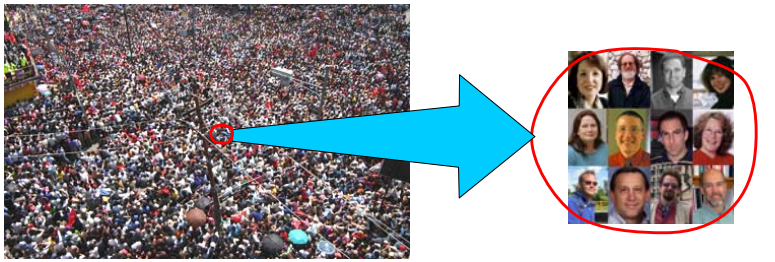
Grundlagen der Biostatistik und Informatik

Statistische Schätzungen,
Konfidenz, Signifikanz

dr. László Smeller

Analytische Statistik



Population

N = „unendlich”

Stichprobe

n = endlich

Theoretische Verteilung
Erwartungswert
Theoretische Streuung

?

Häufigkeitsverteilung
Durchschnitt
Standardabweichung

Repräsentativität der Stichprobe

Quotenstichprobe

Die gleiche Verteilung für wichtige (bekannte) Variablen in der Stichprobe wie in der Grundgesamtheit (z. B. Alter, Geschlecht,...)

Problem:

Sozio-demographische und psychologische, usw. Merkmale sind oft nicht bekannt (Verteilung und Relevanz).

Zufallsstichprobe

Zufällig ausgewählte Elemente der Grundgesamtheit

Problem mit der Repräsentativität, z.B.: 50:50 Männer und Frauen kommt mit 8% Wahrscheinlichkeit vor. (siehe Binomiale Verteilung)

Aufgabe der Schätztheorie

Aus einer Stichprobe Schätzwerte für

- Wahrscheinlichkeit
- Erwartungswert
- Streuung
- oder andere Parametern

einer Verteilung zu ermitteln.

Typen der Schätzungen:

- ***Punktschätzung***
- ***Intervallschätzung***

Punktschätzungen

Wir wollen jetzt die Parameter einer Verteilung (μ,σ) aus den konkreten Werten x_1,...x_n einer Stichprobe „möglichst gut" bestimmen, d.h. einen „Näherungswert" errechnen.

Kriterien:

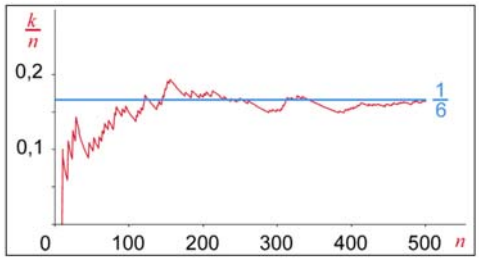
Erwartungstreue (unverzerrt)	Erwartungswert der Schätzwerte = zu schätzender Parameter
Konsistenz	n ↑ bessere Schätzung
Effizienz (wirksam)	kleine Streuung
Exhaustivität (erschöpfend)	berücksichtigt alle Informationen

Punktschätzungen

Der Parameter wird mit einem Wert geschätzt.

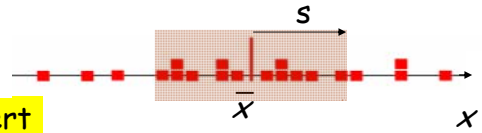
Relative Häufigkeit
ist ein Schätzwert für die Wahrscheinlichkeit

Siehe Definition der statistischen Wahrscheinlichkeit!

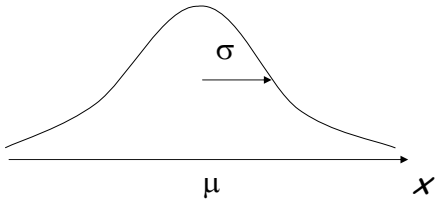


Punktschätzungen

Durchschnitt
ist ein Schätzwert für den Erwartungswert



Standardabweichung
ist ein Schätzwert für die Streuung



Punktschätzungen sagen
***nichts* über die Genauigkeit bzw. Sicherheit** der Schätzung !

Intervallschätzungen

Intervallschätzung oder Konfidenzschätzung* gibt zu einer vorgewählten Sicherheitswahrscheinlichkeit γ, (Konfidenzniveau) ein Intervall (c_1,c_2) an, in dem der unbekannte Parameter (zB. μ oder σ) mit einer Wahrscheinlichkeit von mindestens γ liegt.

Zb.: Erwartungswert der Pulszahl ist bei 95% Konfidenzniveau: 74±6 $^1/_{Min}$

*Konfidenz (Latein): Vertraulichkeit

Intervallschätzungen

Wie großes γ (Konfidenzniveau) soll gewählt werden?
Wie groß sind die Schaden bei einer falschen Schätzung?

Sozialwissenschaft γ=0,9

Medizin γ=0,95

Technik γ=0,99

Einfluss der Streuung und des Stichprobenumfanges

α=1-γ Irrtumswahrscheinlichkeit (Signifikanzniveau)

Konfidenzintervall für den Erwartungswert bei bekannter Streuung

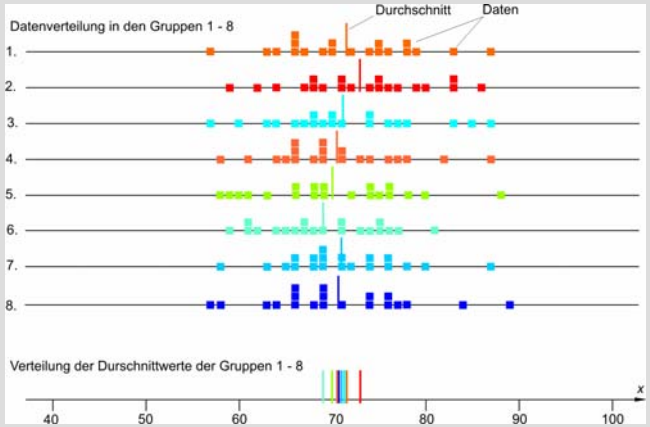
1. Nehmen wir an, dass die Varianz (und damit die theoretische Streuung) bekannt ist.

Gedankenversuch:

Sei *x* eine Zufallsgröße (zB: Pulszahl) mit einer beliebigen Verteilung mit einem Erwartungswert μ und einer Streuung σ.
Nehmen wir jetzt viele Stichproben (zB: viele Studentengruppen), alle mit gleichem Stichprobenumfang *n*.
Sei $\bar{x}_i$ der Durchschnitt der *i*-ten Stichprobe
Wie sieht die Verteilung von $\bar{x}_i$ Werten aus?

Zentraler Grenzwertsatz: bei genügend hohem *n* ist die Verteilung eine Normalverteilung.
Lage ($\mu_{\bar{x}}$) und Breite ($\sigma_{\bar{x}}$) der Verteilung der Durchschnittwerte?

Zur Erinnerung:



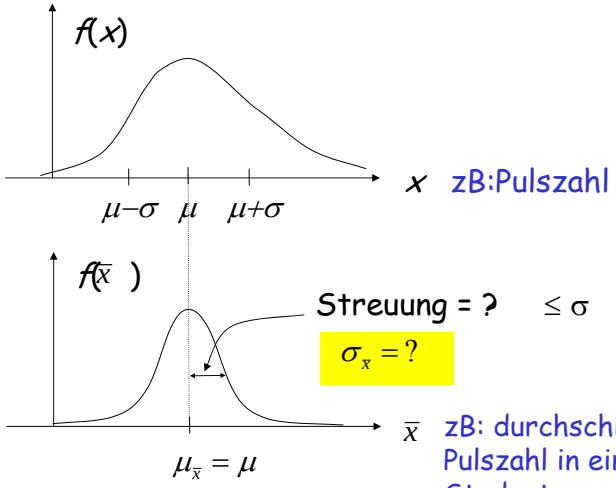
Daten und ihre Durchschnitts-werte

die Daten streuen um den Durchschnittswert

Pulsfrequenzen (1/Min)

die Durchscnhittswerte streuen um den Erwartungswert

Konfidenzintervall für den Erwartungswert



Verteilung von Durchschnitt der Zufallsgrößen

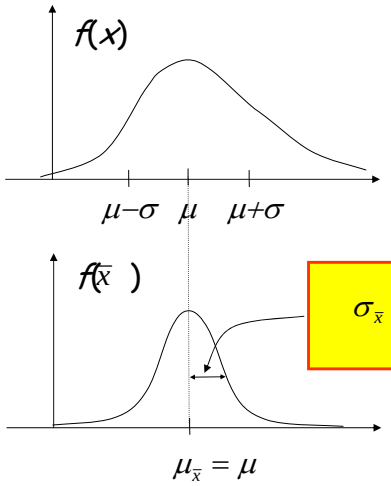
Sei x_1 und x_2 sind unabhängige Zufallsgrößen. Beide folgen einer Normalverteilung mit derselben Erwartungswerten μ und Streuungen σ.

Welche Verteilung folgt der Durchschnitt $x = (x_1 + x_2)/2$?

Normalverteilung, mit den folgenden Parametern:

Messwerte	Summe	Durchschnitt	Allgemein für n Messwerte
x_1 , x_2	x_1+x_2	$x = (x_1 + x_2)/2$	$x = (x_1 + \ldots + x_n)/n$
μ	$\mu+\mu=2\mu$	μ	μ
σ^2	$\sigma^2+\sigma^2=2\sigma^2$		
σ	$\sqrt{2}\,\sigma$	$\sigma/\sqrt{2}$	$\sigma/\sqrt{n}$

Konfidenzintervall für den Erwartungswert

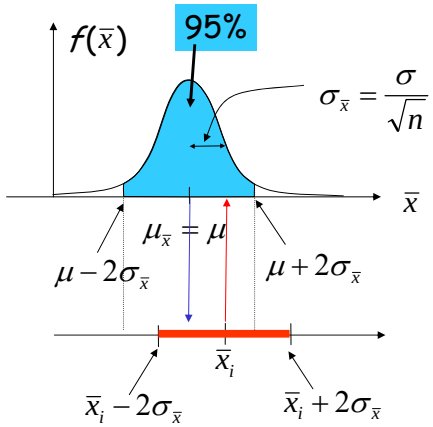


$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

zB:Pulszahl

zB: durchschnittliche Pulszahl in einer Studentengruppe von *n* Studenten

Konfidenzintervall für den Erwartungswert



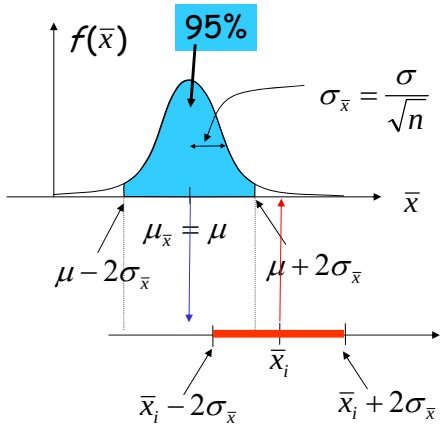
$\bar{x}_i$ liegt mit 95% Wahrscheinlichkeit im Intervall

$\mu - 2\sigma_{\bar{x}} \quad \mu + 2\sigma_{\bar{x}}$

wenn $\mu - 2\sigma_{\bar{x}} \le \bar{x}_i \le \mu + 2\sigma_{\bar{x}}$ dann $\bar{x}_i - 2\sigma_{\bar{x}} \le \mu \le \bar{x}_i + 2\sigma_{\bar{x}}$

95% Wahrsch. 95% Wahrsch.

Konfidenzintervall für den Erwartungswert



$\bar{x}_i$ liegt mit 95% Wahrscheinlichkeit im Intervall

$\mu - 2\sigma_{\bar{x}} \quad \mu + 2\sigma_{\bar{x}}$

d.h.

$\mu + 2\frac{\sigma}{\sqrt{n}} \quad \mu - 2\frac{\sigma}{\sqrt{n}}$

$\bar{x}_i \le \mu - 2\sigma_{\bar{x}}$ oder $\mu + 2\sigma_{\bar{x}} \le \bar{x}_i \Rightarrow \mu \le \bar{x}_i - 2\sigma_{\bar{x}}$ oder $\bar{x}_i + 2\sigma_{\bar{x}} \le \mu$

5% Wahrsch. 5% Wahrsch.

Konfidenzintervall für den Erwartungswert

Wenn die theoretische Streuung bekannt ist, dann kann das Intervall (Konfidenzintervall) $\bar{x}-2\sigma_{\bar{x}},\ \bar{x}+2\sigma_{\bar{x}}$ angegeben werden, in dem der Erwartungswert (μ) mit 95% Wahrscheinlichkeit liegt.

Eine ähnliche Herleitung gibt: μ ist
-mit 68% Wahrscheinlichkeit im Intervall: $\bar{x}-\sigma_{\bar{x}},\ \bar{x}+\sigma_{\bar{x}}$

-- mit 99,7% Wahrscheinlichkeit im Intervall:

$$\bar{x}-3\sigma_{\bar{x}},\ \bar{x}+3\sigma_{\bar{x}}$$

Je größer die ist Sicherheitswahrscheinlichkeit, desto breiter ist das Konfidenzintervall!

Konfidenzintervall für den Erwartungswert

zB: Eine Maschine herstellt Tabletten, je mit einem vorgeschriebenen Wirkstoffgehalt von 20 mg. Der Wirkstoffsgehalt von 10 Tabletten wurde gemessen. Der Durchschnitt beträgt 18,9 mg. Aus einer früheren Messung ist es bekannt, dass die Streuung des Wirkstoffsgehalts 1,6 mg ist. Geben Sie das zur 95% Sicherheitswahrscheinlichkeit gehörende Konfidenzintervall an! (17,9 mg 19,9 mg)
Ist diese Maschine gut eingestellt?

Mit einer sehr langen Mess-Serie haben wir der Erwartungswert und die theoretische Streuung der Blutzuckerkonzentration bestimmt.
Jetzt wird die Blutzuckerkonzentration in 40 Studentengruppen bestimmt. Wir bestimmen das 95% Konfidenzintervall für jede Gruppe. Wieviele Konfidenzintervalle enthalten den Erwartungswert?

Konfidenzintervall für den Erwartungswert bei unbekannter Streuung

2. Am häufigsten ist $\sigma_{\bar{x}}$ nicht bekannt.
   Wie kann man das Konfidenzintervall berechnen?

   Weil die Standardabweichung eine Punktchätzung der Streuung ist:

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

Standardfehler

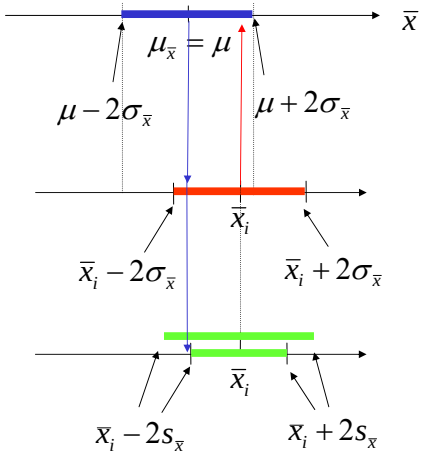
μ liegt mit 95% Wahrscheinlichkeit im $\bar{x}-2s_{\bar{x}},\ \bar{x}+2s_{\bar{x}}$ Bereich.

$\bar{x}\pm 2s_{\bar{x}}$ ist als das zu 95% Konfidenzniveau (Wahrsch.) gehörende Konfidenzintervall genannt.

Bemerkung: wenn n$\to\infty$ dann $s_{\bar{x}}\to 0$

Konfidenzintervall für den Erwartungswert

Liegt μ tatsächlich mit 95% Wahrsch. im $\bar{x}_i-2s_{\bar{x}},\ \bar{x}_i+2s_{\bar{x}}$ Bereich?



Die Ungenauigkeit erhöht sich mit der Schätzung der Streuung.

Die Schätzung ist besonders grob, wenn der Stichprobenumfang (n) klein ist.
⇨Das Konfidenzintervall muss vergrößert werden!

Konfidenzintervall für den Erwartungswert

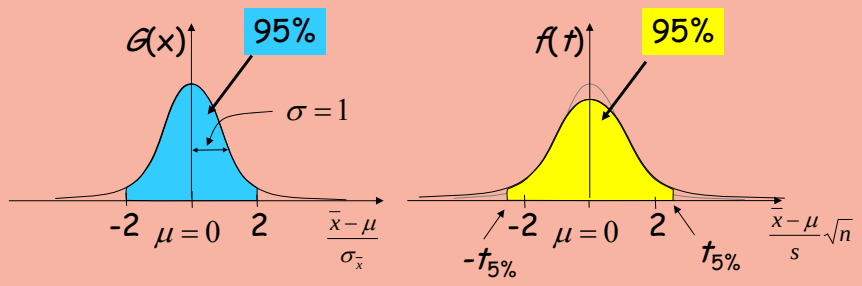
Bei bekanntem σ folgt

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

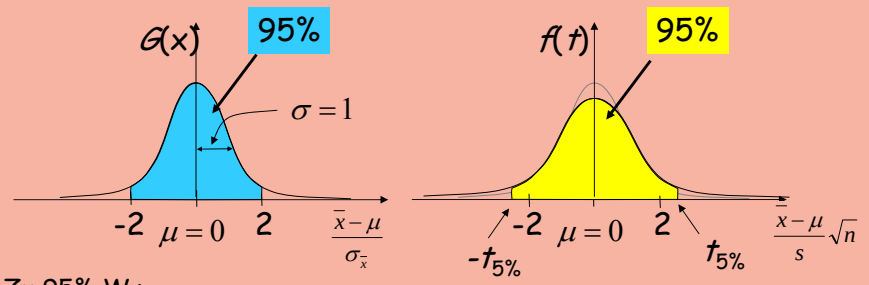
eine Standard-Normalverteilung:

Bei unbekanntem σ aber bekanntem s

folgt $\frac{\bar{x}-\mu}{s}\sqrt{n}$ eine ***t***-Verteilung:



Konfidenzintervall für den Erwartungswert



Zu 95% W.:

$$-2 < \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} < 2$$

$$-2 < \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \qquad \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} < 2$$

$$\mu < \bar{x} + 2\sigma_{\bar{x}} \qquad \mu > \bar{x} - 2\sigma_{\bar{x}}$$

$$\bar{x} + 2\sigma_{\bar{x}} < \mu < \bar{x} + 2\sigma_{\bar{x}}$$

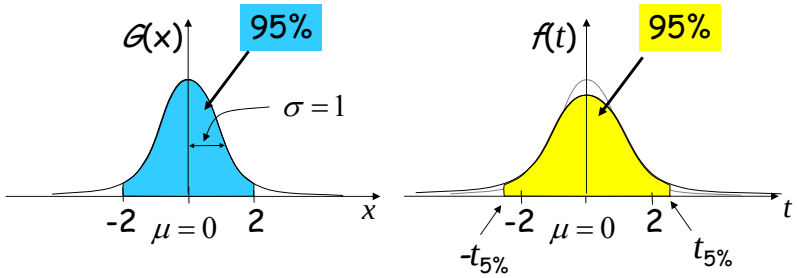
$$-t_{5\%} < \frac{\bar{x}-\mu}{s}\sqrt{n} < t_{5\%}$$

$$-t_{5\%} < \frac{\bar{x}-\mu}{s}\sqrt{n} \qquad \frac{\bar{x}-\mu}{s}\sqrt{n} < t_{5\%}$$

$$\mu < \bar{x} + t_{5\%}\sigma_{\bar{x}} \qquad \mu > \bar{x} - t_{5\%}\sigma_{\bar{x}}$$

$$\bar{x} + t_{5\%}\sigma_{\bar{x}} < \mu < \bar{x} + t_{5\%}\sigma_{\bar{x}}$$

Konfidenzintervall für den Erwartungswert



Statt Normalverteilung, haben wir eine *t*-Verteilung.

Statt „2" müssen wir die *t*-Werte der *t*-Verteilung anwenden.

$$\bar{x} \pm 2s_{\bar{x}}$$

$$\bar{x} \pm ts_{\bar{x}}$$

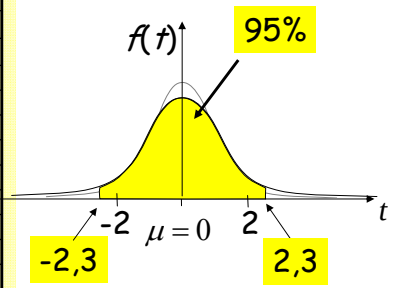
Konfidenzintervall für den Erwartungswert

Die *t*-Tabelle:

Freiheitsgrad	Signifikanzniveau (1-γ)		
	0.05	0.02	0.01
2	4.30266	6.96455	9.92499
3	3.18245	4.54071	5.84085
4	2.77645	3.74694	4.60408
5	2.57058	3.36493	4.03212
6	2.44691	3.14267	3.70743
7	2.36462	2.99795	3.49948
8	2.30601	2.89647	3.35538
9	2.26216	2.82143	3.24984
10	2.22814	2.76377	3.16926
11	2.20099	2.71808	3.10582
12	2.17881	2.68099	3.05454
13	2.16037	2.65030	3.01228
14	2.14479	2.62449	2.97685
15	2.13145	2.60248	2.94673
20	2.08596	2.52798	2.84534
50	2.00856	2.40327	2.67779
70	1.99444	2.38080	2.64790
100	1.98397	2.36421	2.62589
unendlich	1.95996	2.32635	2.57583

n-1

≈ 2



μ liegt in $\bar{x} \pm t_{5\%}s_{\bar{x}}$ mit 95% Wahrscheinlichkeit

Konfidenzintervall für den Erwartungswert

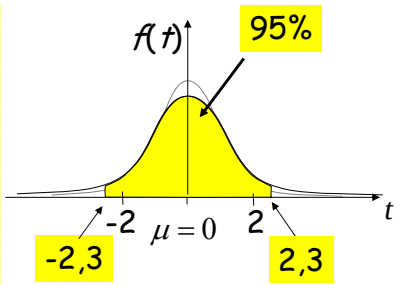
Mit Excel:

TINV(Wahrscheinlichkeit;Freiheitsgrad)

1-γ n-1

z.B.:

TINV(0,05,8)=2,30601≈2,3

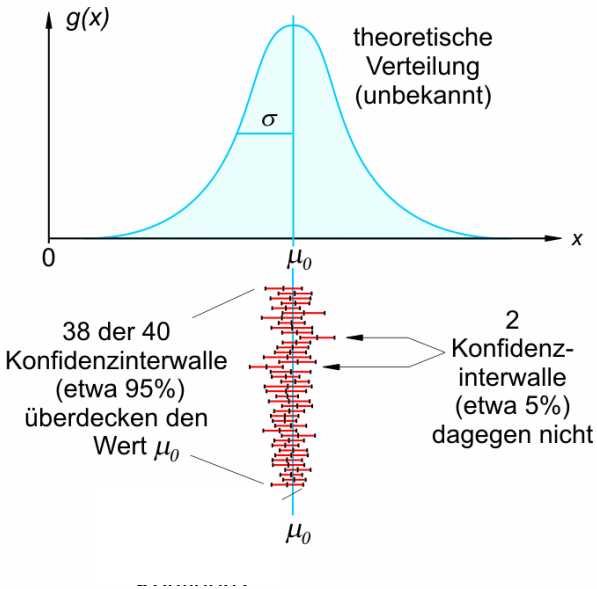


μ liegt in $\bar{x} \pm t_{5\%} s_{\bar{x}}$ mit 95% Wahrscheinlichkeit

z.B. $\bar{x}$=74 $^1/_{Min}$, s=18 $^1/_{Min}$, n=9

$s_{\bar{x}} = 18/\sqrt{9} = 6\ ^1/_{Min}$ $t_{5\%} = 2{,}3$ ⟶ 74±14 $^1/_{Min}$ 95% W.

Bedeutung der Konfidenz



Pr.Buch Abb. 11

Zusammenfassung der Schätzungen

Punktsätzungen:

Stich-probe	Grund-gesamtheit
$\bar{x}$	→ μ
s	→ σ
n	→ ∞

Intervallschätzungen

1. σ ist bekannt:

$\mu \pm 2\frac{\sigma}{\sqrt{n}}$ 95%

2. σ ist unbekannt:

Grob:

$\bar{x} \pm 2 s_{\bar{x}}$ 95%

Genau:

$\bar{x} \pm t_{5\%} s_{\bar{x}}$ 95%

Bestimmung des Stichprobenumfanges

Welcher Stichprobenumfang ist notwendig zu einer bestimmten Genauigkeit?

(z.B.: Körperhöhe mit ±1cm „Genauigkeit" bei 95% Konfidenzniveau)

$$2s_{\bar{x}} = 1\,\text{cm} \quad \Rightarrow \quad s_{\bar{x}} = 0{,}5\ \text{cm}$$

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \quad \Rightarrow \quad s_{\bar{x}}^2 = \frac{s^2}{n} \quad \Rightarrow \quad n = \frac{s^2}{s_{\bar{x}}^2}$$

s = ? s kann aus einer kleineren Stichprobe geschätzt werden.

Z.B.: Körperhöhe in einer Studentengruppe (20 St.): s = 8,3 cm

$$n = \frac{s^2}{s_{\bar{x}}^2} = \frac{8.3^2\,\text{cm}^2}{0.5^2\,\text{cm}^2} \approx 276$$

Konfidenzintervall für Quotienten

Zwei Möglichkeiten: (E/$\bar{E}$, z.B.: Raucher/Nichtraucher)
Binomialverteilung

E kommt mit Wahrscheinlichkeit *p* vor.
Stichprobenumfang: *n*

p wird aus der relativen Häufigkeit geschätzt: $p = n_E/n$
Streuung der Binomialverteilung: $\sigma = \sqrt{np(1-p)}$

Analog zu $\bar{x} \pm 2\sigma/\sqrt{n}$

$p \pm 2\sqrt{p\,(1-p)/n}$ 95% Konfidenzniveau

zB.: 20 Raucher aus 100 ➡ $0{,}2 \pm 2\sqrt{0{,}2 \cdot 0{,}8/100} = 0{,}2 \pm 0{,}08$