
Schätzverfahren II

Dr. Dirk Hasenclever
University of Leipzig

[imise.](#)

Eigenschaften der Normalverteilung

Seien X und Y unabhängig und normalverteilt mit $X \sim N(\mu_X, \sigma_X^2)$ und $Y \sim N(\mu_Y, \sigma_Y^2)$. Sei $a, b \in \mathbf{R}$.

Reproduktivität der Normalverteilung:

- a) $(X \pm Y) \sim N(\mu_X \pm \mu_Y, \sigma_X^2 + \sigma_Y^2)$
- b) $a + b \cdot X \sim N(a + b \cdot \mu_X, b^2 \cdot \sigma_X^2)$
- c) $\frac{(X - \mu_X)}{\sigma_X} \sim N(0, 1)$ (Standardisierung)

Zentraler Grenzwertsatz I

Sei X eine reelle Zufallsvariable mit $E(X) < \infty$ und $\text{var}(X) < \infty$.

Nach dem Gesetz der großen Zahlen ist der arithmetische Mittelwert $\bar{X}_N$ eine erwartungstreue konsistente Schätzfunktion für $E(X)$.

$\bar{X}_N$ ist selbst wieder eine reelle Zufallsvariable.

Frage: Wie sieht die Verteilung von $\bar{X}_N$ aus?

Zentraler Grenzwertsatz II

SATZ:

Sei X eine reelle Zufallsvariable mit $E(X) < \infty$ und $\text{var}(X) < \infty$.

Sei $Z_N = \frac{\bar{X}_N - E(X)}{\sqrt{\text{var}(X)/N}}$ die Standardisierung von $\bar{X}_N$.

Sei $\Phi(z)$ die Verteilungsfunktion der Standardnormalverteilung $N(0,1)$.

Es gilt: $\lim_{N \rightarrow \infty} \text{pr}(Z_N \leq z) = \lim_{N \rightarrow \infty} F_{Z_N}(z) = \Phi(z)$

für alle z in denen F_{Z_N} stetig ist.

Für große N gilt also approximativ:

$$\bar{X}_N \sim N\left(E(X), \frac{\text{var}(X)}{N}\right)$$

Mischungsverteilungen

Seien X und Y reelle Zufallsvariablen mit Dichten f_X und f_Y und endlichen Erwartungswerten und Varianzen.

Sei $Q \sim \text{Bernoulli}(q)$.

Definiere $X_{\langle Q \rangle} Y := \begin{cases} X & \text{falls } Q = 1 \\ Y & \text{falls } Q = 0 \end{cases}$

Notiz:

a) $f_{X_{\langle Q \rangle} Y}(z) = q \cdot f_X(z) + (1 - q) \cdot f_Y(z)$

b) $E(X_{\langle Q \rangle} Y) = q \cdot E(X) + (1 - q) \cdot E(Y)$

c) $\text{var}(X_{\langle Q \rangle} Y) = q \cdot \text{var}(X) + (1 - q) \cdot \text{var}(Y) + q \cdot (1 - q) \cdot (E(X) - E(Y))^2$

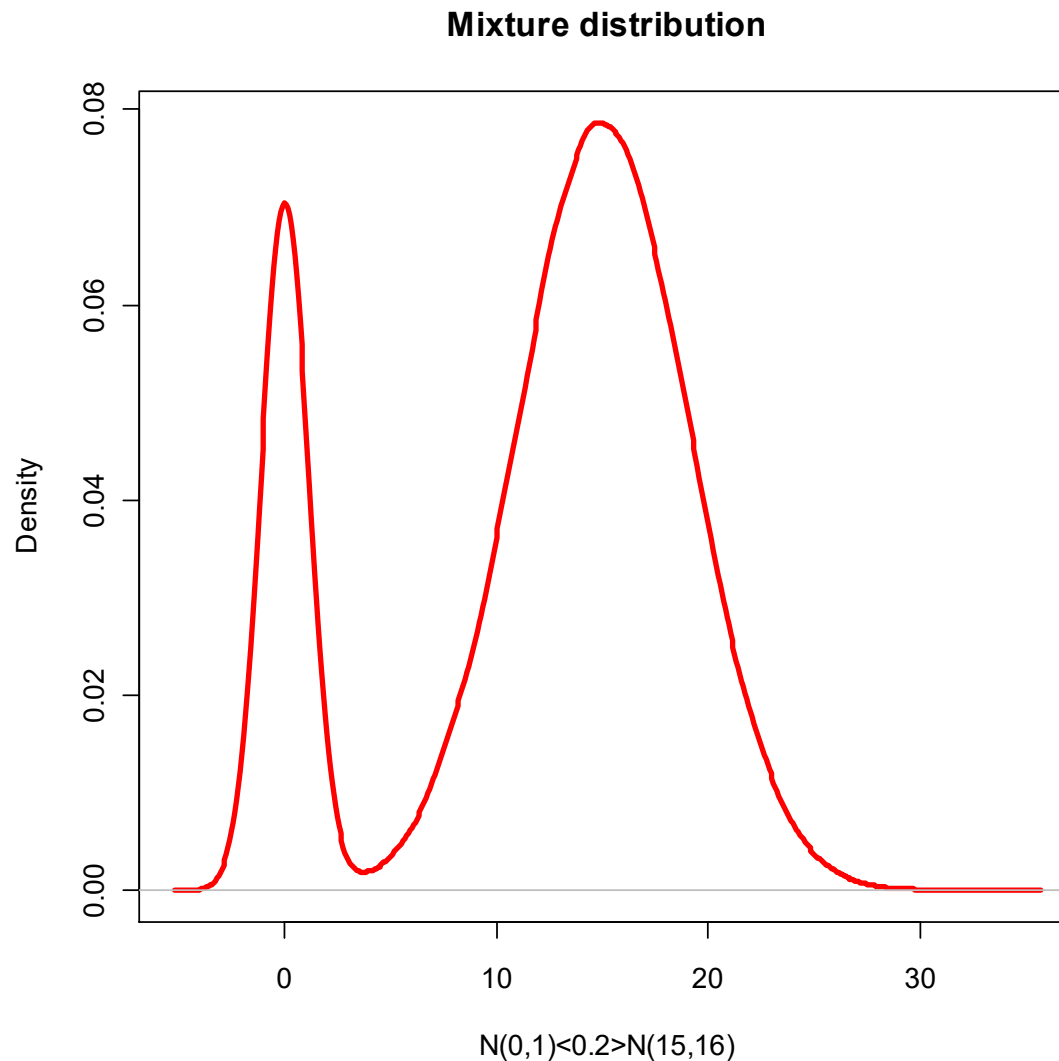
Mischungsverteilungen Varianz

- $\text{Var} (X_Q Y) = E ((X_Q Y)^2) - E(X_Q Y)^2 =$
- $q E(X^2) + (1-q) E(Y^2) - (q E(X) + (1-q) E(Y))^2 =$
- $q \text{var}(X) + q E(X)^2 + (1-q) \text{var}(Y) + (1-q) q E(Y)^2$
 $- q^2 E(X)^2 - (1-q)^2 E(Y)^2 - 2q(1-q)E(X)E(Y) =$
- $q \text{var}(X) + (1-q) \text{var}(Y) +$
 $(q-q^2) E(X)^2 + ((1-q) - (1-q)^2) E(Y)^2 - 2q(1-q)E(X)E(Y)$
- $q \text{var}(X) + (1-q) \text{var}(Y) + q(1-q) (E(X)-E(Y))^2$

Nutze: $E(Z^2) = \text{var}(Z) + E(Z)^2$

$$q(1-q) = (q-q^2) = (1-q) - (1-q^2)$$

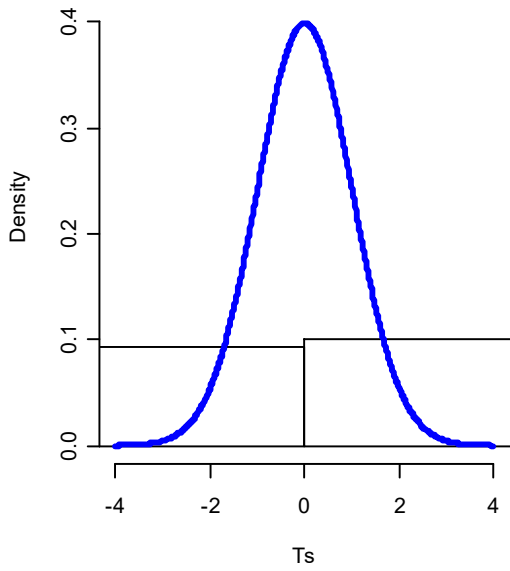
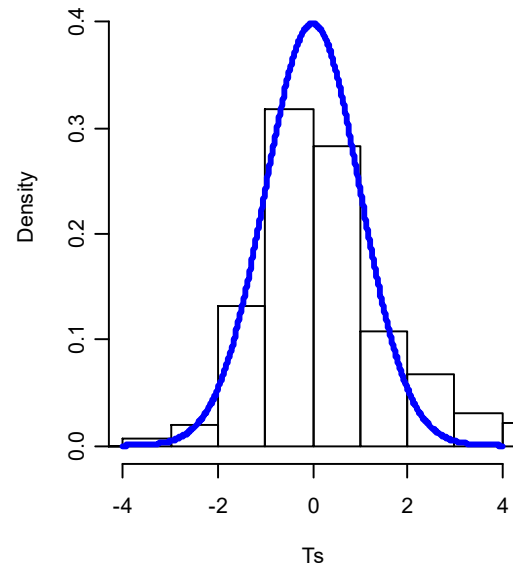
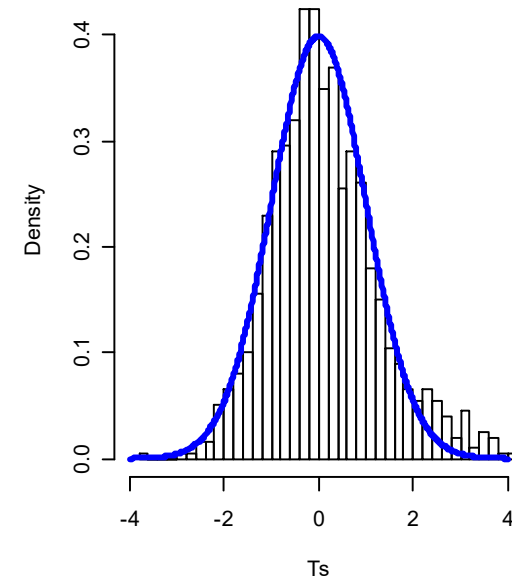
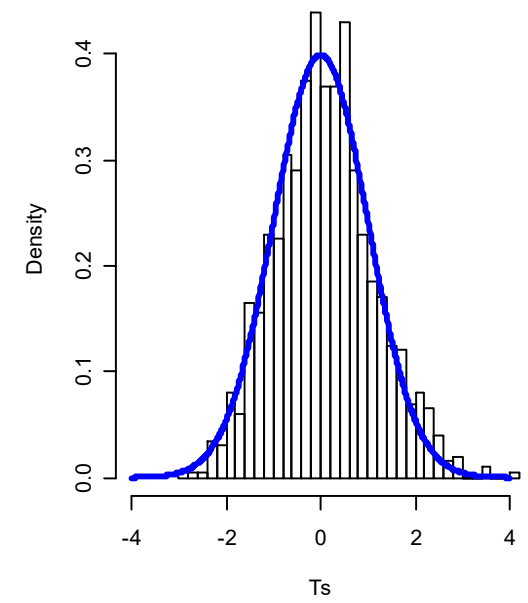
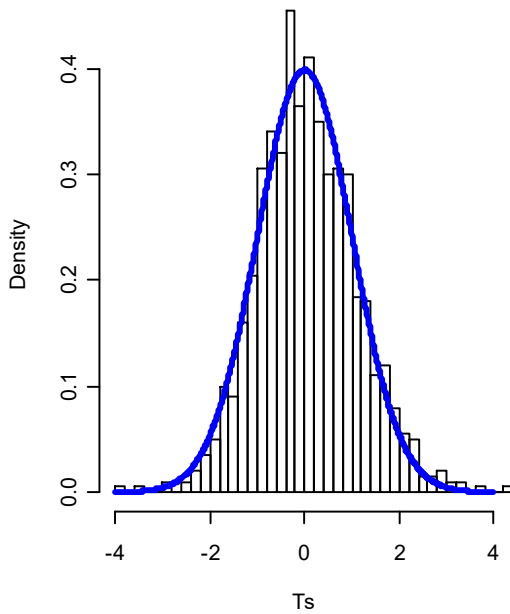
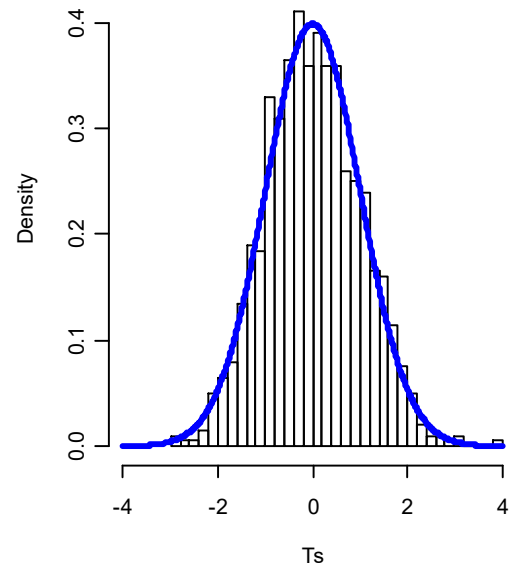
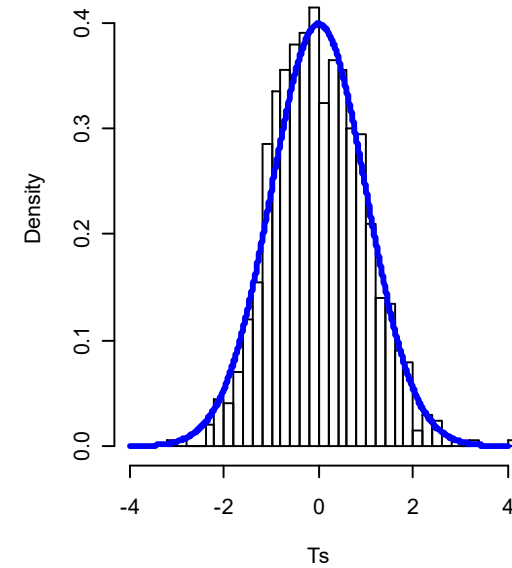
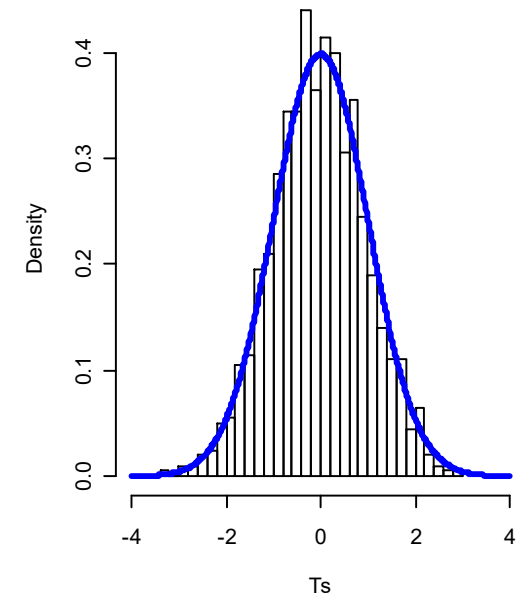
Ein krasses Beispiel

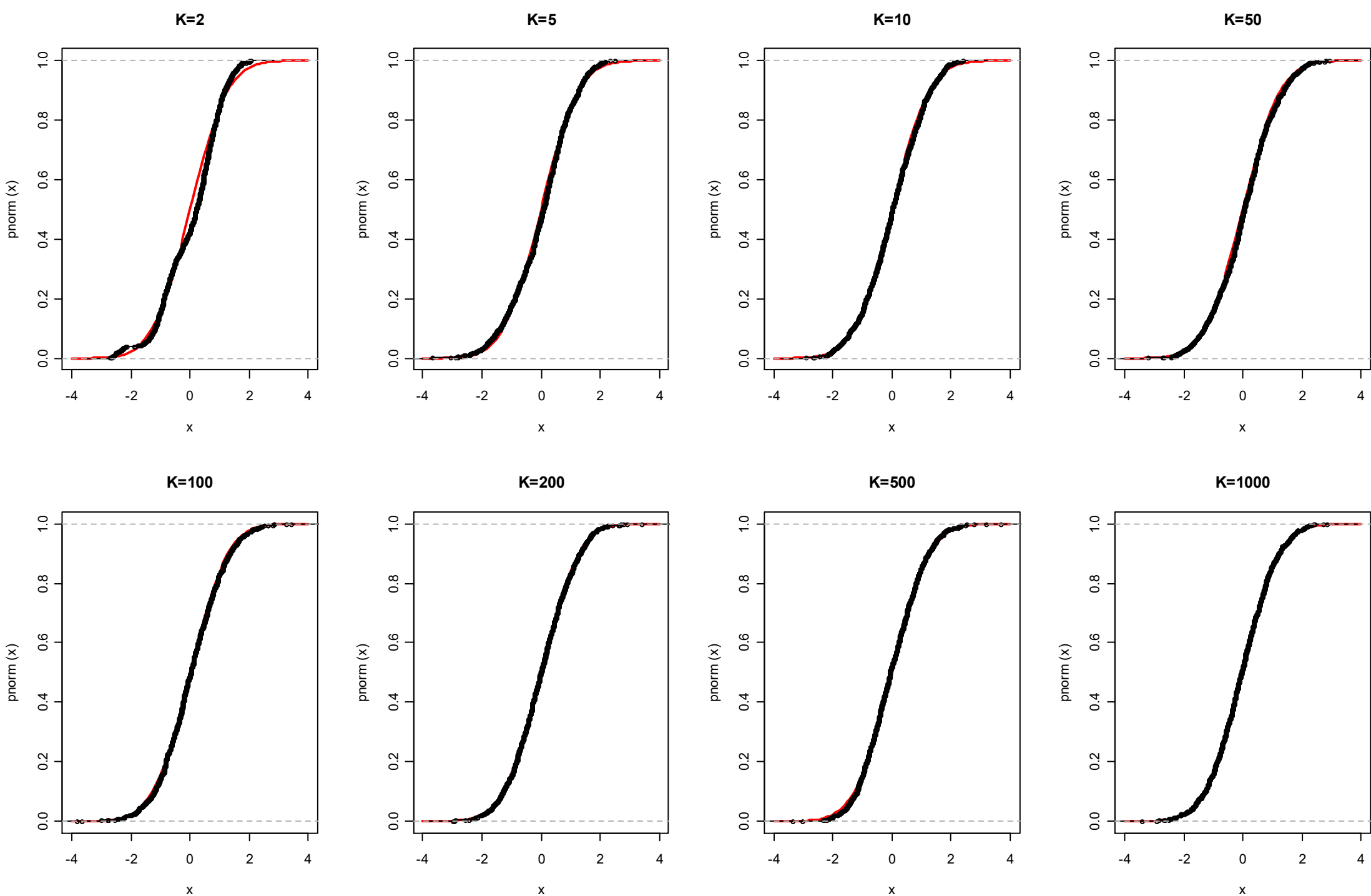


```
M<-q*my1+(1-q)*my2  =12
SD<-sqrt(q*sig1^2+(1-q)*sig2^2
        +q*(1-q)*(my1-my2)^2
        =7
```

**Nun simulieren wir
die Verteilung des
arithmetischen Mittels
dieser
Mischungsverteilung:**

R-Skript siehe Vorlesungsseite

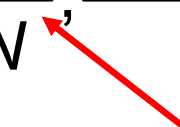
K= 3**K= 5****K= 10****K= 50****K= 100****K= 200****K= 500****K= 10000**



Normalverteilungs-Approximation für die Binomialverteilung

Seien $X_1, \dots, X_N$ unabhängig und identisch verteilt $\sim \text{Bernoulli}(p)$.

Für N genügend groß ist der Häufigkeitsschätzer

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i \sim N \left(p + \frac{0.5}{N}, \frac{p \cdot (1-p)}{N} \right)$$


approximativ normalverteilt.

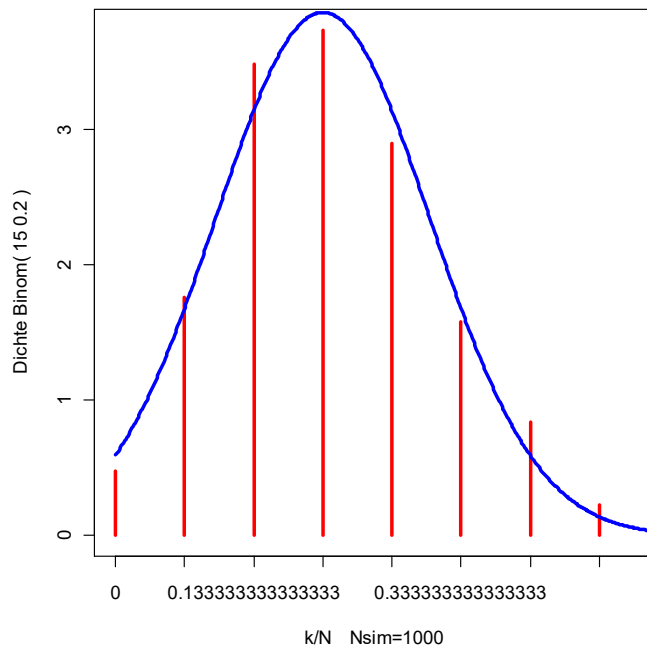
“Stetigkeitskorrektur”

Diese Approximation ist in etwa befriedigend wenn

$$N \cdot p \geq 10 \quad \wedge \quad N \cdot (1-p) \geq 10$$

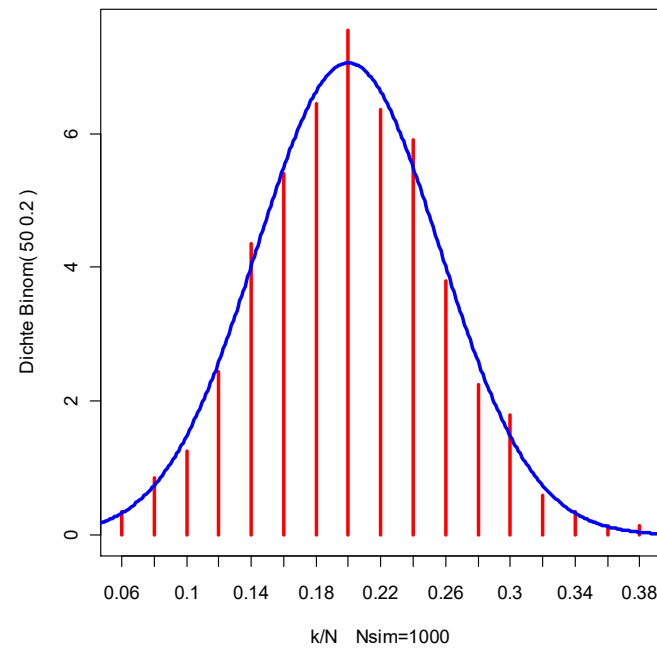
Illustration:

Normalverteilungsapproximation an Häufigkeitsschätzer



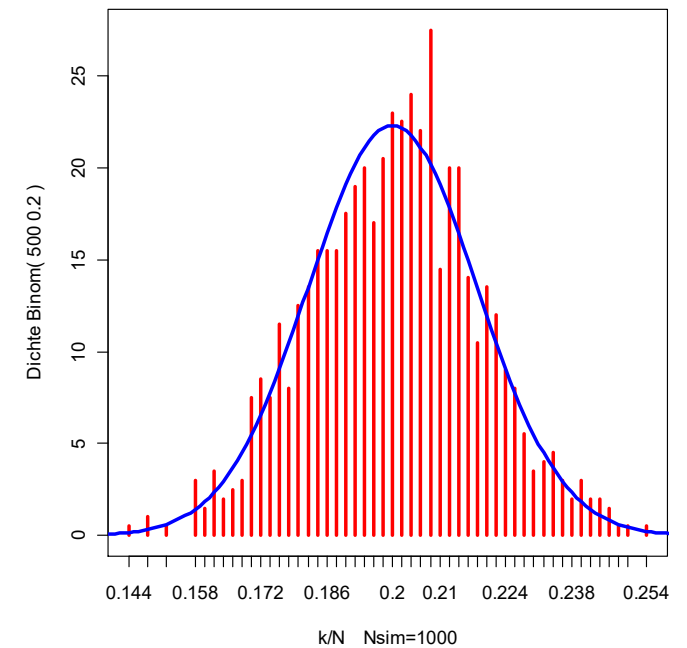
N=15

Normalverteilungsapproximation an Häufigkeitsschätzer



N=50

Normalverteilungsapproximation an Häufigkeitsschätzer



N=500

Biometrische Bedeutung des zentralen Grenzwertsatzes

- Es gibt **diverse Verallgemeinerungen des zentralen Grenzwertsatzes**
 - Die Verteilungen brauchen nicht identisch zu sein, solange nicht eine der Verteilungen dominiert.
 - Die Summanden können schwach korreliert sein.
- Dies erklärt die **biometrische Erfahrung**, dass
 - Messungen, die durch **Überlagerung von vielen kleinen additiven Fehlern** verrauscht werden, sind approximativ **normal verteilt**.
 - Messungen, die durch **Überlagerung von vielen kleinen multiplikativen Fehlern** verrauscht werden, sind approximativ **lognormal verteilt**, d.h. $\log(X)$ ist approximativ normal.

3 σ -Regel

Sei $X \sim N(\mu_X, \sigma_X^2)$.

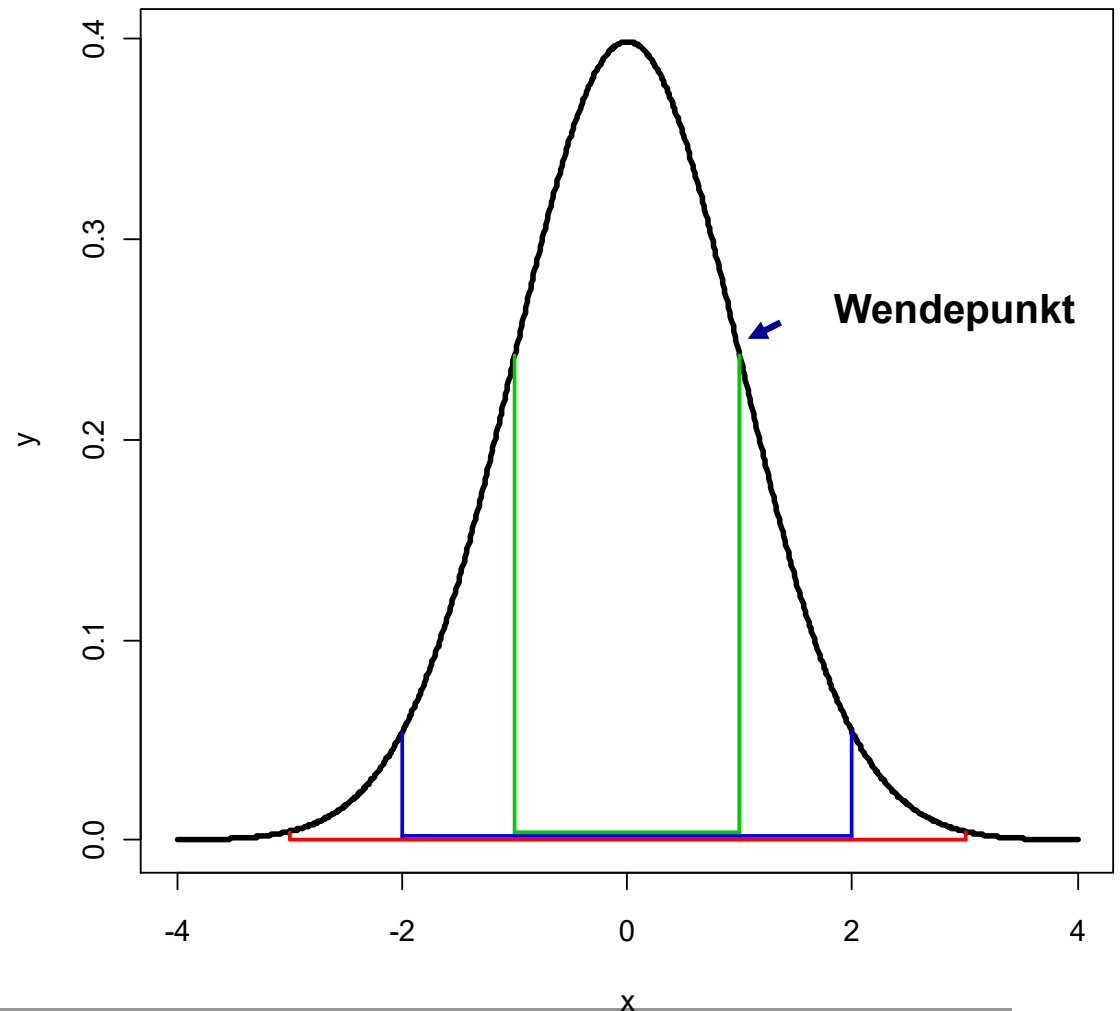
dann gilt:

$$\text{pr}(X \in [\mu_X - \sigma_X, \mu_X + \sigma_X]) \approx 2/3$$

$$\text{pr}(X \in [\mu_X - 2 \cdot \sigma_X, \mu_X + 2 \cdot \sigma_X]) \approx 0.95$$

$$\text{pr}(X \in [\mu_X - 3 \cdot \sigma_X, \mu_X + 3 \cdot \sigma_X]) \approx 0.99$$

Normalverteilung



Idee: Konfidenzintervall für Erwartungswert

Seien $X_1, \dots, X_N$ unabhängig und identisch verteilt.

Sei N groß genug, dass der zentrale Grenzwertsatz greift:

$$\begin{aligned} \Pr(\bar{X}_N \in [E(X) - 2\sqrt{\frac{\text{var}(X)}{N}}, E(X) + 2\sqrt{\frac{\text{var}(X)}{N}}]) &= \\ \Pr(\bar{X}_N \geq E(X) - 2\sqrt{\frac{\text{var}(X)}{N}} \wedge \bar{X}_N \leq E(X) + 2\sqrt{\frac{\text{var}(X)}{N}}) &= \\ \Pr(\bar{X}_N + 2\sqrt{\frac{\text{var}(X)}{N}} \geq E(X) \wedge \bar{X}_N - 2\sqrt{\frac{\text{var}(X)}{N}} \leq E(X)) &= \\ \Pr(E(X) \in [\bar{X}_N - 2\sqrt{\frac{\text{var}(X)}{N}}, \bar{X}_N + 2\sqrt{\frac{\text{var}(X)}{N}}]) &\approx 0.95 \end{aligned}$$

Hat man zudem einen hinreichend guten Schätzer $\hat{\sigma}^2$ für $\text{var}(X)$,
so erhält man approximativ:

$$\Pr(E(X) \in [\bar{X}_N - 2\frac{\hat{\sigma}}{\sqrt{N}}, \bar{X}_N + 2\frac{\hat{\sigma}}{\sqrt{N}}]) \approx 0.95$$

Wie genau ist eine „gegebene“ Schätzung? I

Seien $X_1, \dots, X_N$ unabhängig und identisch verteilt nach einer Grundverteilung F_X .

Sei θ_X ein Parameter der Verteilungsfunktion F_X .

und $\hat{\theta}_N$ eine Schätzfunktion für θ_X .

Gegeben eine Realisierung $x_1, \dots, x_N$ von $X_1, \dots, X_N$.

Wir berechnen den Schätzwert $\hat{\theta}_N(x_1, \dots, x_N)$ für θ_X .

Problem:

$\hat{\theta}_N(x_1, \dots, x_N)$ ist ein zufälliger Wert der Zufallsvariable $\hat{\theta}_N(X_1, \dots, X_N)$.

Wie genau kennen wir θ_X ?

Ideal wäre $|\hat{\theta}_N(x_1, \dots, x_N) - \theta_X|$, aber θ_X ist unbekannt.

Wie genau ist eine „gegebene“ Schätzung? II

Ansatz:

Über die Genauigkeit der singulären konkreten Schätzung $\hat{\theta}_N(x_1, \dots, x_N)$ können wir als Frequentisten keine Aussage machen.

Stattdessen beschreiben wir die Genauigkeit des Schätzverfahrens $\hat{\theta}_N(X_1, \dots, X_N)$ bei wiederholter Anwendung.

Definition: Stichprobenfunktionen $U_{\hat{\theta}_N}(X_1, \dots, X_N)$ und $O_{\hat{\theta}_N}(X_1, \dots, X_N)$ bilden ein Konfidenzintervall-Verfahren zum Konfidenzniveau $(1-\alpha)$ für $\hat{\theta}_N$ falls:

a) $\hat{\theta}_N(X_1, \dots, X_N) \in [U_{\hat{\theta}_N}(X_1, \dots, X_N), O_{\hat{\theta}_N}(X_1, \dots, X_N)]$

b) $pr(\theta_X \in [U_{\hat{\theta}_N}(X_1, \dots, X_N), O_{\hat{\theta}_N}(X_1, \dots, X_N)]) \geq (1-\alpha)$

Bedeutung: 95%- Konfidenzintervall

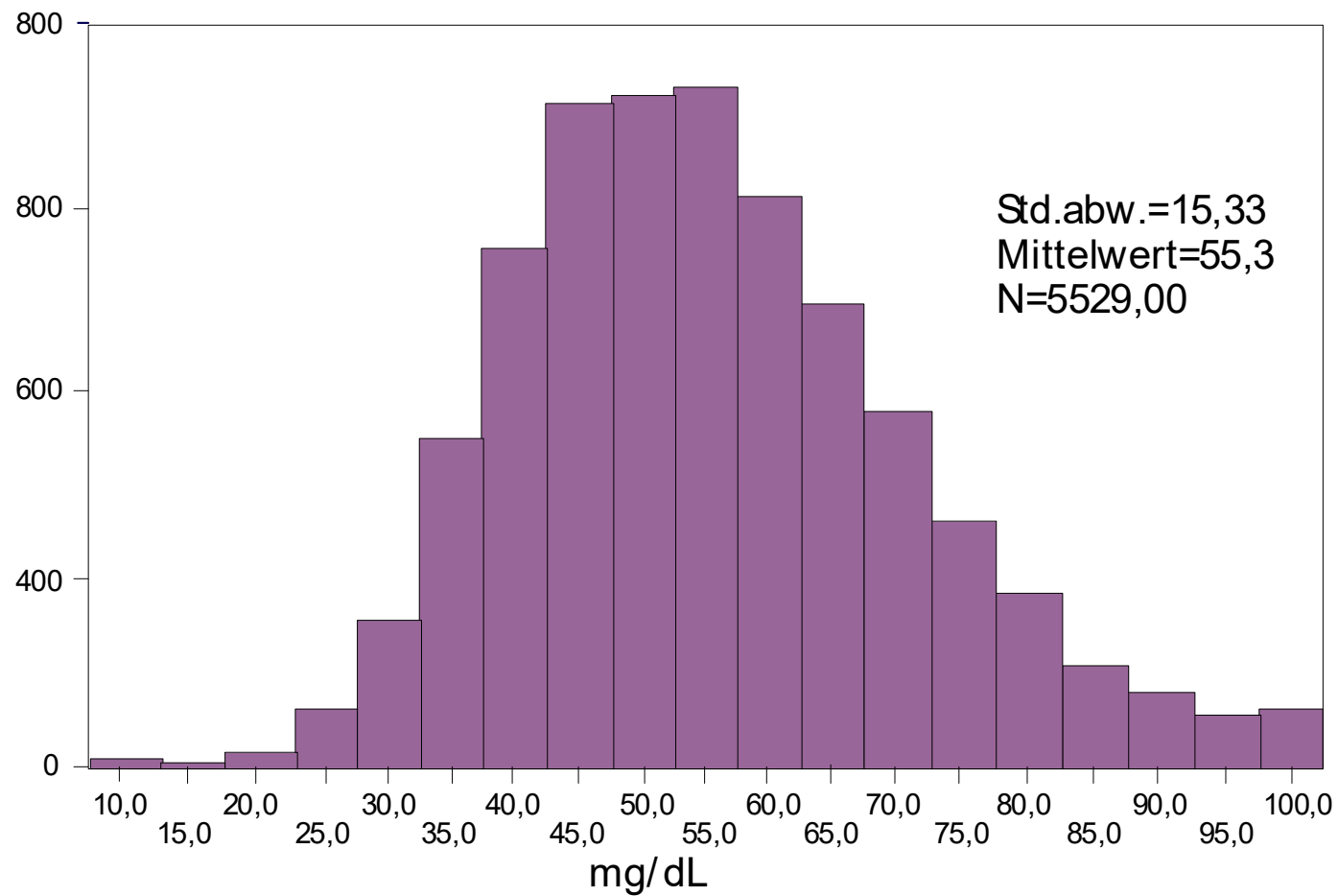
- Bei wiederholter unabhängiger Anwendung des Konstruktionsverfahrens enthält das (zufällige) Konfidenzintervall in 95% aller Fälle den „wahren“ Wert.

- Im **konkret gegebenen** Fall liegt jeweils der wahre Wert **im Intervall oder auch nicht !?!**

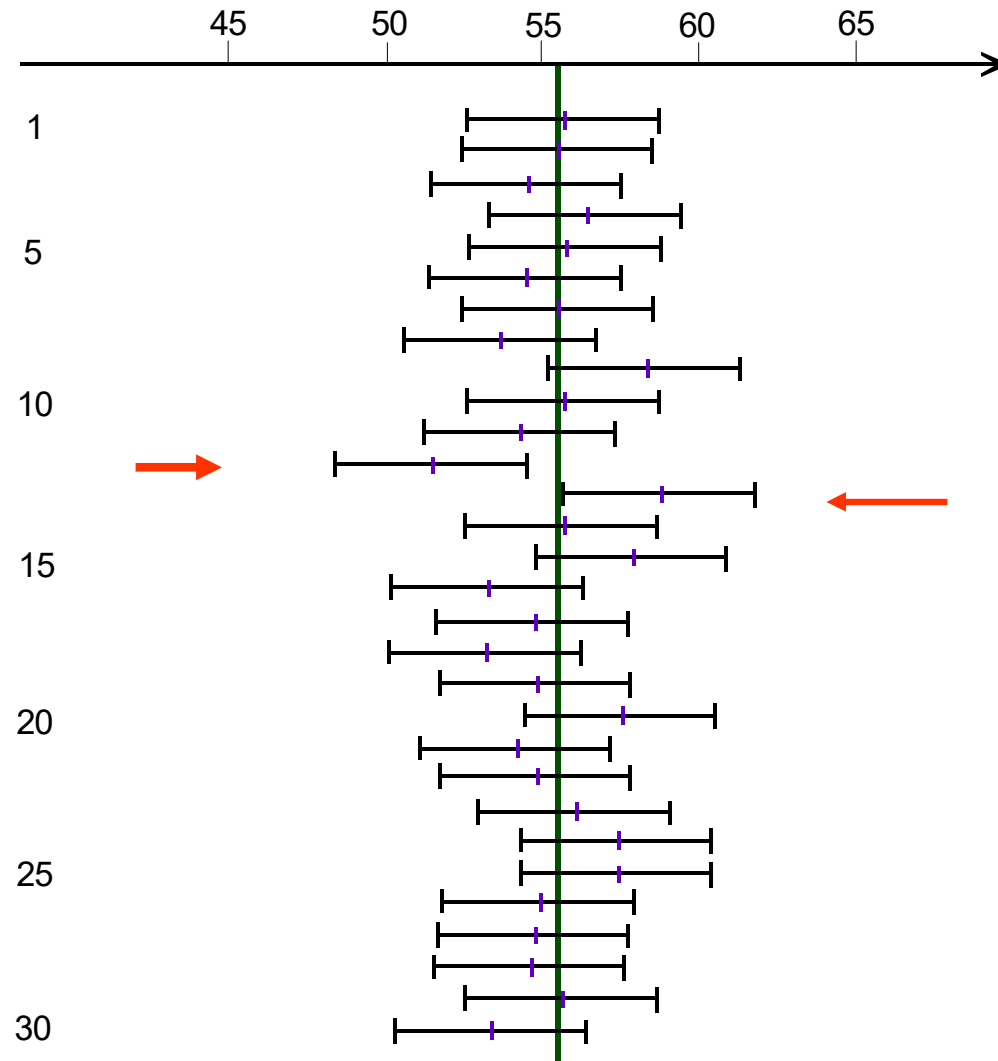
- **Wahrscheinlichkeitsaussagen darüber, ob der wahre Wert in einem gegebenen Intervall liegt, können nur Bayesianer formulieren...**

Lipid-Screening

HDL



Konfidenzintervall: Überdeckungswahrscheinlichkeit bei wiederholten Versuchen



Lipid Screening Daten

30 Stichproben a 100 Probanden

R-Skript auf Vorlesungsseite

Beispiel: Approximatives Konfidenzintervall für den Häufigkeitsschätzer

Wir betrachten den Häufigkeitsschätzer

$$\hat{p} = \frac{k}{N} \quad \text{für einen}$$

Binomialversuch mit Grundwahrscheinlichkeit p .

Falls Normalverteilungsapproximation greift:

$$\text{pr} \left(p \in \left[\hat{p} - 2\sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{N}}, \hat{p} + 2\sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{N}} \right] \right) \approx 0.95$$

$\hat{p} = 0.3$: 95% Konfidenzintervall

$k=3$ $N=10$: [0.155 , 0.445] (N noch zu klein)

$k=30$ $N=100$: [0.254 , 0.346]

$k=300$ $N=1000$: [0.285 , 0.315]

Gute biometrische Praxis

- Keine Messung ohne Rechenschaft über die Genauigkeit der Messung. (**Messfehler?**)
- Keine Schätzung ohne Rechenschaft über die Genauigkeit der Schätzung. (**Konfidenzintervall?**)

Beispiel: Konfidenzintervall für den Erwartungswert einer Normalverteilung I (bzw. falls zentraler Grenzwertsatz greift)

$\bar{X}_N \sim N(\mu, \frac{\sigma^2}{N})$ ist ein konsistenter, erwartungstreuer Schätzer für μ .

Also gilt (nach der 3 σ Regel):

$$[\bar{x}_n - 1.96 \cdot \frac{\sigma}{\sqrt{N}}; \bar{x}_n + 1.96 \cdot \frac{\sigma}{\sqrt{N}}]$$

ist ein 95%-Konfidenzintervall für μ .

Problem: σ im allgemeinen unbekannt.

Ersetzen durch Schätzwert vergrößert die Varianz

-> t-Verteilungen

Beispiel: Konfidenzintervall für den Erwartungswert einer Normalverteilung II (bzw. falls zentraler Grenzwertsatz greift)

Ist σ unbekannt und muss als S_X aus der Stichprobe selbst geschätzt werden, so gilt

$$\frac{\bar{X}_N - \mu}{S_X / \sqrt{N}} \sim t(df = N - 1)$$

t-Verteilung mit N-1 Freiheitsgraden.

$$\left[\bar{X}_n - t_{0.975}(df = N - 1) \cdot \frac{S_X}{\sqrt{N}}; \bar{X}_N + t_{0.975}(df = N - 1) \cdot \frac{S_X}{\sqrt{N}} \right]$$

ist ein 95%-Konfidenzintervall für μ .

$t_{0.975}$ ist das 97.5% Quantil der $t(df = n-1)$ Verteilung.

Z.B. für $n = 30$: $t_{0.975}(df = 29) = 2.05$

t-Verteilung (Eigenschaften)

Dichtefunktion:

m=Freiheitsgrade

$$f(x) = \frac{1}{\sqrt{m\pi}} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \frac{1}{\left(1 + \frac{x^2}{m}\right)^{\frac{m+1}{2}}}$$

Verteilungsfunktion:

$$F(x) = \frac{1}{\sqrt{m\pi}} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \int_{-\infty}^x \frac{1}{\left(1 + \frac{t^2}{m}\right)^{\frac{m+1}{2}}} dt$$

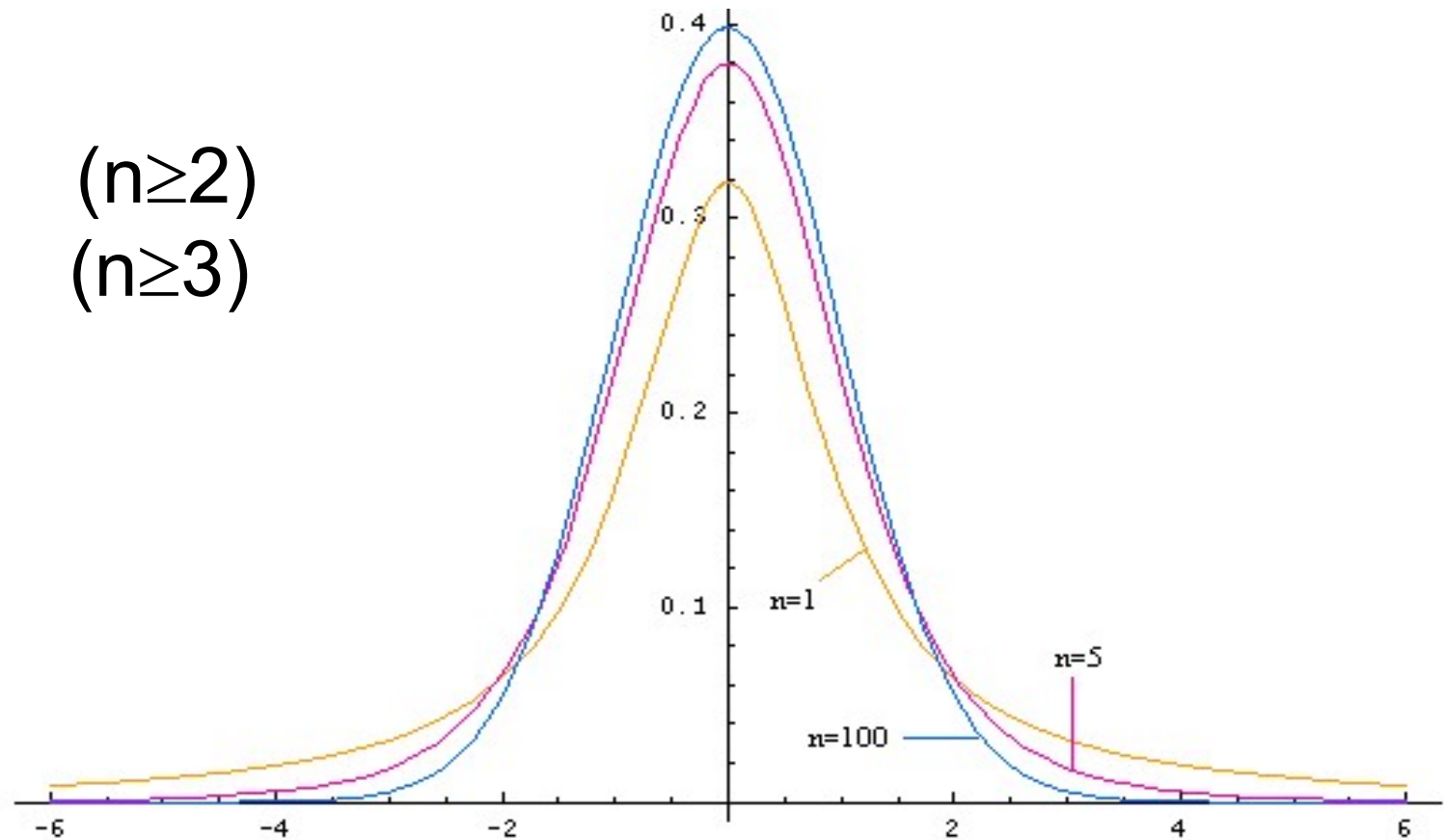
$T = \frac{X}{\sqrt{Y/m}}$ ist t-verteilt mit m Freiheitsgraden, wenn X und Y unabhängig und $X \sim N(0,1)$ und $Y \sim \chi^2_{m=n-1}$ verteilt mit m=n-1 Freiheitsgraden.

Die t-Verteilung berücksichtigt die zusätzliche Streuung bei Verwendung der empirischen Varianz anstatt der wahren.

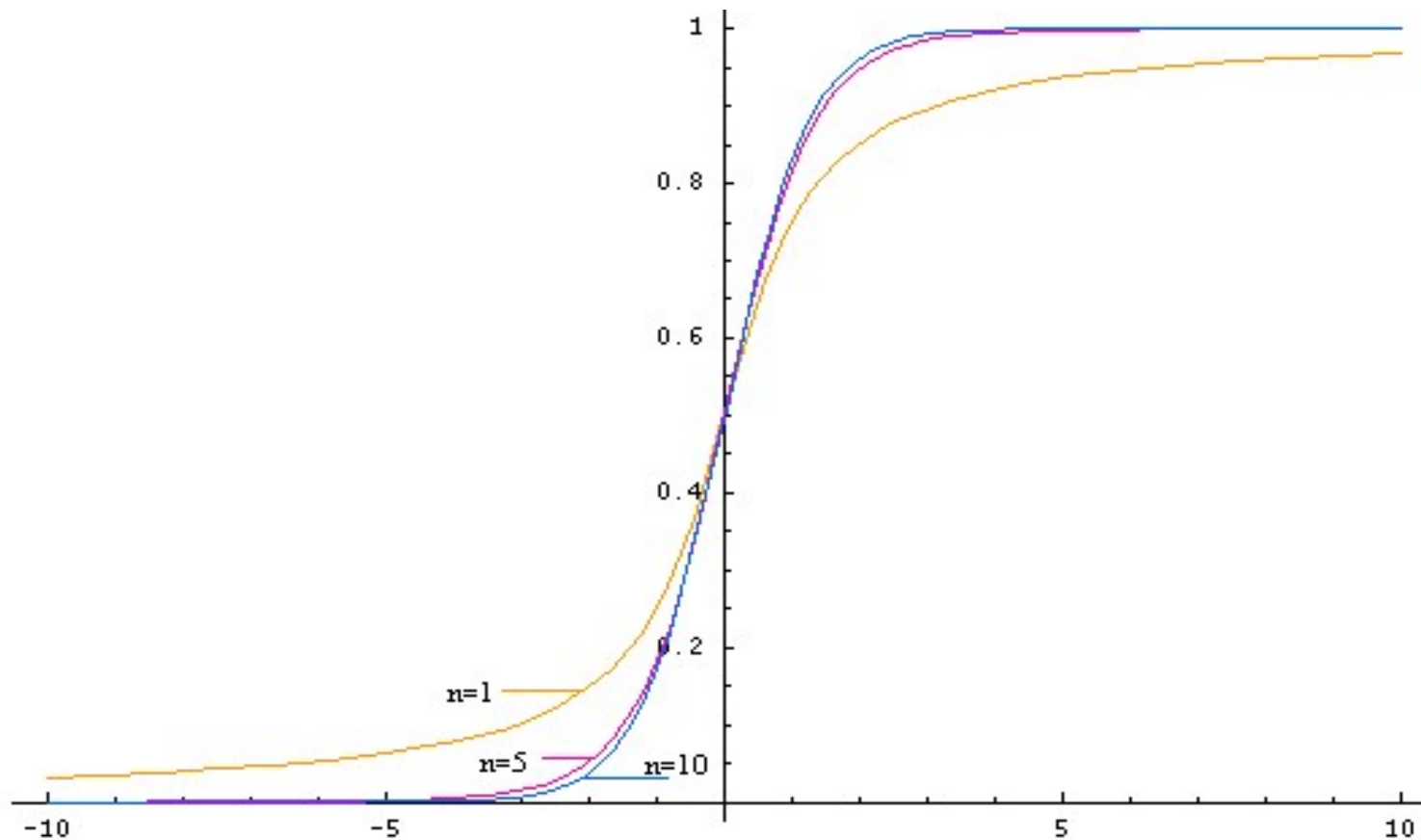
t-Verteilung (Dichte)

$$E(X)=0 \quad (n \geq 2)$$

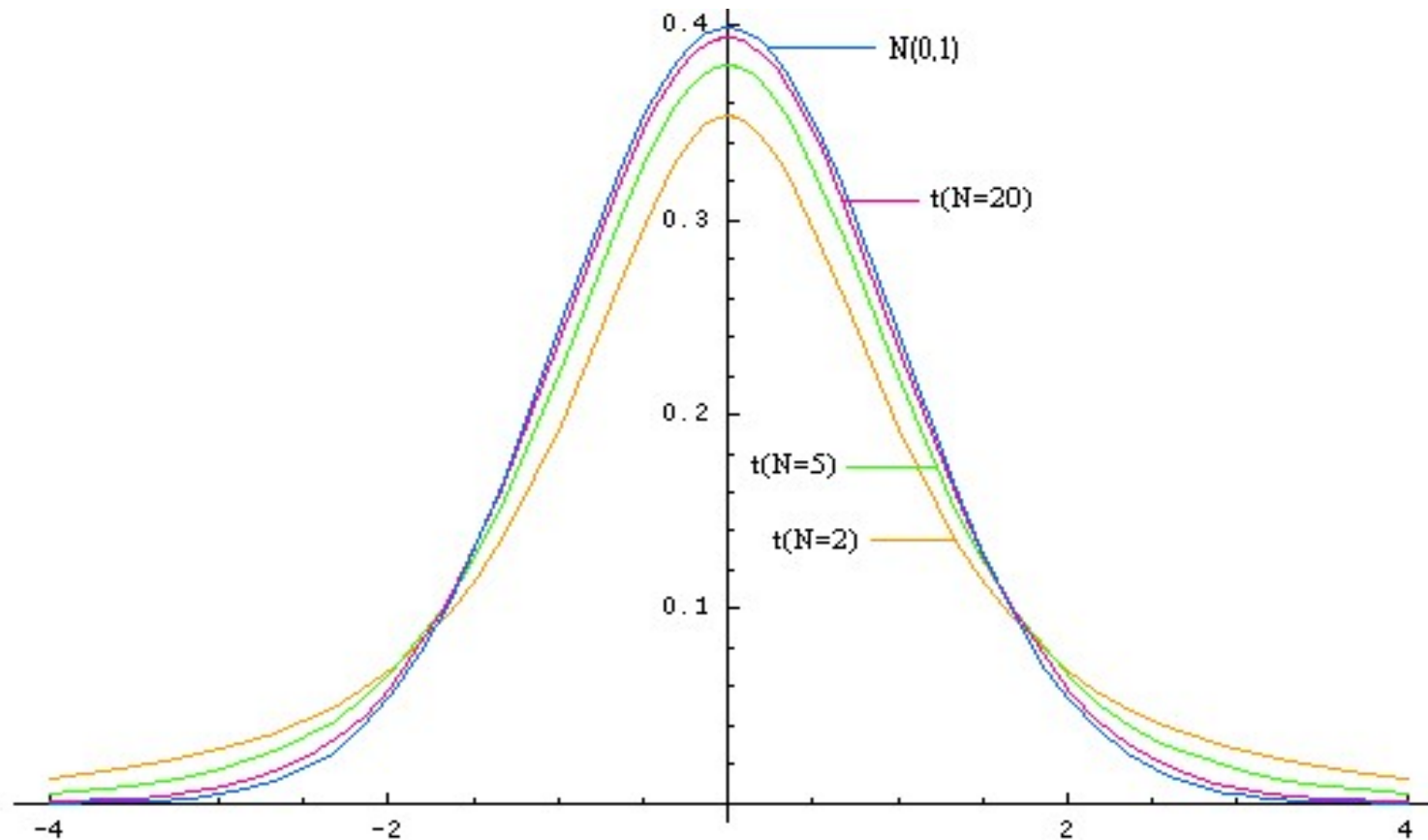
$$V(X)=n/(n-2) \quad (n \geq 3)$$



t-Verteilung (Verteilungsfunktion)

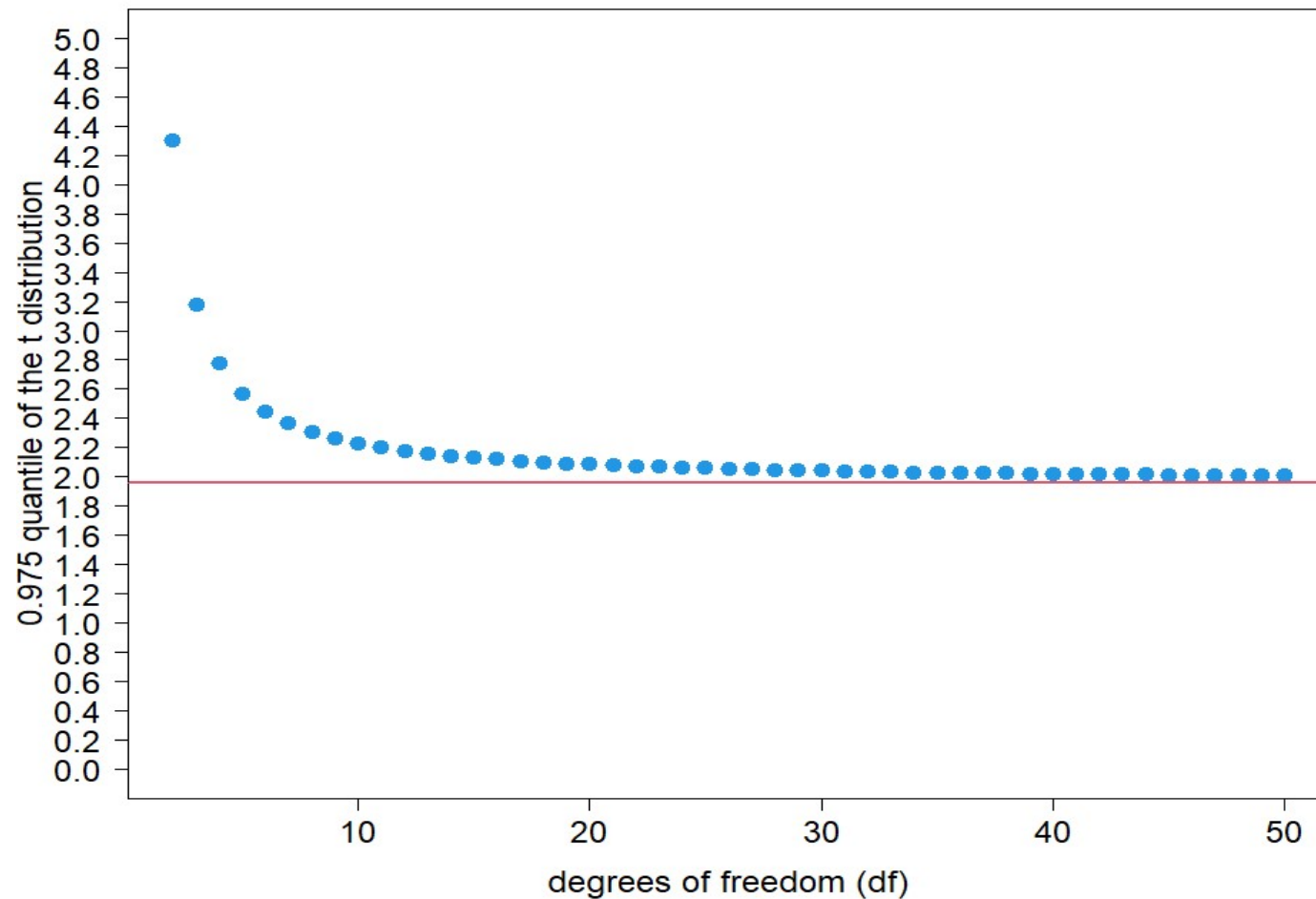


t-Verteilung und $N(0,1)$



Vergleich der Dichten der Standard-Normalverteilung und der t-Verteilung
Für $N > 100$ sind Normalverteilung und t-Verteilung praktisch gleich

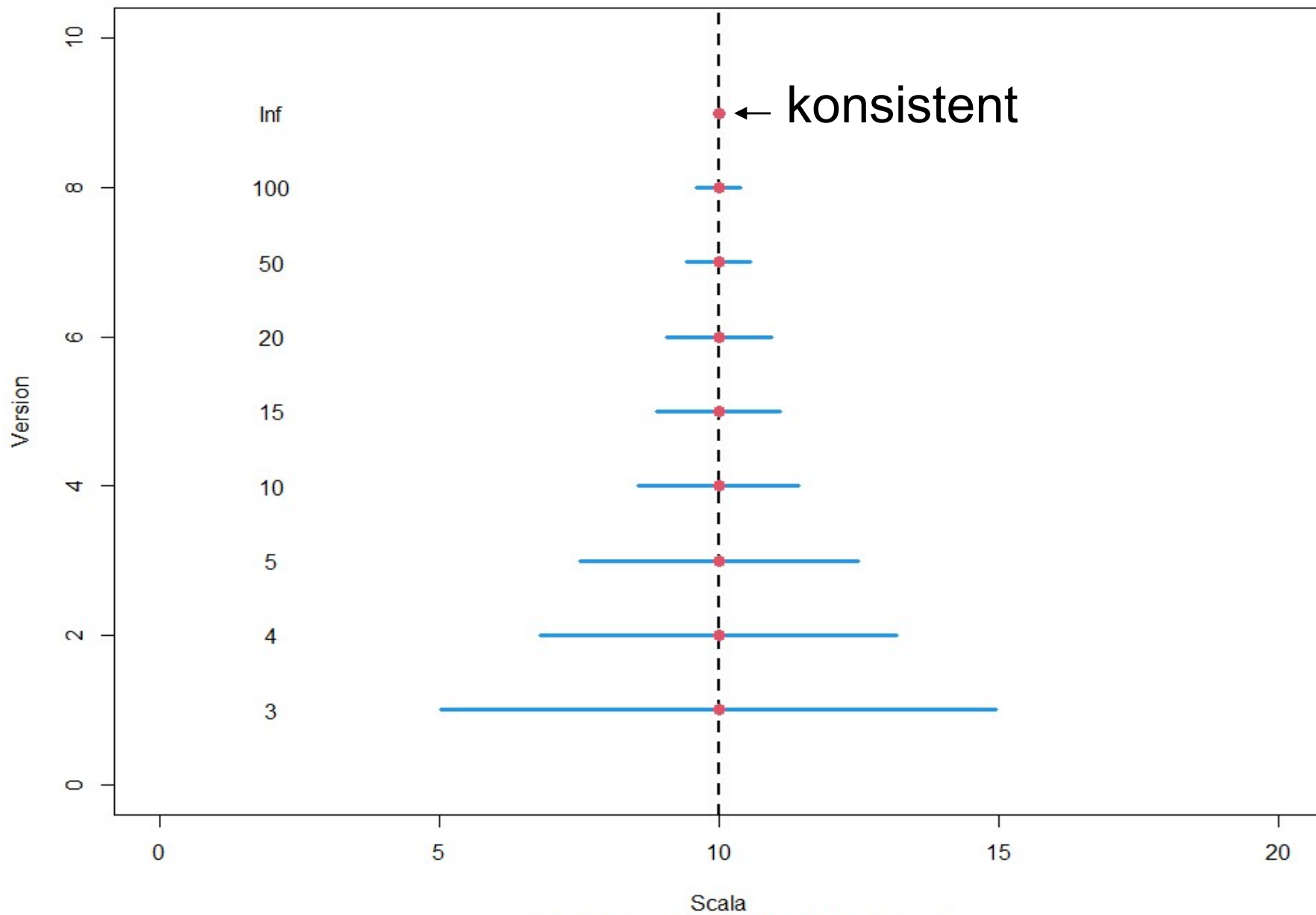
Statt 1.96 als Faktor bei mittleren Fallzahlen ≈ 2 , bei kleinsten Fallzahlen deutlich größer



Qualitative Eigenschaften eines Konfidenzintervalls (allgemein)

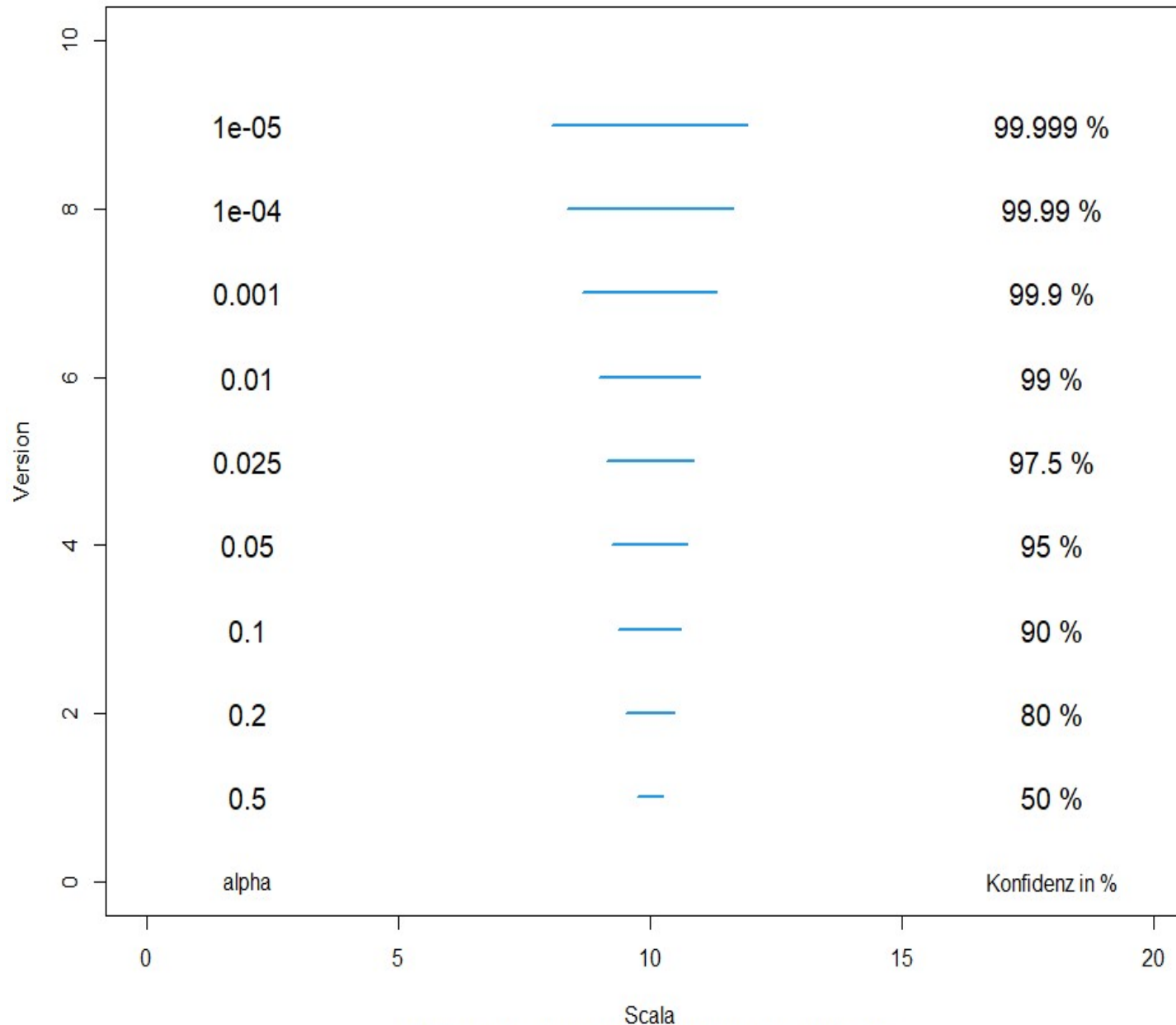
- Breite des CIs $\sim$ Standardfehler $\sim 1/\text{Wurzel}(N)$
 - bitter: doppelt so genau $\rightarrow$ vierfache Fallzahl...
- Standardfehler (SE) bezeichnet allgemein die Standardabweichung der Punktschätzers.
- bei metrischen Daten und arithmetischem Mittel:
Standardfehler = SE = $SD / \text{Wurzel}(N)$, wobei
–SD die Standardabweichung der Einzeldaten

CI-Breite $\sim 1/\sqrt{N}$ + Effekt t-Verteilung gegen Normalverteilung



Example: MeanEstimate= 10, SdEstimate = 2

CI-Breite steigt mit Konfidenzniveau 1-alpha



alpha	qnorm (1-alpha/2)
0.50	0.67
0.20	1.28
0.10	1.64
0.05	1.96
0.025	2.24
0.01	2.58
0.001	3.29
0.0001	3.89
0.00001	4.42

Hauptsatz der Stichprobentheorie

Seien $X_1, \dots, X_N$ unabhängig und identisch normalverteilt $X \sim N(\mu_X, \sigma_X^2)$.

Sei $\bar{X}_N$ das arithmetische Mittel und

$S_X^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X}_N)^2$ der erwartungstreue Varianzschätzer.

SATZ:

a) $\bar{X}_N \sim N(\mu_X, \frac{\sigma_X^2}{N})$

b) $\frac{(N-1) \cdot S_X^2}{\sigma_X^2} \sim \chi^2_{N-1}$ χ^2 -verteilt mit N-1 Freiheitsgraden.

c) $\bar{X}_N$ und S_X^2 sind unabhängig.

Hauptsatz der Stichprobentheorie II

Plausibilisierung:

Verschiebungssatz:

$$\sum_{i=1}^N (X_i - a)^2 = \sum_{i=1}^N (X_i - \bar{X}_N)^2 + N \cdot (\bar{X}_N - a)^2$$

mit $a = \mu_X$ und Erweiterung mit $\frac{1}{\sigma_X^2}$ folgt:

$$\sum_{i=1}^N \left(\frac{X_i - \mu_X}{\sigma_X} \right)^2 = \sum_{i=1}^N \left(\frac{X_i - \bar{X}_N}{\sigma_X} \right)^2 + \left(\frac{\bar{X}_N - \mu_X}{\sigma_X / \sqrt{N}} \right)^2.$$

$$" \quad \chi^2_N \quad = \quad \chi^2_{N-1} \quad + \quad \chi^2_1 \quad "$$

nach Definition Additivität χ^2 nach Definition

Chi²-Verteilung (Eigenschaften)

Dichtefunktion:

$$f(x) = \begin{cases} \frac{1}{2^{n/2} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

Verteilungsfunktion:

$$F(x) = \begin{cases} \frac{1}{2^{n/2} \Gamma\left(\frac{n}{2}\right)} \int_0^x t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

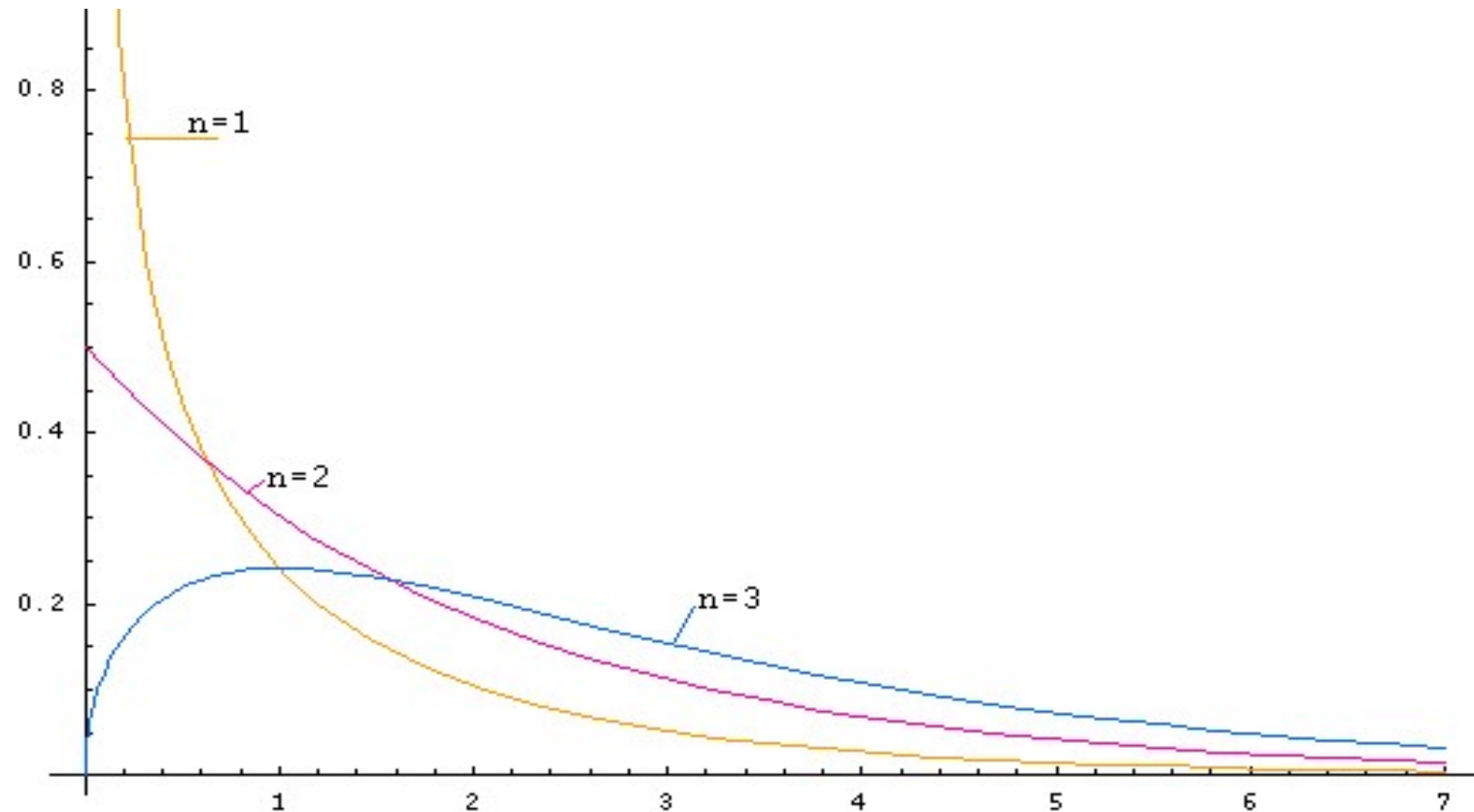
Seien $X_1, \dots, X_n$ unabhängig und standardnormalverteilt,
dann ist: $\sum_{i=1}^n X_i^2 \sim \chi^2_n$ chi²-Verteilt mit n Freiheitsgraden.

Seien $X \sim \chi^2_n$ und $Y \sim \chi^2_m$ unabhängig, dann gilt:

$$(X + Y) \sim \chi^2_{n+m}$$

Chi²-Verteilung (Dichte 1)

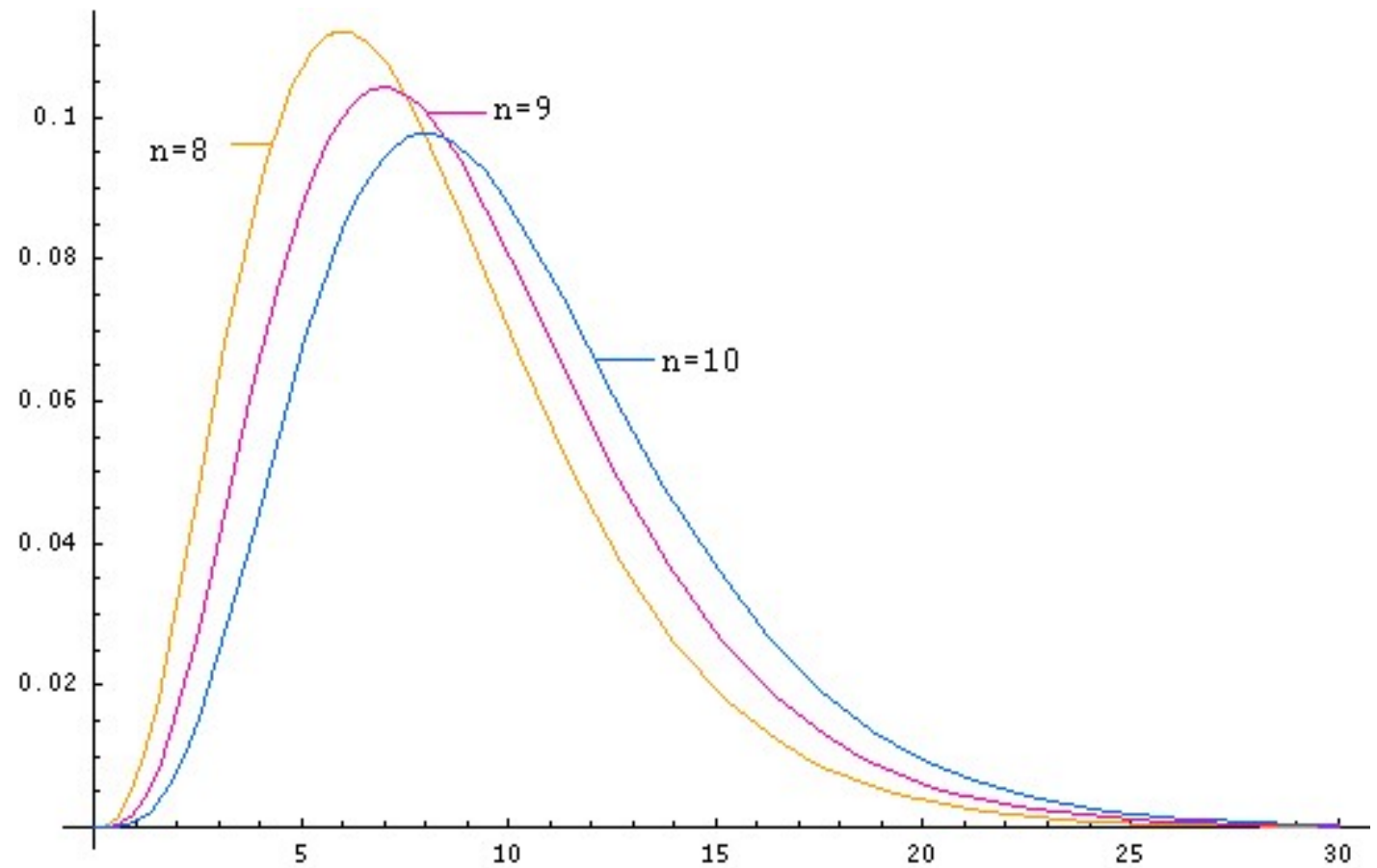
$$E(X) = n$$
$$V(X) = 2n$$



Dichtefunktionen

Chi²-Verteilung (Dichte 2)

$$E(X) = n$$
$$V(X) = 2n$$

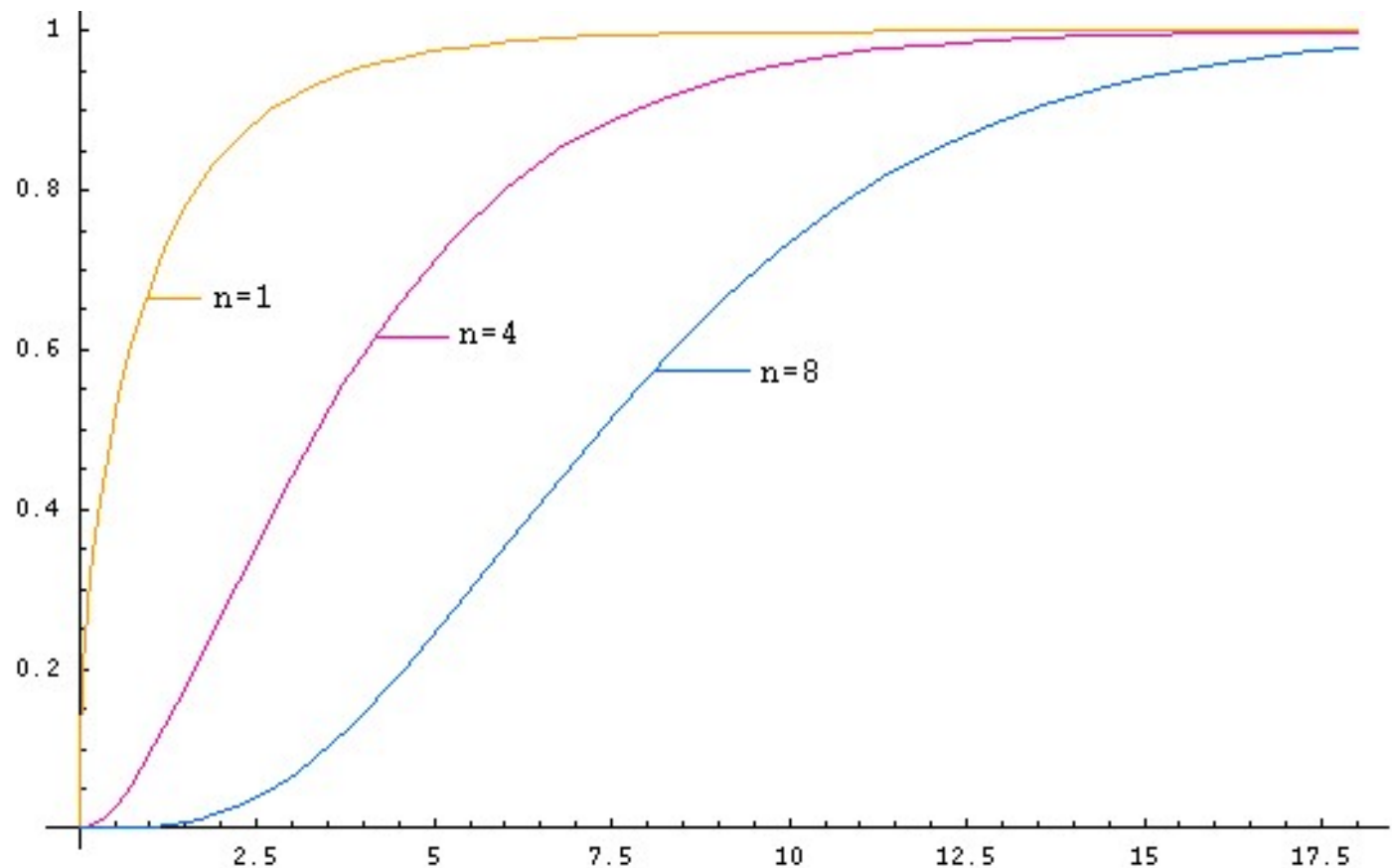


Dichtefunktionen

Chi²-Verteilung (Verteilung)

Die Chi²-Verteilung ist ein Spezialfall der Gamma-Verteilung.

Eine Gamma-Verteilung mit:
 $a=n/2$ und $b=2$ ist
Chi² - Verteilt mit
 n Freiheitsgraden.



Verteilungsfunktionen

Chi²-Verteilung (Approximation)

Die Chi²-Verteilung kann für große n durch die Normalverteilung angenähert werden. In diesem Beispiel: $\text{Chi}^2(100) \approx N(100, 200)$

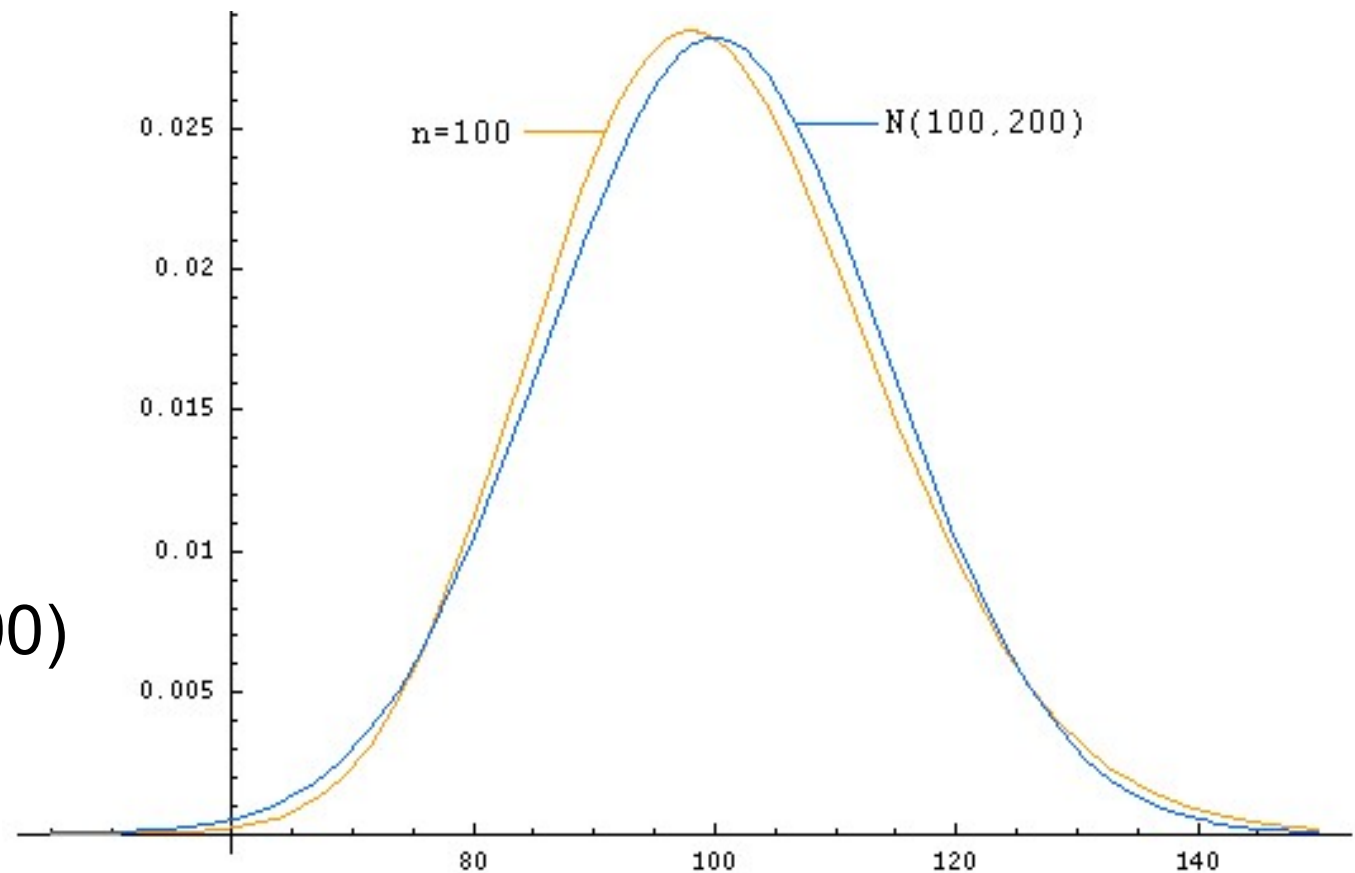
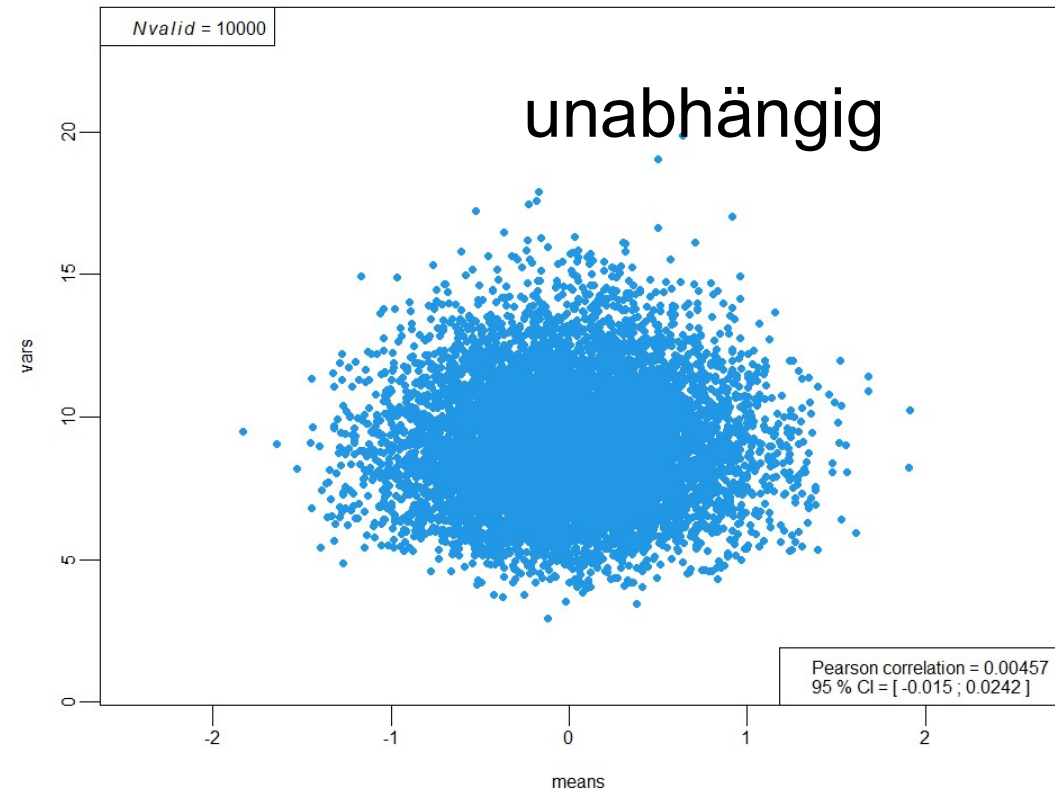
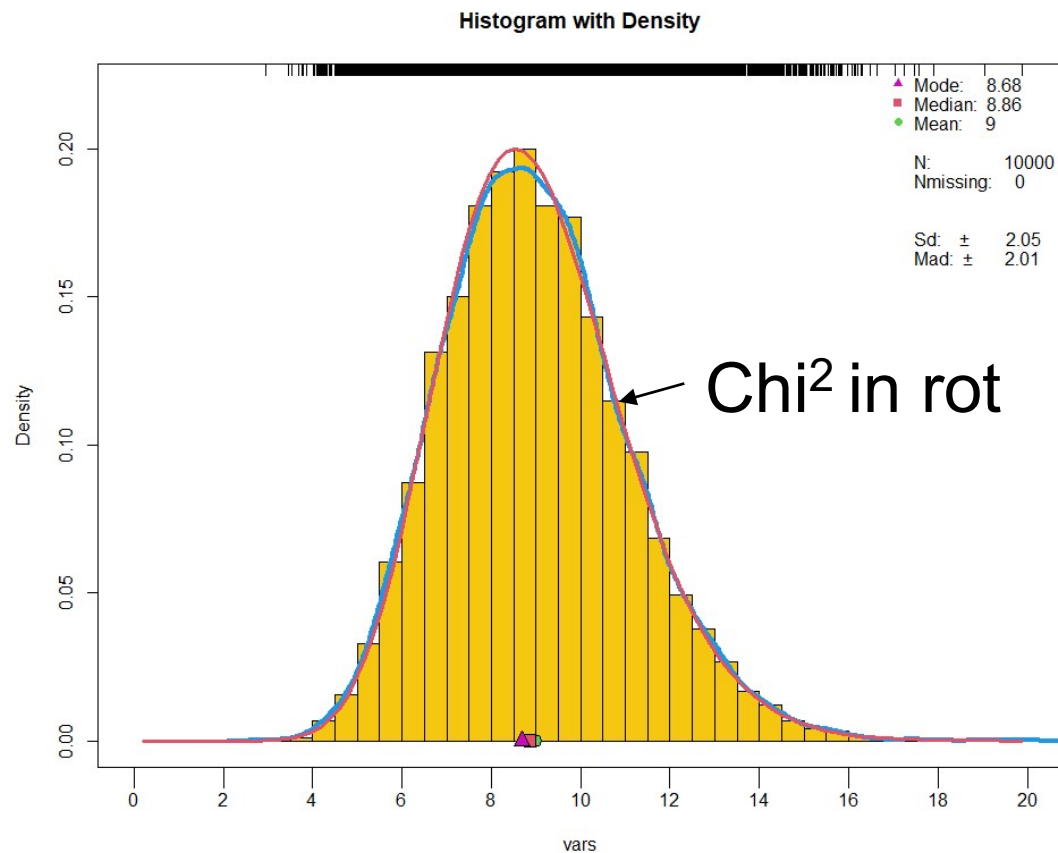
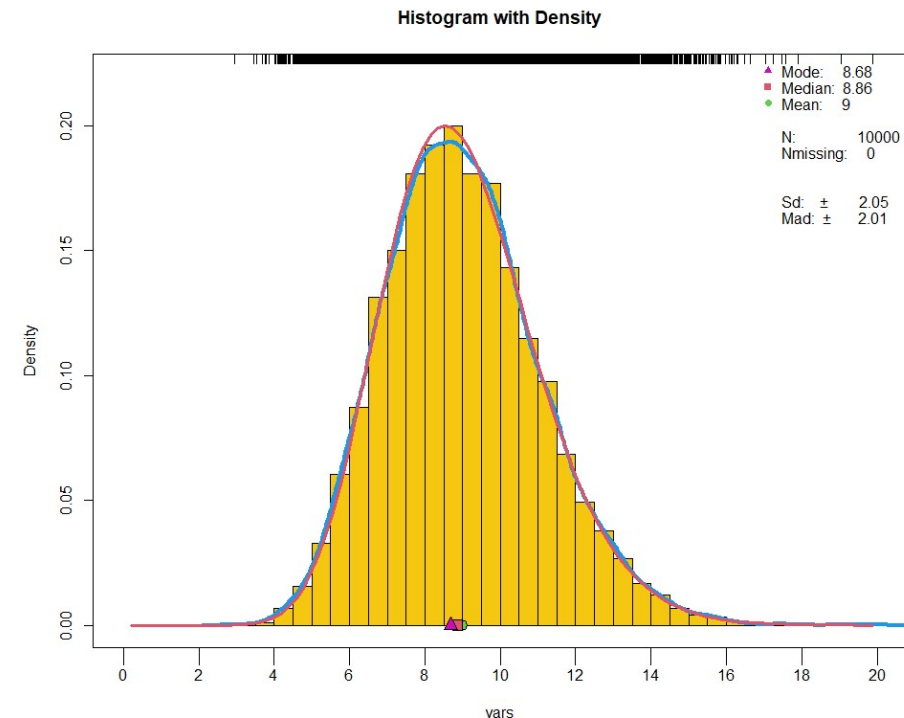


Illustration zum Hauptsatz der Stichprobentheorie Simulation aus Normalverteilung



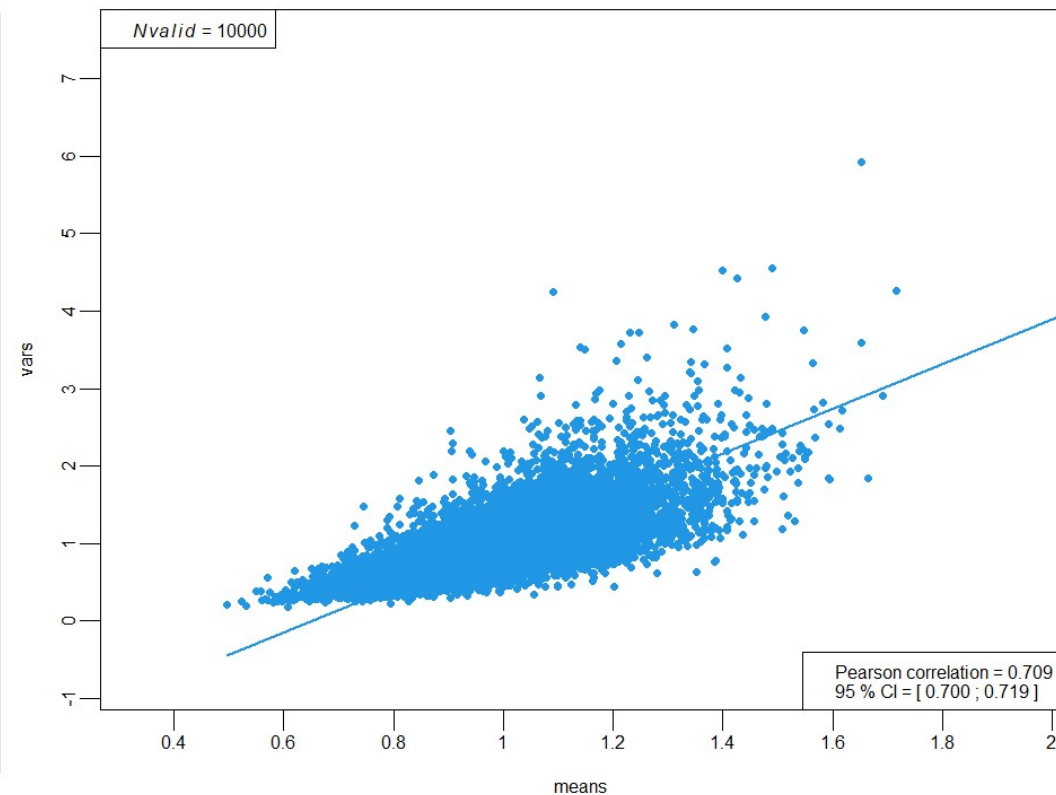
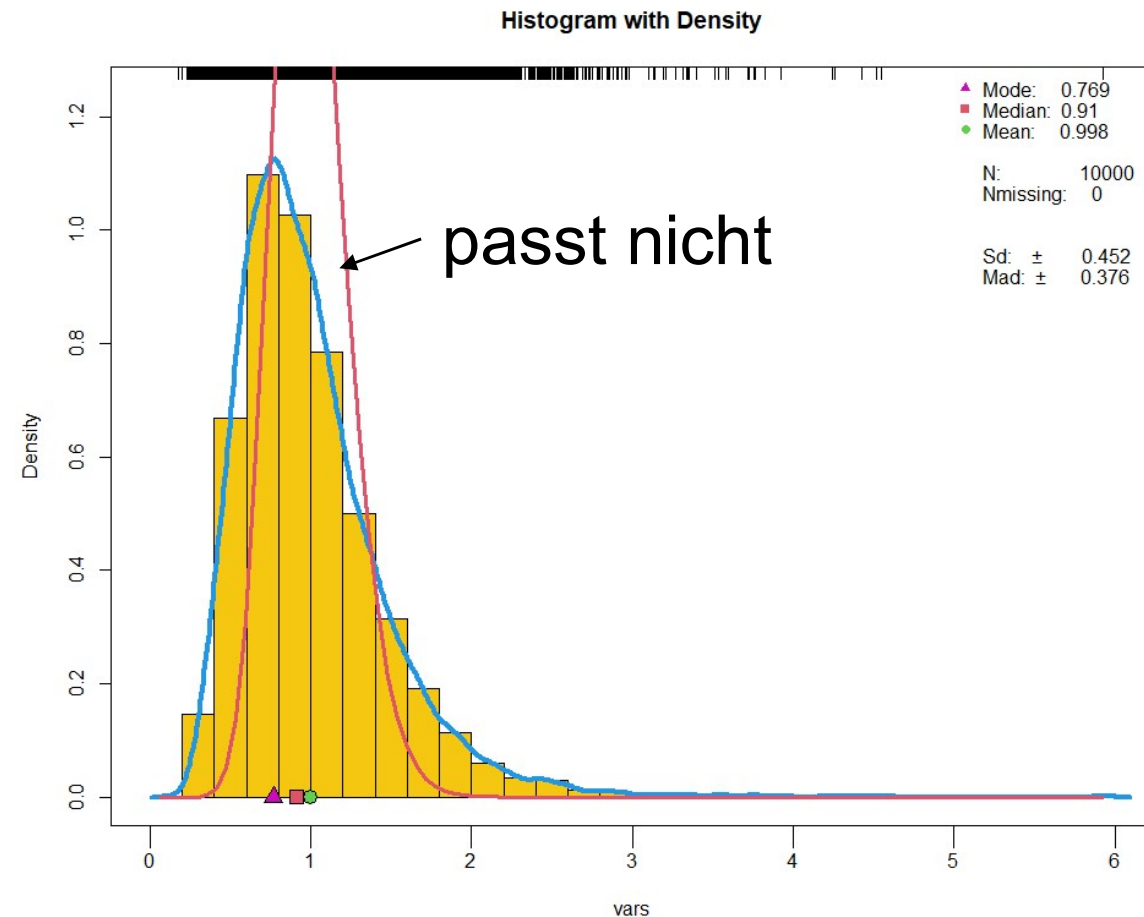
Bei normalverteilten Einzeldaten Konfidenzintervall für die Varianz

- $\hat{\text{var}}(X) = 1/(N-1) * \text{Summe}_{i=1:N} (X_i - \text{Mittelwert})^2$
- Konfidenzintervall für $\hat{\text{var}}(X)$:
- $[\text{qchisq}(0.025, \text{df} = N-1) * \hat{\text{var}}(X) / N),$
- $\text{qchisq}(0.975, \text{df} = N-1) * \hat{\text{var}}(X) / N]$
- wobei qchisq die Quantile der Chi^2 – Verteilung ist.



Achtung: gilt nur für normalverteilte Einzeldaten.

Zum Beispiel: Standardexponentialverteilung



Beachte: $\text{var}(X) = E(X)^2$
bei Exponentialverteilungen

Simulation Konfidenzintervalle mittels t-Verteilung für $N=30$ Stichproben aus $N(0,1)$

**5% irreführende
95% Konfidenzintervalle
sind nach Konstruktion
garantiert!**

