Lineare und logistische Regression

LE2 - Konfidenzintervalle

Dr. Lucia Kleint

Fachhochschule
Nordwestschweiz

Herbstsemester 2020

Inhaltsverzeichnis

Ziel der Kompetenz

Die Statistik ist die Grundlage zur zuverlässigen Auswertung von Daten. Allerdings sind Daten nie perfekt und somit muss ein geeignetes Modell gefunden werden. Das Hauptziel dieser Kompetenz ist zu verstehen, wie Daten in der Data Science durch lineare und logistische Regression praktisch ausgewertet werden können.

Jedes Teil-Skript enthält ein Lernergebnis, inklusive Verweisen auf detailliertere Quellen wie Bücher. Für die Prüfung werden die Themen dieses Skripts, aber so detailliert wie in den Verweisen vorausgesetzt.

last update: 18. August 2020

Verweise

Folgende Bücher/Unterlagen werden für das genauere Verständnis empfohlen:

- L. Fahrmeir et al., Statistik, Der Weg zur Datenanalyse, 8. Auflage, Springer (pdf via FHNW Netzwerk).
- wst Skript, Prof. Dr. M. Steiner-Curtis, 2015 (pdf via FHNW AD)
- `https://ethz.ch/content/dam/ethz/special-interest/math/statistics/sfs/Education/Advanced%20Studies%20in%20Applied%20Statistics/course-material-1921/Regression/reg-script-full.pdf`
- Video D. Jung `https://www.youtube.com/watch?v=HwGhcyDnmG4`
- Video `https://www.youtube.com/watch?v=DdwTa28W4Os`
- Video `https://www.youtube.com/watch?v=jQsYd5-TTmw`

Section 1

Verteilungen

Dichtefunktion und Verteilungsfunktion

Histogramme haben den Nachteil, dass sie nicht stetig sind und vom Binning der Daten abhängen können. Dichtekurven (auch: Dichtefunktion, Wahrscheinlichkeitsdichte, Dichte, PDF für probability density function) sind stetige Funktionen $f(x)$, wenn $f(x) \geq 0 \; \forall \; x \in \mathbb{R}$ und die von $f(x)$ überdeckte Gesamtfläche gleich 1 ist $(\int_{-\infty}^{\infty} f(x)dx = 1)$.

Durch die Integration einer Dichtefunktion kommt man auf die entsprechende Verteilungsfunktion $F(x)$, welche zur Berechnung von Wahrscheinlichkeiten verwendet wird: $F(x) = \int_{-\infty}^{x} f(u)du$. Im Englischen wird sie cumulative distribution function (CDF) genannt.

Die folgenden Folien stellen einige wichtige Dichte- und Verteilungsfunktionen vor und zeigen Beispiele zur Berechnung von Wahrscheinlichkeiten.

Normalverteilung

Die wahrscheinlich wichtigste Dichtekurve ist die Normalverteilung. Normalverteilungen sind symmetrisch und glockenförmig.

Satz (Normalverteilung)

Die Normalverteilung ist definiert durch die Dichtefunktion

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), x \in \mathbb{R}$$

μ ist der Mittelwert und σ die Standardabweichung von $f(x|\mu,\sigma)$.

Die Normalverteilung mit $\mu = 0$ und $\sigma = 1$ heisst Standardnormalverteilung.

Standardnormalverteilung

Durch eine lineare Transformation kann man aus allen normalverteilten Zufallsvariablen X mit einem Mittelwert μ und einer Varianz σ^2 die Standardnormalverteilung erhalten, für deren Zufallsvariablen Z gilt

$$Z = \frac{X - \mu}{\sigma}$$

Deren Mittelwert ist Null und die Varianz(Z) $= \frac{1}{\sigma^2}$Varianz(X) $= 1$.

Satz (Standardnormalverteilung)

Die Standardnormalverteilung hat die Dichtefunktion

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), z \in \mathbb{R}$$

Normalverteilung

Viele Grössen in der Natur sind normalverteilt, z.B. die Körpergrösse oder die Intelligenz. Auch z.B. Punktzahlen in Prüfungen oder die Abweichungen von Schraubengrössen vom Sollwert bei ihrer Produktion folgen oft der Normalverteilung. Wenn man nun Intervalle um den Mittelwert berechnet, kann man aussagen, wie viele Datenpunkte in welchem Intervall liegen.

Normalverteilung

Eine Standardabweichung vom Mittelwert $(\mu \pm \sigma)$ beinhaltet 68% der Daten. 95% der Daten liegen im Intervall $(\mu \pm 2\sigma)$ und 99.7 % der Daten liegen im Intervall $(\mu \pm 3\sigma)$. Dies ist die 68-95-99,7-Regel.

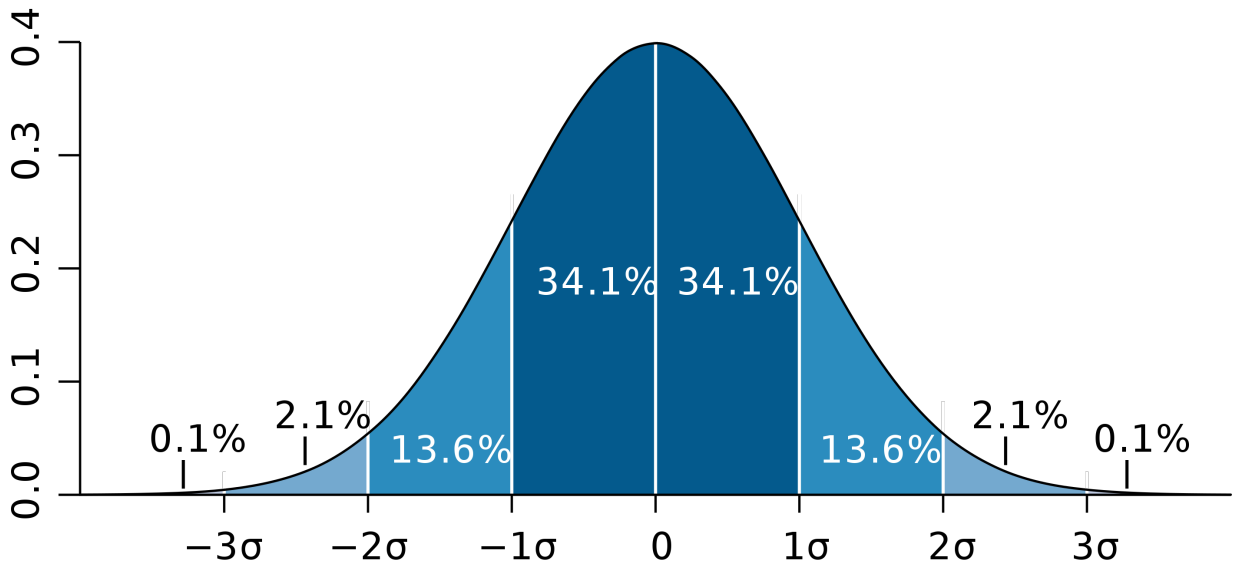

Abbildung: Quelle: Wikipedia

Werte ausserhalb von 2-3 σ werden oft als Ausreisser gesehen, da sie unwahrscheinlich sind. Oft können sie durch systematische Fehler (z.B. falsch kalibrierte Maschine) erklärt werden und nicht durch zufallsbedingte Streuung.

Verteilungsfunktion

Wenn man z.B. das Beispiel der Normalverteilung nimmt, kann man sich fragen, wie viele Schüler mit einem IQ unter 70 erwartet werden, für welche man spezielle Fördermassnahmen vorsehen möchte.
Allgemein gibt die Verteilungsfunktion die Fläche unter der Glockenkurve bis zum Wert x an, d.h., es wird das bestimmte Integral von $-\infty$ bis x berechnet. Dazu transformiert man zuerst in die Standardnormalverteilung, wenn nötig.

Satz (Verteilungsfunktion)

Die Verteilungsfunktion der Standardnormalverteilung ist

$$F(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$$

Beispiel IQ

Beispiel

Wie viele Schüler mit einem IQ unter 70 werden nun erwartet?

Der IQ ist so skaliert, dass die Werte in der Population normalverteilt sind, mit einem Erwartungswert $\mu = 100$ und einer Standardabweichung $\sigma = 15$. Dies heisst, dass 68% der Bevölkerung einen IQ zwischen 85 und 115 haben.

Für unser Beispiel soll man $P(Z < 70)$ bestimmen. Das heisst, man bestimmt die Fläche von $-\infty$ bis 70 in der Dichtefunktion der IQ Verteilung. Damit man nicht jedes Mal das Integral der Normalverteilung berechnen muss, gibt es dafür Tabellen (z.B. `https://de.wikipedia.org/wiki/Standardnormalverteilungstabelle`). Wir bestimmen also $\Phi(\frac{70-100}{15}) = 1 - \Phi(2) = 1 - 0.977 = 0.023$. Somit haben etwa 2.3% der Schüler einen IQ unter 70 (und aus Symmetriegründen auch über 130).

Zentraler Grenzwertsatz

Satz (Zentraler Grenzwertsatz)

Die Summe von vielen (genauer: unendlich vielen) unabhängigen Zufallsgrössen ist näherungsweise normalverteilt.

Dieser Satz ist der Grund für die zentrale Bedeutung der Normalverteilung in der Wahrscheinlichkeitsrechnung und in der Statistik. Deswegen können in den meisten praktischen Anwendungsfällen die Beobachtungsfehler bei Messvorgängen wenigstens näherungsweise als normalverteilt angesehen werden.

Zentraler Grenzwertsatz: Beispiel

Beispiel

Bei der Lottoziehung gibt es 49 Kugeln. Das heisst, man wird eine Zahl zwischen a=1 und b=49 bei jeder Ziehung erhalten und da die Kugeln identisch sind, ist jede Zahl gleich wahrscheinlich. Der Mittelwert ist $\mu = \frac{a+b}{2} = 25$. Wie bestimmt man aber den Mittelwert, falls man weder weiss, wie viele Kugeln es gibt, noch was deren Mittelwert ist?

Zur Lösung braucht man den zentralen Grenzwertsatz. Wir ziehen jeweils 7 Kugeln (man könnte auch weniger oder mehr nehmen) als Stichprobe. Historische Daten zeigen die folgenden Werte:

Tabelle: 5838 Ziehungen der Lottozahlen 6 aus 49 (mit Zusatzzahl)!

Datum	Lottozahlen						Zusatzzahl	Mittelwert
17.06.1956	5	19	25	38	41	46	24	28.29
24.06.1956	14	16	18	32	37	43	21	25.86
⋮				⋮				⋮
1.05.2013	9	12	15	16	22	28	5	15.28

Abbildung: Quelle, auch des Beispiels: M. Oestreich, O. Romberg, Keine Panik vor Statistik, Springer, 2018

Zentraler Grenzwertsatz: Beispiel

Man sieht, dass der Mittelwert der Stichprobe schwankt. Wenn man also wenige Stichproben hat, kann man nicht viel aussagen. Wenn man allerdings die Stichprobennahme oft wiederholt und das Ergebnis in einem Plot zeichnet, dann erhält man die Darstellung rechts.

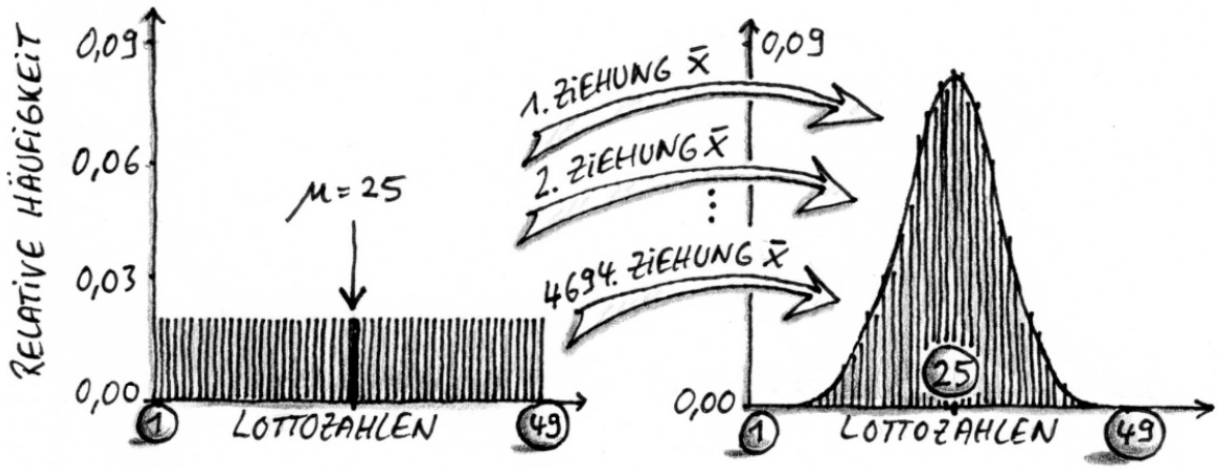

Abbildung 10.1 Die Mittelwerte von 7 Lottozahlen (6 + Zusatzzahl) sind annähernd normalverteilt mit Mittelwert μ und Varianz $\sigma^2_{\bar{x}}$

Abbildung: Quelle, auch des Beispiels: M. Oestreich, O. Romberg, Keine Panik vor Statistik, Springer, 2018

Die Mittelwerte der Stichproben sind normalverteilt und zwar rund um 25, den "wahren" Mittelwert. Dies ist genau die Aussage des zentralen Grenzwertsatzes.

Exponentialverteilung

Die Exponentialverteilung ist eine stetige Wahrscheinlichkeitsverteilung, die durch eine Exponentialfunktion gegeben ist. Sie wird vor allem für die Modellierung der Dauer von zufälligen Zeitintervallen benützt. Beispiele sind die Zeit zwischen zwei ankommenden Kunden, die Zeit zwischen zwei Anrufen, Lebensdauer von Atomen bei radioaktivem Zerfall etc.

Satz (Exponentialverteilung)

Eine stetige Zufallsvariable X is exponentialverteilt mit dem positiven reellen inversen Skalenparameter $\lambda \in \mathbb{R}_{>0}$, wenn sie die Dichtefunktion

$$f_\lambda(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

besitzt.

Exponentialverteilung

Satz (Exponentialverteilung)

Die Verteilungsfunktion der Exponentialverteilung ist

$$F(x) = \int_0^x f_\lambda(t)dt = \begin{cases} 1 - \mathrm{e}^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Die Verteilungsfunktion erlaubt die Berechnung der Wahrscheinlichkeit des Auftretens des nächsten Ereignisses im Intervall von 0 bis x. Der Erwartungswert einer exponentialverteilten Zufallsvariablen ist $1/\lambda$. Er wird auch mittlere Lebensdauer oder mittlere Reichweite genannt. Der Median der Exponentialverteilung ist $\frac{\ln(2)}{\lambda}$.

Exponentialverteilung: Beispiel

Beispiel

Die zufällige Wartezeit von Kunden am Postschalter sei exponentialverteilt mit einem Erwartungswert von 10 Minuten. Wie hoch ist die Wahrscheinlichkeit, dass der Kunde mindestens 15 Minuten warten muss?

Sei X die zufällige Wartezeit, wobei $X \sim \exp(-\lambda)$ mit $\lambda = \frac{1}{10}$. Wir bestimmen $P(X > 15)$.

$P(X > 15) = e^{-15\lambda} = e^{-1.5} \approx 0.22$. 22% der Kunden müssen somit länger als 15 Minuten warten.

Student-t-Verteilung (auch: t-Verteilung)

Die Student-t-Verteilung wurde 1908 von W. Sealy Gosset entwickelt und nach seinem Pseudonym ("Student") benannt. Sie wird benützt, wenn die Anzahl der Messwerte klein ist und/oder wenn die Standardabweichung σ der Grundgesamtheit unbekannt ist. Unter reellen Bedingungen ist dies sehr oft der Fall und das zentrale Grenzwerttheorem kann nicht benützt werden.

Satz (t-Verteilung)

Die t-Verteilung mit n=N-1 Freiheitsgraden hat die Dichtefunktion

$$f_n(t) = c_n \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, t \in \mathbb{R}$$

mit der Konstante $c_n = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}$*, die nur von der Anzahl Freiheitsgrade abhängt.* $\Gamma(x)$ *ist die Gammafunktion. Die Freiheitsgrade beziehen sich auf die Grösse der Stichprobe* N*.*

Student-t-Verteilung (auch: t-Verteilung)

Generell müssen drei Kriterien erfüllt sein, damit die t-Verteilung verwendet werden kann:

- Die Standardabweichung und somit auch die Varianz der Grundgesamtheit sind nicht bekannt
- Die Stichprobe muss zufällig entnommen sein
- Die Grundgesamtheit der Daten, aus welcher die Stichprobe entnommen wurde, muss normalverteilt oder etwa normalverteilt sein oder die Stichprobe muss mindestens 30 Messwerte beinhalten.

In gewissen Fällen reichen auch Stichproben von weniger als 30 Messwerten. Für $n \to \infty$ konvergiert die Dichtekurve der t-Verteilung gegen die Dichtekurve der Standardnormalverteilung. Bei kleineren Freiheitsgraden hat die t-Verteilung breitere Enden, und kann somit für Verteilungen mit einem grösseren Anteil an extremen Werten verwendet werden.

Student-t-Verteilung (auch: t-Verteilung)

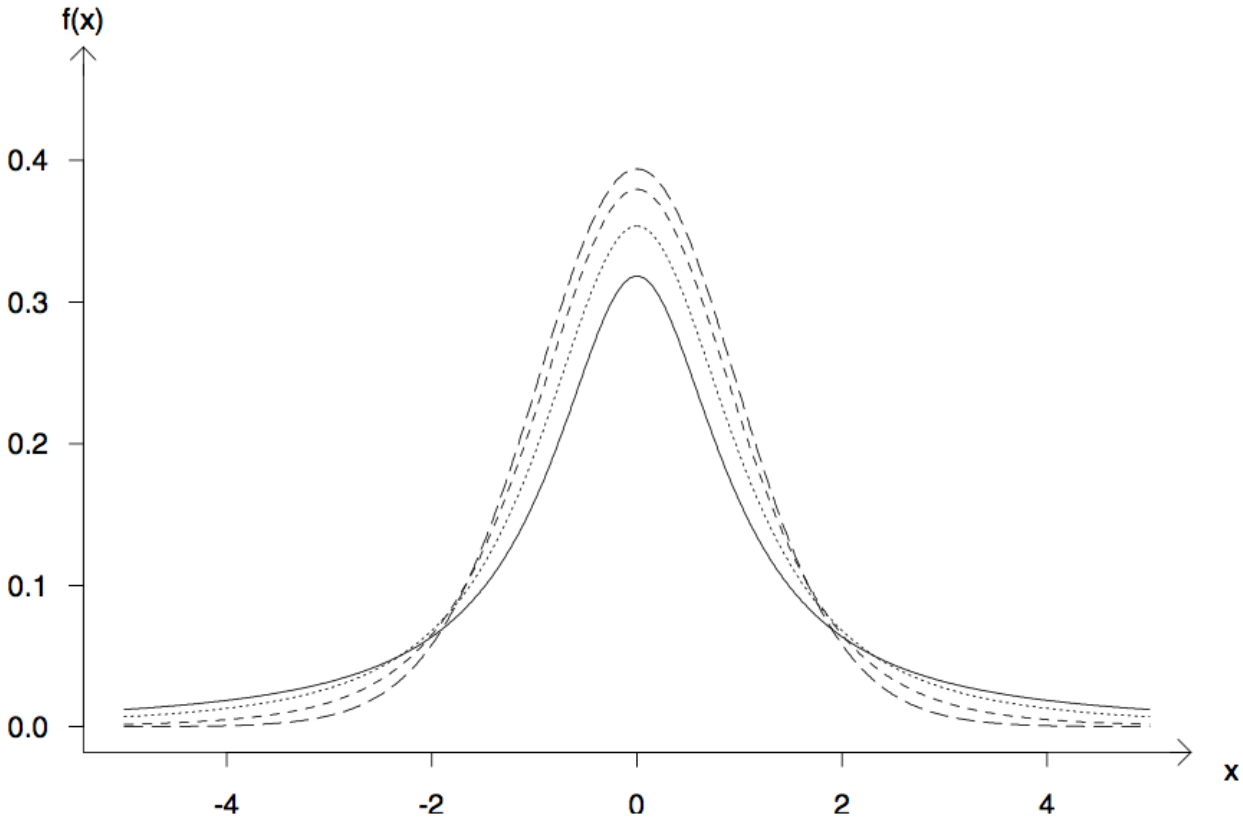

Abbildung 6.19: Dichten von t-Verteilungen für $n = 1$ (—) $n = 2$ $(\cdots)$, $n = 5$ (- - -) und $n = 20$ $(--)$ Freiheitsgrade

Abbildung: Verschiedene t-Verteilungen Quelle: `Fahrmeir,Kap.6`

Student-t-Verteilung (auch: t-Verteilung)

Die t-Verteilung ist symmetrisch um 0, ihr Median und Mittelwert sind 0. Der Erwartungswert und die Varianz einer t-verteilten Zufallsgrösse T mit n Freiheitsgraden ist

$$E(T) = 0 \quad \text{für } n > 1 \quad \text{und} \quad Var(T) = \frac{n}{n-2} \quad \text{für } n > 2.$$

t-Verteilung

Wir haben vorher gesehen, dass man aus einer Normalverteilung in die Standardnormalverteilung transformieren kann, wenn man den Mittelwert und seinen Standardfehler kennt. Nun haben wir aber das Problem, dass der Standardfehler der Grundgesamtheit unbekannt ist und wir nur den Standardfehler der Stichprobe schätzen (oder messen) können.
Wir erhalten dann den sogenannten t-Wert:

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}},$$

wobei $\bar{x}$ der Mittelwert der Stichprobe ist, μ der Mittelwert der Grundgesamtheit, $\sigma_{\bar{x}}$ der Standardfehler des Stichprobenmittelwerts, σ der Standardfehler der Stichprobe und N die Anzahl der Stichproben.
Dies ist der t-Wert einer Stichprobe. Betrachtet man nicht nur einen t-Wert, sondern die t-Werte der gesamten Stichprobenverteilung, erhält man die t-Verteilung.

Beispiel: t-Verteilung

Beispiel

Eine Firma fertigt Glühbirnen, die angeblich durchschnittlich 300 Tage leuchten. Eine Studie testet dies mit 15 zufällig ausgewählten Glühbirnen. Die Glühbirnen in der Studie leuchten durchschnittlich 290 Tage mit einer Standardabweichung von 50 Tagen. Stimmt die Angabe der Firma, oder besser gesagt, was ist die Wahrscheinlichkeit, dass 15 zufällig ausgewählte Glühbirnen eine durchschnittliche Lebensdauer von nicht mehr als 290 Tagen haben?[a]

[a] Das Beispiel stammt von `https://stattrek.com/probability-distributions/t-distribution.aspx`

Wir berechnen den t-Wert $t = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{N}}} = \frac{(290-300)}{50/\sqrt{15}} = -10/12.9 = -0.77$. Wir wissen, dass die Anzahl der Freiheitsgrade 15-1=14 ist. Mit diesen zwei Angaben berechnet man die Wahrscheinlichkeit aus der T-Verteilung (z.B. Computer oder `https://stattrek.com/online-calculator/t-distribution.aspx`) und erhält 0.226. Dies heisst, dass die Wahrscheinlichkeit 22.6% beträgt, dass die Lebensdauer $\leq$290 Tage ist und somit hat die Firma wahrscheinlich bei ihrer Aussage von 300 Tagen übertrieben.

Section 2

Konfidenzintervalle

Konfidenzintervalle

Aus einer Stichprobe kann man z.B. deren Mittelwert oder die Varianz berechnen, aber man weiss noch nicht, wie genau diese Werte sind, oder wie sie von der Stichprobengrösse abhängen. Aus der Stichprobe kann man Intervalle berechnen, in welchen der wahre Wert mit grosser Wahrscheinlichkeit erwartet werden kann. Diesen Prozess nennt man Intervallschätzungen und die Intervalle heissen Vertrauens- oder Konfidenzintervalle.

Das Konfidenzintervall gibt das Intervall und die Wahrscheinlichkeit an, in welchem der Parameter (z.B. der Mittelwert) bei unendlicher Anzahl Wiederholungen eines Zufallsexperiments liegt. Oft verwendet man ein Konfidenzniveau von 95%. Das heisst, wenn man das Zufallsexperiment unendlich mal wiederholen würde, würde ein 95%-Konfidenzintervall in 95% der Fälle den wahren Wert des Parameters umschliessen und in 5% der Fälle nicht.

Konfidenzintervalle mit unbekannter Varianz

Wir nehmen eine Stichprobe $x_1, ..., x_N$ vom Umfang N aus einer normalverteilten Grundgesamtheit. Wir möchten das Konfidenzintervall des unbekannten Erwartungswertes μ der Grundgesamtheit bestimmen.

Zuerst berechnen wir den Mittelwert $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ und die Varianz $\sigma_s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$ aus der Stichprobe.

Der Mittelwert der Stichprobe folgt einer Normalverteilung mit Varianz $\frac{\sigma_s^2}{N}$. Man gibt eine Vertrauenswahrscheinlichkeit γ vor, mit welcher der wahre Erwartungswert innerhalb des Konfidenzintervalls liegt (z.B. 0.90, 0.95, 0.99). Die Irrtumswahrscheinlichkeit α ist $1 - \gamma$ und beträgt üblicherweise 0.10, 0.05, oder 0.01.

Konfidenzintervalle

Wir müssen nun die Grösse $t_{n,1-\alpha/2}$ der t-Verteilung mit $n = N - 1$ Freiheitsgraden berechnen. Der Faktor $1/2$ kommt daher, weil die Verteilung symmetrisch ist, und man somit das Intervall $\gamma/2$ in beide Richtungen zulässt.
Mathematisch geschrieben suchen wir alle μ, so dass

$$P(|T| < t_{n,1-\alpha/2}) = P\left(\left|\frac{\bar{x} - \mu}{\sigma_s}\sqrt{N}\right| < t_{n,1-\alpha/2}\right) = 1 - \alpha$$

gilt. Wenn man die Ungleichung umformt, erhält man das Konfidenzintervall für μ durch

$$\bar{x} - \frac{t_{n,1-\alpha/2}}{\sqrt{N}}\sigma_s \leq \mu \leq \bar{x} + \frac{t_{n,1-\alpha/2}}{\sqrt{N}}\sigma_s$$

Daraus folgt, dass je grösser die Stichprobe (N), desto kleiner wird das Konfidenzintervall.

Konfidenzintervalle mit bekannter Varianz σ^2

Falls die Varianz der Grundgesamtheit bekannt ist, nimmt man statt der t-Verteilung die Normalverteilung, siehe z.B. Kapitel 9.4.1. in Fahrmeir's Buch.
Das Konfidenzintervall ist dann

$$\bar{x} - \frac{z_{1-\alpha/2}}{\sqrt{N}}\sigma \leq \mu \leq \bar{x} + \frac{z_{1-\alpha/2}}{\sqrt{N}}\sigma$$

wobei $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Standardverteilung ist.

Konfidenzintervalle: Beispiel

In einer Schraubenfabrik nehmen wir eine Stichprobe von 10 Schrauben und messen eine mittlere Länge von 5 cm und eine Standardabweichung $\sigma_s = 0.2$ cm. In welchem Konfidenzintervall liegt der wahre aber unbekannte Erwartungswert μ der (normalverteilen) Gesamtheit aller Schrauben mit einer Vertrauenswahrscheinlichkeit von $\gamma = 0.95$?

Wir berechnen die Grösse $t_{9,1-0.025}$ der Student t-Verteilung (Tabelle, oder online) und erhalten 2.262. Somit ergibt sich das Konfidenzintervall

$$5 - \frac{2.262}{\sqrt{10}} 0.2 \leq \mu \leq 5 + \frac{2.262}{\sqrt{10}} 0.2,$$

also $4.86 \leq \mu \leq 5.14$ mit 95% Wahrscheinlichkeit[1].

1) das Beispiel stammt aus dem wst Skript von Prof. Steiner-Curtis, 2015

Einstichproben-t-Test

Problemstellung: Wir haben eine Stichprobe von N Schrauben mit Längen $x_1, ..., x_N$, aber haben vergessen aufzuschreiben, von welcher Maschine wir die Stichprobe genommen haben. Alle Maschinen stellen verschieden lange Schrauben her und deren Erwartungswerte μ sind bekannt. Wie kann man nun bestimmen, ob die Stichprobe zu einer bestimmten Grundgesamtheit G gehört?

Eine Methode ist der sogenannte Einstichproben-t-Test oder auch Student-t-Test. Die Voraussetzungen sind: 1) Der Erwartungswert μ von G ist bekannt. 2) Die Varianz σ^2 von G ist unbekannt. 3) Es sind zufällig N Stichproben aus einer normalverteilten Grundgesamtheit gewählt worden.

t-Test

Da σ^2 unbekannt ist, berechnet man den t-Wert mittels $t = \frac{\bar{x}-\mu}{\sigma_s}\sqrt{N}$. Der t-Wert unterscheidet sich für verschiedene Stichproben und somit gibt es eine Wahrscheinlichkeitsverteilung von t-Werten. Ist die Grundmenge normalverteilt, so sind die t-Werte t-verteilt mit N-1 Freiheitsgraden.

Der t-Test besagt nun, dass wenn $|t| < t_{n,1-\alpha/2}$, dann sind die Abweichungen vom idealen Wert t=0 zufällig. Die Stichprobe stammt somit mit einer Irrtumswahrscheinlichkeit α aus G mit μ. Wenn allerdings $|t| \geq t_{n,1-\alpha/2}$, dann stammt die Stichprobe aus einer anderen Grundgesamtheit.

Der Student-t-Test ist gegenüber Abweichungen der Voraussetzung, dass G normalverteilt sein muss, ziemlich unempfindlich. Deswegen nennt man ihn auch einen robusten Test.

Zweistichproben-t-Test

Problemstellung: Klasse A hat 25 Studenten und in einer Prüfung einen Notendurchschnitt $\bar{x}_A = 3.9$ erreicht. Klasse B, die vom gleichen Dozenten auf gleiche Art unterrichtet wird, hat 28 Studenten und bei der gleichen Prüfung einen Notendurchschnitt $\bar{x}_A = 4.2$ erreicht. Die Standardabweichungen sind identisch $\sigma_A = \sigma_B = 1$. Ist die B-Klasse signifikant besser als die A-Klasse?

Eine Methode, Mittelwerte zu vergleichen ist der Zweistichproben-t-Test. Die Voraussetzungen sind: 1) Die Grundgesamtheiten G_1 und G_2 sind normalverteilt mit unbekannten Erwartungswerten μ_1 und μ_2, und die **gleichen** Varianzen $\sigma_1^2 = \sigma_2^2 = \sigma^2$. σ^2 muss nicht bekannt sein. 2) Es sind zufällig 2 Stichproben $x_1, ..., x_N$ und $y_1, ..., y_N$ aus G_1 und G_2 gewählt worden.

Zweistichproben-t-Test

Wir berechnen die geschätzten Varianzen

$$\sigma_{s1}^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (x_i - \bar{x})^2 \text{ und } \sigma_{s2}^2 = \frac{1}{N_2 - 1} \sum_{i=1}^{N_2} (y_i - \bar{y})^2.$$

und daraus das gewogene Mittel der Varianzen

$$s^2 = \frac{(N_1 - 1)\sigma_{s1}^2 + (N_2 - 1)\sigma_{s2}^2}{N_1 + N_2 - 2}$$

Daraus berechnen wir die Testgrösse

$$t = \frac{\bar{x} - \bar{y}}{s} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}$$

welche unter den Voraussetzungen einer t-Verteilung mit $n = N_1 + N_2 - 2$ Freiheitsgraden genügt. Somit kann man analog zum Einstichprobentest α vorgeben und $t_{n,1-\alpha/2}$ bestimmen. Der t-Test besagt nun, dass wenn $|t| < t_{n,1-\alpha/2}$, dann sind die Unterschiede zwischen $\bar{x}$ und $\bar{y}$ zufällig.

Zweistichproben-t-Test

Wenn die Varianzen nicht gleich sind, also $\sigma_1^2 \neq \sigma_2^2$, hat man das sogenannte Behrens-Fisher-Problem. Der Unterschied zum vorherigen Test ist die Berechnung der gewogenen Varianz:

$$s^2 = \frac{\sigma_{s1}^2}{N_1} + \frac{\sigma_{s2}^2}{N_2}$$

Daraus berechnet man die Testgrösse $t = \frac{\bar{x}-\bar{y}}{s}$, welche einer Student-t-Verteilung gehorcht. Allerdings ist die Berechnung der Freiheitsgrade in diesem Fall komplizierter und wir verweisen dazu auf die Literatur.

Konfidenzintervalle in der linearen Regression

Wenn man mittels $f(X) = \alpha + \beta X$ eine lineare Einfachregression fittet, sollte man sich fragen, wie gut die Koeffizienten α und β stimmen und in welchem Bereich sie mit grosser Wahrscheinlichkeit liegen.

Da α und β geschätzt sind, werden sie oft auch als $\hat{\alpha}$ und $\hat{\beta}$ geschrieben. Ihre Konfidenzintervalle sind (siehe Fahrmeir, Kap. 12):

$$\hat{\alpha} \pm \hat{\sigma}_{\hat{\alpha}} t_{(1-\alpha/2)}(n-2)$$
$$\hat{\beta} \pm \hat{\sigma}_{\hat{\beta}} t_{(1-\alpha/2)}(n-2)$$

wobei $\hat{\sigma}$ die jeweilige Standardabweichung ist und $t_{(1-\alpha/2)}(n-2)$ die Student-t Verteilung mit dem gewünschten Konfidenzniveau (welches üblicherweise mittels α geschrieben wird, aber das hat nichts mit dem Koeffizienten der kleinsten Quadrate zu tun) und $(n-2)$ Freiheitsgraden. $(n-2)$ kommt daher, da die Freiheitsgrade von n auf $n-2$ abnehmen, da zwei Parameter $(\hat{\alpha}, \hat{\beta})$ geschätzt werden. Für $n \geq 30$ sollte man die $t(n-2)$ Verteilung durch die Normalverteilung ersetzen.

Ein Zahlenbeispiel findet man in Beispiel 12.1 in Fahrmeir's Buch.