Vorlesung Statistik – WiSe 2016/17

Teil 1
Einführung und beschreibende Statistik
3.11.2016

Prof. Dr. Peter Martus
UKT Tübingen

Studiengänge
Molekulare Medizin FS 3(BSc)
Biochemie FS 3(BSc)
Medizininformatik FS 3 (BSc)

Vorlesung 1

- Einführende Beispiele
- Graphiktypen
- Statistische Parameter
- Codierung, Skalen und Werteverteilungen
- Normalverteilung und Logarithmustransformation
- Grundsätzliches zur statistischen Denkweise

Teil 1
Einführende Beispiele

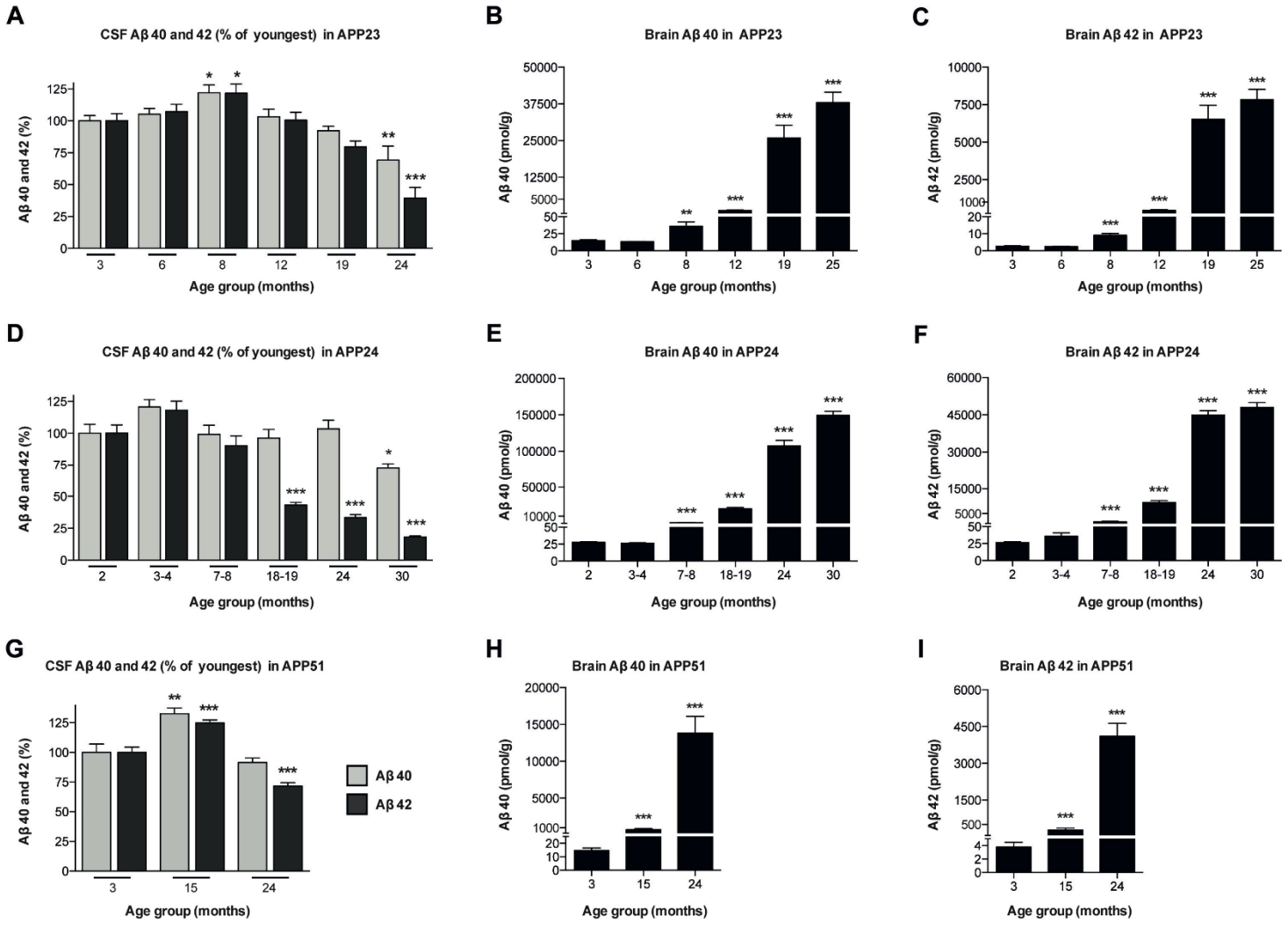
Beispiele für die Anwendung statistischer Methoden in den Lebenswissenschaften

- Laborexperiment
- Diagnosestudie
- Therapiestudie
- Epidemiologische Studie
- Gesundheitsberichtserstattung

Laborexperiment
(Maia et al. EMBO Mol Med 2015; 7:895-903)

- Rolle des Amyloid β Peptid in der Frühphase der Alzheimer Erkrankung im Mausmodell
- Messungen im Liquor und in Gehirnzellen
- Zeitverlauf von Amyloid β Peptid in unterschiedlichen Gruppen von transgenen Mäusen, Bewertung zeitlicher Trends im Gruppenvergleich
- Leitung: M.Jucker, DZNE Tübingen

Laborexperiment
(Maia et al. EMBO Mol Med 2015; 7:895-903)

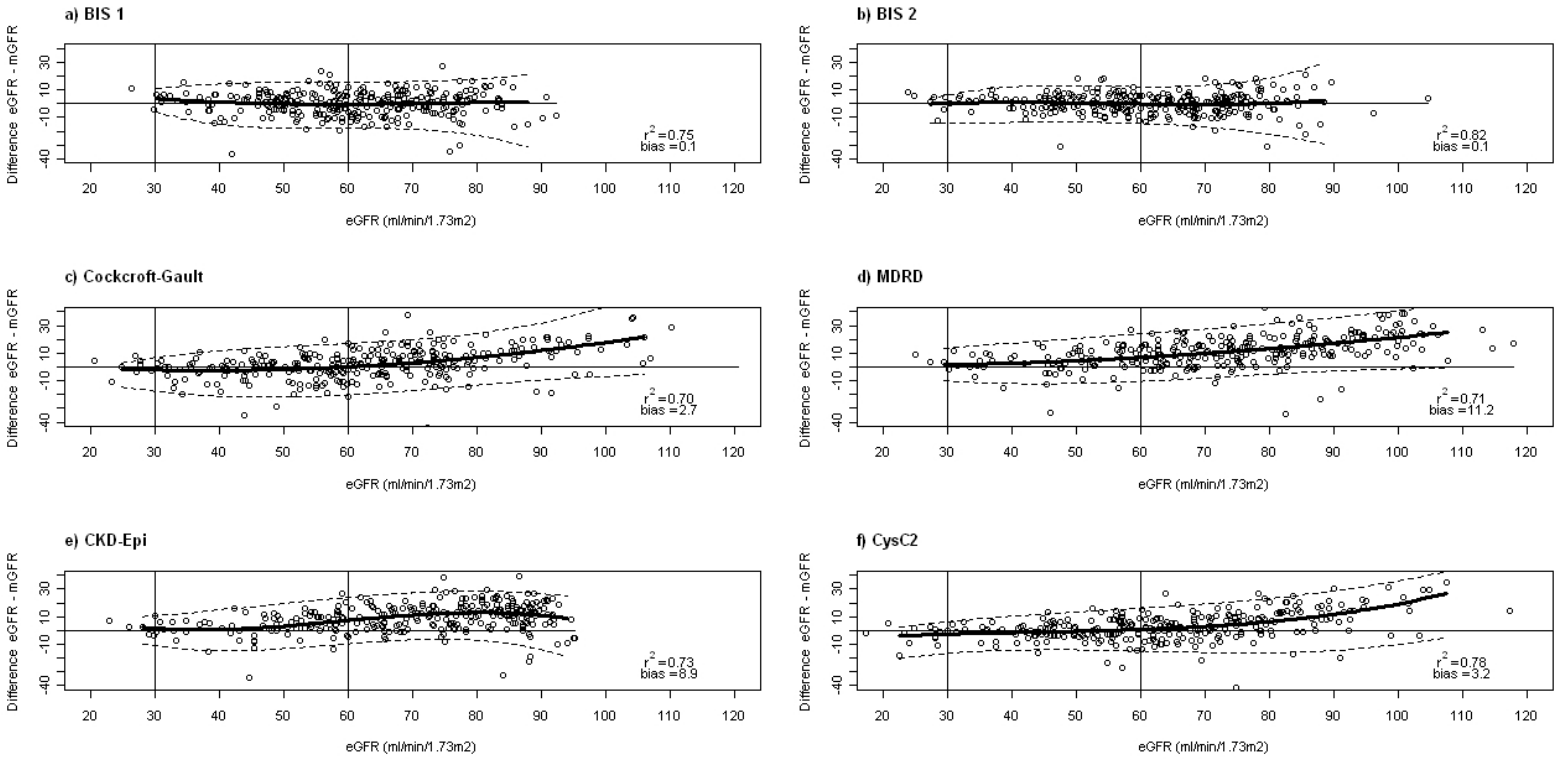


Diagnostische Studie (Schäffner et al., Ann Intern Med. 2012; 157:471-81)

- Verbesserung von Formeln zur Schätzung der Nierenfunktion aus Blutwerten für ältere Menschen
- Studien in der gesunden Normalbevölkerung und an ausgewählten Patientengruppen
- Hier: 600 Freiwillige mit Messung der Nierenfunktion anhand der Eliminationskurve einer Markersubstanz (Iohexol) sowie Blutentnahmen (Kreatinin, Cystatin C).
- Leitung: E.Schäffner, Nephrologie und Public Health, Charité Berlin

Diagnostische Studie (Schäffner et al., Ann Intern Med. 2012; 157:471-81)

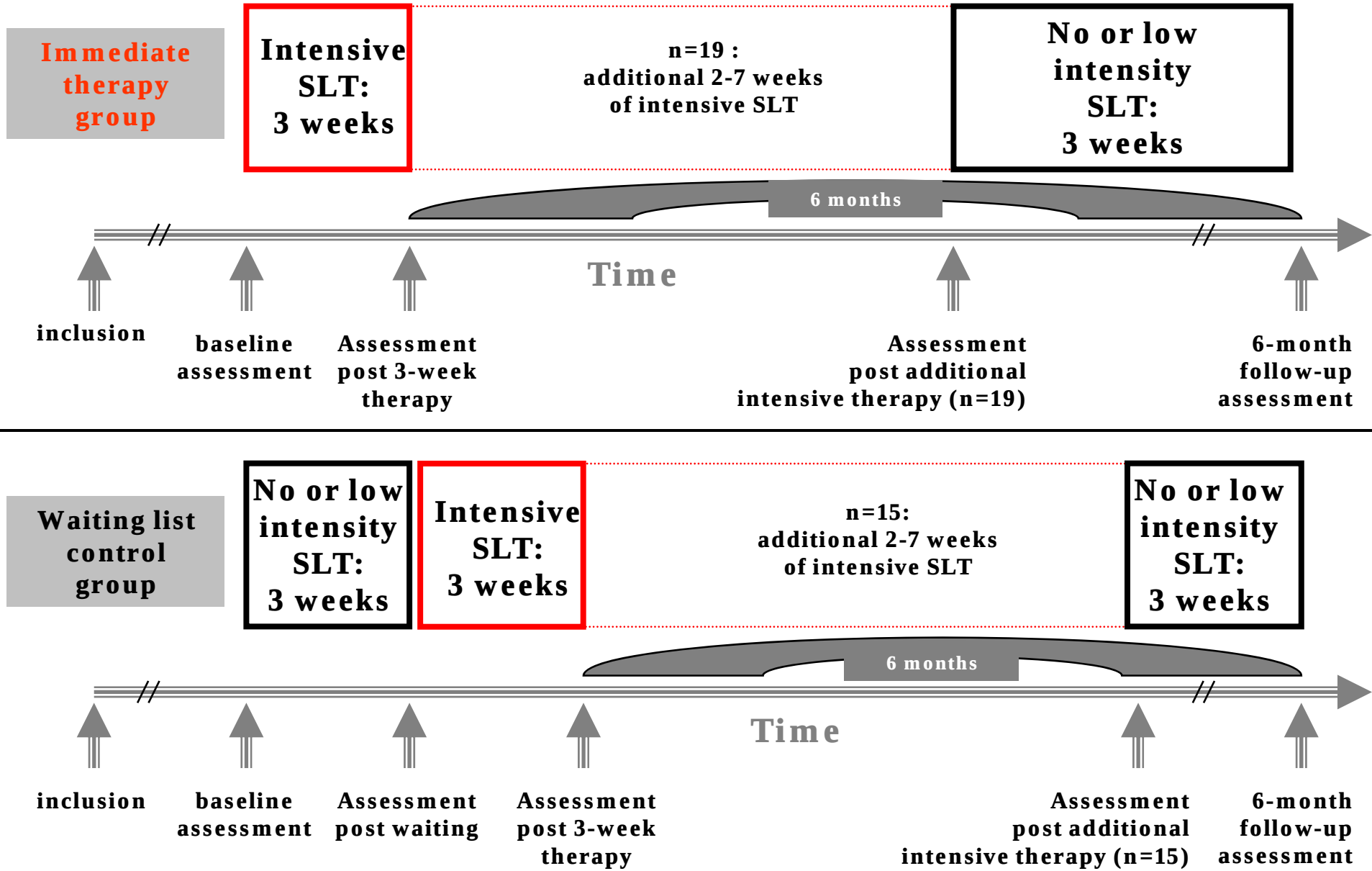
Figure 3. Performance of GFR estimating equations in the validation sample



Therapiestudie
(Breitenstein et al., Lancet, 2016 oder 2017; akzeptiert)

- Wirksamkeit eines intensivierten Sprachtrainings zur Verbesserung der Alltagssprechfähigkeit für Patienten nach Schlaganfall mit chronischer Aphasie
- 156 Patienten, Zufallszuteilung zur intensivierten Therapie oder einer Wartegruppe, die zunächst Standardtherapie erhielt.
- Ziel: Verbesserung der Alltagssprechfähigkeit in während eines Patienteninterviews simulierten Alltagssituationen
- Leitung: C. Breitenstein, Neurologie, Univ. Münster

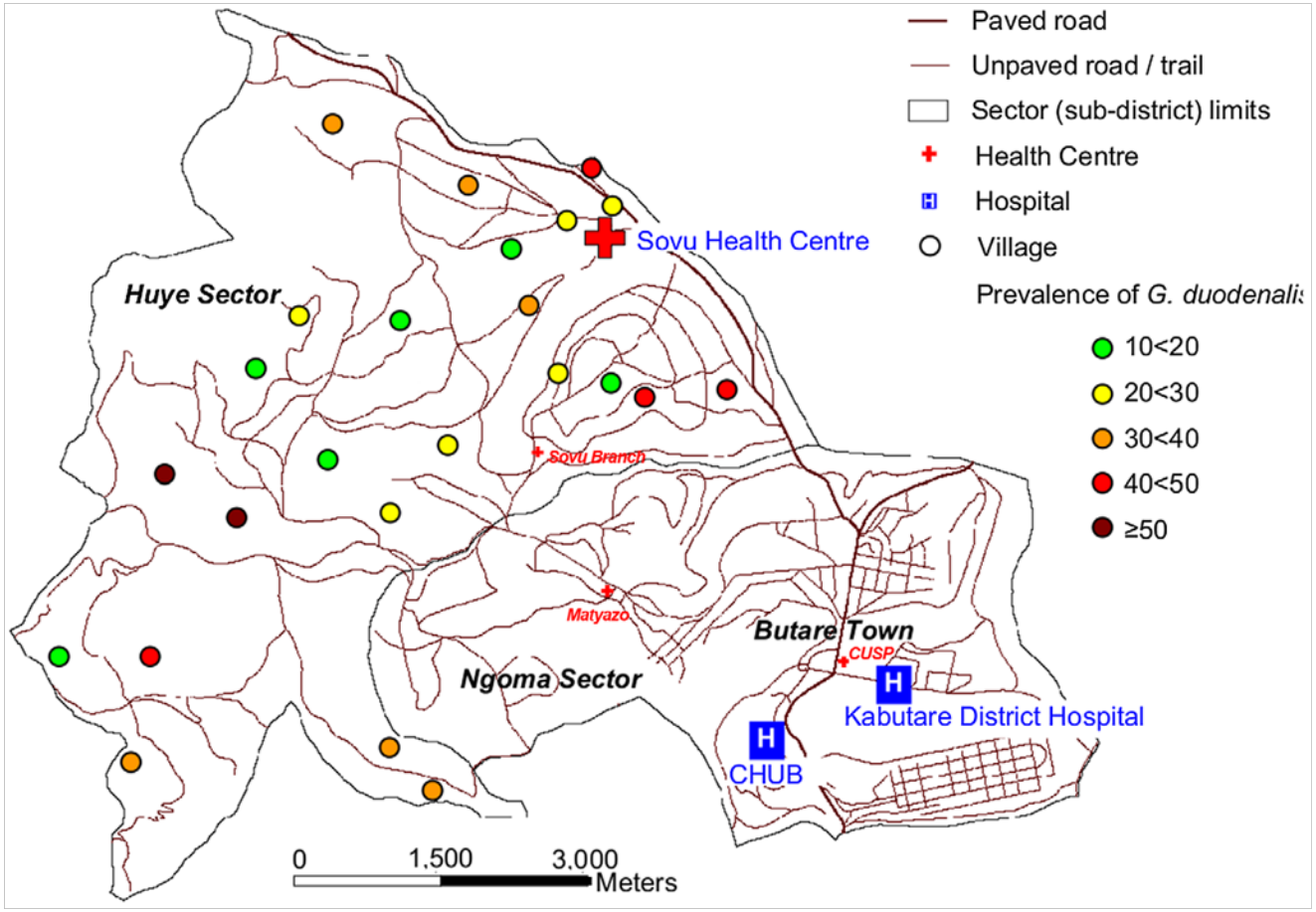
Therapiestudie
(Breitenstein et al., Lancet, 2016 oder 2017; akzeptiert)



Epidemiologische Studie
(Ignatius et al., PLoS Negl Trop Dis. 2012; 6:e1677)

- Giardia Duodenalis Infektion bei Kindern mit Alter unter 5 Jahre in Ruanda
- Stichprobe von 583 Schulkindern, Infektionsstatus, Alter, Geschlecht, Risikofaktoren
- Hypothese: Die Infektion ist ein Risikofaktor für Untergewicht, weitere Fragestellungen: Assoziation Sozialstatus Infektion
- Leitung: F.P. Mockenhaupt, Tropenmedizin, Charité Berlin

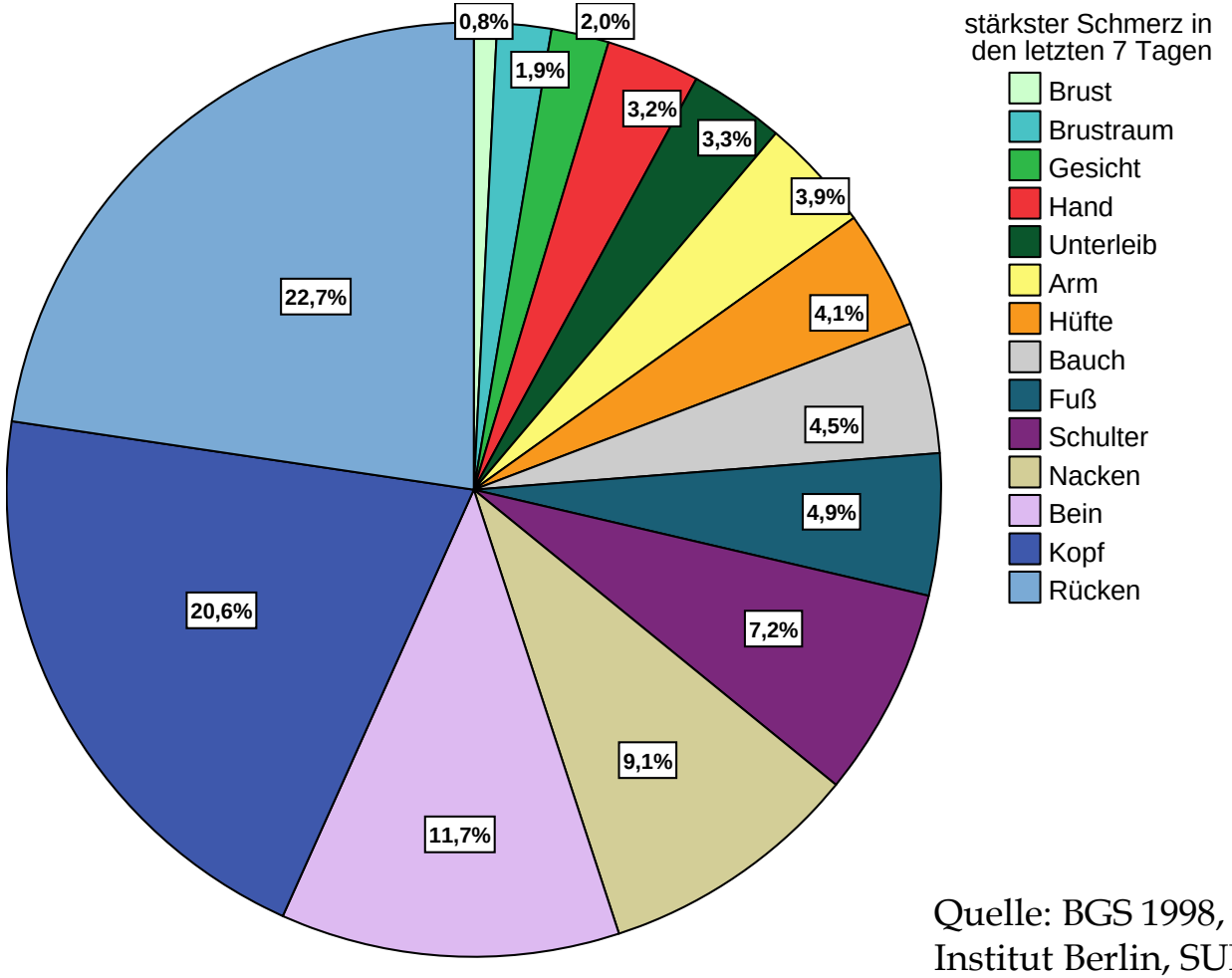
Epidemiologische Studie
(Ignatius et al., PLoS Negl Trop Dis. 2012; 6:e1677)



Gesundheitsberichtserstattung

- Bundesgesundheitssurvey
- Durchgeführt vom Robert-Koch Institut
- Repräsentative Stichprobe
- Ca. 8.000 Teilnehmer
- Alter, Geschlecht, Lebensstil, Gesundheitszustand
- Analyse der Daten primär beschreibend
- Keine wissenschaftlichen Hypothesen

Gesundheitsberichtserstattung

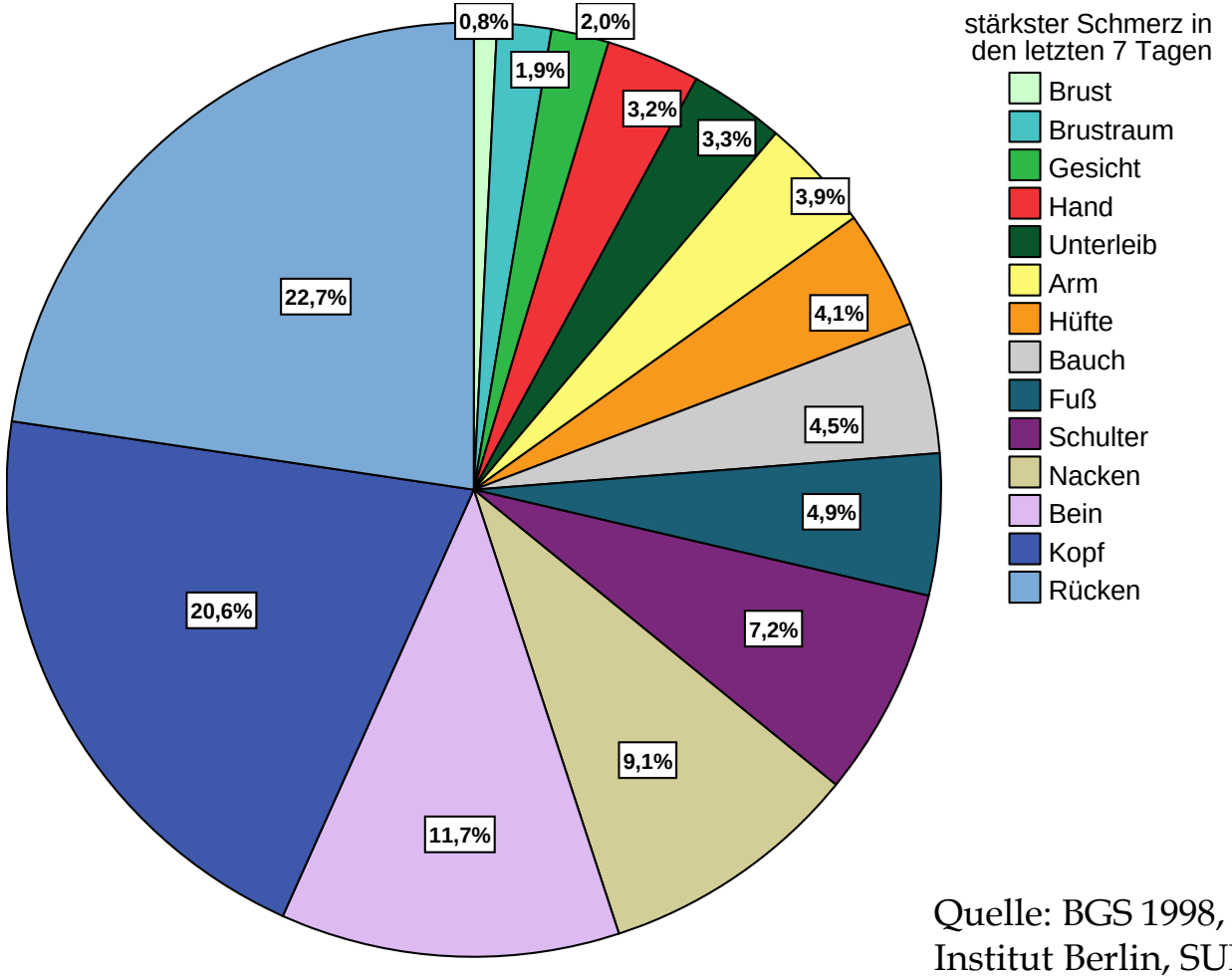


Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Teil 2
Diagrammtypen

Kreis- oder Kuchendiagramm

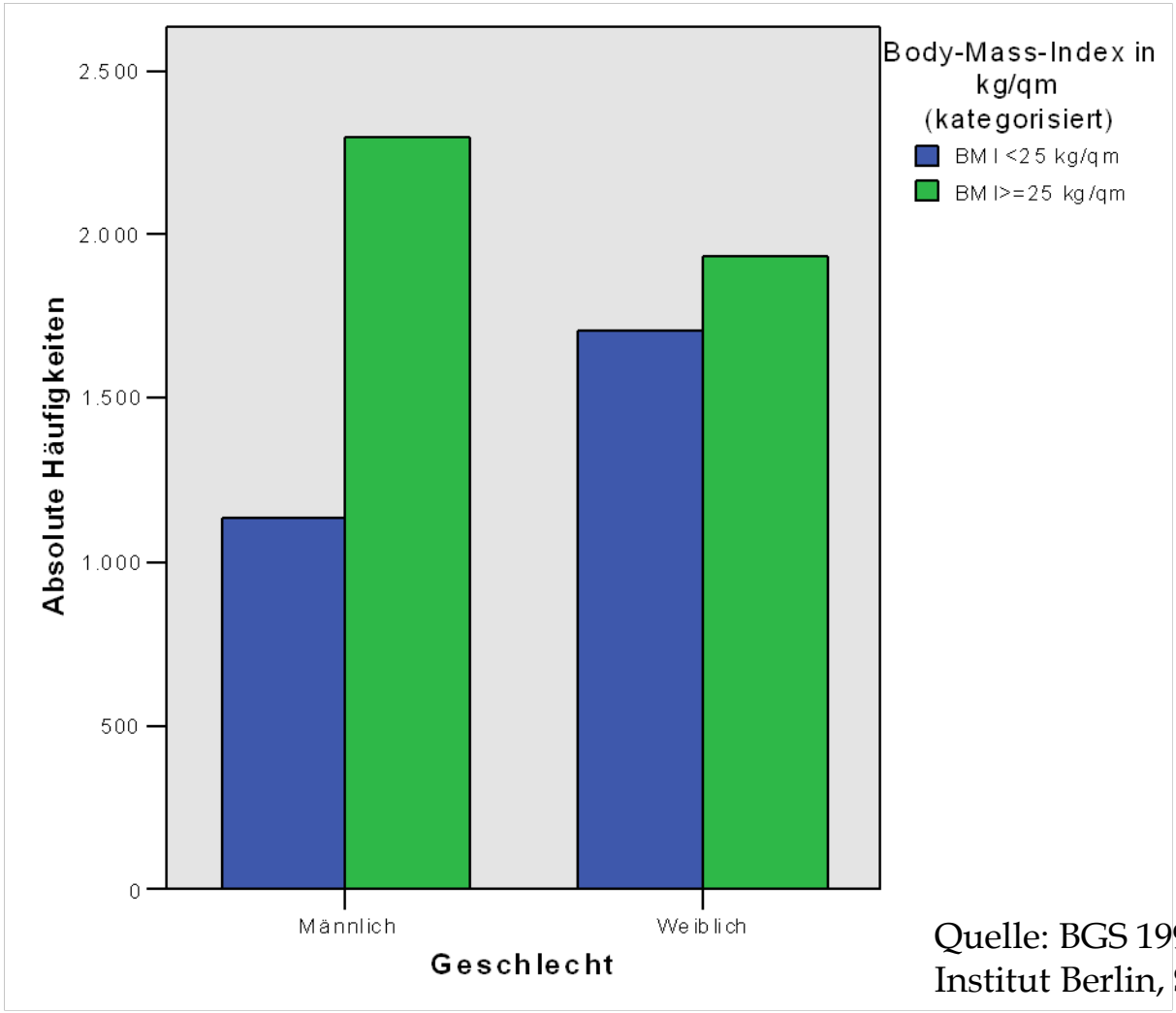
Schmerzlokalisation



Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Balkendiagramm

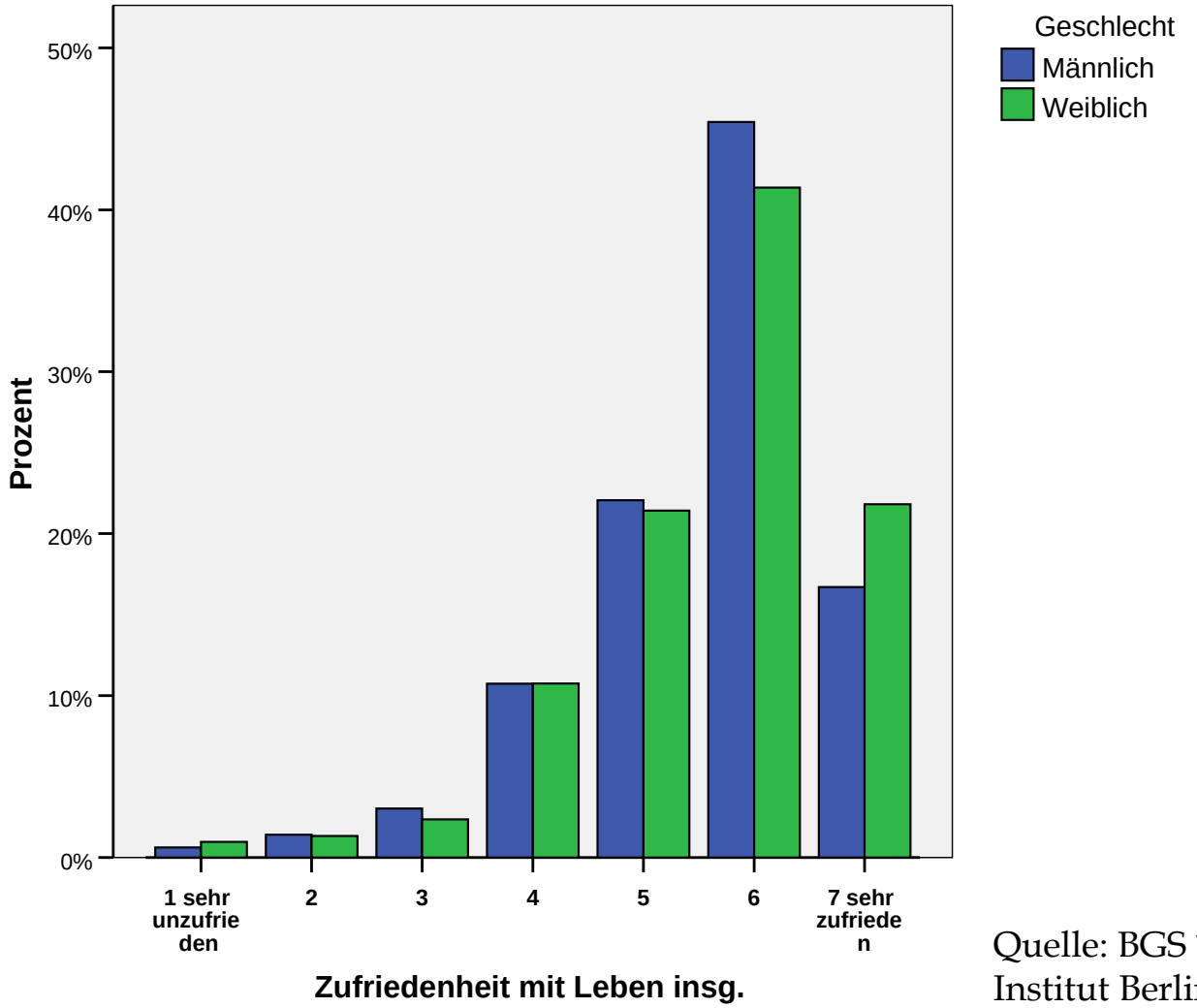
Kategorisierter Body Mass Index und Geschlecht



Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Balkendiagramm

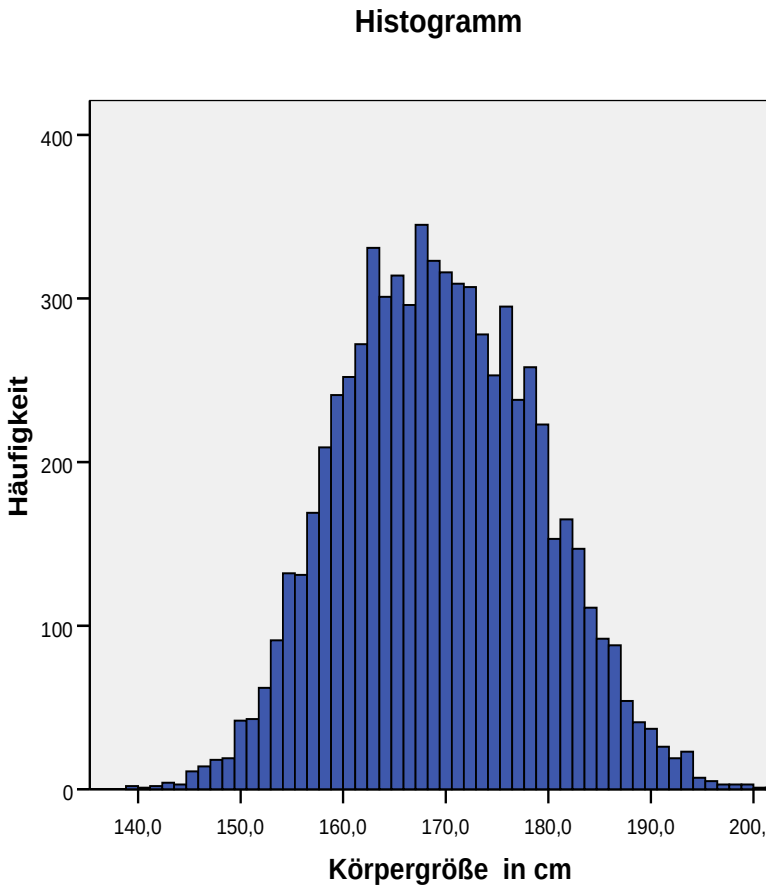
Lebenszufriedenheit und Geschlecht

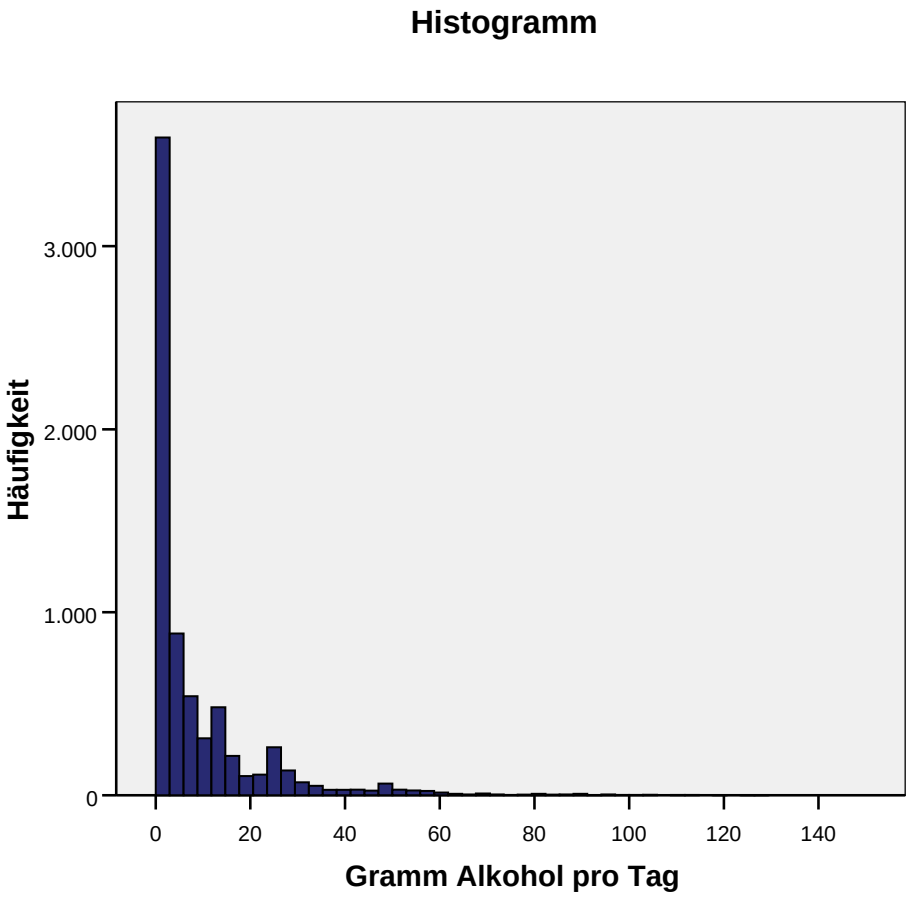


Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Histogramm

Körpergröße und Alkoholkonsum

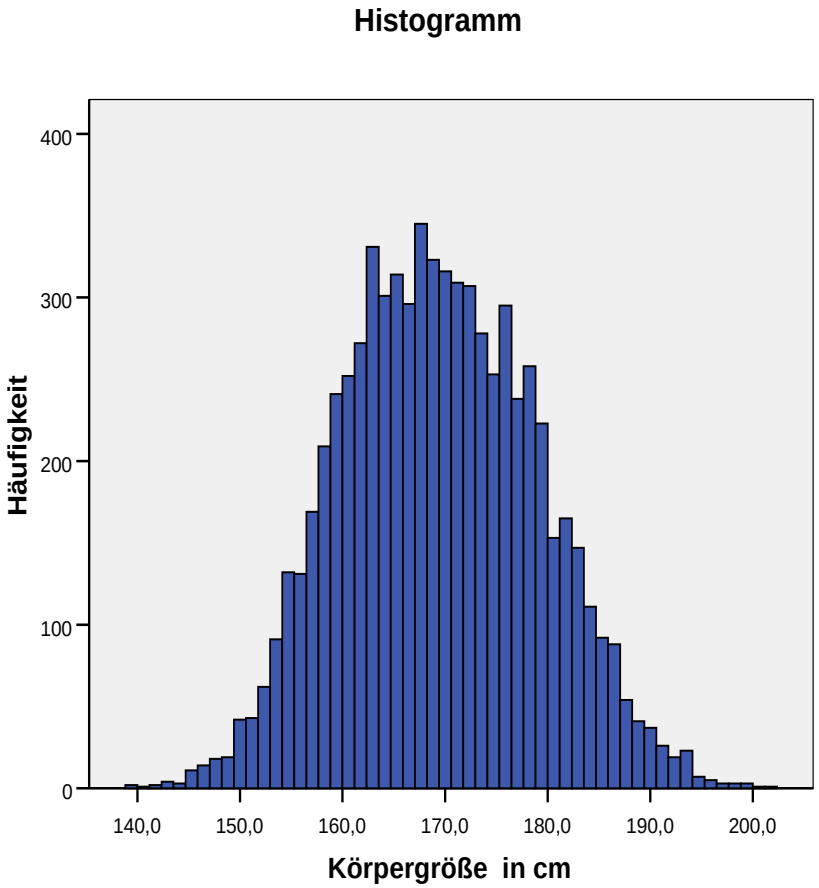


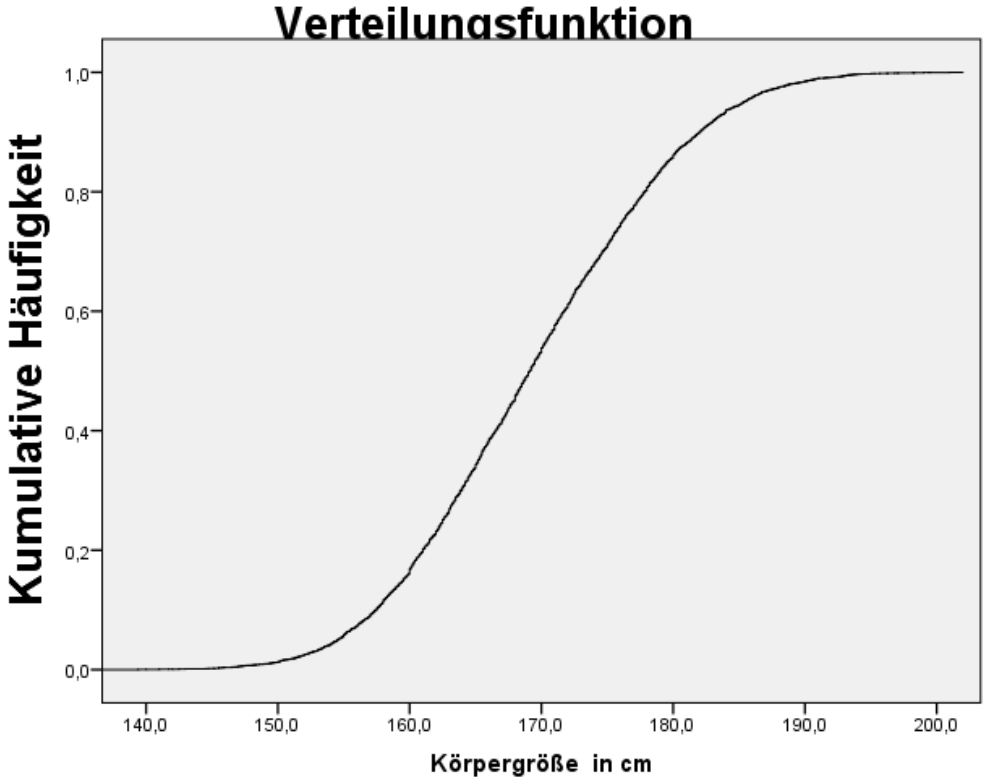


Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Histogramm und Verteilungsfunktion

Körpergröße

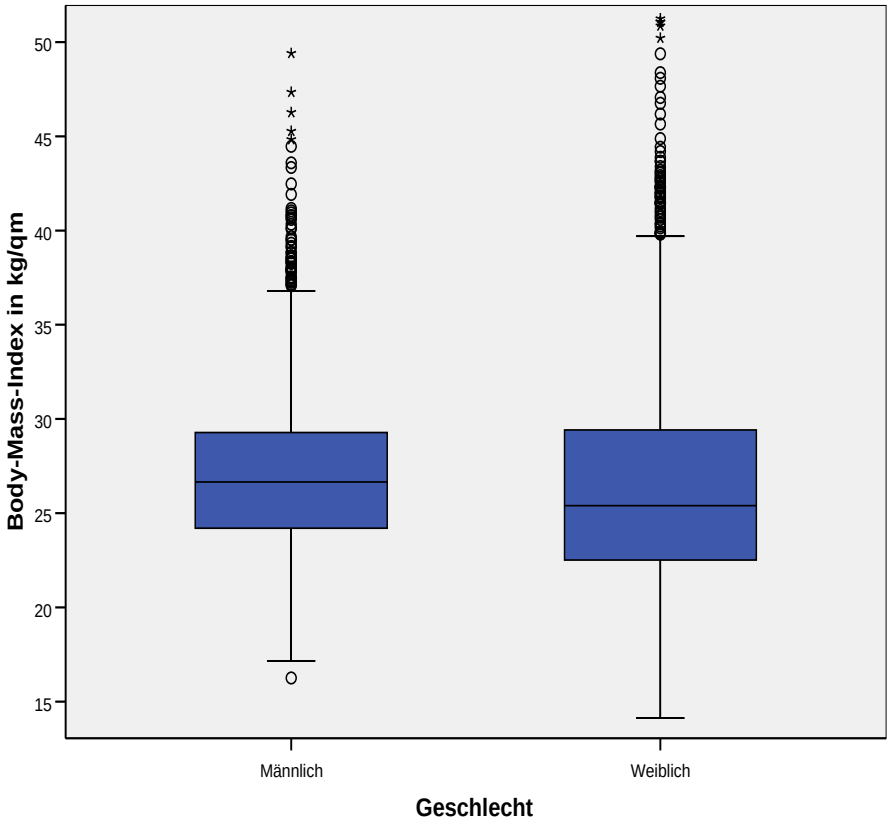




Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

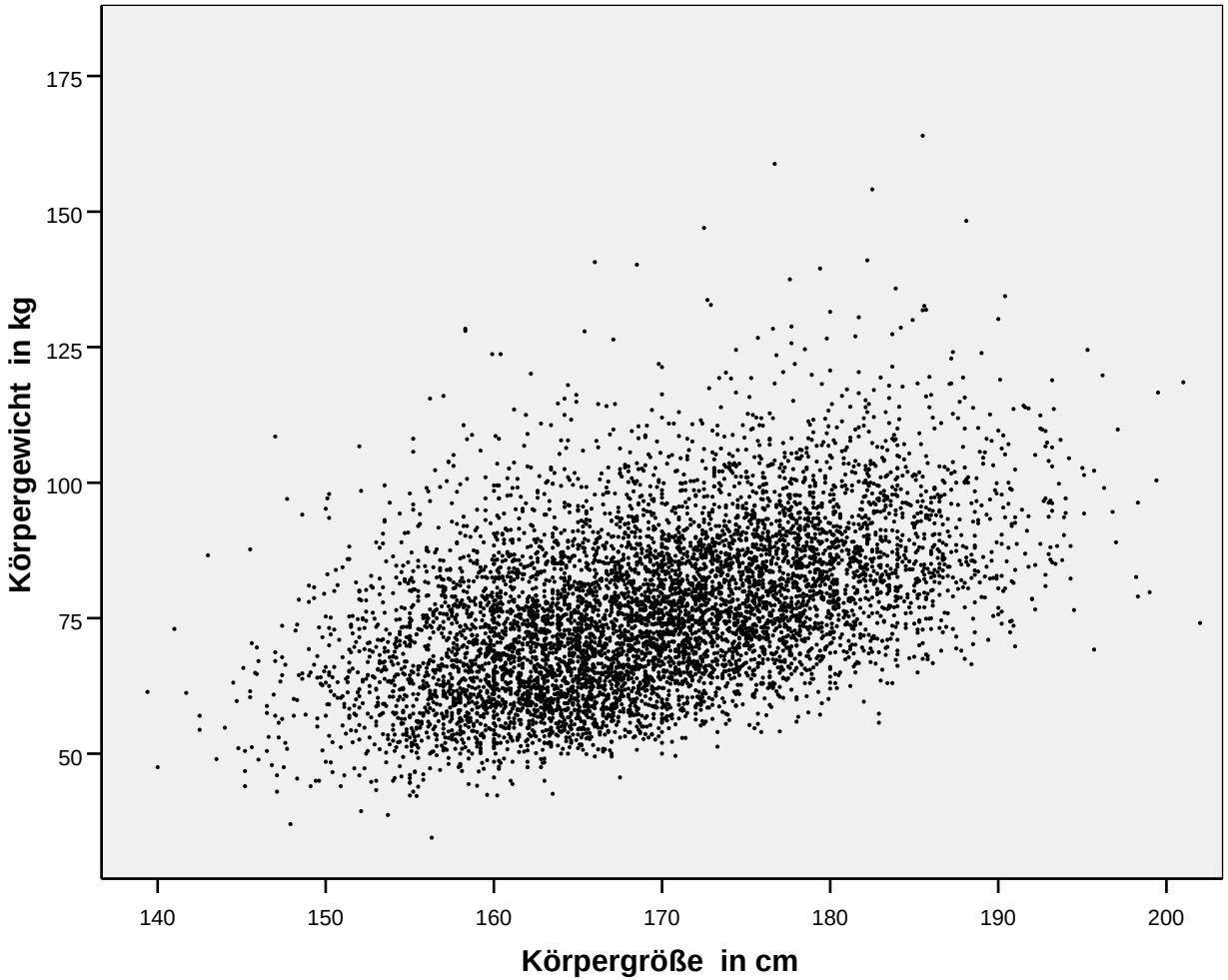
Boxplot

Body-Mass-Index [kg/m²] und Geschlecht



Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Streudiagramm
Körpergröße und Gewicht



Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Teil 3
Parameter

Mittelwert (Mean), Minimum, Maximum, Spannweite

Wir haben n Werte (x_1, x_2, ..., x_n) gemessen.

$$Mean = \bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Die einfachste Beschreibung der Daten enthält neben dem Mittelwert noch den kleinsten und größten Wert sowie deren Differenz (Spanweite)

Übungsaufgabe

Die n gemessenen Werte sind

1 5 7 8 10 12 13 24

Berechnen Sie Mittelwert und Spannweite

Median und Perzentile

Für die folgenden Parameter müssen die Daten sortiert werden:

$x_1 \leq x_2 \leq \ldots \leq x_n$.

Der Median ist für eine ungerade Zahl von Werten der mittlere, für eine gerade Zahl ist er mathematisch nicht eindeutig bestimmt, i.a. wird der der Mittelwert aus den beiden zentralen Werten bestimmt.

Perzentile werden analog für andere Prozentwerte (2.5%, 5%, 25%, 50% (Median), 75%, 95%, 97.5%) berechnet.

Die 25%, 50%, and 75% Perzentile heißen erste, zweite und dritte Quartile. Die Differenz aus 75% und 25% Perzentile heißt Interquartilsabstand.

Übungsaufgabe

Berechnen Sie für die Stichprobe zwei Folien vorher Median und Quartile

Wo ist der Zusammenhang der Perzentile mit dem Boxplot?

Varianz (Var), Standardabweichung (SD) und Variationskoeffizient (VC)

$$Var = \sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$SD = \sigma = \sqrt{Var}$$

$$VC = \frac{\sigma}{\bar{x}}$$

Für die Beschreibung von Daten ist die Standardabweichung (SD von Standard Deviation) anschaulicher als die Varianz.

Der Variationskoeffizient ist nur sinnvoll wenn die Werte nicht negativ werden und der Nullpunkt natürlich ist.

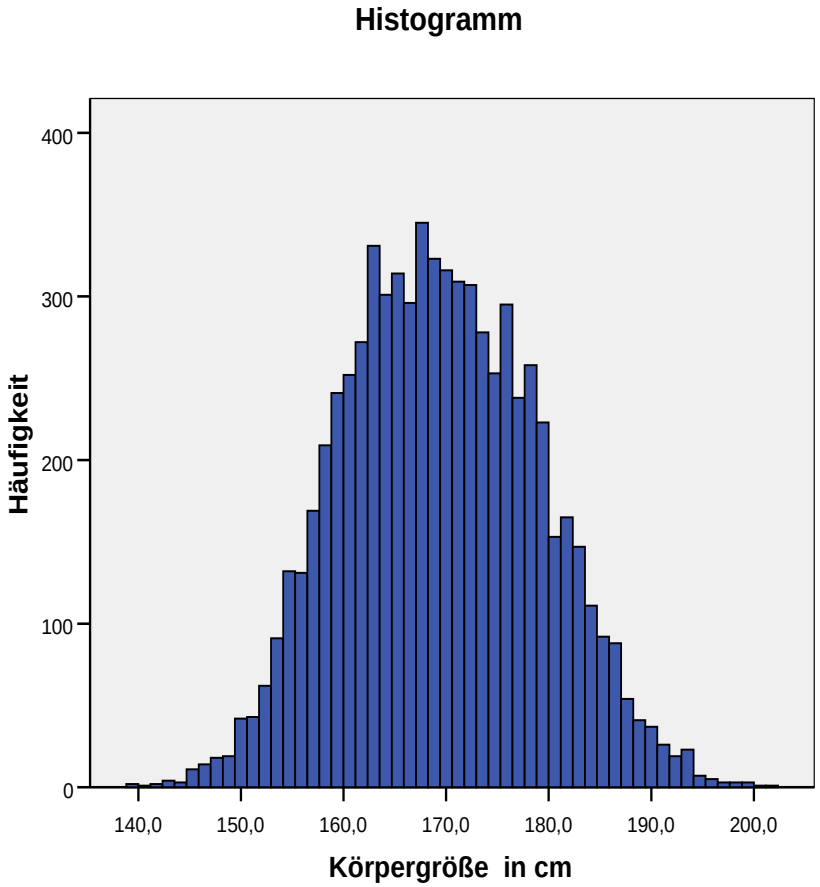
Übungsaufgabe

Berechnen Sie nun noch Varianz, Standardabweichung, Variationskoeffizient sowie den Interquartilsabstand.

Zusammenfassung

Körpergröße [cm}

N		7084
Mittelwert		169,43
Median		169,10
Varianz		89,603
Standardabweichung		9,466
Minimum		139,4
Maximum		202,0
Spannweite		62,6
Perzentile	25	162,5
	50	169,1
	75	176,1



Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Zusammenfassung
Parameter

- Mittelwert und Median nennt man (zentrale) Lageparameter
- Varianz, Streuung, Interquartilsabstand, Spannweite und der Variationskoeffizient heißen auch Streuungsparameter
- Minimum, Maximum und Perzentile sind auch Lageparameter, aber nicht zentrale.
- Ein in der Praxis selten verwendeter Parameter ist der Modalwert (häufigster Wert)

Teil 4
Codierung, Skalen und Werteverteilungen

Beispiele von Variablen und ihrer Codierung im Computer

Anzahl Kinder in einer Familie (1, 2, 3, …)

Temperature (Celsius oder Fahrenheit) (°)

Body Mass Index [BMI] (kg/m^2)

Body-mass-index 2 Categories (BMI < 25 (0), BMI >=25 (1)

Geschlecht (Männer (0), Frauen (1))

Hepatitis Virus (Typen A (1), B (2), C (3), D(4))

Lebenszufriedendheit (7 Stufen: 1=sehr unzufrieden, 7=sehr zufrieden)

Rauchgewohnheiten (täglich (1), manchmal (2), selten (3), Exraucher seit < 1 Jahr (4), Exraucher seit >= 1 Jahr (5), Nichtraucher (6))

Notwendigkeit der Einführung von Skalenniveaus

Die Codierung von Messungen im Computer kann zu Rechnungen „verführen“, die für die jeweiligen Messungen völlig sinnlos sind.

Es ist beispielsweise völliger Unsinn, einen Mittelwert für den Typ der Hepatitisinfektion so zu berechnen, als wäre es ein Messwert wie die Körpergröße.

Um hier Klarheit zu schaffen, hat man eine Einteilung von Messwerten nach unterschiedlichen „Skalen-niveaus“ festgelegt, die jedem Messwert die „erlaubten“ Rechnungen und die empfohlenen Graphiken zuordnet.

Skalenniveaus

Qualitativ

Quantitativ

- Nominal Skala (Kategorien ohne spezifische Ordnung, Sortierreihenfolgen sind mathematisch völlig beliebig)
- Spezialfall: Dichotome = binäre Skala mit zwei Kategorien
- Ordinal Skala (geordnete Kategorien, Abstände nicht quantifizierbar, aber Werte sinnvoll sortierbar)
- Intervall Skala (Differenzen sinnvoll)
- Rational Skala (Differenzen und Quotienten sinnvoll)

Zusätzlich:

- Stetige Messwerte vs Diskrete Skalen

Skalenniveaus und Graphiken

	Kreis-diagramm	Balken-diagramm	Boxplot	Histo-gramm	Verteilungs-funktion
Nominal	+++	+++	---	---	---
Ordinal	+	+++	+++	---	---
Intervall	---	-	+++	+++	+++
Rational	---	-	+++	+++	+++

Skalenniveaus und Lageparameter

	Mittel-wert	Median	Modal-wert	Perzentile	Minimum Maximum
Nominal	---	---	+++	---	---
Ordinal	---	+++	+	+++	+++
Intervall	+++	+++	+	+++	+++
Rational	+++	+++	+	+++	+++

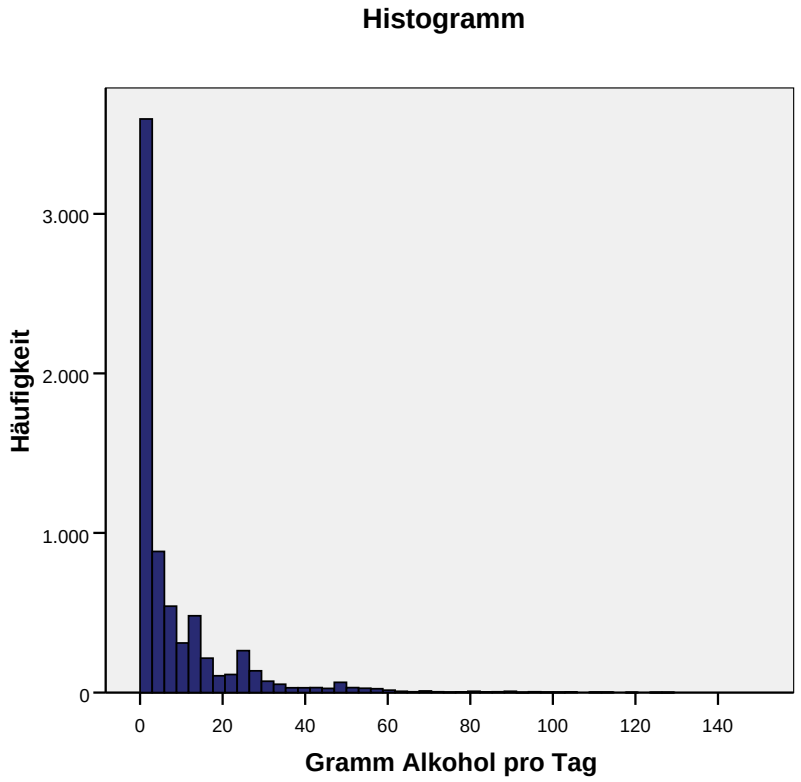
Skalenniveaus und Streuungsparameter

	Varianz Streuung	Variations-koeffizient	Spann-weite	Interquar-tilsabstan	Andere
Nominal	---	---	---	---	+++
Ordinal	---	---	+	+	
Intervall	+++	+	+++	+++	
Rational	+++	+++	+++	+++	

Einfluss der Werteverteilung

Das Skalenniveau alleine reicht manchmal nicht aus, um die geeigneten statistischen Methoden auszuwählen:

N		7124
Mittelwert		8,72
Median		2,81
Varianz		201,34
Standardabweichung		14,19
Minimum		0
Maximum		164,27
Spannweite		164,27
Perzentile	**25**	0,25
	50	2,81
	75	11,88



Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Kriterien für die Auswahl statistischer Methoden bei quantitativen Daten

Im vorherigen Beispiel ist der Median geeigneter als der Mittelwert wenn man den Einfluss von Ausreißern begrenzen will.

Grundsätzlich sollte man quantitative Daten mit Ausreißern oder unsymmetrischer Verteilung lieber durch Mediane / Perzentile /Interquartilsabstände als durch Mittelwerte, Streuungen, Varianzen oder Variationskoeffizienten beschreiben.

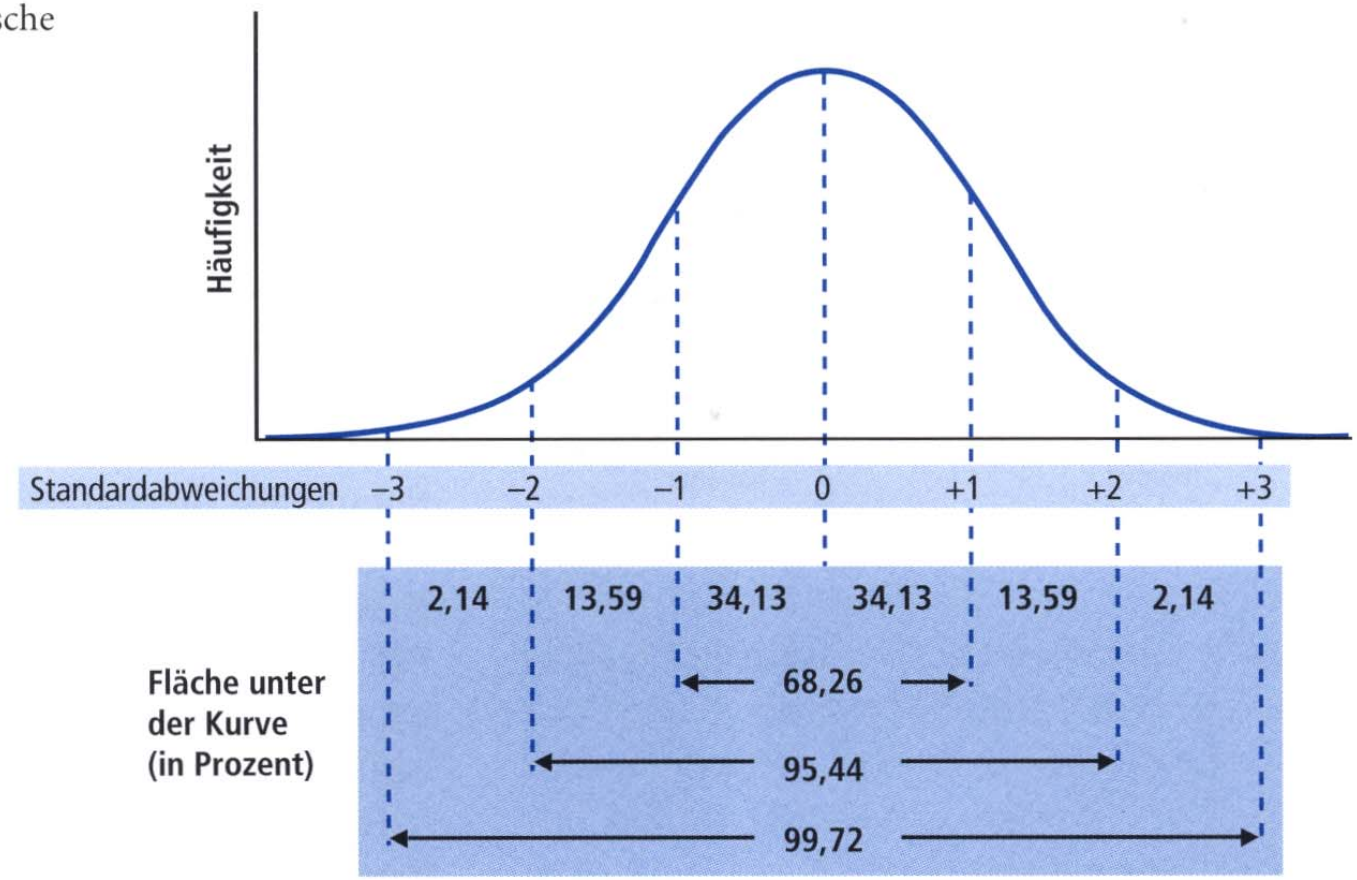
Am meisten Ausreißer-anfällig ist natürlich die Spannweite.

Die „optimale“ Datenverteilung für Mittelwerte, Streuungen, Varianzen ist die Normalverteilung.

Teil 5
Normalverteilung und Logarithmustransformation

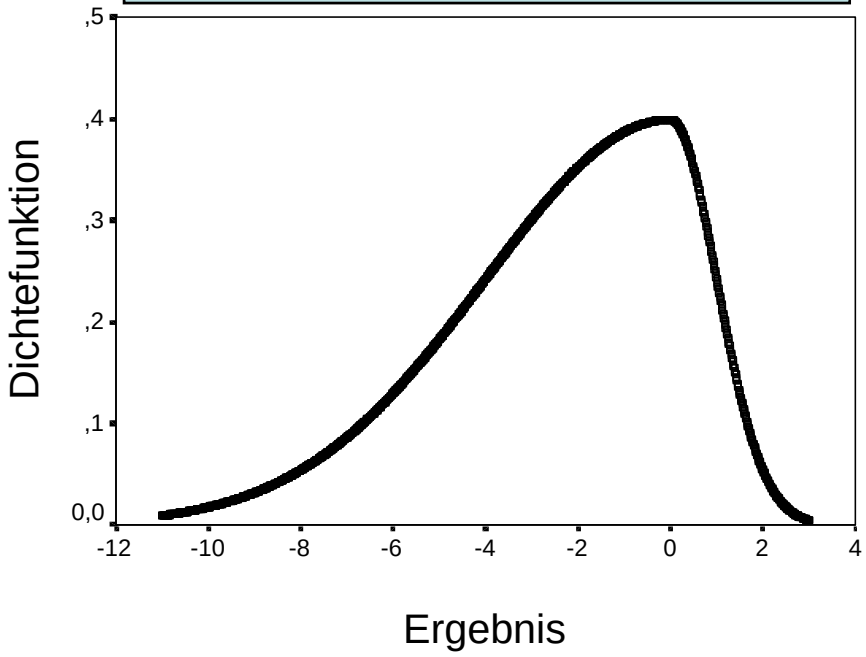
Die Gauss- oder Normalverteilung

Abbildung 2-8: Die Gauß'sche Normalverteilung.

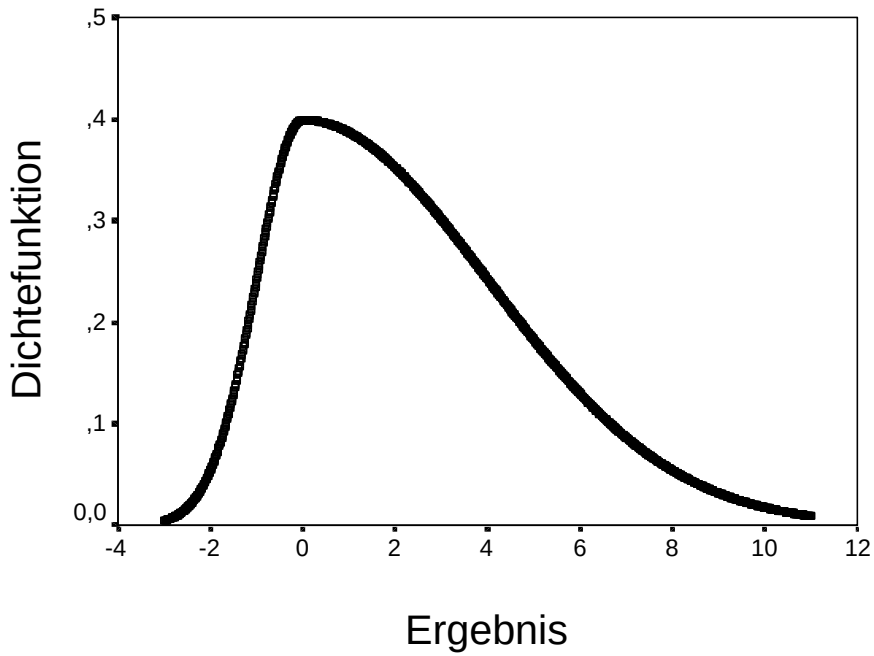


Abweichungen von der Normalverteilung links- und rechtsschiefe Verteilungen

Linksschief, Schiefe < 0



Rechtsschief, Schiefe > 0



„Kür“: Parameter für die Überprüfung der Normalverteilung – Schiefe und Kurtosis

Mit Schiefe und Kurtosis lassen sich Abweichungen von der Normalverteilung durch Zahlen ausdrücken.

Die Schiefe prüft Symmetrie der Werteverteilung. Im Falle perfekter Symmetrie haben die Daten eine Schiefe von 0. Links schiefe Verteilungen haben negative, rechtsschiefe Verteilungen positive Schiefe.

Die Kurtosis vergleicht den Anteil der Daten im Zentrum der Verteilung mit extremen Abweichungen. Perfekt normalverteilte Daten haben eine Kurtosis von 0. Die Gleichverteilung hat negative Kurtosis (zu wenige Extremwerte), eine Verteilung mit vielen Werten im Zentrum aber auch extremen Abweichungen hat negative Kurtosis .

Faustregel: Daten sind zumindest angenähert normalverteilt wenn Schiefe und Kurtosis zwischen -1 und + 1 liegen.

„Kür“: Formeln

$$\text{Schiefe} = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\sigma^3}$$

$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{\sigma^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

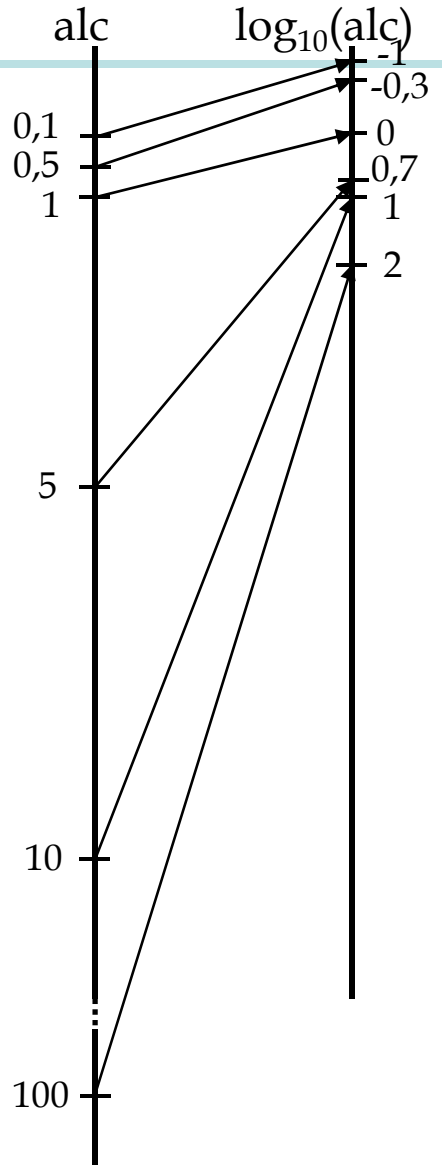
Die Logarithmustransformation

- **Beispiel:**
  Alkoholkonsum pro Tag
- **Berechnung:**

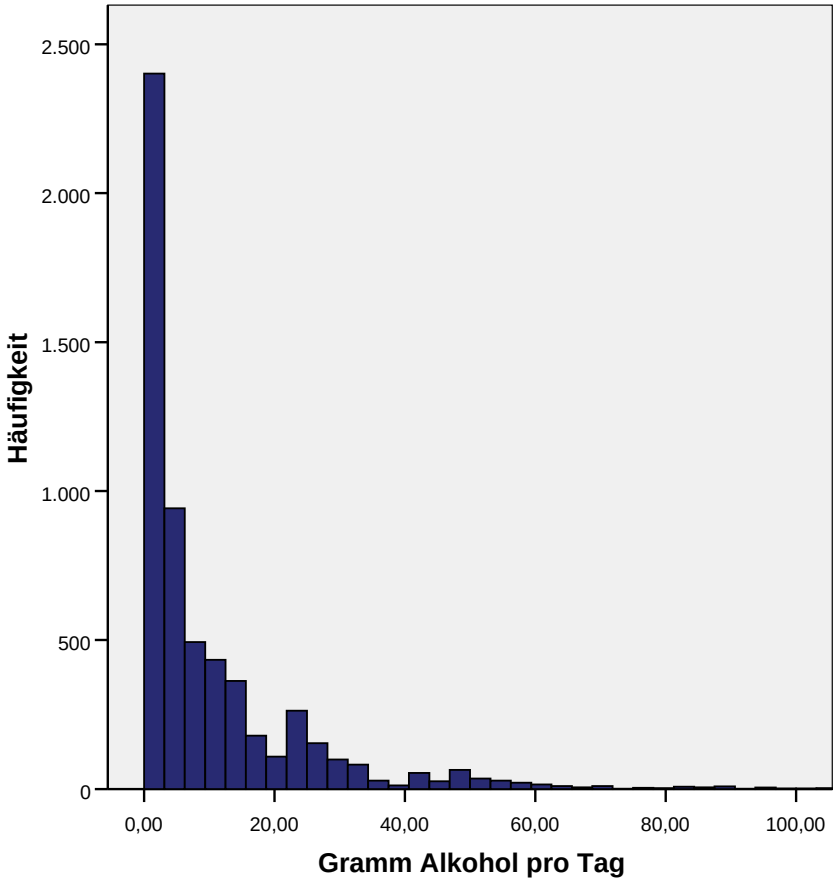
  alc → $\log_{10}$(alc) (definiert für alc > 0)

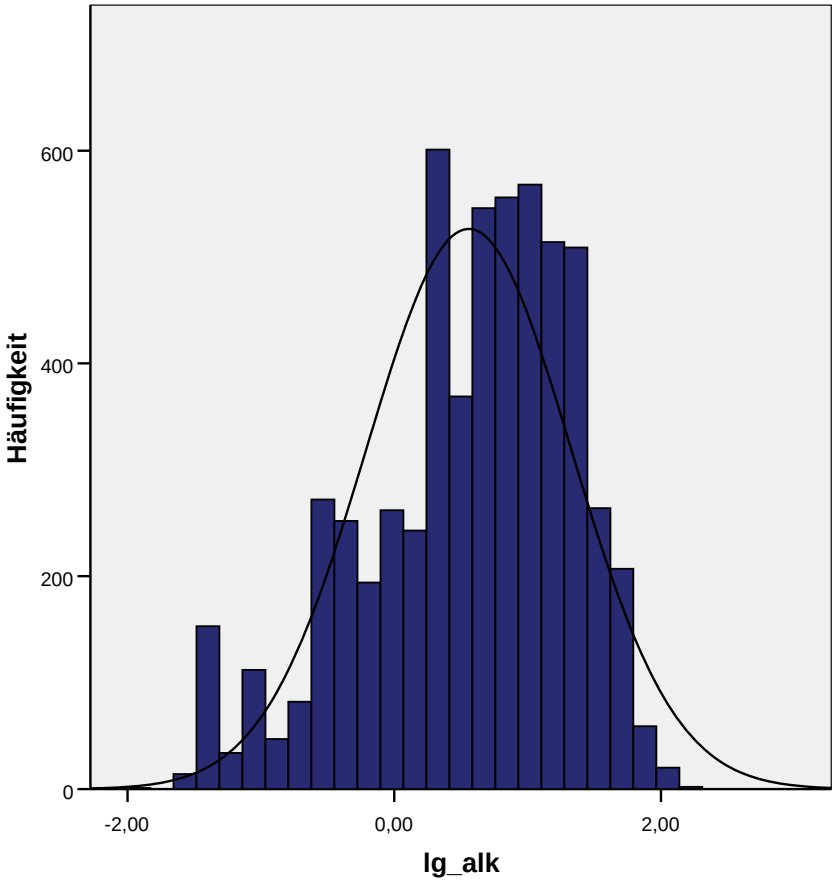
  alc → $\log_{10}$(alc+δ)

alk	**$\log_{10}$(alk)**
0,1	-1
0,5	-0,3
1	0
5	0,7
10	1
100	2



Die Logarithmustransformation





Das geometrische Mittel

Offenbar ist es nach unserem Kriterien „erlaubt“, für die logarithmierten Alkoholwerte einen Mittelwert zu bestimmen.

Dieser Mittelwert liegt bei 0.56.

Der Wert ist nicht aussagekräftig und schwer zu interpretieren. Rücktransformation ergibt aber einen Wert auf der ursprünglichen Skala:

$10^{0.56} = 3.62$

Zum Vergleich:

Der gewöhnliche Mittelwert ist 10.56, der Median 4.53

Der Vorteil der log-Transformation liegt darin, dass für normalverteilte Werte viel mehr und aussagekräftigere statistische Methoden zur Verfügung stehen.

Alternative Berechnung des geometrischen Mittels

Man kann – zumindest theoretisch – das geometrische Mittel auch direkt, also ohne Umweg über Logarithmen berechnen.

Für Werte (x_1, x_2, ..., x_n) ist das geometrische Mittel

$$GeomMean = \sqrt[n]{\left(x_1 \times x_2 \times \ldots \times x_n\right)} = 10^{\frac{\sum_{i=1}^{n} \lg_{10} x_i}{n}}$$

Teil 6
Grundsätzliches zur statistischen Denkweise

Warum können wir Studienergebnisse verallgemeinern?

- Wenn wir über die Population unserer Studie hinaus gehende Aussagen treffen wollen („60% der Männer *in Deutschland* sind übergewichtig“) müssen die Ergebnisse des BGS verallgemeinerbar sein.
- Mit Hilfe der Wahrscheinlichkeitsrechnung können die Ergebnisse für repräsentative Studien abgesichert werden.
- Aber die Repräsentativität der Studie ist oft unklar. Wichtigstes aber nicht einziges Kriterium für Repräsentativität ist die Zufallsauswahl der Studienteilnehmer.

Stichprobe und Zielpopulation

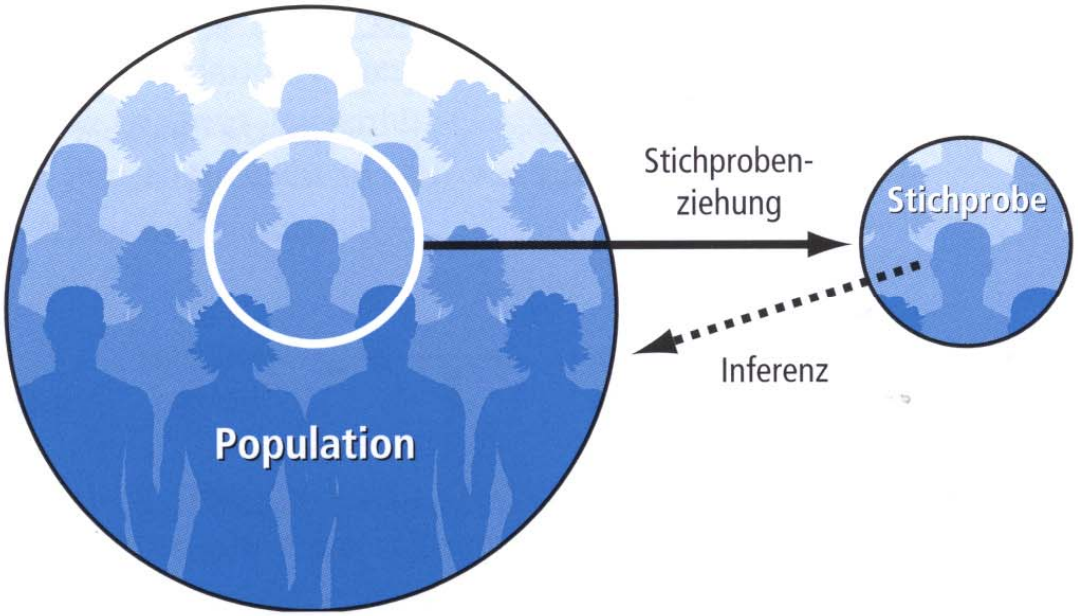


Abbildung 1-3: Population und Stichprobe.

Statistische Terminologie

Auf der vorhergehenden Folie wurde erstmals der Begriff „Stichprobe“ erwähnt.

Er ist einer von mehreren zentralen Begriffen, die zur Basisterminologie der Statistik gehören und deswegen zum Abschluss aufgelistet werden.

Statistische Terminologie

Merkmalsträger = Beobachtungseinheit: Mensch, Tier, oder Objekt, an dem bestimmte Eigenschaften gemessen werden (Person, Tier, Zellkultur, evtl. aber auch Zwillingspaar, Familie, Dorf, Land).

Variable = Merkmal: Messgröße als Ganzes (z.B. Körpergröße)

Ausprägung: einzelner Messwert (z.B. 180 cm)

Studienpopulation = Stichprobe: Population der Merkmalsträger in der Studie

Zielpopulation = Grundgesamtheit: Population auf welche die Studienergebnisse verallgemeinert werden sollen.