von Elementen die gleichen Auftretenswahrscheinlichkeiten aufweisen.
Dies gilt allerdings nur bei Berücksichtigung der Reihenfolge.

Aus diesem Definitionsmerkmal folgt, dass die Wahrscheinlichkeit jeder möglichen Stichprobe von n aus N Elementen gerade gleich dem Kehrwert der möglichen Anordnungen ist.
Bei einer ***einfachen Zufallsauswahl ohne Zurücklegen*** von ***n aus N*** Elementen beträgt daher die Auswahlwahrscheinlichkeit jeder Stichprobe bei Berücksichtigung der Ziehungsreihenfolge:

$$ \Pr(\text{jede Stichprobe}) = ({}_N\mathrm{V}_n)^{-1} = \frac{1}{{}_N\mathrm{V}_n} = \frac{(N-n)!}{N!} $$

Bei einer ***einfachen Zufallsauswahl mit Zurücklegen*** von ***n aus N*** Elementen beträgt dagegen die Auswahlwahrscheinlichkeit jeder Stichprobe bei Berücksichtigung der Ziehungsreihenfolge:

$$ \Pr(\text{jede Stichprobe}) = N^{-n} = \frac{1}{N^n} = \frac{1}{\underbrace{N\cdot N\cdot\ldots\cdot N}_{\text{Produkt aus n Faktoren}}} $$

Realisierungswahrscheinlichkeit von Stichproben bei einfachen Zufallsauswahlen

Spielt die Reihenfolge der Ziehung keine Rolle, so gibt es bei einer einfachen Zufallsauswahl ohne Zurücklegen n! Möglichkeiten, die ausgewählten n von N Elementen in einer unterschiedlichen Reihenfolge zu ziehen. Da keines der ausgewählten Elemente mehrfach in der Stichprobe vorkommen kann, sind alle Auswahlwahrscheinlichkeiten gleich.

Bei einer ***einfachen Zufallsauswahl ohne Zurücklegen*** von ***n aus N*** Elementen ist daher die Auswahlwahrscheinlichkeit jeder Stichprobe ohne Berücksichtigung der Ziehungsreihenfolge gleich dem Kehrwert der Zahl der n aus N Kombinationen ohne Wiederholung:

$$\Pr(\text{jede Stichprobe}) = \left({}_N K_n\right)^{-1} = \frac{1}{{}_N K_n} = \frac{n!\cdot(N-n)!}{N!} = \frac{1}{\binom{N}{n}}$$

Im Unterschied zur Auswahl ohne Zurücklegen, bei der die n ausgewählten Elemente in n! jeweils gleich wahrscheinlichen Anordnungen variieren können, mus bei einer Auswahl mit Zurücklegen die Zahl der mehrfach ausgewählten Elemente berücksichtigt werden.

So gibt es bei der einfachen Zufallsauswahl mit Zurücklegen von n=4 aus N=3 Elementen insgesamt $3^4 = 81$ Stichproben bei Berücksichtigung der Reihenfolge. Wird die Reihenfolge nicht berücksichtigt, gibt es „6 über 4“ = 15 ununterscheidbare Kombinationen. Bei den 3 Kombinationen {a,a,a,a}, {b,b,b,b} und {c,c,c,c} gibt es nur jeweils eine Anordnung, bei den 6 Kombinationen {a,a,a,b}, {a,a,a,c}, {b,b,b,a}, {b,b,b,c}, {c,c,c,a} und {c,c,c,b} gibt es dagegen jeweils 4 Anordnungen, z.B. bei der ersten (a,a,a,b), (a,a,b,a), (a,b,a,a) und (b,a,a,a).

Realisierungswahrscheinlichkeit von Stichproben bei einfachen Zufallsauswahlen

Bei den 3 Kombinationen mit jeweils 2 und 2 gleichen Elementen {a,a,b,b}, {a,a,c,c,} und {b,b,c,c} gibt es jeweils 6 Anordnungen, z.B. bei der ersten (a,a,b,b), (a,b,a,b), (a,b,b,a), (b,a,a,b), (b,a,b,a) und (b,b,a,a). Schließlich gibt es bei den 3 Kombinationen {a,a,b,c}, {b,b,a,c} und {c,c,a,b} mit 2 gleichen und 2 verschiedenen Elementen jeweils 12 Anordnungen, z.B. bei der ersten (a,a,b,c), (a,a,c,b), (a,b,a,c), (a,b,c,a), (a,c,a,b), (a,c,b,a), (b,a,a,c), (b,a,c,a), (b,c,a,a), (c,a,a,b), (c,a,b,a) und (c,b,a,a).

Generell ergibt sich die Zahl der Anordnungen (Permutationen) in einer Menge von n Elementen, von denen jeweils n_1, n_2, ..., n_k gleich und damit ununterscheidbar sind, so dass $n_1 + n_2 + ... + n_k = n$, nach:

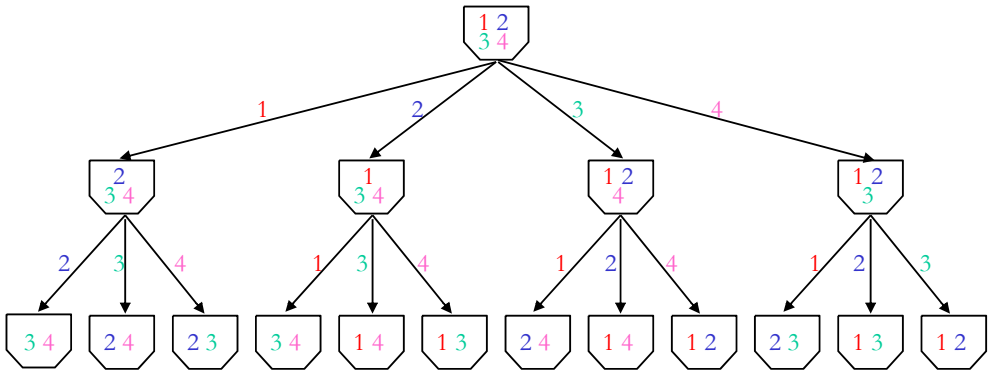
$$P(n_1, n_2, \ldots, n_k) = \frac{(n_1 + n_2 + \ldots + n_k)!}{n_1!\cdot n_2!\cdot \ldots \cdot n_k!}$$

Die Auswahlwahrscheinlichkeit einer Stichprobe mit n_1 Elementen „a“, n_2 Elementen „b“, ... und n_k Elementen „k“ bei einer ***einfachen Zufallsauswahl mit Zurücklegen*** von **n aus N** Elementen und ohne Berücksichtigung der Anordnung beträgt daher:

$$\Pr(\text{Stichprobe}) = \left(\frac{N^n}{P(n_1, n_2, ..., n_k)}\right)^{-1} = \frac{n!}{N^n \cdot n_1!\cdot n_2!\cdot \ldots \cdot n_k!}$$

Zur Demonstration können diese Berechnungen auf die beiden Eingangsbeispiele angewendet werden.

Stichprobenwahrscheinlichkeit bei einfacher Zufallsauswahl ohne Zurücklegen



Vor der ersten Ziehung

Erste Ziehung

Vor der zweiten Ziehung

Zweite Ziehung

Nach der zweiten Ziehung

{1,2} {1,3} {1,4} {2,1} {2,3} {2,4} {3,1} {3,2} {3,4} {4,1} {4,2} {4,3} ***Resultierende Stichprobe***

$$\Pr(\text{jede Stichprobe}) = \frac{1}{{}_N V_n} = \frac{(N-n)!}{N!} = \frac{(4-2)!}{4!} = \frac{2}{24} = \frac{1}{12}$$

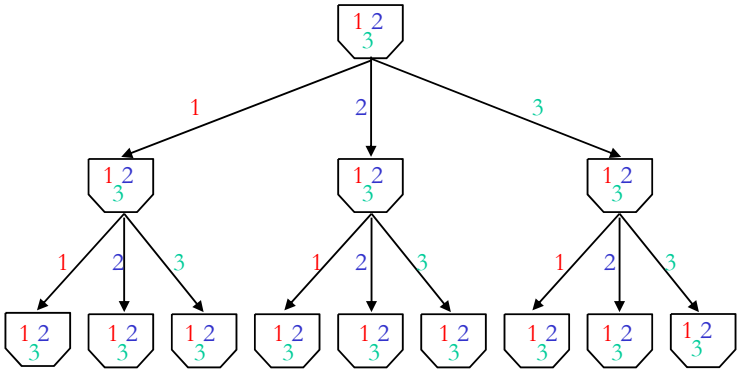
mit Berücksichtigung der Reihenfolge

$$\Pr(\text{jede Stichprobe}) = \frac{1}{{}_N K_n} = \frac{n!\cdot(N-n)!}{N!} = \frac{2!\cdot 2!}{4!} = \frac{4}{24} = \frac{1}{6}$$

ohne Berücksichtigung der Reihenfolge

In Abhängigkeit von der Berücksichtigung bzw. Nichtberücksichtigung der Reihenfolge der Ziehung hat jede der 12 Stichproben eine Wahrscheinlichkeit von 1/12 oder 1/6.

Realisierungswahrscheinlichkeit bei einfacher Zufallsauswahl mit Zurücklegen



Vor der ersten Ziehung

Erste Ziehung

Vor der zweiten Ziehung

Zweite Ziehung

Nach der zweiten Ziehung

{1,1} {1,2} {1,3} {2,1} {2,2} {2,3} {3,1} {3,2} {3,3} ***Resultierende Stichprobe***

$$\Pr(\text{jede Stichprobe}) = \frac{1}{N^n} = \frac{1}{3^2} = \frac{1}{9}$$

mit Berücksichtigung der Reihenfolge

$$\binom{N+n-1}{n} = \binom{3+2-1}{2} = 6 \text{ Stichproben,}$$

ohne Berücksichtigung der Reihenfolge

wobei: $\Pr(\text{Stichprobe mit identischen Elementen}) = \frac{n!}{N^n \cdot n!} = \frac{2!}{3^2 \cdot 2!} = \frac{1}{9}$

und: $\Pr(\text{Stichprobe mit verschiedenen Elementen}) = \frac{n!}{N^n \cdot n_1! \cdot n_2!} = \frac{2!}{3^2 \cdot 1! \cdot 1!} = \frac{2}{9}$

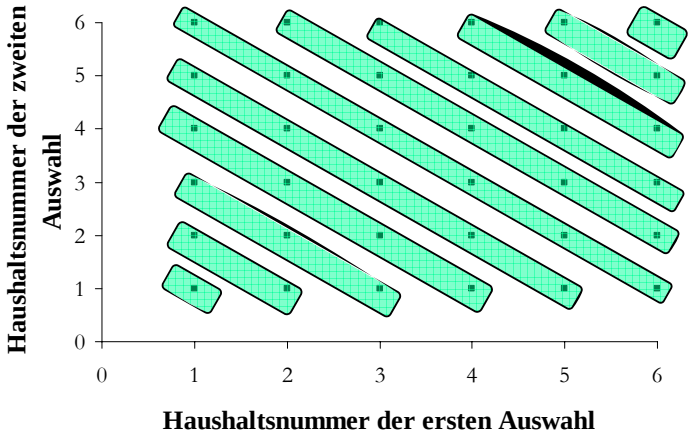
Lerneinheit 10:
Zufallsvariablen und Wahrscheinlichkeitsverteilungen

Die Berechnung der Ziehungswahrscheinlichkeit einer Stichprobe ist nur der erste Schritt bei der Abschätzung der Risiken von Fehlentscheidungen bei Induktionsschlüssen von einer Stichprobe auf die Population, aus der die Stichprobe kommt.
Von Interesse sind i.a. nämlich nicht die Stichproben an sich, sondern Kennwerte, die aus der resultierenden Verteilung in einer Stichprobe berechnet werden und als Schätzung entsprechender Kennwerte in der Population herangezogen werden.

Als Eingangsbeispiel zur Wahrscheinlichkeitsrechnung (in L08) wurde die Auswahl von Haushalten betrachtet. Das Beispiel beinhaltet eine einfache Zufallsauswahl mit Zurücklegen von n=2 aus N=6 Haushalten. Von Interesse könnte etwa das mittlere Einkommen in der Population der N=6 Haushalte sein.
Der Einfachheit halber sei angenommen, dass die Haushaltsnummer das Haushaltseinkommen in 1000 € pro Monat angibt. Haushalt 1 hat also ein Einkommen von 1 Tsd. €, Haushalt 2 ein Einkommen von 2 Tsd € ... und Haushalt 6 schließlich von 6 Tsd. €.

Bei Berücksichtigung der Anordnung gibt es $6^2 = 36$ unterscheidbare Stichproben. Wenn zur Schätzung des Einkommens in jeder Stichprobe das durchschnittliche Einkommen der Haushalte in den einzelnen Stichproben berechnet wird, ergeben sich allerdings nur 11 verschiedene Mittelwerte.

Zufallsvariablen und Wahrscheinlichkeitsverteilungen

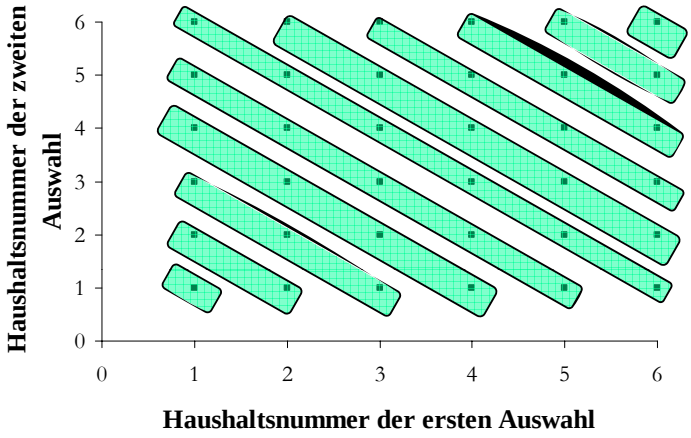


Wird zweimal Haushalt Nr. 1 ausgewählt, beträgt das Durchschnittseinkommen in der Stchprobe 1000 €.

Wird zuerst Haushalt Nr. 1 und dann Haushalt Nr 2 ausgewählt oder erst Haushalt Nr. 2 und dann Haushalt Nr. 1, beträgt das Durchschnittseinkommen in den beiden Stichproben jeweils 1500 €.

Wird zweimal Haushalt Nr. 2 ausgewählt oder Haushalt Nr. 1 und Haushalt Nr. 3, ergibt sich jeweils ein Durchschnittseinkommen in den Stichproben von 2000 €.

Wahrscheinlichkeitsverteilungen

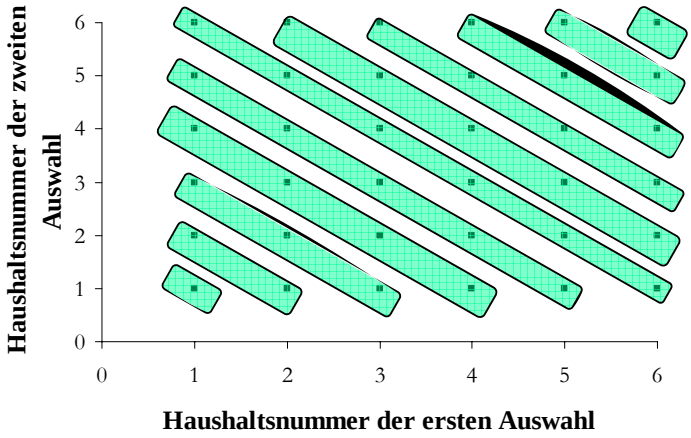


Elemente in Stichprobe	Realisierungswahr-scheinlichkeit	Mittleres Einkommen
{1,1}	1/36	1000 €
{2,1}	2/36	1500 €
{3,1}{2,2}	3/36	2000 €
{4,1}{3,2}	4/36	2500 €
{5,1}{4,2}{3,3}	5/36	3000 €
{6,1}{5,2}{4,3}	6/36	3500 €
{6,2}{5,3}{4,4}	5/36	4000 €
{6,3}{5,4}	4/36	4500 €
{6,4}{5,5}	3/36	5000 €
{6,5}	2/36	5500 €
{6,6}	1/36	6000 €
Summe:	36/36	

Da jede Stichprobe ein Ereignis eines Zufallsexperiments ist, gilt dies auch für das in einer Stichprobe berechnete Durchschnittseinkommen. Die Realisierungswahrscheinlichkeit eines Durchschnittseinkommens ergibt sich als Summe der Realisierungswahrscheinlichkeiten der Stichproben, die zu diesem Durchschnittseinkommen führen.

Die Tabelle zeigt so für jedes der 11 möglichen Durchschnittseinkommen in den Stichproben die jeweilige Auftretenswahrscheinlichkeit. Die Realisierungswahrscheinlichkeiten ergeben eine unimodale, symmetrische Wahrscheinlichkeitsverteilung um den Wert 3500 €.

Zufallsvariablen



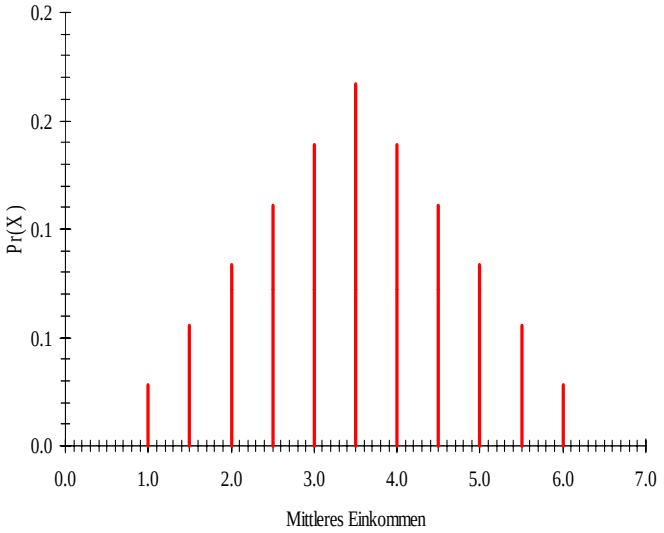
Elemente in Stichprobe	Realisierungswahr-scheinlichkeit	Mittleres Einkommen
{1,1}	1/36	1000 €
{2,1}	2/36	1500 €
{3,1}{2,2}	3/36	2000 €
{4,1}{3,2}	4/36	2500 €
{5,1}{4,2}{3,3}	5/36	3000 €
{6,1}{5,2}{4,3}	6/36	3500 €
{6,2}{5,3}{4,4}	5/36	4000 €
{6,3}{5,4}	4/36	4500 €
{6,4}{5,5}	3/36	5000 €
{6,5}	2/36	5500 €
{6,6}	1/36	6000 €
Summe:	36/36	

Die unterschiedlichen Ausprägungen des mittleren Einkommens in den Stichproben können zu einer Variable „Durchschnittseinkommen in den Stichproben“ zusammengefasst werden.
Im Unterschied zu empirischen Variablen weist diese Variable keine empirischen Auftretenshäufigkeiten ihrer Ausprägungen auf, sondern Auftretetenswahrscheinlichkeiten.

Variablen, deren Ausprägungen mit (im Prinzip berechenbaren) Auftretenswahrscheinlichkeiten realisiert werden, heißen ***Zufallsvariablen***.

Zufallsvariablen und Wahrscheinlichkeitsverteilungen

Elemente in Stichprobe	Realisierungswahr-scheinlichkeit	Mittleres Einkommen
{1,1}	1/36	1000 €
{2,1}	2/36	1500 €
{3,1}{2,2}	3/36	2000 €
{4,1}{3,2}	4/36	2500 €
{5,1}{4,2}{3,3}	5/36	3000 €
{6,1}{5,2}{4,3}	6/36	3500 €
{6,2}{5,3}{4,4}	5/36	4000 €
{6,3}{5,4}	4/36	4500 €
{6,4}{5,5}	3/36	5000 €
{6,5}	2/36	5500 €
{6,6}	1/36	6000 €
Summe:	36/36	

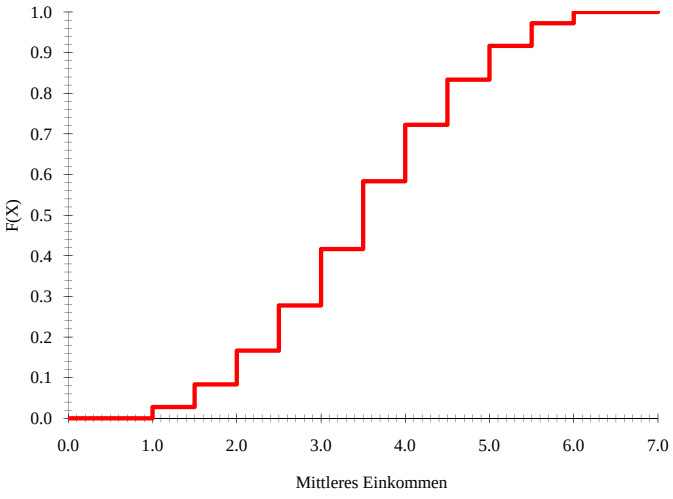


Die Auftretenswahrscheinlichkeiten der Ausprägungen definieren die ***Wahrscheinlichkeits-funktion*** einer Zufallsvariablen X, die jeder Ausprägung ihre Realisierungswahrscheinlichkeit zuordnet. Wahrscheinlichkeitsfunktionen der Ausprägungen einer Zufallsvariable X werden durch **Pr(X)** oder **f(x)** symbolisiert.

Die ***Auftretenswahrscheinlichkeiten*** der Ausprägungen einer Zufallsvariablen entsprechen den ***relativen Auftretenshäufigkeiten*** der Ausprägungen einer empirischen Verteilung.

Zufallsvariablen und Wahrscheinlichkeitsverteilungen

X (mittleres Einkomen in €)	Wahrscheinlich-keitsfunktion	Verteilungs-funktion
1000	1/36	1/36
1500	2/36	3/36
2000	3/36	6/36
2500	4/36	10/36
3000	5/36	15/36
3500	6/36	21/36
4000	5/36	26/36
4500	4/36	30/36
5000	3/36	33/36
5500	2/36	35/36
6000	1/36	36/36
Summe:	36/36	



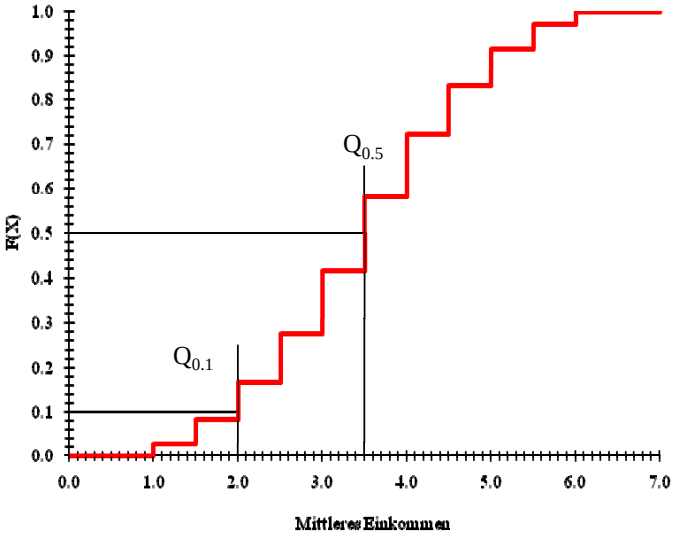
Die Aufsummierung der Wahrscheinlichkeitsfunktion ergibt die ***Verteilungsfunktion*** F(X), die für jede Zahl die Wahrscheinlichkeit angibt, dass eine Realisierung kleiner oder gleich dieser Zahl ist:

$$F(X = x) = Pr(X \leq x)$$

Die Verteilungsfunktion von Zufallsvariablen entspricht der empirischen Verteilungsfunktion empirischer Variablen, also der Aufsummierung der relativen Häufigkeiten, mit denen eine Ausprägung vorkommt.

Zufallsvariablen und Wahrscheinlichkeitsverteilungen

X (mittleres Einkomen in €)	Wahrscheinlich-keitsfunktion	Verteilungs-funktion
1000	1/36	1/36
1500	2/36	3/36
2000	3/36	6/36
2500	4/36	10/36
3000	5/36	15/36
3500	6/36	21/36
4000	5/36	26/36
4500	4/36	30/36
5000	3/36	33/36
5500	2/36	35/36
6000	1/36	36/36
Summe:	36/36	



Analog zu empirischen Verteilungsfunktionen lassen sich auch für Zufallsvariablen aus der Umkehrung der Verteilungsfunktion ***Quantilwerte*** berechnen, die wie empirische Quantilen bei empirischen Verteilungen berechnet werden.
So ist das z.B. das 10%-Quantil der Wert, bei dem die Verteilungsfunktion erstmals den Anteil 0.1 erreicht oder überschreitet.
Das 50%-Quantil ist bei Zufallsvariablen immer gleichzeitig der ***Median***, da es bei Wahrscheinlichkeiten keine geraden und ungeraden Fallzahlen gibt.

Im Beispiel beträgt der Median 3500 €.

Zufallsvariablen und Wahrscheinlichkeitsverteilungen

X (mittleres Einkomen in €)	Wahrscheinlich-keitsfunktion	Verteilungs-funktion	Quan-tile	$X \cdot Pr(X)$	$X^2 \cdot Pr(X)$
1000	1/36 = 0.0278	1/36 = 0.0278		1000/36	1000000/36
1500	2/36 = 0.0555	3/36 = 0.0833		3000/36	4500000/36
2000	3/36 = 0.0833	6/36 = 0.1667	10%	6000/36	12000000/36
2500	4/36 = 0.1111	10/36 = 0.2778	25%	10000/36	25000000/36
3000	5/36 = 0.1389	15/36 = 0.4167		15000/36	45000000/36
3500	6/36 = 0.1667	21/36 = 0.5833	50%	21000/36	73500000/36
4000	5/36 = 0.1389	26/36 = 0.7222		20000/36	80000000/36
4500	4/36 = 0.1111	30/36 = 0.8333	75%	18000/36	81000000/36
5000	3/36 = 0.0833	33/36 = 0.9167	90%	15000/36	75000000/36
5500	2/36 = 0.0555	35/36 = 0.9722		11000/36	60500000/36
6000	1/36 = 0.0278	36/36 = 1.0000		6000/36	36000000/36
Summe:	36/36 = 1.0000			126000/36	493500000/36
				3500	13708333.33

Analog zu empirischen Verteilungen lassen sich auch für Zufallsvariablen weitere Kennwerte berechnen. Das arithmetische Mittel heißt bei Zufallsvariablen ***Erwartungswert*** μ_X („mü von X“) und ist die Summe aus den Ausprägungen mal deren Auftretenswahrscheinlichkeiten:

$$\mu(X) = \mu_X = \sum_{k=1}^{K} Pr\left(x_{(k)}\right) \cdot x_{(k)}$$

Im Beispiel ergibt sich ein Erwartungswert von 3500 €.

X (mittleres Einkomen in €)	Wahrscheinlich-keitsfunktion	Verteilungs-funktion	Quan-tile	$X \cdot Pr(X)$	$X^2 \cdot Pr(X)$
1000	1/36 = 0.0278	1/36 = 0.0278		1000/36	1000000/36
1500	2/36 = 0.0555	3/36 = 0.0833		3000/36	4500000/36
2000	3/36 = 0.0833	6/36 = 0.1667	10%	6000/36	12000000/36
2500	4/36 = 0.1111	10/36 = 0.2778	25%	10000/36	25000000/36
3000	5/36 = 0.1389	15/36 = 0.4167		15000/36	45000000/36
3500	6/36 = 0.1667	21/36 = 0.5833	50%	21000/36	73500000/36
4000	5/36 = 0.1389	26/36 = 0.7222		20000/36	80000000/36
4500	4/36 = 0.1111	30/36 = 0.8333	75%	18000/36	81000000/36
5000	3/36 = 0.0833	33/36 = 0.9167	90%	15000/36	75000000/36
5500	2/36 = 0.0555	35/36 = 0.9722		11000/36	60500000/36
6000	1/36 = 0.0278	36/36 = 1.0000		6000/36	36000000/36
Summe:	36/36 = 1.0000			126000/36	493500000/36
				3500	13708333.33

Die ***Varianz*** σ^2_X (ausgesprochen „sigma-Quadrat von X“) ist der Erwartungswert der quadrierten Abweichungen vom Erwartungswert:

$$\sigma^2(X) = \sigma^2_X = \sum_{k=1}^{K} Pr\left(x_{(k)}\right) \cdot \left(x_{(k)} - \mu_X\right)^2 = \sum_{k=1}^{K} Pr\left(x_{(k)}\right) \cdot x^2_{(k)} - \mu^2_X$$

Im Beispiel beträgt die Varianz $\sigma^2(X)$ = 1 458 333.33 €² (=13708333.33–3500²) und die Standardabweichung $\sigma(X)$ = 1207.61 €.

Das Gesetz der großen Zahl

Die Auftretenswahrscheinlichkeiten der Ausprägungen von Zufallsvariablen in Wahrscheinlichkeitsverteilungen entsprechen den relativen Häufigkeiten von Realisierungen in empirischen Verteilungen.
Es scheint also eine Ähnlichkeit zwischen relativen Häufigkeiten und Wahrscheinlichkeiten zu geben.

Diese Ähnlichkeit wird in der ***frequentistischen Definition der Wahrscheinlichkeit*** genutzt:

> Die Wahrscheinlichkeit Pr(A) eines Ereignisses A ist gleich dem Grenzwert der relativen Auftretenshäufigkeit n_A / n dieses Ereignisses, wenn die Zahl der Wiederholungen n des Zufallsexperiments, zu dessen Ereignissen A gehört, über alle Grenzen wächst:

$$\lim_{n \to \infty} \left(\frac{n_A}{n} \right) = Pr(A)$$

Die frequentistische Wahrscheinlichkeitsdefinition führt zu einem scheinbar empirischen Wahrscheinlichkeitsbegriff, da Wahrscheinlichkeiten nach dieser Definition praktisch relative Häufigkeiten sind.
Da es aber empirisch unmöglich ist, Zufallsexperimente tatsächlich unendlich oft zu wiederholen, können sie nicht direkt beobachtet werden.

Das Gesetz der großen Zahl

Begründet wird die frequentistische Sicht auf Wahrscheinlichkeit oft durch das

Gesetz der großen Zahl:

Wenn die Zahl n der Wiederholungen eines Zufallsexperiments über alle Grenzen steigt, dann nähert sich die Wahrscheinlichkeit, dass der Abstand der relativen Häufigkeit n_A/n eines Ereignisses A von der Wahrscheinlichkeit Pr(A) dieses Ereignisses im einfachen Zufallsexperiment kleiner oder gleich einer beliebig kleinen positiven Zahl ε ist, dem Wert Eins an.

$$\lim_{n\to\infty}\left(\Pr\left(\left|\frac{n_A}{n}-\Pr(A)\right|<\varepsilon\right)\right)=1$$

Das Gesetz der großen Zahl lässt sich formal beweisen, soll hier allerdings nur an einem Beispiel verdeutlicht werden.

Für das Beispiel wird eine Münze wiederholt geworfen. Ein solcher Münzwurf lässt sich als Zufallsexperiment mit zwei möglichen Ergebnissen „Kopf“ und „Zahl“auffassen, die im folgenden durch die Buchstaben A für „Kopf“ und B für „Zahl“ symbolisiert werden. Entsprechend der klassischen Wahrscheinlichkeitsdefinition wird unterstellt, dass die Realisierungswahrscheinlichkeit jedes der beiden Ergebnisse 0.5 beträgt. Denkbar sind aber auch beliebige andere Werte, die sich zu 1.0 summieren.

Bei z.B. 3 Wiederholungen sind 8 (= 2·2·2) Ergebnisse möglich:
(A,A,A), (A,A,B), (A,B,A), (B,A,A), (A,B,B), (B,A,B), (B,B,A) und (B,B,B)

Das Gesetz der großen Zahl

Das Zufallsexperiment des Münzwurfs entspricht einer einfachen Zufallsauswahl von n aus N=2 Elementen (nämlich A und B) mit Zurücklegen.

Die Auftretenswahrscheinlichkeit jeder Stichprobe bei Berücksichtigung der Anordnung beträgt daher N^{-n}, hier also bei 3 Wiederholungen $2^{-3} = 1/8$.

Soll die relative Häufigkeit des Ereignisses A („Kopf“) berechnet werden, interessiert allerdings nicht die Wahrscheinlichkeit einer beliebigen Stichprobe, sondern die Wahrscheinlichkeit, dass bei n Wiederholungen genau n_A mal „Kopf“ vorkommt. Wenn dies der Fall ist, muss $n–n_A = n_B$ mal „Zahl“ vorkommen.

Die Wahrscheinlichkeit, dass bei n Wiederholungen insgesamt n_A mal „Kopf“ vorkommt, ist dann gleich der Auftretenswahrscheinlichkeit einer einfachen Zufallsauswahl von n aus N=2 mit Zurücklegen und n_A und n_B Wiederholungen. Die Wahrscheinlichkeit berechnet sich somit nach:

$$\Pr(n_A)=\frac{n!}{N^n\cdot n_A!\cdot n_B!}=\frac{n!}{n_A!\cdot n_B!}\cdot\left(\frac{1}{N}\right)^n=\frac{n!}{n_A!\cdot(n-n_A)!}\cdot\left(\frac{1}{2}\right)^n=\binom{n}{n_A}\cdot 0.5^n$$

Da die relative Auftretenshäufigkeit p_A von A („Kopf“) der Quotient n_A/n ist, lassen sich auch die Auftretenswahrscheinlichkeiten aller realisierbaren relativen Häufigkeiten von A über diese Formel berechnen:

$$\Pr\left(\frac{n_A}{n}\right)=\frac{n!}{N^n\cdot n_A!\cdot n_B!}=\frac{n!}{n_A!\cdot(n-n_A)!}\cdot\left(\frac{1}{2}\right)^n=\binom{n}{n_A}\cdot 0.5^n$$

Das Gesetz der großen Zahl

Als Beispiel werden die Wahrscheinlichkeiten bei n=3 berechnet. Jede der 8 (=2^3) Stichproben (A,A,A), (A,A,B), (A,B,A), (B,A,A), (A,B,B), (B,A,B), (B,B,A) und (B,B,B) hat eine Auftretenswahrscheinlchkeit von 1/8.

Die Wahrscheinlichkeiten der relativen Häufigkeiten des Ereignisses A („Kopf“) berechnen sich nach:

Ereignis keinmal „Kopf“ (B,B,B):

n=3, n_A= 0 $\Rightarrow Pr(p_A = 0/3) = 3!/(0!\cdot 3!) \cdot 0.5^3 = 1/8 = 0.125$

Ereignisse einmal „Kopf“ (A,B,B), (B,A,B), (B,B,A):

n=3, n_A= 1 $\Rightarrow Pr(p_A = 1/3) = 3!/(1!\cdot 2!) \cdot 0.5^3 = 3/8 = 0.375$

Ergebnisse zweimal „Kopf“ (A,A,B), (B,A,A), (A,B,A):

n=3, n_A= 2 $\Rightarrow Pr(p_A = 2/3) = 3!/(2!\cdot 1!) \cdot 0.5^3 = 3/8 = 0.375$

Ereignis dreimal „Kopf“ (A,A,A):

n=3, n_A= 3 $\Rightarrow PR(p_A = 3/3) = 3!/(3!\cdot 0!) \cdot 0.5^3 = 1/8 = 0.125$

Über die Wahrscheinlichkeiten der Anteile lässt sich anschließend ausrechnen, wie wahrscheinlich es ist, dass die realisierte relative Häufigkeit innerhalb eines Intervalls liegt.

So kann z.B. berechnet werden, wie wahrscheinlich es ist, dass die relative Häufigkeit des Ereignisses A („Kopf“) beim dreimaligen Werfen einer Münze im Intervall 0.5 ± 0.2 = 0.3 bis 0.7 liegt.

Bei n=3 Würfen beträgt diese Wahrscheinlichkeit 6/8, nämlich die Summe der Auftretenswahrscheinlichkeit des Ereignissen 1 mal A (→ rel. Häufigkeit: 1/3 > 0.3) und 2 mal A (→ rel. Häufigkeit: 2/3 < 0.7).

Das Gesetz der großen Zahl

Wahrscheinlichkeit einer relativen Häufigkeit zwischen 0.3 und 0.7 bei n=1 bis n=18 Wiederholungen eines ausgewogenen Müpze.

n	$Pr(0.3 \le p_A \le 0.7)$	n	$Pr(0.3 \le p_A \le 0.7)$	n	$Pr(0.3 \le p_A \le 0.7)$
1	0.00	7	0.55	13	0.91
2	0.50	8	0.71	14	0.82
3	0.75	9	0.82	15	0.88
4	0.38	10	0.66	16	0.92
5	0.63	11	0.77	17	0.86
6	0.78	12	0.85	18	0.90

Simulation von 50000 Würfen einer Münze

n	p_A	p_A–0.5
10	.200	–.300
100	.500	.000
500	.524	.024
1000	.474	–.026
5000	.495	–.005
10000	.507	.007
50000	.504	.004

Die Tabelle zeigt die entsprechenden Wahrscheinlichkeiten von n=1 bis n=18 Wiederholungen des Münzwurfs. Ganz entsprechend dem Gesetz der großen Zahl steigt – mit gewissen Schwankungen – die Wahrscheinlichkeit immer mehr an, dass die relative Häufigkeit Kopf im Intervall 0.5±0.2 liegt.

Ein ähnliches Egebnis ergibt auch der empirische Versuch.

So zeigt die Tabelle rechts oben den Anteil des Ereignisses A, wenn in einer Computersimulation tatsächlich wiederholt eine Münze geworfen wird, deren Auftretenswahrscheinlichkeit von „Kopf“ 0.5 ist. Je mehr die Anzahl der Wiederholungen ansteigt, um so mehr nähert sich die relative Häufigkeit von „Kopf“ der Auftretenswahrscheinlichkeit 0.5 an.

Wahrscheinlichkeit und relative Häufigkeit

Obwohl es also eine Beziehung zwischen empirischen relativen Häufigkeiten und Wahrscheinlichkeiten zu geben scheint, sollte doch klar sein, dass der statistische Begriff der „Wahrscheinlichkeit“ eine theoretische Modellvorstellung ist und kein reales (empirisches) Phänomen bezeichnet.

Im Gesetz der großen Zahl wird zwar eine Beziehung zwischen empirischen relativen Häufigkeiten und Wahrscheinlichkeiten hergestellt.
Gleichwohl beinhaltet der frequentistische Wahrscheinlichkeitsbegriff einen (fehlerhaften) Zirkelschluss, falls er mit dem Gesetz der großen Zahl begründet wird: Im Gesetz der großen Zahl taucht nämlich bereits der Begriff der Wahrscheinlichkeit auf, der doch erst durch den frequentistische Begriff definiert werden soll.

> Die frequentistische Definition wäre erst dann nicht zirkulär, wenn es gelänge, die Forderung der „Wiederholung eines Zufallsexperiments unter gleichen Bedingungen“ unabhängig vom Begriff der statistischen Unabhängigkeit zweier Ereignisse zu definieren.

Ungeachtet dieses logischen Problems führt der frequentistische Wahrscheinlichkeitsbegriff jedoch zu einer intuitiven und hilfreichen Vorstellung der Bedeutung des Wortes „Wahrscheinlichkeit“.
Ein Vorteil gegenüber dem klassischen Wahrscheinlichkeitsbegriff liegt insbesondere auch darin, dass nicht unterstellt werden muss, dass Elementarereignisse mit gleicher Wahrscheinlichkeit auftreten müssen. Stattdessen kann durch Wiederholen gewissermaßen empirisch „geprüft“ werden, ob z.B. eine Münze oder ein Würfel ausgewogen ist, d.h. zu gleichwahrscheinlichen Ergebnissen führt.

Kennwerteverteilungen als Verbindungsglied zwischen Populationsparameter und Stichprobenstatistiken

Das Beispiel der zufälligen Auswahl von n=2 aus N=6 Haushalten zeigt, dass bezogen auf eine konkrete Stichprobe das durchschnittliche Haushaltseinkommen in dieser Stichprobe ein Kennwert der empirischen Einkommensverteilung in der Stichprobe ist, bezogen auf die Wahrscheinlichkeitsverteilung der durchschnittlichen Haushaltseinkommen in den Stichproben dagegen eine Realisierung einer Zufallsvariable.

Ziel der Berechnung eines Stichprobenmittelwerts ist i.a. die Schätzung eines Populationskennwertes, im Beispiel des durchschnittlichen Haushaltseinkommens in der Population. Die Kennwerte einer Population heißen auch ***Populationsparameter***. Da die Werte von Populationsparametern in der Regel unbekannt sind, sollen sie mit Hilfe von Stichprobendaten ***geschätzt*** werden. Hierzu werden aus den Realisierungen einer Stichprobe möglichst geeignete Kennwerte berechnet, die in der Statistik „***Stichprobenstatistiken***“ oder einfach „***Statistiken***“ heißen.
Bezogen auf die Menge der möglichen Stichproben ist eine Statistik eine Zufallsvariable, deren Wahrscheinlichkeitsverteilung auch als ***Kennwerteverteilung*** bezeichnet wird, da es sich um die (Wahrscheinlichkeits-) Verteilung von Stichprobenkennwerten über verschiedene Stichproben handelt.

Insgesamt müssen daher drei Arten von Verteilungen unterschieden werden:
1. die Verteilung aller Elemente in einer Population,
2. die Verteilung von Realisationen in einer Zufallsstichprobe aus dieser Population und
3. die Kennwerteverteilung von Stichprobenstatistiken über alle möglichen Stichproben.

Stichprobenkennwerte, Kennwerteverteilungen und Populationsparameter

Populationsverteilung:
Haush.

einkom.	n_k	p_k	cp_k	$p_k \cdot x_k$	$p_k \cdot (x_k)^2$
1000	1	1/6	1/6	166.67	166666.67
2000	1	1/6	2/6	333.33	666666.67
3000	1	1/6	3/6	500.00	1500000.00
4000	1	1/6	4/6	666.67	2666666.67
5000	1	1/6	5/6	833.3	4166666.67
6000	1	1/6	6/6	1000.00	6000000.00
Summe:	6	6/6		3500.00	15166666.67

$$\bar{X} = 3500.00$$

$$S_X^2 = 15166666.67 - 3500^2$$

$$= 2916666.67$$

$$S_X = \sqrt{2916666.67} = 1707.83$$

Als Beispiel wird wieder die Population von N=6 Haushalten herangezogen, aus denen in einer einfachen Zufallsauswahl mit Zurücklegen n=2 Haushalte ausgewählt werden. Die blau unterlegte Tabelle gibt die empirische Häufigkeitsverteilung dieser Population wieder.

Für diese Verteilung ergibt sich ein Mittelwert von $\bar{X}$= 3 500 € und eine Varianz von S^2_X = 2 916 666.67 €^{2} bzw. eine Standardabweichung von S_X = 1 707.83 €.
Zur Unterscheidung der drei Populationen sind hier die Populationsparameter durch große lateinische Buchstaben symbolisiert, also X-quer für den Mittelwert und S^2_X bzw S_X für die Varianz bzw. Standardabweichung.

Stichprobenkennwerte, Kennwerteverteilungen und Populationsparameter

Populationsverteilung:
Haush.

einkom.	n_k	p_k	cp_k
1000	1	1/6	1/6
2000	1	1/6	2/6
3000	1	1/6	3/6
4000	1	1/6	4/6
5000	1	1/6	5/6
6000	1	1/6	6/6
Summe:	6	6/6	

$$\bar{X} = 3500$$

$$S_X^2 = 2916666.67$$

$$S_X = 1707.83$$

Stichprobenverteilung 1
Haush. {1,1}

einkom.	n_k	p_k	cp_k	$p_k \cdot x_k$	$p_k \cdot (x_k)^2$
1000	1	0.5	0.5	500.00	500000.00
1000	1	0.5	1.0	500.00	500000.00
Summe:	2	1.0		1000.00	1000000.00

$$\bar{x}_1 = 1000 \; ; \; s_1^2 = 1000000 - 1000^2 = 0 \; ; \; s_1 = 0$$

Stichprobenverteilung 2
Haush. {1,2}

einkom.	n_k	p_k	cp_k	$p_k \cdot x_k$	$p_k \cdot (x_k)^2$
1000	1	0.5	0.5	500.00	500000.00
2000	1	0.5	1.0	1000.00	2000000.00
Summe:	2	1.0		1000.00	2500000.00

$$\bar{x}_2 = 1500 \; ; \; s_1^2 = 2500000 - 1500^2 = 250000 \; ; \; s_1 = 500$$

Die **Populationsverteilung** und deren Parameter sind in der Realität nicht oder nur bei sehr hohem Aufwand beobachtbar. Als Ersatz wird eine Stichprobe gezogen und aus der **Stichprobenverteilung** werden Stichprobenstatistiken berechnet.

Als Beispiel sind grün unterlegt zwei mögliche Stichproben dargestellt und deren Stichprobenmittelwerte und -varianzen berechnet, symbolisiert durch kleine mit der Stichprobennummer indizierte lateinische Buchstaben (z.B. s_1 und s_2 für Standardabweichungen).

Stichprobenkennwerte, Kennwerteverteilungen und Populationsparameter

Populationsverteilung:
Haush.

einkom.	n_k	p_k	cp_k
1000	1	1/6	1/6
2000	1	1/6	2/6
3000	1	1/6	3/6
4000	1	1/6	4/6
5000	1	1/6	5/6
6000	1	1/6	6/6
Summe:	6	6/6	

$\bar{X} = 3500$

$S_X^2 = 2916666.67$

$S_X = 1707.83$

Stichprobenverteilung 1
Haush. {1,1}

einkom.	n_k	p_k	cp_k
1000	1	0.5	0.5
1000	1	0.5	1.0
Summe:	2	1.0	

$\bar{x}_1 = 1000$

Stichprobenverteilung 2
Haush. {1,2}

einkom.	n_k	p_k	cp_k
1000	1	0.5	0.5
2000	1	0.5	1.0
Summe:	2	1.0	

$\bar{x}_2 = 1500$

Kennwerteverteilung:

$\bar{x}_i$ (mittleres Einkomen in €)	Wahrscheinlich-keitsfunktion	Verteilungs-funktion
1000	1/36	1/36
1500	2/36	3/36
2000	3/36	6/36
2500	4/36	10/36
3000	5/36	15/36
3500	6/36 =1/6	21/36
4000	5/36	26/36
4500	4/36	30/36
5000	3/36	33/36
5500	2/36	35/36
6000	1/36	36/36
Summe:	36/36	

$\mu(\bar{x}) = 3500 \; ; \; \sigma^2(\bar{x}) = 1458333.33 \; ; \; \sigma(\bar{x}) = 1207.61$

Die Wahrscheinlichkeitsverteilung der Stichprobenmittelwerte über alle Stichproben ergibt schließlich die ***Kennwerteverteilung*** (grau unterlegt), die das Verbindungsglied zwischen Stichprobe und Population ist.

Die Wahrscheinlichkeitsverteilung der Stichprobenmittelwerte wurde bereits als Beispiel für die Verteilung einer Zufallsvariable berechnet. Der Mittelwert der Kennwerteverteilung beträgt 3500 € und die Standardabweichung 1207.61 €

Stichprobenkennwerte, Kennwerteverteilungen und Populationsparameter

Populationsverteilung:
Haush.

einkom.	n_k	p_k	cp_k
1000	1	1/6	1/6
2000	1	1/6	2/6
3000	1	1/6	3/6
4000	1	1/6	4/6
5000	1	1/6	5/6
6000	1	1/6	6/6
Summe:	6	6/6	

$\bar{X} = 3500$

$S_X^2 = 2916666.67$

$S_X = 1707.83$

Stichprobenverteilung 1
Haush. {1,1}

einkom.	n_k	p_k	cp_k
1000	1	0.5	0.5
1000	1	0.5	1.0
Summe:	2	1.0	

$\bar{x}_1 = 1000$

Stichprobenverteilung 2
Haush. {1,2}

einkom.	n_k	p_k	cp_k
1000	1	0.5	0.5
2000	1	0.5	1.0
Summe:	2	1.0	

$\bar{x}_2 = 1500$

Kennwerteverteilung:

$\bar{x}_i$ (mittleres Einkomen in €)	Wahrscheinlich-keitsfunktion	Verteilungs-funktion
1000	1/36	1/36
1500	2/36	3/36
2000	3/36	6/36
2500	4/36	10/36
3000	5/36	15/36
3500	6/36 =1/6	21/36
4000	5/36	26/36
4500	4/36	30/36
5000	3/36	33/36
5500	2/36	35/36
6000	1/36	36/36
Summe:	36/36	

2/3 = { 4/36 (2500), 5/36 (3000), 6/36 =1/6 (3500), 5/36 (4000), 4/36 (4500) }

$\mu(\bar{x}) = 3500 \; ; \; \sigma^2(\bar{x}) = 1458333.33 \; ; \; \sigma(\bar{x}) = 1207.61$

Die ***Kennwerteverteilung*** ermöglicht Aussagen über die Risiken des Induktionsschlusses.

Im Beispiel lässt sich so aus der Kennwerteverteilung ablesen, dass mit einer Wahrscheinlichkeit von 1/6 ein Stichprobenmittelwert genau mit dem Populationsmittelwert (3500 €) übereinstimmt und mit einer Wahrscheinlichkeit von 2/3 der Stichprobenmittelwert um maximal 1000 € vom Populationsmittelwert abweicht.

Stichprobenkennwerte, Kennwerteverteilungen und Populationsparameter

Populationsverteilung:
Haush.

einkom.	n_k	p_k	cp_k
1000	1	1/6	1/6
2000	1	1/6	2/6
3000	1	1/6	3/6
4000	1	1/6	4/6
5000	1	1/6	5/6
6000	1	1/6	6/6
Summe:	6	6/6	

$\bar{X} = 3500$

$S_X^2 = 2916666.67$

$S_X = 1707.83$

Stichprobenverteilung 1
Haush. {1,1}

einkom.	n_k	p_k	cp_k
1000	1	0.5	0.5
1000	1	0.5	1.0
Summe:	2	1.0	

$\bar{x}_1 = 1000$

Stichprobenverteilung 2
Haush. {1,2}

einkom.	n_k	p_k	cp_k
1000	1	0.5	0.5
2000	1	0.5	1.0
Summe:	2	1.0	

$\bar{x}_2 = 1500$

Kennwerteverteilung:

$\bar{x}_i$ (mittleres Einkomen in €)	Wahrscheinlichkeitsfunktion	Verteilungsfunktion
1000	1/36	1/36
1500	2/36	3/36
2000	3/36	6/36
2500	4/36	10/36
3000	5/36	15/36
3500	2/3 = 6/36 = 1/6	21/36
4000	5/36	26/36
4500	4/36	30/36
5000	3/36	33/36
5500	2/36	35/36
6000	1/36	36/36
Summe:	36/36	

$\mu(\bar{x}) = 3500 \; ; \; \sigma^2(\bar{x}) = 1458333.33 \; ; \; \sigma(\bar{x}) = 1207.61$

Wenn in der Statistik Aussagen über die Güte von Schätzungen getroffen werden, beziehen sich diese nicht auf eine konkrete Stichprobe, sondern immer auf die Kennwerteverteilung der Statistik, die zur Schätzung eines Populationsparameters herangezogen wird.

Stichprobenkennwerte, Kennwerteverteilungen und Populationsparameter

Populationsverteilung:
Haush.

einkom.	n_k	p_k	cp_k
1000	1	1/6	1/6
2000	1	1/6	2/6
3000	1	1/6	3/6
4000	1	1/6	4/6
5000	1	1/6	5/6
6000	1	1/6	6/6
Summe:	6	6/6	

$\bar{X} = 3500$

$S_X^2 = 2916666.67$

$S_X = 1707.83$

Stichprobenverteilung 1
Haush. {1,1}

einkom.	n_k	p_k	cp_k
1000	1	0.5	0.5
1000	1	0.5	1.0
Summe:	2	1.0	

$\bar{x}_1 = 1000$

Stichprobenverteilung 2
Haush. {1,2}

einkom.	n_k	p_k	cp_k
1000	1	0.5	0.5
2000	1	0.5	1.0
Summe:	2	1.0	

$\bar{x}_2 = 1500$

Kennwerteverteilung:

$\bar{x}_i$ (mittleres Einkomen in €)	Wahrscheinlichkeitsfunktion	Verteilungsfunktion
1000	1/36	1/36
1500	2/36	3/36
2000	3/36	6/36
2500	4/36	10/36
3000	5/36	15/36
3500	2/3 = 6/36 = 1/6	21/36
4000	5/36	26/36
4500	4/36	30/36
5000	3/36	33/36
5500	2/36	35/36
6000	1/36	36/36
Summe:	36/36	

$\mu(\bar{x}) = 3500 \; ; \; \sigma^2(\bar{x}) = 1458333.33 \; ; \; \sigma(\bar{x}) = 1207.61$

Ein konkreter Stichprobenmittelwert kann vom gesuchten Populationsparameter trotz hoher Stichprobengüte sehr stark abweichen.

So sind in den beiden aufgelisteten Stichprobenverteilungen 1 und 2 die Stichprobenmittelwerte mit Werten von 1000€ und 1500€ deutlich vom Populationsmittelwert mit 3500€ entfernt, obwohl der Erwartungswert der Kennwerteverteilung mit dem Populationsmittelwert übereinstimmt.

Stichprobenkennwerte, Kennwerteverteilungen und Populationsparameter

Populationsverteilung:

Haush. einkom.	n_k	p_k	cp_k
1000	1	1/6	1/6
2000	1	1/6	2/6
3000	1	1/6	3/6
4000	1	1/6	4/6
5000	1	1/6	5/6
6000	1	1/6	6/6
Summe:	6	6/6	

$\bar{X} = 3500$

$S_X^2 = 2916666.67$

$S_X = 1707.83$

Kennwerteverteilung:

$\bar{x}_i$ (mittleres Einkomen in €)	Wahrscheinlich-keitsfunktion	Verteilungs-funktion
1000	1/36	1/36
1500	2/36	3/36
2000	3/36	6/36
2500	4/36	10/36
3000	5/36	15/36
3500	6/36 =1/6	21/36
4000	5/36	26/36
4500	4/36	30/36
5000	3/36	33/36
5500	2/36	35/36
6000	1/36	36/36
Summe:	36/36	

$$\sigma^2(\bar{x}) = \frac{S_X^2}{2} = \frac{2916666.67}{2} = 1458333.33$$

$\mu(\bar{x}) = 3500$

$\sigma^2(\bar{x}) = 1458333.33$

$\sigma(\bar{x}) = 1207.61$

Die Kennwerteverteilung von Statistiken zur Schätzung von Popualtionsparamtern weist i.a. Beziehungen zur Populationsverteilung auf.

So gilt für das Beispiel, das der Erwartungswert der Kennwerteverteilung gleich dem Populationsmittelwert ist und dass die Varianz der Kennwerteverteilung gleich 1/2 mal der Populationsvarianz ist.

Lerneinheit 11 : Diskrete Wahrscheinlichkeitsverteilungen

Die Binomialverteilung

Im Beispiel zum Gesetz der großen Zahl (s. Lerneinheit L10) wurde die Wahrscheinlichkeit berechnet, mit der bei n Würfen einer Münze n_A bzw $p_A = n_A/n$ mal das Ereignis A („Kopf“) auftritt. Dabei wurde unterstellt, dass die Wahrscheinlichkeit von „Kopf“ wie „Zahl“ jeweils 0.5 beträgt. Wie ändert sich die Wahrscheinlichkeiten, wenn die Auftretenswahrscheinlichkeit Pr(A) nicht 0.5, sondern eine beliebige Zahl π_A zwischen 0 und 1 ist?

Wenn die Wahrscheinlichkeit von A Pr(A), im folgenden als π_A bezeichnet, z. B. 0.4 beträgt, dann ist die Wahrscheinlichkeit des komplementären Ereignissen $\neg A = B$ (im Beispiel des Münzwurfs also „Zahl“) $Pr(B) = \pi_B = 1 - 0.4 = 0.6$.

Generell gilt bei der Betrachtung komplementärer Ereignisse A und B: $\pi_B = 1 - \pi_A$.

Wenn nun in den n Wiederholungen des Zufallsexperiemnts n_A mal A auftritt, muss entsprechend $n_B = n - n_A$ mal B auftreten. Die n Wiederholungen des Zufallsexperiments sind statistisch unabhängig voneinander. Somit ist die Realisierungswahrcheinlichkeit einer beliebigen Folge von Ergebnissen gleich dem Produkt der Realisierungswahrscheinlichkeiten jeder einzelnen Wiederholung.

Wenn z.B. bei n=5 Wiederholungen die Folge (A,B,A,A,B) auftritt, dann ist bei $\pi_A = 0.4$ die Realisierungswahrscheinlichkeit dieser Folge:
$Pr(A,B,A,A,B) = 0.4 \cdot 0.6 \cdot 0.4 \cdot 0.4 \cdot 0.6 = 0.4^3 \cdot 0.6^2$.
Die gleiche Realisierungswahrscheinlichkeit hat jede andere Folge, bei der bei n=5 Wiederholungen $n_A = 3$ mal A und $n_B = 2$ mal B auftritt.

Die Binomialverteilung

Im allgemeinen Fall tritt das Ereignis A eines Zufallsexperiments mit Realisierungswahrscheinlichkeit π_A und das komplementäre Ereignis $B = \neg A$ mit Wahrscheinlichkeit π_B auf, so dass $\pi_A + \pi_B = 1$.

Wenn in einer Folge von n Wiederholungen des Zufallsexperiments n_A mal A und n_B mal B realisiert wird, beträgt die Realisierungswahrscheinlichkeit dieser Folge bei Berücksichtigung der Reihenfolge der Ziehung entsprechend:

$$\Pr(n_A) = (\pi_A)^{n_A} \cdot (\pi_B)^{n_B} = \pi_A^{n_A} \cdot (1-\pi_A)^{n-n_A}$$
$$= (1-\pi_B)^{n-n_B} \cdot \pi_B^{n_B} = \Pr(n_B)$$

Solange nur die Häufigkeiten n_A und n_B interessieren, ist die Reihenfolge der Ziehung der Elemente irrelevant. Die Zahl der ununterscheidbaren Stichproben ist dann gleich der Zahl der Permutationen $P(n_A,n_B)$, unterschiedliche Folgen (Anordnungen) von n_A mal A und n_B mal B zu realisieren:

$$P(n_A, n_B) = \frac{n!}{n_A! \cdot n_B!} = \frac{n!}{n_A! \cdot (1-n_A)!} = \binom{n}{n_A}$$
$$= \frac{n!}{(1-n_B)! \cdot n_B!} = \binom{n}{n_B}$$

Die Binomialverteilung

Die gesuchte Wahrscheinlichkeit n_A mal A und n_B mal B ohne Berücksichtigung der Anordnungsreihenfolge zu erhalten, wobei $n_A + n_B = n$, ist wegen der Disjunktheit der verschiedenen Anordnungen die Summe der Wahrscheinlichkeiten aller Anordnungen mit gleichem n_A und n_B damit:

$$\Pr(n_A) = \binom{n}{n_A} \cdot \pi_A^{n_A} \cdot (1-\pi_A)^{n-n_A} = \frac{n!}{n_A! \cdot (1-n_A)!} \cdot \pi_A^{n_A} \cdot (1-\pi_A)^{n-n_A}$$

$$= \binom{n}{n_B} \cdot \pi_B^{n_B} \cdot (1-\pi_B)^{n-n_B} = \frac{n!}{n_B! \cdot (1-n_B)!} \cdot \pi_B^{n_B} \cdot (1-\pi_B)^{n-n_B} = \Pr(n_B)$$

Diese Wahrscheinlichkeit gilt für alle Zufallsexperimente, bei denen bei n Wiederholungen, die Auftretenshäufigkeit eines Ereignisses A interessiert.

Ein Beispiel ist das Urnenmodell mit N Kugeln, von denen N_A als A gekennzeichnet und N_B als B gekennzeichnet sind, wenn wiederholt jeweils eine Kugel ausgewählt und diese anschließend wieder zurück in die Urne gelegt wird. Die Wahrscheinlichkeit, bei einer Ziehung eine als A gekennzeichnete Kugel zu ziehen ist bei jeder Ziehung $\pi_A = N_A/N$, die komplementäre Wahrscheinlichkeit $\pi_B = N_B/N = 1 - \pi_A$.

Da in der statistischen Datenanalyse Ereignisse i.a. durch Zahlenwerte numerischer Variablen bezeichnet werden, kann das Auftreten des Ereignisses A auch durch eine dichotome Variable A beschrieben werden, die den Wert A=1 aufweist, wenn A auftritt, und den Wert A=0, wenn nicht A, sondern B auftritt.

Die Binomialverteilung

Anstelle von π_A und π_B werden die Realisierungwahrscheinlichkeiten daher meist durch π_1 für das Ereignis A=1 und $\pi_0 = 1 - \pi_1$ für das Ereignis A=0 bezeichnet.
Die Wahrscheinlichkeit, dass bei n Wiederholungen n_1 mal das Ereignis A=1 und enstprechend $n_0 = n - n_1$ mal das Ereignis A=0 auftritt, kann dann als Wert der Wahrscheinlichkeitsfunktion einer Zufallsvariable X mit den möglichen Ausprägungen x = 0, 1, 2, ..., n modelliert werden.

Die Realisierungswahrscheinlichkeiten der Ausprägungen dieser Zufallsvariable X berechnen sich dann als Wahrscheinlichkeiten, dass in den n Wiederholungen X = x mal das Ereignis A=1 auftritt:

$$\Pr(X = x \mid n, \pi_1) = \binom{n}{x} \cdot \pi_1^x \cdot (1-\pi_1)^{n-x} \quad \text{mit } x = 0,1,2,\ldots,n$$

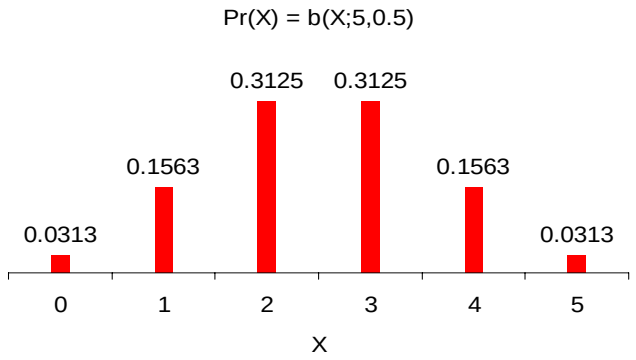
Die nach dieser Formel berechnete Wahrscheinlichkeitsverteilung einer Zufallsvariablen X wird ***Binomialverteilung*** genannt, wobei

- π_1 die Wahrscheinlichkeit ist, mit der ein interessierende Ereignis (A=1) im Zufallsexperiment auftritt und
- n die Zahl der unabhängigen Wiederholungen des Zufallsexperiments ist.

X ist dann ***binomialverteilt*** mit den ***Parametern*** **n** und **π_1**, was auch als **b(X; n, π_1)** symbolisiert wird.

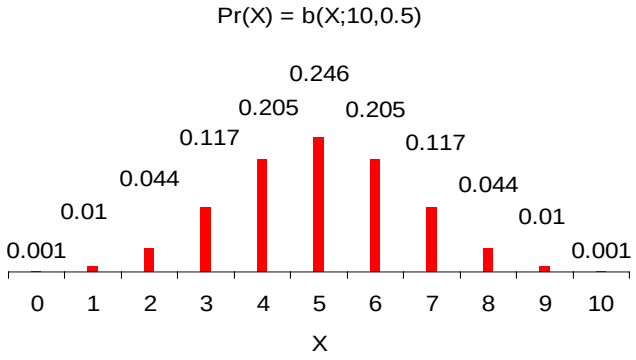
Als Beispiele werden im folgenden die Binomialverteilungen b(X; 5, 0.5), b(X; 10, 0.5), b(X; 10, 0.4) und b(X; 10, 0.7) betrachtet.

Die Binomialverteilung



$\mu_X = 2.5\ ;\ \sigma_X^2 = 7.5 - 2.5^2 = 1.25$

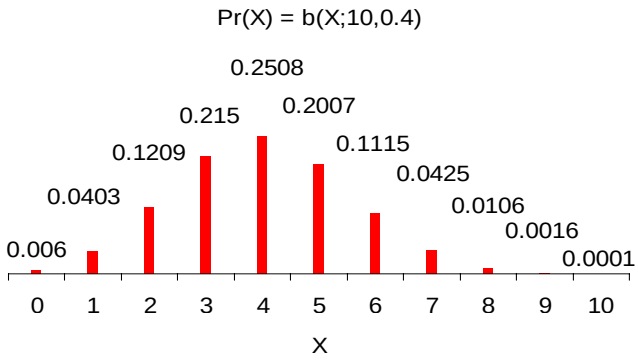
X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.03125	0.03125	0.00000	0.00000
1	0.15625	0.18750	0.15625	0.15625
2	0.31250	0.50000	0.62500	1.25000
3	0.31250	0.81250	0.93750	2.81250
4	0.15625	0.96875	0.62500	2.50000
5	0.03125	1.00000	0.15625	0.78125
			2.50000	7.50000



$\mu_X = 5.0\ ;\ \sigma_X^2 = 27.5 - 5.0^2 = 2.5$

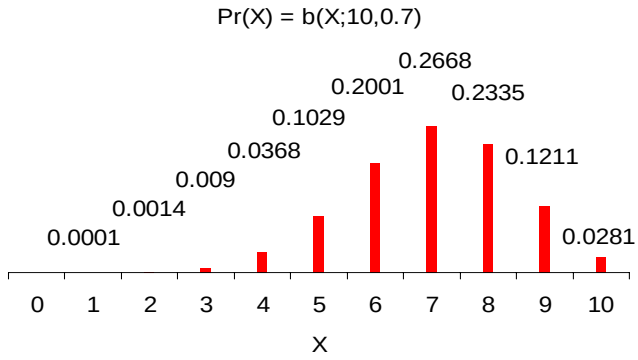
X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.00098	0.00098	0.00000	0.00000
1	0.00977	0.01074	0.00977	0.00977
2	0.04395	0.05469	0.08789	0.17578
3	0.11719	0.17188	0.35156	1.05469
4	0.20508	0.37695	0.82031	3.28125
5	0.24609	0.62305	1.23047	6.15234
6	0.20508	0.82813	1.23047	7.38281
7	0.11719	0.94531	0.82031	5.74219
8	0.04395	0.98926	0.35156	2.81250
9	0.00977	0.99902	0.08789	0.79102
10	0.00098	1.00000	0.00977	0.09766
			5.0000	27.50000

Die Binomialverteilung



$\mu_X = 4.0\ ;\ \sigma_X^2 = 18.4 - 4.0^2 = 2.4$

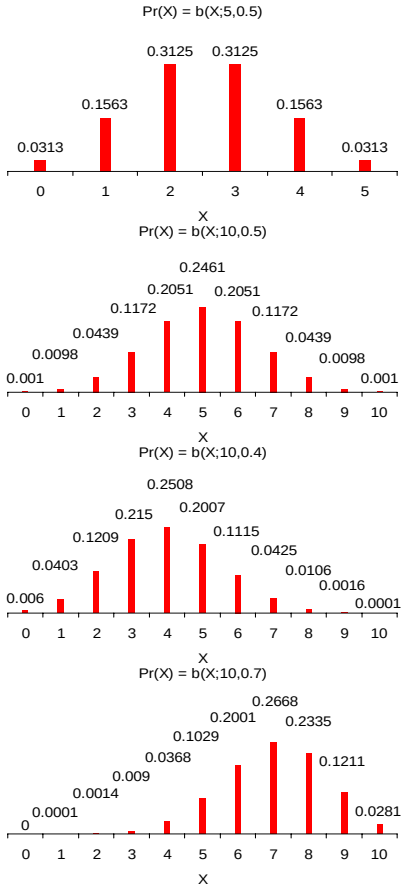
X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.00605	0.00605	0.00000	0.00000
1	0.04031	0.04636	0.04031	0.04031
2	0.12093	0.16729	0.24186	0.48373
3	0.21499	0.38228	0.64497	1.93492
4	0.25082	0.63310	1.00329	4.01316
5	0.20066	0.83376	1.00329	5.01645
6	0.11148	0.94524	0.66886	4.01316
7	0.04247	0.98771	0.29727	2.08090
8	0.01062	0.99832	0.08493	0.67948
9	0.00157	0.99990	0.01416	0.12740
10	0.00010	1.00000	0.00105	0.01049
			4.00000	18.40000



$\mu_X = 7.0\ ;\ \sigma_X^2 = 51.1 - 7.0^2 = 2.1$

X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.00001	0.00001	0.00000	0.00000
1	0.00014	0.00014	0.00014	0.00014
2	0.00145	0.00159	0.00289	0.00579
3	0.00900	0.01059	0.02701	0.08102
4	0.03676	0.04735	0.14703	0.58811
5	0.10292	0.15025	0.51460	2.57298
6	0.20012	0.35039	1.20073	7.20435
7	0.26683	0.61722	1.86780	13.07457
8	0.23347	0.85069	1.86780	14.94236
9	0.12106	0.97175	1.08955	9.80593
10	0.02825	1.00000	0.28248	2.82475
			7.0000	51.10000

Die Binomialverteilung



Wie die grafische Darstellung verdeutlicht, hängt die Verteilungsform vor allem von der Wahrscheinlichkeit π_1 ab, mit der das Ereignis A (bzw. die Ausprägung A=1 der dichotomen Variable A) auftritt.
Ist $\pi_1 = 0.5$ und dann wegen der Bedingung $\pi_1 + \pi_0 = 1$ also auch $\pi_0 = 0.5$, dann ist die Verteilung symmetrisch. Ist dagegen $\pi_1 < 0.5$ (und entsprechend $\pi_0 > 0.5$), dann ist die Verteilungs rechtsschief, ist $\pi_1 > 0.5$ (und entsprechend $\pi_0 < 0.5$), ist die Verteilungs linksschief.
Mit steigender Zahl der Wiederhoungen n erhöht sich die Zahl der möglichen Ausprägungen. Es kann gezeigt werden, dass sich dabei bei $\pi_1 \neq 0.5$ die Schiefe der Verteilung reduziert.

Aus der Wahrscheinlichkeitsfunktion lässt sich durch Aufsummieren die Verteilungsfunktion einer Binomialverteilung berechnen:

$$F\left(X = x \mid X \sim b\left(X; n, \pi_1\right)\right) = \Pr(X \le x)$$

$$= \sum_{j=0}^{x} \binom{n}{j} \cdot \pi_1^j \cdot \left(1 - \pi_1\right)^{n-j}$$

Die Bernoulli-Verteilung

Der einfachste Fall einer Binomialverteilung ergibt sich, wenn n=1 ist, also das Zufallsexperiment nur einmal ausgeführt wird. Eine solche Binomialverteilung b(X; 1, π_1) wird auch ***Punkt-Binomialverteilung*** oder nach dem Mathematiker Bernoulli ***Bernoulli-Verteilung*** genannt.

Für die Bernoulli-Verteilung gilt also:
$\Pr(X=1) = \pi_1$ und $\Pr(X=0) = \pi_0 = 1 - \pi_1$. und entsprechend $F(X=0) = 1 - \pi_1$ und $F(X=1) = 1$.
Erwartungswert und Varianz der Bernoulli-Verteilung sind dann:

$$\mu_X = \pi_0 \cdot 0 + \pi_1 \cdot 1 = \pi_1 \text{ und } \sigma_X^2 = \left(\pi_0 \cdot 0^2 + \pi_1 \cdot 1^2\right) - \pi_1^2 = \pi_1 \cdot \left(1 - \pi_1\right) = \pi_1 \cdot \pi_0$$

Da bei einer Binomialverteilung mit n > 1 n mal die Realisierung einer Bernoulli-Verteilung beobachtet wird und dabei die Zahl der Realisierungen A=1 aufsummiert wird, kann eine Binomialverteilung mit den Parametern n und π_1 und auch als Summe statistisch unabhängiger Bernoulli-Verteilungen mit jeweils gleichen Parameterwert π_1 aufgefasst werden. Die Binomialverteilung mit n > 1 ist also die Summe von n voneinander statistisch unabhängiger Bernoulli-Verteilungen mit gleichem Parameter π_1.

Dies kann noch verallgemeinert werden.

Wird das Zufallsexperiment zunächst n_1 mal durchgeführt, ergibt sich für die Summe X_1 der Ereignisse A=1 eine Binomialverteilung mit dem Parameter n_1 und π_1.
Werden anschließend weitere n_2 Versuche durchgeführt, ist die Summe X_2 der Ereignisse A=1 dieser zweiten Wiederholungsreihe binomialverteilt mit dem Parameter n_2 und π_1.
Außerdem ist dann die Gesamtsumme $X = X_1 + X_2$ der Ereignisse A=1 binomialverteilt mit dem Parameter $n=n_1+n_2$ und π_1.

Summen voneinander unabhängiger Binomialverteilungen

Das Beispiel verdeutlicht, dass jede Summe von statistisch unabhängigen Binomialverteilungen mit gleichem Parameter π_1 wiederum binomialverteilt ist:

Wenn X_1 binomialverteilt ist mit $b(X_1; n_1, \pi_1)$ und X_2 binomialverteilt mit $b(X_2; n_2, \pi_1)$, und X_1 und X_2 statistisch unabhängig voneinander sind, dann ist die Summe $Y = X_1 + X_2$ binomialverteilt mit $b(Y; n_1+n_2, \pi_1)$.

Dabei gibt es einen interessanten Zusammenhang zwischen den Erwartungswerten und Varianzen der Ausgangsverteilungen.

Dies kann am Beispiel einer Bernoulli-Verteilung mit $\pi_1 = 0.5$ verdeutlicht werden:
Wenn $\pi_1 = 0.5$, dann ist $\mu_X = 0.5$ und $\sigma^2_X = 0.5 \cdot(1-0.5) = 0.5^2 = 0.25$.
Die oben dargestellten Binomialverteilungen b(X; 5, 0.5) und b(X; 10, 0.5) haben Erwartungswerte $\mu(X|b(X; 5, 0.5)) = 2.5$ und $\mu(X|b(X;10, 0.5)) = 5$.
Da $2.5 = 5 \cdot 0.5$ und $5=10 \cdot 0.5$ weist dies darauf hin, dass der Erwartungswert der Summe von n Bernoulli-Verteilungen gleich n mal dem Erwartungswert der Bernoulli-Verteilung ist.
Gleiches zeigt sich bei den Varianzen: Die Varianz der Bernoulli-Verteilung mit $\pi_1 = 0.5$ ist 0.25, die der Binomialverteilung b(X; 5, 0.5) beträgt $1.25 = 5 \cdot 0.25$ und die der Binomialverteilung b(X;10, 0.5) beträgt $2.5 = 10 \cdot 0.25$.

Erwartungswert, Varianz und Schiefe von Binomialverteilungen

Das Beispiel demonstriert, dass Erwartungswert und Varianz einer Binomialverteilung eine einfache Funktion der Modellparameter n und π_1 sind:

$$\mu_X = n \cdot \pi_1 \text{ und } \sigma^2_X = n \cdot \pi_1 \cdot (1-\pi_1) \text{ wenn } X \sim b(X; n, \pi_1)$$

Das Symbol „~“ bedeutet hier, das X entsprechend der rechts folgenden Wahrscheinlichkeitsverteilung verteilt ist.

Da für das dritte zentrale Moment einer Binomialverteilung gilt:

$$\mu_3 = n \cdot \pi_1 \cdot (1-\pi_1) \cdot (1-2 \cdot \pi_1) = n \cdot \pi_1 \cdot \pi_0 \cdot (\pi_0 - \pi_1)$$

folgt für die Schiefe der Verteilung:

$$\frac{\mu_3}{\sigma^3} = \frac{1-2 \cdot \pi_1}{\sqrt{n \cdot \pi_1 \cdot (1-\pi_1)}} = \frac{\pi_0 - \pi_1}{\sqrt{n \cdot \pi_1 \cdot \pi_0}}$$

Aus der Gleichung für die Schiefe folgen die bereits erwähnten Eigenschaften der Binomialverteilung:

- ist $\pi_1 = \pi_0$, ist der Schiefekoeffizient Null und die Verteilung symmetrisch.
- Steigt die Zahl der Wiederholungen n an, wird die Schiefe geringer, da im Schiefekoeffizienten durch die Quadratwurzel von n geteilt wird und damit der Quotient bei steigendem n immer kleiner wird.

Erwartungswert und Varianz von Linearkombinationen unabhängiger Zufallsvariablen

Die Berechnung von Erwartungswerten und Varianzen von Summen unabhängiger Zufallsvariablen aus den Erwartungswerten und Varianzen der Summanden gilt nicht nur für die Binomialverteilung, sondern generell.

Wenn Y die Summe von statistisch unabhängigen Zufallsvariablen ist, dann ist der Erwartungswert und die Varianz von Y gleich der Summe der Erwartungswerte bzw. der Varianzen der Summanden:

$$Y = \sum_{i=1}^{n} X_i \Rightarrow \mu_Y = \sum_{i=1}^{n} \mu(X_i) \text{ und } \sigma_Y^2 = \sum_{i=1}^{n} \sigma^2(X_i)$$

Da zudem die Berechnung von Mittelwerten und Varinanzen von ***Lineartransformationen*** auch für Zufallsvariablen gilt:

$$Y = a + b \cdot X \Rightarrow \mu_Y = a + b \cdot \mu_X \text{ und } \sigma_Y^2 = b^2 \cdot \sigma_X^2$$

folgt bei Lineartransformationen der X_i für den Erwartungswert und die Varianz einer beliebigen ***Linearkombination*** von n statistisch unabhängigen Zufallsvariablen X_1, X_2, ..., X_n:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_n \cdot X_n$$

$$\Rightarrow \mu_Y = b_0 + \sum_{i=1}^{n} b_i \cdot \mu(X_i) \text{ und } \sigma_Y^2 = \sum_{i=1}^{n} b_i^2 \cdot \sigma^2(X_i)$$

Erwartungswert und Varianz von Linearkombinationen unabhängiger Zufallsvariablen

Der Erwartungswert und die Varianz einer Summe ist ein Spezialfall einer Linearkombination, bei der die Konstante $b_0 = 0$ und alle Gewichte $b_i = 1$ (i=1, 2, ..., n) sind.

Diese Eigenschaft gilt nur für ***statistisch unabhängige Variablen***.
Entsprechend der Definition, dass zwei Ereignisse eines Zufallsexperiment statistisch unabhängig sind, wenn die bedingten Wahrscheinlichkeiten gleich den unbedingten Wahrscheinlichkeiten sind oder – was gleichbedeutend ist – wenn die Wahrscheinlichkeit des gemeinsamen Auftretens der beiden Ereignisse gleich dem Produkt der Auftretenswahrscheinlichkeiten der Einzelereignisse ist, gilt für zwei Zufallsvariablen X und W:

Zwei Zufallsvariablen X und W sind *statistisch unabhängig* voneinander, wenn die Wahrscheinlichkeit des gemeinsamen Auftretens jeder beliebigen Kombination der Ausprägungen beider Variablen gerade das Produkt der Wahrscheinlichkeitsfunktionen ist:
$Pr(X \& Y) = Pr(X=x \cap Y=y) = Pr(X=x) \cdot Pr(Y=y)$ für alle Ausprägungen x und y.

Diese Definition ist für die Binomialverteilung als Summe von unabhängigen Bernoulli-Verteilungen erfüllt.

So ist die Auftretenswahrscheinlichkeit der Folge (0,0) = $\pi_0 \cdot \pi_0$, der Folge (0,1)= $\pi_0 \cdot \pi_1$, und der Folge (1,1) = $\pi_1 \cdot \pi_1$.

Aus der Binomialverteilung abgeleitete Wahrscheinlichkeitsverteilungen

Auf die Berechnung der Wahrscheinlichkeiten einer Binomialverteilung können weitere Wahrscheinlichkeitsverteilungen zurückgeführt werden. So interessiert oft die Wahrscheinlichkeit, wie viele Wiederholungen notwendig sind, um das erste Element mit der Eigenschaft A=1 zu erhalten.

> *So mag ein Biologe an der Untersuchung einer Spezies interessiert sein, die in einem Biotop mit einer Wahrscheinlichkeit von nur $\pi_1 = 0.01$ vorkommt. Wenn der Biologie eine einfache Zufallsauswahl von Organismen in dem Biotop zieht, wie groß sollte die Stichprobe sein, damit mit einer vorgegebenen Wahrscheinlichkeit damit zu rechen ist, dass mindestens ein Exemplar der interessierenden Spezies vorkommt.*

Im Unterschied zur Binomialverteilung wird hier also nicht bei gegebener Fallzahl nach der Zahl der Elemente mit der interessierenden Eigenschaft, sondern nach der Größe der Stichprobe gefragt, um genau 1 Element mit dieser Eigenschaft zu erhalten.

> *Wenn erst nach X=x Ziehungen das erste Mal ein Element mit der interessierenden Eigenschaft ausgewählt wird, dann müssen zunächst in x–1 Ziehungen x–1 Elemente ohne diese Eigenschaft ausgewählt worden sein.*

Da die Wahrscheinlichkeit π_1 beträgt, ein Element mit der interessierenden Eigenschaft auszuwählen, berechnet sich die gesuchte Wahrscheinlichkeit nach:

$$\Pr\left(X = x \middle| \pi_1\right) = \left(1 - \pi_1\right)^{x-1} \cdot \pi_1$$

Die geometrische Verteilung

Diese Wahrscheinlichkeitsverteilung wird als ***geometrische Verteilung*** bezeichnet. Die Wahrscheinlickeiten berechnen sich als Folgen der Ausprägung X=1 von Binomialverteilungen mit den Parametern n=1, 2, ... und π_1.

Der Erwartungswert der geometrischen Verteilung berechnet sich dann nach:

$$\mu_X = \pi_1 \cdot 1 + \left(1 - \pi_1\right) \cdot \pi_1 \cdot 2 + \left(1 - \pi_1\right)^2 \cdot \pi_1 \cdot 3 + \ldots + \left(1 - \pi_1\right)^{x-1} \cdot \pi_1 \cdot x + \ldots = \frac{1}{\pi_1}$$

> *Wenn in einer Population der Anteil der Elemente mit der Eigenschaft A=1 gleich $\pi_1 = 0.01$ ist, dann ist bei einer einfachen Zufallsauswahl mit Zurücklegen mit 1/0.01 = 100 Ziehungen zu rechnen, bevor das erste Mal ein Element mit der Eigenschaft A=1 ausgewählt wird.*

Die Varianz der geometrischen Verteilung berechnet sich nach:

$$\sigma_X^2 = \frac{1 - \pi_1}{\pi_1^2}$$

Wenn x Wiederholungen bis zum ersten Auftretens der Eigenschaft A=1 notwendig sind, dann wird in den ersten x – 1 Wiederholungen die Eigenschaft A=0 beobachtet. Interessiert die Wahrscheinlichkeitsverteilung dieser Zahl der „erfolglosen“ Wiederholungen bis zum ersten „Erfolg“, ergeben sich die gleichen Wahrscheinlichkeiten. Erwartungswert und Varianz von X mit den Ausprägungen 0, 1, 2, ... sind hier: $\mu_X = 1/(1–\pi_1)$ und $\sigma^2_X = (1–\pi_1)/\pi^2_1$.

Die Pascal-Verteilung

Eine Verallgemeinerung der geometrischen Verteilung besteht darin, die Wahrscheinlichkeit zu erhalten, dass X = x Ziehungen mit Zurücklegen notwendig sind, um genau k Elemente mit der Eigenschaft A=1 zu erhalten. Der Wertebereich von X ist dann k, k+1, k+2, ...

Die Auftretenswahrscheinlichkeiten können ebenfalls mit Hilfe von Folgen von binomialverteilten Realisierungen berechnet werden:

$$\Pr\left(X = x \middle| k, \pi_1\right) = \binom{x-1}{k-1} \cdot \pi_1^k \cdot \left(1-\pi_1\right)^{x-k}$$

Diese Verteilung ergibt sich, weil die ersten k–1 Elemente mit der Eigenschaft A=1 einer Binomialverteilung mit den Parametern n = k–1 und π_1 folgen und das letzte Element einer Bernoulli-Verteilung mit π_1.

Diese Verteilung wird nach dem Mathematiker Pascal ***Pascal-Verteilung*** genannt.
Erwartungswert und Varianz berechnen sich nach:

$$\mu_X = \frac{k}{\pi_1} \text{ und } \sigma_X^2 = k \cdot \frac{1-\pi_1}{\pi_1^2}$$

Als Beispiel kann interessieren, wie viele Versuche notwendig sind, um beim wiederholten Würfeln zweimal eine „6“ zu erreichen. Es ist damit zu rechnen, dass im Durchschnitt in jeweils 2/(1/6) = 12 Würfen zweimal eine „6“ vorkommt.

Die hypergeometrische Verteilung

Wenn in einer Population von N Elementen N_1 Elemente die Eigenschaft A=1 aufweisen und entsprechend $N_0 = N - N_1$ Elemente die komplementäre Eigenschaft A=0 und in einer einfachen Zufallsauswahl mit Zurücklegen n Elemente ausgewählt werden, ist die Häufigkeit, dass die Stichprobe x Elemente mit der Eigenschaft A=1 enthält, binomialverteilt mit den Parametern n und $\pi_1 = N_1/N$.
Wie ändert sich die Wahrscheinlichkeitsverteilung, wenn dieAuswahl der n Elemente in einer einfachen Zufallsauswahl ohne Zurücklegen erfolgt?

In L09 wurde bereits gezeigt, dass bei einer einfachen Zufallsauswahl ohne Zurücklegen die Wahrscheinlichkeit jeder Stichprobe ohne Berücksichtigung der Reihenfolge $1/{}_N K_n$, also Eins durch „N über n“ beträgt.
Von Interesse ist nun die Wahrscheinlichkeit, dass eine Stichprobe genau x Elemente mit der Eigenschaft A=1 aufweist. Diese Wahrscheinlichkeit ist die Summe der Auftretenswahrscheinlichkeiten aller Stichproben, in denen jeweils x Elemente mit dieser Eigenschaft vorkommen.

Insgesamt weisen N_1 Elemente der Population die Eigenschaft A=1 auf. Daher gibt es „N_1 über x“ Möglichkeiten (Kombinationen), aus der Menge im Umfangs N_1 x Elemente auszuwählen. Analog gibt es unter den $N_0 = N - N_1$ Elementen mit der Eigenschaft A=0 genau „N_0 über n–x“ Möglichkeiten, n – x Elemente auszuwählen.

Die hypergeometrische Verteilung

Die Auswahl von x Elementen aus N_1 und von n – x aus N_0 erfolgt unabhängig voneinander. Daher gibt es insgesamt „N_1 über x“ mal „N_0 über n – x“ Möglichkeiten, eine Stichprobe des Umfangs n zuziehen, in der x Elemente die Eigenschaft A=1 haben und n–x die Eigenschaft A=0.

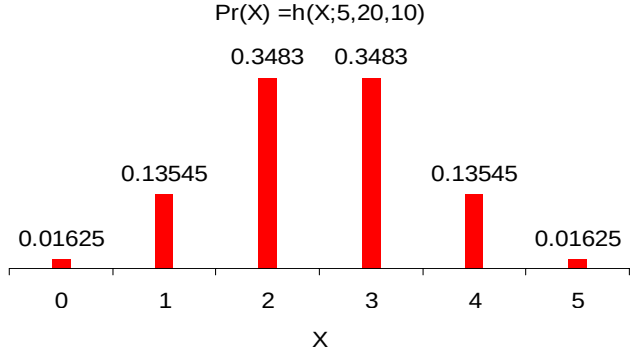
Die gesuchte Wahrscheinlichkeit bei einer einfachen Zufallsauswahl ohne Zurücklegen genau x von n Elementen mit der Eigenschaft A=1 zu haben, berechnet sich daher als Produkt der Zahl der möglichen Anordnungen (Ziehungsreihenfolgen) mal der Wahrscheinlichkeit jeder einzelnen Stichprobe:

$$\Pr\left(X = x \mid n, N, N_1\right) = \frac{\binom{N_1}{x} \cdot \binom{N - N_1}{n - x}}{\binom{N}{n}}$$

Eine Zufallsvariable X mit den möglichen Ausprägungen x = 0, 1, ..., N_1 wobei x ≤ n heißt ***hypergeometrisch*** verteilt, wenn sich die Wahrscheinlichkeitsfunktion nach dieser Formel berechnet. Die ***hypergeometrische Verteilung*** hat drei Parameter **n**, **N** und **N_1** und wird im Folgenden durch **h(X; n, N, N_1)** symbolisiert.

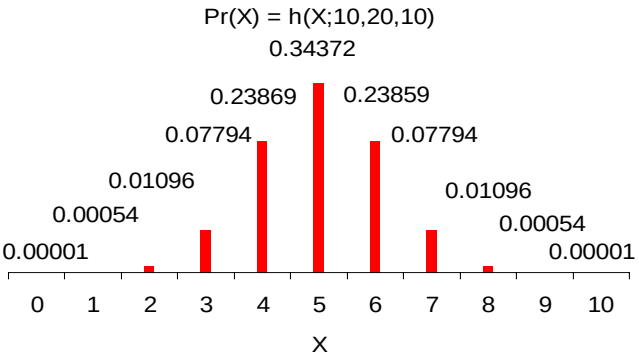
Als Beispiele werden die hypergeometrischen Verteilungen h(X; 5, 20, 10), h(X;10, 20, 10), h(X;10, 20, 8) und b(X;10, 20, 14) betrachtet.

Die hypergeometrische Verteilung



X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.01625	0.01625	0.00000	0.00000
1	0.13545	0.15170	0.13543	0.13545
2	0.34830	0.50000	0.69659	1.39319
3	0.34830	0.84830	1.04489	3.13467
4	0.13545	0.98375	0.54180	2.16718
5	0.01625	1.00000	0.08127	0.40635
			2.50000	7.23684

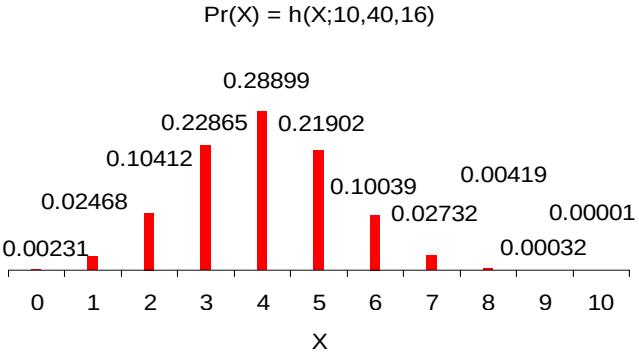
$$\mu_X = 2.5 \; ; \; \sigma_X^2 = 7.23684 - 2.5^2 = 0.98684$$



X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.00001	0.00001	0.00000	0.00000
1	0.00054	0.00055	0.00054	0.00054
2	0.01096	0.01151	0.02192	0.04384
3	0.07794	0.08945	0.23382	0.70147
4	0.23869	0.32814	0.95477	3.81909
5	0.34372	0.67186	1.71859	8.59296
6	0.23859	0.91055	1.43216	8.59296
7	0.07794	0.98849	0.54558	3.81909
8	0.01096	0.99945	0.08768	0.70147
9	0.00054	0.99999	0.00487	0.04384
10	0.00001	1.00000	0.00005	0.00054
			5.0000	26.31579

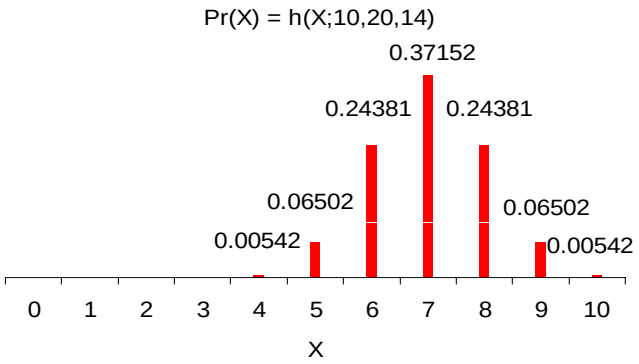
$$\mu_X = 5.0 \; ; \; \sigma_X^2 = 26.31579 - 5.0^2 = 1.31579$$

Die hypergemetrische Verteilung



$$\mu_X = 4.0\ ;\ \sigma_X^2 = 17.84615 - 4.0^2 = 1.84615$$

X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.00231	0.00231	0.00000	0.00000
1	0.02468	0.02699	0.02468	0.02468
2	0.10412	0.13111	0.20824	0.41647
3	0.22865	0.35976	0.68595	2.05786
4	0.28899	0.64875	1.15596	4.62383
5	0.21902	0.86777	1.09512	5.47558
6	0.10039	0.96816	0.60231	3.61389
7	0.02732	0.99548	0.19121	1.33848
8	0.00419	0.99967	0.03352	0.26819
9	0.00032	0.99999	0.00292	0.02624
10	0.00001	1.00000	0.00009	0.00094
			4.00000	17.84615



$$\mu_X = 7.0\ ;\ \sigma_X^2 = 50.1056 - 7.0^2 = 1.10526$$

X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.00000	0.00000	0.00000	0.00000
1	0.00000	0.00000	0.00000	0.00014
2	0.00000	0.00000	0.00000	0.00579
3	0.00000	0.00000	0.00000	0.08102
4	0.00542	0.00542	0.02167	0.08669
5	0.06502	0.07043	0.32508	1.62539
6	0.24381	0.31424	1.46285	8.77709
7	0.37152	0.68576	2.60062	18.20433
8	0.24381	0.92957	1.95046	15.60372
9	0.06502	0.99458	0.58514	5.26625
10	0.00542	1.00000	0.05418	0.54180
			7.0000	50.10526

Die hypergeometrische Verteilung

Wie die Beispiele zeigen, hat die ***hypergeometrische Verteilung*** eine ähnliche Form wie die Binomialverteilung. Bei einem Anteil $N_1/N = N_0/N = 0.5$ ist die Verteilung symmetrisch; bei $N_1/N < 0.5$ ist sie tendenziell rechtsschief, bei $N_1/N > 0.5$ kann sie entsprechend linksschief sein.

Wie die Beispiele auch zeigen, kann der Wertebereich verglichen mit einer Binomialverteilung eingeschränkter sein: Wenn $N_1 < n$ kann die Zufallsvariable X nur die Ausprägungen 0, 1, 2, ..., N_1 annehmen; ist $N_0 < n$ nur die Ausprägungen $n - N_0$, $n - N_0+1$, ..., n.

So liegt der Wertebereich von h(X;10,20,14) zwischen 4 und 10, da hier $N_0 = 20 - 14 = 6$ und $n - N_0 = 10 - 6 = 4$.

Die Verteilungsfunktion der hypergeometrischen Verteilung ergibt sich wieder über Aufsummieren:

$$F\left(X = x \middle| h\left(X; n, N, N_1\right)\right) = \Pr\left(X \le x\right) = \sum_{j=0}^{x} \frac{\binom{N_1}{j} \cdot \binom{N - N_1}{n - j}}{\binom{N}{n}}$$

Wie bei der Binomialverteilung sind auch Erwartungswert und Varianz der hypergeometrischen Verteilung Funktionen der Verteilungsparameter.

So zeigen die Beipsiele, dass der Erwartungswert $n \cdot N_1/N$ beträgt, bei h(X;5,20,10) also 5·10/20 = 2.5, bei h(X;10,20,14) = 10·14/20 = 7.

Erwartungswert, Varianz und Schiefe der hypergeometrischen Verteilung

Generell gilt für Erwartungswert und Varianz einer hypergeometrischen Verteilung:

$$\mu_X = n \cdot \frac{N_1}{N} \text{ und } \sigma_X^2 = n \cdot \frac{N_1}{N} \cdot \left(1 - \frac{N_1}{N}\right) \cdot \frac{N-n}{N-1} \text{ wenn } X \sim h\left(X; n, N, N_1\right)$$

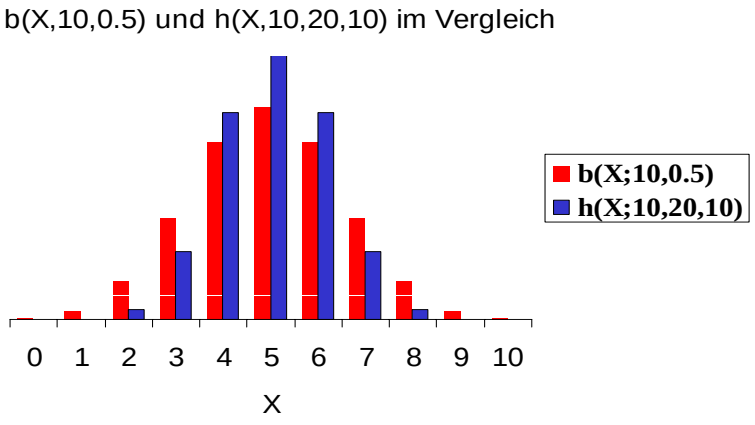
Für die wiedergegebenen Beispiele hypergeometrischer Verteilungen gilt also:
h(X;5,20,10) $\Rightarrow \mu_X = 5 \cdot 10/20 = 2.5$; $\sigma^2_X = 5 \cdot 10/20 \cdot (1-10/20) \cdot (20-5)/(20-1) = 0.98684$
h(X;10,20,10) $\Rightarrow \mu_X = 10 \cdot 10/20 = 5$; $\sigma^2_X = 10 \cdot 10/20 \cdot (1-10/20) \cdot (20-10)/(20-1) = 1.31579$
h(X;10,40,16) $\Rightarrow \mu_X = 10 \cdot 16/40 = 4$; $\sigma^2_X = 10 \cdot 16/40 \cdot (1-16/40) \cdot (40-10)/40-1) = 1.84615$
h(X;10,20,14) $\Rightarrow \mu_X = 10 \cdot 14/20 = 7$; $\sigma^2_X = 10 \cdot 14/20 \cdot (1-14/20) \cdot (20-10)/(20-1) = 1.10526$

Auch die Schiefe ist eine Funktion der Modellparameter:

$$\frac{\mu_3}{\sigma^3} = \frac{\left(N - 2 \cdot N_1\right) \cdot \sqrt{N-1} \cdot \left(N - 2 \cdot n\right)}{\sqrt{n \cdot N_1 \cdot \left(N - N_1\right) \cdot \left(N - n\right)} \cdot \left(N - 2\right)}$$

Aus der Gleichung für die Schiefe folgt:
- ist $N_1 = N_0$ oder $N = 2 \cdot n$, dann ist die Verteilung symmetrisch, da dann der Zähler in der Formel des Schiefekoeffizienten Null ist.
- Steigt die Zahl der Wiederholungen n an, wird die Schiefe geringer.

Beziehung zwischen hypergeometrischer Verteilung und Binomialverteilung

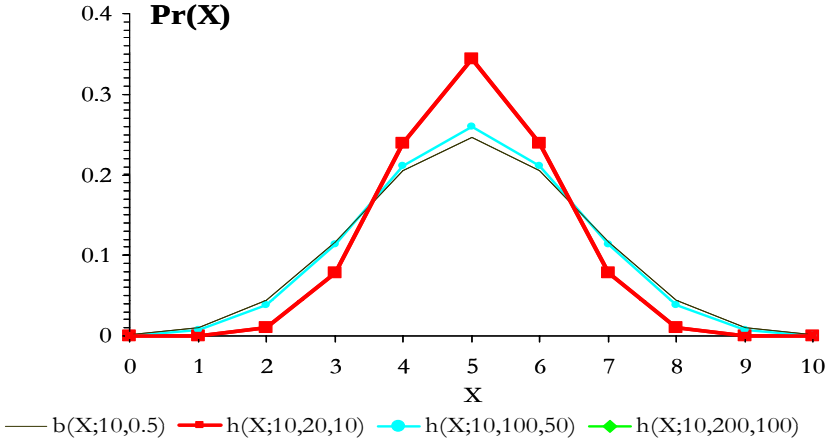


Binomialverteilung und hypergeometrische Verteilung geben die Wahrscheinlichkeiten an, dass in einer einfachen Zufallsauswahl mit bzw. ohne Zurücklegen x von n Elementen eine Eigenschaft A=1 aufweisen, wenn in der Population N_1 von N Elementen diese Eigenschaft haben. Der Vergleich der Wahrscheinlichkeitsfunktionen von b(X; 10, 0.5) und h(X; 10, 20, 10) zeigt, dass bei gleicher Fallzahl n und gleichem Anteil $\pi_1 = N_1/N$ beide Verteilungen eine ähnliche Form haben, die Binomialverteilung aber eine größere Streuung aufweist.

Tatsächlich sind die Erwartungswerte bei beiden Verteilungen gleich $n \cdot \pi_1$ während sich die Varinzen nur durch den Faktor (N–n)/(N–1) unterscheiden, um den die Varianz der hypergeometrischen Verteilung kleiner ist:

$$\sigma^2\left(X \middle| b(X; n, \pi_1\right) = n \cdot \pi_1 \cdot \left(1 - \pi_1\right) \text{ und } \sigma^2\left(X \middle| h(X; n, N, \pi_1 \cdot N\right) = n \cdot \pi_1 \cdot \left(1 - \pi_1\right) \cdot \frac{N-n}{N-1}$$

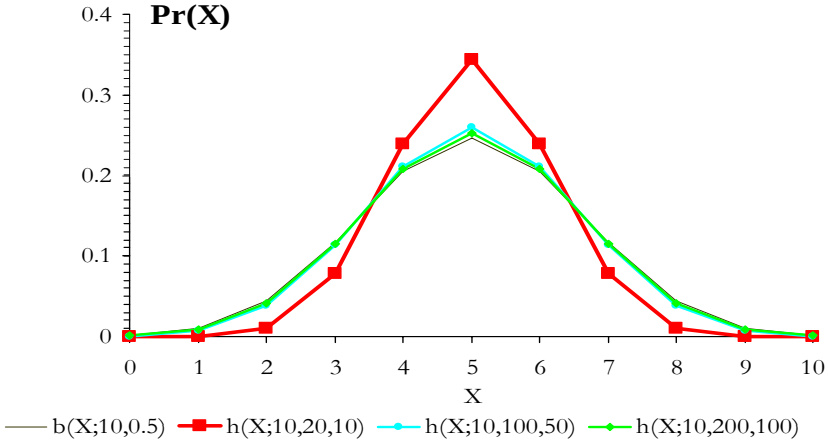
Beziehung zwischen hypergeometrischer Verteilung und Binomialverteilung



Wenn der Populationsumfang N relativ zum Stichprobenumfang n ansteigt, dann nähert sich der Faktor (N–n)/(N–1)immer mehr dem Wert eins an.
Tatsächlich nähern sich auch die Wahrscheinlichkeiten der Ausprägungen von Binomialverteilung und hypergeometrischer Verteilung dann immer mehr an.

Die Abbildung zeigt exemplarisch die Auftretetenswahrscheinlichkeiten von hypergeometrischen Verteilungen mit den Parametern h(X;10, 20, 10), h(X;10, 100, 50) und h(X;10, 200, 100) sowie die Binomialverteilung mit den Parametern b(X,10, 0.5). Gemeinsam ist allen Verteilungen, dass der Populationsanteil des betrachteten Merkmals A=1 stets $\pi_1=N_1/N=0.5$ beträgt. Je größer der Populationsumfang N bei der hypergeometrischen Verteilung, desto ähnlicher sind sich die Verteilungen.

Asymptotische Annäherung der hypergeometrischer Verteilung an die Binomialverteilung



Im Extremfall einer unendlich großen Population sind die beiden Verteilungen identisch.
Wenn sich zwei Wahrscheinlichkeitsverteilungen unter bestimmten Bedingungen immer ähnlicher werden, spricht man von einer ***asymptotischen Annäherung***.

Die hypergeometrische Verteilung nähert sich somit der Binomialverteilung asymptotisch an, wenn der Populationsumfang N über alle Grenzen ansteigt und dabei der Anteil $\pi_1 = N_1/N$ konstant bleibt:

$$\lim_{N\to\infty}\left(h\left(X;n,N,N_1\middle|\pi_1=\frac{N_1}{N}\right)\right)=b\left(X;n,\pi_1=\frac{N_1}{N}\right)$$

Beziehung zwischen hypergeometrischer Verteilung und Binomialverteilung

Die Wahrscheinlichkeiten von $h(X;n,\pi_1)$ und $b(X;n,N,\pi_1 \cdot N)$ sind sich für praktische Zwecke hinreichend ähnlich sind, wenn der Quotient $N/n > 20$ ist.

Wenn der Stichprobenumfang also kleiner als 1/20 des Populationsumfangs ist, kann bei einer einfachen Zufallsauswahl ohne Zurücklegen die Auftretenswahrscheinlichkeit von X=x eines interessierenden Merkmals A=1 mit hinreichender Genauigkeit auch über die Binomialverteilung anstelle der hypergeometrischen Verteilung berechnet werden

Die Poisson-Verteilung

Mit Hilfe der Pascal-Verteilung lässt sich berechnen, wie groß die Wahrscheinlichkeit ist, dass in x = 1, 2, ... Wiederholungen eines Zufallsexperiments genau k mal das interessierende Ereignis A=1 vorkommt.
Tatsächlich besteht eine wichtige Anwendung statistischer Modellierung darin, die Wahrscheinlichkeit zu berechnen, mit der 0, 1, 2, ... interessierende Ereignisse auftreten. Im Unterschied zur Pascal-Verteilung, die die Wahrscheinlichkeitsverteilungen der Anzahl der notwendigen Wiederholungen eines Zufallsexperiments erfasst, interessiert hier also die Zahl der Ereignisse.

Die Herleitung der Verteilung aus der Binomialverteilung lässt sich an einem Beispiel verdeutlichen:

Angenommen, die Wahrscheinlichkeit, dass in einer Stadt in einem Zeitraum von n Tagen X Unfälle auftreten, sei binomialverteilt mit $b(X; n, \pi_1)$. Wie wahrscheinlich ist es dann, dass Y = 0, 1, 2, ... Unfälle auftreten, wenn einerseits n beliebig ansteigt, andererseits das Produkt $\lambda = n \cdot \pi_1$ dabei konstant bleibt?

Die Poisson-Verteilung

Die gesuchte Wahrscheinlichkeit ergibt sich mathematisch über die Grenzwertbetrachtung der Wahrscheinlichkeitsfunktion einer Binomialverteilung :

$$\lim_{n\to\infty}\left(b\left(X=x;n,\pi_1 \middle| n\cdot\pi_1=\lambda\right)\right)=\lim_{n\to\infty}\left(\binom{n}{x}\cdot\left(\frac{\lambda}{n}\right)^x\cdot\left(1-\frac{\lambda}{n}\right)^{n-x}\right)=\frac{\lambda^x}{x!}\cdot e^{-\lambda}$$

Da die Zahl der Wiederholungen über alle Grenzen wächst, können auch die Häufigkeiten X eine beliebige ganze Zahl zwischen 0 und +∞ annehmen.

Die resultierende Wahrscheinlichkeitsverteilung ist die ***Poisson-Verteilung*** mit dem Parameter **λ**, im Folgenden symbolisiert durch **p(X; λ)** wobei X die Ausprägungen 0, 1, 2, ... hat:

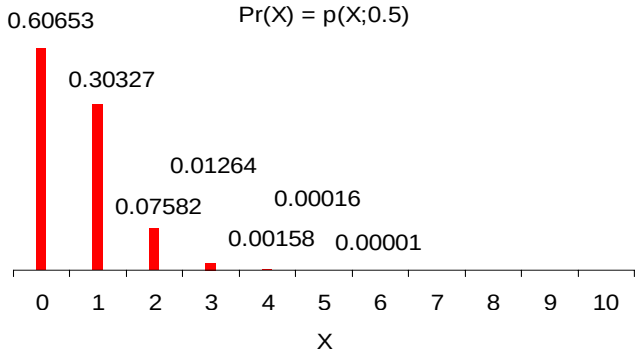
$$\Pr\left(X=x \middle| p\left(X;\lambda\right)\right)=\frac{\lambda^x}{x!}\cdot e^{-\lambda}$$

Obwohl es eine unendliche Zahl von möglichen Realisierungen gibt, lassen sich für alle Ausprägungen einer Poisson-Verteilung Wahrscheinlichkeiten berechnen, die jedoch für sehr große Werte schnell gegen Null gehen, da bei der Berechnung durch die Fakultät des Wertes einer Ausprägung geteilt wird.

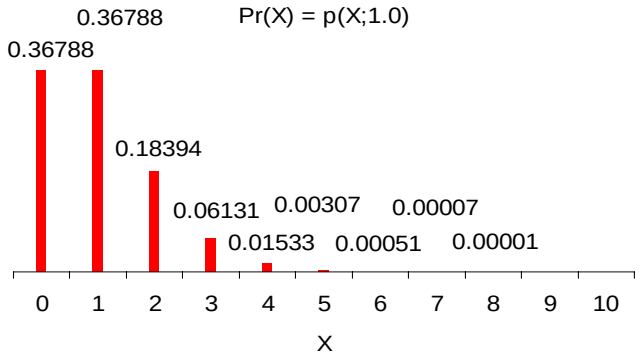
Die Wahrscheinlichkeitsfunktion der Poisson-Verteilung berechnet sich nach:

$$F\left(X=x \middle| P\left(X;\lambda\right)\right)=\Pr\left(X\le x\right)=\sum_{j=0}^{x}\frac{\lambda^j}{j!}\cdot e^{-\lambda}$$

Die Poisson-Verteilung



$$\mu_X = 0.5 \; ; \; \sigma_X^2 = 0.75 - 0.5^2 = 0.5$$

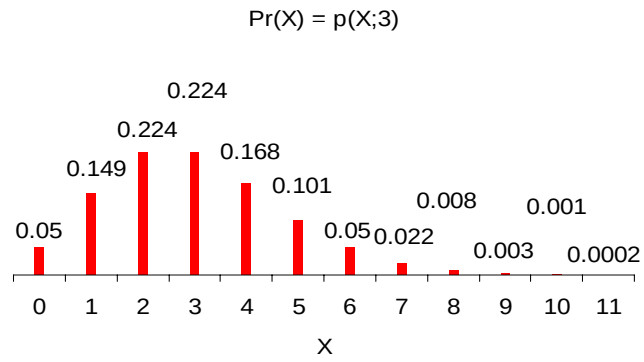


$$\mu_X = 1.0 \; ; \; \sigma_X^2 = 2 - 1 = 1$$

X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.60653	060653	0.00000	0.00000
1	0.30327	0.90980	0.30327	0.30327
2	0.07582	0.98561	0.15163	0.30327
3	0.01264	0.99825	0.03791	0.11327
4	0.00158	0.99983	0.00632	0.02527
5	0.00016	0.99999	0.00079	0.00395
6	0.00001	1.00000	0.00008	0.00047
7	0.00000	1.00000	0.00001	0.00005
8	0.00000	1.00000	0.00000	0.00000
9	0.00000	1.00000	0.00000	0.00000
10	0.00000	1.00000	0.00000	0.00000
			0.50000	0.75000

X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.36788	0.36788	0.00000	0.00000
1	0.36788	0.73576	0.36788	0.36788
2	0.18394	0.91970	0.36788	0.73576
3	0.06131	0.98101	0.18394	0.55182
4	0.01533	0.99634	0.06131	0.24525
5	0.00307	0.99941	0.01533	0.07664
6	0.00051	0.99992	0.00307	0.01839
7	0.00007	0.99999	0.00051	0.00358
8	0.00001	1.00000	0.00007	0.00058
9	0.00000	1.00000	0.0.0001	0.00008
10	0.00000	1.00000	0.00000	0.00001
			1.00000	2.00000

Die Poisson-Verteilung



$$\mu_X = 3 \; ; \; \sigma_X^2 = 12 - 3^2 = 3$$

X	Pr(X)	F(X)	Pr(X)·X	Pr(X)·X²
0	0.04979	0.04979	0.00000	0.00000
1	0.14936	0.19915	0.14936	0.14936
2	0.22404	0.42319	0.44808	0.89617
3	0.22404	0.64723	0.67213	2.01638
4	0.16803	0.81526	0.67213	2.68850
5	0.10082	0.91608	0.50409	2.52047
6	0.05041	0.96649	0.30246	1.81474
7	0.02160	0.98810	0.15123	1.05860
8	0.00810	0.99620	0.06481	0.51850
9	0.00270	0.99890	0.02430	0.21874
10	0.00081	0.99971	0.00810	0.08102
11	0.00022	0.99993	0.00243	0.02673
12	0.00006	0.99998	0.00066	0.00795
13	0.00001	1.00000	0.00017	0.00215
14	0.00000	1.00000	0.00004	0.00054
15	0.00000	1.00000	0.00001	0.00012
16	0.00000	1.00000	0.00000	0.00003
17	0.00000	1.00000	0.00000	0.00001
18	0.00000	1.00000	0.00000	0.00000
19	0.00000	1.00000	0.00000	0.00000
20	0.00000	1.00000	0.00000	0.00000
			3.00000	12.00000

Die Poisson-Verteilung

Die Beispiele demonstrieren, dass der Modellparameter λ der Poisson-Verteilung gleich dem Erwartungswert und der Varianz der Verteilung ist:

$$\mu_X = \lambda \; ; \; \sigma_X^2 = \lambda \text{ wenn } X \sim p(X;\lambda)$$

Da auch das dritte zentrale Moment der Poisson-Verteilung gleich dem Parameter λ ist, ist die Verteilung rechtsschief. Wie der Schiefekoeffizeint zeigt, nimmt die Schiefe allerdings mit steigendem Wert von λ ab:

$$\frac{\mu_3}{\sigma^3} = \frac{1}{\sqrt{\lambda}}$$

Inhaltlich lässt sich λ als Intensität deuten. Je größer der Wert, desto eher ist mit dem Auftreten von Ereignissen zu rechnen. So beträgt die Wahrscheinlichkeit, dass in einem betrachteten Zufallsprozess kein Ereignis auftritt: Pr(X=0) = exp(–λ), was bei einem Wert von λ=1 knapp 36.8% ist, bei einem Wert von λ=2 auf 13.5% sinkt und bei λ=3 nur 5.0% beträgt. Die komplementären Wahrscheinlichkeiten geben dann an, dass mindestens 1 Ereignis auftritt.

Wie bereits die Binomialverteilung ist auch die Poisson-Verteilung additiv:

Wenn X_1 poissonverteilt ist mit $p(X_1; \lambda_1)$ und X_2 poissonverteilt mit $p(X_2; \lambda_2)$, und X_1 und X_2 statistisch unabhängig voneinander sind, dann ist die Summe $Y = X_1 + X_2$ poissonverteilt mit $p(Y; \lambda_1 + \lambda_2)$.

Anwendungen diskreter Wahrscheinlichkeitsverteilungen

Die vorgestellten Wahrscheinlichkeitsverteilungen lassen sich in einer Vielzahl von Anwendungen nutzen. Im folgenden werden einige Möglichkeiten vorgestellt.

Berechnung von Risiken

Ein Seminar wird von n=45 Studierenden besucht. Eine Studentin, die gerade Geburtstag hat, möchte herausfinden, ob ein anderer Teilnehmer am selben Tag wie sie Geburtstag hat. Wie wahrscheinlich ist dieses Ereignis?

Wenn unterstellt wird, dass sich die Wahrscheinlichkeit von Geburtstagen gleichmäßig über alle 365 Tage des Jahres verteilt, dann beträgt die Wahrscheinlichkeit, dass zwei zufällig herausgegriffene Personen am gleichen Tag Geburtstag haben $365 \cdot 1/365^2 = 1/365 = 0.00274$.
Die gesuchte Wahrscheinlichkeit kann dann über eine Binomialverteilung berechnet werden. Es ist die Wahrscheinlichkeit, dass bei n = 44 Mitstudierenden und einer Wahrscheinlichkeit $\pi_1 = 1/365$ die Anzahl der Fälle mit der interessierenden Eigenschaft, am gleichen Tag Geburtstag zu haben, mindestens 1 beträgt:

$$\Pr\left(X \geq 1 \middle| b\left(X; 44, 365^{-1}\right)\right) = 1 - \Pr\left(X = 0 \middle| b\left(X; 44, 365^{-1}\right)\right)$$

$$= 1 - \binom{44}{0} \cdot \left(\frac{1}{365}\right)^0 \cdot \left(1 - \frac{1}{365}\right)^{44} = 1 - 0.886 = 0.114$$

Die Wahrscheinlichkeit, dass noch eine zweite Person Geburtstag hat, beträgt ungefähr 11.4%.

Anwendungen diskreter Wahrscheinlichkeitsverteilungen

Anders sieht es aus, wenn die Wahrscheinlichkeit gesucht ist, dass irgendwelche zwei der 45 Studierenden am selben Tag Geburtstag haben.

Für die Möglichkeit, am selben Tag Geburtstag zu haben, stehen dann nämlich alle möglichen nichtredundanten Paare aus der Gruppe der 45 Personen zur Verfügung. Bei n Personen gibt es „n über 2“ Kombinatioen aus n Personen 2 für Paarvergleiche auszuwählen.

Im Beispiel beträgt die Zahl der Paarvergleiche also „45 über 2“ = 45!/(2!·43!) =990.

Die Wahrscheinlichkeit, dass mindestes eines der 990 Paare am gleichen Tag Geburtstag hat, ist dann:

$$\Pr\left(X \geq 1 \middle| b\left(X;990,365^{-1}\right)\right) = 1 - \Pr\left(X = 0 \middle| b\left(X;990,365^{-1}\right)\right)$$

$$= 1 - \binom{990}{0} \cdot \left(\frac{1}{365}\right)^{0} \cdot \left(1 - \frac{1}{365}\right)^{990} = 1 - 0.066 = 0.934$$

Die Wahrscheinlichkeit, dass es in einer Gruppe von 45 Personen (mindestens einmal) zwei Personen gibt, die am gleichen Tag Geburtstag haben, beträgt also 93.4% und ist damit sehr hoch.

Anwendungen diskreter Wahrscheinlichkeitsverteilungen

Berechnung der notwendigen Stichprobengröße

In einem Staat sei der Anteil der Millionäre 5%. Wie groß muss bei einer einfachen Zufallsauswahl die Stichprobe sein, damit in der Stichprobe mit 5 Millionären zu rechnen ist?

Die gesuchte Fallzahl kann als Erwartungswert einer Pascal-Verteilung mit den Parametern k=5 und π_1= 0.05 aufgefasst werden. Erwartungswert und Varianz betragen dann:

$$\mu\left(X \middle| k = 5, \pi_1 = 0.05\right) = \frac{5}{0.05} = 100 \; ; \; \sigma^2\left(X \middle| k = 5, \pi_1 = 0.05\right) = \frac{5 \cdot 0.95}{0.05^2} = 1900 \Rightarrow \sigma_X = 43.6$$

Wenn die Stichprobe n=100 Fälle umfässt, ist mit Mittel mit 5 Millionären zu rechnen. Allerdings ist die Streuung mit einer Standardabweichung von 43.6 Fällen sehr groß. Tatsächlich beträgt die Wahrscheinlichkeit, bei einer Stichprobe von n=100 mindestens 5 Millionäre in der Stichprobe zu haben, nur gut 56.4%:

$$\Pr\left(X \geq 5 \middle| b\left(X;100,0.05\right)\right) = 1 - \Pr\left(X \leq 4 \middle| b\left(X;100,0.05\right)\right)$$

$$= 1 - 0.95^{100} - \binom{100}{1} \cdot 0.05 \cdot 0.95^{99} - \binom{100}{2} \cdot 0.05^2 \cdot 0.95^{98} - \binom{100}{3} \cdot 0.05^3 \cdot 0.95^{97}$$

$$- \binom{100}{4} \cdot 0.05^4 \cdot 0.95^{96} = 1 - 0.00592 - .03112 - 0.08118 - 0.13958 - .17814 = 56.4\%$$

Anwendungen diskreter Wahrscheinlichkeitsverteilungen

Wird die Stichprobe um gut 1 Standardabweichung der Pascal-Verteilung auf n=150 Fälle erhöht, beträgt die Wahrscheinlichkeit, dass mindestens 5 Millionäre in der Stichprobe sind, schon 96.6%.

Wahrscheinlichkeit von Ereignissen

Es habe sich gezeigt, dass an einer Ampel im Schnitt etwa alle 2 Jahre ein Unfall geschieht. Im vergangenen Jahr sind allerdings 2 Unfälle an der Ampel geschehen. Ist diese Häufung außergewöhnlich oder erwartbar?

Es wird von einer Poisson-Verteilung ausgegangen, die bei einem Zeitraum von 2 Jahren einen Erwartungswert von $\lambda=1$ aufweist. Wird von Unabhängigkeit der einzelnen Ereignisse ausgegangen, dann beträgt der Erwartungswert bei einer Halbierung des betrachteten Zeitraums $\lambda=0.5$. Die Wahrscheinlichkeit, dass bei diesem Parameter X=2 Ereignisse auftreten, beträgt dann:

$$\Pr\left(X=2\middle|p\left(X;0.5\right)\right)=\frac{0.5^2}{2!}\cdot e^{-0.5}=0.076$$

Die Wahrscheinlichkeit, dass in einem Jahr 2 Unfälle geschehen, beträgt immerhin 7.6%. Wie die oben aufgelistete Wahrscheinlichkeitsverteilung der Poisson-Verteilung mit $\lambda=0.5$ zeigt, beträgt die Wahrscheinlichkeit sogar 8.9% (= 1–0.9098), dass sich in einem Jahr mindestens 2 Unfälle ereignen.

Die Häufung ist eher nicht als außergewöhnlich zu bezeichnen, sondern etwa alle 11 Jahre zu erwarten.

Anwendungen diskreter Wahrscheinlichkeitsverteilungen

Wahrscheinlichkeiten von Anteilen

Mit Hilfe der hypergeometrischen bzw. der Binomialverteilung kann auch die Wahrscheinlichkeitsverteilung von Anteilen in einfachen Zufallsauswahlen berechnet werden.
Formal berechnet sich ein Anteil $p_1 = n_1/n$ als eine einfache Lineartransformation aus den absoluten Anteilen. Da jedem möglichen Anteilswert in einer Stichprobe genau eine absolute Häufigkeit entspricht, sind die Auftretenswahrscheinlichkeiten der Anteile gleich den Auftretenswahrscheinlichkeiten der korrespondieren Häufigkeiten.

Bei einer **einfachen Zufallsauswahlen mit Zurücklegen** berechnet sich daher die Wahrscheinlichkeit einer relative Häufigkeit $p_1 = n_1/n$ in einer Stichprobe der Größe n, wenn der interessierende Anteil in der Population gleich $\pi_1 = N_1/N$ ist, nach:

$$\Pr\left(p_1\right)=\Pr\left(X=n\cdot p_1\middle|b\left(X,n,\pi_1\right)\right)=\binom{n}{p_1\cdot n}\cdot\left(\pi_1\right)^{p_1\cdot n}\cdot\left(1-\pi_1\right)^{n\cdot\left(1-p_1\right)}$$

Die Verteilungsfunktion ist dann:

$$F\left(p_1\right)=\Pr\left(X\le n\cdot p_1\middle|b\left(X,n,\pi_1\right)\right)=\sum_{i=1}^{n\cdot p_1}\binom{n}{i}\cdot\left(\pi_1\right)^{i}\cdot\left(1-\pi_1\right)^{n-i}$$

Wahrscheinlichkeiten von Anteilen bei einfachen Zufallsauswahlen mit Zurücklegen

Zu beachten ist, dass nur solche Anteilswerte Realisierungswahrscheinlichkeiten > 0 haben können, bei denen $n \cdot p_1$ eine ganze Zahl ist.
Bei der Berechnung der Verteilungsfunktion erfolgt entsprechend die Summierung bis der Summierungsindex i die nächste ganze Zahl größer/gleich $n \cdot p_1$ist.

Angenommen, eine Stichprobe umfasse n=15 Fälle. Dann können nur die Stichprobenanteile 1/15, 2/15, ..., 15/15 Auftretenswahrscheinlichkeiten ungleich Null haben. Soll berechnet werden, wie wahrscheinlich es ist, dass ein Anteil kleiner oder gleich 0.5 ist, muss die Binomialverteilung bis zum Wert i=8 aufsummiert werden, weil 0.5·15 = 7.5 keine ganze Zahl ist.

Der Erwartungswert und die Varianz der Kennwerteverteilung eines Stichprobenanteils berechnet sich über Regeln für Lineartransformationen nach:

$$\mu(p_1) = \frac{1}{n} \cdot \mu\left(X \middle| b(X; n, \pi_1)\right) = \frac{1}{n} \cdot n \cdot \pi_1 = \pi_1 = \frac{N_1}{N}$$

$$\sigma^2(p_1) = \left(\frac{1}{n}\right)^2 \cdot \sigma^2\left(X \middle| b(X; n, \pi_1)\right) = \frac{1}{n^2} \cdot n \cdot \pi_1 \cdot (1 - \pi_1) = \frac{\pi_1 \cdot (1 - \pi_1)}{n} = \frac{\frac{N_1}{N} \cdot \frac{N - N_1}{N}}{n}$$

Wahrscheinlichkeiten von Anteilen bei einfachen Zufallsauswahlen ohne Zurücklegen

Bei einer ***einfachen Zufallsauswahl ohne Zurücklegen*** wird anstelle der Binomialverteilung die hypergeometrische Verteilung herangezogen. Die Wahrscheinlichkeitsverteilung berechnet sich nach:

$$Pr(p_1) = Pr\left(X = n \cdot p_1 \middle| h(X, n, N, N_1)\right) = \frac{\binom{N_1}{n \cdot p_1} \cdot \binom{N - N_1}{n \cdot (1 - p_1)}}{\binom{N}{n}}$$

Die Verteilungsfunktion ist:

$$F(p_1) = Pr\left(X \le n \cdot p_1 \middle| h(X, n, N, N_1)\right) = \sum_{i}^{n \cdot p_1} \frac{\binom{N_1}{i} \cdot \binom{N - N_1}{n - i}}{\binom{N}{n}}$$

Wiederum muss darauf geachtet werden, dass $n \cdot p_1$ eine ganze Zahl zwischen 0 und n ist.
Falls nur der Populationsanteil π_1 bekannt ist, ergibt sich N_1 nach $N_1 = N \cdot \pi_1$.

Wahrscheinlichkeiten von Anteilen bei einfachen Zufallsauswahlen ohne Zurücklegen

Erwartungswert und Varianz der Kennwerteverteilung berechnen sich wiederum über die Regeln für Lineartransformationen aus Erwartungswert und Varianz der hypergeometrischen Verteilung und betragen nun:

$$\mu(p_1) = \pi_1 = \frac{N_1}{N}$$

$$\sigma^2(p_1) = \frac{\pi_1 \cdot (1-\pi_1)}{n} \cdot \frac{N-n}{N-1} = \frac{\frac{N_1}{N} \cdot \frac{N-N_1}{N}}{n} \cdot \frac{N-n}{N-1}$$

Vor einer Anwendung der Verteilungen ist zu überlegen, ob die hypergeometrische Verteilung oder die Binomialverteilung herangezogen werden sollte:

- Bei einer einfachen Zufallsauswahl ohne Zurücklegen und bekanntem Populationsumfang N ist die hypergeometrische Verteilung die korrekte Verteilung.
- Wenn der Populationsanteil relativ zum Stichprobenanteil sehr groß ist, N > 20·n, kann als Näherung auch bei einer einfachen Zufallsauswahl ohne Zurücklegen die Binomialverteilung herangezogen werden, wenn die Berechnung anderenfalls zu aufwändig ist.
- Die Binomialverteilung wird als Näherung auch dann herangezogen, wenn der Populationsumfang N nicht bekannt ist.
- Bei einer einfachen Zufallsauswahl mit Zurücklegen wird die Binomialverteilung herangezogen.

Wahrscheinlichkeiten von Anteilen bei einfachen Zufallsauswahlen ohne Zurücklegen

Anwendungsbeispiel:
Wie wahrscheinlich ist es, dass bei der einfachen Zufallsauswahl mit bzw. ohne Zurücklegen von n=2 aus N=9 Haushalten der Anteil der ausgewählten Haushalte, die maximal 2000 € Monatseinkommen haben, 0.0, 0.5 bzw. 1.0 beträgt, wenn der Populationsanteil von Haushalten mit maximal 2000 € Monatseinkommen 1/3 beträgt.

Bei einer einfachen Zufallsauswahl mit Zurücklegen betragen die Wahrscheinlichkeiten:

$$\Pr(p_1 = 0.0) = \binom{2}{0} \cdot \left(\frac{1}{3}\right)^0 \cdot \left(\frac{2}{3}\right)^2 = 0.444 \; ; \; \Pr(p_1 = 0.5) = \binom{2}{1} \cdot \left(\frac{1}{3}\right)^1 \cdot \left(\frac{2}{3}\right)^1 = 0.444 \; ;$$

$$\Pr(p_1 = 1.0) = \binom{2}{2} \cdot \left(\frac{1}{3}\right)^2 \cdot \left(\frac{2}{3}\right)^0 = 0.111$$

Bei einer einfachen Zufallsauswahl ohne Zurücklegen ist zunächst $N_1 = 1/3 \cdot 9 = 3$ zu berechnen. Die Wahrscheinlichkeiten betragen dann:

$$\Pr(p_1 = 0.0) = \frac{\binom{3}{0} \cdot \binom{6}{2}}{\binom{9}{2}} = 0.417 \; ; \; \Pr(p_1 = 0.5) = \frac{\binom{3}{1} \cdot \binom{6}{1}}{\binom{9}{2}} = 0.5 \; ; \; \Pr(p_1 = 1.0) = \frac{\binom{3}{2} \cdot \binom{6}{0}}{\binom{9}{2}} = 0.083$$

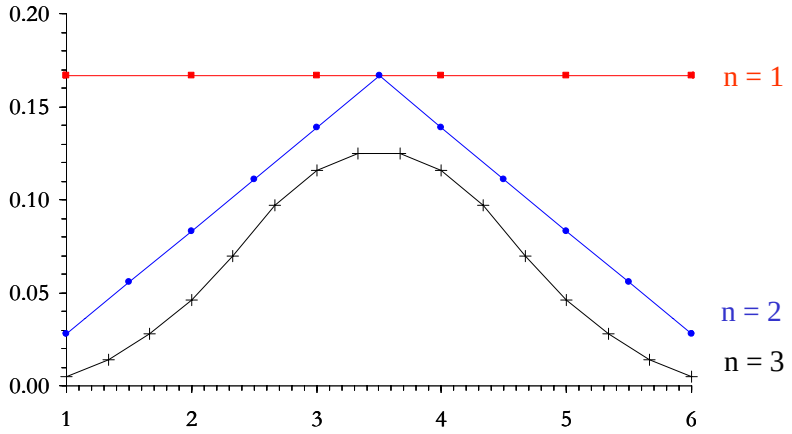
Lerneinheit 12: Stetige Wahrscheinlichkeitsverteilungen

Mit Hilfe der in Lerneinheit L11 vorgestellten Wahrscheinlichkeitsverteilungen können bei einfachen Zufallsauswahlen die Wahrscheinlichkeiten von Stichprobenkennwerten wie absolute oder relative Häufigkeiten eines Merkmals berechnet werden.
Wenn die Stichprobengrößen ansteigen oder eine Variable in der Population sehr viele Ausprägungen hat, werden die Berechnungen allerdings sehr aufwendig, da einerseits einzelne Ereignisse sehr geringe Auftretenswahrscheinlichkeiten haben und andererseits sehr viele Einzelereignisse zu einem interessierenden Ereignis zusammengefasst werden müssen.

Ein Beispiel ist die Berechnung der Kennwerteverteilung von Stichprobenmittelwerten. Bei der Vorstellung des Begriffs der Zufallsvariable wurde die Verteilung des Stichprobenmittelwerts bei einer einfachen Zufallsauswahl mit Zurücklegen des Umfangs n=2 aus einer Population des Umfangs N=6 betrachtet, wobei die sechs Elemente (im Beispiel Haushalte) jeweils einen unterschiedlichen Wert der interessierenden Variable, nämlich 1, 2, ..., 6 Tausend Euro Einkommen aufwiesen (vgl. L10).
Wenn ein einziges Elemens ausgewählt wird, gibt es sechs mögliche Stichproben mit 6 verschiedenen Ausprägungen des Haushaltseinkommens. Bei dem betrachteten Fall von n=2 ausgewählten Elementen gibt es bereits 36 (=6^2) Stichproben, die zu 11 verschiedenen Ausprägungen des Durchschnittseinkommens in der Stichprobe führen. Steigt die Fallzahl weiter an, gibt es bei n=3 Fällen in der Stichprobe bereits 216 (=6^3) Stichproben mit 16 unterschiedlichen Werten des Durchschnittseinkommens.

Stetige Wahrscheinlichkeitsverteilungen

Wahrscheinlichkeitsverteilungen des Mittelwerts bei einfacher Zufallauswahl mit Zurücklegen bei n = 1, 2 und 3 aus N=6



Die Abbildung zeigt für dieses Beispiel die Wahrscheinlichkeitsverteilungen der Stichprobenmittelwerte, wobei die Realisierungswahrscheinlichkeiten jeder Verteilung durch eine durchgezogene Linie verbunden sind.
Je größer der Stichprobenumfang ansteigt, desto mehr Ausprägungen gibt es. Da sich alle Wahrscheinlichkeiten zu eins addieren, sinken tendenziell die Auftretenswahrscheinlichkeiten bei steigender Zahl der Ausprägungen.
An der Abbildung fällt zudem auf, dass sich die Form der Verteilung ändert und mit steigendem Stichprobenumfang einer unimodalen symmetrischen Glockenform nähert.

Der zentrale Grenzwertsatz

Diese Annäherung der Verteilungsform ist keine Besonderheit, sondern notwendige Folge des nach dem Gesetz der großen Zahl wichtigsten Theorem der Statistik, dem sogenannten ***zentralen Grenzwertsatz***.

Der zentrale Gremzwertsatz besagt, dass sich unter sehr allgemeinen Bedingungen unabhängig von der Ausgangsverteilung die Wahrscheinlichkeitsverteilung einer Summe statistisch unabhängiger und identisch verteilter Zufallsvariablen einer Normalverteilung annähert:

> **Die Summe unabhängiger und identisch verteilter Zufallsvariablen nähert sich bei steigender Zahl von Summanden asymptotisch einer Normalverteilung an.**

Ein Stichprobenmittelwert ist nun bei jeder einfachen Zufallsauswahl mit Zurücklegen die Summe identisch verteiler Zufallsvariablen:

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = \frac{X_1}{n} + \frac{X_2}{n} + \ldots + \frac{X_n}{n}$$

Da bei jeder Auswahl die Wahrscheinlichkeitsverteilung der mögliche Wert der Realisierung die gleiche Wahrscheinlichkeitsverteilung hat – im Beispiel kann jeder der 6 möglichen Werte des Haushaltseinkommens in der Population mit gleicher Wahrscheinlichkeit ausgewählt werden – und die Auswahlen unabhängig voneinander erfolgen, ist der Stichprobenmittelwert die Summe identisch verteilter unabhängiger Zufallsvariablen.

Der zentrale Grenzwertsatz

Die Voraussetzungen für die Anwendbarkeit des zentralen Grenzwertsatzes beziehen sich auf die Ausgangsverteilung. Für diese muss gelten, dass ihr Erwartungswert eine beliebige reelle Zahl ungleich $\pm\infty$ ist und dass die Varianz ebenfalls eine beliebige reelle Zahl ungleich Null und ungleich $+\infty$ ist.

Diese Bedingungen werden für empirische Populationen, aus denen eine einfache Zufallsauswahl mit Zurücklegen gezogen wird, stets erfüllt.

Der zentrale Grenzwertsatz kann sogar insofern verallgemeinert werden, als die Wahrscheinlichkeitsverteilungen der aufsummierten Zufallsvariablen nicht wirklich identisch und strikt statistisch unabhängig voneinander sein müssen. Allerdings ist es dann schwierig, die Bedingungen zu klären, die erfüllt sein müssen. Außerdem kann die Annäherung an eine Normalverteilung dann sehr langsam erfolgen.

Da für eine Summe aus unabhängigen Zufallsvariablen gilt, dass ihr Erwartungswert die Summe der Erwartungswerte und ihre Varianz die Summe der Varianzen der Summanden ist, können Erwartungswert und Varianz bei größeren Summen leicht sehr große Werte annehmen.
Um dies zu vermeiden, kann es sinnvoll sein, die Ausgangsvariablen zu standardisieren und die Summe durch die Quaratwurzel aus der Zahl der Summanden zu teilen:

$$Z = \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^{n} \frac{X_i - \mu_X}{\sigma_X} = \frac{\left(\sum_{i=1}^{n} x_i\right) - n \cdot \mu_X}{\sqrt{n \cdot \sigma_X^2}} \quad \Rightarrow \quad \mu_Z = 0 \text{ und } \sigma_Z^2 = 1$$

Der zentrale Grenzwertsatz

Die transformierten Summe erfüllt weiterhin die Bedingungen des zentralen Grenzwertsatzes, nähert sich also einer Normalverteilung an, die dann zu einer ***Standardnormalverteilung*** wird, d.h. standardisiert ist und daher einen Erwartungswert von Null und eine Varianz von Eins aufweist.

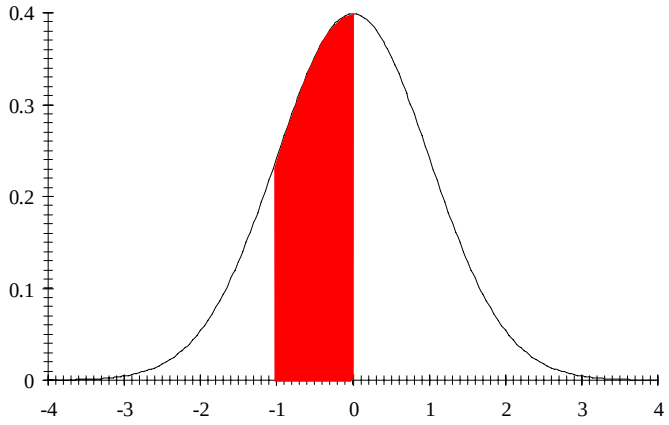
Aufgrund ihrer großen Bedeutung wird für die **Standardnormalverteilung** ein eigenes Symbol verwendet und zwar der kleine griechisches Buchstabe **ϕ** („phi“) für die **Wahrscheinlichkeits-verteilung** und der entsprechende große griechische Buchstabe **Φ** („Phi“) für die **Verteilungs-funktion**.

Der zentrale Grenzwertsatz kann daher so formuliert werden, dass sich die standardisierte Summe unabhängiger und identisch verteilter Zufallsvariablen asymptotisch einer Standardnormal-verteilung annähert:

$$\lim_{n\to\infty}\left(\Pr\left(\frac{\sum_{i=1}^{n} X_i - n\cdot\mu_X}{\sqrt{n\cdot\sigma_X^2}} \right)\right) = \phi$$

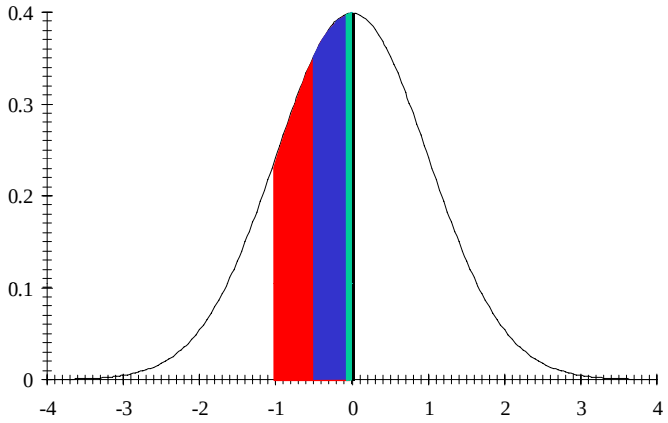
Im Unterschied zu allen bisher betrachteten Wahrscheinlichkeitsverteilungen ist die Standard-normalverteilung ein Beispiel für eine stetige Wahrscheinlichkeitsverteilung.

Stetige Wahrscheinlichkeitsverteilungen



Der Wertebereich ***stetiger (kontinuierlicher) Wahrscheinlichkeitsverteilungen*** umfasst alle reellen Zahlen bzw. Intervalle von reellen Zahlen.
Anders als bei der Poisson-Verteilung, die abzählbar unendlich viele Ausprägungen aufweist, gibt es bei stetigen Verteilungen in jedem beliebig kleinen Intervall des Wertebereichs eine unbegrenzte Zahl möglicher Realisierungen. Da die Wahrscheinlichkeit des Auftretens der Gesamtheit aller Realisierungen Eins sein muss, kann bei stetigen Wahrscheinlichkeitsverteilungen die Realisierungswahrscheinlichkeit jeder einer einzelnen Ausprägung nur Null sein. Angebbar ist daher immer nur die Wahrscheinlichkeit, mit der eine Realisierung in ein Intervall fällt.

Wahrscheinlichkeitsdichten



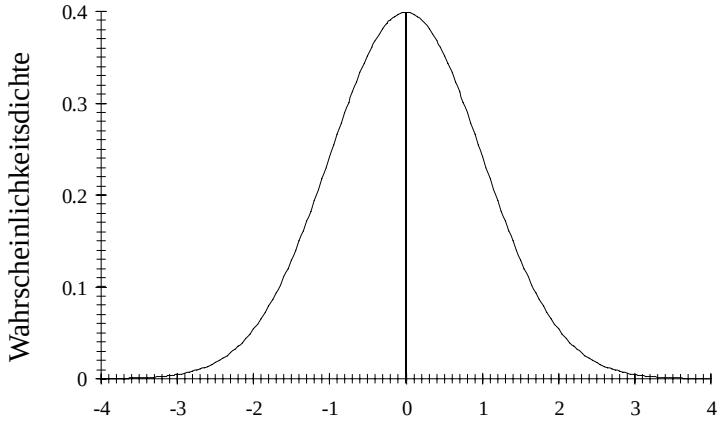
So zeigt der rot markierte Bereich, die Wahrscheinlichkeit, mit der bei der abgebildeteten Standardnormalverteilung eine Realisierung zwischen –1 und 0 liegt.

Je „dünner“ ein solches Intervall wird, desto geringer ist die Wahrscheinlichkeit, dass eine Realisierung in das Intervall fällt.

So sinkt im Beispiel die Realisierungswahrscheinlichkeit immer mehr, je mehr sich die Untergrenze des Intervalls der Obergrenze 1 nähert.

Im Extremfall hat das Intervall die Dicke null, d.h. die zweidimensionale Fläche wird zu einer eindimensionalen Linie von der Kurve bis zur unteren waagerechten Achse.

Wahrscheinlichkeitsdichten

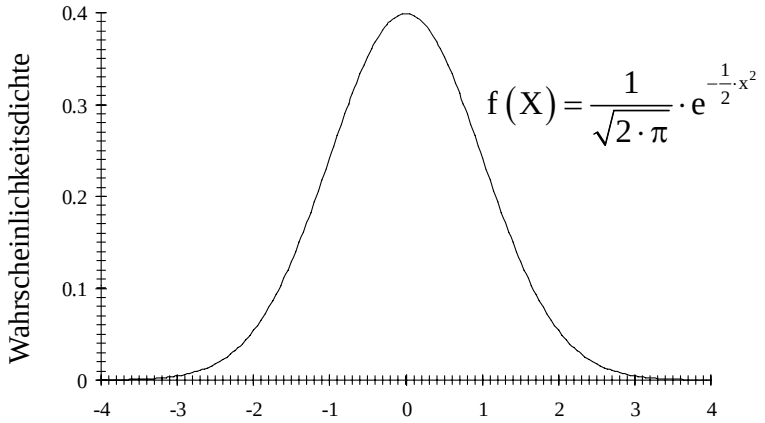


Die Länge dieser Linie wird als ***Wahrscheinlichkeitsdichte*** *(engl.* ***density****)* **f(X)** bezeichnet. Ihr jeweiliger Wert ergibt sich durch die Funktion, die den Kurvenverlauf der Wahrscheinlichkeitsverteilung beschreibt.

Im Beispiel der Kurve der Standardnormalverteilung berechnet sich die Dichte über die folgende Funktion:

$$f(X) = \phi(X) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot x^2}$$

Wahrscheinlichkeitsdichten



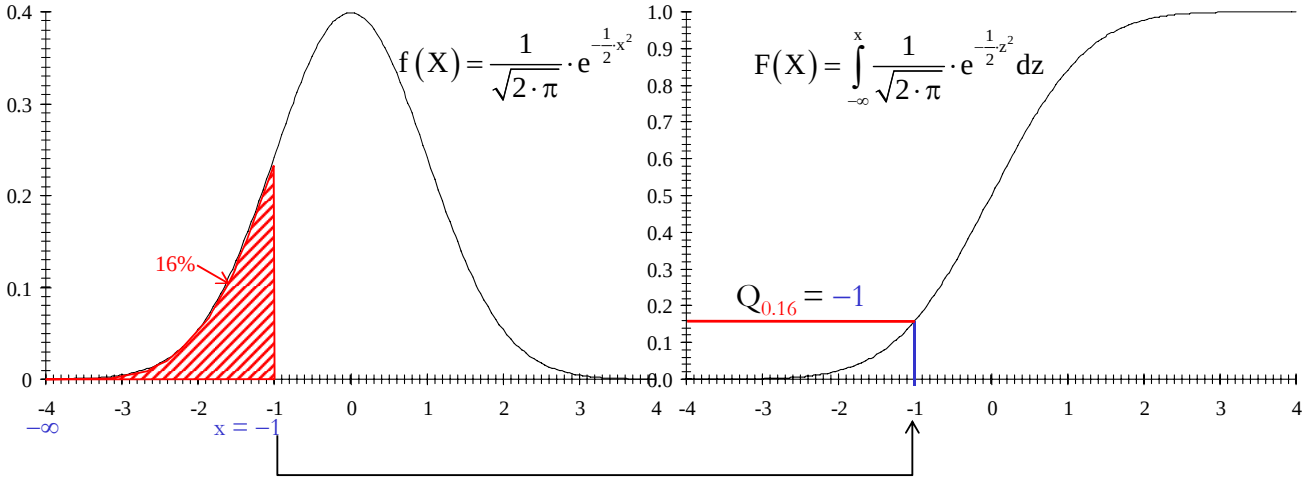
Obwohl die Realisierungswahrscheinlichkeit eines spezifischen reellen Wertes bei stetigen Verteilungen stets Null ist, kann die Wahrscheinlichkeitsdichte von Null verschieden sein. Wenn das der Fall ist, kann der Wert realisiert werden.
Eine Wahrscheinlichkeit von Null bedeutet bei einer stetigen Verteilung also nur dann ein unmögliches Ereignis, wenn auch die Wahrscheinlichkeitsdichte Null ist.

Wahrscheinlichkeitsdichten haben ähnliche Eigenschaften wie Wahrscheinlichkeiten,

> *So gibt das Verhältnis der Dichtewerte zweier Ausprägungen einer stetigen Variablen die relative Chance des Auftretens der beiden Ausprägungen an.*

Sie *summieren* sich allerdings nicht zum Wert Eins, sondern ***integrieren*** sich über den gesamten Wertebereich zu diesem Wert.

Verteilungsfunktion einer stetigen Zufallsvariablen

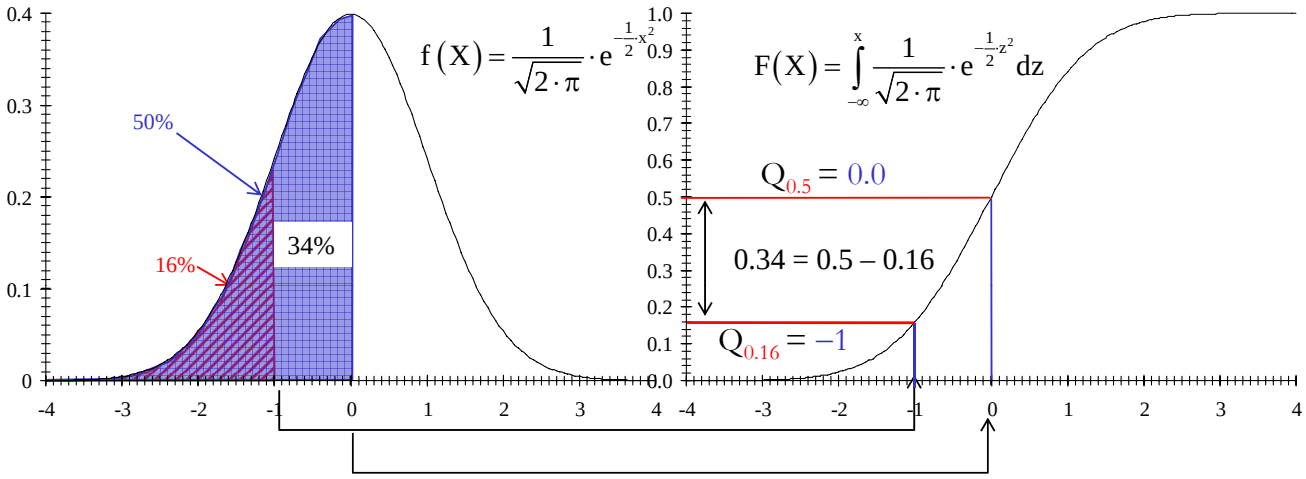


Die ***Verteilungsfunktion*** F(X=x) ist bei einer stetigen Wahrscheinlichkeitsverteilung die Fläche vom linken Rand der Verteilung (bzw. –∞) bis zum Wert X.
Mathematisch ist diese Fäche das ***bestimmte Integral*** über die Dichtefunktion f(X) von minus unendlich bis zum Wert x.

> *So ist z.B., die Wahrscheinlichkeit, dass eine standardnormalverteilte Größe kleiner gleich –1 ist, die Fläche unter der Kurve vom linken Extrem bis zur Stelle minus eins. Diese Fläche beträgt 16% der Gesamtfläche von 1.0.*

Die Verteilungsgfunktion lässt sich auch grafisch darstellen und ergibt bei einer Normalverteilung eine s-förmige Kurve.

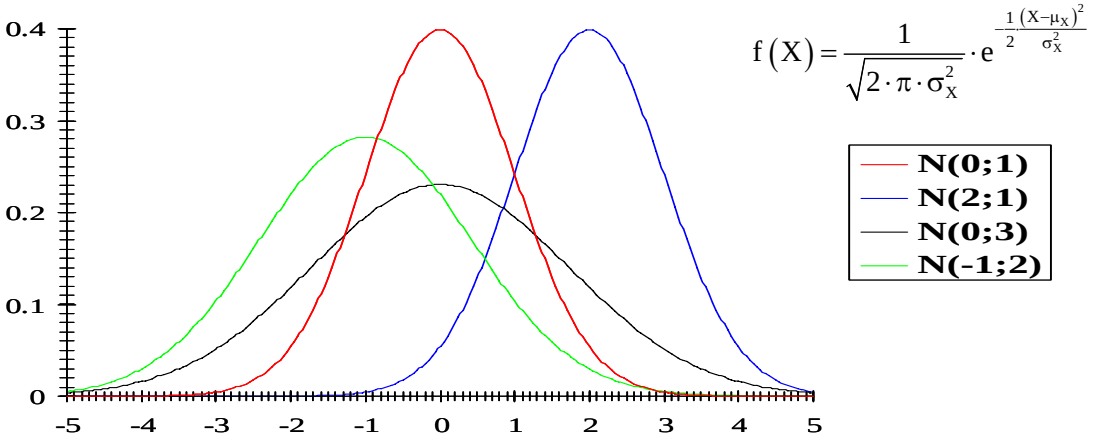
Verteilungsfunktion einer stetigen Zufallsvariablen



Über die Verteilungsfunktion einer stetigen Zufallsvariablen lassen sich für beliebige Intervalle des Wertebereichs Realisierungswahrscheinlichkeiten berechnen.

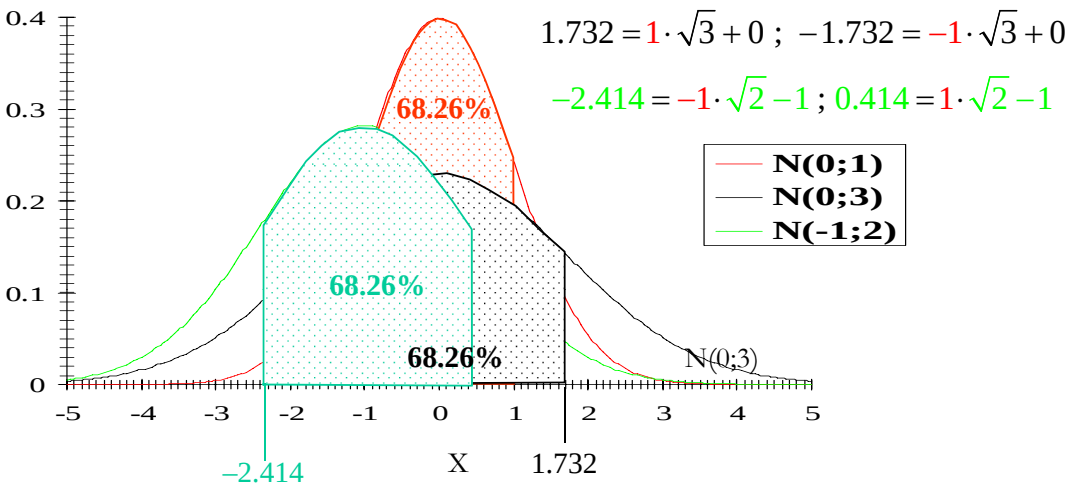
So ist z.B. die Quantilwahrscheinlichkeit des Quantilwerts 0 der Standardnormalverteilung 0.5 oder 50%.
Die Quantilwahrscheinlichkeit des Quantilwerts –1 der Standardnormalverteilung beträgt 0.16 oder 16%.
Dann ist die Wahrscheinlichkeit, dass eine standardnormalverteilte Zufallsvariable zwischen –1 und 0 liegt, 34% (= 50% – 16%).

Die Normalverteilung



Aufgrund des zentralen Grenzwertsatzes ist die ***Normalverteilung*** die wichtigste stetige Wahrscheinlichkeitsverteilung, was auch ihren Namen begründet, obgleich nur die wenigsten Verteilungen tatsächlich normalverteilt sind.
Alle Normalverteilungen haben eine glockenförmige Dichtefunktion, wobei die Dichtefunktion einer normalverteilten Zufallsvariable X eine Funktion ihres Erwartungswertes und ihrer Varianz ist. Erwartungswert μ_X und Varianz σ^2_X (bzw. Standardabweichung σ_X) sind die Parameter einer Normalverteilung.
Um auszudrücken, dass eine Zufallsvariable X mit dem Erwartungswert μ und der Varianz σ^2 normalverteilt ist, wird dass Symbol „$N(\mu ; \sigma^2)$“ oder „$N(\mu , \sigma)$“ verwendet.
Je größer die Varianz, desto flacher ist der Kurvenverlauf.

Die Normalverteilung



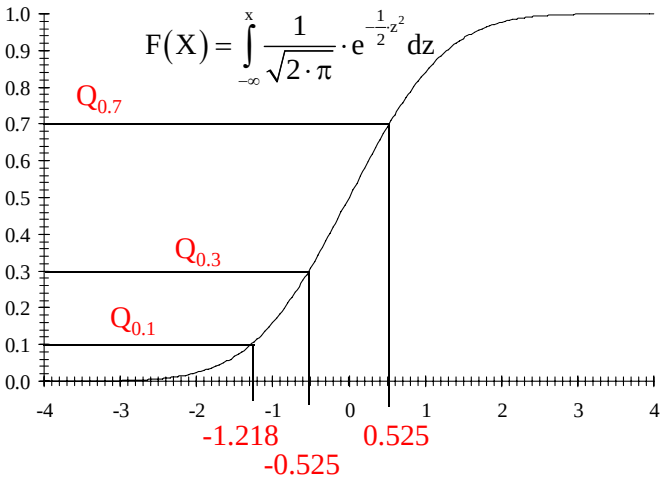
Kennzeichen einer Normalverteilung ist, dass in einem Abstand von ± 1 Standardabweichung vom Erwartungswert, der wegen der Symmetrie der Verteilung gleichzeitig Median und Modus ist, immer 68.26% aller Realisationen liegen, dass in einem Abstand von ± 2 Standardabweichungen vom Erwartungswert immer 95.44% aller Realisationen liegen, in einem Abstand von ± 3 Standardabweichung vom Erwartungswert immer 99.72%, usw..

Aufgrund dieser Eigenschaft ist es leicht möglich, Quantile von Normalverteilungen ineinander umzurechnen:

$$Q_{\alpha;N(\mu,\sigma)} = Q_{\alpha;N(0,1)} \cdot \sigma + \mu \quad \text{bzw.} \quad Q_{\alpha;N(0,1)} = \frac{Q_{\alpha;N(\mu,\sigma)} - \mu}{\sigma}$$

Die Normalverteilung

α	z_α
0.100	–1.218
0.200	–0.842
0.250	–0.674
0.300	–0.524
0.400	–0.253
0.500	0.000
0.600	0.253
0.700	0.524
0.800	0.674
0.900	0.842
1.000	1.218

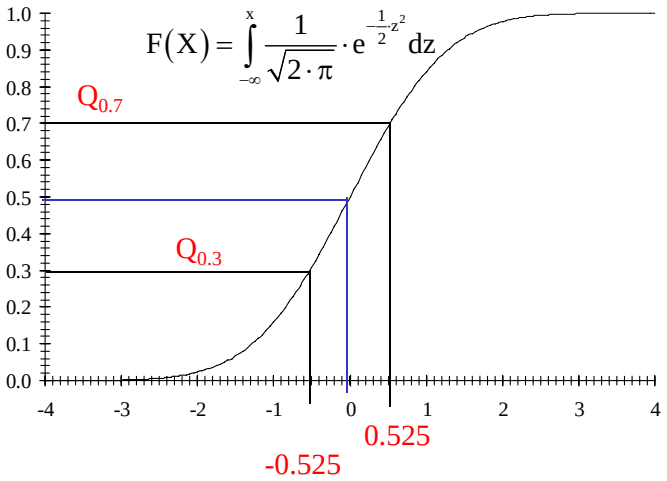


In vermutlich allen Statistiklehrbüchern finden sich Tabellen, in denen die Quantilanteile und Quantilwerte der Standardnormalverteilung aufgelistet sind. Da die Standardnormalverteilung standardisiert ist, werden die Quantilwerte oft als ***Z-Werte*** bezeichnet, wobei allerdings zu beachten ist, dass bisweilen auch nichtnormalverteilte Realisierungen so bezeichnet werden, wenn sie sich auf standardisierte Variablen beziehen.

Die Z-Werte bzw. die entsprechenden Anteilswerte oder Wahrscheinlichkeiten entsprechen den Argumenten und Funktionswerten der Verteilungsfunktion der Normalverteilung.

Die Normalverteilung

α	z_α
0.100	–1.218
0.200	–0.842
0.250	–0.674
0.300	–0.524
0.400	–0.253
0.500	0.000
0.600	0.253
0.700	0.524
0.800	0.674
0.900	0.842
1.000	1.218



Aufgrund der Symmetrie der Standdardnormalverteilung ist es hinreichend, nur die unteren oder die oberen 50% der Verteilung aufzulisten.

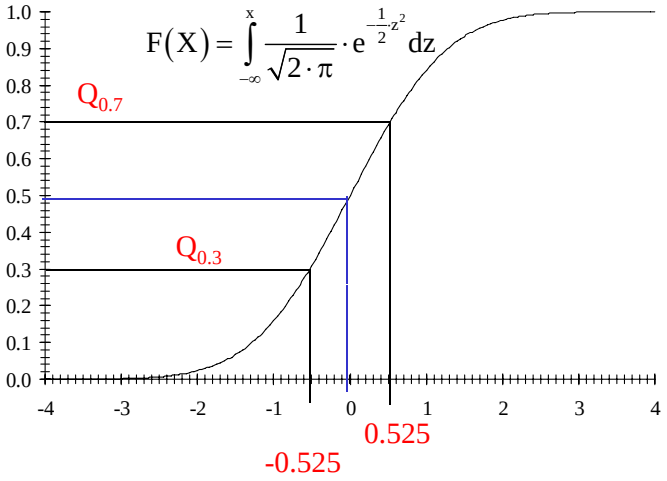
So ist der Wert des 30%-Quantils der Standardnormalverteilung gleich dem Negativen des 70%-Quantils bzw. umgekehrt das 70%-Quantil gleich minus Eins mal dem 30%-Quantil.

Generelle ist bei symmetrischen Verteilungen das (1–α)-Quantil mit umgekehrten Vorzeichen gleich weit vom Median und Mittelwert entfernt wie das α-Quantil:

$$(Q_{1-\alpha}-\mu)=-(Q_\alpha-\mu) \text{ bzw. bei } \mu=0\text{: } Q_{1-\alpha}=-Q_\alpha$$

Die Normalverteilung

α	z_α
0.100	–1.218
0.200	–0.842
0.250	–0.674
0.300	–0.524
0.400	–0.253
0.500	0.000
0.600	0.253
0.700	0.524
0.800	0.674
0.900	0.842
1.000	1.218



α	z_α
75.0%	0.674
90.0%	1.282
95.0%	1.645
97.5%	1.960
99.0%	2.326
99.5%	2.576
99.9%	3.090
99.95%	3.291

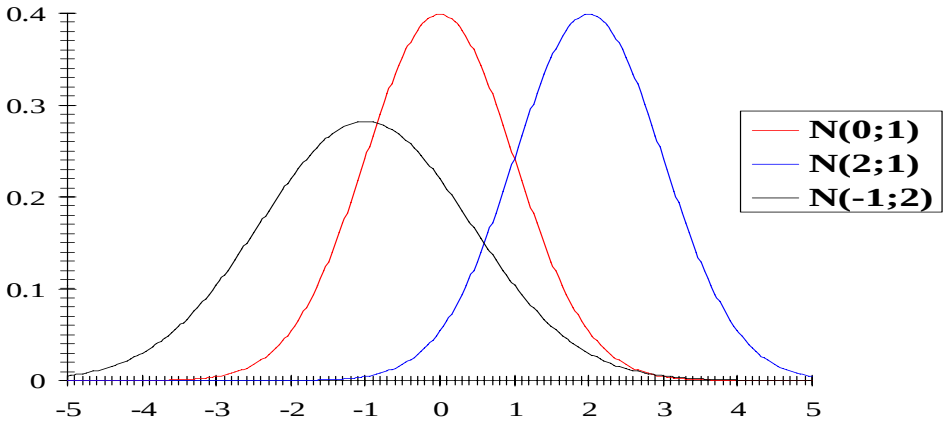
Da die Verteilungsfunktion der Standardnormalverteilung durch das große griechische Φ symbolisiert wird, lassen sich Quantilanteile und Quantilanteile hierüber symbolisieren:

$$\Phi(z_\alpha)=\alpha \text{ bzw. } \Phi^{-1}(\alpha)=z_\alpha.$$

Neben den links aufgeführten Perzentilen der Standardnormalverteilung werden in Anwendungen vor allem die rechts aufgeführten Quantile mit den Anteilen 75%, 90%, 95%, 97.5%, 99%, 99.5%, 99,9% und 99.95 % bzw. die entsprechenden Werte am linken Rand der Verteilung 25%, 10%, 5%, 2.5%, 1%, 1/2%, 0.1% und 0.05% für die Berechnung der Realisierungswahrscheinlichkeiten von Intervallen der Normalverteilung benötigt.

Die Normalverteilung

α	z_α
75.0%	0.674
90.0%	1.282
95.0%	1.645
97.5%	1.960
99.0%	2.326
99.5%	2.576
99.9%	3.090
99.95%	3.291



Aufgrund der erwähnten Eigenschaft von Normalverteilungen ist die Umrechnung von der Standardnormalverteilung in andere Normalverteilung sehr einfach.

So ist das 75-Quantil der Standardnormalverteilung $\Phi^{-1}(0.75) = 0.674$. Aus diesem Wert lässt sich das 75%-Quantil jeder beliebigen Normalverteilung mit Erwartungswert μ und Standardabweichung σ durch Umkehrung der Z-Transformation berechnen:

$$Q_\alpha = \Phi^{-1}(\alpha)\cdot\sigma + \mu$$

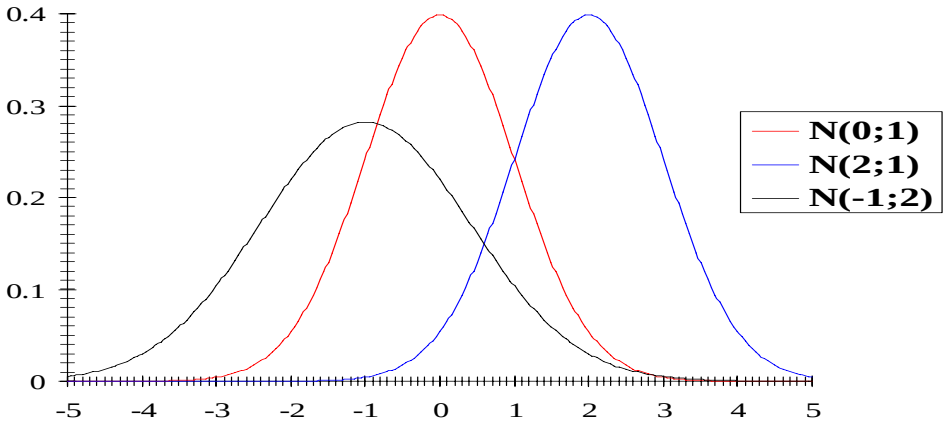
Für die abgebildeten Normalverteilungen ergibt sich so z.B:

$\Phi^{-1}(0.75) = 0.674 \Rightarrow Q_{0.75;N(2;1)} = \Phi^{-1}(0.75)\cdot 1 + 2 = 2.674;$

$Q_{0.75;N(-1;2)} = \Phi^{-1}(0.75)\cdot\sqrt{2} - 1 = 0.674\cdot 1.414 - 1 = -0.047.$

Die Normalverteilung

α	z_α
75.0%	0.674
90.0%	1.282
95.0%	1.645
97.5%	1.960
99.0%	2.326
99.5%	2.576
99.9%	3.090
99.95%	3.291



Umgekehrt ergibt sich der korrespondierende Quantilanteil für jeden Wert x einer Normalverteilung durch die Z-Transformation des Wertes:

$$\alpha = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Für die drei abgebildeten Normalverteilungen ergibt sich so z.B:

Wenn $X \sim N(-1;2)$ und $x = 0.813 \Rightarrow F(0.813) = \Phi((0.813+1)/\sqrt{2}) = \Phi(1.282) = 0.90;$

Wenn $X \sim N(2;1)$ und $x = -4.326 \Rightarrow F(-4.326) = \Phi((-4.326-2)/1) = \Phi(-2.326)$

$= 1 - \Phi(+2.326) = 1 - 0.99 = 0.01.$

Anwendungen der Normalverteilung

Als Folge des zentralen Grenzwertsatzes gibt es sehr viele Anwendungen der Normalverteilung.

Asymptotische Annäherungen an eine Normalverteilung

Viele zunächst sehr unterschiedliche Wahrscheinlichkeitsverteilungen nähern sich asymptotisch der Normalverteilung an.

Dies gilt auch für die Binomialverteilung, die hypergeometrische Verteilung und die Poisson-Vertreilung.

a) Annäherung der Binomialverteilung an die Normalverteilung

Die Binomialverteilung ergibt sich als Summe von statistisch unabhängigen identisch verteilten Bernoulli-Verteilungen. Solange der Parameter π_1 größer 0 und kleiner 1 ist die Voraussetzung für den zentralen Grenzwertsatz gegeben. Bei steigender Zahl der Wiederholungen n nähert sich daher die Binomialverteilung einer Normalverteilung an.

Erwartungswert und Varianz bleiben dabei unverändert:

$$\lim_{n\to\infty}\left(b\left(X;n,\pi_1\right)\right) = N\left(n\cdot\pi_1; n\cdot\pi_1\cdot\left(1-\pi_1\right)\right)$$

In der Literatur finden sich unterschiedliche Kriterien, ab wann die Annäherung einer Binomialverteilung an die Normalverteilung für praktische Berechnungen hinreichend genau ist.

Asymptotische Annäherungen an eine Normalverteilung

So soll das Produkt aus Anzahl der Wiederholungen und den Quotienten der Wahrscheinlichkeiten der Binomialverteilung größer 9 sein:

$$n\cdot\frac{\pi_i}{1-\pi_i} > 9 \quad \text{und} \quad n\cdot\frac{1-\pi_i}{\pi_i} > 9$$

Eine andere Faustregel besagt, dass das Produkt $n\cdot\pi_1\cdot(1-\pi_1) > 25$ sein muss oder auch nur > 5, wenn π_1 zwischen 0.1 und 0.9 liegt.

Die praktische Anwendung der Normalverteilung auf Binomialverteilungen hat das Problem zu lösen, dass die Binomialverteilung diskret, die Normalverteilung dagegen stetig ist. Bei Berechnungen werden daher die diskreten Ausprägungen einer Binomialverteilung als Klassenmittelwerte von Intervallen der Breite 1 mit exakten Klassengrenzen aufgefasst.
Dann berechnet sich die Wahrscheinlichkeit einer Realisation der Binomialverteilung über die Normalverteilung nach:

$$\Pr\left(X=x\left|b\left(X;n,\pi_1\right)\right.\right) = \binom{n}{x}\cdot\pi_1^x\cdot\left(1-\pi_1\right)^{n-x} \approx \Phi\left(\frac{x+0.5-n\cdot\pi_1}{\sqrt{n\cdot\pi_1\cdot\left(1-\pi_1\right)}}\right) - \Phi\left(\frac{x-0.5-n\cdot\pi_1}{\sqrt{n\cdot\pi_1\cdot\left(1-\pi_1\right)}}\right)$$

Bei den Extremwerten X = 0 und X = 1 ist zu beachten, dass bei X = 0 die Untergenze des Intervalls der Normalverteilung immer bei $-\infty$ liegt, das $\Phi(-\infty) = 0$, und bei X = 1 die Obergrenze des Intervalls immer bei $+\infty$ liegt, da $\Phi(\infty) = 1$.

Asymptotische Annäherungen an eine Normalverteilung

Die Verteilungsfunktion bzw. beliebige Intervalle binomialverteilter Größen berechnen sich dann entsprechend über passende Intervalle der Standardnormalverteilung

Als Beispiel sollen die Werte der Wahrscheinlichkeitsfunktion und der Verteilungsfunktion einer nach b(X; 16, 0.5) binomialverteilten Zufallsvariable für die Ausprägungen 0, 4, 8, 12 und 16 berechnet werden. Der Erwartungswert der Binomialverteilung ist dann $\mu_X=8$ und die Standardabweichung beträgt $\sigma_X=2$.
Für die Berechnung der Wahrscheinlichkeiten über die Quantilwerte der Standardnormalverteilung müssen zunächst die Unter- und Obergrenzen (Z_1, Z_2) der Intervalle berechnet und dann für diese Z-Werte die Quantilanteile von N(0;1) berechnet werden.

	Z=(X±0.5–8)/2 Z-Werte		Quantilanteile		Berechnung über Normalverteilung		Exakte Berechnung über Binomialverteilung	
X	Z_1	Z_2	$\Phi(Z_1)$	$\Phi(Z_2)$	Pr(X)	F(X)	Pr(X)	F(X)
0	–3.75	–∞	0.0001	0.0000	0.0001	0.0001	0.00002	0.00002
4	–1.75	–2.25	0.0401	0.0122	0.0279	0.0401	0.02777	0.03841
8	0.25	–0.25	0.5987	0.4013	0.1974	0.5987	0.19638	0.59819
12	2.25	1.75	0.9878	0.9599	0.0279	0.9878	0.02777	0.98936
16	+∞	3.75	1.0000	0.9999	0.0001	1.0000	0.00002	1.00000

Das Beispiel zeigt, dass die Berechnung bei allen Werten fast bis auf die dritte Nachkommastelle genau ist.

Asymptotische Annäherungen an eine Normalverteilung

b) Annäherung der hypergeometrischen Verteilung an die Normalverteilung

Obwohl bei einer einfachen Zufallsauswahl ohne Zurücklegen die insgesamt n Auswahlen nicht unabhängig voneinander sind, gilt der zentrale Grenzwertsatz auch für hypergeometrische Verteilungen:

$$\lim_{n\to\infty}\left(h\left(X;n,N,N_1\right)\right)=N\left(n\cdot\frac{N_1}{N};n\cdot\frac{N_1}{N}\cdot\frac{1-N_1}{N}\cdot\frac{N-n}{N-1}\right)$$

Wie bei der Binomialverteilung bestimmt der Erwartungswert und die Varianz der hypergeometrischen Verteilung die Parameter der asymptotischen Normalverteilung. Die Auftretenswahrscheinlichkeiten der Ausprägungen berechnen sich wieder über Intervalle der Standardnormalverteilung:

$$\Pr\left(X=x\middle|h\left(X;n,N,N_1\right)\right)=\frac{\binom{N_1}{x}\cdot\binom{N-N_1}{n-x}}{\binom{N}{n}}$$

$$\approx\Phi\left(\frac{x+0.5-n\cdot\frac{N_1}{N}}{\sqrt{n\cdot\frac{N_1}{N}\cdot\frac{N-N_1}{N}\cdot\frac{N-n}{N-1}}}\right)-\Phi\left(\frac{x-0.5-n\cdot\frac{N_1}{N}}{\sqrt{n\cdot\frac{N_1}{N}\cdot\frac{N-N_1}{N}\cdot\frac{N-n}{N-1}}}\right)$$

Anwendungen der Normalverteilung als Kennwerteverteilung in Stichproben

Für die Sozialforschung liegt die Bedeutung der Normalverteilung vor allem darin, dass sie als Kennwerteverteilung beim Schätzen und Testen von Populationsparametern eingesetzt wird.

a) Normalverteilung als Kennwerteverteilung von Stichprobenanteilen

Da Stichprobenanteile Lineartransformationen von absoluten Häufigkeiten sind, lassen sich die Kennwerteverteilungen von Stichprobenanteilen in einfachen Zufallsauswahlen über die hypergeometrische Verteilung bzw. die Binomialverteilung berechnen.
Die asymptotische Annäherungen dieser Verteilungen an die Nornalverteilung gelten dann auch für Stichprobenanteile.

Die bei der Berechnung von absoluten Häufigkeiten verwendete ***Stetigkeitskorrektur*** durch die Ersetzung einer einzelnen Häufigkeit durch das entsprechende Intervall von ±0.5 um den Häufigkeitswert wird bei relativen Häufigkeiten zum Intervall von ±0.5/n um den Stichprobenanteil.
Bei großen Stichproben, wie sie in der Umfrageforschung üblich sind, haben Stichprobenanteile aber auch bei einer dichotomen Variable in der Grundgesamtheit so viele mögliche Ausprägungen, dass sie praktisch wie stetige Variablen behandelt werden können. Auf die Stetigkeitskorrektur wird daher meist verzichtet.
Die (asymptotische) Kennwerteverteilung von Stichprobenanteile berechnet sich daher nach:

$$f\left(p_1\right) \approx N\left(n \cdot \pi_1; n \cdot \pi_1 \cdot \left(1-\pi_1\right)\right)$$

Anwendungen der Normalverteilung als Kennwerteverteilung in Stichproben

$$f\left(p_1\right) \approx N\left(n \cdot \pi_1; n \cdot \pi_1 \cdot \left(1-\pi_1\right)\right)$$

In der Gleichung steht p_1 für die Realisierungen der Stichprobenanteile, n für die Fallzahl in der Stichprobe und π_1 für den Anteil N_1/N der betrachteten Eigenschaft in der Population.

Wenn der Stichprobenumfang n relativ zum Populationsumfang N groß ist ($N/n \leq 20$) und die Stichprobe eine einfache Zufallsauswahl ohne Zurücklegen ist, wird für die Standardabweichung der Kennwerteverteilung die Varianz der hypergeometrischen Verteilung herangezogen, so dass dann gilt:

$$f\left(p_1\right) \approx N\left(n \cdot \pi_1; n \cdot \pi_1 \cdot \left(1-\pi_1\right) \cdot \frac{N-n}{N-1}\right)$$

Die Anwendung der Kennwerteverteilung setzt voraus, dass der Populationsanteil π_1 bekannt ist. Wenn dies nicht der Fall ist, gilt die asymptotische Annäherung bei hinreichend großen Stichproben auch dann, wenn bei der Berechnung der Varianz in der Wurzel unter dem Bruchstrich der Stichprobenanteil p_1 eingesetzt wird.

Als Faustregel gilt, dass der Stichprobenumfang $n > 60$ sein soll. Dann gilt:

$$f\left(p_1\right) \approx N\left(n \cdot \pi_1; n \cdot p_1 \cdot \left(1-p_1\right)\right)$$

Anwendungen der Normalverteilung als Kennwerteverteilung in Stichproben

b) Normalverteilung als Kennwerteverteilung von Stichprobenmittelwerten

Der zentrale Grenzwertsatz wurde am Beispiel von Stichprobenmittelwerten erläutert. Tatsächlich gilt, dass bei einfachen Zufallsauswahlen aus empirischen Populationen Stichprobenmittelwerte stets asymptotisch normalverteilt sind, wenn die betrachtete Größe in der Population eine Varianz ungleich Null aufweist.

Für Anwendungen ist es notwendig, Erwartungswert und Varianz der Kennwerteverteilung zu kennen. Da bei einfachen Zufallsauswahlen mit Zurücklegen der Erwartungswert jeder Realisierung einer Variablen X gleich dem Mittelwert μ_X von X in der Population ist und die Varianz der Realisierung gleich der Varianz σ^2_X von X in der Population, folgt aus der Regel für Erwartungswerte und Varianzen von Linearkombinationen unabhängiger Zufallsvariablen:

$$\mu\left(\bar{X}\right) = \mu\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = n \cdot \frac{\mu_X}{n} = \mu_X \text{ und } \sigma^2\left(\bar{X}\right) = \sigma^2\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = n \cdot \frac{\sigma_X^2}{n^2} = \frac{\sigma_X^2}{n}$$

In einfachen Zufallsauswahlen mit Zurücklegen gilt daher für die asymptotische Kennwerteverteilung von Stichprobenmittelwerten:

$$f\left(\bar{X}\right) \approx N\left(\mu_X; \frac{\sigma_X^2}{n}\right)$$

Kennwerteverteilung von Stichprobenmittelwerten

Wie schon bei Stichprobenanteilen gilt auch hier, dass bei einer einfachen Zufallsauswahl ohne Zurücklegen der Faktor, um den sich die Varianz der hypergeometrischen Verteilung von der Varianz der Binomialverteilung verringert, berücksichtigt werden muss, wenn die Population verglichen mit dem Stichprobenumfang nicht sehr groß ist ($N/n \leq 20$). Dann gilt:

$$f\left(\bar{X}\right) \approx N\left(\mu_X; \frac{\sigma_X^2}{n} \cdot \frac{N-n}{N-1}\right)$$

Das Eingangsbeispiel mit der Population von N = 6 Haushalten zeigte, dass die asymptotische Annäherung an die Normalverteilung bei Stichprobenmittelwerten oft sehr schnell erfolgt. Tatsächlich hängt es von der Verteilungsform der betrachteten Variable X in der Population ab, wie schnell die Annäherung erfolgt.
Für praktische Berechnungen hat sich gezeigt, dass die Annäherung nahezu immer genau genug ist, wenn der Stichprobenanteil mindestens 30 erreicht:

$$\mathbf{n \geq 30}$$

Wie schon bei der Kennwerteverteilung von Stichprobenanteilen zeigt sich auch hier, dass die Kennwerteverteilung von den Populationsparametern, hier μ_X und σ^2_X, abhängt. Und wie bei den Anteilen gilt auch hier, dass die asymptotische Annäherung an eine Normalverteilung auch dann funktioniert, wenn anstelle der Populationsvarianz eine Schätzung dieses Populationsparameters eingesetzt wird.

Kennwerteverteilung von Stichprobenmittelwerten

Für die Schätzung der Populationsvarianz wird allerdings anstelle der Stichprobenvarianz eine leicht modifizierte Formel eingesetzt:

$$f\left(\overline{X}\right) \approx N\left(\mu_X; \frac{\hat{\sigma}_X^2}{n}\right) = N\left(\mu_X; \frac{\sum_{i=1}^{n}\left(X_i - \overline{x}\right)}{n \cdot \left(n-1\right)}\right)$$

Kennwerteverteilung von Stichprobenmittelwerten bei Variablen aus normalverteilen Populationen

Die asymptotische Kennwerteverteilung für Stichprobenmittelwert gilt praktisch unabhängig davon, wie die interessierende Variable X in der Popuation, aus der die Stichprobe kommt, verteilt ist.
Da Normalverteilungen auch die Eigenschaft haben, dass alle Linearkombinationen von Normalverteilungen wiederum normalverteilt sind, folgt für Stichprobenmittelwerte aus normalverteilten Populationen, dass sie nicht nur asymptotisch, sondern exakt normalverteilt sind.

Für die Kennwerteverteilung bei einfachen Zufallsauswahlen gilt also:

$$f\left(\overline{X}\right) = N\left(\mu_X; \frac{\sigma_X^2}{n}\right) \text{ wenn X in der Population normalverteilt.}$$

Lerneinheit 13: Schätzen von Anteilen, Mittelwerten und Varianzen

Populationsverteilung:

Haush. einkom.	n_k	p_k	cp_k
1000	1	1/6	1/6
2000	1	1/6	2/6
3000	1	1/6	3/6
4000	1	1/6	4/6
5000	1	1/6	5/6
6000	1	1/6	6/6
Summe:	6	6/6	

$\mu_X = 3500 \; ; \; \sigma^2_X = 2916666.67$

$\mu_{\bar{x}} = 3500 \; ; \; \sigma^2_{\bar{x}} = 1458333.33$

Kennwerteverteilung:

$\bar{x}_i$ (mittleres Einkomen in €)	Wahrscheinlich-keitsfunktion	Verteilungs-funktion
1000	1/36	1/36
1500	2/36	3/36
2000	3/36	6/36
2500	4/36	10/36
3000	5/36	15/36
3500	6/36	21/36
4000	5/36	26/36
4500	4/36	30/36
5000	3/36	33/36
5500	2/36	35/36
6000	1/36	36/36
Summe:	36/36	

Die wichtigsten Anwendungen von Wahrscheinlichkeitsverteilungen in der Sozialforschung ergeben sich bei der Beurteilung von Schätzungen von Populationsparametern und der Prüfung von Hypothesen über ihre Werte auf der Basis von Stichprobendaten.

So ist in Lerneinheit L10 der Begriff der Wahrscheinlichkeitsverteilung einer Zufallsvariable am Beispiel der hier noch einmal wiedergegebenen Kennwerteverteilung des Stichprobenmittelwerts des Haushaltseinkommens bei einer einfachen Zufallsauswahl mit Zurücklegens des Umfangs n=2 aus einer Populaion von N=6 vorgestellt worden.

Schätzer und Schätzungen

Populationsverteilung:

Haush. einkom.	n_k	p_k	cp_k
1000	1	1/6	1/6
2000	1	1/6	2/6
3000	1	1/6	3/6
4000	1	1/6	4/6
5000	1	1/6	5/6
6000	1	1/6	6/6
Summe:	6	6/6	

$\mu_X = 3500 \; ; \; \sigma^2_X = 2916666.67$

$\mu_{\bar{x}} = 3500 \; ; \; \sigma^2_{\bar{x}} = 1458333.33$

Kennwerteverteilung:

$\bar{x}_i$ (mittleres Einkomen in €)	Wahrscheinlich-keitsfunktion	Verteilungs-funktion
1000	1/36	1/36
1500	2/36	3/36
2000	3/36	6/36
2500	4/36	10/36
3000	5/36	15/36
3500	6/36	21/36
4000	5/36	26/36
4500	4/36	30/36
5000	3/36	33/36
5500	2/36	35/36
6000	1/36	36/36
Summe:	36/36	

In der Statistik versteht man unter einer Schätzung die Bestimmung des Wertes eines interessierenden Parameters einer (Populations-) Verteilung durch die Berechnung von geeigneten Statistiken aus den Realisierungen einer Zufallsstichprobe.

So werden im Beispiel die Stichprobenmittelwerte verwendet, um den Populationsmittelwert zu schätzen.

Schätzer und Schätzungen

Ziel ist es, möglichst gute Schätzungen zu erhalten. Da die Werte von Stichprobenstatistiken Realierungen von Zufallsvariablen sind, macht es allerdings wenig Sinn, Aussagen über eine konkrete Schätzung zu machen.

> *So zeigt das Eingangsbeispiel, dass einige Realisierungen der Zufallsvariable „Stichprobenmittelwert" den Populationswert von 3500€ exakt schätzen, andere dagegen deulich andere Werte aufweisen. Welcher Wert in einer Stichprobe aber vorkomt, ist völlig unvorhersehbar.*

Stattdessen beziehen sich Aussagen über die Qualität von Schätzungen stets auf die Wahrscheinlichkeitsverteilung der Statistik, die zur Schätzung herangezogen wird.
Um diesen Unterschied zu verdeutlichen, wird zwischen Schätzung und Schätzer unterschieden Ein ***Schätzer*** (*engl:* ***estimator***, bisweilen auch als ***Schätzfunktion*** bezeichnet) ist eine Zufallsvariable, die für Schätzungen eines Parameters herangezogen wird, eine ***Schätzung*** *(engl:* ***estimate****)* ist eine Realisation dieser Zufallsvariable in einer konkreten Stichprobe.

Durch die Betrachtung der Kennwerteverteilung eines Schätzers ist es möglich, dessen Eigenschaften zu bestimmen. Gesucht sind Schätzer mit erwünschten Eigenschaften. Die wichtigsten erwünschten Eigenschaften sind

- Erwartungstreue oder Unverzerrtheit
- Konsistenz und
- Effizienz.

Erwartungstreue

Ein Schätzer ist ***erwartungstreu*** oder ***unverzerrt*** *(engl:* ***unbiased****)*, wenn der Erwartungswert der Kennwerteverteilung des Schätzers mit dem zu schätzenden Populationswert übereinstimmt.

In der Statistik wird oft das griechische kleine Theta („θ") als Symbol für einen beliebigen Parameter verwendet. Ein kleines Dach („^") über dem Symbol kennzeichnet dann einen Schätzer dieses Paramters

Erwartungstreue lässt sich dann ausdrücken als:

$$\mu\left(\hat{\theta}\right) = \theta$$

> *In L12 wurde gezeigt, dass der Ewartungswert der Kennwerteverteilung von Stichproben bei einfachen Zufallsauswahlen gleich dem Populationsmittelwert der Variable sind, für die der Mittelwert berechnet wird. Stichprobenmittelwerte sind daher erwartungstreue Schätzer von Populationsmittelwerten. So ist auch im Beispiel der Erwartungswert der Kennwerteverteilung des mittleren Haushaltseinkommen in den Stichproben mit 3500€ gleich dem mittleren Einkommen über die N=6 Haushalte.*

Ist ein Schätzer nicht erwartungstreu, sondern verzerrt, wird die Differenz zwischen seinem Erwartungswert und dem zu schätzenden Parameter als ***Verzerrung*** *(engl:* ***bias****)* bezeichnet:

$$\text{bias} = \mu_{\hat{\theta}} - \theta = \mu\left(\hat{\theta} - \theta\right)$$

Konsistenz

Ein Schätzer ist ***konsistent*** *(engl:* ***consistent****)*, wenn mit steigendem Stichprobenumfang die Wahrscheinlichkeit gegen Eins geht, dass Abweichungen zwischen den Schätzungen und dem zu schätzenden Parameter einen beliebig kleinen positiven Wert ε nicht überschreiten:

$$\lim_{n\to\infty} \Pr\left(\left(\hat{\theta}_n - \theta\right) < \varepsilon\right) = 1$$

Um hierbei klarzustellen, dass ein Schätzer eine Funktion des Stichprobenumfangs n ist, ist er in der Gleichung durch den Buchstaben n indiziert: $\hat{\theta}_n$ bezeichnet also den Schätzer von θ in einer Stichprobe mit n Fällen,

Formal folgt aus der Konsistenz auch, dass der Erwartungswert der quadrierten Abweichungen des Schätzers von dem zu schätzenden Populationsparameter bei steigender Fallzahl gegen Null geht:

$$\lim_{n\to\infty} \mu\left(\left(\hat{\theta}_n - \theta\right)^2\right) = 0$$

Der Erwartungswert der quadrierten Abweichungen eines Schätzers vom zu schätzenden Parameter wird in der Statistik als ***mittlerer quadrierter Fehler*** bezeichnet und nach dem englischen Ausdruck ***mean squared error*** durch das Symbol **MSE** abgekürzt. Ein Schätzer ist also konsistent, wenn sein MSE mit wachsender Fallzahl gegen Null geht.

Konsistenz

Aus der Eigenschaft von Mittelwerten und Erwartungswerten, dass die Summe der quadrierten Abweichungen vom Mittelwert bzw. der Erwartungswert der quadrierten Abweichungen vom Erwartungswert minimal ist, folgt, dass der mittlere quadrierte Fehler stets als Summe aus der Varianz des Schätzers plus der quadrierten Verzerrung dargestellt werden kann:

$$\mathrm{MSE} = \mu\left(\hat{\theta} - \theta\right)^2 = \sigma^2\left(\hat{\theta}\right) + \left(\mu_{\hat{\theta}} - \theta\right)^2$$

Bei erwartungstreuen Schätzern ist der MSE gleich der Schätzervarianz, da die Verzerrung Null ist.

In L12 wurde gezeigt, dass die Varianz der Kennwerteverteilung von Stichprobenmittelwerten in einfachen Zufallsauswahlen die Populationsvarianz geteilt durch den Stichprobenumfang n ist, wobei bei einer Auswahl ohne Zurücklegen noch der Faktor (N–n)/(N–1) hinzukommt:

$$\sigma^2\left(\bar{X}\right) = \begin{cases} \dfrac{\sigma_X^2}{n} & \text{in einfachen Zufallsauswahlen mit Zurücklegen} \\ \dfrac{\sigma_X^2}{n} \cdot \dfrac{N-n}{N-1} & \text{in Zufallsauswahlen ohne Zurücklegen} \end{cases}$$

Da mit steigender Fallzahl die Varianz des Stichprobenmittelwerts gegen Null geht und er ein erwartungstreuer Schätzer ist, ist er auch ein konsistenter Schätzer.

Effizienz

Das dritte Kriterium für die Beurteilung eines Schätzers ist die Effizienz. Ein Kennwert ist ***effizient***, wenn es keinen anderen Schätzer mit geringerem quadrierten Fehler gibt:

$$\mathrm{MSE} = \mu\left(\left(\hat{\theta} - \theta\right)^2\right) \overset{!}{=} \min$$

Im Unterschied zur Unverzerrtheit und Konsistenz ist die Effizienz ein Gütemaß, das relativ bezogen auf andere Schätzer gilt.
Tatsächlich ist der Nachweis sehr schwer, dass ein Schätzer effizient ist. Leichter ist der Nachweis, dass der Schätzer in bestimmten Situationen (z.B. bei bestimmten Verteilungen in der Population) oder verglichen mit einer bestimmten Klasse von Schätzern effizient ist.

Wenn eine Verteilung in der Population symmetrisch ist, kann zur Schätzung des Populationsmittelwerts auch der Median herangezogen werden. Dies trifft z.B. auf Normalverteilungen zu. Es kann gezeigt werden, dass der Median bei einer einfachen Zufallsauswahl aus normalverteilten Populationen erwartungstreu ist und die Varianz der Kennwerteverteilung des Median $0.5 \cdot \pi \cdot \sigma^2_X / n$ ist. Daher gilt:

$$\mathrm{MSE}\left(\tilde{\mathrm{X}}\right) = \frac{\pi \cdot \sigma^2_{\mathrm{X}}}{2 \cdot \mathrm{n}} > \frac{\sigma^2_{\mathrm{X}}}{\mathrm{n}} = \mathrm{MSE}\left(\bar{\mathrm{X}}\right)$$

Bei normalverteilten Populationen ist der Stichprobenmittelwert ein effizienterer Schätzer des Populationsmittelwerts als der Stichprobenmedian.

Standardfehler

Um die Güte eines Schätzers nach den Kriterien Unverzerrtheit, Konsistenz und Effizienz zu beurteilen, muss der Erwartungswert und die Varianz eines Schätzers bekannt sein.
Für die praktische Nutzung ist insbesondere seine Varianz relevant. So mag zwar ein Schätzer erwartungstreu sein. Wenn seine Varianz aber groß ist, ist es gleichwohl möglich, dass viele Schätzungen den zu schätzenden Parameterwert deutlich verfehlen.

Da die Varianz nicht in der gleichen Einheit wie die zu schätzende Größe gemessen wird, wird anstelle der Varianz meist die Standardabweichung betrachtet. Die Standardabweichung der Kennwerteverteilung eines Schätzers wird als ***Standardschätzfehler*** oder einfach ***Standardfehler*** bezeichnet *(engl: **standard error**,* oft durch **SE** symbolisiert). Wie später gezeigt wird, ist die Kenntnis des Standardfehlers notwendig, um Konfidenzintervalle zu berechnen und statistische Tests durchzuführen.

Da der Standardfehler die Wurzel aus der Varianz der Kennwerteverteilung ist, berechnet sich der Standardfehler des Stichprobenmittelwerts bei einfachen Zufallsauswahlen mit Zurücklegen bzw. aus verglichen zur Stichprobengröße sehr großen Populationen nach:

$$\mathrm{SE}\left(\bar{\mathrm{X}}\right) = \sigma_{\bar{\mathrm{X}}} = \sqrt{\frac{\sigma^2_{\mathrm{X}}}{\mathrm{n}}} = \frac{\sigma_{\mathrm{X}}}{\sqrt{\mathrm{n}}}$$

Da die Populationsvarianz in der Regel unbekannt ist, wird meist der geschätzte Standardfehler verwendet:

$$\hat{\sigma}_{\bar{\mathrm{X}}} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{x}_{\mathrm{i}} - \bar{\mathrm{x}}\right)^2}{\mathrm{n} \cdot \left(\mathrm{n}-1\right)}} = \sqrt{\frac{\mathrm{SS}_{\mathrm{X}}}{\mathrm{n} \cdot \left(\mathrm{n}-1\right)}}$$

Schätzer von Anteilen, Mittelwerten und Varianzen

a) Schätzer der Populationsvarianz bzw. Standardabweichung

Die Populationsvarianz ist nicht nur eine zentrale Kenngröße, die bei metrischen und dichotomen Variablen die Unterschiedlichkeit der Ausprägungen der Elemente in der Population beschreibt, sie ist auch insofern wichtig, als sie die Kennwerteverteilung von Stichprobenmittelwerten beeinflusst.

Es liegt nahe, zur Schätzung einer Populationsvarianz die Stichprobenvarianz s^2_X zu verwenden. Tatsächlich ist die Stichprobenvarianz jedoch kein erwartungstreuer Schätzer, da aufgrund der Minimierungseigenschaft des Stichprobenmittelwerts (vgl. L06 und L07) folgendes gilt:

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2+\left(\bar{X}-\mu_X\right)^2=s_X^2+\left(\bar{X}-\mu_X\right)^2$$

$$\Rightarrow s_X^2=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2-\left(\bar{X}-\mu_X\right)^2$$

$$\Rightarrow \mu\left(s_X^2\right)=\frac{\sum_{i=1}^{n}\mu\left(\left(X_i-\mu_X\right)^2\right)}{n}-\mu\left(\left(\bar{X}-\mu_X\right)^2\right)=\frac{n\cdot\sigma_X^2}{n}-\sigma_{\bar{X}}^2=\sigma_X^2-\sigma_{\bar{X}}^2$$

$$\Rightarrow \mu\left(s_X^2\right)=\sigma_X^2-\frac{\sigma_X^2}{n}=\sigma_X^2\cdot\left(\frac{n-1}{n}\right)$$

Schätzer von Anteilen, Mittelwerten und Varianzen

$$\mu\left(s_X^2\right)=\mu\left(\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right)=\sigma_X^2-\frac{\sigma_X^2}{n}=\sigma_X^2\cdot\left(\frac{n-1}{n}\right)$$

Da (n–1)/n kleiner 1 ist, unterschätzt die Stichprobenvarianz die Populationsvarianz. Mit steigendem n wird die Verzerrung allerdings immer kleiner. Die Stichprobenvarianz ist daher ein ***asymptotisch erwartungstreuer Schätzer*** der Populationsvarianz.

Da die Höhe der Verzerrung bekannt ist, lässt sich ein erwartungstreuer Schätzer der Populationsvarianz finden, indem die Stichprobenvarianz mit dem Kehrwert der Verzerrung multipliziert wird.

Der erwartungstreue Schätzer der Populationsvarianz ist daher die ***geschätzte Populationsvarianz***

$$\hat{\sigma}_X^2=\frac{n}{n-1}\cdot s_X^2=\frac{SS_X}{n-1}=\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$$

Die (asymptotische) Varianz der Stichprobenvarianz ist eine Funktion der vierten zentralen Momente:

$$\sigma^2\left(s_X^2\right)=\frac{\mu_4-\left(\mu_2\right)^2}{n}=\frac{\mu\left(\left(X-\mu_X\right)^4\right)-\left(\sigma_X^2\right)^2}{n}$$

Da sowohl die Varianz der Kennwerteverteilung der Stichprobenvarianz wie auch ihre Verzerrung mit steigendem Stichprobenumfang gegen Null geht, ist die Stichprobenvarianz ein konsistenter Schätzer der Populationsvarianz.

Schätzer von Anteilen, Mittelwerten und Varianzen

Es kann gezeigt werden, dass die Kennwerteverteilung der Stichprobenvarianz und damit auch der geschätzten Popuationsvarianz bei steigender Fallzahl n asymptotisch normalverteilt ist. Allerdings verläuft die Annäherung an die Normalverteilung in Abhängigkeit von der Verteilung in der Population bisweilen sehr langsam.

Deutlich schneller ist die Annäherung, wenn die betrachtete Variable X in der Population normalverteilt ist. Bei Normalverteilungen ist zudem das vierte zentrale Moment eine Funktion der Varianz bzw. Standardabweichung:

$$\mu\left(\left(X_i - \mu_X\right)^4\right) = 3\left(\sigma^2\right)^2$$

Die Varianz der Kennwerteverteilung der Stichprobenvarianz ist bei normalverteilter Variable in der Population daher

$$\sigma^2\left(s_X^2\right) = \frac{3\cdot\left(\sigma_X^2\right)^2 - \left(\sigma_X^2\right)^2}{n} = \frac{2\cdot\sigma_X^4}{n}$$

Dann gilt für den Standardfehler des erwartungstreuen Schätzers der Populationsvarianz:

$$\sigma\left(\hat{\sigma}_X^2\right) = \sigma_X^2 \cdot \sqrt{\frac{2}{n-1}}$$

Schätzer von Anteilen, Mittelwerten und Varianzen

Die Kennwerteverteilung der Stichprobenvarianz ist bei normalverteilten Populationen proportional zur Chiquadratverteilung mit df = n–1 Freiheitsgraden. Die Chiquadratverteilung wird bei der Zusammenhangsanalyse zwischen zwei nominalskalierten Variablen vorgestellt. Hier reicht der Hinweis, dass ab etwa n > 30 Fällen eine hinreichend genaue Annäherung an die Normalverteilung berechnet werden kann.

Die ***geschätzte Standardabweichung*** in der Population ist die positive Quadratwurzel der geschätzten Populationsvarianz:

$$\hat{\sigma}_X = \sqrt{\hat{\sigma}_X^2} = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Wie die Stichprobenvarianz ist der geschätzte Populationsstandardabweichung nur ein asymptotisch erwartungstreuer Schätzer der Standardabweichung in der Population. Anders als bei der Stichprobenvarianz gibt es allerdings keinen einfachen Korrekturfaktor, um einen erwartungstreuen Schätzer zu erhalten.

Schätzer von Anteilen, Mittelwerten und Varianzen

b) Schätzer von Populationsmittelwerten

Als Schätzer für Populationsmittelwerte bietet sich der Stichprobenmittelwert an. Es wurde bereits gezeigt, dass die Stichprobenmittelwert in einfachen Zufallsauswahlen mit und ohne Zurücklegen ein erwartungstreuer Schätzer des Populationsmittelwertes ist. Der Standardfehler beträgt dann:

$$\sigma_{\bar{X}} = \begin{cases} \dfrac{\sigma_X}{\sqrt{n}} & \text{mit Zurücklegen} \\ \dfrac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\dfrac{N-n}{N-1}} & \text{ohne Zurücklegen} \end{cases}$$

Der geschätzte Standardfehler unterscheidet sich nur dadurch, dass anstelle der Populationsvarianz die geschätzte Populationsvarianz verwendet wird:

$$\sigma_{\bar{X}} = \begin{cases} \dfrac{\hat{\sigma}_X}{\sqrt{n}} = \sqrt{\dfrac{SS_X}{n \cdot (n-1)}} & \text{mit Zurücklegen} \\ \dfrac{\hat{\sigma}_X}{\sqrt{n}} \cdot \sqrt{\dfrac{N-n}{N-1}} = \sqrt{\dfrac{SS_X}{n \cdot (n-1)} \cdot \dfrac{N-n}{N-1}} & \text{ohne Zurücklegen} \end{cases}$$

Schätzer von Anteilen, Mittelwerten und Varianzen

Die Kennwerteverteilung ist des Stichprobenmittelwerts ist asymptotisch normalverteilt, wobei die Annäherung bei $n \geq 30$ in der Regel hinreichend genau ist.

Bei normalverteilten Populationen ist der Stichprobenmittelwert exakt normalverteilt. Dann ist die Z-Transformation des Stichprobenmittelwert standardnormalverteilt:

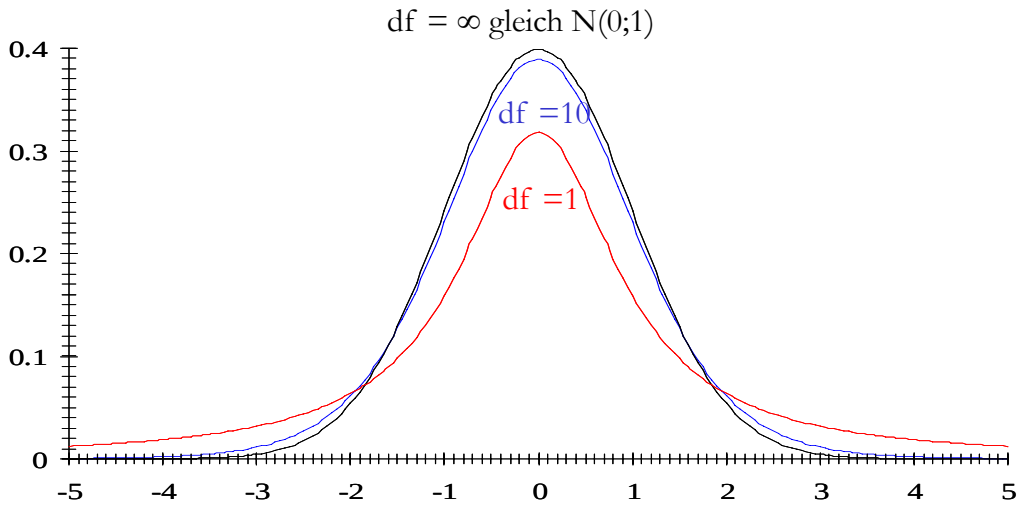
$$\frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}} \sim \phi$$

Wird dagegen die geschätzte Populationsvarianz für die Z-Tranformation herangezogen:

$$\frac{\bar{X} - \mu_X}{\hat{\sigma}_X / \sqrt{n}} = \frac{\bar{X} - \mu_X}{\sqrt{\dfrac{\sum_{i=1}^{n} (X_i - \bar{X})}{n \cdot (n-1)}}} \sim t_{df=n-1}$$

ist die resultierende Größe auch bei normalverteilten Populationen nur asymptotisch standardnormalverteilt, was Folge der Ersetzung des Populationsmittelwerts durch den Stichprobenmittelwert bei der Schätzung der Varianz ist. Es kann allerdings gezeigt werden, dass der Quotient dann einer ***T-Verteilung*** mit ***df = n–1 Freiheitsgraden*** folgt.

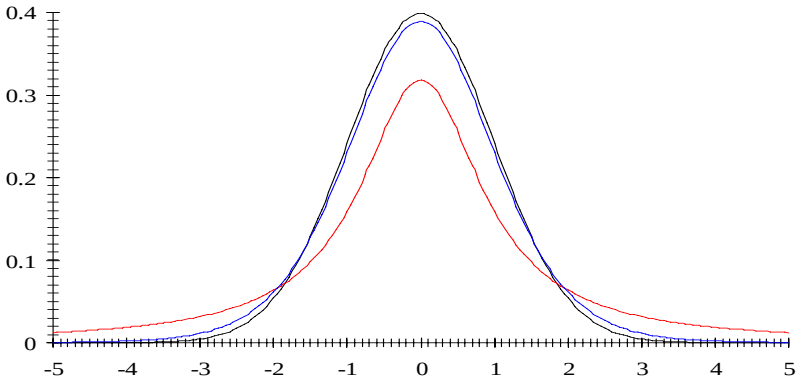
Die T-Verteilung



Die ***T-Verteilung*** ist eine symmetrische, unimodale Verteilung, die der Standardnormalverteilung sehr ähnlich ist, aber eine größere Varianz hat und insbesondere an den Enden größere Dichten aufweist.

Dies hat zur Folge, dass die Quantilwerte der T-Verteilung bei gleicher Quantilwahrscheinlichkeit weiter vom Median Null entfernt sind als die entsprechenden Quantilwerte der Standardnormalverteilung. Mit steigender Zahl der sogenannten Freiheitsgraden nähert sich die T-Verteilung asymptotisch der Standardnormalverteilung an, so dass $t_{df=\infty} = N(0;1)$.

Die T-Verteilung



Die Dichtefunktion der T-Verteilung ist eine sehr komplexe Funktion, die in Abhängigkeit von einem Parameter variiert, der als ***Freiheitsgrad*** *(engl:* ***degree of freedom***, **df**) bezeichnet wird.

Die Verteilung ist wie die Standardnormalverteilung unimodal symmetrisch um Null verteilt. Bei df > 1 ist auch der Erwartungswert Null, bei df=1 ist er dagegen nicht definiert. Ähnlich ist auch die Varianz erst ab df=3 Freiheitsgraden mit $\sigma^2(t_{df}) = df / (df–2)$ definiert.
Ähnlich wie die Standardnormalverteilung sind auch Quantile der T-Verteilung in Statistikbüchern tabelliert. Die wichtigsten Quantilanteile sind auf den nächsten beiden Folien wiedergegeben.
Ab etwa df > 30 unterscheiden sich die Quantilwerte der T-Vereilung kaum noch von denen der Standardnormalverteilung.

Quantile der T-Verteilung

df	75.0%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%	99.95%
1	1.000	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	0.816	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.765	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850

Quantile der T-Verteilung

df	75.0%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%	99.95%
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Aus der Tabelle ist ersichtlich, dass z.B. das 95%-Quantil der T-Verteilung mit 60 Freiheitsgraden den Quantilwert 1.671 aufweist.
Die unterste Zeile enthält die Quantile der Standardnormalverteilung, das ist gleichzeitig die T-Verteilung mit df=∞ Freiheitsgraden.
Da T-Verteilungen um 0 symmetrisch verteilt sind, können aus der Tabelle auch Quantile mit Wahrscheinlichkeiten < 50% abgelesen werden. So ist das 5%-Quantil der t-Verteilung mit df=60 minus eins mal dem 95%-Quantil (5% = 100% – 95%) und daher gleich -1.671.

c) Schätzer von Populationsanteilen

Populationsanteile können als Mittelwerte von 0/1-kodierten dichotomen Variablen aufgefasst werden. Daher gelten im Prinzip die Aussagen über die Kennwerteverteilung von Mittelwerten auch für Populationsanteile.
Eine Abweichung besteht darin, dass bei dichotomen 0/1-kodierten Variablen sowohl der Mittelwert wie die Varianz eine Funktion desselben Parameters p_1 bzw. π_1 sind. Dies hat die Konsequenz, dass bei der Schätzung der Varianz einer 0/1-kodierten Variable die Populationsvarianz direkt durch die Stichprobenvarianz geschätzt wird:

$$\hat{\sigma}_X^2 = s_X^2 = p_1 \cdot (1 - p_1) \text{ wenn } \sigma_X^2 = \pi_1 \cdot (1 - \pi_1)$$

In einfachen Zufallsstichproben ist der Stichprobenanteil p_1 ein konsistenter und erwartungstreuer Schätzer des Populationsanteils p_1. Für den (geschätzten) Standardfehler gilt:

$$\sigma(p_1) = \begin{cases} \sqrt{\dfrac{\pi_1 \cdot (1-\pi_1)}{n}} \\ \sqrt{\dfrac{\pi_1 \cdot (1-\pi_1)}{n} \cdot \dfrac{N-n}{N-1}} \end{cases} \text{ bzw. } \hat{\sigma}(p_1) = \begin{cases} \sqrt{\dfrac{p_1 \cdot (1-p_1)}{n}} & \text{mit Zurücklegen} \\ \sqrt{\dfrac{p_1 \cdot (1-p_1)}{n} \cdot \dfrac{N-n}{N-1}} & \text{ohne Zurücklegen} \end{cases}$$

Der geschätzte Standardfehler ist ab etwa n > 60 hinreichend genau.
Da das Maximum der Populationsvarianz 0.25 an der Stelle $\pi_1 = 1 - \pi_1 = 0.5$ beträgt, kann bei kleineren Stichproben auch einfach die maximale Populationsvarianz 0.25 im Zähler der Formel des Standardfehlers verwendet werden, was zu einer leichten Überschätzung des Standardfehlers führen kann.

Die Kennwerteverteilung von Stichprobenanateile ist bei einfachen Zufallsauswahlen asymptotisch normalverteilt (vgl. L12). Wenn die Annäherung bei kleinen Stichproben nicht hinreichend genau ist, sollte die exakte Binomialverteilung bzw. hypergeometrische Verteilung verwendet werden.

Punktschätzung und Intervallschätzung

Von ***Punktschätzung*** spricht man, wenn die Realisation eines Schätzers als konkrete Schätzung des unbekannten Wertes eines Populationsparameters verwendet wird.

Es ist allerdings sehr unwahrscheinlich, dass eine einzelne Schätzung exakt mit dem unbekannten Populationsparameter übereinstimmt.

> *Wenn z.B. bei einer Population von N=100 000 der Populationsanteil π_1=0.60 ist, beträgt die Wahrscheinlichkeit nur etwa 8%, dass bei einer Stichprobengröße von n=100 ein Stichprobenanteil ebenfalls p_1=0.6 (=60/100) ist.*

Umgekehrt ist dann in 92% aller Stichproben mit Abweichungen zu rechnen.
Die Wahrscheinlichkeit von 8% ergibt sich bei Anwendung der asymptotischen Normalverteilung auf die Kennwerteverteilung des Stichprobenanteils:

$$\Pr(p_1 = 0.6) \approx \Phi\left(\frac{60 + 0.5 - 0.6 \cdot 100}{\sqrt{100 \cdot 0.6 \cdot 0.4 \cdot \frac{100000 - 100}{100000 - 1}}}\right) - \Phi\left(\frac{60 - 0.5 - 0.6 \cdot 100}{\sqrt{100 \cdot 0.6 \cdot 0.4 \cdot \frac{100000 - 100}{100000 - 1}}}\right)$$
$$= \Phi(0.102) - \Phi(-0.102) \approx 0.08$$

Da der gesuchte Wert vermutlich nur in der Nähe der Schätzung liegt, ist es oft sinnvoller, statt eines exakten Wertes ein Intervall anzugeben, in dem der gesuchte Wert vermutlich liegt.
Statt von Punktschätzung spricht man dann von ***Intervallschätzung***.

Bei der Intervallschätzung werden sogenannte ***Konfidenzintervalle*** für den unbekannten Populationsparameter geschätzt.

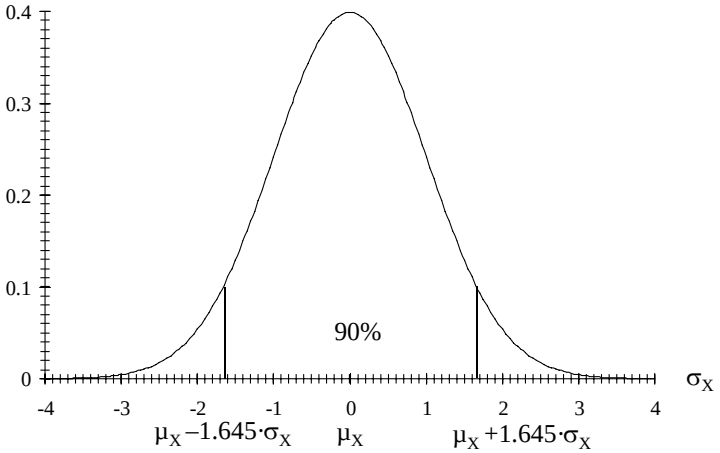
Die Logik von Konfidenzintervallen

DieVorgehensweise kann am Beispiel der Berechnung eines Konfidenzintervalls zur Schätzung eines Populationsmittelwertss verdeutlicht werden.

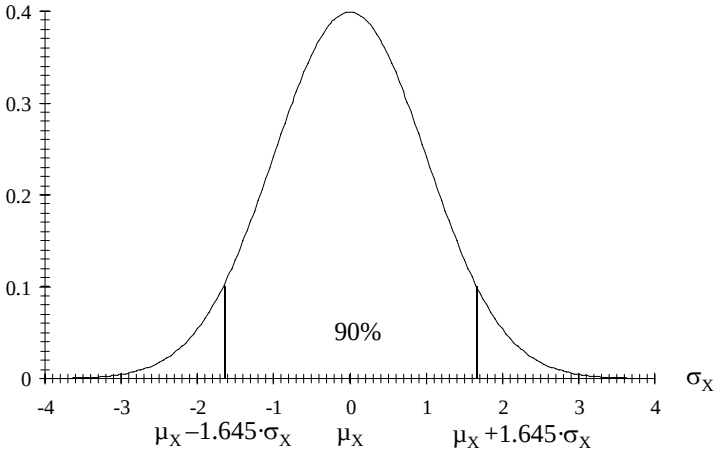
Bei einer einfachen Zufallsauswahl aus einer normalverteilten Population ist der Stichprobenmittelwert um den zu schätzenden Populationsmittelwert normalverteilt:

$$f\left(\bar{X}\right) = N\left(\mu_X; \frac{\sigma_X^2}{n}\right) \Rightarrow f\left(\frac{\bar{X} - \mu_X}{\sigma\left(\bar{X}\right)}\right) = \phi$$

Mit Hilfe der Standardnormalvereilung lässt sich dann ein Intervall berechnen, in dem die Stichprobenmittelwerte mit einer Wahrscheinlichkeit von z.B. 90% realisiert werden:



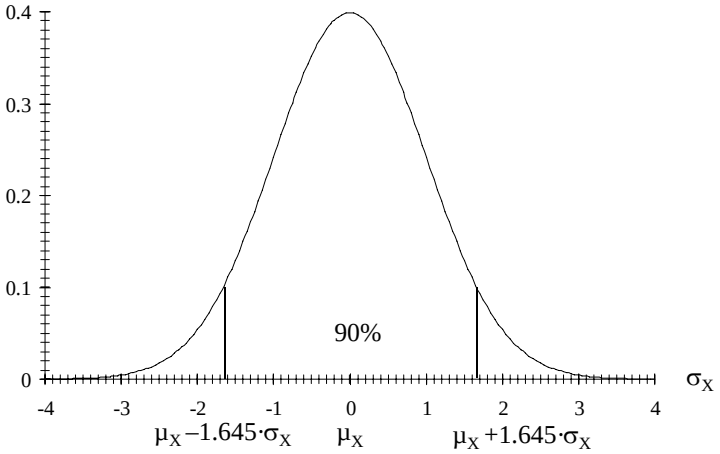
Die Logik von Konfidenzintervallen



90% alle Realisierungen einer Standardnormalverteilung liegen zwischen den 5%-Quantil = –1.645 und dem 95%-Quantil = +1.645:

$$\begin{aligned} 0.9 &= 0.95 - 0.05 \\ &= \Phi(1.645) - \Phi(-1.645) \\ &= \Pr(-1.645 \le Z \le 1.645) \\ &= \Pr\left(-1.645 \le \frac{\bar{X} - \mu_X}{\sigma(\bar{X})} \le 1.645\right) \end{aligned}$$

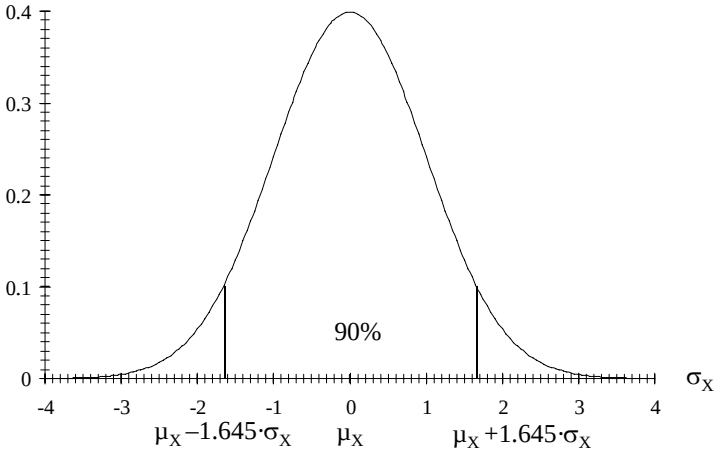
Die Logik von Konfidenzintervallen



Durch Umformen ergeben sich Intervallgrenzen eines Intervalls um den Stichprobenmittelwert:

$$\begin{aligned} 0.90 &= \Pr\left(-1.645 \le \left(\bar{X} - \mu_X\right) / \sigma(\bar{X}) \le 1.645\right) \\ &= \Pr\left(-1.645 \cdot \sigma(\bar{X}) \le \bar{X} - \mu_X \le 1.645 \cdot \sigma(\bar{X})\right) \\ &= \Pr\left(-\bar{X} - 1.645 \cdot \sigma(\bar{X}) \le -\mu_X \le -\bar{X} + 1.645 \cdot \sigma(\bar{X})\right) \\ &= \Pr\left(\bar{X} + 1.645 \cdot \sigma(\bar{X}) \ge \mu_X \ge \bar{X} - 1.645 \cdot \sigma(\bar{X})\right) \\ &= \Pr\left(\bar{X} - 1.645 \cdot \sigma(\bar{X}) \le \mu_X \le \bar{X} + 1.645 \cdot \sigma(\bar{X})\right) \end{aligned}$$

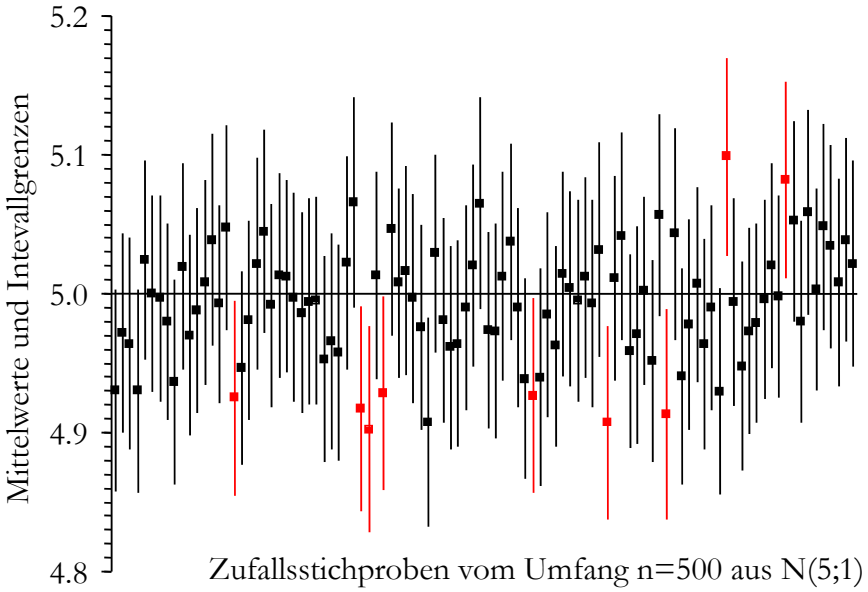
Die Logik von Konfidenzintervallen



$$\left(\bar{X} - 1.645 \cdot \sigma\left(\bar{X}\right) \leq \mu_X \leq \bar{X} + 1.645 \cdot \sigma\left(\bar{X}\right)\right) = 90\%$$

Bei gegebenen Stichprobenmittelwert und bekanntem bzw. geschätzten Standardfehler lässt sich also ein Intervall berechnen, das mit einer Wahrscheinlichkeit von 90% den zu schätzenden Parameter enthält. Ein so berechnetes Intervall, dass mit einer bestimmten Wahrscheinlichkeit zu beobachten ist, wird als ***Konfidenzintervall*** bezeichnet.

Interpretation von Konfidenzintervallen



Für die Interpretation ist es wichtig, sich klar zu machen, das das Konfidenzintervall selbst eine Zufallsvariable ist.

Die Abbildung zeigt so 90%-Konfidenintervalle um die Stichprobenmittelwerte von 100 Stichproben des Umfangs n=500 aus einer normalverteilten Population mit dem Populationsmittelwert 5 und einer Varianz von 1.
Von den 100 Intervallen enthalten 91 den Populationswert, neun dagegen nicht.

Interpretation von Konfidenzintervallen

Da das Konfidenzintervall die Zufallsvariable ist und der zu schätzenden Populationsmittelwert eine konstante empirische Größe, wäre es somit falsch zu behaupten, dass der Populationsmittelwert mit der Wahrscheinlichkeit von – im Beispiel – 90% im Konfidenzintervall liegt.

> *Die Wahrscheinlichkeitsaussage bezieht sich also nicht auf den unbekannten Parameter, sondern auf die Zufallsvariable „Konfidenzintervall“.*

Mit einer Wahrscheinlichkeit von 90% enthält das Intervall den Populationsmittelwert. Ob der Populationsmittelwert aber tatsächlich im Intervall liegt oder nicht, bleibt unbekannt.
Da die Wahrscheinlichkeit, dass das Intervall den Populationswert enthält, aber mit 90% recht hoch ist, kann begründet angenommen werden, dass es vermutlich den Populationsmittelwert enthält. Es ist also vernünftig anzunehmen, dass der gesuchte Populationswert innerhalb der Grenzen des Konfidenzintervalls liegt.

Die Wahrscheinlichkeit, dass das Konfidenzintervall den zu schätzenden Parameter enthält, ist die sogenannte ***Vertrauenswahrscheinlichkeit***.
Im Beispiel beträgt die Vertrauenswahrscheinlichkeit 90%.

Umgekehrt ist die Wahrscheinlichkeit 1 minus der Vertrauenswahrscheinlichkeit, dass ein Intervall den Populationswert nicht enthält. Diese Wahrscheinlichkeit wird als ***Irrtumswahrscheinlichkeit*** bezeichnet und durch den kleinen griechischen Buchstaben α (alpha) gekennzeichnet.
Im Beispiel beträgt die Irrtumswahrscheinlichkeit 10%.

Vorgehensweise bei Intervallschätzung

Anhand des Beispiel lässt sich die generelle Vorgehensweise bei der Berechnung von Konfidenzintervallen verdeutlichen:

Schritt 1: Auswahl einer geeigneten Stichprobenstatistik
Im ersten Schritt ist eine Stichprobenstatistik auszuwählen, für dessen Kennwerteverteilung gilt, dass der zu schätzende Populationsparameter Erwartungswert bzw. Median der Kennwerteverteilung ist.

Schritt 2: Festlegung der Irrtumswahrscheinlichkeit
Im zweiten Schritt wird die Vertrauenswahrscheinlichkeit bzw. tatsächlich i.a. umgekehrt die Irrtumswahrscheinlichkeit festgelegt.

In der Sozialforschung werden in der Regel Irrtumswahrscheinlichkeiten von 5% oder von 1% akzeptiert und entsprechend 95%- oder 99%-Konfidenzintervalle berechnet.

Im Prinzip ist diese Festlegung rein willkürlich. Zwar sinkt mir der Irrtumswahrscheinlichkeit die Chance von Fehlern. Allerdings steigen gleichzeitig die Längen von Konfidenzintervallen Wenn ein Konfidenzintervall aber sehr lang ist, hat es kaum Aussagekraft.

Schritt 3: Berechnung der Intervallgrenzen
Nach der Festlegung der Irrtumswahrscheinlichkeit α kann das Intervall berechnet werden. Dazu werden Quantile der Kennwerteverteilung benötigt. Die Intervallgrenzen berechnen sich üblicherweise nach:

$$\text{Untergrenze} = \text{Punktschätzung} - (1 - \alpha/2)\text{-Quantil der Kennwerteverteilung}$$
$$\text{Obergrenze} = \text{Punktschätzung} + (1 - \alpha/2)\text{-Quantil der Kennwerteverteilung}$$

Konfidenzintervalle für Populationsanteile

Es liegt nahe, zur Schätzung von Populationsanteilen Stichprobenanteile zu verwenden, da diese in einfachen Zufallsauswahlen erwartungstreue Schätzer der Populationsanteile sind.
Als Kennwerteverteilung wird die asymptotisch gültige Normalverteilung verwendet. Die Berechnung der Intervallgrenzen erfolgt dann über die Umkehrung der Z-Transformation der Standardnormalverteilung.

Bei einer einfachen Zufallsauswahl aus einer relativ zum Stichprobenumfang großen Population berechnet sich das (1–α)-Konfidenzintervall eines Populationsanteils dann nach:

$$\text{c.i.}(\pi_1) = p_1 \pm \sqrt{\frac{p_1 \cdot (1-p_1)}{n}} \cdot z_{\alpha/2} = p_1 \pm \sqrt{\frac{p_1 \cdot (1-p_1)}{n}} \cdot z_{1-\alpha/2}$$

$z_{\alpha/2}$ bzw. $z_{1-\alpha/2}$ sind die α/2- und (1–α/2)-Quantile der Standardnormalerteilung.

Bei einer einfachen Zufallsauswahl ohne Zurücklegen aus einer relativ kleinen Population (N/n ≤ 20) berechnet sich das Konfidenzintervall nach:

$$\text{c.i.}(\pi_1) = p_1 \pm \sqrt{\frac{p_1 \cdot (1-p_1)}{n} \cdot \frac{N-n}{N-1}} \cdot z_{\alpha/2} = p_1 \pm \sqrt{\frac{p_1 \cdot (1-p_1)}{n} \cdot \frac{N-n}{N-1}} \cdot z_{1-\alpha/2}$$

Die Schätzung des Standardfehlers der Kennwerteverteilung über den Stichprobenanteil setzt hinreichend große Stichproben (n > 60) voraus.

Konfidenzintervalle für Populationsanteile

Ist diese Bedingung nicht gegeben, kann ein maximales Konfidenzintervall berechnet werden:

$$\text{c.i.}(\pi_1) = p_1 \pm \frac{0.5 \cdot z_{\alpha/2}}{\sqrt{n}} = p_1 \pm \frac{0.5 \cdot z_{1-\alpha/2}}{\sqrt{n}} \text{ bzw. } \text{c.i.}(\pi_1) = p_1 \pm \frac{0.5 \cdot z_{\alpha/2}}{\sqrt{n \cdot \frac{N-n}{N-1}}} = p_1 \pm \frac{0.5 \cdot z_{1-\alpha/2}}{\sqrt{n \cdot \frac{N-n}{N-1}}}$$

Das maximale Konfidenzintervall ist konservativ, da der Standardfehler tendenziell überschätzt wird, und das Intervall entsprechend vermutlich ein wenig zu lang ist.

Bei sehr kleinen Stichprobenfallzahlen ist möglicherweise die asymptotische Annäherung an die Normalverteilung nicht gegeben. Dann können als Alternative auch das α/2- und das (1–α/2)-Quantile der Binomialverteilung bzw. der hypergeometrischen Verteilung berechnet werden, wobei im Sinne eines konservativen Vorgehens die Quantile für einen Populationsanteil von $\pi_1 = N_1/N = 0.5$ berechnet werden.

Als Beispiel soll ein 95%-Konfidenzintervall für einen Populationsanteil berechnet werden, wenn in einer einfachen Zufallsauswahl ohne Zurücklegen von n=100 Fällen aus N = 100 000 Fällen ein Stichprobenanteil von $p_1 = 0.6$ (=60/100) resultiert.
Da N/n = 1000 > 20 und $z_{1-\alpha/2} = z_{0.975} = 1.96$ berechnet sich die Intervallgrenzen nach:

$$\text{c.i.}(\pi_1) = 0.6 \pm \sqrt{\frac{0.6 \cdot 0.4}{100}} \cdot 1.96 = 0.6 \pm 0.096 = [0.504, 0,696]$$

Konfidenzintervalle für Populationsanteile

Bei einer Irrtumswahrscheinlichkeit von 5% ist damit zu rechnen, dass der Populationsanteil vermutlich zwischen 0.50 und 0.70 liegt.
Das Konfidenzntervall ist relativ lang. Wird daher die Irrtumswahrscheinlichkeit auf 10% heraufgesetzt, beträgt $z_{1-\alpha/2} = z_{0.950} = 1.645$. Die Grenzen des 90%-Konfidenzintervalls betragen dann:

$$\text{c.i.}(\pi_1) = 0.6 \pm \sqrt{\frac{0.6 \cdot 0.4}{100}} \cdot 1.645 = 0.6 \pm 0.081 = \left[0.52, 0, 68\right]$$

Wird berücksichtigt, dass die Auswahl ohne Zurücklegen erfolgte, ändern sch die Intervallgrenzen bei drei Nachkommastellen nicht:

$$\text{c.i.}(\pi_1) = 0.6 \pm \sqrt{\frac{0.6 \cdot 0.4}{100} \cdot \frac{100000 - 100}{100000 - 1}} \cdot 1.645 = 0.6 \pm 0.081 = \left[0.52, 0, 68\right]$$

Konfidenzintervalle für Populationsmittelwerte

Die Berechnung von Konfidenzintervallen für Mittelwerte erfolgt nach der gleichen Logik. Unabhängig von der Verteilung in der Population kann bei einer Fallzahl ab n=30 die Annäherung an die Normalverteilung genutzt werden.

Das (1–α)-Konfidenzintervall berechnet sich dann nach:

$$\text{c.i.}(\mu_X) = \bar{x} \pm \hat{\sigma}\left(\bar{x}\right) \cdot z_{1-\alpha/2} = \bar{x} \pm \sqrt{\frac{SS_X}{n \cdot \left(n-1\right)}} \cdot z_{1-\alpha/2} = \bar{x} \pm \frac{\hat{\sigma}_X}{\sqrt{n}} \cdot z_{1-\alpha/2} = \bar{x} \pm \frac{s_X}{\sqrt{n-1}} \cdot z_{1-\alpha/2}$$

Bei einer einfachen Zufallsauswahl ohne Zurücklegen aus einer relativ kleinen Population (N/n ≤ 20) wird der Korrekturfaktor (N–n)/(N–1) verwendet. Das Konfidenzintervall ist nach:

$$\text{c.i.}(\mu_X) = \bar{x} \pm \sqrt{\frac{SS_X}{n \cdot \left(n-1\right)} \cdot \frac{N-n}{N-1}} \cdot z_{1-\alpha/2} = \bar{x} \pm \frac{\hat{\sigma}_X}{\sqrt{n \cdot \frac{N-n}{N-1}}} \cdot z_{1-\alpha/2} = \bar{x} \pm \frac{s_X}{\sqrt{\left(n-1\right) \cdot \frac{N-n}{N-1}}} \cdot z_{1-\alpha/2}$$

Wenn die interessierende Größe X in der Population normalverteilt ist, wird statt der Standardnormalverteilung die T-Verteilung mit df = n – 1 Freiheitsgraden verwendet. Das Konfidenzintervall ergibt sich dann nach:

$$\text{c.i.}(\mu_X) = \bar{x} \pm \sqrt{\frac{SS_X}{n \cdot \left(n-1\right)}} \cdot t_{1-\alpha/2;df=n-1} = \bar{x} \pm \frac{\hat{\sigma}_X}{\sqrt{n}} \cdot t_{1-\alpha/2;df=n-1} = \bar{x} \pm \frac{s_X}{\sqrt{n-1}} \cdot t_{1-\alpha/2;df=n-1}$$

Konfidenzintervalle für Populationsmittelwerte

Da Konfidenzintervalle, die über die T-Verteilung berechnet werden, länger sind als Konfidenzintervalle mit gleicher Irrtumswahrscheinlichkeit, die auf der Standardnormalverteilung beruhen, wird oft auch dann die T-Verteilung verwendet, wenn die Verteilung von X in der Population unbekannt oder nicht normalverteilt ist.
Es besteht dann eine größere Chance, dass die Konfidenzintervalle den zu schätzenden Populationsmittelwert tatsächlich enthalten. Man bezeichnet dieses vorsichtigere Vorgehen in der Statistik als ***konservatives Schätzen***.

Das ***konservative Schätzen*** wird insbesondere auch dann verwendet, wenn die Stichprobe sehr klein (n < 30) und die asymptotische Annäherung an die Normalverteilung nicht garantiert ist.

Als Beispiel soll ein 95%-Konfidenzintervall für das Haushaltseinkommen berechnet werden, wenn in einer einfachen Zufallsauswahl von n=1024 Haushalten ein mittleres Einkommen von 3500€ beobachtet wurde und die geschätzte Populationsstandardabweichung 1500€ beträgt.
Das Konfidenzintervall berechnet sich dann nach:

$$\text{c.i.}(\mu_X) = \bar{x} \pm \frac{\hat{\sigma}_X}{\sqrt{n}} \cdot z_{1-\alpha/2} = 3500 \pm \frac{1500}{\sqrt{1024}} \cdot 1.96 = 3500 \pm 91.875 = [3408, 3592]$$

Das mittlere Einkommen liegt vermutlich zwischen 3400 und 3600 €.

Konfidenzintervalle für Populationsmittelwerte aus normalverteilten Populationen

Wird anstelle der standardnormalverteilung die T-Verteilung herangezogen, wird das 97.5%-Quantil der T-Verteilung mit df=1023 Freiheitsgraden benötigt.
Aus der vorne wiedergegebenen Tabelle ist zu entnehmen, dass der Wert bei df=120 bei 1.98 liegt und bei df=∞ bei dem verwendeten Wert 1.96. Der exakte Wert liegt irgendwo dazwischen.
Wird im Sinne eines konservativen Schätzens der Wert 1.98 verwendet, vergrößern sich das Intervall von 3500 ± 91.875 auf 3500 ± 92.8125.

Nutzung von Konfidenzintervallen zur Berechnung der Fallzahl

Mit Hilfe von Konfidenzintervallen kann auch die notwendige Fallzahl für eine Untersuchung bestimmt werden. Wenn bei einer Irrtumswahrscheinlichkeit α eine Genauigkeit von ε verlangt wird, wobei ε die halbe Länge des Konfidenzintervalls ist, dann folgt aus der Rechenformel für das Konfidenzintervall bei der Betrachtung von Populationsanteilen:

$$\varepsilon = z_{1-\alpha/2} \cdot \sqrt{\frac{\pi_1 \cdot (1-\pi_1)}{n}}$$

Durch Auflösen der Gleichung ergibt sich die notwendige Fallzahl:

$$n = \frac{\left(z_{1-\alpha/2}\right)^2 \cdot \pi_1 \cdot (1-\pi_1)}{\varepsilon^2}$$

Wenn α=5% und eine Genauigkeit von ε = ±3% verlangt wird, und von einem Populationsanteil von π_1 = 0.5 ausgegangen wird, dann benötigt man eine Fallzahl von:

$$n = \frac{\left(z_{1-\alpha/2}\right)^2 \cdot \pi_1 \cdot (1-\pi_1)}{\varepsilon^2} = \frac{1.96^2 \cdot 0.5^2}{0.03^2} = 1067.111 \approx 1068$$

Bei Mittelwerten muss eine ungefähre Schätzung der Populationsvarianz vorliegen. Die Fallzahl brechnet sich hier nach:

$$n = \frac{\left(z_{1-\alpha/2}\right)^2 \cdot \hat{\sigma}_X^2}{\varepsilon^2}$$

Lerneinheit 14: Die Logik statistischen Testens

In vielen sozialwissenschaftlichen Fragestellungen sollen Vermutungen über Eigenschaften einer Population überprüft werden.

Es soll z.B. geprüft werden, ob in einer Stadt eine Mehrheit der Bürger für die Einrichtung einer Ganztagsschule ist. In einer einfachen Zufallsauswahl von n=100 Bürgern sprechen sich 60% für die Einrichtung der Schule aus. Aus dem Ergebnis wird geschlossen, dass es tatsächlich eine Mehrheit für die Einrichtung der Ganztagsschule gibt.

Das Beispiel weist auf die Ähnlichkeit der Fragestellung beim statistischen Schätzen und beim statistischen Testen hin:

- Beim Schätzen wird aufgrund von Stichprobendaten in einem Induktionsschluss auf eine Eigenschaft der Population geschlossen;
- beim Testen wird anhand von Stichprobendaten induktiv entschieden, ob eine Vermutung über eine Eigenschaft der Population zutrifft oder nicht zutrifft.

Beim statistischen Testen wird also immer eine Entscheidung über die empirische Gültigkeit einer vermuteten (postulierten) Populationseigenschaft getroffen.
Als Entscheidungsgrundlage werden Informationen aus einer Zufallsstichprobe verwendet.
⇒ **Statistischer Test sind Entscheidungsregeln, die Stichprobendaten nutzen.**

Statistische Test über Konfidenzintervalle

Eine naheliegende und nicht selten auch genutzte Möglichkeit zur Prüfung einer Hypothese besteht darin, ein (1–α)-Konfidenzintervall zu berechnen.

Wenn ein gegen die zu prüfende Hypothese sprechender Wert innerhalb des Konfidenzintervalls liegt, ist die Hypothese mit der Irrtumswahrscheinlichkeit α abzulehnen.

Im Beispiel der Fragestellung, ob es eine Mehrheit für die Einrichtung der Ganztagsschule gibt, berechnen sich die Grenzen des 95%-Konfidenzintervalls bei einer Stichprobenfallzahl von n=100 und n_1=60 Befürwortern der Ganztagsschule nach:

$$\text{c.i.}(p_1) = \frac{n_1}{n} \pm \sqrt{\frac{n_1}{n} \cdot \frac{n-n_1}{n}} \cdot z_{1-\alpha/2} = \frac{n_1}{n} \pm \sqrt{\frac{n_1 \cdot (n-n_1)}{n^3}} \cdot z_{1-\alpha/2}$$

$$= \frac{60}{100} \pm \sqrt{\frac{60 \cdot 40}{100^3}} \cdot 1.96 = 0.6 \pm 0.096 = [0.504, 0.696]$$

Da die Untergrenze des Konfidenzintervalls über 50% liegt, liegt kein Wert, der gegen die Hypothese spricht, im Konfidenzintervall. Daher kann bei einer Irrtumswahrscheinlichkeit von 5% ausgeschlossen werden, dass es keine Mehrheit für die Einrichtung der Schule gibt.
Dann gibt es umgekehrt vermutlich eine Mehrheit für die Einrichtung der Schule.

Statistische Test über Konfidenzintervalle

Beim Testen über Konfidenzintervalle werden keine Informationen über Populationsparameter verwendet, die die zu prüfenden Hypothese enthält.
Es ist jedoch oft möglich, zusätzliche Annahmen über die Population zu nutzen, die zutreffen müssen, wenn die Hypothese richtig sein sollte.

Im Beispiel der Einrichtung einer Ganztagsschule ist die Vermutung, dass es eine Mehrheit für die Schule gibt, falsch, wenn in der Population höchstens ein Anteil π_1=0.5 für die Einrichtung der Schule ist.
Wenn dies gerade noch zutrifft, dann ist es nicht nötig, den Standardfehler aus den Stichprobendaten zu schätzen. Sein Wert folgt dann nämlich direkt aus der Stichprobengröße und der Annahme der Hypothese, dass ein Anteil von π_1=0.5 für die Einrichtung der Schule ist. Der Standardfehler beträgt dann nämlich unter dieser Bedingung:

$$\sigma(p_1) = \sqrt{\frac{\pi_1 \cdot (1-\pi_1)}{n}} = \sqrt{\frac{0.5 \cdot (1-0.5)}{100}} = 0.05$$

Falls in der Population 50% für die Einrichtung der Schule sind, dann ist zudem der Erwartungswert des Stichprobenanteils $\mu(p_1)$=0.5.
Zusammen mit dem Standardfehler kann diese Information genutzt werden, um über eine Z-Transformation des Stichprobenanteils eine Statistik zu berechnen, die asymptotisch standardnormalverteilt ist, wenn in der Population tatsächlich genau 50% für die Einrichtung der Schule sind.

Signifikanztest

$$Z = \frac{p_1 - \mu(p_1)}{\sigma(p_1)} = \frac{p_1 - \pi_1}{\sqrt{\frac{\pi_1 \cdot (1-\pi_1)}{n}}} = \frac{p_1 - 0.5}{0.05}$$

Falls es doch eine Mehrheit für die Einrichtung der Schule gibt, dann ist der Populationsanteil $\pi_1 > 0.5$.
Da dann der Abzug von 0.5 im Zähler der Teststatistik zu gering ist, ist dann eher mit positiven Werten zu rechen.
Daher wird die Hypothese, dass es keine Mehrheit gibt, bei einer Irrtumswahrscheinlichkeit von 5% abgelehnt, wenn die Teststatistik größer ist als das 95%-Quantil der Standardnormalverteilung, also größer ist als 1.645. Aus den Stichprobendaten p_1= 0.6 folgt:

$$Z = \frac{p_1 - 0.5}{0.05} = \frac{0.6 - 0.5}{0.05} = 2.0$$

Da 2.0 > 1.645, wird die Hypothese, dass es keine Mehrheit gibt, mit einer Irrtumswahrscheinlichkeit von 5% abgelehnt.

Die Idee dieser Art der Hypothesenprüfung geht auf den britischen Statistiker ***Fisher*** zurück und wird als ***Signifikanztest*** bezeichnet.

Signifikanztest

Vor der Durchführung eines ***Signifikanztests*** wird zwischen der inhaltlich interessierenden ***Forschungshypothese*** und ihrem Gegenteil, der sogenannten ***Nullhypothese H_0*** unterschieden.

Im Beispiel besagt die Forschungshypothese, dass es eine Mehrheit für die Einrichtung der Gesamtschule gibt, während die Nullhypothese H_0 behauptet, dass es keine Mehrheit gibt, also maximal 50% der Bevölkerung für die Einrichtung sind:

$$H_0: \pi_1 \leq 0.5$$

Der Signifikanztest prüft dann anhand der Stichprobendaten und unter Ausnutzung aller Informationen, die in der Nullhypothese enthalten sind, ob die Nullhypothese bei einer vorgegebenen Irrtumswahrscheinlichkeit noch mit der empirischen Wirklichkeit vereinbar ist.

Ist dies nicht der Fall, gilt die Nullhypothese als widerlegt und entsprechend ihr Gegenteil, also die Forschungshypothese als bestätigt.
Man spricht dann davon, dass das Ergebnis mit der ***Irrtumswahrscheinlichkeit α*** (auch als ***Signifikanzniveau*** bezeichnet) ***signifikant*** ist.

Wenn die Nullypothese dagegen mit den empirischen Daten vereinbar ist und entsprechend nicht ausgeschlossen werden kann, ist das Ergebnis nicht signifikant und die Forschungshypothese kann nicht bestätigt werden.

Alpha- und Betafehler (Fehler erster und zweiter Art)

Die eigentlich interessierende Forschungshypothese wird also nur dann bestätigt, wenn mit großer Sicherheit ausgeschlossen werden kann, dass die Nullhypothese falsch ist. Je kleiner die Irrtumswahrscheinlichkeit α, desto größer die Sicherheit $1-\alpha$, dass die Forschungshypothese zutrifft, wenn die Nullhypothese abgelehnt wird.

Man könnte nun auf die Idee kommen, die Irrtumswahrscheinlichkeit möglichst klein zu halten. Dabei wird jedoch übersehen, dass es auch die Möglichkeit gibt, dass die Nullhypothese beibehalten wird, obwohl sie falsch ist.
Insgesamt sind nämlich vier Situationen zu unterscheiden:

	H_0 ist richtig	H_0 ist falsch
Akzeptanz von H_0	richtige Entscheidung	falsche Entscheidung = β-Fehler (Fehler zweiter Art)
Verwerfen von H_0	falsche Entscheidung = α-Fehler (Fehler erster Art)	richtige Entscheidung

Neben dem Alphafehler, die Nullhypothese fälschlicherweise zu verwerfen, ist auch der Betafehler, das ist die fehlerhafte Beibehaltung der Nullhypothese, zu berücksichtigen.

Alpha- und Betafehler (Fehler erster und zweiter Art)

	H_0 ist richtig	H_0 ist falsch
Akzeptanz von H_0	richtige Entscheidung	falsche Entscheidung = β-Fehler (Fehler zweiter Art)
Verwerfen von H_0	falsche Entscheidung = α-Fehler (Fehler erster Art)	richtige Entscheidung

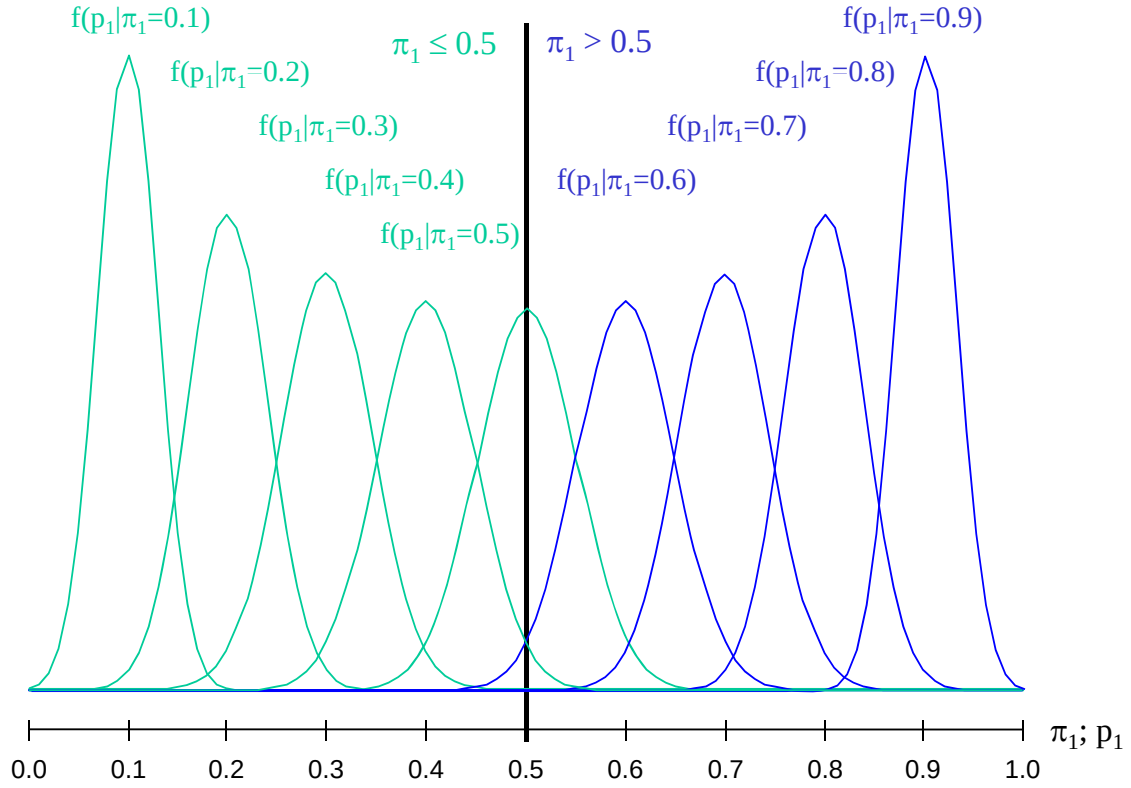
Statt von Alpha- und Betafehlern spricht man auch von Fehler erster und zweiter Art.

Der Signifikanztest, wie er bisher vorgestellt wurde, berücksichtigt nicht die Möglichkeit des Fehlers zweiter Art.
Anders ist es beim Hypothesentesten nach ***Neyman*** und ***Pearson***.
Die Grundidee des ***Neyman-Pearson-Tests*** besteht darin, zunächst symmetrisch zu denken und den gesamten Wertebereich der interessierenden Populationseigenschaft in zwei disjunkte Bereiche aufzuteilen, den Bereich der ***Nullhypothese H_0*** und den Bereich der ***Alternativhypothese H_1***.
Im Beispiel der Einrichtung der Schule würde etwa folgendes Hypothesenpaar formuliert:

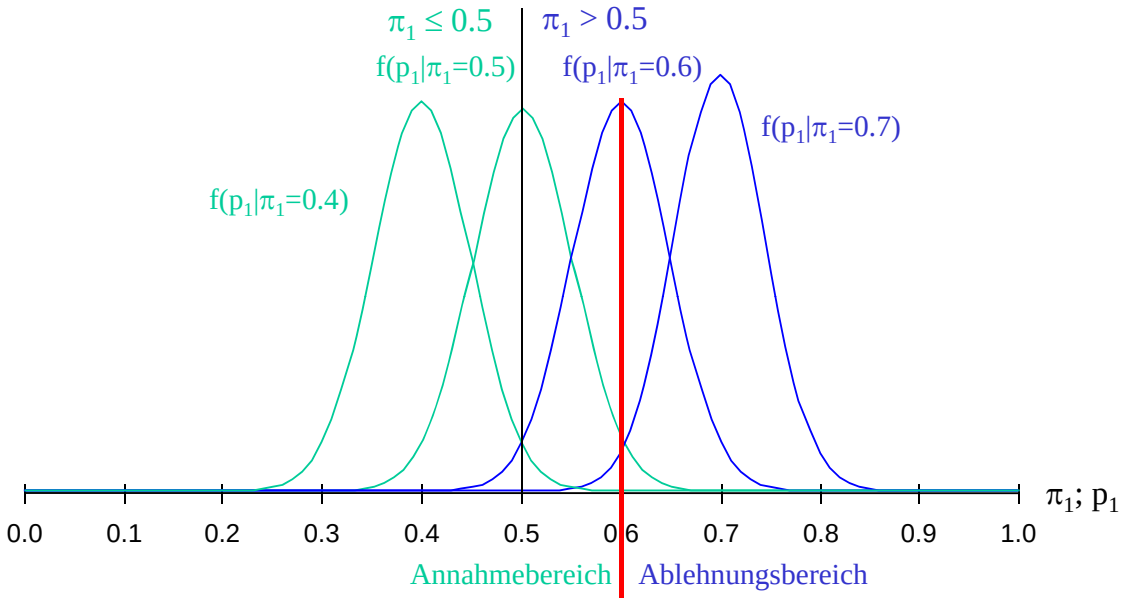
$$H_0: \pi_1 \leq 0.5 \text{ versus } H_1: \pi_1 > 0.5.$$

Verteilungen der Stichprobenstatistik bei Null- und Alternativhypothese

In Abhängigkeit von der Gültigkeit der Null- bzw. der Alternativhypothese unterscheiden sich die Stichprobenverteilungen des Stichprobenanteils.



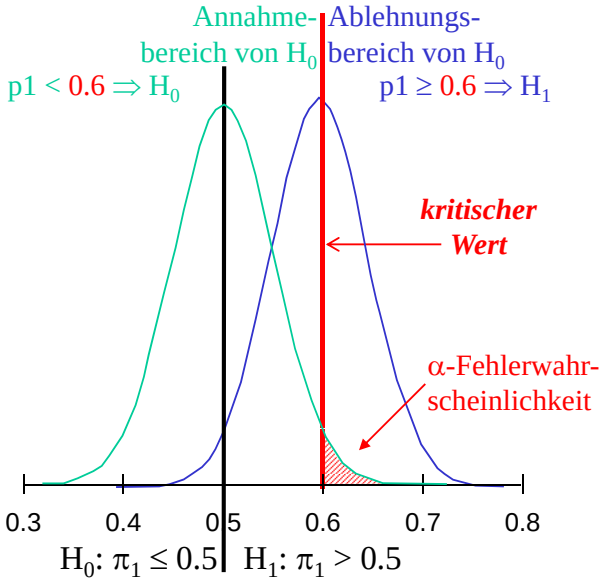
Annahmebereich und Ablehnungsbereich der Nullhypothese



Da die Kennwerteverteilung des Stichprobenanteils in Abhängigkeit von der Gültigkeit der Null- bzw. Alternativhypothese variiert, kann der **Stichprobenanteil als *Teststatistik*** verwendet werden.

Analog zur Zweiteilung des Wertebereichs des zu testenden Populationsparamters wird auch der Wertebreich des Teststatistik in zwei Bereiche geteilt: den ***Annahmebereich*** und den ***Ablehungsbereich*** der Nullhypothese.

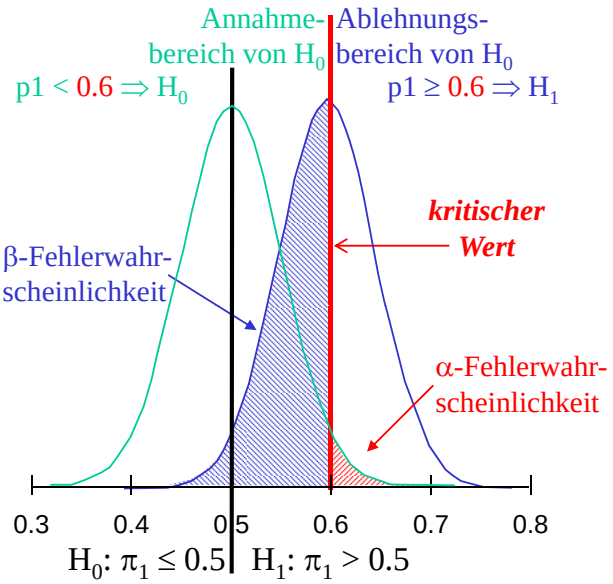
Festlegung der (maximalen) Irrtumswahrscheinlichkeit



Der Wert, der den Annahmebereich vom Ablehnungsbereich trennt, wird als ***kritischer Wert*** bezeichnet

In der Abbildung wird die Nullhypothese abgelehnt, wenn die Teststatistik den kritischen Wert 0.6 erreicht oder überschreitet.

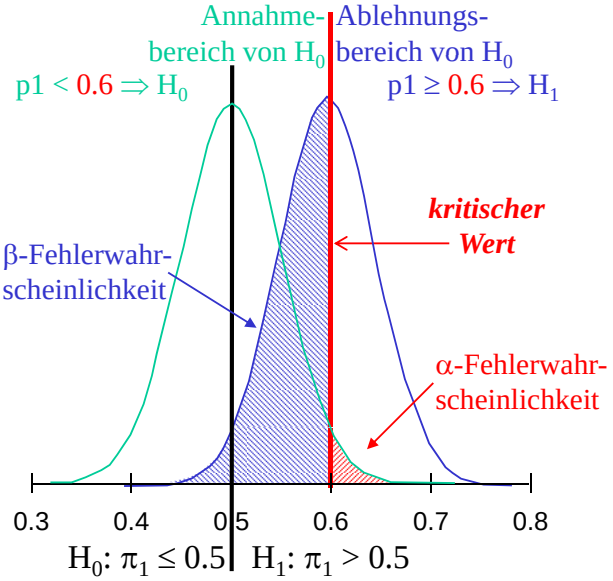
Festlegung der (maximalen) Irrtumswahrscheinlichkeit



Die α-Fehlerwahrscheinlichkeit ist dann die Wahrscheinlichkeit, dass die Teststatistik in den Ablehnungsbereich fällt, wenn die Nullhypothese richtig ist.

Umgekehrt ist die β-Fehlerwahrscheinlichkeit die Wahrscheinlichkeit, dass die Teststatistik in den Annahmebereich fällt, obwohl die Nullhypothese falsch ist.

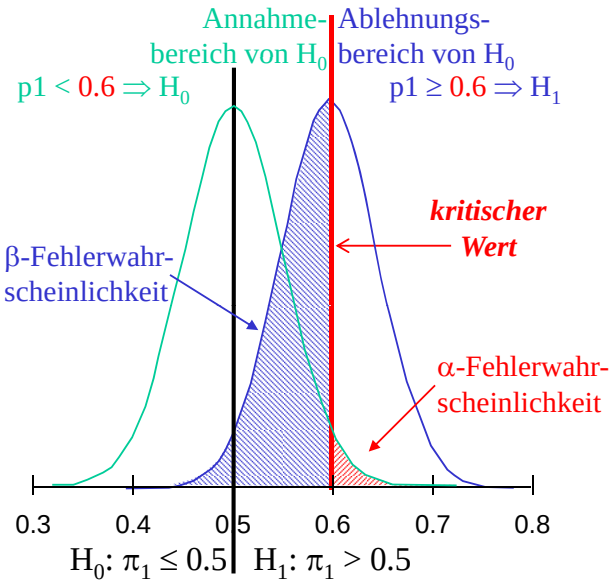
Festlegung der (maximalen) Irrtumswahrscheinlichkeit



Im Beispiel beträgt die maximale α-Fehlerwahrscheinlichkeit die Wahrscheinlichkeit, dass dass ein Stichprobenanteil größer oder gleich 0.6 ist, wenn der Populationsanteil 0.5 ist. Da der Stichprobenanteil asymptotisch normalverteilt ist, berechnet sich diese Wahrscheinlichkeit nach:

$$\Pr\left(p_1 \geq 0.6 \middle| \pi_1 = 0.5; n = 100\right) = 1 - \Phi\left((0.6 - 0.5)/0.05\right) = 1 - \Phi\left(2.0\right) = \Phi\left(-2.0\right) = 0.0228$$

Festlegung der (maximalen) Irrtumswahrscheinlichkeit

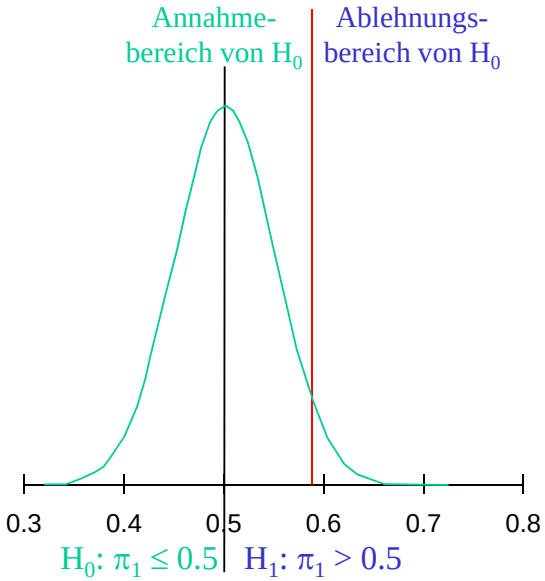


Aus den Zweiteilungen des Wertebereichs des zu testenden Populationsparameters in den Bereich der Null- und der Alternativhypothese und des Wertebereichs der Teststatistik in den Annahme- und den Ablehnungsbereich ist die maximale α-Fehlerwahrscheinlichkeit gleichzeitig 1.0 minus der maximalen β-Fehlerwahrscheinlichkeit.

*Im Beispiel beträgt die maximale **β-Fehlerwahrscheinlichkeit daher 1 – 0.0228 = 0.9772**, wenn der Populationsanteil unmerklich über 0.5 liegt.*

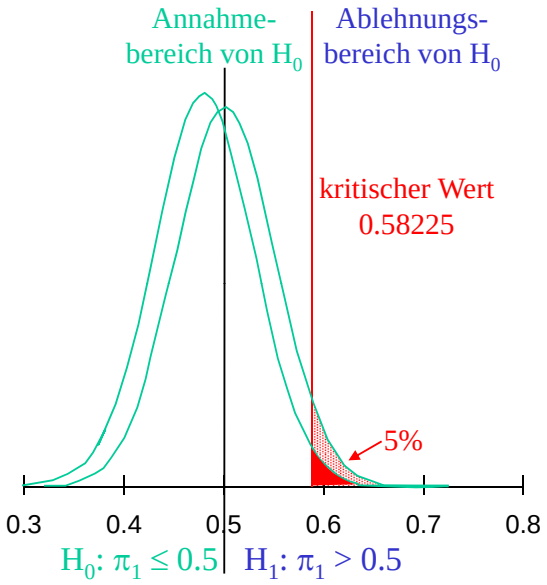
Die Symmetrie zwischen Null- und Alternativ-hypothese wird verlassen, wenn eine der beiden Hypothesen die eigentliche Forschungshypothese ist, also die Vermutung des Forscher beinhaltet.
In der Regel ist die Forschungshypothese die Alternativhypothese.

Festlegung der (maximalen) Irrtumswahrscheinlichkeit



Im Sinne eines möglichst ***strengen Testens*** soll die Wahrscheinlichkeit der fälschlichen Akzeptanz der Forschungshypothese einen Maximalwert nicht überschreiten. In der Sozialforschung wird üblicherweise die maximale α-Fehlerwahrscheinlichkeit auf 5% oder 1% festgelegt.
Diese Forderung kann realisiert werden, wenn die Forschungshypothese die Alternativhypothese ist, da der α-Fehler dann die falsche Annahme der Forschungshypothese beinhaltet.

Festlegung der (maximalen) Irrtumswahrscheinlichkeit



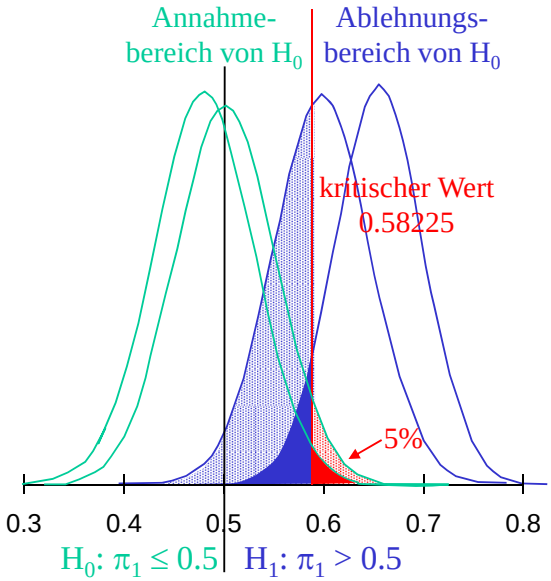
$$1-\Phi(Z)=0.05$$
$$\Rightarrow \Phi(Z)=0.95$$
$$\Rightarrow Z=\Phi^{-1}(0.95)=1.645$$

$$1.645=\frac{\text{krit}-0.5}{\sqrt{0.5\cdot(1-0.5)/100}}$$
$$\Rightarrow \text{krit}=z_{1-\alpha}\cdot\sigma\left(p_{1|H_0}\right)+\mu\left(p_{1|H_0}\right)$$
$$=1.645\cdot 0.05+0.5$$
$$=0.58225$$

Bei einer maximalen α-Fehlerwahrscheinlichkeit von 5% beträgt der kritische Wert 0.58225: Wenn die Nullhypothese gerade noch zutrifft (π_1=0.5), beträgt die Wahrscheinlichkeit 5%, dass der kritische Wert 0.58225 erreicht oder überschritten wird.

Ist $\pi_1 < 0.5$, ist die Wahrscheinlichkeit kleiner 5%, dass der kritische Wert erreicht oder überschritten wird. Die α-Fehlerwahrscheinlichkeit ist dann kleiner.

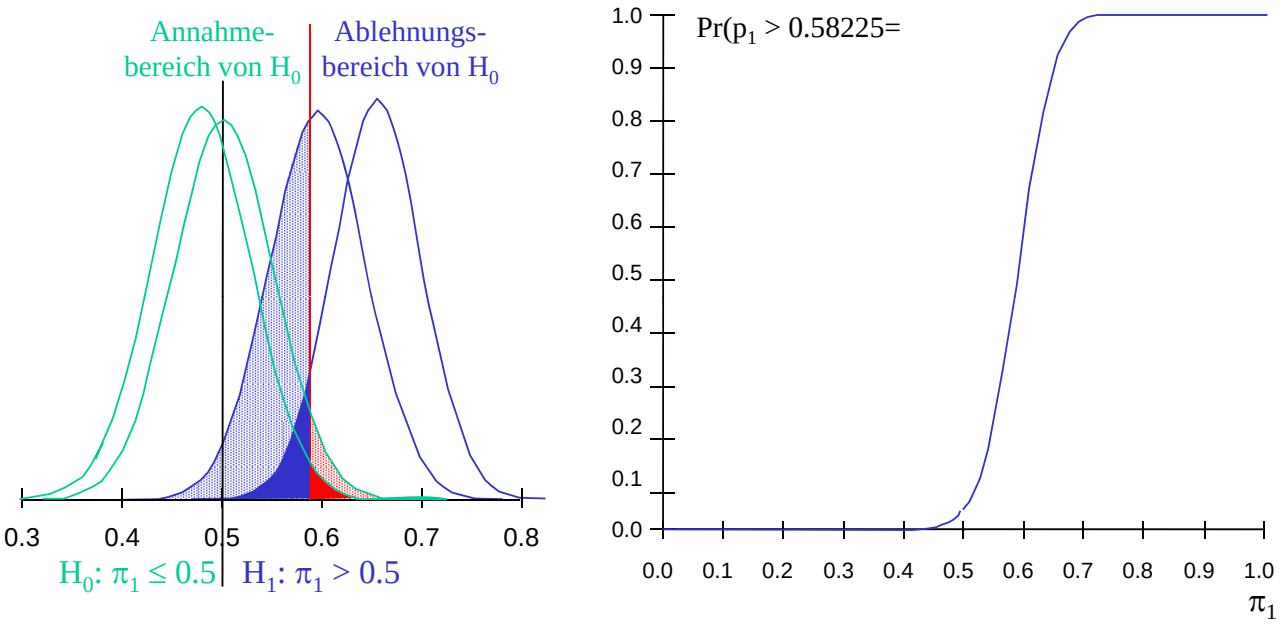
Festlegung der (maximalen) Irrtumswahrscheinlichkeit



Wenn umgekehrt die Forschungshypothese gerade noch zutrifft, der Populationsanteil also ganz geringfügig über 50% liegt, beträgt die β-Fehlerwahrscheinlichkeit immerhin fast 95%.

Je dichter der tatsächliche Populationswert nahe der Trennline zwischen Null- u. Alternativhypothese liegt, desto größer ist die β-Fehlerwahrscheinlichkeit.

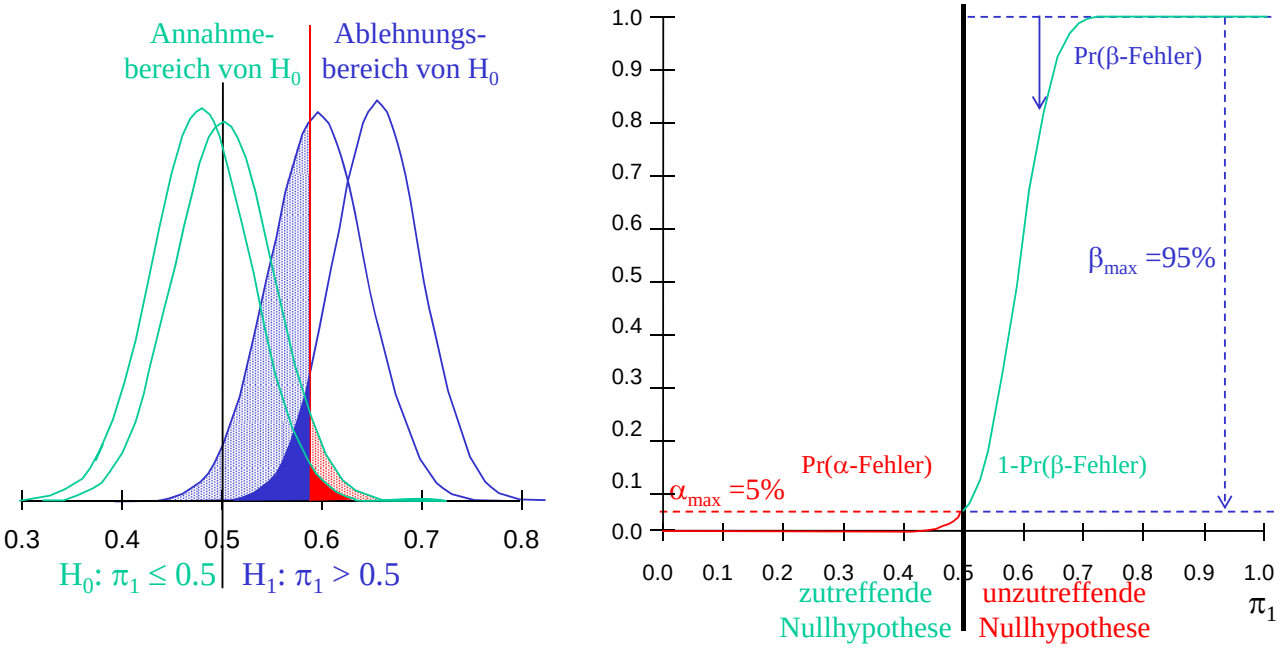
Festlegung der (maximalen) Irrtumswahrscheinlichkeit



Um einen Eindruck über die Fehlerrisiken insgesamt zu erhalten, kann für den gesamten Wertebereich der Teststatistik die Wahrscheinlichkeit berechnet werden, mit der die Teststatistik in den Ablehnungsbereich fällt.

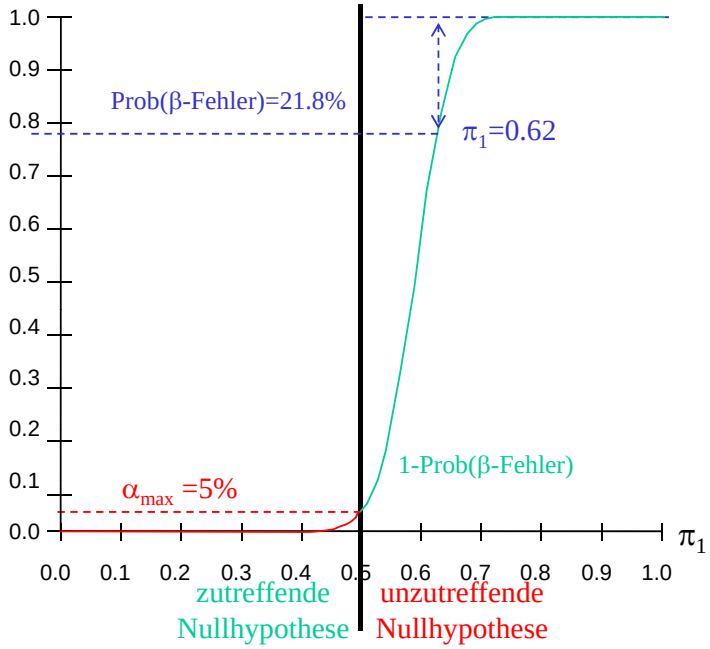
Die so berechnete Funktion heißt ***Teststärkefunktion*** *(engl.* ***power function****).*

Teststärkefunktion



Im Bereich der Nullhypothese gibt die Teststärkefunktion die Irrtumswahrscheinlichkeit α an.
Im Bereich der Alternativhypothese gibt die Teststärkefunktion Eins minus der β-Fehlerwahrscheinlichkeit an.

Trennschärfe



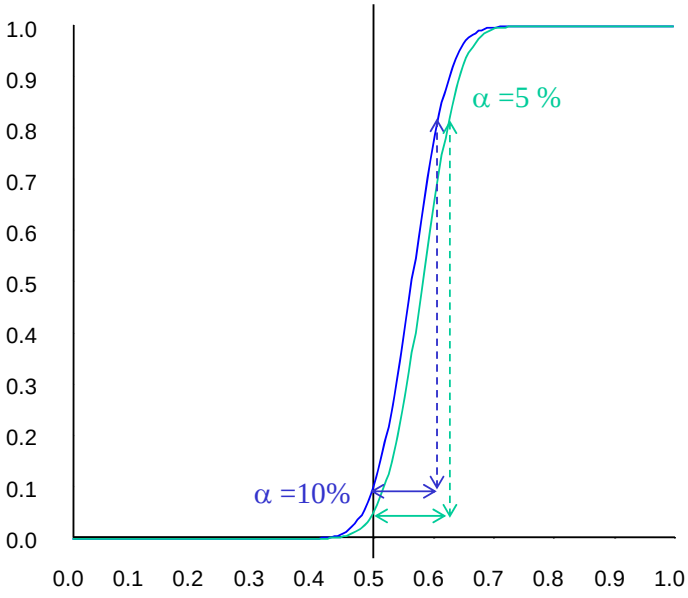
Wenn $\pi_1 \geq 0.62$ wird eine (dann falsche) Nullhypothese mit einer Wahrscheinlichkeit von $1-\beta \geq 78.2\%$ entdeckt.

Bei einem Wert von π_1 zwischen >0.5 und <0.62 liegt die (β-) Fehlerwahrscheinlichkeit zwischen 95% und 21.8 %. Der Test ist in diesem Bereich nicht ***trennscharf****.*

Wenn $\pi_1 \leq 0.5$, wird die (dann zutreffende) Nullhypothese mit einer Irrtumswahrscheinlichkeit von maximal $\alpha = 5\%$ entdeckt.

Die Teststärkefunktion sollte im Bereich der Nullhypothese möglichst geringe Werte nahe 0 und im Bereich der Alternativhypothese möglichst große Werte nahe 1 aufweisen. Es gibt jedoch immer einen Bereich, in dem ein Test sehr hohe Fehlerwahrscheinlichkeiten aufweist. In diesem nicht trennscharfen Bereich kann der Test nur schlecht zwischen H_0 und H_1 diskriminieren.

Einfluss des maximalen Irrtumswahrscheinlichkeit α auf die Trennschärfe



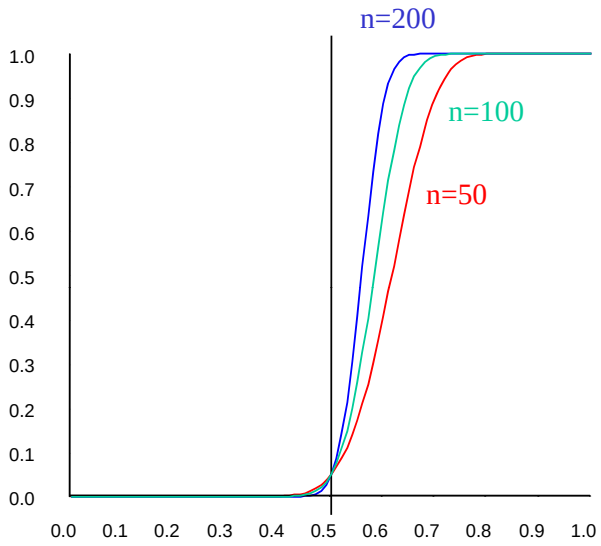
So ist bei einer Irrtumswahrscheinlichkeit von $\alpha \leq 10$ % der Bereich, in dem der Test nicht trennscharf ist, kleiner als bei einer Irrtumswahrscheinlichkeit von $\alpha \leq 5$ %.

Bei gegebener Fallzahl und zu geringer Trennschärfe muss daher gegebenenfalls die Irrtumswahrscheinlichkeit α heraufgesetzt werden.

Der „Preis“ für die größere Trennschärfe bei zutreffender Alternativhypothese ist allerdings, dass eher eine richtige Nullhypothese fälschlicherweise abgelehnt wird, also die maximale α-Fehlerwahrscheinlichkeit steigt.

Der Bereich geringer Trennschärfe ist kleiner, wenn die maximale α-Fehlerwahrscheinlichkeit heraufgesetzt wird.

Einfluss des Stichprobenumfangs auf die Trennschärfe



Bei einer Fallzahl von nur n=50 ist der Test im Bereich zwischen $\pi_1 > 0.5$ und etwa $\pi_1 < 0.68$ nicht sehr trennschaft

Bei einer Fallzahl von nur n=200 ist der Test im Bereich zwischen $\pi_1 > 0.5$ und etwa $\pi_1 < 0.57$ nicht sehr trennschaft

Ist die Teststärke nicht hoch genug, sollte - wenn möglich - die Fallzahl erhöht werden.

Bei gegebener Irrtumswahrscheinlichkeit hängt die ***Trennschärfe*** (***Teststärke***) von der Stichprobengröße ab:
je kleiner die Stichprobe, desto größer der Standardschätzfehler und desto weniger steil und damit weniger trennscharf verläuft die Teststärkefunktion.
Umgekehrt gilt: je größer die Stichprobe, desto kleiner der Standardschätzfehler und desto steiler und damit trennschärfer verläuft die Teststärkefunktion.

Vorgehensschritte beim statistischen Testen

Das Beispiel des Tests der Forschungshypothese, dass ein Populationseinteil einen vorgegebenen Wert überschreitet, zeigt die generelle Vorgehensweise bei der Hypothesenprüfung nach ***Neyman*** und ***Pearson***.

Schritt 1: Formulierung von Null- und Alternativhypothese
Im ersten Schritt wird der Wertebereich der betrachteten Populationseigenschaft in zwei disjunkte Bereiche zerteilt, die der Nullhypothese H_0 und der Alternativhypothese H_1 entsprechen. Dabei wird die Nullhypothese H_0 so formuliert, dass sie möglichst das Gegenteil der eigentlich interessierenden Forschungshypothese ist, die dann der Alternativhypothese H_1 entspricht.

Schritt 2: Auswahl der statistischen Prüfgröße (Teststatistik)
Für die Durchführung des Tests wird als empirische Prüfgröße eine Teststatistik benötigt, die bei richtiger und falscher Nullhypothese unterschiedliche Kennwerteverteilungen aufweist.

> *Im Beispiel des Prüfens einer Vermutung über einen Populationsanteil wurde als Teststatistik der Stichprobenanteil herangezogen, dessen Kennwerteverteilung asymptotisch normalverteilt ist.*

In der Regel wird eine Teststatistik so ausgewählt, dass ihre Kennwerteverteilung bei (gerade noch) zutreffender Nullhypothese leicht zu berechnen ist.

Vorgehensschritte beim statistischen Testen

Da die Quantile von Normalverteilungen i.a. über die Standardnormalverteilung berechnet werden, bietet es sich an, anstelle des Stichprobenanteils eine Teststatistik zu berechnen, die bei gerade noch gültiger Nullhypothese standardnormalverteilt ist.
Im Beispiel ist dies der Fall, wenn $\pi_1 = 0.5$ ist:

$$Z = \frac{p_1 - \mu\left(p_{1|H_o}\right)}{\sigma\left(p_{1|H_o}\right)} = \frac{p_1 - \pi_{1|H_o}}{\sqrt{\pi_{1|H_o} \cdot \left(1 - \pi_{1|H_o}\right)/n}} = \frac{p_1 - 0.5}{\sqrt{0.5 \cdot (1 - 0.5)/100}} = \frac{p_1 - 0.5}{0.05}$$

Schritt 3: Festlegung von Irrtumswahrscheinlichkeit α und Ablehnungsbereich

Die Entscheidungsregel des Tests sollte so sein, dass die Wahrscheinlichkeit einer Fehlentscheidung minimiert wird. Da aber α- und β-Fehlerwahrscheinlichkeiten nicht unabhängig voneinander sind, wird nur die maximale α-Fehlerwahrscheinlichkeit festgelegt.
Der Wert beträgt in der Regel 5% oder 1%.
Allerdings sollte die Trennschärfe des Tests berücksichtigt werden. Bei sehr hohen Fallzahlen ist eine sehr kleine α-Fehlerwahrscheinlichkeit angemessen. Wenn bei geringen Fallzahlen die Trennschärfe des Tests zu gering ist und die Fallzahl nicht heraufgesetzt werden kann, ist es sinnvoll, auch größere α-Fehlerwahrscheinlichkeiten von 10% oder sogar darüber zu akzeptieren.

Vorgehensschritte beim statistischen Testen

Ob die Trennschärfe hinreichend hoch ist, ist eine Frage, die vor dem Hintergrund der inhaltlichen Anwendung entschieden werden sollte.

Im Beispiel der Prüfung der Forschungshypothese, ob es eine Mehrheit zugunsten der Gesamtschule gibt, wäre vermutlich die Trennschärfe nicht hoch genug, wenn erst bei einem Populationswert, der sehr deutlich über $\pi_1 = 0.5$ liegt, die β-Fehlerwahrscheinlichkeit unter 20% läge:
Wenn erst bei einer 2/3-Mehrheit in der Population auch die Stichprobendaten mit hoher Wahrscheinlichkeit für die Forschungshypothese sprechen, wäre der Test nicht sehr sinnvoll.

Mit der Festlegung der Irrtumswahrscheinlichkeit wird auch der Ablehnungsbereich festgelegt, der auch als kritischer Bereich bezeichnet wird. Der Ablehnungsbereich wird dabei so gewählt, dass die maximale α-Fehlerwahrscheinlichkeit nicht überschritten wird.

Im Beispiel ist die Teststatistik nur bei dem Populationswert $\pi_1 = 0.5$ standardnormalverteilt. Aus den Regeln der Lineartransformation einer Variablen ergibt sich generell für den Erwartungswert der Teststatistik:

$$\mu(Z) = \mu\left(\frac{p_1 - 0.5}{0.05}\right) = \mu\left(\frac{-0.5}{0.05} + \frac{1}{0.05} \cdot p_1\right) = \frac{-0.5}{0.05} + \frac{1}{0.05} \cdot \mu(\pi_1) = \frac{\pi_1 - 0.5}{0.05}$$

Bei einem Populationsanteil $\pi_1 < 0$ ist der Erwartungswert von Z $\mu(Z) < 0$, bei einem Populationsanteil $\pi_1 > 0$ ist der Erwartungswert von Z $\mu(Z) > 0$.

Vorgehensschritte beim statistischen Testen

Da nur im letzten Fall die Nullhypothese falsch wäre, ist bei falscher Nullhypothese eher mit hohen (positiven) Werten der Teststatistik zu rechnen.
Bei einer Irrtumswahrscheinlichkeit von maximal 5% sollte die Nullhypothese daher abgelehnt werden, wenn die Teststatistik in einer Stichprobe das 95%-Quantil der Standardnormalverteilung erreicht oder überschreitet. Dies ist der Fall, wenn $Z \geq 1.645$.
Bei gerade noch gültiger Nullhypothese, wenn also $\pi_1 = 0.5$, wird dieser Wert nur in 5% aller Stichproben erreicht. Ist $\pi_1 < 0.5$, dann ist dies noch seltener der Fall.

Formal gesehen spielt es keine Rolle, ob der kritische Wert für die Teststatistik Z oder für den Stichprobenanteil p berechnet wird.

Wird anstelle der Teststatistik Z der Stichprobenanteil p_1 als Teststatistik verwendet, ist der kritische Wert das 95%-Quantil der Stichprobenverteilung von p_1, wenn $\pi_1 = 0.5$.
Der so bestimmte Wert 0.58225 ergibt sich auch durch die Umkehrung der Z-Transformation aus den kritischen Wert von Z:

$$z_{0.95} = 1.645 = \frac{p_{krit} - \mu\left(p_{1|H_0}\right)}{\sigma\left(p_{1|H_0}\right)} = \frac{p_{krit} - 0.5}{0.05} \Rightarrow p_{krit} = 1.645 \cdot 0.05 + 0.5 = 0.58225$$

Generelle Vorgehensweise beim statistischen Testen

Schritt 4: Berechnung der Prüfgröße und Entscheidung

Nachdem die Teststatistik ausgewählt und über die Festlegung der Irrtumswahrscheinlichkeit auch der Ablehnungsbereich bzw. der kritische Wert berechnet ist, kann der Test anhand der empirischen Stichprobendaten durchgeführt werden.

Im Beispiel betrug der Stichprobenanteil 0.6. Daraus ergibt sich für die Teststatistik:

$$Z = \frac{p_1 - 0.5}{0.05} = \frac{0.6 - 0.5}{0.05} = 2.0$$

Da der Wert 2.0 größer ist als der kritischen Wert 1.645 , liegt die Teststatistik im Ablehnungsbereich. Die Nullhypothese ist daher bei einer Irrtumswahrscheinlichkeit von maximal 5% zu verwerfen.
Umgekehrt ist dann zu erwarten, dass die Alternativhypothese und damit die Forschungshypothese vermutlich zutrifft. Der Test führt also zu der Entscheidung, dass von einer Mehrheit für die Einrichtung der Ganztagsschule auszugehen ist.

Schritt 5: Überprüfung der Anwendungsvoraussetzungen

Die Durchführung eines Tests ist in der Regel an Anwendungsvoraussetzungen gebunden. Diese sollten am besten vor der Durchführung des Tests geprüft werden. Da die Prüfung bisweilen aber bereits die Berechnung der Teststatistik beinhaltet, erfolgt die Prüfung oft erst im letzten Schritt.

Generelle Vorgehensweise beim statistischen Testen

Im Beispiel wird neben der Voraussetzung einer (einfachen) Zufallsauswahl die asymptotische Annäherung des Stichprobenanteils an die Normalverteilung vorausgesetzt. Als Faustregel gilt, dass diese Annäherung hinreichend genau ist, wenn die Quotienten aus den beiden Populationsanteilen π_1 und π_0 (= 1 – π_1) mal der Fallzahl größer 9 sind. Bei der Anwendung wird dabei als Populationsanteil wieder der Wert genommen, bei dem die Nullhypothese gerade noch zutrifft, also $\pi_1 = 0.5$. Da n=100 , ergibt sich hier:

$$n \cdot \frac{\pi_1}{1-\pi_1} = n \cdot \frac{1-\pi_1}{\pi_1} = n = 100 > 9$$

Die Anwendungsvoraussetzungen sind also erfüllt.

Wenn die Anwendungsvoraussetzungen nicht erfüllt sind und es keine anderen Test gibt, dessen Voraussetzungen erfüllt sind, dann ist das Ergebnis des Tests nur mit Vorsicht zu verwenden. Es kann zwar sein, dass das Testergebnis gegenüber der Verletzung einer Anwendungsvoraussetzung ***robust*** ist und dann das Testergebnis trotz Verletzung einer Annahme korrekt ist. Die ist aber nicht immer der Fall.

Einseitige und zweiseitige Tests

Im Beispiel des Tests der Forschungshypothese, dass eine Mehrheit für die Einführung einer Ganztagesschule ist, ist die Nullhypothese falsch, wenn ein Populationswert einen vorgegebenen Wert (im Beispiel: $\pi_1 > 0.5$) erreicht oder überschreitet.

Ein solcher Test heißt ***einseitiger Hypothesentest***, da der von der Nullhypothese postulierte Wertebereich eines Populationsparameters entweder gegen ein Überschreiten (wie im Beispiel) oder gegen ein Unterschreiten geprüft wird.

In einem ***zweiseitigen Hypothesentest*** postuliert die Nullhypothese dagegen, dass der zu testende Populationsparameter einen bestimmten Wert aufweist. Die Nullhypothese ist dann falsch, sowohl, wenn dieser Wert überschritten, als auch, wenn er unterschritten wird.
Die generelle Vorgehensweise unterscheidet sich in der Schrittfolge nicht von der Vorgehensweise bei einem einseitigen Test.

Schritt 1: Formulierung von Null- und Alternativhypothese

Beispiel: Ein Spieler ist der Ansicht, dass er beim geschickten Werfen einer Münze entscheiden kann, ob eher „Kopf“ oder eher „Zahl“ oben liegen wird. Die Forschungshypothese behauptet dann, dass die Wahrscheinlichkeit (von z.B. „Kopf“) ungleich 0.5 ist:

$$H_0: \pi_1 = 0.5 \text{ versus } H_1: \pi_1 \neq 0.5$$

In einem Experiment soll dies geprüft werden, indem der Spieler in insgesamt n= 250-maligen Werfen eines Würfels jedes Mal „Kopf“ erreichen soll.

Zweiseitige Tests

Schritt 2: Auswahl der statistischen Prüfgröße:

Der Stichprobenanteil ist bei einer einfachen Zufallsauswahl um den Populationsanteil normalverteilt.

Wenn $\pi_1 = 0.5$, *dann ist* $Z = \dfrac{p_1 - \pi_1}{\sqrt{\dfrac{\pi_1 \cdot (1-\pi_1)}{n}}} = \dfrac{p_1 - 0.5}{\sqrt{\dfrac{0.5 \cdot (1-0.5)}{250}}}$

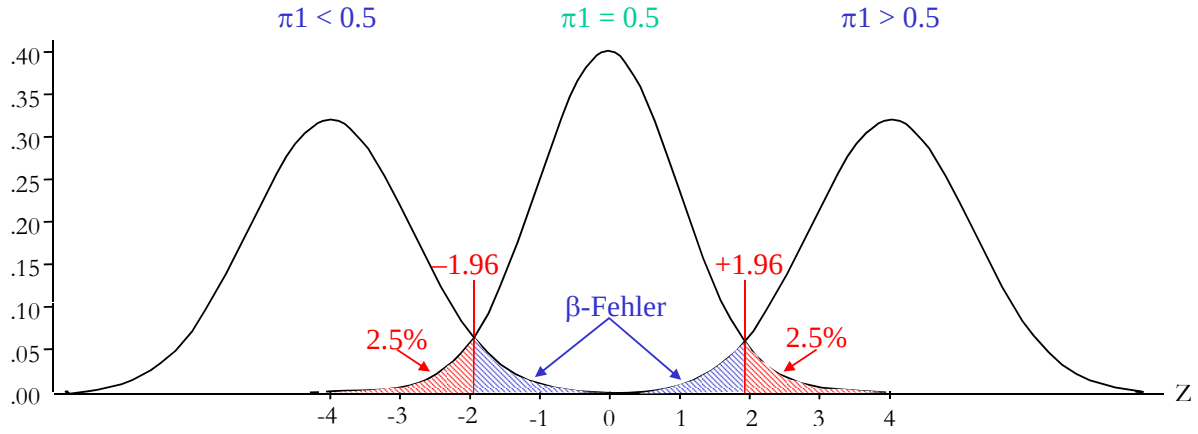
standardnormalverteilt.

Wenn die Nullhypothese falsch ist, $\pi_1 \neq 0.5$, *dann ist entweder eher mit* ***kleinen*** *Werten (wenn* $\pi_1 < 0.5$*) oder aber eher mit* ***großen*** *Werten (wenn* $\pi_1 > 0.5$*) der Teststatistik zu rechnen.*
Wenn die Nullhypothese zutrifft, ist dagegen mit Werten um 0.0 zu rechnen.

Schritt 3: Festlegung von Irrtumswahrscheinlichkeit und kritischen Werten:

Die Irrtumswahrscheinlichkeit soll $\alpha = 5\%$ *betragen. Da bei zutreffender Nullhypothese mit Werten der Teststatistik um den Erwartungswert 0.0 zu rechnen ist, ist die Nullhypothese abzulehnen, wenn die Werte der Teststatistik sehr viel kleiner oder sehr viel gößer als 0.0 sind.*
Daher wird die Nullhypothese abgelehnt, wenn die Teststatistik kleiner als das 2.5%-Quantil oder aber größer als das 97.5%-Quantil der Standardnormalverteilung ist. Die kritischen Werte sind daher ±1.96.

Zweiseitige Tests



Bei einem zweiseitigen Hypothesentest gibt es auch ***zwei kritische Werte***, die den Bereich der Akzeptanz der Nullhypothese gegen die Teilbereiche der Ablehnung der Nullhypothese abgrenzen.

Die Wahrscheinlichkeit bei gültiger Nullhypothese den unteren kritischen Wert zu erreichen oder zu unterschreiten oder aber den oberen kritischen Wert zu erreichen oder zu überschreiten, ist jeweils $\alpha/2$, so dass die Irrtumswahrscheinlichkeit insgesamt $2 \cdot \alpha/2 = \alpha$ beträgt.

Bei Populationswerten im Bereich der Alternativhypothese H_1 ist dann die β-Fehlerwahrscheinlichkeit die Wahrscheinlichkeit, dass die Teststatistik größer als der untere kritische Wert und kleiner als der obere kritische Wert ist.

Zweiseitige Tests

Schritt 4: Berechnung der Teststatistik und Entscheidung

Wenn der Spieler bei den 250-maligen Werfen der Münze 130-mal das Ergebnis „Kopf“ erziehlt, berechnet sich für die Teststatistik der Wert

$$Z = \frac{\frac{130}{250} - 0.5}{\sqrt{\frac{0.5 \cdot (1-0.5)}{250}}} = \frac{0.02}{0.0316} = 0.632$$

Da –1.96 < 0.632 < +1.96, kann die Nullhypothese bei einer Irrtumswahrscheinlichkeit von 5% nicht verworfen werden. Obwohl der Spieler in mehr als 50% der Fälle „Kopf“ erzielt, kommt der Test zu dem Ergebnis, dass die Abweichung von 50% auch rein zufällig sein kann. Es kann daher nicht davon ausgegangen werden, dass der Spieler tatsächlich das Ergebnis durch sein Werfen beeinflussen kann.

Schritt 5: Überprüfung der Anwendungsvoraussetzungen

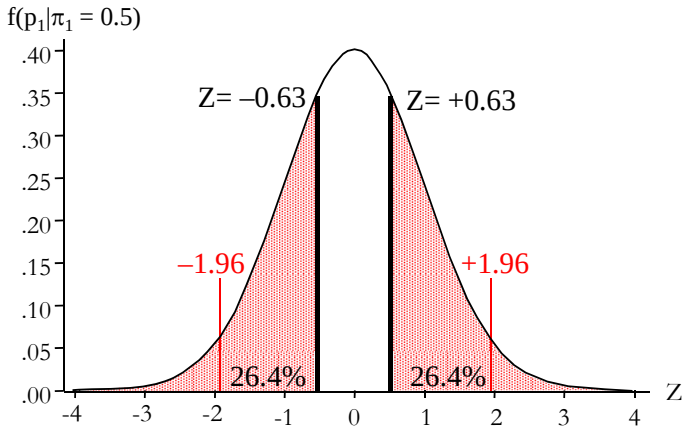
Die Prüfung der Anwendungsvoraussetzung ergibt für π_1=0.5:

$$n \cdot \frac{\pi_1}{1-\pi_1} = n \cdot \frac{1-\pi_1}{\pi_1} = n = 250 > 9$$

Die Anwendungsvoraussetzungen sind erfüllt.

Empirisches Signifikanzniveau

Zusätzlich zum Wert der Teststatistik wird vor allem bei Signifikanztests oft das ***empirische Signifikanzniveau*** berichtet, das die Wahrscheinlichkeit angibt, dass eine Teststatistik bei (gerade noch) zutreffender Nullhypothese den beobachteten Wert annimmt oder einen Wert, der noch stärker gegen die Nullhypothe spricht.



Im Beispiel des zweiseitigen Tests der Nullhypothese H_0: π_1 = 0.5 beträgt der Wert der Teststatistik 0.632.
Diesem Wert entspricht im zweiseitigen Test ein empirisches Signifikanzniveau von 52.8%:

$Pr(Z \geq 0.632) = 1 - \Phi(0.632) = 26.4\%$
$Pr(Z \leq -0.632) = \Phi(-0.632) = 26.4\%$

$Pr(-0.632 \geq Z \;\&\; Z \geq 0.632) = 52.8\%$

Ist das empirische Signifikanzniveau kleiner als die maximale Irrtumswahrscheinlichkeit α, dann ist die Nullhypothese zu verwerfen;
ist das empirische Signifikanzniveau größer oder gleich der maximale Irrtumswahrscheinlichkeit α, dann ist die Nullhypothese beizubehalten.

Signifikanztest und Neyman-Pearson-Test

Die Vorgehensweise beim Signifikanztest nach Fisher und dem Hypothesentest nach Neyman und Pearson ist bei der praktischen Anwendung meistens identisch:
Die Nullhypothese wird abgelehnt, wenn der Wert der Teststatistik in der Stichprobe bei gegebener maximaler α-Fehlerwahrscheinlichkeit im Ablehnungsbereich liegt, wobei die Nullhypothese (möglichst) das Gegenteil der eigentlich interessierenden Forschungshypothese postuliert.

Der Unterschied liegt vor allem darin, dass beim Neyman-Pearson-Test von vornherein auch die Möglichkeit eines β-Fehlers berücksichtigt wird. Daraus folgt, dass als Teststatistik nur eine Statistik in Frage kommt, deren Kennwerteverteilung sich bei Zutreffen der Null- bzw. der Alternativhypothese unterscheidet. Die Berücksichtigung der Trennschärfe führt zudem dazu, dass vor der Durchführung eines Tests der notwendige Stichprobenumfang und der Wert der maximalen α-Fehlerwahrscheinlichkeit stärker ins Blickfeld gerät.

Dies ermöglicht es, Kritik an einer rein mechanischen Anwendung von Signifikanztests zu begegnen. Die ***Kritik an Signifikanztest*** bezieht sich vor allem darauf, dass bei sehr großen Fallzahlen praktisch jede empirische Teststatistik signifikant wird, die Nullhypothese also abgelehnt wird. Umgekehrt ist es bei sehr kleinen Fallzahlen oft kaum möglich ein signifikantes Ergebnis zu erhalten. Beide Ergebnisse sind Folge der eigentlich erwünschten Eigenschaft, dass die **Höhe des Standardfehlers einer Teststatistik mit steigender Fallzahl sinkt**.

Signifikanztest und Neyman-Pearson-Test

Beim Neyman-Pearson-Test kann dies berücksichtigt werden, indem entsprechend die α-Fehlerwahrscheinlichkeit an die Fallzahl angepasst wird, bei großen Fallzahlen also nur eine sehr kleine maximale α-Fehlerwahrscheinlichkeit akzeptiert wird.
Bei sehr kleinen Fallzahlen kann dagegen auch bei großen α-Fehlerwahrscheinlichkeiten die Trennschärfe so gering sein, dass der Test nicht informativ ist. Dann ist die Duchführung eines Tests gar nicht mehr sinnvoll.

Bei der Anwendung eines statistischen Tests kommt es zudem darauf an, auch inhaltlich relevante Hypothesenpaare zu formulieren. Wird etwa die Nullhypothese, dass es keinen Zusammenhang zwischen zwei Variablen gibt, gegen die Alternativhypothese geprüft, dass es einen Zusammenhang gibt, dann führt ein Test korrekterweise bei großen Fallzahlen vermutlich selbst dann zu einem signifikanten Ergebnis, wenn der Zusammenhang sehr sehr gering ist. Wenn aber ein nur geringer Zusammenhang inhaltlich nicht interessiert, wäre eine Hypothesenformulierung angemessener, bei der die Nullhypothese behauptet, dass ein Zusammenhang geringer oder gleich einem Minimalwert ist, der als praktisch irrelevant betrachtet wird.

Beziehung zwischen Hypothesentest und Konfidenzintervallen

Es wurde bereits darauf hingewiesen, dass ein Hypothesentest auch über die Berechnung eines Konfidenzintervalls durchgeführt werden kann. Tatsächlich entspricht die Berechnung eines Konfidenzintervalls einem ***zweiseitigen Hypothesentest***, bei dem die Nullhypothese immer dann abgelehnt wird, wenn der durch die Nullhypothese postulierter Wert außerhalb der Grenzen des Konfidenzintervalls liegt. Die Nullhypothese wird umgekehrt beibehalten, wenn der durch die Nullhypothese postulierter Wert innerhalb des Konfidenzintervalls liegt.

Entsprechend kann der Annahmebereich beim zweiseitigen Hypothesentest analog der Berechnung eines Konfidenzintervalls als Berechnung eines Intervalls um den durch die Nullhypothese postulierten Wert erfolgen.

Dies kann am Beispiel des Test eines Populationsanteils verdeutlicht werden, wenn als Teststatistik nicht Z, sondern der Stichprobenanteil verwendet wird. Der Annahmebereich des Hypothesenpaars H_0: $\pi_1 = \pi$ vs. H_1: $\pi_1 \neq \pi$ berechnet sich dann nach:

$$\text{Annahmebereich von } H_0\text{: c.i.}\left(\pi_1 \middle| H_0\right) = \pi_{1|H_0} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\pi_{1|H_0} \cdot \left(1-\pi_{1|H_0}\right)}{n}}$$

Der Unterschied zum üblichen Konfidenzintervall besteht allein darin, dass beim Konfidenzintervall das Intervall um den Stichprobenwert berechnet wird:

$$\text{Konfidenzintervall von } \pi_1\text{: c.i.}\left(\pi_1\right) = p_1 \pm z_{1-\alpha/2} \cdot \sqrt{\frac{p_1 \cdot \left(1-p_1\right)}{n}}$$

Hypothesentests über Populationsmittelwerte

Das Anwendungsbeispiel zur Logik des Hypothesentestens bezieht sich auf den Test eines Populationsanteils. Mit der gleichen Logik können auch andere Populationsparameter geprüft werden. Notwendig ist stets die Kenntnis der Kennwerteverteilung.
Um Mittelwerte zu testen, wird entsprechend die Kennwerteverteilung von Stichprobenmittelwerten benötigt. In L19 wurde die entsprechenden Verteilungen bereits für die Berechnung von Konfidenzintervallen herangezogen. Dabei wurde danach unterschieden, ob die betrachtete Variable in der Population normalverteilt ist oder nicht.

Asymptotische Z-Test bei beliebigen Populationsverteilungen
Ist dies nicht der Fall, wird die unabhängig von der Populationsverteilung asymptotisch gültige Normalverteilung herangezogen. Zur Prüfung der Hypothesen
a) H_0: $\mu_X = \mu_0$ vs. H_1: $\mu_X \neq \mu_0$ (zweiseitiger Test) bzw.
b) H_0: $\mu_X \leq \mu_0$ vs. H_1: $\mu_X > \mu_0$ (einseitiger Test nach oben) bzw.
c) H_0: $\mu_X \geq \mu_0$ vs. H_1: $\mu_X < \mu_0$ (zweiseitger Test nach unten)
wird die Teststatistik:

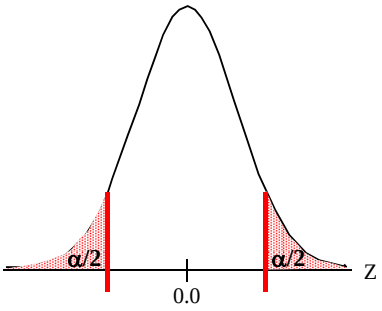
$$Z = \frac{\bar{x} - \mu_0}{\hat{\sigma}\left(\bar{x}\right)} = \frac{\bar{x} - \hat{\sigma}}{\sqrt{\hat{\sigma}_X^2 / n}} = \frac{\bar{x} - \hat{\sigma}}{\sqrt{s_X^2 / \left(n-1\right)}} = \frac{\bar{x} - \hat{\sigma}}{\sqrt{\frac{SS_X}{n \cdot \left(n-1\right)}}}$$

berechnet.
Trifft die Nullhypothese (gerade noch) zu, ist also μ_X in der Population gleich dem postulierten Wert μ_0, dann ist die Teststatistik asymptotisch standardnormalverteilt.

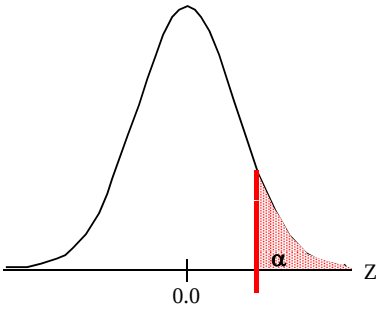
Hypothesentests über Populationsmittelwerte

Ist die Nullhypothese falsch, dann ist die Teststatistik dagegen asymptotisch normalverteilt, wobei der Erwartungswert gleich der Differenz $\mu_X - \mu_0$ ist.
Daraus folgt als Entscheidungsregel für Hypothesentests mit der Irrtumswahrscheinlichkeit α:
a) Bei H_0: $\mu_X = \mu_0$: Wenn $Z \leq z_{\alpha/2}$ oder $Z \geq z_{1-\alpha/2} \Rightarrow H_1$;
b) Bei H_0: $\mu_X \leq \mu_0$: Wenn $Z \geq z_{1-\alpha} \Rightarrow H_1$;
c) Bei H_0: $\mu_X \geq \mu_0$: Wenn $Z \leq z_{\alpha} \Rightarrow H_1$.

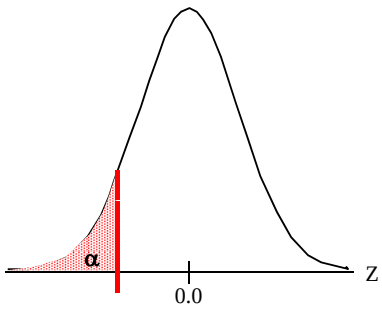
Ablehnungsbereich bei
H_0: $\mu_X = \mu_0$
sehr kleine oder sehr große
Werte sprechen gegen H_0



Ablehnungsbereich bei
H_0: $\mu_X \leq \mu_0$
sehr große Werte
sprechen gegen H_0



Ablehnungsbereich bei
H_0: $\mu_X \geq \mu_0$
sehr kleine Werte
sprechen gegen H_0



Hypothesentests über Populationsmittelwerte

Hinweis:
Da Anteile als Mittelwerte von 0/1-kodierten dichotomen Variablen aufgefasst werden können, gilt dieser Test auch für einen beliebigen Anteil p_1, wenn bei der Berechnung des Standardfehlers berücksichtigt wird, dass die Stichprobenvarianz ein erwartungstreuer Schätzer der Populationsvarianz ist.
Die generelle Teststatistik zur Prüfung von Anteilen ist daher:

$$Z = \frac{p_1 - \pi_0}{\hat{\sigma}(p_1)} = \frac{p - \pi_0}{\sqrt{\frac{p_1 \cdot (1 - p_1)}{n}}}$$

Exakter T-Test bei in der Population normalverteilten Variablen

Wenn bekannt ist, dass die Variable, über die ein Test des Populationsmittelwerts durchgeführt wird, in der Population normalverteilt wird, wird statt der asymptotisch gültigen Normalverteilung die T-Verteilung mit df = n–1 Freiheitsgraden für die Berechnung der kritischen Werte verwendet. Die Teststatistik ändert sich nicht:

$$T = \frac{\bar{x} - \mu_0}{\hat{\sigma}(\bar{x})} = \frac{\bar{x} - \hat{\sigma}}{\sqrt{\hat{\sigma}_X^2 / n}} = \frac{\bar{x} - \hat{\sigma}}{\sqrt{s_X^2 / (n-1)}} = \frac{\bar{x} - \hat{\sigma}}{\sqrt{\frac{SS_X}{n \cdot (n-1)}}}$$

Hypothesentests über Populationsmittelwerte

Die Entscheidungsregel ist hier:
a) Bei H_0: $\mu_X = \mu_0$: Wenn $T \leq t_{\alpha/2;df=n-1}$ oder $T \geq t_{1-\alpha/2;df=n-1} \Rightarrow H_1$;
b) Bei H_0: $\mu_X \leq \mu_0$: Wenn $T \geq t_{1-\alpha;df=n-1} \Rightarrow H_1$;
c) Bei H_0: $\mu_X \geq \mu_0$: Wenn $T \leq t_{\alpha;df=n-1} \Rightarrow H_1$.

Da der Z-Test nur asymptotisch gültig ist, setzt er hinreichend große Fallzahlen voraus ($n \geq 30$ besser $n \geq 50$ bzw. bei Anteilen $n \geq 60$ und $n \cdot p_1/(1-p_1) > 9$ und $n \cdot (1-p_1)/p_1 > 9$). Der T-Test ist dagegen auch bei kleinen Fallzahlen gültig, wenn die Normalverteilungsannahme zutrifft.

Da bei gleichen Irrtumswahrscheinlichkeiten der Ablehnungsbereich beim T-Test kleiner ist als beim Z-Test, die Nullhypothese also nicht so leicht abgelehnt wird, wird der T-Test beim Testen von Mittelwerten (aber nicht von Anteilen!) oft auch dann herangezogen, wenn die getestete Variable in der Population nicht normalverteilt ist bzw. deren Verteilung unbekannt ist.
Im Sinne eines vorsichtigen (konservativen oder strengen) Testens ist das allerdings nur sinnvoll, wenn die Nullhypothese tatsächlich das Gegenteil der eigentlich interessierenden Forschungshypothese ist.

Lerneinheit 15: Von der Anteilsdifferenz zur Vierfeldertabelle

Eine der wichtigsten Aufgaben der Statistik in den Sozialwissenschaften besteht in der Analyse von Zusammenhängen.

So mag sich z.B. ein Sozialwissenschaftler dafür interessieren, ob sich in den alten Bundesländern Männer und Frauen bei der Einstellung zum Schwangerschaftsabbruch unterscheiden.
Als empirische Datenbasis finden sich im Allbus 2006 die Antworten von Befragten auf die Frage, ob Schwangerschaftsabbruch entsprechend dem Willen der Frau erlaubt oder verboten sein sollte.

Um diese Fragen zu beantworten, müssen die Antworten der Männer auf diese Frage mit den Antworten der Frauen verglichen werden.

Berechnet man getrennt die Häufigkeitsverteilung von Männern und Frauen ergibt sich folgendes Bild:

Antworten männlicher Befragter:

Schwangerschaftsabbruch nach Willen der Frau	n_k	p_k	cp_k
- sollte erlaubt sein	344	0.330	0.330
- sollte verboten sein	700	0.670	1.000
Summe	1044	1.000	

(Quelle: Allbus 2006, nur Westen)

Antworten weiblicher Befragter:

Schwangerschaftsabbruch nach Willen der Frau	n_k	p_k	cp_k
- sollte erlaubt sein	403	0.365	0.365
- sollte verboten sein	701	0.635	1.000
Summe	1104	1.000	

(Quelle: Allbus 2006, nur Westen)

Von der Anteilsdifferenz zur Vierfeldertabelle

Antworten männlicher Befragter

Schwangerschaftsabbruch nach Willen der Frau	n_k	p_k	cp_k
- sollte erlaubt sein	344	0.330	0.330
- sollte verboten sein	700	0.670	1.000
Summe	1044	1.000	

(Quelle: Allbus 2006, nur Westen)

Antworten weiblicher Befragter

Schwangerschaftsabbruch nach Willen der Frau	n_k	p_k	cp_k
- sollte erlaubt sein	403	0.365	0.365
- sollte verboten sein	701	0.635	1.000
Summe	1104	1.000	

(Quelle: Allbus 2006, nur Westen)

Der Vergleich der beiden Verteilungen zeigt, dass die weiblichen Befragten in der Allbus-Stichprobe sich zu einem geringfügig größeren Anteil für die Erlaubnis des Schwangerschaftsbbruchs aussprechen als die männlichen Befragten:
Die Differenz der entsprechenden Anteile beträgt 0.365 – 0.330 = 0.035.

Die Darstellung der Häufigkeitsverteilungen der Antworten in zwei getrennten Tabellen für Männer und Frauen erscheint nicht sehr sinnvoll, wenn die Zahlen für die Interpretation wieder zusammengestellt werden müssen.

Tatsächlich wird in der ***bivariaten Zusammenhangsanalyse*** die ***gemeinsame Häufigkeitsverteilung*** von zwei Variablen analysiert.
Dazu kann die gemeinsame Verteilung von zwei kategorialen Variablen in einer ***Kreuztabelle*** zusammengefasst werden, die die Häufigkeitsverteilung der ***Auftretenskombinationen*** der beiden Variablen wiedergibt.

Von der Anteilsdifferenz zur Vierfeldertabelle

Antworten männlicher Befragter Schwangerschaftsabbruch nach Willen der Frau	n_k	p_k	cp_k
- sollte erlaubt sein	344	0.330	0.330
- sollte verboten sein	700	0.670	1.000
Summe	1044	1.000	

Antworten weiblicher Befragter Schwangerschaftsabbruch nach Willen der Frau	n_k	p_k	cp_k
- sollte erlaubt sein	403	0.365	0.365
- sollte verboten sein	701	0.635	1.000
Summe	1104	1.000	

Kreuztabelle von „Haltung zum Schwangerschaftsabbruch“ und Geschlecht:

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	344	403	747
- sollte verboten sein	700	701	1401
Summe	1044	1104	2148

Die Daten in der Kreuztabelle enthalten die gleichen Zahlen wie die getrennten univariaten Häufigkeitstabellen.

So ist erkennbar, dass von den 1044 männlichen Befragten 344 für eine Erlaubnis und 624 für ein Verbot des Schwangerschaftsabbruchs sin und von den 1104 Frauen 403 für eine Erlaubnis und 701 für ein Verbot.

Zusätzlich enthält die Kreuztabelle in der unteren Zeile bzw. der rechten Randspalte Informationen über die univariaten Häufigkeitsverteilungen der beiden betrachteten Variablen X (“Geschlecht”) und Y (“Schwangerschaftsabbruch”).

Von der Anteilsdifferenz zur Vierfeldertabelle

Antworten männlicher Befragter Schwangerschaftsabbruch nach Willen der Frau	n_k	p_k	cp_k
- sollte erlaubt sein	344	0.330	0.330
- sollte verboten sein	700	0.670	1.000
Summe	1044	1.000	

Antworten weiblicher Befragter Schwangerschaftsabbruch nach Willen der Frau	n_k	p_k	cp_k
- sollte erlaubt sein	403	0.365	0.365
- sollte verboten sein	701	0.635	1.000
Summe	1104	1.000	

Spaltenvariable

Zeilenvariable

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	344	403	747
- sollte verboten sein	700	701	1401
Summe	1044	1104	2148

Die Variable, deren Ausprägungen die Zeilen der Kreuztabelle festlegen, heißt ***Zeilenvariable*.**

Im Beispiel ist die Variable Y “Haltung zum Schwangerschaftsabbruch” Zeilenvariable.

Die Variable, deren Ausprägungen die Spalten der Kreuztabelle festlegen, heißt ***Spaltenvariable*.**

Im Beispiel ist die Variable X “Geschlecht” Spaltenvariable.

Von der Anteilsdifferenz zur Vierfeldertabelle

Spaltenvariable

Zeilenvariable	Geschlecht des Befragten (X)		Summe
Schwangerschaftsabbruch nach Willen der Frau (Y)	männlich	weiblich	
- sollte erlaubt sein	n_{11} 344	403	747
- sollte verboten sein	700	701	1401
Summe	1044	1104	2148

Entsprechend der Zahl der Ausprägungen der Zeilen- und der Spaltenvariable spricht man von ***I×J-Tabellen*** *(engl.* ***r by c-tables****)*, wenn die Zeilenvariable I *(engl: r* für *"number of* ***r****ows")* Ausprägungen und die Spaltenvariable J *(engl: c* für *"number of* ***c****olumns")* Ausprägungen hat.

Im Beispiel liegt eine "2 mal 2"-Tabelle vor, da beide Variablen dichotom sind, also nur 2 Ausprägungen haben.

Die 2×2-Tabelle ist die kleinstmögliche Kreuztabelle von zwei Variablen. Sie hat 2×2 = 4 (innere) Zellen.
Man bezeichnet eine solche Kreuztabelle auch als ***Vierfeldertabelle*** (in älteren Texten *Vierfeldertafel*).

Um die einzelnen Zellen einer Kreuztabelle eindeutig zu identifizieren, werden Indizes verwendet, die die Nummer der Ausprägung der Zeilen- und Spaltenvariablen angeben.

Im Beispiel gibt es 344 Fälle mit der ***Ausprägungskombination*** *"männlich" und "sollte erlaubt sein", d.h.* $n_{11} = 344$

Von der Anteilsdifferenz zur Vierfeldertabelle

Spaltenvariable

Zeilenvariable	Geschlecht des Befragten (X)		Summe
Schwangerschaftsabbruch nach Willen der Frau (Y)	männlich	weiblich	
- sollte erlaubt sein	n_{11} 344	n_{12} 403	$n_{1\bullet}$ 747
- sollte verboten sein	n_{21} 700	n_{22} 701	$n_{2\bullet}$ 1401
Summe	$n_{\bullet 1}$ 1044	$n_{\bullet 2}$ 1104	$n_{\bullet\bullet}$ 2148 n

An erster Stelle steht immer der Zeilenindex, an zweiter Stelle der Spaltenindex.

n_{21} ist daher die gemeinsame Häufigkeit der zweiten Ausprägung der Zeilenvariable und der ersten Ausprägung der Spaltenvariable.

n_{12} ist dagegen die gemeinsame Häufigkeit der ersten Ausprägung der Zeilenvariable und der zweiten Ausprägung der Spaltenvariable und n_{22} die gemeinsame Häufigkeit jeweils der zweiten Ausprägung der Zeilen- und der Spaltenvariable.

Die univariaten Verteilungen am rechten und unteren Rand, die sich auch durch Aufsummieren der inneren Tabellenzellen ergeben, werden dadurch gekennzeichnet, dass ein "•" oder ein "+" für den Index steht, über den aufsummiert wird.

$n_{1\bullet}$ oder n_{1+} ist daher die Häufigkeit der ersten Ausprägung der Zeilenvariable;

$n_{\bullet 1}$ oder n_{+1} ist entsprechend die Häufigkeit der ersten Ausprägung der Spaltenvariable.

In der untersten rechten Zelle steht dann die Gesamtfallzahl $n_{\bullet\bullet}$ (oder einfach n).

Im Beispiel ist n=2148.

Von der Anteilsdifferenz zur Vierfeldertabelle

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	a	344	b	403	747	a+b
- sollte verboten sein	c	700	d	701	1401	c+d
Summe	a+c	1044	b+d	1104	2148	

Nur in Vierfeldertabellen gibt es die Besonderheit, dass die vier inneren Tabellenzellen auch durch die ersten vier kleinen Buchstaben des Alphabets bezeichnet werden.

Im Beispiel ist $a = n_{11} = 344$, $b = n_{12} = 403$, $c = n_{21} = 700$ *und* $d = n_{22} = 701$.

Wenn wie im Beispiel die Zellen einer Kreuztabelle die absoluten Auftretenshäufigkeiten enthalten, dann zeigt die Tabelle die ***gemeinsame*** oder ***bivariate absolute Häufigkeitsverteilung*** der Zeilen- und der Spaltenvariable in der Stichprobe.

Da die univariate Häufigkeitsverteilungen der beiden Variablen in den rechten bzw. unteren Randzellen der Tabelle wiedergegeben werden, werden die univariaten Verteilungen in diesem Kontext auch als ***Randverteilungen*** bezeichnet.
Formal ergeben sich Randverteilungen durch Aggregation über die Ausprägungen anderer Variablen.

Die Randverteilung der Zeilenvariable Geschlechts ergibt sich im Beispiel durch Aufsummieren über die Ausprägungen der Spaltenvariable. So ergeben sich die 1044 Fälle, die männlich sind, durch Aufsummieren der Häufigkeiten in den Zellen a und c.

Von der Anteilsdifferenz zur Vierfeldertabelle

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	a	344	b	403	747	a+b
- sollte verboten sein	c	700	d	701	1401	c+d
Summe	a+c	1044	b+d	1104	2148	n

In der Regel werden in einer Kreuztabelle Ausprägungen für ungültige Fälle (*missing values*) nicht aufgeführt. Nur wenn es keine ungültigen Fälle bei den kreuztabellierten Variablen gibt, ist die Gesamtfallzahl gleich dem Stichprobenumfang.

Geschlecht	n_k
- männlich	1115
- weiblich	1184
- k. A.	0
Summe	2299

Abtreibung	n_k
- erlaubt	747
- verboten	1401
- w. n.	139
- k. A.	12
Summe	2299

Im Allbus 2006-Beispiel enthält die Weststichprobe 2299 Fälle, von denen 1115 männlich und 1184 weiblich sind.
Bei der Frage nach dem Schwangerschaftsabbruch nach Willen der Frau gibt es jedoch 151 ungültige Angaben, wobei 139 Befragte mit "weiß nicht" antworteten und von 12 Befragten keine Angabe vorliegt. In der Tabelle werden daher nur die 2148 (=2299 – 151) Fälle aufgeführt, die bei beiden Variablen gültige Antworten aufweisen.

Zusammenhangsanalyse in der Vierfeldertabelle

Kreuztabelle von „Haltung zum Schwangerschaftsabbruch“ und Geschlecht:

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	344	403	747
- sollte verboten sein	700	701	1401
Summe	1044	1104	2148

Ziel der Betrachtung einer bivariaten Verteilung ist die Beantwortung der Frage, ob, und wenn, welcher Zusammenhang zwischen den beiden Variablen besteht.

Im Beispiel soll der Frage nachgegangen werden, ob sich die Einstellung zum Schwangerschaftsabbruch bei Männern und Frauen unterscheidet.
Dazu werden die relativen Häufigkeiten von Männern und Frauen verglichen.

Statistisch gesehen ist der Vergleich der relativen Antworthäufigkeiten der Männern mit der der Frauen ein Vergleich von ***bedingten (konditionalen) Verteilungen***.

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	0.330 (344)	0.365 (403)	0.348 (747)
- sollte verboten sein	0.670 (700)	0.635 (701)	0.652 (1401)
Summe	1.000 (1044)	1.000 (1104)	1.000 (2148)

(Quelle: Allbus 2006, nur Westen)

Zusammenhangsanalyse in der Vierfeldertabelle

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	0.033 (344)	0.365 (403)	0.348 (747)
- sollte verboten sein	0.670 (700)	0.635 (701)	0.652 (1401)
Summe	1.000 (1044)	1.000 (1104)	1.000 (2148)

(Quelle: Allbus 2006, nur Westen)

Bei der Berechnung werden die Zellenhäufigkeiten in jeder Spalte durch die Spaltensumme in der unteren Zeile geteilt:

$$p_{i(j)} = \frac{n_{ij}}{n_{\bullet j}}$$

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	$p_{1(1)} = n_{11}/n_{\bullet 1}$	$p_{1(2)} = n_{12}/n_{\bullet 2}$	$p_{1\bullet} = n_{1\bullet}/n$
- sollte verboten sein	$p_{2(1)} = n_{21}/n_{\bullet 1}$	$p_{2(2)} = n_{22}/n_{\bullet 2}$	$p_{2\bullet} = n_{2\bullet}/n$
Summe	1.000 ($n_{\bullet 1}$)	1.000 ($n_{\bullet 2}$)	1.000 (n)

Um die bedingende Variable von der bedingten zu unterscheiden, wird im Folgenden der Index der bedingenden Variable – im Beispiel die Spaltenvariable Geschlecht – in Klammern gesetzt. $p_{i(j)}$ steht also für die (konditionale) relative Häufigkeit der i-ten Ausprägung der Zeilenvariable, wenn die Spaltenvariable die j-te Ausprägung aufweist.

Zusammenhangsanalyse in der Vierfeldertabelle

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	33.0% (344)	36.5% (403)	34.8% (747)
- sollte verboten sein	67.0% (700)	63.5% (701)	65.2% (1401)
Summe	100.0% (1044)	100.0% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

Anstelle der Anteile werden oft Prozentwerte angegeben.

Wärend 33.0% der Männer der Ansicht sind, Schwangerschaftsabbruch nach dem Willen der Frau sollte erlaubt sein, sind es 35.5% der Frauen, die diese Ansicht teilen.

Zwei Zufallsvariablen sind statistisch unabhängig voneinander, wenn bedingte und unbedingte Verteilungen gleich sind.

Bei Unabhängigkeit der beiden Variablen sollten daher in der Kreuztabelle die relativen Häufigkeiten der konditionalen Verteilungen gleich den Randverteilungen sein.

Im Beispiel müssten dann die Prozentwerte in der ersten Zeile stets 34.8% betragen und in der zweiten Zeile 65.2%.

Da sich relative Häufigkeiten und absolute Häufigkeiten ineinander umrechnen lassen, lässt sich berechnen, welche absoluten Häufigkeiten zu erwarten wären, wenn Unabhängigkeit zwischen den Variablen bestünde, indem die relativen Häufigkeiten mit der jeweiligen Bezugszahl multipliziert werden.

Zusammenhangsanalyse in der Vierfeldertabelle

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	33.0% (344)	36.5% (403)	34.8% (747)
- sollte verboten sein	67.0% (700)	63.5% (701)	65.2% (1401)
Summe	100.0% (1044)	100.0% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

Bei Unabhängigkeit ergäbe sich für a = 0. 348 ·1044 = 363.1, für b = 0.348 ·1104 = 383.9, für c = 0.652 ·1044 = 608.9 und für d = 0.652 ·1104 = 720.1.
Um nicht zu große Rundungsfehler zu erhalten, ist bei der Berechnung anstelle von 0.348 der Quotient 747/2148 und statt 0.652 der Quotient 1401/2148 eingesetzt.
Die so berechneten erwarteten absoluten Häufigkeiten sind auf eine Nachkommastelle gerundet.

Bei Unabhängigkeit erwartete Häufigkeiten

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	34.8% (363.1)	34.8% (383.9)	34.8% (747)
- sollte verboten sein	65.2% (680.9)	65.2% (720.1)	65.2% (1401)
Summe	100.0% (1044)	100.0% (1104)	100.0% (2148)

Zusammenhangsanalyse in der Vierfeldertabelle

Beobachtete Häufigkeiten

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	33.0% (344)	36.5% (403)	34.8% (747)
- sollte verboten sein	67.0% (700)	63.5% (701)	65.2% (1401)
Summe	100.0% (1044)	100.0% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

Bei Unabhängigkeit erwartete Häufigkeiten

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	34.8% (363.1)	34.8% (383.9)	34.8% (747)
- sollte verboten sein	65.2% (680.9)	65.2% (720.1)	65.2% (1401)
Summe	100.0% (1044)	100.0% (1104)	100.0% (2148)

In der Realität sind absolute Häufigkeiten mit Nachkommastellen unmöglich.Tatsächlich können die **bei Unabhängigkeit erwarteten absoluten Häufigkeiten** als **Schätzung der Erwartungswerte** von binomialverteilten Zufallsvariablen interpretiert werden.

Wenn es in einer Population sowohl unter den Männern wie den Frauen einen Anteil von $\pi_1 = 0.348$ gibt, die für die Erlaubnis des Schwangerschaftsabbruchs sind, und in einfachen Zufallsauswahlen jeweils 1044 Männer und 1104 Frauen ausgewählt werden, dann wäre der Erwartungswert der Männer 363.1 ($= n \cdot \pi_1 = 1044 \cdot 0.348$) und der der Frauen 383.9 ($= n \cdot \pi_1 = 1104 \cdot 0.348$).

Zusammenhangsanalyse in der Vierfeldertabelle

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	33.0% (344)	36.5% (403)	34.8% (747)
- sollte verboten sein	67.0% (700)	63.5% (701)	65.2% (1401)
Summe	100.0% (1044)	100.0% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

Neben dem Extrem statistischer Unabhängigkeit (kein Zusammenhang) kann auch der umgekehrte Fall eines ***maximalen (perfekten) Zusammenhangs*** bestehen.
Dabei können zwei Situationen unterschieden werden.

Zum einen bestünde ein maximaler Zusammenhang wenn im Beispiel alle Männer für die Erlaubnis und alle Frauen für ein Verbot von Schwangerschaftsabbrüchen wären.

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	100% (1044)	0% (0)	48.6% (1044)
- sollte verboten sein	0% (0)	100% (1104)	51.2% (1104)
Summe	100% (1044)	100% (1104)	100.0% (2148)

Zusammenhangsanalyse in der Vierfeldertabelle

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	33.0%	(344)	36.5%	(403)	34.8%	(747)
- sollte verboten sein	67.0%	(700)	63.5%	(701)	65.2%	(1401)
Summe	100.0%	(1044)	100.0%	(1104)	100.0%	(2148)

(Quelle: Allbus 2006, nur Westen)

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe
- sollte erlaubt sein	0%	(0)	100%	(1104)	51.2% (1104)
- sollte verboten sein	100%	(1044)	0%	(0)	48.6% (1044)
Summe	100%	(1044)	100%	(1104)	100.0% (2148)

Zum anderen bestünde auch ein maximaler Zusammenhang wenn im Beispiel alle Frauen für die Erlaubnis und alle Männer für ein Verbot von Schwangerschaftsabbrüchen wären.

Zur Erfassung der Stärke des Zusammenhang bietet es sich daher an, die Differenz der Prozentwerte der Ausprägungen der Haltung zum Schwangerschaftsabruch zwischen den beiden Geschlechtern als ein Maß für die Stärke des Zusammenhangs zwischen zwei dichotomen Variablen zu verwenden.

Prozentsatzdifferenz

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	33.0%	(344)	36.5%	(403)	34.8%	(747)
- sollte verboten sein	67.0%	(700)	63.5%	(701)	65.2%	(1401)
Summe	100.0%	(1044)	100.0%	(1104)	100.0%	(2148)

(Quelle: Allbus 2006, nur Westen)

Dieses Zusammenhangsmaß heißt ***Prozentsatzdifferenz*** $d_{YX}\%$ und gibt die Differenz der bedingten relativen Häufigkeiten in ***Prozentpunkten*** an:

$$d_{YX}\% = 100 \cdot \left(p_{1(1)} - p_{1(2)} \right) = 100 \cdot \left(\frac{n_{11}}{n_{\bullet 1}} - \frac{n_{12}}{n_{\bullet 2}} \right) = 100 \cdot \left(\frac{a}{a+c} - \frac{b}{b+d} \right)$$

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	a	344	b	403	747	a+b
- sollte verboten sein	c	700	d	701	1401	c+d
Summe	a+c	1044	b+d	1104	2148	n

Im Beispiel beträgt die Prozentsatzdifferenz 100 ·(344/1044 – 403/1104) = –3.5 Prozentpunkte.

Die Differenz wird in „***Prozentpunkten***“ und nicht in „Prozent“ gemessen, da sich die Prozentwerte in den beiden Spalten auf **verschiedene Bezugszahlen** beziehen,

im Beispiel 33.0% auf n_{+1} = 1044 und 36.5% auf n_{+2} = 1104 Fälle.

Prozentsatzdifferenz

Der Wertebereich der Prozentsatzdifferenz liegt zwischen –100 Prozentpunkten und +100 Prozentpunkten.
Besteht kein Zusammenhang, beträgt der Wert 0 Prozentpunkte.

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	34.8%	(363.1)	34.8%	(383.9)	34.8%	(747)
- sollte verboten sein	65.2%	(680.9)	65.2%	(720.1)	65.2%	(1401)
Summe	100.0%	(1044)	100.0%	(1104)	100.0%	(2148)

d_{YX} = 0.0% kein Zusam-menhang

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	100%	(1044)	0%	(0)	48.6%	(1044)
- sollte verboten sein	0%	(0)	100%	(1104)	51.2%	(1104)
Summe	100%	(1044)	100%	(1104)	100.0%	(2148)

d_{YX} = + 100.0% perfekter positiver Zusammenhang

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	0%	(0)	100%	(1104)	51.2%	(1104)
- sollte verboten sein	100%	(1044)	0%	(0)	48.6%	(1044)
Summe	100%	(1044)	100%	(1104)	100.0%	(2148)

d_{YX} = – 100.0% perfekter negativer Zusammenhang

Prozentsatzdifferenz

Als Hilfestellung bei der Interpretation einer Prozentsatzdifferenz wird der Wertebereich nach einer *Faustregel* in Regionen eingeteilt:

	$-5\% < d_{YX}\% < +5\%$	praktisch kein Zusammenhang
	$+5\% \le d_{YX}\% < +20\%$	geringer Zusammenhang
bzw.	$-20\% < d_{YX}\% \le -5\%$	
	$+20\% \le d_{YX}\% < +50\%$	mittlerer Zusammenhang
bzw.	$-50\% < d_{YX}\% \le -20\%$	
	$+50\% \le d_{YX}\%$	starker Zusammenhang
bzw.	$-50\% \le d_{YX}\%$	

Das Vorzeichen der Prozentsatzdifferenz ist ab ordinalem Skalenniveau der beiden Variablen interpretierbar. Dabei ist Vorsicht angebracht, da es von der Kodierung der Variablen abhängt, ob eine Prozentsatzdifferenz positiv oder negativ ist.
Das Vorzeichen der Prozentsatzdifferenz stimmt mit der Richtung der Beziehung (positiv: je mehr X desto mehr Y bzw. negativ: je mehr X desto weniger Y) nur dann überein, wenn sowohl die erste Ausprägung der Spalten- wie auch der Zeilenvariablen entweder für ein “mehr” oder für ein “weniger” einer Eigenschaft stehen.
Im Zweifelsfall sollte nur der Absolutbetrag der Prozentsatzdifferenz berichtet werden.

Prozentsatzdifferenz

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	33.0% (344)	36.5% (403)	34.8% (747)
- sollte verboten sein	67.0% (700)	63.5% (701)	65.2% (1401)
Summe	100.0% (1044)	100.0% (1104)	100.0% (2148)
(Quelle: Allbus 2006, nur Westen)			

Stärke eines Zusammenhangs	
praktisch kein	$0 \leq \lvert d_{YX}\% \rvert < 5$
geringer	$5 \leq \lvert d_{YX}\% \rvert < 20$
mittlerer	$20 \leq \lvert d_{YX}\% \rvert < 50$
starker	$50 \leq \lvert d_{YX}\% \rvert$
perfekter Zus.	$100 = \lvert d_{YX}\% \rvert$

Die Prozentsatzdifferenz von nur 3.5 Prozentpunkten weist darauf hin, dass es praktisch keinen Unterschied zwischen Männern und Frauen bei der Frage gibt, ob ein Schwangerschaftsabbruch bei finanzieller Notlage erlaubt oder verboten sein sollte.

Das negative Vorzeichen (33.0 – 36.5 = –3.5) besagt aufgrund der Kodierung des Geschlechts ("weiblich" = höherer Wert) und des Schwangerschaftsabbruchs ("Verbot" = höherer Wert), dass Frauen sich seltener für ein ein Verbot von Schwangerschaftsabbrüchen aussprechen als Männer.

Prozentsatzdifferenz

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	33.0% (344)	36.5% (403)	34.8% (747)
- sollte verboten sein	67.0% (700)	63.5% (701)	65.2% (1401)
Summe	100.0% (1044)	100.0% (1104)	100.0% (2148)
(Quelle: Allbus 2006, nur Westen)	**$d_{YX\%}$ = – 3.5 Prozentpunkte**		

Wären die Werte für weibliche Befragte in der ersten Spalte oder wären in der ersten Zeile der ersten Zeile die Werte derjenigen aufgetragen, die für ein Verbot sind, dann wäre die Prozentsatzdifferenz bei gleichem Absolutwert positiv.

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) weiblich	männlich	Summe
- sollte erlaubt sein	36.5% (403)	33.0% (344)	34.8% (747)
- sollte verboten sein	63.5% (701)	67.0% (700)	65.2% (1401)
Summe	100.0% (1104)	100.0% (1044)	100.0% (2148)
(Quelle: Allbus 2006, nur Westen)	**$d_{YX\%}$ = + 3.5 Prozentpunkte**		

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte verboten sein	67.0% (700)	63.5% (701)	65.2% (1401)
- sollte erlaubt sein	33.0% (344)	36.5% (403)	34.8% (747)
Summe	100.0% (1044)	100.0% (1104)	100.0% (2148)
(Quelle: Allbus 2006, nur Westen)	**$d_{YX\%}$ = + 3.5 Prozentpunkte**		

Kennwerteverteilung der Prozentsatzdifferenz

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	33.0%	(344)	36.5%	(403)	34.8%	(747)
- sollte verboten sein	67.0%	(700)	63.5%	(701)	65.2%	(1401)
Summe	100.0%	(1044)	100.0%	(1104)	100.0%	(2148)

(Quelle: Allbus 2006, nur Westen)

Die Prozentsatzdifferenz ist eine Linearkombination von zwei Anteilen $p_{1(1)}$ und $p_{1(2)}$ in einer Kreuztabelle.
In einer einfachen Zufallsauswahl sind die beiden Anteile jeweils asymptotisch normalverteilt und statistisch unabhängig voneinander, wenn entweder getrennte Stichproben für die beiden Ausprägungen der Spaltenvariable gezogen werden, oder aber es dem Zufall der Auswahl überlassen bleibt, welche Ausprägung bei der Spaltenvariable realisiert wird.

Die Kennwerteverteilung ist dann ebenfalls asymptotisch normalverteilt, wobei sich Erwartungswert und Varianz nach den Regeln für Linearkombinationen berechnen lassen:

$$d_{YX}\% = 100 \cdot p_{1(1)} + (-100) \cdot p_{1(2)}$$

$$f\left(d_{YX}\%\right) \approx N\left(100 \cdot \left(\pi_{1(1)} - \pi_{1(2)}\right); 100^2 \cdot \left(\frac{\pi_{1(1)} \cdot \pi_{2(1)}}{n_{\bullet 1}} + \frac{\pi_{1(2)} \cdot \pi_{2(2)}}{n_{\bullet 2}}\right)\right)$$

Bei der Varianz wird von einer einfachen Zufallsauswahl mit Zurücklegen ausgegangen. Bei kleinen Population ist zusätzlich der Faktor (N–n)/(N–1) zu berücksichtigen.

Kennwerteverteilung der Prozentsatzdifferenz

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	a	344	b	403	747	a+b
- sollte verboten sein	c	700	d	701	1401	c+d
Summe	a+c	1044	b+d	1104	2148	n

Da die Populationsanteile $\pi_{1(1)}$, $\pi_{2(1)}$, $\pi_{1(2)}$ und $\pi_{2(2)}$ unbekannt sind, werden sie durch die Stichprobenanteile $p_{1(1)}$, $p_{2(1)}$, $p_{1(2)}$ und $p_{2(2)}$ geschätzt.

Der Standardfehler der Prozentsatzdifferenz beträgt dann:

$$\hat{\sigma}\left(d_{YX}\%\right) = 100 \cdot \sqrt{\frac{p_{1(1)} \cdot p_{2(1)}}{n_{\bullet 1}} + \frac{p_{1(2)} \cdot p_{2(2)}}{n_{\bullet 2}}} = 100 \cdot \sqrt{\frac{\frac{a}{a+c} \cdot \frac{c}{a+c}}{a+c} + \frac{\frac{b}{b+d} \cdot \frac{d}{b+d}}{b+d}}$$

$$= 100 \cdot \sqrt{\frac{a \cdot c}{(a+c)^3} + \frac{b \cdot d}{(b+d)^3}}$$

Die asymptotische Annäherung an die Normalverteilung ist hinreichend genau, wenn
(a) $n_{\bullet 1} \cdot p_{1(1)}/p_{2(1)} = (a+c) \cdot a/c > 9$ und $n_{\bullet 1} \cdot p_{2(1)}/p_{1(1)} = (a+c) \cdot c/a > 9$,
(b) $n_{\bullet 2} \cdot p_{1(2)}/p_{2(2)} = (b+d) \cdot b/d > 9$ und $n_{\bullet 2} \cdot p_{2(2)}/p_{1(2)} = (b+d) \cdot d/b > 9$,
(c) $n_{\bullet 1} > 60$ und
(d) $n_{\bullet 2} > 60$

Konfidenzintervall für die Prozentsatzdifferenz

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	a	344	b	403	747	a+b
- sollte verboten sein	c	700	d	701	1401	c+d
Summe	a+c	1044	b+d	1104	2148	n

Für das Beispiel des Zusammenhangs zwischen Haltung zum Schwangerschaftsabbruch und Geschlecht ergibt sich anhand der Allbus-Daten ein Standardfehler von:

$$\hat{\sigma}(d_{YX}\%) = 100 \cdot \sqrt{\frac{344 \cdot 700}{(1044)^3} + \frac{403 \cdot 701}{(1104)^3}} = 2.053$$

Analog zum Vorgehen bei einfachen Anteilen lässt sich das (1–α)-Konfidenzintervall für die Prozentsatzdifferenz berechnen nach:

$$\text{c.i.}(\delta_{YX}\%) = d_{YX}\% \pm z_{1-\alpha/2} \cdot \hat{\sigma}(d_{YX}\%) = d_{YX}\% \pm z_{1-\alpha/2} \cdot 100 \cdot \sqrt{\frac{a \cdot c}{(a+c)^3} + \frac{b \cdot d}{(b+d)^3}}$$

Die Grenzen des 95%-Konfidenzintervalls berechnen sich für das Beispiel nach:

$$-3.5 \pm 1.96 \cdot 2.053 = [-7.52 \; ; \; 0.52]$$

Mit einer Irrtumswahscheinlichkeit von 5% ist damit zurechnen, dass die Prozentsatzdifferenz in der Population zwischen –7.5 und +0.5 Punkte beträgt.

Hypothesentests über Prozentsatzdifferenzen

Die Kennwerteverteilung lässt sich auch für Hypothesentests über Prozentsatzdifferenzen nutzen.

Schritt 1: Formulierung von Null- und Alternativhypothese

Wie bei einfachen Anteilen lassen sich drei Hypothesenpaare unterscheiden:

(a) $H_0: \delta_{YX}\% = d\%$ versus $H_1: \delta_{YX}\% \neq d\%$

(b) $H_0: \delta_{YX}\% \leq d\%$ versus $H_1: \delta_{YX}\% > d\%$

(c) $H_0: \delta_{YX}\% \geq d\%$ versus $H_1: \delta_{YX}\% < d\%$

In den Hypothesen steht d% für einen vorgegebenen Wert, den die Prozentsatzdifferenz nach der Nullhypothese einnimmt (a), nicht überschreitet (b) oder nicht unterschreitet (c).
Das erste Hypothesenpaar führt zu einem zweiseitigen, das zweite und dritte zu einseitigen Tests.

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung

Für die Teststatistik wird die asymptotische Normalverteilung der Kennwerteverteilung ausgenutzt und die Prozentsatzdifferenz in der Stichprobe unter der Annahme, dass $d_{YX}\% = d\%$ ist, standardisiert:

$$Z = \frac{d_{YX}\% - d\%}{\hat{\sigma}(d_{YX}\%)} = \frac{\left(\frac{a}{a+c} - \frac{b}{b+d}\right) - \frac{d\%}{100}}{\sqrt{\frac{a \cdot c}{(a+c)^3} \cdot \frac{b \cdot d}{(b+d)^3}}}$$

Y \ X	1	2	Σ
1	a	b	a+b
2	c	d	c+d
Σ	a+c	b+d	n

Hypothesentests über Prozentsatzdifferenzen

Wenn die Prozentsatzdifferenz d_{YX}% tatsächlich gleich d% ist, dann ist die Teststatistik asymptotisch standardnormalverteilt.
Trifft diese Annahme nicht zu, ist die Teststatistik asymptotisch normalverteilt mit einem Erwartungswert größer Null, wenn die Prozentsatzdifferenz δ_{YX}% in der Population größer d% ist, bzw. mit einem Erwartungswert kleiner Null, wenn die Prozentsatzdifferenz δ_{YX}% in der Population kleiner d% ist.

Spezieller Standardfehler, wenn d% = 0

Ein Wert von d% = 0 bedeutet, dass bei zutreffender Nullhypothesen die Prozentsatzdifferenz in der Population 0 ist, 0 nicht überschreitet oder 0 nicht unterschreitet.
Falls in der Population tatsächlich δ_{YX}% = 0, dann sind bedingte und unbedingte relative Häufigkeiten gleich groß.

Dies kann bei der Berechnung des Standardfehlers ausgenutzt werden, in dem bei der Schätzung der Wahrscheinlichkeiten anstelle der bedingten Anteile aus den beiden Tabellenspalten die unbedingten Anteile aus der Randverteilung verwendet werden.
Der Standardfehler der Teststatistik berechnet sich dann also nach:

$$\hat{\sigma}\left(Z\middle|\delta_{YX}\%=0\right)=\sqrt{\frac{p_{1\bullet}\cdot p_{2\bullet}}{n_{\bullet 1}}+\frac{p_{1\bullet}\cdot p_{2\bullet}}{n_{\bullet 2}}}=\sqrt{p_{1\bullet}\cdot p_{2\bullet}\cdot\left(\frac{1}{n_{\bullet 1}}+\frac{1}{n_{\bullet 2}}\right)}$$

$$=\sqrt{\frac{(a+b)}{n}\cdot\frac{(c+d)}{n}\cdot\left(\frac{1}{a+c}+\frac{1}{b+d}\right)}$$

	X		
Y	1	2	∑
1	a	b	a+b
2	c	d	c+d
∑	a+c	b+d	n

Hypothesentests über Prozentsatzdifferenzen

Die Teststatistik ergibt sich dann nach:

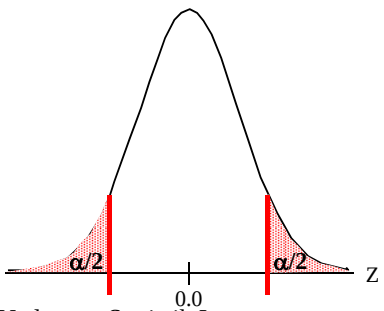
$$Z=\frac{p_{1(1)}-p_{2(1)}}{\sqrt{p_{1\bullet}\cdot p_{2\bullet}\cdot\left(\frac{1}{n_{\bullet 1}}+\frac{1}{n_{\bullet 2}}\right)}}=\frac{\frac{a}{a+c}-\frac{b}{b+d}}{\sqrt{\frac{(a+b)\cdot(c+d)}{n^2}\cdot\left(\frac{1}{a+c}+\frac{1}{b+d}\right)}}$$

	X		
Y	1	2	∑
1	a	b	a+b
2	c	d	c+d
∑	a+c	b+d	n

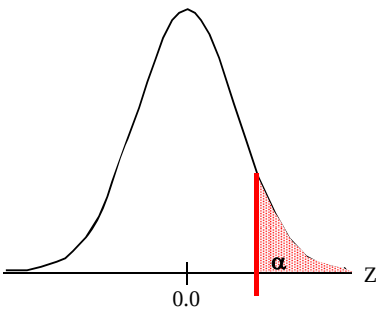
Schritt 3: Festlegung von Irrtumswahrscheinlichkeit und kritischen Werten

Die Ablehnungsbereiche ergeben sich jeweils danach, wann eine Teststatistik gegen die Nullhypothese spricht:

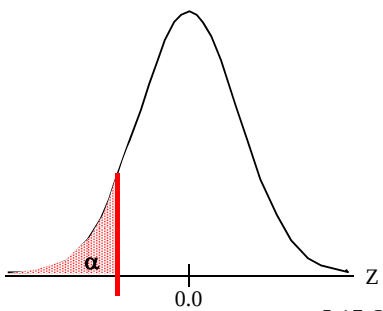
Ablehnungsbereich bei
H_0: δ_{YX}% = d%
sehr kleine oder sehr große Werte sprechen gegen H_0



Ablehnungsbereich bei
H_0: δ_{YX}% ≤ d%
sehr große Werte sprechen gegen H_0

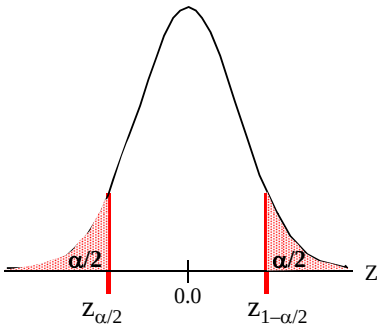


Ablehnungsbereich bei
H_0: δ_{YX}% ≥ d%
sehr kleine Werte sprechen gegen H_0

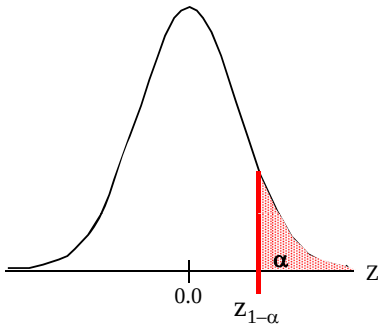


Hypothesentests über Prozentsatzdifferenzen

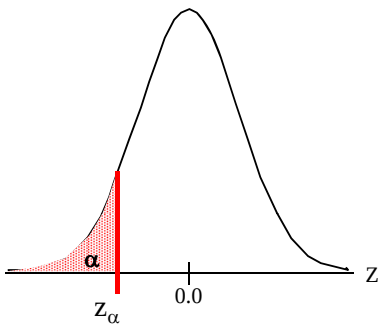
Ablehnungsbereich bei
a) H_0: $\delta_{YX}\% = d\%$



Ablehnungsbereich bei
b) H_0: $\delta_{YX}\% \leq d\%$



Ablehnungsbereich bei
c) H_0: $\delta_{YX}\% \geq d\%$



Beim Hypothesenpaar a) sind die kritischen Werte daher das $(\alpha/2)$- und das $(1-\alpha/2)$-Quantil, bei b) ist der kritische Wert das $(1-\alpha)$-Quantil und bei c) das α-Quantil der Standardnormalverteilung.

Schritt 4: Berechnung der Teststatistik und Entscheidung

Im vierten Schritt wird die Teststatistik berechnet und anhand des resultierenden Wertes die Nullhypothese beibehalten bzw. verworfen.

Die Nullhypothese H_0 wird mit einer Irrtumswahrscheinlichkeit α abgelehnt, wenn

(a) beim Test von H_0: $\delta_{YX}\% = d\%$ gilt: $Z \leq z_{\alpha/2}$ oder $Z \geq z_{1-\alpha/2}$,
(b) beim Test von H_0: $\delta_{YX}\% \leq d\%$ gilt: $Z \geq z_{1-\alpha}$ bzw.
(c) beim Test von H_0: $\delta_{YX}\% \geq d\%$ gilt: $Z \leq z_{\alpha}$.

Hypothesentests über Prozentsatzdifferenzen

Schritt 5: Prüfung der Anwendungsvoraussetzungen

Vor der endgültigen Interpretation der Ergebnisse ist zu prüfen, ob die Anwendungsvoraussetzungen für die Durchführung des Tests erfüllt sind.

	X		
Y	1	2	∑
1	a	b	a+b
2	c	d	c+d
∑	a+c	b+d	n

Die Annäherung an die Normalverteilung ist hinreichend genau, wenn

(a) $n_{\bullet 1} \cdot p_{1(1)}/p_{2(1)} = (a+c) \cdot a/c > 9$ und $n_{\bullet 1} \cdot p_{2(1)}/p_{1(1)} = (a+c) \cdot c/a > 9$,
(b) $n_{\bullet 2} \cdot p_{1(2)}/p_{2(2)} = (b+d) \cdot b/d > 9$ und $n_{\bullet 2} \cdot p_{2(2)}/p_{1(2)} = (b+d) \cdot d/b > 9$,
(c) $n_{\bullet 1} = a+c > 60$ und
(d) $n_{\bullet 2} = b + d > 60$.

Wird der Standardfehler unter der Annahme $\delta_{YX}\% = 0\%$ berechnet, ist die Annäherung an die Normalverteilung hinreichend genau, wenn

(a) $n \cdot p_{1\bullet}/p_{2\bullet} = n \cdot (a+b)/(c+d) > 9$ und $n \cdot p_{2\bullet}/p_{1\bullet} = n \cdot (c+d)/(a+b) > 9$
(b) $n > 60$
(c) $n_{\bullet 1} \cdot p_{1\bullet}/p_{2\bullet} = (a+c) \cdot (a+b)/(c+d) > 9$ und $n_{\bullet 1} \cdot p_{2\bullet}/p_{1\bullet} = (a+c) \cdot (c+d)/(a+b) > 9$
(d) $n_{\bullet 2} \cdot p_{1\bullet}/p_{2\bullet} = (b+d) \cdot (a+b)/(c+d) > 9$ und $n_{\bullet 2} \cdot p_{2\bullet}/p_{1\bullet} = (b+d) \cdot (c+d)/(a+b) > 9$

Anwendungsbeispiel:

Als Beispiel soll anhand der Daten der Stichprobe des Allbus 2006 geprüft werden, dass Frauen in den alten Bundesländern signifikant häufiger gegen ein Verbot von Schwangerschaftsabbrüchen sind als Männer.
Die Irrtumswahrscheinlichkeit soll 5% betragen.

Hypothesentests über Prozentsatzdifferenzen

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	a	344	b	403	747	a+b
- sollte verboten sein	c	700	d	701	1401	c+d
Summe	a+c	1044	b+d	1104	2148	n

Schritt 1: Formulierung von Null und Alternativhypothese

Aus der Forschungshypothese und der Anordnung der Variablen in der Tabelle folgt als zu testendes Hypothesenpaar:

$$H_0: \delta_{YX}\% \geq 0\% \text{ versus } H_1: \delta_{YX}\% < 0\%$$

*Die **Forschungshypothese** postuliert also eine **negative Prozentsatzdifferenz**, wobei die Forschungshypothese als Alternativhypothese spezifiziert wird. Zur Prüfung wird entsprechend ein einseitiger Test der Prozentsatzdifferenz nach unten durchgeführt.*

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung

Als Teststatistik wird die Prozentsatzdifferenz in der Stichprobe an der Trennstelle zwischen Null- und Alternativhypothese standardisiert. Die Teststatistik Z ist dann standardnormalverteilt, wenn die Nullhypothese gerade noch richtig ist.
Da nach der Nullhypothese $\delta_{YX}\% = 0\%$, wird der Standardfehler der Anteilsdifferenz über die Randverteilung der Zeilenvariable geschätzt.

Hypothesentests über Prozentsatzdifferenzen

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	a	344	b	403	747	a+b
- sollte verboten sein	c	700	d	701	1401	c+d
Summe	a+c	1044	b+d	1104	2148	n

	X		
Y	1	2	Σ
1	a	b	a+b
2	c	d	c+d
Σ	a+c	b+d	n

Unter der Annahme $\delta_{YX}\% = 0$ berechnet sich der Standardfehler für die Allbusdaten nach:

$$\hat{\sigma}\left(d_{YX}\%/100 \middle| \delta_{YX}\% = 0\right) = \sqrt{\frac{(a+b)\cdot(c+d)}{n^2}\cdot\left(\frac{1}{a+c}+\frac{1}{b+d}\right)}$$

$$= \sqrt{\frac{747\cdot 1401}{2148^2}\cdot\left(\frac{1}{1044}+\frac{1}{1104}\right)} = 0.0205$$

Zum Vergleich wird der Standardfehler der Anteilsdifferenz auch unterder Annahme ungleicher Antreile in der Popualtion berechnet.
Ist $\delta_{YX}\% \neq 0$ ergibt sich der Standardfehler für die Allbusdaten nach:

$$\hat{\sigma}\left(d_{YX}\%/100 \middle| \delta_{YX}\% \neq 0\right) = \sqrt{\frac{a\cdot c}{(a+c)^3}+\frac{b\cdot d}{(b+d)^3}} = \sqrt{\frac{344\cdot 700}{1044^3}+\frac{403\cdot 701}{1104^3}} = 0.0205$$

Aufgrund der hohen Fallzahlen und ähnlicher konditionaler Anteile in der Stichprobe sind die Standardfehler bei diesen Daten bis auf die vierte Nachkommastelle gleich.

Hypothesentests über Prozentsatzdifferenzen

Die Teststatstik berechnet sich für diesen Test nach:

$$Z = \frac{\frac{a}{a+c} - \frac{b}{b+d}}{\sqrt{\frac{(a+b)\cdot(c+d)}{n^2}\cdot\left(\frac{1}{a+c}+\frac{1}{b+d}\right)}}$$

	X		
Y	1	2	∑
1	a	b	a+b
2	c	d	c+d
∑	a+c	b+d	n

Ist die Prozentsatzdifferenz in der Population null, ist Z standardnormalverteilt. Bei einer positiven Prozentsatzdifferenz in der Population ist die Teststatistik mit einem Erwartungswert größer Null, bei negativer Prozentsatzdifferenz in der Population mit einem Erwartungswert kleiner Null normalverteilt.

Schritt 3: Festlegung von Irrtumswahrscheinlichkeit und kritischen Werten

Die Irrtumswahrscheinlichkeit wird auf die üblichen 5% gesetzt. Die Nullhypothese wird bei dem einseitigen Test nach unten abgelehnt, wenn die Teststatistik kleiner oder gleich dem 5%-Quantil der Standardnormalverteilung ist, also den Wert –1.645 erreicht oder unterschreitet.

Hypothesentests über Prozentsatzdifferenzen

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	a	344	b	403	747	a+b
- sollte verboten sein	c	700	d	701	1401	c+d
Summe	a+c	1044	b+d	1104	2148	n

Schritt 4: Berechnung der Teststatistik und Entscheidung

$$Z = \frac{\frac{a}{a+c} - \frac{b}{b+d}}{\sqrt{\frac{(a+b)\cdot(c+d)}{n^2}\cdot\left(\frac{1}{a+c}+\frac{1}{b+d}\right)}} = \frac{\frac{344}{1044} - \frac{403}{1104}}{\sqrt{\frac{747\cdot 1401}{2148^2}\cdot\left(\frac{1}{1044}+\frac{1}{1104}\right)}} = -1.7$$

Da Z= –1.7 < –1.645, ist die Nullhypothese zu verwerfen.
Bei einer Irrtumswahrscheinlichkeit von 5% ist davon auszugehen, dass Frauen eher als Männer gegen ein Verbot von Schwangerschaftsabbrüchen nach dem Willen der Frau sind.

Hinweis:

Da das 95%-Konfidenzintervall den Wert 0.0 enthält, würde bei einen zweiseitigen Test der Nullhypothese einer Prozenzsatzdifferenz von Null das Ergebnis nicht signifikant sein!

Generell gilt, dass einseitige Hypothesentests eher zur Ablehnung der Nullhypothese führen als zweiseitige Tests!

Hypothesentests über Prozentsatzdifferenzen

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich		weiblich		Summe	
- sollte erlaubt sein	a	344	b	403	747	a+b
- sollte verboten sein	c	700	d	701	1401	c+d
Summe	a+c	1044	b+d	1104	2148	n

	X		
Y	1	2	∑
1	a	b	a+b
2	c	d	c+d
∑	a+c	b+d	n

Schritt 5: Prüfung der Anwendungsvoraussetzungen

Die Anwendungsvoraussetzungen sind erfüllt:

(a) 2148 ·747/1401 = 1145 > 9 & 2148 > 60

(b) 1044 ·747/1401 = 556.7 > 9 & 1104 ·747/1401 =588.6 > 9

Lerneinheit 16: Symmetrische und asymmetrische Beziehungen

Haltung zum Schwangerschaftsabbruch nach Geschlecht

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	33.0% a (344)	36.5% b (403)	34.8% (747)
- sollte verboten sein	67.0% c (700)	63.5% d (701)	65.2% (1401)
Summe	100% (1044)	100% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

$d_{YX}\% = -3.5$ Prozentpunkte

Bei der Betrachtung des Zusammenhangs zwischen der Haltung zur Erlaubnis oder Verbot eines Schwangerschaftsabbruchs bei finanzieller Notlage und dem Geschlecht wurde Geschlecht als bedingende und die Haltung zum Schwangerschaftsabbruch als bedingte Variable betrachtet.
Formal möglich ist auch, das Geschlecht als bedingte und die Haltung zum Schwangerschaftsabbruch als bedingende Variable zu betrachten:

Geschlecht nach Haltung zum Schwangerschaftsabbruch

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	46.1% a (344)	53.9% b (403)	100.0% (747)
- sollte verboten sein	50.0% c (700)	50.0% d (701)	100.0% (1401)
Summe	48.6% (1044)	51.4% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

$d_{XY}\% = 46.1 - 50.0 = -3.9$ Prozentpunkte

Asymmetrische Beziehungen

Geschlecht nach Haltung zum Schwangerschaftsabbruch

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	46.1% a (344)	53.9% b (403)	100.0% (747)
- sollte verboten sein	50.0% c (700)	50.0% d (701)	100.0% (1401)
Summe	48.6% (1044)	51.4% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

$d_{XY}\% = 46.1 - 50.0 = -3.9$ Prozentpunkte

Aufgrund der Vertauschung von bedingender und bedingter Variable ändert sich die Interpretation:
Während unter denen, die für die Erlaubnis eines Schwangerschaftsabbruchs sind, 46.1% Männer sind, sind unter denen, die für ein Verbot eintreten, 50.0% Männer.
Die Prozentsatzdifferenz beträgt 3.9 Prozentpunkte.

Was bedingende und was bedingte Verteilung ist, ist in der Regel eine Frage der Zielsetzung der Analyse:
Wird der Zusammenhang im Sinne einer ***kausalen Beziehung*** interpretiert, ist die bedingende Variable die Ursachenvariable und die bedingte Variable die kausal abhängige Effektvariable.
Generell wird bei ***asymmetrischen Beziehungen*** unterschieden zwischen der ***abhängigen Variablen***, deren Verteilung in Abhängigkeit von der ***unabhängigen*** oder ***erklärenden Variablen*** betrachtet wird.

Asymmetrische Beziehungen

Haltung zum Schwangerschaftsabbruch nach Geschlecht

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	33.0% a (344)	36.5% b (403)	34.8% (747)
- sollte verboten sein	67.0% c (700)	63.5% d (701)	65.2% (1401)
Summe	100% (1044)	100% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

$d_{YX}\% = -3.5$ Prozentpunkte

Da das Geschlecht nicht durch die Haltung zum Schwangerschaftsabbruch kausal beeinflusst werden kann, liegt es nahe, Geschlecht als erklärende und die Haltung als unabhängige Variable aufzufassen.

Tatsächlich sind Kreuztabellen in der Regel sehr oft so aufgebaut, dass die Spaltenvariable die erklärende Variable und die Zeilenvariable die abhängige Variable ist.
Als eine weitere Konvention wird in statistischen Formeln in der Regel die unabhängige Variable als X und die abhängige Variable als Y bezeichnet.

“Erklärend” bedeutet jedoch nicht immer “kausal verursachend”. So dürfte auch im Beispiel nicht das biologische Geschlecht eine spezifische Haltung zum Schwangerschaftsabbruch auslösen. Vielmehr dürften eher unterschiedliche soziale Lagen (Situationen) oder unterschiedliche Wertorientierungen zwischen den *sozialen Geschlechtern* den (geringen und praktisch zu vernachlässenden Effekt) verursacht haben.

Asymmetrische Beziehungen

Haltung zum Schwangerschaftsabbruch nach Geschlecht

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	33.0% a (344)	36.5% b (403)	34.8% (747)
- sollte verboten sein	67.0% c (700)	63.5% d (701)	65.2% (1401)
Summe	100% (1044)	100% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

$d_{YX}\% = -3.5$ Prozentpunkte

Darüber hinaus kann es auch rein praktische Gründe haben, eine Variable als abhängige und die andere als erklärende Variable aufzufassen.
Ein Grund liegt oft darin, dass die Ausprägung einer Variable eher bekannt oder leichter zu messen ist und dies genutzt wird, um - ohne jede kausale Interpretation - die Ausprägung der anderen Variable vorherzusagen.

Die ***prognostizierende*** Variable ist dann ***unabhängig***e Variable, und die Variable, deren Werte ***prognostiziert*** werden, die ***abhängige*** Variable.
An diese Betrachtungsweise erinnert die alternative Bezeichnung ***Prädiktorvariable*** für die unabhängigeVariable und ***Kriteriumsvariable*** für die abhängige Variable.

Asymmetrische Beziehungen

Geschlecht nach Haltung zum Schwangerschaftsabbruch

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	46.1% a (344)	53.9% b (403)	100.0% (747)
- sollte verboten sein	50.0% c (700)	50.0% d (701)	100.0% (1401)
Summe	48.6% (1044)	51.4% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

$d_{XY}\% = 46.1 - 50.0 = -3.9$ Prozentpunkte

Wenn die Spaltenvariable die abhängige Variable und die Zeilenvariable die erklärende Variable ist, müssen bei allen Berechnungsformeln jeweils Spalten- und Zeilenindizes vertauscht werden.

So berechnet sich die Prozentsatzdifferenz $d_{XY}\%$ nach:

$$d_{XY}\% = 100 \cdot \left(\frac{a}{a+b} - \frac{c}{c+d} \right)$$

im Beispiel: $$d_{XY}\% = 100 \cdot \left(\frac{344}{747} - \frac{700}{1401} \right) = -3.9$$

Auf zusätzliche Formeln für Standardfehler, Konfidenzintervalle und Teststatistiken kann verzichtet werden, da auch einfach die Position der Variablen in der Kreuztabelle vertauscht werden kann, im Beispiel also Geschlecht zur Zeilen- und die Haltung zum Schwangerschaftsabbruch Spaltenvariable wird.

Asymmetrische Beziehungen

Geschlecht nach Haltung zum Schwangerschaftsabbruch

Schwangerschaftsabbruch nach Willen der Frau (Y)	Geschlecht des Befragten (X) männlich	weiblich	Summe
- sollte erlaubt sein	46.1% a (344)	53.9% b (403)	100.0% (747)
- sollte verboten sein	50.0% c (700)	50.0% d (701)	100.0% (1401)
Summe	48.6% (1044)	51.4% (1104)	100.0% (2148)

(Quelle: Allbus 2006, nur Westen)

$$d_{XY}\% = 100 \cdot \left(\frac{a}{a+b} - \frac{c}{c+d} \right)$$

$d_{XY}\% = 46.1 - 50.0 = -3.9$ Prozentpunkte

Spaltenvariable als unabhängige Variable:

Spaltenvariable

Zeilenvariable: Geschlecht des Befragten	Schwangerschaftsabbruch sollte erlaubt sein	verboten sein	Summe
- männlich	46.1% a (344)	50.0% b (700)	48.6% (1044)
- weiblich	53.9% c (403)	50.1% d (701)	51.4% (1104)
Summe	100.0% (747)	100.0% (1401)	100.0% (2148)

(Quelle: Allbus 1996)

$$d_{YX}\% = 100 \cdot \left(\frac{a}{a+c} - \frac{b}{b+d} \right)$$

$d_{YX}\% = 46.1 - 50.0 = -3.9$ Prozentpunkte

Symmetrische Beziehungen

Im Unterschied zu einer asymmetrischen Beziehung wird bei einer ***symmetrischen Beziehung*** gar nicht zwischen abhängiger und unabhängiger Variable unterschieden.

Grund kann z.B. sein, dass sich die beiden Variablen gegenseitig beeinflussen, oder dass weder eine Kausalbeziehung unterstellt wird, noch eine Prognose einer Variablen durch die andere angestrebt wird.

So kann z.B. vermutet werden, dass die Beurteilung der eigenen wirtschaftlichen Lage (EWL) die Beurteilung der allgemeinen Wirtschaftlage im Staat (AWL) beeinflusst, aber umgekehrt auch die Beurteilung der eigenen Lage durch die (medienvermittelte) Beurteilung der allgemeinen wirtschaftlichen Lage beeinflusst wird.

Ausgangspunkt der Analyse kann in dieser Situation die Betrachtung der auf die ***Gesamttabelle bezogenen relativen Häufigkeiten*** p_{ij} bzw. der korrespondierenden Prozentwerte sein:

$$p_{ij} = \frac{n_{ij}}{n}$$

Auf die Gesamtfallzahl bezogenen relativen Häufigkeiten in Prozent

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Symmetrische Beziehungen

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Da nach dem Multiplikationssatz der Wahrscheinlichkeitstheorie statistische Unabhängigkeit bedeutet, dass die gemeinsame Auftretenswahrscheinlichkeit gleich dem Produkt der Ausgangswahrscheinlichkeiten ist, können analog zur asymmetrischen Betrachtung auch bei symmetrischer Betrachtung die bei Unabhängigkeit erwarteten relativen und absoluten Häufigkeiten berechnet werden:

$$\hat{\pi}_{ij} = p_{i\cdot} \cdot p_{\cdot j} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} = \text{bzw. } e_{ij} = n \cdot \hat{\pi}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

In den Formeln steht $\hat{\pi}_{ij}$ für die bei statistischer Unabhängigkeit erwarteten relativen Häufigkeiten (geschätzten Populationsanteile) und e_{ij} für die bei Unabhängigkeit erwarteten absoluten Häufigkeiten.

Für die erste Zelle (a) berechnet sich so die bei Unabhängigkeit erwarteten Häufigkeit als:

$$\hat{\pi}_{11} = 0.434 \cdot 0.135 = \frac{990 \cdot 307}{2282^2} = 0.058 \text{ bzw. } e_{11} = 2282 \cdot 0.058 = \frac{990 \cdot 307}{2282} = 133.2$$

Symmetrische Beziehungen

Beobachtete erwartete auf die Gesamtfallzahl bezogenen relative Häufigkeiten in Prozent:

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Bei Unabhängigkeit erwartete auf die Gesamtfallzahl bezogenen relative Häufigkeiten in Prozent:

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	5.8% (133.2)	37.5% (856.8)	43.4% (990)
- nicht gut	7.6% (173.8)	49.0% (1118.2)	56.6% (1292)
Summe	13.5% (307.0)	86.5% (1975.0)	100.0% (2282)

$$\hat{\pi}_{ij} = p_{i\cdot} \cdot p_{\cdot j} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} \text{ bzw. } e_{ij} = n \cdot \hat{\pi}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

Die bei Unabhängigkeit erwarteten absoluten Häufigkeiten sind identisch mit denen, die sich ergeben würden, wenn AWL oder wenn EWL erklärende Variable wäre.
Bei statistischer Unabhängigkeit unterscheiden sich nämlich die erwarteten Häufigkeiten bei asymmetrischen und bei symmetrischen Beziehungen nicht.

Pearsons Chiquadrat-Statistik

Residuen

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	3.9% (88.8)	–3.9% (–88.8)	0.0% (0)
- nicht gut	–3.9% (–88.8)	3.9% (88.8)	0.0% (0)
Summe	0.0% (0)	0.0% (0)	(n = 2282)

Die Differenzen aus den tatsächlichen und den bei Unabhängigkeit erwarteten (relativen) Häufigkeiten werden als Residuen bezeichnet:

$$r_{ij} = n_{ij} - e_{ij} \text{ bzw. } r_{ij}\% = 100 \cdot \frac{n_{ij} - e_{ij}}{n} = 100 \cdot \left(p_{ij} - \hat{\pi}_{ij}\right)$$

Es gibt mehr Befragte, die sowohl die eigene wie die allgemeine Lage für gut oder aber für nicht gut halten, als bei Unabhängigkeit der beiden Variablen zu erwarten wären. Umgekehrt gibt es weniger Personen als bei Unabhängigkeit erwartet, die die eigene Lage für gut und die allgemeine Lage für nicht gut bzw. die allgemeine Lage für gut und die eigene Lage für nicht gut halten.

Nach dem Statistiker Pearson ist eine Statistik benannt, die alle Abweichungen zwischen beobachteten und erwarteten Häufigkeiten in einer I×J-Tabelle zusammenfasst und als ***Pearsons Chiquadrat-Statistik*** bezeichnet wird:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

Chi-Quadrat

Beobachtete erwartete auf die Gesamtfallzahl bezogenen relative Häufigkeiten in Prozent:

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Bei Unabhängigkeit erwartete auf die Gesamtfallzahl bezogenen relative Häufigkeiten in Prozent:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij}-e_{ij}\right)^2}{e_{ij}}$$

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	5.8% (133.2)	37.5% (856.8)	43.4% (990)
- nicht gut	7.6% (173.8)	49.0% (1118.2)	56.6% (1292)
Summe	13.5% (307.0)	86.5% (1975.0)	100.0% (2282)

i	j	n_{ij}	e_{ij}	$(n_{ij}-e_{ij})^2/e_{ij}$
1	1	222	133.2	59.200
1	2	768	856.8	9.203
2	1	85	173.8	45.371
2	2	1207	1118.2	7.052
∑		2283	2283	120.826

$$\chi^2 = \frac{(222-133.2)^2}{133.2} + \frac{(768-856.8)^2}{856.8} + \frac{(85-173.8)^2}{173.8} + \frac{(1207-1118.2)^2}{1118.2} = 120.826$$

Phi-Quadrat

Beobachtete Häufigkeiten:

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Erwartete Häufigkeiten:

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	5.8% (133.2)	37.5% (856.8)	43.4% (990)
- nicht gut	7.6% (173.8)	49.0% (1118.2)	56.6% (1292)
Summe	13.5% (307.0)	86.5% (1975.0)	100.0% (2282)

Werden statt der absoluten, die relativen Häufigkeiten bei der Berechnung herangezogen, ergibt sich der Kennwert Φ^2 (Phi-Quadrat):

i	j	p_{ij}	e_{ij}/n	$n^{-1}(n_{ij}-e_{ij})^2/e_{ij}$
1	1	.097	.058	.0262
1	2	.337	.375	.0041
2	1	.037	.076	.0200
2	2	.529	.490	.0031
∑		1.000	0.999	0.0534

$$\Phi^2 = \frac{(.097-.058)^2}{.058} + \frac{(337-.375)^2}{.375} + \frac{(.037-.076)^2}{.076} + \frac{(.529-.490)^2}{.490} = .0534 \approx 120.826/2282$$

Phi-Quadrat und Phi

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut		Allgemeine Wirtschaftslage nicht gut		Summe	
- gut	a	9.7% (222)	b	33.7% (768)	43.4% (990)	a+b
- nicht gut	c	3.7% (85)	d	52.9% (1207)	56.6% (1292)	c+d
Summe	a+c	13.5% (307)	b+d	86.5% (1975)	100.0% (2282)	n

(Quelle: Allbus 2006, nur Westen)

Die Formeln zur Berechnung von χ^2 und Φ^2 gelten für Tabellen beliebiger Größe.
Nur in der Vierfeldertabelle gelten alternative Berechnungsformeln:

$$\Phi^2 = \frac{(a \cdot d - b \cdot c)^2}{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)} \quad \text{bzw.} \quad \chi^2 = n \cdot \frac{(a \cdot d - b \cdot c)^2}{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}$$

Bei den Beispieldaten ergeben sich:

$\Phi^2 = (222 \cdot 1207 - 768 \cdot 85)^2 / (990 \cdot 1292 \cdot 307 \cdot 1975) = 0.0530$

$\chi^2 = 2282 \cdot (222 \cdot 1207 - 768 \cdot 85)^2 / (990 \cdot 1292 \cdot 307 \cdot 1975) = 120.867$

Abweichungen bei den alternativen Berechnunggen ergeben sich dadurch, dass bei der ersten Berechnung über die allgemeine Formel von Chi-Quadrat bzw. Phi-Quadrat nur mit einer Nachkommastelle gerechnet wurde.

Phi-Quadrat und Phi

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	Allgemeine Wirtschaftslage nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Der Wertebereich von Φ^2 liegt in einer Vierfeldertabelle zwischen 0 und 1, wobei 0 bei statistischer Unabhängigkeit und 1 bei einem perfekten Zusammenhang erreicht wird. Φ^2 kann daher als ein Maß für die Stärke eines symmetrischen Zusammenhangs genutzt werden.

Anstelle von Φ^2 wird jedoch meisten dessen Quadratwurzel Φ ***(Phi)*** verwendet:

$$\Phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}}$$

Im Beispiel beträgt $\Phi = +0.230$

Ein Vorteil von Φ ist, dass sein Wertebereich von –1 bis +1 läuft, so dass (ab ordinalem Messniveau) zwischen positiven und negativen Beziehungen unterschieden werden kann.

Darüber hinaus kann Φ auch als geometrisches Mittel der beiden asymmetrischen Anteilsdifferenzen in einer Vierfeldertabelle definiert werden:

$$\Phi = \sqrt{\frac{d_{YX}\%}{100} \cdot \frac{d_{XY}\%}{100}}$$

Phi-Quadrat und Phi

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Im Beispiel ergibt sich:

$$d_{YX}\% / 100 = 222/307 - 768/1975 = 0.334$$
$$d_{XY}\% / 100 = 222/90 - 85/1292 = 0.158$$
$$\Phi = (0.334 \cdot 0.158)^{0.5} = 0.230$$

Da die Vorzeichen der Prozentsatz- bzw. Anteilsdifferenzen sich bei Vertauschung von abhängiger und unabhängiger Variable nicht ändern, hat das Produkt der beiden Werte stets ein positives Vorzeichen. Bei der Berechnung von Φ über das geometrische Mittel der Anteilsdifferenzen kann das Vorzeichen der Ausgangswerte übernommen werden.

Stärke eines Zusammenhangs	
praktisch kein	$0.00 \le \lvert \Phi \rvert < 0.05$
geringer	$0.05 \le \lvert \Phi \rvert < 0.20$
mittlerer	$0.20 \le \lvert \Phi \rvert < 0.50$
starker	$0.50 \le \lvert \Phi \rvert$
Perfekter Zus.	$1.00 = \lvert \Phi \rvert$

Die Interpretation des Wertes von Φ erfolgt analog zur Prozentsatzdifferenz.

Der Wert von 0.230 weist somit auf einen mittelstarken Zusammenhang zwischen der Beurteilung der eigenen und der allgemeinen wirtschaftlichen Lage hin.

Yules Q

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Eine Alternative zu Φ ist das nach dem Statistiker Yules benannte und sehr leicht zu berechnende Zusammanhangsmaß ***Yules Q***:

$$Q = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}$$

Für die Beispieldaten beträgt

$$Q = (222 \cdot 1207 - 768 \cdot 85) / (222 \cdot 1207 + 768 \cdot 85) = 0.608.$$

Obwohl Q in der Regel deutlich höhere Werte aufweist als Φ, ist der Wertebereich gleich.
Bei einem perfekten negativen Zusammenhang ist der Wert –1, bei Unabhängigkeit 0 und bei perfektem positiven Zusammenhang +1.
Da Q auch bei einem nur geringen Zusammenhang relativ große Werte annehmen kann, hat sich in der Sozialforschung Φ als Zusammenhangsmaß für symmetrische Beziehungen in einer Vierfeldertabelle durchgesetzt.

Chiquadrat-Test auf statistische Unabhängigkeit

Ähnlich wie bei der Prozentsatzdifferenz lassen sich auch für Φ und Q asymptotisch gültige Standardfehler berechnen. Die Berechnungsformeln sind allerdings komplex, so dass sie i.a nur durch Statistikprogramme berechnet werden. Es gibt auch keine akzeptierten Faustregeln, unter welchen Bedingungen eine asymptotische Annäherung an die Normalverteilung hinreichend ge-nau ist.

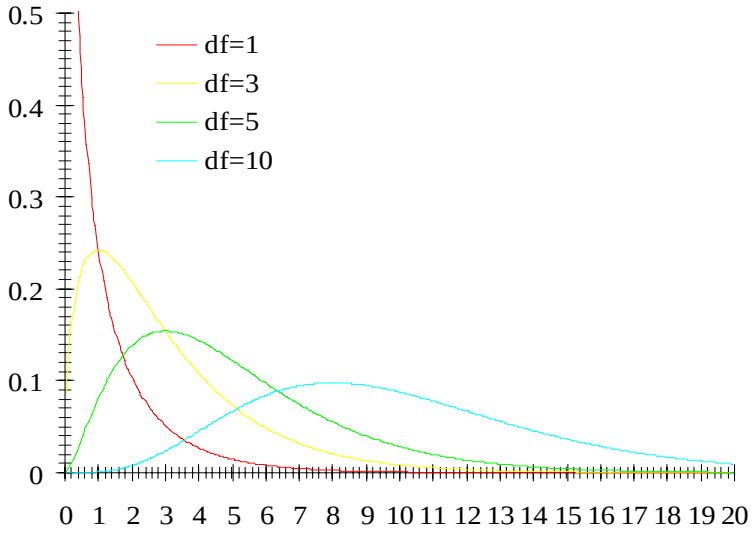
Leichter zu berechnen sind Tests der Nullhypothese, dass der Wert von Φ in der Population gleich (bzw. kleiner/gleich oder größer/gleich) Null ist gegen die Alternativhypothese, dass Φ in der Population ungleich (bzw. größer oder kleiner) Null sind.
Da in den Formeln von Φ und Q die Zähler identisch sind, gilt der Test gleichzeitig auch zur Prüfung der entsprechenden Nullhypothese für Yules Q.

Als Teststatistik wird bei beiden Zusammenhangsmaßen Pearsons Chiquadrat-Statistik herangezogen, die bei statistischer Unabhängigkeit von Zeilen- und Spaltenvariable in der Population (zentral) chiquadratverteilt ist.
Die Chiquadrat-Verteilung ist wie die T-Verteilung eine Verteilungsfamilie, wobei sich die einzelnen Verteilungen entsprechend ihrer Freiheitsgrade unterscheiden.

Es lässt sich zeigen, dass die Summe der Quadrate von k statistisch unabhängigen Standardnormalverteilungen einer Chiquadrat-Verteilung mit df = k Freiheitsgraden folgt:

$$f\left(\sum_{i=1}^{k} z_i^2 \,\middle|\, z_i \sim N(0;1)\right) = \chi^2_{df=k}$$

Chiquadratverteilung



Die Abbildung zeigt die Dichten von Chiquadrat-Verteilungen mit 1, 3, 5 und 10 Freiheitsgraden. Erkennbar ist, dass Chiquadrat-Verteilungen rechtsschief sind, wobei die Schiefe mit steigender Zahl an Freiheitsgraden abnimmt

Erwartungswert, Varianz und Schiefe sind Funktionen der Freiheitsgrade:

$$\mu\left(\chi^2_{df}\right) = df \;;\; \sigma^2\left(\chi^2_{df}\right) = 2 \cdot df \;;\; \frac{\mu_3}{\sigma^3} = \sqrt{8} \cdot \frac{1}{\sqrt{df}}$$

Chiquadratverteilung

Quantile von χ^2:

α	90%	95%	99%	99.9%
df=1	2.706	3.841	6.635	10.83
df=2	4.605	5.991	9.210	13.82
df=3	6.251	7.815	11.34	16.27
df=4	7.779	9.488	13.28	18.47
df=5	9.236	11.07	15.08	20.51
df=6	10.64	12.59	16.81	22.46
df=7	12.03	14.07	18.48	24.32
df=8	13.36	15.51	20.09	26.12
df=9	14.68	16.92	21.67	27.88
´df=10	15.99	18.31	23.21	29.59

Quantile von χ^2:

α	90%	95%	99%	99.9%
df=11	17.28	19.68	24.73	31.26
df=12	18.55	21.03	26.22	32.91
df=13	19.81	22.36	27.69	34.53
df=14	21.06	23.68	29.14	36.12
df=15	22.31	25.00	30.58	37.70
df=16	23.54	26.30	32.00	39.25
df=17	24.77	27.59	33.41	40.79
df=18	25.99	28.87	34.81	42.31
df=19	27.20	30.14	36.19	43.82
df=20	28.41	31.41	37.57	45.31

Quantile von χ^2:

α	90%	95%	99%	99.9%
df=21	29.62	32.67	38.93	46.80
df=22	30.81	33.92	40.29	48.27
df=23	32.01	35.17	41.64	49.73
df=24	33.20	36.42	42.98	51.18
df=25	34.38	37.65	44.31	52.62
df=26	35.56	38.89	45.64	54.05
df=27	36.74	40.11	46.96	55.48
df=28	37.92	41.34	48.28	56.89
df=29	39.09	42.56	49.59	58.30
df=30	40.26	43.77	50.89	59.70

Wie bei der T-Verteilung liegen auch die für statistische Tests am häufigsten genutzten Quantile der Chiquadrat-Verteilung tabelliert vor.

Aus dem zentralen Grenzwertsatz folgt, dass sich die Chiquadrat-Verteilung bei steigender Zahl von Freiheitsgraden asymptotisch einer Normalverteilung mit $\mu = df$ und $\sigma^2 = 2 \cdot df$ annähert. Die Annäherung verläuft langsam, ist aber ab etwa 30 Freiheitsgraden bei Nutzung der folgenden Näherungsformel für praktische Zwecke hinreichend genau:

$$\chi^2_{\alpha;df} \approx 0.5 \cdot \left(z_\alpha + \sqrt{2 \cdot df - 1} \right)^2$$

$$\text{z.B.: } \chi^2_{\alpha=0.95;df=20} = 31.41 \approx 0.5 \cdot \left(z_{0.95} + \sqrt{2 \cdot 20 - 1} \right)^2 = 0.5 \cdot \left(1.645 + \sqrt{39} \right)^2 = 31.13$$

Chiquadratverteilung

Da die Summe von k quadrierten statistisch unabhängiger Standardnormalverteilungen mit k Freiheitsgraden chiquadratverteilt ist, folgt:

- die α-Quantile der Chiquadratverteilung mit df=1 Freiheitsgrade sind gleich dem Quadrat des α/2-Quantils bzw. des (1–α/2)-Quantils der Standardnormalverteilung.
  z.B. $\chi^2_{\alpha=0.95;df=1} = 3.84 = 1.96^2 = (z_{0.975})^2 = (-1.96)^2 = (z_{0.025})^2$.
- Die Summe statistisch unabhängiger Chiquadrat-Verteilungen mit df=k und df=n Freiheitsgraden ist chiquadratverteilt mit df=k+n Freiheitsgraden.

Interessanterweise besteht auch eine Beziehung zwischer der Chiquadrat-Verteilung und der diskreten Poisson-Verteilung. Für die Verteilungsfunktionen der beiden Verteilungen gilt nämlich:

$$F\left(X = x \middle| X \sim P(X;\lambda) \right) = 1 - F\left(Y = 2 \cdot \lambda \middle| Y \sim \chi^2_{df=2\cdot(1+x)} \right)$$

Für den Chiquadrat-Test auf statistische Unabhängigkeit wird allerdings eine andere Beziehung zur Poisson-Verteilung genutzt:
Aufgrund der Verwandtschaft der Poisson-Verteilung mit der Binomialverteilung ist es nämlich möglich, in einfachen Zufallsauswahlen die absolute Auftretenshäufigkeiten in den Zellen der Kreuztabelle als (voneinander statistisch unabhängige) Realisationen von Poisson-Verteilungen aufzufassen, wobei der Parameter der Poisson-Verteilung in einer Zelle $\lambda = n \cdot \pi_{ij}$ ist, mit n gleich der Fallzahl der Stichprobe insgesamt und π_{ij} gleich dem Populationsanteil der Ausprägungskombination i und j der Zeilen- und Spaltenvariable, falls n als Realisierung einer Zufallsvariable interpretiert werden kann.

Chiquadrat-Test auf statistische Unabhängigkeit

Da auch die Poisson-Verteilung nach dem zentralen Grenzwertsatz asymptotisch normalverteilt ist, folgt für jede einzelne Zellenhäufigkeit:

$$\lim_{n\to\infty}\left(\frac{f_{ij}-n\cdot\pi_{ij}}{\sqrt{n\cdot\pi_{ij}}}\right)=N(0:1) \text{ bzw. } \lim_{n\to\infty}\left(\frac{\left(f_{ij}-n\cdot\pi_{ij}\right)^2}{n\cdot\pi_{ij}}\right)=\chi^2_{df=1}$$

Die asymptotische Annäherung gilt auch, wenn für π_{ij} eine konsistente Schätzung des Populationsanteils verwendet wird und entsprechend statt $n\cdot\pi_{ij}$ die erwarteten absoluten Häufigkeiten. Pearsons Chiquadrat-Statistik kann daher auch als Summe standardisierter (z-transformierter) und quadrierter Realisationen von Poisson-Verteilungen aufgefasst werden und ist somit asymptotisch chiquadratverteilt, wenn die bei der Berechnung verwendeten erwarteten Häufigkeiten konsistente Schätzer der λ-Parameter der Poisson-Verteilungen in den Zellen der Tabelle sind.

Beim Chiquadrat-Test auf statistische Unabhängigkeit werden die λ-Parameter der Zellenhäufigkeiten aus dem Produkt aus der Fallzahl und den relativen Häufigkeiten der beiden Randverteilungen geschätzt. Wenn in der Population tatsächlich statistische Unabhängigkeit zwischen Zeilen- und Spaltenvarible besteht, dann handelt es sich um konsistente Schätzungen der λ-Parameter. Pearsons Chiquadrat-Statistik ist dann asymptotisch chiquadratverteilt.

Die Schätzung der λ-Parameter hat Auswirkungen auf die Zahl der Freiheitsgrade. Würde man die λ-Werte kennen und nicht schätzen, hätte Pearsons Chqiadrat-Statistik genau so viele Freiheitsgrade wie die Tabelle Zellen aufweist, in der Vierfeldertabelle also auch vier Freiheitsgrade.

Chiquadrat-Test auf statistische Unabhängigkeit

Durch die Schätzung der λ-Parameter verliert man jedoch Freiheitsgrade.Die tatsächliche Zahl der Freiheitsgrade ergibt sich dadurch, dass von der Zahl der (inneren) Tabellenzellen die Zahl der für die Schätzung verwendeten Informationen abgezogen wird.

In der Vierfeldertabelle werden für die Schätzung der bei Unabhängigkeit erwarteten Häufigkeiten drei empirische Informationen verwendet:
- die Fallzahl n der Stichprobe insgesamt,
- die Schätzung $p_{1\bullet}$ für den Populationsanteil $\pi_{1\bullet}$ der Zeilenvariable und
- die Schätzung $p_{\bullet 1}$ für den Populationsanteil $\pi_{\bullet 1}$ der Spaltenvariable.

Da sich die Summen der Spalten- und Zeilenanteile jeweils zu 1.0 summieren, ergeben sich die Schätzungen der Komplementäranteile $\pi_{2\bullet}$ und $\pi_{\bullet 2}$ als $(1-p_{1\bullet})$ und $(1-p_{\bullet 1})$.

Da drei Stichprobeninformationen für die Schätzung der erwarteten Häufigkeiten genutzt werden, ist Pearsons Chiquadrat-Statistik in der Vierfeldertabelle asymptotisch chiquadratverteilt mit df = 4 –3 = 1 Freiheitsgrad, wenn in der Population die Zeilen- und die Spaltenvariable statistisch unabhängig voneinander sind.

Der Verwendung des Wortes „Freiheitsgrad“ für den Parameter der Chiquadrat-Verteilung lässt sich anhand der Vierfeldertabelle nachvollziehen. Der Test auf statistische Unabhängigkeit in der Vierfeldertabelle hat df=1 Freiheitsgrad. Tatsächlich reicht bei vorgegebener Randverteilung eine einzige Zellenhäufigkeit aus, um die übrigen drei und damit alle vier Zellenwerte eindeutig festzulegen.

Chiquadrat-Test auf statistische Unabhängigkeit

	X		
Y	1	2	∑
1	a	b	a+b
2	c	d	c+d
∑	a+c	b+d	n

	X		
Y	1	2	∑
1	a	b	990
2	c	d	c+d
∑	307	b+d	2282

	X		
Y	1	2	∑
1	200	790	990
2	107	1185	1292
∑	307	1975	2282

So kann z.B. die Fallzahl n=2282 betragen und die jeweils erste Ausprägung der Randverteilungen $n_{1\cdot} = 307$ und $n_{\cdot 1} = 990$.

Ist nun z.B. $n_{11} = 200$, dann folgt für die übrigen Häufigkeiten $n_{12} = 790$, $n_{21} = 107$ und $n_{22} = 1185$.

Pearsons Chiquadrat-Statistik ist nur dann asymptotisch chiquadratverteilt, wenn in der Population die Unabhängigkeitsannahme von Zeilen- und Spaltenvariable zutrifft und die erwarteten Häufigkeiten dann konsistenten Schätzer der λ-Parameter sind.
Aber auch wenn dies nicht der Fall ist, lässt sich die Verteilung der Statistik angeben. Bei einfachen Zufallsauswahlen ist die Chiquadrat-Statistik dann nichtzentral chiquadratverteilt. Die nichtzentrale Chiquadrat-Verteilung hat eine ähnliche Form wie die (zentrale) Chiquadrat-Verteilung, jedoch einen Parameter mehr, den Nichtzentralitätsparameter ν, der bei der zentralen Chiquadrat-Verteilung Null ist. Ihr Erwartungswert ist gleich df + ν und ihre Varianz ist gleich 2·(df+2·ν).
Trifft die Unabhängigkeitsannahme nicht zu, ist ν eine Funktion der Fallzahl und der quadrierten Abweichungen der bei Unabhängigkeit geschätzten Populationsanteile von den tatsächlichen Populationsanteilen π_{ij}.

Chiquadrat-Test auf statistische Unabhängigkeit

Diese Eigenschaften von Pearsons Chiquadrat-Statistik können für Tests genutzt werden.
Die Vorgehensweise beim Chiquadratest folgt dabei der generellen Vorgehensweise beim statistischen Hypothesentesten.

Schritt 1: Formulierung von Null- und Alternativhypothese

Getestet wird bei Pearsons Chiquadratest, dass bei einer einfachen Zufallsauswahl Zeilen- und Spaltenvariable in der Grundgesamtheit statistisch unabhängig voneinander sind:

$$H_0: \pi_{ij} = \pi_{i\cdot} \cdot \pi_{\cdot j} \text{ für alle i, j versus } H_1: \pi_{ij} \neq \pi_{i\cdot} \cdot \pi_{\cdot j} \text{ für mindestens ein i,j}$$

π_{ij} ist die relative Häufigkeit der Ausprägungskombination der i-ten Ausprägung der Zeilen- und der j-ten Ausprägung der Spaltenvariablen in der Population.
Da bei statistischer Unabhängigkeit alle Zusammenhangsmaße null sind, können alternative Hypothesenpaare formuliert werden:

$$H_0: \Phi = 0 \text{ bzw. } H_0: Q = 0 \text{ bzw. } H_0: \delta_{YX}\% = 0 \text{ bzw. } H_0: \delta_{XY}\% = 0$$
$$\text{versus } H_1: \Phi \neq 0 \text{ bzw. } H_1: Q \neq 0 \text{ bzw. } H_1: \delta_{YX}\% \neq 0 \text{ bzw. } H_1: \delta_{XY}\% \neq 0.$$

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung

Als Teststatistik wird Pearsons Chiquadrat-Statistik berechnet:

$$\chi^2 = n \cdot \frac{(a \cdot d - b \cdot c)^2}{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}$$

	X		
Y	1	2	∑
1	a	b	a+b
2	c	d	c+d
∑	a+c	b+d	n

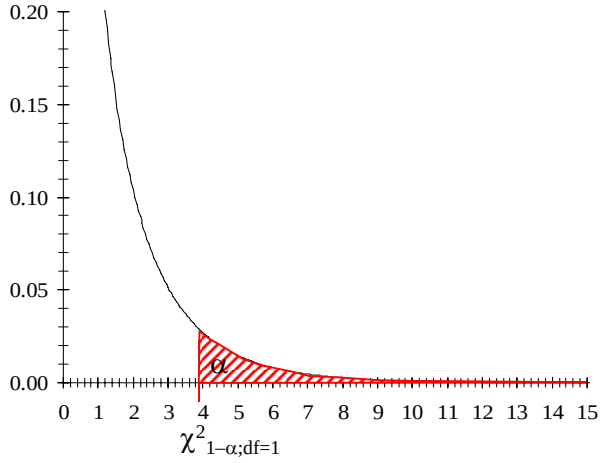
Chiquadrat-Test auf statistische Unabhängigkeit

Bei gültiger Nullhypothese ist die Teststatistik asymptotisch chiquadratverteilt mit df=1 Freiheitsgraden.

Wenn die Nulhypothese nicht zutrifft, also keine statistische Unabhängigkeit zwischen Zeilen- und Spaltenvariable besteht, dann ist die Teststatistik nichtzentral chiquadratverteilt. Da eine nichtzentrale Chiquadratverteilung einen größeren Erwartungswert hat als eine zentrale Chiquadratverteilung, ist bei unzutreffender Nullhypothese mit größeren Werten der Teststatistik als bei Gültigkeit der Nullhypothese zu rechnen.

Schritt 3: Festlegung von Irrtumswahrscheinlichkeit und kritischen Werten

Bei gegebener Irrtumswahrscheinlichkeit α (i.a. 5% oder 1%) ergibt sich der kritische Wert daher als das $(1-\alpha)$-Quantil der Chiquadratverteilung mit df=1 Freiheitsgraden.



Chiquadrat-Test auf statistische Unabhängigkeit

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Schritt 4: Berechnung der Teststatistik und Entscheidung

Schließlich wird die Teststatistik berechnet und anhand des resultierenden Wertes die Nullhypothese beibehalten bzw. verworfen.

Die Nullhypothese H_0 wird mit einer Irrtumswahrscheinlichkeit α abgelehnt, wenn gilt:

$$\chi^2 \geq \chi^2_{1-\alpha;df=1}.$$

Für die Beispieltabelle hat sich ein Wert vom $\chi^2 = 120.9$ ergeben.

Bei einer Irrtumswahrscheinlichkeit von z.B. 5%, beträgt der Wert des 95%-Quantils der Chiquadratverteilung mit df=1 Freiheitsgraden 3.841.

Da 120.9 > 3.481, ist die Nullhypothese zu verwerfen.

Bei einer Irrtumswahrscheinlichkeit von 5% kann davon ausgegangen werden, dass ein Zusammenhang zwischen der Beurteilung der allgemeinen wirtschaftlichen Lage und der Beurteilung der eigenen Lage besteht.

Chiquadrat-Test auf statistische Unabhängigkeit

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut	nicht gut	Summe
- gut	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Schritt 5: Überprüfung der Anwendungsvoraussetzungen

Da der Chiquadrattest wie der Z-Test von Anteilen und Anteilsdifferenzen nur asymptotisch gilt, sind die Anwendungsvoraussetzungen zu prüfen.

Im Prinzip können die gleichen Regeln wie beim Z-Test der Prozentsatzdifferenz unter der Annahme $\delta_{YX\%}=0$ getroffen werden. Allerdings hat sich beim Chiquadrattest ein alternatives Kriterium durchgesetzt. Danach sollen die erwarteten Häufigkeiten in allen vier Zellen größer fünf sein. Zur Überprüfung kann die geringste erwartete Häufigkeit berechnet werden:

$$\min\left(n_{i\bullet}\right)\cdot\min\left(n_{\bullet j}\right)/n > 5$$

Im Beispiel ergibt sich: 990 ·307 / 2282 = 133.2 > 5.
Die Anwendungsvoraussetzung ist erfüllt.

Beziehung zwischen Pearsons Chiquadrat-Test und Z-Test von Prozentsatzdifferenzen

Wenn die Nullhypothese geprüft werden soll, dass es keinen Zusammenhang zwischen Spalten- und Zeilenvariable gibt, kann sowohl Pearsons Chiquadrat-Test als auch der Z-Test der Prozentsatzdifferenz verwendet werden.

Tatsächlich führen beide Tests bei einem zweiseitigen Hypothesentest und d% = 0 stets zum identischen Ergebnis, da Pearsons Chiquadratstatistik in dieser Situation gerade das Quadrat der Teststatistik des Z-Tests ist:

$$\chi^2 = n\cdot\frac{\left(a\cdot d - b\cdot c\right)^2}{\left(a+b\right)\cdot\left(c+d\right)\cdot\left(a+c\right)\cdot\left(b+d\right)} = \left(\frac{\frac{a}{a+c}-\frac{b}{b+d}}{\sqrt{\frac{\left(a+b\right)\cdot\left(c+d\right)}{n}\cdot\left(\frac{1}{a+c}+\frac{1}{b+d}\right)}}\right)^2 = Z^2$$

Da $\chi^2 = n\cdot\Phi^2$, folgt auch, dass Φ multipliziert mit der Wurzel aus der Fallzahl asymptotisch standardnormalverteilt ist, wenn in der Population Unabhängigkeit zwischen Zeilen- und Spaltenvariable besteht. Daher kann die Statistik

$$Z = \Phi\cdot\sqrt{n} = \frac{a\cdot d - b\cdot c}{\sqrt{\left(a+b\right)\cdot\left(c+d\right)\cdot\left(a+c\right)\cdot\left(b+d\right)}}\cdot\sqrt{n}$$

in Vierfeldertabellen ebenfalls zum Testen der Hypothesenpaare a) H_0: $\Phi = 0$, b) H_0: $\Phi \le 0$ und c) H_0: $\Phi \ge 0$ versus a) H_1: $\Phi \ne 0$, b) $H_!$: $\Phi > 0$ und c) $H_!$: $\Phi < 0$ herangezogen werden.

Beziehung zwischen Pearsons Chiquadrat-Test und Z-Test von Prozentsatzdifferenzen

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut		Allgemeine Wirtschaftslage nicht gut		Summe	
- gut	a	9.7% (222)	b	33.7% (768)	43.4% (990)	a+b
- nicht gut	c	3.7% (85)	d	52.9% (1207)	56.6% (1292)	c+d
Summe	a+c	13.5% (307)	b+d	86.5% (1975)	100.0% (2282)	n

(Quelle: Allbus 2006, nur Westen)

Diese Teststatistik hat stets den gleichen Wert wie die Teststatistik Z beim Test der Prozentsatzdifferenz auf Null. Die Nullhypothese wird entsprechend mit Irrtumswahrscheinlichkeit α verworfen, wenn a) $|Z| \geq z_{1-\alpha/2}$, b) $Z \geq z_{1-\alpha}$ bzw. c) $Z \leq z_{\alpha}$.

Für die Beispieldaten aus dem Allbus berechnen sich die Statistiken nach:

$$Z = \Phi \cdot \sqrt{n} = \frac{222 \cdot 1207 - 768 \cdot 85}{\sqrt{990 \cdot 1292 \cdot 307 \cdot 1975}} \cdot \sqrt{2282} = 10.99$$

$$= \frac{\frac{d_{YX}\%}{100}}{\hat{\sigma}\left(\frac{d_{YX}\%}{100}\right)} = \frac{\frac{222}{307} - \frac{768}{1975}}{\sqrt{\frac{990 \cdot 1292}{2282^2} \cdot \left(\frac{1}{307} + \frac{1}{1975}\right)}} = 10.99$$

$$= \frac{\frac{d_{XY}\%}{100}}{\hat{\sigma}\left(\frac{d_{XY}\%}{100}\right)} = \frac{\frac{222}{990} - \frac{85}{1292}}{\sqrt{\frac{307 \cdot 1975}{2282^2} \cdot \left(\frac{1}{990} + \frac{1}{1292}\right)}} = 10.99$$

Lerneinheit 17:
Bivariate Zusammenhänge zwischen nominalskalierten Variablen

	Konfession			
Wahlabsicht	katholisch	evangelisch	keine	Summe
CDU	48.4% (327)	35.6% (306)	22.3% (141)	35.7% (774)
SPD	29.3% (198)	34.9% (300)	34.2% (216)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Die Vierfeldertabelle ergibt sich bei der Kreuztabellierung von zwei dichotomen Variablen. Hat eine Variable mehr als zwei Ausprägungen, so hat die resultierende Kreuztabelle mehr als vier Zellen.

Das Beispiel zeigt die 5×3-Tabelle der Wahlabsicht (abhängige Zeilenvariable) nach Konfession (unabhängige Spaltenvariable).

Dichotome Variablen können stet auch als Ratio-Skalen mit den Ausprägungen „Eigenschaft vorhanden“ und „Eigenschaft nicht vorhanden“ interpretiert werden. Bei Variablen mit mehr als zwei Ausprägungen gilt dies nicht. Bei Zusammenhangsanalysen muss dann das Messniveau berücksichtigt werden.

Bivariate Zusammenhänge in der Mehrfeldertabelle

	Konfession			
Wahlabsicht	keine	katholisch	evangelisch	Summe
CDU	22.3% (141)	48.4% (327)	35.6% (306)	35.7% (774)
SPD	34.2% (216)	29.3% (198)	34.9% (300)	32.9% (714)
Grüne	21.2% (134)	13.6% (92)	15.0% (129)	16.4% (355)
FDP	6.5% (41)	7.2% (49)	12.7% (109)	9.2% (199)
PDS	15.8% (100)	1.5% (10)	1.9% (16)	5.8% (126)
Summe	100.0% (632)	100.0% (676)	100.0% (860)	100.0% (2168)

Quelle: Allbus 1996

So macht es bei zwei polytomen nominalskalierten Variablen keinen Sinn, von einem positiven bzw. negativen Zusammenhang zu sprechen, da es kein mehr oder weniger an einer Eigenschaft gibt, und daher auch keine sinnvolle Aussage der Form „je mehr von X, desto mehr (bzw. weniger) von Y“.

Bezogen auf den Zusammenhang in Kreuztabellen heißt dies, dass eine beliebige Permutation von Zeilen oder von Spalten die Stärke des Zusammenhangs nicht verändern darf.

So muss sich der selbe Zusammenhang ergeben, wenn im Beispiel die Parteien z.B. nach dem relativen Anteil ihrer Stimmen angeordnet werden und bei der Konfession zuerst die Konfessionslosen in der Tabelle aufgeführt werden.

Bei einem asymmetrischen Zusammenhang und Prozentuierung innerhalb der Kategorien der unabhängigen Variable kann die Interpretation aber trotzdem analog zur Vierfeldertabelle erfolgen; es sind jedoch mehr Prozentwertvergleiche notwendig.

Bivariate Zusammenhänge in der Mehrfeldertabelle

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (327)	35.6% (306)	22.3% (141)	35.7% (774)
SPD	29.3% (198)	34.9% (300)	34.2% (216)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Verglichen werden wiederum die relativen Häufigkeiten bzw. Prozentwerte einer Ausprägung der abhängigen Variablen zwischen den Ausprägungen der unabhängigen Variablen.

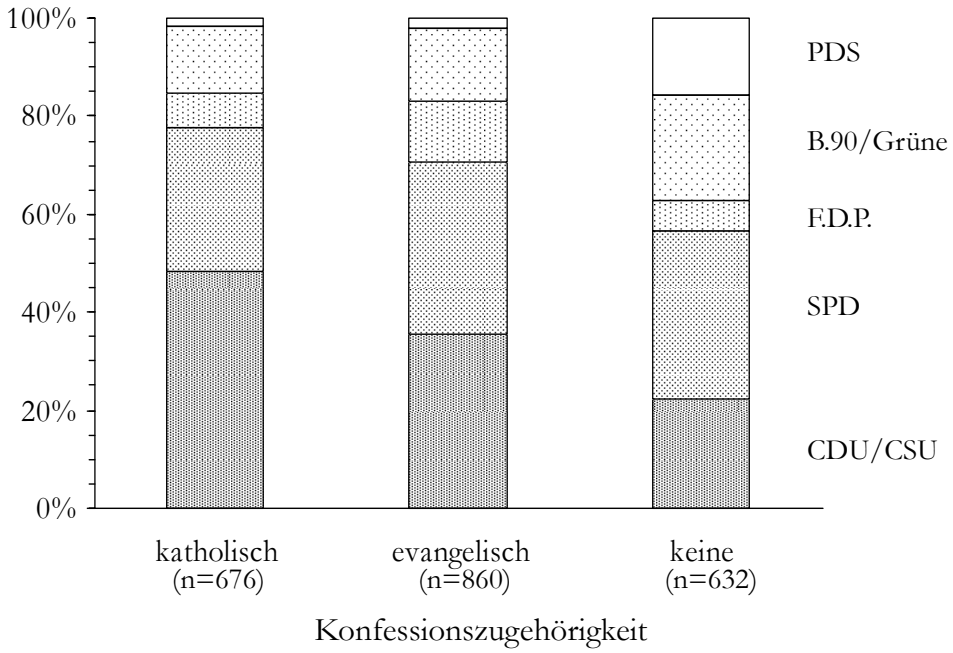
Sichtbar wird, dass Katholiken zu einem höheren Anteil CDU wählen als Protestanten und diese mehr als Konfessionslose.

Die SPD wird von Protestanten am häufigsten gewählt, dicht gefolgt von Konfessionslosen und dann von Katholiken. Die Prozentwertunterschiede sind hier aber nicht sehr groß.

Die FDP wird vor allem von Protestanten gewählt, die Grünen und die PDS von Konfessionslosen.

Hinweis: Aufgrund der disproportionalen Schichtung nach alten und neuen Bundesländern lassen sich die Ergebnisse allerdings nicht einfach auf die Bundesrepublik insgesamt verallgemeinern.

Bivariate Zusammenhänge in der Mehrfeldertabelle



Eine grafische Darstellung über Säulendiagramme der bedingten Verteilungen ist meist übersichtlicher, solange die abhängige Variable nicht sehr viele Ausprägungen hat.

Chiquadrat-Test auf statistische Unabhängigkeit in der I×J-Kreuztabelle

Pearsons Chiquadrat-Test auf Unabhängigkeit der Zeilen- und Spaltenvariablen in der Population kann auch bei Mehrfeldertabellen angewendet werden.
Der einzige Unterschied zur Vierfeldertabelle besteht darin, dass sich die Berechnung der Teststatistik über mehr Zellen erstreckt und die Zahl der Freiheitsgrade größer ist.
Dies kann am Beispiel des Zusammnehangs zwischen Wahlabsicht und Konfession verdeutlich werden.

Schritt 1: Formulierung von Null- und Alternativhypothese

H_0: $\pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j}$ für i=1 bis 5, j =1 bis 3
versus H_1: $\pi_{ij} \neq \pi_{i\bullet} \cdot \pi_{\bullet j}$ für mindestens ein i,j

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung

Als Teststatistik wird für alle Hypothesentests Pearsons Chiquadrat-Statistik herangezogen:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

Bei gültiger Nullhypothese ist die Teststatistik asymptotisch chiquadratverteilt.
Die Zahl der Freiheitsgrade ist bei einer Kreuztabelle mit I Zeilen und J Spalten df=(I–1)·(J–1).
Wenn die Nulhypothese nicht zutrifft, also eine statistische Abhängigkeit zwischen Zeilen- und Spaltenvariable besteht, dann ist die Teststatistik nichtzentral chiquadratverteilt.

Chiquadrat-Test auf statistische Unabhängigkeit in der I×J-Kreuztabelle

Schritt 3: Festlegung von Irrtumswahrscheinlichkeit und kritischen Werten

Bei gegebener Irrtumswahrscheinlichkeit α (i.a. 5% oder 1%) ergibt sich der kritische Wert als das (1–α)-Quantil der Chiquadrat-Verteilung mit df=(I–1)·(J–1) Freiheitsgraden.

Im Beispiel ist df=(5–1)·(3–1) = 8.

Die Nullhypothese H_0 wird mit einer Irrtumswahrscheinlichkeit α abgelehnt, wenn gilt:

$$\chi^2 \geq \chi^2_{1-\alpha;\ df=(I-1)\cdot(J-1)}$$

Bei einer Irrtumswahrscheinlichkeit von 5%, beträgt der Wert des 95%-Quantils der Chiquadrat-Verteilung mit df=8 Freiheitsgraden 15.51.
Im Anwendungsbeispiel ist die Nullhypothese als vermutlich falsch zu verwerfen, wenn die Teststatistik einen Wert von mindestens 15.51 erreicht.

Schritt 4: Berechnung der Teststatistik und Entscheidung

Im letzten Schritt wird die Teststatistik berechnet und anhand des resultierenden Wertes die Nullhypothese beibehalten bzw. verworfen.

Chiquadrat-Test auf statistische Unabhängigkeit in der I×J-Kreuztabelle

Bei Unabhängigkeit erwartete und tatsächlich beobachtete Häufigkeiten

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	241.3 (327)	307.0 (306)	225.6 (141)	(774)
SPD	222.6 (198)	283.2 (300)	208.1 (216)	(714)
FDP	62.0 (49)	78.9 (109)	58.0 (41)	(199)
Grüne	110.7 (92)	140.8 (129)	103.5 (134)	(355)
PDS	39.3 (10)	50.0 (16)	36.7 (100)	(126)
Summe	(676)	(860)	(632)	(2168)

(beobachtete Häufigkeiten in Klammen)

$$\chi^2 = \left(\frac{(327-241.3)^2}{241.3} + \frac{(306-307.0)^2}{307.0} + \frac{(141-225.6)^2}{225.6} + \frac{(198-222.6)^2}{222.6} + \frac{(300-283.2)^2}{283.2} \right.$$
$$+ \frac{(216-208.1)^2}{208.1} + \frac{(49-62.0)^2}{62.0} + \frac{(109-78.9)^2}{78.9} + \frac{(41-58.0)^2}{58.0} + \frac{(92-110.7)^2}{110.7}$$
$$\left. + \frac{(129-140.8)^2}{140.8} + \frac{(134-103.5)^2}{103.5} + \frac{(10-39.3)^2}{39.3} + \frac{(16-50.0)^2}{50.0} + \frac{(100-36.7)^2}{36.7} \right) = 252.4$$

Chiquadrat-Test auf statistische Unabhängigkeit in der I×J-Kreuztabelle

Für das Beispiel hatte sich ein Wert vom $\chi^2 = 252.4$ ergeben.
Da 252.4 > 15.51, ist die Nullhypothese zu verwerfen.
Bei einer Irrtumswahrscheinlichkeit von 5% kann davon ausgegangen werden, dass ein Zusammenhang zwischen der Wahlabsicht und der Konfession besteht.

Schritt 5: Prüfung der Anwendungsvoraussetzungen

Der Chiquadrat-Test ist nur asymptotisch gültig.
Die Annäherung ist hinreichend genau, wenn die erwarteten Häufigkeiten größer 5 sind.
Als Faustregel gilt bei größeren Tabellen, dass
a) $e_{ij} > 1$ für alle i, j und
b) $e_{ij} > 5$ für mindestens 80% (4/5) aller Zellen.

Da im Beispiel die kleinste erwartete Häufigkeit 36.7 ist, ist die Anwendungsvoraussetzung erfüllt.

Standardisierte Residuen

Standardisierte Residuen

Wahlabsicht	Konfession katholisch	evangelisch	keine
CDU	5.5	–0.1	–5.6
SPD	–1.7	1.0	0.5
FDP	–1.7	3.4	–2.2
Grüne	–1.8	–1.0	3.0
PDS	–4.7	–4.8	10.4

Wird der Chiquadratanteil jeder Tabellenzelle berechnet, die Wurzel daraus gezogen und als Vorzeichen die Differenz zwischen beobachteter und erwarteter Häufigkeit verwendet, dann ergeben sich die ***standardisierten Residuen***:

$$sr_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

Die Werte sind bei gültiger H_0 asymptotisch standardnormalverteilt.
Werte ≥ 1.96 oder ≤ -1.96 weisen also darauf hin, dass es bei einer Irrtumswahrscheinlichkeit von 5% überzufällige Abweichungen von Unabhängigkeit in der entsprechenden Tabellenzelle gibt.

Im Beispiel zeigt, sich, dass es unter den CDU-Wählern überzufällig viele Katholiken und zu wenig Konfessionslose gibt.
Bei der PDS sind beide Konfessionen unter- und die Konfessionslosen überrepräsentiert.

Cramérs V

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (327)	35.6% (306)	22.3% (141)	35.7% (774)
SPD	29.3% (198)	34.9% (300)	34.2% (216)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Analog der Berechnung von Φ kann auch bei größeren Tabellen aus der Chiquadrat-Statistik ein symmetrisches Zusammenhangsmaß konstruiert werden.
Dabei wird die Teststatistik durch ihren Maximalwert geteilt und aus dem Quotienten die Quadratwurzel gezogen. Dieses Zusammenhangsmaß heißt nach dem Statistiker Cramér Cramérs V.

In einer I×J-Kreuztabelle ist der Maximalwert von χ^2 gleich dem Produkt aus der Fallzahl und dem Maximum der Spalten- oder Zeilenzahl minus eins:

$$\chi^2 \leq n \cdot \min(I-1, J-1)$$

Im Beispiel mit 5 mal 3 Tabellenzellen ist das Maximum von Chiquadrat 2168 ·2 = 4336.

Cramérs V

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (327)	35.6% (306)	22.3% (141)	35.7% (774)
SPD	29.3% (198)	34.9% (300)	34.2% (216)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Die Berechnungsformel für Cramérs V ist dann:

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n \cdot \min(I-1, J-1)}}$$

ImBeispiel ergibt sich ein Wert von √(252,4/4336) = 0.24

Die Interpretation entspricht der von Φ. Tatsächlich sind beide Maße in der Vierfeldertabelle (evtl. bis auf das Vorzeichen) identisch, da das Maximum von χ^2 dann n ist.

Es besteht somit eine mittelstarke Beziehung zwischen Wahlabsicht und Konfession.

Im Unterschied zu Φ hat V kein Vorzeichen, da das Maß für nominalskalierte Variablen mit mehr als zwei Ausprägungen konstruiert ist.

Die Logik von PRE-Maßen: Lambda und relative Devianzreduktion

Asymmetrische Zusammenhangsmaße basieren oft auf der Logik der Reduzierung von Vorhersagefehlern:
Die Voraussage einer Realisationen einer abhängigen Variable kann fehlerhaft sein. Die Anzahl der Fehler sollte sich reduzieren, wenn die Zielvariable mit einer Prädiktorvariable zusammenhängt und die Wert der Prädiktorvariablen bei den Fällen bekannt sind. Auf dieser Idee basiert die Logik von Zusammenhangsmaßen, die die Vorhersagefehlerreduktion erfassen, sogenannte PRE-Maße. (PRE steht für *proportional reduction in error*).

E_0 soll das Ausmaß der Fehler bezeichnen, mit denen zu rechnen ist, wenn keine Zusatzinformationen vorliegen. E_1 ist das Ausmaß der Fehler, wenn bekannt ist, welchen Wert eine Prädiktorvariable aufweist. Das Ausmaß, indem sich die Fehler bei Kenntnis einer erklärenden Variable reduzieren, ergibt sich dann über die Formel

$$PRE = \frac{E_0 - E_1}{E_0} = 1 - \frac{E_1}{E_0}$$

Der resultierende Wert lässt sich leicht interpretieren, da er den Anteil der Fehlerreduktion angibt. Ein Wert von 0 bedeutet keinerlei Reduktion, ein Wert von 0.5 oder 50% eine Halbierung der Fehler und ein Wert von 1 bzw. 100% eine maximale Fehlerreduktion, d.h. eine perfekte Vorhersage.

Um ein PRE-Maß zu konstruieren, muss zunächst festgelegt werden, was Vorhersagefehler sind. Bei nominalskalierten Variablen liegt es nahe, den Modalwert als Vorhersagewert zu verwenden und als Fehler zu zählen, mit welcher Häufigkeit Abweichungen vom Modalwert auftreten.

Lambda

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (**327**)	35.6% (**306**)	22.3% (141)	35.7% (**774**)
SPD	29.3% (198)	34.9% (300)	34.2% (**216**)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Wenn die Zeilenvariable abhängige Variable ist, ergibt sich die Höhe der Fehler ohne Kenntnis der erklärenden Variable aus der Fallzahl in der Tabelle minus dem Modalwert der abhängigen Zeilenvariablen, also dem Maximalwert in der rechten Randspalte:

$$E_0 = n - \max_i\left(n_{i\bullet}\right) \qquad \textit{im Beispiel: 2168–774=1394}$$

Analog berechnen sich die Fehler für alle Ausprägungen der erklärenden Variablen, also der Spalten durch die Differenz der jeweiligen Spaltensumme minus dem Maximum der Spalte:

$$E_1 = \sum_{j=1}^{J}\left(n_{\bullet j} - \max_i\left(n_{ij}\right)\right) = n - \sum_{j=1}^{J}\max_i\left(n_{ij}\right) \qquad \textit{im Beispiel: 2168–(327+306+216)=1319}$$

Das resultierende Zusammenhangsmaß heißt λ_{YX} (lambda-YX). Obwohl das gleiche Symbol verwendet wird, sollte das Zusammenhangsmaß λ_{YX} nicht mit dem Parameter λ der Poisson-Verteilung verwechselt werden.

Lambda

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (**327**)	35.6% (**306**)	22.3% (141)	35.7% (**774**)
SPD	29.3% (198)	34.9% (300)	34.2% (**216**)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Für das Beispiel ergibt sich:

$$\lambda_{YX} = 1 - \frac{E_1}{E_0} = 1 - \frac{\sum_{j=1}\left(n_{\bullet j} - \max_i\left(n_{ij}\right)\right) \text{ in Spalte } j}{n_{\bullet\bullet} - \max_i\left(n_{i\bullet}\right)}$$

$$= 1 - \frac{E_1}{E_0} = 1 - \frac{(676-327)+(860-306)+(632-216)}{2168-774} = 1 - \frac{1319}{1394} = 0.054$$

Bei Kenntnis der Konfession lässt sich die Wahlabsicht mit einer um 5.4% geringeren Fehlerquote voraussagen als ohne Kenntnis der Konfession.

Wenn die Wahlabsicht als erklärende Variable und Konfession als abhängige Variable betrachtet werden, ergibt sich E_0 als Fallzahl minus dem Modalwert der Randzeile und E_1 als Summe der Spaltensummen minus den jeweiligen Modalwerten in den Reihen.

Lambda

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (**327**)	35.6% (**306**)	22.3% (141)	35.7% (**774**)
SPD	29.3% (198)	34.9% (300)	34.2% (**216**)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Bei einem perfekten Zusammenhang ist entweder λ_{YX} =1 oder λ_{XY} =1. Dann ist auch Pearsons Chiquadrat-Statistik maximal. Dies ist immer dann der Fall, wenn entweder in allen Zeilen oder in allen Spalten nur eine einzige Zelle ungleich 0 ist.

Bei den Beispieldaten ergäbe sich z.B. bei einem perfekten Zusammenhang:

Wahlabsicht	katholisch	evangelisch	keine
CDU	774	0	0
SPD	0	714	0
FDP	0	199	0
Grüne	0	0	355
PDS	0	0	126

λ_{YX} = 0.767
λ_{XY} = 1.000
χ^2 = 4336.0 = 2 · 2168

Devianzreduktion

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (327)	35.6% (306)	22.3% (141)	35.7% (774**)**
SPD	29.3% (198)	34.9% (300)	34.2% (216**)**	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Ein Nachteil von λ_{YX} ist, dass der Modalwert in der Regel nur eine sehr ungenaue Prognose erlaubt. Daher kann λ_{YX} selbst dann Null sein, wenn der Chiquadrat-Test auf statistische Unabhängigkeit einen signifikanten Zusammenhang anzeigt.

Die Konzeption der proportionalen Fehlerreduktion kann aber auch bei anderen Fehlerdefinitionen angewendet werden.
So kann die Devianz, d.h. die Streuung nominalskalierter Variablen, als Maß für den Vorhersagefehler verwendet werden. Der Fehler E_0 ist dann die Devianz der Randverteilung der abhängigen Variable und E_1 die Summe der Devianzen in den konditionalen Verteilungen der abhängigen Variable gegeben die Ausprägungen der unabhängigen Variable.

Devianzreduktion

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (327)	35.6% (306)	22.3% (141)	35.7% (774)
SPD	29.3% (198)	34.9% (300)	34.2% (216)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Wenn im Beispiel die Spaltenvariable „Wahlabsicht“ (Y) abhängige Variable ist, berechnet sich die Devianz der Randverteilung nach:

$$D_Y = -2\sum_{i=1}^{I} n_{i\bullet} \cdot \ln\left(\frac{n_{i\bullet}}{n_{\bullet\bullet}}\right)$$

$$= -2 \cdot \left(774 \cdot \ln\left(\frac{774}{2168}\right) + 714 \cdot \ln\left(\frac{714}{2168}\right) + 199 \cdot \ln\left(\frac{199}{2168}\right) + 355 \cdot \ln\left(\frac{355}{2168}\right) + 126 \cdot \ln\left(\frac{126}{2168}\right) \right)$$

$$= 6132.71$$

Devianzreduktion

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (327)	35.6% (306)	22.3% (141)	35.7% (774)
SPD	29.3% (198)	34.9% (300)	34.2% (216)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

E_1 berechnet sich dann nach: $D_{YX} = -2\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} \cdot \ln\left(\frac{n_{ij}}{n_{\bullet j}}\right)$

$$D_{YX} = -2 \cdot \left(\begin{array}{l} 327 \cdot \ln\left(\frac{327}{676}\right) + 198 \cdot \ln\left(\frac{198}{676}\right) + 49 \cdot \ln\left(\frac{49}{676}\right) + 92 \cdot \ln\left(\frac{92}{676}\right) + 10 \cdot \ln\left(\frac{10}{676}\right) \\ +306 \cdot \ln\left(\frac{306}{860}\right) + 300 \cdot \ln\left(\frac{300}{860}\right) + 109 \cdot \ln\left(\frac{109}{860}\right) + 129 \cdot \ln\left(\frac{129}{860}\right) + 16 \cdot \ln\left(\frac{16}{860}\right) \\ +141 \cdot \ln\left(\frac{141}{632}\right) + 216 \cdot \ln\left(\frac{216}{632}\right) + 41 \cdot \ln\left(\frac{41}{632}\right) + 134 \cdot \ln\left(\frac{134}{632}\right) + 100 \cdot \ln\left(\frac{100}{632}\right) \end{array} \right)$$

$$= 5895.04$$

Devianzreduktion

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (327)	35.6% (306)	22.3% (141)	35.7% (774)
SPD	29.3% (198)	34.9% (300)	34.2% (216)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Das resultierende PRE-Maß wird als ***relative Devianzreduktion***, ***Likelihood-Ratio-Index*** oder (Mc Faddens) *Pseudo-R-Quadrat* $R^{/}$ bezeichnet.
In einer bivariaten Kreuztabelle wird das Maß auch als ***Unsicherheitskoeffizient*** bezeichnet.

$$R^{/}_{YX} = 1 - \frac{D_{YX}}{D_Y} = 1 - \frac{-2\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} \cdot \ln\left(\frac{n_{ij}}{n_{\bullet j}}\right)}{-2\sum_{i=1}^{I} n_{i\bullet} \cdot \ln\left(\frac{n_{i\bullet}}{n_{\bullet\bullet}}\right)} = 1 - \frac{5895.04}{6132.71} = 0.038$$

Im Beispiel reduziert sich die Devianz der Wahlabsicht bei Kenntnis der Konfession um 3.8%.

Stärke des Zusammenhangs

Wahlabsicht	Konfession katholisch	evangelisch	keine	Summe
CDU	48.4% (327)	35.6% (306)	22.3% (141)	35.7% (774)
SPD	29.3% (198)	34.9% (300)	34.2% (216)	32.9% (714)
FDP	7.2% (49)	12.7% (109)	6.5% (41)	9.2% (199)
Grüne	13.6% (92)	15.0% (129)	21.2% (134)	16.4% (355)
PDS	1.5% (10)	1.9% (16)	15.8% (100)	5.8% (126)
Summe	100.0% (676)	100.0% (860)	100.0% (632)	100.0% (2168)

Quelle: Allbus 1996

Verglichen mit den Zusammenhangsmaßen Φ bzw. Crámers V und der Anteilsdifferenz d_{YX}%/100 weisen die Pre-Maße meist deutlich geringere Werte auf. Tatsächlich sind sie in der Größenordnung eher mit Φ^2 zu vergleichen. Als Anhaltswerte für die Interpretation ergibt sich damit:

Stärke eines Zusammenhangs	
praktisch kein	$R^{/}$ bzw. $\lambda < 0.0025$
geringer	$0.0025 \leq R^{/}$ bzw. $\lambda < 0.04$
mittlerer	$0.0400 \leq R^{/}$ bzw. $\lambda < 0.25$
starker	$0.2500 \leq R^{/}$ bzw. λ

Für die Beispieldaten ergeben sich geringe bis mittlere Zusammenhangsstärken:
$R^{/}_{YX} = 0.038 \quad \lambda_{YX} = 0.054$
$R^{/}_{XY} = 0.050 \quad \lambda_{XY} = 0.084$

Bei der Interpretation asymmetrischer nominalskalierten Zusammenhangsmaßen ist zu beachten, dass die Maße tendenziell um so geringer sind, je stärker die Randverteilung der abhängigen Variable von einer Gleichverteilung abweicht.

LR-Test auf statistische Unabhängigkeit

Zur Prüfung der statistischen Unabhängigkeit von Zeilen- und Spaltenvariablen kann anstelle von Pearsons Chiquadrat-Test auch geprüft werden, ob die relative Devianzreduktion signifikant von null verschieden ist.
Dieser Test wird als ***Likelihood-Ratio-Test*** bezeichnet. Die Teststatistik wird durch $\mathbf{L^2}$ symbolisiert.
Die Teststatistik L^2 ist die Differenz der bedingten Devianz D_{YX} von der unbedingten Devianz D_Y. Alternativ kann die Teststatistik ähnlich wie Pearsons Chiquadrat-Statistik über die beobachteten Zellenhäufigkeiten n_{ij} und die bei Unabhängigkeit erwarteten Häufigkeiten e_{ij} berechnet werden:

$$L^2 = D_Y - D_{YX} = 2 \cdot \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \cdot \ln\left(\frac{n_{ij}}{e_{ij}}\right)$$

Im Beispiel ergibt sich ein Wert von $L^2 = D_Y - D_{YX} = 6132.71 - 5895.04 = 237.67$.

Wenn die Nullhypothese zutrifft, dass kein Zusammenhang besteht, dann ist die LR-Statistik asymptotisch chiquadratverteilt.
Die Zahl der Freiheitsgrade berechnet sich wie bei Pearsons Chiquadrat: df=(I–1)·(J–1).
Ist die Nullhypothese falsch, ist L^2 nichtzentral chiquadratverteilt.

Pearsons Test und der LR-Test sind asymptotisch äquivalent, so dass beide Teststatistiken i.a sehr ähnliche Werte aufweisen. Große Abweichungen können ein Hinweis sein, dass die asymptotische Annäherung nicht hinreichend ist.

Lerneinheit 18:
Symmetrische Beziehungen zwischen zwei metrischen Variablen

Da die Messniveaus eine hierarchische Ordnung bilden, ist es stets möglich, nur die Informationen eines niedrigeren Messniveaus zu nutzen. Die Logik der Zusammenhangsanalyse zwischen zwei nominalskalierten Variablen kann daher auch angewendet werden, wenn die betrachteten Variablen ordinales oder metrisches Messniveau aufweisen.
Dabei werden allerdings Informationen verschenkt. So ist es in der Tabellenanalyse von zwei nominalskalierten Variablen (mit mehr als zwei Ausprägungen) nicht möglich, zwischen einer positiven und einer negativen Je-desto-Beziehung zu unterscheiden.
Wenn die betrachteten Variablen zudem sehr viele Ausprägungen haben, was bei metrischem Messniveau in der Regel der Fall ist, sind die resultierenden Tabellen oft unübersichtlich und enthalten keine oder sehr wenige Fälle in den einzelnen Tabellenzellen. Prozentuierungen sind dann weder sinnvoll noch aussagekräftig und Tests können wegen der geringen erwarteten Häufigkeiten nicht angewendet werden.

Als Beispiel zeigt die Tabelle auf der nächsten Seite die gemeinsame Häufigkeitsverteilung des Alters von Partnerinnen bzw. Partner der insgesamt 185 Befragten aus dem Allbus 2006, die eine Lebenspartnerin bzw. einen Lebenspartner haben, mit dem sie nicht zusammenleben. Da nur das Geschlecht der befragten Person erfasst wurde, wurde bei der Erstellung der Tabelle angenommen, dass es sich um heterosexuelle Partnerschaften handelt.

Beziehungen zwischen zwei metrischen Variablen

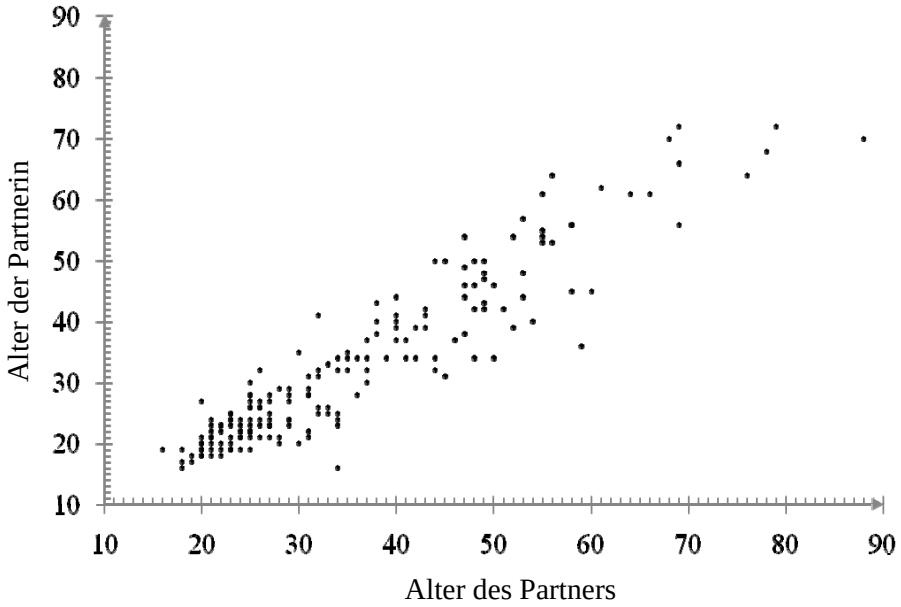
Alter der Partnerin (in Zeilen) nach Alter des Partners (in Spalten) bei nicht zusammen lebenden Partnern

	16	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	58	59	60	61	64	66	68	69	76	78	79	88	Σ
	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
17	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
18	0	0	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
19	1	1	0	2	1	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
20	0	0	0	3	1	2	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
21	0	0	0	1	2	0	1	3	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
22	0	0	0	0	1	1	0	2	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
23	0	0	0	0	1	2	1	1	1	1	2	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
24	0	0	0	0	1	0	3	2	1	1	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
25	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
26	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
27	0	0	0	1	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
28	0	0	0	0	0	0	0	0	2	0	1	0	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
29	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
30	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
32	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	1	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
35	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	2
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	3
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	3
62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	2
66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2
Σ	1	3	3	9	8	7	10	9	13	7	7	3	6	2	6	5	3	6	3	2	4	3	1	5	2	2	3	4	2	1	5	4	5	2	1	2	3	1	4	2	3	1	1	1	1	1	1	3	1	1	1	1	185

Beziehungen zwischen zwei metrischen Variablen

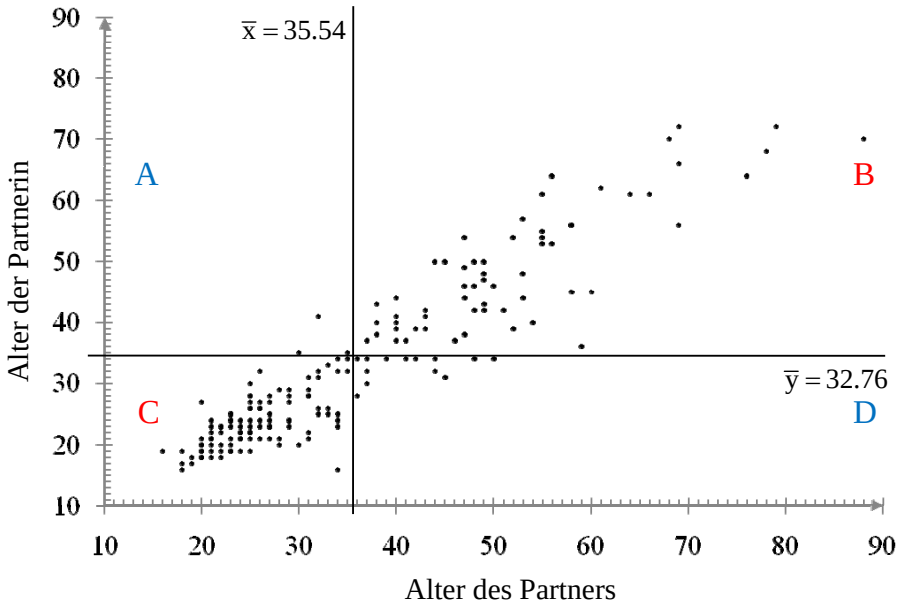
Zwar ist es möglich, durch Zusammenfassen in Ausprägungsklassen Kreuztabellen mit weniger Zellen zu generieren, doch geht dabei Informationen verloren.

Bei metrischem Messniveau der beiden Variablen besteht stattdessen die Möglichkeit, alle Realisierungen als Punkte in ein Koordinatensystem einzutragen.



Im Beispiel aus dem Allbus 2006 verläuft die resultierende Punktewolke tendenziell von links unten nach rechts oben: Je älter (jünger) eine Person ist, desto älter (jünger) ist also auch tendenziell auch sein Partner bzw. seine Partnerin.

Beziehungen zwischen zwei metrischen Variablen

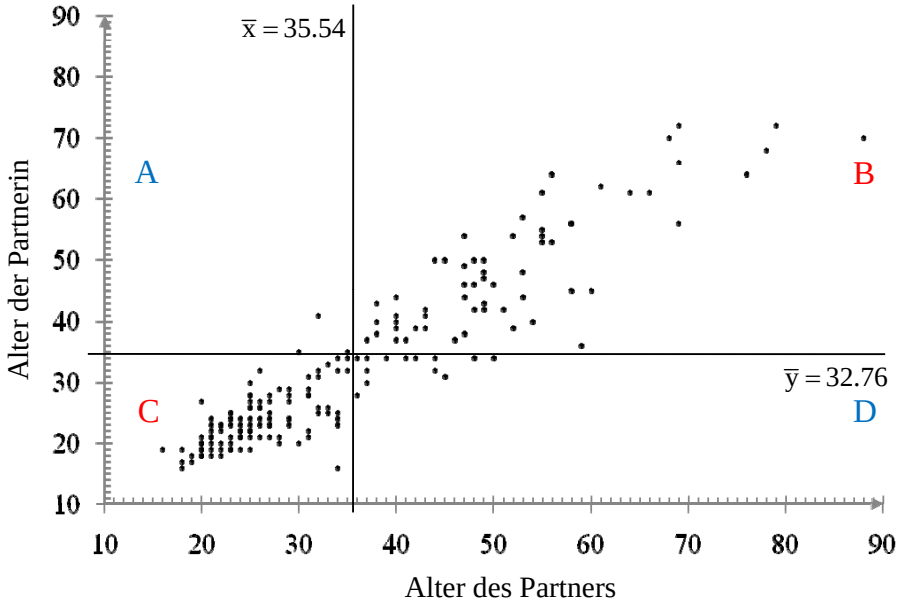


Deutlicher wird das Vorzeichen der Beziehung zwischen X (Alter des Partners) und Y (Alter der Partnerin), wenn zusätzlich die Mittelwerte der beiden Variablen als senkrechte (Mittelwert der X-Werte) bzw. waagerechte Geraden (Mittelwert der Y-Werte) eingezeichnet werden.

Der Schnittpunkt der beiden Mittelwerte ergibt den Schwerpunkt der Punktewolke.

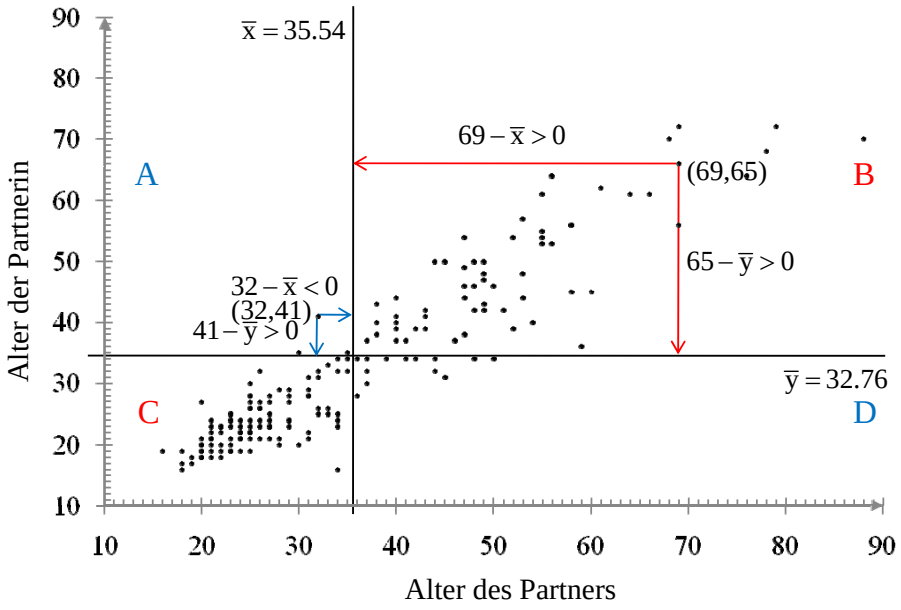
Durch die beiden Geraden wird das Koordinatensystem in vier Regionen eingeteilt.

Beziehungen zwischen zwei metrischen Variablen



- In der Region A liegen Punkte, deren X-Werte kleiner als der Mittelwert von X und deren Y-Werte größer als der Mittelwert von Y sind.
- In Region B liegen Punkte, deren X- und Y-Werte größer als die jeweiligen Mittelwerte sind.
- In Region C liegen Punkte, deren X-und Y-Werte kleiner als die jeweiligen Mittelwerte sind.
- In Region D liegen Punkte, deren X- Werte größer als der Mittelwert von X und deren Y-Werte kleiner als der Mittelwert von Y sind.

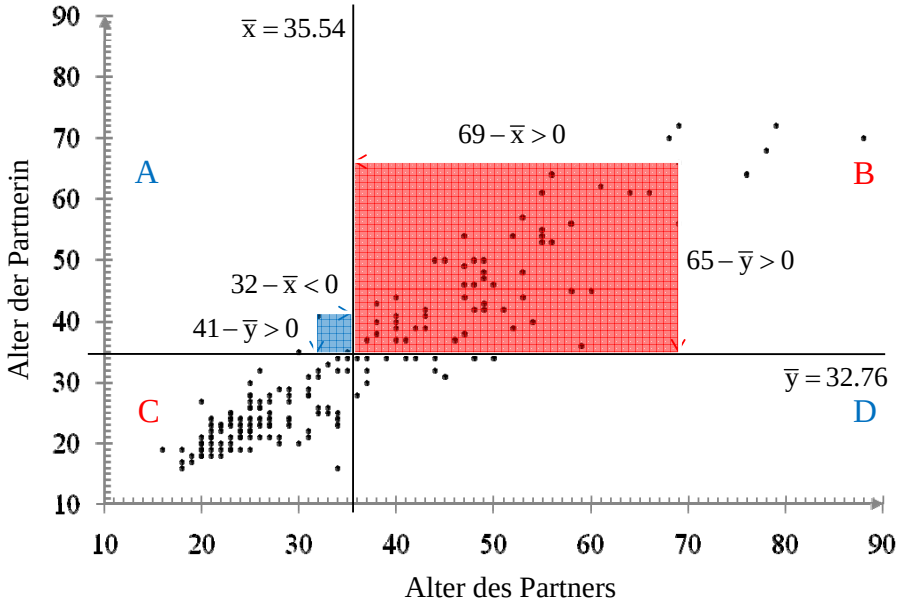
Beziehungen zwischen zwei metrischen Variablen



Wenn die meisten Punkte in B oder C liegen, besteht eine positive Beziehung, wenn die meisten Punkte in A oder D liegen, besteht dagegen eine negative Beziehung zwischen den beiden Variablen.

Betrachtet man den Schwerpunkt der Verteilung als Mittelpunkt, so ergeben sich Koordinaten relativ zu diesem Mittelpunkt, wenn von den Werten der ursprünglichen Koordinaten jeweils der Mittelwert von X bzw. Y abgezogen wird.

Beziehungen zwischen zwei metrischen Variablen

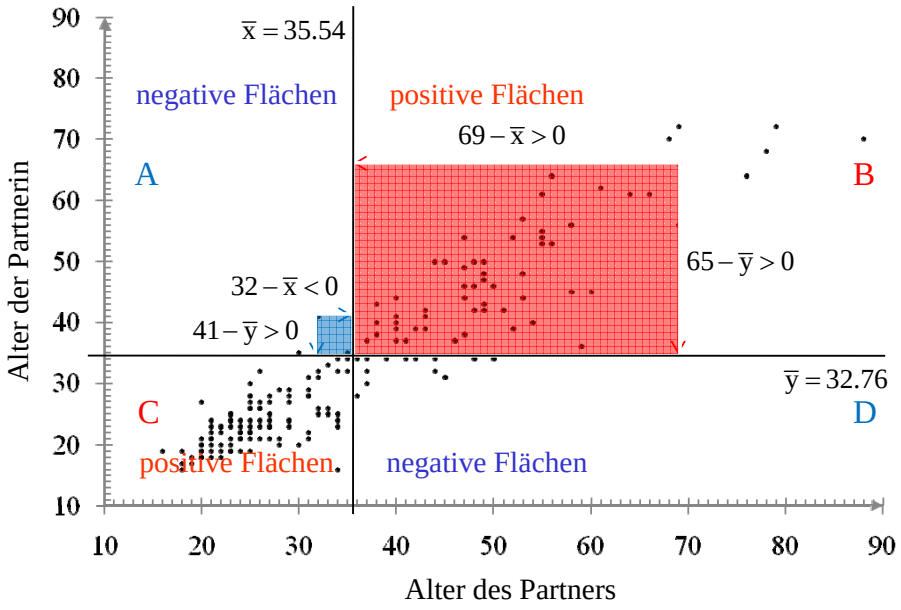


Das Produkt der Abstände eines Punktes zum Schwerpunkt definiert ein Reckteck, wobei die Fläche des Rechtecks gerade dem Wert des Produktes entspricht.

So definiert der Punkt (69,65), der für ein Paar steht, bei dem der Partner 69 Jahre und die Partnerin 65 Jahre alt ist, ein Rechteck in Region B mit einer Fläche von 1078.75 = (69–35.54)·(65–32.76).

Analog definiert der Punkt (32,41) ein Rechteck der Fläche (32–35.54)·(41–32.76) in A.

Kovariation als symmetrisches Zusammehangsmaß zwischen zwei metrischen Variablen

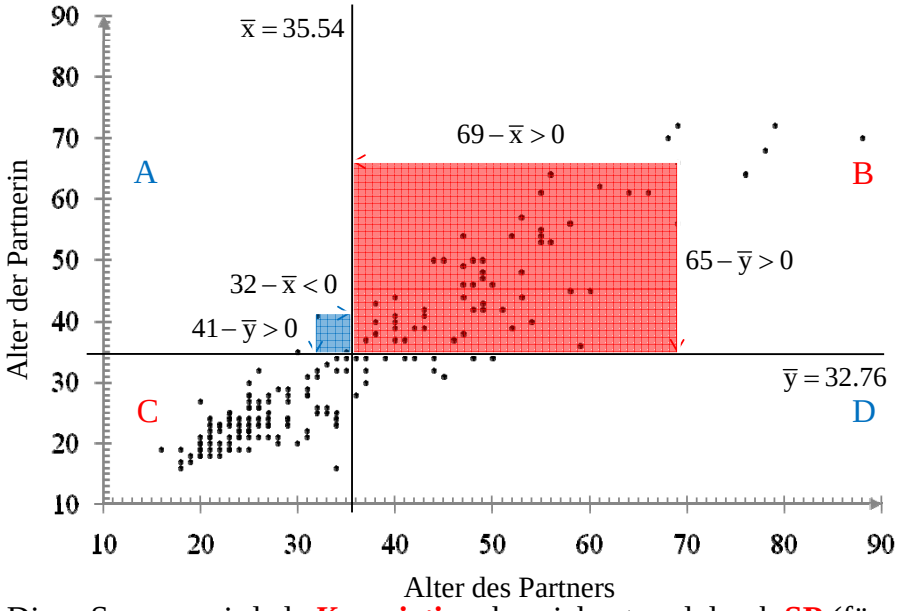


Berücksichtig man die Vorzeichen des Abstands zum Mittelwert, ergeben sich in den Regionen B und C positive und in den Regionen A und D negative Flächen.

Die Summe der Flächen kann dann als Maß für die symmetrische Beziehung (positiv, negativ, keine) zwischen zwei metrischen Variablen genutzt werden:

$$SP_{XY} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Kovariation als symmetrisches Zusammehangsmaß zwischen zwei metrischen Variablen



$$SP_{XY} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$SP_{XY} > 0 \Rightarrow$ positive Beziehung

$SP_{XY} < 0 \Rightarrow$ negative Beziehung

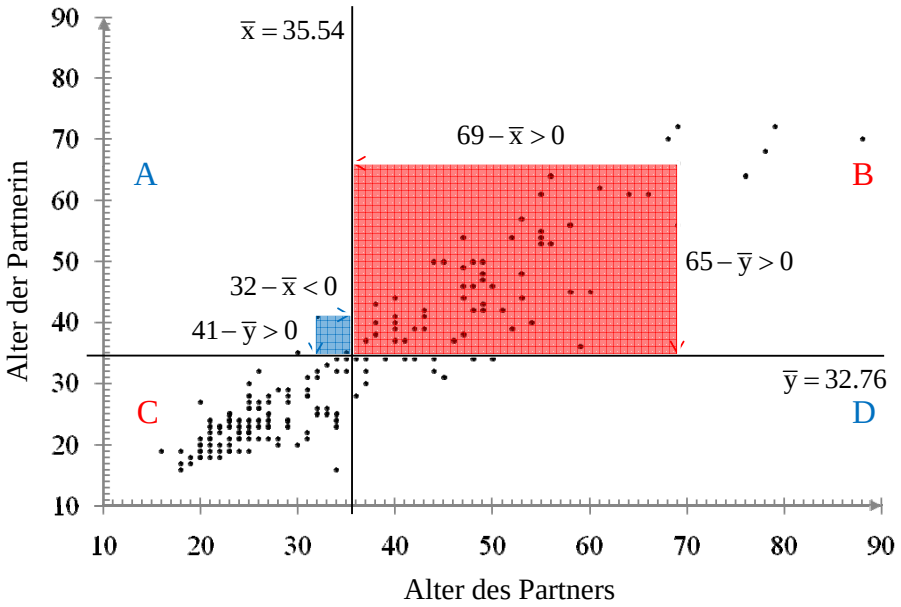
$SP_{XY} = 0 \Rightarrow$ keine monotone Beziehung

Diese Summe wird als ***Kovariation*** bezeichnet und durch **SP** (für: *sum of products*) symbolisiert.

Die Kovariation SP_{XY} gibt an, ob zwischen zwei metrischen Variablen X und Y eine positive Je-desto-Beziehung, eine negative Je-desto-Beziehung oder gar keine monotone Beziehung besteht.

Ein Nachteil der Kovariation ist, dass ihr Wert allein aufgrund steigender Fallzahl zunehmen kann.

Kovarianz als symmetrisches Zusammehangsmaß zwischen zwei metrischen Variablen



$$s_{XY} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$s_{XY} > 0 \Rightarrow$ positive Beziehung

$s_{XY} < 0 \Rightarrow$ negative Beziehung

$s_{XY} = 0 \Rightarrow$ keine monotone Beziehung

Dies gilt nicht, wenn anstelle der Summe der Flächen der Durchschnittswert berechnet wird, indem die Kovariation durch die Fallzahl geteilt wird:

$$s_{XY} = \frac{SP_{XY}}{n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n}$$

Die resultierende Statistik heißt ***Kovarianz***.

Beziehungen zwischen Kovariation und Variation bzw. Kovarianz und Varianz

Die Formeln zur Berechnung von Kovariation und Kovarianz ähneln den Formeln für die Variation und die Varianz:

$$SP_{XY} = \sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) \qquad s_{XY} = \frac{SP_{XY}}{n}$$

$$SS_X = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x}) \cdot (x_i - \bar{x}) \qquad s_X^2 = \frac{SS_X}{n}$$

Aufgrund dieser Ähnlichkeit kann die Variation bzw. Varianz einer Variable auch als Kovariation bzw. Kovarianz einer Variable mit sich selbst betrachtet werden. Entsprechend wird in Formeln die Varianz s^2_X einer Variable X bisweilen auch durch s_{XX} symbolisiert.

Wie bei der Berechnung von Variation und Varianz können auch bei der Berechnung von Kovariation und Kovarianz alternative Formeln verwendet werden:

$$SP_{XY} = \sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) = \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y} = \sum_{i=1}^{n} x_i \cdot y_i - \frac{\left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{n}$$

$$s_{XY} = \frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \frac{\left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{n^2}$$

Produktmomentkorrelation als symmetrisches Zusammehangsmaß

Es gibt eine weitere Beziehung zwischen Variationen und Kovariation bzw. Varianzen und Kovarianz:

- Das Maximum der Kovariation bzw. Kovarianz zwischen zwei Variablen X und Y ist gleich dem geometrischen Mittel der Variationen bzw. Varianzen der beiden Variablen!

$$SP_{XY} \le \sqrt{SS_X \cdot SS_Y} \; ; \; s_{XY} \le \sqrt{s_X^2 \cdot s_Y^2} = s_X \cdot s_Y$$

Diese Eigenschaft kann genutzt werden, um ein symmetrisches Zusammenhangsmaß zwischen zwei metrischen Variablen zu definieren, dass analog zu Φ in der Vierfeldertabelle zwischen –1 und +1 variiert.

Dieses Zusammenhangsmaß wird ***Produktmomentkorrelation*** r_{XY} oder nach dem Statistiker Pearson auch ***Pearsons Korrelationskoeffizient*** genannt und ist der Quotient aus der Kovarianz bzw. der Kovariation geteilt durch das geometrische Mittel der Varianzen bzw. Variationen:

$$r_{XY} = \frac{SP_{XY}}{\sqrt{SS_X \cdot SS_Y}} = \frac{s_{XY}}{\sqrt{s_X^2 \cdot s_Y^2}} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y}}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2\right)}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2\right) \cdot \left(\sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}^2\right)}}$$

Produktmomentkorrelation als symmetrisches Zusammehangsmaß

Die Bezeichnung Produktmomentkorrelation kommt daher, dass die Kovarianz auch als zentriertes Produktmoment m_{11} bezeichnet wird.
Wie bei univariaten Momenten kann auch zwischen Rohproduktmomenten und zentrierten Produktmomenten unterschieden werden:

$$m'_{jk} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^j \cdot y_i^k \; ; \; m_{jk} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^j \cdot (y_i - \bar{y})^k$$

$$\mu'_{jk} = \mu\left(X^j \cdot Y^k\right) \quad ; \; \mu_{jk} = \mu\left((X - \mu_X)^j \cdot (Y - \mu_Y)^k\right)$$

Für die Produktmomentkorrelation r_{XY} gilt dann:

$$r_{XY} = \frac{m_{11}(XY)}{\sqrt{m_2(X) \cdot m_2(Y)}} = \frac{m'_{11}(XY) - m'_1(X) \cdot m'_1(Y)}{\sqrt{\left(m'_2(X) - \left(m'_1(X)\right)^2\right) \cdot \left(m'_2(Y) - \left(m'_1(Y)\right)^2\right)}}$$

Analog ist die Produktmomentkorrelation ρ_{XY} in der Population bzw. zwischen zwei Zufallsvariablen definiert.

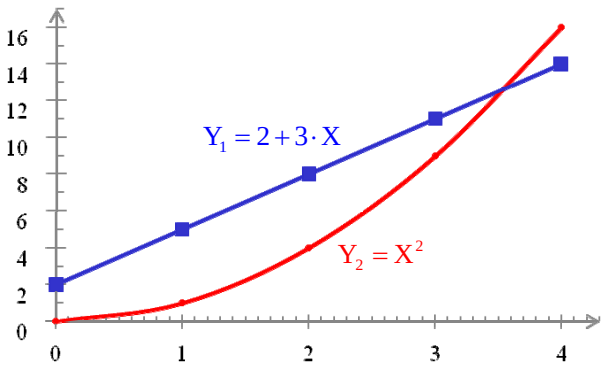
Stärke eines Zusammenhangs	
praktisch kein	$.00 \le \lvert r_{YX}\% \rvert < .05$
geringer	$.05 \le \lvert r_{YX}\% \rvert < .20$
mittlerer	$.20 \le \lvert r_{YX}\% \rvert < .50$
starker	$.50 \le \lvert r_{YX}\% \rvert$
perfekter Zus.	$1.0 = \lvert r_{YX}\% \rvert$

Für die Interpretation der Produktmomentkorrelation gelten analoge Faustregeln wie für Φ: Vorzeichenbereinigte Werte < 0.05 sind vernachlässigbar, Werte bis <0.2 gering, Werte bis < .5 mittelstark und darüber liegende Werte stark.

Produktmomentkorrelation als symmetrisches Zusammehangsmaß

Der Maximalwert einer perfekten Produktmomentkorrelation wird nur erreicht, wenn alle Datenpunkte exakt auf einer Gerade liegen.
Da sich jede Gerade durch eine lineare Gleichung der Form Y= a + b·X beschreiben lässt, bedeutet dies, dass die Produktmomentkorrelation das Ausmaß der Linearität des Zusammenhangs zwischen zwei Variablen erfasst.

Darüber hinaus erfasst die Produktmomentkorrelation aber auch nichtlineare Beziehungen, soweit sie insgesamt einen positiven (ansteigenden) bzw. negativen (fallenden) Trend beinhalten.
Wenn etwa eine perfekte quadratische Beziehung der Form $Y=X^2$ vorliegt und X nur nicht-negative Werte aufweist, liegt eine positive Korrelation vor, die aber nicht perfekt ist, da der Anstieg nicht-linear ist:



	X	X^2	Y_1	Y_1^2	$X \cdot Y_1$	Y_2	Y_2^2	$X \cdot Y_2$
	0	0	2	4	0	0	0	0
	1	1	5	25	5	1	1	1
	2	4	8	64	16	4	16	8
	3	9	11	121	33	9	81	27
	4	16	14	196	56	16	256	64
Σ	10	30	40	410	110	30	354	100

$$r_{XY_1} = \frac{110 - 5 \cdot 2 \cdot 8}{\sqrt{\left(30 - 5 \cdot 2^2\right) \cdot \left(410 - 5 \cdot 8^2\right)}} = 1$$

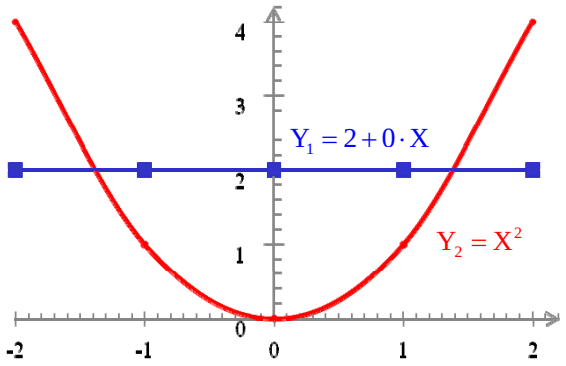
$$r_{XY_2} = \frac{100 - 5 \cdot 2 \cdot 6}{\sqrt{\left(30 - 5 \cdot 2^2\right) \cdot \left(354 - 5 \cdot 6^2\right)}} = 0.959$$

Produktmomentkorrelation als symmetrisches Zusammehangsmaß

Wenn es keinen insgesamt ansteigenden oder fallenden Trend in der Beziehung zwischen den Realisierungen zweier Variablen gibt, ist die Produktmomentkorrelation Null.
Dies ist insbesondere immer dann der Fall, wenn die beiden Variablen statistisch unabhängig voneinander sind:

- Bei statistischer Unabhängigkeit gilt: $r_{YX} = 0$.

Das Umgekehrte gilt nicht:
Selbst bei einer perfekten Beziehung kann die Produktmomentkorrelation Null sein. Dies gilt etwa bei einem strikt nichtmonotonen Zusammenhang, wie er etwa durch die quadratische Gleichung $Y=X^2$ ausgedrückt wird, wenn die Verteilung symmetrisch um Null variiert.



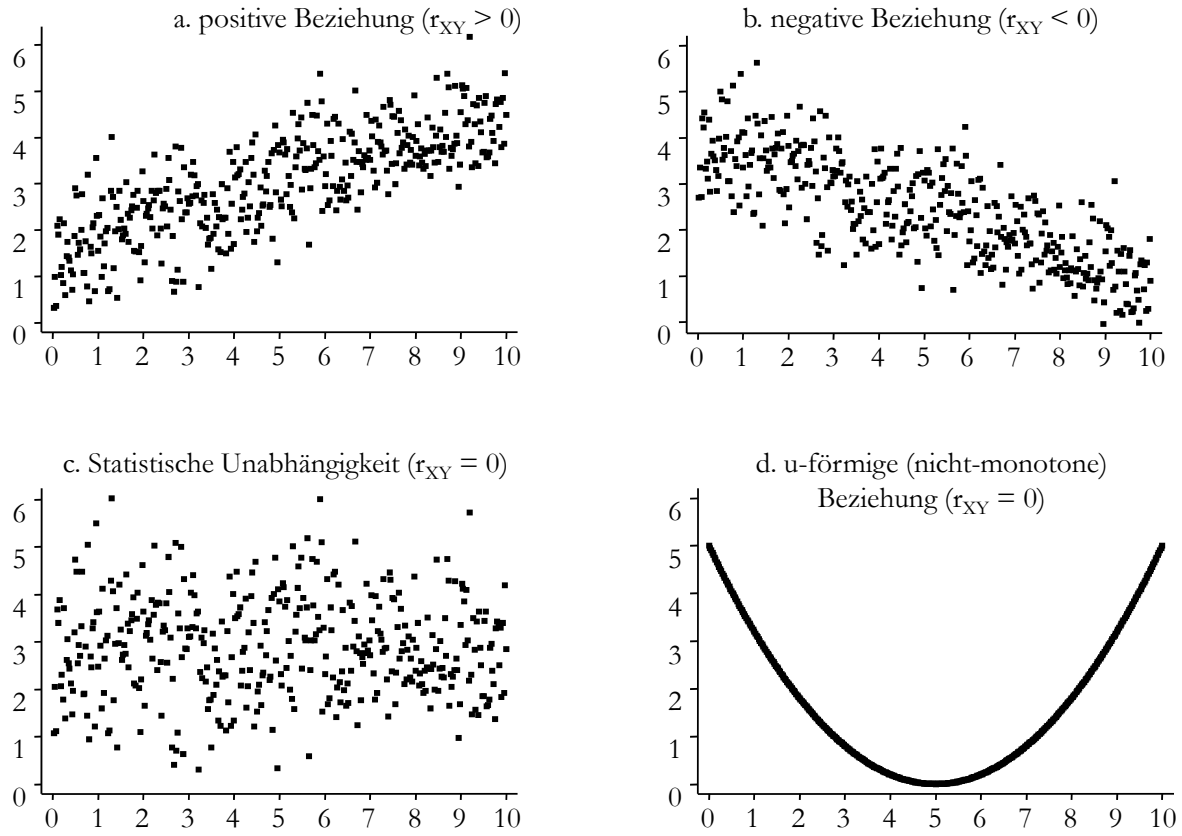
	X	X^2	Y_1	Y_1^2	$X \cdot Y_1$	Y_2	Y_2^2	$X \cdot Y_2$
	−2	4	2	4	−4	4	16	−8
	−1	1	2	4	−2	1	1	−1
	0	0	2	4	0	0	0	0
	1	1	2	4	2	1	1	1
	2	4	2	4	4	4	16	8
$\sum$	0	10	10	20	0	10	34	0

$$r_{XY_1} = \frac{0 - 5 \cdot 0 \cdot 2}{\sqrt{\left(10 - 5 \cdot 0^2\right) \cdot \left(20 - 5 \cdot 2^2\right)}} = 0$$

$$r_{XY_2} = \frac{0 - 5 \cdot 0 \cdot 2}{\sqrt{\left(10 - 5 \cdot 0^2\right) \cdot \left(34 - 5 \cdot 2^2\right)}} = 0$$

Die folgende Folie zeigt Punktewolken für verschiedene Produktmomentkorrelationen.

Produktmomentkorrelation als symmetrisches Zusammehangsmaß



Schätzung und Tests von symmetrischen Zusammenhängen

Erwartungstreuer Schätzer der Populationskovarianz

Ähnlich wie die Stichprobenvarianz s^2_X kein erwartungstreuer Schätzer der Populationsvarianz σ^2_X ist, ist auch die Stichprobenkovarianz s_{XY} kein erwartungstreuer Schätzer der Populationskovarianz σ_{XY}.

Allerdings ergibt sich durch Anwendung des bereits bei der Varianz verwendeten Korrekturfaktors n/(n–1) ein bei einfachen Zufallsauswahlen erwartungstreuer Schätzer :

$$\hat{\sigma}_{XY} = \frac{n}{n-1} \cdot s_{XY} = \frac{SP_{XY}}{n-1} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Schätzung und Test von Populationskorrelationen

Der Korrekturfaktor hilft nicht bei der Schätzung der Produktmomentkorrelation. Da die Produktmomentkorrelation ein Quotient ist, bleibt der Wert unverändert, wenn im Zähler und Nenner jeweils erwartungstreue Schätzer der Kovarianz und der Varianzen verwendet werden. Tatsächlich gibt es keine leicht berechenbare Statistik, die ein erwartungstreuer Schätzer der Populationskorrelation ist. Allerdings ist Pearsons Korrelationskoeffizient bei einfachen Zufallsauswahlen ein konsistenter und asymptotisch erwartungstreuer Schätzer.

Zum Prüfen der Nullhypothese, dass die Produktmomentkorrelation in der Population gleich Null, kleiner/gleich Null oder größer/gleich Null ist, kann folgende Teststatistik verwendet werden:

$$T = r_{XY} \cdot \sqrt{\frac{n-2}{1-r^2_{XY}}}$$

Schätzung und Tests von Produktmomentkorrelationen

Wenn die Nullhypothese zutrifft, die Populationkorrelation also Null ist, ist die Teststatistik mit df=n–2 Freiheitsgraden t-verteilt.
Die Nullhypothese wird daher abgelehnt, wenn die Teststatistik beim zweiseitigen Test kleiner oder gleich dem (α/2)-Quantil bzw. größer oder gleich dem (1–α/2)-Quantil der T-Verteilung ist. Bei einem einseitigen Test nach unten ist der kritische Wert das α-Quantil, bei einem einseitigen Test nach oben das (1–α)-Quantil der T-Verteilung .

Der Test basiert auf der Annahme, dass die beiden Variablen X und Y in der Population normalverteilt sind, ist allerding recht robust gegenüber Verletzung der Verteilungsannahmen.

Anwendungsbeispiel

Als Beispiel werden die bereits grafisch vorgestellten Daten des Alters der 185 Fälle von nicht zusammen lebenden Paaren aus dem Allbus 2006 analysiert.
Zur Berechnung werden neben der Zahl der gültigen Fälle, die Summen über alle Realisationen, die Quadratsummen und die Produktsummen der beiden Variablen benötigt.

Fall	Alter Mann		Alter Frau		Produkt
ID	X	X^2	Y	Y^2	X·Y
1	21	441	21	441	441
2	55	3025	53	2809	2915
…	…	…	…	…	…
185	61	3721	62	3844	3782
∑	6575	272579	6060	232676	249462

Anwendungsbeispiel

Fall	Alter Mann		Alter Frau		Produkt
ID	X	X^2	Y	Y^2	X·Y
1	21	441	21	441	441
2	55	3025	53	2809	2915
…	…	…	…	…	…
185	61	3721	62	3844	3782
∑	6575	272579	6060	232676	249462

$$\bar{x} = \frac{6575}{185} = 35.5405 \; ; \; \bar{y} = \frac{6060}{185} = 32.7568$$

$$s_X^2 = \frac{272579}{185} - \left(\frac{6575}{185}\right)^2 = 210.2700$$

$$s_Y^2 = \frac{232676}{185} - \left(\frac{6060}{185}\right)^2 = 184.7030$$

$$s_{XY} = \frac{249462}{185} - \frac{6575 \cdot 6060}{185^2} = 184.2504$$

Zunächst werden Mittelwerte und Variationen und die Kovariation bzw. Varianzen und die Kovarianz berechnet.

Aus den Mittelwerten ist ersichtlich, dass in den Allbus-Daten der männliche Partner mit 35.5 Jahren im Mittel fast 3 Jahre älter ist als seine Partnerin.
Gleichzeitig weist die positive Kovarianz auf einen positiven Zusammenhang hin, was sich bereits in der Punktewolke gezeigt hatte.

Aus den berechneten Statistiken wird dann die Korrelation berechnet.

$$r_{XY} = \frac{184.2504}{\sqrt{210.2700 \cdot 184.7030}} = 0.935$$

Die Korrelation ist mit einem Wert von 0.935 sehr hoch. Es besteht also eine sehr enge Beziehung zwischen den beiden Variablen.

Anwendungsbeispiel

Fall	Alter Mann		Alter Frau		Produkt
ID	X	X^2	Y	Y^2	X·Y
1	21	441	21	441	441
2	55	3025	53	2809	2915
…	…	…	…	…	…
185	61	3721	62	3844	3782
∑	6575	272579	6060	232676	249462

$$\bar{x} = \frac{6575}{185} = 35.5405 \; ; \; \bar{y} = \frac{6060}{185} = 32.7568$$

$$s_X^2 = \frac{272579}{185} - \left(\frac{6575}{185}\right)^2 = 210.2700$$

$$s_Y^2 = \frac{232676}{185} - \left(\frac{6060}{185}\right)^2 = 184.7030$$

$$r_{XY} = \frac{184.2504}{\sqrt{210.2700 \cdot 184.7030}} = 0.935$$

$$s_{XY} = \frac{249462}{185} - \frac{6575 \cdot 6060}{185^2} = 184.2504$$

Das Beispiel der Allbus-Daten verdeutlicht, dass eine hohe Korrelation zwischen zwei Variablen nicht bedeutet, dass die Verteilungen sich auch in ihren Mittelwerten und Streuungen ähnlich sein müssen.

Im Beispiel unterscheiden sich sowohl die Mittelwerte wie auch die Varianzen.

Mit einer Irrtumswahrscheinlichkeit von 1% soll geprüft werden, dass die Korrelation auch in der Population größer Null ist.

Schritt 1: Formulierung der Hypothesen

Im Sinne eines strengen Testens wird die Forschungshypothese, dass es einen positiven Zusammenhang gibt, als Alternativhypothese formuliert.

$$H_0: \rho_{XY} \leq 0 \text{ vs. } H_1: \rho_{XY} > 0$$

Anwendungsbeispiel

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung

Beim Test einer Produktmomentkorrelation auf Null ist die Teststatistik

$$T = r_{XY} \cdot \sqrt{\frac{n-2}{1-r_{XY}^2}}$$

mit df=n–2 Freiheitsgraden t-verteilt.
Ist die Populationskorrelation ungleich Null, ist mit einem von Null abweichendem Erwartungswert zu rechnen.

Schritt 3: Festlegung von Irrtumswahrscheinlichkeit und kritischen Werten

Wenn die Alternativhypothese (Forschungshypothese) zutrifft, ist eher mit positiven Werten der Teststatistik zu rechnen. Bei einer Irrtumswahrscheinlichkeit von 1% wird die Nullhypothese daher abgelehnt, wenn die Teststatistik kleiner oder gleich dem 99%-Quantil der T-Verteilung mit df=183 Freiheitsgraden ist. Der kritische Wert liegt zwischen 2.326 bei df=∞ und 2.358 bei df=120.

Schritt 4: Berechnung der Teststatistik und Entscheidung

$$T = r_{XY} \cdot \sqrt{\frac{n-2}{1-r_{XY}^2}} = 0.935 \cdot \sqrt{\frac{183}{1-0.935^2}} = 38.14$$

Da der Wert deutlich größer ist als die Teststatistik, kann davon ausgegangen werden, dass es auch in der Grundgesamtheit eine positive Beziehung zwischen dem Alter der beiden Partner gibt.

Anwendungsbeispiel

Schritt 5: Anwendungsvoraussetzungen

Der Test setzt Normalverteilung der beiden Variablen voraus. Die Randverteilungen in der oben wiedergegebene Kreuztabelle der beiden Variablen spricht eher gegen diese Annahme. Allerdings ist der Test i.a. robust.
Der Allbus ist auch keine einfache Zufallsauswahl. Da eine disproportionale Stichprobe gezogen wurde, muss unterstellt werden, dass die Beziehung zwischen den beiden Variablen in den beiden Erhebungsgebieten alte und neue Bundesländer gleich ist.

Lerneinheit 19:
Asymmetrische Beziehung zwischen zwei metrischen Variablen

In asymmetrischen Beziehungen werden die bedingten Verteilungen der abhängigen Variable bei verschiedenen Ausprägungen der erklärenden Variable verglichen.
Die vollständige Darstellung aller Ausprägungskombinationen in einer bivariaten Kreuztabelle führt bei metrischen Variablen dazu, dass bei den üblichen Stichprobenumfängen viele Zellen nur wenige oder gar keine Fälle enthalten.

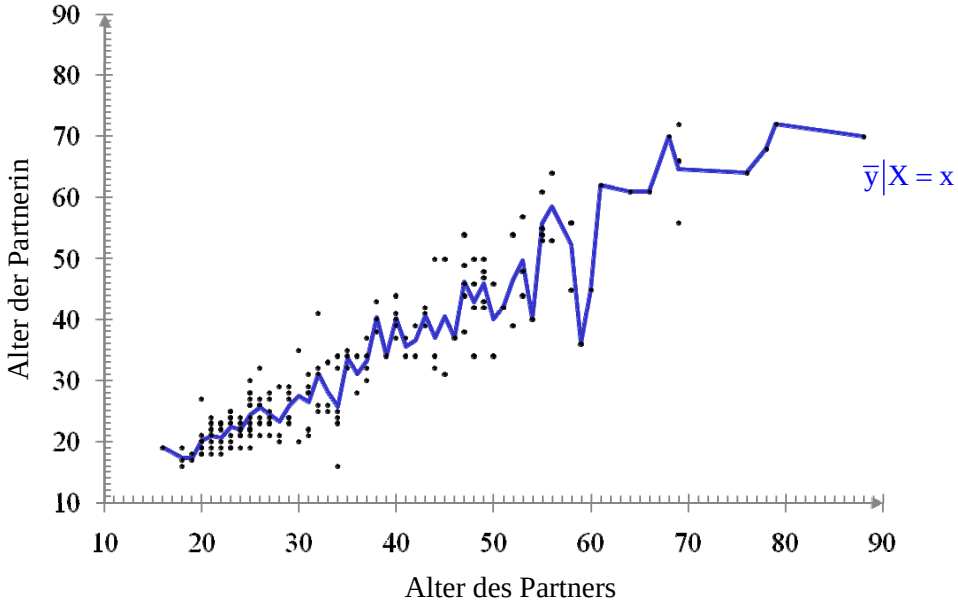
Daher wird bei der Analyse asymmetrischer Beziehungen zwischen zwei metrischen Variablen davon abgesehen, die gesamte bedingte Verteilung zu betrachten. Stattdessen werden Kennwerte der bedingten Verteilungen analysiert. So werden im ***Regressionsmodell*** die ***bedingten Mittelwerte*** der abhängigen Variable als ***Funktion*** der Ausprägungen ***der erklärenden Variable*** dargestellt.

Bezogen auf die bedingten Populationsmittelwerte gilt also:

$$\mu(Y|X) = \mu_{Y|X} = g(X = x)$$

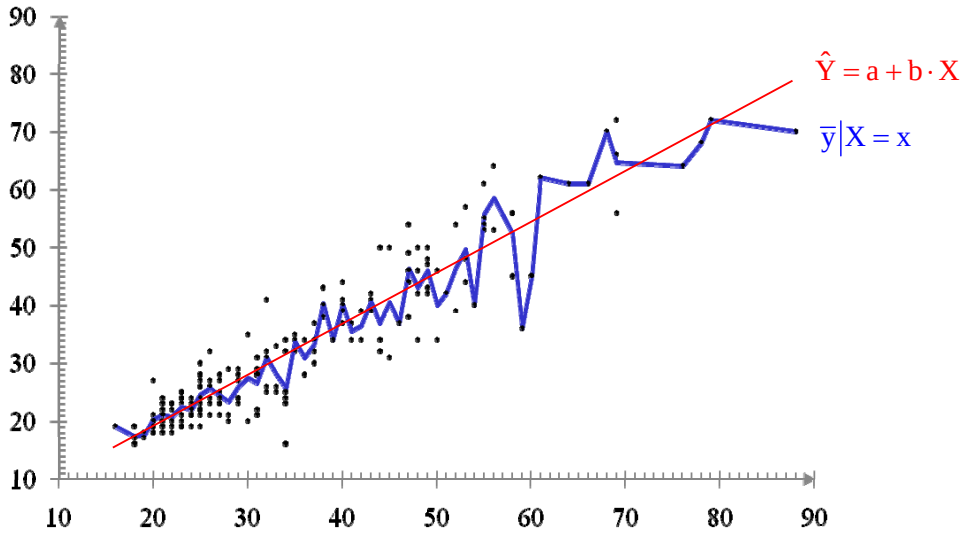
Als Beispiel wird der gerichtete Zusammenhang zwischen dem Alter der Partner von insgesamt 185 Befragten aus dem Allbus 2006 betrachtet, die eine Lebenspartnerin bzw. einen Lebenspartner haben, mit dem sie nicht zusammenleben. Da nur das Geschlecht der befragten Person und nicht ihres Partners erfasst wurde, wird in den folgenden Analysen unterstellt, dass es sich um heterosexuelle Partnerschaften handelt.

Empirische Regressionsfunktion



Die resultierende ***empirische Regressionsfunktion*** steigt tendenziell an, hat aber eine sehr unregelmäßige Form, was vermutlich darauf zurückzuführen ist, dass für eine Ausprägung der erklärenden Variablen oft nur ein Fall oder sehr wenige Fälle für die Berechnung des jeweiligen bedingten Mittelwerts zur Verfügung stehen. Lägen Populationsdaten vor, würde sich möglicherweise eine glatte Kurve ergeben.

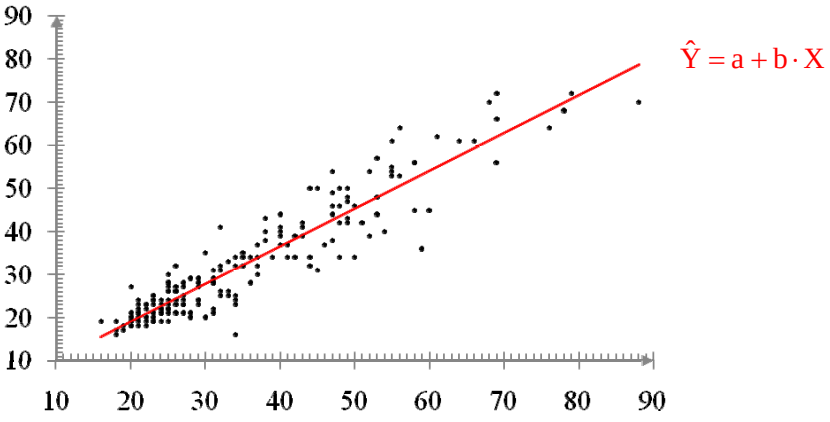
Lineare Regression



Es liegt nahe, anstelle der bedingten Stichprobenmittelwerte aus den Daten eine solche „glatte“ Regressionsfunktion zu schätzen.
Dazu muss die Kurvenform bekannt sein. Sehr oft angewendet und möglicherweise auch hier angemessen ist die Schätzung einer linearen Regressionsfunktion. Die mathematische Gleichung der ***linearen Regressionsfunktion*** lautet:

$$\hat{\mu}_{Y|X} = \hat{Y} = a + b \cdot X$$

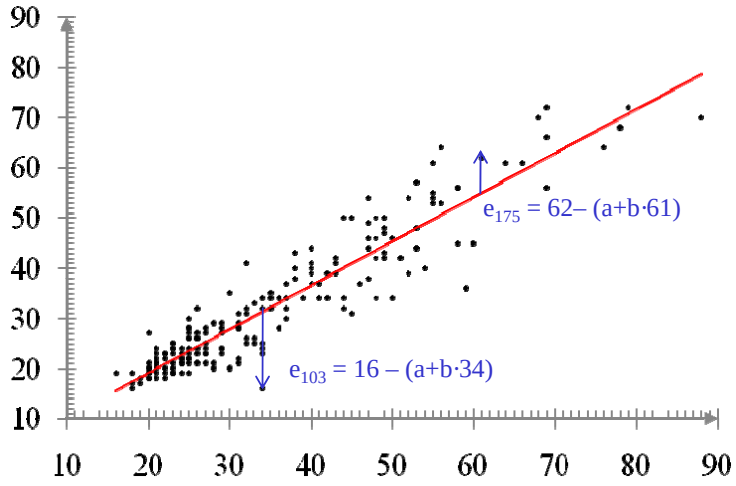
OLS-Schätzung der linearen Regressionsfunktion



Die geschätzten bedingten Mittelwerte der Regressionsfunktion werden im Regressionsmodell auch als ***Vorhersagewerte*** $\hat{Y}$ bezeichnet.

Bei der Berechnung der ***Regressionskoeffizienten*** **a** und **b** dieser ***linearen Regressionsfunktion*** aus Stichprobendaten wird die Eigenschaft eines Mittelwerts, dass er die Summe der quadrierten Abweichungen aller Fälle von ihm minimiert, auf die gesamte Regressionsfunktion übertragen. In diesem Sinne wird in der sog. ***Kleinstquadratmethode*** (nach der englischen Bezeichnung *„ordinary least squares method“* oft ***OLS-Methode*** oder ***OLS- Schätzung*** genannt) die Summe der quadrierten Residuen minimiert.

OLS-Schätzung der linearen Regressionsfunktion



Beim 103-ten Fall in der Stichprobe ist der Partner 34 Jahre alt und die Partnerin 16 Jahre.
$\Rightarrow e_{103} = 16 - (a+b \cdot 34)$

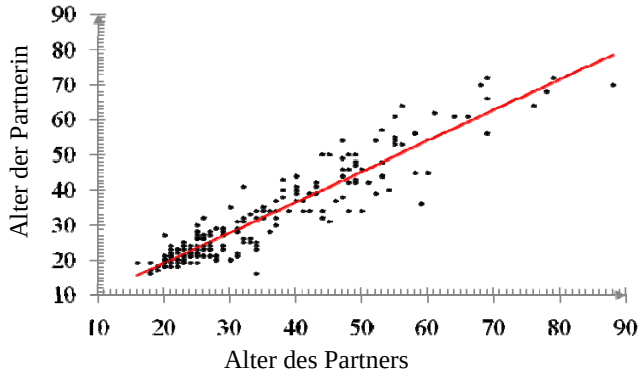
Ein ***Residuum*** e_i ist die Differenz zwischen der Realisation des i-ten Falles der abhängigen Variable von seinem Vorhersagewert:

$$e_i = y_i - \hat{y}_i = y_i - (a + b \cdot x_i)$$

Bei der OLS-Schätzung werden die Regressionskoeffizienten a und b so bestimmt, dass die Summe der quadrierten Residuen minimal ist:

$$Q(a,b) = \sum_{i=1}^{n} \left(y_i - (a + b \cdot x_i)\right)^2 = \sum_{i=1}^{n} e_i^2 \overset{!}{=} \text{minimal}$$

OLS-Schätzung der Regressionskoeffizienten



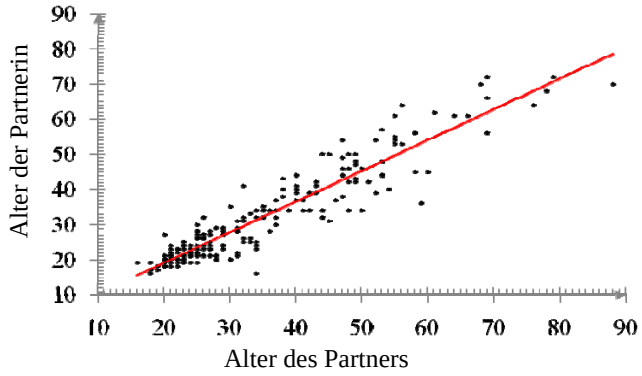
Die Regressionskonstante a ist dann die Differenz von b mal dem Stichprobenmittelwert der erklärenden Variable vom Stichprobenmittelwert der abhängigen Variable:

$$a = \bar{y} - b \cdot \bar{x}$$

Mit Hilfe der Differentialrechnung kann gezeigt werden, dass die ***Minimierungsfunktion Q(a,b)*** genau dann ein Minimum aufweist, wenn b der Quotient aus der Kovariation bzw. Kovarianz zwischen abhängiger und erklärender Variable geteilt durch die Variation bzw. Varianz der erklärenden Variable ist:

$$b = \frac{SP_{YX}}{SS_X} = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} y_i \cdot x_i - \frac{\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} y_i \cdot x_i - n \cdot \bar{y} \cdot \bar{x}}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2} = \frac{s_{YX}}{s_X^2} = \frac{\hat{\sigma}_{YX}}{\hat{\sigma}_X^2}$$

OLS-Schätzung der Regressionskoeffizienten



Fall	X	Y	X^2	Y^2	X·Y
1	21	21	441	441	441
...	...	...	...	…	...
185	61	62	3721	3844	3782
∑	6575	6060	272579	232676	249463

Für die Berechnung werden die Summen, Quadratsummen und Produktsummen der Realisierungen von abhängiger und unabhängiger Variable benötigt.

Aus ihnen können dann die Mittelwerte, Variationen und Kovariationen bzw. anstelle der Variationen und Kovariationen die Varianzen und Kovarianzen berechnet werden:

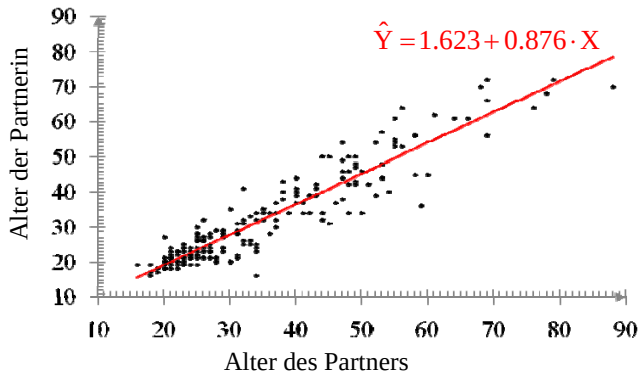
$$\bar{x} = 6575/185 = 35.5405 \; ; \; \bar{y} = 6060/185 = 32.7568$$

$$SS_X = 272579 - 6575^2/185 = 38899.9460 \qquad s_X^2 = 38899.9460/185 = 210.2700$$

$$SS_Y = 232676 - 6060^2/185 = 34170.0541 \qquad s_Y^2 = 34170.0541/185 = 184.7030$$

$$SP_{YX} = 249463 - 6575 \cdot 6060/185 = 34087.3243 \qquad s_{YX} = 34087.3243/185 = 184.2558$$

OLS-Schätzung der Regressionskoeffizienten



$$\bar{x} = 5.5405 \; ; \; \bar{y} = 32.7568$$

$$SS_X = 38899.9460$$

$$SS_Y = 34170.0541$$

$$SP_{YX} = 34087.3243$$

$$s_X^2 = 38899.9460/185 = 210.2700$$

$$s_Y^2 = 34170.0541/185 = 184.7030$$

$$s_{YX} = 34087.3243/185 = 184.2558$$

Aus den Statistiken werden schließlich die Regressionskoeffizienten berechnet:

$$b = \frac{SP_{YX}}{SS_X} = \frac{34087.3243}{38899.9460} = 0.876 = \frac{s_{YX}}{s_X^2} = \frac{184.2558}{210.2700} = 0.876$$

$$a = 32.7568 - 0.876 \cdot 35.5405 \approx 1.623$$

Die gleichen Werte ergeben sich auch, wenn anstelle der Stichprobenvarianzen und Kovarianzen die geschätzten Populationsvarianzen und Kovarianzen verwendet werden:

$$\hat{\sigma}_X^2 = 38899.9460/184 = 211.4127$$

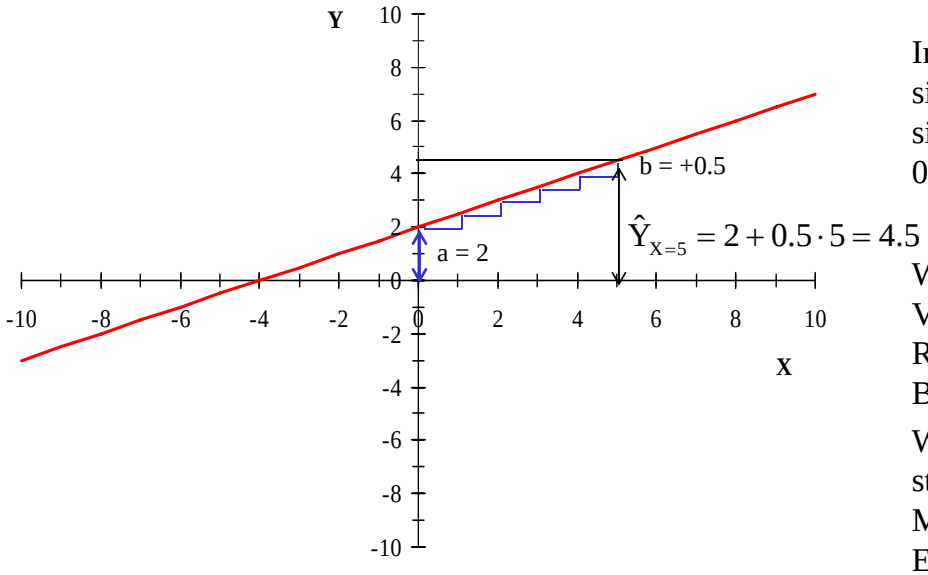
$$\hat{\sigma}_Y^2 = 34170.0541/184 = 185.7068$$

$$\hat{\sigma}_{YX} = 34087.3243/184 = 185.2572$$

$$b = \frac{\hat{\sigma}_{YX}}{\hat{\sigma}_X^2} = \frac{185.2572}{211.4127} = 0.876$$

Bedeutung der Regressionskoeffizienten

Die geschätzten ***Regressionskoeffizienten a*** und ***b*** der linearen Regressionsfunktion bestimmen die Lage der Regressionsgerade und damit der Vorhersagewerte im Koordinatensystem.



Im Beispiel ist eine Regressionsgerade mit der Regressionskonstante a = 2 und b = 0.5 eingezeichnet:

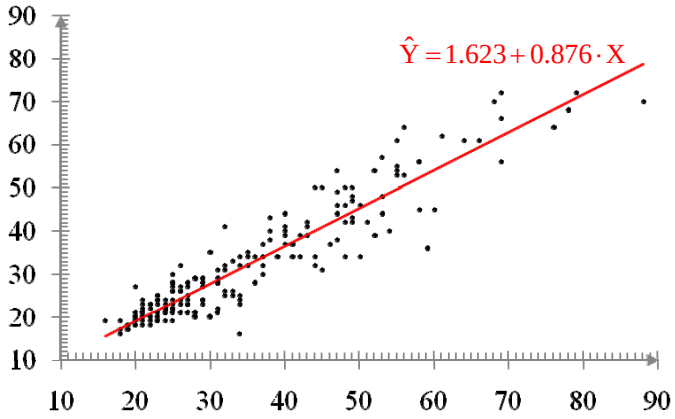
$$\hat{Y} = 2 + 0.5 \cdot X$$

Wenn X = 0, dann ist der Vorhersagewert gleich der Regressionskonstante, im Beispiel = 2

Wenn X um +1 Einheit ansteigt, steigt der bedingte Mittelwert von Y um +0.5 Einheiten an.

- Die ***Regressionskonstante*** (auch als ***Interzept*** bezeichnet) **a** gibt den geschätzten Mittelwert von Y an, wenn X = 0 ist.
- Das ***Regressionsgewicht*** **b** gibt die ***Steigung*** der Geraden an.

Interpretation der linearen Regression



$$b_{Y.X} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} = \frac{s_{XY}}{s_X^2} = \frac{SP_{XY}}{SS_X}$$

$$= \frac{184.2504}{210.2700} = 0.876$$

$$a_{Y.X} = \bar{y} - b \cdot \bar{x}$$

$$= 32.76 - 0.876 \cdot 35.54 = 1.623$$

Die geschätzte lineare Regressionsfunktion, die bisweilen auch als ***Trendlinie*** bezeichnet wird, kann daher zur ***Vorhersage*** (Prognose) der Ausprägungen der abhängigen Variable herangezogen werden.

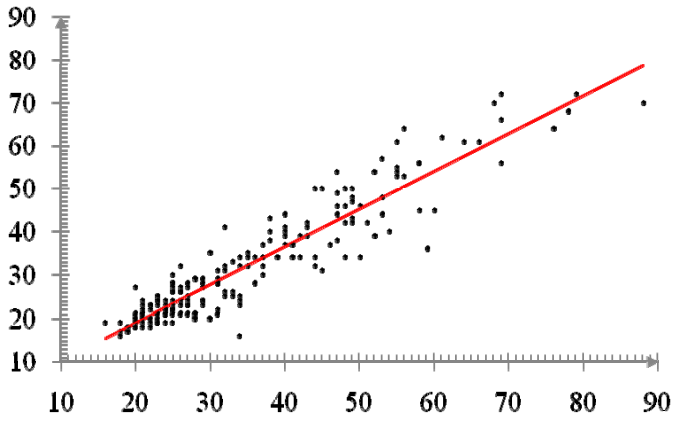
Aus den Schätzungen a=1.623 und b=0.876 des Allbus-Beispiels berechnet sich z.B. das erwartete durchschnittliche Alter der Partnerinnen von 35jährigen Männern als:

1.623 + 0.876·35 = 32.29 Jahre.

Über die Regressionsfunktion können auch Vorhersagewerte für nicht im Datensatz vorkommende Ausprägungen der unabhängigen Variable berechnet werden.

So gibt es im Allbus keinen Fall mit einem 85jährigen Partner. Erwartbar wäre für einen solchen Fall ein Alter der Partnerin von 1.623 + 0.876·85 = 76.1 Jahren.

Interpretation der linearen Regression

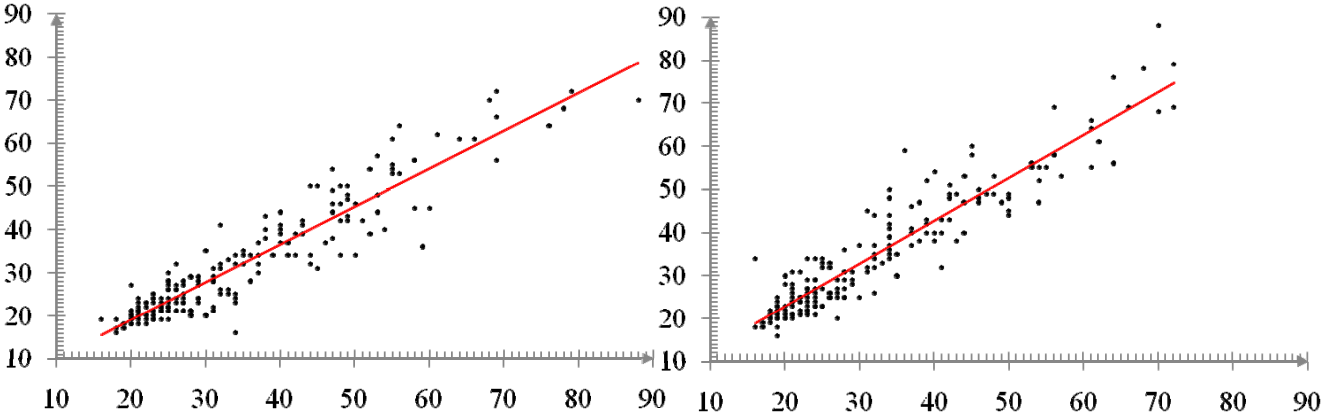


$$b_{Y.X} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} = \frac{s_{XY}}{s_X^2} = \frac{SP_{XY}}{SS_X}$$
$$= \frac{184.2504}{210.2700} = 0.876$$
$$a_{Y.X} = \bar{y} - b \cdot \bar{x}$$
$$= 32.76 - 0.876 \cdot 35.54 = 1.623$$

Bei einer ***kausalen Interpretation*** ist in erster Linie das Regressionsgewicht b von Interesse. Wenn X um +1 Einheit ansteigt, dann steigt im Durchschnitt Y um b Einheiten an. Das Regressionsgewicht **b** wird daher auch als ***Effektstärke*** bezeichnet.

Da es nicht sein kann, dass das Alter der Partnerin um b = 0.876 Jahre ansteigt, wenn das Alter des Partners um +1 Jahr steigt, ist eine direkte Kausalinterpretation im Beispiel allerdings nicht sinnvoll.
Denkbar ist jedoch, dass mit steigendem Alter männliche Partner eher relativ jüngere Partnerinnen präferieren bzw. umgekehrt Frauen mit zunehmenden Alter einen größeren Altersabstand präferieren.

Interpretation der linearen Regression



Regression von Y auf X

$$b_{Y.X} = \frac{s_{XY}}{s_X^2} = \frac{184.2504}{210.2700} = 0.876$$

$$a_{Y.X} = 32.76 - 0.876 \cdot 35.54 = 1.623$$

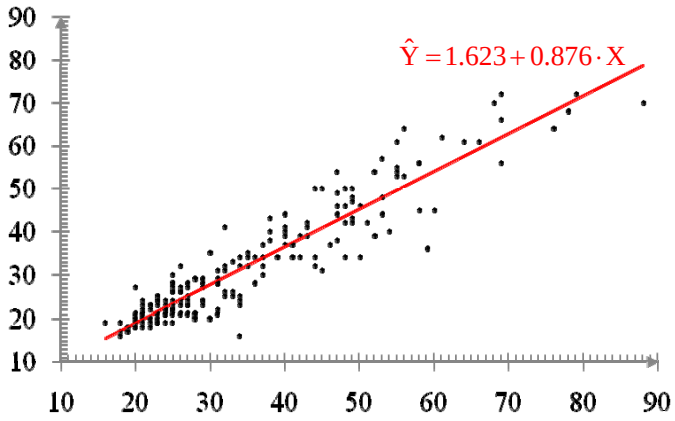
Regression von X auf Y

$$b_{X.Y} = \frac{s_{XY}}{s_Y^2} = \frac{184.2504}{184.7030} = 0.998$$

$$a_{X.Y} = 35.54 - 0.876 \cdot 32.76 = 2.864$$

Die Umkehrung der Beziehung durch die Regression des Partners (X) auf das Alter der Partnerin (Y) zeigt jedoch, dass das Regressionsgewicht $b_{X.Y}$ ebenfalls (geringfügig) < 1 ist, was gegen diese Vermutung spricht. Ein Regressionsgewicht ungleich 1.0 bei den beiden Altersvariablen kann aber auch ein Ergebnis der Regression zur Mitte sein, also durch Messfehler (mangelnde Reliabilität) hervorgerufen sein.

Interpretation der linearen Regression

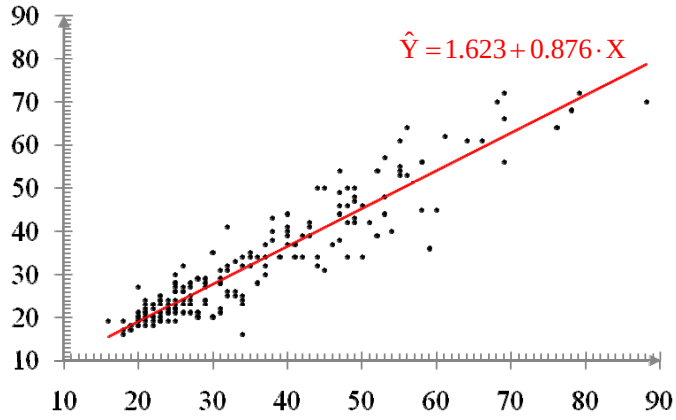


$$b_{Y.X} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} = \frac{s_{XY}}{s_X^2} = \frac{SP_{XY}}{SS_X}$$
$$= \frac{184.2504}{210.2700} = 0.876$$
$$a_{Y.X} = \bar{y} - b \cdot \bar{x}$$
$$= 32.76 - 0.876 \cdot 35.54 = 1.623$$

Die Kleinstquadratschätzung ermöglicht sowohl eine rein deskriptive auf die Stichprobe bezogene als auch eine inferenzstatistische auf die Population bezogene Interpretation.

- Bezogen auf die Stichprobe beschreibt die Regressionsfunktion, wie die asymmetrische Beziehung zwischen abhängiger und unabhängiger Variable bestmöglich durch eine lineare Gleichung beschrieben werden kann.
- Bezogen auf die Population ist die Regressionsfunktion eine Schätzung der bedingten Populationsmittelwerte, wobei dann allerdings **angenommen** werden muss, dass die bedingten Mittelwerte in der Population tatsächlich eine lineare Funktion der Ausprägungen der unabhängigen Variable sind.

Interpretation der linearen Regression

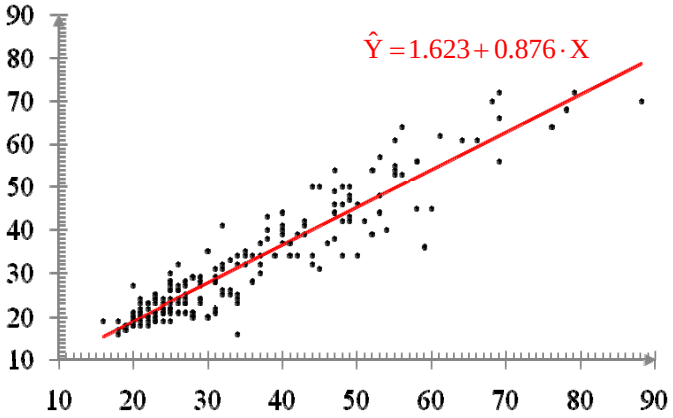


$$b_{Y.X} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} = \frac{s_{XY}}{s_X^2} = \frac{SP_{XY}}{SS_X}$$
$$= \frac{184.2504}{210.2700} = 0.876$$
$$a_{Y.X} = \bar{y} - b \cdot \bar{x}$$
$$= 32.76 - 0.876 \cdot 35.54 = 1.623$$

Bei beiden Sichtweisen bedeutet ein Regressionsgewicht b>0, dass eine positive Beziehung zwischen den beiden Variablen besteht, ein Regressionsgewicht b<0, dass eine negative Beziehung besteht und eine Regressionsgewicht von b=0, dass keine bzw. keine monotone Beziehung besteht.

Es kann gezeigt werden, dass die Regressionskoeffizienten a und b konsistente und erwartungstreue Schätzer der korrespondierenden Parameter in der Population sind, wenn die Regressionsfunktion in der Population tatsächlich linear ist und eine einfache Zufallsauswahl vorliegt. Anstelle der zweiten Bedingung ist es hinreichend anzunehmen, dass die Populationsresiduen, das sind die Differenzen der Realisierungen von den bedingten Populationsmittelwerten, nicht mit der erklärenden Variable korreliert sind.

Berechnung der Standardfehler



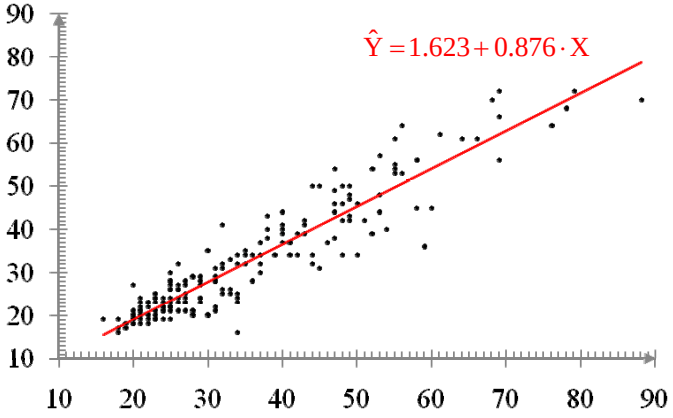
$$b_{Y.X} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} = \frac{s_{XY}}{s_X^2} = \frac{SP_{XY}}{SS_X}$$
$$= \frac{184.2504}{210.2700} = 0.876$$
$$a_{Y.X} = \bar{y} - b \cdot \bar{x}$$
$$= 32.76 - 0.876 \cdot 35.54 = 1.623$$

Wenn zusätzlich die Populationsresiduen untereinander unkorreliert sind und die bedingten Varianzen in der Population für alle Ausprägungen der erklärenden Variablen konstant sind, dann ergibt die Kleinstquadratmethode zudem effiziente lineare Schätzer der Regressionskoeffizienten.

Die OLS-Schätzer sind dann zudem asymptotisch um die zu schätzenden Populationsparameter normalverteilt mit den geschätzten Standardfehlern:

$$\hat{\sigma}(b_{Y.X}) = \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n-2}} = \sqrt{\frac{1}{s_X^2} \cdot \frac{s_Y^2 - b_{Y.X}^2 \cdot s_X^2}{n-2}}$$

Berechnung der Standardfehler



$$\bar{x} = \frac{6575}{185} = 35.5405 \; ; \; \bar{y} = \frac{6060}{185} = 32.7568$$
$$s_X^2 = \frac{272579}{185} - \left(\frac{6575}{185}\right)^2 = 210.2700$$
$$s_Y^2 = \frac{232676}{185} - \left(\frac{6060}{185}\right)^2 = 184.7030$$
$$s_{XY} = \frac{249462}{185} - \frac{6575 \cdot 6060}{185^2} = 184.2504$$

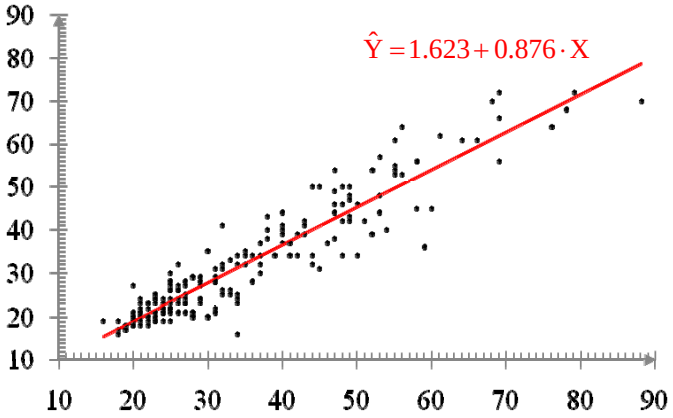
und $$\hat{\sigma}(a_{Y.X}) = \sqrt{\frac{\frac{1}{n}\cdot\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n-2}} = \sqrt{\frac{s_X^2 + \bar{x}^2}{s_X^2} \cdot \frac{s_Y^2 - b_{Y.X}^2 \cdot s_X^2}{n-2}}$$

Die Annäherung an die Normalverteilung ist bei n ≥30 besser n ≥ 50 hinreichend genau.

Für die Allbus-Daten ergibt sich:

$$\hat{\sigma}(b) = \sqrt{\frac{1}{210.2700} \cdot \frac{184.7030 - 0.876^2 \cdot 210.2700}{183}} = 0.025$$

Berechnung der Standardfehler



$$\bar{x} = \frac{6575}{185} = 35.5405 \; ; \; \bar{y} = \frac{6060}{185} = 32.7568$$

$$s_X^2 = \frac{272579}{185} - \left(\frac{6575}{185}\right)^2 = 210.2700$$

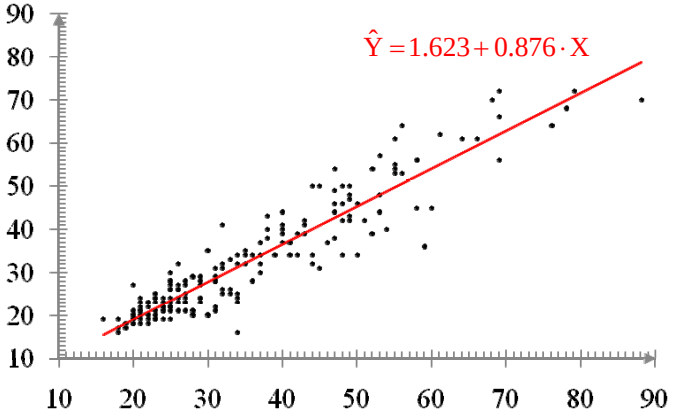
$$s_Y^2 = \frac{232676}{185} - \left(\frac{6060}{185}\right)^2 = 184.7030$$

$$s_{XY} = \frac{249462}{185} - \frac{6575 \cdot 6060}{185^2} = 184.2504$$

$$\hat{\sigma}(a) = \sqrt{\frac{210.2700 + 35.54^2}{210.2700} \cdot \frac{184.7030 - 0.876^2 \cdot 210.2700}{183}} = 0.945$$

Wenn die Differenzen der Realisierungen der abhängigen Variable von den bedingten Populationsmittelwerten normalverteilt sind, dann ist die Standardisierung mit Hilfe des Populationswerts μ(b) = β und des geschätzten Standardfehler nicht nur asymptotisch standardnormalverteilt, sondern exakt t-verteilt mit df = n–2 Freiheitsgraden.
Die asymptotische Normalverteilung bzw. die T-Verteilung kann genutzt werden, um Konfidenzintervalle zu berechnen oder um statistische Tests durchzuführen.

Tests der Regressionskoeffizienten



$$\bar{x} = 35.5405 \; ; \; \bar{y} = 32.7568$$

$$s_X^2 = 210.2700 \; ; \; s_Y^2 = 184.7030$$

$$s_{XY} = 184.2504$$

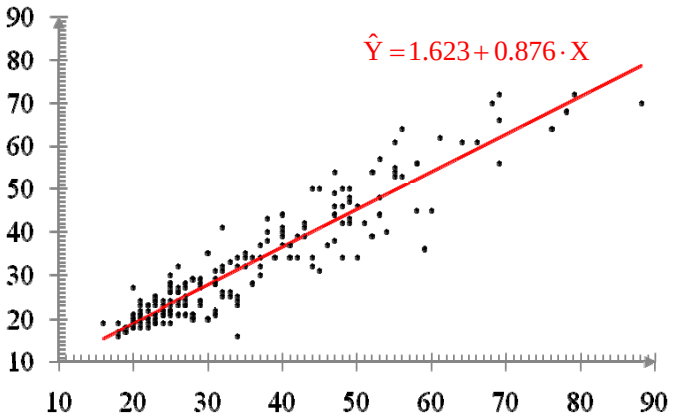
$$\hat{\sigma}(a) = 0.945 \quad \hat{\sigma}(b) = 0.025$$

Da die T-Verteilung verglichen mit der Standardnormalverteilung bei gleicher Irrtunswahrscheinlichkeit zu längeren Konfidenzintervallen und größeren Annahmebereichen der Nullhypothese führt, wird im Sinne eines vorsichtigen (konservativen) Schätzens und Testens auch bei nicht normalverteilten Populationsresiduen die T-Verteilung herangezogen.

Als Beispiel soll in einem einseitigen Hypothesentest geprüft werden, dass das Regressionsgewicht $b_{YX} > 0$ ist. Die Teststatistik beträgt dann:

$$T = \frac{b_{Y.X} - \beta}{\hat{\sigma}(b_{Y.X})} = \frac{0.876 - 0}{0.025} = 35.646$$

Tests der Regressionskoeffizienten



$\bar{x} = 35.5405 \; ; \; \bar{y} = 32.7568$

$s_X^2 = 210.2700 \; ; \; s_Y^2 = 184.7030$

$s_{XY} = 184.2504$

$\hat{\sigma}(a) = 0.945 \quad \hat{\sigma}(b) = 0.025$

Wenn das Regressionsgewicht in der Grundgesamtheit Null ist, ist T asymptotisch normal-verteilt bzw. bei normalverteilten Residuen t-verteilt mit df = n–2 = 183 Freiheitsgraden. Bei einer Irrtumswahrscheinlichkeit von 1% liegt der kritische Wert beim einseitigen Test nach oben zwischen 2.326 und 2.358. Die Nullhypothese ist somit abzulehnen. Vermutlich besteht auch in der Population eine positive Beziehung.

Da die Produktmomentkorrelation und das Regressionsgewicht den gleichen Zähler aufweisen, ist der ein- oder zweiseitige Hypothesentest des Regressionsgewichts auf Null gleichbedeutend mit einem ein- oder zweiseitigem Hypothesentest der Korrelation auf Null. Tatsächlich sind die beiden T-Teststatistiken bis auf Rundungsfehler identisch.

Eigenschaften der Stichprobenresiduen bei der OLS-Methode

Die Stichprobenresiduen e_i können zu einer Variable E zusammengefasst werden.
Aus der Kleinstquadratmethode folgt dann.

(1) Die Summe der Stichprobenresiduen ist Null:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - a - b \cdot x_i = 0$$

(2) Dann ist auch der Mittelwert der Residualvariable E = 0 und die Stichprobenvarianz gleich dem Mittelwert der quadrierten Residuen:

$$\bar{e} = \frac{1}{n} \cdot \sum_{i=1}^{n} e_i = 0 \; ; \; s_E^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} e_i^2$$

Da die Varianz der Stichprobenresiduen gleich dem Mittelwert der quadrierten Residuen ist, minimiert die OLS-Methode also auch die Residualvarianz in der Stichprobe.

(3) Die Residualvariable E ist mit der erklärenden Variablen und den Vorhersagewerten unkor-reliert:

$$SP_{XE} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot e_i = \sum_{i=1}^{n} x_i \cdot e_i = 0 \Rightarrow s_{XE} = \frac{SP_{XE}}{n} = 0 \; ; \; r_{XE} = \frac{s_{XE}}{s_X \cdot s_E} = 0$$

$$SP_{\hat{Y}E} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}) \cdot e_i = \sum_{i=1}^{n} y_i \cdot e_i = 0 \Rightarrow s_{\hat{Y}E} = \frac{SP_{\hat{Y}E}}{n} = 0 \; ; \; r_{\hat{Y}E} = \frac{s_{\hat{Y}E}}{s_{\hat{Y}} \cdot s_E} = 0$$

Varianzzerlegung der abhängigen Variable

Bei der OLS-Schätzung der Regressionskoeffizienten kann die abhängige Variable Y auch als eine Linearkombination aus X und E aufgefasst werden:

$$Y = \hat{Y} + E = a + b \cdot X + E$$

(4) Aus der Unkorreliertheit der Stichprobenresiduen mit den Vorhersagewerten und den Regeln für Linearkombinationen von unabhängigen Variablen folgt dann, dass die Varianz bzw. Variation der abhängigen Variable gleich der Summe aus den Varianzen bzw. Variationen der Vorhersagewerte und der Stichprobenresiduen ist:

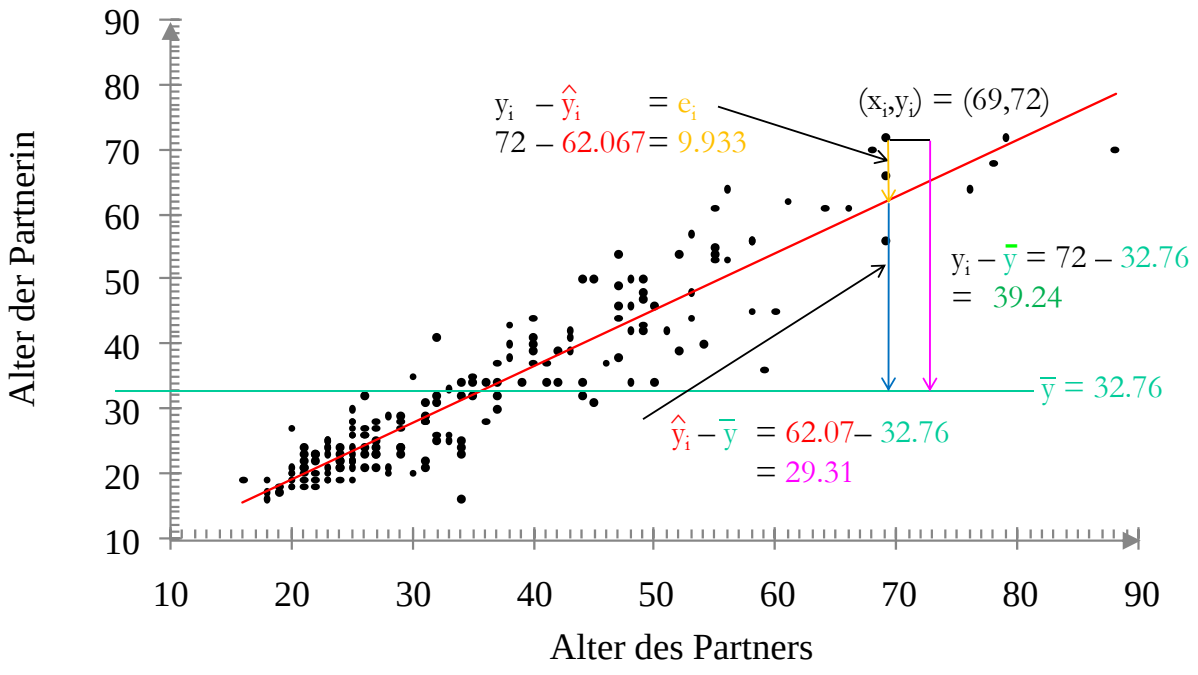
$$SS_Y = SS_{\hat{Y}} + SS_E = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$s_Y^2 = s_{\hat{Y}}^2 + s_E^2 = \frac{1}{n} \cdot \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n} \cdot \sum_{i=1}^{n} e_i^2 = \frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Als PRE-Maß für die Stärke der Beziehung bietet es sich daher an, die Variationsverhältnisse in Beziehung zu setzen. Das so gebildete asymmetrische Zusammenhangsmaß ist der ***Determinationskoeffizient*** R^2*:*

$$R^2 = 1 - \frac{s_E^2}{s_Y^2} = 1 - \frac{SS_E}{SS_Y} = \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad = \quad \frac{s_{\hat{Y}}^2}{s_Y^2} = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Variationszerlegung



In der grafischen Darstellung ist gut erkennbar, wie bei den Fällen die Distanzen der abhängigen Variable zur Regressionsgerade meistens geringer sind als zum Mittelwert der abhängigen Variable.

Determinationskoeffizient, Korrelation und Regressionsgewichte

Zwischen dem Determinationskoeffizient R^2, der Produktmomentkorrelation r_{YX} und den beiden Regressionsgewichten $b_{Y.X}$ der Regressionen von Y auf X und $b_{X.Y}$ der Regression von X auf Y gibt es Zusammenhänge:

- Der Determinationskoeffizient ist das Quadrat der Produktmomentkorrelation:

$$R^2 = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(a + b \cdot x_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}((\bar{y} - b \cdot \bar{x}) + b \cdot x_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(b \cdot (x_i - \bar{x}))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= b^2 \cdot \frac{SS_X}{SS_Y} = b^2 \cdot \frac{s_X^2}{s_Y^2} = \frac{(s_{YX})^2}{(s_X^2)^2} \cdot \frac{s_X^2}{s_Y^2} = \frac{(s_{YX})^2}{s_x^2 \cdot s_Y^2} = \left(\frac{s_{YX}}{s_X \cdot s_Y}\right)^2 = (r_{YX})^2$$

- Der Determinationskoeffizient ist das Produkt und die Produktmomentkorrelation das geometrische Mittel der Regressionsgewichte der Regressionen von Y auf X und von X auf Y.
  Um die Richtung der Regression zu unterscheiden, werden die Regressionskoeffizienten mit Indizes versehen:

$$Y = a_{Y.X} + b_{Y.X} \cdot X + E_Y \Rightarrow b_{Y.X} = \frac{SP_{YX}}{SS_X} = \frac{s_{YX}}{s_X^2} = \frac{\hat{\sigma}_{YX}}{\hat{\sigma}_X^2} \; ; \; a_{Y.X} = \bar{y} - b_{Y.X} \cdot \bar{x}$$

$$X = a_{X.Y} + b_{X.Y} \cdot Y + E_X \Rightarrow b_{X.Y} = \frac{SP_{XY}}{SS_Y} = \frac{s_{XY}}{s_Y^2} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_Y^2} \; ; \; a_{X.Y} = \bar{x} - b_{X.Y} \cdot \bar{y}$$

Determinationskoeffizient, Korrelation und Regressionsgewichte

Für den Determinationskoeffizient gilt dann:

$$R^2 = \frac{(s_{YX})^2}{s_X^2 \cdot s_Y^2} = \frac{s_{YX}}{s_X^2} \cdot \frac{s_{YX}}{s_Y^2} = b_{Y.X} \cdot b_{X.Y} \; ; \; r_{YX} = \sqrt{b_{Y.x} \cdot b_{X.Y}}$$

- Werden abhängige und erklärende Variable über die Z-Transformation standardisiert, dann sind die Regressionskonstanten Null und die resultierenden ***standardisierten Regressionsgewichte*** gleich der Produktmomentkorrelation.
  Dies gilt sowohl für die Regression von Y auf X als auch für die Regression von X auf Y.

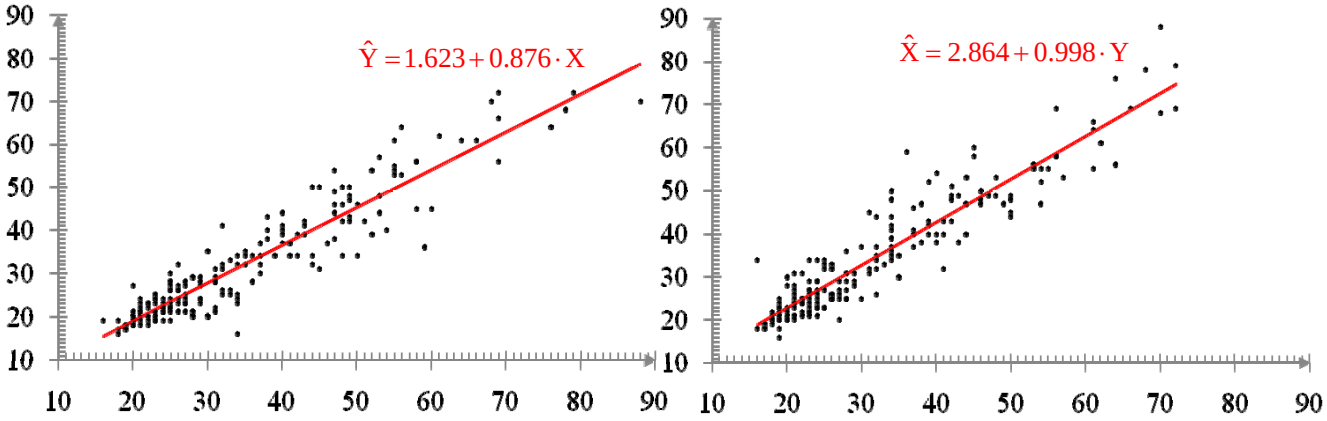
$$Z_Y = \frac{Y - \bar{y}}{s_Y} \; ; \; Z_X = \frac{X - \bar{x}}{s_X} \; ; \; Z_Y = a_{Z_Y.Z_X} + b_{Z_Y.Z_X} \cdot Z_X + E_{Z_Y} \; ; \; Z_X = a_{Z_X.Z_Y} + b_{Z_X.Z_Y} \cdot Z_Y + E_{Z_X}$$

$$\Rightarrow b_{Z_Y.Z_X} = \frac{s(Z_Y, Z_X)}{s_{Z_X}^2} = \frac{r_{YX}}{1} = r_{YX} \; ; \; a_{Z_Y.Z_X} = \bar{z}_Y - b_{Z_Y.Z_X} \cdot \bar{z}_X = 0 - b_{Z_Y.Z_X} \cdot 0 = 0$$

$$\Rightarrow b_{Z_X.Z_Y} = \frac{s(Z_X, Z_Y)}{s_{Z_Y}^2} = \frac{r_{YX}}{1} = r_{YX} \; ; \; a_{Z_X.Z_Y} = \bar{z}_X - b_{Z_X.Z_Y} \cdot \bar{z}_Y = 0 - b_{Z_X.Z_Y} \cdot 0 = 0$$

In der bivariaten Regression sind daher die standardisierten Regressionsgewichte der Regression von Y auf X und von X auf Y identisch und gleich der Produktmomentkorrelation.
Daher sind dann auch die Determinationskoeffizienten beider Regressionen gleich.

Determinationskoeffizient, Korrelation und Regressionsgewichte



$\overline{x} = 35.5405$ $\overline{y} = 32.7568$ $s_X^2 = 210.2700$ $s_Y^2 = 184.7030$ $s_{XY} = 184.2504$

Für das Anwendungsbeispiel gilt so:

$$R^2 = \frac{(s_{YX})^2}{s_X^2 \cdot s_Y^2} = \frac{s_{YX}}{s_X^2} \cdot \frac{s_{YX}}{s_Y^2} = \frac{184.2504^2}{210.270 \cdot 184.703} = b_{Y.X} \cdot b_{X.Y} = 0.876 \cdot 0.998 = 0.874$$

Mit 87.4% wird ein sehr großer Anteil der Variation der abhängigen Variable erklärt. Im Beispiel gibt es also eine sehr enge Beziehung zwischen dem Alter der Partnerin und dem Alter des Partners.

Lerneinheit 20:
Bivariate Beziehungen zwischen ordinalen Variablen

In die Berechnung von Regressionskoeffizienten und Produktmomentkorrelation gehen Abstandsinformationen zwischen den Messwerten ein. Wenn die betrachteten Variablen ordinal sind, ist die Berechnung daher nicht sinnvoll. Auf der anderen Seite interessiert bei ordinalen Variablen nicht nur wie bei nominalskalierten Variablen, ob und wie stark ein Zusammenhang ist, sondern auch, ob die Beziehung positiv oder negativ ist.

Für ordinale Variablen sind daher spezielle Zusammenhangsmaße konstruiert. Dabei werden unterschiedliche Strategien verwendet, um die ordinalen Informationen zu nutzen:

- Bei der Strategie des Paarvergleichs werden alle möglichen Paare der Stichprobe bzw. der Population betrachtet und dabei geprüft, ob die Realisierungen der beiden betrachteten Variablen gleiche, kleinere oder größere Ausprägungen aufweisen.
- Bei der Strategie des Rangreihenvergleichs werden die Rangwerte der beiden Variablen verglichen.
- Bei der Strategie der ungenauen metrischen Messung wird angenommen, dass die ordinalen Variablen ungenaue Messungen von unbeobachteten metrischen Variablen sind, wobei die Beziehung zwischen den unbeobachteten Variablen geschätzt wird.

Ordinale Zusammenhangsanalyse auf der Basis von Paarvergleichen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Bei der Strategie des Paarvergleichs werden für die n Fälle eines Datensatzes alle $n \cdot (n-1)/2$ möglichen ordinalen Paarvergleiche durchgeführt.

Als Beispiel wird im folgenden der Zusammenhang zwischen der trichotomisierten Bewertung der allgemeinen Wirtschaftslage (AWL) und der eigenen Wirtschaftslage (EWL) anhand der westdeutschen Teilstichprobe des Allbus 2006 analysiert.

Bei jedem Paarvergleich können verschiedene Resultate auftreten:

- Bei beiden Variablen hat einer der beiden Fälle einen höheren oder aber niedrigeren Wert als der andere Fall. Man spricht dann von einem ***konkordanten*** (übereinstimmenden) ***Ergebnis***.
- Bei einer Variable hat der eine, bei der anderen Variable der andere Fall einen jeweils höheren oder niedrigeren Wert. Man spricht dann von einem ***diskordanten*** (gegenläufigen) ***Ergebnis***.

Ein konkordantes Ergebnis spricht für eine positive Beziehung, da hier ein höherer Wert bei einer Variable mit einem höheren Wert bei einer anderen Variable ein hergeht. Umgekehrt spricht ein diskordantes Ergebnis für eine negative Beziehung.

Die Logik des Paarvergleichs

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Neben einem konkordanten oder einem diskordanten Ergebnis ist es aber auch möglich, dass die Realisierungen bei einer oder beiden Variablen bei den beiden Fällen eines Paares gleiche Ausprägungen aufweisen. Man spricht dann von einem verbundenen Paar *(engl. Tie)*.

- Wenn die Werte bei X gleich sind, ist dass Paar ***x-verbunden***,
- wenn die Werte bei Y gleich sind, ist das Paar ***y-verbunden***.
- Schließlich können die Realisierungen der beiden Fälle bei beiden Variablen gleich sein. Das Paar ist dann ***x,y-verbunden***.

Da bei allen Paarvergleichen nur die Information „ist gleich“, „ist kleiner“ oder „ist größer“ betrachtet wird, werden zwar ordinale, aber keine metrischen Informationen genutzt.
Für die Erfassung des ordinalen Zusammenhangs wird gezählt, wie viele der Paarvergleiche konkordant, diskordant, x-verbunden, y-verbunden bzw. x,y-verbunden sind. Diese Zahlen gehen in die Berechnung der ordinalen Zusammenhangsmaße ein. In einem ersten Berechnungsschritt müssen daher diese Zahlen für alle Paare des Datensatzes berechnet werden.

Die Logik des Paarvergleichs

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Wenn zwei Variable als Kreuztabelle vorliegen, muss die Anordnung berücksichtigt werden.
*Im Beispiel sind beide Variablen in die **gleiche Richtung** kodiert. Von links nach rechts (AWL) bzw. von oben nach unten (EWL) wird die Bewertung jeweils schlechter.*
Für die Paarvergleiche folgt dann:

C	D	T_X	T_Y	T_{XY}
konkordant	diskordant	x-verbunden	y-verbunden	xy-verbunden
$x_a > x_b$ & $y_a > y_b$	$x_a > x_b$ & $y_a < y_b$	$x_a = x_b$ & $y_a \neq y_b$	$x_a \neq x_b$ & $y_a = y_b$	$x_a = x_b$ & $y_a = y_b$
rechts unten	links unten	unten in Spalte	rechts in Zeile	gleiche Zelle

Beispiel:

Fall:	a	b	a	b	a	b	a	b	a	b
AWL	gut	schlecht	gut	schlecht	gut	schlecht	gut	schlecht	gut	gut
EWL	gut	schlecht	schlecht	gut	gut	gut	schlecht	schlecht	gut	gut

Die Logik des Paarvergleichs

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

C
konkordant
$x_a > x_b$ & $y_a > y_b$
rechts unten

Die Zahl der konkordanten Paare berechnet sich im Beispiel nach:
222·(438+394+101+274) + 571·(394+274)
+ 68·(101+274) +438·274 = 794 894.

D
diskordant
$x_a > x_b$ & $y_a < y_b$
links unten

Die Zahl der diskordanten Paare berechnet sich im Beispiel nach:
571 (68+17) + 197·(68+438+17+101)
+ 438·17 + 394·(17+101) = 225 401.

Wäre die Anordnung entgegengesetzt, so dass bei der Spaltenvariable von links nach rechts aufsteigende und bei der Zeilenvariable von oben nach unten absteigende Werte angeordnet werden bzw. bei der Spaltenvariable von links nach rechts absteigende und bei der Zeilenvariable von oben nach unten aufsteigende Werte, dann wäre die Berechnung vertauscht, so dass sich nach rechts unten diskordante und nach links unten konkordante Paare ergäben.

Die Logik des Paarvergleichs

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

T_X
x-verbunden
$x_a = x_b$ & $y_a \neq y_b$
unten in Spalte

Die Zahl der x-verbundenen Paare berechnet sich im Beispiel nach:
222·(68+17) + 571·(438+101) + 197·(394+274)
+ 68·17 + 438·101 + 394·274 = 611 585.

T_Y
y-verbunden
$x_a \neq x_b$ & $y_a = y_b$
rechts in Zeile

Die Zahl der y-verbundenen Paare berechnet sich nach:
222·(571+197) + 68·(438+394) + 17·(101+274)
+ 571·197 + 438·394 + 101·274 = 546 180.

T_{XY}
xy-verbunden
$x_a = x_b$ & $y_a = y_b$
gleiche Zelle

Die Zahl der x,y-verbundenen Paare berechnet sich dann nach:
222·221/2 + 571·570/2 + 197·196/2 + 68·67/2 + 438·437/2
394·393/2 + 17·16/2 + 101·100/2 + 274·273/2 = 424 561.

Symmetrische Zusammenhangsmaße auf der Basis von Paarvergleichen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Die Summe aus $C+D+T_X+T_Y+T_{XY}$ muss mit der Gesamtzahl aller Paarvergleiche übereinstimmen:

794 894 + 225 401 + 611 585 + 546 180 + 424 561 = 2 602 621 = 2282 · 2281 / 2.

Bei einer (perfekten) positiven Beziehung sollte es nur konkordante, aber keine diskordanten Paare geben, bei einer (perfekten) negativen Beziehung nur diskordante, aber keine konkordanten Paare. Die Differenz aus der Anzahl der konkordanten und diskordanten Paare gibt daher an, ob eine Beziehung eher positiv oder eher negativ ist.

Offen ist dabei, wie verbundene Paare (Ties) berücksichtigt werden sollen:
- Wenn man von strikten „je-desto“-Beziehung ausgeht, sprechen Ties gegen einen positiven bzw. negativen Zusammenhang.
- Auf der anderen Seite sprechen zumindest gleiche Werte eines Paares bei beiden Variablen (T_{XY}) nicht gegen einen positiven Zusammenhang.
- Schließlich lässt sich auch argumentieren, dass Ties gar keine Rolle spielen.

Symmetrische Zusammenhangsmaße auf der Basis von Paarvergleichen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

In Abhängigkeit von der Berücksichtigung von Ties sind unterschiedliche **symmetrische Zusammenhangsmaße** für ordinale Variablen auf der Basis von Paarvergleichen definiert.

- Das Zusammenhangsmaß **γ** (gamma) berücksichtigt gar keine Ties:

$$\gamma = \frac{C - D}{C + D} \text{ ; im Beispiel: } \gamma = \frac{794894 - 225401}{794894 - 225401} = 0.558$$

- beim Maß τ_b (tau-b) werden nur x,y-verbundene Paare nicht berücksichtig,

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_X) \cdot (C + D + T_Y)}}$$

$$\text{im Beispiel: } \tau_b = \frac{794894 - 225401}{\sqrt{(794894 + 225401 + 611585)(794894 + 225401 + 546180)}} = 0.356$$

- und beim Maß τ_a (tau-a) sprechen alle Ties gegen einen Zusammenhang.

$$\tau_a = \frac{C - D}{n \cdot (n-1) / 2} \text{ ; im Beispiel: } \tau_a = \frac{794894 - 225401}{2282 \cdot 2281 / 2} = 0.219$$

Asymmetrische Zusammenhangsmaße auf der Basis von Paarvergleichen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Es lässt sich argumentieren, dass bei einer **asymmetrischen (gerichteten) Beziehung** der Spaltenvariable X auf die Zeilenvariable Y die y-verbundene Paare gegen eine Beziehung sprechen, da dann die Bedingung „Wenn die erklärende Variable verschiedene Werte aufweist, dann muss auch die abhängige Variable verschiedene Werte aufweisen“ verletzt ist:
X-verbundene Paaren sprechen dagegen nicht gegen eine gerichtete (asymmetrische) Beziehung von X auf Y.
Umgekehrt ist es, wenn die Zeilenvariable Y erklärende und die Spaltenvariable X abhängige Variable ist.
Nach dieser Logik ist das gerichtete Zusammenhangsmaß Somers $\mathbf{d_{YX}}$ bzw. $\mathbf{d_{XY}}$ konstruiert:

$$d_{YX} = \frac{C - D}{C + D + T_Y} = \frac{794894 - 225401}{794894 + 225401 + 546180} = 0.364$$

$$d_{XY} = \frac{C - D}{C + D + T_X} = \frac{794894 - 225401}{794894 + 225401 + 611585} = 0.349$$

Ordinale Zusammenhangsmaße auf der Basis von Paarvergleichen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

$\gamma = 0.558$; $\tau_b = 0.356$; $\tau_a = 0.172$; $d_{YX} = 0.364$; $d_{XY} = 0.349$

Der Vergleich der Formeln zeigt, dass Tau-b gleich dem geometrischen Mittel von Somers d_{YX} und d_{XY} ist.

Da sich die Berechnungsformeln der Zusammenhangsmaße nur im Nenner unterscheiden, sind die Vorzeichen bei allen Zusammenhangsmaßen auf der Basis von Paarvergleichen stets gleich.

Für das Beispiel gilt somit: Je besser die eigene wirtschaftliche Lage eingeschätzt wird, desto besser wird auch die allgemeine Lage eingeschätzt und umgekehrt.

Es stellt sich die Frage, welches Maß verwendet werden sollte:

- Bei gerichteten Beziehungen kommt nur Somers d_{YX} bzw. d_{XY} in Frage.
- Bei symmetrischen Beziehungen wird am häufigsten τ_b verwendet,
- während τ_a i.a. nur dann herangezogen wird, wenn jede Ausprägung auch (theoretisch) einen unterschiedlichen Wert aufweisen kann.
- Das Maß γ wird eher selten verwendet, da dieses Maß vor allem bei wenigen Ausprägungen der Variablen die Tendenz hat, leicht recht hohe Werte anzunehmen.

Beziehungen zwischen den Zusammenhangsmaßen in der Vierfeldertabelle

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut (X=1)	nicht gut (X=0)	Summe
- gut (Y=1)	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut (Y=0)	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Bei dichotomen Variablen gibt es Beziehungen zwischen den ordinalen Zusammenhangsmaßen d_{YX}, d_{XY}, τ_b und γ mit den Regressionsgewichten der linearen Regression und der Produktmomentkorrelation, wenn die kreuztabellierten Variablen 1/0-kodiert sind, sowie mit den Prozentsatzdifferenzen d_{YX}%, d_{XY}% und den symmetrischen Zusammenhangsmaßen φ und Q.

Die Zahl konkordanten, diskordanten und verbundenen Paare beträgt in der Vierfeldertabelle:

$C = n_{11} \cdot n_{22}$ im Beispiel: $C = 222 \cdot 1207 = 267954$, $D = n_{12} \cdot n_{21}$ im Beispiel: $D = 768 \cdot 85 = 65280$

$T_X = n_{11} \cdot n_{21} + n_{12} \cdot n_{22}$ im Beispiel: $T_X = 222 \cdot 85 + 768 \cdot 1207 = 945846$

$T_Y = n_{11} \cdot n_{12} + n_{21} \cdot n_{22}$ im Beispiel: $T_Y = 222 \cdot 768 + 85 \cdot 1207 = 273091$

Die ordinalen Zusammenhangsmaße sind dann:

$$d_{YX} = \frac{267954 - 65280}{267954 + 65280 + 273091} = 0.334 \; ; \; d_{XY} = \frac{267954 - 65280}{267954 + 65280 + 945846} = 0.158$$

$$\gamma = \frac{267954 - 65280}{267954 + 65280} = 0.608 \; ; \; \tau_b = \sqrt{0.334 \cdot 0.158} = 0.230$$

Beziehungen zwischen Zusammenhangsmaßen in der Vierfeldertabelle

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut (X=1)	nicht gut (X=0)	Summe
- gut (Y=1)	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut (Y=0)	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Bei 1/0-Kodierung der dichotomen Variablen berechnen sich die Mittelwerte, Varianzen und Kovarianzen der Spaltenvariable X und der Zeilenvariable Y nach:

$\bar{x} = \frac{n_{+1}}{n} = p_{+1}$ im Beispiel: $\bar{x} = \frac{307}{2282} = 0.135$, $\bar{y} = \frac{n_{1+}}{n} = p_{1+}$ im Beispiel: $\bar{y} = \frac{990}{2282} = 0.434$,

$s_X^2 = p_{+1} \cdot p_{+2}$ im Beispiel: $s_X^2 = 0.135 \cdot 0.865 = 0.116$,

$s_Y^2 = p_{1+} \cdot p_{2+}$ im Beispiel: $s_Y^2 = 0.434 \cdot 0.566 = 0.246$,

$s_{YX} = p_{11} - p_{+1} \cdot p_{1+}$ im Beispiel: $s_{YX} = 0.097 - 0.135 \cdot 0.434 = 0.0389$

Daraus folgt für Regressionsgewichte $b_{Y.X}$ der linearen Regression von Y auf X bzw. $b_{X.Y}$ von X auf Y sowie der Produktmomentkorrelation r_{xy}:

$$b_{Y.X} = \frac{0.0389}{0.116} = 0.335 \; , \; b_{X.Y} = \frac{0.0389}{0.246} = 0.158 \; ; \; r_{YX} = \frac{0.0389}{\sqrt{.116 \cdot .246}} = 0.230$$

Beziehungen zwischen Zusammenhangsmaßen in der Vierfeldertabelle

Eigene wirtschaftliche Lage des Befragten	Allgemeine Wirtschaftslage gut (X=1)	nicht gut (X=0)	Summe
- gut (Y=1)	9.7% (222)	33.7% (768)	43.4% (990)
- nicht gut (Y=0)	3.7% (85)	52.9% (1207)	56.6% (1292)
Summe	13.5% (307)	86.5% (1975)	100.0% (2282)

(Quelle: Allbus 2006, nur Westen)

Die Prozentsatzdifferenzen, Phi und Yules Q betragen im Beispiel:

$$d_{YX}\% = 100 \cdot \left(\frac{a}{a+c} - \frac{b}{b+d} \right) = 100 \cdot \left(\frac{222}{307} - \frac{768}{1975} \right) = 33.4$$

$$d_{XY}\% = 100 \cdot \left(\frac{a}{a+b} - \frac{c}{c+d} \right) = 100 \cdot \left(\frac{222}{990} - \frac{85}{1282} \right) = 15.8$$

$$\phi = \sqrt{\frac{d_{YX}\%}{100} \cdot \frac{d_{XY}\%}{100}} = \sqrt{0.334 \cdot 0.158} = 0.230 \qquad Q = \frac{a \cdot d - b \cdot c}{a \cdot d - b \cdot c} = \frac{267954 - 65280}{267954 + 65280} = 0.608$$

Der Vergleich der Berechnungsformeln bzw. der berechneten Werte im Beispiel zeigt, dass für dichotome Variablen gilt:

(1) $d_{YX} = d_{YX}\%/100 = b_{Y.X}$ und $d_{XY} = d_{XY}\%/100 = b_{X.Y}$;

(2) $\tau_b = \varphi = r_{YX}$;

(3) $\gamma = Q$.

Ordinale Zusammenhangsmaße auf der Basis von Paarvergleichen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Für inferenzstatistische Anwendungen wird die Verteilung der Koeffizienten benötigt. Bei sehr großen Fallzahlen sind alle Maße asymptotisch normalverteilt, doch ist die Annäherung sehr langsam. Zudem ist auch die Berechnung (asymptotisch gültiger) geschätzter Standardfehler sehr aufwendig, so dass diese Berechnung in der Praxis nur über Statistikprogramme erfolgt.

Koeffizient	Wert	S.E.	Z
γ	0.558	0.024	20.582
τ_b	0.356	0.017	20.582
d_{YX}	0.364	0.017	20.582
d_{XY}	0.349	0.017	20.582

(Berechnung mit SPSS)

Die Tabelle zeigt die mit SPSS berechneten Werte.
SPSS berechnet nicht τ_a und daher auch keine Standardfehler und Teststatistik für diesen Koeffizienten.

Die Teststatistiken Z (im originalen SPSS-Ausdruck als „näherungswiese T“ bezeichnet) beziehen sich auf die Nullhypothese, dass das jeweilige Zusammenhangsmaß Null ist.

Ordinale Zusammenhangsmaße auf der Basis von Paarvergleichen

Koeffizient	Wert	S.E.	Z
γ	0.558	0.024	20.582
τ_b	0.356	0.017	20.582
d_{YX}	0.364	0.017	20.582
d_{XY}	0.349	0.017	20.582

(Berechnung mit SPSS)

Da die alle Maße genau dann Null sind, wenn die Zahl der diskordanten und konkordanten Paare gleich ist, sind alle Z-Wert identisch. Der Wert ist der Quotient aus der Differenz C–D geteilt durch den Standardfehler von C–D unter der Annahme, dass die Nullhypothese H_0: C=D in der Population zutrifft.

Im Beispiel kann der Wert 20.582 als Teststatistik für die Nullhypothese verwendet werden, dass es keinen monotonen Zusammenhang gibt, H_0: C = D, dass die Zahl der konkordanten Paare kleiner/gleich der Zahl der diskordanten Paare ist, H_0: C ≤ D (kein positiver Zusammenhang), oder dass die Zahl der konkordanten Paare größer/gleich der diskordanten Paare ist: H_0: C ≥ D (kein negativer Zusammenhang).
Die Nullhypothese wird mit einer Irrtumswahrsacheinlichkleit α abgelehnt, wenn $Z \leq z_{\alpha/2}$ bzw. $Z \geq z_{1-\alpha/2}$ im zweiseitigen Test von H_0: C=D bzw. $Z \geq z_{1-\alpha}$ bei H_0: C ≤ D oder $Z \leq z_{\alpha}$ bei H_0: C ≥ D.

Ordinale Zusammenhangsmaße auf der Basis von Paarvergleichen

Koeffizient	Wert	S.E.	Z
γ	0.558	0.024	20.582
τ_b	0.356	0.017	20.582
d_{YX}	0.364	0.017	20.582
d_{XY}	0.349	0.017	20.582

(Berechnung mit SPSS)

Es ist möglich, dass ein ordinales Zusammenhangsmaß nicht signifikant ist, gleichzeitig aber der Chiquadrat-Test bzw. LR-Test auf statistische Unabhängigkeit auf einen signifikanten Zusammenhang hinweist. Dies spricht dann für einen nicht-monotonen Zusammenhang zwischen den beiden ordinalen Variablen.
Da ordinale bzw. metrische Zusammenhänge mehr Informationen nutzen, können Zusammenhangstests für ordinale oder metrische Variablen trennschärfer sein als entsprechende Tests bei nominalskalierten Variablen. Es kann daher auch vorkommen, dass ein Test auf statistische Unabhängigkeit ein nichtsignifikantes Resultat, ein Test auf monotonen oder linearen Zusammenhang dagegen ein signifikantes Resultat ergibt.

Die geschätzten Standardfehler in der dritten (mit „S.E.“ überschriebenen) Spalte sind asymptotisch gültiger Standardfehler bei beliebigen Werten der Koeffizienten. Sie können z.B. für die Berechnung von Konfidenzintervallen genutzt werden.

So berechnet sich das asymptotisch gültige 95%-Konfidenzintervall für τ_b im Beispiel nach: $\tau_b \pm z_{0.975} \cdot S.E. = 0.356 \pm 1.96 \cdot 0.017 = 0.32$ bis 0.39.

Ordinale Zusammenhangsmaße auf der Basis von Rangreihen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Die Logik der Paarvergleiche nutzt nicht alle ordinalen Informationen aus. Wenn X bei einem Fall A den Rangplatz 1, bei einem Fall B den Rangplatz 2 und bei einem Fall C den Rangplatz 3 aufweist, dann nutzen Paarvergleiche für die Eigenschaft X, dass A < B, A < C und B < C. Nicht genutzt wird dagegen, dass der Abstand zwischen A und C kleiner als der Abstand zwischen A und B sowie zwischen B und C sein muss.

Die zweite Strategie für die Zusammenhangsanalyse bei ordinalen Zusammenhängen nutzt diese zusätzliche Informationen, indem sie die Rangplatzwerte der Realisierungen bei den beiden ordinalen Variablen vergleicht. Dabei stellt sich die Frage, wie mit gleichen Rangwerten umzugehen ist.
In der Regel werden bei gleichen Ausprägungen mittlere Ränge verwendet.

Für die Berechnung müssen zunächst für die Randverteilungen die kumulierten Ränge berechnet werden. Für die Spaltenvariable (allgemeine Lage, AWL) ergibt sich 307, 1724 (=307+1110) und 2282 (= 1724+865), für die Zeilenvariable (eigene Lage, EWL) 990, 1890 (=990+900) und 2282 (=1890+392).

Ordinale Zusammenhangsmaße auf der Basis von Rangreihen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe	Rangplatz
(sehr) gut	222	571	197	990	495.5
teils/teils	68	438	394	900	1440.5
(sehr) schlecht	17	101	274	392	2086.5
Summe	307	1110	865	2282	
Rangplatz	154	862.5	1850		

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Der mittleren Rangplatz einer Kategorien ist die Hälfte der Summe der jeweils ersten und der letzten Rangzahl einer Realisierung in dieser Kategorie:

Variable	*Bewertung*	*(sehr) gut*	*teils/teils*	*(sehr) schlecht*
AWL	*Rangplatz*	*(1+307)/2 = 154*	*(308+1417)/2 = 862.5*	*(1418+2282)/2 = 1850*
EWL	*Rangplatz*	*(1+990)/2 = 495.5*	*(991+1890)/2 = 1440.5*	*(1891+2282)/2 = 2086.5*

Jedem Fall werden so für beide Variablen Rangplätze zugeordnet. Fälle in der ersten Tabellenzelle AWL=EWL=sehr gut erhalten die Rangplatzwerte 154 und 495.5, Fälle in der letzten Zelle die Rangplatzwerte 1850 und 2086.5.

Ordinale Zusammenhangsmaße auf der Basis von Rangreihen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe	Rangplatz
(sehr) gut	222	571	197	990	495.5
teils/teils	68	438	394	900	1440.5
(sehr) schlecht	17	101	274	392	2086.5
Summe	307	1110	865	2282	
Rangplatz	154	862.5	1850		

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Spearmans Rangkorrelation R_s ist die Produktmomentkorrelation über die Ränge.
Für die Berechnung werden zunächst die Summen, Quadratsummen und die Produktsumme der Ränge über alle Realisierungen benötigt:

$$\sum \text{Rang}(\text{AWL}) = \sum \text{Rang}(\text{EWL}) = 2604903$$

$$\sum \left(\text{Rang}(\text{AWL})\right)^2 = 3793479249.5 \; ; \; \sum \left(\text{Rang}(\text{EWL})\right)^2 = 3817166314.5$$

$$\sum \text{Rang}(\text{AWL}) \cdot \text{Rang}(\text{EWL}) = 3295672477$$

Die Rangkorrelation für die Allbus-Daten berechnet sich dann nach:

$$r_s(X,Y) = \frac{3295672477/2282 - 2604903^2/2282^2}{\sqrt{\left(3793479249.5/2282 - 2604903^2/2282^2\right)\left(3817166314.5/2282 - 2604903^2/2282^2\right)}}$$
$$= 0.387$$

Ordinale Zusammenhangsmaße auf der Basis von Rangreihen

Da vor der Berechnung die Ausprägungen in Rangwerte transformiert werden, ergeben sich stets die gleichen Korrelationen, wenn vor der Berechnung der Ränge eine beliebige für Ordinalskalen zulässige monotone Transformation der Werte durchgeführt wird.

Dies gilt aber auch, wenn bei der Berechnung der Ränge bei gleichen Ausprägungen stets der kleinste Rangplatz verwendet wird.

Die drei Ausprägungen (sehr) gut, teils/teils und (sehr) schlecht würden dann die Kodierungen 1,2,3 oder bei umgekehrter Polung 3,2,1 erhalten. Die Produktmomentkorrelation beträgt bei dieser Kodierung 0.380.

Im Beispiel führt die Berechnung über mittlere Rangwerte statt kleinster Rangplätze zu einem geringfügig geringeren Wert. Ursache abweichender Werte ist letztlich die Distanz der resultierenden Kodierungen.

Bei der Kodierung 1,2,3 ist der Abstand zwischen den drei Ausprägungen gleich, bei den Kodierungen 154, 862.5 und 1850 für AWL sowie 495.5, 1440.5 und 2086.5 für EWL beträgt dagegen der Abstand zwischen teils/teils und (sehr) schlecht das 0.383-fache (AWL) bzw. 0.498-fache (EWL) des Abstandes von (sehr) gut und teils/teils. Die ungleichen Abstände ergeben bei den Allbus-Daten offenbar eine etwas stärke Angleichung an eine lineare Beziehung.

Das Beispiel zeigt, dass ungleiche (Rang-) Abstände zwischen den Kategorien nicht notwendigerweise zu sehr unterschiedlichen Zusammenhangsstärken führen. Dies weist darauf hin, dass die Produktmomentkorrelation offenbar relativ unmempfindlich (robust) gegenüber beliebigen monotonen Transformationen ist.

Ordinale Zusammenhangsmaße auf der Basis von Rangreihen

Anmerkung:
Dies und Ergebnisse von Simulationsstudien zur Robustheit von Zusammenhangsanalysen haben dazu geführt, dass ordinale Variablen mit mindestens drei oder vier Ausprägungen in vielen statistischen Datenanalysen wie metrische Variablen behandelt werden, obwohl dies streng genommen nicht zulässig ist.

Ein Vorteil von Spearmans Rangkorrelation besteht darin, dass die gleichen Signfikanztests wie bei der Produktmomentkorrelation angewendet werden.

Soll im Beispiel getestet werden, ob die Rangkorrelation bei einer Irrtumswahrscheinlichkeit von 1% signifikant von Null verschieden ist, wird die Teststatistik T der Produktmomentkorrelation auf den Wert von Spearmans Rangkorrelation angewendet:

$$T = 0.387 \cdot \sqrt{\frac{2282-2}{1-0.387^2}} = 20.0$$

Die kritischen Werte von T liegen bei einem zweiseitigen Test und einer Irrtumswahrscheinlichkeit α=0.01 zwischen ±2.617 (df=120) und ±2.576 (df=∞).
Bei df=2280 Freiheitsgraden ist die Nullhypothese, dass es keinen oder gar einen negativen Zusammenhang gibt, bei einer Irrtumswahrscheinlichkeit von 1% zu verwerfen.

Ordinale Messungen als ungenaue Erfassung von metrischen Variablen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Eine dritte Analysestrategie für ordinale Variablen geht davon aus, dass die ordinalen Messungen als ungenaue Messungen eigentlich metrischer Variablen aufgefasst werden können.
Ein Zusammenhangsmaß, dass dieser Logik folgt, ist die Berechnung ***polychorischer Korrelationen***.

Die Berechnung basiert auf einem sogenannten ***Schwellenwertmodell***. Es wird angenommen, dass die eigentlich interessierenden Größen stetige metrische Variablen sind, die allerdings aufgrund des Messverfahren nur sehr ungenau gemessen werden können. Alle Realisierungen unterhalb eines ersten Schwellenwertes können bei der Messung nicht unterschieden werden und erhalten daher den kleinsten (ordinalen) Wert. Ganz analog können auch alle Realisierungen zwischen dem ersten und dem zweiten Schwellenwert nicht unterschieden werden. Ihnen wird daher bei der Messung die zweitkleinste Ausprägung zugewiesen. Bei insgesamt K Schwellenwerten ergeben sich so K+1 ordinale Ausprägungen.

Polychorische Korrelationen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Die Ordinalität ist in diesem Modell letztlich Folge davon, dass die Schwellenwerte unbekannt sind und daher die Abstände zwischen den ordinalen Ausprägungen unbekannt sind und zudem Unterschiede zwischen den Realisierungen innerhalb einer ordinalen Ausprägungen nicht messbar sind.

Die polychorischen Korrelationen sind dann Schätzungen der unbekannten Produktmoment-korrelation zwischen den eigentlich interessierenden, aber nicht direkt gemessenen stetigen und metrischen Variablen auf der Basis der Kreuztabellierung der (ungenauen) ordinalen Messungen.

Um zu einer Schätzung zu kommen wird üblicherweise angenommen, dass die nicht direkt gemessenen stetigen metrischen Variablen normalverteilt sind. Unter dieser Annahme können bei einer einfachen Zufallsauswahl sowohl die Schwellenwerte als auch die Produktmomentkorrelation geschätzt werden. Die Berechnung ist allerdings so aufwendig, dass sie ausschließlich mit Computerprogrammen realisiert wird.

Polychorische Korrelationen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Polychorische Korrelation: 0.483 (SE: 0.0228)

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Für die Beispieldaten aus dem Allbus 2006 berechnet das Programm LISREL 8 eine polychorische Korrelation von 0.483 mit einem asymptotischen Standardfehler von 0.0228.

Der Standardfehler kann zur Berechnung von Konfidenzintervallen oder asymptotisch gültigen Z-Tests verwendet werden.

So berechnet sich das asymptotisch gültige 95%-Konfidenzintervall für die polychorische Korrelation im Beispiel nach: $0.483 \pm 1.96 \cdot 0.0228 = 0.438$ bis 0.528.
Der Quotient $Z = 0.483/0.0228 = 21.2$ kann als Teststatistik für einen asymptotisch gültigen Z-Test zur Prüfung der Nullhypothese, dass die Korrelation in der Population Null ist, herangezogen werden.

Bei der Berechnung polychorischer Korrelationen werden auch Chiquadrattests zur Prüfung der Nullhypothese berechnet, dass die Korrelation in der Population Null ist.

Für die Beispieldaten, ergibt sich ein Chquadratwert von 11.6, der bei df=3 Freiheitsgraden ein empirisches Signifikanzniveau von 0.009 aufweist.

Polychorische Korrelationen

Eigene wirtschaftl. Lage	Allgemeine wirtschaftliche Lage (sehr) gut	teils/teils	(sehr) schlecht	Summe
(sehr) gut	222	571	197	990
teils/teils	68	438	394	900
(sehr) schlecht	17	101	274	392
Summe	307	1110	865	2282

Polychorische Korrelation: 0.483 (SE: 0.0228)

Quelle: Westdeutsche Teilstichprobe, Allbus 2006

Polychorischen Korrelationen sind in der Regel größer als Produktmomentkorrelationen auf der Basis der gleichen Daten und Ignorierung der Ordinalität. Im Beispiel beträgt so die Produktmomentkorrelation nur $r_{YX} = 0.380$.

Bei der Berechnung ist allerdings zu berücksichtigen, dass das zugrundeliegende Schwellenwertmodell theoretisch sinnvoll sein muss.

> *Im Beispiel aus dem Allbus kann etwa angenommen werden, dass die Beurteilung der wirtschaftlichen Lage eine – in den Köpfen der Befragten – kontinuierliche metrische Größe ist und die Ordinalität Folge der Vorgabe von „groben“ Antwortkategorien.*