

# 8. Vorlesung Statistik I

## Intervallschätzung II



We are happy to share our materials openly:

The content of these Open Educational Resources by Lehrstuhl für Psychologische Methodenlehre und Diagnostik, Ludwig-Maximilians-Universität München is licensed under CC BY-SA 4.0. The CC Attribution-ShareAlike 4.0 International license means that you can reuse or transform the content of our materials for any purpose as long as you cite our original materials and share your derivatives under the same license.

- Bislang (letzte Vorlesung):
  - Motivation und Grundlagen der Intervallschätzung
  - Konfidenzintervalle und Gütekriterien
  - Konfidenzintervalle für den Parameter  $\mu$  einer Normalverteilung
- Jetzt:
  - Konfidenzintervalle für den Parameter  $\pi$  einer Bernoulli-Verteilung

# Konfidenzintervalle für den Parameter $\pi$ einer Bernoulli- Verteilung

# Wdh.: Konstruktion eines Konfidenzintervalls für $\mu$

- Wir sind von einer geeigneten Schätzfunktion für den Parameter  $\mu$  ausgegangen:

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Wir haben diese Schätzfunktion z-standardisiert, dann für alle unbekannten Parameter im Nenner die entsprechende Schätzfunktion eingesetzt und so eine Zufallsvariable  $T$  erhalten:
  - z-Standardisierung:

$$Z = \frac{\bar{X} - E(\bar{X})}{SD(\bar{X})} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

- Ersetzen von  $\sigma^2$  durch die Schätzfunktion  $S^2$ :

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}}$$

- Es hat sich herausgestellt, dass die Wahrscheinlichkeitsverteilung  $P_T$  dieser Zufallsvariable  $T$  genau bestimmbar ist, auch wenn der Wert des Parameters  $\mu$  selbst nicht bekannt ist:

$$T \sim t(n - 1)$$

- Wir haben auf Basis dieser t-Verteilung die Quantile  $t_{\frac{\alpha}{2}}$  und  $t_{1-\frac{\alpha}{2}}$  berechnet, so dass

$$P\left(t_{\frac{\alpha}{2}} \leq T \leq t_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

- Wir haben  $T$  eingesetzt und die Ungleichungen nach  $\mu$  aufgelöst:

$$P\left(\bar{X} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}\right) = 1 - \alpha$$

- Dies hat das zufällige Konfidenzintervall

$$I(X_1, \dots, X_n) = [U, O] = \left[ \bar{X} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}} \right]$$

und das konkrete in unserer Stichprobe berechenbare Konfidenzintervall

$$I(x_1, \dots, x_n) = [u, o] = \left[ \bar{x} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}, \bar{x} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}} \right]$$

ergeben.

# Konstruktion eines Konfidenzintervalls für $\pi$

- Ausgangssituation:  $X_i \stackrel{\text{iid}}{\sim} Be(\pi)$  für  $i = 1, 2, \dots, n$
- Ziel: Konstruktion eines Konfidenzintervalls  $I(X_1, \dots, X_n)$  mit Konfidenzniveau  $1 - \alpha$  mit minimaler erwarteter Länge für den Parameter  $\pi$ .
- Wir werden versuchen, das Vorgehen bei der Konstruktion des Konfidenzintervalls für  $\mu$  auf die Konstruktion eines Konfidenzintervall für  $\pi$  zu übertragen.

## Schätzfunktion für $\pi$

- Wir wissen, dass

$$\hat{\pi} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

eine erwartungstreue, effiziente und konsistente Schätzfunktion für  $\pi$  ist.

## z-Standardisierung der Schätzfunktion

- Wir wissen außerdem (aus VL6), dass

$$E(\bar{X}) = \pi$$

$$SD(\bar{X}) = SE(\bar{X}) = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

- Als z-Standardisierung von  $\bar{X}$  ergibt sich daher:

$$Z = \frac{\bar{X} - E(\bar{X})}{SD(\bar{X})} = \frac{\bar{X} - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$$

## Ersetzen der unbekannten Parameter im Nenner

- Wir ersetzen  $\pi$  im Nenner durch die entsprechende Schätzfunktion  $\hat{\pi} = \bar{X}$  :
  - statt

$$Z = \frac{\bar{X} - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$$

- jetzt

$$Z^* = \frac{\bar{X} - \pi}{\sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}} = \frac{\bar{X} - \pi}{\sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}}$$

- Bemerkung: Das  $\pi$  im Zähler muss stehen bleiben, da wir ja später nach diesem auflösen wollen.

- Anders als beim Konfidenzintervall für  $\mu$  gilt nun ja  $X_i \stackrel{\text{iid}}{\sim} Be(\pi)$  und **nicht mehr**  $X_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$  und entsprechend folgt  $\bar{X}$  auch **keiner Normalverteilung** mehr.
- Problem: Deshalb ist die Wahrscheinlichkeitsverteilung von

$$Z^* = \frac{\bar{X} - \pi}{\sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}}$$

**nicht bekannt.**

- Lösung: Wir bestimmen die **approximative Wahrscheinlichkeitsverteilung** von  $Z^*$ .

# Approximative Wahrscheinlichkeitsverteilungen

## Definition: Approximative Wahrscheinlichkeitsverteilung

- Sei  $X$  eine Zufallsvariable mit Träger  $T_X$  und Wahrscheinlichkeitsverteilung  $P_X$  und Verteilungsfunktion  $F$ .
- Sei  $P_X^*$  eine weitere Wahrscheinlichkeitsverteilung mit Verteilungsfunktion  $F^*$ .
- Man sagt dann, dass  $P_X^*$  eine **approximative Wahrscheinlichkeitsverteilung** von  $X$  ist, falls

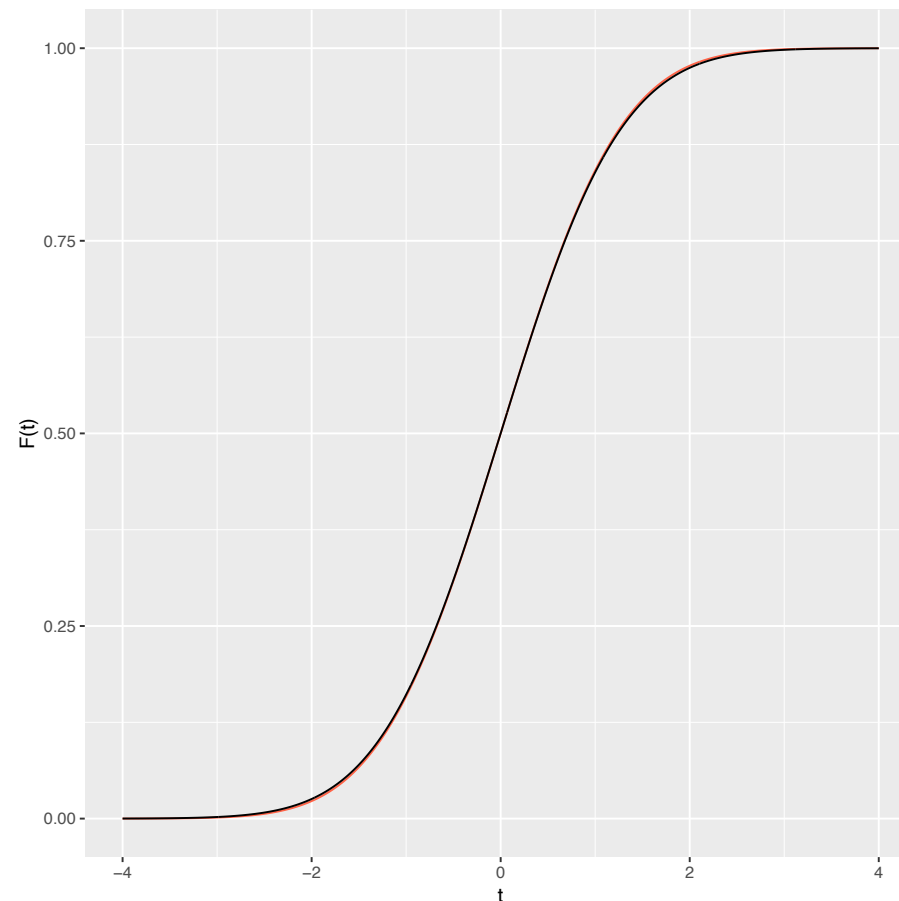
$$F^*(x) \approx F(x) = P(X \leq x)$$

für alle  $x \in T_X$  gilt (vereinfachte Definition).

- Das heißt: Falls wir die Wahrscheinlichkeit  $P(X \leq x)$  mithilfe der „falschen“ Verteilungsfunktion  $F^*$  berechnen, liegt diese Wahrscheinlichkeit sehr nahe an der tatsächlichen Wahrscheinlichkeit, die wir mithilfe der tatsächlichen Verteilungsfunktion  $F$  von  $X$  berechnen würden.
- Wir schreiben in diesem Fall auch:  $X \stackrel{a}{\sim} P_X^*$

## Beispiel I

- Die Standardnormalverteilung ist eine approximative Wahrscheinlichkeitsverteilung einer Zufallsvariable deren tatsächliche Wahrscheinlichkeitsverteilung eine t-Verteilung mit hohem  $\nu$  ist:



rot:  $N(0, 1)$

schwarz:  $t(50)$

- Vergleich einiger Werte der Verteilungsfunktionen:

- $P(T \leq 1)$ :

```
> pnorm(1, 0, 1)
[1] 0.8413447
```

←  $F^*(1) \approx P(T \leq 1)$

```
> pt(1, 50)
[1] 0.8389372
```

←  $F(1) = P(T \leq 1)$

- $P(T \leq -0.5)$ :

```
> pnorm(-0.5, 0, 1)
[1] 0.3085375
```

←  $F^*(-0.5) \approx P(T \leq -0.5)$

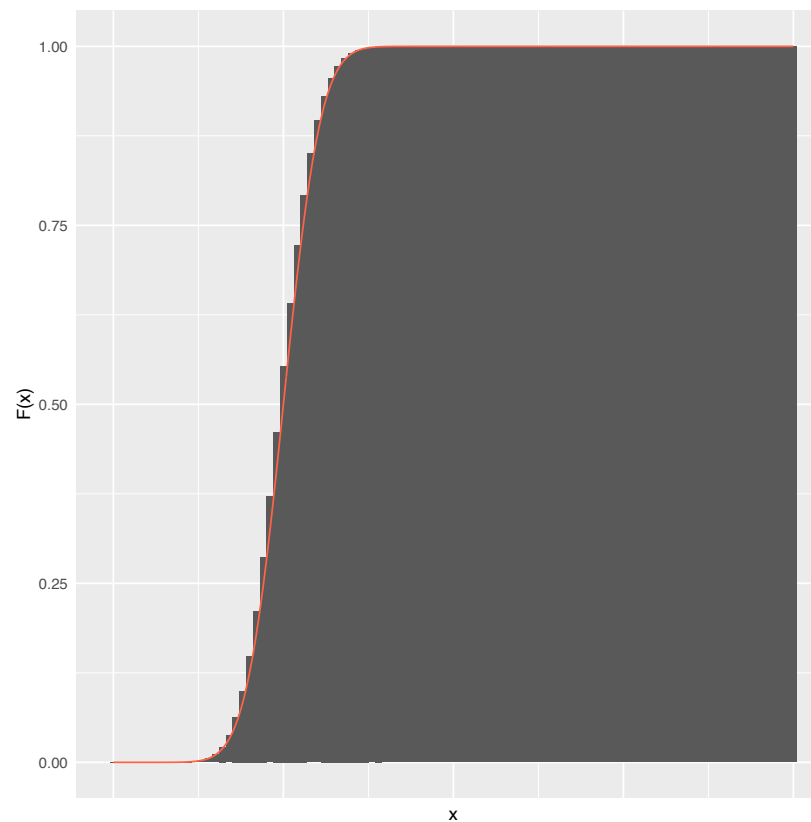
```
> pt(-0.5, 50)
[1] 0.3096343
```

←  $F(-0.5) = P(T \leq -0.5)$

- Ist also eine Zufallsvariable  $T$  t-verteilt mit hohem  $\nu$ , dann ist  $T \stackrel{a}{\sim} N(0, 1)$

## Beispiel II

- Die Normalverteilung mit Parametern  $\mu = n\pi$  und  $\sigma^2 = n\pi(1 - \pi)$  ist eine approximative Wahrscheinlichkeitsverteilung einer Zufallsvariable  $X$ , deren tatsächliche Wahrscheinlichkeitsverteilung eine Binomialverteilung mit hohem  $n$  ist:



rot:  $N(25, 18.75)$

schwarz:  $B(100, 0.25)$

- Vergleich einiger Werte der Verteilungsfunktionen:

- $P(X \leq 20)$ :

```
> pnorm(20, 25, sqrt(18.75))  
[1] 0.1241065 ←————  $F^*(20) \approx P(X \leq 20)$ 
```

```
> pbinom(20, 100, 0.25)  
[1] 0.148831 ←————  $F(20) = P(X \leq 20)$ 
```

- $P(X \leq 30)$ :

```
> pnorm(30, 25, sqrt(18.75))  
[1] 0.8758935 ←————  $F^*(30) \approx P(X \leq 30)$ 
```

```
> pbinom(30, 100, 0.25)  
[1] 0.8962128 ←————  $F(30) = P(X \leq 30)$ 
```

- Ist also eine Zufallsvariable  $X$  binomialverteilt mit hohem  $n$ , dann ist  $X \overset{a}{\sim} N(n\pi, n\pi(1 - \pi))$

# Konvergenz von Wahrscheinlichkeitsverteilungen

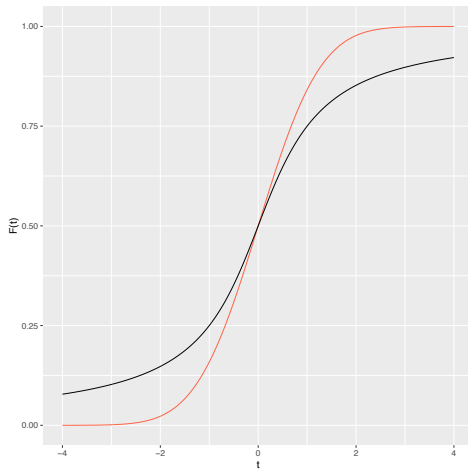
- In vielen Fällen wird die Approximation einer Wahrscheinlichkeitsverteilung  $P$  durch eine Wahrscheinlichkeitsverteilung  $P^*$  umso genauer, je größer der Wert eines bestimmten Parameters  $\theta$  von  $P$  wird.
- In diesen Fällen sagt man auch, dass die Wahrscheinlichkeitsverteilung  $P$  für  $\theta \rightarrow \infty$  gegen die Wahrscheinlichkeitsverteilung  $P^*$  **konvergiert**.
- Notation:

$$\lim_{\theta \rightarrow \infty} P = P^*$$

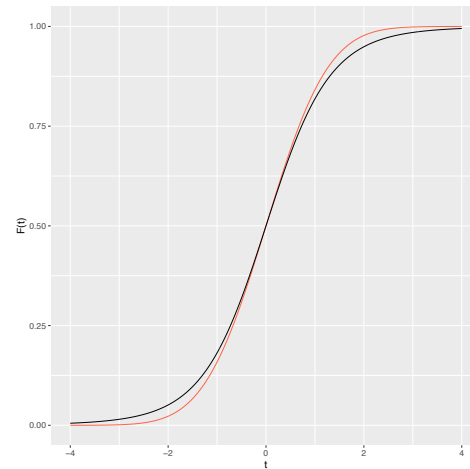
# Beispiel I

- t-Verteilung und Standardnormalverteilung:

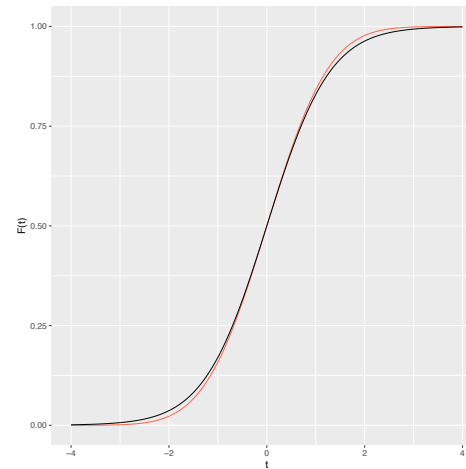
$$\lim_{\nu \rightarrow \infty} t(\nu) = N(0, 1)$$



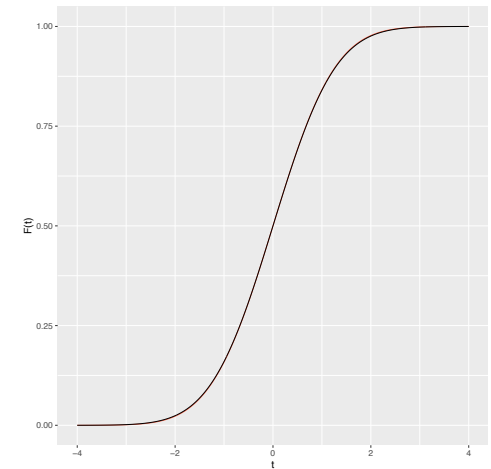
rot:  $N(0, 1)$   
schwarz:  $t(1)$



rot:  $N(0, 1)$   
schwarz:  $t(5)$



rot:  $N(0, 1)$   
schwarz:  $t(10)$

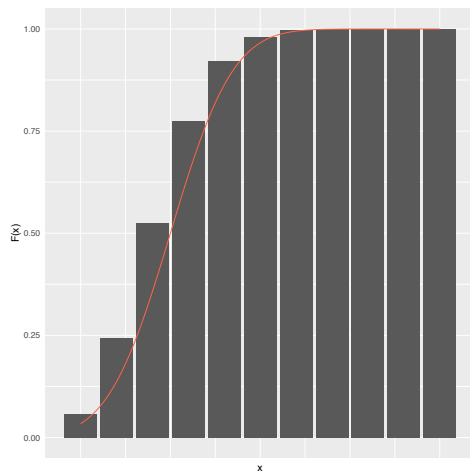


rot:  $N(0, 1)$   
schwarz:  $t(100)$

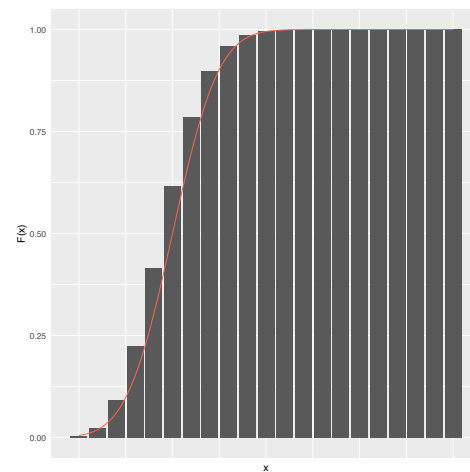
## Beispiel II

- Binomialverteilung und Normalverteilung:

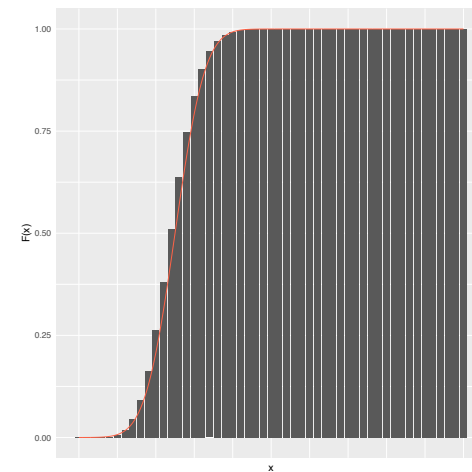
$$\lim_{n \rightarrow \infty} B(n, \pi) = N(n\pi, n\pi(1 - \pi))$$



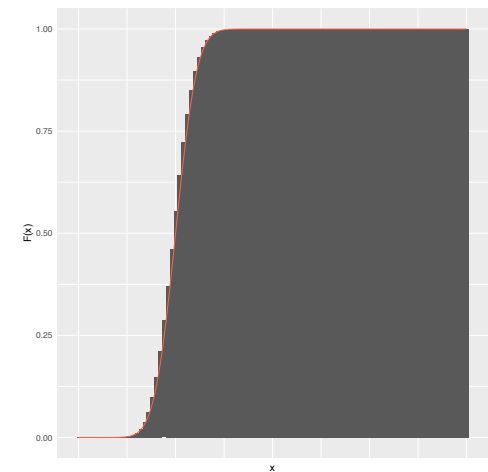
rot:  $N(2.5, 1.875)$   
schwarz:  $B(10, 0.25)$



rot:  $N(5, 3.75)$   
schwarz:  $B(20, 0.25)$



rot:  $N(12.5, 9.375)$   
schwarz:  $B(50, 0.25)$



rot:  $N(25, 18.75)$   
schwarz:  $B(100, 0.25)$

# Weiter mit der Konstruktion eines Konfidenzintervalls für $\pi$

- Problem: Die Wahrscheinlichkeitsverteilung von

$$Z^* = \frac{\bar{X} - \pi}{\sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}}$$

ist nicht bekannt.

- Lösung: Wir bestimmen die **approximative Wahrscheinlichkeitsverteilung** von  $Z^*$ .

## Approximative Verteilung von $Z^*$

- Es lässt sich zeigen, dass die Wahrscheinlichkeitsverteilung von

$$Z^* = \frac{\bar{X} - E(\bar{X})}{\widehat{SD}(\bar{X})} = \frac{\bar{X} - \pi}{\sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}}$$

für  $n \rightarrow \infty$  gegen die Standardnormalverteilung konvergiert:

$$\lim_{n \rightarrow \infty} P_{Z^*} = N(0, 1)$$

- Das heißt, für große Stichproben ist die Standardnormalverteilung eine approximative Wahrscheinlichkeitsverteilung von  $Z^*$ :

$$Z^* \stackrel{a}{\sim} N(0, 1)$$

- Je größer die Stichprobe ist (also je größer  $n$ ), desto besser ist diese Approximation.

## Quantile der Standardnormalverteilung I

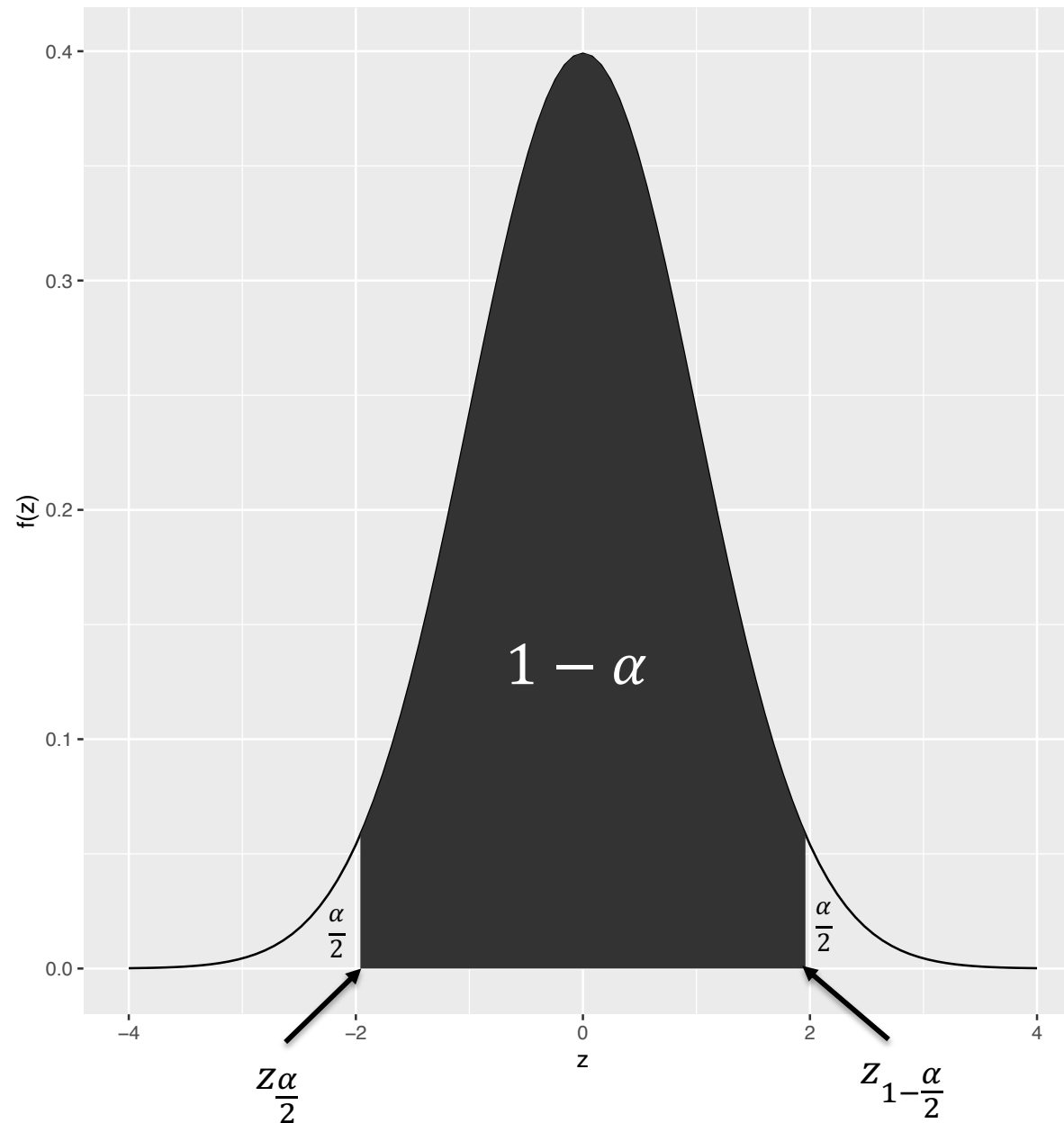
- Ab hier ist das Vorgehen wieder analog zur Konstruktion des KIs für  $\mu$ , nur eben mit der approximativen statt der genauen Wahrscheinlichkeitsverteilung von  $Z^*$  :
- Wir wissen, dass  $Z^*$  approximativ einer Standardnormalverteilung folgt und können mithilfe der Verteilungsfunktion  $F$  der Standardnormalverteilung diejenigen Werte  $z_{\frac{\alpha}{2}}$  und  $z_{1-\frac{\alpha}{2}}$  bestimmen, für die jeweils  $F\left(z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$  und  $F\left(z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$  gilt.

- Dann gilt:

$$P\left(z_{\frac{\alpha}{2}} \leq Z^* \leq z_{1-\frac{\alpha}{2}}\right) \approx F\left(z_{1-\frac{\alpha}{2}}\right) - F\left(z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

- Nochmal: Je größer der Stichprobenumfang  $n$ , desto besser ist diese Approximation.

## Quantile der Standardnormalverteilung II



- Zusammenfassung bis jetzt:
- Wir wissen, dass

$$Z^* = \frac{\bar{X} - \pi}{\sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}}$$

approximativ einer Standardnormalverteilung folgt.

- Wir können deshalb auf der Basis der Standardnormalverteilung die Quantile  $z_{\frac{\alpha}{2}}$  und  $z_{1-\frac{\alpha}{2}}$  bestimmen, so dass

$$P\left(z_{\frac{\alpha}{2}} \leq Z^* \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

gilt.

- Einsetzen von

$$Z^* = \frac{\bar{X} - \pi}{\sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}}$$

in  $P\left(z_{\frac{\alpha}{2}} \leq Z^* \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$  und Umstellen (genau wie beim KI für  $\mu$ ):

$$P\left(z_{\frac{\alpha}{2}} \leq Z^* \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

$$P\left(z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \pi}{\sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

$$P\left(z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1 - \bar{X})}{n}} \leq \bar{X} - \pi \leq z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}\right) \approx 1 - \alpha$$

## Konstruktion des Konfidenzintervalls III

$$P\left(-\bar{X} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq -\pi \leq -\bar{X} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right) \approx 1 - \alpha$$

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \geq \pi \geq \bar{X} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right) \approx 1 - \alpha$$

$$P\left(\bar{X} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \geq \pi \geq \bar{X} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right) \approx 1 - \alpha$$

$$P\left(\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq \pi \leq \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right) \approx 1 - \alpha$$

- Also ist

$$I(X_1, \dots, X_n) = [U, O] = \left[ \bar{X} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right]$$

ein Konfidenzintervall für  $\pi$  mit approximativem Konfidenzniveau  $1 - \alpha$ , da

$$P\left( \bar{X} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq \pi \leq \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right) \approx 1 - \alpha$$

- Anmerkung: Man kann leider nicht zeigen, dass dieses Konfidenzintervall eine minimale erwartete Länge hat. Aber die Länge ist auf jeden Fall kürzer als die von  $[-\infty, \infty]$ .

- Für dieses Konfidenzintervall können wir die Realisationen  $u$  und  $o$  und somit das konkrete Konfidenzintervall

$$I(x_1, \dots, x_n) = [u, o] = \left[ \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right]$$

aus unserer Stichprobe berechnen.

- Wir müssen hierfür lediglich den Mittelwert  $\bar{x}$  (= die relative Häufigkeit  $h(1)$  der mit 1 kodierten Messwertausprägung), den Stichprobenumfang  $n$  und das Quantil  $z_{1-\frac{\alpha}{2}}$  bestimmen.

- Da wir die Wahrscheinlichkeitsverteilung von  $Z^*$  nur approximiert haben, ist das tatsächliche Konfidenzniveau unseres Konfidenzintervalls nicht exakt gleich  $1 - \alpha$ .
- Für kleine Stichproben ( $n < 100$ ) kann das tatsächliche Konfidenzniveau unter Umständen deutlich niedriger als  $1 - \alpha$  sein.
- Aber: Je größer der Stichprobenumfang  $n$ , desto näher liegt die tatsächliche Wahrscheinlichkeit

$$P\left(\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq \pi \leq \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right)$$

an dem von uns vorgegebenen Wert  $1 - \alpha$ , und desto näher liegt somit das tatsächliche Konfidenzniveau unseres Konfidenzintervalls an dem von uns vorgegebenen Wert  $1 - \alpha$ .

- Und: Wir benötigen sowieso große Stichproben. Warum?

- Zur Erinnerung:
  - Je länger das konkrete Konfidenzintervall, desto ungenauer unsere Schätzung.
  - Je kürzer das konkrete Konfidenzintervall, desto genauer unsere Schätzung.
- Drei Faktoren beeinflussen die Länge des konkreten Konfidenzintervalls:
  1. Das Konfidenzniveau  $1 - \alpha$  (indirekt über  $z_{1-\frac{\alpha}{2}}$ )
    - Je höher das Konfidenzniveau, desto länger das Intervall.
  2. Der Stichprobenumfang  $n$ 
    - je größer der Stichprobenumfang, desto kürzer das Intervall.
  3. Der Schätzwert  $\bar{x}$  für  $\pi$ 
    - Je größer  $\bar{x}(1 - \bar{x})$ , desto länger das Intervall.

### Sehr wichtig:

- Da das Konfidenzniveau  $1 - \alpha$  auf jeden Fall hoch ( $> 0.95$ ) gewählt werden sollte und wir keinen Einfluss auf den Schätzwert  $\bar{x}$  haben, bleibt als einzige Größe, die wir beeinflussen können, der Stichprobenumfang  $n$ .
- Um kleine Konfidenzintervalle und somit genaue Schätzungen zu erhalten, müssen wir also eine **große Stichprobe** erheben.

- Bemerkung I: Es gibt neben der hier vorgestellten Methode noch mehrere alternative Methoden, um Konfidenzintervalle für  $\pi$  zu konstruieren. Im Gegensatz zum Normalverteilungsfall gibt es hier nicht ein einziges Konfidenzintervall, dass alle Gütekriterien erfüllt. Die verschiedenen Konfidenzintervalle haben jeweils Vor- und Nachteile.
- Bemerkung II: Das hier besprochene approximative Konfidenzintervall wird auch **Wald-Konfidenzintervall** genannt.
- Bemerkung III: Es ist auch möglich, auf Basis der Binomialverteilung ein Konfidenzintervall mit exaktem (d.h. nicht nur approximativem) Konfidenzniveau für  $\pi$  zu konstruieren. Dieses hat aber schlechtere Eigenschaften als das approximative Intervall (u.a. größere erwartete Länge und Probleme, die sich aus der Diskretheit der Binomialverteilung ergeben).

## Beispiel Depression I

- Wir interessieren uns für die relative Häufigkeit  $h_D$  der Depression in einer Population.
- Wir ziehen eine einfache Zufallsstichprobe mit  $n = 100$  Personen aus dieser Population.
- $X_1, X_2, \dots, X_i, \dots, X_{100}$  sind Bernoulli-Variablen, die jeweils den Wert 1 annehmen, falls die  $i$ -te Person in der Stichprobe depressiv ist.
- $X_i \stackrel{\text{iid}}{\sim} Be(\pi)$
- $\pi = h_D$
- Wir wollen nun ein 0.95-Konfidenzintervall für  $\pi$  berechnen.

## Beispiel Depression II

- Das konkrete Konfidenzintervall hat die Form

$$I(x_1, \dots, x_n) = \left[ \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right]$$

- Um dieses berechnen zu können, benötigen wir  $\bar{x}$ ,  $n$  und  $z_{1-\frac{\alpha}{2}}$ .
- Zum Beispiel könnten 20% der Personen in der von uns gezogenen Stichprobe depressiv sein, d.h.:

$$\bar{x} = h(1) = 0.2$$

- Wir wissen zudem, dass  $n = 100$  ist.

## Beispiel Depression III

- Wir müssen also nur noch  $z_{1-\frac{\alpha}{2}}$  berechnen:
- Das von uns gewünschte Konfidenzniveau ist  $1 - \alpha = 0.95$ .
- Damit ist  $\alpha = 1 - 0.95 = 0.05$ ,  $\frac{\alpha}{2} = 0.025$  und somit  $z_{1-\frac{\alpha}{2}} = z_{1-0.025} = z_{0.975}$ .
- Damit können wir  $z_{0.975}$  in R berechnen:

```
> qnorm(0.975, 0, 1)  
[1] 1.959964
```

- Also ist  $z_{0.975} \approx 1.96$

## Beispiel Depression IV

- Einsetzen von  $\bar{x} = 0.2$ ,  $n = 100$  und  $z_{0.975} \approx 1.96$  in das konkrete Konfidenzintervall:

$$I(x_1, \dots, x_n) = \left[ \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right]$$

$$= \left[ 0.2 - 1.96 \cdot \sqrt{\frac{0.2 \cdot 0.8}{100}}, 0.2 + 1.96 \cdot \sqrt{\frac{0.2 \cdot 0.8}{100}} \right]$$

$$\approx [0.12, 0.28]$$

## Beispiel Depression V

- Ergebnis der Berechnung des konkreten Konfidenzintervalls:

$$I(x_1, \dots, x_n) = [0.12, 0.28]$$

- Interpretation:
  - Auf Basis unserer Stichprobe sind die Werte in dem Intervall  $[0.12, 0.28]$  die plausiblen Werte für  $\pi$  und somit für die relative Häufigkeit  $h_D$  der Depression in der Population.

- Bislang:
  - Motivation und Grundlagen der Intervallschätzung
  - Konfidenzintervalle und Gütekriterien
  - Konfidenzintervalle für den Parameter  $\mu$  einer Normalverteilung
  - Konfidenzintervalle für den Parameter  $\pi$  einer Bernoulli-Verteilung
- Jetzt:
  - Zusammenfassung
  - Übersicht

- Mit Konfidenzintervallen versuchen wir, die Genauigkeit einer Schätzung des Parameters  $\pi$  interpretierbarer zu machen.
- Zur Konstruktion eines Konfidenzintervalls für  $\pi$  wenden wir die gleiche Logik wie beim KI für  $\mu$  an.
- Da die Schätzfunktion  $\bar{X}$  hier allerdings **nicht normalverteilt** ist, ist die wahre Verteilung der nach Logik der z-Standardisierung umgeformten Zufallsvariable unbekannt.
- Glücklicherweise können wir jedoch davon ausgehen, dass mit größeren Stichproben diese unbekannte Verteilung **durch die Standardnormalverteilung approximiert** werden kann.
- Durch diese Approximation gelten unsere Anforderungen an das KI wie das Konfidenzniveau zwar nicht exakt, wir können uns jedoch darauf verlassen, dass je größer die Stichprobe ist, das Problem immer kleiner wird.

# Übersicht

# Punktschätzung

| Interessierende<br>Größe in der<br>Population                                                                                                                                           | Wahrscheinlich-<br>keitsverteilung der<br>Zufallsvariablen $X_1, X_2, \dots, X_n$ in der<br>einfachen Zufalls-<br>stichprobe | Parameter, der<br>der interessier-<br>enden Größe in<br>der Population<br>entspricht | Erwartungs-<br>treue,<br>effiziente und<br>konsistente<br>Schätz-<br>funktion | Standard-<br>fehler dieser<br>Schätz-<br>funktion | Schätzwert       |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|---------------------------------------------------|------------------|
| $h_{Pop}$<br><br>(relative Häufigkeit einer<br>Messwertausprägung einer<br>diskreten Variable in der<br>Population)                                                                     | $X_i \stackrel{iid}{\sim} Be(\pi)$                                                                                           | $\pi$                                                                                | $\bar{X}$                                                                     | $\sqrt{\frac{\pi(1-\pi)}{n}}$                     | $\bar{x} = h(1)$ |
| $\bar{x}_{Pop}$<br><br>(Mittelwert einer stetigen<br>Variable in der Population,<br>deren Histogramm durch die<br>Dichte einer Normalverteilung<br>approximiert werden kann)            | $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$                                                                                  | $\mu$                                                                                | $\bar{X}$                                                                     | $\sqrt{\frac{\sigma^2}{n}}$                       | $\bar{x}$        |
| $s_{emp Pop}^2$<br><br>(Empirische Varianz einer<br>stetigen Variable in der<br>Population, deren<br>Histogramm durch die Dichte<br>einer Normalverteilung<br>approximiert werden kann) | $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$                                                                                  | $\sigma^2$                                                                           | $S^2$                                                                         | $\sqrt{\frac{2\sigma^4}{n-1}}$                    | $s^2$            |

# Intervallschätzung

| Interessierende<br>Größe in der<br>Population                                                                                                                                               | Wahrscheinlich-<br>keitsverteilung<br>der Zufalls-<br>variablen $X_1, X_2, \dots, X_n$ in der<br>einfachen<br>Zufalls-<br>stichprobe | Parameter,<br>der der interes-<br>sierenden<br>Größe in<br>der<br>Population<br>entspricht | Zufälliges Konfidenzintervall                                                                                                                                           | Konkretes Konfidenzintervall                                                                                                                                            |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $h_{Pop}$<br><br>(relative<br>Häufigkeit einer<br>Messwert-<br>ausprägung einer<br>diskreten Variable<br>in der Population)                                                                 | $X_i \stackrel{\text{iid}}{\sim} Be(\pi)$                                                                                            | $\pi$                                                                                      | $\left[ \bar{X} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right]$ | $\left[ \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right]$ |
| $\bar{x}_{Pop}$<br><br>(Mittelwert einer<br>stetigen Variable<br>in der Population,<br>deren<br>Histogramm<br>durch die Dichte<br>einer<br>Normalverteilung<br>approximiert<br>werden kann) | $X_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$                                                                                   | $\mu$                                                                                      | $\left[ \bar{X} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}} \right]$                               | $\left[ \bar{x} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}, \bar{x} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}} \right]$                               |