

## Modul 8

---

Einstieg in die Wissenschaft  
Methoden  
Ergebnisse und Interpretationen

**Besser nicht lügen mit Statistik –  
Einführung in die beschreibende  
Statistik**

**Fachvorlesung (90 min)**

1

## M8 Lehrveranstaltungen Biometrie

---

- VL Besser nicht lügen mit Statistik –  
Einführung in die beschreibende Statistik ↔
- FS Einführung in die Beschreibende Statistik
- VL Lieber auf Nummer sicher gehen –  
Einführung in den Statistischen Test
- FS Prinzip des Statistischen Tests
- VL Loslegen Können – Überblick über  
Statistische Testverfahren
- FS Praktische Anwendung Statistischer Tests

2

## Lernziele

---

Die Studierenden sollen

den Begriff der Stichprobe definieren können,

Skalenniveaus am Beispiel erkennen und die Unterschiede verschiedener Skalenniveaus benennen können,

die wesentlichen elementaren Techniken der deskriptiven Statistik benennen können,

für eine randomisierte Therapiestudie die Stichprobenziehung beurteilen und eine deskriptive Analyse grob konzipieren können,

Scheu vor statistischer Denkweise abbauen, Interesse für das Gebiet entwickeln.

3

## Warum dürfen wir Studienergebnisse verallgemeinern?

---

- I.a. werden statistische Erhebungen nicht an allen Betroffenen durchgeführt (Ausnahme: Amtliche Todesursachenstatistik)
- Meist beschränken sich Studien auf *Stichproben* wie z.B. den Bundes-Gesundheitssurvey des Robert Koch Instituts
- Man hofft, die Studienergebnisse auf die Gesamtheit aller Betroffenen (die „Grundgesamtheit“) übertragen zu können.
- Voraussetzung hierfür ist, dass die Stichprobe (1) *repräsentativ* für die Grundgesamtheit und (2) *hinreichend groß* ist.
- Eine Möglichkeit, Repräsentativität zu erreichen, besteht in der zufälligen Auswahl der Stichprobe aus der Grundgesamtheit.
- In der Praxis kann dieses Ziel aber nur angenähert werden.

4

## Definition Stichprobe

---

- Eine Gruppe von Personen (Patienten / gesunde Probanden) bezeichnet man als *Stichprobe*, wenn sie für eine größere Grundgesamtheit repräsentativ ist.
- Andernfalls sollte man von einer *Fallserie* sprechen.
- Es ist sozusagen das Kerngeschäft von Epidemiologie und Biometrie die *Verallgemeinerungsfähigkeit* von Stichprobenergebnissen zu sichern.

5

## Deskriptive Statistik

---

### **Problem:**

Wie lassen sich die erhobenen Daten übersichtlich darstellen und beschreiben?

### **Lösung:**

durch geeignete Grafiken und Maßzahlen, die die Daten repräsentieren und zusammenfassen

6

## Beispieldatensatz Bundes-Gesundheitssurvey

---

Bundes-Gesundheitssurvey 1998 (BGS98)

- N=7124, Altersspanne: 17-79 Jahre
- Informationen zu u.a.:
  - Alter, Geschlecht, Größe, Gewicht, BMI
  - Krankheiten, physischem und emotionalem Gesundheitszustand
  - Ernährung, Rauchen, Alkoholkonsum
  - Bewegung, Sport, Stress
  - Zufriedenheit, Zufriedenheit mit Ärzten
  - Bildungsstand, Beruf, Einkommen
  - Familiensituation

7

## Wodurch unterscheiden sich die folgenden Merkmale:

---

Familienstand in fünf Kategorien

1 = Verheiratet, zusammen lebend, 2 = Verheiratet, getrennt lebend, 3 = ledig, 4 = geschieden, 5 = verwitwet

Schulabschluss in fünf Kategorien

1 = kein Abschluss, 2 = Hauptschule, 3 = Realschule, 4 = Fachoberschule, 5 = Gymnasium

Körpergröße

Messung mit Genauigkeit von 1 cm

Merkmale bezeichnet man auch als Variablen, mögliche Werte als Ausprägungen

8

## Skalierung von Variablen

Kategorielle Variablen Beispiel Familienstand	Ordinale Variablen Beispiel Schulabschluss	Quantitative Variablen Beispiel Körpergröße
Mit den Werten sind <b>weder Größenvergleiche noch Rechenoperationen</b> möglich.	Mit den Werten sind <b>nur Größenvergleiche jedoch keine Rechenoperationen</b> möglich.	Mit den Werten sind <b>sowohl Größenvergleiche als auch Rechenoperationen</b> möglich.

9

## Deskription kategorieller und ordinaler Variablen

10

## Eine Kategorielle Variable Häufigkeitstabelle

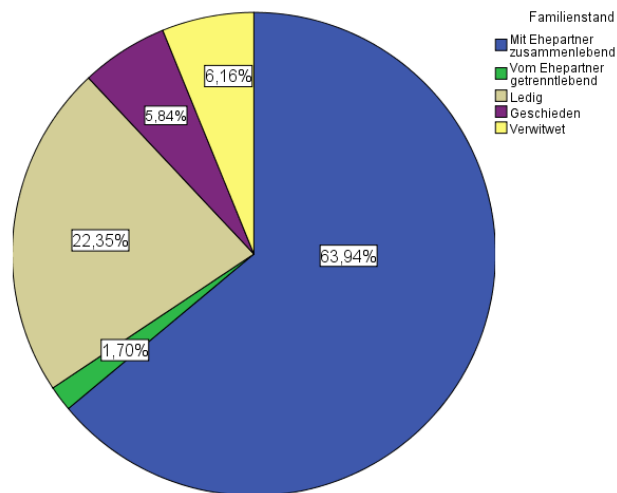
Absolute und relative Häufigkeiten

f096 Familienstand

		Häufigkeit	Prozent	Gültige Prozente
Gültig	1 Mit Ehepartner zusammenlebend	4432	62,2	63,9
	2 Vom Ehepartner getrenntlebend	118	1,7	1,7
	3 Ledig	1549	21,7	22,3
	4 Geschieden	405	5,7	5,8
	5 Verwitwet	427	6,0	6,2
	Gesamt	6931	97,3	100,0
Fehlend	System	193	2,7	
Gesamt		7124	100,0	

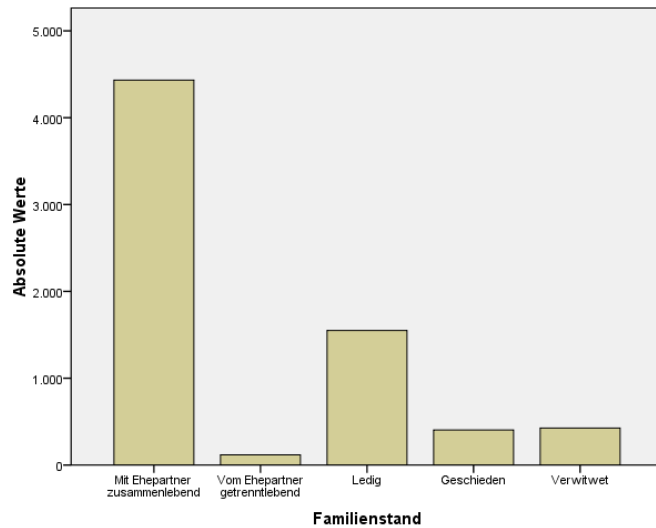
11

## Eine Kategorielle Variable Kreisdiagramm



12

## Eine Kategorielle Variable Balkendiagramm



13

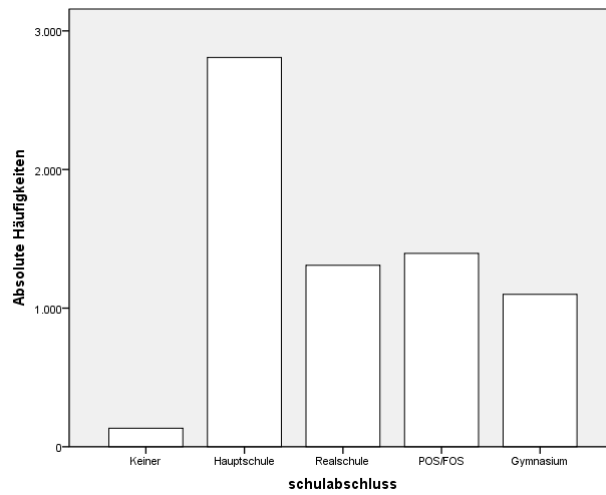
## Eine Ordinale Variable Häufigkeitstabelle

Absolute, relative und *kumulative* Häufigkeiten

		schulabschluss			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00 Keiner	134	1,9	2,0	2,0
	2,00 Hauptschule	2808	39,4	41,6	43,6
	3,00 Realschule	1309	18,4	19,4	63,0
	4,00 POS/FOS	1395	19,6	20,7	83,7
	5,00 Gymnasium	1100	15,4	16,3	100,0
	Total	6746	94,7	100,0	
Missing	99,00 andere	181	2,5		
	System	197	2,8		
	Total	378	5,3		
Total		7124	100,0		

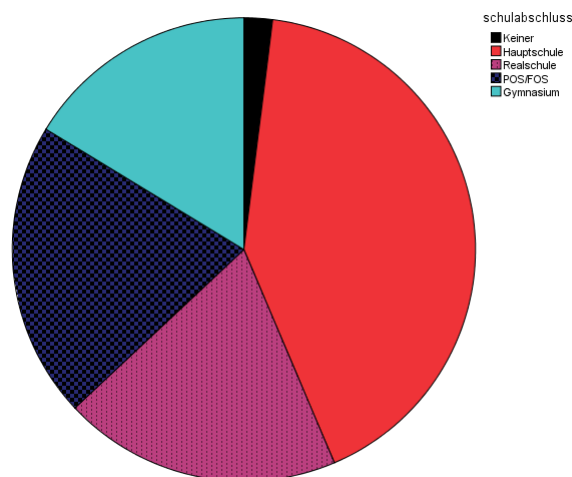
14

## Eine Ordinale Variable Balkendiagramm



15

## Schulabschluss

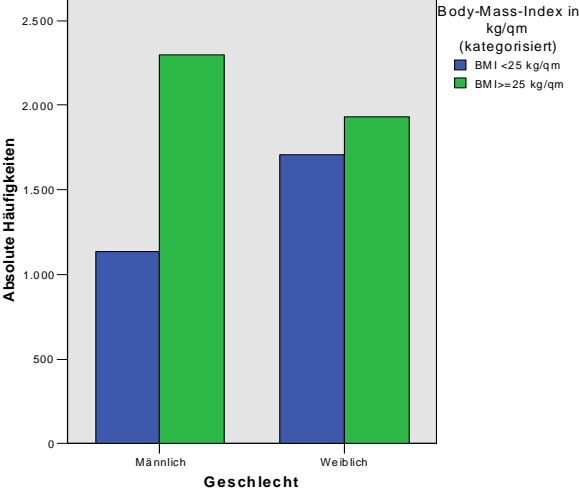


Ist das Kreisdiagramm geeignet???

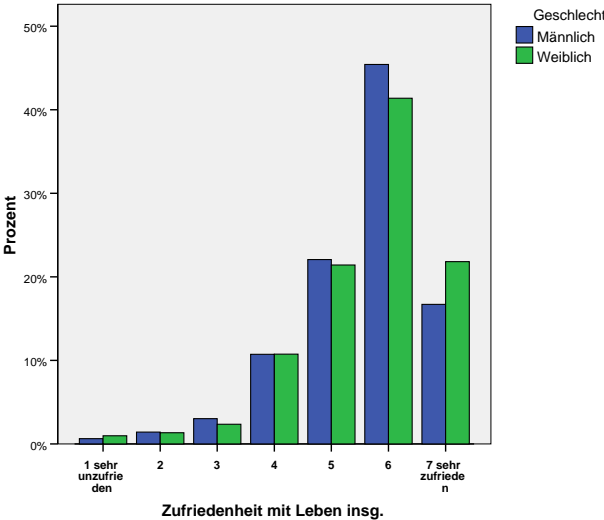
16

# Zwei Kategoriale Variablen Gruppiertes Balkendiagramm

(BGS98, n=7072)



# Eine kategoriale und eine ordinale Variablen Gruppiertes Balkendiagramm



## Zwei kategoriale Variablen Kreuztabelle

- Absolute und relative Häufigkeiten

**sex Geschlecht \* bmi\_kat BMI in Kategorien Kreuztabelle**

				bmi_kat BMI in Kategorien		
				1,00 BMI unter 25	2,00 BMI größer oder gleich 25	Gesamt
sex Geschlecht	1 Männlich	Anzahl		1138	2297	3435
		% von sex Geschlecht		33,1%	66,9%	100,0%
	2 Weiblich	Anzahl		1706	1931	3637
		% von sex Geschlecht		46,9%	53,1%	100,0%
Gesamt		Anzahl		2844	4228	7072
		% von sex Geschlecht		40,2%	59,8%	100,0%

19

## Eine kategorielle und eine ordinale Variablen Kreuztabelle

Geschlecht \* Zufriedenheit mit Leben insg. Crosstabulation

		Zufriedenheit mit Leben insg.								
		1 sehr unzufrieden	2	3	4	5	6	7 sehr zufrieden	Total	
Geschlecht	Männlich	Count	21	47	101	358	736	1515	557	3335
		% within Geschlecht	,6%	1,4%	3,0%	10,7%	22,1%	45,4%	16,7%	100,0%
	Weiblich	Count	34	47	83	379	756	1460	770	3529
		% within Geschlecht	1,0%	1,3%	2,4%	10,7%	21,4%	41,4%	21,8%	100,0%
Total		Count	55	94	184	737	1492	2975	1327	6864
		% within Geschlecht	,8%	1,4%	2,7%	10,7%	21,7%	43,3%	19,3%	100,0%

---

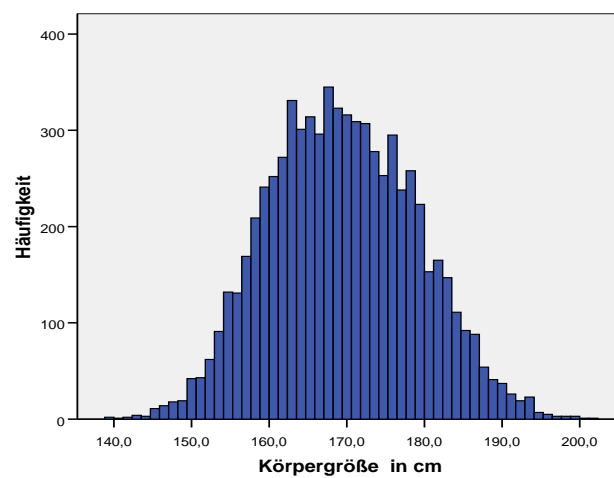
# Deskription quantitativer Variablen

21

## Eine Quantitative Variable Histogramm

---

Körpergröße in cm



22

## Eine Quantitative Variable Mittelwert, Varianz und Streuung

Mittelwert, Varianz und Streuung für Messwerte  $x_1, x_2, \dots, x_n$

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

Die folgenden Formeln sind nur der Vollständigkeit halber angegeben und kein Lernstoff.

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

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

Die Varianz (Var) ist nur von theoretischer Bedeutung, für die Datenbeschreibung verwendet man fast ausschließlich die Streuung (= Standardabweichung, SD)

23

## Eine Quantitative Variable Median und Quartile

Eine weitere Möglichkeit zur Beschreibung der Daten beruht auf der Sortierung nach Größe:

Der **Median** ist der mittlere Wert der sortierten Stichprobe  
m.a.W. 50% der Werte sind jeweils kleiner/gleich oder größer gleich dem Median

Die **Quartile** werden analog für 25% und 75% der Stichprobe definiert

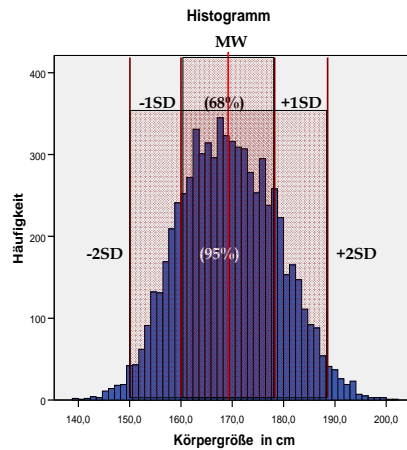
25% der Werte sind kleiner/gleich der ersten Quartile = 25% Perzentile

75% der Werte sind kleiner/gleich der dritten Quartile = 75% Perzentile

24

## Eine Quantitative Variable Histogramm

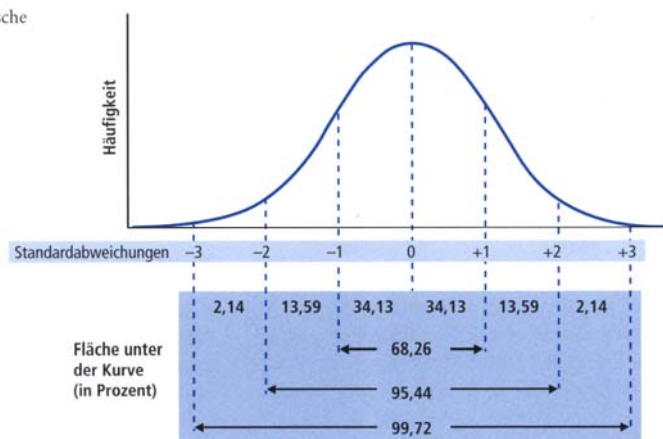
N	7084
Mittelwert	169,43
Median	169,10
Varianz	89,603
Standardabweichung	9,466
Minimum	139,4
Maximum	202,0
Spannweite	62,6
Perzentile	
25	162,5
50	169,1
75	176,1



25

## Die Normalverteilung Idealisierung quantitativer Variablen

Abbildung 2-8: Die Gauß'sche Normalverteilung.



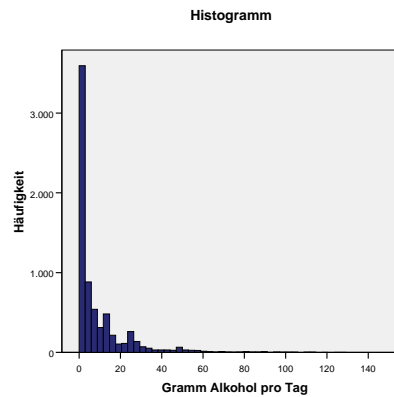
Aus Fletcher et al. Klinische Epidemiologie

26

## Alkoholkonsum pro Tag in Gramm reinem Alkohol

### Normalverteilung???

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

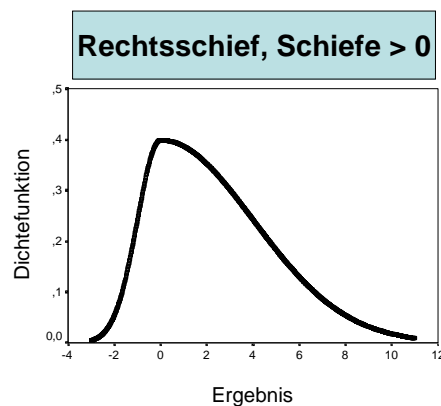
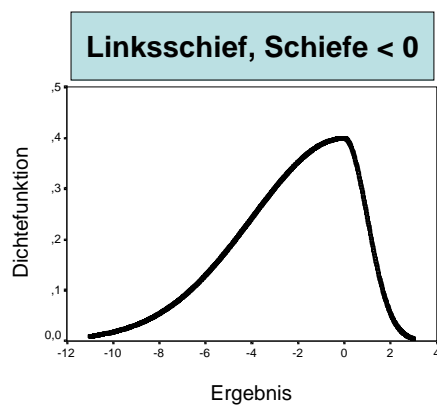


8,72g /Tag entspricht etwa 1,5 Liter Bier pro Woche...

2,81g /Tag entspricht etwa 0,5 l Bier pro Woche

27

## Abweichungen von der Normalverteilung Links- und Rechtsschiefe Verteilungen



28

## Histogramm vs. Balkendiagramm

Balkendiagramme geben nur diejenigen Kategorien wieder, die in den Daten tatsächlich vorkommen.

Balkendiagramme können also verwendet werden

-*Immer* bei kategoriellen Daten

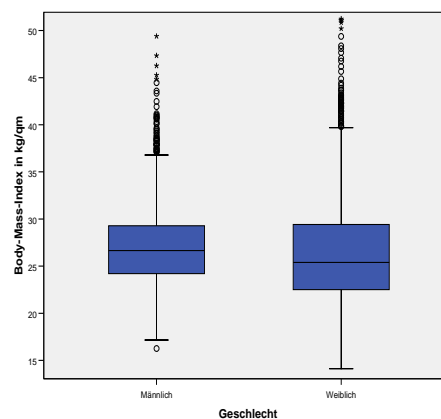
-Bei ordinalen Daten, *nur* wenn alle Kategorien vorkommen und die Zahl der Kategorien nicht zu groß ist (vgl. Bundesgesundheitsurvey, Zufriedenheit)

Bei Balkendiagrammen entspricht die Höhe der Balken der Häufigkeit, bei Histogrammen entspricht dagegen die Fläche der Rechtecke der Häufigkeit.

29

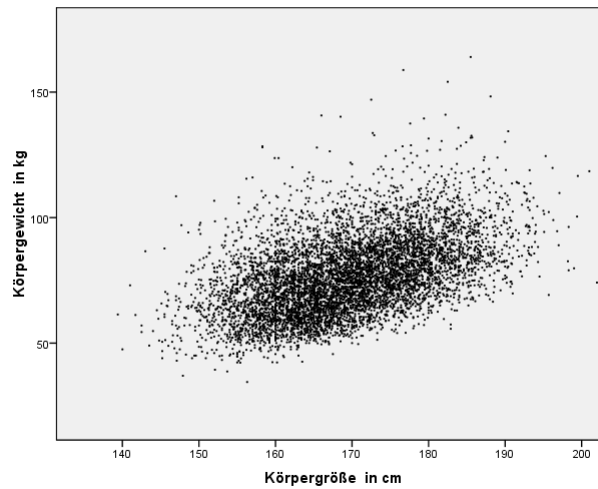
## Eine quantitative und eine kategorielle Variable Boxplot

<b>Männlich</b>	<b>N</b>	3435
	<b>Mittelwert</b>	27,0
	<b>Median</b>	26,7
	<b>Perzentile</b>	
	25	24,2
	50	26,7
	75	29,3
<b>Weiblich</b>	<b>N</b>	3637
	<b>Mittelwert</b>	26,4
	<b>Median</b>	25,4
	<b>Perzentile</b>	
	25	22,5
	50	25,4
	75	29,4



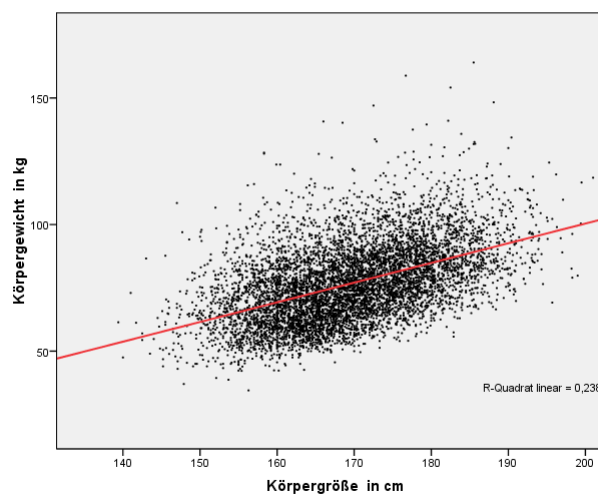
30

## Zwei quantitative Variablen Streudiagramm



31

## Zwei quantitative Variablen Streudiagramm mit Regressionsgerade

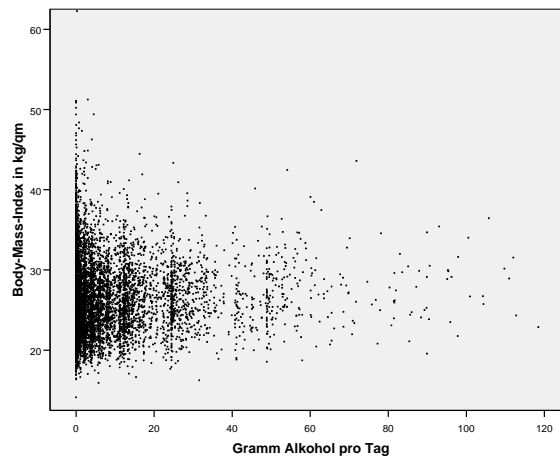


R=0.487  
(Korrelations-  
koeffizient  
nach Pearson)

32

## Alkoholkonsum und BMI

---



Wäre hier eine Regressionsgerade geeignet ???

33

## Lernziele

---

Die Studierenden sollen

den Begriff der Stichprobe definieren können,

Skalenniveaus am Beispiel erkennen und die Unterschiede verschiedener Skalenniveaus benennen können,

die wesentlichen elementaren Techniken der deskriptiven Statistik benennen können,

für eine randomisierte Therapiestudie die Stichprobenziehung beurteilen und eine deskriptive Analyse grob konzipieren können,

Scheu vor statistischer Denkweise abbauen, Interesse für das Gebiet entwickeln.

34

## Wiederholung aus Modul 1

---

Statistik für die Randomisierte Klinische Studie (RCT)

Damit beim RCT der Vergleich „fair“ ausfällt, dürfen sich die Gruppen vor Therapie nicht unterscheiden (Vermeidung von Selektionsbias).

I.a. werden bestimmte Risikopatienten ausgeschlossen. Dies verletzt zwar die Repräsentativität, wenn es für beide Gruppen gleichermaßen gilt, ist der Therapievergleich dennoch fair (Spektrumbias)

35

## Aufbau der statistischen Analyse für RCTs

---

Schritt 1

Beschreibung der Daten vor Therapie in der Gesamtstichprobe  
Repräsentativität - Spektrumbias

Schritt 2

Beschreibung der Daten vor Therapie getrennt nach Behandlung  
Vergleichbarkeit - Selektionsbias

Schritt 3

Beschreibung der Ergebnisse nach Therapie, getrennt nach Behandlung  
Ausmaß und Relevanz eventueller Unterschiede

Schritt 4

Statistische Testung der Ergebnisse  
→ nächste Vorlesung

36

**Besser nicht lügen mit Statistik - Einführung in die beschreibende Statistik**  
Fachvorlesung (90 min)

Einrichtung

CC04 - Institut für Biometrie und klinische Epidemiologie - CBF/CCM

Kurzbeschreibung

Ziel dieser Vorlesung ist es, den Studierenden ein erstes Gespür für die statistische Aufbereitung wissenschaftlicher Daten zu vermitteln. Die für alle folgenden Lehrveranstaltungen grundlegenden Begriffe (Stichprobe, Merkmalsträger, Merkmal, Skalenniveau) und Techniken (Deskription mit Hilfe von Tabellen, Grafiken, deskriptiven Parametern) werden vermittelt.



Übergeordnetes Lernziel

Die Studierenden sollen im Rahmen des eigenen wissenschaftlichen Arbeitens statistische Methoden als nützliches Werkzeug erkennen können. Es soll vermittelt werden, dass es im Umgang mit Daten nicht möglich ist, „keine Statistik zu treiben“ und dass es gute und schlechte statistische Methodiken gibt.



Lernziele

Die Studierenden sollen...

- den Begriff der Stichprobe definieren können.
- Skalenniveaus am Beispiel erkennen und die Unterschiede verschiedener Skalenniveaus benennen können.
- die wesentlichen elementaren Techniken der deskriptiven Statistik benennen können.
- ▶ für eine randomisierte Therapiestudie die Stichprobenziehung beurteilen und eine deskriptive Analyse grob konzipieren können.
- Scheu vor statistischer Denkweise abbauen, Interesse für das Gebiet wecken.

Zeitaufwand

60 Minuten für Vor- und Nachbereitung.

Lernspirale

Diese Vorlesung versteht sich als erste Einführung in statistisches Denken. Insofern sind die Ansprüche bewusst moderat gehalten. Den Studierenden sollen die elementaren Grundbegriffe statistischen Denkens - beschränkt auf deskriptive Methoden - vermittelt werden, um die Grundlage für weitergehende Veranstaltungen in diesem Modul (insbesondere der wissenschaftlichen Arbeit) aber auch den beiden weiteren Wissenschaftsmodulen zu erwerben.

Empfehlungen

Vor- und Nachbereitung:

Den Studierenden werden via Blackboard Beispiele „falscher“ und „richtiger“ statistischer Darstellungen aus wissenschaftlicher und populärer Literatur zur Einarbeitung vorgeschlagen. Die Beispiele werden kontinuierlich aktualisiert, um das Interesse der Studierenden zu wecken.

Notizen Evaluation

---