

UNIVERSITÄT LEIPZIG

Medizinische Fakultät

QSB1

Epidemiologie, Medizinische Biometrie und Medizinische Informatik

Deskriptive Statistik II

Dr. rer. med. habil. Marita Ziepert /

Dr. rer. nat. Meinhard Mende

IMISE, Universität Leipzig

imise.

Lernziele Deskriptive Statistik II

Am Ende der Vorlesung werden Sie

- absolute / relative Häufigkeiten sowie absolute / relative Häufigkeitssummen angeben,
- Lage- und Streuungsmaße berechnen;
- die empirische Verteilungsfunktion interpretieren und statistische Maßzahlen aus ihr ablesen und
- einen Boxplot zeichnen und interpretieren können.

Deskriptive Statistik II

Unter deskriptiver Statistik versteht man die Aufbereitung, Systematisierung und empirisch gewonnener Daten zur Beschreibung von Kohorten.

Methoden:

1. Tabellarische Darstellung
2. Statistische Kenngrößen
3. Grafische Darstellung

3. Deskriptive Statistik II

3.1 Kenngrößen für kategoriale Größen

Beispiel 3.1: die LAP-01-Studie

EUROPEAN UROLOGY 79 (2021) 750–759

available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology



Platinum Priority – Prostate Cancer

Editorial by John W. Yaxley, Geoff Coughlin and Vipul Patel on pp. 760–761 of this issue

Robotic-assisted Versus Laparoscopic Surgery: Outcomes from the First Multicentre, Randomised, Patient-blinded Controlled Trial in Radical Prostatectomy (LAP-01)

Jens-Uwe Stolzenburg[a,†,], Sigrun Holze[a,†], Petra Neuhaus[b], Iason Kyriazis[a,c], Hoang Minh Do[a,d], Anja Dietel[a], Michael C. Truss[e], Corinn I. Grzella[e], Dogu Teber[f,g], Markus Hohenfellner[f], Robert Rabenalt[h,i], Peter Albers[h], Meinhard Mende[b,j]*

erste multizentrische randomisierte Studie, welche laparoskopische radikale Prostatektomie ohne / mit Roboterunterstützung verglich

Kategoriale Daten I

Das sind hier:

- die Gleason-Summe: 5, 6, …, 10
- die geplante Nervschonung: keine, unilateral, bilateral

Table 1 – Baseline characteristics and perioperative data

	RARP N = 547	LRP N = 171
Baseline characteristics (full analysis set)		
Sociodemographic data		
Age at surgery (yr), median (quartiles)	65 (59; 69)	65 (59; 70)
Body size (cm), median (quartiles)	177 (173; 181)	176 (172; 181)
Body weight (kg), median (quartiles)	84 (78; 93)	84 (77; 92)
BMI (kg/m^2), median (quartiles)	27.2 (25.2; 29.4)	27.0 (25.0; 29.1)
PSA, preoperative (ng/ml), median (IQR)	7.7 (5.6, 12.1)	8.1 (6.0, 11.0)
Gleason sum (preop), *n* (%)		
5	1 (0.2)	0 (0)
6	188 (34)	55 (32)
7	258 (47)	83 (49)
8	61 (11)	23 (13)
9	35 (6.4)	9 (5.3)
10	4 (0.7)	1 (0.6)
Planned nerve sparing, *n* (%)		
None	216 (39)	67 (39)
Unilateral	80 (15)	24 (14)
Bilateral	251 (46)	80 (47)

Kategoriale Daten II

Für qualitative Merkmale werden
die absoluten und die relativen Häufigkeiten berechnet.

Das qualitative Merkmal A habe die Ausprägungen A_1, ..., A_k.
Nun beobachtet man in einer Stichprobe vom Umfang n die Ausprägung A_i mit der absoluten Häufigkeit n_i (i = 1, 2, ..., k).

$$n_1 + n_2 + \cdots + n_k = n$$

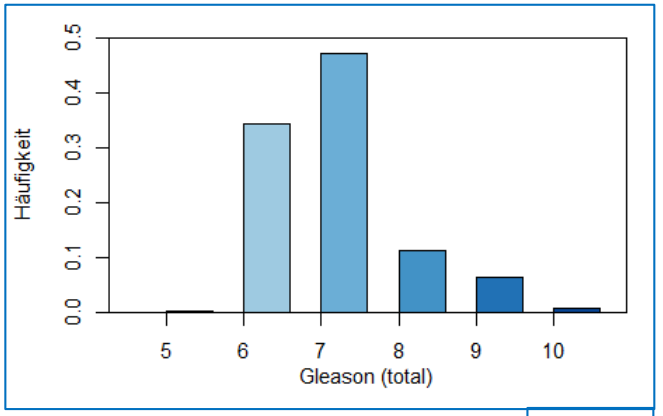
Die relative Häufigkeit ist dann gegeben durch:

$$h_i = \frac{n_i}{n} \quad \text{(i = 1, 2, ..., k)}$$

Meist wird h_i in % angegeben: $h_i \times 100\%$

Relative Häufigkeiten → Wahrscheinlichkeiten

Gleason sum (preop), *n* (%)	
5	1 (0.2)
6	188 (34)
7	258 (47)
8	61 (11)
9	35 (6.4)
10	4 (0.7)



- Ordnet man den Ausprägungen die relativen Häufigkeiten zu, erhält man die empirische Verteilungsdichte

- Relative Häufigkeiten (auf der empirischen Seite) werden auf der theoretischen / mathematischen Seite durch Wahrscheinlichkeiten modelliert.

- Umgekehrt interpretiert man relative Häufigkeiten als Wahrscheinlichkeiten.

Kategoriale Daten III

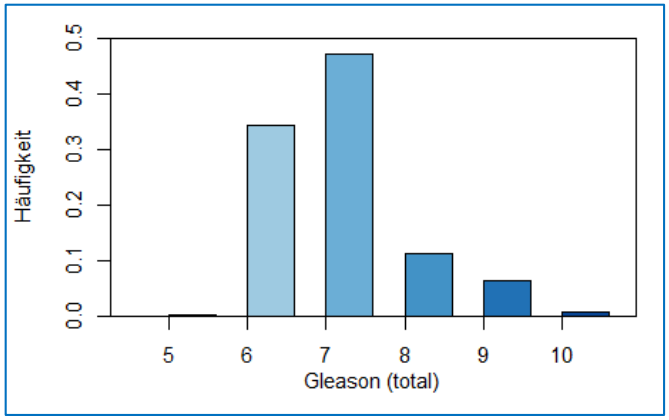
Für quantitative diskrete Merkmale kann man zusätzlich berechnen:

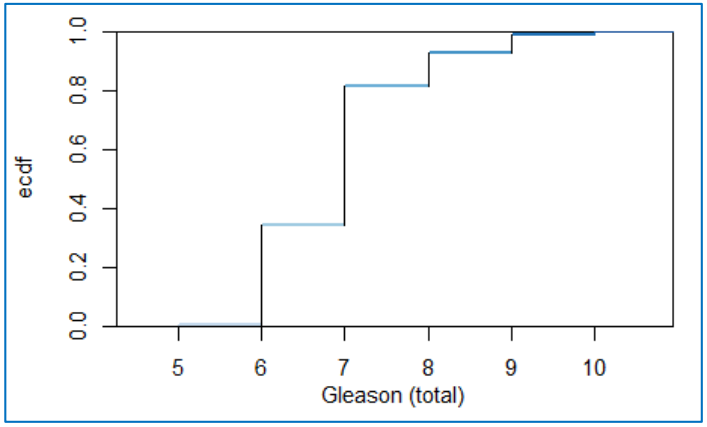
- die absolute Häufigkeitssumme $N_i = \sum_{j=1}^{i} n_j \, , i = 1,2, \ldots, k$
- und die relative Häufigkeitssumme $H_i = \sum_{j=1}^{i} h_j \, , 1,2, \ldots, k$

Sind die Merkmale aufsteigend (von den kleinen zu den großen Werten) geordnet, gibt

- N_i die Anzahl und
- H_i den Anteil der Beobachtungseinheiten mit Ausprägungen $\leq$ der i-ten an (i = 1, 2, …, k), also kumuliert.
- Meist wird H_i in % angegeben: $H_i \times 100\%$

(Empirische) Kumulative Verteilungsfunktion





- Die relativen Häufigkeitssummen stellen die empirische kumulative Verteilungsfunktion (ecdf) dar.
- Sie zeigt für jedes x im Wertebereich an, welcher Anteil der Werte höchstens x ist.
- Sie enthält alle Informationen der Stichprobe.

Kategoriale Daten IV

Frage: Wie viele Probanden im LRP-Arm haben eine Gleason-Summe ≤ 7?

Die absolute Häufigkeitssumme

$$N_3 =$$

bzw.

- die relative Häufigkeitssumme

$$H_3 =$$

Gleason-Summe im LRP-Arm			
	Absolute Häufigkeit	Relative Häufigkeit	Relative Häufig-keitssumme
5	1	0,2%	0,2%
6	188	34%	34%
7	258	47%	82%
8	61	11%	93%
9	35	6,4%	99,3%
10	4	0,7%	100%
n	547	100%	

3. Deskriptive Statistik II

3.2 Kenngrößen für metrische Größen

3.2.1 Nichtparametrisch (Rangverfahren)

3.2.2 Parametrisch

Zwei Ansätze

Parametrisch

- Annahme einer Verteilung (meist Normalverteilung)
- Verteilungsformel mit freien Parametern
- Schätzung der Parameter aus der Stichprobe
  → Kenngrößen

Beispiele:

- Mittelwert
- Standardabweichung

Nichtparametrisch
(rangbasiert)

- Stichprobe sortieren
- Ränge bilden
- Nutzung der Stichproben-perzentile als Kenngröße

Beispiele

- Median
- Quartile (IQR)
- Spannweite

Statistische Kenngrößen

Für quantitative Merkmale kann man
Lagemaße (syn.: Lokalisationsmaße) und Streuungsmaße für Verteilungen angeben.

Die Lagemaße sagen etwas über die Lage der Verteilung, einen Durchschnittswert der Daten aus.

z.B.: arithmetischer Mittelwert, Median

Streuungsmaße sagen etwas über die Abweichung der Messwerte vom Durchschnitt, die Variation, aus.

z.B.: Varianz, Standardabweichung,
Spannweite, Interquartilsabstand

Beispiel 3.2: die Ex-DHF-Studie



nature medicine

Edelmann F et al. Nature Medicine 2025;31:306-314

Article

https://doi.org/10.1038/s41591-024-03342-7

Combined endurance and resistance exercise training in heart failure with preserved ejection fraction: a randomized controlled trial

randomisierte, kontrollierte Studie

- P Patienten mit HFpEF
- I Usual Care + überwachtes Kraft- und Ausdauertraining (ET)
- C Usual Care (UC)
- O 3-stufiger Endpunkt (hierarchisch): verbessert / unverändert / verschlechtert

Tabellarische Darstellung metrischer Größen I

- Für quantitative stetige Merkmale werden k Klassen gebildet, mit den Klassengrenzen $a_0 < a_1 < \cdots < a_k$.
- $a_i - a_{i-1}$ $(i = 1, 2, \ldots, k)$ ist die Breite der i-ten Klasse, Standard: für alle Klassen gleich groß.
- Ist das Merkmal nach links oder rechts unbegrenzt, dann führt man eine Restklasse ein.
- Faustregel zur Wahl der Klassenintervalle:

1. Intervalle gleicher Länge mit Mittelpunkten an geeigneten Werten.
2. kleine Datensätze → kleine Anzahl Intervalle, großen Datensätze → mehr Intervalle

- Die hierzu passende grafische Darstellung ist das Histogramm.

Tabellarische Darstellung II

Aufgabe: Wie viele Probanden sind jünger als 60 Jahre?

Die absolute Häufigkeitssumme

$$N_2 =$$

bzw.

- die relative Häufigkeitssumme

$$H_2 =$$

Alter (J.) (Klassiert)

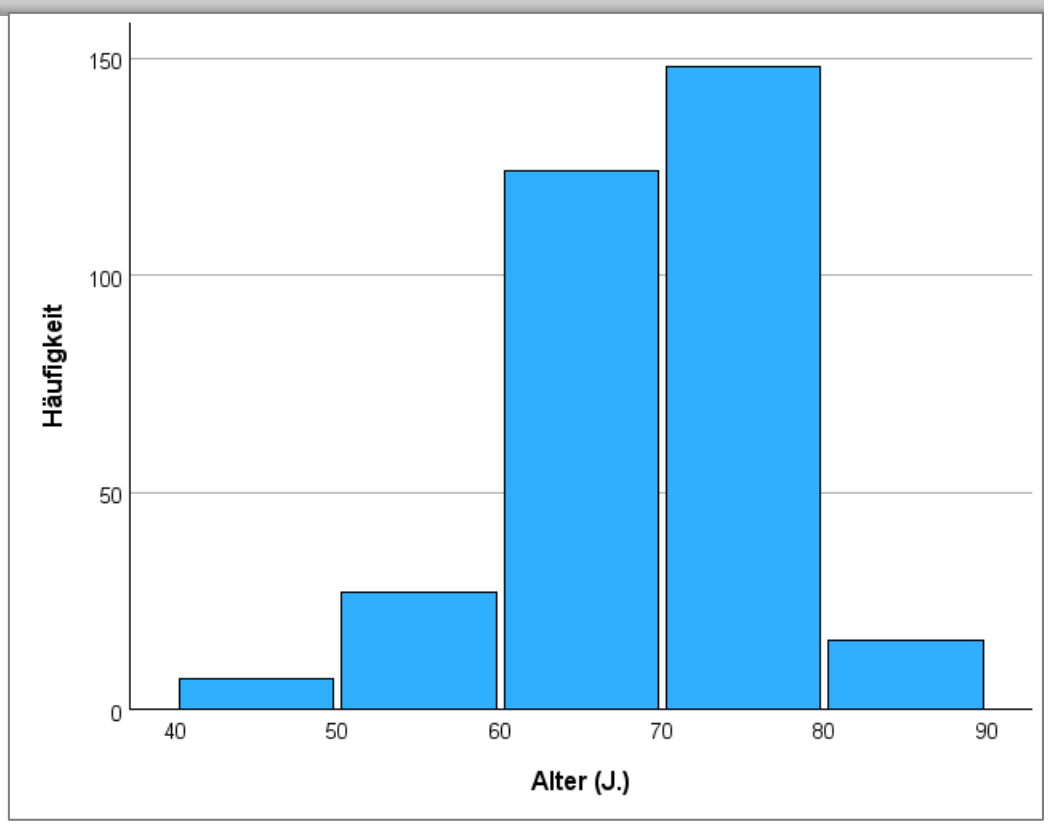
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	40 - 50	7	2,2	2,2	2,2
	50 - 60	27	8,4	8,4	10,6
	60 - 70	124	38,5	38,5	49,1
	70 - 80	148	46,0	46,0	95,0
	80 - 90	16	5,0	5,0	100,0
	Gesamt	322	100,0	100,0	

Histogramm

- Häufung zwischen 60 und 80

Deskriptive Statistik:

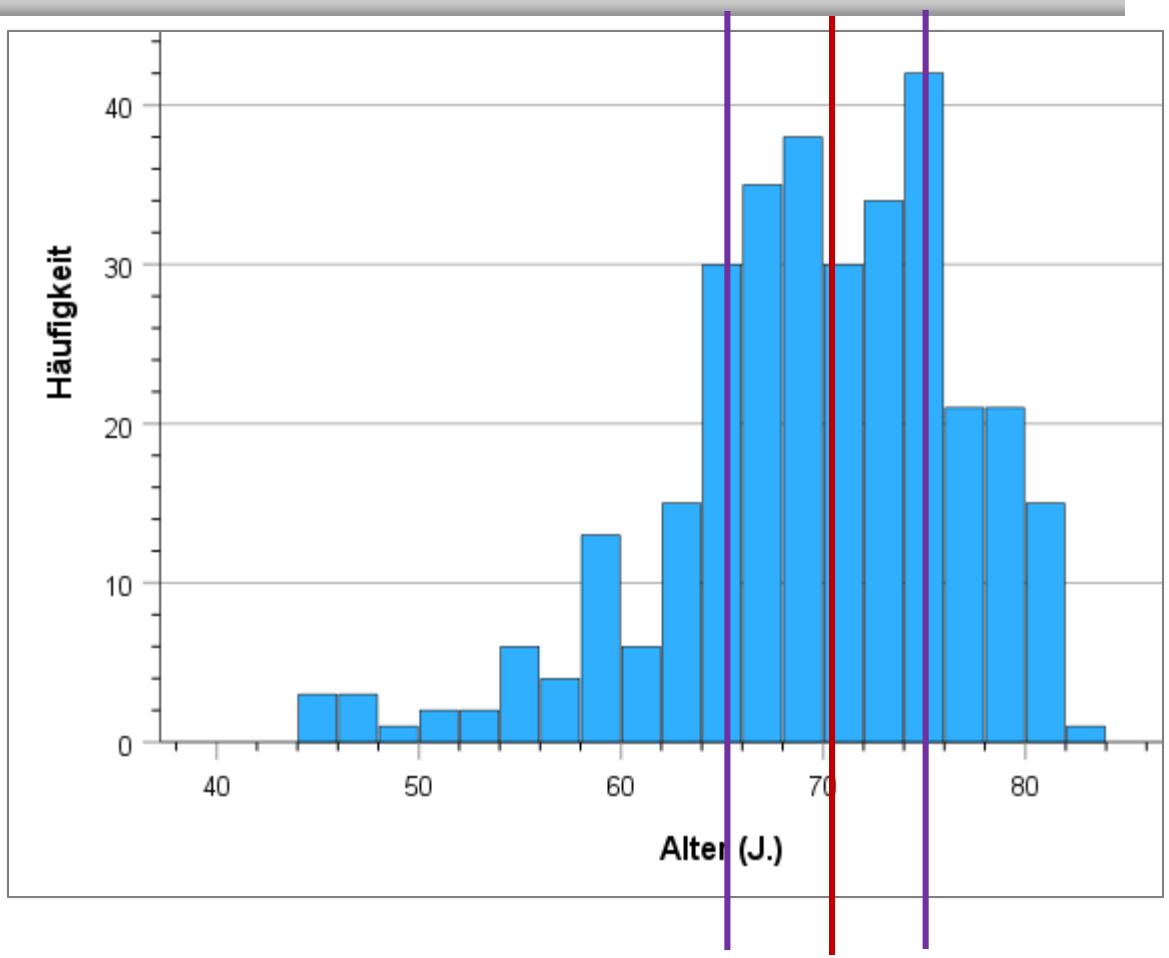
- Wie kann man diese Werte mit ein – zwei typischen Kenngrößen beschreiben?



3.2.1 Rangbasierte Kenngrößen

Idee:

- Man sortiert die realisierten Werte der Größe nach und unterteilt sie in
- 2 gleichmächtige Mengen: Median
- 4 gleichmächtige Mengen: Quartile



Lagemaß I: Median

Median $\tilde{x}$:

Die Daten werden der Größe nach geordnet, d. h. man geht von der Urliste der Daten $x_1, x_2, \ldots, x_n$ zur Rangliste $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ über; (i) heißt Rangzahl

z. B.: 1 Rangzahl des kleinsten Wertes,

n Rangzahl des größten Wertes

n ungerade: $\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$ n gerade: $\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$

Es gibt andere Definitionen, z. B. $\tilde{x} = x_{\left(\frac{n}{2}\right)}$

Median II

- Allgemeiner: für einen Prozentsatz p:
- Der Median ist der Wert in der Mitte der Rangliste $\tilde{x}_{0,50}$.

$$\tilde{x}_p = \begin{cases} x_{(i)} \\ \frac{1}{2}\left(x_{(i)} + x_{(i+1)}\right) \end{cases}$$

falls $n \times p$ eine ganze Zahl ist, gilt:

$$i = n \times p$$

falls $n \times a$ keine ganze Zahl ist, gilt:
i ist die nächst höhere Zahl

z. B.: n = 7: 7 x 0,5 = 3,5 ⇨

n = 8: 8 x 0,5 = 4 ⇨

Median III

Median bei ungerader Anzahl von Beobachtungen

z. B.: Ex-DHF-Studie, NT-proBNP baseline, n = 11

189, 51, 141, 36, 173, 130, 95, 26, 41, 103, 43 ng / ml

Schritt 1: Beobachtungen ordnen:

1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11.

Schritt 2: Mittlere Zahl bestimmen:

Median IV

Median bei gerader Anzahl von Beobachtungen

z. B.: Ex-DHF-Studie, NT-proBNP baseline, n = 10

189, 51, 141, 36, 173, 130, 95, 26, 41, 103 ng / ml

Schritt 1: Beobachtungen ordnen:

1. 2. 3. 4. 5. 6. 7. 8. 9. 10.

Schritt 2: Zwei mittlere Zahlen bestimmen:

Schritt 3: Daraus der Mittelwert

Streuungsmaße I: Quartile

Quartile $\tilde{x}_{0,25}$ und $\tilde{x}_{0,75}$:

- Die gesamte Rangliste wird vier Viertel geteilt. Das erste (untere) Quartil $\tilde{x}_{0,25}$ teilt mittig die untere Hälfte der Rangliste, das dritte (obere) Quartil $\tilde{x}_{0,75}$ halbiert etwa die obere Hälfte der Rangliste.
- Das beim Median (= zweites Quartil) erläuterte Verfahren wird auch zur Berechnung des unteren und oberen Quartils benutzt.

Synonyme Bezeichnungen:

1. Quartil = unteres Quartil = 25%-Quantil = 25%-Perzentil
2. Quartil = mittleres Quartil = 50%-Quantil = 50%-Perzentil = Median
3. Quartil = oberes Quartil = 75%-Quantil = 75%-Perzentil

Quartile II

z. B.: Ex-DHF, NT-proBNP baseline, n = 12

1. Ordnen (Rangplatz zuordnen)

26	36	41	43	51	67	95	103	130	141	173	189
$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$X_{(9)}$	$X_{(10)}$	$X_{(11)}$	$X_{(12)}$

2. Für Q1: $12 \cdot 0{,}25 = 3$ → Mittel aus Wert des 3. und 4. Elementes

$$Q1 =$$

3. Für Q3: $12 \cdot 0{,}75 = 9$ → Mittel aus Wert des 9. und 10. Elementes

$$Q3 =$$

CAVE! Statistische Programme (SAS, SPSS, R, ...)
nutzen verschiedene Berechnungsformeln



Jotquadrat: Grafitto in Bonn. commons. wikipedia.org

Streuungsmaß III: Interquartilsabstand

IQR (interquartile range)

z. B.: Ex-DHF, NT-proBNP baseline, n = 12

1. Ordnen (Rangplatz zuordnen)

26	36	41	43	51	67	95	103	130	141	173	189
$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$X_{(9)}$	$X_{(10)}$	$X_{(11)}$	$X_{(12)}$

2. Intervall [Q1, Q3]:

hier: [42 – 135,5] ng / ml

Streuungsmaß IV: Spannweite

Spannweite (range):

- Differenz zwischen größtem und kleinstem Wert
- $R = x_{(max)} - x_{(min)} = x_{(n)} - x_{(1)}$

z. B.: Ex-DHF, NT-proBNP baseline

189, 51, 141, 36, 173, 130, 95, 26, 41, 103, 43 ng / ml

Spannweite:

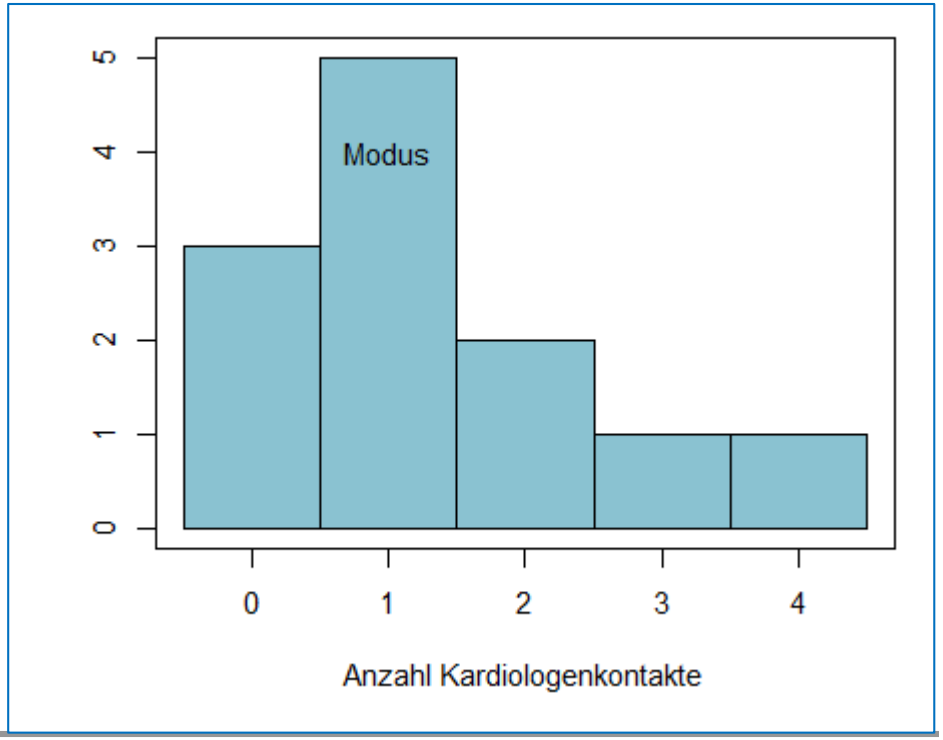
Hinweis: in Literatur oft Angabe der Spannweite als geschlossenes Intervall: z. B. [26, 189] ng / ml

Lagemaß II: Modalwert

Der Modalwert (syn.: Modus) ist der Wert, welcher in der Datenerhebung eines Merkmales am häufigsten vorkommt.

- Beispiel 3.3: Ex-DHF, Anzahl der Kardiologen-kontakte im letzten Jahr:

- Modalwert =



3.2.2 Parametrische Kenngrößen

Wurde ein Merkmal X mit den Beobachtungen x_i $(i = 1, \ldots, n)$ gemessen, so ist der

- arithmetische Mittelwert: $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Beispiel 3.2 DIAST-CHF, NT-proBNP, n = 11

189, 51, 141, 36, 173, 130, 95, 26, 41, 103, 43 ng / ml

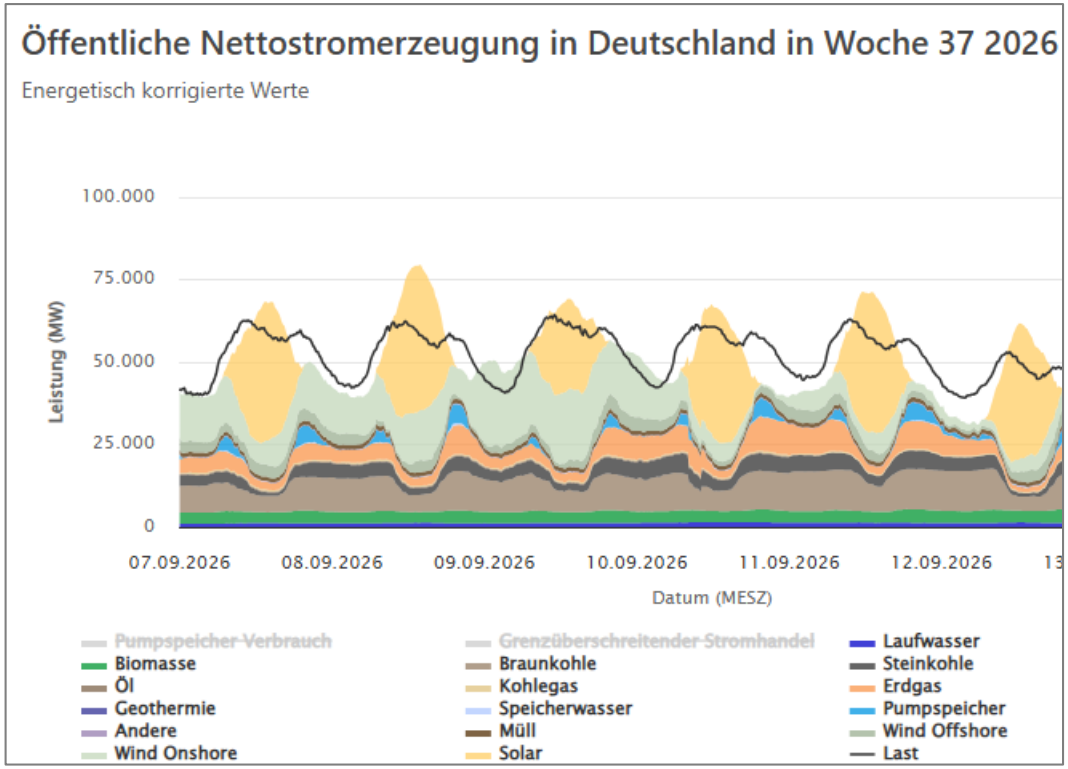
$$\bar{x} = \frac{189+51+141+36+173+130+95+26+41+103+43}{11} = \qquad \frac{ng}{ml}$$



Der Mittelwert genügt nicht!

Im Mittel wird in Deutschland genügend Elektroenergie erzeugt.

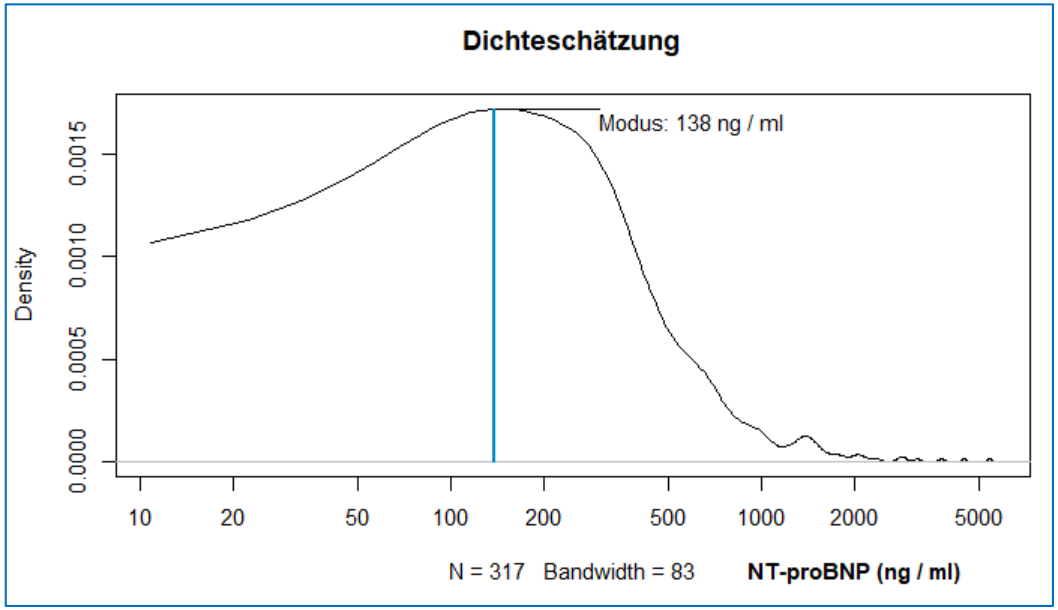
- Genügt das?



Quelle:
https://www.energy-charts.info/charts/power/chart.htm?l=de&c=DE&week=37 (2026-09-15)

Lagemaß III: Modalwert

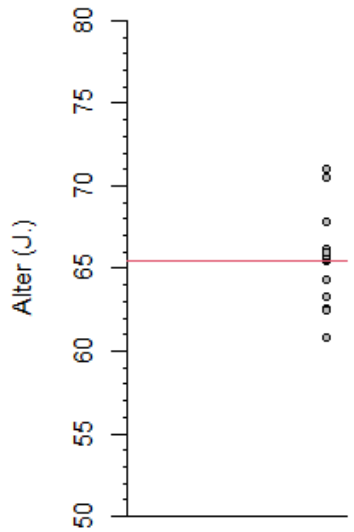
- Bei metrischen Merkmalen ist der Modalwert als das Maximum der Verteilungsdichte definiert.
- Beispiel: Ex-DHF-Studie, NT-proBNP baseline, n = 317



Ein Lagemaß allein genügt nicht, …

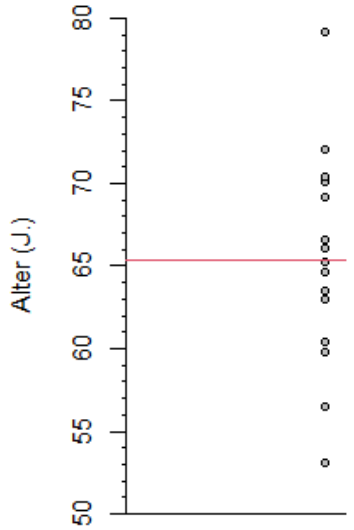
um eine Verteilung zu charakterisieren

Beispiel 1:



Beide Datenreihen haben denselben Mittelwert $\bar{x} = 65$

Beispiel 2:

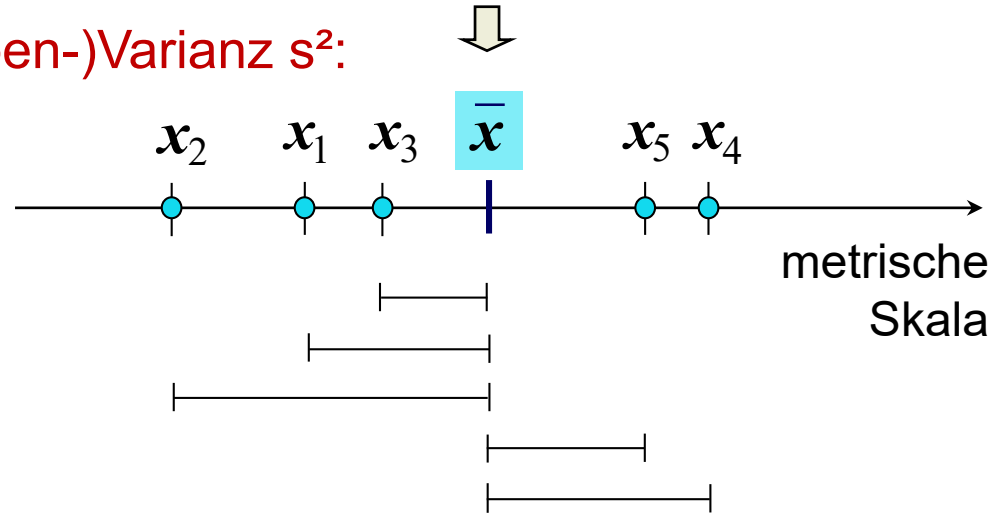


Aber die Datenreihen unterscheiden sich erheblich in der Variation / Streuung:

$s = 2{,}8 \quad s = 6{,}5$

Streuungsmaß II: Varianz

Die (Stichproben-)Varianz s^2:



… misst die mittlere quadrierte Abweichung der Einzelbeobachtungen vom arithmetischen Mittelwert:

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

$$\textit{Varianz (x)} = s^2 = \frac{SQ}{n-1}$$



CAVE: $n - 1$!

Streuungsmaß III: Standardabweichung

Standardabweichung s:

$$s = \sqrt{s^2} = \sqrt{\frac{SQ}{n-1}} = \sqrt{\frac{\sum_{i-1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- Mittleres Abstandsmaß für die Beurteilung der Abweichung der Einzelbeobachtungen vom Mittelwert
- Dimension: (Maßeinheit des Merkmals)

Hinweis: Nutzen Sie die Standardabweichung nur bei annähernd symmetrischen Häufigkeitsverteilungen!

Welches Lagemaß?

Welches Lagemaß drückt das adäquat aus, was ich sagen möchte?

Eine Tageszeitung berichtete, dass der Verband der Haus- und Grundstücksbesitzer die Verhandlungen über den neuen Mietspiegel abgebrochen habe:

Der Mietspiegel orientiert sich an Durchschnittsmieten, wobei Lage, Alter, Größe und Ausstattung berücksichtigt werden. Der Streit der Hauseigentümer mit der Senatsbauverwaltung entzündete sich an der Frage, wie denn durchschnittliche Mieten zu berechnen seien:

- Die Hauseigentümer forderten den arithmetischen Mittelwert,
- die Senatsverwaltung bestand auf dem Median.

Warum?

Mittelwert oder Median

Mittelwert $\overline{x}$

- alle Messwerte gehen in die Berechnung ein
- hat theoretische Vorteile (siehe später)
- ausreißerabhängig
  → nicht robust!

Wenige sehr hohe Mieten treiben den Durchschnitt in die Höhe.

Median $\tilde{x}$

- nur Rangordnung geht ein
- ist einfacher interpretierbar
- das heißt robust gegen wenige Ausreißer

Wenige sehr hohe Mieten beeinflussen den Median nur unwesentlich.

Statistische Kenngrößen – Zusammenfassung

	Lagemaße (measures of location)	Streuungsmaße (measures of dispersion)
Parametrisch	▪ Arithmetisches Mittel - symmetrische Verteilung - Fallzahl moderat bis groß	▪ Standardabweichung
Nichtparametrisch (rangbasiert)	▪ Modus ▪ Median - asymmetrische Verteilung - Ausreißer - kleine Fallzahl	▪ Spannweite ▪ Quartile, Quartilsabstand - für quantitative und ordinale Merkmale

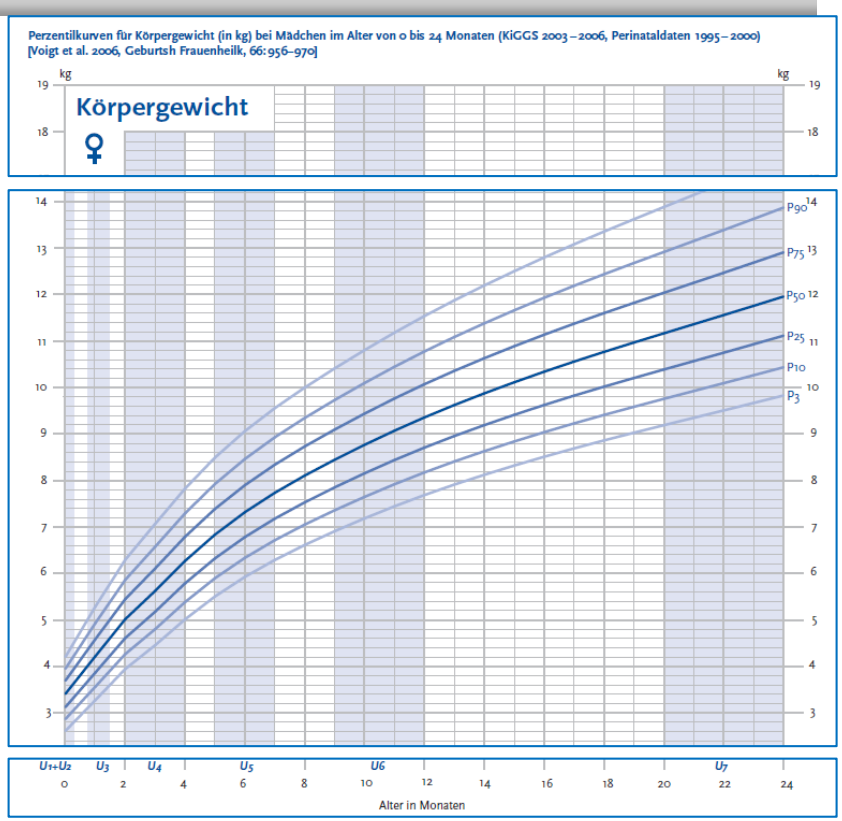
Statistische Kenngrößen – englisch

Englische Bezeichnungen:

Arithmetisches Mittel:	mean
Median:	median
Spannweite:	range
Varianz:	variance
Standardabweichung:	standard deviation (SD)
unteres Quartil:	(first /) lower quartile
oberes Quartil:	(third /) upper quartile
Interquartilsabstand:	interquartile range (IQR)



Der Normpatient?



- Den mittleren Patienten gibt es nicht.
- Vielmehr zeichnen sich Lebewesen durch große individuelle Variabilität aus.

- individuus [unteilbar]
  (Dorsch: Lexikon der Psychologie, https://dorsch.hogrefe.com/stichwort/individuum, 2025-07-23)
- Neuhauser H et al. Referenzperzentile … (KiGGS). Beiträge zur Gesundheitsberichterstattung des Bundes, 2. Auflage 2013

Der Normpatient?

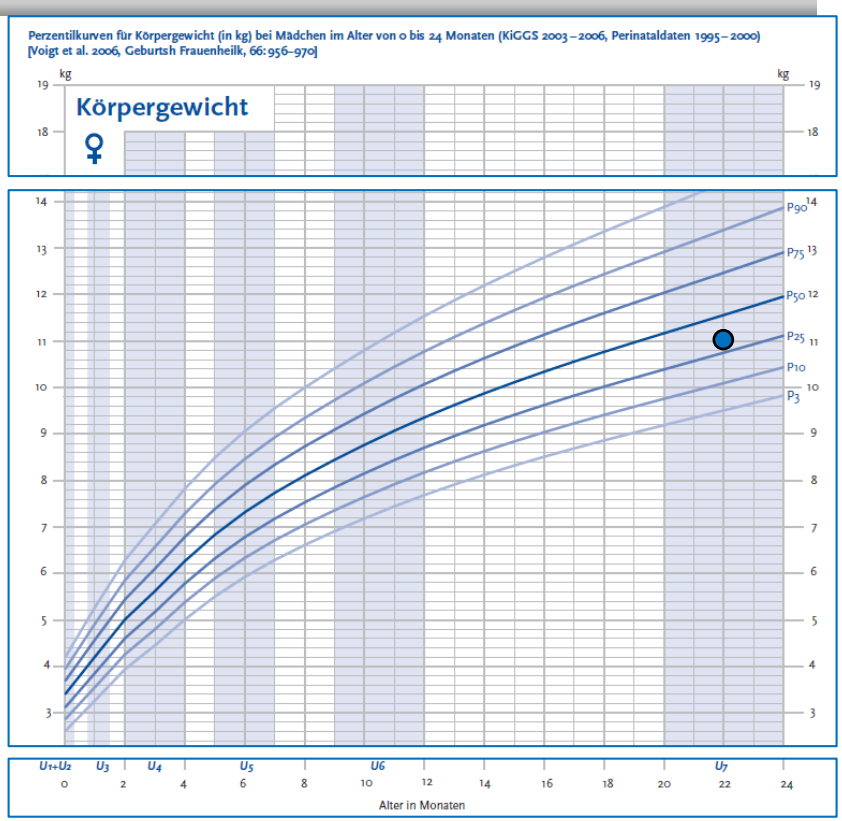
Eine besorgte Kinderärztin zur U7 (22 Monate):

- *Frau Mu, ich mache mir Sorgen, Ihre Tochter ist zu leicht. Sie wiegt nur 11 kg.*

Wie bewerten Sie diese Messung?

-

-



Statistische Kenngrößen – Baselinetabelle

Standard medizinischer Publikationen:

LAP01:

Baseline-Tabelle: Studienkohorte beschreiben.

- Merkmale welchen Typs erkennen Sie?
- Welche statistischen Kenngrößen wurden hier verwendet?

Stolzenburg U et al. European Urology 2021;79:741-750

Table 1 – Baseline characteristics and perioperative data

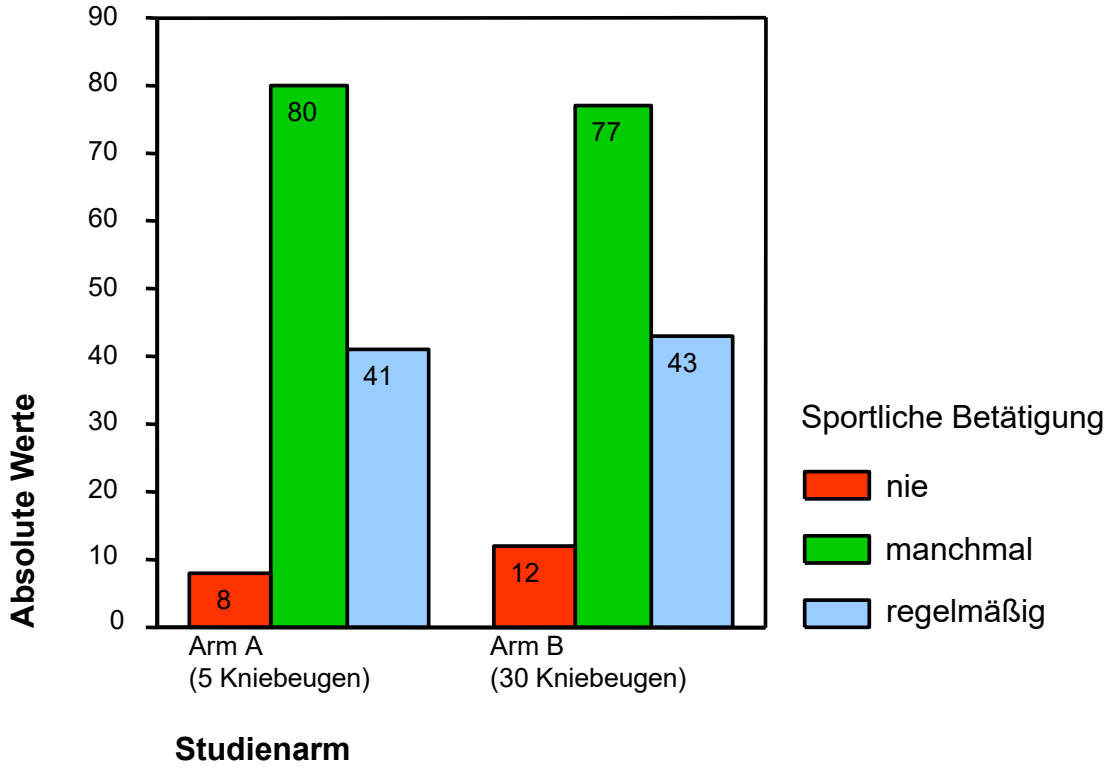
		RARP N = 547
Baseline characteristics (full analysis set)		
Sociodemographic data		
Age at surgery (yr), median (quartiles)		65 (59; 69)
Body size (cm), median (quartiles)		177 (173; 181)
Body weight (kg), median (quartiles)		84 (78; 93)
BMI (kg/m^2), median (quartiles)		27.2 (25.2; 29.4)
Karnofsky Index, median (quartiles)		100 (100, 100)
Family status, number (%)		
Married		444 (81)
Smoker, *n* (%)		86 (16)
Urinary track medical history, *n* (%)		
History of urinary tract infection		14 (2.6)
Transurethral resection of bladder cancer		2 (0.4)
Transurethral resection of the prostate		13 (2.4)
Other interventions on the urinary tract		123 (23)
Primary disease: carcinoma of the prostate		
Diagnosis since (mo), median (IQR)		2.1 (1.5, 3.0)
PSA, preoperative (ng/ml), median (IQR)		7.7 (5.6, 12.1)
Gleason sum (preop), *n* (%)		
5	1 (0.2)	
6	188 (34)	
7	258 (47)	
8	61 (11)	
9	35 (6.4)	
10	4 (0.7)	
Planned nerve sparing, *n* (%)		
None	216 (39)	
Unilateral	80 (15)	
Bilateral	251 (46)	
Diabetes mellitus	82 (15)	
Renal failure, *n* (%)	86 (16)	

3. Deskriptive Statistik II

3.3 Repräsentative Grafik

Balkendiagramm

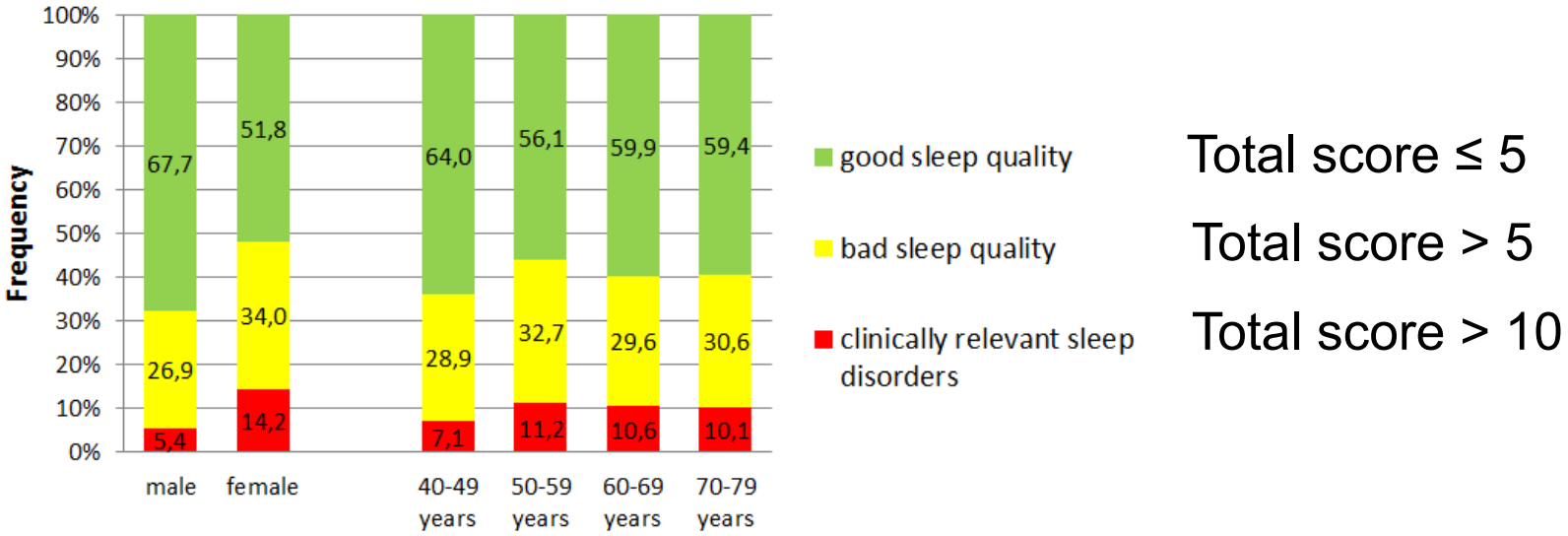
- Balkendiagramm (gruppiert)



geeignet für qualitative Merkmale, aber auch für Mittelwerte

Balkendiagramm II

- LIFE-Kohorte: Pittsburgh Sleep Quality Index



Total score ≤ 5

Total score > 5

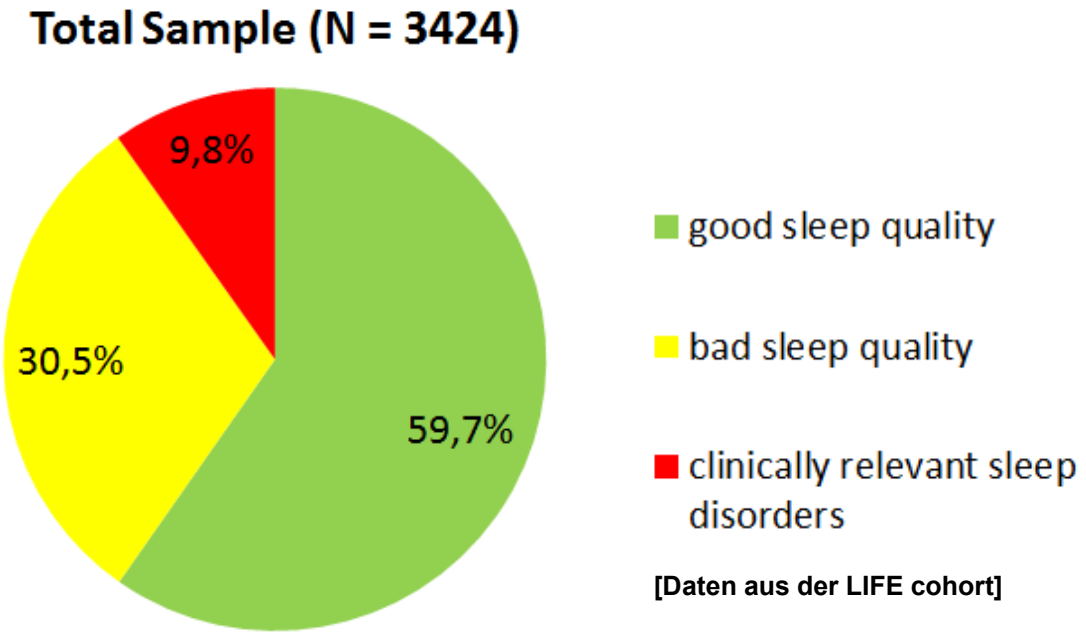
Total score > 10

- Subjective sleep quality (PSQI): Total score = Sum of 7 component-scores (range 0-21 points)

geeignet für kategoriale / ordinale Daten

Kreisdiagramm

Kreisdiagramm



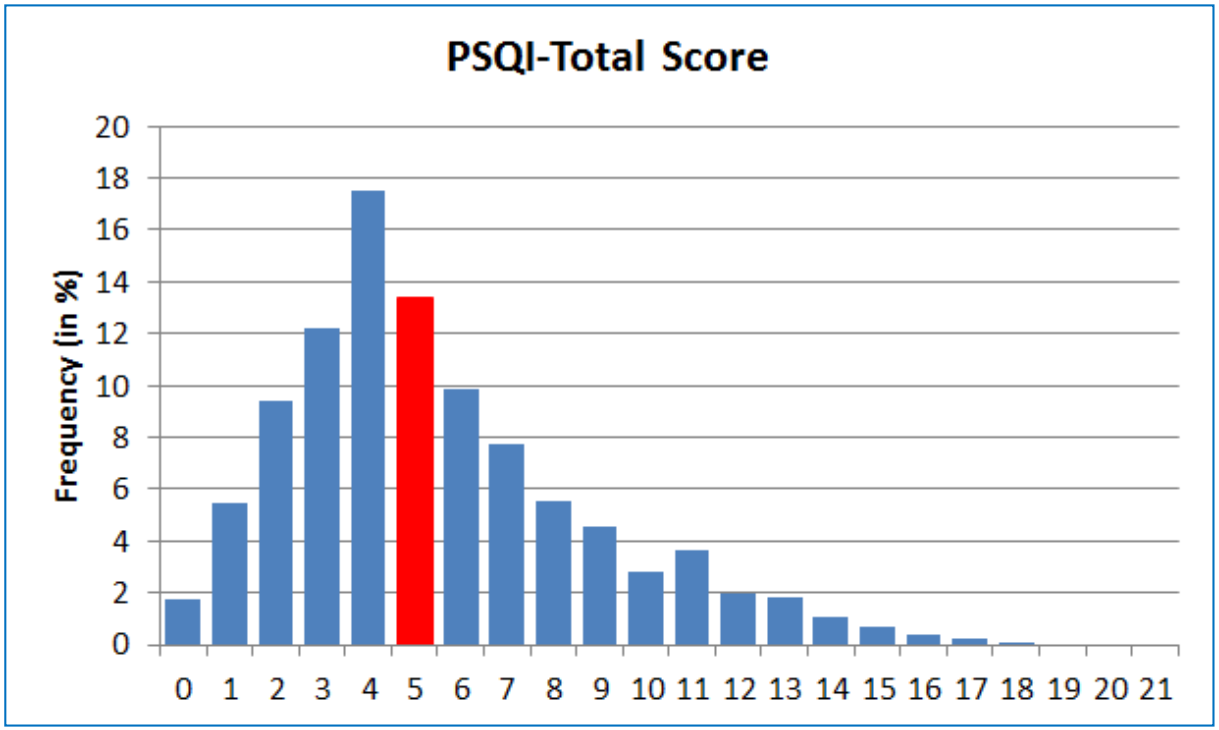
[Daten aus der LIFE cohort]

⇨ geeignet für qualitative Merkmale

CAVE: Manche Journals lehnen Kreisdiagramme ab!

Balkendiagramm / Histogramm

- Säulen- / Balkendiagramm (Pittsburgh Sleep Quality Index)



- geeignet für quantitativ diskrete (auch metrische) Merkmale

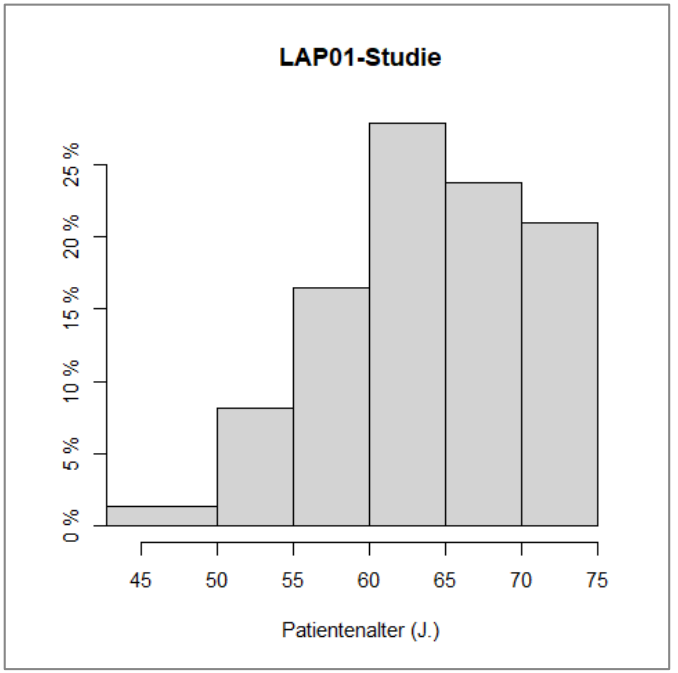
Histogramm II

- Histogramm Probandenalter in der LAP01-Studie
- Zusammenfassung der Beobachtungen in Klassen (vgl. Folie 17):

Klassen-intervall	Klassen-mitte	absolute Häufigkeit	[%]
unter 50	47,5	17	2,2
50 – 54	52	55	7,0
55 – 59	57	119	15,2
60 – 64	62	241	30,8
65 – 69	67	186	23,8
70 – 74	72	164	21,0

- geeignet für quantitativ stetige Merkmale

Häufigkeiten pro Klasse:



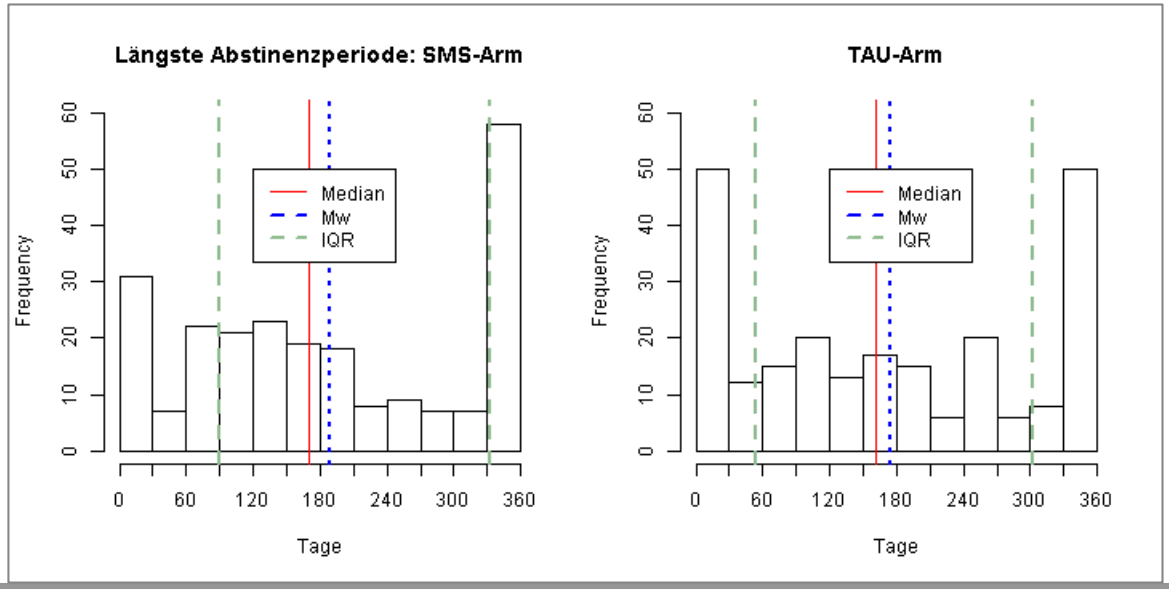
Beispiel Histogramm: CAPS-Studie

Lucht M et al. Addiction 2021;116(6):1431-1442

- Kohorte: Alkoholabhängige nach Entgiftung
- Randomisation 1:1 zu den Armen
  - Treatment as usual (TAU)
  - SMS-Intervention

Längste Abstinenzperiode:

Weil sich Werte an den Rändern häufen, sind weder Mittelwert noch Median repräsentativ.



Welcher Studienarm erscheint besser?

Fehlerbalkenplot I

Fehlerbalken (error bar)

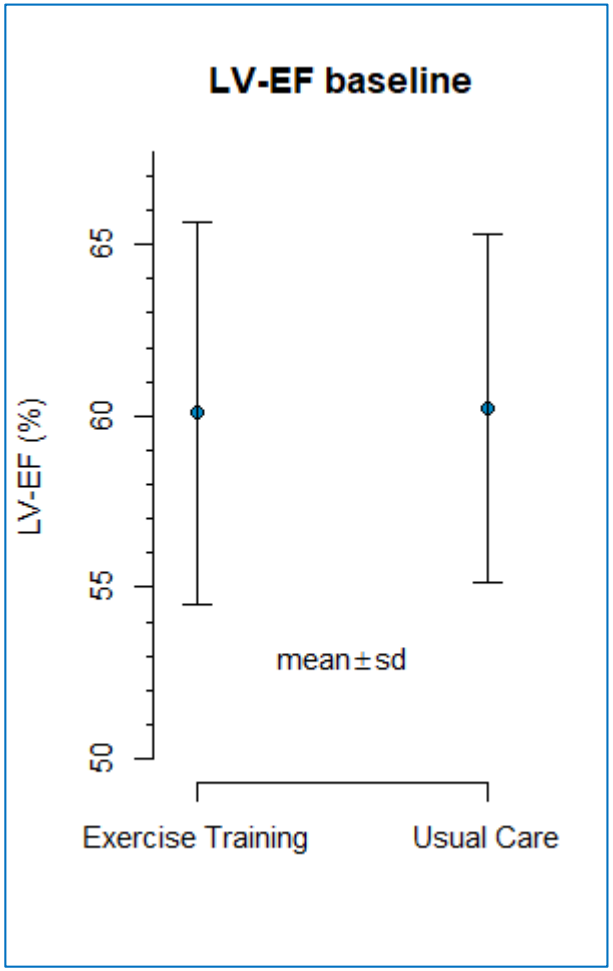
- Deskriptive Statistik:

  Darstellung von: $\bar{x} \pm s$

- geeignet für metrische Merkmale

Beispiel: Ex-DHF-Studie,
LV-EF baseline

- Hintergrund: Bei normalverteilten Merkmalen enthält der Bereich $[\bar{x} - s, \bar{x} + s]$ reichlich 2/3 aller Werte.
  (→ Vorlesung 5)



Fehlerbalkenplot II

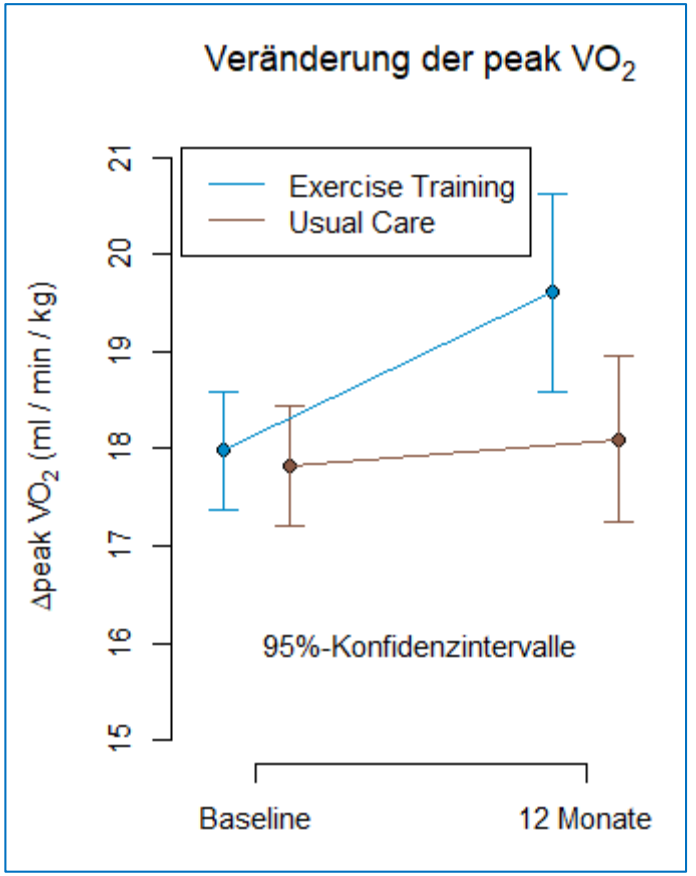
Fehlerbalken (error bar)

- Darstellung von Studienergebnissen

  dann $\Delta\bar{x} \pm S.E.M.$ oder
  95%-Konfidenzintervalle
- geeignet für metrische Merkmale

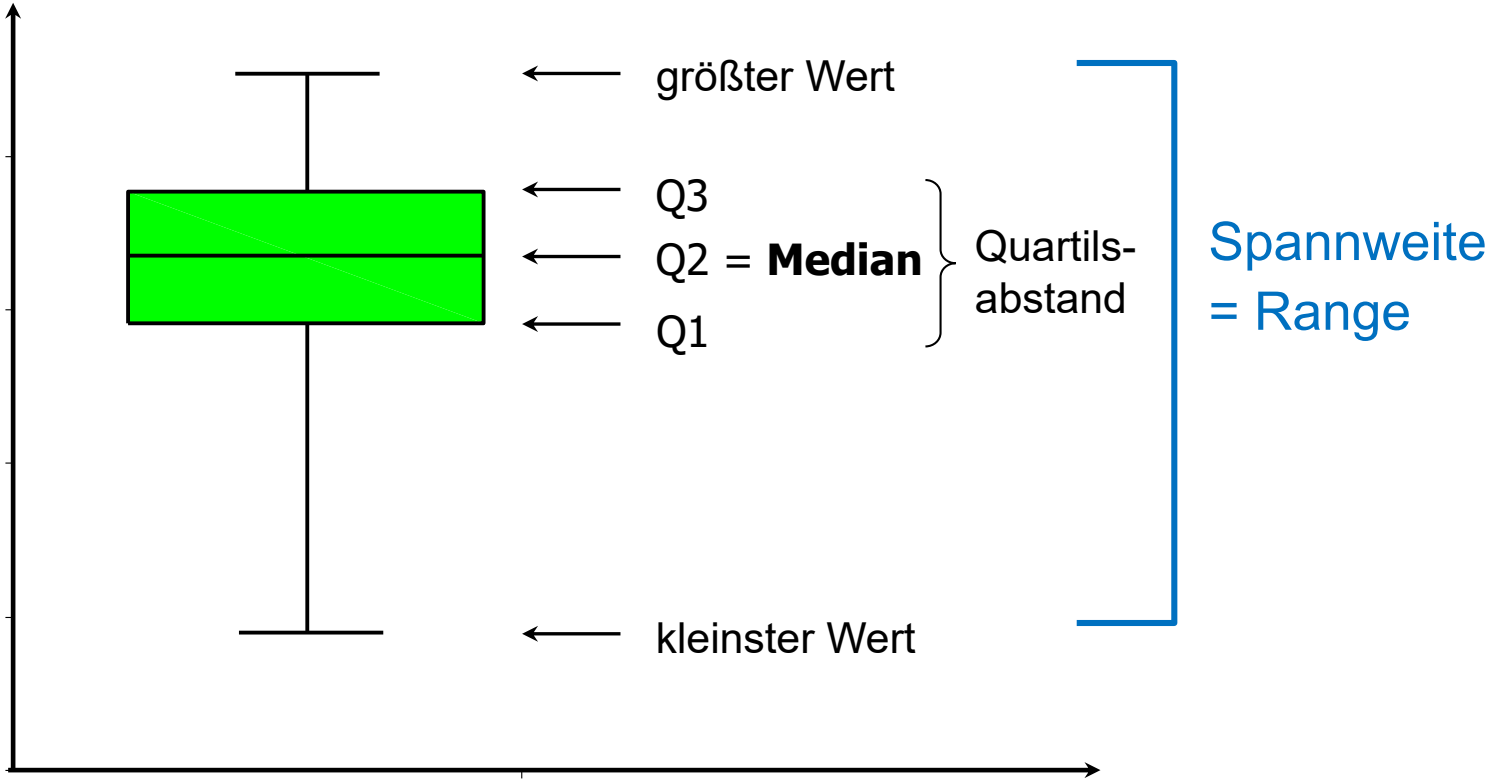
Beispiel: Ex-DHF-Studie,
Veränderung der peak VO_2

- Hinweis: Die Verbindungslinien sind streng genommen nicht korrekt, helfen aber, den Trend zu erkennen.



Boxplot I

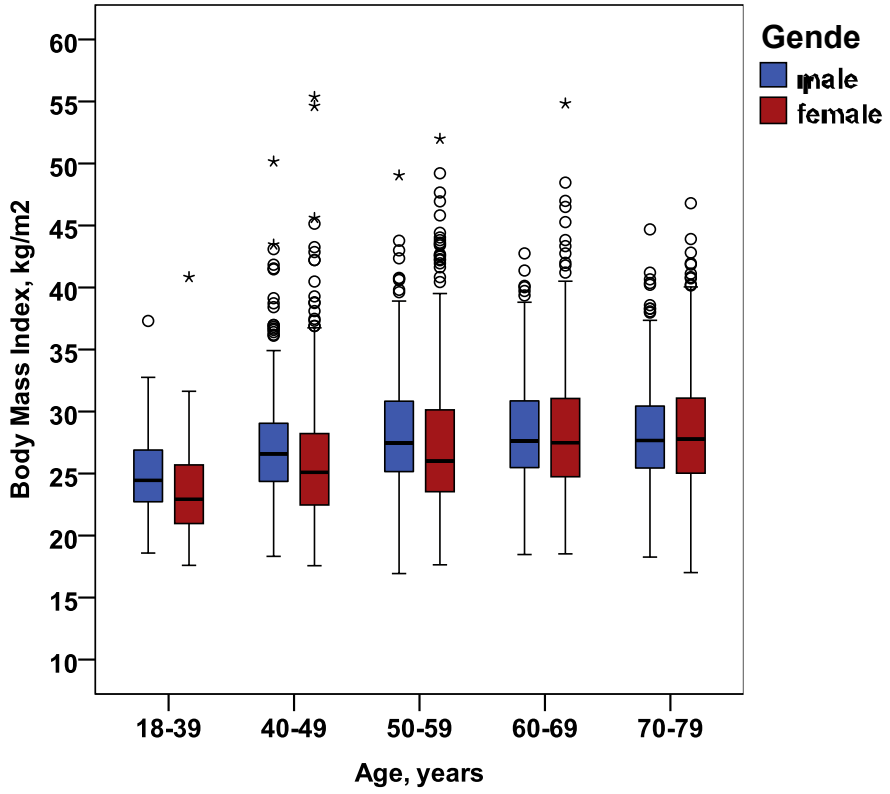
– Box-Plot / Box- and Whisker(s)-Plot



– geeignet insbesondere für quantitative Merkmale

Boxplot II

Box-Plot (SPSS)



Bemerkung: SPSS markiert Ausreißer (o) und Extremwerte (*)

Daten aus der LIFE cohort (n = 4194)

Ausreißer:

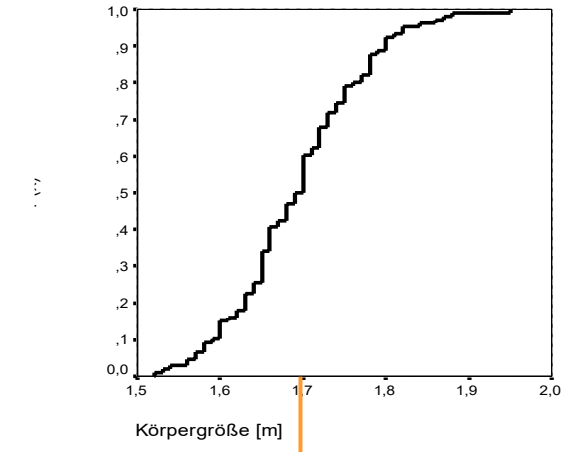
> 1.5-facher Abstand des entsprechenden Quartils vom Median

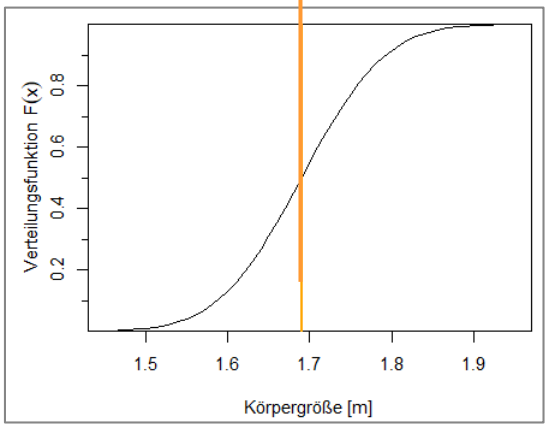
Extremwert:

> 3-fache Abstand des entspr. Quartils vom Median

Whiskers an der Box: zeigen unterhalb der Box den kleinsten Wert und oberhalb der Box den größten Wert an, der noch kein Ausreißer ist

Empirische und theoretische Verteilungsfunktion





- Kumuliert man von links die Häufigkeiten, erhält man die empirische Verteilungsfunktion.
- Für jedes x kann man aus der Treppenfunktion ablesen, welcher Anteil der Werte der Stichprobe ≤ x ist.
- Auf der Theorie- (Modell-)ebene entspricht ihr die Verteilungsfunktion, bei metrischen Zufallsvariablen eine Kurve.

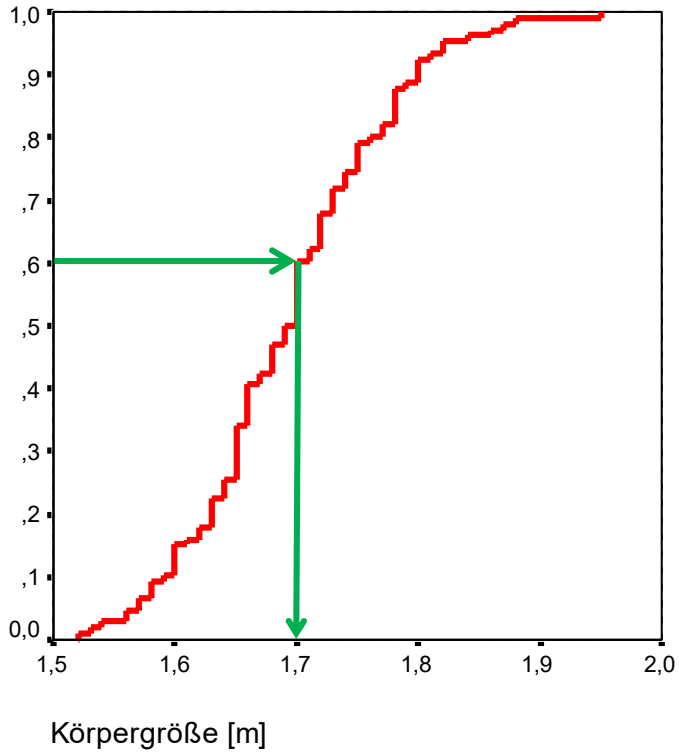
Empirische Verteilungsfunktion - Beispiel

- geeignet für quantitativ stetige Merkmale

Körpergröße [m]

	Häufigkeit	Prozent	Kumulierte Prozente
1,52	1	,9	,9
1,53	1	,9	1,9
1,54	1	,9	2,8
1,56	2	1,9	4,7
1,57	2	1,9	6,6
1,58	3	2,8	9,4
1,59	1	,9	10,4
1,60	5	4,7	15,1
1,61	1	,9	16,0
1,62	2	1,9	17,9
1,63	5	4,7	22,6
1,64	3	2,8	25,5
1,65	9	8,5	34,0
1,66	7	6,6	40,6
1,67	2	1,9	42,5
1,68	5	4,7	47,2
1,69	3	2,8	50,0
1,70	11	10,4	60,4
1,71	2	1,9	62,3
1,72	6	5,7	67,9
1,73	4	3,8	71,7
1,74	3	2,8	74,5
1,75	5	4,7	79,2
....	....	....	....

60%-Quantil



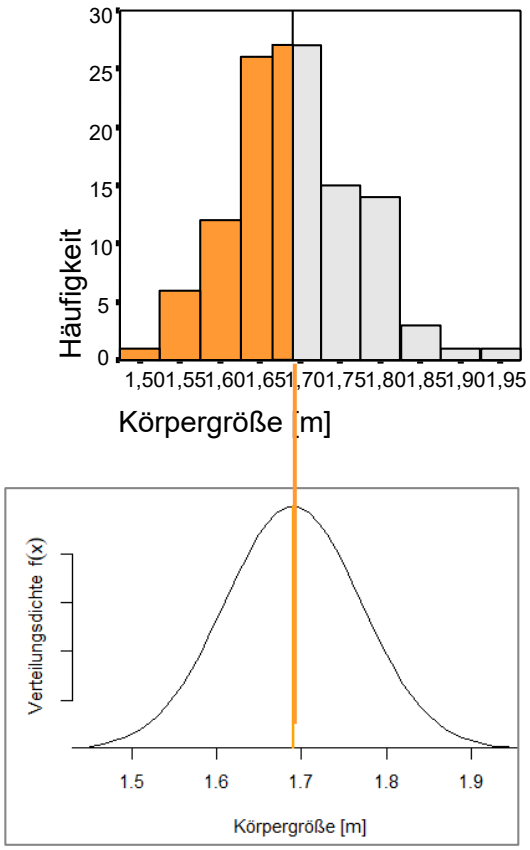
Empirische Verteilungsfunktion II

Empirische Verteilungsfunktion F_n :

Falls die Daten voneinander verschieden sind, gibt F_n für jedes x auf der Zahlengeraden an, wie groß der Anteil der Daten ist, die kleiner oder gleich x sind:

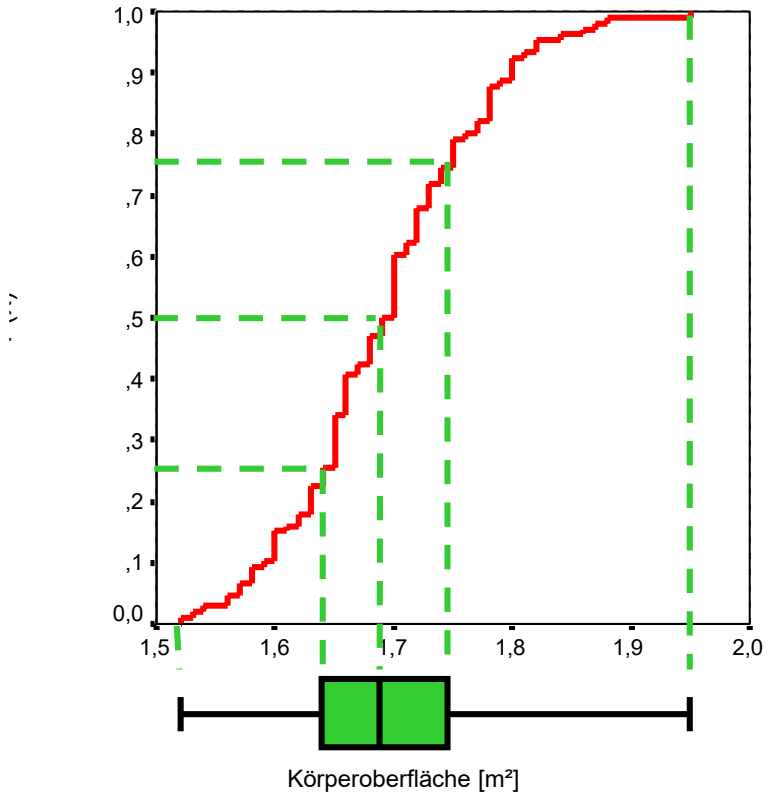
- $F_n(x_{(1)}) = \frac{1}{n}$ $\quad F_n(x_{(n)}) = \frac{n}{n} = 1$ $\quad F_n(x) = 0 \quad f\ddot{u}r \quad x < x_{(1)}$
- Der Median ist definiert als kleinster Wert $\tilde{x}$ für den gilt: $F_n(\tilde{x}) \geq 0{,}5$.
- Das p-Quantil (p-Perzentil, $0 < p < 1$) ist der kleinste Wert x_p, für den gilt: $F_n(x_p) \geq p$

Histogramm und Verteilungsdichte



- Das Histogramm aus (empirisch gewonnenen) Daten repräsentiert
- (auf der theoretischen Ebene) die Darstellung der Verteilungsdichte.

Box-Plot und empirische Verteilungsfunktion



Körpergröße (m)

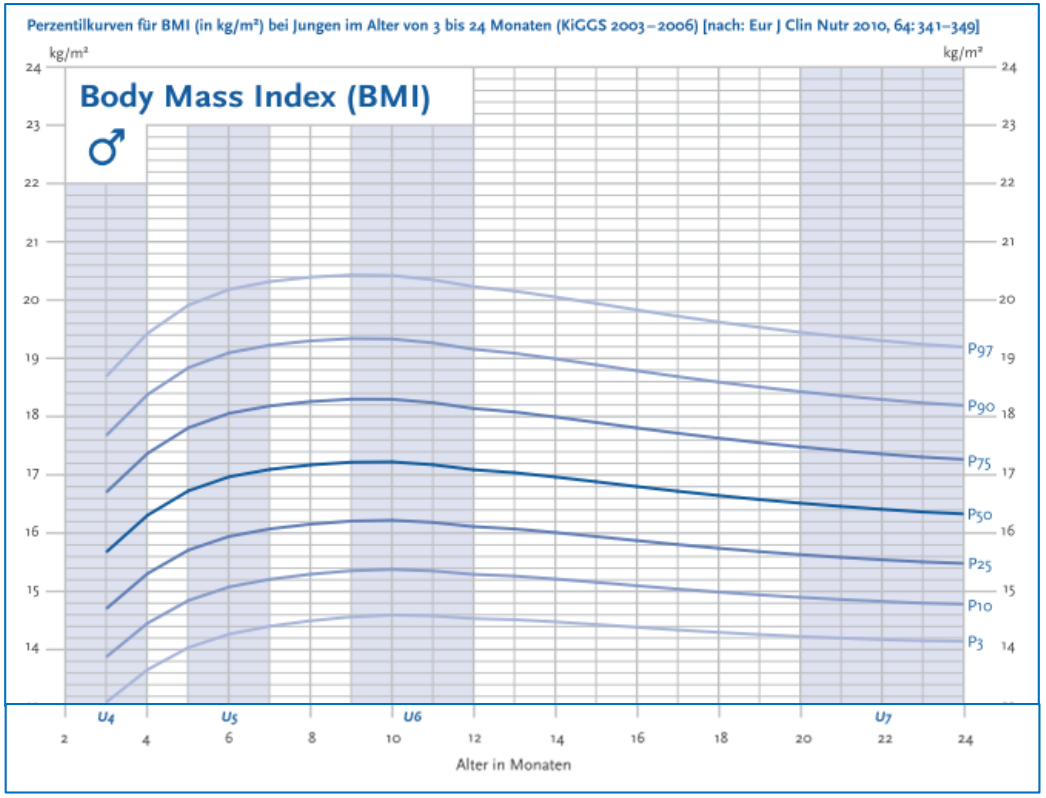
Minimum:	1,52
Maximum:	1,95
Q1:	1,64
Q2 (Median):	1,69
Q3:	1,75

Die vereinfachte Darstellung des Boxplots zeigt keine Ausreißer und Extremwerte.

Die Enden der Whiskers sind Minimum und Maximum.

Grafische Darstellung X

- Somatogramm (Perzentilkurven)



Neuhauser H et al. Referenzperzentile … (KiGGS). Beiträge zur Gesundheitsberichterstattung des Bundes, 2. Auflage 2013

- geeignet zur Darstellung von Normbereichen



UNIVERSITÄT LEIPZIG

Medizinische Fakultät

Vielen Dank!

Dr. rer. nat. Meinhard Mende

Senior-Wissenschaftler, Institut für Medizinische Informatik, Statistik und Epidemiologie und Zentrum für Klinische Studien

Härtelstraße 16-18, 04107 Leipzig

Meinhard.mende@zks.uni-leipzig.de