Datenanalyse – Sitzung 4

Konfidenzintervall und Normalverteilung

Institut für Kommunikationswissenschaft und Medienforschung
Ludwig-Maximilians-Universität München

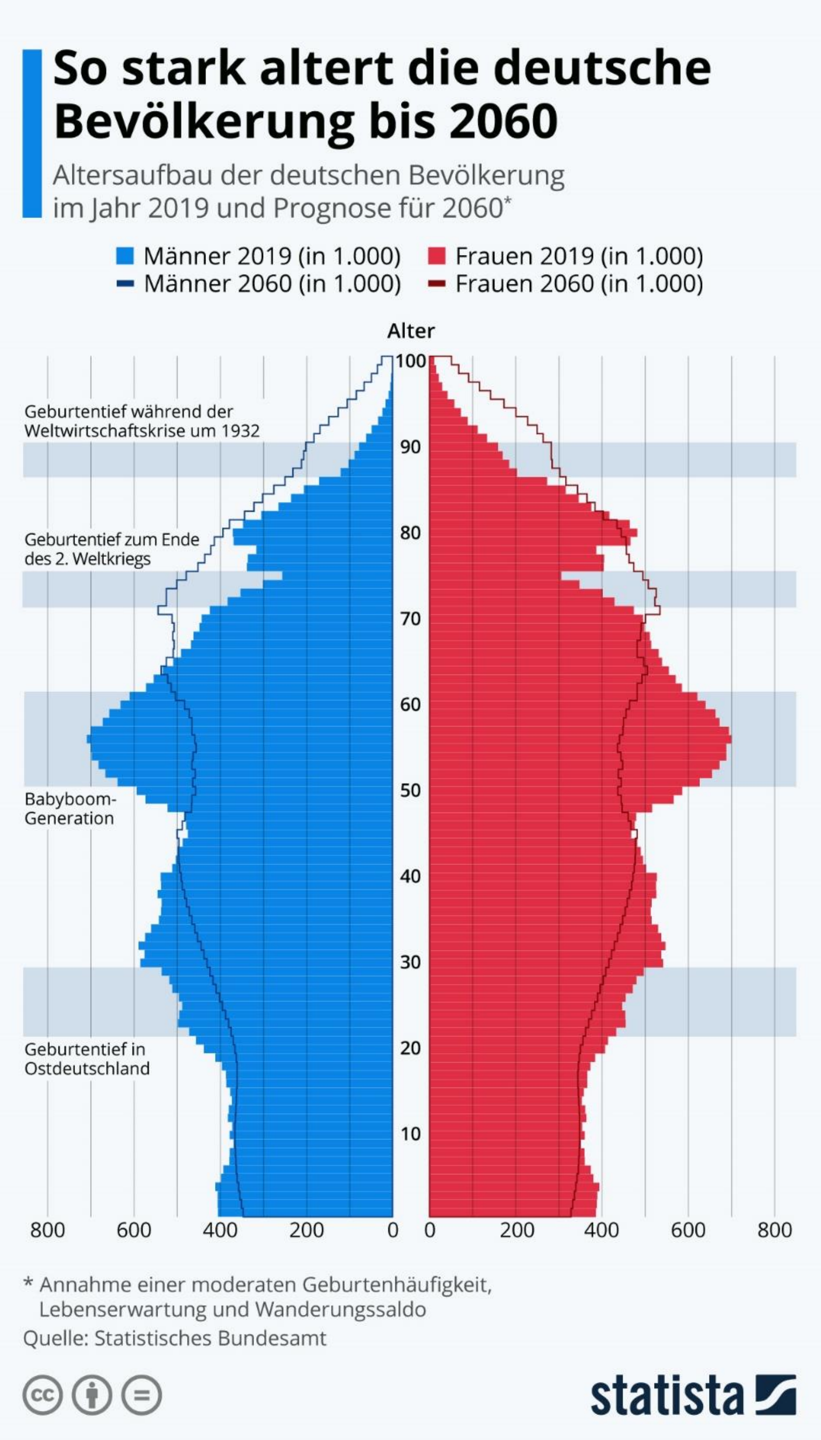
Ablauf der Sitzung

1. Datenverteilung und Histogramme
2. Konfidenzintervalle – von der anderen Seite betrachtet
3. Normalverteilung – warum und wie schief

VERTEILUNG UND DARSTELLUNG IM HISTOGRAMM

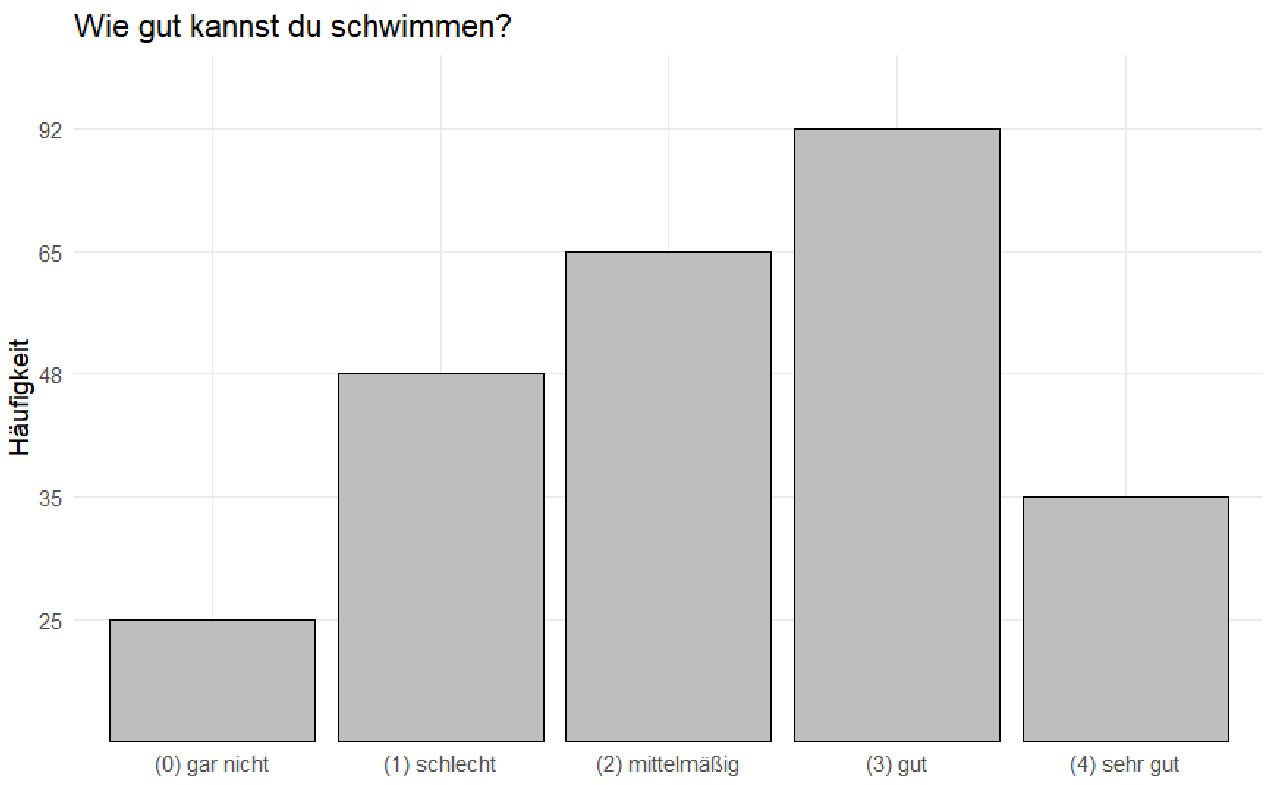
Verteilung

- Die Verteilung eines Merkmals …
  - ist mehr als Mittelwert und Standardverteilung
- Die Verteilung ist relevant …
  - für die Interpretation von Ergebnissen
  - für die Prüfung der Voraussetzung „Normalverteilung“



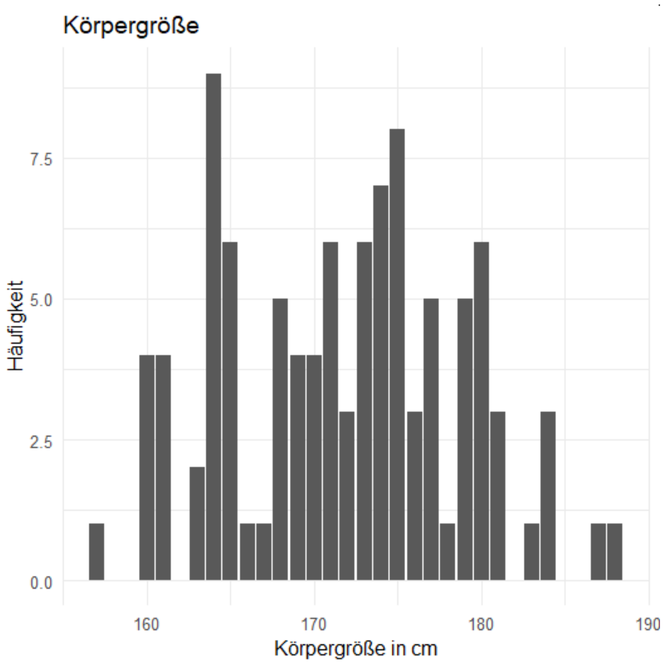
Wiederholung: Säulendiagramm

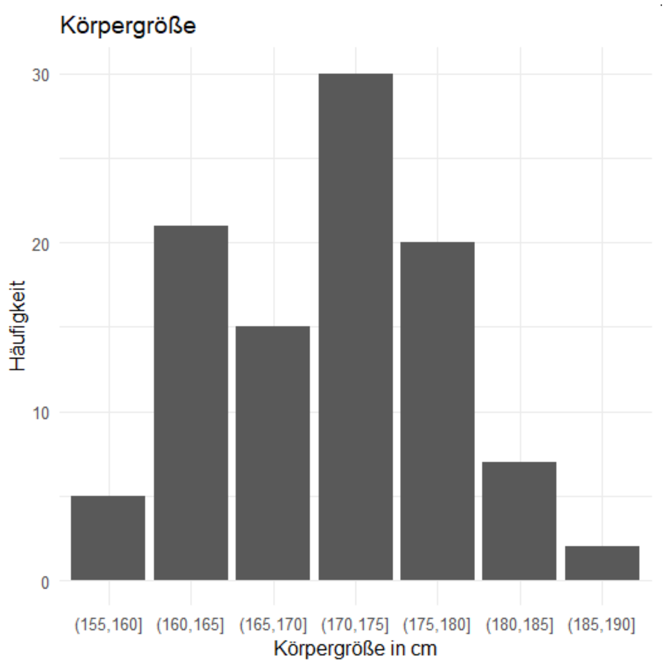
- Das Säulendiagramm visualisiert die Häufigkeit eines diskreten (z.B. **nominalen**) Merkmals
- Nach rechts ist die Ausprägung des Merkmals angetragen
- Nach oben wird angetragen, wie viele Personen das Merkmal in der jeweiligen Ausprägung tragen

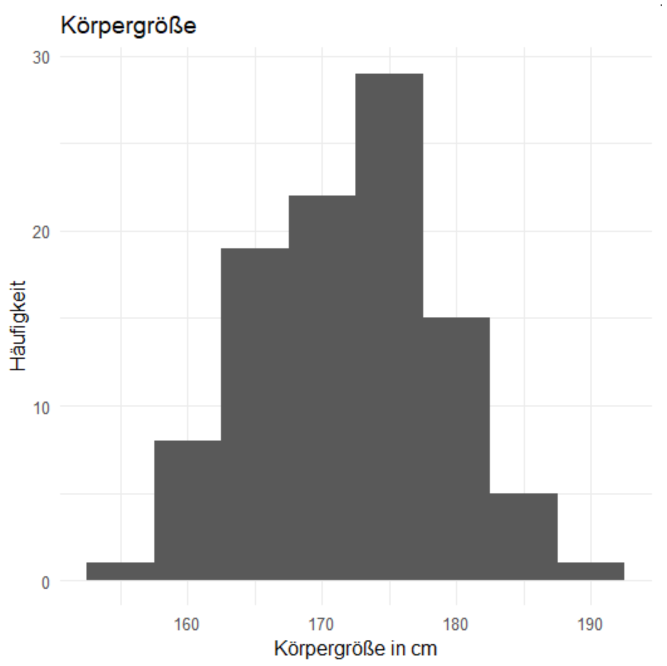


Wiederholung: Histogramm

- Das Histogramm visualisiert die Verteilung eines nicht-diskreten Merkmals
  - viele (Alter in Jahren) oder
    unendlich viele (Körpergröße, ganz genau gemessen) Ausprägungen
- Bereiche (z.B. 0-10, >10-20, …) werden als Kategorien zusammengefasst





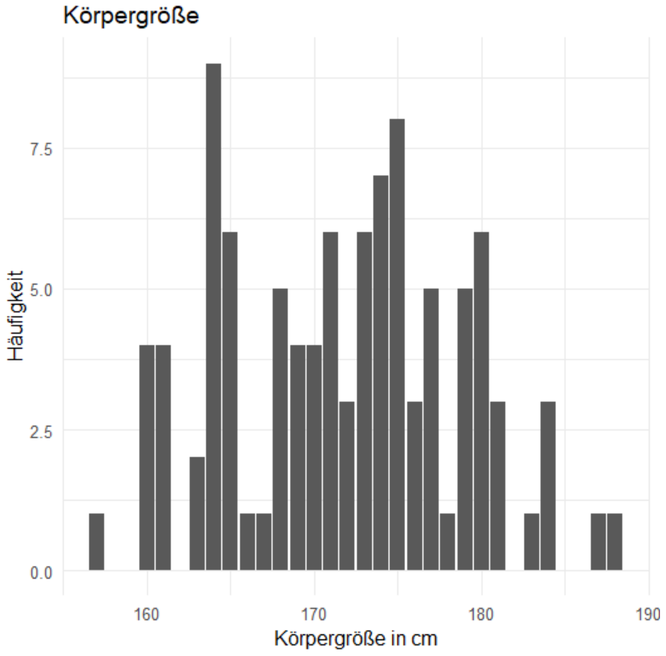


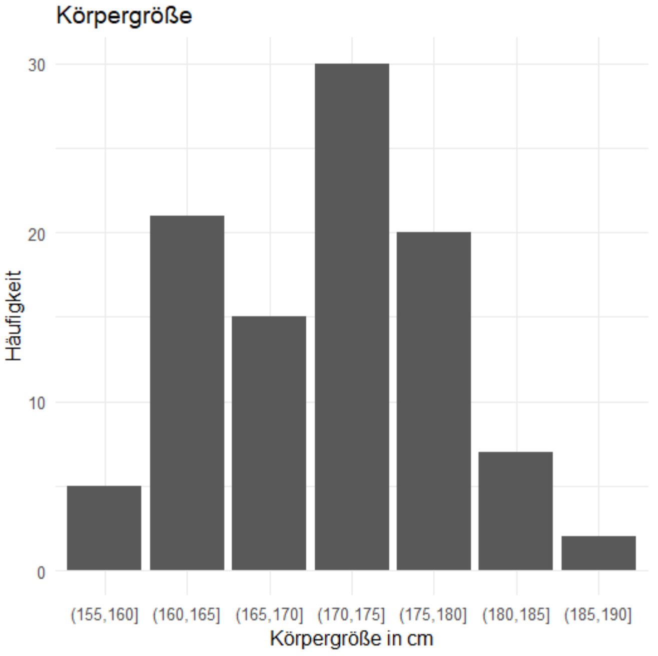
Wiederholung: Histogramm

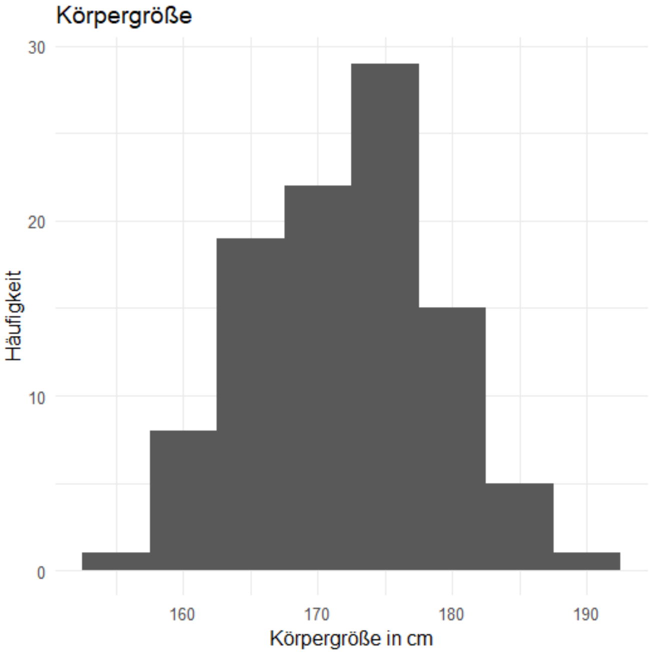
```
groesse %>%
    as.data.frame() %>%
    ggplot(aes(x = .)) +
    theme_minimal() +
    geom_bar() +
    labs(title = "Körpergröße",
       x = "Körpergröße in cm",
       y = "Häufigkeit")
```

```
groesse %>%
    as.data.frame() %>%
    mutate(categories = cut(.,
        breaks = c(-Inf, 150, 155, 160,
                   165, 170, 175, 180,
                   185, 190, +Inf))) %>%
    ggplot(aes(x = categories)) +
    theme_minimal() +
    geom_bar() +
    labs(title = "Körpergröße",
       x = "Körpergröße in cm",
       y = "Häufigkeit")
```

```
groesse %>%
    as.data.frame() %>%
    ggplot(aes(x = .)) +
    geom_histogram(binwidth=5) +
    theme_minimal() +
    labs(title = "Körpergröße",
       x = "Körpergröße in cm",
       y = "Häufigkeit")
```

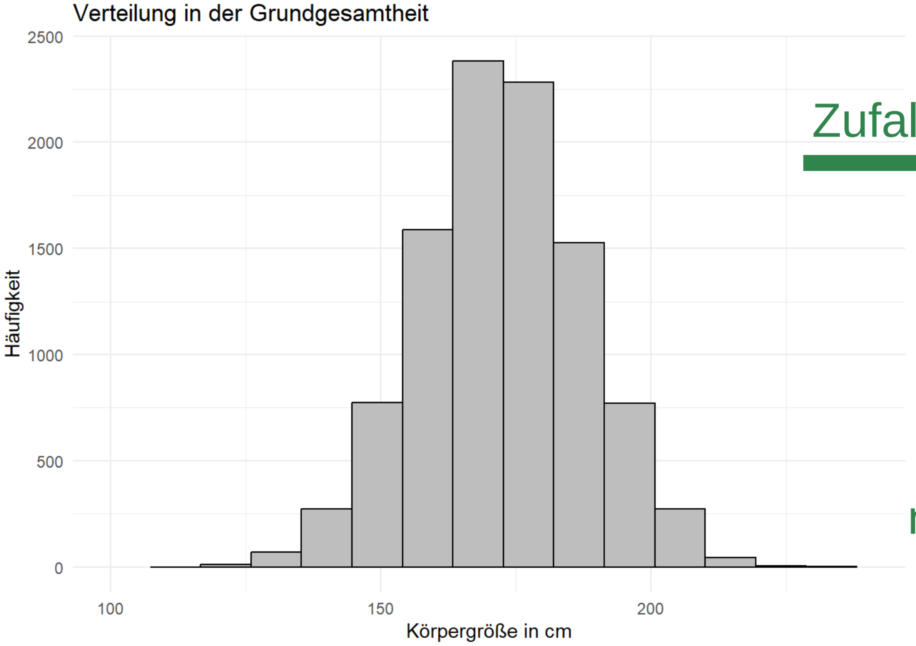






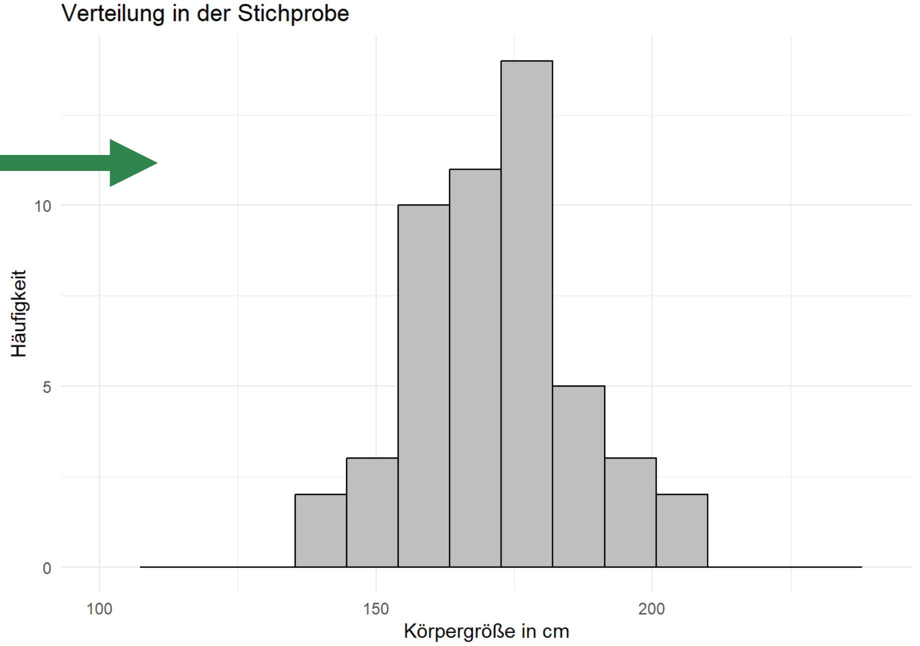
KONFIDENZINTERVALL VON HINTEN BETRACHTET

Verteilung und Stichprobe



Zufallsstichprobe

(N=50)



die wahre Verteilung kennen wir normalerweise nicht

```
## Erzeugen der Grundgesamtheit
grundgesamtheit <- (rnorm(10000) * 15 + 172.5) %>% round(digits = 1) %>% as_tibble()

## Erstellen des Histogramms für die Verteilung der Grundgesamtheit
grundgesamtheit %>%
  ggplot(aes(x = value)) + theme_minimal() +
  geom_histogram(bins = 16, fill = "grey", color = "black") + xlim(100, 240) +
  labs(title = "Verteilung in der Grundgesamtheit", x = "Körpergröße in cm", y = "Häufigkeit")
```

```
## Ziehen einer einzigen Stichprobe
stichprobe <- grundgesamtheit %>%
  slice_sample(n = 50)

## Erstellen des Histogramms für die Verteilung der gezogenen Stichprobe
stichprobe %>%
  ggplot(aes(x = value)) + theme_minimal() +
  geom_histogram(bins = 16, fill = "grey", color = "black") + xlim(100, 240) +
  labs(title = "Verteilung in der Stichprobe", x = "Körpergröße in cm", y = "Häufigkeit")
```

```
grundgesamtheit %>%
  summarize(M = mean(value), SD = sd(value)) %>%
  # sprintf() sorgt dafür, dass genau eine Nachkommastelle
  mutate(M = sprintf("%.1f", M), SD = sprintf("%.1f", SD))
```

$\bar{x} = \mu = 172{,}4\ cm$

$s = 15{,}2\ cm$

$\bar{x} = 171{,}6\ cm$

$s = 15{,}0\ cm$

Standardabweichung und Standardfehler

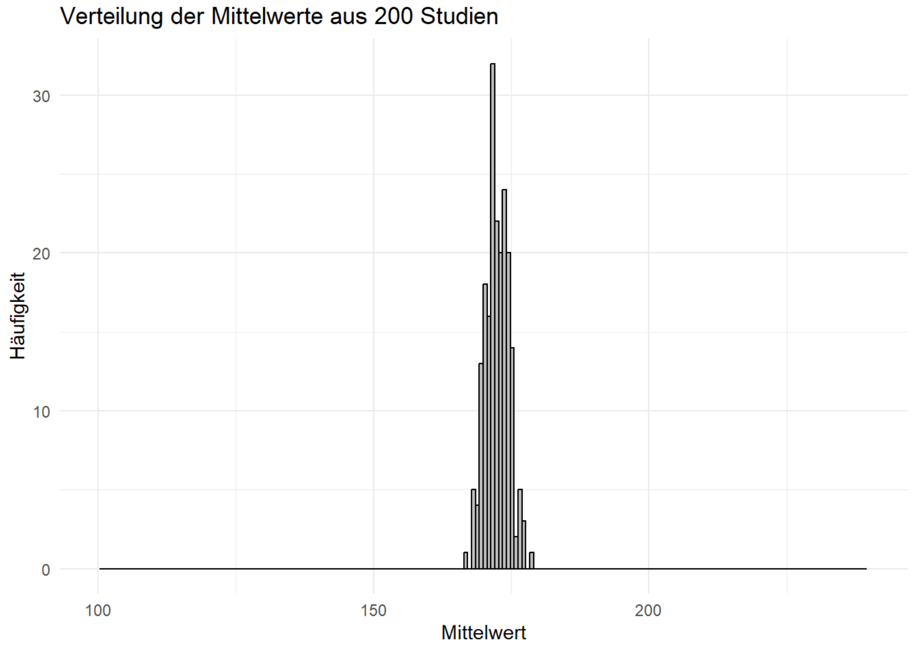
- Wie gut ist eine Schätzung auf Basis N=50 ?
- Der Standardfehler beschreibt die (Un-)Genauigkeit

```r
## Erstellen des Histogramms für alle berechneten Mittelwerte
mittelwerte <- as_tibble(list(value = mittelwerte_von_200studien))

mittelwerte %>%
  ggplot(aes(x = value)) + theme_minimal() +
  geom_histogram(bins = 16, fill = "grey", color = "black") + xlim(100, 240) +
  labs(title = "Verteilung der Mittelwerte aus 200 Studien", x = "Mittelwert", y = "Häufigkeit")
```

```r
## Schreiben einer Funktion, die wiederholt Stichproben zieht und Mittelwerte berechnet
studie_durchfuehren <- function(grundgesamtheit) {
  grundgesamtheit %>%
    slice_sample(n = 50) %>%
    summarize(mean = mean(value)) %>%
    pull(mean) %>%
    return()
}

## 200 Studien durchführen, also 200 Stichproben ziehen und deren Mittelwerte berechnen
mittelwerte_von_200studien <- map_dbl(1:200, ~studie_durchfuehren(grundgesamtheit))
```

```
173.2 169.4 174.1 170.1 173.0 172.0 175.3 172.4 171.9 171.7 171.1 173.5
171.5 177.1 171.5 174.2 171.6 169.8 171.8 171.9 173.2 177.4 172.5 173.6
174.4 171.7 172.7 172.8 174.8 171.3 172.8 171.9 176.4 172.1 175.0 172.6
171.7 172.3 174.6 173.1 171.7 175.6 175.2 172.0 172.8 173.1 173.4 170.0
172.7 170.7 174.3 175.2 171.9 174.4 171.3 169.3 174.0 174.5 174.2 175.4
175.1 169.5 175.7 172.0 173.5 172.9 170.5 173.6 174.1 170.5 173.8 171.5
175.2 169.7 170.8 172.0 171.8 173.8 170.6 174.4 169.3 172.1 176.9 173.9
```

```r
mittelwerte_von_200studien %>% print(digits = 4, nsmall=1)
```

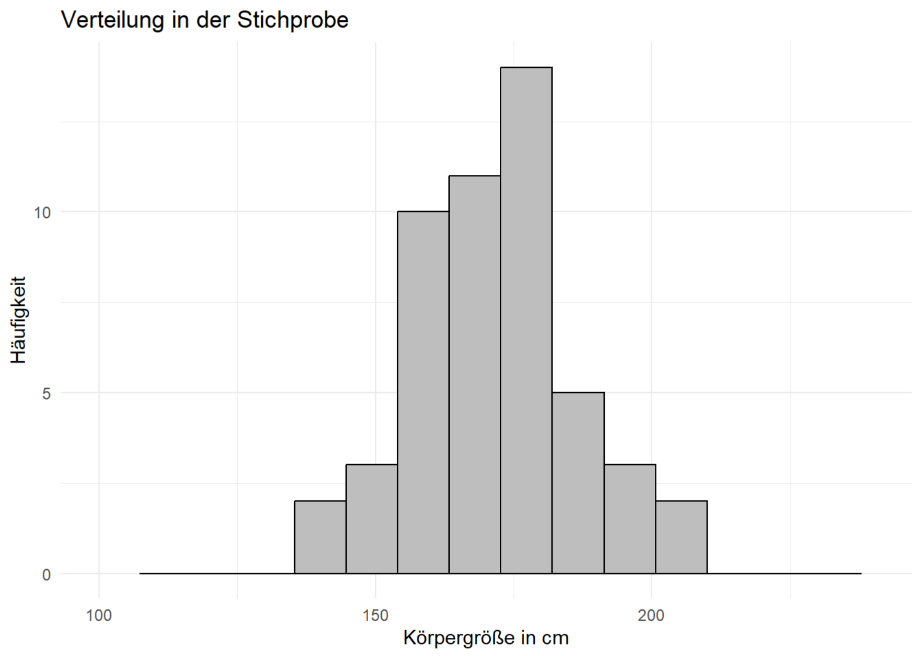


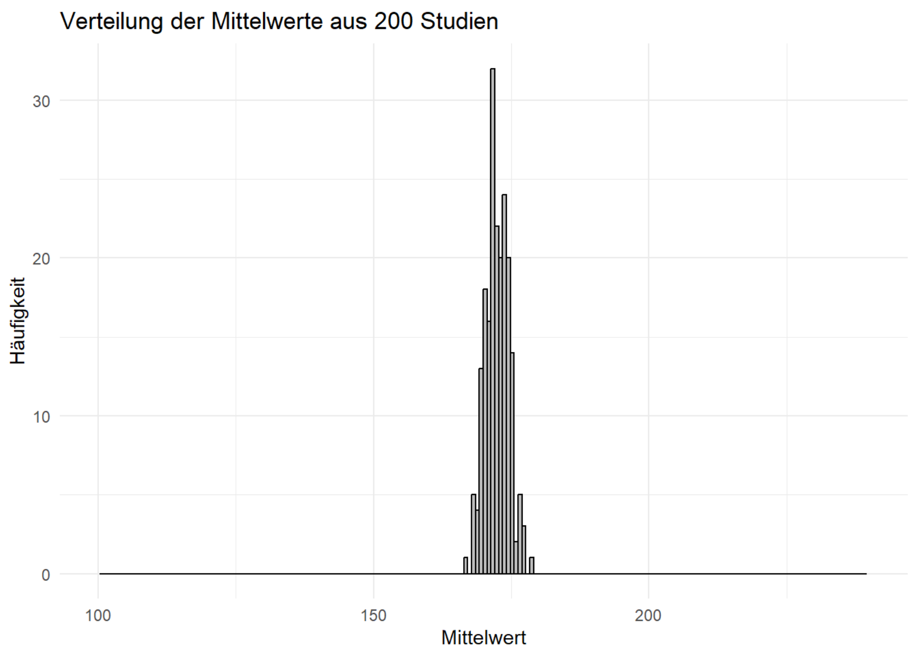
s = 2,08 cm = SE

Standardabweichung der Mittelwerte = Standardfehler

Standardabweichung und Standardfehler

Wir können auf Basis der Standardabweichung einer Stichprobe übrigens den Standardfehler der Mittelwert-Schätzungen schätzen





s = 15,0 cm
(aus einer Stichprobe)

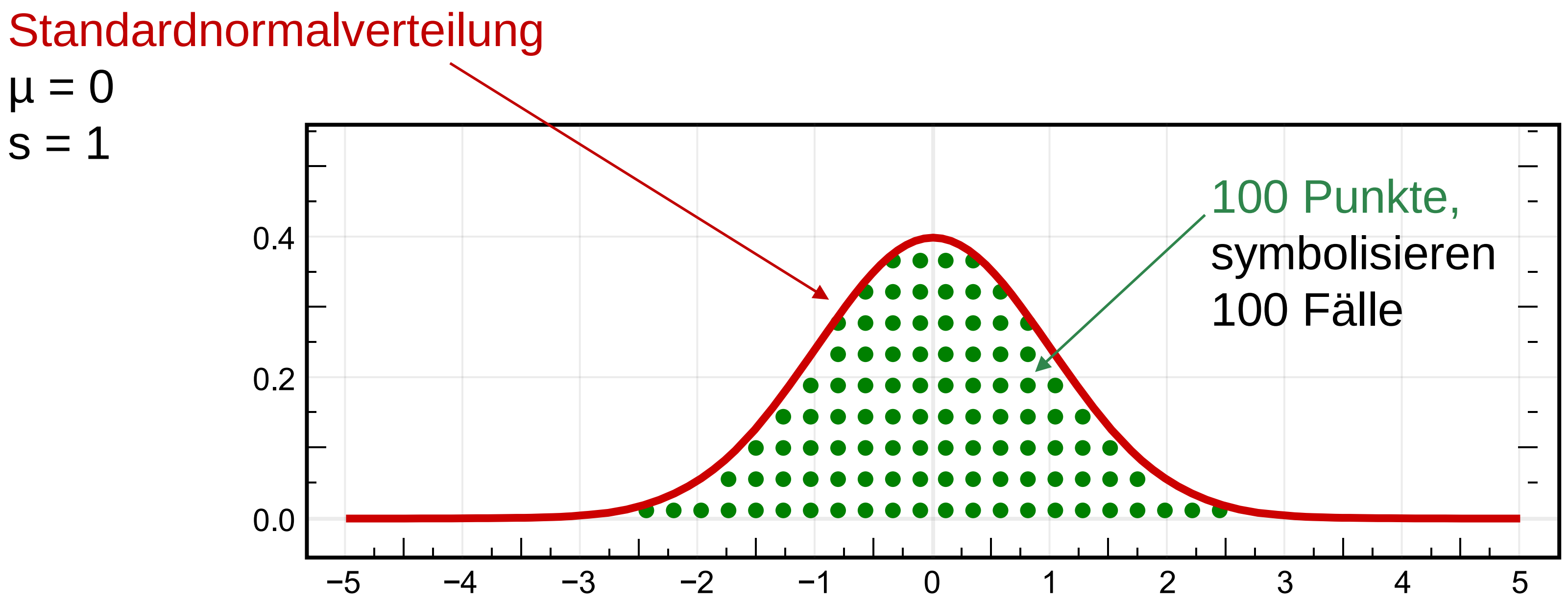
$$\hat{s}_{\bar{x}} = \frac{s}{\sqrt{N}} = \frac{15\ cm}{\sqrt{50}} = 2{,}12\ cm$$

$$SE = s_{\bar{x}} = 2{,}08\ cm$$

Für N=50 gar nicht so schlecht …

z-Werte als Hilfsmittel

- Wenn ich eine Normalverteilung habe (✓), dann liegen ca. 95% der Werte im Bereich von zwei Standardabweichungen um den Mittelwert.
- Genauer bekommen wir es mit z-Werten.

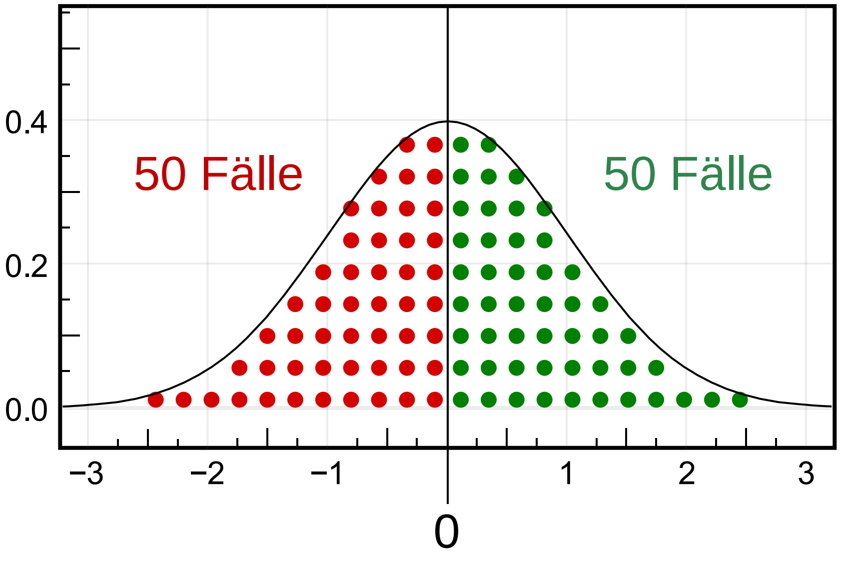


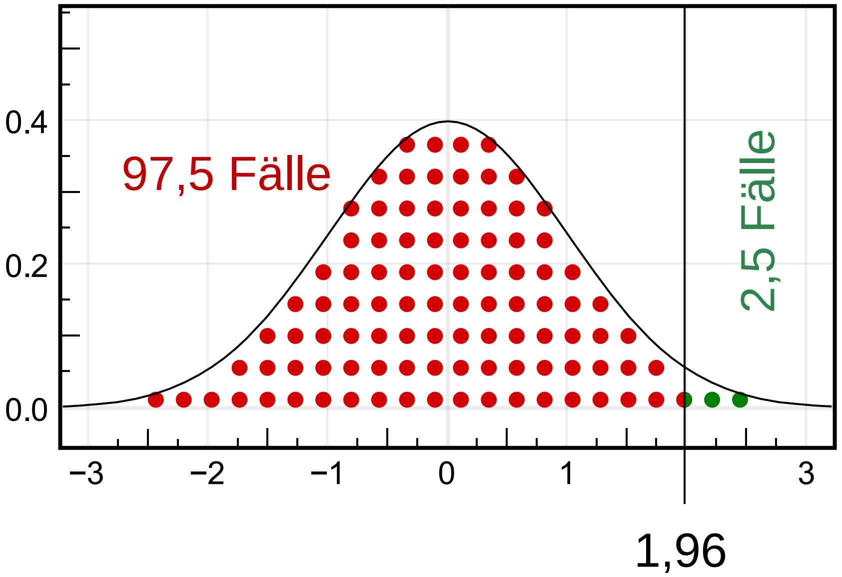
z-Werte als Hilfsmittel

- Die z-Wert-Tabelle verrät uns, wo wir auf der X-Achse sind, wenn wir einen bestimmten Anteil in der Normalverteilung haben.
- 50% der Fälle haben wir z.B. bei x=0 (trivial)
- 97,5% der Fälle haben wir bei x=1,96 (Standardabweichungen)

Tabellenset
(Download)

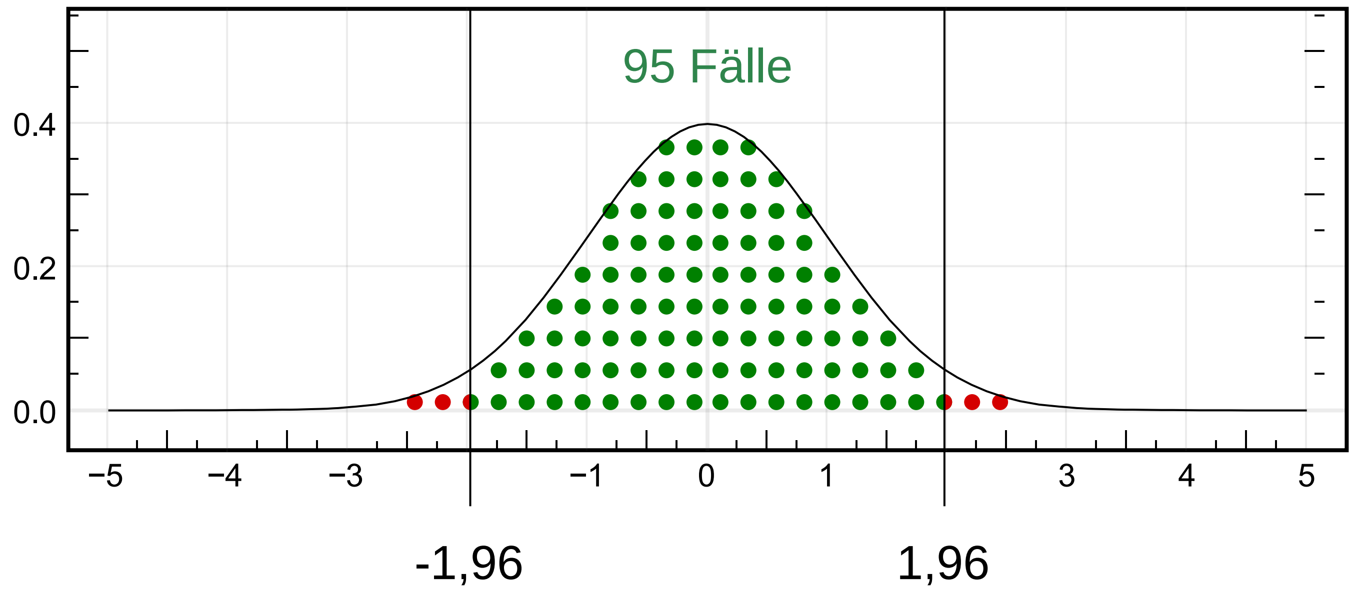
$P(x \leq z)$	z
0,5%	−2,58
1,0%	−2,33
2,5%	−1,96
5,0%	−1,64
95,0%	1,64
97,5%	**1,96**
99,0%	2,33
99,5%	2,58





z-Werte als Hilfsmittel

- Wenn wir wissen wollen, in welchem Bereich 95% der Fälle liegen, brauchen wir die z-Werte für 2,5% (links) und 97,5% (rechts)



$P(x \leq z)$	z
0,5%	−2,58
1,0%	−2,33
2,5%	−1,96
5,0%	−1,64
95,0%	1,64
97,5%	1,96
99,0%	2,33
99,5%	2,58

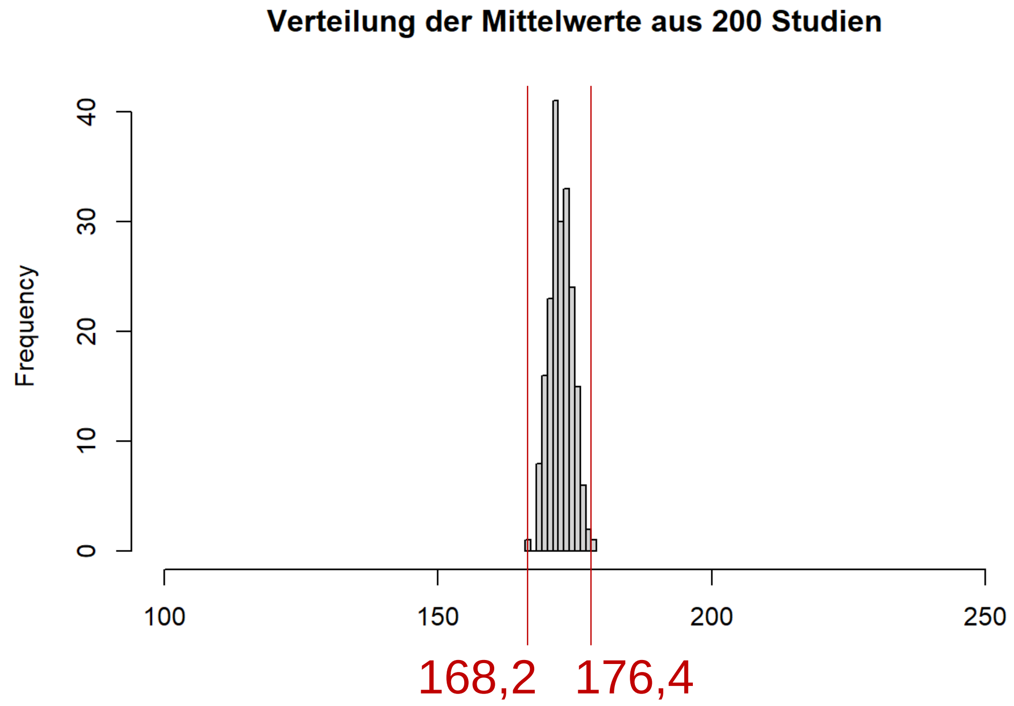
Standardfehler mal z-Wert

- Als Standardfehler hatten wir bei 200 Ziehungen à N=50 SE=2,08 beobachtet
- Und wir wissen, dass 95% der Mittelwerte zw. -1,96 und +1,96 Standardabweichungen um den Mittelwert herum liegen, also zw.

$$M - 1{,}96\ SE = $$
$$172{,}3\ cm - 1{,}96 \times 2{,}08\ cm = 168{,}2\ cm$$

und

$$M + 1{,}96\ SE = $$
$$172{,}3\ cm + 1{,}96 \times 2{,}08\ cm = 176{,}4\ cm$$

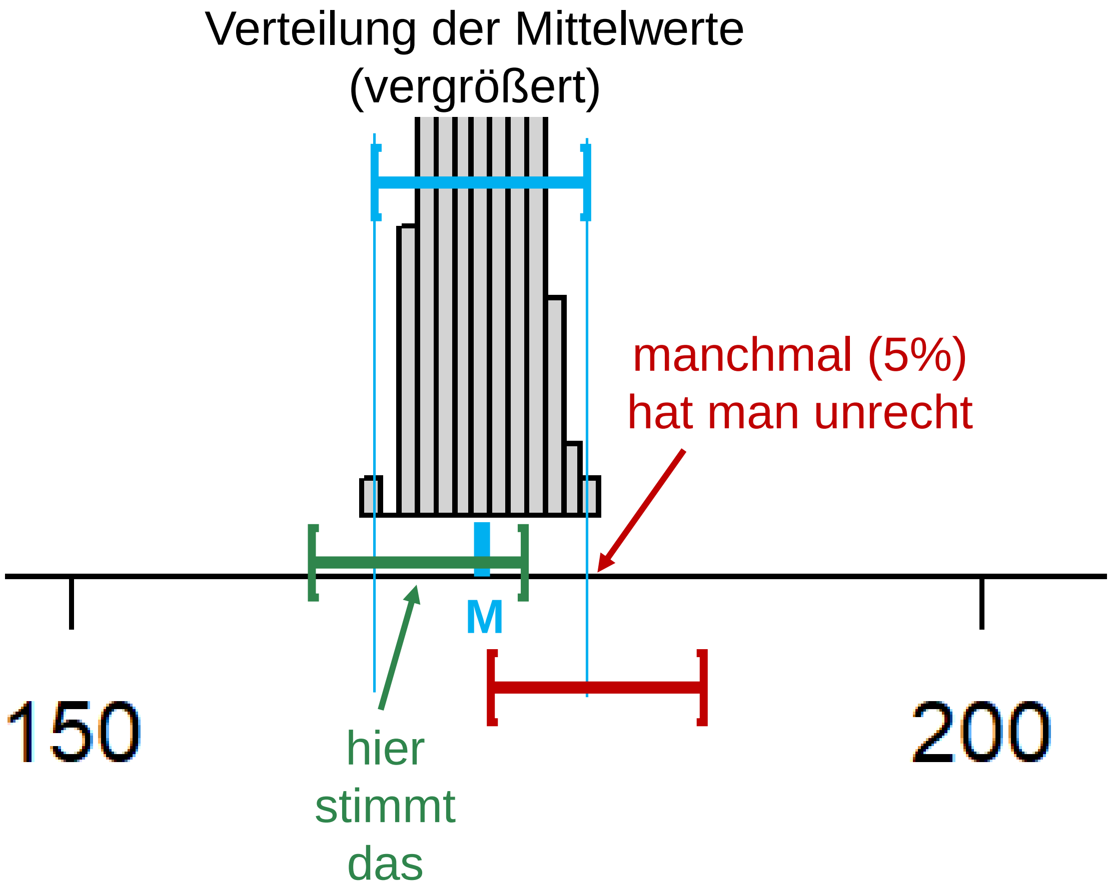


$$\bar{\bar{x}} = \bar{M} = 172{,}3\ cm$$
$$s_{\bar{x}} = s_M = 2{,}08\ cm$$

Standardfehler mal z-Wert

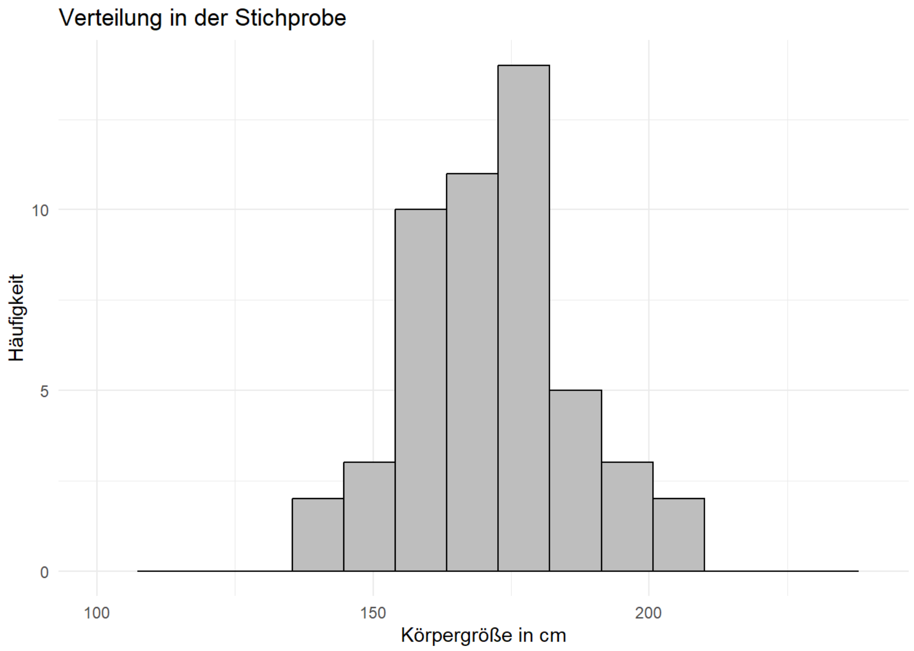
- Wenn wir wissen, dass die beobachteten Mittelwerte um $\pm 1{,}96 \times 2{,}08\ cm$ schwanken,

- Dann können wir davon ausgehen, dass der wahre Mittelwert in eben diesem Bereich um einen beobachteten Mittelwert liegt.



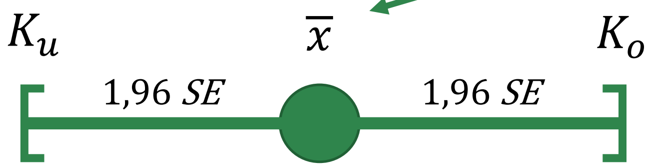
Konfidenzintervall

- Das Konfidenzintervall für den Mittelwert (Punktschätzung) ist also $\overline{x} \pm 1{,}96\ SE$ (für 95% Konfidenz)
- Das ist der Bereich, in dem wahrscheinlich (z.B. mit 95%) der wahre Mittelwert der Grundgesamtheit liegt



In der Stichprobe beobachtet

$\bar{x} = 171{,}6\ cm$
$s = 15{,}0\ cm$



Irgendwo hier liegt wahrscheinlich (95%) der wahre Mittelwert

$\bar{x} \pm z \cdot SE$

Konfidenzintervall für Mittelwerte µ

Hinweis: Nur bei genügend großem Stichprobenumfang ($N \geq 30$) kann der Zentrale Grenzwertsatz auf den Mittelwert angewendet werden

Das ist die Formel für den Standardfehler

$$[K_u; K_o] = \left[\bar{x} - z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}}\, ;\, \bar{x} + z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}}\right]$$

Das heißt: Die Mitte zwischen γ (z.B. 95%) und 100%

(weil wir die $(1-\gamma) = 5\%$ halbieren und auf beide Seiten der Normalverteilung aufteilen)

$$z_{\frac{1+95\%}{2}} = z_{\frac{1{,}95}{2}} = z_{0{,}975}$$

$P(x \leq z)$	z
0,5%	−2,58
1,0%	−2,33
2,5%	−1,96
5,0%	−1,64
95,0%	1,64
97,5%	1,96
99,0%	2,33
99,5%	2,58

Konfidenzintervall für Anteilswerte (p)

- Anteilswerte (p = probability oder percentage) sind Mittelwerte für 0/1-Variablen, also nur „vorhanden“ oder „nicht vorhanden“.
- $p = \mu$ = Mittelwert = Anteilswert
- Die Standardabweichung lässt sich hier einfacher berechnen:

$$s^2 = p \cdot (1 - p)$$

- Daraus folgt der Standardfehler

$$\mathrm{SE} = \frac{s}{\sqrt{N}} = \sqrt{\frac{p \cdot (1 - p)}{N}}$$

Konfidenzintervall für Anteilswerte (p)

- Wenn wir den Standardfehler direkt aus p berechnen können …

$$\mathrm{SE} = \frac{s}{\sqrt{N}} = \sqrt{\frac{p \cdot (1-p)}{N}}$$

Zur Erinnerung
$\bar{x} \pm z \cdot SE$

- Ergibt sich folgende Formel für das Konfidenzintervall

$$[K_u;\ K_o] = \left[p - z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{N}}\ ;\ p + z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{N}}\right]$$

µ geschätzt als p = $\overline{x}$

Das ist die Formel für den Standardfehler (s. oben)

Wichtige Take-Aways

- Wenn Stichproben durch eine echten Zufallsziehung gezogen werden, dann unterliegen sie (und ihre Kennwerte) einem **Stichprobenfehler**.
- Der **Standardfehler** ist die Standardabweichung für Kennwerte (z.B. Mittelwert, Anteilswert) über mehrere (gedachte) Stichproben hinweg.
- Das **Konfidenzintervall** ist ein Bereich um den beobachteten Kennwert herum, in welchem der wahre Kennwert (z.B. Mittelwert oder Anteilswert) wahrscheinlich liegt.
- Die akzeptierte **Irrtumwahrscheinlichkeit** besagt, in wie viele Fällen (z.B. Studien) die Aussage falsch ist, dass der wahre Kennwert im Konfidenzintervall liegt.

ÜBUNGSBLATT: AUFGABE 1

Aufgabe 1a – Approximatives Konfidenzintervall

Eine repräsentative Studie an 200 zufällig ausgewählten Personen* hat ergeben, dass Paare in Deutschland durchschnittlich 8 Minuten (s = 2,0 Minuten) pro Tag miteinander kommunizieren.

Konstruieren Sie ein 95%-Konfidenzintervall für den tatsächlichen Durchschnittswert in der deutschen Bevölkerung.

* Zur Erinnerung: Wir können mit den Annahmen zur Normalverteilung nur den Auswahl-/Stichprobenfehler kontrollieren, keine systematischen Verzerrungen!

Aufgabe 1a – Approximatives Konfidenzintervall

- Verfügbare Informationen
  - $N = 200$
  - $\gamma = 95\% = 0{,}95$
  - $\bar{x} = 8{,}0\ Min.$
  - $s = 2{,}0\ Min.$
- Formel
  - $CI = \bar{x} \pm z \cdot SE$ $\longrightarrow$ $[K_u; K_o] = \left[\bar{x} - z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}}\ ;\ \bar{x} + z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}}\right]$

Aufgabe 1a – Approximatives Konfidenzintervall

$P(x \leq z)$	z
0,5%	−2,58
1,0%	−2,33
2,5%	−1,96
5,0%	−1,64
95,0%	1,64
97,5%	1,96
99,0%	2,33
99,5%	2,58

- Verfügbare Informationen
  - $N = 200$
  - $\gamma = 95\% = 0,95$
  - $\bar{x} = 8,0\ Min.$
  - $s = 2,0\ Min.$
- Formel & Tabelle
  - $CI = \bar{x} \pm z \cdot SE \longrightarrow [K_u; K_o] = \left[\bar{x} - z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}}\ ;\ \bar{x} + z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}}\right]$
- Werte einsetzen

lt. Tabelle

  - $z_{\frac{1+\gamma}{2}} = z_{\frac{1+0,95}{2}} = z_{0,975} = 1,96$
  - $[K_u; K_o] = \left[8 - 1,96 \cdot \frac{2}{\sqrt{200}}\ ; 8 + 1,96 \cdot \frac{2}{\sqrt{200}}\right] = [\,8 - 0,277\ ; 8 + 0,277] = [\,7,72\ ; 8,28]$

Aufgabe 1a – Interpretation

In der Stichprobe kommunizierten die Paare im Durchschnitt 8 Minuten pro Tag miteinander (s = 2,0). Der wahre Mittelwert liegt mit hoher Wahrscheinlichkeit* (95%) im Bereich 7,7 bis 8,3 Minuten pro Tag.

*Interpretation der Interpretation

Der wahre Wert liegt entweder im Intervall oder eben nicht. Das Risiko, dass er aufgrund des Strichprobenfehlers außerhalb des Intervalls liegt, ist 5%.

Wenn wir unendlich viele Stichproben aus der Grundgesamtheit ziehen und 95%-Konfidenzintervalle um den Stichprobenmittelwert konstruieren, dann enthalten 95% dieser Intervalle den wahren Wert der Grundgesamtheit und 5% nicht.

Aufgabe 1b – Approximatives Konfidenzintervall

Wie verändert sich das Konfidenzintervall, wenn Sie die Sicherheit auf 99% erhöhen (also das Risiko auf 1% reduzieren) möchten?

Aufgabe 1b – Approximatives Konfidenzintervall

- Was ändert sich?
  - $\gamma = 99\% = 0{,}99$
- Werte einsetzen
  - $z_{\frac{1+\gamma}{2}} = z_{\frac{1+0{,}99}{2}} = z_{0{,}995} = 2{,}58$

Der z-Wert wird größer für $\gamma = 99\%$ (vorher: 1,96)

$P(x \leq z)$	z
0,5%	−2,58
1,0%	−2,33
2,5%	−1,96
5,0%	−1,64
95,0%	1,64
97,5%	1,96
99,0%	2,33
99,5%	2,58

Aufgabe 1b – Approximatives Konfidenzintervall

- Was ändert sich?
  - $\gamma = 99\% = 0{,}99$
- Werte einsetzen
  - $z_{\frac{1+\gamma}{2}} = z_{\frac{1+0{,}99}{2}} = z_{0{,}995} = 2{,}58$
  - $[K_u; K_o] = \left[8 - 2{,}58 \cdot \frac{2}{\sqrt{200}}\ ; 8 + 2{,}58 \cdot \frac{2}{\sqrt{200}}\right] =$
  - $[\,8 - 0{,}365\ ; 8 + 0{,}365] = [\,7{,}64\ ; 8{,}36]$

Der z-Wert wird größer für $\gamma = 99\%$ (vorher: 1,96)

Entsprechend wird auch das Intervall größer
(vorher [7,72 ;8,28])

$P(x \leq z)$	z
0,5%	−2,58
1,0%	−2,33
2,5%	−1,96
5,0%	−1,64
95,0%	1,64
97,5%	1,96
99,0%	2,33
99,5%	2,58

Wenn wir seltener falsch liegen möchten (1% der Studien statt 5% der Studien), dann müssen wir mehr Sicherheit einkalkulieren, und dadurch wird die Schätzung ungenauer, also das Intervall größer.

Konfidenzintervall in R

Konfidenzintervall in R

- Die Funktion `tidycomm::describe()` liefert u.a. das 95% Konfidenzintervall

```
> stichprobe %>% describe()
# A tibble: 1 × 15
  Variable      N Missing      M     SD    Min    Q25    Mdn    Q75    Max Range CI_95_LL CI_95_UL Skewness Kurtosis
* <chr>     <int>   <int>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 value        50       0   172.   15.0   140.   160.   171.   180.   208.  68.1     167.     176.    0.272     2.95
```

- Alternativen
  - Formel verwenden mit `qt()` für den z-Wert (mit df gegen ∞ oder n-1)
  - `stats::confint()`

Übungsblatt: Aufgabe 2

Aufgabe 2 – Konfidenzintervall für Anteilswerte

Bei der Bundestagswahl 2021 lag die Wahlbeteiligung deutschlandweit bei 76,4%. Für eine Studie wurden zufällig 120 Personen ausgewählt, welche regelmäßig die Fernsehsendung „Hart aber Fair“ sehen. Von diesen gaben 102 Personen an, dass sie 2021 gewählt haben.

Können Sie auf Basis dieser Studie mit einer Sicherheit von 95% behaupten, dass sich die Wahlbeteiligung bei Zuschauerinnen und Zuschauern der Sendung „Hart aber Fair“ von der allgemeinen Wahlbeteiligung unterscheidet?*

*anders formuliert: Können Sie mit 95% Sicherheit sagen, dass der wahre Mittelwert aller Zuschauerinnen und Zuschauern der Sendung „Hart aber Fair“ nicht 76,4% ist?

Aufgabe 2 – Approximatives Konfidenzintervall für $\boldsymbol{p}$

- Verfügbare Informationen
  - $N = 120$
  - $\gamma = 95\% = 0{,}95$
  - $p = \frac{102}{120} = 85\%$
- Formel

$$CI = \bar{x} \pm z \cdot SE \longrightarrow [K_u; K_o] = \left[p - z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{N}}\ ;\ p + z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{N}}\right]$$

Aufgabe 2 – Approximatives Konfidenzintervall für $\boldsymbol{p}$

$P(x \leq z)$	z
95,0%	1,64
97,5%	1,96
99,0%	2,33
99,5%	2,58

- Verfügbare Informationen
  - $N = 120$
  - $\gamma = 95\% = 0{,}95$
  - $p = \frac{102}{120} = 85\%$
- Formel & Tabelle

$$CI = \bar{x} \pm z \cdot SE \longrightarrow [K_u; K_o] = \left[p - z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{N}}\ ;\ p + z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{N}}\right]$$

- Werte einsetzen
  - $z_{\frac{1+\gamma}{2}} = z_{\frac{1+0,95}{2}} = z_{0,975} = 1{,}96$
  - $margin = z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{N}} = 1{,}96 \cdot \sqrt{\frac{0{,}85\cdot(1-0{,}85)}{120}} = 1{,}96 \cdot 0{,}0326 = 0{,}0639 = 6{,}39\%$
  - $[K_u; K_o] = [\,0{,}85 - 0{,}0639\ ; 0{,}825 + 0{,}0639] = [\,78{,}6\%\ ; 91{,}4\%]$

Aufgabe 2 – Ergebnis

Bei der Bundestagswahl 2021 lag die Wahlbeteiligung deutschlandweit bei 76,4%. Für eine Studie wurden zufällig 120 Personen ausgewählt, welche regelmäßig die Fernsehsendung „Hart aber Fair" sehen. Von diesen gaben 102 Personen an, dass sie 2021 gewählt haben.

Auf Basis der Befragung können wir mit einer Sicherheit von 95% davon ausgehen, dass die Wahlbeteiligung unter den Zuschauerinnen und Zuschauern von „Hart aber Fair" zwischen 78,6% und 91,4% liegt.

Aufgabe 2 – Interpretation

Bei der Bundestagswahl 2021 lag die Wahlbeteiligung deutschlandweit bei 76,4%. Für eine Studie wurden zufällig 120 Personen ausgewählt, welche regelmäßig die Fernsehsendung „Hart aber Fair" sehen. Von diesen gaben 102 Personen an, dass sie 2021 gewählt haben.

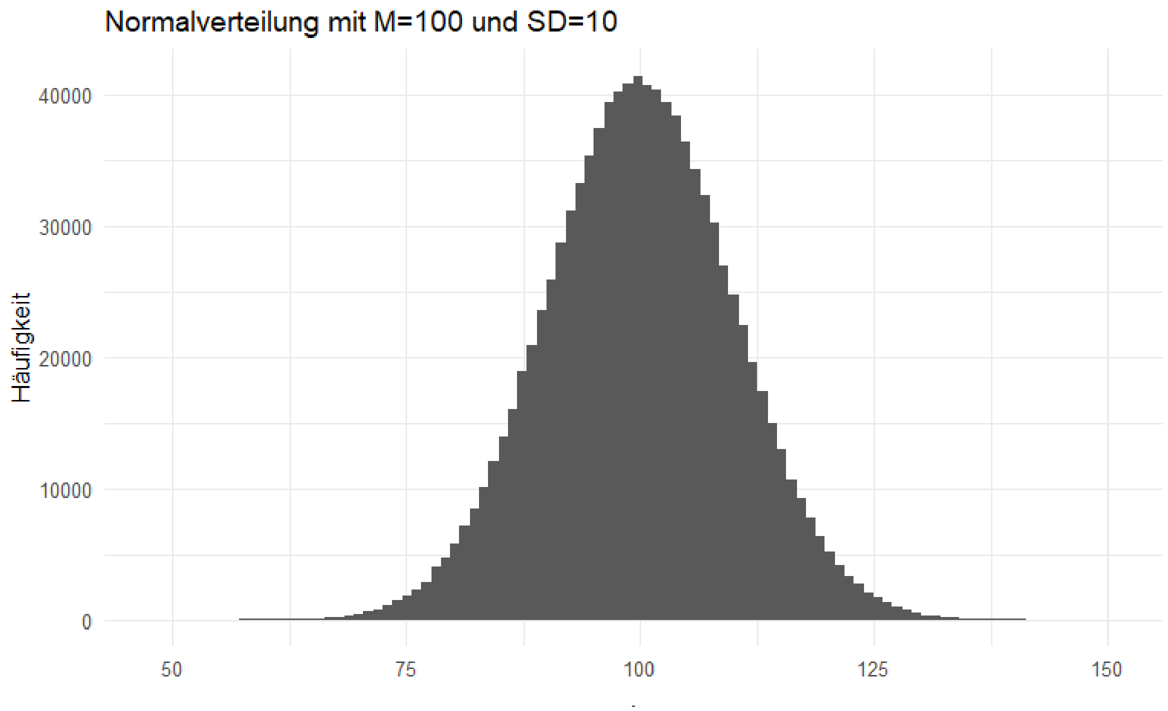
Auf Basis der Befragung können wir mit einer Sicherheit von 95% davon ausgehen, dass die Wahlbeteiligung unter den Zuschauerinnen und Zuschauern von „Hart aber Fair" zwischen 78,6% und 91,4% liegt.

Da sich der Wert der Gesamtbevölkerung (p = 76,4%) nicht in diesem Intervall befindet, gehen wir davon aus, dass sich die Zuschauerinnen und Zuschauer von „Hart aber Fair" in ihrer Wahlbeteiligung systematisch von der Gesamtbevölkerung unterscheiden (bei dieser Aussage nehmen wir eine Irrtumswahrscheinlichkeit von 5% in Kauf).

Verteilung und Normalverteilung

Normalverteilung

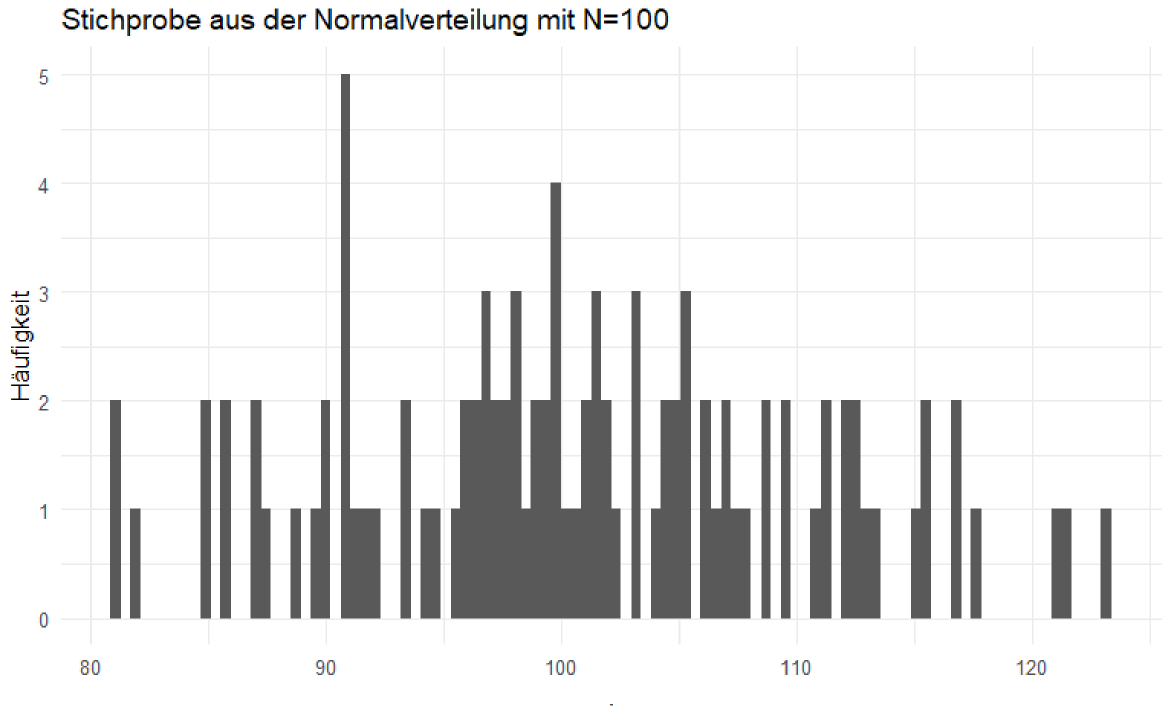
- Die Normalverteilung ist Kernelement vieler statistischer Verfahren
  (zur Herkunft siehe: Binomialverteilung)
- In der Praxis entsteht eine Normalverteilung dadurch, dass es einen Soll-Wert gibt, und dieser wird von individuellen Abweichungen überlagert.
- Die Körpergröße ist (nach Kontrolle des Geschlechts) normalverteilt
- Das Alter ist <u>nicht</u> normalverteilt



Prüfung auf Normalverteilung

- Viele statistische Analysetechniken, bei denen Verteilungsannahmen zugrunde liegen (z.B. t-Tests, Varianzanalyse, Korrelation und Regression), gehen davon aus, dass die untersuchten Merkmale in der Grundgesamtheit normalverteilt sind.

- Die tatsächlichen Verteilungen in der Grundgesamtheit ist häufig unbekannt.
- Wir müssen also anhand unserer Stichprobe abschätzen, ob das Merkmal in der Grundgesamtheit (wahrscheinlich) normalverteilt ist.

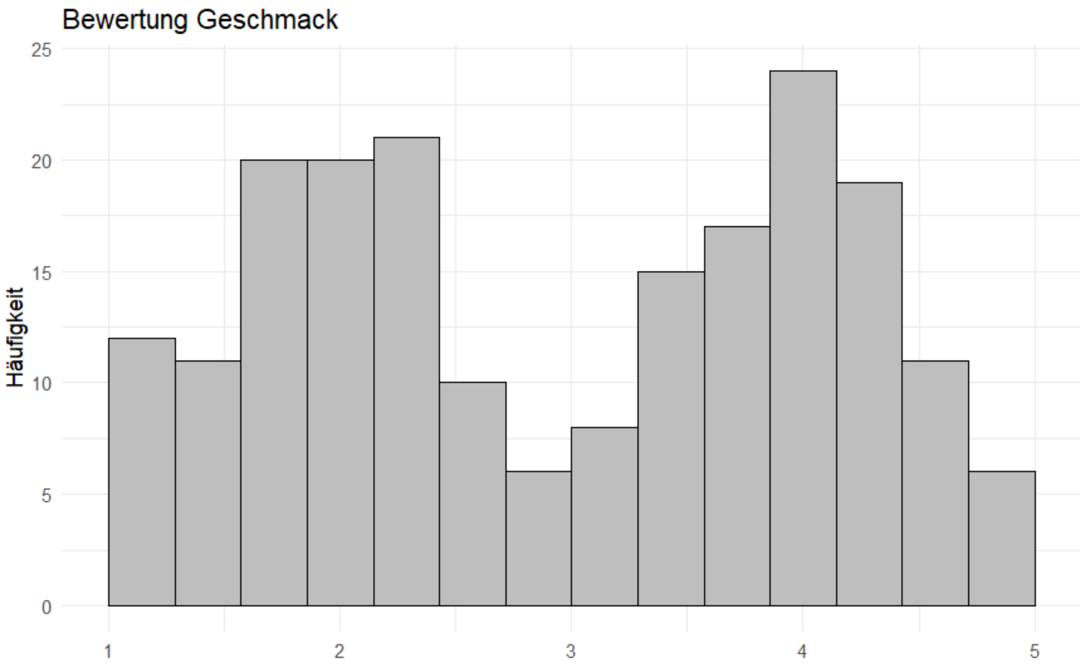


Warum die Prüfung?

Merkmal ist in der Grundgesamtheit normalverteilt	Annahmen korrekt		Ergebnis korrekt (z.B. p-Wert)
Kleine Abweichung gegenüber der Normalverteilung	Annahmen nicht ganz korrekt		Ergebnis nicht ganz korrekt
Große Abweichung gegenüber der Normalverteilung	Annahmen ziemlich **falsch**		Ergebnis ziemlich **falsch**

Bimodale Verteilung

- „Zweigipflige Verteilung"
- Vorsicht mit der Interpretation von Mittelwerten und Standardabweichung
- Möglicherweise zwei Teilpopulationen



Überprüfung der Normalverteilung

- Es sieht aus wie eine Normalverteilung
  - Histogramm
  - Q-Q-Plot
  - Schiefe und Kurtosis

- Wir sind nicht sicher ($p > 5\%$) dass es <u>keine</u> Normalverteilung ist
  - Shapiro-Wilk Test
  - Kolmogorov-Smirnov Test

Überprüfung der Normalverteilung

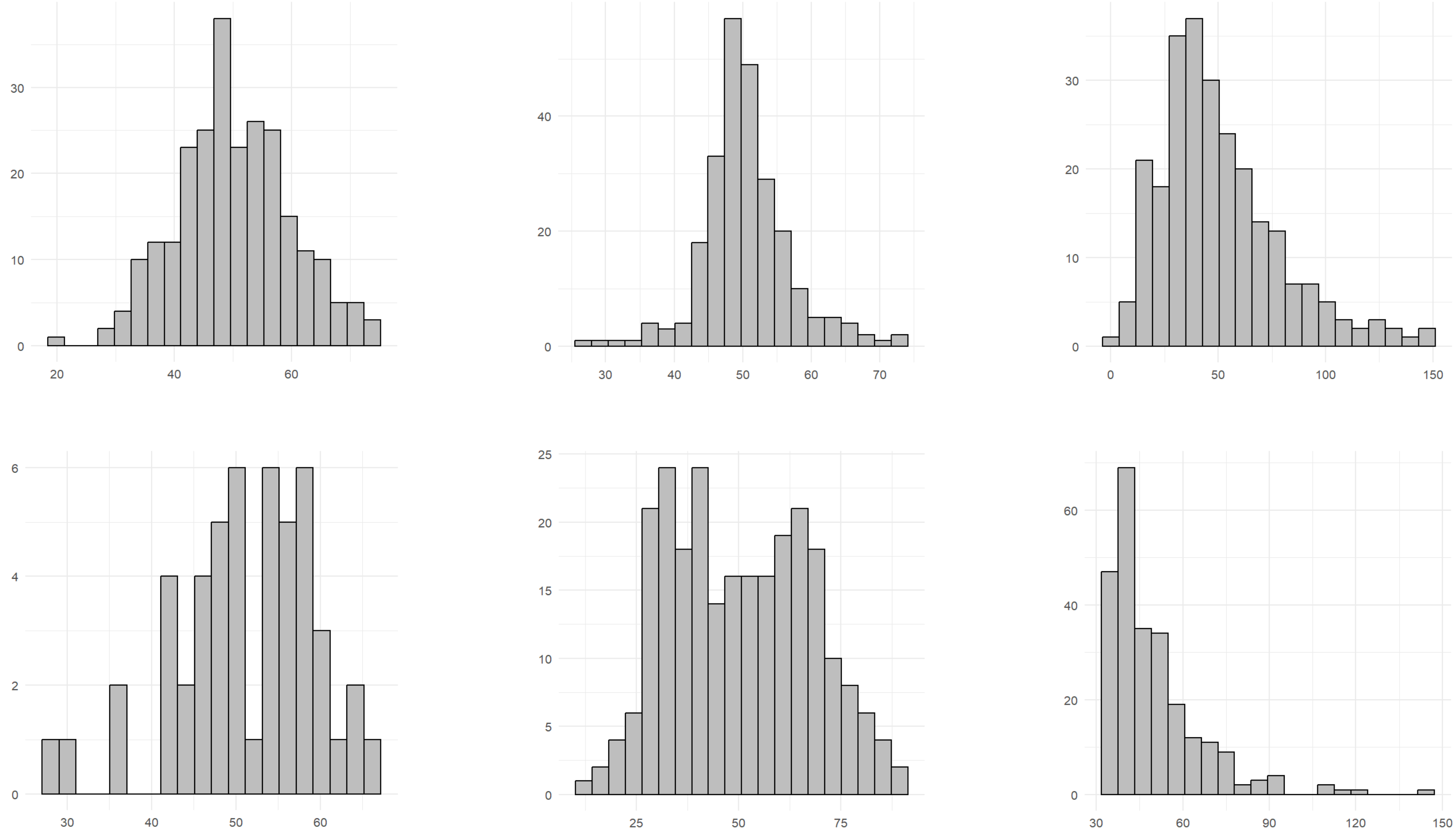
- Es sieht aus wie eine Normalverteilung
  - Histogramm
  - Q-Q-Plot
  - Schiefe und Kurtosis

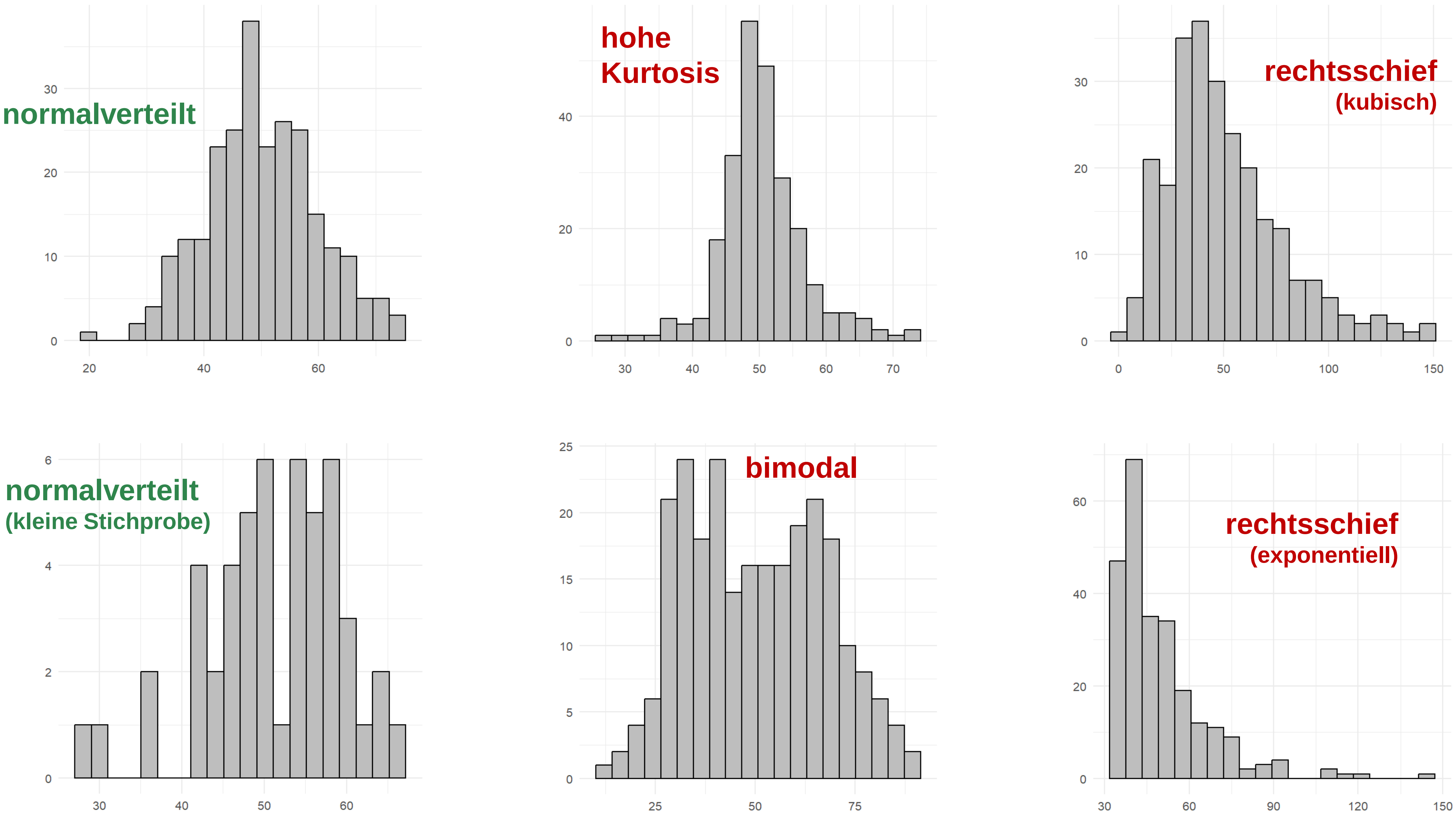
- Wir sind nicht sicher ($p > 5\%$) dass es <u>keine</u> Normalverteilung ist
  - Shapiro-Wilk Test
  - Kolmogorov-Smirnov Test

Das Ergebnis dieser
statischen Tests ist abhängig von

Stärke der Abweichung
×
Anzahl der Fälle

nur diese würde uns interessieren


0
10
20
30
20
40
60
0
20
40
30
40
50
60
70
0
10
20
30
0
50
100
150
0
2
4
6
30
40
50
60
0
5
10
15
20
25
25
50
75
0
20
40
60
30
60
90
120
150


normalverteilt
hohe Kurtosis
rechtsschief (kubisch)
normalverteilt (kleine Stichprobe)
bimodal
rechtsschief (exponentiell)

Überprüfung der Normalverteilung (1) Histogramm

Bitte erzeugen Sie
mit dem R-Code
die fünf Verteilungen

```r
# Normalverteilung

## Erzeugen normalverteilter Werte
set.seed(1)
data_norm = rnorm(250, 50, 10)

## Erzeugen nicht-normalverteilter Werte
### Spitze Verteilung
data_curtosis = c(data_norm[1:125], (data_norm[126:250] - 50) / 4 + 50)
### Zweigipflige Verteilungen
data_bimodal = c(data_norm[1:125] - 15, data_norm[126:250] + 15)
### Kubische Verteilung
data_cubic = data_norm ^ 3 / 2800
### Exponentielle Verteilung
data_exponential = rexp(250, rate=3) * 50 + 34
```

Und diesen Code können Sie kopieren, wenn Sie ein Histogramm benötigen

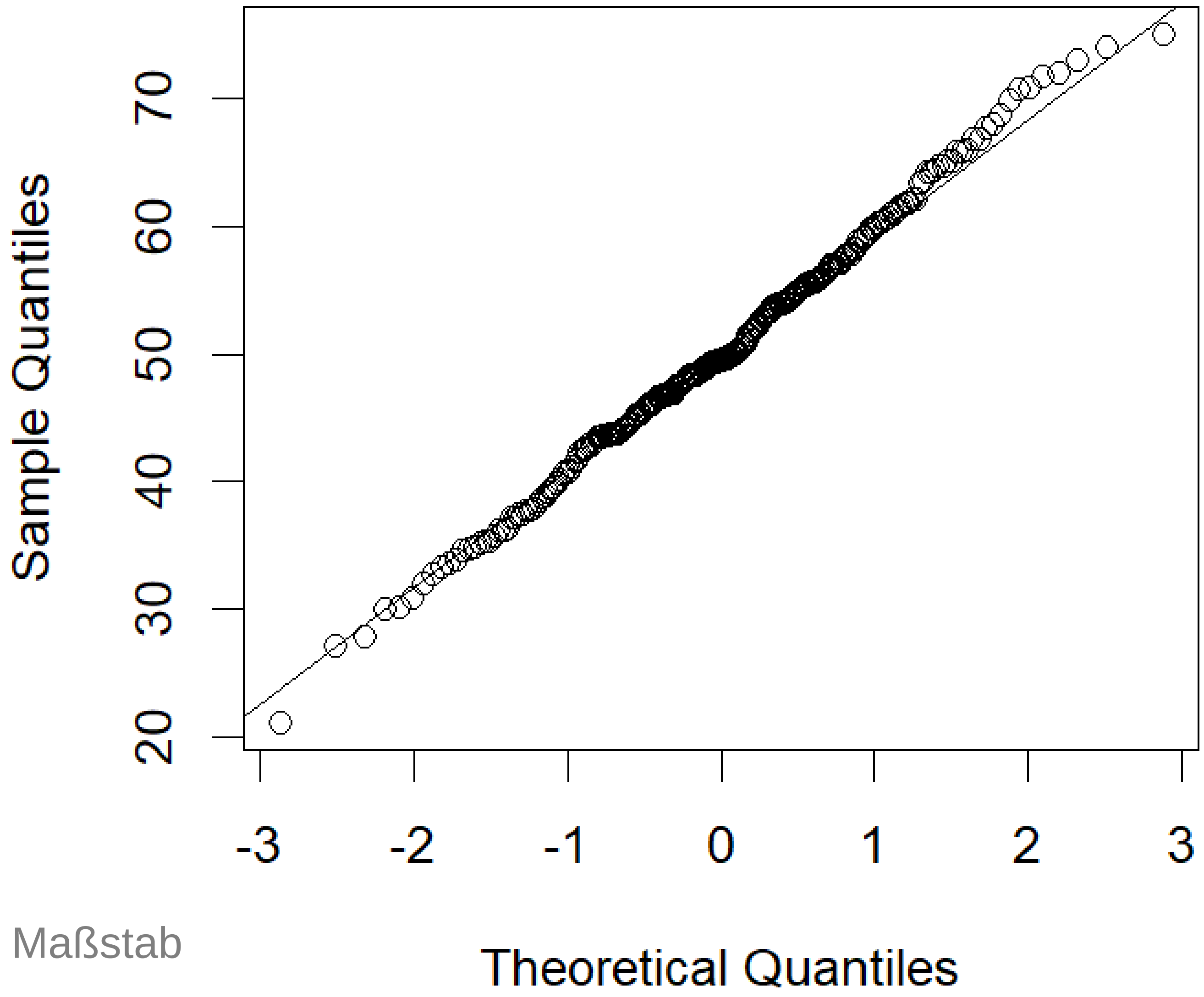
```r
data_norm %>%
    as_tibble() %>% ggplot(aes(x = value)) + theme_minimal() +
    geom_histogram(bins = 20, fill = "grey", color = "black")
```

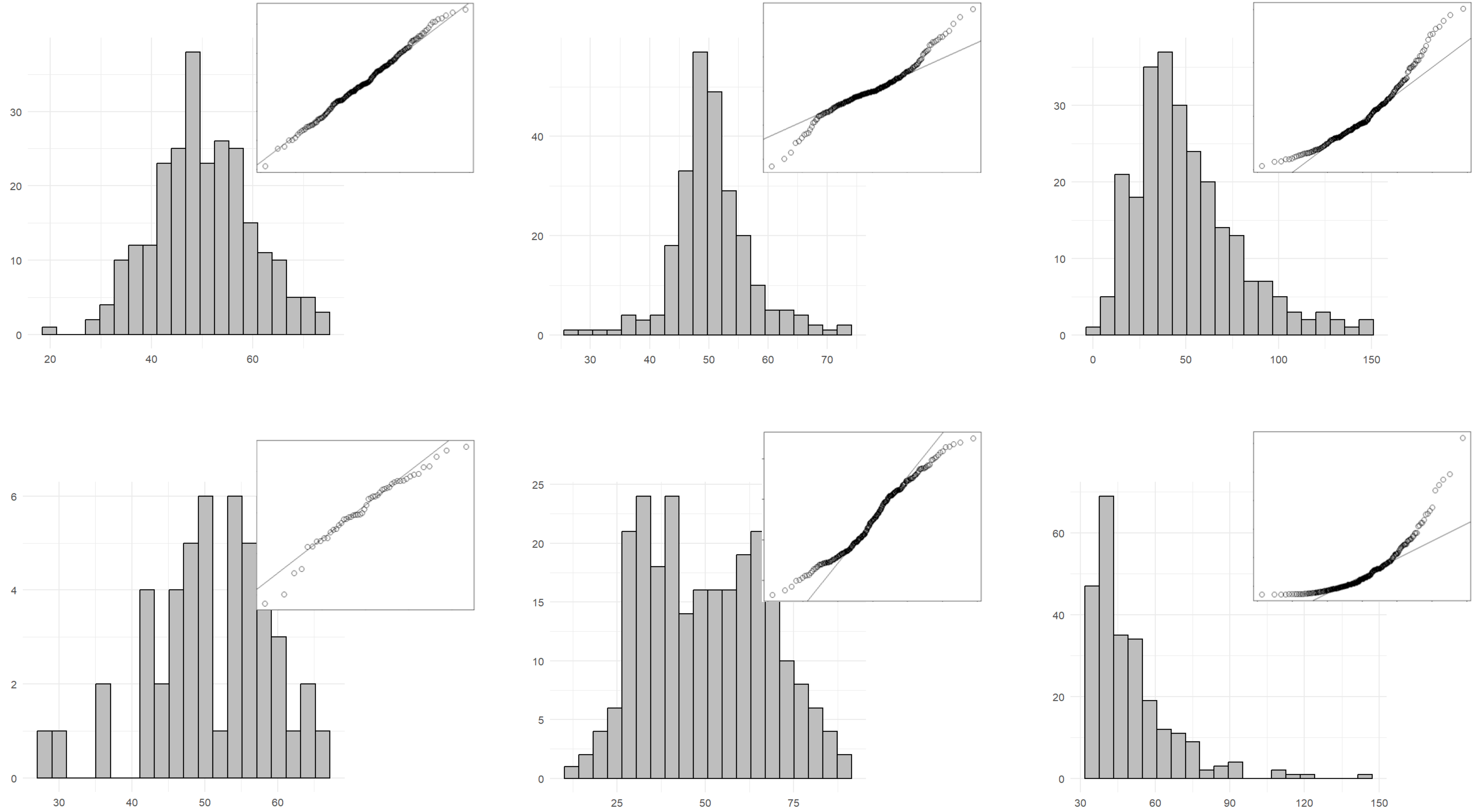
Überprüfung der Normalverteilung (2) Q-Q-Plot

Der Q-Q-Plot trägt die Einzelwerte in einem variierenden Abstand an, sodass sich bei einer Normalverteilung eine Gerade ergibt

```
data_norm %>% qqnorm()

data_norm %>% qqline()
```

optionale Linie als Maßstab




0
10
20
30
20
40
60
0
20
40
30
40
50
60
70
0
10
20
30
0
50
100
150
0
2
4
6
30
40
50
60
0
5
10
15
20
25
25
50
75
0
20
40
60
30
60
90
120
150

Überprüfung der Normalverteilung (3) Schiefe und Wölbung

Quantifizieren der Form einer Verteilung – relativ zur Normalverteilung

- **Schiefe (Skewness)**
  - Liegt der Großteil der Werte nach rechts (>0) versetzt zur Mitte?
  - Die Normalverteilung hat eine Schiefe von 0
  - Problematisch (im Sinne der Voraussetzung*) ist eine Abweichung von mehr als ±2
- **Wölbung (Kurtosis)**
  - Ist die Verteilung spitzer (>3) als eine Normalverteilung?
  - Die Normalverteilung hat eine Wölbung von 3
  - Der „Exzess“ ist die Wölbung minus 3 (damit die Normalverteilung den Wert 0 hat)
  - Problematisch (im Sinne der Voraussetzung*) ist eine Abweichung von mehr als ±2

* Trochim, W. M., & Donnelly, J. P. (2006). The research methods knowledge base (3rd ed.). Cincinnati, OH:Atomic Dog.
Gravetter, F., & Wallnau, L. (2014). Essentials of statistics for the behavioral sciences (8th ed.). Belmont, CA: Wadsworth.
Field, A. (2009). Discovering statistics using SPSS. London: SAGE.

Überprüfung der Normalverteilung (3) Schiefe und Wölbung

```
data %>% describe()
```

	Variable	N	Missing	M	SD		Skewness	Kurtosis
*	<chr>	<int>	<int>	<dbl>	<dbl>		<dbl>	<dbl>
1	data_norm	250	0	50.2	9.63		0.0545	2.97
2	data_curtosis	250	0	50.5	6.46	…	0.318	5.08
3	data_bimodal	250	0	50.2	17.1		0.153	2.09
4	data_cubic	250	0	50.2	28.1		1.08	4.16
5	data_exponential	250	0	50.0	16.4		2.14	9.30

```
data_norm %>%
    as_tibble() %>%
    describe()

data_norm[1:50] %>%
    as_tibble() %>%
    describe()

data_curtosis %>%
    as_tibble() %>%
    describe()

data_bimodal %>%
    as_tibble() %>%
    describe()

data_cubic %>%
    as_tibble() %>%
    describe()

data_exponential %>%
    as_tibble() %>%
    describe()
```

Überprüfung der Normalverteilung (3) Schiefe und Wölbung

```
data %>% describe()
```

	Variable	N	Missing	M	SD		Skewness	Kurtosis
*	<chr>	<int>	<int>	<dbl>	<dbl>		<dbl>	<dbl>
1	data_norm	250	0	50.2	9.63		0.0545	2.97
2	data_curtosis	250	0	50.5	6.46	…	0.318	5.08
3	data_bimodal	250	0	50.2	17.1		0.153	2.09
4	data_cubic	250	0	50.2	28.1		1.08	4.16
5	data_exponential	250	0	50.0	16.4		2.14	9.30

Schiefe akzeptabel von −2 bis +2

Kurtosis akzeptabel von 1 bis 5

```
data_norm %>%
    as_tibble() %>%
    describe()

data_norm[1:50] %>%
    as_tibble() %>%
    describe()

data_curtosis %>%
    as_tibble() %>%
    describe()

data_bimodal %>%
    as_tibble() %>%
    describe()

data_cubic %>%
    as_tibble() %>%
    describe()

data_exponential %>%
    as_tibble() %>%
    describe()
```

Umgang mit nicht-normalverteilten Merkmalen

- Transformation der Werte (z.B. Logarithmus, Wurzel)
- Nicht-parametrische Verfahren verwenden
- Größe des Fehlers abschätzen, evtl. ignorieren
- Ausreißer separat behandeln (z.B. entfernen)

- Regelmäßiges Problem bei …
  - Netzwerkdaten (z.B. Internet, Social Media)
  - Nutzungsmenge und sonstiges, das eskalieren kann
  - Daten über „seltene“ Ereignisse
  - Merkmale ohne „natürlichen“ Normalwert

nicht Gegenstand dieses Seminars

ÜBUNGSBLATT: AUFGABE 3

Aufgabe 3 – Normalverteilung

Prüfen Sie im Datensatz *Worlds of Journalism** anhand von Schiefe und Kurtosis, ob die Merkmale „Autonomy in news story selection” und „Work experience as a journalist” die Voraussetzung der Normalverteilung für parametrische statistische Verfahren erfüllen. Interpretieren Sie die Kennwerte.

* Zur Erinnerung: Den Datensatz bekommen Sie mit dem tidycomm-Package, z.B. via `tidycomm::WoJ`

Bonusaufgabe:

Erstellen Sie Histogramm und Q-Q-Plot

Aufgabe 3 – Normalverteilung

Autonomy in news story selection

`autonomy_selection`

Work experience as a journalist

`work_experience`

```
WoJ %>% select(autonomy_selection, work_experience) %>% describe()
```

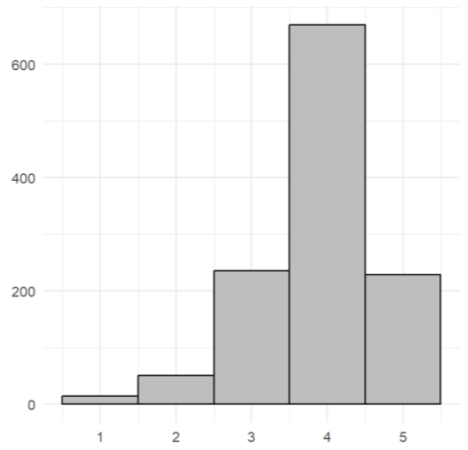
Variable	N	Missing	M	SD	Min	Q25	Mdn	Q75	Max	Range	CI_95_LL	CI_95_UL	Skewness	Kurtosis
<chr>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
autonomy_selection	1197	3	3.88	0.803	1	4	4	4	5	4	3.83	3.92	-0.801	4.19
work_experience	1187	13	17.8	10.9	1	8	17	25	53	52	17.2	18.5	0.427	2.41

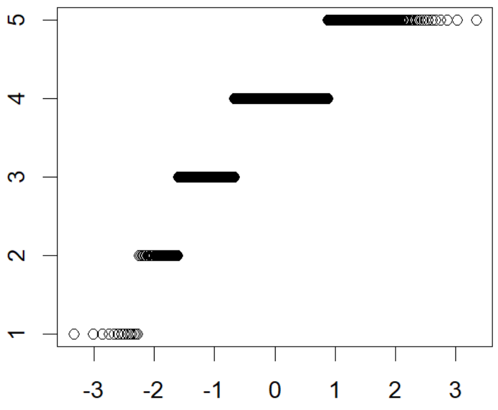
Beide Variablen sind unauffällig, die Voraussetzung der Normalverteilung ist ausreichend erfüllt. Die Autonomie-Einschätzung ist leicht links-schief und zeigt eine erhöhte Kurtosis, die Arbeitserfahrung (in Jahren) ist leicht rechts-schief und etwas abgeflacht. Letzteres könnte daran liegen, dass es für die Arbeitserfahrung keinen natürlich „Normal-Wert“ gibt.

Aufgabe 3 – Normalverteilung

Autonomy in news story selection

`autonomy_selection`

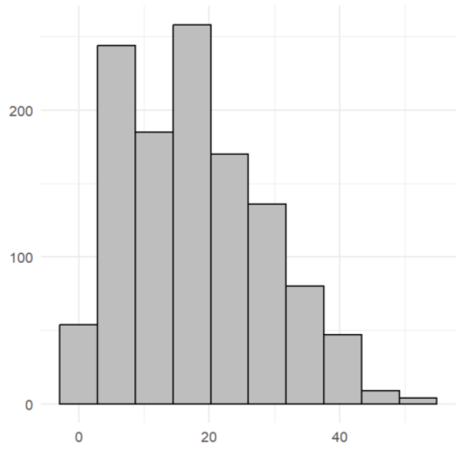


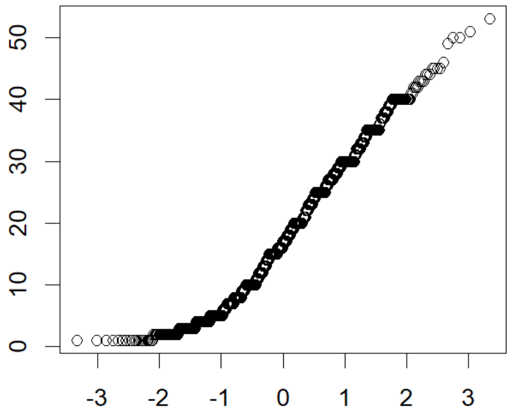



```
WoJ %>%
    ggplot(aes(x = autonomy_selection)) + theme_minimal() +
    geom_histogram(bins = 5, fill = "grey", color = "black")

qqnorm(WoJ$autonomy_selection)
```

Work experience as a journalist

`work_experience`






```
WoJ %>%
    ggplot(aes(x = work_experience)) + theme_minimal() +
    geom_histogram(bins = 10, fill = "grey", color = "black")

qqnorm(WoJ$work_experience)
```

Danke für Ihre Aufmerksamkeit!