Statistik - Review Session

Julian Koller

5/19/2021

Aufbau

- Fragen Sammeln
- (Eignige) Kernkonzepte der Vorlesung wiederholen
- Repetition nützlicher R Gundlagen
- Weitere Fragerunde

Geläufige Messskalen

- Nomminalskala
- Ordinalskala
- Intervallskala
- Verhältnisskala

$\rightarrow$ Beispiele nennen! (Auf Visualisierungen können wir in der R Repetition zurückkommen)

Wichtigste Masse zentraler Tendenz

- Modus
- Median
- Mittelwert

$\rightarrow$ Welches Mass eignet sich für welche Skala?

Robustheit: Median vs. Mittelwert

Wenn man sich für ein Mass zentraler Tendenz entscheidet, sollte man deren Robustheit im Hinterkopf behalten!

$\rightarrow$ Ist der Median oder der Mittelwert Robuster gegenüber ausreissern? Weshalb?
Optimalitätseigenschaft des Mittelwerts:

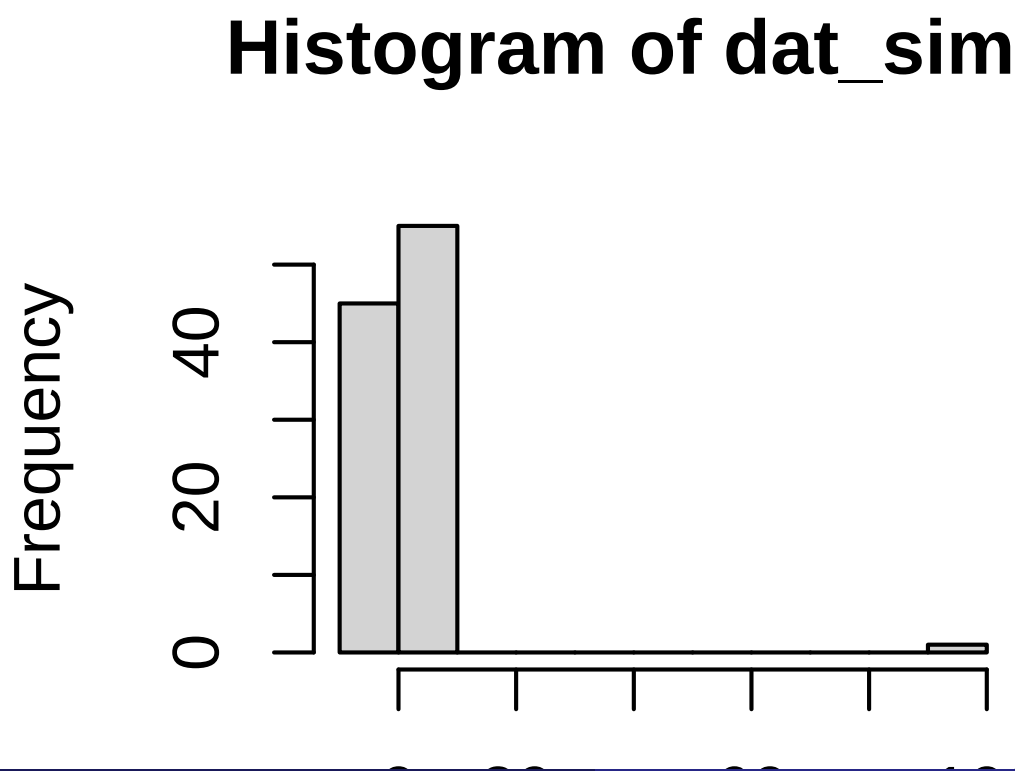
$$arg\ min \sum_{i=1}^{n} (x_i - z)^2 = \bar{x}$$

- Intuition erklären

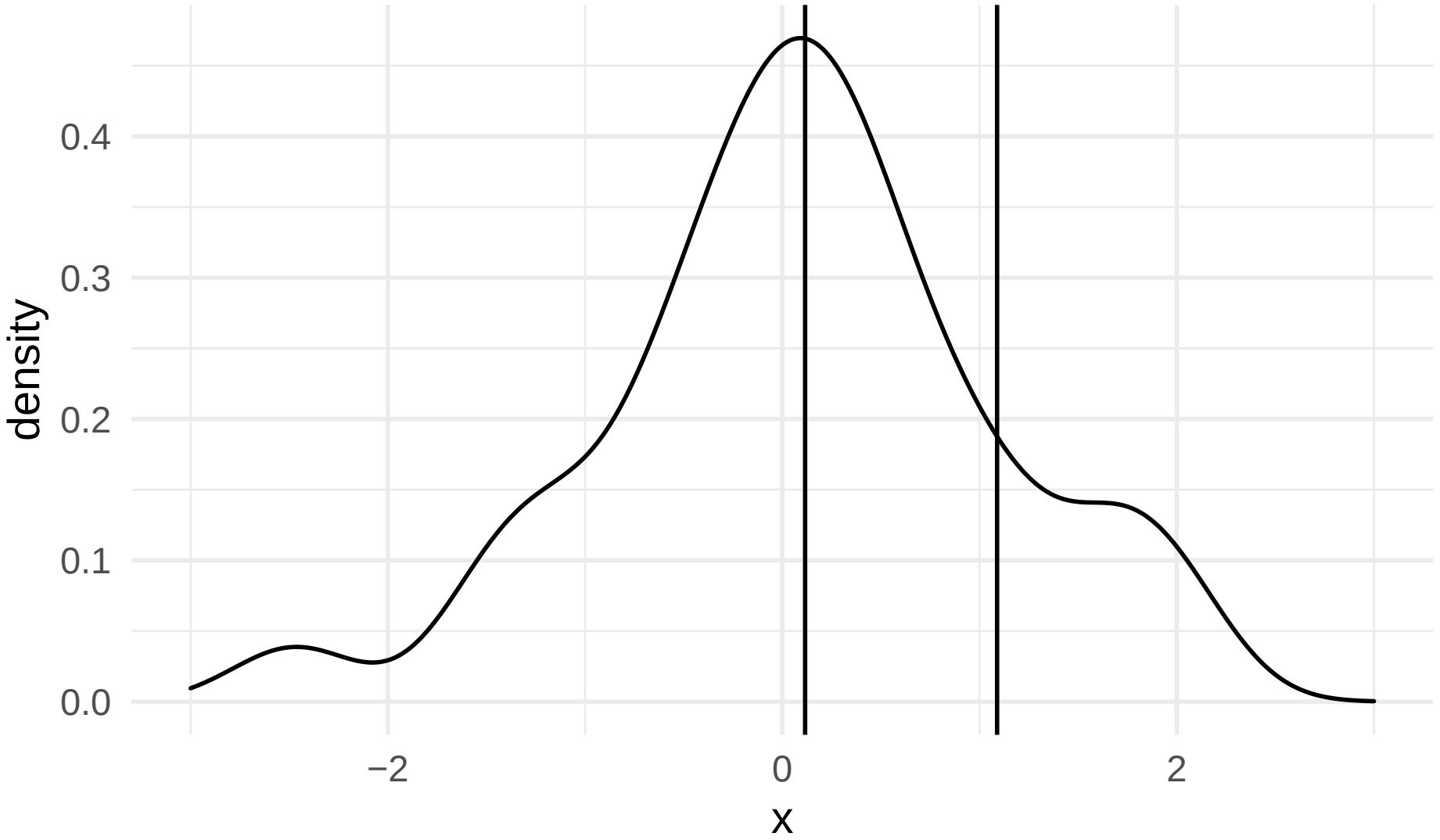
Robustheit: Median vs. Mittelwert

```
library(ggplot2)

dat_sim <- data.frame(x = c(rnorm(100), 100))
hist(dat_sim$x)
```



Robustheit: Median vs. Mittelwert



Welches ist der Mittelwert und Median?

Momente einer Verteilung

- Erwartungswert (empirisch: Mittelwert)
- Varianz
- Schiefe
- Wölbung

$\rightarrow$ Beispiele zeichnen mit Terminologie!

Erwartungswert

$$E(\hat{X}) = \frac{1}{n-1} \sum_{i=1}^{n} x_i$$

Weshalb korrigieren nehmen wir die korrektur um den einen Freiheitsgrad vor?
Wir erinnern uns:

$$arg\ min \sum_{i=1}^{n} (x_i - z)^2 = \bar{x}$$

$\rightarrow$ Intuition zeichnen!

Varianz

- Wir schätzen:

$$V(\hat{X}) = \frac{1}{n-1}(x_i - \bar{x})^2$$

- Würde ohne die Korrektur zwangsläufig unterschätzt (Optimalitätsbedingung) - Weshalb quadrieren wir die Abweichungen?

Test auf Assoziation zweier Kategorialer Variablen:

H0: Keine Assoziation
H1: Assoziation

Idee des Chi-quadrat tests: Vergleich der Häufigkeiten unter H0 mit tatsächlich beobachteten Häufigkeiten.

$$\chi^2_{(R-1)(C-1)} = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

wobei:

$$e_{ij} = \frac{f_i * f_j}{n^2} * n$$

Was ist hier die Idee?
$\rightarrow$ Es gibt viele Zusammenhangsmasse für Kategoriale Variabeln, aber Chi^2 ist einfach afu signifikanz zu testen, aber schwer interpretierbar.

Beispiel aus der Vorlesung

Tabelle 8: Parteiidentifikation und Stimmentscheid zur Hornkuh-Initiative (Zellenhäufigkeiten und erwartete Häufigkeiten (rot) bei statistischer Unabhängigkeit)

	Parteiidentifikation								
	SVP	SP	FDP	CVP	GPS	glp	andere	keine	
Nein	68	86	119	84	15	25	36	119	552
Nein erw.	81.4	102.6	98.8	66.8	34.2	23.3	33.7	111.3	
Ja	82	103	63	39	48	18	26	86	465
Ja erw.	68.6	86.4	83.2	56.2	28.8	19.7	28.3	93.7	
	150	189	182	123	63	43	62	205	1017

Quelle: VOTO (ungewichtete Werte).

Figure 1: Beispiel aus der Vorlesung geklaut

Zusammenhänge numerischer Variablen

- Intuitive Quantifizierung einer Assoziation: Kovarianz!

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- Was ist die Intuition hinter dieser Formel? - Was ist ein Nachteil dieses Zusammenhangsmasses?

Korrelation: Eine Standardisierung der Kovarianz

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_Y^2 \sigma_X^2}$$

Wir wissen schon, wie man alle Bestandteile dieser Formel schätzt. Wir haben also:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})}$$

$\rightarrow$ Hat den entscheidenden Vorteil, dass -1 $<$ r $<$ 1

$\rightarrow$ Leicht interpretierbar!

Z-Standardisierung

- Z Standardiesierung ist ein Weg, numerische Variablen aller Skalen vergleichbar zu machen.

$$Z = \frac{x - \mu}{\sigma}$$

Wird aus den Daten geschätzt als:

$$Z = \frac{x - \bar{x}}{\hat{\sigma}}$$

Stellt sicher, dass Variablen Mittelwert 0 und Varianz 1 aufweisen.
rightarrow Kovarianz zweier Z-standardiserter Variablen ist gleich der Korrelation der original skalierten Variablen!

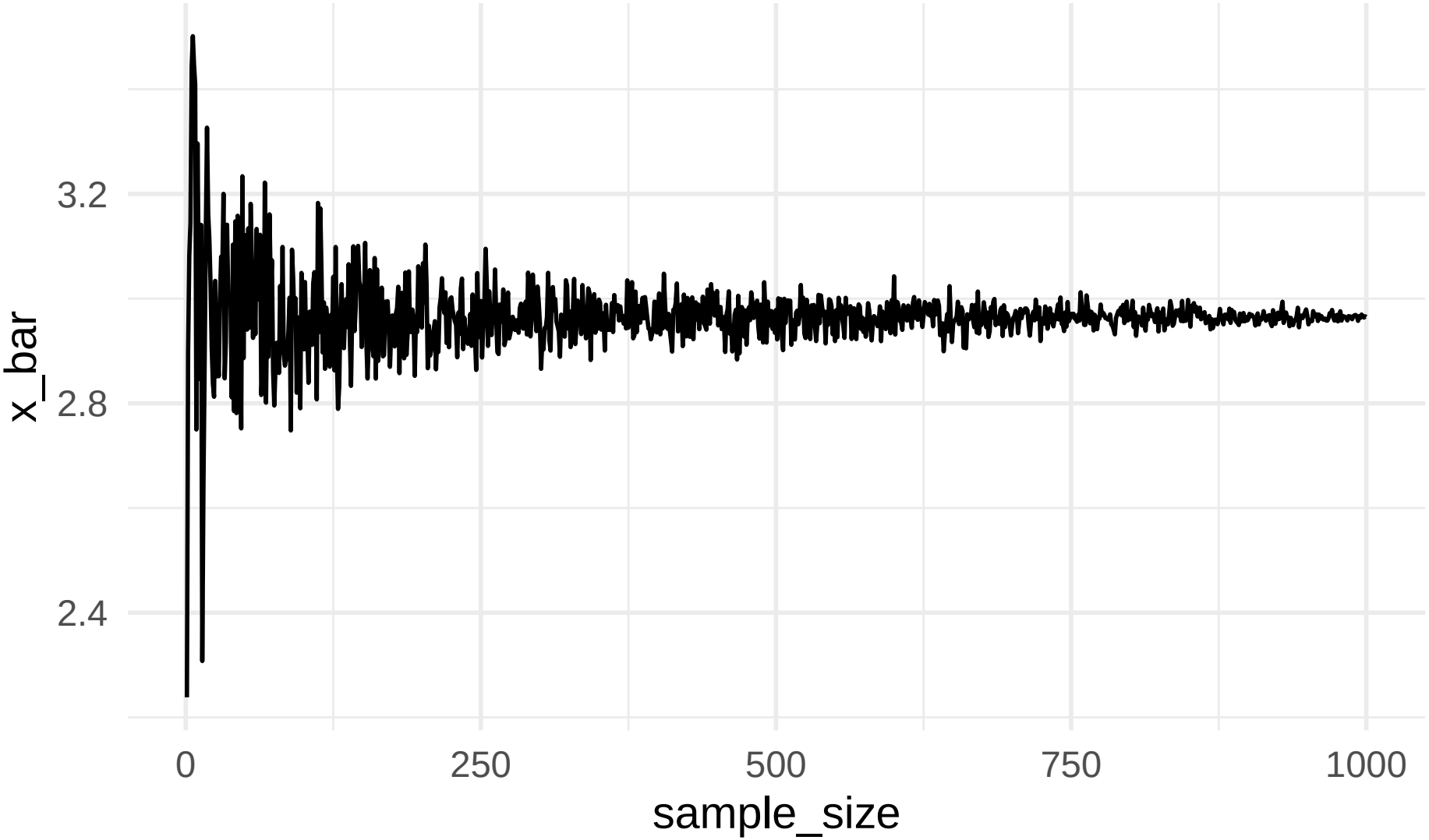
Wahrscheinlichkeiten: Das Gesetz der Grossen Zahlen

In der Statistik ist man sich eines Ergebnisses immer sicherer, wenn es auf mehr beobachtungen beruht. Dergrund dafür: Das Gesetz der grossen Zahlen.
Kurze Demonstration anhand einer Simulation.

```
set.seed(2251)
x <- rnorm(1000, mean = 3, sd = 1)
x_bar <- c()
for (i in 1:1000) {
  sample <- x[sample(1:length(x), size = i)]
  x_bar[i] <- mean(sample)
}
```

Was macht dieser Code genau und wie?

Wahrscheinlichkeiten: Das Gesetz der Grossen Zahlen



Je mehr beobachtungen wir zur Hand haben, umso genauer unsere Schätzung. Dies gilt auch für (frequentistische) Wahrscheinlichkeiten!

Ein Paar Grundregeln der Wahrscheinlichkeitsrechnung:

Was ist bedingte Wahrscheinlichkeit?

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Unter Unabhängigkeit von A und B, sollte, P(A|B) = P(A) sein. (Siehe erwartete Häufigkeiten des Chi^2)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Der Satz von Bayes

Wenn wir $P(A|B)$ haben, wie können wir dann auf $P(B|A)$ schliessen?

$$P(B|A) = \frac{P(A \cap B)}{P(B)} = \frac{P(A|B)P(A)}{P(B)}$$

$\rightarrow$ Falls Nachfrage vorhanden: Beispiel auf Folie 55 Vorlesung 6 kurz lösen!

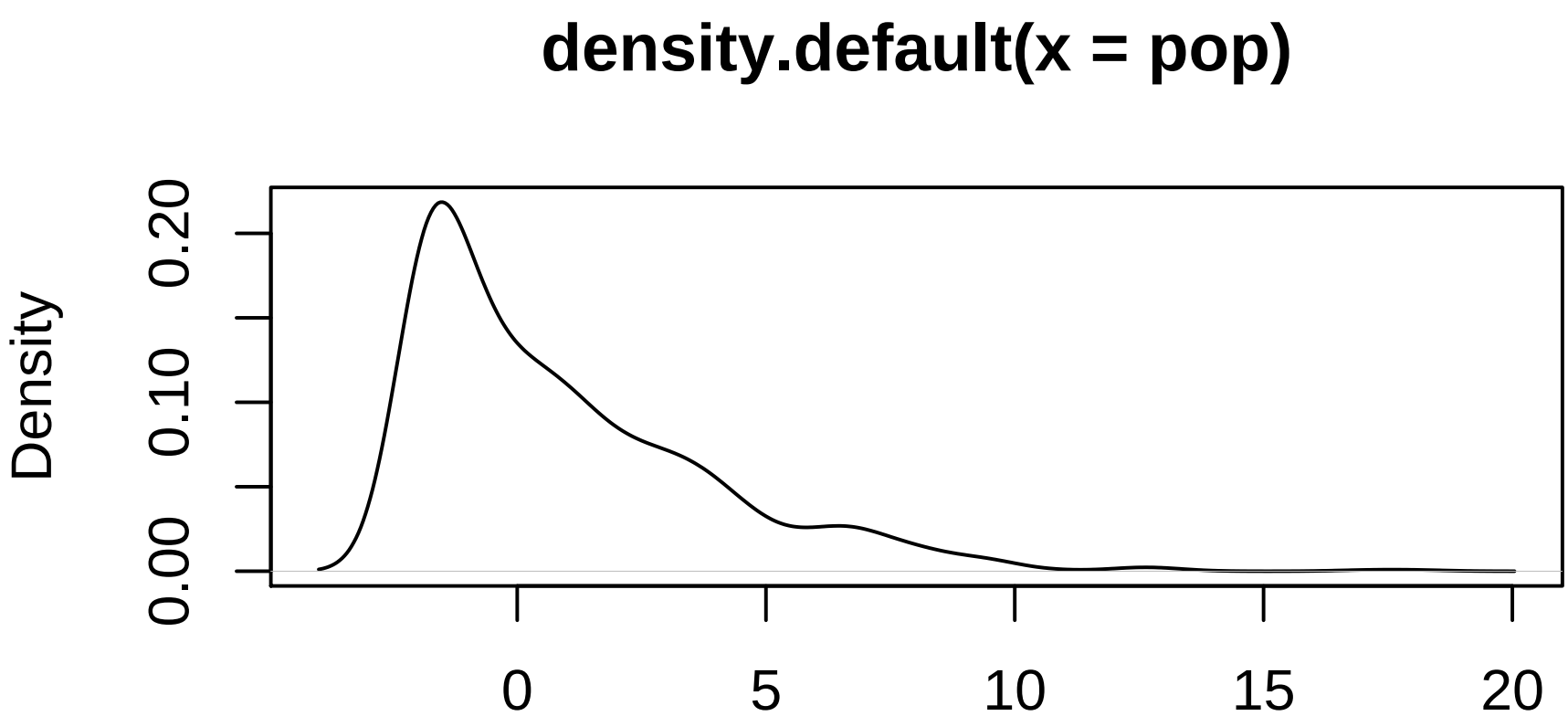
Statistische Inferenz: Zentraler Grenzwertsatz

- Fundament der statistischen Inferenz!
- Wenn wir einen Parameter einer Grundgesamtheit mit einer Stichprobe schätzen, so ist der Schätzer bei ausreichender Stichprobengrösse annähernd Normalverteilt.
- Dies erlaubt uns, die Unsicherheit eines Schätzers zu quantifizieren!
- Zum Verständnis: Nochmals eine Demonstration mittels Simulation...

Zentraler Grenzwertsatz Simulationsbeispiel:

```r
# Angenommen, wir haben eine Grundpopulation von 1000:
pop <- rchisq(n = 1000, df = c(5, 1))-2

plot(density(pop))
```



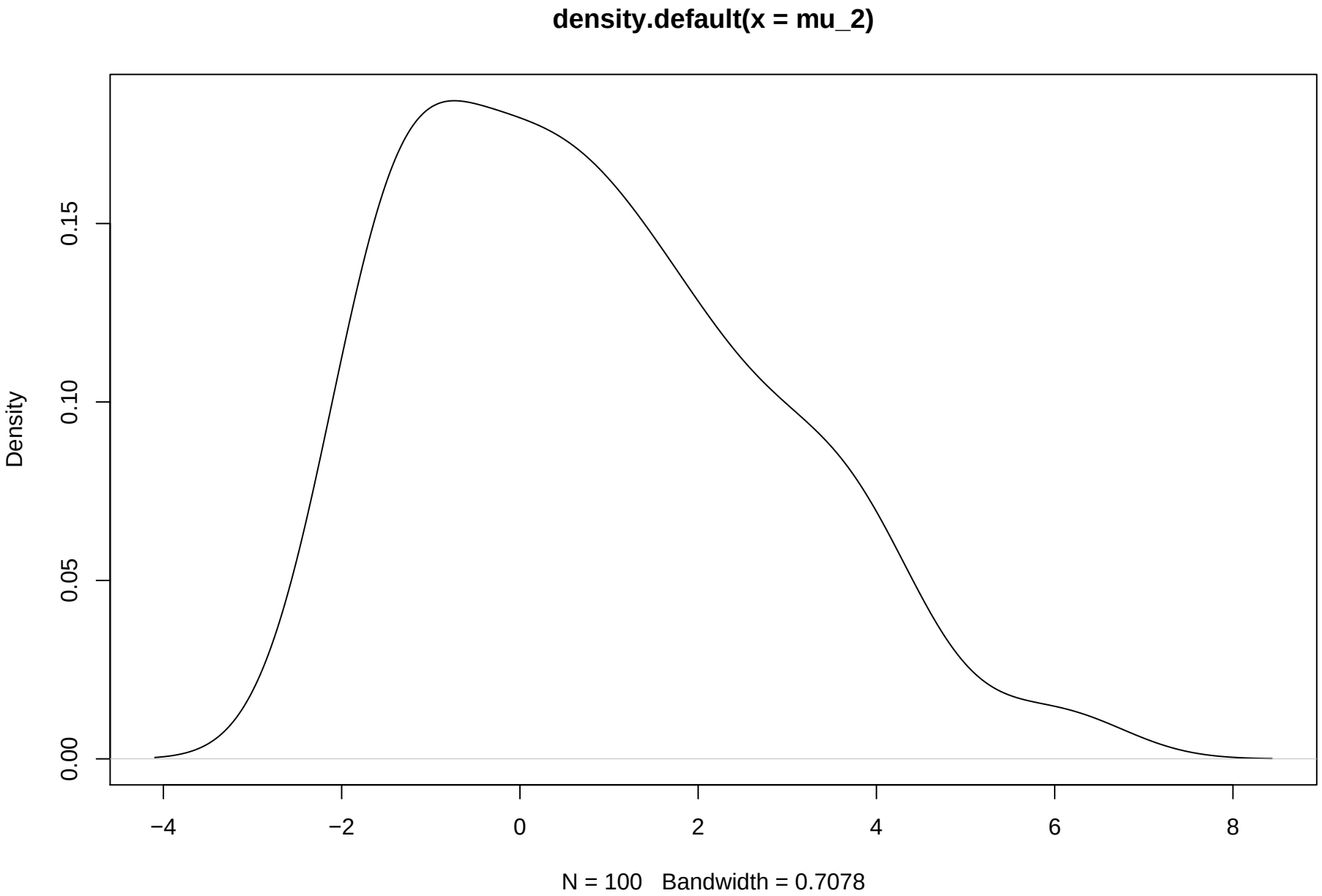
Simulation Fortsetzung

→ Eindeutig nicht normalverteilt!

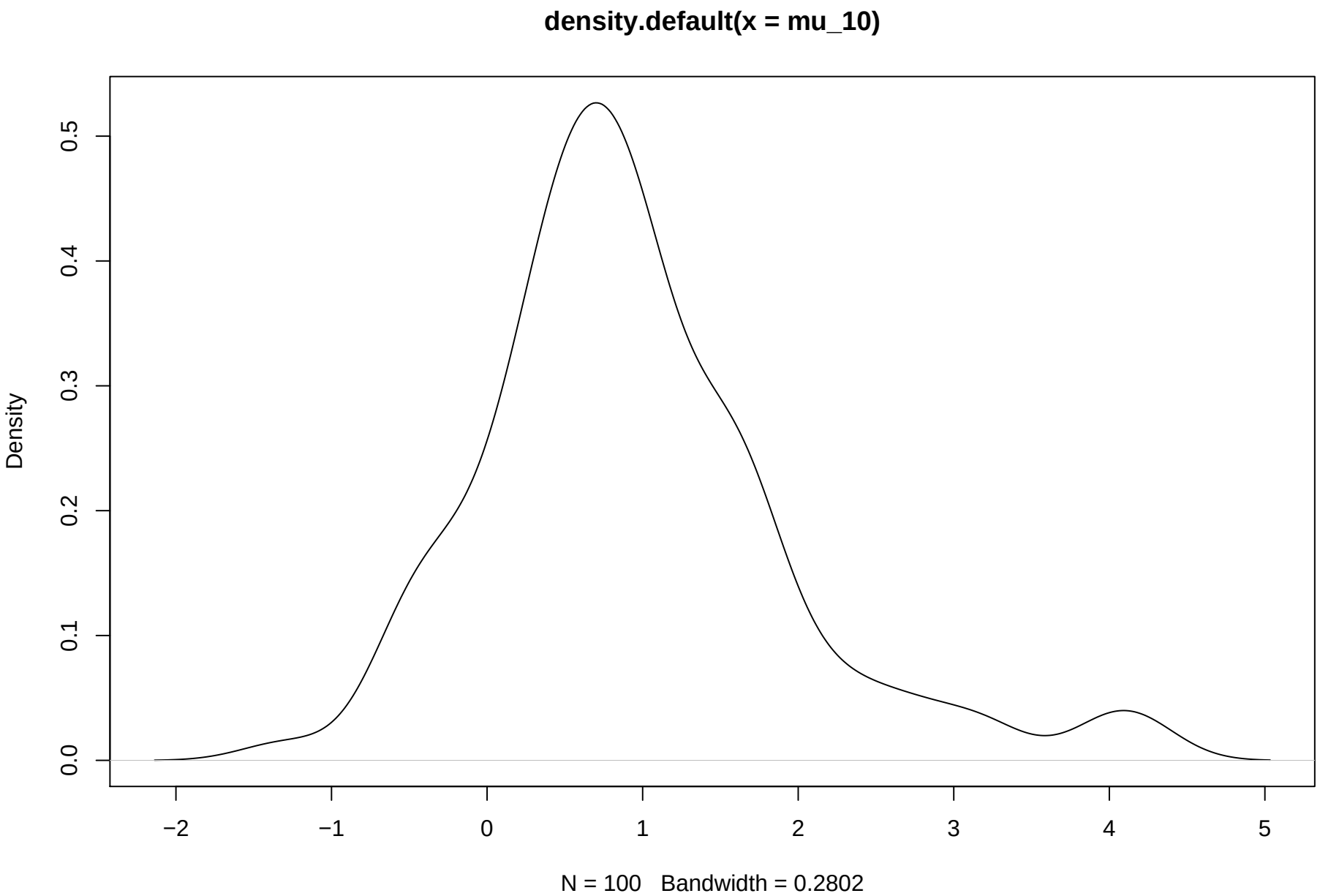
- Wir können uns die Verteilung eines geschätzten Parameters (Mittelwert) under verschiedenen Stichprobengrössen anschauen:

```
mu_2 <- c()
for (i in 1:100) {
  mu_2[i] <- mean(sample(pop, size = 2))
}
mu_10 <- c()
for (i in 1:100) {
  mu_10[i] <- mean(sample(pop, size = 10))
}
mu_30 <- c()
for (i in 1:100) {
  mu_30[i] <- mean(sample(pop, size = 30))
}
mu_200 <- c()
```

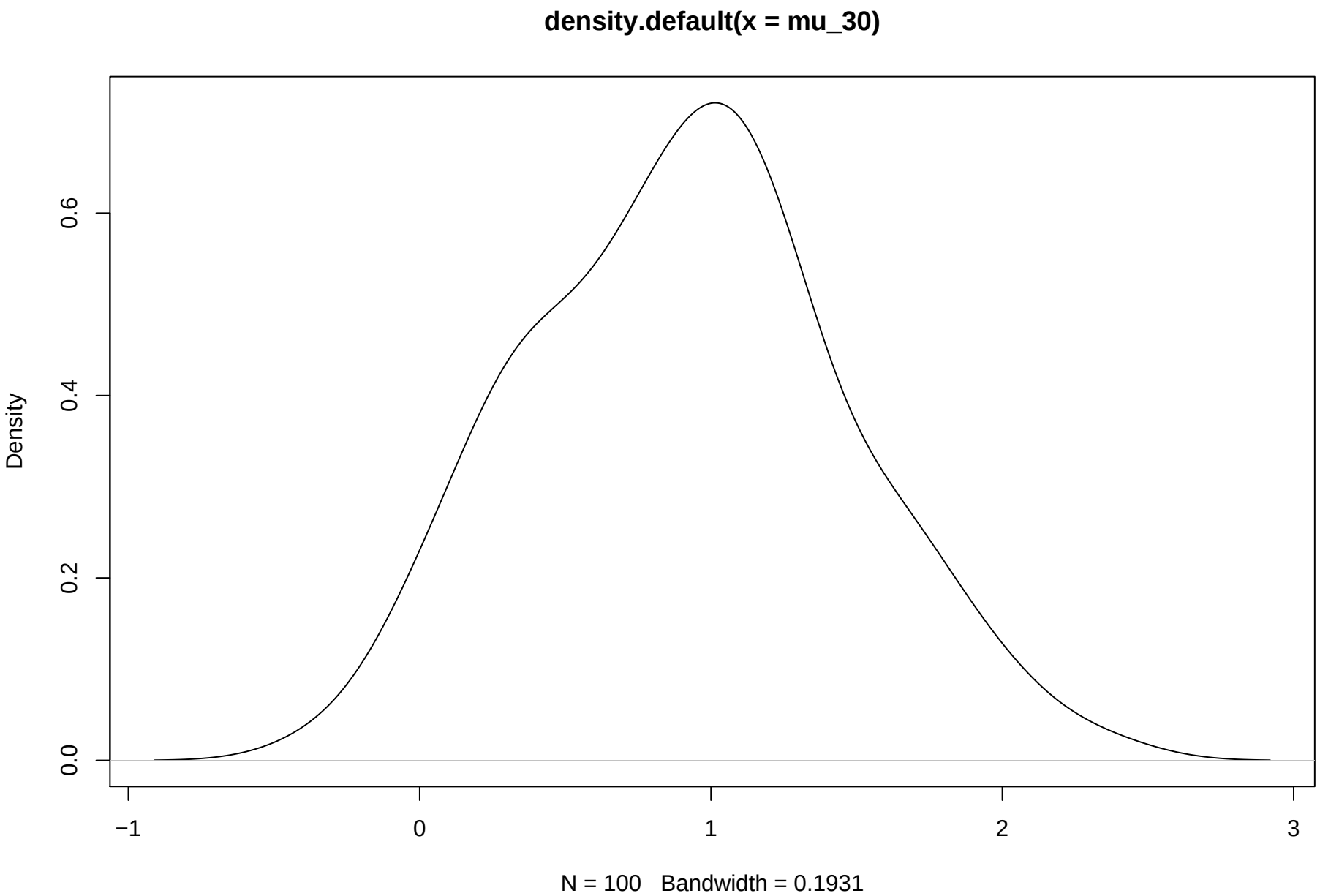
Wir beobachten:



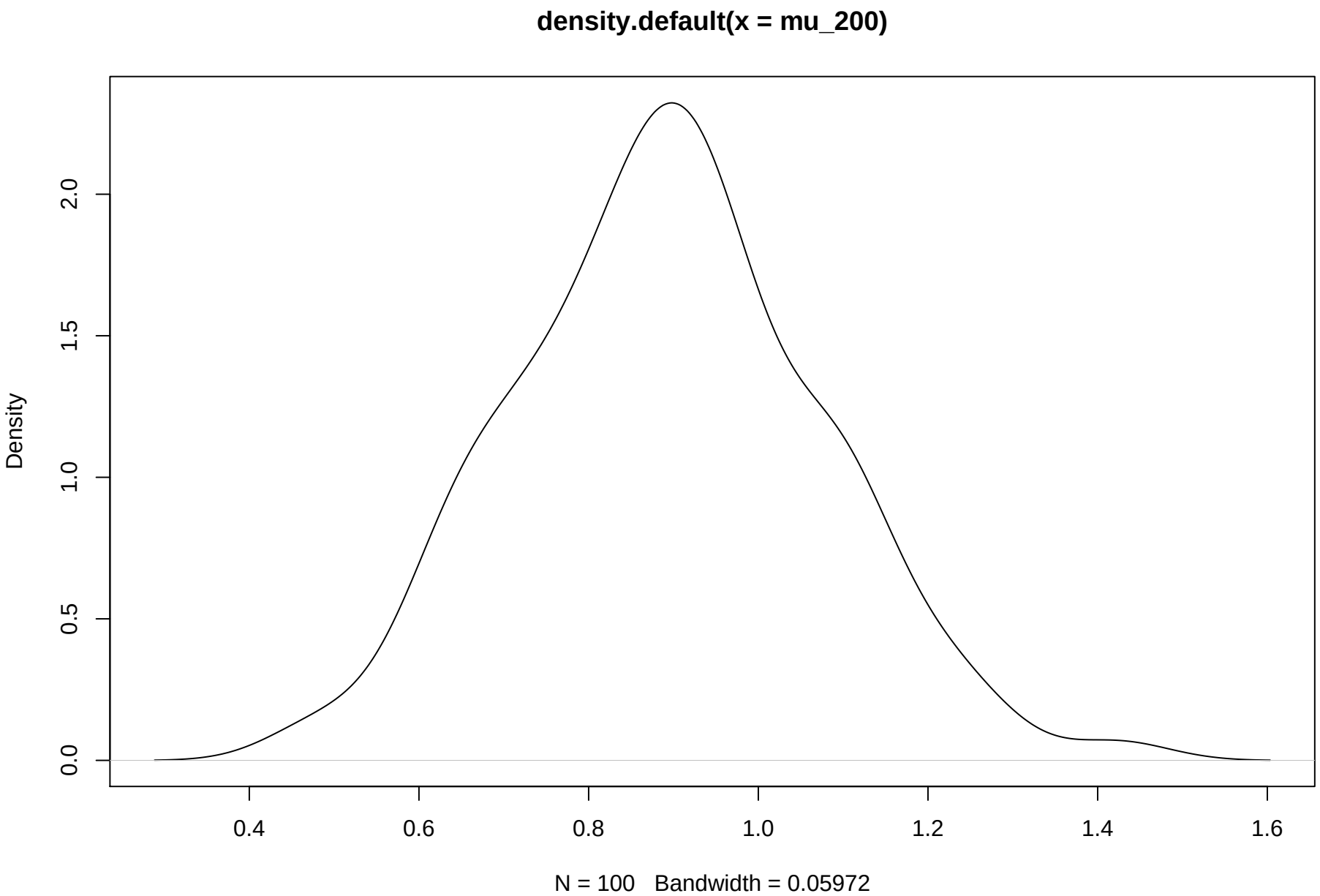
Wir beobachten:



Wir beobachten:



Wir beobachten:



Warum ist das nützzlich?

- Wir sehen (hoffentlich), dass die Verteilung der Schätzer sich mit zunehmender Stichprobengrösse einer Normalverteilung annähert.
- Dies ist die theoretische Verteilung des Schätzers
- Wir können Quantile der theoretischen Verteilung nutzen, um die Unsicherheit des Schätzers zu quantifizieren
- Dies wird uns erlauben, Hypothesentests durchzuführen.

Hypothesentests

- Nehmen wir dieselbe Population als Beispiel.
- Wir wollen testen $H_0 : \mu = 0$ gegen $H1 : \mu \neq 0$
- Gedankenexperiment: Was für eine Verteilung des Schätzers würden wir erwarten, wenn H_0 wahr wäre?
- Eine Normalverteilung mit Mittelwert 0! (Zentraler Grenzwertsatz)

Kurzer Reminder an die Z-Standardisierung

- Z Standardiesierung ist ein Weg, numerische Variablen aller Skalen vergleichbar zu machen.

$$Z = \frac{x - \mu}{\sigma}$$

Wird aus den Daten geschätzt als:

$$Z = \frac{x - \bar{x}}{\hat{\sigma}}$$

Stellt sicher, dass Variablen Mittelwert 0 und Varianz 1 aufweisen.
$\rightarrow$ Kovarianz zweier Z-standardiserter Variablen ist gleich der Korrelation der original skalierten Variablen!

Hypothesentest Fortsetzung

- Wir können uns die Frage stellen: Auf welcher Quantile der Z-standardisierten theoretischen Verteilung des Schätzers unter H_0 liegt unser geschätzter Parameter (Mittelwert).
- Um dies zu berechnen brauchen wir die Standardabweichung der theoretischen Verteilung des Schätzers
- Dies ist genau die Interpretation des Standardfehlers!

$$SE = \sqrt{\frac{\sigma X}{n}}$$

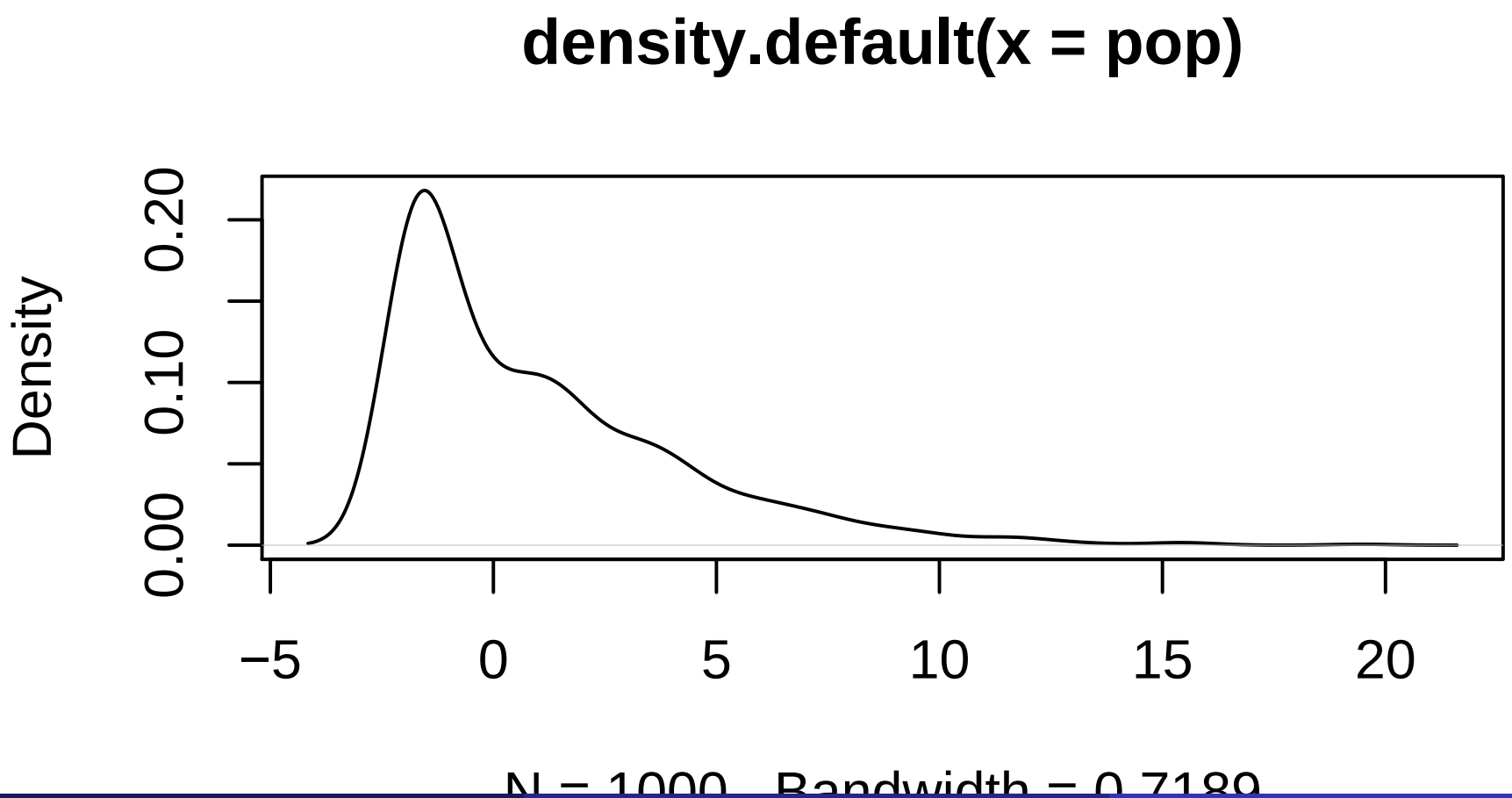
$$Z = \frac{\hat{\mu} - \mu_{H0}}{SE}$$

Interpretation des Z-Werts

- Der Z-Wert gibt an, wie viele Standardabweichungen vom Mittelwert der Verteilung des Schätzers under H0 unser geschätzter Parameter entfernet liegt.
- Je grösser der absolute Z-Wert, umso extremer wäre dieser Parameter unter H0
- Unter H0 wären hohe Z-Werte unwahrscheinlich zu beobachten
- Dank des zentralen Grenzwertsatzes können wir quantifizieren, wie unwahrscheinlich genau! (P-Wert)
- Ab einem gewissen "Grad der Unwahrscheinlichkeit", einen Z-Wert unter H0 zu beobachten verwerfen wir H0!

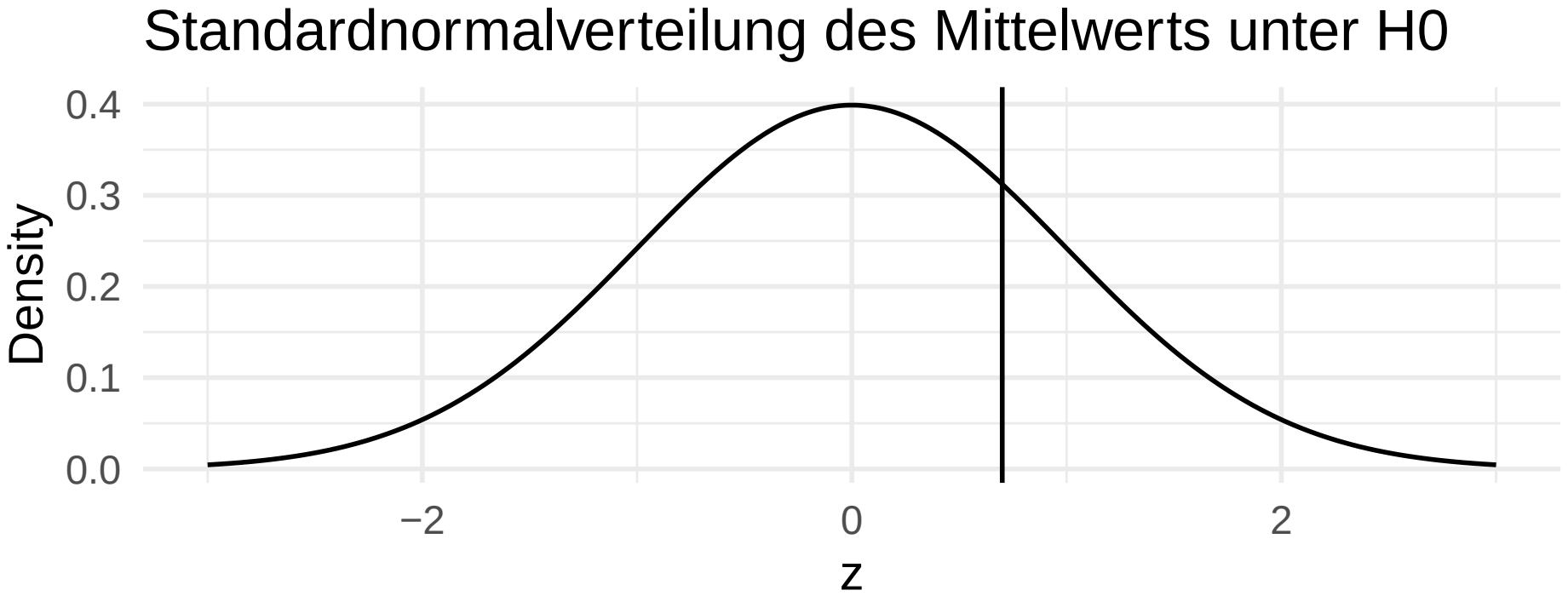
Beispiel

- Nehmen wir zum Beispiel unseren Schätzer mit 100 beobachtungen oben. Wir wollen testen, ob sich dessen Mittelwert signifikant von 0 unterscheidet. Wir erinnern uns an die Verteilung der Grundpopulation:

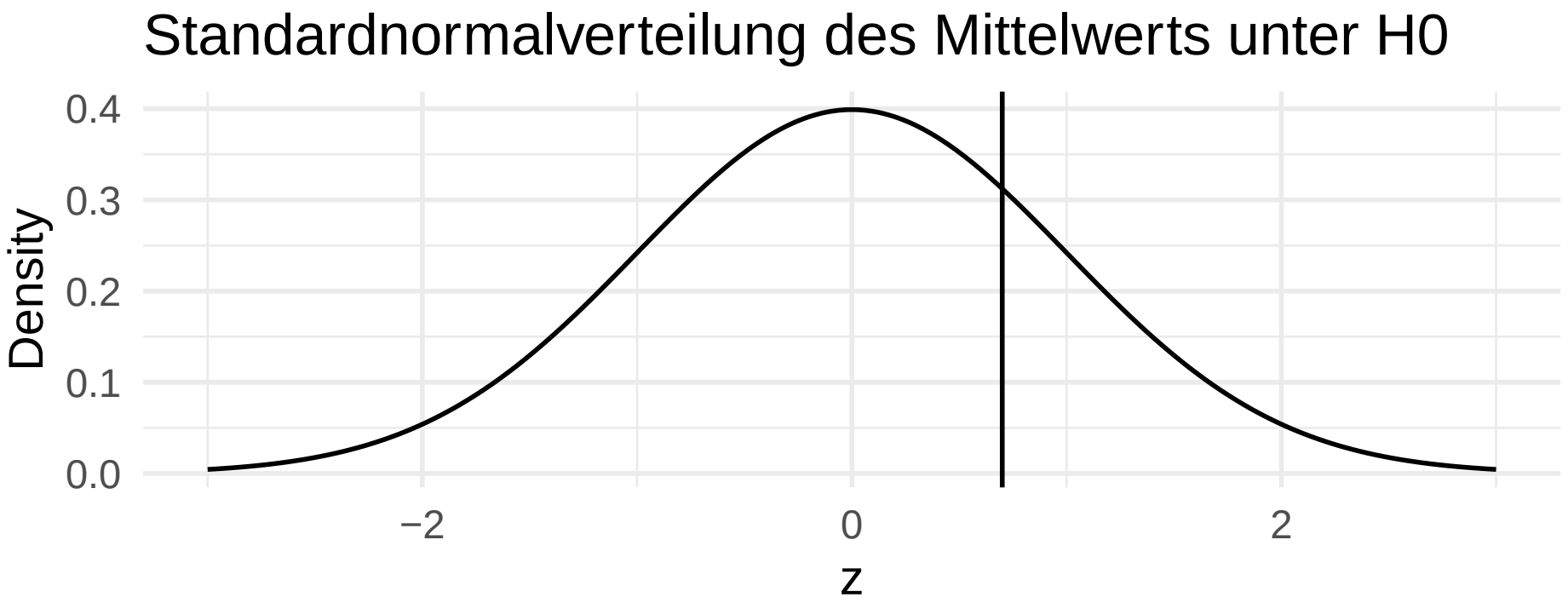


Beispiel Fortsetzung

Von Auge lässt sich kaum abschätzen, ob der Mittelwert sich von 0 unterscheidet.
Für einen Test von $H_0 : \mu = 0$ und $H_1 : \mu \neq 0$ ziehen wir 35 stichproben und erhalten einen Mittelwert von 0.7. Dies führt zu einem Z-Wert von ca. 1.4.



vom Z-Wert über den P-wert zur Entscheidung des Tests



Wie identifizieren wir hier den P-Wert und was ist die Logik dahinter? → Intuition zeichnen!

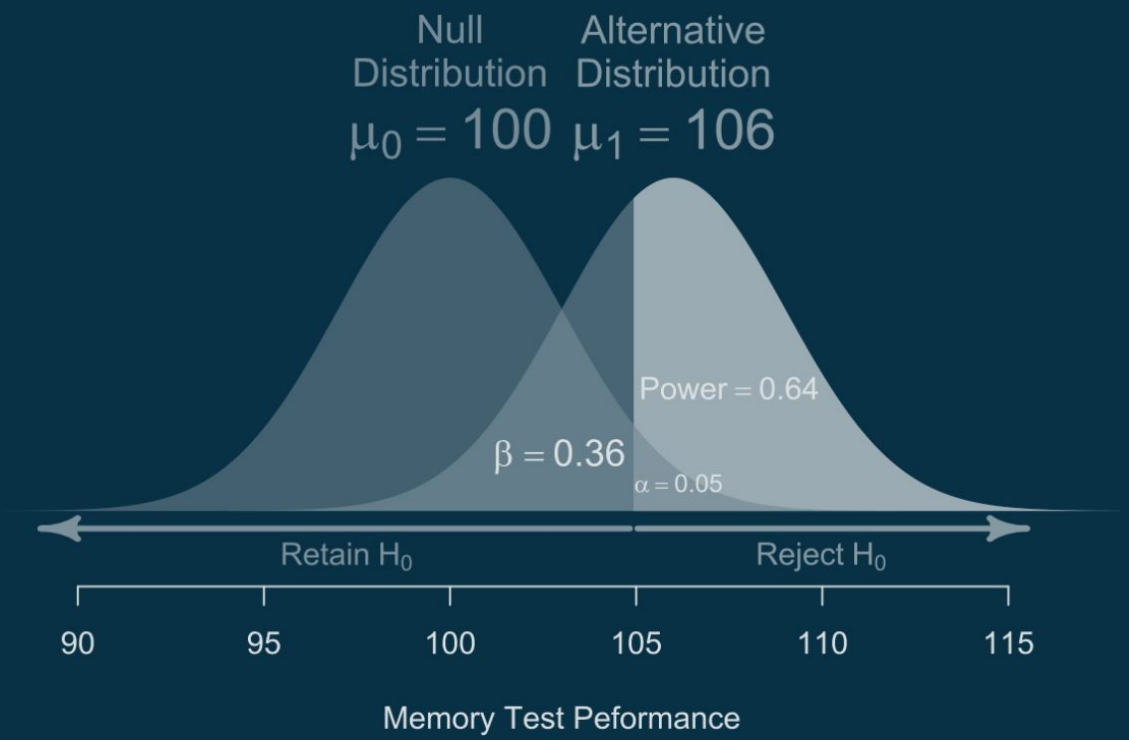
Konfidenzintervalle

- Wenn man die vorhergehende Logik des Hypothesentests verstanden hat, sind KIs einfach!
- Kehrseite derselben probabilistischen Münze
- Wir können um unseren Schätzer oder um den Nullwert ein theoretisches Konfidenzintervall berechnen. Wenn der jeweils andere Wert für ein bestimmtes Konfidenzniveau nicht im Intervall enthalten ist, kann man schliessen, dass ein statistisch signifikanter Unterschied zwischen den beiden besteht.
- Diese Berechnung führt zum selben Resultat, wie die Vorgehensweise oben.
- Nochmals Intuition zeichnen!

Fehler 1. und 2. Art

Unser probabilistisches Modell kann auf zwei arten falsch liegen:

Abbildung 5: Güte eines Tests: Beispiel



Quelle: `http://my.ilstu.edu/~wjschne/138/Psychology138Lab14.html`

Schwere Entscheidung...

- Wir können H0 fälschlicherweise verwerfen (1. Art) oder fälschlicherweise nicth verwerfen (2. Art).
- Bei der Wahl des Signifikanzniveaus sind wir mit einem Trade-off zwischen den beiden Fehlerarten konfrontiert. Weshalb? Zeichnen!
- Welchwe Fehlerart man minimieren will hängt in der Regel stark von der Anwendung ab. In gewissen Gebieten ist ein Fehler 1. Art besonder unerwünscht, in anderen ist eine möglichst hohe Güte wichtig.
- Angenommen, wir versuchen statistisch das Auftreten von Bürgerkriegen vorherzusagen. Wie würden die obigen Ausführungen die Wahl des Konfidenzniveaus beinflussen?

Weshalb eine t-Verteilung verwenden?

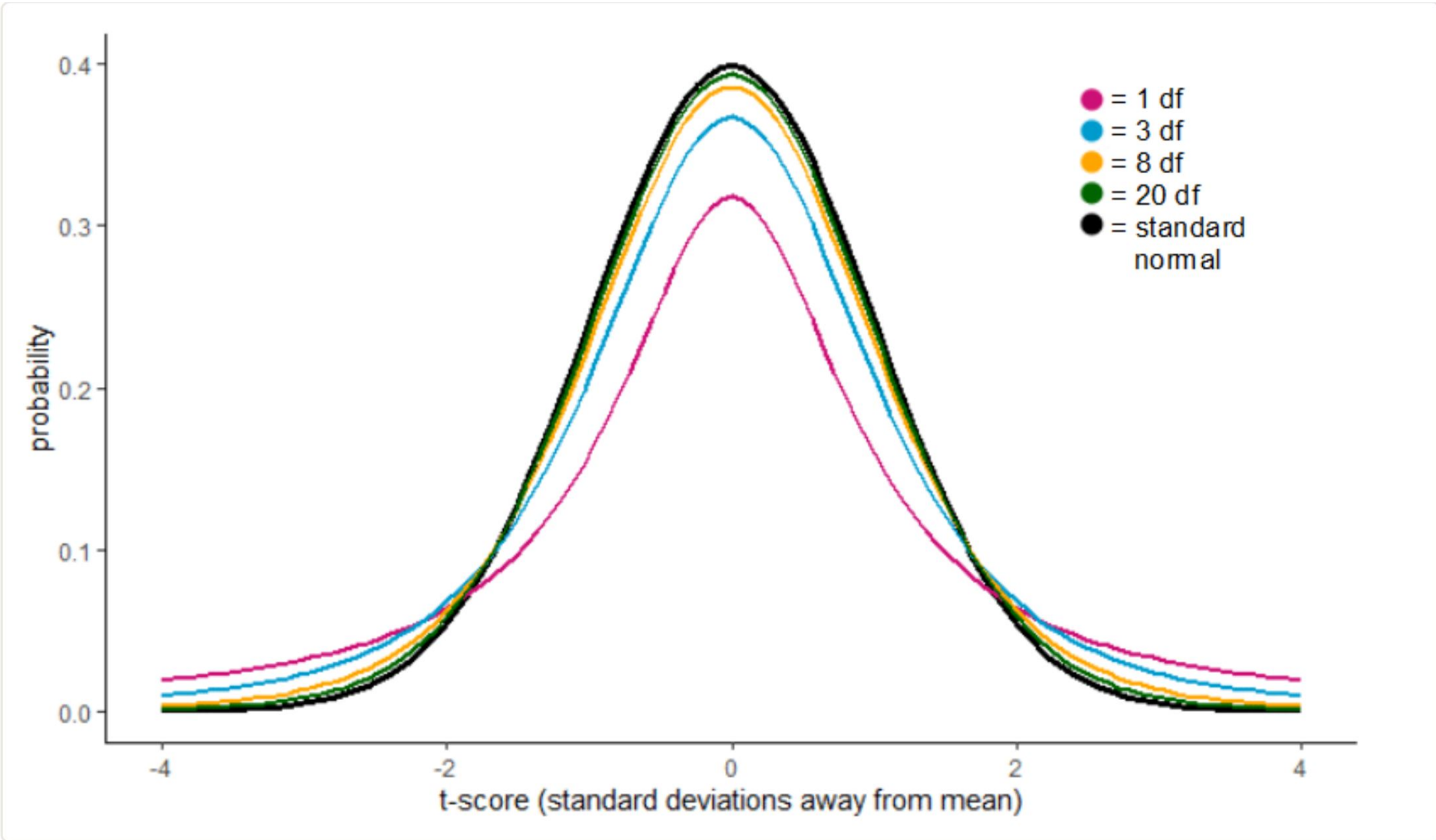


Figure 2: Die Wölbung macht es aus! Erhöhte Wahrscheinlichkeit extremer Werte unter der t-Verteilung

Übergang

Je nach dem wie viel Zeit noch vorhanden –> Ins R wechseln zur Repetition nützlicher Grundlagen