
**Forschungsmethodik:
Interventions- und Evaluationsforschung
theoretisch begründen**

Vorlesung Biostatistik, WS 2026/2027

Dr. Marcus Wagner, IMISE

Vorlesung I am 13.10.2026

Inhaltsverzeichnis

0. Zur Organisation des Kurses

1. Einführung

1.1. Motivation und Ziele für den Kurs

1.2. Zum Gegenstand der Statistik

1.3. Grundkonzepte der Statistik; Beispiele

2. Deskriptive Statistik

2.1. Variablentypen

2.2. Kategoriale Daten

2.3. Stichprobe: metrische Variablen

Anmerkungen und Literaturhinweise

Die nächste Hürde?

Es trainieren und assistieren:

- **Vorlesung:**

Marcus Wagner (Lehrbeauftragter)

- **Übungen:**

Miroslav Bacak

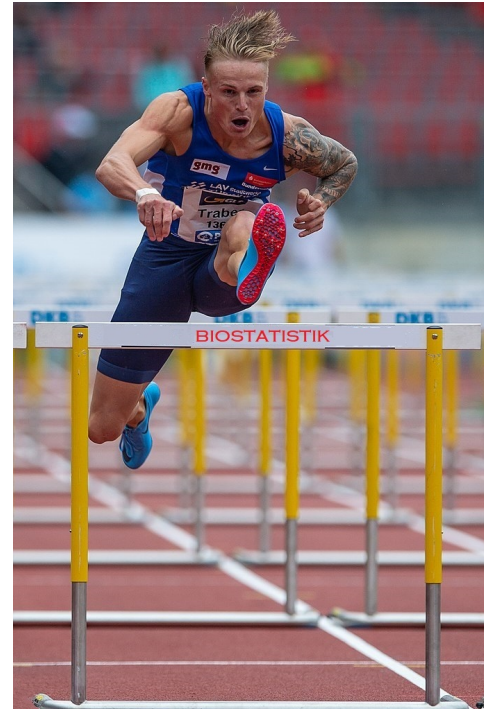
Marcus Wagner

Silke Zachariae

- **Organisation:**

Peter Ahnert (Lehrbeauftragter)

Volkmar Abel (Lehrassistent)



Zur Organisation: Ablauf

- **Vorlesung:** 13.10. – 01.12.2026, 8 Veranstaltungen
dienstags, 08:00 – 09:30 Uhr
Härtelstr. 16-18, Kleiner Hörsaal (3. Etage, Eingang im Hof)
Skript zur Vorlesung jeweils am vorhergehenden Donnerstag auf moodle2
- **Übungen:** 20.10. – 04.12.2026, 7 Veranstaltungen in vier festen Gruppen
Gruppe A: dienstags, 09:45 – 11:15 Uhr Gruppe C: freitags, 09:15 – 10:45 Uhr
Gruppe B: dienstags, 11:30 – 13:00 Uhr Gruppe D: freitags, 11:00 – 12:30 Uhr
Härtelstr. 16-18, PC-Pool IMISE, Raum 006 (Erdgeschoß, Eingang im Hof)
- **Zusammenfassung und Wiederholung:** Dienstag, 08.12.2026, 08:00 – 09:30 Uhr
Härtelstr. 16-18, Kleiner Hörsaal
- **Klausur:** Dienstag, 15.12.2026, ab 08:00 Uhr, in Präsenz
Ort wird noch bekanntgegeben

Zur Organisation: Kontakt und Rückfragen

- **Inhaltliche Fragen zum Kurs:**

Marcus Wagner, `marcus.wagner(a)imise.uni-leipzig.de`

Webseite: `www.thecityto come.de`

Gesprächstermine nach Vereinbarung per Mail

- **Organisatorische Fragen zum Kurs:**

Volkmar Abel, `volkmar.abel(a)uni-leipzig.de`

Sprechzeit: dienstags, 14:00 – 17:00 Uhr, Härtelstr. 16-18, Raum 241

Statistik — verrufen oder maßlos überschätzt

Zwei verfälschte Zitate

“Ich glaube nur der Statistik, die ich selbst gefälscht habe.”

(Zitat wird Winston Churchill (1874 – 1965) zugeschrieben)

“For an educated citizenship in a modern democracy, statistical thinking would be as indispensable as reading and writing.”

(Zitat wird Herbert George Wells (1866 – 1946) zugeschrieben)

Und doch eine Wissenschaft?

Ja, Statistik ist eine wissenschaftliche Disziplin mit exakter und begründeter Methodik, aber gerade deshalb auch mit klar begrenzter Tragweite.

Statistische Fehlinterpretationen und Täuschungen werden sowohl durch Unkenntnis der Methodik als auch durch Unkenntnis ihrer Grenzen begünstigt.

Einführung: Motivation und Ziele für den Kurs

Wir wollen:

- 1) *grundlegende Begriffe, Konzepte, Verfahren und Schlußweisen der Statistik kennenlernen,*
- 2) *"statistical illiteracy" überwinden: statistische Aussagen und Argumente verstehen und beurteilen lernen,*
- 3) *Grundzüge der Planung statistischer Experimente in den Humanwissenschaften ("Studien") erlernen, und*
- 4) *last but not least: ein einfaches Statistik-Softwarepaket bedienen lernen (jamovi).*

Warum?

Einführung: Zum Gegenstand der Statistik

Gegenstand der Statistik in den Humanwissenschaften

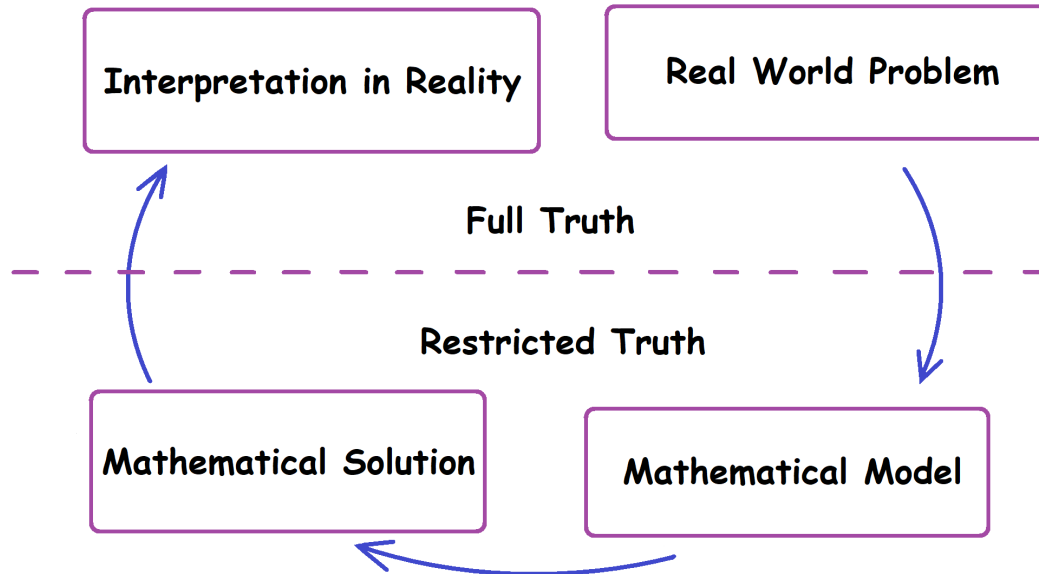
Sie untersucht innerhalb von Gruppen von Menschen, die objektiv abgegrenzt werden können, die Verteilung von Merkmalen, die objektiv ermittelt werden können, vergleicht derartige Verteilungen oder stellt Zusammenhänge (Assoziationen) zwischen ihnen her.

Grenzen der Statistik: methodischer Verzicht

- *Statistik kann (im Normalfall) keine sicheren Aussagen/Voraussagen über einzelne Personen innerhalb der untersuchten Gruppe machen.*
- *Statistik kann Assoziationen zwischen Merkmalen aufdecken/beschreiben, aber keine Ursache-Wirkungs-Zusammenhänge beweisen.*

⇒ Eine sachgerechte Interpretation statistischer Ergebnisse muß stets "von außerhalb" der Statistik kommen!

Wirklichkeit und mathematisches Modell



- *Grundkonzepte der Statistik beruhen auf mathematischen Modellen der Maßtheorie und Wahrscheinlichkeitsrechnung. Innerhalb dieser Modelle sind exakte Aussagen und Schlüsse möglich.*

Erstes Grundkonzept

Statistik für Merkmale mit vollständiger Information

- *sogenannte deskriptive (beschreibende) Statistik*
- *egal, mit welcher Bestimmtheit die Merkmale zustandekommen, werden sie als Zufallsgrößen betrachtet und mit Begriffen der Wahrscheinlichkeitstheorie beschrieben: Verteilung des Merkmals, relative Häufigkeiten, Mittelwert, Standardabweichung, ...*
- *Datenreduktion: wir wollen die ursprüngliche Verteilung des Merkmals mit wenigen Begriffen und Kenngrößen charakterisieren und visualisieren.*

Merke: *Datenreduktion bedeutet Informationsverlust.*

“Im Durchschnitt war der Dorfteich nur einen Meter tief, und trotzdem ist die Kuh ertrunken.”

Beispiel: Vergleich von Geburtenraten

Beispiel 1.1. *(einzelnes Merkmal, vollständige Information)*

Welche der beiden Städte Leipzig und Görlitz hatte im Kalenderjahr 2025 die höhere Geburtenrate?



- **Leipzig:**
632 522 Einwohner per 31.12.2024
Lebendgeborene 2025: 4476,
Rate: 0.71 %
- **Görlitz:**
57 511 Einwohner per 31.12.2024,
Lebendgeborene 2025: 309,
Rate: 0.54 %

Richtig verstanden? *Kristian und Peggy wohnen in Görlitz und sind seit 5 Jahren verheiratet, aber ihr Kinderwunsch hat sich bisher nicht erfüllt. Sollten sie nach Leipzig umziehen?*

Gegenbeispiel: Kundenbindung

Beispiel 1.2. *Wieviel Prozent der Fans des Fußballclubs RB Leipzig haben im Jahr 2025 eine Jahreskarte erworben?*



- *Kann diese Fragestellung Gegenstand der Statistik sein?*

Nein, denn die Gesamtpopulation kann nicht objektiv abgegrenzt werden. Der Club kann zwar belegen, wieviel Jahreskarten im Kalenderjahr 2025 verkauft worden sind, aber wer in diesem Zeitraum "Fan" der Mannschaft war, läßt sich nicht objektivieren: Heute ja, morgen nein, aber vielleicht übermorgen wieder!

Zweites Grundkonzept

Statistik für Merkmale mit unvollständiger Information

- *Wir haben eine Gesamtpopulation und objektiv bestimmbare Merkmale, können sie aber aus praktischen Gründen nicht für jede Person ermitteln. Die vollständige Information ist nur für einen kleinen Teil der Gesamtpopulation ("Stichprobe") verfügbar; für die Stichprobe können wir jedoch deskriptive Statistik durchführen.*
- **Frage:** *Können wir aus der Merkmalsverteilung in der Stichprobe auf die Merkmalsverteilung in der Gesamtpopulation schließen?*
- **Antwort:** *Ja, aber nur mit begrenzter Tragweite. Alle Schlüsse von der Stichprobe auf die Merkmalsverteilungen in der Gesamtpopulation sind mit prinzipieller, unvermeidlicher Unsicherheit behaftet. Diese Unsicherheit können wir allerdings — wieder mit Begriffen der Wahrscheinlichkeitstheorie — zahlenmäßig beschreiben.*

Beispiel: Übergewicht

Beispiel 1.3. (einzelnes Merkmal, unvollständige Information)

Anhand einer Stichprobe von etwa $N = 10\,000$ zufällig ausgewählten Einwohnern (LIFE-ADULT) soll ermittelt werden, wieviel Prozent der erwachsenen Einwohner von Leipzig Übergewicht haben. Indikator dafür: Body-Mass-Index $\geq 25 \text{ kg/m}^2$



- LIFE-ADULT-Studie (Baseline):

Gesamtzahl der Probanden mit BMI-Messung: 10 289 ,
davon Anzahl der Probanden mit $\text{BMI} \geq 25$: 6 716 ,
das sind 65.3 Prozent.

Richtig oder falsch? “Aufgrund dieses Ergebnisses wäre zu erwarten, daß etwa $2/3$ der Teilnehmer unseres Kurses Übergewicht haben.”

Zweites Grundkonzept: Stichproben

Drei Grundsätze

- 1) *Je größer die Stichprobe, desto größer ist der Anteil der Gesamtinformation, der schon in der Stichprobe steckt $\implies$ desto zahlenmäßig kleiner wird die Unsicherheit. Zu kleine Stichproben sagen fast nichts oder gar nichts über die Gesamtpopulation.*
- 2) *Ein statistisches Experiment sollte (mindestens gedanklich) beliebig oft wiederholbar sein. Wenn wir immer wieder gleichgroße Stichproben ziehen und das sehr oft tun, bekommen wir ebenfalls mehr und mehr Information über die Gesamtpopulation.*
- 3) *In der Praxis können wir das Ziehen einer Stichprobe nicht beliebig oft wiederholen, sondern wir haben meistens nur eine einzige Gelegenheit. Die dabei erhaltene Stichprobe sollte dann möglichst repräsentativ für die Gesamtpopulation sein.*

Merke: *Es kommt sowohl auf den Umfang als auch auf die Qualität der Stichprobe an!*

Gegenbeispiel: Verliebtheit

Beispiel 1.4. *Wieviele Studenten an der Universität Leipzig sind derzeit verliebt?*

Dazu wird drei Tage lang eine Umfrage in der Mensa durchgeführt. Die Frage lautet: Sind Sie derzeit in eine andere Person verliebt? Möglich ist eine der drei Antworten: ja / nein / ich weiß nicht.



- *Können diese Fragestellungen Gegenstand der Statistik sein?*

Die erste Fragestellung nicht, denn das Merkmal "Verliebtheit" kann man nicht objektivieren. Die Fragestellung aus der Umfrage schon, denn ob sich jemand als verliebt bezeichnet oder nicht, ist eine objektive, protokollierbare Aussage. Man kann nur nicht, auch nicht nach Anwendung aller irgend denkbaren statistischen Methoden, vom zweiten aufs erste schließen.

Drittes Grundkonzept

Vergleich zwischen Merkmalen mit unvollständiger Information

- *Wir haben in einer Gesamtpopulation zwei Untergruppen und ein objektiv bestimmbares Merkmal, das wiederum aus praktischen Gründen nicht für jede Person ermittelt werden kann. Die vollständige Information ist für zwei Stichproben verfügbar, die jeweils aus einer der beiden Untergruppen kommen.*
- **Frage:** *Zum Merkmalsvergleich führen wir anhand der beiden Stichproben einen statistischen Hypothesentest durch. Können wir das Ergebnis des Vergleichs auf die vollständigen Untergruppen übertragen?*
- **Antwort:** *Ja, aber wieder nur mit begrenzter Tragweite. Die prinzipielle Unsicherheit bei der Auswertung eines statistischen Hypothesentests läßt sich allerdings — wieder mit Begriffen der Wahrscheinlichkeitstheorie — zahlenmäßig beschreiben.*

Beispiel: Der unausgeschlafene Tennisspieler

Beispiel 1.5. (Merkmalsvergleich für Stichproben) [REYNER/HORNE (2013)]

Wird die Zielgenauigkeit beim Aufschlag im Tennis durch Schlafmangel beeinflusst?



- Gruppe A: 8 College-Studenten (Halbprofis), 18 – 22 Jahre, normaler Schlaf (im Mittel 6.5 h)
- Gruppe B: 8 College-Studenten (Halbprofis), 18 – 22 Jahre, verkürzter Schlaf (im Mittel 4.3 h)
- Mittlere Anzahl der Treffer bei 30 Aufschlägen in einen festen Zielbereich:
 $T = 15.75$ (Gruppe A) / $T = 11.25$ (Gruppe B)
- Welch-Test gegen folgende “Nullhypothese” $\mathcal{H}_0$:
“Mittlere Trefferzahl ist in beiden Situationen gleich”
ergibt Kennzahl $p = 0.0007$, “signifikante Ablehnung”

Frage: Kann man daraus schließen, daß bei Tennisspielern im Alter von 18 – 22 Jahren generell das Fehlen von 2 h Nachtschlaf die Trefferquote um ca. 28% reduziert?

Viertes Grundkonzept

Statistik für Assoziationen zwischen mehreren Merkmalen

- Wir verknüpfen ein “Zielmerkmal” Z und “Erklärungsmerkmale” (sog. Kovariaten) $K_1, K_2, K_3, \dots$ durch eine Modellgleichung, zum Beispiel

$$Z = c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + c_3 \cdot K_3 \dots ,$$

und ermitteln die unbekannten Koeffizienten $c_0, c_1, c_2, c_3, \dots$ als Lösung des überbestimmten Gleichungssystems

$$Z(1) \approx c_0 + c_1 \cdot K_1(1) + c_2 \cdot K_2(1) + c_3 \cdot K_3(1) \dots \quad (\text{Daten von Person \#1})$$

$$Z(2) \approx c_0 + c_1 \cdot K_1(2) + c_2 \cdot K_2(2) + c_3 \cdot K_3(2) \dots \quad (\text{Daten von Person \#2})$$

$$\vdots$$
$$\vdots$$

im Sinne des kleinsten quadratischen Fehlers.

- Kommen die Daten aus einer Stichprobe, so können wir wieder nur mit prinzipieller (aber zahlenmäßig bestimmbarer) Unsicherheit schließen, daß die Modellgleichung mit den errechneten Koeffizienten auch für die Gesamtpopulation gilt.

Beispiel: Basketball bei Schulkindern

Beispiel 1.6. (*Assoziation zwischen Merkmalen*) [STÖCKEL ET AL. (2011)]

Dribbeln soll sowohl mit der linken als auch mit der rechten Hand erlernt werden. Bringt es einen Vorteil, wenn man zuerst mit der führenden Hand trainiert?



- Trainingsgruppe A ("ND-D"):
*N = 52 Kinder (alle Rechtshänder),
zuerst 4 Einheiten linke Hand,
dann 4 Einheiten rechte Hand*
- Trainingsgruppe B ("D-ND"):
*N = 52 Kinder (alle Rechtshänder),
zuerst 4 Einheiten rechte Hand,
dann 4 Einheiten linke Hand*

- *Mittlere Verbesserung beim Dribbeln durch Standard-Slalomstrecke:*

$$\Delta = -1.18 \text{ s (Gruppe A)} / \Delta = -0.77 \text{ s (Gruppe B)}$$

Frage: *Ist der Gruppenunterschied nur mit dem Trainingsprogramm oder noch mit weiteren Merkmalen (z.B. Körpergröße, Fitneß, Vortraining) assoziiert?*

Deskriptive Statistik: Variablentypen

Wir unterscheiden kategoriale und metrische Variablen.

- **kategorial:**

binär



nominal

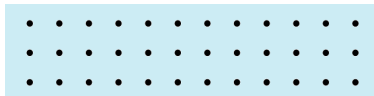


ordinal



- **metrisch:**

ganzzahlig



kontinuierlich



Variablentypen (II)

Beispiel 1.7. *“Gehört die Person zur Interventionsgruppe?” — kategorial, binär*

Beispiel 1.8. *Geschlecht (im Sinne des Chromosomengeschlechts) — kategorial, binär*

Beispiel 1.9. *Augenfarbe — kategorial, nominal*

Beispiel 1.10. *Personenstand: ledig / verheiratet / verwitwet / geschieden
— kategorial, nominal*

Beispiel 1.11. *Prüfungsnoten: sehr gut / gut / befriedigend / genügend / ungenügend
— kategorial, ordinal*

Beispiel 1.12. *HADS-Fragebogen (Item 1): “Ich fühle mich angespannt oder überreizt”:
meistens / oft / gelegentlich / überhaupt nicht — kategorial, ordinal*

Beispiel 1.13. *Geburtsjahr — metrisch, ganzzahlig*

Beispiel 1.14. *Body-Mass-Index — metrisch, kontinuierlich*

Kategoriale Variablen

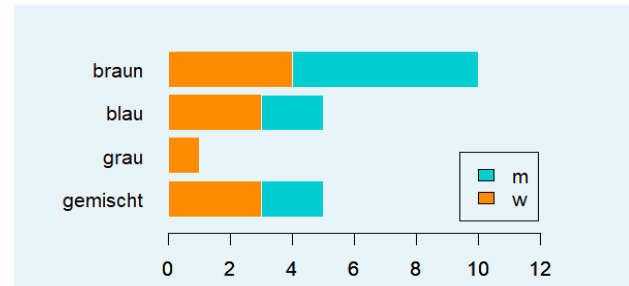
• Beispiel 1.15.: Daten

PROBAND-ID	GESCHLECHT	AUGENFARBE	GRÖSSE (m)
PR-001-1	männlich	blau	1,41
PR-002-2	weiblich	gemischt	1,56
PR-003-0	männlich	braun	1,53
PR-004-1	männlich	blau	1,45
PR-005-2	weiblich	blau	1,54
PR-006-0	männlich	braun	1,49
PR-007-1	weiblich	gemischt	1,31
PR-008-2	weiblich	braun	1,64
PR-009-0	männlich	braun	1,43
PR-010-1	weiblich	grau	1,51
PR-011-2	weiblich	blau	1,50
PR-012-0	weiblich	braun	1,55
PR-013-1	männlich	gemischt	1,57
PR-014-2	männlich	braun	1,60
PR-015-0	weiblich	braun	1,38
PR-016-1	männlich	gemischt	1,58
PR-017-2	weiblich	braun	1,57
PR-018-0	weiblich	blau	1,45
PR-019-1	männlich	braun	1,58
PR-020-2	weiblich	gemischt	1,49
PR-021-0	männlich	braun	1,61

• Häufigkeitstabelle (Augenfarbe)

braun	blau	grau	gemischt	Gesamt
10	5	1	5	21

• Balkendiagramm (Augenfarbe)



• Kontingenztafel (Geschlecht vs. Farbe)

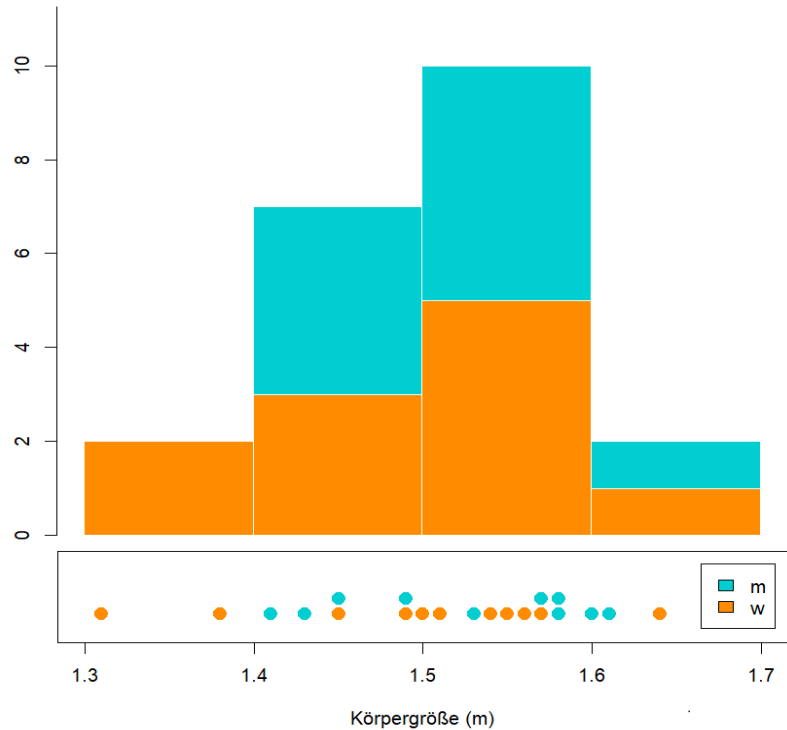
	braun	blau	grau	ge- mischt	Zeilen- summe
männlich	6	2	0	2	10
weiblich	4	3	1	3	11
Spalten- summe	10	5	1	5	21

Metrische Variablen: Histogramm

- **Daten** (aus Beispiel 1.15.)

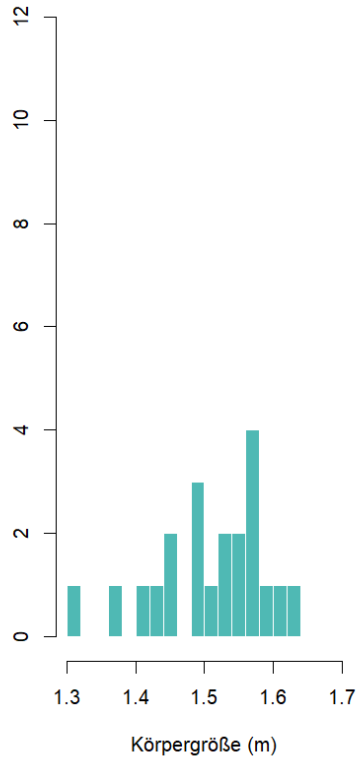
PROBAND-ID	...	GRÖSSE (m)
PR-001-1		1,41
PR-002-2		1,56
PR-003-0		1,53
PR-004-1		1,45
PR-005-2		1,54
PR-006-0		1,49
PR-007-1		1,31
PR-008-2		1,64
PR-009-0		1,43
PR-010-1		1,51
PR-011-2		1,50
PR-012-0		1,55
PR-013-1		1,57
PR-014-2		1,60
PR-015-0		1,38
PR-016-1		1,58
PR-017-2		1,57
PR-018-0		1,45
PR-019-1		1,58
PR-020-2		1,49
PR-021-0		1,61

- **Histogramm** (mit 4 Klassen)

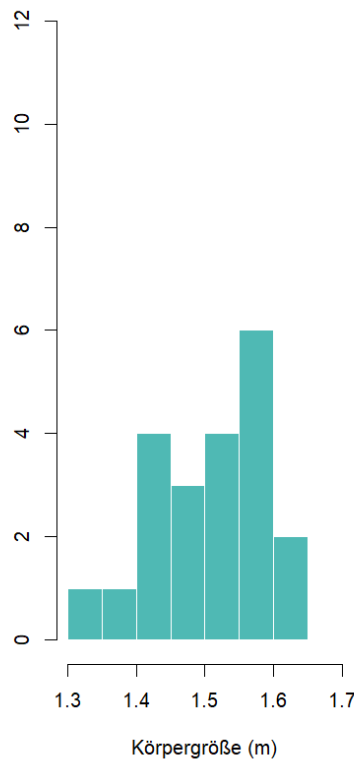


Datenreduktion im Histogramm

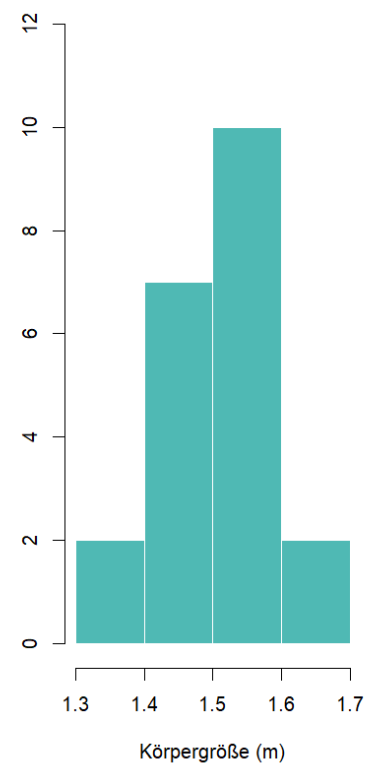
• *Histogramm mit 20 Klassen*



... mit 8 Klassen



... mit 4 Klassen



Mittelwert, Standardabweichung, Spannweite

- **Definition 1.16.:** (Stichproben-Mittelwert) Gegeben ist eine metrische Variable x mit N Messungen $x_1, \dots, x_N$. Dann ist der Stichproben-Mittelwert $\hat{\mu}$ von x als das arithmetische Mittel der Messungen erklärt:

$$\hat{\mu} = \frac{1}{N} \cdot (x_1 + x_2 + \dots + x_{N-1} + x_N) = \frac{1}{N} \cdot \sum_{k=1}^N x_k.$$

- **Definition 1.17.:** (empirische Standardabweichung) Für eine metrische Variable x mit N Messungen ist die empirische Standardabweichung $\hat{\sigma}$ von x erklärt durch

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \cdot \sum_{k=1}^N (x_k - \hat{\mu})^2}; \text{ darin ist } \hat{\mu} \text{ der Stichproben-Mittelwert von } x.$$

- **Definition 1.18.:** (Spannweite, Range) Für eine metrische Variable x mit N Messungen mögen $x_{\min}$ und $x_{\max}$ die kleinste und die größte Messung bezeichnen. Die Spannweite (Range) von x ist erklärt durch $\hat{R} = x_{\max} - x_{\min}$.

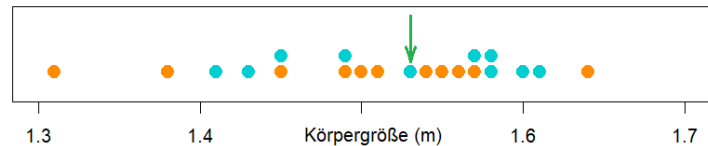
- Der Stichproben-Mittelwert $\hat{\mu}$ ist ein Lagemaß. Empirische Standardabweichung $\hat{\sigma}$ und Spannweite $\hat{R}$ sind Streuungsmaße, die in der gleichen Einheit wie x und $\hat{\mu}$ gemessen werden.

Der Median (als alternatives Lagemaß)

- *Motivation für ein anderes Lagemaß: Rangordnung der Meßwerte statt Mittelbildung*

Frage: Welcher Wert der Variablen x halbiert die geordnete Meßreihe aus Beispiel 1.15.?

Antwort: Es gibt insgesamt 21 Messungen, also die Messung Nr. 11 mit $m = 1.53$.



- **Definition 1.19.:** (Stichproben-Median) Gegeben ist eine metrische Variable x mit N der Größe nach geordneten Messungen $x_1, \dots, x_N$. Der Stichproben-Median von x ist erklärt als

$$\hat{m} = \begin{cases} x_{k+1}, & \text{falls } N = 2 \cdot k + 1 \text{ ungerade ist,} \\ (x_k + x_{k+1})/2 & \text{falls } N = 2 \cdot k \text{ gerade ist.} \end{cases}$$

- Stichproben-Mittelwert und Stichproben-Median einer metrischen Variable sind im allgemeinen verschieden. Im Beispiel 1.15.: $\hat{\mu} = 1.51$, $\hat{m} = 1.53$
- Im Unterschied zum Stichproben-Mittelwert ist der Stichproben-Median unempfindlich gegen Ausreißer.

Quartile und Interquartil-Range

- *Motivation: Weitere Unterteilung einer geordneten Meßreihe durch die sog. Quartile $\hat{Q}_{25}$, $\hat{Q}_{50}$ und $\hat{Q}_{75}$ in vier etwa gleichgrosse Teile.*

- **Definition 1.20.:** *(Quartile) Gegeben ist eine metrische Variable x mit N der Größe nach geordneten Messungen. Dazu ist*

$$\hat{Q}_{25} = \begin{cases} \text{Stichproben-Median von } \{x_1, \dots, x_{k+1}\}, & \text{falls } N = 2 \cdot k + 1 \text{ ungerade ist,} \\ \text{Stichproben-Median von } \{x_1, \dots, x_k\} & , \text{ falls } N = 2 \cdot k \text{ gerade ist,} \end{cases}$$

das 1. Quartil; der Stichproben-Median $\hat{Q}_{50} = \hat{m}$ von $\{x_1, \dots, x_N\}$ das 2. Quartil;

$$\hat{Q}_{75} = \text{Stichproben-Median von } \{x_{k+1}, \dots, x_N\},$$

unabhängig davon, ob N ungerade oder gerade ist, das 3. Quartil.

- **Definition 1.21.:** *(IQR) Das Intervall $[\hat{Q}_{25}, \hat{Q}_{75}]$, der Interquartil-Range (IQR), ist das zum Stichproben-Median passende Streuungsmaß. Seine Länge ist $|IQR|$.*

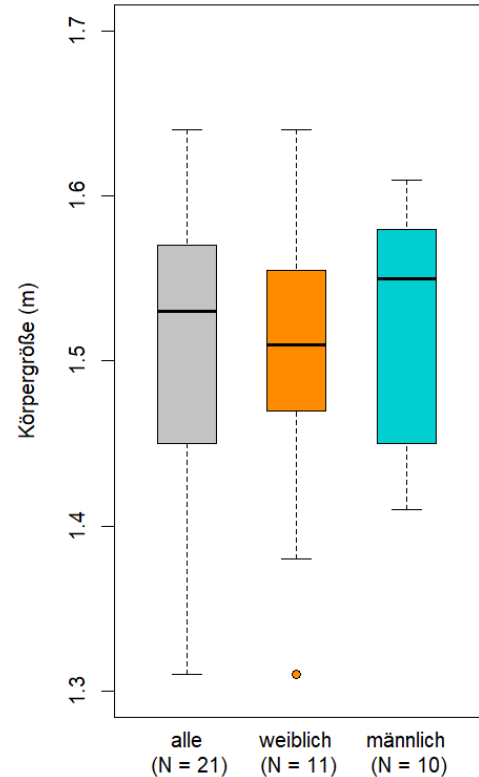
- *In Beispiel 1.15. ist $\hat{Q}_{25} = 1.45$, $\hat{Q}_{75} = 1.57$, $IQR = [1.45, 1.57]$ und $|IQR| = 0.12$.*

Visualisierung mit Boxplots

- **Elemente im Boxplot:**

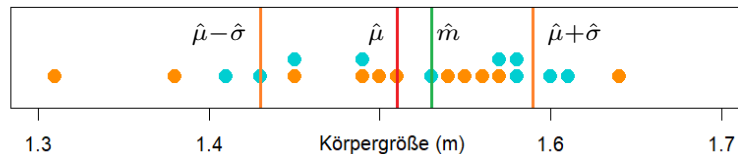
- *Skala: Wertebereich der metrischen Variable*
- *Box: reicht von $\hat{Q}_{25}$ bis $\hat{Q}_{75}$, also $|IQR|$ lang, repräsentiert 50 % der Meßwerte*
- *fette Linie in der Box: Stichproben-Median $\hat{m}$*
- *Antennen ("Whiskers"): beginnen bei $\hat{Q}_{25}$ bzw. $\hat{Q}_{75}$, höchstens $1.5 \cdot |IQR|$ lang, Ende beim minimalen bzw. maximalen Wert in diesem Bereich*
- *Ausreißer: alle Datenpunkte im Abstand von $1.5 \cdot |IQR|$ bis $3 \cdot |IQR|$ von $\hat{Q}_{25}$ bzw. $\hat{Q}_{75}$*
- *extreme Ausreißer: alle Datenpunkte mit Abständen größer als $3 \cdot |IQR|$ von $\hat{Q}_{25}$ bzw. $\hat{Q}_{75}$ (anderes Symbol, im Beispiel nicht vorhanden)*

- **Achtung: extreme Datenreduktion!**

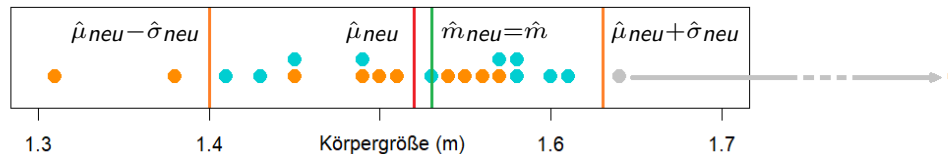


Robustheit gegen Ausreißer

- Mittelwert, empirische Standardabweichung und Spannweite sind empfindlich gegen Ausreißer!
- Beispiel 1.15.: ursprüngliche Werte mit Kenngrößen $\hat{\mu} = 1.51$, $\hat{\sigma} = 0.08$, $\hat{R} = 0.33$



- Fehler: statt 1,64 m wird der Wert 1,94 m aufgezeichnet. Was geschieht?



- Auswirkung: $\hat{\mu}_{neu} = 1.52$, $\hat{\sigma}_{neu} = 0.12$, $\hat{R}_{neu} = 0.63$, während die robusten Kenngrößen Median und IQR hier sogar unverändert bleiben.

Zusatz: Die empirische Verteilungsfunktion

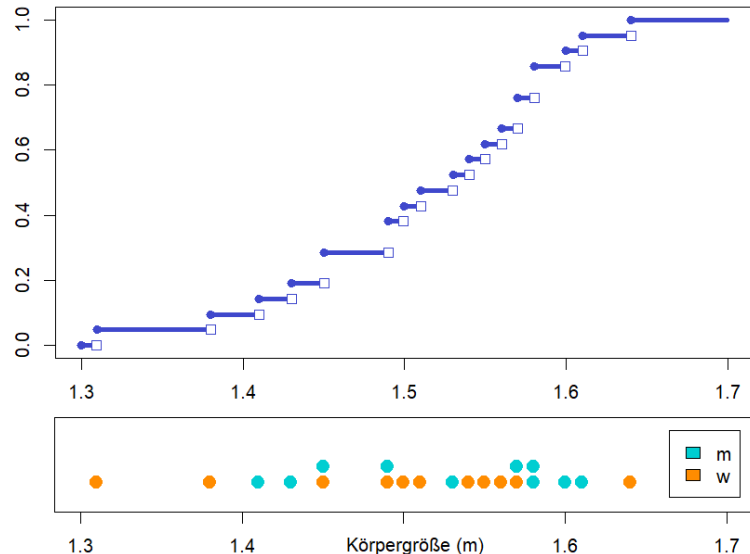
• **Definition 1.22.:** (empirische Verteilungsfunktion $\hat{F}$ für eine metrische Variable x)

Diese ist erklärt als: $\hat{F}(x) = \frac{\# \text{ Anzahl Messungen } \leq x}{\# \text{ Anzahl Messungen gesamt}}$

• **In Worten:** Die Messungen werden der Größe nach geordnet, die Häufigkeiten werden von links summiert und auf die Gesamtzahl der Messungen normiert.

• $\hat{F}$ ist eine Sprungfunktion, wegen " $\leq$ " in der Definition gehören immer die linken Endpunkte der Strecken zum Funktionsgraphen, die rechten nicht.

• $\hat{F}$ enthält die vollständige Information über die Variable x .



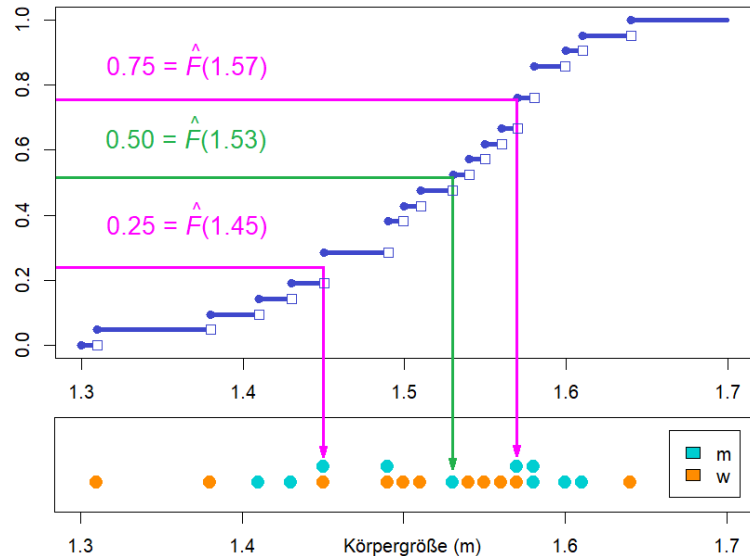
Quartile und empirische Verteilungsfunktion

- Mit Hilfe von $\hat{F}$ können wir die Quartile $\hat{Q}_{25}$, $\hat{Q}_{50}$ und $\hat{Q}_{75}$ neu definieren und unmittelbar veranschaulichen.

- Zur Erinnerung: In Beispiel 1.15. war $\hat{Q}_{25} = 1.45$, $\hat{Q}_{50} = 1.53$, $\hat{Q}_{75} = 1.57$ und $IQR = [1.45, 1.57]$.

- **Definition 1.20.a:** (andere Definition für Quartile) Gegeben ist eine metrische Variable x mit N der Größe nach geordneten Messungen.

Dazu ist $\hat{Q}_{25} = \min\{x \mid \hat{F}(x) \geq 0.25\}$ das 1. Quartil, $\hat{Q}_{50} = \min\{x \mid \hat{F}(x) \geq 0.5\}$ das 2. Quartil und $\hat{Q}_{75} = \min\{x \mid \hat{F}(x) \geq 0.75\}$ das 3. Quartil.



Zusammenfassung

- **Kenngößen für metrische Variablen in einer Stichprobe:** *ergeben sich aus den empirischen Häufigkeiten (sog. diskrete Wahrscheinlichkeitsverteilung)*

	<i>“parametrisch”</i>	<i>“nichtparametrisch”</i>
<i>Definition</i>	<i>aufgrund der Werte</i>	<i>aufgrund der Ränge</i>
<i>Lagemaß</i>	<i>Stichproben-Mittelwert $\hat{\mu}$</i>	<i>Stichproben-Median $\hat{m}$</i>
<i>Streuungsmaße</i>	<i>emp. Standardabweichung $\hat{\sigma}$, (Varianz $\hat{\sigma}^2$, Spannweite $\hat{R}$)</i>	<i>Quartile $\hat{Q}_{25}$ und $\hat{Q}_{75}$, Interquartil-Range IQR</i>
<i>Robustheit gegen Ausreißer</i>	<i>nicht robust</i>	<i>robust geg. wenige Ausreißer</i>

- **Kenngößen für metrische Variablen in der Gesamtpopulation:** *Hier sind neue Berechnungsvorschriften nötig, weil die Merkmale in der Gesamtpopulation nicht mehr mit diskreten, sondern mit stetigen Wahrscheinlichkeitsverteilungen beschrieben werden.*