

Lineare Regression

JiTT Ziele

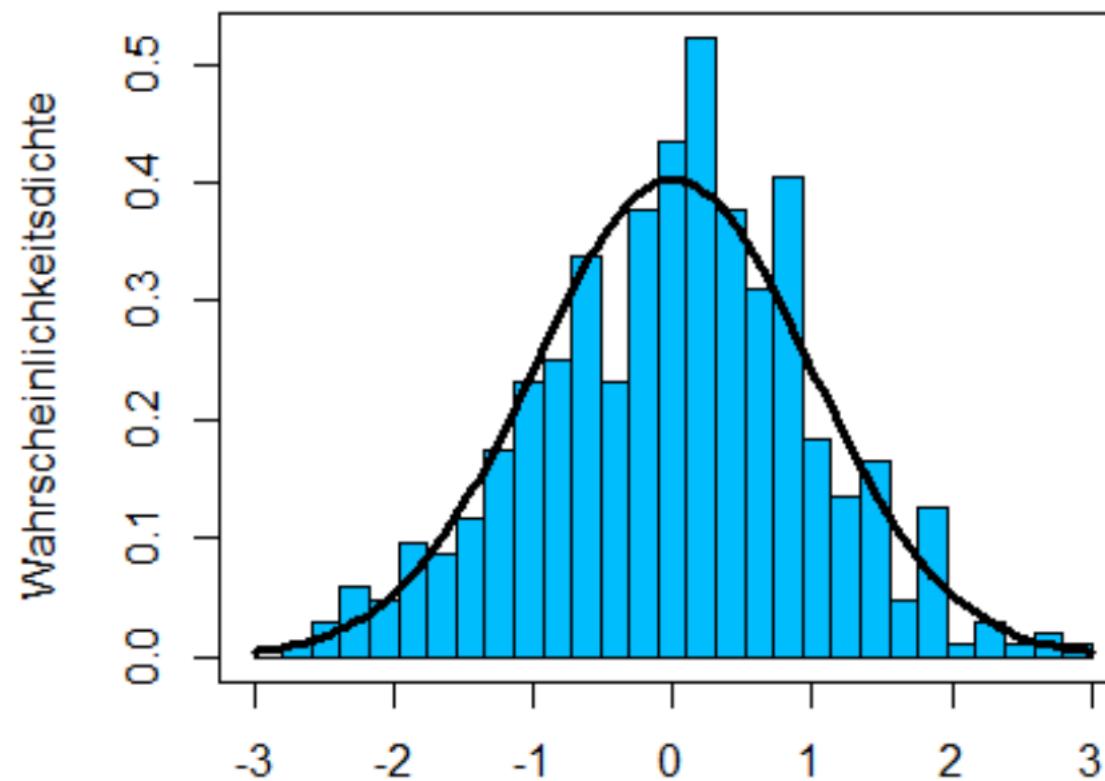
- Ihr kennt die Normalverteilung und die t-Verteilung und deren Unterschiede.
- Ihr kennt den Einstichproben- und Zweistichproben-t-Test
- Ihr könnt Konfidenzintervalle berechnen, sowohl für statistische Fragestellungen, wie auch für Koeffizienten bei der Regression.

Normalverteilung

Histogramme hängen vom Binning ab.

Verteilungsfunktionen eignen sich für Rechnungen.

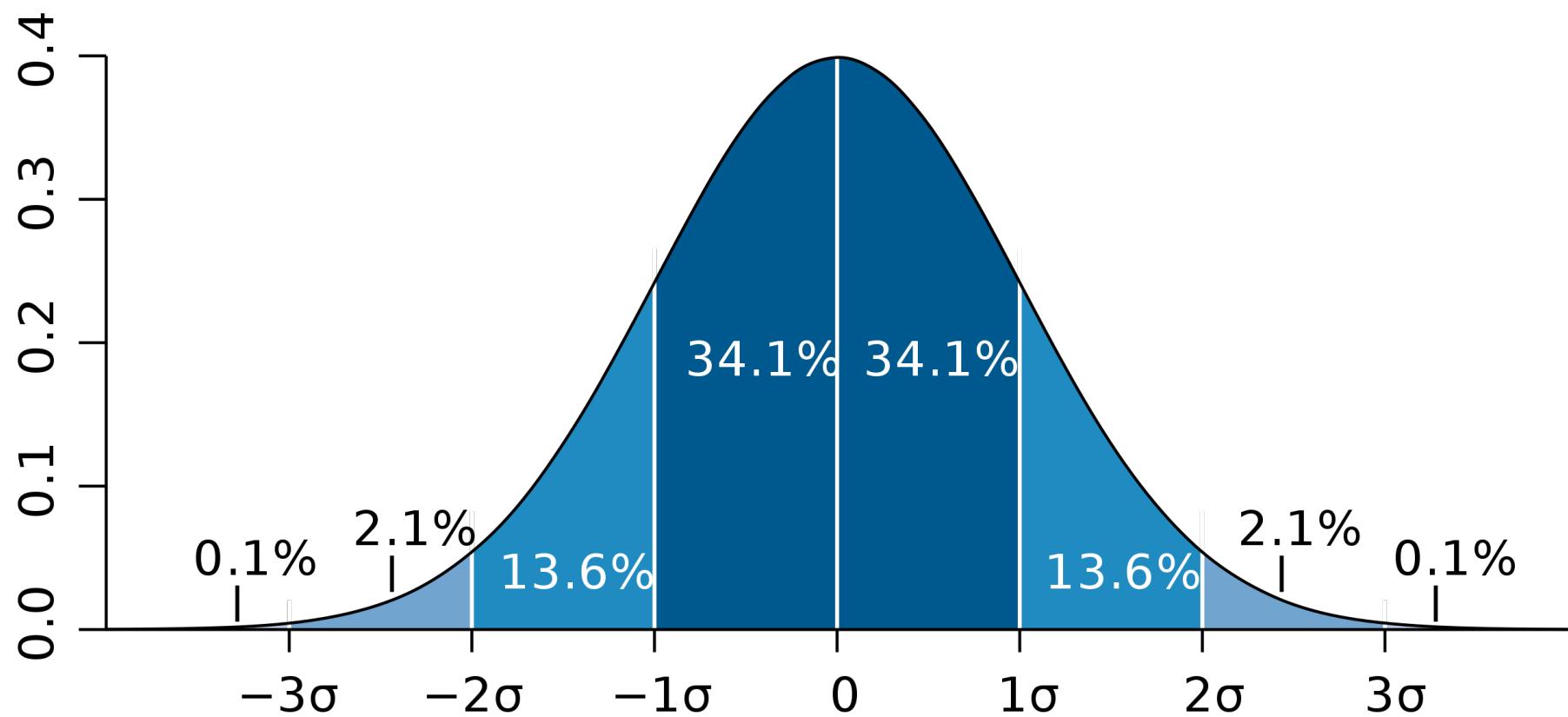
Beispiel Histogramm



Normalverteilung

z.B. IQ, Körpergrösse, Produktionsabweichungen, ...

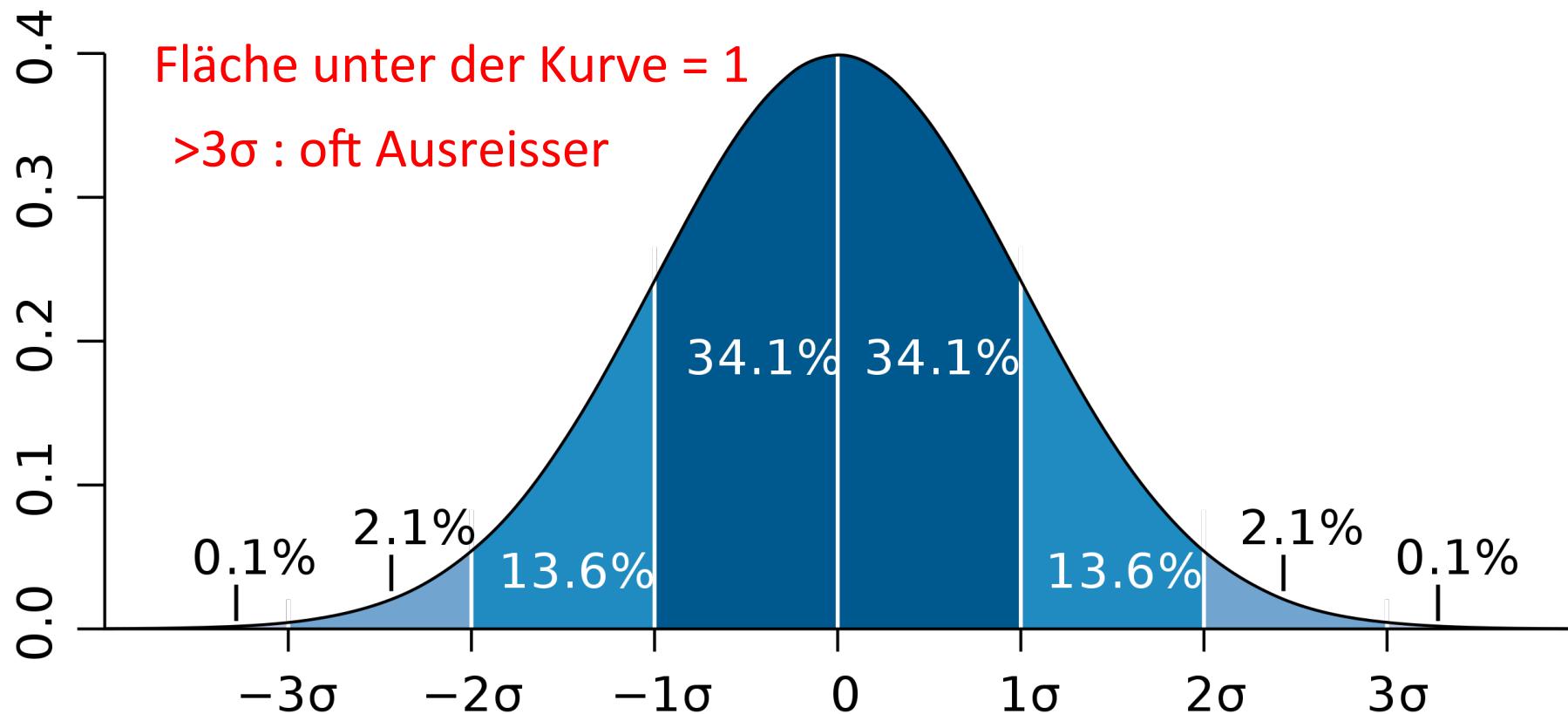
Wenn man Intervalle um den Mittelwert berechnet, kann man aussagen, wie viele Datenpunkte in welchem Intervall liegen.



Bildquelle: wikipedia

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.

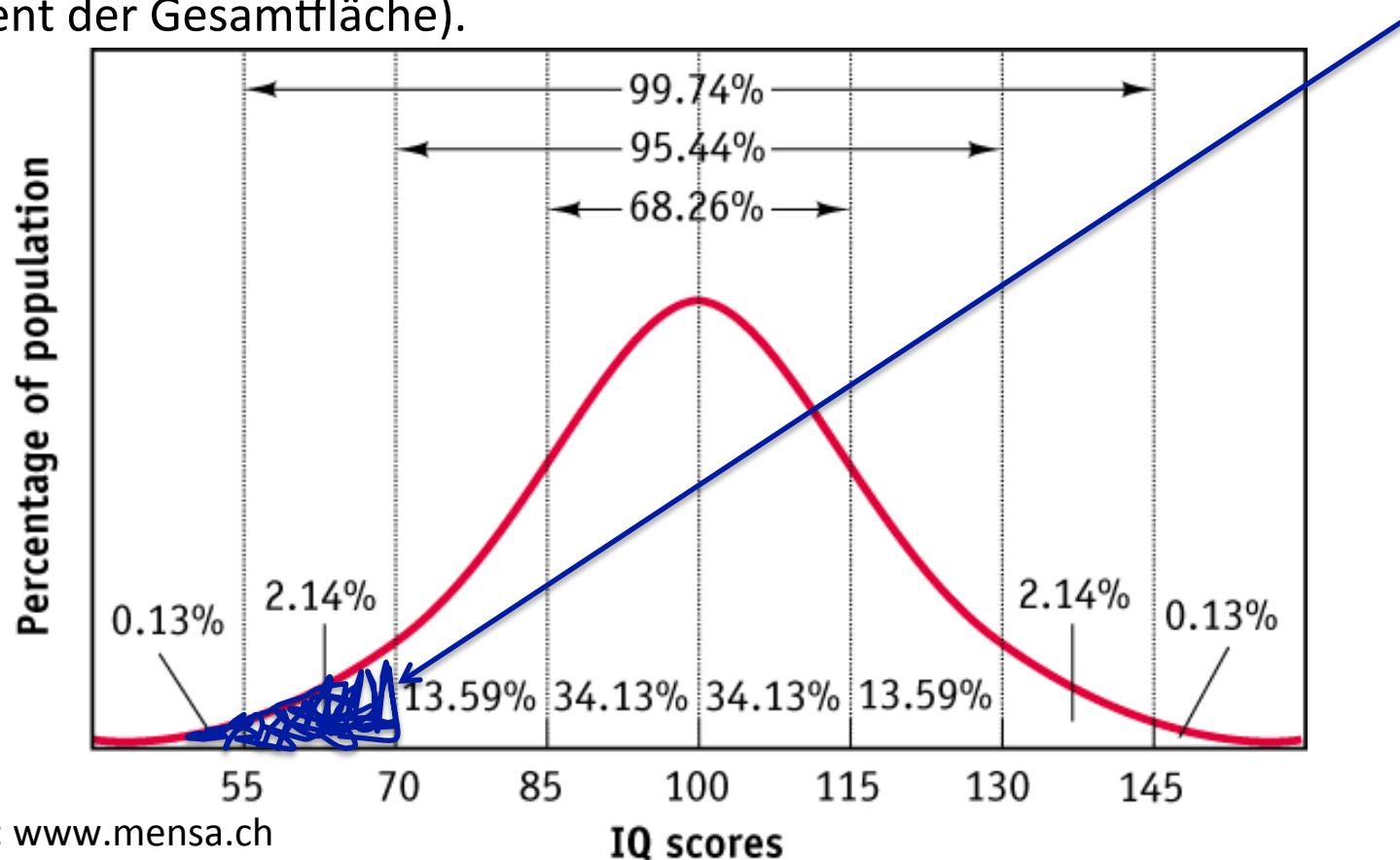


Normalverteilung

Beispiel: Wieviele Schüler mit IQ < 70 werden erwartet für Fördermassnahmen?

Bekannt: Mittelwert 100, Standardabweichung 15

Wir wissen, dass die totale Fläche = 1 ist. D.h. wir suchen das Integral von 0 bis 70 (als Prozent der Gesamtfläche).



Bildquelle: www.mensa.ch

Normalverteilung

Da das Integral dieser Funktion kompliziert ist, nimmt man Tabellen (oder z.B. Python).

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

Zufallsvariable
(Standardnormal-
verteilung)

Zufallsvariable
(beliebige
Normalverteilung)

Mittelwert

Standardabweichung

The diagram illustrates the formula for standardizing a random variable X from a general normal distribution into a standard normal distribution. The formula is $Z = \frac{X - \mu}{\sigma}$. Red arrows connect the text labels to the corresponding variables in the formula: 'Zufallsvariable (Standardnormal-verteilung)' connects to X, 'Zufallsvariable (beliebige Normalverteilung)' connects to μ , and 'Standardabweichung' connects to σ .

Also hier $Z = (70 - 100) / 15 = -2$

Normalverteilung

Also hier $Z = (70 - 100) / 15 = -2$

1 Standardnormalverteilung

Tabelliert sind die Werte der Verteilungsfunktion $\Phi(z) = P(Z \leq z)$ für $z \geq 0$.

Ablesebeispiel: $\Phi(1.75) = 0.9599$

Funktionswerte für negative Argumente: $\Phi(-z) = 1 - \Phi(z)$

Die z -Quantile ergeben sich genau umgekehrt.

Beispielsweise ist $z(0.9599) = 1.75$ und $z(0.9750) = 1.96$.

0.9772 abgelesen.

$1 - 0.9772 = 0.023$

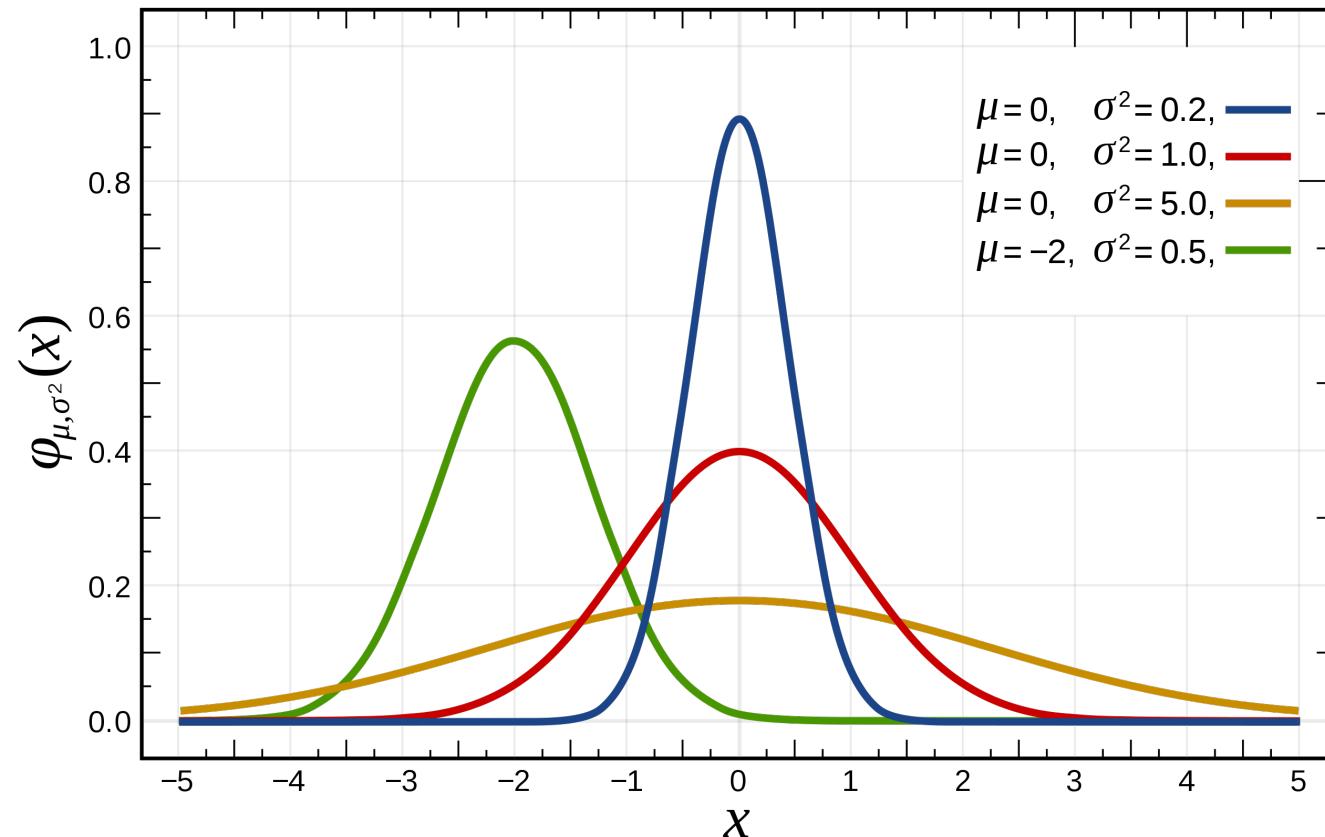
=> 2.3% der Schüler haben IQ < 70.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

Normalverteilung

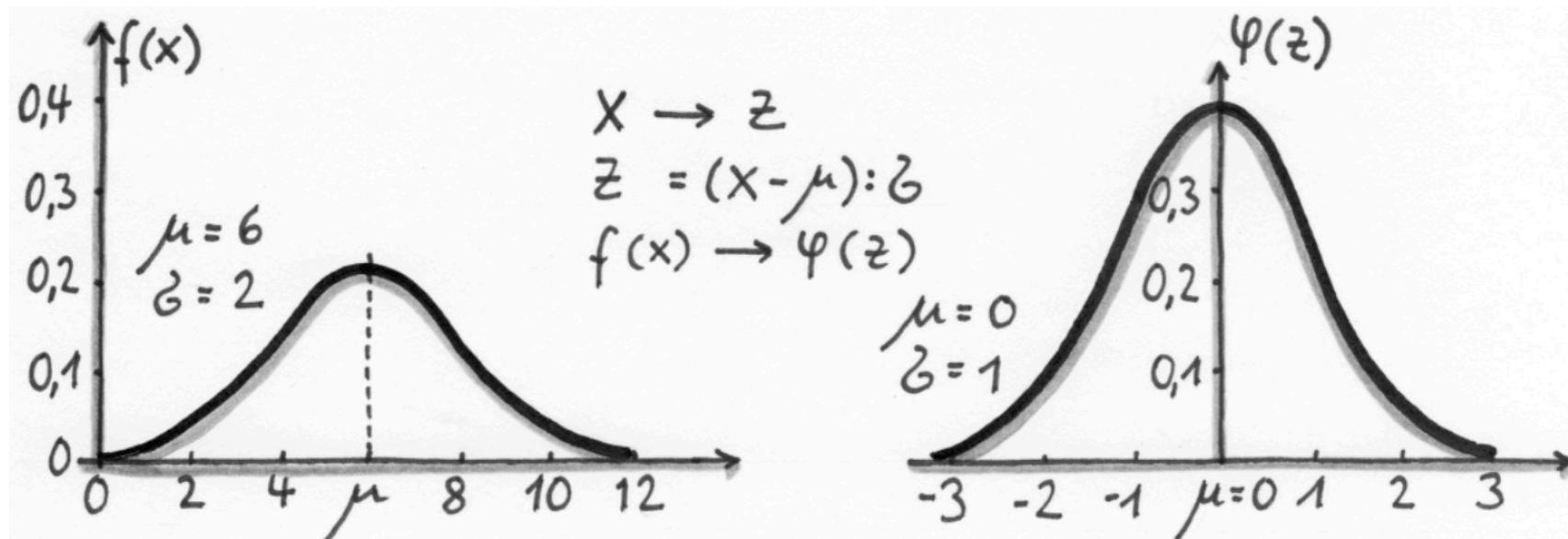
Warum haben wir nun zuerst den Z-Wert berechnet?

Oft hat man beliebige Normalverteilungen und man kann nicht für jede eine separate Tabelle haben. Für Berechnungen transformiert man deshalb in die Standardnormalverteilung ($\mu=0$, $\sigma=1$).



Bildquelle: wikipedia

Normalverteilung



Bildquelle: Buch "keine Panik vor Statistik"

Normalverteilung: Transformation

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

Zufallsvariable
(Standardnormal-verteilung)

Zufallsvariable
(beliebige
Normalverteilung)

Mittelwert

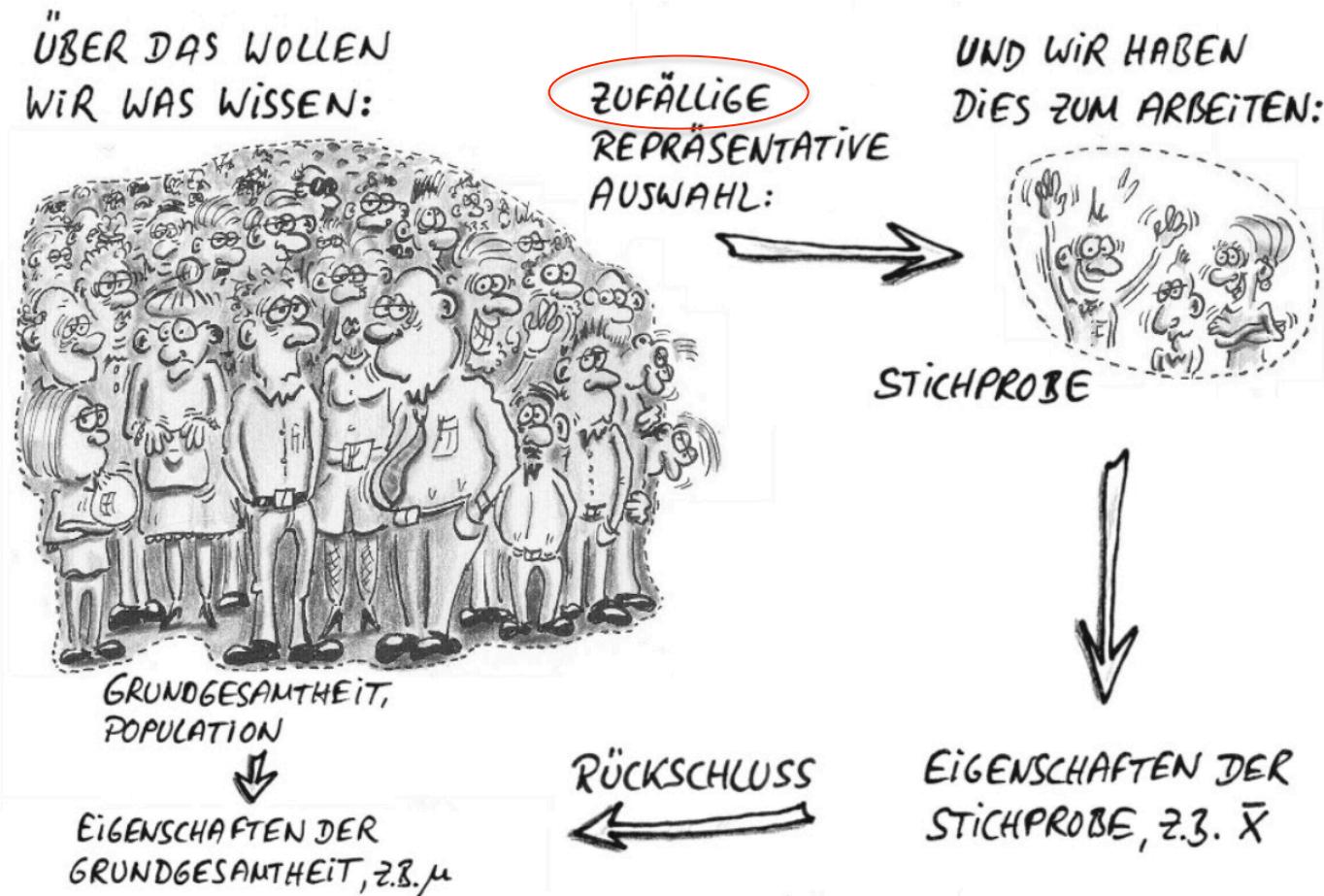
Standardabweichung

The diagram illustrates the transformation of a random variable X from a general normal distribution to a standard normal distribution. The formula is $Z = \frac{X - \mu}{\sigma}$. Red arrows connect the text labels to their respective terms in the formula: 'Zufallsvariable (Standardnormal-verteilung)' connects to X , 'Zufallsvariable (beliebige Normalverteilung)' connects to μ , 'Mittelwert' connects to μ , and 'Standardabweichung' connects to σ .

- 1) Standardabweichung muss bekannt sein
- 2) Bei Stichproben: Anzahl der Messwerte muss genügend gross sein (ca. >30).

Stichproben

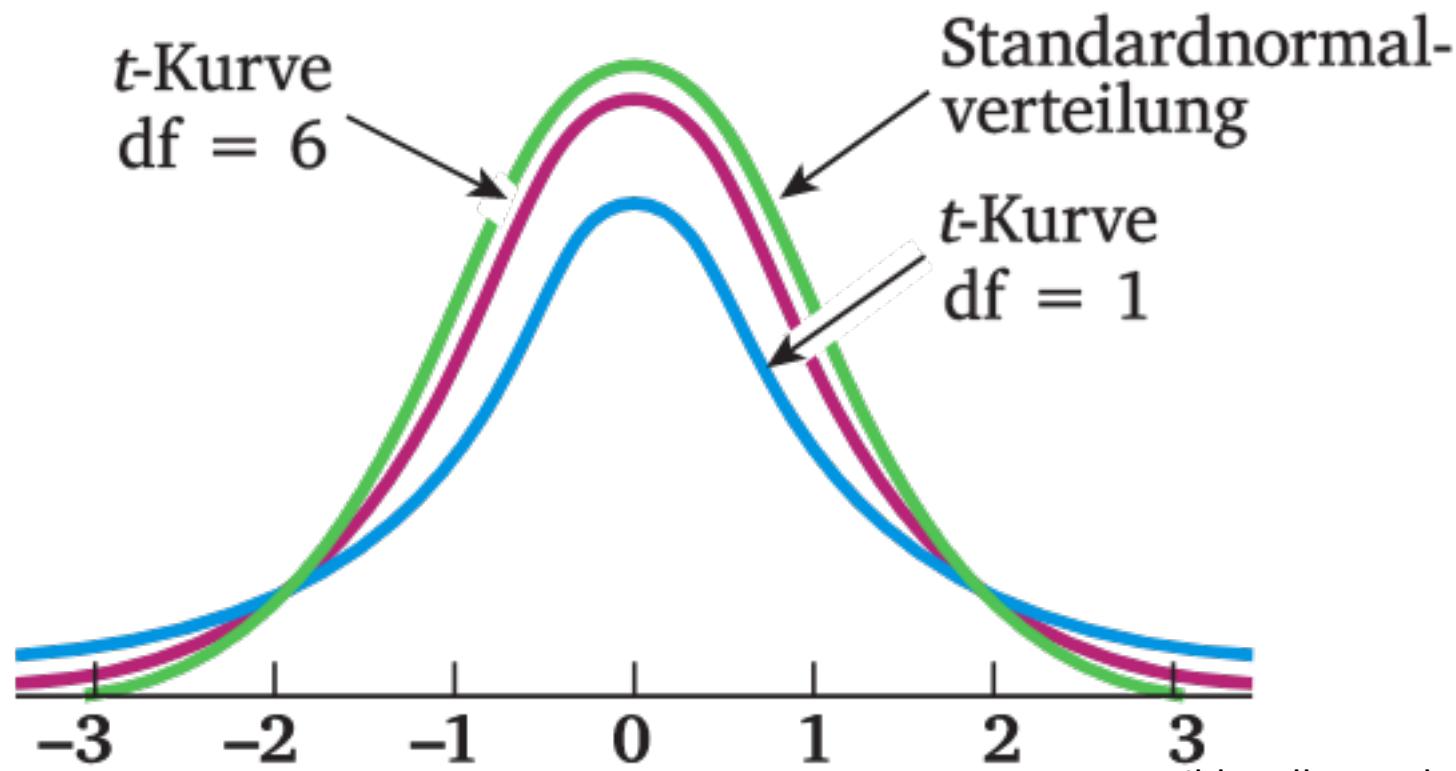
Oft hat man nur Stichproben, aber will etwas über die Grundgesamtheit aussagen.



Bildquelle: Buch "keine Panik vor Statistik"

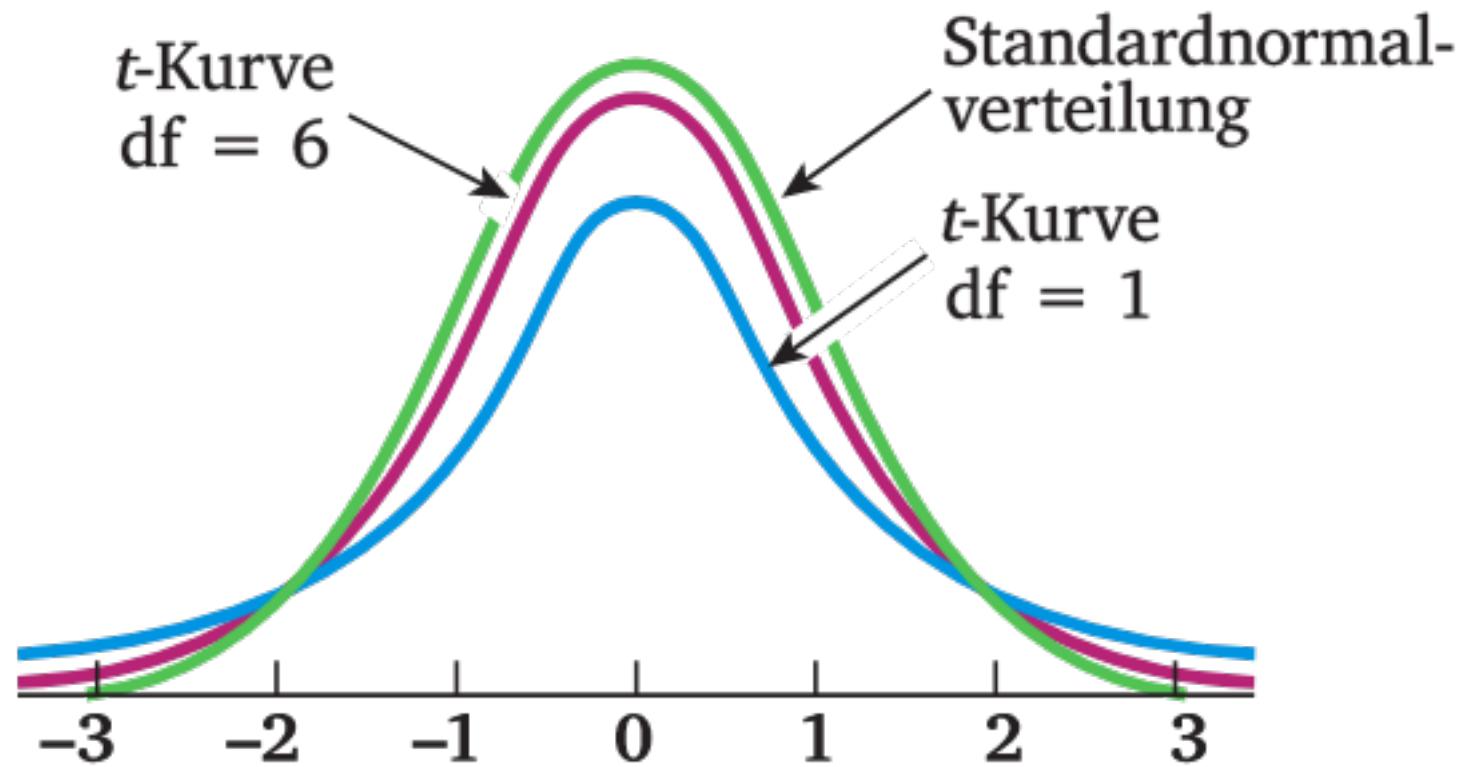
t-Verteilung

- 1) Standardabweichung der Grundgesamtheit ist unbekannt
- 2) Anzahl der Messwerte ist klein (ca. <30), aber man nimmt an, dass die Grundgesamtheit normalverteilt ist. ODER: Man hat eine grosse (ca. >30) Stichprobe.
- 3) Die Stichprobe wurde zufällig entnommen.
=> Normalverteilung kann nicht gebraucht werden.



t-Verteilung

Wie man es sich merken kann: Bei kleinen Stichproben erwischt man eher extreme Werte, deswegen ist die t-Verteilung höher an ihren Enden.



t-Verteilung

Analog zur Standardnormalverteilung erhält man den t-Wert:

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

Annotations for the t-value formula:

- "t-Wert" points to the leftmost t .
- "Mittelwert der Stichprobe" points to \bar{x} .
- "Mittelwert der Grundgesamtheit" points to μ .
- "Standardabweichung der Stichprobe" points to σ/\sqrt{N} .
- "Anzahl der Stichproben" points to N .
- "Standardabweichung des Stichprobenmittelwerts" points to $\sigma_{\bar{x}}$.

t-Verteilung: Beispiel

Glühbirnen X leuchten angeblich 300 Tage. Test mit 15 zufällig gewählten Glühbirnen: sie leuchten 290 Tage mit Std.abweichung 50 Tage. Lügt die Firma?

Berechnung t-Wert:

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}} = \frac{(290 - 300)}{50/\sqrt{15}} = -10/12.9 = -0.77.$$

Freiheitsgrade: 15-1=14

t-Verteilung: Beispiel

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}} = \frac{(290 - 300)}{50/\sqrt{15}} = -10/12.9 = -0.77.$$

To learn more about Student's t distribution, go to Stat Trek's [tutorial on the t distribution](#).

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Random variable

Degrees of freedom

t score

Probability: $P(T \leq -0.77)$

Calculate

Fläche unter der definierten Kurve wird berechnet.

t-Verteilung: Beispiel

Glühbirnen X leuchten angeblich 300 Tage. Test mit 15 zufällig gewählten Glühbirnen: sie leuchten 290 Tage mit Std.abweichung 50 Tage. Lügt die Firma?

Berechnung t-Wert:

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}} = \frac{(290 - 300)}{50/\sqrt{15}} = -10/12.9 = -0.77.$$

Wahrscheinlichkeit aus t-Verteilung mit Freiheitsgraden 15-1=14: 0.226.

Dies heisst, dass die Wahrscheinlichkeit 22.6% beträgt, dass die Lebensdauer ≤ 290 Tage ist und somit hat die Firma wahrscheinlich bei ihrer Aussage von 300 Tagen übertrieben.

Einstichproben t-Test

Man testet Eigenschaften von Mittelwerten

2 Varianten:

- Einstichproben t-Test

eine Stichprobe



© Can Stock Photo

Mittlere Länge $\bar{x}=3$ cm

?



Mittlere Längen μ_i bekannt,
Varianz unbekannt

- Berechne Standardabweichung der Stichprobe
- Berechne t-Wert
- wähle gewünschtes Signifikanzniveau (z.B. 5%). t dafür in Tabelle nachschauen
- $|t| < t_{n,1-\alpha/2} \Rightarrow$ Stichprobe aus G, $|t| \geq t_{n,1-\alpha/2} \Rightarrow$ Stichprobe nicht aus G

Zweistichproben t-Test

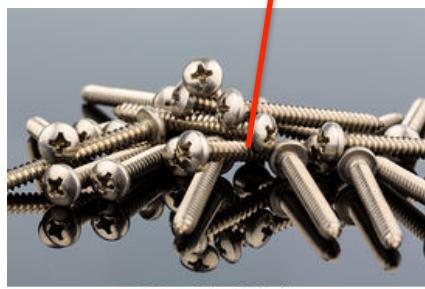
2 Varianten:

- Zweistichproben t-Test

Maschine 1



Stichprobe 1

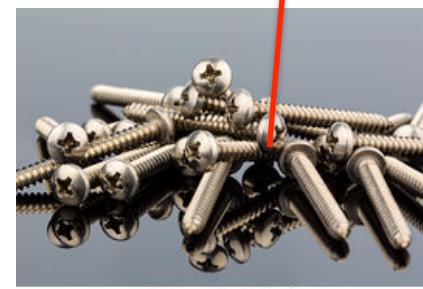


Mittlere Länge 3 cm. $\sigma=0.5$ cm

Maschine 2



Stichprobe 2



Mittlere Länge 3.5 cm. $\sigma=0.5$ cm

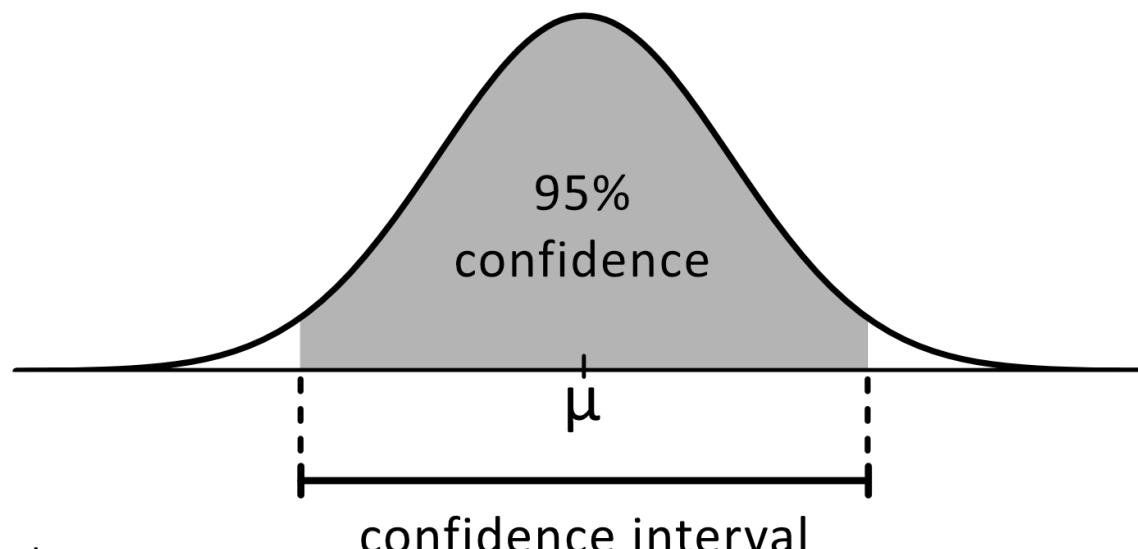
Produzieren Maschinen 1 und 2 die gleichen Schraubenlängen?

Konfidenzintervalle

Wir möchten wissen, wie gut unsere Schätzung des Mittelwerts war. Also welche Werte fallen in einen gewissen (gegebenen) Wahrscheinlichkeitsbereich?

=> Effektiv muss man die Fläche der Verteilung benützen, deswegen haben wir vorher zur Standardnormalverteilung transformiert.

Distribution of sample means (\bar{x})
around population mean (μ)



Konfidenzintervalle

Varianz bekannt oder grosse Stichprobe (=Normalverteilung):

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

Varianz unbekannt (=t-Verteilung):

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

Konfidenzintervalle: Beispiel

Beispiel: Eine Stichprobe von 10 Schrauben hat 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 (normalverteilten) Gesamtheit aller Schrauben mit einer Vertrauenswahrscheinlichkeit von $\gamma = 0.95$?

Konfidenzintervalle: Beispiel

Beispiel: Eine Stichprobe von 10 Schrauben hat 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 (normalverteilten) 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 $4.86 \text{ cm} \leq \mu \leq 5.14 \text{ cm}$ mit 95% Wahrscheinlichkeit

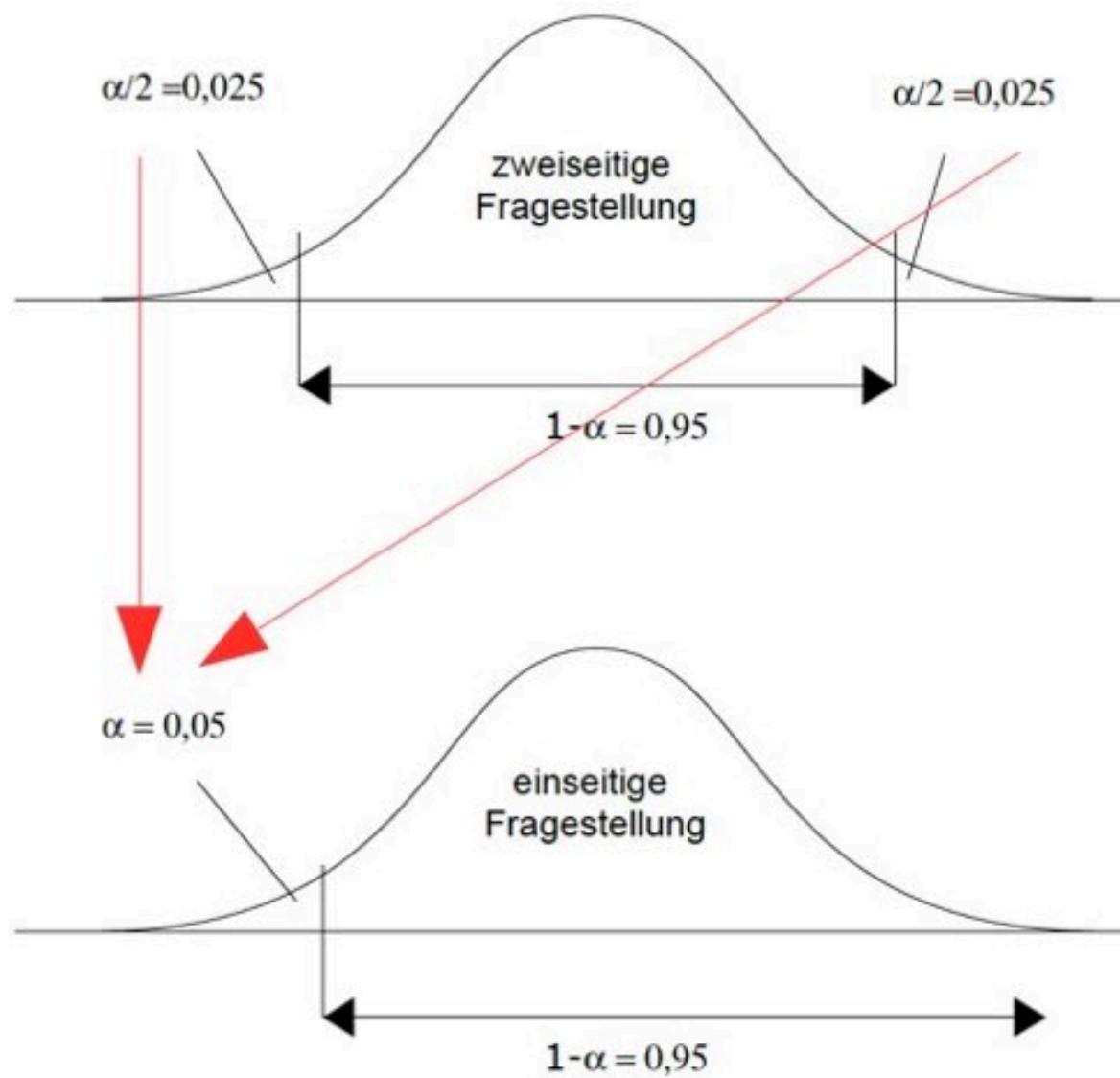
Konfidenzintervalle

t Tabellen haben 2 Optionen für alpha:

n	α	0,10	0,05	0,025	0,01	0,005	0,001	0,0005	einseitig zweiseitig
		0,20	0,10	0,05	0,02	0,01	0,002	0,001	
1		3,078	6,314	12,71	31,82	63,66	318,3	636,6	
2		1,886	2,920	4,303	6,965	9,925	22,33	31,56	
3		1,638	2,353	3,182	4,541	5,841	10,22	12,92	
4		1,533	2,132	2,776	3,747	4,604	7,173	8,610	
5		1,476	2,015	2,571	3,365	4,032	5,893	6,869	
6		1,440	1,943	2,447	3,143	3,707	5,208	5,959	
7		1,415	1,895	2,365	2,998	3,499	4,785	5,408	
8		1,397	1,860	2,306	2,896	3,355	4,501	5,041	
9		1,383	1,833	2,262	2,821	3,250	4,297	4,781	
10		1,372	1,812	2,228	2,764	3,169	4,144	4,587	
11		1,363	1,796	2,201	2,718	3,106	4,025	4,437	
12		1,356	1,782	2,179	2,681	3,055	3,930	4,318	
13		1,350	1,771	2,160	2,650	3,012	3,852	4,221	
14		1,345	1,761	2,145	2,624	2,977	3,787	4,140	
15		1,341	1,753	2,131	2,602	2,947	3,733	4,073	
16		1,337	1,746	2,120	2,583	2,921	3,686	4,015	
17		1,333	1,740	2,110	2,567	2,898	3,646	3,965	
18		1,330	1,734	2,101	2,552	2,878	3,610	3,922	
19		1,328	1,729	2,093	2,539	2,861	3,579	3,883	
20		1,325	1,725	2,086	2,528	2,845	3,552	3,850	
21		1,322	1,721	2,080	2,519	2,821	3,527	3,810	

Warum?

Konfidenzintervalle



Konfidenzintervalle für Regression

Regression füttet eine lineare Funktion:

$$f(X) = \alpha + \beta X$$

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

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

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

für ein Zahlenbeispiel,
siehe Serie LE2, Aufg. 9

n-2 Freiheitsgrade!

nicht der Achsenabschnitt, sondern Irrtumswahrscheinlichkeit