Statistik

TEIL 4

Hans-Hermann Thulke
ba @ thulke-statistics.de
0172-3449934

Statistik

- Daten erheben, verstehen, werten
- Hypothesen prüfen
- Modellieren von Zusammenhängen

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Grundgesamtheit + Stichprobe

Wahrscheinlichkeit

Datentypen, Merkmalsskalen

Häufigkeits- & Punktediagramm

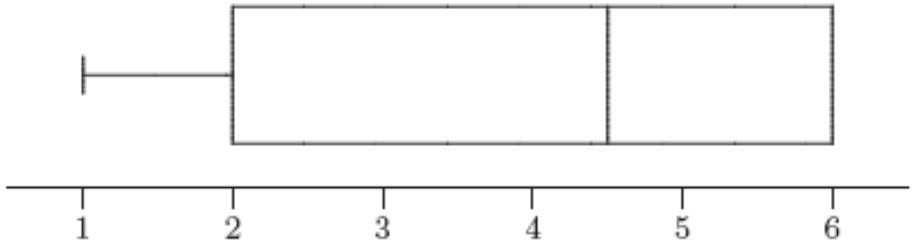
Lagemaße & Streuungsmaße

Box-Whisker-Plot

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Aufgabe

Ein Würfel wurde sechsmal geworfen, wobei genau eine Augenzahl doppelt auftrat. Sei X das Merkmal "Augenzahl bei einem Würfelwurf". Aus den $n = 6$ Beobachtungen $x_1, \ldots, x_6$ ergab sich der folgende (einfache) Boxplot:



Welche Augenzahl wurde nicht beobachtet? Erstellen Sie die geordnete Urliste $x_{(1)}, \ldots, x_{(6)}$. Bestimmen Sie außerdem den Modus x_{mod} und den Interquartilsabstand IQR.

Ein Ladenbesitzer bestellt seine Produkte bei 5 unterschiedlichen Händlern. Folgende Tabelle zeigt, wieviele Produkteinheiten (n_i) er von den einzelnen Händlern bezieht, wobei Y_i der Preis für eine Produkteinheit von Händler i bezeichnet.

Händler i	1	2	3	4	5
n_i	5	5	8	3	4
Y_i (€)	7,50	15,82	5,15	7,00	5,30

1. Zur Charakterisierung der Preise pro Produkteinheit sollen arithmetisches Mittel, empirische Varianz und Standardabweichung berechnet werden.

2. Zeichnen Sie einen einfachen Boxplot für die Preise pro Produkteinheit und leiten Sie alle dafür nötigen Größen her. Berechnen Sie auch den Interquartilsabstand.

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Neu

Überblicksbeschreibung von Datensätzen

1. Box-Whisker-Plot

2. Empirische und theoretische Verteilungen

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

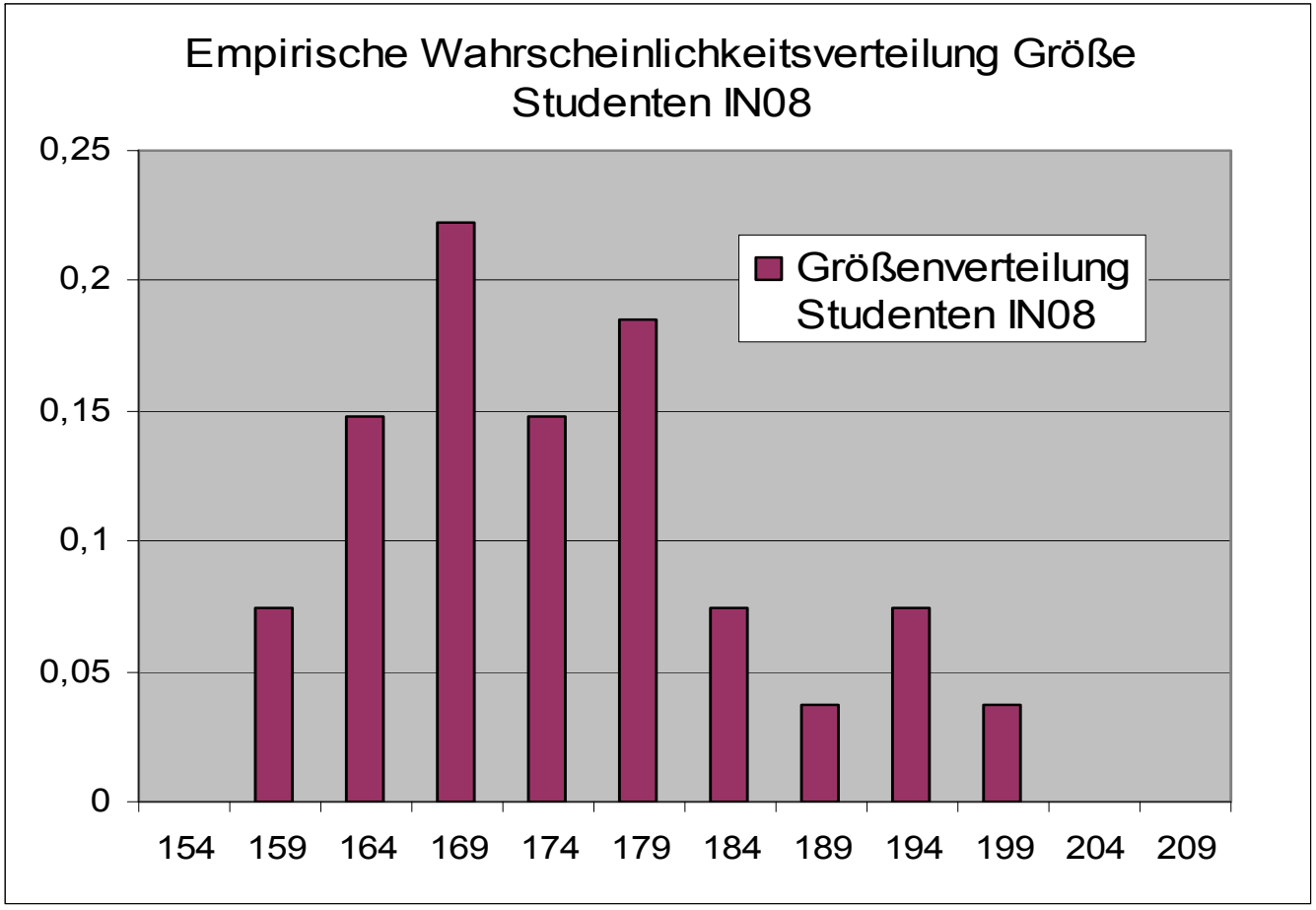
Verteilungen

1. Empirische Verteilung:

2. Theoretische Verteilung:
    1. Gleichverteilung
    2. Normalverteilung
    3. Poissonverteilung
    4. Exponentialverteilung
    5. Chi-Quadrat-Verteilung
    6. Binomialverteilung
    7. Hypergeometrische Verteilung

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

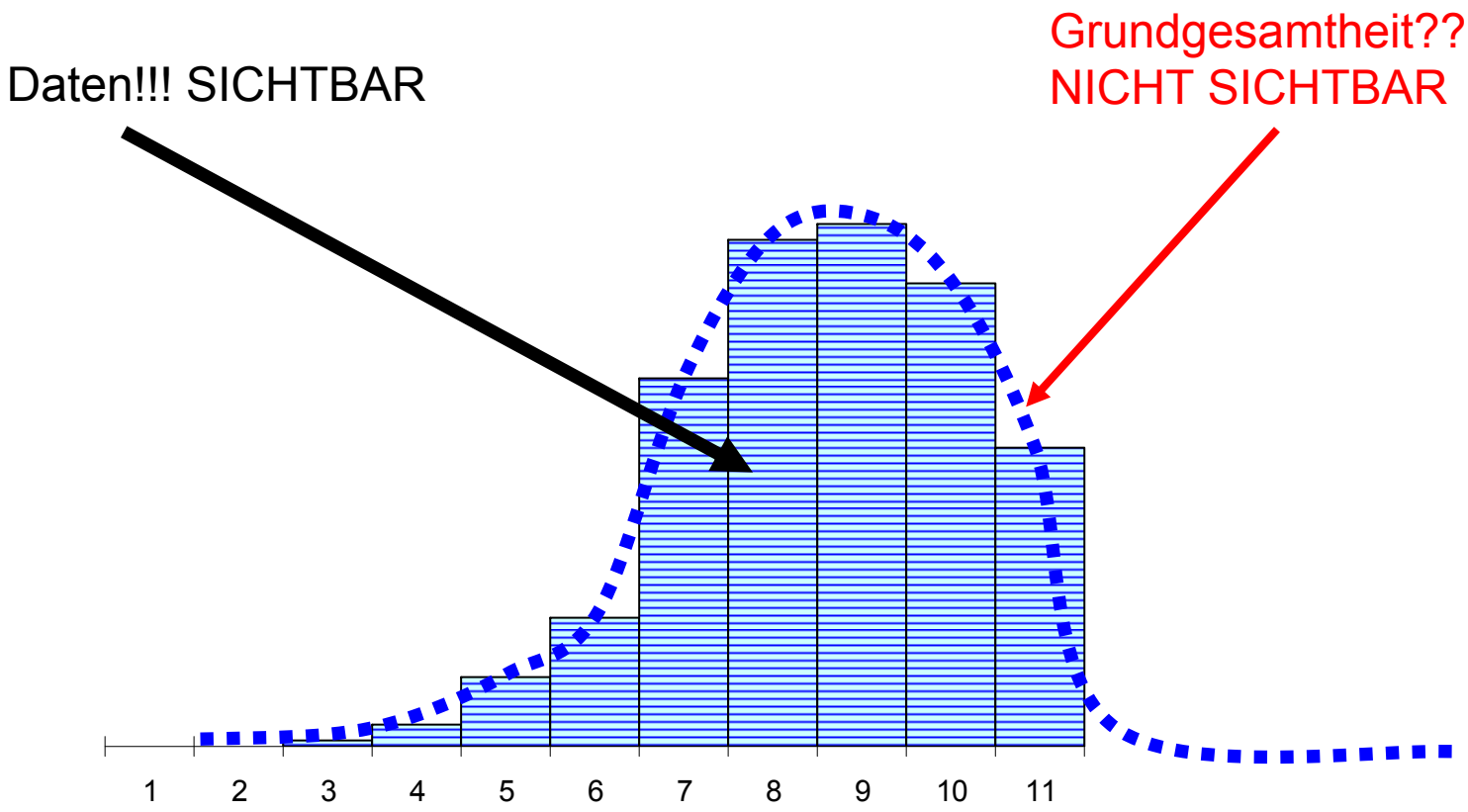
Empirische Verteilung



N = 26

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Wahrscheinlichkeitsfunktion



thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

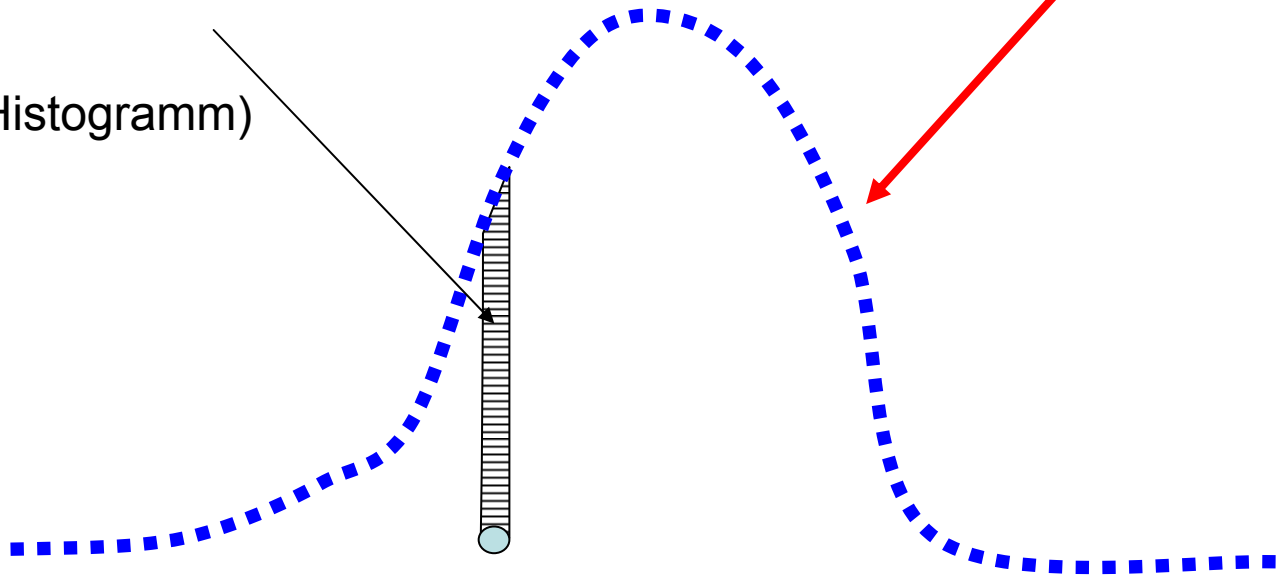
Wahrscheinlichkeitsfunktion

Wahrscheinlichkeitsfunktion

Grundgesamtheit

Wert hat eine Wahrscheinlichkeit aufzutreten

(messbar, siehe Histogramm)



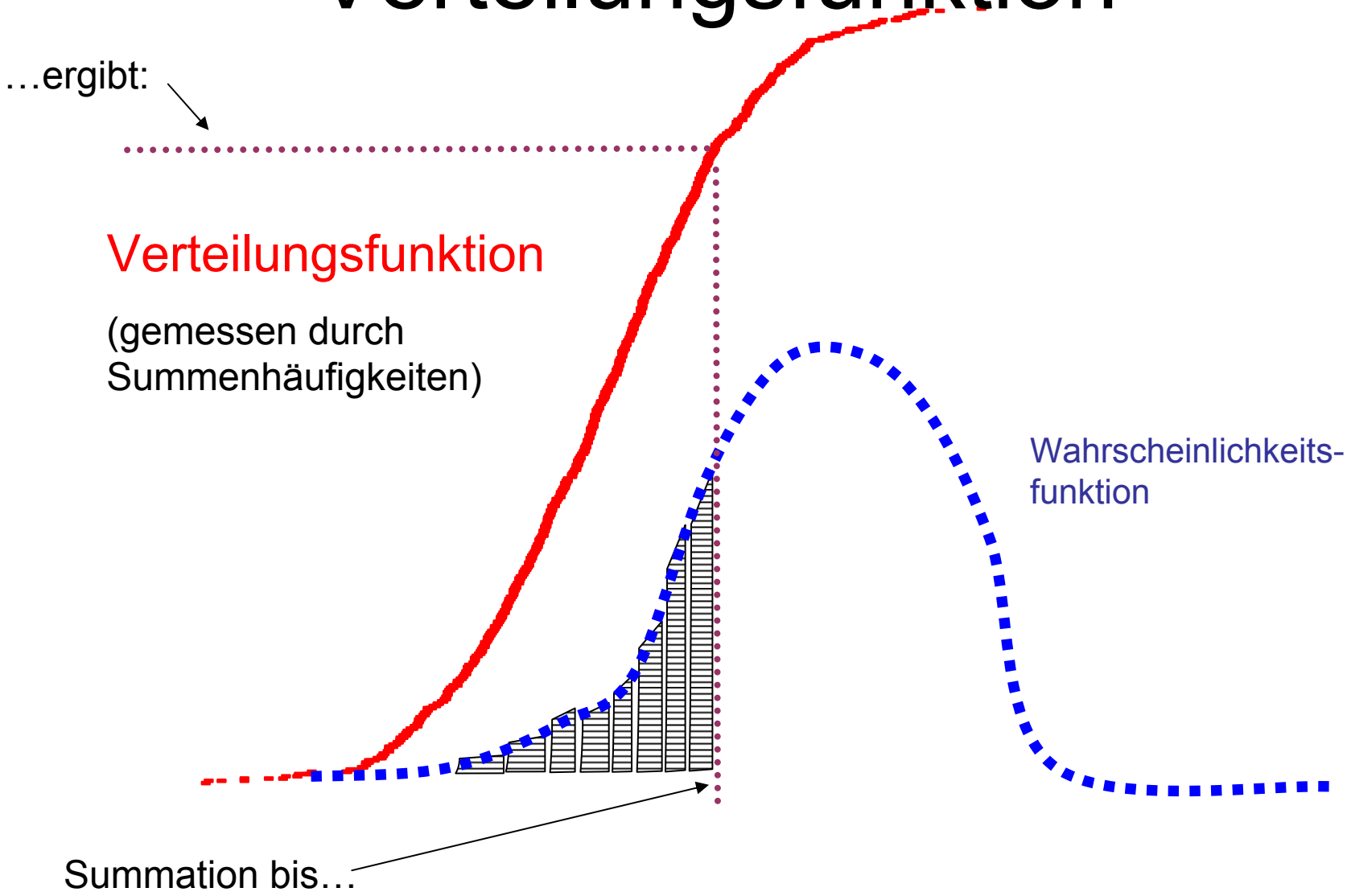
Wert z.B. 5 beim Würfeln

Interval z.B. 4,1 beim Messen

d.h. >4 & <4,2

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Verteilungsfunktion

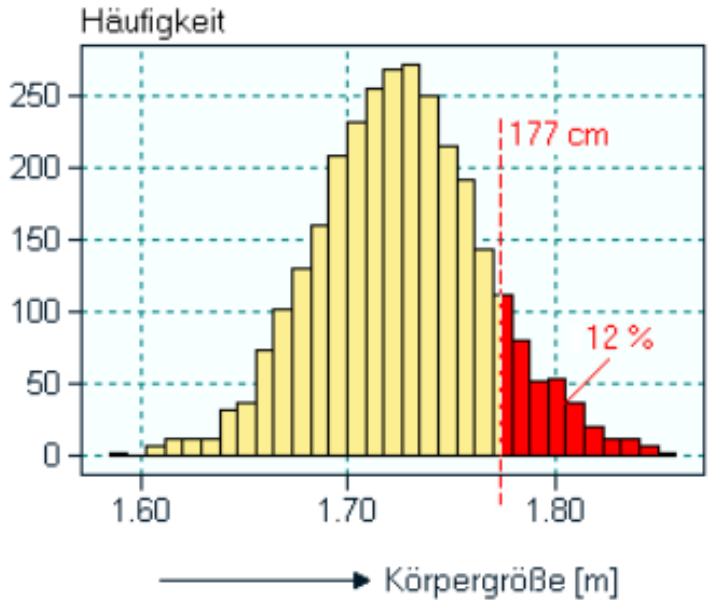


Beispiel: WS beim Würfeln weniger als 5 Augen zu bekommen
ist die Summe der WS eine 1, 2, 3 oder 4 zu erhalten.

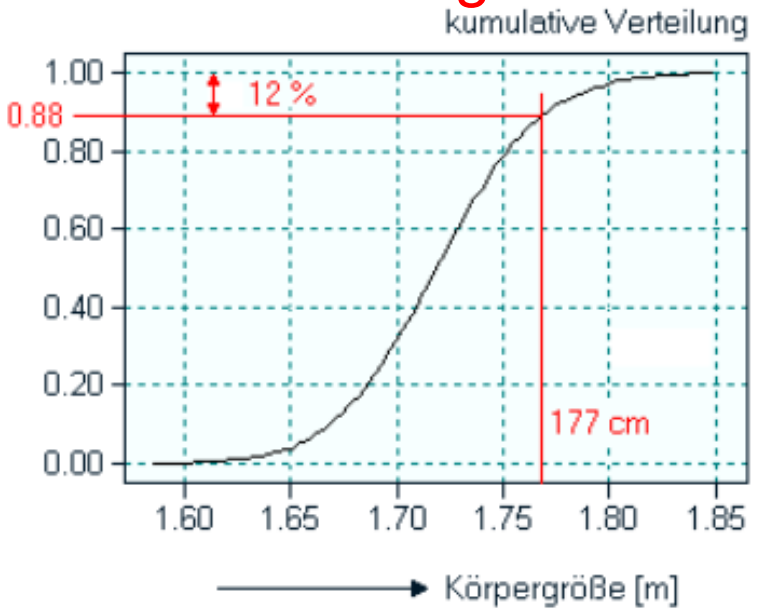
thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Verteilungsfunktion

Wahrscheinlichkeitsfunktion



Verteilungsfunktion



Während man den Prozentanteil der Frauen, die größer als 177 cm sind, durch Integration der Häufigkeitsverteilung berechnen muss (linke Abbildung, rotes Gebiet), erhält man dieselbe Zahl aus der kumulativen Häufigkeitsverteilung einfach durch Setzen eines Grenzwerts (rechte Abbildung).

Quelle: http://www.statistics4u.info/fundstat_germ/index.html

Funktion der Verteilungsfunktion

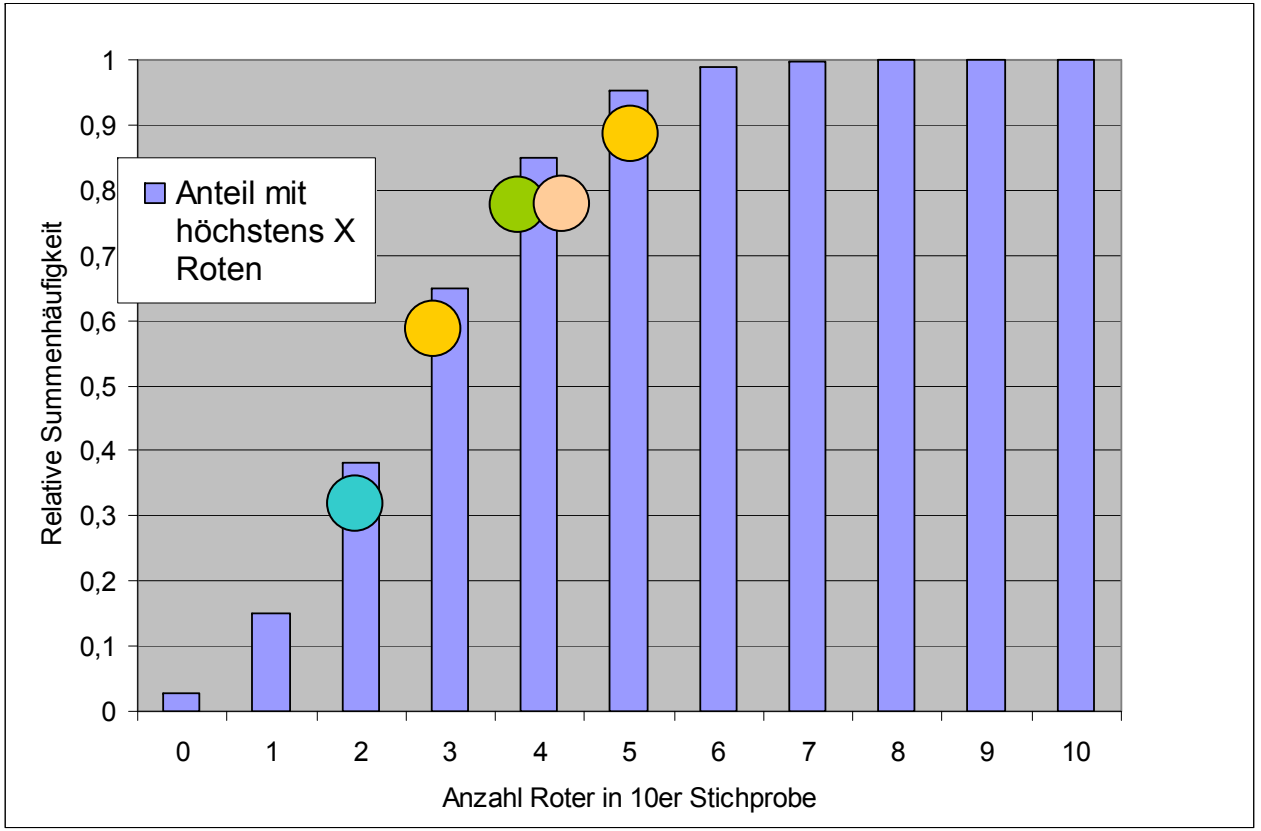
Verteilungsfunktion

WS(X≤4) = 85%

WS(3<X≤5) =
WS(X≤5) – WS(X≤3)
= 95% - 65% = 30%

WS(X>4) =
1 – WS(X≤4) =
1 - 85% = 15%

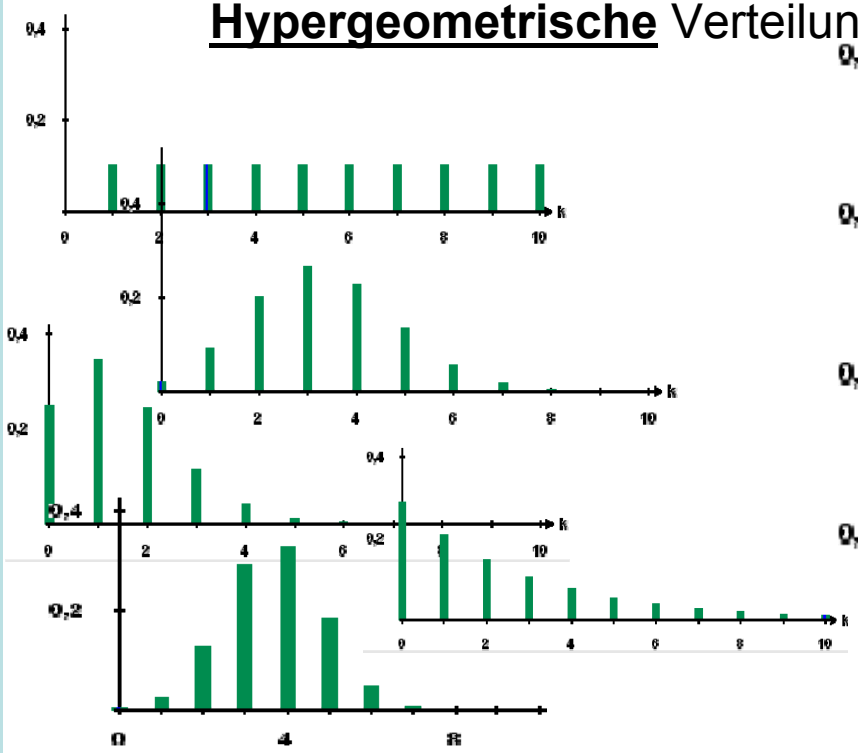
WS(X≤??) > 38% ergibt 2



Theoretische Verteilungsfunktionen

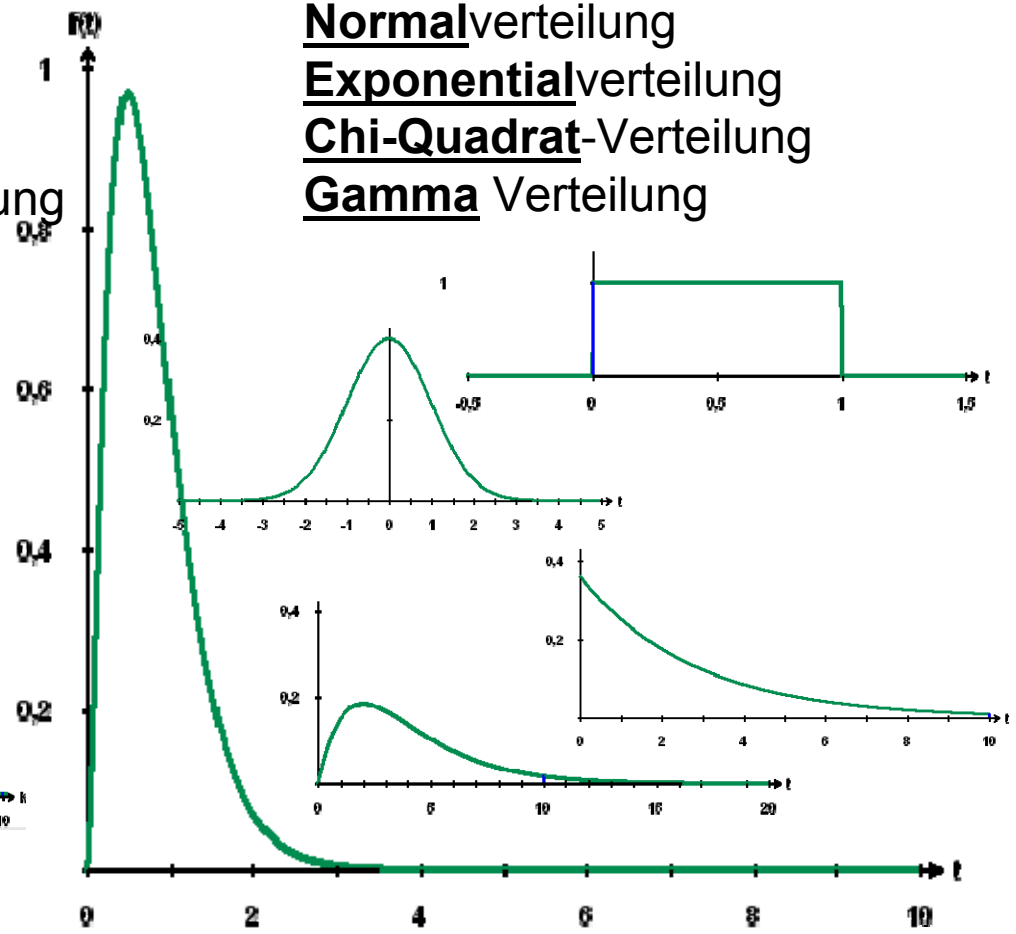
„Diskret“ – Kategoriales Merkmal

Gleichverteilung
Binomialvertailung
Poissonverteilung
Geometrische Verteilung
Hypergeometrische Verteilung



„Stetig“ – Metrisches Merkmal

Gleichverteilung
Normalverteilung
Exponentialverteilung
Chi-Quadrat-Verteilung
Gamma Verteilung



Daten erheben, verstehen, werten

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Bernoulli-Verteilung

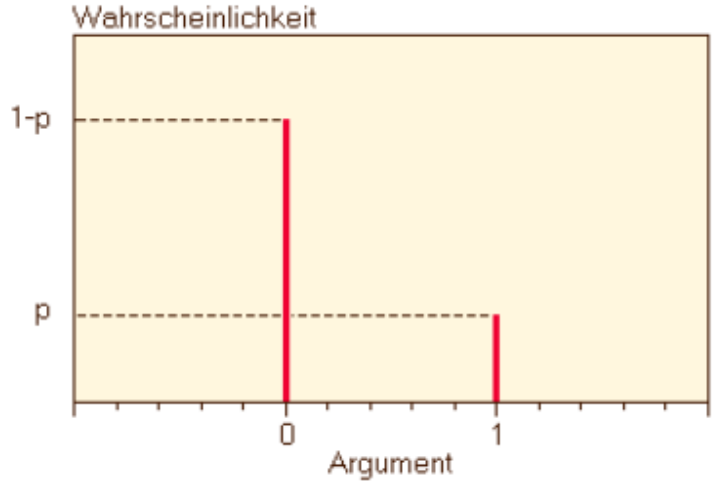
Definition

$$f(x) = \begin{cases} p & \text{für } x=1 \\ 1-p & \text{für } x=0 \\ 0 & \text{sonst} \end{cases}$$

Eine Bernoulli-Verteilung muss die folgenden Bedingungen erfüllen:

- das Ergebnis eines Experiments ist binär (nur zwei mögliche Werte)
- die N Versuche sind unabhängig voneinander
- die Wahrscheinlichkeiten der Ergebnisse bleiben konstant

Grafische Darstellung



Anwendungen Bernoulli-Verteilungen können beobachtet werden, wenn ein zufälliger Prozess exakt zwei Ergebnisse hat, wie z.B. in der Qualitätssicherung, wo ein Produkt als gut oder schlecht klassifiziert werden kann.

Erstes Moment $E(x) = p$

Zweites Moment $VAR(x) = p(1-p)$

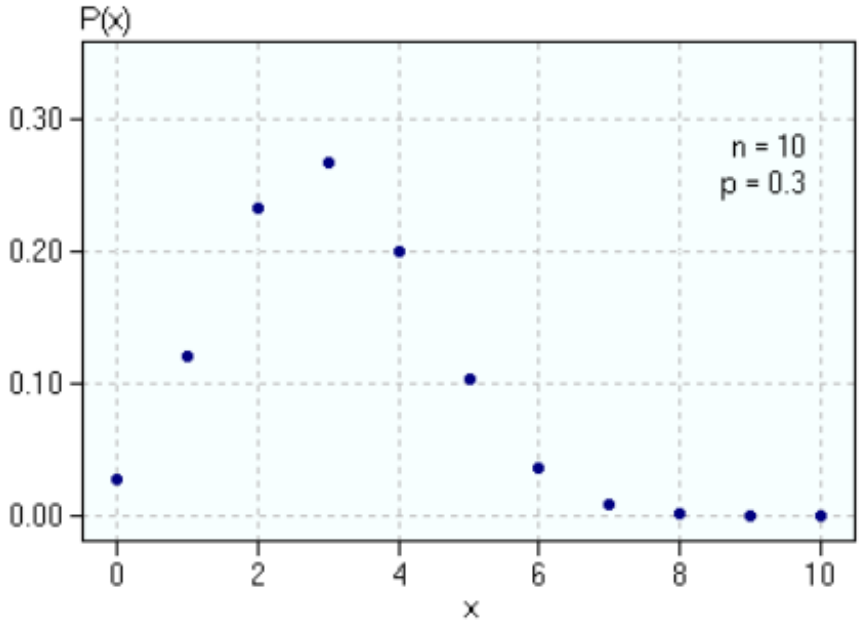
Binomialverteilung

Definition

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{für } x = 0,1,\ldots,n;\ \ 0 < p < 1$$

Die Binomialverteilung beschreibt eine diskrete Wahrscheinlichkeitsverteilung.

Grafische Darstellung



Erstes Moment $E(X) = np$

Zweites Moment $VAR(X) = np(1-p)$

Anwendungen Die Binomialverteilung beschreibt den wahrscheinlichen Ausgang einer Folge von gleichartigen Versuchen, die jeweils nur zwei mögliche Ergebnisse haben (also eine Bernoulli-Verteilung aufweisen). Wenn das gewünschte Ergebnis E eines Versuches die Wahrscheinlichkeit p besitzt, und die Zahl der Versuche n ist, dann gibt die Binomialverteilung an, mit welcher Wahrscheinlichkeit sich insgesamt x-mal das Ereignis E einstellt.

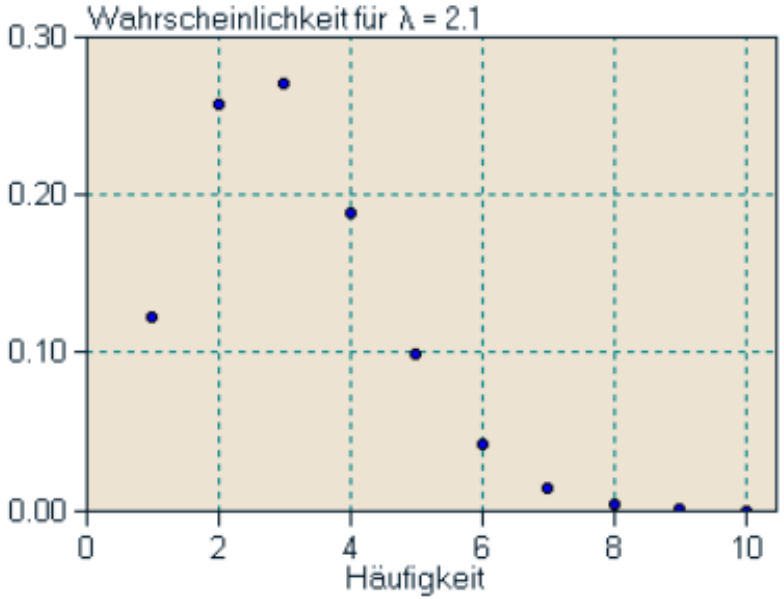
Poissonverteilung

Definition

$$P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda}, \text{ für } x = 0,1,2,\ldots$$

Die Poissonverteilung ist der Binomialverteilung sehr ähnlich. Sie wird angewendet, wenn die Wahrscheinlichkeit, dass das beobachtete Ereignis eintrifft, im Gegensatz zur Grundgesamtheit sehr klein ist.

Grafische Darstellung



Erstes Moment $m = 1$

Zweites Moment $s = 1$

Anwendungen Die Poissonverteilung ist die Verteilung der seltenen Ereignisse. Eine typische Anwendung wäre die Wahrscheinlichkeit, dass eine bestimmte Anzahl an Unfällen in einem vorgegebenen Intervall auftritt oder die Wahrscheinlichkeit eines radioaktiven Zerfalls. Die Zahl der Atome, die zerfallen können, ist im Vergleich zur Anzahl der vorhandenen Atome sehr klein.

Hypergeometrische Verteilung

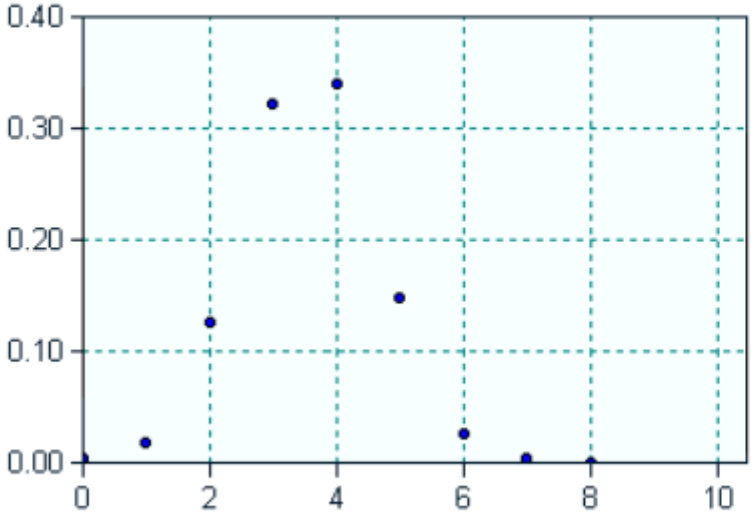
Die hypergeometrische Verteilung ist eine diskrete Verteilung und wird verwendet, um die Wahrscheinlichkeit zu beschreiben, k Beobachtungen der Klasse 1 zu finden, wenn n Stichproben aus einer Grundgesamtheit von N Objekten gezogen worden sind und die Wahrscheinlichkeit eines einzelnen Elements der Klasse 1 gleich p ist.

Definition

$$P(k) = \frac{\binom{pN}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}} \qquad pN \ldots \text{ganzzahlig}$$

N ... Anzahl der Elemente in der Grundgesamtheit
n ... Zahl an entnommenen Elemente
p ... Wahrscheinlichkeit der Klasse 1
k ... Zahl der erwarteten Elemente der Klasse 1

Grafische Darstellung



Erstes Moment $m = np$

Zweites Moment $\sigma^2 = np(1-p)\frac{N-n}{N-1}$

Anwendungen Wird oft in der Qualitätskontrolle verwendet.

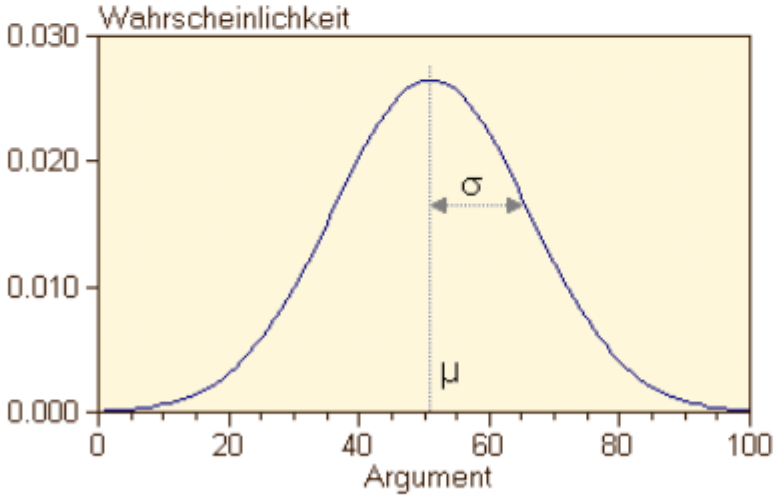
Normalverteilung

Definition

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right)$$

- Die Zufallsvariable x kann jeden Wert zwischen -unendlich und +unendlich annehmen.
- Die Verteilung ist um das erste Moment (= Mittelwert) symmetrisch.
- Der Begriff "Normalverteilung" wird oft für eine Verteilung verwendet, die wie eine Normalverteilung aussieht. Um Fehler zu vermeiden, sollte der Begriff "Standardnormalverteilung" für Normalverteilungen verwendet werden, die eine Standardabweichung von 1.0 und einen Mittelwert von 0.0 aufweisen.

Grafische Darstellung



Sie können die Form der Normalverteilung für verschiedene Standardabweichungen testen, indem Sie das folgende `interaktive Beispiel` starten.

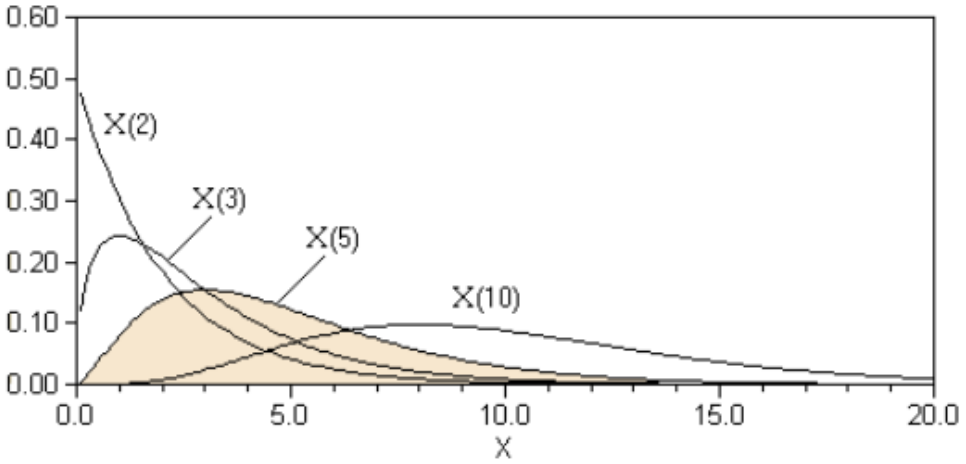
Anwendungen Eine der wichtigsten Verteilungen der Statistiktheorie, die aber nicht so häufig vorkommt, wie man erwarten würde. Die Bedeutung der Normalverteilung in der Statistik gründet sich auf dem zentralen Grenzwertsatz. Beispiele:

Chi-Quadrat-Verteilung

Um Rückschlüsse über die Verteilung der Grundgesamtheit auf der Basis der Probenvarianz zu ziehen, müssen wir eine spezielle Verteilung berücksichtigen, die Chi-Quadrat-Verteilung (χ^2): Ist eine zufällige Variable Y normal verteilt (mit dem Mittelwert μ und der Varianz s^2), dann zeigt die Größe

$$(n-1)\frac{s^2}{\sigma^2}$$

eine χ^2-Verteilung mit n-1 Freiheitsgraden für eine zufällige Stichprobe der Größe n. Einige Beispiele für die χ^2-Verteilung für verschiedene Freiheitsgrade werden in der Darstellung unten gezeigt.



Wie Sie sehen können, ist die χ^2-Verteilung asymmetrisch und immer positiv. Der Mittelwert der χ^2-Verteilung ist gleich der Anzahl der Freiheitsgrade n-1; die Varianz ist gleich der doppelten Anzahl an Freiheitsgraden. Die χ^2-Verteilung ist in statistischen Tabellen aufgelistet oder kann online mit Hilfe des Verteilungsrechners berechnet werden. Die χ^2-Verteilung wird zum Testen der Unterschiede zwischen Grundgesamtheiten und Probenvarianzen sowie zwischen theoretischen und beobachteten Verteilungen verwendet.

Eine wichtige Eigenschaft der χ^2-Verteilung ist deren Additivität: Wenn zwei unabhängige Größen χ^2-verteilt sind (mit den Freiheitsgraden f_1 und f_2), dann ist die Summe der beiden Größen χ^2-verteilt mit dem Freiheitsgrad f_1+f_2.

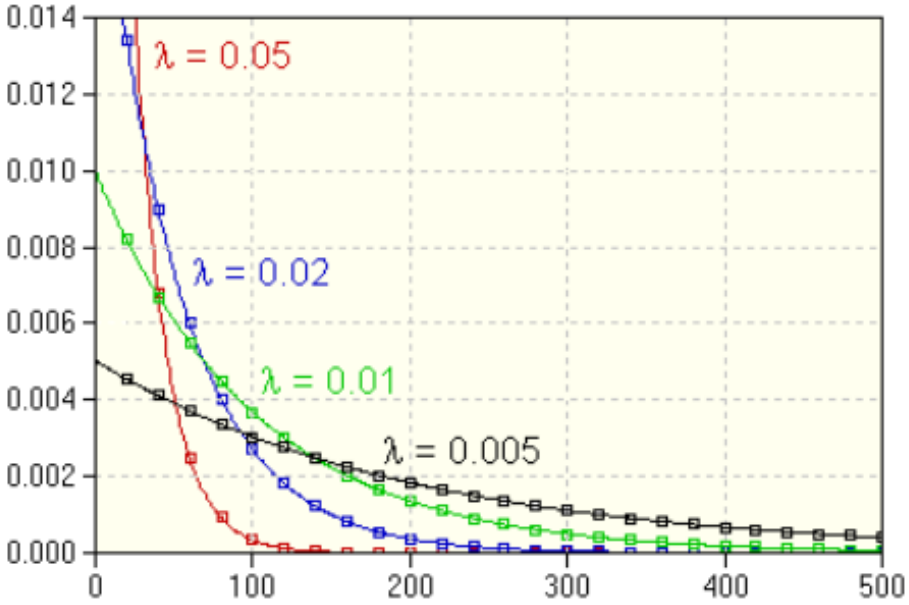
Exponentialverteilung

Definition

$f(x) = \lambda e^{-\lambda x}$

Die Exponentialverteilung hat den reellen Parameter λ, der den Charakter einer Ausfallrate besitzt. Der Kehrwert von λ ist als Lebensdauer zu verstehen. λ muss immer größer als null sein.

Grafische Darstellung



Anwendungen Die Exponentialverteilung beantwortet die Frage nach der Dauer von zufälligen Zeitintervallen, wie z.B. der Dauer von Telefongesprächen, der Zeit zwischen zwei Anrufen, oder der Lebensdauer von Bauteilen (ohne Alterungserscheinungen).

Erstes Moment $m = 1/\lambda$

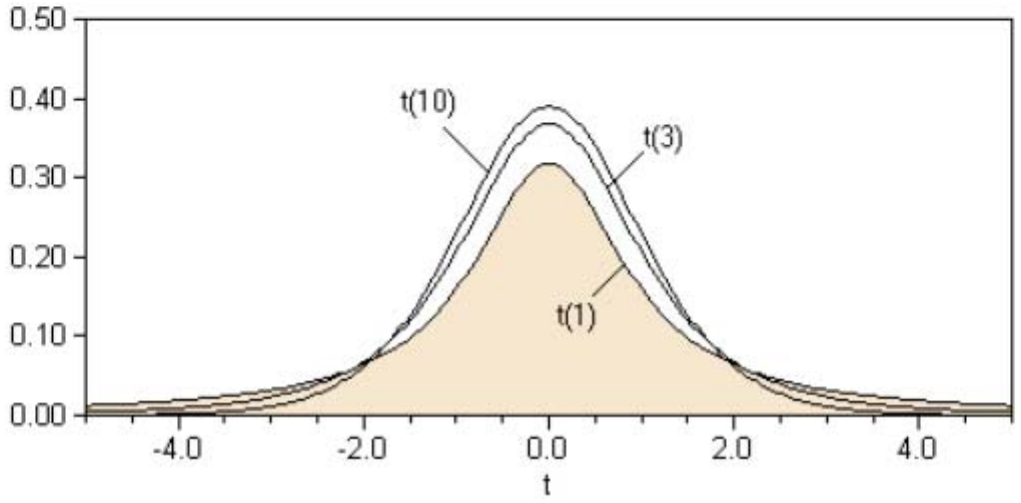
Zweites Moment $s^2 = 1/\lambda^2$

t-Verteilung

W.S. Gosset demonstrierte 1908, dass das Verhältnis **t** zwischen der Differenz des Stichprobenmittelwerts und des Populationsmittelwerts und dem Standardfehler des Mittelwerts nicht normal verteilt ist, wenn die Populationsparameter unbekannt sind:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Dieses Verhältnis folgt einer speziellen Verteilung, der so genannten t-Verteilung. Die t-Verteilung ist für eine kleine Stichprobenanzahl breiter als die Normalverteilung und nähert sich der Normalverteilung für eine große Stichprobenmenge. In der Darstellung unten sehen Sie die t-Verteilung für 1, 3 und 10 Freiheitsgrade.

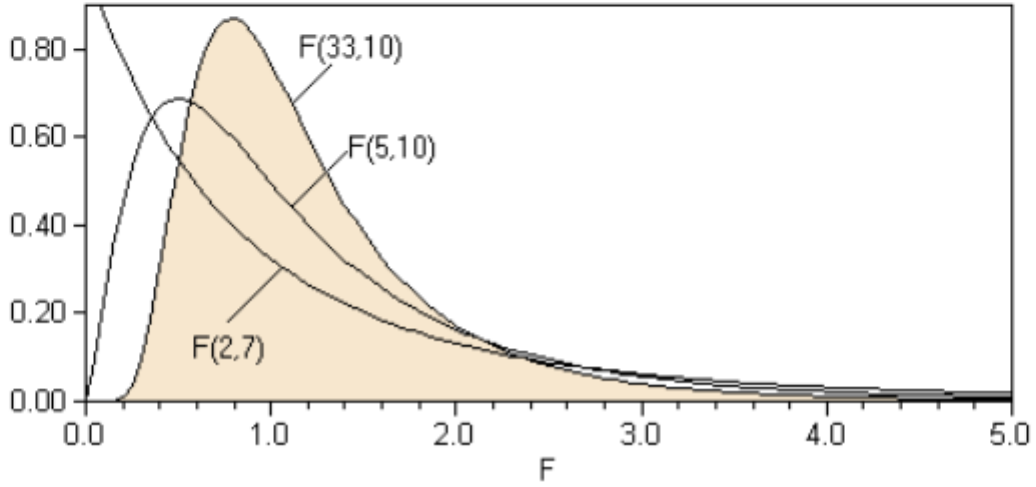


F-Verteilung

Die F-Verteilung (nach R.A. Fisher benannt) ist relevant für die Berechnung der Verhältnisse der Varianzen von normal verteilten Statistiken. Nehmen wir an, wir haben zwei Proben mit n_1 und n_2 Beobachtungen. Das Verhältnis

$$F = \frac{s_1^2}{s_2^2}$$

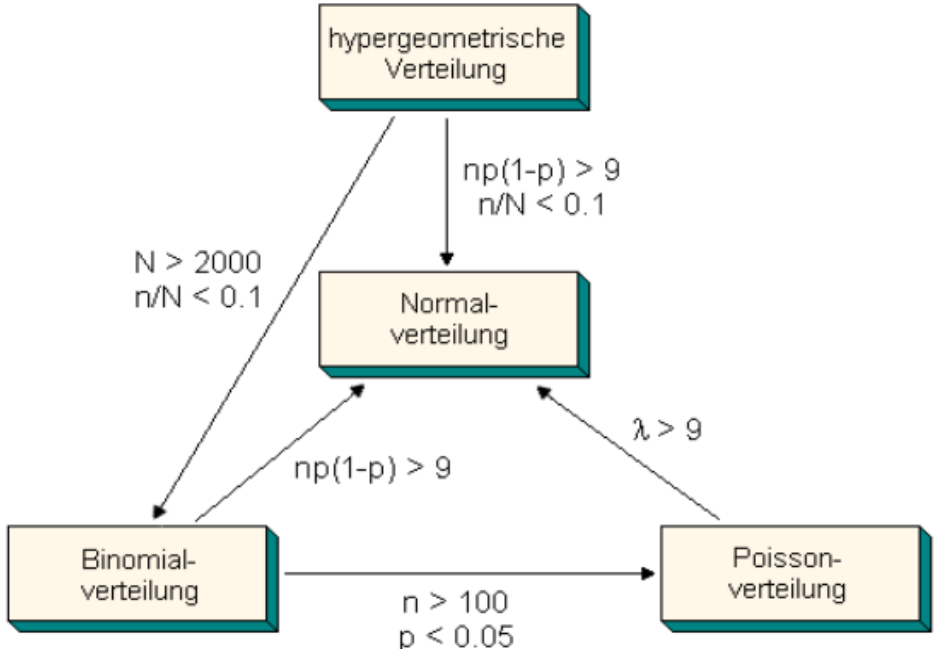
ist nach einer F-Verteilung verteilt, mit $df_1 = n_1-1$ Freiheitsgraden für den Zähler des Quotienten und mit $df_2 = n_2-1$ Freiheitsgraden für den Nenner. Die F-Verteilung ist nach rechts verschoben und die F-Werte können nur positiv sein.



Veranschaulichung

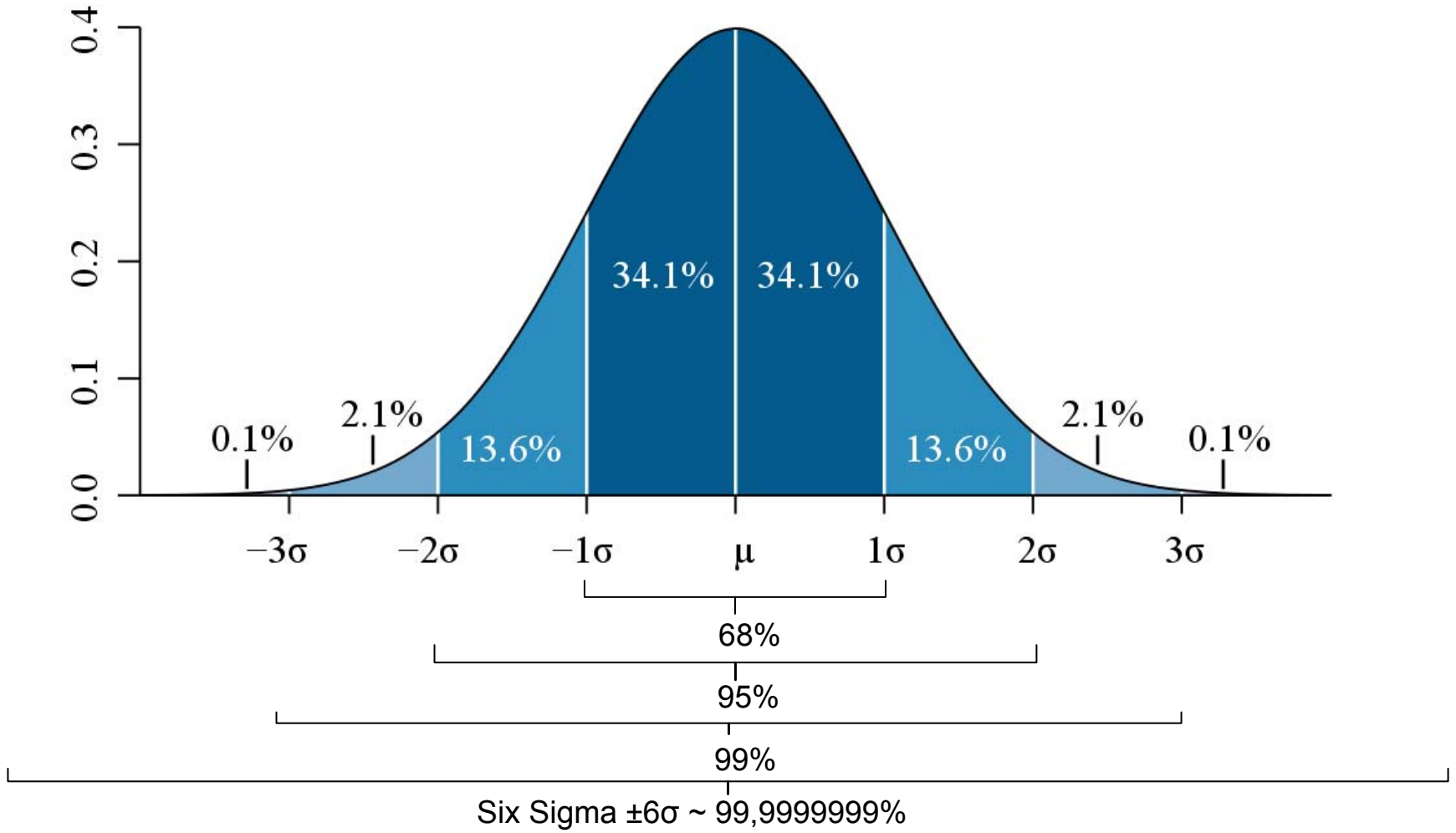
Bestehende Wahrscheinlichkeitsfunktionen bzw. Verteilungen

1. Verteilungen: http://www.uni-konstanz.de/FuF/wiwi/heiler/os/vt-index.html
2. http://www.math.tu-clausthal.de/Arbeitsgruppen/Stochastische-Optimierung/Appletsammlung/distrtable/
3. http://www.statistics4u.info/fundstat_germ/index.html



„Normal“ z.B. Fehler

Prozessgenauigkeit ±3σ entspräche:
740,000 fehlerhafter Kreditkartenbuchungen pro Tag
in den USA



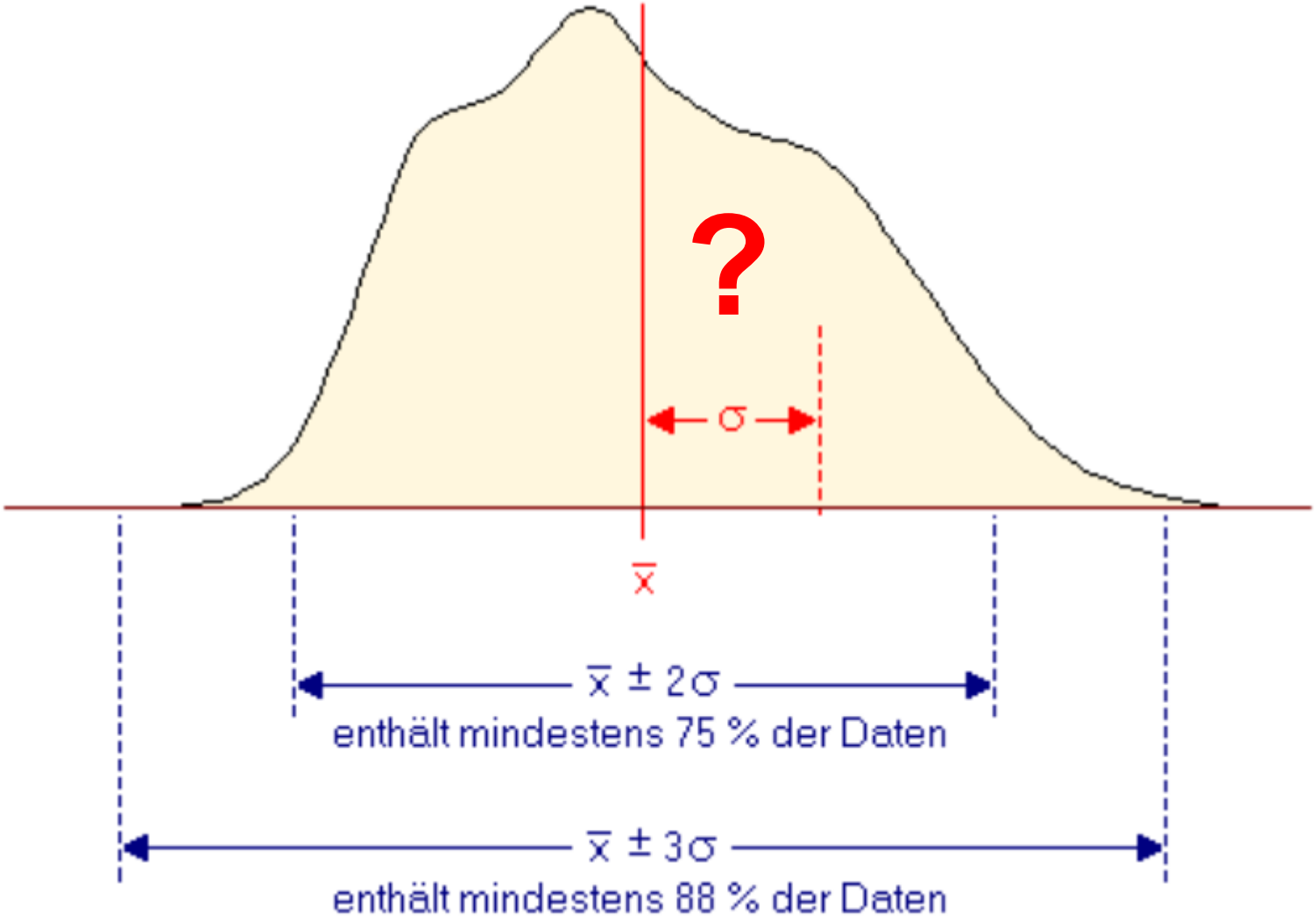
thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Theorem von Tschebyscheff
Russische Mathematiker (1821-1894)

$(1 - 1/k^2)\%$ in +/- k * σ für k > 1!!

97,2% aller Daten in +/- 6 σ

„Allgemein“



thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Grundgesamtheit + Stichprobe

Wahrscheinlichkeit

Datentypen, Merkmalsskalen

Häufigkeits- & Punktediagramm

Lagemaße & Streuungsmaße

Box-Whisker-Plot

Verteilungen

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Stichprobenverteilung !!!

Schlüssel zur schließenden Statistik

Stichprobenverteilung des Mittelwerts

1. Auch eine Verteilung!
2. Wie entsteht sie?
3. Was gewinnt man mit ihr?
4. Wofür lässt sie sich ausnutzen?

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Stichprobenverteilung des MW

1. **Auch eine Verteilung**!
2. Wie entsteht sie?
3. Was gewinnt man mit ihr?
4. Wofür lässt sie sich ausnutzen?

„Verteilung“: Erfasst wie häufig die möglichen Werte einer zufälligen Größe auftreten d.h. wie wahrscheinlich es ist einen bestimmten Wert zu beobachten.

Grundgesamtheit → zufällige Stichprobe → Mittelwert

d.h. der konkrete Wert des Stichprobenmittelwerts ist eine zufällige Größe!!!

Bei gedachter Wiederholung des Stichprobenziehens (mit zurücklegen) und jeweiliger Berechnung des Stichprobenmittelwerts, erhält man die empirische Verteilung der Werte des Stichprobenmittelwerts

die **STICHPROBENVERTEILUNG** des Mittelwerts

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Stichprobenverteilung des MW

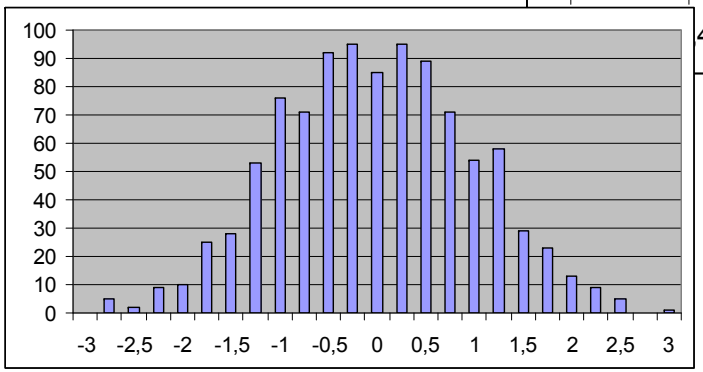
1. Auch eine Verteilung!
2. **Wie entsteht sie?**
3. Was gewinnt man mit ihr?
4. Wofür lässt sie sich ausnutzen?

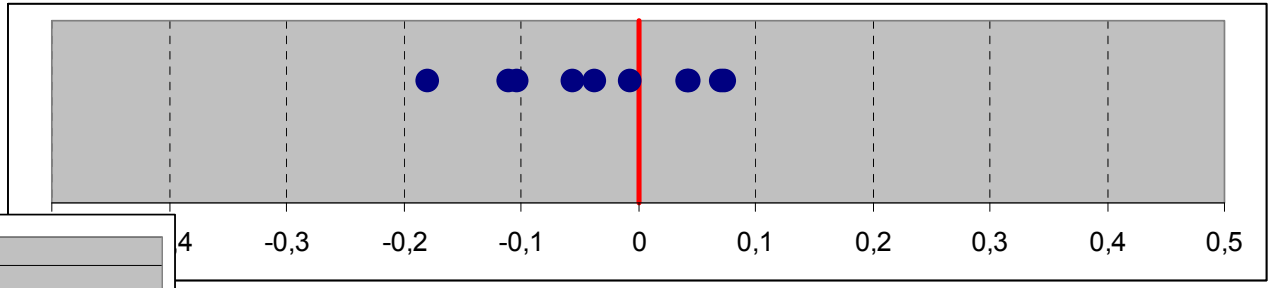
Grundgesamtheit → zufällige Stichprobe → Mittelwert

„Unendlich oft“ eine gleichgroße Stichprobe ziehen (mit zurücklegen)

Jeweils den Mittelwert errechnen.

a) Daten z.B. generiert





b) Stichproben ziehen und auswerten.

Jeder blaue Punkt entspricht dem MW einer zufälligen Stichprobe vom Umfang 100.

Der tatsächliche MW ist -0,008.

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Stichprobenverteilung des MW

Zentraler Grenzwertsatz

Für eine **Grundgesamtheit mit Varianz σ^2 und Mittelwert μ**
wird die **Verteilung der Stichprobenmittelwerte** aus n unabhängigen Werten
mit zunehmenden n (n > 30)
einer **Normalverteilung mit Varianz σ^2/n und Mittelwert μ** immer ähnlicher.

Berechnen:

Mittelwert der SPV des MW

Wie gehabt: Summe / Anzahl

Standardabweichung der SPV

des MW = „Standardfehler“

σ / WURZEL(n)

Bzw.

S / WURZEL(n)

Kleiner mit wachsendem n !

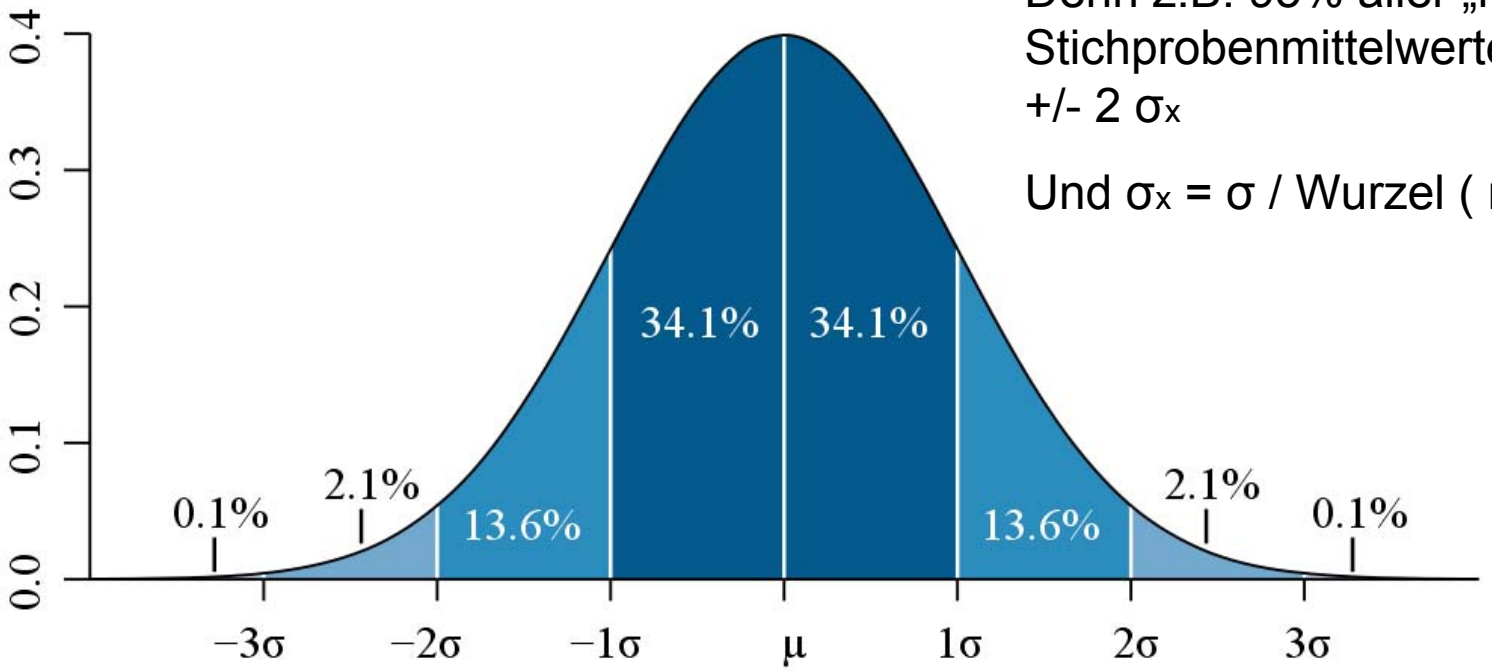
Die Standardabweichung des Stichprobenmittels wird **Standardfehler** genannt

Aus dem Zentralen Grenzwertsatz leitet sich also ab, dass bei genügend großem Stichprobenumfang, die Verteilung der Stichprobenmittelwerte vollständig bekannt ist und die Eigenschaften der Normalverteilung ausgenutzt werden können. Insbesondere nimmt die Variabilität = Unsicherheit der Berechnung mit zunehmendem Stichprobenumfang n ab.

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Stichprobenverteilung des MW

1. Auch eine Verteilung!
2. Wie entsteht sie?
3. **Was gewinnt man mit ihr?**
4. Wofür lässt sie sich ausnutzen?



Genauigkeit:

Denn z.B. 95% aller „möglichen“ Stichprobenmittelwerte liegen in +/- 2 σ_x

Und $\sigma_x = \sigma$ / Wurzel (n)

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Stichprobenverteilung

Standardfehler = Standardabweichung eines Stichprobenkennwertes z.B. MW

Mittelwert

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Andere Standardfehler

Median

$$\hat{\sigma}_{Md} = 1.25 \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

Standardabweichung

$$\hat{\sigma}_{s} = \frac{\hat{\sigma}}{\sqrt{2n}}$$

Prozentwert

$$\hat{\sigma}_{\%} = \sqrt{\frac{P \cdot Q}{n}}$$

P: Prozentsatz mit dem das Ereignis „Erfolg" eintritt.

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Grundgesamtheit + Stichprobe

Wahrscheinlichkeit

Datentypen, Merkmalsskalen

Häufigkeits- & Punktediagramm

Lagemaße & Streuungsmaße

Box-Whisker-Plot

Verteilungen

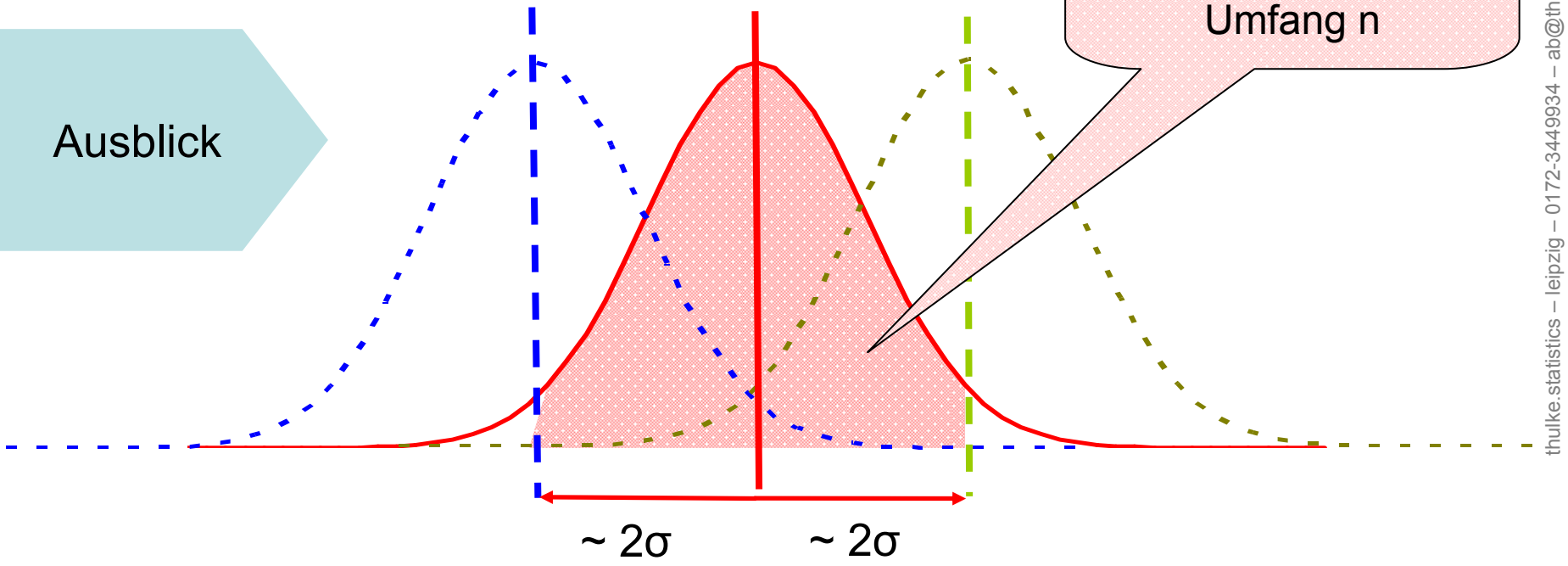
Stichprobenverteilung

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de

Stichprobenverteilung des MW

1. Auch eine Verteilung!
2. Wie entsteht sie?
3. Was gewinnt man mit ihr?
4. **Wofür lässt sie sich ausnutzen?**

Ausblick



1. 95% der Stichprobenmittelwerte zu **ROT** liegen zwischen **BLAU** und **GRÜN**.
2. Zu jedem berechnete Mittelwert zwischen **BLAU** und **GRÜN** (also in 95% aller Fälle) überdeckt der zugehörige 95% Bereich den wahren Mittelwert also **ROT**.

Grundgesamtheit + Stichprobe

Wahrscheinlichkeit

Datentypen, Merkmalsskalen

Häufigkeits- & Punktediagramm

Lagemaße & Streuungsmaße

Box-Whisker-Plot

Verteilungen

Stichprobenverteilung

Konfidenzinterval

thulke.statistics – leipzig – 0172-3449934 – ab@thulke-statistics.de