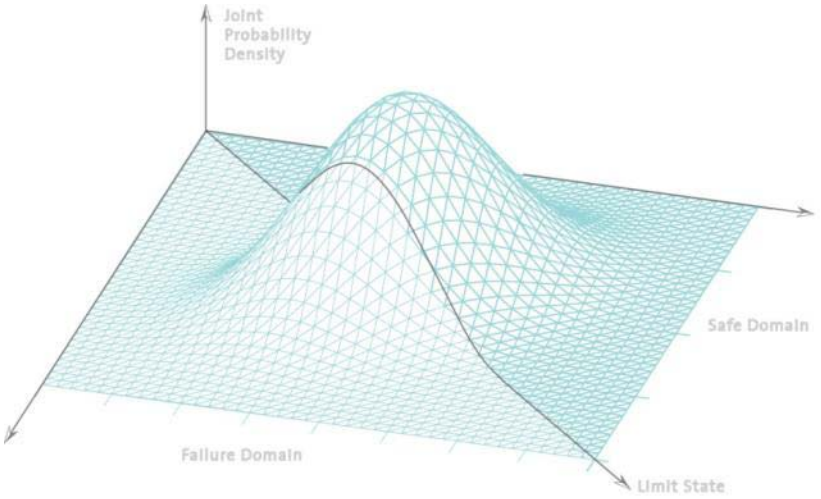


Statistik und Wahrscheinlichkeitsrechnung

Prof. Dr. Michael H. Faber

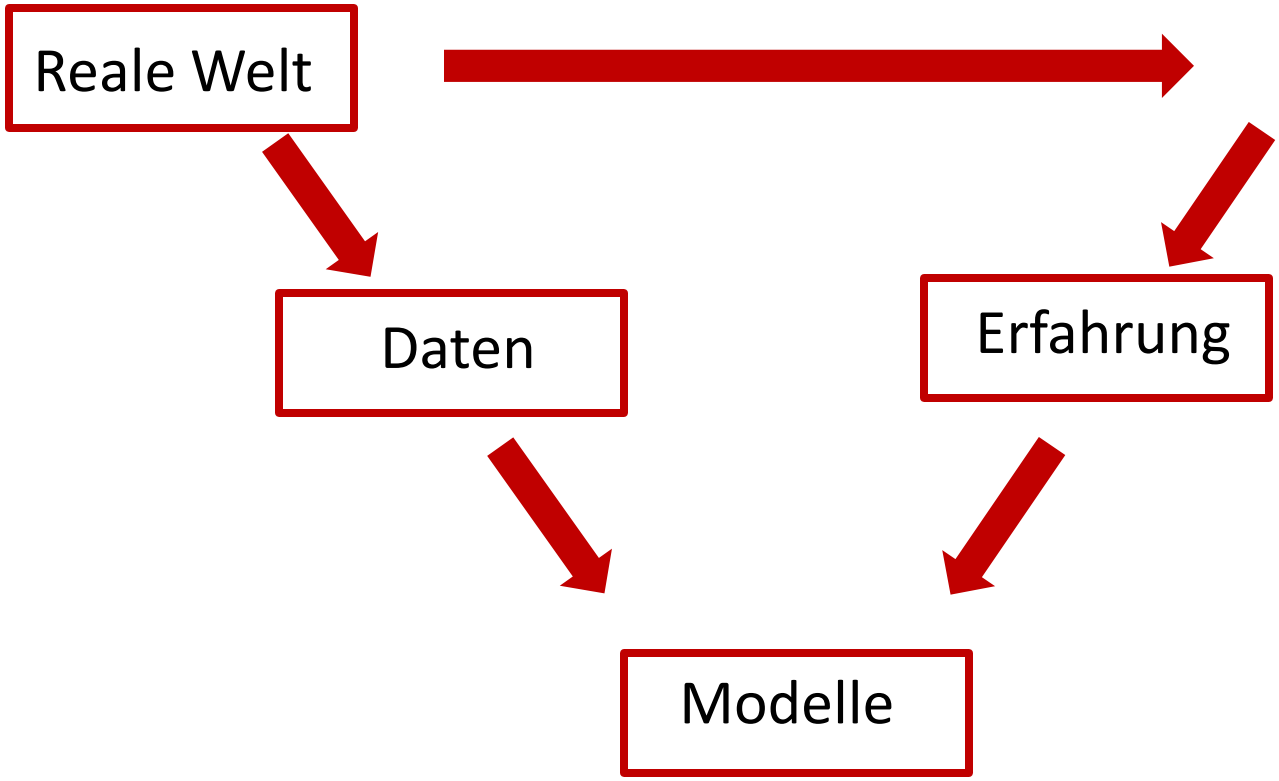
8. Vorlesung

Inhalte der heutigen Vorlesung

- Überblick über Schätzung und Modellbildung
- Wahrscheinlichkeitsverteilungen in der Statistik
- Parameterschätzung
  - Statistische Charakteristiken von Stichproben: Mittelwert
  - Statistische Charakteristiken von Stichproben: Varianz
  - Konfidenzintervalle der Schätzer

Überblick Schätzung und Modellbildung

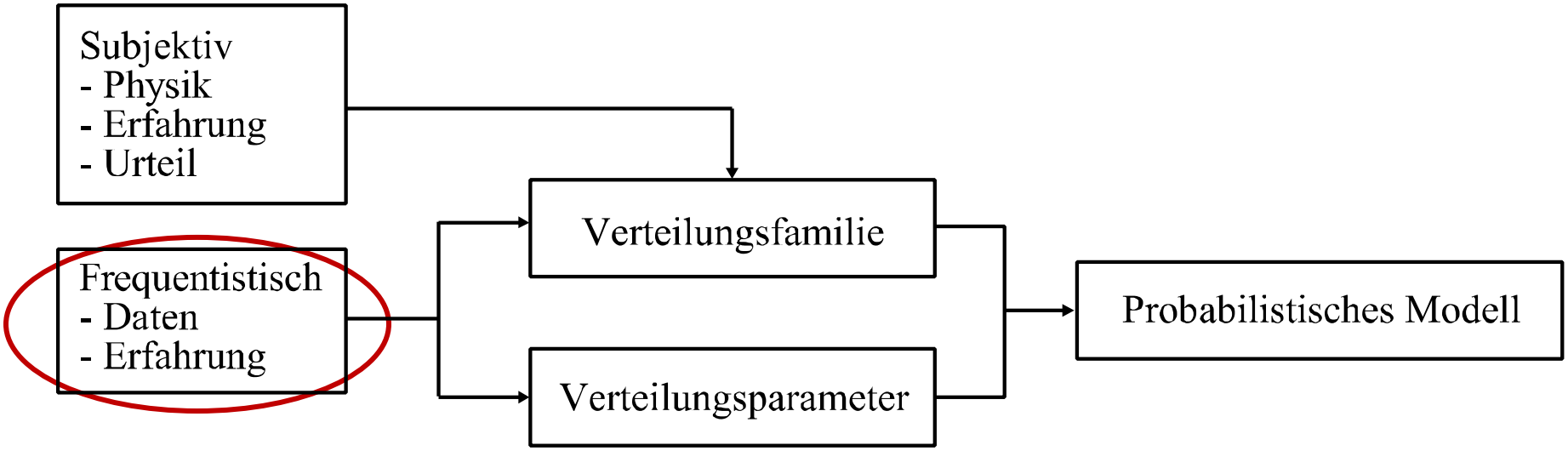
Wie kommen Ingenieure zu Wissen?



Überblick Schätzung und Modellbildung

Unterschiedliche Typen an Informationen werden zur Bildung von Ingenieurmodellen verwendet

- Subjektive Information
- Frequentistische Information



Überblick Schätzung und Modellbildung

Die Modellbildung kann in fünf Schritten erfolgen:

1) Bewertung und statistische Erfassung verfügbarer Daten
2) Wahl einer Verteilungsfunktion
3) Schätzung der Verteilungsparameter
4) Testen des Modells
5) Aktualisierung der Parameter des Modells

Wahrscheinlichkeitsverteilungsfunktionen

In der klassischen Statistik werden häufig bestimmte Wahrscheinlichkeitsverteilungsfunktionen, welche alle von der Normalverteilung abgeleitet werden können, verwendet und zur Bewertung und zum Testen verwendet.

Diese Wahrscheinlichkeitsverteilungsfunktionen sind:

- Chi-Quadrat Verteilung
- Chi-Verteilung
- t-Verteilung
- F-Verteilung

Wahrscheinlichkeitsverteilungsfunktionen

Chi-Quadrat Verteilung (χ^2- Verteilung)

Wenn $X_i,\ i = 1, 2, \ldots n$ standardnormalverteilte und unabhängige Zufallsvariablen sind , dann ist die Summe der Quadrate der Zufallsvariablen, also:

$$Y_n = \sum_{i=1}^{n} X_i^2$$ Chi-Quadrat verteilt.

Die Chi-Quadrat Verteilung ist regenerativ, d.h. die Summe der Chi-Quadrat verteilten Zufallsvariablen ist auch wieder Chi-Quadrat-verteilt.

Wahrscheinlichkeitsverteilungsfunktionen

Chi-Quadrat Verteilung (χ^2- Verteilung)

Betrachte den einfachsten Fall mit $n=1$, d.h. $Y_1 = X^2$
dann können wir schreiben

$$F_{Y_1}(y) = P(Y_1 \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le +\sqrt{y})$$
$$= F_X(\sqrt{y}) - F_X(-\sqrt{y}) = F_X(\sqrt{y}) - (1 - F_X(\sqrt{y})) =$$
$$= 2F_X(\sqrt{y}) - 1$$

und bekommen

$$f_{Y_1}(y) = \frac{dF_{Y_1}(y)}{dy} = \frac{d(2F_X(\sqrt{y})-1)}{dy} = y^{-\frac{1}{2}} f_X(\sqrt{y}) = \frac{1}{\sqrt{2\pi y}} \exp(-\frac{1}{2}y)$$

Wahrscheinlichkeitsverteilungsfunktionen

Chi-Quadrat Verteilung (χ^2- Verteilung)

- Chi-Quadrat Wahrscheinlichkeitsverteilung ist gegeben durch

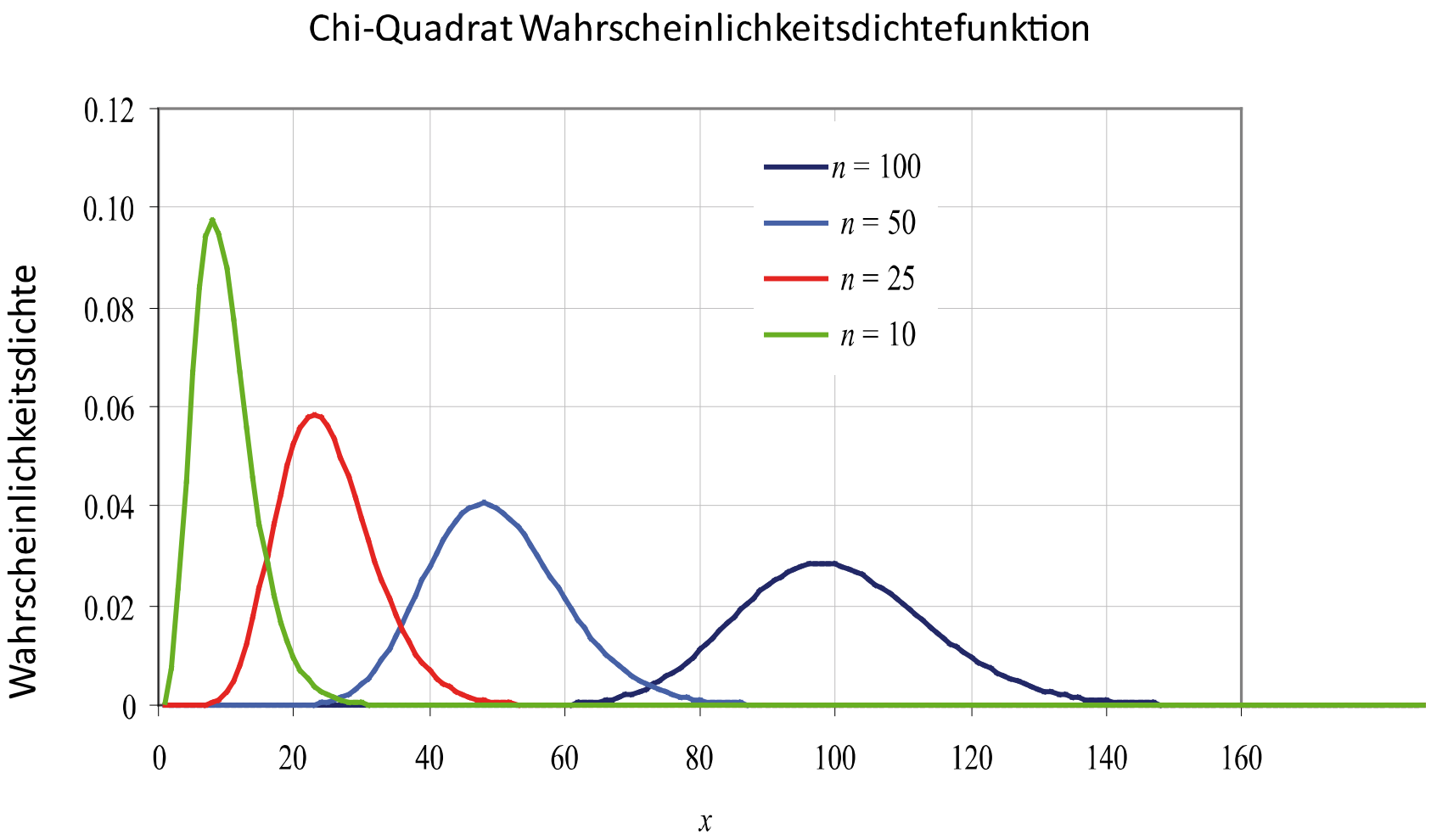
$$f_{Y_n}(y_n) = \frac{y_n^{\left(\frac{n}{2}-1\right)}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} exp\left(\frac{-y_n}{2}\right), \qquad y_n \geq 0$$

- Der Mittelwert ist $\mu_{Y_n} = n$
- Die Varianz ist $\sigma^2_{Y_n} = 2n$

Freiheitsgrade

- $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$ ist die komplette Gamma Funktion.
- Für grosse n konvergiert die Chi-Quadrat Verteilung zu einer Normalverteilung.

Wahrscheinlichkeitsverteilungsfunktionen



Wahrscheinlichkeitsverteilungsfunktionen

Chi-Verteilung (χ - Verteilung)

Wenn die Zufallsvariable Z durch die Wurzel von der Chi-Quadrat verteilten Zufallsvariable gegeben ist, d.h.

$$Z = \sqrt{Y_n} = \sqrt{\sum_{i=1}^{n} X_i^2}$$

dann ist sie Chi-verteilt mit n Freiheitsgraden.

Wahrscheinlichkeitsverteilungsfunktionen

Chi-Verteilung (χ - Verteilung)

Angenommen, dass Y_n Chi-Quadrat verteilt ist mit n Freiheitsgraden.

Mit $Z = \sqrt{Y_n}$ können wir schreiben

$$F_Z(z) = P(Z \le z) = P(\sqrt{Y_n} \le z) = P(Y_n \le z^2) = F_{Y_n}(z^2)$$

Und wir bekommen

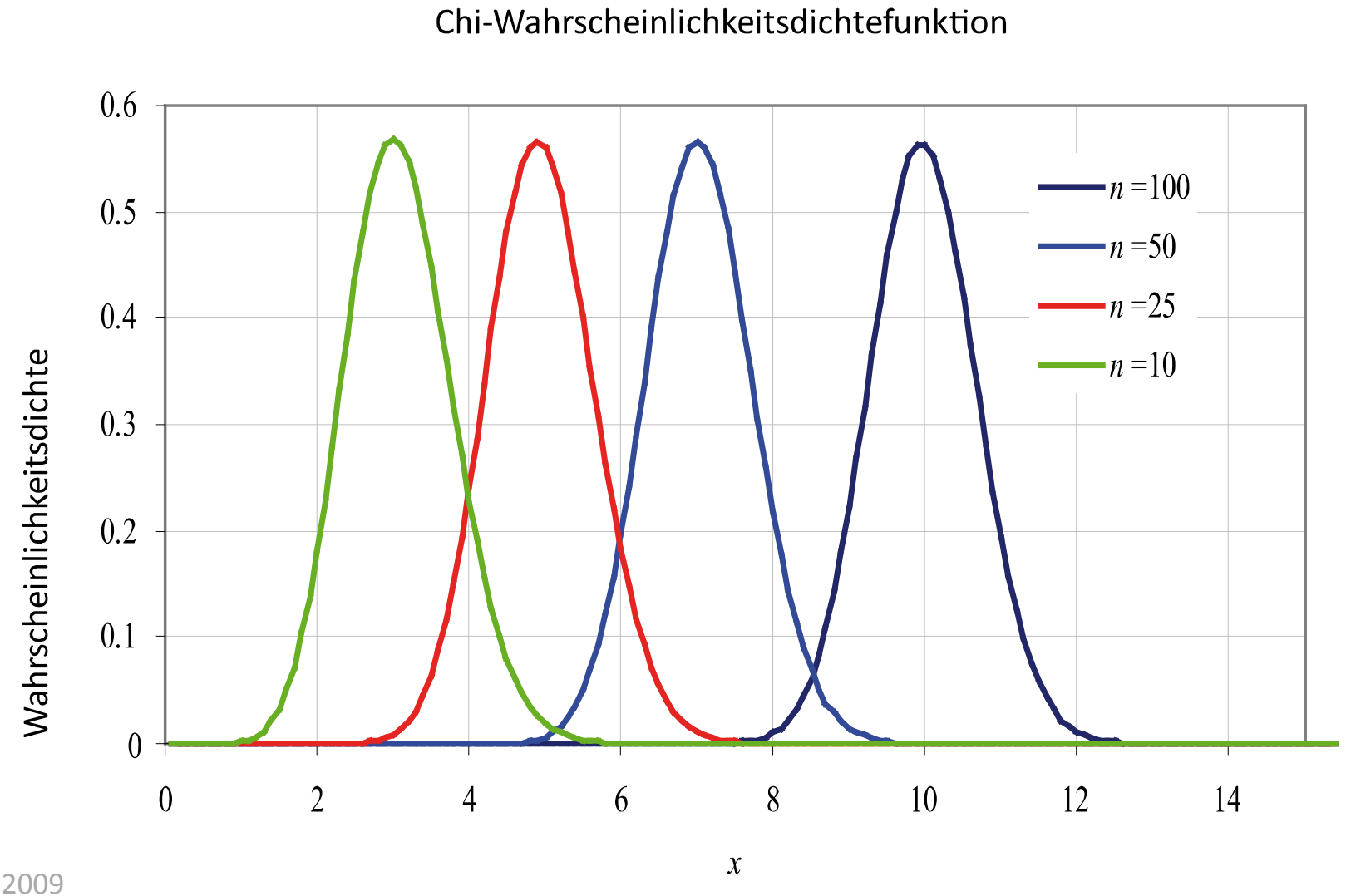
$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{dF_{Y_n}(z^2)}{dz} = 2zf_{Y_n}(z^2) = \frac{z^{n-1}}{2^{\left(\frac{n}{2}-1\right)}\Gamma\left(\frac{n}{2}\right)}\exp(-\frac{z^2}{2})$$

Wahrscheinlichkeitsverteilungsfunktionen

Chi-Verteilung (χ - Verteilung)

- Ist gegeben durch $$f_Z(z) = \frac{z^{(n-1)}}{2^{\left(\frac{n}{2}-1\right)}\Gamma\left(\frac{n}{2}\right)} \exp\left(\frac{-z^2}{2}\right), \qquad z \geq 0$$
- Der Mittelwert ist $$\mu_z = \sqrt{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$
- Die Varianz ist $$\sigma_z^2 = n - 2\frac{\Gamma^2\left(\frac{n+1}{2}\right)}{\Gamma^2\left(\frac{n}{2}\right)}$$

Wahrscheinlichkeitsverteilungsfunktionen



Wahrscheinlichkeitsverteilungsfunktionen

t-Verteilung (Student-Verteilung)

- Wenn eine standardnormalverteilte Zufallsvariable X durch eine Chi-verteilte Zufallsvariable geteilt wird, d.h.

$$S = \frac{X}{\frac{\sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{n}}} = \frac{X}{\frac{\sqrt{Y_n}}{\sqrt{n}}} = \frac{X}{\frac{Z}{\sqrt{n}}} = \frac{\sqrt{n}X}{Z}$$

  dann heisst die Verteilung von S *t*-Verteilung bzw. Student-Verteilung mit n Freiheitsgraden.

- Für grosse n konvergiert die *t*-Verteilung zu einer Normalverteilung

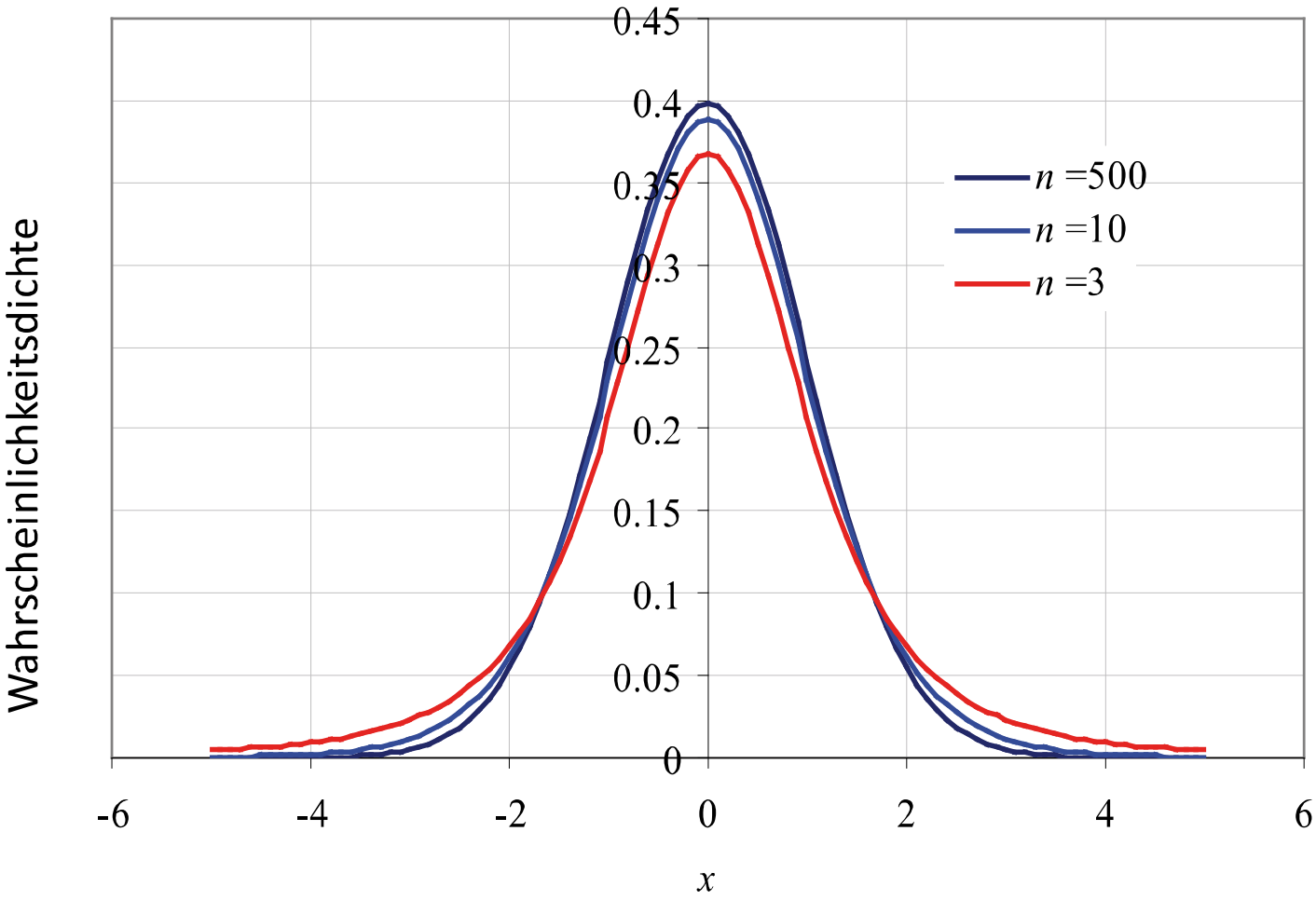
Wahrscheinlichkeitsverteilungsfunktionen

t-Verteilung (Student-Verteilung)

- Ist gegeben durch $$f_S(s) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\ \Gamma\left(\frac{n}{2}\right)}\left(1+\frac{s^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, \qquad -\infty \le s \le \infty$$
- Der Mittelwert ist 0
- Die Varianz ist $\sigma_S^2 = \frac{n}{n-2}$

Wahrscheinlichkeitsverteilungsfunktionen

t- Wahrscheinlichkeitsdichtefunktion



Wahrscheinlichkeitsverteilungsfunktionen

F-Verteilung

- Wenn eine Zufallvariable Q gegeben ist als das Verhältnis zwischen zwei Chi-Quadrat-verteilten Zufallsvariablen, d.h.

$$Q = \frac{Y_{n_1}}{Y_{n_2}}$$

- Dann ist Q F-verteilt mit den Freiheitsgraden n_1, n_2.

Wahrscheinlichkeitsverteilungsfunktionen

F-Verteilung

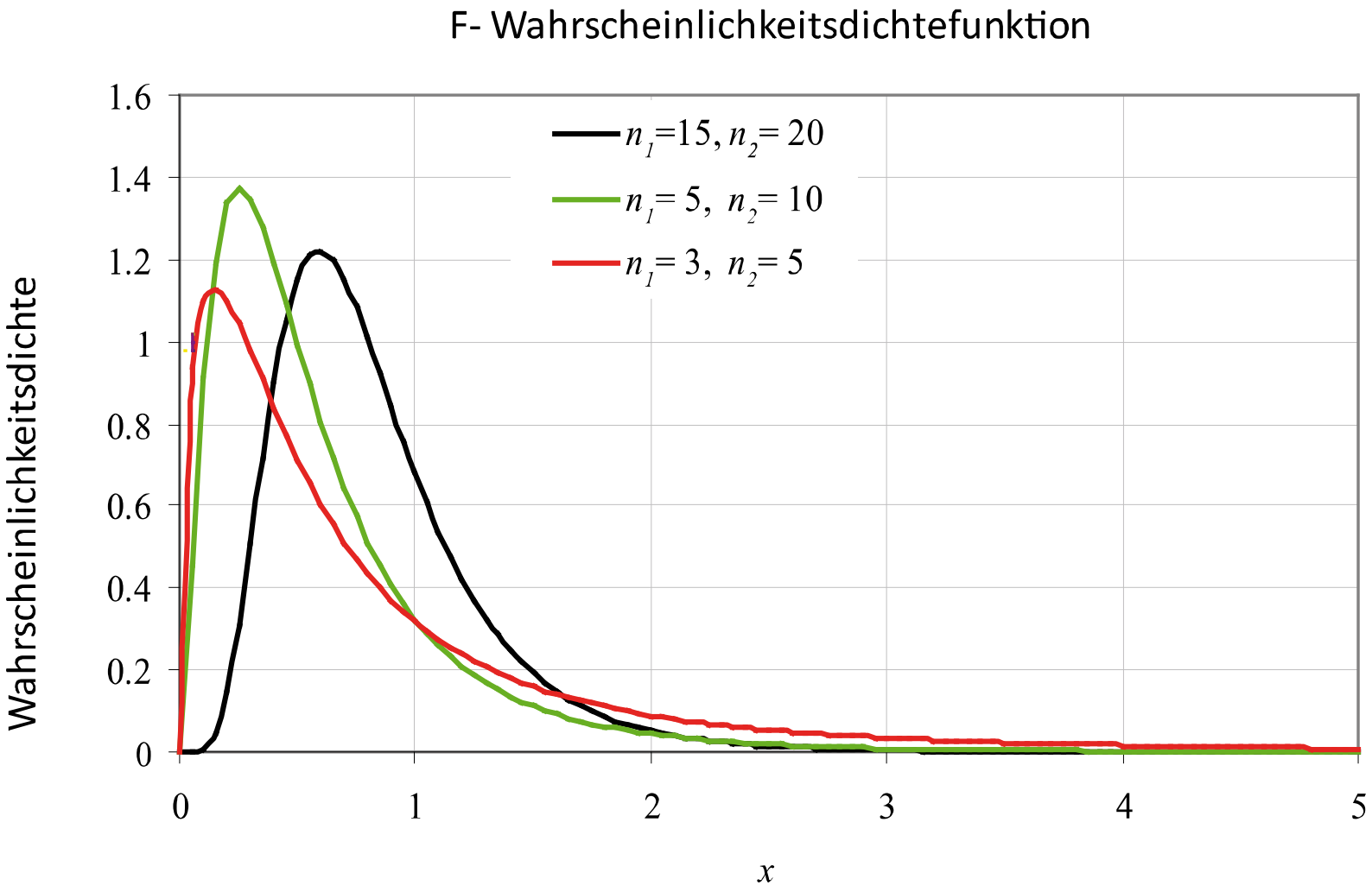
- Die F-Wahrscheinlichkeitsdichtefunktion ist gegeben als

$$f_Q(q) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right) q^{\left(\frac{n_1-2}{2}\right)} (1+q)^{-\left(\frac{n_1+n_2}{2}\right)}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}, \qquad q \geq 0$$

- Der Mittelwert ist $\mu_Q = \frac{n_2}{n_2 - 2}, \qquad n_2 > 2$

- Die Varianz ist $\sigma_Q^2 = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \qquad n_2 > 4$

Wahrscheinlichkeitsverteilungsfunktionen



Wahrscheinlichkeitsverteilungsfunktionen

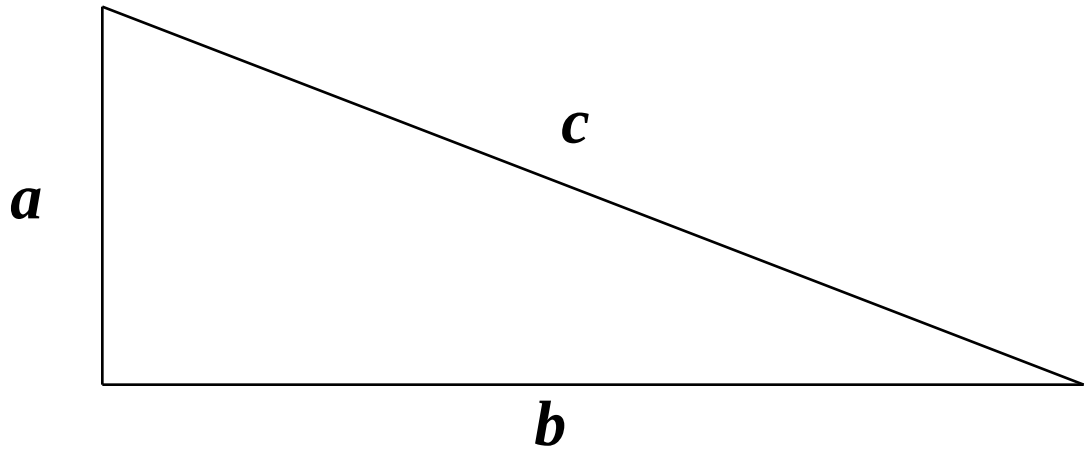
Zusammenfassung: Basierend auf unabhängigen normal verteilten Zufallsvariablen lassen sich folgende Verteilungen ableiten

Verteilungstyp	Wann
Chi-Quadrat-Verteilung	Summe der Quadrate $N(0;1)$
Chi-Verteilung	Wurzel von Chi-Quadrat
t-Verteilung	Verhältnis von $N(0;1)$ zu Chi/n
F-Verteilung	Verhältnis von zwei Chi-Quadrat

Wahrscheinlichkeitsverteilungsfunktionen

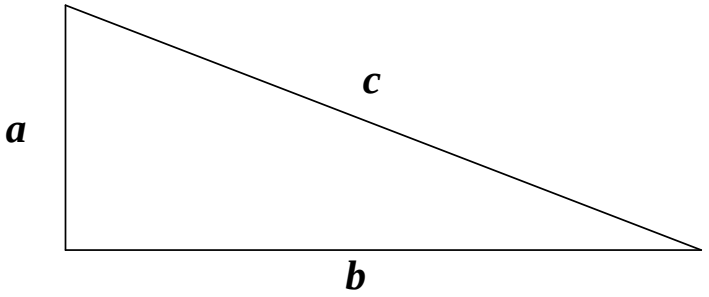
Beispiel: Chi-Verteilung

- Es wurden Messungen von den Seiten a und b durchgeführt, mit der Absicht, die Seite c bestimmen zu können.



Wahrscheinlichkeitsverteilungsfunktionen

Beispiel: Chi-Verteilung

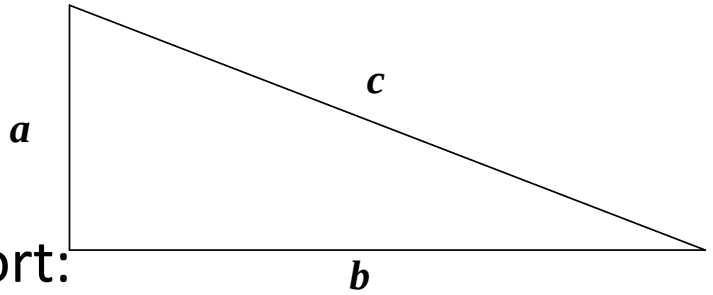


- Es wird angenommen, dass die Messungen von a und b mit dem selben absoluten Fehler ε durchgeführt werden, welcher als $N(0; \sigma_\varepsilon)$ angenommen wird (Normalverteilt, erwartungstreu -> d. h. ohne systematischen Fehler und mit einer Standardabweichung σ_ε.).

- Bestimme die statistischen Charakteristiken des Fehlers in c, welcher durch a und b bestimmt wurde.

Wahrscheinlichkeitsverteilungsfunktionen

Beispiel: Chi-Verteilung



- Der Fehler setzt sich folgendermassen fort:

$$\varepsilon_c = \sqrt{\varepsilon_a^2 + \varepsilon_b^2}$$

- Daraus lässt sich folgen, dass

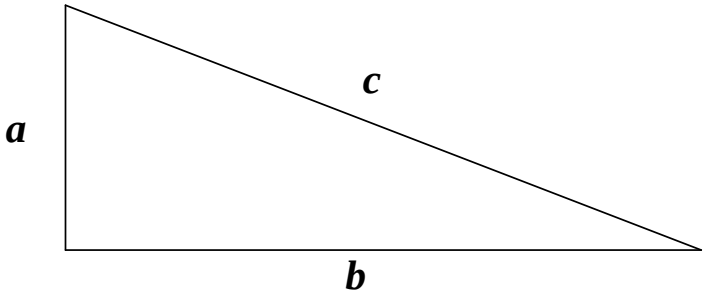
$$\frac{\varepsilon_c}{\sigma_\varepsilon} = \sqrt{\left(\frac{\varepsilon_a}{\sigma_\varepsilon}\right)^2 + \left(\frac{\varepsilon_b}{\sigma_\varepsilon}\right)^2}$$

Chi-verteilt ist,

mit zwei Freiheitsgraden.

Wahrscheinlichkeitsverteilungsfunktionen

Beispiel: Chi-Verteilung



- Die Wahrscheinlichkeitsdichtefunktion von $Z = \frac{\varepsilon_c}{\sigma_\varepsilon}$ kann bestimmt werden durch

$$f_Z(z) = z\exp(-0.5z^2), \qquad z \geq 0$$

Unter Einhaltung von $$f_{\varepsilon_c}(\varepsilon_c) = \frac{\varepsilon_c}{\sigma_\varepsilon}\exp\left(\frac{-0.5\cdot\varepsilon_c^2}{\sigma_\varepsilon^2}\right), \qquad \varepsilon_c \geq 0$$

Parameterschätzung für Stichproben

Wenn neue Daten verfügbar sind, besteht der erste Schritt darin, diese zu beurteilen.

Daten/Beobachtungen

n	$\hat{x}_n^o$	$F_X(x_n)$
1	24.4	0.047619048
2	27.6	0.095238095
3	27.8	0.142857143
4	27.9	0.19047619
5	28.5	0.238095238
6	30.1	0.285714286
7	30.3	0.333333333
8	31.7	0.380952381
9	32.2	0.428571429
10	32.8	0.476190476
11	33.3	0.523809524
12	33.5	0.571428571
13	34.1	0.619047619
14	34.6	0.666666667
15	35.8	0.714285714
16	35.9	0.761904762
17	36.8	0.80952381
18	37.1	0.857142857
19	39.2	0.904761905
20	39.7	0.952380952

Mittelwert

Varianz

Median

Usw.

Funktion von Stichproben

Stichproben-charakteristik

oder

Stichproben-statistik

Parameterschätzung für Stichproben

- Die statistischen Eigenschaften von Stichprobenstatistiken werden im folgenden genauer betrachtet, um die darin enthaltenen Informationen besser zu verstehen.

- Angenommen wir haben noch unbekannte Stichproben $X_i,\ i = 1,2,..n$ aus einem Experimentergebnis generiert durch die kumulative Verteilungsfunktion $F_{X_i}(x_i, \mathbf{p}) = F_X(x, \mathbf{p}), i = 1,2,..n$

  Dann können wir die Stichprobenstatistiken beschreiben für den Stichprobenmittelwert und die Stichprobenvarianz.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Parameterschätzung für Stichproben

Die Stichprobenstatistiken sind Zufallsvariablen, solange die Ergebnisse das Experiments noch nicht realisiert sind.

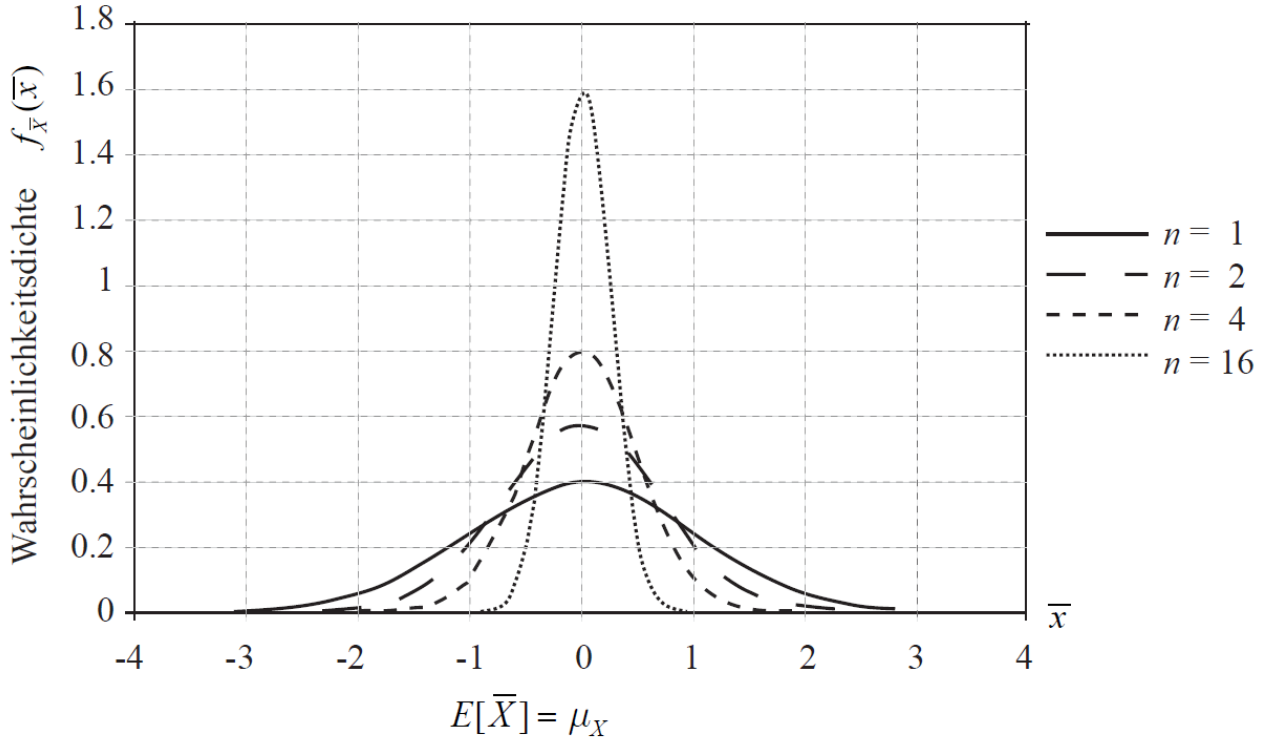
Daher kann der Erwartungswert und die Varianz für den Stichprobenmittelwert folgendermassen bestimmt werden:

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n} n\mu_X = \mu_X$$

$$Var\left[\bar{X}\right] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var\left[X_i\right] = \frac{1}{n}\sigma_X^2$$

Parameterschätzung für Stichproben

Die Wahrscheinlichkeitsdichtefunktion für den Stichprobenmittelwert kann als eine Normalverteilung angenommen werden – Zentraler Grenzwertsatz



Parameterschätzung für Stichproben

Beispiel: Körpergewicht der Studierenden.

	1. Probe
	G [kg]
1	75
2	75
3	80
4	72
5	84
6	90
7	55
8	85
9	69
10	70

Mittelwert	75.5
Standardabweichung	8.99

Parameterschätzung für Stichproben

Beispiel: Körpergewicht der Studierenden.

	1. Probe	2. Probe
	G [kg]	G [kg]
1	75	65
2	75	77
3	80	68
4	72	85
5	84	71
6	90	76
7	55	79
8	85	80
9	69	75
10	70	80

Mittelwert	75.5	75.6
Standardabweichung	8.99	5.47

Parameterschätzung für Stichproben

Beispiel: Körpergewicht der Studierenden.

	1. Probe	2. Probe	3. Probe	4. Probe	5. Probe
	G [kg]	G [kg]	G [kg]	G [kg]	G [kg]
1	75	65	63	72	59
2	75	77	62	78	73
3	80	68	58	59	73
4	72	85	76	65	69
5	84	71	93	90	56
6	90	76	72	76	60
7	55	79	58	62	71
8	85	80	76	77	75
9	69	75	58	57	60
10	70	80	79	63	70
Mittelwert	75.5	75.6	69.5	69.9	66.6
Standardabweichung	8.99	5.47	10.51	9.40	6.34

Parameterschätzung für Stichproben

Für die Stichprobenvarianz erhalten wir:

$$E\left[S^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2\right]$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n}E\left[(X_i - \mu)^2\right] - n\,E\left[(\overline{X} - \mu)^2\right]\right)$$

$$= \frac{1}{n}\left(n \cdot E\left[(X_i - \mu)^2\right] - n\,E\left[(\overline{X} - \mu)^2\right]\right) =$$

$$= \frac{1}{n}\left(n \cdot \sigma_X^2 \; - \; n\frac{\sigma_X^2}{n}\right)$$

$$= \sigma_X^2 - \frac{1}{n}\sigma_X^2 = \frac{(n-1)}{n}\sigma_X^2$$

Der Schätzer der Stichprobenvarianz ist nicht erwartungstreu (biased).

Parameterschätzung für Stichproben

$$E\left[S^2\right] = \frac{(n-1)}{n}\sigma_X^2$$

Wir können nun einfach erwartungstreue (unbiased) Schätzer für die Varianz bestimmen:

$$\breve{S}^2 = \frac{n}{n-1}S^2$$

$$= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Parameterschätzung für Stichproben

$$E\left[S^2\right] = \frac{(n-1)}{n}\sigma_X^2$$

Wir können nun einfach erwartungstreue (unbiased) Schätzer für die Varianz bestimmen:

$$\breve{S}^2 = \frac{n}{n-1}S^2$$

Nicht n wie bei der Varianz in der beschreibenden Statistik!

$$= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Parameterschätzung für Stichproben

Die Qualität eines Schätzers kann nicht alleine dadurch bewertet werden, ob er erwartungstreu ist. Es spielen auch andere Eigenschaften eine wichtige Rolle wie:

- Effizienz — kleinste Fehlerquadrate
- Invarianz — $h(\bar{\theta}) = \overline{h(\theta)}$
- Konsistenz — Konvergenz zu wahren Werten
- Hinlänglichkeit — Maximaler Nutzen aus den Daten
- Robustheit — Sensitivität bei Weglassen individueller Daten

Wir werden dies nicht im Detail anschauen – Merken Sie sich, dass diese Überlegungen von Bedeutung sind.

Konfidenzintervalle für Schätzer

- Wir haben gesehen, dass Schätzer z.B. des Mittelwertes mit statistischen Unsicherheiten assoziiert sind, und wir haben ihren Mittelwert und ihre Varianz bestimmt.

- Basierend auf diesen Informationen ist es uns möglich, ein Konfidenzintervall für die Schätzer zu bestimmen.

- Konfidenzintervalle können als Intervalle verstanden werden, innerhalb welcher z.B. der Mittelwert mit einer bestimmten Wahrscheinlichkeit gefunden werden kann.

Konfidenzintervalle für Schätzer

- Wir können ein Konfidenzintervall z.B. für den Mittelwert erstellen.
- Für den Fall, dass der Mittelwert unsicher und die Varianz bekannt ist:

  Aufgrund von n Beobachtungen lässt sich der Mittelwert schätzen als (normalverteilte) Zufallsvariable mit Mittelwert gleich $\overline{X}$ und Standardabweichung $\sigma_X \frac{1}{\sqrt{n}}$.

- Durch Transformation erhalten wir die standardnormalverteilte Zufallsvariable

$$\frac{\overline{X} - \mu_X}{\sigma_X \frac{1}{\sqrt{n}}}$$

Konfidenzintervalle für Schätzer

Das zweiseitige und symmetrische Konfidenzintervall des Mittelwertes ist gegeben durch:

Stichprobenmittelwert

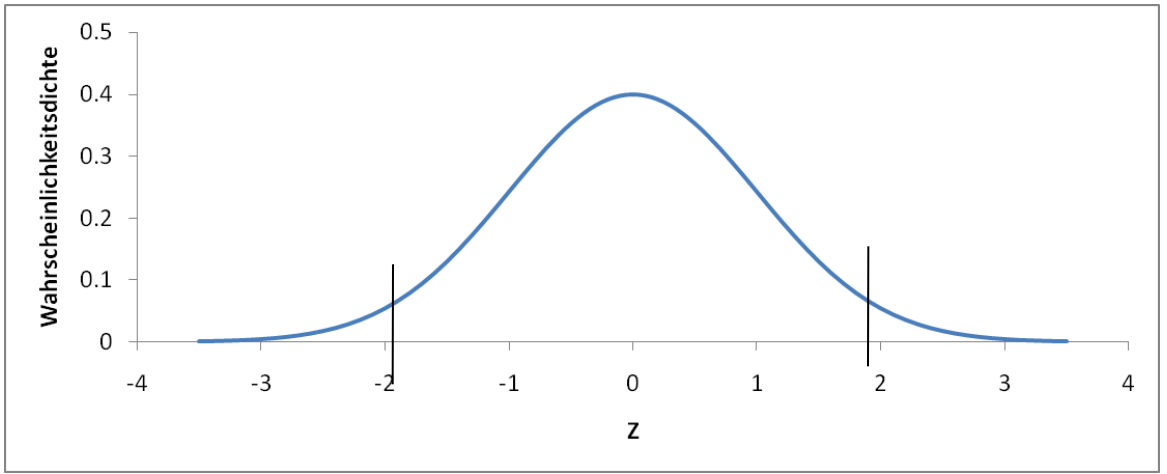
wahrer Mittelwert

$$P\left[-k_{\alpha/2} < \frac{\bar{X} - \mu_X}{\sigma_X \frac{1}{\sqrt{n}}} < k_{\alpha/2}\right] = P\left[-k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}} < \bar{X} - \mu_X < k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}}\right] = 1 - \alpha$$

Anzahl Stichproben

bekannte Standardabweichung

Signifikanzniveau



Konfidenzintervalle für Schätzer

Das Konfidenzintervall definiert ein Intervall, in dem der Stichprobenmittelwert mit der Wahrscheinlichkeit $1-\alpha$ liegt.

$$P\left[-k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}} < \overline{X} - \mu_X < k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}}\right] = 1-\alpha$$

Bekannte Standardabweichung

Stichprobenmittelwert

Wahrer Mittelwert

Anzahl Stichproben

Das Konfidenzintervall kann, durch die Annahme, dass der Mittelwert normalverteilt ist, wie folgt bestimmt werden:

$$k_{\alpha/2} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right) = \Phi^{-1}\left(1-\frac{0.05}{2}\right) = 1.96$$

Konfidenzintervalle für Schätzer

Für den Fall, dass $\alpha = 0.05$, $n = 16$ und $\sigma_X = 20$ erhalten wir

$$P\left[-1.96 < \frac{\bar{X} - \mu_X}{20\frac{1}{\sqrt{n}}} < 1.96\right] = 1 - 0.05$$

$$P\left[-9.8 < \bar{X} - \mu_X < 9.8\right] = 0.95$$

Konfidenzintervalle für Schätzer

- Wenn wir beobachten, dass der Stichprobenmittelwert z.B. gleich 400 ist, wissen wir, dass der wahre Mittelwert mit einer Wahrscheinlichkeit von 0.95 innerhalb des Intervalles liegt.

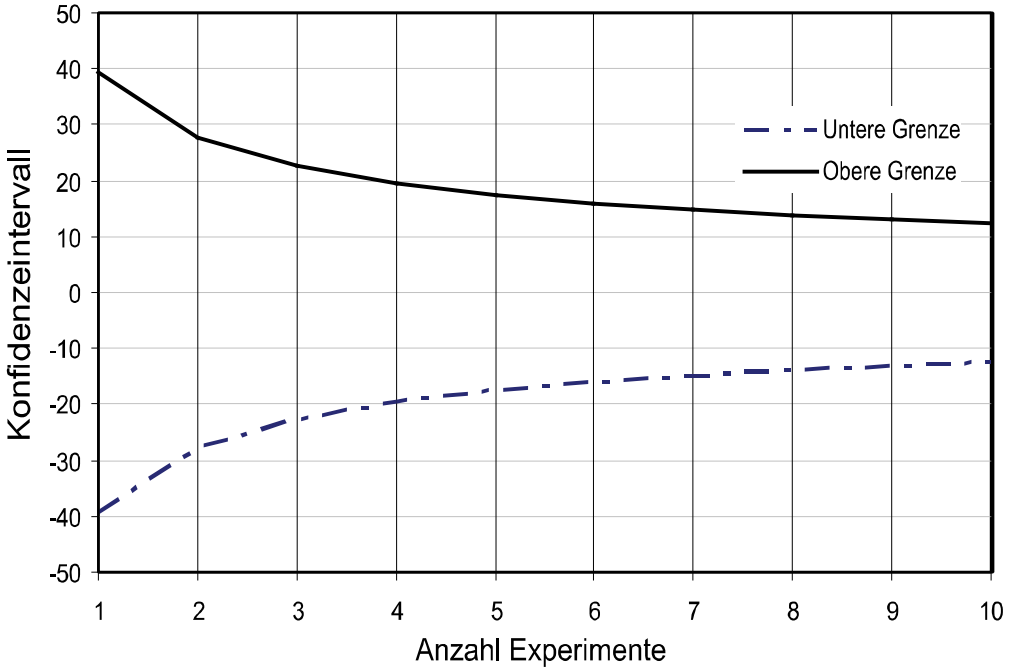
$$P\left[-9.8 < \overline{X} - \mu_X < 9.8\right] = 0.95$$

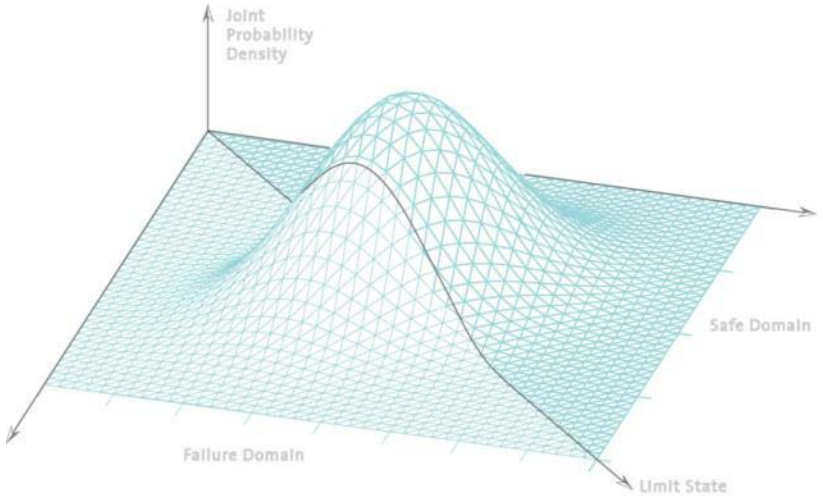
$$390.2 < \mu_X < 409.8$$

- Normalerweise werden Konfidenzintervalle für Mittelwert, Varianz und charakteristische Werte (Fraktilwerte) in Betracht gezogen.
- Das Konfidenzintervall repräsentiert / beschreibt die (statistische) Unsicherheit, welche durch zu wenig Daten entsteht.

Konfidenzintervalle für Schätzer

- Die Anzahl verfügbarer Daten hat einen signifikanten Einfluss auf das Konfidenzintervall.
- Unter Verwendung des vorherigen Beispiels ($\sigma_X = 20$) ist in der folgenden Graphik die Abhängigkeit des Konfidenzintervalls von der Anzahl der Experimente n illustriert.





Statistik und Wahrscheinlichkeitsrechnung

Prof. Dr. Michael Havbro Faber