Statistik I im Sommersemester 2007

Themen am 4.6.2007:

Wahrscheinlichkeitstheorie und Inferenzstatistik

- Zentraler Grenzwertsatz und stetige Verteilungen
- Wahrscheinlichkeitsverteilungen von Mittelwerten
- Schätzer, Schätzungen und Eigenschaften von Schätzern
- Intervallschätzung mit Konfidenzintervallen
- Punkt- und Intervallschätzung von Anteilen, Mittelwerten und Varianzen

Lernziele:

1. Stetige Wahrscheinlichkeitsverteilungen
2. Standardnormalverteilung
3. Bedeutung des zentralen Grenzwertsatzes
4. Erwartungstreue, Konsistenz und Effizienz als erwünschte Schätzereigenschaften
5. Die Bedeutung von Standardfehlern bei der Schätzung von Populationsparametern
6. Interpretation von Konfidenzintervallen und Irrtumswahrscheinlichkeiten
7. Anwendung der T-Verteilung

Wiederholung

Frequentistischen Definition der Wahrscheinlichkeit: $\lim_{n\to\infty}\left(\frac{n_A}{n}\right)=\Pr(A)$

Gesetz der großen Zahl: $\lim_{n\to\infty}\left(\Pr\left(\left|\frac{n_A}{n}-\Pr(A)\right|<\varepsilon\right)\right)=1$

Abweichungen von einfachen Zufallsauswahlen in empirischen Stichproben:
- Geschichtete Zufallsauswahlen und mehrstufige Zufallsauswahlen
- Konsequenzen von systemaischen Ausfällen: Verzerrung

Wahrscheinlichkeitsverteilung von Häufigkeiten eines Merkmals A bei einfacher Zufallsauswahl ***mit Zurücklegen*** des Umfangs n, das in der Population die relative Häufigkeit $\pi_1 = N_1/N$: aufweist: die **Binomialverteilung**

$$\Pr(X=n_1)=b(X;n,\pi_1)=\binom{n}{n_1}\cdot\pi_1^{n_1}\cdot(1-\pi_1)^{n-n_1}=\frac{n!}{(n-n_1)!\cdot n!}\cdot\pi_1^{n_1}\cdot(1-\pi_1)^{n-n_1}$$

$$\mu_X=n\cdot\pi_1 \text{ und } \sigma_X^2=n\cdot\pi_1\cdot(1-\pi_1)$$

Wiederholung

Wahrscheinlichkeitsverteilung von Häufigkeiten eines Merkmals A bei einfacher Zufallsauswahl ***ohne Zurücklegen*** des Umfangs n, das in der Population die relative Häufigkeit N_1/N: aufweist: die **hypergemoetrische Verteilung**

$$\Pr\left(X = n_1\right) = h\left(X = n_1; n, N, N_1\right) = \frac{\binom{N_1}{n_1} \cdot \binom{N - N_1}{N - n_1}}{\binom{N}{n}} = \frac{\frac{N_1!}{n_1! \cdot \left(N_1 - n_1\right)!} \cdot \frac{\left(N - N_1\right)!}{\left(n - n_1\right)! \cdot \left(N - N_1 - n + n_1\right)!}}{\frac{N!}{n! \cdot \left(N - n\right)!}}$$

$$\mu\left(n_1\right) = n \cdot \frac{N_1}{N} \text{ und } \sigma^2\left(n_1\right) = n \cdot \frac{N_1}{N} \cdot \left(1 - \frac{N_1}{N}\right)$$

Annäherung der hypergeometrischen Verteilung mit den Parametern n, N_1 und N an die Binomialverteilung mit den Parametern n und $\pi_1 = N_1/N$.
Die Annäherung ist hinreichend genau, wenn $N/n > 20$.

Wahrscheinlichkeiten von Anteilen bei einfachen Zufallsauswahlen mit Zurücklegen

Die hypergeometrische Verteilung bzw. die Binomialverteilung kann genutzt werden, um die Wahrscheinlichkeitsverteilungen von (absoluten) Häufigkeiten eines Merkmals A in einer Stichprobe bei einfachen Zufallsauswahlen mit bzw. ohne Zurücklegen zu berechnen. Über diese Wahrscheinlichkeitsverteilungen können aber auch die ***relative Häufgkeiten*** in der Stichprobe berechnet werden.

Die Wahrscheinlichkeitsverteilung einer relative Häufigkeit $p_1 = n_1/n$ lässt sich aus der Verteilung der absoluten Häufigkeit berechnen, da es sich um eine Lineartransformation handelt:

$$p_1 = 0 + 1/n \cdot n_1$$

Ausgangspunkt ist eine Population mit insgesamt N Elementen, von denen N_1 eine interessierende Eigenschaft aufweisen. Wenn zufällig n=1 Element aus dieser Population ausgewählt wird, beträgt die Wahrscheinlichkeit, dass das Element die interessierende Eigenschaft aufweist $\pi_1 = N_1/N$. Die Wahrscheinlichkeitsverteilung ist dann bernoulliverteilt. Bei einer größeren Fallzahl mit n > 1 Elementen ist dann die Häufigkeit bei einer einfachen Zufallsauswahl mit Zurücklegen binomialverteilt mit den Parametern n und $\pi_1 = N_1/N$.

Bei einfachen Zufallsauswahlen mit Zurücklegen berechnet sich daher die Wahrscheinlichkeit einer relative Häufigkeit $p_1 = n_1/n$ über die Binomialverteilung nach:

$$\Pr(p_1) = b\left(X = n \cdot p_1; n, \frac{N_1}{N}\right) = \binom{n}{p_1 \cdot n} \cdot \left(\frac{N_1}{N}\right)^{p_1 \cdot n} \cdot \left(1 - \frac{N_1}{N}\right)^{n \cdot (1 - p_1)}$$

Wahrscheinlichkeiten von Anteilen bei einfachen Zufallsauswahlen mit Zurücklegen

Die Binomialverteilung kann annäherungsweise auch bei einer einfachen Zufallsauswahl ohne Zurücklegen verwendet werden, wenn der Populationsumfang N relativ zur Stichprobengröße n sehr groß ist: N/n >20.
Ist diese Bedingung nicht erfüllt, berechnet sich die Wahrscheinlichkeit einer relative Häufigkeit $p_1 = n_1/n$ bei einer einfachen Zufallsauswahle ohne Zurücklegen über die hypergeometrische Verteilung nach:

$$\Pr(p_1) = b(X = n \cdot p_1; n, N, N_1) = \frac{\binom{N_1}{n \cdot p_1} \cdot \binom{N - N_1}{n - n \cdot p_1}}{\binom{n}{N}}$$

Die Gleichungen für die Wahrscheinlichkeitsfunktion eines Anteils gelten nur unter der Bedingung $p_1 = n_1/n$. Bei allen Werten $p_1 \neq n_1/n$ sind die Auftretenswahrscheinlichkeiten stets null.

Wiederholung

Linearkombination von K statistisch unabhängigen Zufallsavariablen:

Wenn (1) $Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_K \cdot X_K$, dann folgt:

$$\mu_Y = \mu(Y) = b_0 + b_1 \cdot \mu(X_1) + b_2 \cdot \mu(X_2) + \ldots + b_K \cdot \mu(X_k) = b_0 + \sum_{k=1}^{K} b_k \cdot \mu(X_k)$$

$$\sigma_Y^2 = \sigma^2(Y) = b_1^2 \cdot \sigma^2(X_1) + b_2^2 \cdot \sigma^2(X_2) + \ldots + b_K^2 \cdot \sigma^2(X_K) = \sum_{k}^{K} b_k^2 \cdot \sigma^2(X_k)$$

Relative Häufigkeit $p_1 = n_1/n$ sind Linearkombinationen absoluter Häufigkeiten:

$$p_1 = 0 + 1/n \cdot n_1$$

Daher Mittelwert und Varianz der Wahrscheinlichkeitsverteilung relativer Häufigkeitren:

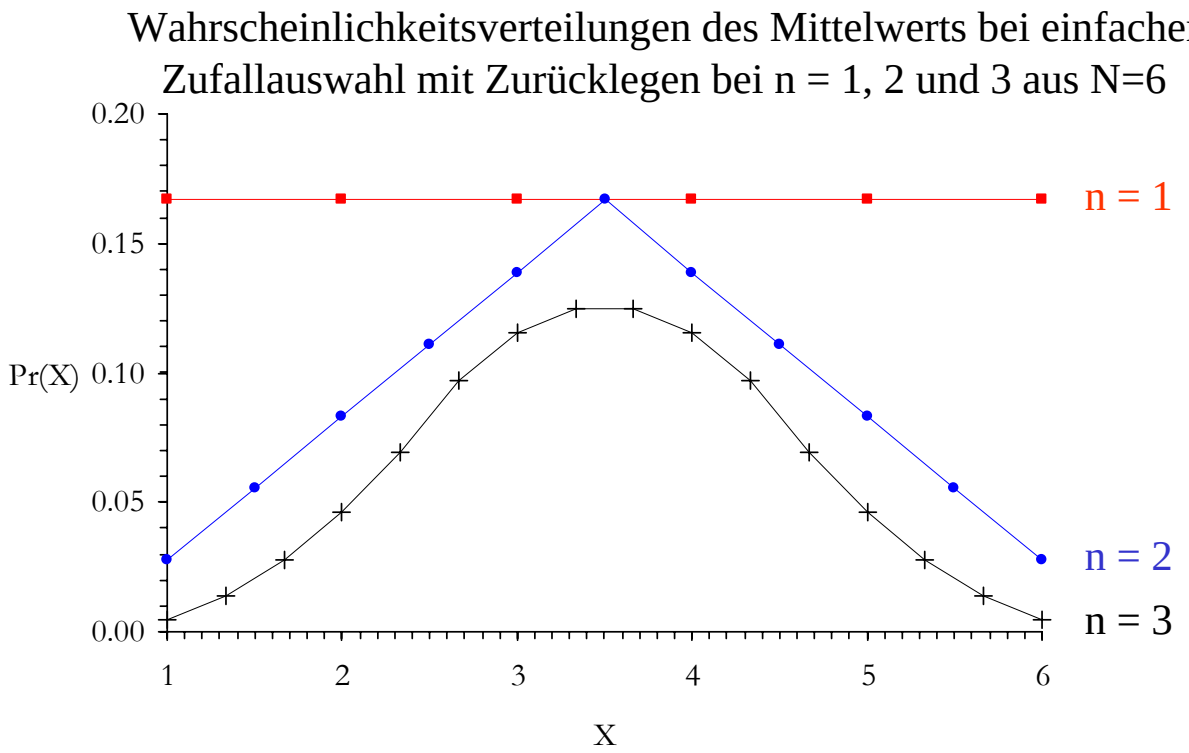
• bei einer einfachen Zufallsauswahl mit Zurücklegen:

$$\mu(p_1) = \frac{1}{n} \cdot \left(n \cdot \frac{N_1}{N} \right) = \frac{N_1}{N} \text{ und } \sigma^2(p_1) = \frac{1}{n^2} \cdot \left(n \cdot \frac{N_1}{N} \cdot \left(1 - \frac{N_1}{N} \right) \right) = \frac{1}{n} \cdot \frac{N_1}{N} \cdot \left(1 - \frac{N_1}{N} \right)$$

• bei einer einfachen Zufallsauswahl ohne Zurücklegen:

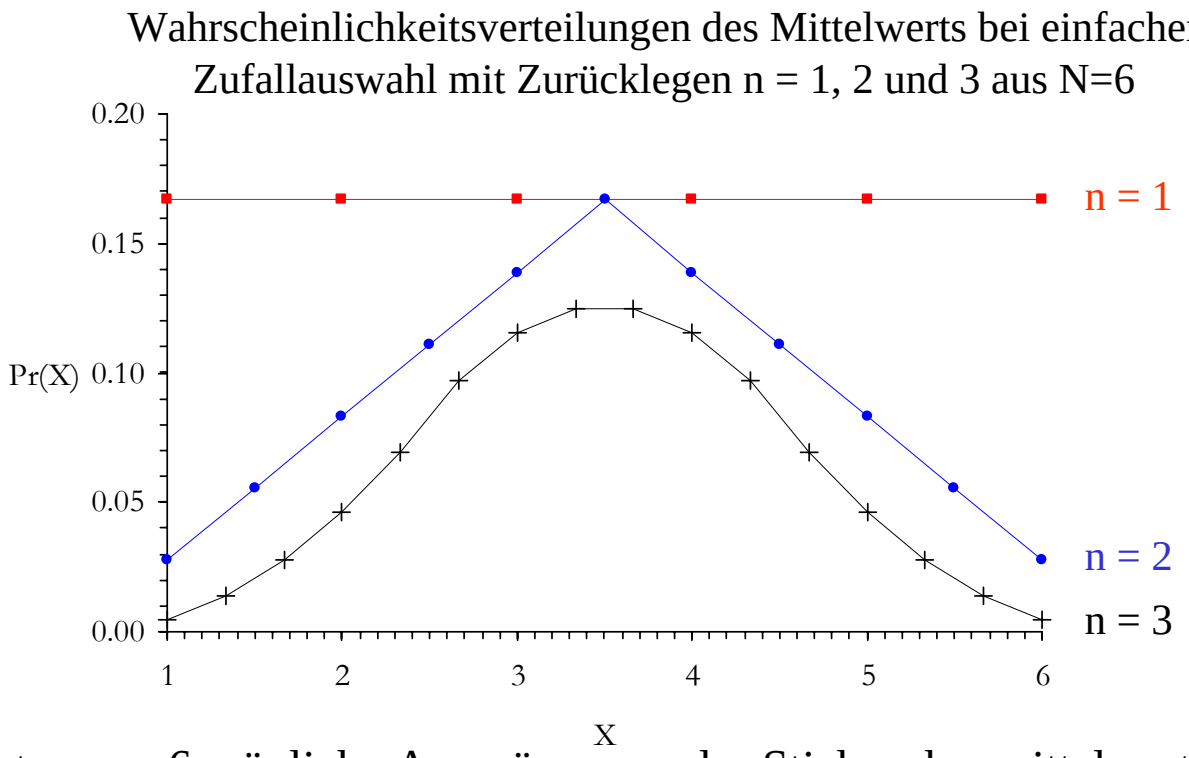
$$\mu(p_1) = \frac{1}{n} \cdot \left(n \cdot \frac{N_1}{N} \right) = \frac{N_1}{N} \text{ und } \sigma^2(p_1) = \frac{1}{n^2} \cdot \left(n \cdot \frac{N_1}{N} \cdot \left(1 - \frac{N_1}{N} \right) \cdot \frac{N-n}{N-1} \right) = \frac{1}{n} \cdot \frac{N_1}{N} \cdot \left(1 - \frac{N_1}{N} \right) \cdot \frac{N-n}{N-1}$$

Wahrscheinlichkeitsverteilungen von Stichprobenmittelwerten



Wenn eine Stichprobe verwendet wird, um einen Populationsmittelwert zu schätzen, wird die Kennwerteverteilung des Stichprobenmittelwerts über alle Stichproben benötigt.
Die obige Abbildung zeigt die Wahrscheinlichkeitsvereilung des Stichprobenmittelwerts für das Beispiel einer einfachen Zufallsauswahl mit Zurücklegen aus einer Grundgesamtheit von N=6 Haushalten, die 1000, 2000, 3000, 4000, 5000 und 6000 € pro Monat verdienen.
Für jede Wahrscheinlichkeitsverteilung sind die Realisierungswahrscheinlichkeiten durch eine durchgezogene Linie verbunden.

Der Zentrale Grenzwertsatz



Bei n=1 gibt es nur 6 mögliche Ausprägungen des Stichprobenmittelwerts, bei n=2 sind es 11 und bei n=3 sind es bereits 16.
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.
Dies ist nicht zufällig, sondern Folge des ***zentralen Grenzwertsatzes***.

Der Zentrale Grenzwertsatz

Der ***zentrale Grenzwertsatz*** ist die neben dem Gesetz der großen Zahl vielleicht wichtigste Aussage der Wahrscheinlichkeitstheorie:

> Die Summe unabhängiger und identisch verteilter Zufallsvariablen nähert sich bei steigender Zahl von Summanden asymptotisch einer Normalverteilung an:

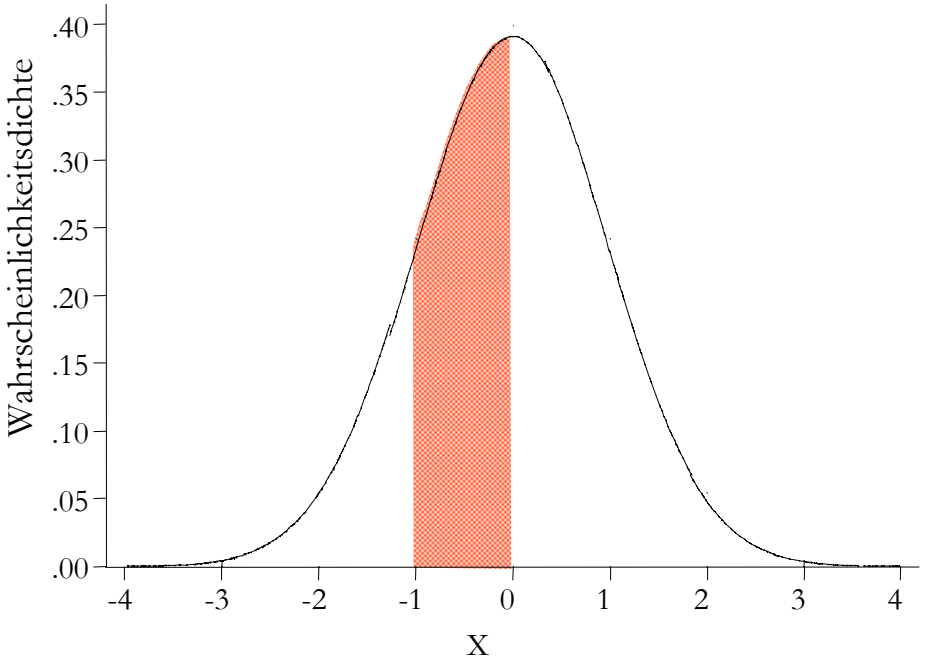
$$\lim_{n\to\infty}\left(\Pr\left(\sum_{i=1}^{n} X_i\right)\right) = N\left(n\cdot\mu_X; n\cdot\sigma_X^2\right)$$

Da bei steigender Zahl von Summanden Erwartungswert und Varianz der Summe ansteigen, wird der zentrale Grenzwertsatz in der Regel für standardisierte (Z-transformierte) Zufallsvariablen formuliert:

$$\lim_{n\to\infty}\left(\Pr\left(\frac{\sum_{i=1}^{n} X_i - n\cdot\mu_X}{\sqrt{n\cdot\sigma_X^2}}\right)\right) = N(0;1)$$

Das Symbol $N(\mu; \sigma^2)$ bzw. $N(\mu, \sigma)$ steht für eine normalverteilte Zufallsvariable mit Erwartungswert μ und Varianz σ^2 bzw. einer Standardabweichung σ.
Entsprechend steht $N(0;1)$ für eine standardisierte Normalverteilung mit Erwartungswert 0 und Varianz 1, die als ***Standardnormalverteilung*** bezeichnet wird.

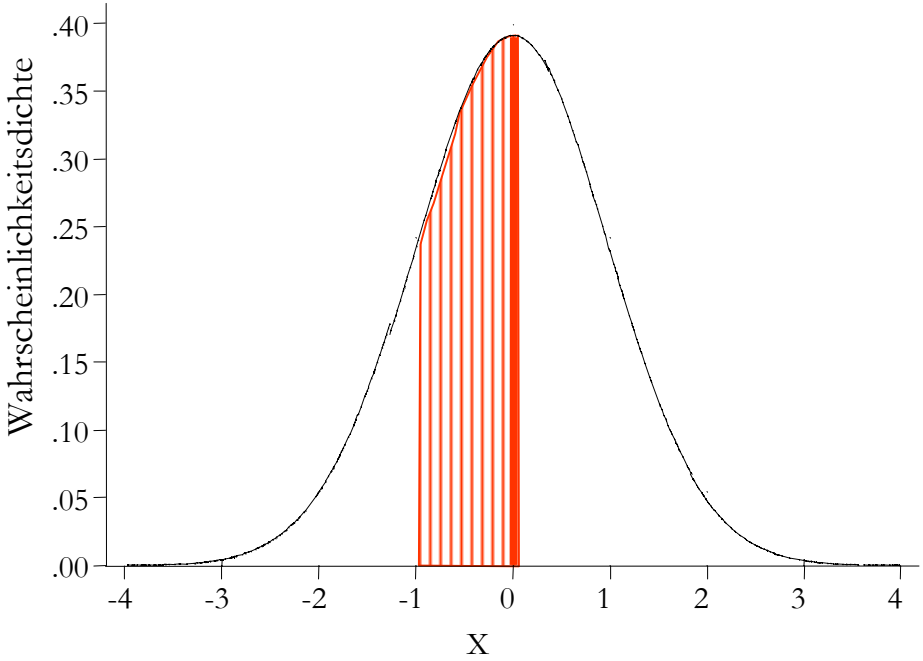
Stetige Wahrscheinlichkeitsverteilungen



So ist in der abgebildeten Wahrscheinlichkeitsverteilung einer normalverteilten Variablen die Wahrscheinlichkeit, dass eine Realisierung in das Intervall zwischen –1 und 0 fällt, die rot eingetragene Fäche unter der Kurve.

Die Normalverteilung ist ein Beispiel für eine ***stetige*** *(kontinuierliche)* ***Wahrscheinlichkeitsverteilung***, bei der der Wertebereich der Realisierungen nicht nur wenige (diskrete) Ausprägungen, sondern unendlich viele reelle Zahlen umfasst.
Da die Wahrscheinlichkeit des Auftretens der Gesamtheit aller Realisierungen eins ist, ist bei stetigen Wahrscheinlichkeitsverteilungen die Wahrscheinlichkeit des Auftretens einer einzelnen Ausprägung immer null.
Angebbar ist immer nur die Wahrscheinlichkeit, mit der eine Realisierung in ein vorgegebenes Intervall fällt.

Wahrscheinlichkeitsdichten

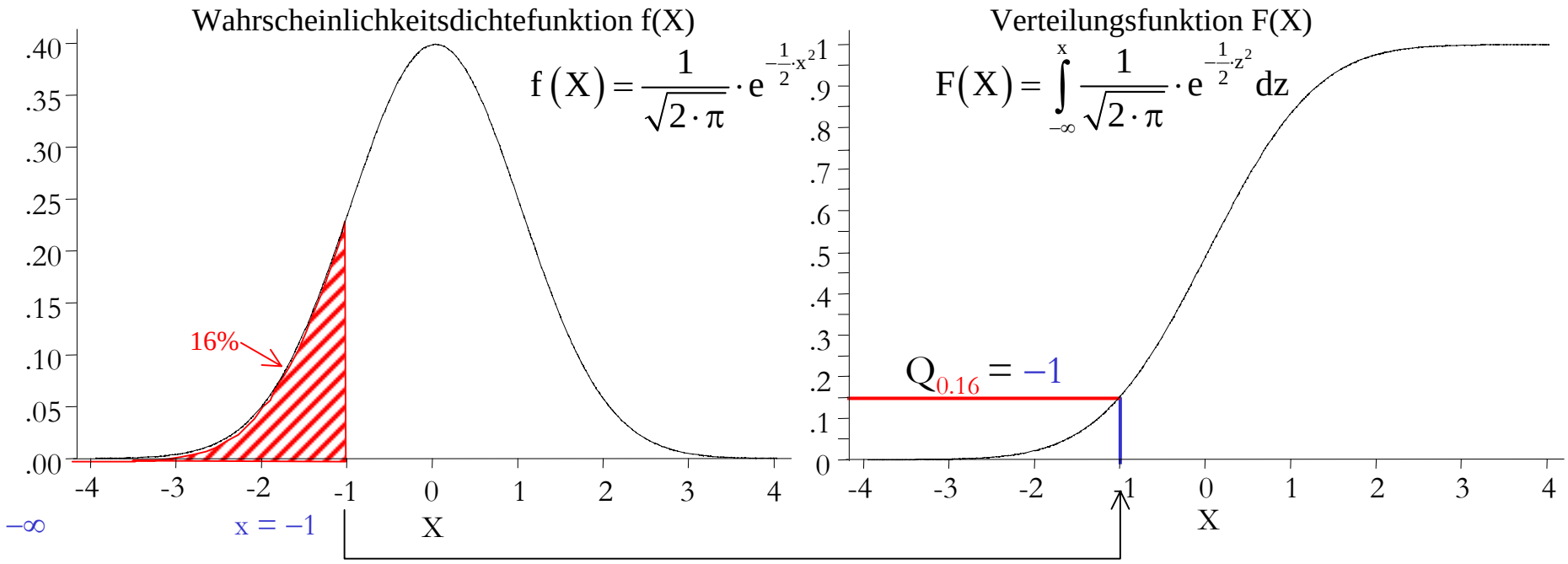


Die Wahrscheinlichkeitsdichte der abgebildeten ***Standardnormalverteilung*** ist folgende Funktion der Ausprägungen x einer Variaben X:

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

Je „dünner“ ein solches Intervall wir, desto geringer ist die Wahrscheinlichkeit, dass eine Realisierung in das Intervall fällt.
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.
Die Länge dieser Linie ist genau der Wert der Funktion, die als Kurvenverlauf in der Abbildung eingezeichnet ist.
Sie wird als ***Wahrscheinlichkeitsdichte*** *(engl.* ***density****)* f(X) bezeichnet.
Das Verhältnis der Dichtewerte zweier Ausprägungen einer stetigen Variablen gibt die relative Chance des Auftretens der beiden Ausprägungen an.

Verteilungsfunktion einer stetigen Zufallsvariablen

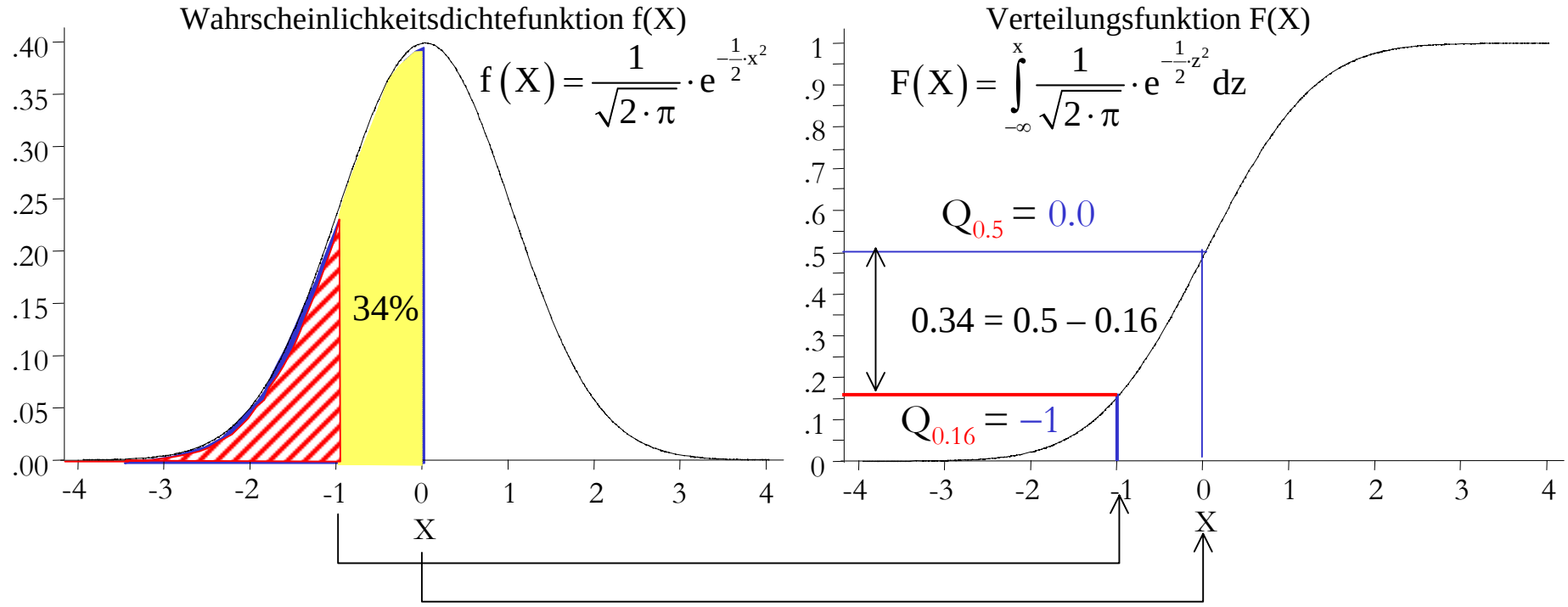


Die ***Verteilungsfunktion*** F(X=x) ist bei einer stetigen Wahrscheinlichkeitsverteilung die Fläche vom linken Rand der Verteilung (bzw. –∞) bis zum Wert X.
Mathematisch ist diese Fäche das bestimmte Integral über die Dichtefunktion von minus unendlich bis x.

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 Verteilunsgfunktion lässt sich auch grafisch darstellen und ergibt bei einer Normalverteilung eine s-förmige Kurve.

Verteilungsfunktion einer stetigen Zufallsvariablen



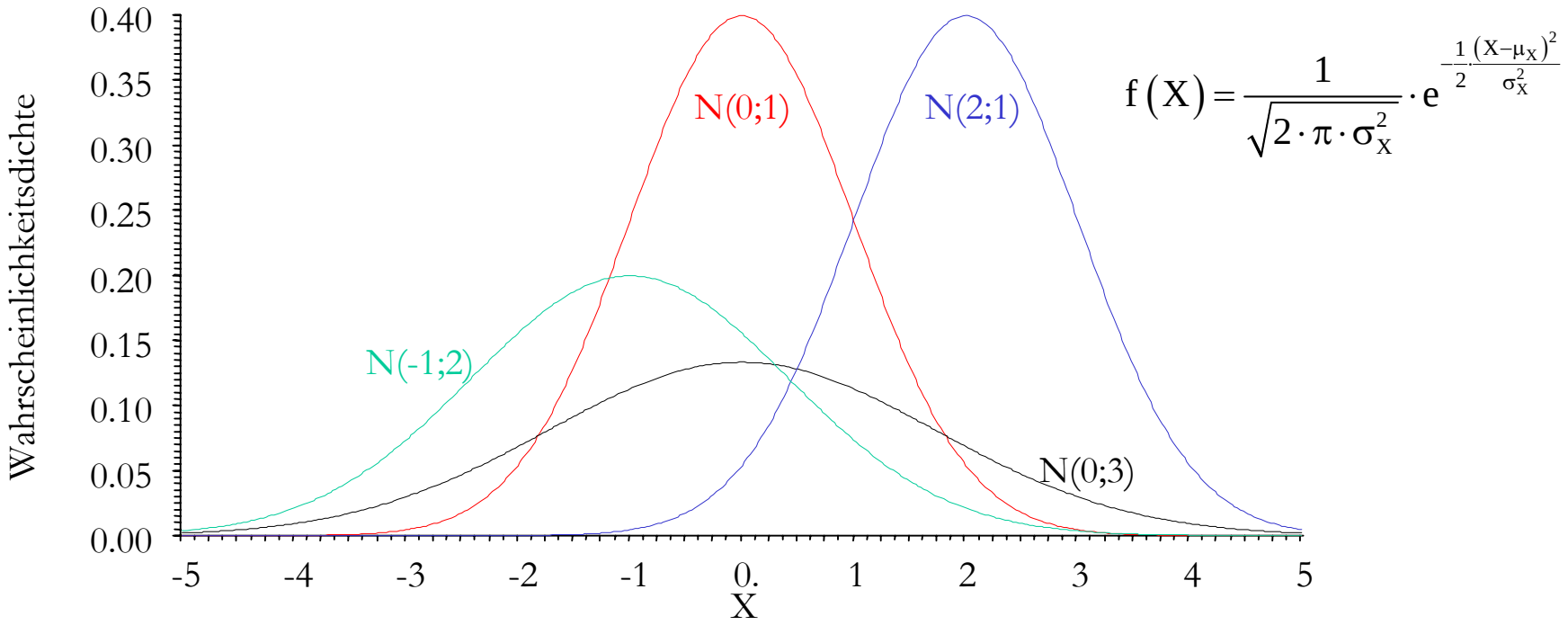
Über die Verteilungsfunktion einer stetigen Zufallsvariablen lassen sich für beliebige Intervalle des Wertebereichs Realisierungswahrscheinlichkeiten berechnen.

Die Quantilwahrscheinlichkeit des Quantilwerts 0 der Standardnormalverteilung ist 0.5 oder 50%.

Die Quantilwahrscheinlichkeit des Quantilwerts –1 der Standardnormalverteilung beträgt 0.16 oder 16%.

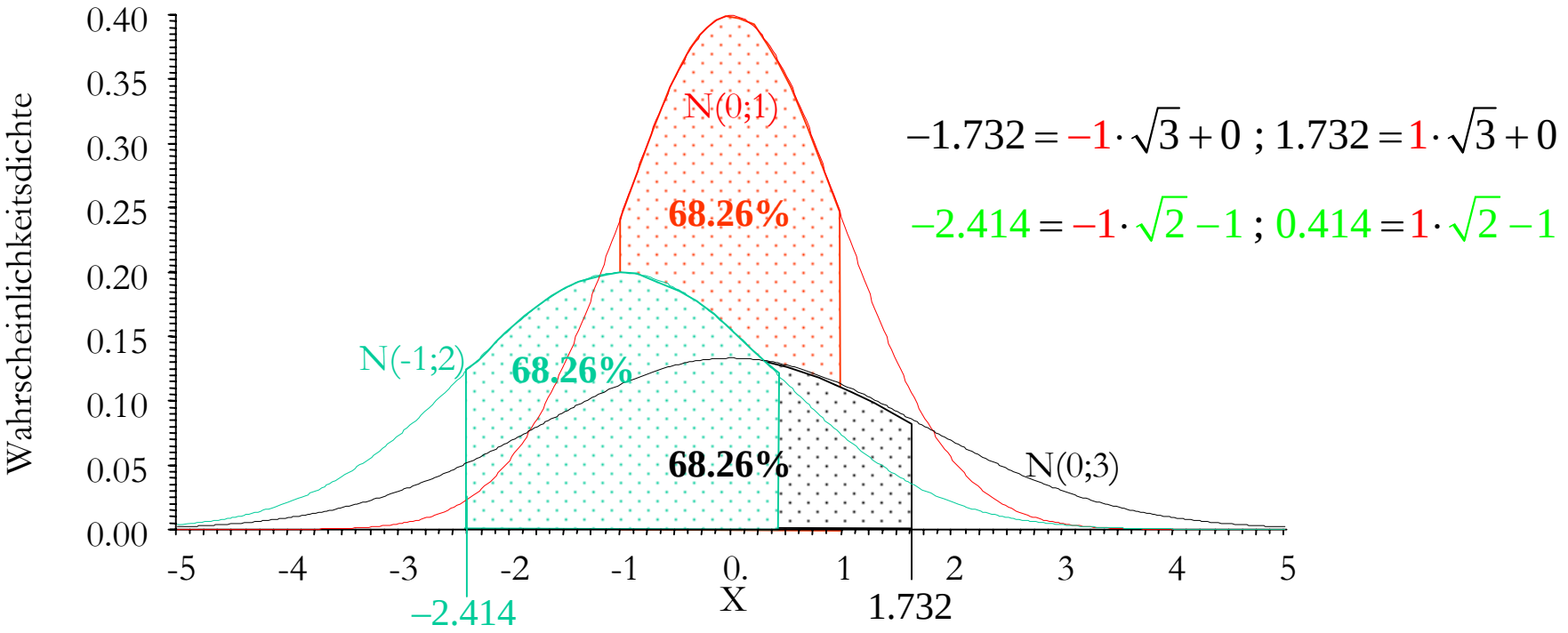
Dann ist die Wahrscheinlichkeit, dass eine standardnormalverteilte Zufallsvariable zwischen –1 und 0 liegt, 34% (= 50% – 16%).

Die Normalverteilung



Die bekannteste stetige Wahrscheinlichkeitsverteilung ist die ***Normalverteilung***. Normalverteilungen haben eine glockenförmige Dichtefunktion.
Die Dichtefunktion einer normalverteilten Zufallsvariable X ist eine Funktion ihres Erwartungswertes und ihrer Varianz. Daher sind Erwartungswert und Varianz (bzw. Standardabweichung) die Parameter einer Normalverteilung.
Um auszudrücken, dass eine Zufallsvariable X mit dem Erwartungswert μ uund 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
0.000	-∞
0.005	-2.57
0.010	-2.326
0.015	-2.170
0.020	-2.054
0.025	-1.960
...	...
0.050	-1.645
...	...
0.100	-1.282
...	...

α	z_α
0.200	-0.842
...	...
0.250	-0.674
...	...
0.300	-0.524
...	...
0.400	-0.253
...	...
0.500	0.000
...	...
0.600	0.253

α	z_α
...	...
0.700	0.524
...	...
0.750	0.674
...	...
0.800	0.842
...	...
0.900	1.282
...	...
0.995	2.576
1.00	∞

Aus den abgebildeten Ausschnitten einer Tabelle mit Z-Werten lässt sich so etwa ablesen,

- dass das 1%-Quantil der Standardnormalverteilung $Q_{0.01;N(0;1)} = -2.326$ ist
- das das 5%-Quantil $Q_{0.05;N(0;1)} = -1.645$ beträgt.

Umgekehrt lässt sich der Tabelle entnehmen,

- dass der Wert –1.96 das 2.5%-Quantil ist, $\Phi(-1.96) = 0.025$
- und der Wert +1.282 das 90%-Quantil, $\Phi(1.282) = 0.90$.

Jede Normalverteilung kann also durch eine einfache Lineartransformation in eine beliebige andere Normalverteilung umgeformt werden.
Darüber hinaus gilt, dass Linearkombinationen von normalverteilten Zufallsvariablen wiederum normalverteilt sind.

Die ***Standardnormalverteilung*** ist eine Normalverteilung mit Erwartungswert null und einer Varianz von eins. Die Quantilwerte einer Standardnormalverteilung werden bisweilen auch als „***Z-Werte***“ bezeichnet.
Aufgrund ihrer Bedeutung gibt es spezifische Symbole. So steht $\varphi(x)$ für die Dichtefunktion und $\Phi(x)$ für die Verteilungsfunktion der Standardnormalverteilung an der Stelle X=x.

Die Normalverteilung

α	z_α
0.000	-∞
0.005	-2.57
0.010	-2.326
0.015	-2.170
0.020	-2.054
0.025	-1.960
...	...
0.050	-1.645
...	...
0.100	-1.282
...	...

α	z_α
0.200	-0.842
...	...
0.250	-0.674
...	...
0.300	-0.524
...	...
0.400	-0.253
...	...
0.500	0.000
...	...
0.600	0.253

α	z_α
...	...
0.700	0.524
...	...
0.750	0.674
...	...
0.800	0.842
...	...
0.900	1.282
...	...
0.995	2.576
1.00	∞

Da Normalverteilungen symmetrisch sind, können bereits aus einer Hälfte der Verteilung alle Quantilwerte berechnet werden.

Generell gilt bei symmetrischen Verteilungen:

$$(Q_{1-\alpha} - \mu) = -(Q_\alpha - \mu) \text{ bzw. bei } \mu=0\text{: } Q_{1-\alpha} = -Q_\alpha$$

$$Q_{0.1} = -1.282 \Rightarrow Q_{0.9} = +1.282$$
$$Q_{0.2} = -0.842 \Rightarrow Q_{0.8} = +0.842$$
$$Q_{0.4} = -0.253 \Rightarrow Q_{0.6} = +0.253$$

So ist der Wert des 10%-Quantils –1.282. Daraus folgt, dass das 90%-Quantil +1.282 sein muss.

Das 90%-Quantil einer Normalverteilung mit Erwartungswert 3 und Varianz 4 berechnet sich dann mit Hilfe der Umkehrung der Z-Transformation $X = Z \cdot \sigma_X + \mu_X$ nach:

$$Q_{0.9;N(3;4)} = Q_{0.9;N(0;1)} \cdot 2 + 3 = 1.282 \cdot 2 + 3 = 5.564$$

bzw.

$$= -Q_{0.1;N(0;1)} \cdot 2 + 3 = -(-1.282) \cdot 2 + 3 = 5.564.$$

Dem Quantilwert 6.29 einer Normalverteilung mit Erwartungswert 3 und Varianz 4 entspricht die Wahrscheinlichkeit:

$$\Phi\left(\frac{6.29-3}{2}\right) = \Phi(1.645) = 1 - \Phi(-1.645) = 1.00 - 0.05 = 0.95$$

Stichprobenverteilungen von Mittelwerten

Der zentrale Grenzwertsatz besagt, dass Summen unabhängiger identisch verteilter Zufallsvariablen mit steigender Zahl der Summanden asymptotisch normalverteilt sind.
Dies gilt unabhängig von der Wahrscheinlichkeitsverteilung der Ausgangsvariablen.
Vorausgesetzt wird allerdings, dass die ersten Momente der Wahrscheinlichkeitsverteilungen der Ausgangsvariablen existieren, d.h. berechenbare reelle Zahlen sind. Diese Bedingung ist bei diskreten Variablen mit begrenzten Wertebereich stets erfüllt.

Jeder Stichprobenmittelwert kann als Summe gleichartiger Summanden dargestellt werden:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{x_i}{n}$$

In einfachen Zufallsauswahlen ohne Zurücklegen können die Summanden als identisch verteilte statistisch unabhängige Zufallsvariablen aufgefasst werden.
Also ist die Kennwerteverteilung von Stichprobenmittelwerten bei einfachen Zufallsauswahlen mit Zurücklegen unabhängig von der Verteilung der interessierenden Größe in der Population asymptotisch normalverteilt.
Bei relativ zur Stichprobengröße sehr großen Populationen gilt dies auch für einfache Zufallsauswahlen ohne Zurücklegen.

Stichprobenverteilungen von Mittelwerten

Aus den Regeln für Linearkombinationen von Zufallsvariablen folgt dann, dass der Erwartungswert und die Varianz der Kennwerteverteilung eine Funktion des Populationsmittelwerts μ(X), der Populationsvarianz $\sigma^2(X)$ und der Stichprobengröße n sind.

Erwartungswert und Varianz der Kennwerteverteilung von Stichprobenmittelwerten sind daher bei ***einfachen Zufallsauswahlen mit Zurücklegen***:

$$\mu(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}\mu(x_i) = \frac{1}{n}\cdot n\cdot\mu(X) = \mu(X) = \mu_X$$

$$\sigma^2(\bar{x}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2(x_i) = \frac{1}{n^2}\cdot n\cdot\sigma^2(X) = \frac{\sigma^2(X)}{n} = \frac{1}{n}\cdot\sigma_X^2$$

Bei ***einfachen Zufallsauswahlen ohne Zurücklegen*** gilt für den Erwartungswert und die Varianz der Kennwerteverteilung von Stichprobenmittelwerten:

$$\mu(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}\mu(x_i) = \frac{1}{n}\cdot n\cdot\mu(X) = \mu_X$$

$$\sigma^2(\bar{x}) = \frac{N-n}{N-1}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2(x_i) = \frac{N-n}{N-1}\cdot\frac{\sigma^2(X)}{n} = \frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\sigma_X^2$$

In beiden Situationen sind die Kennwerteverteilungen asymptotisch normalverteilt.

Stichprobenverteilungen von Mittelwerten

Als Kennwerteverteilung für Stichprobenmittelwerte kann also die Normalverteilung herangezogen werden.

Zu beachten ist allerdings, dass die Normalverteilung in der Regel nur asymptotisch gilt, d.h. bei hinreichend großen Stichproben. Als Faustregel hat die Erfahrung gezeigt, dass die Normalverteilung als Kennwerteverteilung von Stichprobenmittelwerten bereits bei einer Fallzahl ab etwa 30 Fällen hinreichend genau ist:

$$\mathbf{n \geq 30}.$$

Genau genommen kommt es auf die Zahl der Ausprägungen und die Verteilungsform an, ab welcher Fallzahl eine hinreichend genaue Annäherung an die Normalverteilung vorliegt. Bei diskreten Verteilingen mit wenigen Ausprägungen und bei schiefen Verteilungen ist die Annäherung langsamer als bei symmetrischen und unimodalen Verteilungen mit vielen Ausprägungen.

Unabhängig von der Stichprobengröße sind Stichprobenmittelwerte über verschiedene Stichproben hinweg immer dann exakt normalverteilt, wenn bereits die interessierende Variable in der Population normalverteilt ist. Stichprobenmittelwerte sind dann Linearkombinationen von Normalverteilungen und müssen daher ebenfalls normalverteilt sein.

Asymptotische Normalverteilung von Anteilen und Häufigkeiten

Die Binomialverteilung mit den Parametern $b(X; n, \pi_1)$ kann als Summe von n unabhängigen Bernoulli-Verteilungen mit gleicher Wahrscheinlichkeit π_1 aufgefasst werden.
Nach dem zentralen Grenzwertsatz muss sich daher die Binomialverteilung asymptotisch einer Normalverteilung annähern.
Dies gilt tatsächlich. Die Annäherung ist hinreichend genau, wenn gilt:

$$n \cdot \frac{\pi_i}{1-\pi_i} > 9 \quad \text{und} \quad n \cdot \frac{1-\pi_i}{\pi_i} > 9$$

Ist diese Bedingung erfüllt kann anstelle der Binomialverteilung auch eine Normalverteilung mit dem Erwartungswert $\mu_X = n \cdot \pi_1$ und Varianz $\sigma^2_X = n \cdot \pi_1 \cdot (1-\pi_1)$ zur Berechnung von Häufigkeiten bzw. Anteilen bei Zufallsauswahlen mit Zurücklegen verwendet werden.
Da sich die hypergeometrische Verteilung der Binomialverteilung annähert, gilt dies auch bei Zufallsauswahlen ohne Zurücklegen, wobei die Varianz der asymptotischen Normalverteilung dann $\sigma^2_X = (N-n)/(N-1) \cdot n \cdot \pi_1 \cdot (1-\pi_1)$ beträgt.

Asymptotische Normalverteilung von Anteilen und Häufigkeiten

Die hypergeometrische und die Binomialverteilung sind diskret, die Normalverteilung dagegen stetig.
Um dies zu berücksichtigen, wird bei der Berechnung der Wahrscheinlichkeiten von Häufigkeiten jeweils 0.5 zu den ganzzahligen Ausprägungen der Binomialverteilung bzw. der hypergeometrischen Verteilung abgezogen bzw. addiert. Die Wahrscheinlichkeit, dass bei einer Fallzahl von n die Anzahl n_1 der Fälle mit der Ausprägung 1 im Intervall von a bis b liegt, berechnet sich somit bei einfachen Zufallsauswahlen mit Zurücklegen nach:

$$P(a \le n_1 \le b) = \sum_{n_1=a}^{n_1=b} \binom{n}{n_1} \cdot \pi_1^{n_1} \cdot (1-\pi_1)^{n-n_1} \approx \Phi\left(\frac{b+0.5-(n\cdot\pi_1)}{\sqrt{n\cdot\pi_1\cdot(1-\pi_1)}}\right) - \Phi\left(\frac{a-0.5-(n\cdot\pi_1)}{\sqrt{n\cdot\pi_1\cdot(1-\pi_1)}}\right)$$

Bei einfachen Zufallsauswahlen ohne Zurücklegen und großen Populationen nähert sich analog die hypergeometrische Verteilung an die Normalverteilung an nach:

$$P(a \le n_1 \le b) = \sum_{n_1=a}^{n_1=b} \frac{\binom{N_1}{n_1} \cdot \binom{N-N_1}{n-n_1}}{\binom{N}{n}} \approx \Phi\left(\frac{b+0.5-(n\cdot\pi_1)}{\sqrt{n\cdot\pi_1\cdot(1-\pi_1)\cdot\frac{N-n}{N-1}}}\right) - \Phi\left(\frac{a-0.5-(n\cdot\pi_1)}{\sqrt{n\cdot\pi_1\cdot(1-\pi_1)\cdot\frac{N-n}{N-1}}}\right)$$

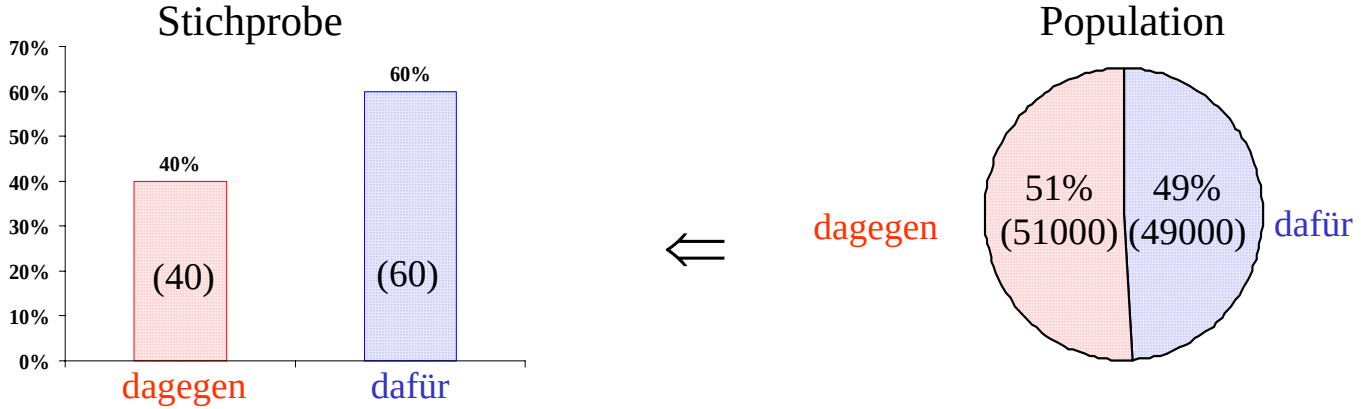
Asymptotische Normalverteilung von Anteilen und Häufigkeiten

Bei der asymptotischen Berechnung der Kennwerteverteilung von Anteilen wird in der Regel auf die Stetigkeitskorrektur ±0.5/n verzichtet.
Die asymptotische Kennwerteverteilung ist dann:

$$f\left(p_1\right) \approx N\left(\pi_1; \frac{\pi_1 \cdot \left(1-\pi_1\right)}{n}\right) \qquad \text{mit Zurücklegen}$$

$$\approx N\left(\pi_1; \frac{\pi_1 \cdot \left(1-\pi_1\right)}{n} \cdot \frac{N-n}{n-1}\right) \qquad \text{ohne Zurücklegen}$$

Schätzen von Anteilen, Mittelwerten und Varianzen



Eine der wichtigsten Anwendungen der Statistik in den Sozialwissenschaften besteht darin, anhand von Stichprobendaten Aussagen über eine Grundgesamtheit (Population) treffen zu können.

Es interessiert z.B. der Anteil π_1 der jenigen Personen in einer Stadt, die für die Einrichtung einer Ganztagsschule sind. In einer einfachen Zufallsauswahl von n=100 Personen sind p_1=60% für die Einrichtung.

Es liegt nahe, den Stichprobenanteil p_1 als Schätzung des unbekannten Populationsanteils π_1 zu verwenden.

Da aber von einer Teilmenge (der Stichprobe) auf eine umfassendere Allgemeinheit (die Population) geschlossen wird, handelt es sich bei der Schätzung um einen Induktionsschluss der prinzipiell unsicher ist und fehlerhaft sein kann.

So ist es im Beispiel denkbar, dass in der Grundgesamtheit von N=100 000 nicht eine Mehrheit von 60%, sondern nur eine Minderheit von 49% für die Einrichtung der Ganztagsschule ist.

Schätzer und Schätzung

Für eine einzelne Schätzung lässt sich grundsäzlich nicht angeben, ob ihr Wert mit dem zu schätzenden Populationswert übereinstimmt oder ob sie sehr vom gesuchten Wert abweicht.

Bei Zufallsauswahlen ist jede Schätzung ein Zufallsexperiment und jede Schätzung ein mögliches Ereignis dieses Zufallsexperiments.Eine Schätzung kann daher als Realisierung einer Zufallsvariable aufgefasst werden.
Zufallsvariablen, die für Schätzungen verwendet werden, heißen ***Schätzer***.
Eine konkrete ***Schätzung*** ist dann eine von vielen möglichen Realisierungen eines Schätzers.

Die Kennwerteverteilung des Schätzers, d.h. die Wahrscheinlichkeits(dichte)verteilung der Zufallsvariable erlaubt Aussagen darüber, wie wahrscheinlich Schätzungen sind, die nahe beim zu schätzenden Populationswert liegen.

Die theoretische Statistik versucht Schätzer zu konstruieren, die möglichst gute Eigenschaften aufweisen. Die wichtigsten Eigenschaften sind dabei

- Erwartungstreue oder Unverzerrtheit
- Konsistenz und
- Effizienz.

Eigenschaften von Schätzern

Erwartungstreue oder ***Unverzerrtheit***
Ein Schätzer ist **unverzerrt** oder ***erwartungstreu*** (engl. *unbiased*), wenn der Erwartungswert der Kennwerteverteilung des Schätzers mit dem zu schätzenden Populationswert übereinstimmt;

$$\mu\left(\hat{\theta}\right) = \theta$$

In der Statistik wird das griechische kleine Theta („θ“) oft als allgemeinses Symbol für einen beliebigen Parameter verwendet.
Ein kleines Dach („^“) über dem Symbol kennzeichnet dann einen Schätzer oder eine Schätzung.

Bei einfachen Zufallsauswahlen lässt sich die Kennwerteverteilung des Stichprobenanteils aus der Binomialverteilung oder der hypergeometrischen Verteilung berechnen.
In beiden Fällen ist der Erwartungswert der Kennwerteverteilung genau der Anteil $\pi_1 = N_1/N$ *der Elemente in der Population, die die betrachtete Eigenschaft aufweisen.*
Der Stichprobenanteil ist daher bei einfachen Zufallsauswahlen ein erwartungstreuer Schätzer.

Eigenschaften von Schätzern

Konsistenz

Ein Schätzer ist ***konsistent***, wenn bei steigender Stichprobenfallzahl die Wahrscheinlichkeit gegen eins geht, dass der Abstand zwischen dem zu schätzenden Parameter und dem Stichprobenkennwert gegen null geht.

$$\lim_{n\to\infty}\left(\Pr\left(\left|\hat{\theta}-\theta\right|=0\right)\right)=1$$

Aus dem Gesetz der großen Zahl folgt, dass die Wahrscheinlichkeit einer beliebig kleinen Abweichung zwischen Stichprobenanteil und Populationsanteil bei einfachen Zufallsauswahlen gegen eins geht, wenn die Fallzahl über alle Grenzen wächst.

Der Stichprobenanteil ist daher bei einfachen Zufallsauswahlen mit Zurücklegen ein konsistenter Schätzer des Populationsanteils.

Bei einfachen Zufallsauswahlen ohne Zurücklegen ist der Anteil der ausgewählten Fälle ebenfalls gleich dem Populationsanteil, wenn im Extremfall alle Fälle ausgewählt werden.

Eigenschaften von Schätzern

Effizienz

Die Realisationen der Kennwerteverteilung sollen möglichst gering um den zu schätzenden Populationsparameter streuen. Ein Kennwert ist ***effizient***, wenn es keinen anderen Schätzer gibt, der mit einer geringeren Streuung um den zu schätzenden Parameter streut.

Als Maß für die Effizienz wird üblicherweise der Erwartungswert der quadrierten Abstände vom zu schätzenden Parameterwert herangezogen, der nach der englischen Bezeichnung ***mean squared error*** (MSE) heißt:

$$\mathrm{MSE} = \mu\left(\left(\hat{\theta} - \theta\right)^2\right) = \sigma^2\left(\hat{\theta}\right) + \left(\mu\left(\hat{\theta}\right) - \theta\right)^2$$

Die Gleichung zeigt, dass MSE auch als Summe der Varianz der Kennwerteverteilung eines Schätzers plus der quadrierten Verzerrung (engl. *bias*), das ist der quadrierte Abstand zwischen dem Erwartungswert des Schätzers und dem zu schätzendem Parameter dargestellt werden kann.

Zur Schätzung des Populationsmittelwertes kann bei einer symmetrischen, unimodalen Verteilung sowohl der Stichprobenmittelwert als auch der Stichprobenmedian herangezogen werden. Effizienter ist die Kenngröße, deren Kennwerteverteilung mit einer geringeren Streuung um den Populationsmittelwert variiert. Welche das ist, hängt von Verteilung ab.
Bei einfachen Zufallsstichproben aus normalverteilten Populationen ist der Stichprobenmittelwert ein effizienterer Schätzer des Erwartungswert als der Stichprobenmedian.

Standardfehler

Bei unverzerrten Schätzern ist die quadrierte Verzerrung definitionsgemäß null, so dass die Effizienz in diesem Fall über die Varianz der Kennwerteverteilung gemessen werden kann. Anstelle der Varianz wird meist die Standardabweichung einer Kennwerteverteilung als Streuungsmaß verwendet.
Die Standardabweichung eines Schätzers wird als ***Standardschätzfehler*** oder ***Standardfehler*** bezeichnet.

Da bei einfachen Zufallsauswahlen ohne Zurücklegen der Stichprobenanteil ein unverzerrter Schätzer des Populationsanteils ist, ist der Standardfehler die Quadratwurzel aus dem MSE. Sie ist aus der Standardabweichung der hypergeometrischen Verteilung berechenbar:

$$\sigma(p_1) = \sqrt{\frac{1}{n} \cdot \left(\frac{N_1}{N}\right) \cdot \left(1 - \frac{N_1}{N}\right) \cdot \frac{N-n}{N-1}}$$

Üblicherweise wird neben der Schätzung eines Populationsparameters auch der Standardfehler der Kennwerteverteilung aus den Stichprobendaten geschätzt.

Schätzung von Häufigkeiten, Anteilen, Mittelwerten und Varianzen

Absolute (N_1) und relative Häufigkeiten ($\pi_1=N_1/N$) eines Merkmals A in der Population bzw. einer dichotomen Variable mit den Ausprägungen „1“ (für „Merkmal vorhanden“) und „0“ (für „Merkmal nicht vorhanden“) können in einfachen Zufallsauswahlen mit und ohne Zurücklegen über die korrespondierenden Stichprobenstatisiken n_1 und p_1 konsistent und unverzerrt geschätzt werden. Die Kennwerteverteilung der Schätzer ist dann binomialverteilt bzw. hypergeometrisch verteilt und nähert sich bei hinreichend großen Fallzahlen asymptotisch einer Normalverteilung an.

Analog können auch Populationsmittelwerte $\mu(X)$ einer Variable X in einfachen Zufallsauswahlen erwartungstreu und konsistent über den Stichprobenmittelwert $\bar{x}$ geschätzt werden. Die Kennwerteverteilung ist bei in der Population normalverteilten X normalverteilt. Bei anderen Verteilungen ist die Kennwerteverteilung asymptotisch normalverteilt.

Es liegt nahe, zur Schätzung einer Populationsvarianz die Stichprobenvarianz zu verwenden. Tatsächlich ist diese ein zwar konsistenter, allerdings kein erwartungstreuer Schätzer. Der Erwartungswert der Stichprobenvarianz ist bei einfachen Zufallsauswahlen (ohne Zurücklegen) nämlich:

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

Schätzung von Populationsvarianzen und Standardabweichungen

Die Höhe des Verzerrungsfaktors (n-1)/n nähert sich 1, wenn die Stichprobenfallzahl n ansteigt. Der Schätzer ist daher zumindest *asymptotisch erwartungstreu*.

Zur Schätzung einer Populationsvarianz wird i.a. ein bei jeder Fallzahl erwartungstreuer Schätzer verwendet, der sich aus der Stichprobenvarianz mal dem Kehrwert des Verzerrungsfaktors ergibt. Der erwartungstreue Schätzer der Populationsvarianz ist daher:

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

Der Standardfehler des erwartungstreuen Schätzers der Populationsvarianz hängt von der Verteilung in der Population ab. Ist diese (annähernd) normalverteilt, gilt:

$$\sigma\left(\hat{\sigma}_X^2\right) = \sigma_X^2 \cdot \sqrt{\frac{2}{n-1}}$$

Die Kennwerteverteilung ist bei normalverteilten Populationen proportional zur sogenannten Chiquadratverteilung. Konfidenzintervalle auf der Basis von Chiquadratverteilungen werden aber meistens nicht berechnet.
Für die Schätzung der Populationsstandardabweichung wird meistens die Wurzel aus der geschätzten Populationsvarianz benutzt, die allerdings nicht erwartungstreu ist:

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

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.
So ist die Wahrscheinlichkeit, dass ein Stichprobenanteil p_1=0.6 (=60/100) bei einer Population von N=100 000 und einer Stichprobengröße von n=100 einem Populationsanteil π_1=0.60 entspricht nur etwa 8%:

$$\Pr(p_1 = 0.6) \approx \Phi\left(\frac{60 + 0.5 - 0.6 \cdot 100}{\sqrt{100 \cdot 0.6 \cdot 0.4 \cdot \frac{100000 - 100}{100000 - 1}}}\right) - \Phi\left(\frac{60 - 0.5 - 0.6 \cdot 100}{\sqrt{100 \cdot 0.6 \cdot 0.4 \cdot \frac{100000 - 100}{100000 - 1}}}\right)$$

$$= \Phi(0.102) - \Phi(-0.102) \approx 0.08$$

In 92% aller Stichproben ist also mit Abweichungen zu rechnen.

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***.

Vorgehensweise bei Intervallschätzung

Mit Hilfe der Kennwerteverteilung eines Schätzers können Intervallschätzungen berechnet werden. Dies kann am Beispiel der Schätzung eines Stichprobenmittelwerts verdeutlicht werden.
Bei einer einfachen Zufallsauswahl aus einer normalverteilten Population ist der Stichprobenmittelwert um den zu schätzenden Populationsmittelwert normalverteilt:

$$f\left(\bar{X}\right) = N\left(\mu_X; \frac{\sigma_X^2}{n}\right)$$

Aus der Normalverteilung lässt sich ein Intervall berechnen, in dem der Stichprobenmittelwert mit einer Wahrscheinlichkeit von z.B. 90% liegt:

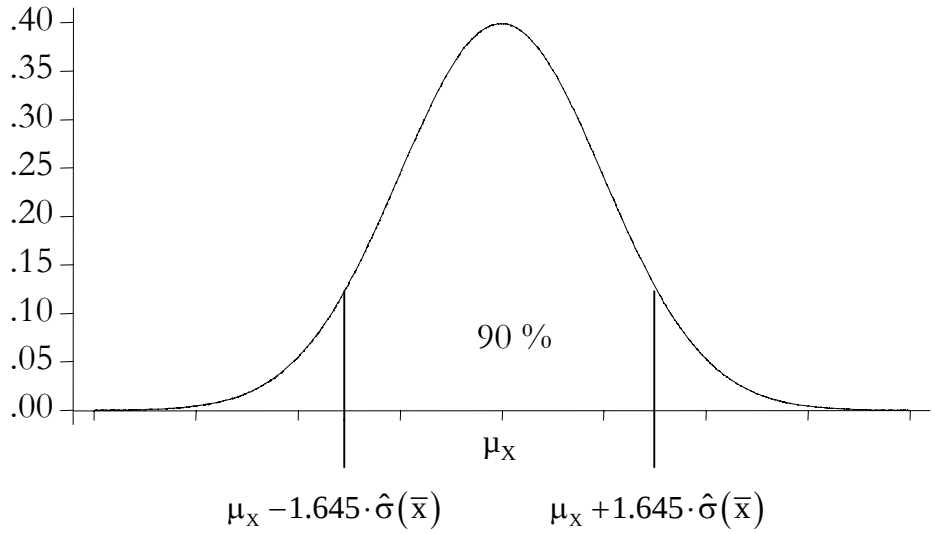
90% aller Realisationen liegen zwischen dem 95%-Quantil und dem 5%-Quantil der Standardnormalverteilung.

$$\begin{aligned} 0.9 &= 0.95 - 0.05 = \Phi\left(1.645\right) - \Phi\left(-1.645\right) \\ &= \Pr(-1.645 \le Z \le 1.645) \\ &= \Pr\left(-1.645 \le \frac{\bar{X} - \mu_X}{\sigma\left(\bar{X}\right)} \le 1.645\right) \end{aligned}$$

Durch Umlehrung der Z-Transformation werden die Quantilgrenzen für die Normalverteilung des Stichprobenmittelwerts berechnet.

$$\begin{aligned} 0.9 &= \Pr\left(-1.645 \cdot \sigma\left(\bar{X}\right) \le \bar{X} - \mu_X \le 1.645 \cdot \hat{\sigma}\left(\bar{X}\right)\right) \\ &= \Pr\left(\mu_X - 1.645 \cdot \sigma\left(\bar{X}\right) \le \bar{X} \le \mu_X + 1.645 \cdot \sigma\left(\bar{X}\right)\right) \\ &= \Pr\left(\mu_X - 1.645 \cdot \sqrt{\frac{\sigma_X^2}{n}} \le \bar{X} \le \mu_X + 1.645 \cdot \sqrt{\frac{\sigma_X^2}{n}}\right) \end{aligned}$$

Vorgehensweise bei Intervallschätzung



Mit einer Wahrscheinlichkeit von 90% wird ein Stichprobenmittelwert also in einem Intervall realisiert, das ±1.645 Standardfehler um den gesuchten Erwartungswert liegt.

Das Intervall lässt sich so umformen, dass es zu einem Intervall um den Populationsmittelwert wird:

$$0.9 = \Pr\left(\mu_X - 1.645 \cdot \sigma\left(\bar{X}\right) \le \bar{X} \le \mu_X + 1.645 \cdot \sigma\left(\bar{X}\right)\right)$$

$$= \Pr\left(-\bar{X} - 1.645 \cdot \hat{\sigma}\left(\bar{X}\right) \le -\mu_X \le -\bar{X} + 1.645 \cdot \sigma\left(\bar{X}\right)\right)$$

$$= \Pr\left(\bar{X} + 1.645 \cdot \sigma\left(\bar{X}\right) \ge \mu_X \ge \bar{X} - 1.645 \cdot \sigma\left(\bar{X}\right)\right)$$

$$= \Pr\left(\bar{X} - 1.645 \cdot \sigma\left(\bar{X}\right) \le \mu_X \le \bar{X} + 1.645 \cdot \sigma\left(\bar{X}\right)\right)$$

Ein solches Intervall, dass mit einer bestimmten Wahrscheinlichkeit zu beobachten ist, wird als ***Konfidenzintervall*** bezeichnet.

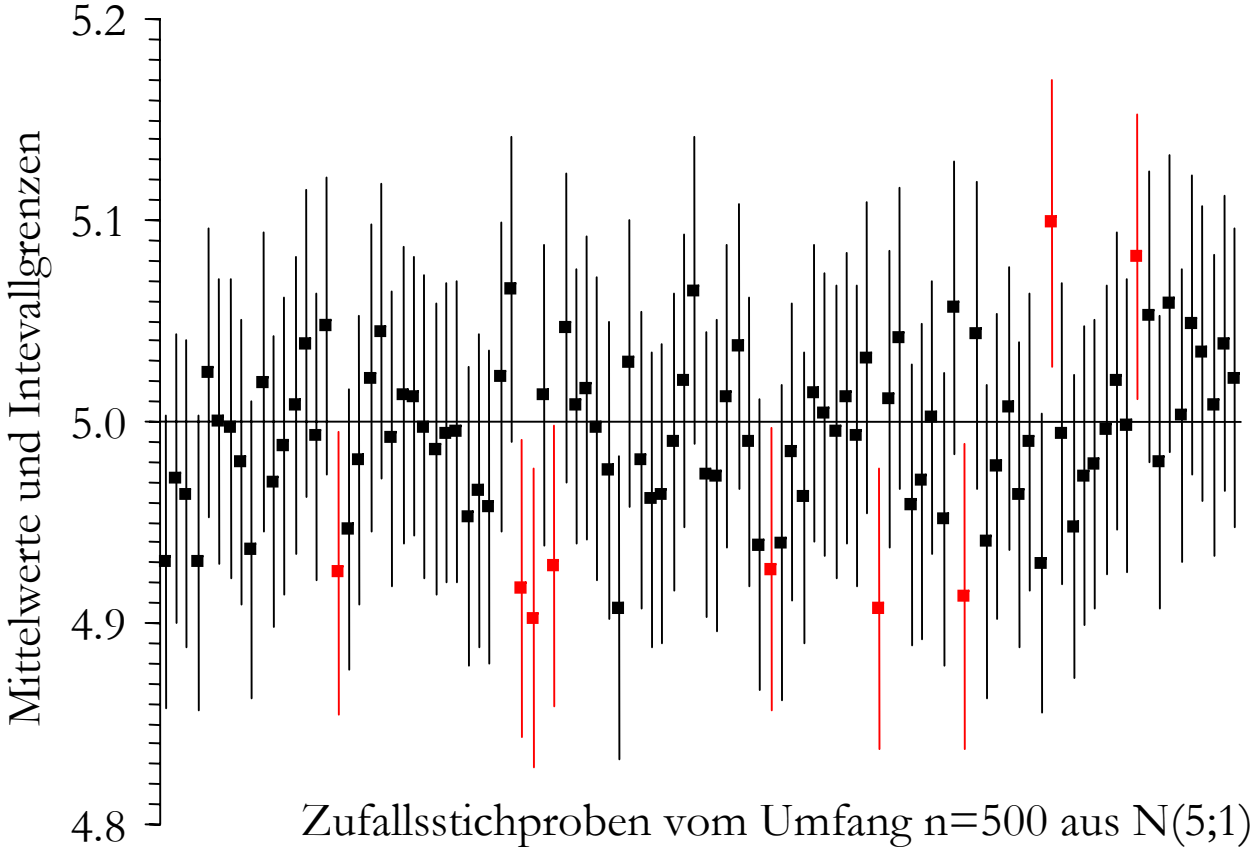
Interpretation von Konfidenzintervallen

Durch die Umformung ist das Intervall selbst bzw. sind seine Intervallgrenzen Zufallsvariablen. Mit einer vorgegebenen Wahrscheinlichkeit, im Beispiel 90%, liegen die Intervallgrenzen so, dass der zu schätzende Populationsmittelwert innerhalb der Intervallgrenzen ist.
Die Wahrscheinlichkeitsaussage bezieht sich nicht auf den unbekannten Parameter, sondern auf die Zufallsvariable „Konfidenzintervall“.
Die Behauptung, dass der unbekannte Populationsparameter mit bekannter Wahrscheinlichkeit in einem berechneten Intervall liegt, wäre daher falsch.

Wenn die Realisationen von Konfidenzintervallen mit einer bekannten Wahrscheinlichkeit den zu schätzenden Parameter überdecken, dann ist die Wahrscheinlichkeit, dass dies nicht der Fall ist, gleich eins minus dieser Wahrscheinlichkeit, im Beispiel also 100%-90% = 10%. Die Wahrscheinlichkeit eines Fehlers wird als ***Irrtumswahrscheinlichkeit*** bezeichnet und durch den kleinen griechischen Buchstaben α (alpha) gekennzeichnet.

Die Berechnungsart von Konfidenzintervallen führt also dazu, dass der Anteil aller Konfidenzintervalle, die den Populationsparameter überdecken, gleich der vorgegebenen Wahrscheinlichkeit ist.
Wenn diese Vertrauenswahrscheinlichkeit hoch bzw. die Irrtumswahrscheinlichkeit klein ist, dann ist das Vertrauen berechtigt, dass auch ein konkret berechnetes Intervall den zu schätzenden Wert tatsächlich enthält, auch wenn unbekannt bleibt, ob dies tatsächlich der Fall ist.

Interpretation von Konfidenzintervallen



Die Abbildung zeigt 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 5.0, neun dagegen nicht.

Vorgehensweise bei Intervallschätzung

Aus dem Beispiel lässt sich die generelle Vorgehensweise bei der Berechnung von Konfidenzintervallen verallgemeinern:

Schritt 1:

Im ersten Schritt ist eine Stichprobenstatistik auszuwählen, für dessen Kennwerteverteilung gilt, dass der zu schätzende Populationsparameter der Erwartungswert der Kennwerteverteilung ist.

Schritt 2:

Im zweiten Schritt wird die Irrtumswahrscheinlichkeit bzw. umgekehrt die Vertrauenswahrscheinlichkeit festgelegt.
In der Sozialforschung werden üblicherweise Irrtumswahrscheinlichkeiten von 5% oder 1% akzeptiert und entsprechend 95%- oder 99%-Konfidenzintervalle berechnet.
Je kleiner die Irrtumswahrscheinlichkeit, desto größer sind die Längen der Konfidenzintervalle. Wenn ein Konfidenzintervall zu lang ist, hat es kaum Aussagekraft.

Schritt 3:

Nach der Festlegung der Irrtumswahrscheinlichkeit α kann das Intervall berechnet werden. Dazu werden Quantile der Kennwerteverteilung benötigt. Die Intervallgrenzen ergeben sich dann nach:

Untergrenze = Punktschätzung – $\alpha/2$-Quantil der Kennwerteverteilung
Obergrenze = Punktschätzung + $(1 – \alpha/2)$-Quantil der Kennwerteverteilung

Konfidenzintervalle für Populationsanteile

Bei einfachen Zufallsauswahlen ist der Stichprobenanteil ein konsistenter und erwartungstreuer Schätzer des entsprechenden Populationsanteils. Die Kennwerteverteilung lässt sich bei einfachen Zufallsauswahlen ohne Zurücklegen über die hypergeometrische Verteilung und bei einfachen Zufallsauswahlen mit Zurücklegen (bzw. wenn die Population um ein Vielfaches größer ist als die Stichprobe: $N > 20 \cdot n$) über die Binomialverteilung berechnen. Bei den in der Umfrageforschung üblichen großen Stichproben nähern sich beide Kennwerteverteilungen asymptotisch einer Normalverteilung an.

Die Annäherung ist für praktische Zwecke hinreichend genau, wenn:

$$n \cdot \frac{\pi_i}{1-\pi_i} > 9 \quad \text{und} \quad n \cdot \frac{1-\pi_i}{\pi_i} > 9$$

Der Standardfehler $\sigma(p_1)$ des Schätzers ist:

$$\sigma(p_1) = \sqrt{\frac{\pi_1 \cdot (1-\pi_1)}{n}}$$ bei einfacher Zufallsauswahl mit Zurücklegen

$$= \sqrt{\frac{\pi_1 \cdot (1-\pi_1)}{n} \cdot \frac{N-n}{N-1}}$$ bei einfacher Zufallsauswahl ohne Zurücklegen

Da die Berechnung des Standardfehlers die Kenntnis des zu schätzenden Popualtionsanteils π_1 voraussetzt, wird in der Praxis oft der geschätzte Standardfehler verwendet, bei dem in der Gleichung der Populationsanteil durch seinen Schätzer ersetzt wird.

Konfidenzintervalle für Populationsanteile

Der geschätzte Standardfehler der kennwerteverteilung eines Stichprobenanteils berechnet sich dann nach:

$$\hat{\sigma}(p_1) = \sqrt{\frac{p_1 \cdot (1 - p_1)}{n}}$$ bei einfacher Zufallsauswahl mit Zurücklegen

$$= \sqrt{\frac{p_1 \cdot (1 - p_1)}{n} \cdot \frac{N - n}{N - 1}}$$ bei einfacher Zufallsauswahl ohne Zurücklegen

Als Faustregel gilt: Wenn n > 60, dann ist die Schätzung des Standardfehlers für praktische Anwendungen genau genug.
Bei kleineren Fallzahlen kann der maximal mögliche Standardfehler verwendet werden, der sich ergibt, wenn der Populationsanteil π_1=0.5 ist:

$$\sigma(p_1) \leq \frac{0.5}{\sqrt{n}}$$ bei einfacher Zufallsauswahl mit Zurücklegen

$$\leq \frac{0.5}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N - 1}}$$ bei einfacher Zufallsauswahl ohne Zurücklegen

Bei der Berechnung von Konfidenzintervalle für Anteile wird die asymptotische Annäherung der Kennwerteverteilung an die Normalverteilung genutzt.

Konfidenzintervalle für Populationsanteile

Da die Normalverteilung symmetrisch um den Erwartungswert verteilt ist, ergeben sich durch Anwendung der Z-Transformation die Grenzen des (1-α)-Konfidenzintervalls nach:

$$\text{c.i.}(\pi_1) = p_1 \pm \sqrt{\frac{p_1 \cdot (1-p_1)}{n}} \cdot z_{\alpha/2} = p_1 \pm \sqrt{\frac{p_1 \cdot (1-p_1)}{n}} \cdot z_{1-\alpha/2}$$

α	z_α
0.000	-∞
0.005	-2.57
0.010	-2.326
0.015	-2.170
0.020	-2.054
0.025	**-1.960**
0.050	-1.645
0.100	-1.282

Die Berechnung ist hinreichend genau, wenn gilt:

(a) $n \cdot p_1 / (1-p_1) > 9$ bzw. $n \cdot (1-p_1) / (p_1) > 9$

(b) $n > 60$

Als Beispiel soll für das Eingangsbeispiel der Stichprobe von n=100 und einem Stichprobenanteil von p_1 = 60% Befürwortern von Ganztagsschulen ein 95%Konfidenzintervall berechnet werden.

Bei einem 95%-Konfidenzintervall beträgt die Irrtumswahrscheinlichkeit (100% – 95%) = 5% = 0.05. Benötigt wird somit das Quantil der Standardnormalverteilung mit der Qauntilwahrscheinlicheit 0.05/2 = 0.025. Der Quantilwert ist –1.96.

Da die Normalverteilung symmetrisch ist, kann anstelle des 2.5%-Qiantils auch das 97.5%-Quantil (0.975 = 1 – α/2) verwendet werden, dessen Wert +1.96 beträgt.

Die Grenzen des 95%-Konfidenintervalls berechnen sich dann 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

$$\text{c.i.}(\pi_1) = 0.6 \pm \sqrt{\frac{0.6 \cdot 0.4}{100}} \cdot 1.96 = 0.6 \pm 0.096 = \left[0.504, 0, 696\right]$$

Mit einer Irrtumswahrscheinlichkeit von 5% liegen die Intervallgrenzen des Konfidenzintervalls so, dass der Populationsmittelwert dazischen liegt.
Vermutlich liegt der Anteil der Befürworter somit zwischen 50.4% und 69.6%.
Die Anwendungsvoraussetzungen sind erfüllt, da gilt:

$$100 \cdot 0.4 / 0.6 = 66.7 > 9 \text{ und } 100 > 60$$

Konfidenzintervalle für Populationsmittelwerte

Bei einfachen Zufallsauswahlen ist der Stichprobenmittelwert ein konsistenter und erwartungstreuer Schätzer des entsprechenden Populationsmittelwerts.
Aus dem zentralen Grenzwertsatz folgt, dass unabhängig von der Verteilung in der Population ein Stichprobenmittelwert asymptotisch normalverteilt ist.
Die Annäherung ist für praktische Anwendungen genau genug, wenn $n > 30$.

Der Standardfehler des Schätzers beträgt:

$$\sigma(\bar{x}) = \sqrt{\frac{\sigma_X^2}{n}} = \frac{\sigma_X}{\sqrt{n}}$$ bei einfacher Zufallsauswahl mit Zurücklegen

$$= \sqrt{\frac{\sigma_X^2}{n} \cdot \frac{N-n}{N-1}} = \frac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$ bei einfacher Zufallsauswahl ohne Zurücklegen

In der Regel ist die Populationsstandardabweichung σ_X unbekannt. Der Standardfehler muss dann aus den Stichprobendaten berechnet werden, wozu der erwartungstreue Schätzer der Populationsvarianz verwendet wird.
Die Kennwerteverteilung bleibt auch dann asymptotisch normalverteilt.

Konfidenzintervalle für Populationsmittelwerte

Der geschätzte Standardfehler berechnet sich bei einfachen Zufallsauswahlen nach:

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

mit Zurücklegen

$$\hat{\sigma}(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n \cdot (n-1)}} \cdot \sqrt{\frac{N-n}{N-1}} \dot{=} \sqrt{\frac{SS_X}{n \cdot (n-1)}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{s_X}{\sqrt{n-1}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{\hat{\sigma}_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

ohne Zurücklegen

Das $(1-\alpha/2)$-Konfidenzintervall des Mittelwerts mit der Irrtumswahrscheinlichkeit α ergibt sich dann nach:

$$c.i.(\mu_X) = \bar{x} \pm \hat{\sigma}(\bar{x}) \cdot z_{\alpha/2} = \bar{x} \pm \hat{\sigma}(\bar{x}) \cdot z_{1-\alpha/2}$$

Die asymptotische Annäherung an die Normalverteilung ist in der Regel hinreichend genau, wenn der Stichprobenumfang größer oder gleich 30 ist:

$$n \geq 30.$$

Konfidenzintervalle für Populationsmittelwerte aus normalverteilten Populationen

Ist die interessierende Variable X in der Population normalverteilt, dann ist bei einer einfachen Zufallsauswahl der Stichprobenmittelwert nicht nur asymptotisch, sondern bei jeder Stichprobengröße normalverteilt.

Allerdings ist die aus dem Mittelwert über die Z-Transformation berechnete Variable:

$$Z = \frac{\bar{x} - \mu_X}{\sigma(\bar{x})} = \frac{\bar{x} - \mu_X}{\sigma(x)/\sqrt{n}}$$

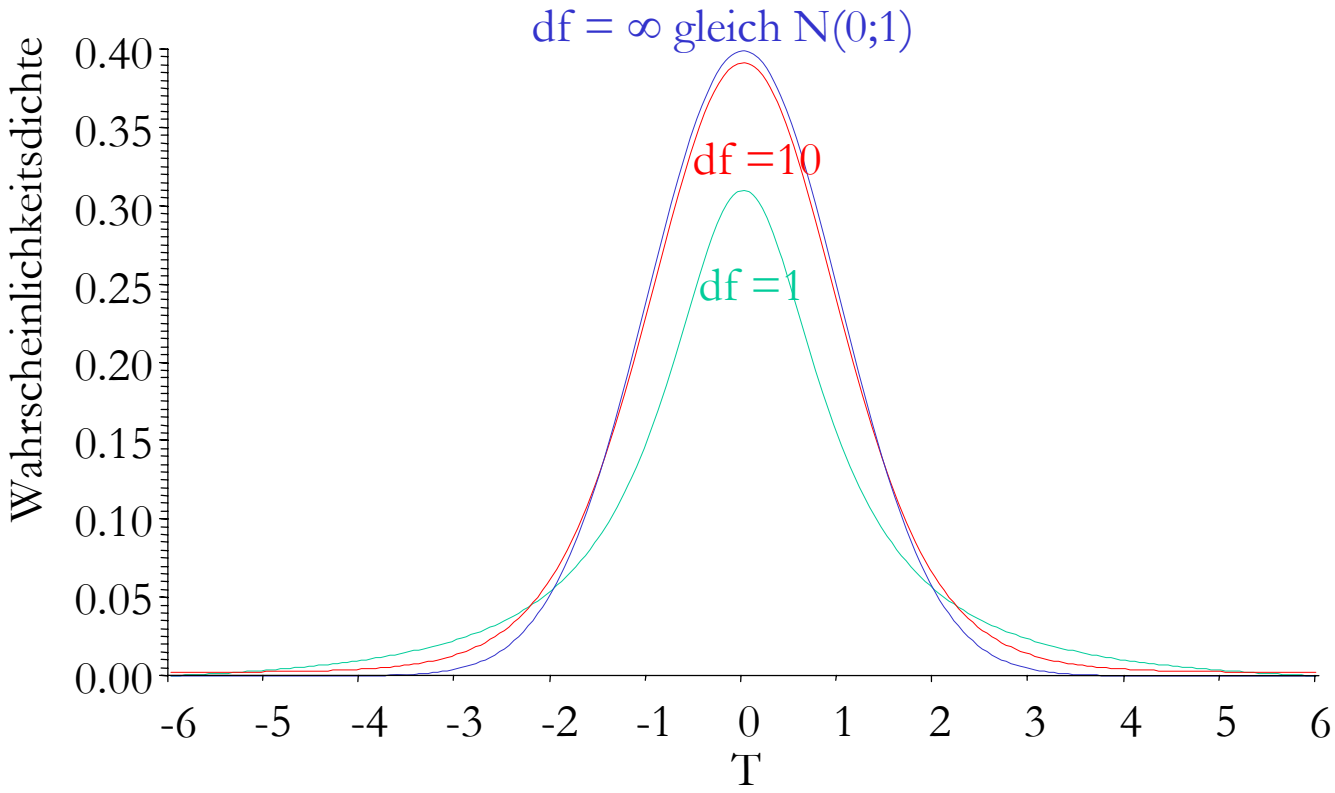
nur dann exakt standardnormalverteilt, wenn im Nenner des Quotienten bei der Berechnung des Standardfehlers des Mittelwert die tatsächliche Populationsvarianz von X eingesetzt wird.

Ist diese unbekannt und wird stattdessen die aus den Stichprobendaten geschätzte erwartungstreue Populationsvarianz eingesetzt, ergibt sich wie bei nicht normalverteilten Populationen nur eine asymptotisch gültige Standardnormalverteilung.

Es kann jedoch gezeigt werden, dass die Z-Transformation des Stichprobenmittelwerts auf der Basis der geschätzten Populationsvarianz dann zu einer T-Verteilung mit df = n–1 Freiheitsgraden führt, wobei df der Parameter einer T-Verteilung ist:

$$f(X_i) = N(\mu_X; \sigma_X^2) \Rightarrow f\left(\frac{\bar{x} - \mu_X}{\sqrt{\frac{1}{n \cdot (n-1)} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}}\right) = t_{df=n-1}$$

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 Nullpunkt entfernt sind als die entsprechenden Quantilwerte der Standardnormalverteilung. Mit steigender Zahl von Freiheitsgraden nähert sich die T-Verteilung asymptotisch der Standardnormalverteilung an, so dass $t_{df=\infty}$ = N(0;1):

Quantile der T-Verteilung

In Tabellen werden Quantilwerte von T-Verteilungen für wichtige Quantilwahrscheinlichkeiten und unterschiedliche Freiheitsgrade tabelliert:

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.

Konfidenzintervalle für Populationsmittelwerte

Die T-Verteilung wird für die Berechnung des (1−α)-Konfidenzintervallen für Mittelwerte aus normalverteilten Populationen herangezogen. Bei Irrtumswahrscheinlichkeit α und unbekannter Standardabweichung berechnet sich das (1−α)-Konfidenzintervall nach:

$$\begin{aligned} \text{c.i.}(\mu_X) &= \bar{x} \pm \hat{\sigma}\left(\bar{x}\right) \cdot t_{1-\alpha/2,df=n-1;} \\ &= \bar{x} \pm \frac{\hat{\sigma}_X}{\sqrt{n}} \cdot t_{1-\alpha/2,df=n-1} \\ &= \bar{x} \pm \frac{s_X}{\sqrt{n-1}} \cdot t_{1-\alpha/2,df=n-1} \end{aligned}$$

Quantile von T	
df	97.5%
120	1.980
∞	1.960

In der Stichprobe des Allbus 1996 beträgt der Mittelwert der Befragten 46.117 Jahren, die Stichprobenvarianz ist 281.112 und die Fallzahl beträgt 3510 Personen.
Gesucht ist das 95%-Konfidenintervall für den Populationsmittelwert:

$$\text{c.i.}(\mu_X) = \bar{x} \pm \sqrt{\frac{s_X^2}{n-1}} \cdot t_{0.975,df=3509} = 46.117 \pm \sqrt{\frac{281.112}{3509}} \cdot 1.96 = 46.117 \pm 0.555$$

Da nur Personen ab 18 Jahren befragt wurden ist zu schließen, dass 1996 das durchschnittliche Alter von volljährigen Personen in Deuschland vermutlich zwischen 45.562 und 46.672 Jahren lag.

Asymptotische Konfidenzintervalle für Mittelwerte bei beliebiger Verteilung

Wenn die Variable X in der Gundgesamtheit nicht normalverteilt ist, kann anstelle eines exakten Konfidenzintervall das bereits vorgestellte asymptotisches Konfidenzintervall berechnet werden.

Da Konfidenzintervalle, die über die T-Verteilung berechnet werden, länger sind als Konfidenzintervalle mit gleicher Irrtumswahrscheinlichkeit, die auf der Standardnormalverteilung beruhen, wird üblicherweise 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 überdecken. Dieses vorsichtigere Vorgehen wird als *konservatives Schätzen* bezeichnet.

Quantile von T	
df	97.5%
120	1.980
∞	1.960

Die Zahl der Freiheitsgrade ist gleich der Fallzahl minus eins.
Bei über 120 Freiheitsgraden unterscheidet sich die T-Verteilung aber kaum noch von der Standardnormalverteilung.
Bei Fallzahlen über n=121 kann daher stets die Standardnormalverteilung zur Berechnung der Quantilwerte herangezogen werden.

Statistikprogramme berechnen auch T-Verteilungen bei sehr vielen Freiheitsgraden.

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, dann benötigt man eine Fallzahl von:

$$n = \frac{\left(z_{1-\alpha/2}\right)^2 \cdot \pi_1 \cdot (1-\pi_1)}{\varepsilon^2} = \frac{1.96^2 \cdot 0.5^2}{0.03^2} = 1067.111 \approx 1068$$