Statistik

TEIL 5

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

Verteilungen

Stichprobenverteilung

Konfidenzinterval

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

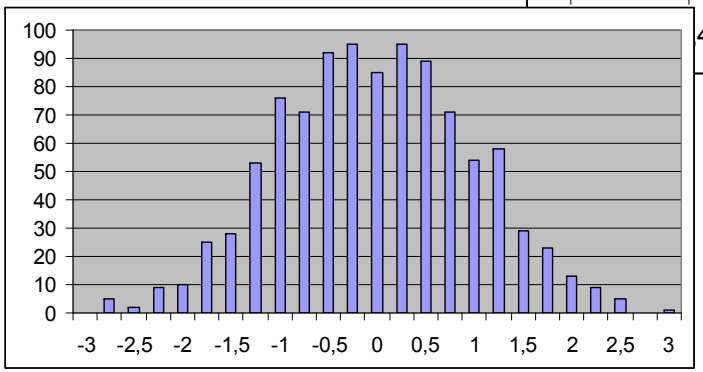
1. Auch eine Verteilung!
2. **Wie entsteht sie?**
3. Was gewinnt man mit ihr?
4. Wofür lässt sie sich ausnutzen?

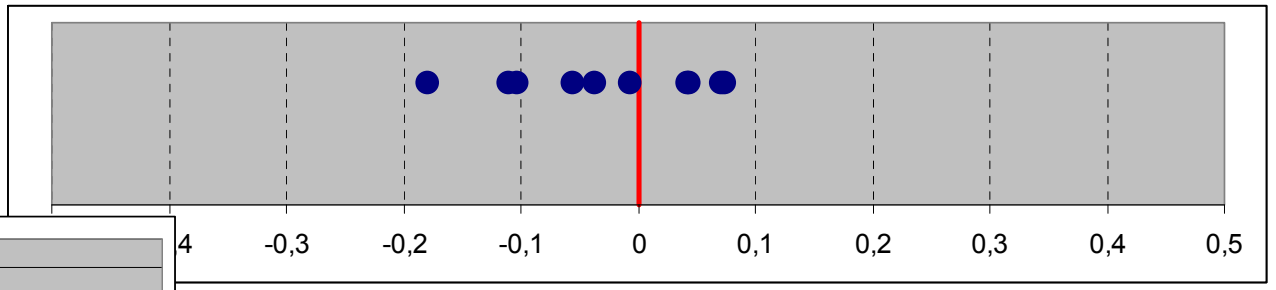
Grundgesamtheit → zufällige Stichprobe → Mittelwert

„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.

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)

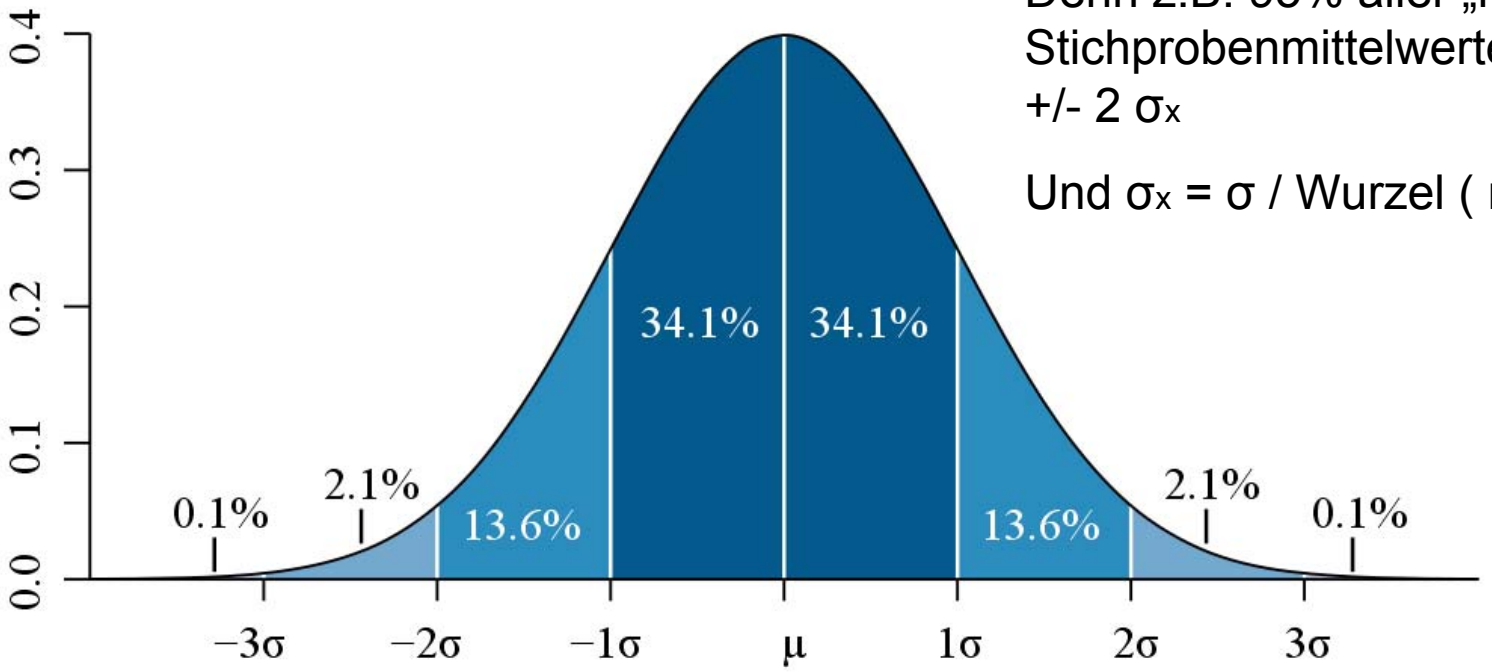
Kleiner mit wachsendem n !

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.

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

Standardabweichung des

Mittelwert

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

Standardabweichung des Merkmals in der GGS

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

Standardabweichung des Merkmals geschätzt aus einer Stichprobe

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

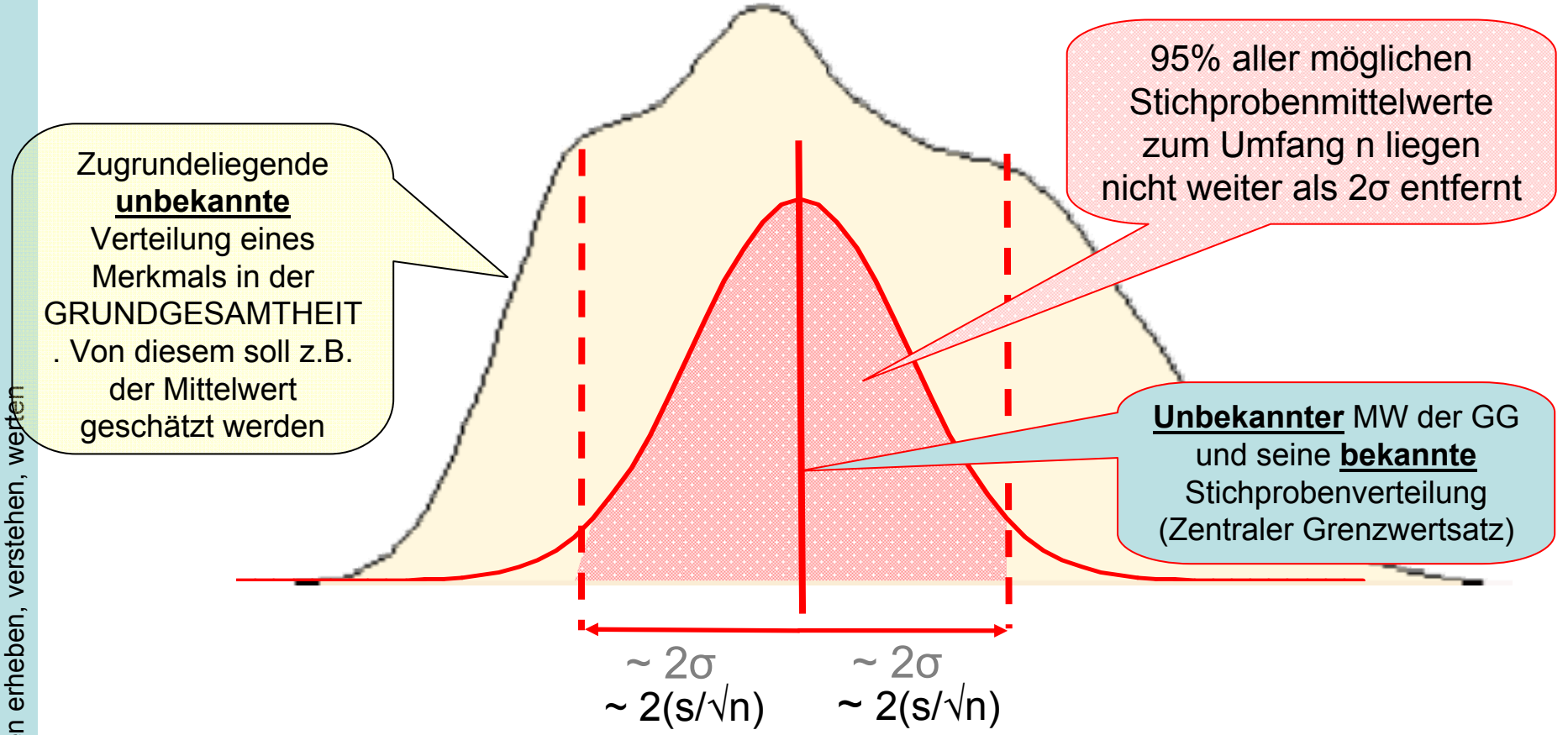
Verteilungen

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?**

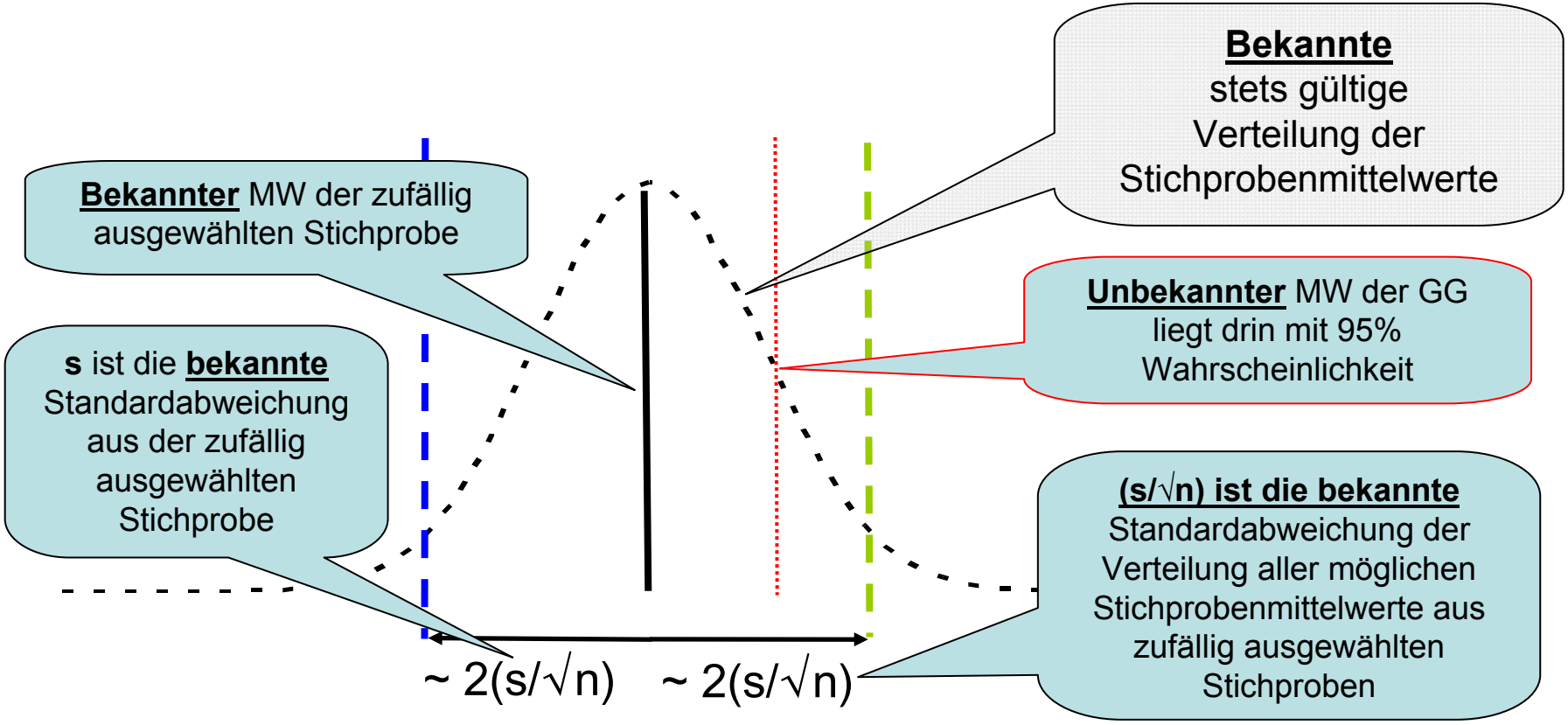


1. 95% der Stichprobenmittelwerte zu **ROT** würden zwischen den gestrichelten Linien liegen

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?**



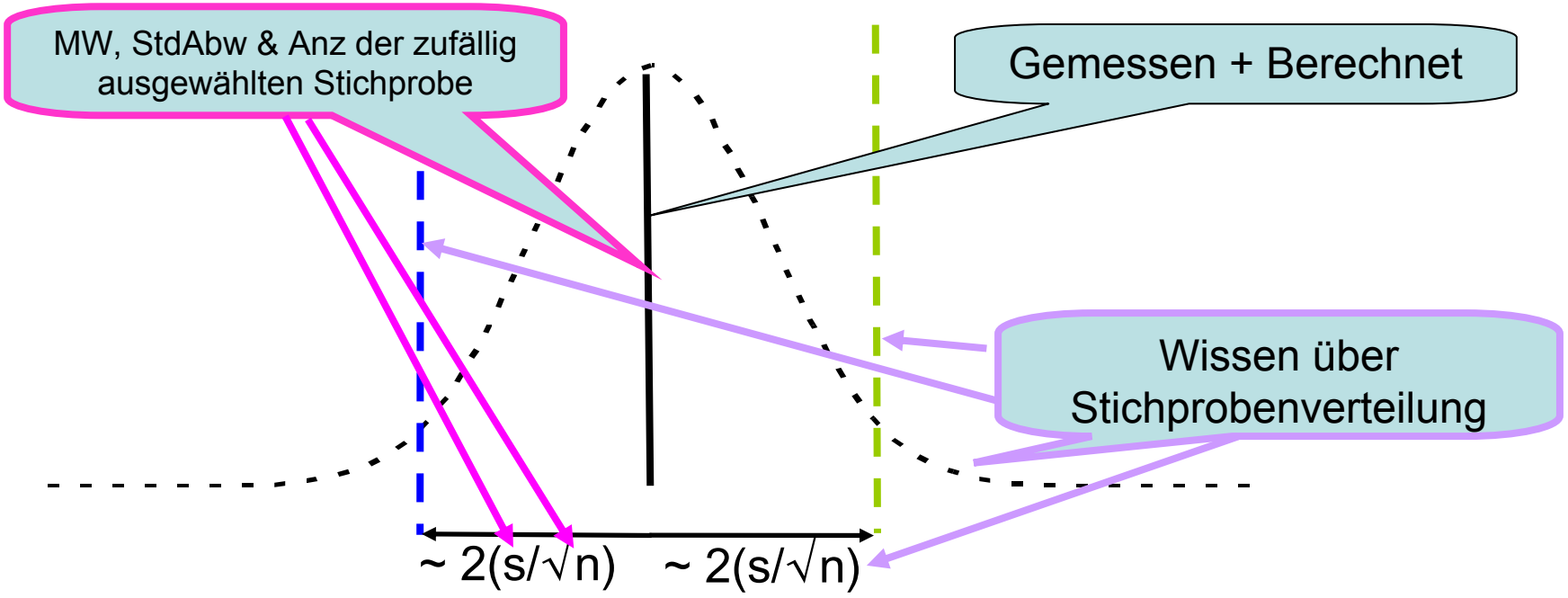
1. Zum **berechneten Mittelwert** überdeckt das zugehörige 95% Intervall zwischen **BLAU** und **GRÜN** den wahren Mittelwert, also **ROT**, mit 95% Wahrscheinlichkeit.

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?**

Was bleibt in der Praxis von den ersten Abbildungen?



1. Errechnete Stichprobenmittelwert, Standardabweichung, Anzahl
2. Das Wissen, daß mit 95% Wahrscheinlichkeit der **ROT**e Strich irgendwo zwischen **BLAU** und GRÜN liegt

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

Vertrauensbereich

Oder Konfidenzintervall eines geschätzten Parameters

Anders gesagt: Mit 95% Wahrscheinlichkeit liegt der gesuchte Mittelwert **ROT** im Konfidenzintervall oder Vertrauensbereich um das **berechnete Stichprobenmittel** zwischen der **BLAUen** und der **GRÜNen** Linie.

Das gleiche nochmal anders…

- Eine Punktschätzung (z.B. unser Stichprobenmittelwert) liefert einen „best guess“ über die GG. Wegen der Stichprobenunsicherheit kann man sich fast sicher sein, dass die Punktschätzung falsch ist (also nicht exakt **ROT** ergibt)
- Eine verbesserte Methode ist, einen Bereich anzugeben, in dem der gesuchte Parameter mit 95% Wahrscheinlichkeit enthalten ist.
- Einen solchen Bereich nennt man **Konfidenzintervall** - seine Breite ergibt sich durch die gewünschte Sicherheit also die Anzahl an Standardabweichungen z.B. $2\sigma \sim 95\%$; $3\sigma \sim 99\%$ usw.

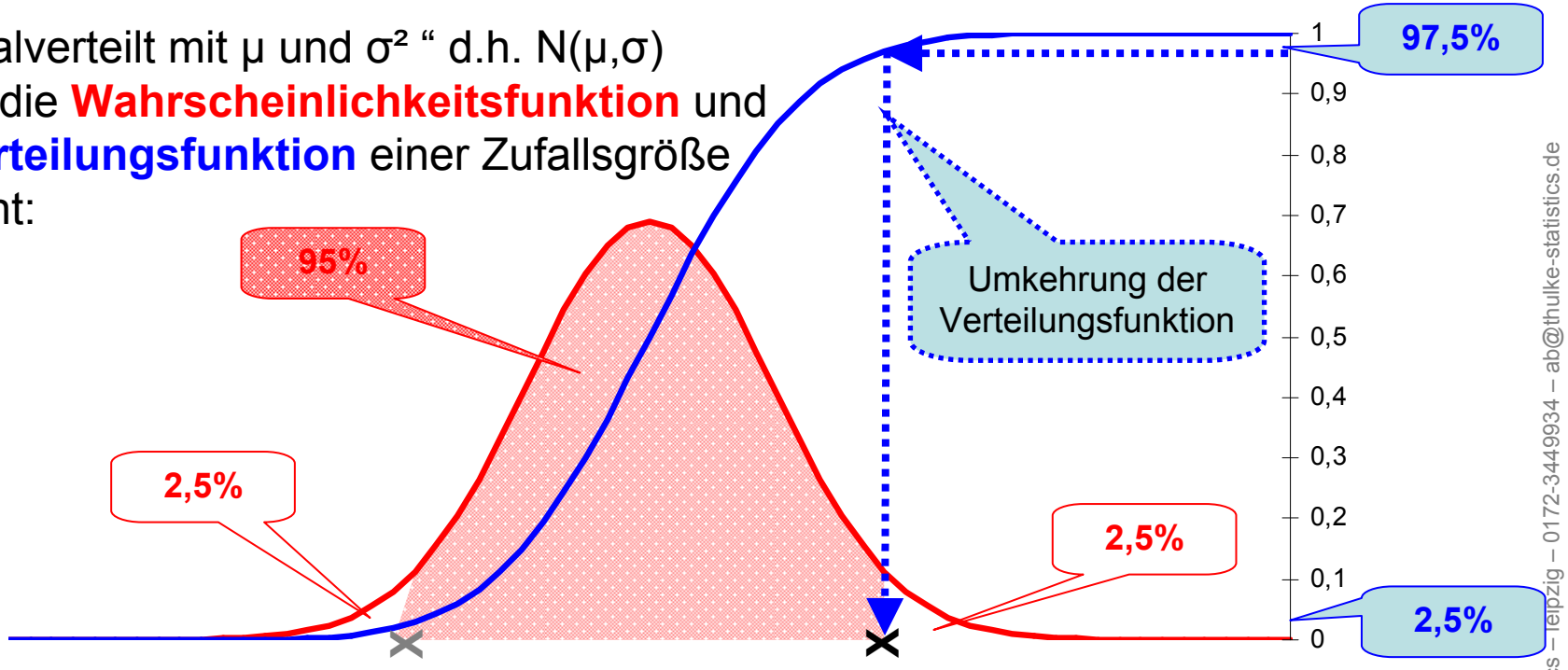
Die 95% Wahrscheinlichkeit heißt Sicherheitswahrscheinlichkeit und ihr Wert kann beliebige gewählt werden (üblich 90%, 95%, 99%).

Mit (100 - 95)% Wahrscheinlichkeit wird der tatsächliche Wert nicht enthalten sein – die 5% heißen daher Irrtumswahrscheinlichkeit.

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

Stichprobenverteilung des MW

„Normalverteilt mit μ und σ² “ d.h. N(μ,σ) macht die **Wahrscheinlichkeitsfunktion** und die **Verteilungsfunktion** einer Zufallsgröße bekannt:



Welche Werte der X-Achse grenzen den ±2σ Bereich (rot schraffiert) ein?

Bekannt, er soll 95% der gesamten Fläche unter der Wahrscheinlichkeitskurve überdecken d.h. rechts und links müssten 2,5% Fläche angeschnitten werden.

Verteilungsfunktion beschreibt die Fläche.

X_{oben} = 1,95996 für N(0,1)

Bzw. für alle anderen

1,96*σ für N(0,σ)

z.B. Für Stichprobenverteilung MW:

$1{,}96 * \frac{\sigma}{\sqrt{n}}$ wegen $N(0, \frac{\sigma}{\sqrt{n}})$ ZGWS

Vertrauensbereich

Konfidenzintervall eines geschätzten Parameters

Damit hängt die Breite des Konfidenzintervalls von 3 Faktoren ab

1. von der **Sicherheit** 1-α, mit der man den Parameter einfangen will (z.B. 95%)
2. von der **Varianz** in der Grundgesamtheit σ^2 und
3. von der **Stichprobengröße** n

(2. und 3. ergeben den Standardfehler)

Z(97,5%) :: **Wert auf der X-Achse**, bei dem die kumulierte Standardnormalverteilungsfunktion bei 97,5% ist.
Oder auch – z(2,5%) der **(negative) Wert auf der X-Achse**, bei dem die kumulierte Standardnormalverteilungsfunktion bei 2,5% ist.

$$\left[\bar{x} - z\left(1-\frac{\alpha}{2}\right)\cdot\frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} + z\left(1-\frac{\alpha}{2}\right)\cdot\frac{\sigma}{\sqrt{n}}\right]$$

Varianz GG bekannt

$$\left[\bar{x} - t\left(1-\frac{\alpha}{2}, n-1\right)\frac{s}{\sqrt{n}}\ ;\ \bar{x} + t\left(1-\frac{\alpha}{2}, n-1\right)\frac{s}{\sqrt{n}}\right]$$

Varianz SP berechnet

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

Beispiel 1

Aus Langzeiterfahrungen ist bekannt, das die Kraftstoffverbrauchsmessung bei Schwerlasttransportern technisch bedingt mit einer Standardabweichung von 1,5 Litern/100km variiert. Berechnen Sie aus einer Messreihe von 25 Werten mit dem errechneten Durchschnittswert von 22,3 Litern/100km das symmetrische 95% Konfidenzintervall für den zu erwartenden mittleren Verbrauch

$$\left[\bar{x} - z\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + z\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Sicherheit 95%

$z_{0,975} = 1,96$

$21,7 \leq \mu \leq 22,89$

Das Intervall schließt mit 95%iger Wahrscheinlichkeit den zu erwartenden Durchschnittsverbrauch ein.

Bei bekannter Varianz der GG ist die SPV des Mittelwerts normalverteilt, daher z

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

Beispiel 2

Ein Unternehmen möchte flächendeckend auf dem Markt ein neues Spülmittel einführen. Um die Käuferakzeptanz auszuloten, wird in einem Supermarkt dieses Produkt mit hohem Werbeaufwand platziert. Es soll mit dieser Aktion der durchschnittliche tägliche Absatz in einem Supermarkt dieser Größe geschätzt werden. Man definiert nun den täglichen Absatz als Zufallsvariable X [Stück] mit den unbekannten Parametern Erwartungswert μ und der Varianz σ^2. Man geht auf Grund langjähriger Beobachtungen hier davon aus, dass X annähernd normalverteilt ist. Die Marktforschungsabteilung hat einen Konfidenzkoeffizienten (=Sicherheitswahrscheinlichkeit) von 0,95 als ausreichend erachtet. Es wird nun 16 Tage lang der tägliche Absatz erfasst. Es hat sich beispielsweise ergeben

$$\left[\bar{x} - t(1-\tfrac{\alpha}{2}, n-1)\frac{s}{\sqrt{n}} \; ; \; \bar{x} + t(1-\tfrac{\alpha}{2}, n-1)\frac{s}{\sqrt{n}}\right]$$

Bei unbekannter Varianz der GG ist die SPV des Mittelwerts t-verteilt, daher t (für n>100 kein Unterschied)

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

Absatz x	110	112	106	90	96	118	108	114	107	90	85	84	113	105	90	104

16 Messwerte

Bei normalverteilter Grundgesamtheit mit unbekannter Varianz wird das Konfidenzintervall für den Erwartungswert angegeben als

$$\left[\bar{x} - t\left(1-\frac{\alpha}{2};n-1\right)\frac{s}{\sqrt{n}}\ ;\ \bar{x} + t\left(1-\frac{\alpha}{2};n-1\right)\frac{s}{\sqrt{n}}\right]$$

Es ist

$$\bar{x} = \frac{1}{16}\cdot(110+112+\cdots+104) = \frac{1}{16}\cdot 1632 = 102$$

und

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum(x_i-\bar{x})^2 \\ &= \frac{1}{15}\left((110-102)^2+(112-102)^2+\cdots+(104-102)^2\right) \\ &= \frac{1}{15}\cdot 1856 = 123{,}73 \end{aligned}$$

Es ist das (1-α/2)-Quantil der t-Verteilung mit 15 Freiheitsgraden

$$t\left(1-\frac{\alpha}{2};n-1\right) = t\,(0{,}975;15) = 2{,}13$$

Das 95 %-Konfidenzintervall berechnet sich dann als

$$\left[102-2{,}13\frac{\sqrt{123{,}73}}{\sqrt{16}};102+2{,}13\frac{\sqrt{123{,}73}}{\sqrt{16}}\right] = [102-5{,}92;\,102+5{,}92] = [96{,}08;\,107{,}92]$$

Beispiel 3

Ermitteln Sie den Anteil roter Steine im Beutel und schätzen Sie das Konfidenzintervall.

$$\left[\hat{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},\ \hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right]$$

Für n groß nähert sich die Anteilsverteilung der Normalverteilung, daher z.

Für kleine n gibt es exakte Rechenvorschriften zur Intervallbestimmung, F Verteilung

Das Merkmal ist Binomialverteilt X~B(π;n);

Laut Übersicht: μ = n*π und σ² = n*π*(1-π)

FRAGE: Warum z-Quantil, obwohl die Varianz doch aus der Stichprobe berechnet wurde?

Nach schätzen von π (Mittelwert) ist bei dieser Verteilung die Varianz bekannt!!!

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

Grundgesamtheit + Stichprobe

Wahrscheinlichkeit

Datentypen, Merkmalsskalen

Häufigkeits- & Punktediagramm

Lagemaße & Streuungsmaße

Box-Whisker-Plot

Verteilungen

Stichprobenverteilung

Konfidenzinterval

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