Datenauswertung

Prof. Dr. Josef Brüderl
Universität Mannheim

Frühjahrssemester 2007

Methoden-Curriculum B.A. Soziologie

Basismodul: Methoden und Statistik: 22

VL Datenerhebung (2): 5
ÜK (2): 3

VL Datenauswertung (2): 5
Ü (2): 2

VL Multivariate Analyse (2): 5
Ü (2): 2

Aufbaumodul: Methoden der empirischen Sozialforschung: 13

ÜK Datenerhebungsseminar (2): 5

ÜK Datenanalyseseminar (4): 8

Was ist Statistik?

- Statistik ist ein Teilgebiet der angewandten Mathematik
- Statistik ist ein wichtiges Hilfsmittel für die empirische Sozialforschung (Datenauswertung)
- Herkunft des Begriffs
  - Neulateinisch „statista“ etwa „Staatsmann“
  - Gottfried Achenwall (1749) Staatsverfassung der europäischen Reiche. Statistik als Lehre der „Staatsmerkwürdigkeiten“.
- Die zwei Bedeutungen
  - Sammlung numerischer Informationen über Tatbestände (amtliche Statistik)
  - Verfahren zur Auswertung numerischer Daten
    - Informationsgewinnung (explorative Statistik)
    - Informationsreduktion (deskriptive Statistik)
    - Verallgemeinerung (induktive Statistik, Inferenzstatistik)

Der Forschungsprozess

- Problemfindung und -präzisierung
- Theoriebildung (Hypothesen)
- Wahl des Forschungsdesigns

  Design, Erhebungsverfahren, Stichprobe

VL Datenerhebung

- Operationalisierung (Indikatoren, Erhebungsinstrument)
- Datenerhebung

  Pretest, Schulung der Erhebenden, Feldarbeit (intern/extern)

- Datenerfassung

  Codieren, Übertragen auf Datenträger, Bereinigen

- Datenanalyse

  Grundauszählung, Datenaufbereitung, Hypothesentests

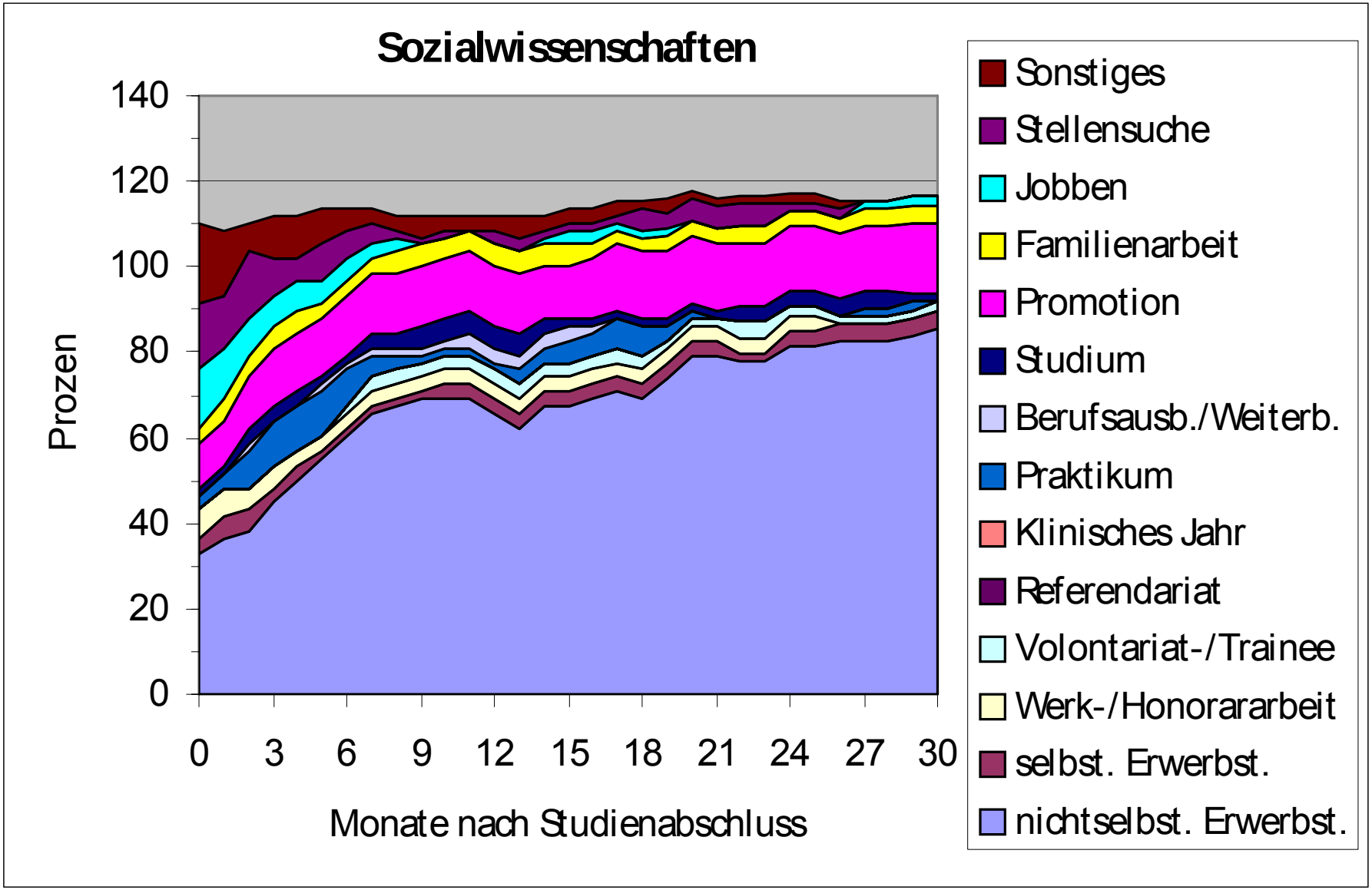
VL Datenauswertung

- Publikation

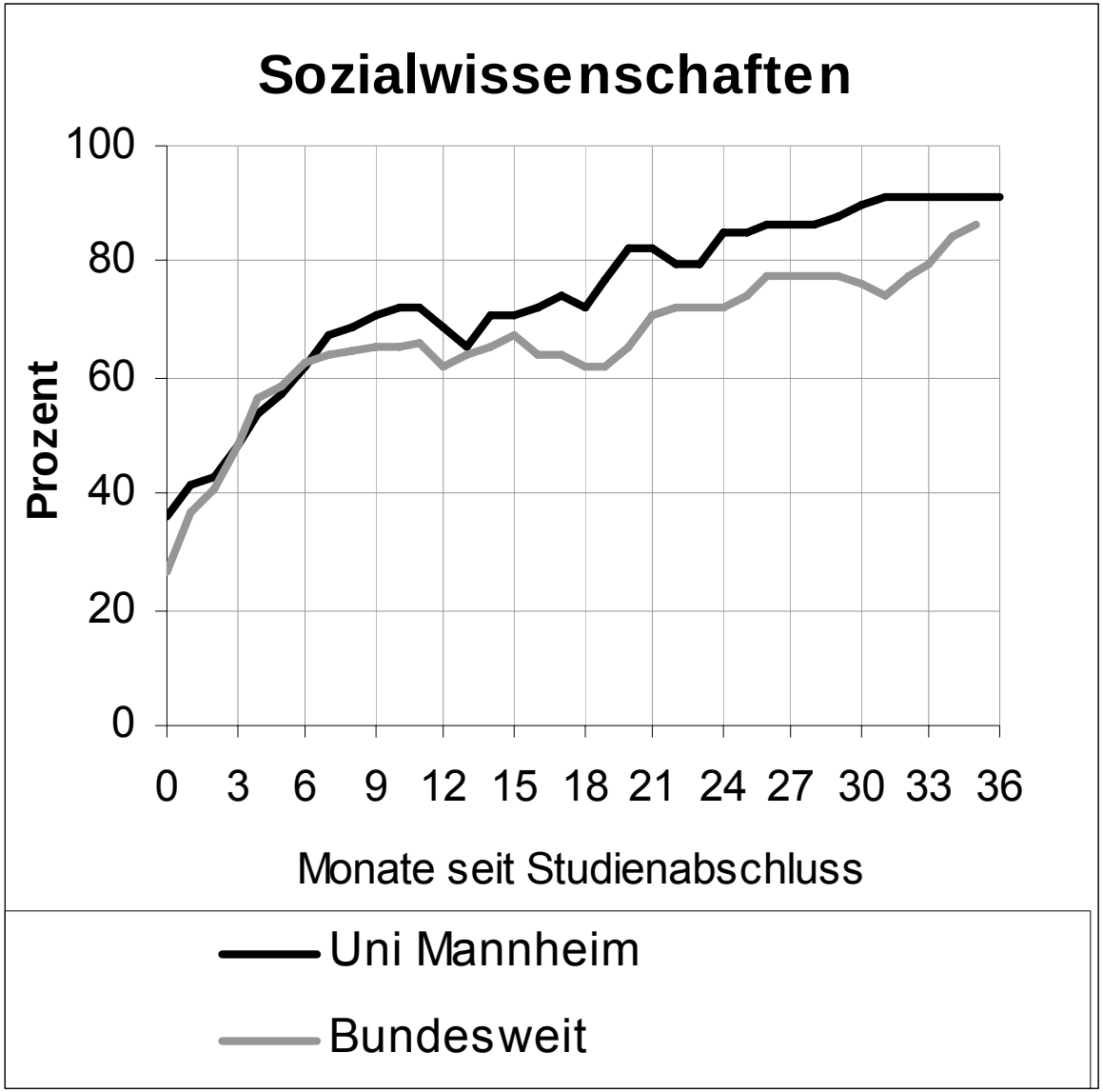
Mannheimer Absolventenstudie 2003

- Problemfindung: Was wird aus unseren Absolventen?
- Theorie: Na ja
- Forschungsdesign
  - Retrospektive Befragung über Berufsverlauf
  - Schriftliche Befragung (Kosten!)
  - Grundgesamtheit: Absolventen WS 1997/98 bis SS 2001 (N=4200)
- Operationalisierung
  - Z.B. Erfolg (Suchdauer, Einkommen, Fachadäquanz, Zufriedenheit)
  - Fragebogenentwicklung
- Datenerhebung
  - Adressrecherche (2900 von 4200 = 68 %)
  - Versand der Fragebögen (Juni 2003)
  - Rücklaufkontrolle (1400 von 2900 = 48 %; Ausschöpfungsquote = 33 %)
- Datenerfassung: durch externes Institut
- Datenauswertung: 1 Jahr
- Bericht: Dezember 2004 im Internet

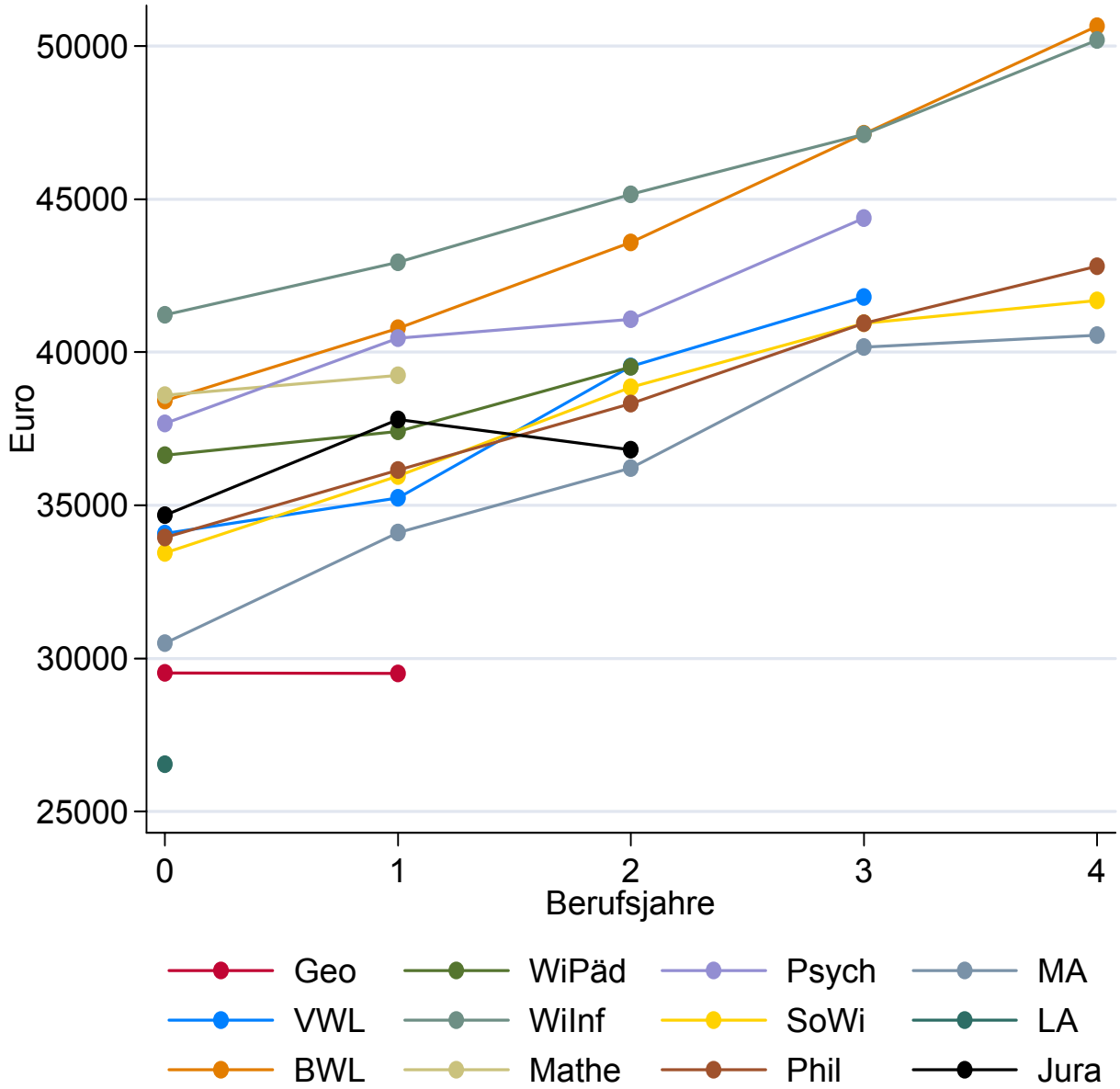
Die ersten 30 Monate: Dipl Sowi



Reguläre Erwerbsarbeit: Dipl Sowi



Bruttojahreseinkommen



Grundlegende Begriffe

- Grundgesamtheit und Stichprobe
- Untersuchungseinheit
  - Fall, „case“
- Merkmal
  - Variable, „variable“
- Merkmalsausprägung
  - (möglicher) Wert, „value“
- Realisation
  - (beobachteter) Wert, „code“
- Messen
  - Zuordnung von Werten zu den Fällen mit Messinstrument

Die Datenmatrix

	Variable 1	Variable 2
Fall 1	Wert von Fall 1 auf Variable 1	Wert von Fall 1 auf Variable 2
Fall 2	Wert von Fall 2 auf Variable 1	Wert von Fall 2 auf Variable 2
Fall 3	Wert von Fall 3 auf Variable 1	Wert von Fall 3 auf Variable 2
Fall 4	Wert von Fall 4 auf Variable 1	Wert von Fall 4 auf Variable 2
Fall 5	Wert von Fall 5 auf Variable 1	Wert von Fall 5 auf Variable 2

Fallnr. ID	Geburtsjahr X	Einkommen Y
1	1960	2400
2	1956	1898
3	1968	4000
4	1956	3450
5	1960	2050

Skalenniveaus

- Nominalskala: Äquivalenzrelation (gleich, ungleich)
  - Geschlecht, Familienstand, Beruf, Partei
  - Zulässige Berechnungen: auszählen
- Ordinalskala: plus Ordnungsrelation (größer, kleiner)
  - Schicht, Schulnoten, Psycho-Skalen
  - Zulässige Berechnungen: auszählen, ordnen
- Intervallskala: zusätzlich Abstände definiert
  - Geburtsjahr, Schulnoten (?), Psycho-Skalen (?)
  - Zulässige Berechnungen: auszählen, ordnen, Differenz
- Ratioskala: zusätzlich Nullpunkt definiert
  - Alter, Einkommen, Schuldbildung
  - Zulässige Berechnungen: auszählen, ordnen, Differenz, Verhältnis

Weitere Begriffe

- Kategoriale Variablen
  - Nominal- bzw. ordinalskalierte Variablen
- Metrische Variablen
  - Intervall- bzw. ratioskalierte Variablen
- Stetige und diskrete Variablen
  - Diskret: nicht alle Zahlenwerte im Wertebereich möglich
    - Schulbildung, Kinderzahl
  - Stetig: alle Zahlenwerte im Wertebereich möglich
    - Alter, Einkommen (eigentlich nur „quasi“ stetig)
- Gruppierte Daten
  - Durch Klassenbildung aus einer (diskreten oder) stetigen Variable
  - Zweck: Informationsreduktion
  - Gruppierte Variable: diskret und ordinalskaliert

Kapitel II
Univariate Datenanalyse

Prof. Dr. Josef Brüderl
Universität Mannheim

Frühjahrssemester 2007

Notation

- n die Anzahl der Untersuchungseinheiten
- X eine Variable
- i eine einzelne, aber keine bestimmte Untersuchungseinheit ($i \in \{1, ..., n\}$),
- x_i der Wert der Variable X für die Untersuchungseinheit i

- $x_1, \ldots, x_i, ..., x_n$ die (Roh-) Daten

- k die Anzahl der Ausprägungen ($k \leq n$)
- $a_1 < a_2 < ... < a_k$ die in den Daten vorkommenden Ausprägungen

Häufigkeitsverteilungen

- $h(a_j)$ bzw. h_j die <u>absolute</u> Häufigkeit der Ausprägung a_j, d.h. die Anzahl der x_i aus $x_1,...,x_n$ mit $x_i = a_j$ ($j \in \{1,...,k\}$)

- $f(a_j)$ bzw. f_j die <u>relative</u> Häufigkeit der Ausprägung a_j, d.h. $f_j = \frac{h_j}{n}$

- $F(a_j)$ bzw. F_j die <u>kumulierte relative</u> Häufigkeit der Ausprägung a_j, d.h.

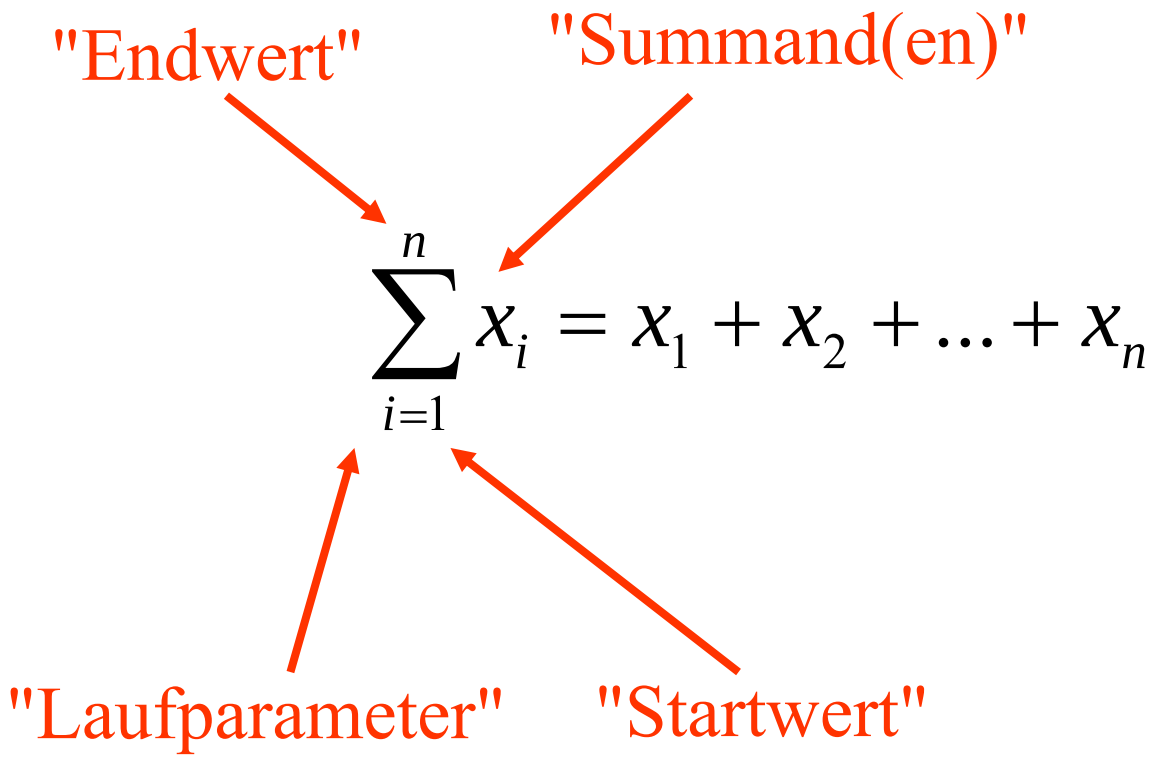
$$F_j = \sum_{l=1}^{j} f_l$$

Häufigkeitstabelle

- Eine Häufigkeitstabelle (frequencies) listet die absolute, relative und kumulierte Häufigkeitsverteilung auf

a_1	h_1	f_1	F_1
a_2	h_2	f_2	F_2
...	...	...	...
a_k	h_k	f_k	F_k

Exkurs: Zum Gebrauch des Summenzeichens



Exkurs: Einige Regeln

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \ldots + x_n^2$$

$$\left(\sum_{i=1}^{n} x_i\right)^2 = (x_1 + x_2 + \ldots + x_n)^2$$

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots x_n y_n$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} (x_i + y_i)^2 = \sum x_i^2 + \sum 2 x_i y_i + \sum y_i^2$$

$$\sum_{i=1}^{n} k x_i = k \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} k = nk$$

$$\sum_{j=1}^{n} y_i x_j = y_i \sum_{j=1}^{n} x_j$$

$$\sum_{i=1}^{2} \sum_{j=1}^{3} x_i y_j = \left(\sum_{i=1}^{2} x_i\right)\left(\sum_{j=1}^{3} y_j\right) = (x_1 + x_2)(y_1 + y_2 + y_3)$$

Beispiel: Mathenote

Die Rohdaten (Erstsemesterbefragung Jg. 2001/02):

```
2  3  4  4  2  3  4  3  3  4  4  4  5  3  3  2  2
4  2  4  2  5  3  1  3  3  3  3  3  2  3  4  4  3
3  3  5  3  3  4  3  .  5  4  3  4  3  4  3  3  2
3  4  3  4  3  .  2  3  2  4  1  3  4  2  2  2  3
3  3  5  3  .  2  2  4  2  5  4  4  4  3  4  4  4
2  2  3  3  3  3  4  .  .  3  3  3  3  4  3  5  4
4  5  4  4  4  2  4  3  4  4  5  5  4  .  5  3  4
4  3  2  2  4  2  3  4  3  4  3  2  4  5  4  5  3
2  5  5  3  2  4  5  3  3  2  5  5  4  5  3  4  4
1  3  2  4  3  1  3  3  3  4  3  .  4  2  .  3  3
3  4  4  5  3  5  4  4  4  2  4  3  4  .  3  5  2
4  5  4  3  4  5
```

Beispiel: Mathenote

- Häufigkeitstabelle
- n = 184 (193)
- k = 5

a_j h_j f_j F_j

mathenote oberstufe	Freq.	Percent	Cum.
1	4	2.17	2.17
2	30	16.30	18.48
3	66	35.87	54.35
4	60	32.61	86.96
5	24	13.04	100.00
Total	184	100.00	

Beispiel: Mathenote

Balkendiagramm (Jg. 1998/99)

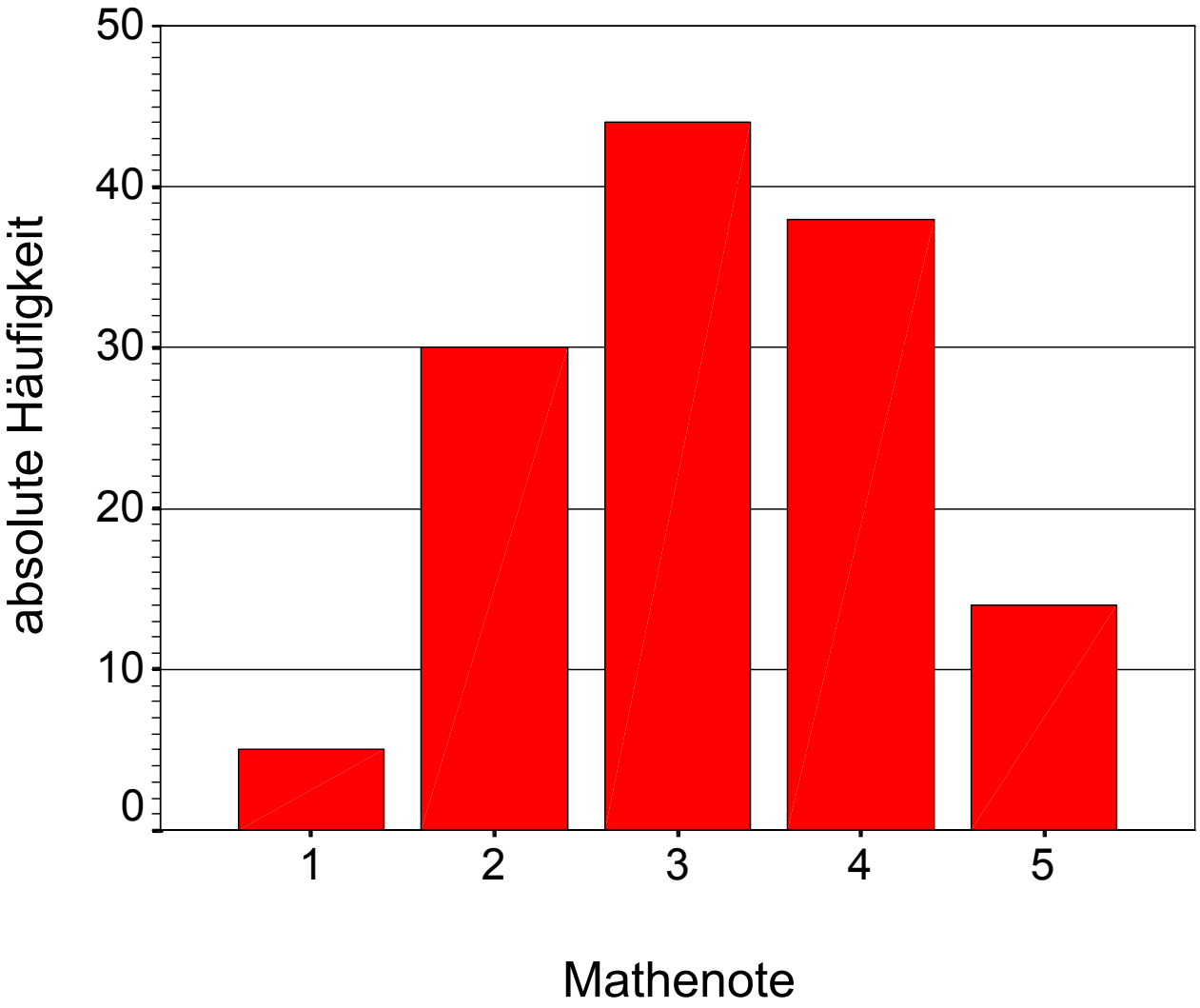


Chart-Junk !

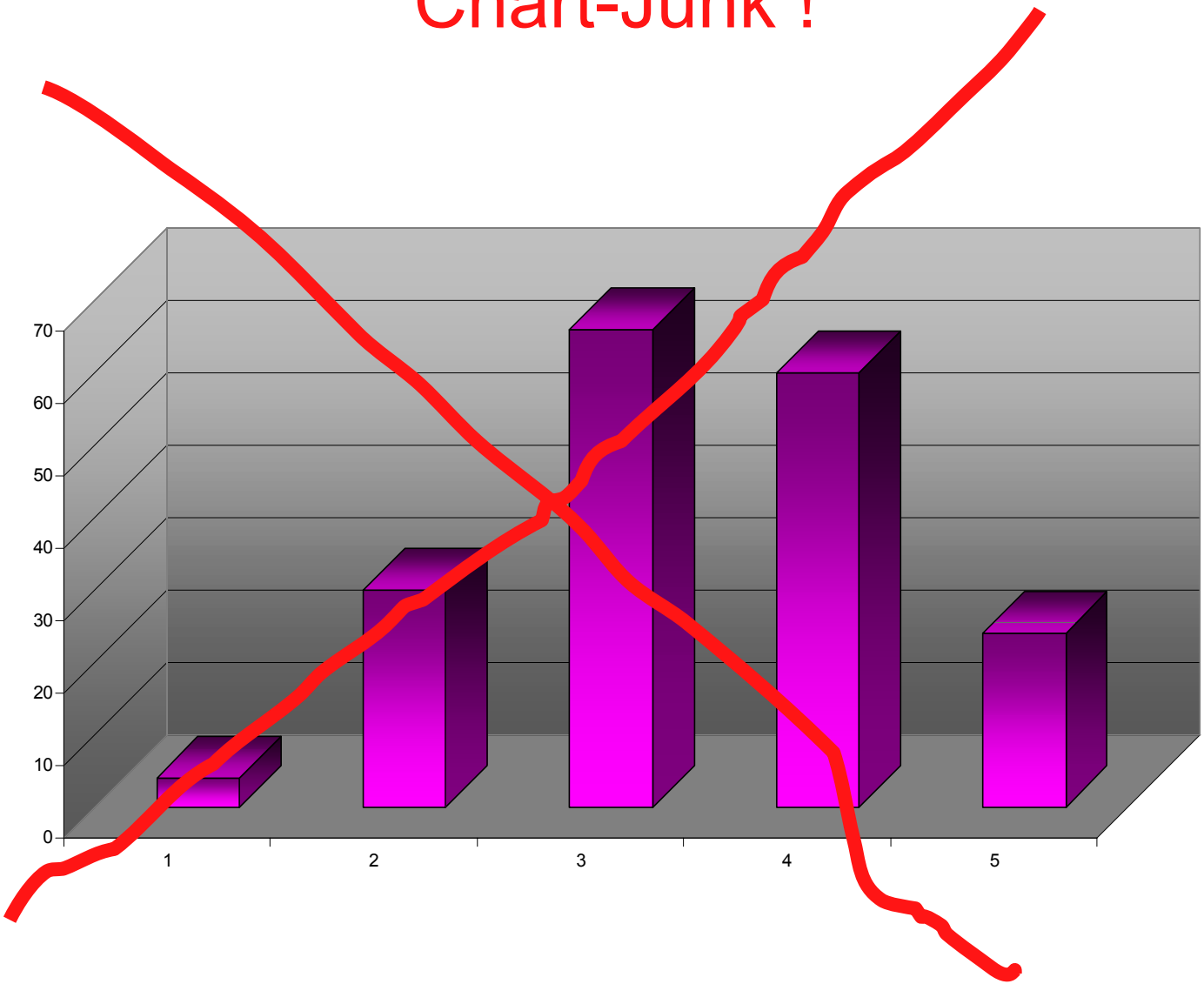
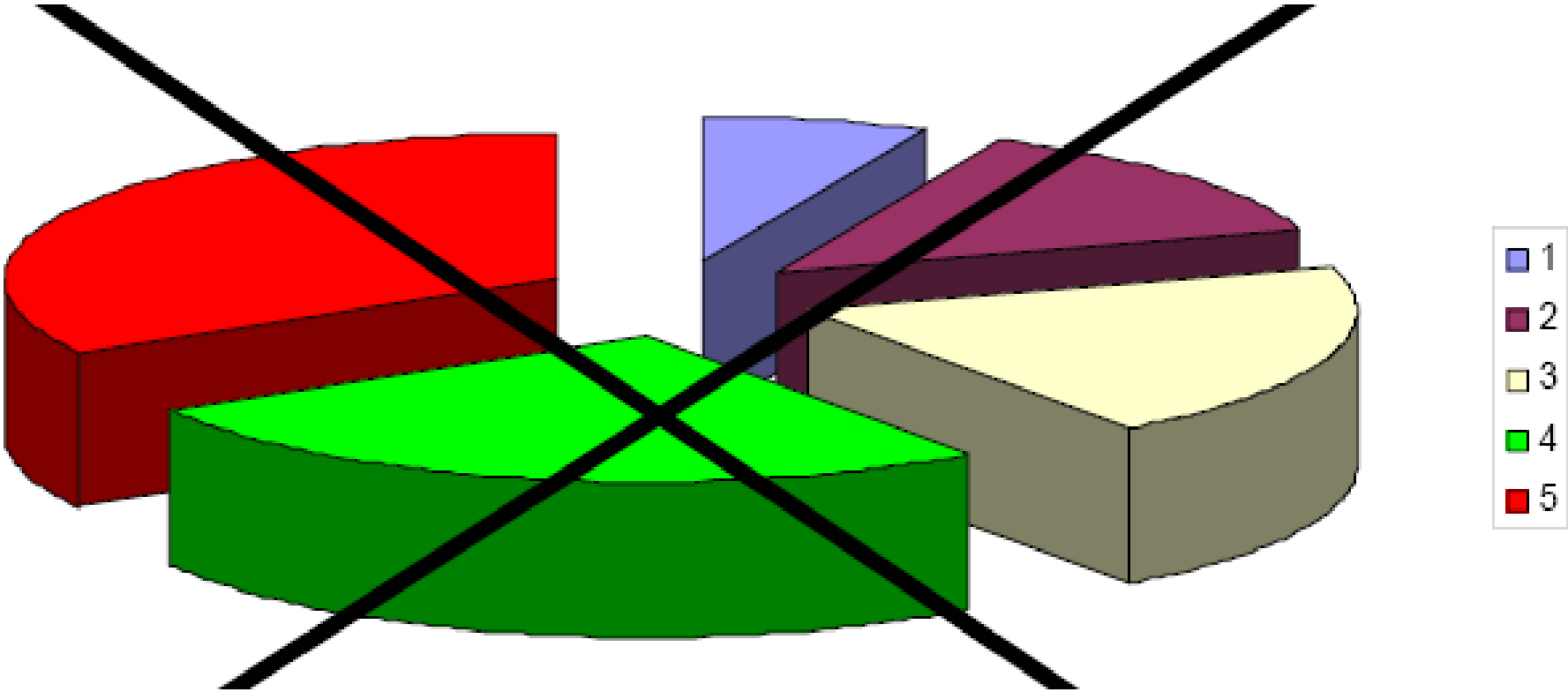


Chart-Junk !

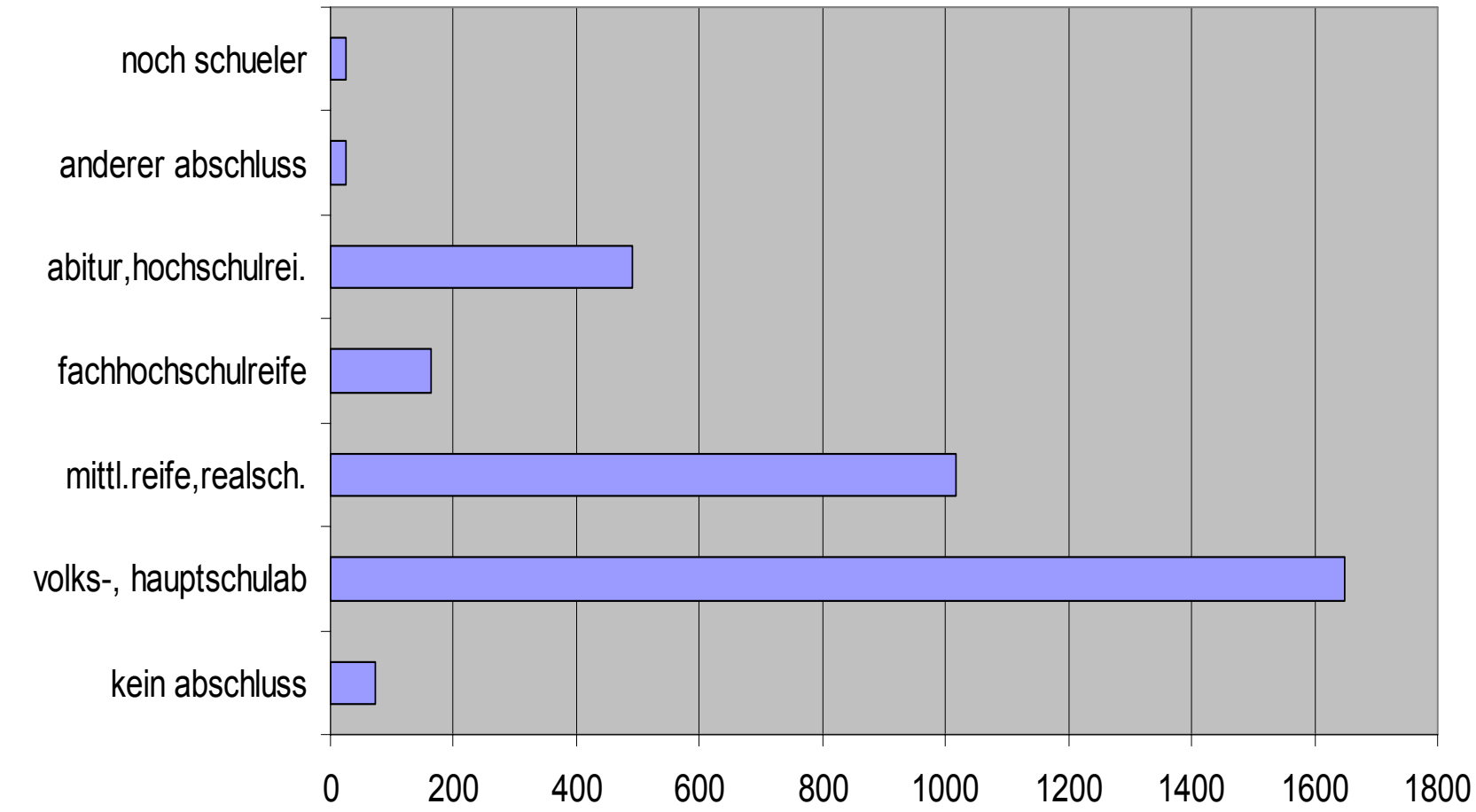


Beispiel: Schulabschluss im Allbus 1994

allgemeiner schulabschluss	Freq.	Percent	Cum.
kein abschluss	74	2.15	2.15
volks-, hauptschulab	1650	47.90	50.04
mittl.reife,realsch.	1017	29.52	79.56
fachhochschulreife	164	4.76	84.33
abitur,hochschulrei.	490	14.22	98.55
anderer abschluss	24	0.70	99.25
noch schueler	26	0.75	100.00
Total	3445	100.00	

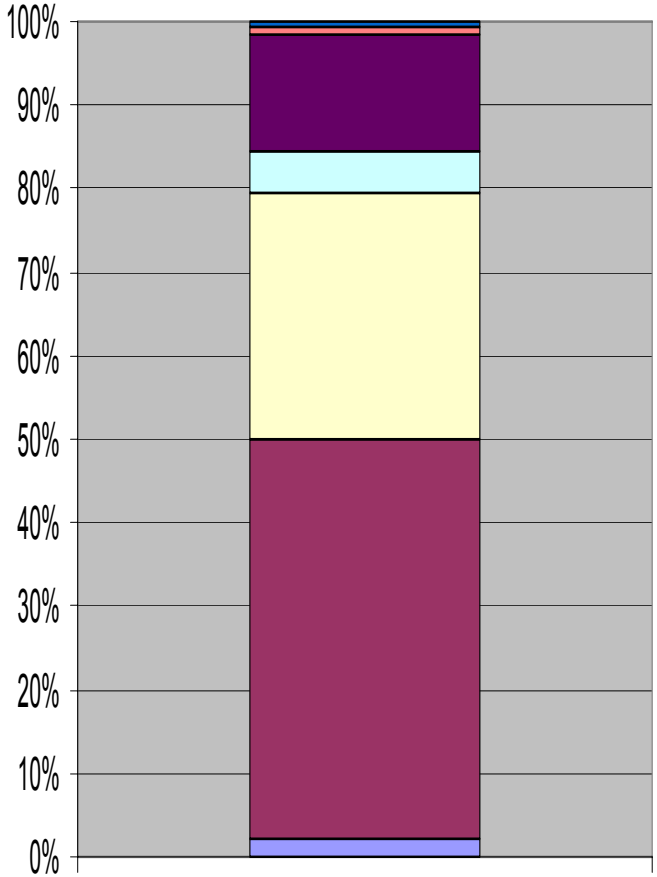
Beispiel: Schulabschluss im Allbus 1994

Balkendiagramm (Querformat)

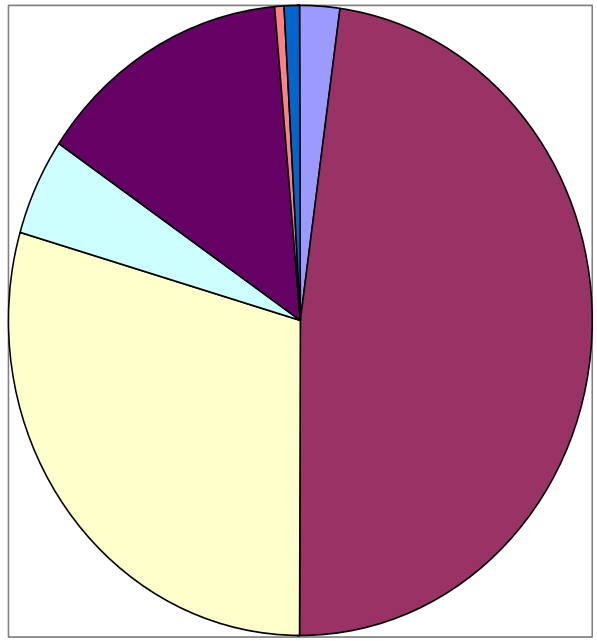


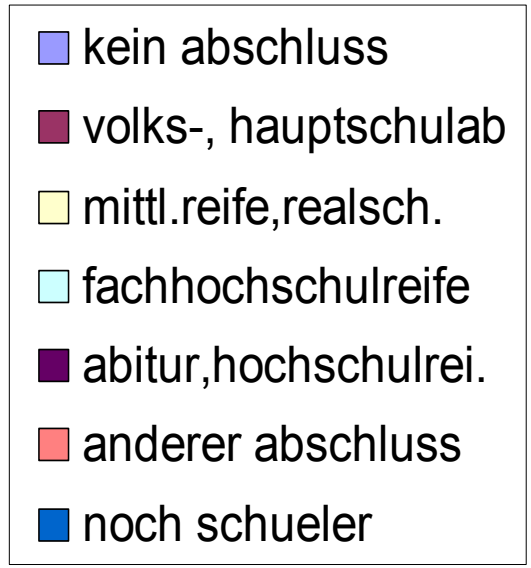
Beispiel: Schulabschluss im Allbus 1994

Säulendiagramm



Kreisdiagramm





Bsp.: Abiturnote Erstsemesterbefr. Jg. 2001/02

Problem: zu viele Ausprägungen

abinote	Freq.	Percent	Cum.
1.1	1	0.54	0.54
1.2	1	0.54	1.08
1.5	1	0.54	1.61
1.6	2	1.08	2.69
1.7	2	1.08	3.76
1.8	5	2.69	6.45
1.9	5	2.69	9.14
2	8	4.30	13.44
2.1	5	2.69	16.13
2.2	6	3.23	19.35
2.3	14	7.53	26.88
2.4	6	3.23	30.11
2.5	7	3.76	33.87
2.6	6	3.23	37.10
2.7	16	8.60	45.70
2.8	18	9.68	55.38
2.9	9	4.84	60.22
3	19	10.22	70.43
3.1	10	5.38	75.81
3.2	12	6.45	82.26
3.3	13	6.99	89.25
3.4	9	4.84	94.09
3.5	8	4.30	98.39
3.6	1	0.54	98.92
3.7	2	1.08	100.00
Total	186	100.00	

Bsp.: Abiturnote Erstsemesterbefr. Jg. 2001/02

Lösung: Gruppierung der Daten

abitkat	Freq.	Percent	Cum.
1.0 - 1.4	2	1.08	1.08
1.5 - 1.9	15	8.06	9.14
2.0 - 2.4	39	20.97	30.11
2.5 - 2.9	56	30.11	60.22
3.0 - 3.4	63	33.87	94.09
3.5 - 3.9	11	5.91	100.00
Total	186	100.00	

Bsp.: Abiturnote Erstsemesterbefr. Jg. 2001/02

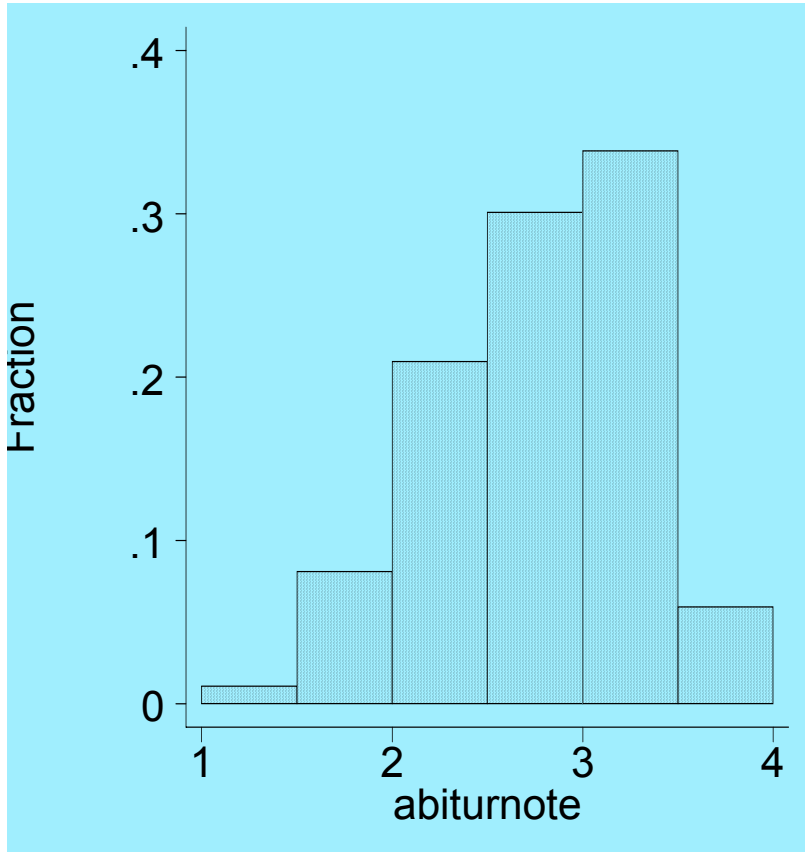
Graphische Lösung: Histogramm

- Bilde Gruppierung in k Klassen, d.h. $[c_0,c_1),[c_1,c_2),...,[c_{k-1},c_k)$
- Zeichne Rechtecke mit:

  Breite: $d_j = c_j - c_{j-1}$

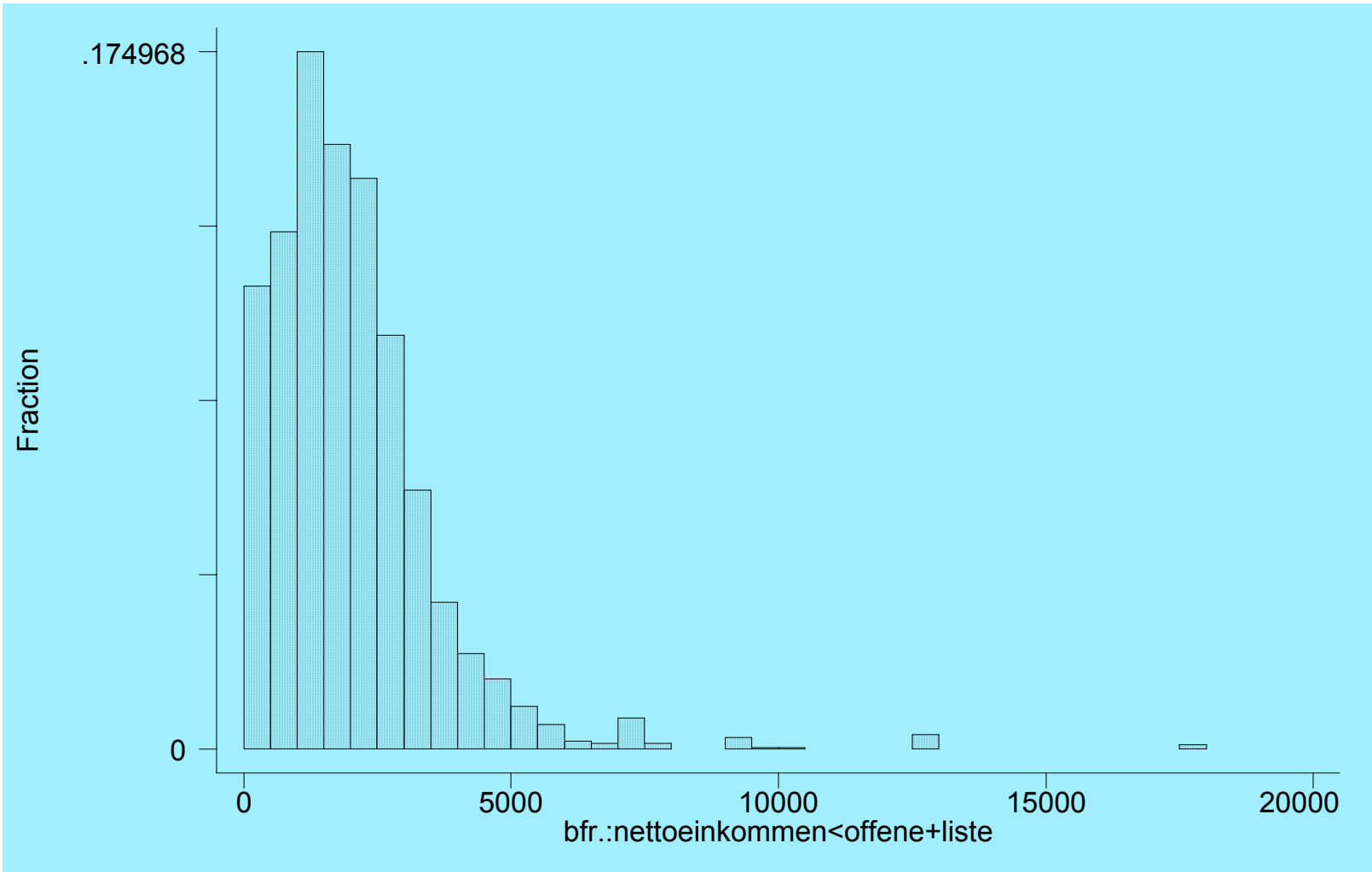
  Höhe $= h_j/d_j$ bzw. f_j/d_j

  Fläche $= h_j$ bzw. f_j
- Prinzip der Flächentreue! Bei verscheiden breiten Klassen ist die Höhe der Balken schwer interpretierbar (Dichte)
- Deshalb: verwende gleich breite Klassen (wie im Bsp. nebenan) und trage h_j bzw. f_j ab



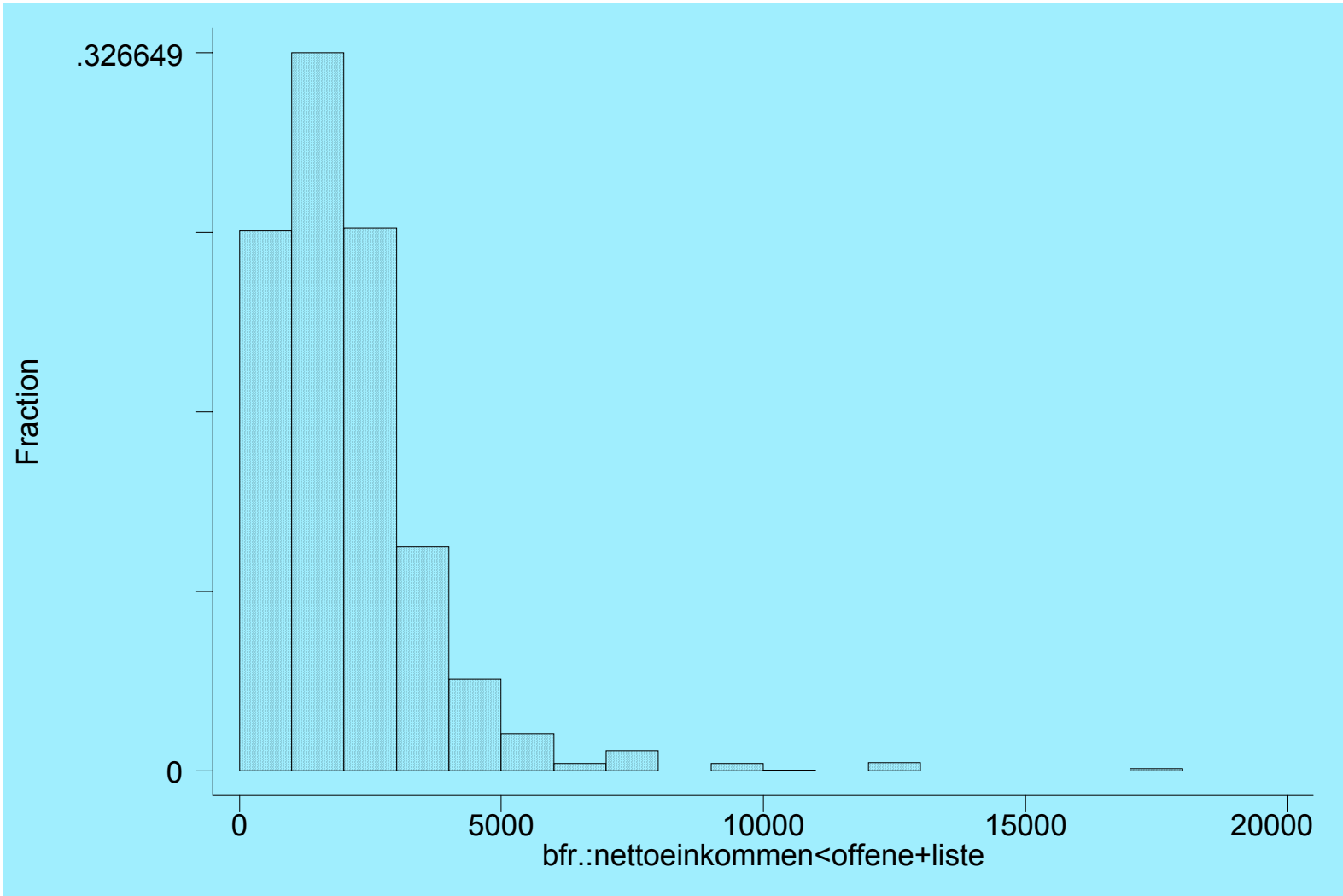
Beispiel: Einkommen im Allbus 1994

Monatliches Netto-Einkommen in DM; hier 40 Klassen



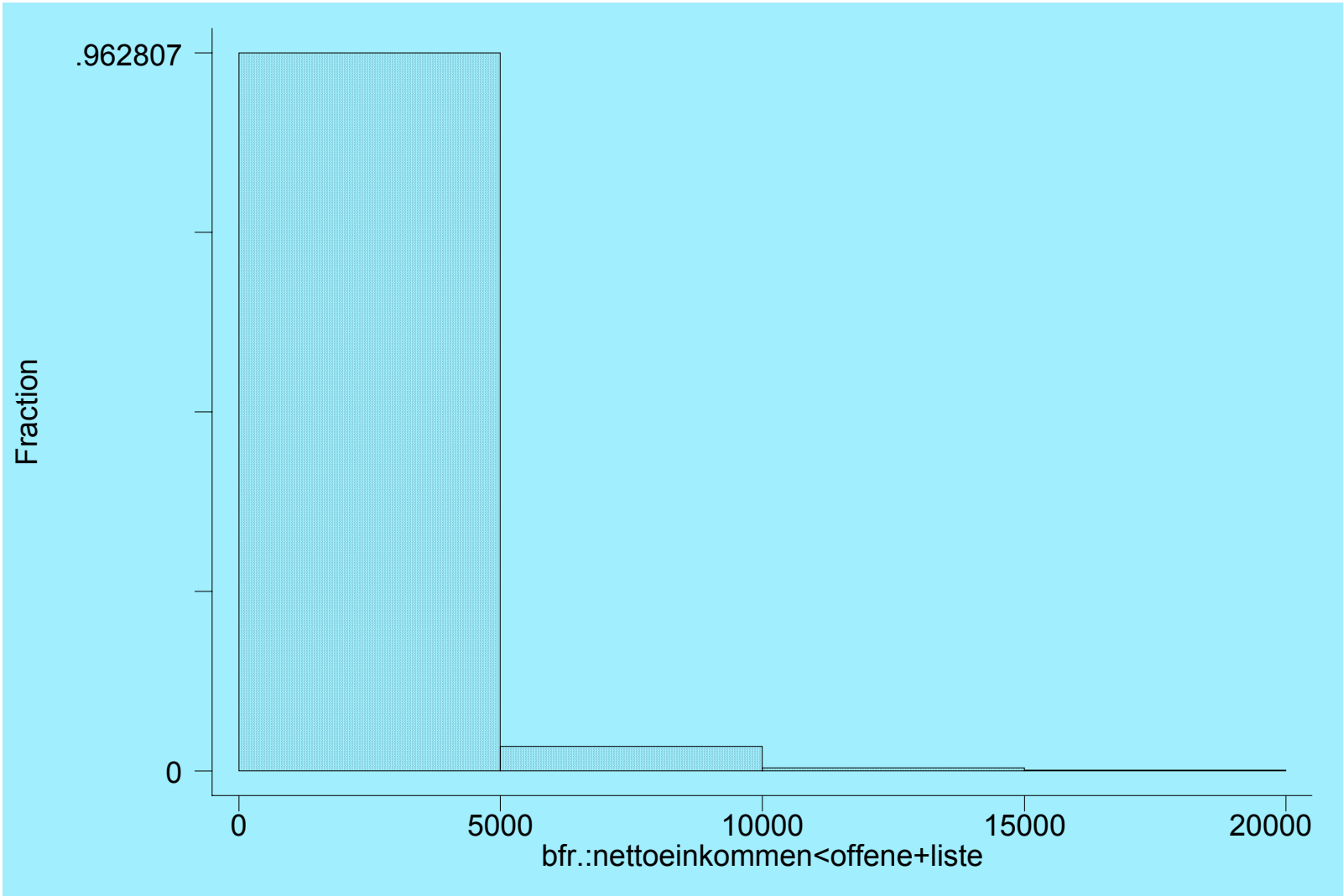
Beispiel: Einkommen im Allbus 1994

Monatliches Netto-Einkommen in DM; hier 20 Klassen



Beispiel: Einkommen im Allbus 1994

Monatliches Netto-Einkommen in DM; hier 4 Klassen



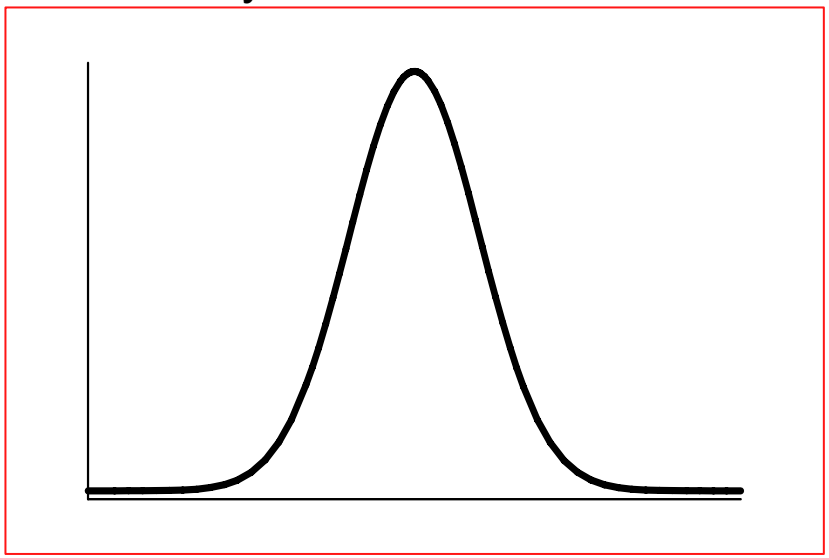
Verteilungstypen

- Gipfel: unimodal $\leftrightarrow$ bimodal $\leftrightarrow$ multimodal
  - ein (zwei, mehrere) 'deutliche(r)' Gipfel
- Symmetrie: symmetrisch $\leftrightarrow$ asymmetrisch
  - es gibt eine Spiegelachse und beide Hälften sind 'annähernd' gleich
- Schiefe: linksschief $\leftrightarrow$ rechtsschief
  - Daten sind rechtsseitig oder linksseitig konzentriert

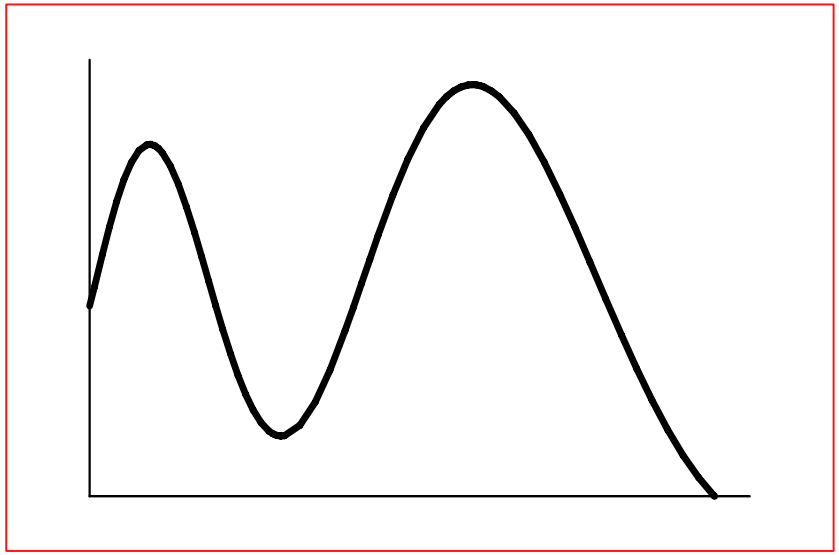
!Vorsicht: 'unscharfe' Kriterien, genauere Kriterien folgen!

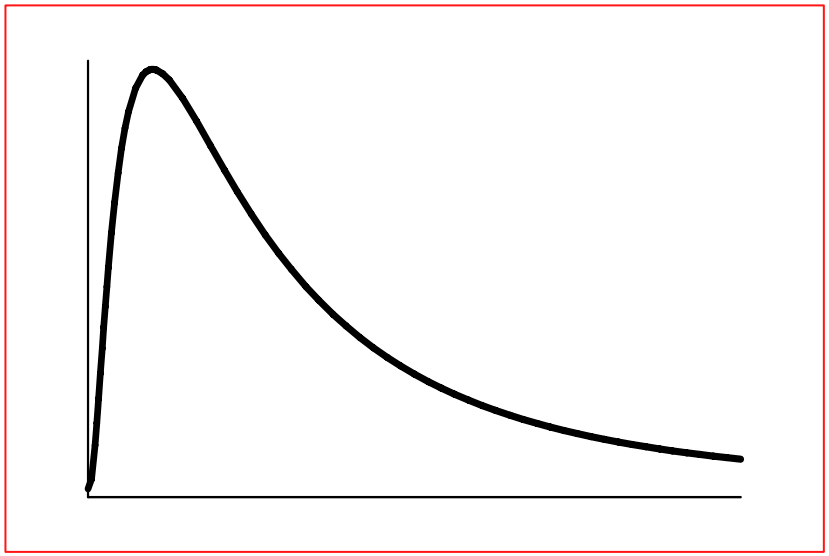
Verteilungstypen

unimodal-symmetrisch

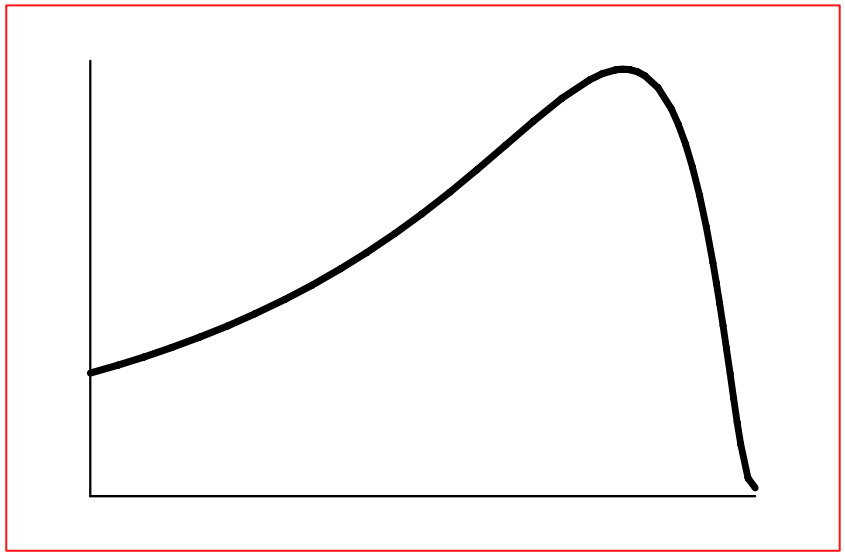


bimodal-asymmetrisch





rechtsschief-asymmetrisch



linksschief-asymmetrisch

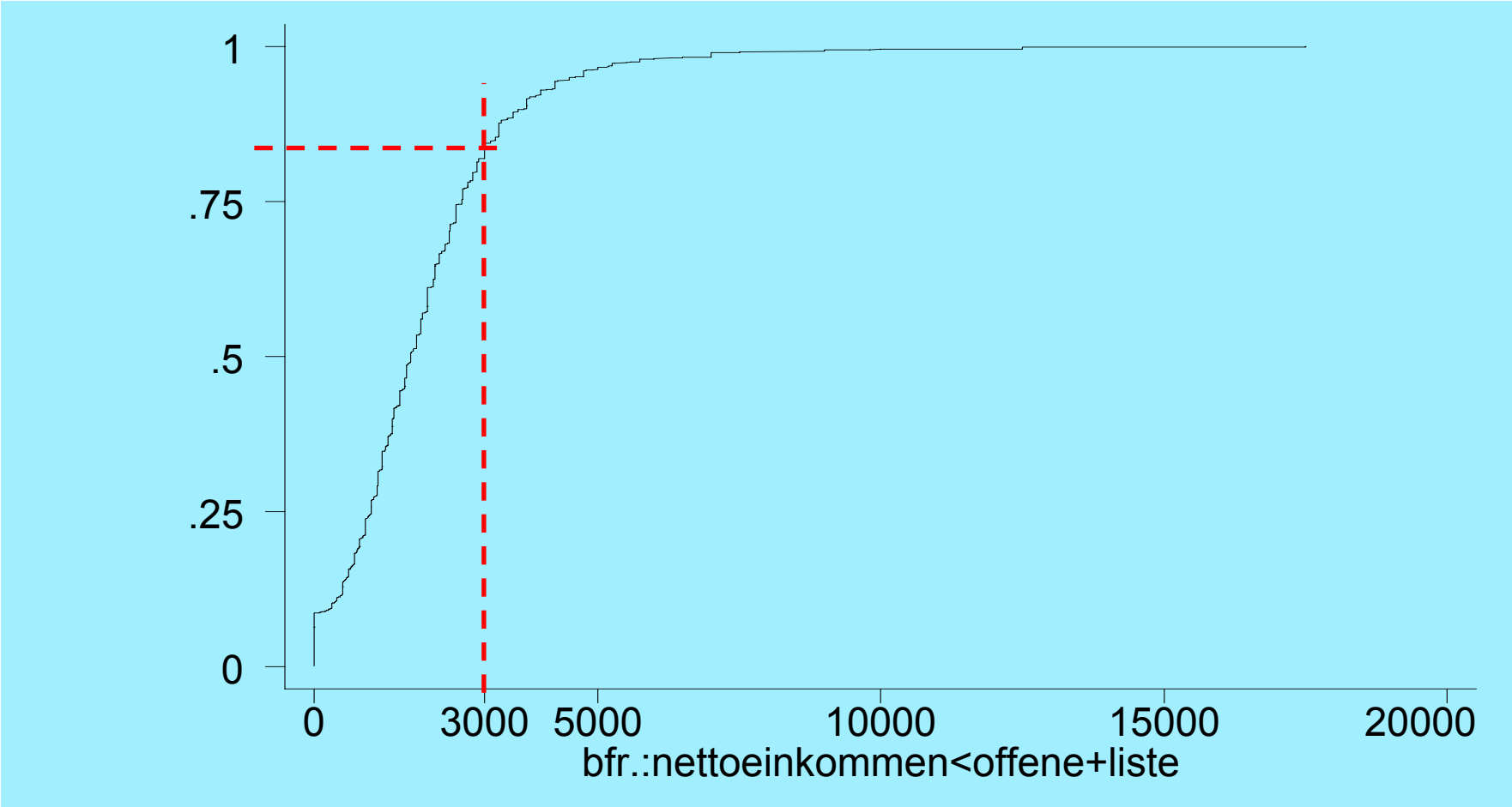
Empirische Verteilungsfunktion

- Auch kumulierte Häufigkeitsverteilung genannt
- Gibt für jeden Wert der Verteilung einer ordinalen oder metrischen Variablen den Anteil der Fälle an, deren Ausprägungen kleiner oder gleich diesem Wert sind

$$F(x) = F(a_j), \ \ falls \ a_j \leq x < a_{j+1}$$

Zur Erinnerung: $F(a_j) = \sum_{l=1}^{j} f(a_l)$

Beispiel: Einkommen im Allbus 1994



Exkurs: Achte auf die Untersuchungseinheit

UE: Haushalt

HHgrösse	Freq.	Percent
1	50	50.00
2	20	20.00
3	10	10.00
4	20	20.00
Total	100	100.00

Zeitungsmeldung:
Hälfte aller Deutschen leben Single

UE: Person

HHgrösse	Freq.	Percent
1	50	25.00
2	40	20.00
3	30	15.00
4	80	40.00
Total	200	100.00

Ist aber falsch!
Ein Viertel aller Deutschen lebt Single.

Lagemaße

- Parameter bzw. Kennwerte zur Beschreibung des Zentrums (der "Lage") einer Häufigkeitsverteilung

- Die wichtigsten Lageparameter sind:
  - Der Modus (Modalwert)
  - Der Median (Zentralwert)
  - Das arithmetische Mittel (Mittelwert)

Der Modus

- Der Modus x_{MOD} ist die Ausprägung mit der größten Häufigkeit, d.h. $x_{MOD} = a_j$ mit $h(a_j) = \max\{h(a_l) \mid l = 1, ..., k\}$
- Der Modus ist eindeutig, falls die Häufigkeitsverteilung ein eindeutiges Maximum besitzt

- Beispiel 1: Urliste 2, 4, 5, 4, 3, 1, 3, 4, 2, 5

  $X_{MOD} = 4$

- Beispiel 2: Studierende der Statistik I, WS 01/02

studiengang	Freq.	Percent	Cum.
diplom sozialw.	146	76.44	76.44
magister soziologie	34	17.80	94.24
magister erzieh.w.	10	5.24	99.48
sonstiges	1	0.52	100.00
Total	191	100.00	

Berechnung mit gruppierten Daten

- Klassenmitte der häufigsten Klasse

abitkat	Freq.	Percent	Cum.
1.0 - 1.4	2	1.08	1.08
1.5 - 1.9	15	8.06	9.14
2.0 - 2.4	39	20.97	30.11
2.5 - 2.9	56	30.11	60.22
3.0 - 3.4	63	33.87	94.09
3.5 - 3.9	11	5.91	100.00
Total	186	100.00	

$x_{MOD} = 3{,}2$

Zum Vergleich: aus den ungruppierten Daten: $x_{MOD} = 3{,}0$
(s. Folie 45)

Eigenschaften des Modus

- Berechenbar schon ab Nominalskalenniveau
- Problematisch bei
  - bi- und multimodalen Verteilungen
  - allg. bei sehr vielen, ähnlich besetzten (dünn besetzten) Kategorien
  - insb. bei stetigen Merkmalen
- Qualitätseigenschaft: Der Modus maximiert die Summe der Indikatorfunktionen ("meiste Treffer"), d.h.

$$\sum_{i=1}^{n} I(x_i, x_{MOD}) \geq \sum_{i=1}^{n} I(x_i, z) \quad \textit{für alle } z$$

$$\textit{wobei } I(a,b) = 1 \textit{, falls } a = b, \textit{ und } I(a,b) = 0, \textit{ falls } a \neq b$$

Der Median

- Der Median x_{MED} ist die mittlere Beobachtung der geordneten (!) Urliste, d.h.

$$x_{MED} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \textit{für } n \textit{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \textit{für } n \textit{ gerade} \end{cases}$$

$$\textit{auch}:$$

$$\tilde{x} = x_{MED}$$

Median: Beispiele

- Beispiel 1: Gegeben sei folgende Urliste

  2, 4, 5, 4, 3, 1, 3, 4, 2, 5

  geordnet:

  1, 2, 2, 3, 3, 4, 4, 4, 5, 5

  $X_{MED} = 3.5$

- Beispiel 2: Mathenoten

$X_{MED} = 3$

mathenote oberstufe	Freq.	Percent	Cum.
1	4	2.17	2.17
2	30	16.30	18.48
3	66	35.87	54.35
4	60	32.61	86.96
5	24	13.04	100.00
Total	184	100.00	

Berechnung bei gruppierten Daten

- Ist $[c_{i-1},c_i]$ die 'Einfallsklasse' des mittleren Wertes, so interpoliere x_{MED} nach:

$$x_{MED} = c_{i-1} + d_i \frac{0.5 - F(c_{i-1})}{f_i}$$

Beispiel: Abiturnote

abitkat	Freq.	Percent	Cum.
1.0 - 1.4	2	1.08	1.08
1.5 - 1.9	15	8.06	9.14
2.0 - 2.4	39	20.97	30.11
2.5 - 2.9	56	30.11	60.22
3.0 - 3.4	63	33.87	94.09
3.5 - 3.9	11	5.91	100.00
Total	186	100.00	

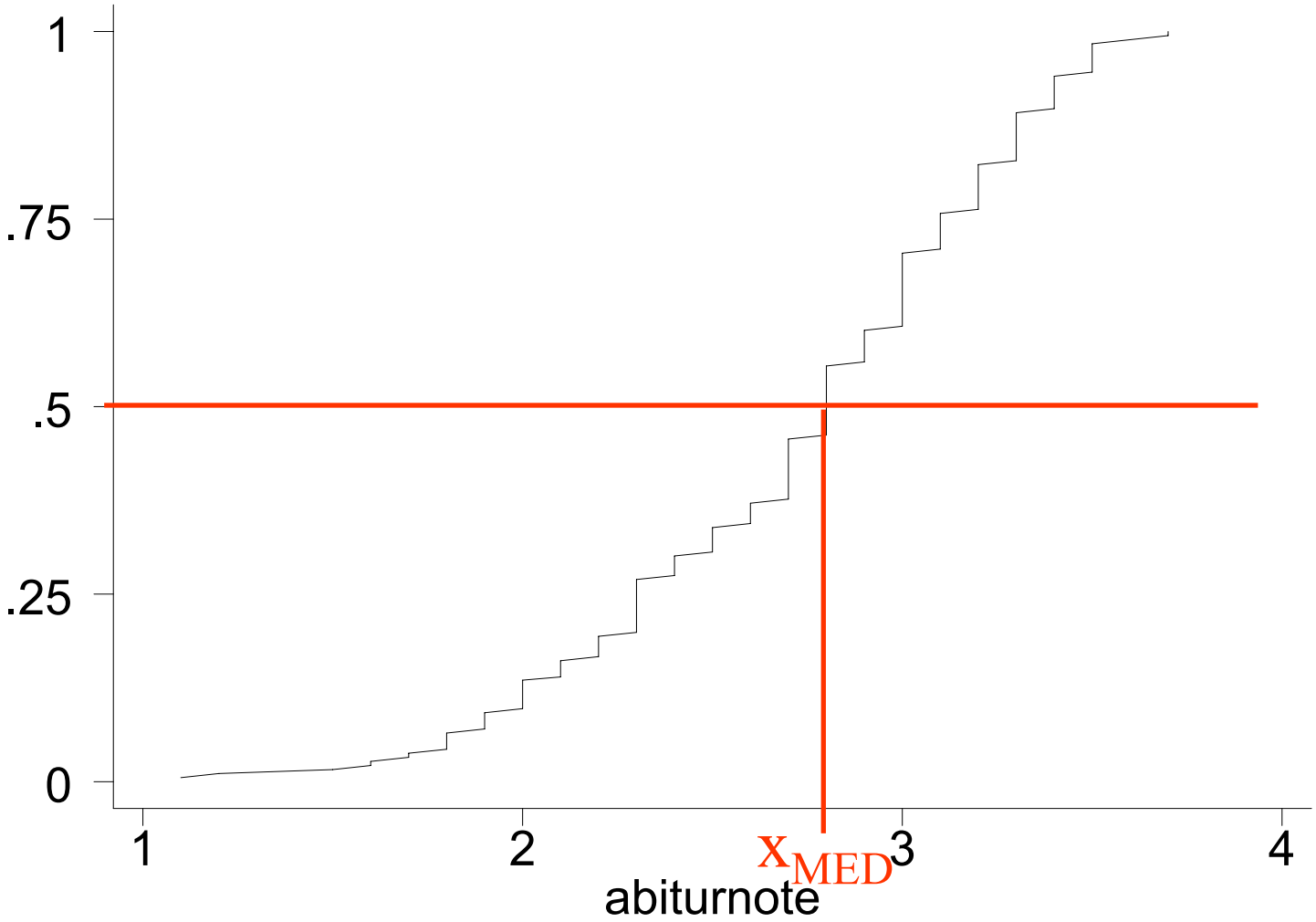
- Einfallsklasse $[c_{i-1}, c_i] = [2.5, 2.9]$, $d_i = 0.4$
  $F(c_{i-1}) = F(2.5) = 30.11$, $f_i = 30.11$

$$x_{MED} = 2.5 + 0.4 \cdot \frac{0.5 - 0.3011}{0.3011} = 2.76$$

Zum Vergleich: Abiturnote ungruppiert

abinote	Freq.	Percent	Cum.	
1.1	1	0.54	0.54	
1.2	1	0.54	1.08	
1.5	1	0.54	1.61	
1.6	2	1.08	2.69	
1.7	2	1.08	3.76	
1.8	5	2.69	6.45	
1.9	5	2.69	9.14	
2	8	4.30	13.44	
2.1	5	2.69	16.13	
2.2	6	3.23	19.35	
2.3	14	7.53	26.88	
2.4	6	3.23	30.11	
2.5	7	3.76	33.87	
2.6	6	3.23	37.10	
2.7	16	8.60	45.70	
2.8	18	9.68	55.38	MED
2.9	9	4.84	60.22	
3	19	10.22	70.43	MOD
3.1	10	5.38	75.81	
3.2	12	6.45	82.26	
3.3	13	6.99	89.25	
3.4	9	4.84	94.09	
3.5	8	4.30	98.39	
3.6	1	0.54	98.92	
3.7	2	1.08	100.00	
Total	186	100.00		

Graphische Bestimmung des Medians anhand der Verteilungsfunktion



Eigenschaften des Medians

- sinnvoll ab Ordinalskalenniveau
- unempfindlich gegen 'Ausreißer'
- mindestens 50% der Fälle sind kleiner oder gleich dem Median
- mindestens 50% der Fälle sind größer oder gleich dem Median
- Qualitätseigenschaft: kleinster Gesamtabstand zu allen anderen Werten, d.h.

$$\sum_{i=1}^{n} \left| x_i - x_{MED} \right| \leq \sum_{i=1}^{n} \left| x_i - z \right| \quad \textit{für alle } z$$

Das arithmetische Mittel

Das arithmetische Mittel ist gleich der Summe aller Fälle geteilt durch die Anzahl aller Fälle, d.h.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Beispiel 1: Gegeben sei wieder die Urliste

2, 4, 5, 4, 3, 1, 3, 4, 2, 5

$$\bar{x} = \frac{1}{10}(2 + 4 + 5 + 4 + 3 + 1 + 3 + 4 + 2 + 5) = 3.3$$

Berechnung aus Häufigkeitstabelle

$$\bar{x} = \sum_{j=1}^{k} a_j f_j$$

Beispiel

mathenote oberstufe	Freq.	Percent	Cum.
1	4	2.17	2.17
2	30	16.30	18.48
3	66	35.87	54.35
4	60	32.61	86.96
5	24	13.04	100.00
Total	184	100.00	

$$\bar{x} = (1 \cdot 0.0217 + 2 \cdot 0.1630 + 3 \cdot 0.3587 + 4 \cdot 0.3261 + 5 \cdot 0.1304) \approx 3.38$$

Berechnung aus gruppierten Daten

- mit den relativen Häufigkeiten der Klassen gewichtete Summe der Klassenmitten

abitkat	Freq.	Percent	Cum.
1.0 - 1.4	2	1.08	1.08
1.5 - 1.9	15	8.06	9.14
2.0 - 2.4	39	20.97	30.11
2.5 - 2.9	56	30.11	60.22
3.0 - 3.4	63	33.87	94.09
3.5 - 3.9	11	5.91	100.00
Total	186	100.00	

$$\bar{x} = \frac{1}{186}(2 \cdot 1.2 + 15 \cdot 1.7 + 39 \cdot 2.2 + 56 \cdot 2.7 + 63 \cdot 3.2 + 11 \cdot 3.7) \approx 2.72$$

Berechnung aus Schichten

Der Gesamtmittelwert lässt sich auch aus getrennten Mittelwerten für einzelne Schichten ("Teilgesamtheiten") berechnen durch

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{r} n_j \bar{x}_j$$

Beispiel (fiktiv)

Wohnort	N(Einkommen)	mean(Einkommen)
Westdeutschland	600	3000
Ostdeutschland	200	2000
Total	800	2750

$$\bar{x} = \frac{1}{800} \cdot (600 \cdot 3000 + 200 \cdot 2000) = 2750$$

Eigenschaften des arithmetischen Mittels

- sinnvoll für metrische Daten
- empfindlich gegen „Ausreißer“
- "Schwerpunkteigenschaft":

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$$

- Qualitätseigenschaft: Minimierung der 'Abstandsquadrate', d.h.

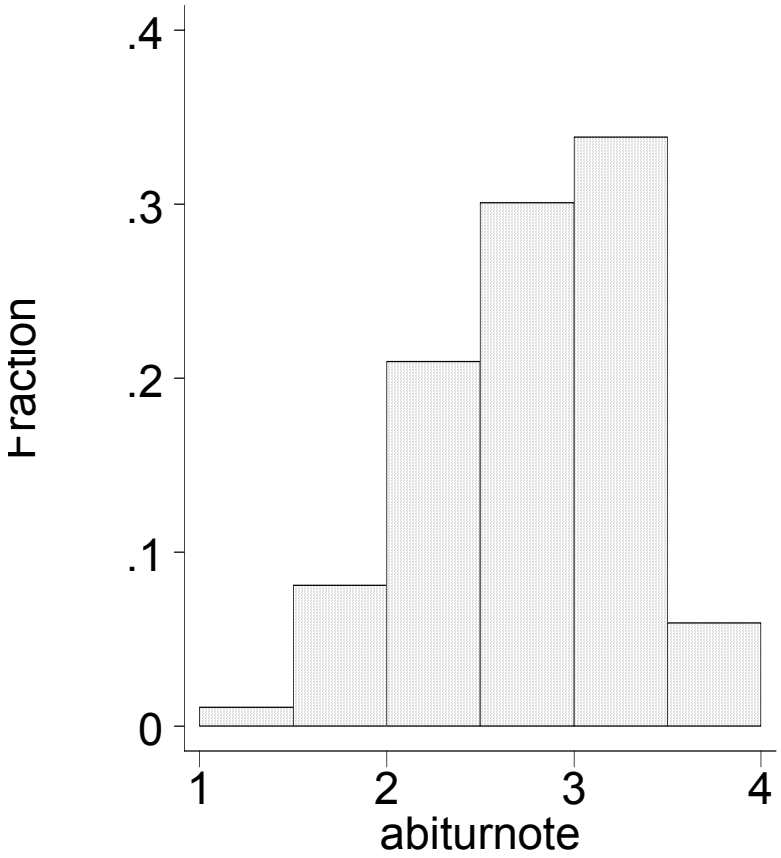
$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \leq \sum_{i=1}^{n}\left(x_i - z\right)^2 \quad \textit{für alle } z$$

Lageregeln für unimodale Verteilungen

$$symmetrisch: \quad x_{MOD} \approx x_{MED} \approx \bar{x}$$

$$rechtsschief: \quad x_{MOD} < x_{MED} < \bar{x}$$

$$linksschief: \quad x_{MOD} > x_{MED} > \bar{x}$$

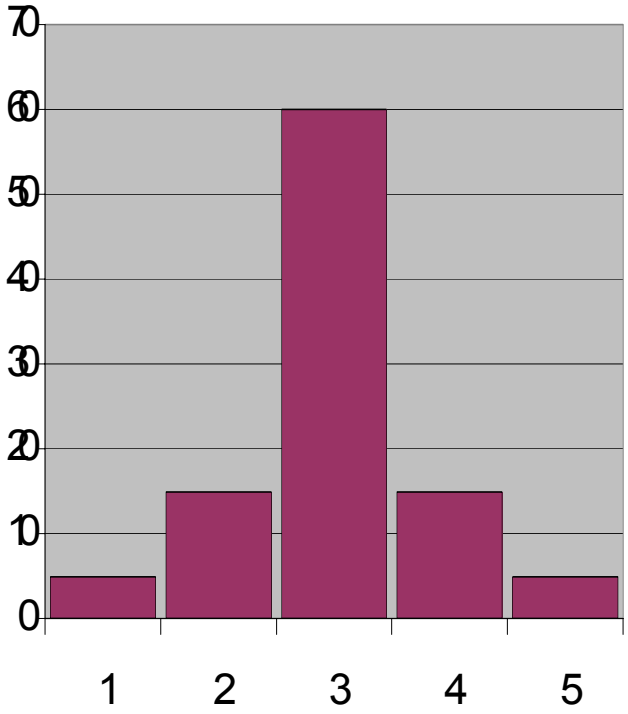


- Lageparameter aus den ungruppierten Daten (s. Folie 45):
  - Modus: 3.0
  - Median: 2.8
  - Mittelwert: 2.73

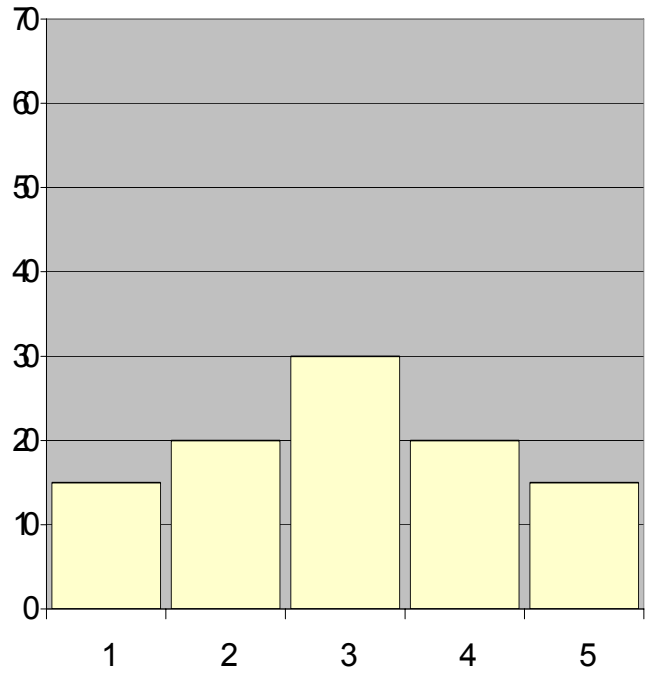
=> linksschief

Streuungsmaße

- Maße der zentralen Tendenz (Modus, Median, arithmetisches Mittel) können bestimmte Unterschiede von Verteilungen nicht erfassen



Modus = 3
Median = 3
arithm. Mittel = 3



Modus = 3
Median = 3
arithm. Mittel = 3

Die Spannweite (Range)

- Die Spannweite R einer Verteilung ist der Abstand zwischen dem kleinstem und dem größtem Wert, d.h. $R = x_{max} - x_{min}$

Beispiel 1: Gegeben sei folgende Urliste

2, 4, 5, 4, 3, 1, 3, 4, 2, 5

R = 5 - 1 = 4

- Probleme:
  - Die Spannweite wächst tendenziell mit n
  - und ist empfindlich gegenüber Ausreißern
    - berücksichtigt evtl. nur zwei Ausreißer!

Der Interquartilsabstand

- Quantile:

  Für jeden Wert p mit $0 < p < 1$ ist der Wert x_p das "p-Quantil", wenn mindestens ein Anteil von p der Daten kleiner oder gleich x_p und mindestens ein Anteil von 1-p der Daten größer oder gleich x_p sind.

  Quantile sind in gewisser Weise Verallgemeinerungen des Medians. Der Median ist das 50%-Quantil $x_{0.50}$

- Der Interquartilsabstand d_Q ist die Distanz zwischen dem 25%-Quantil ("unteres Quartil") und dem 75%-Quantil ("oberes Quartil"), d.h.

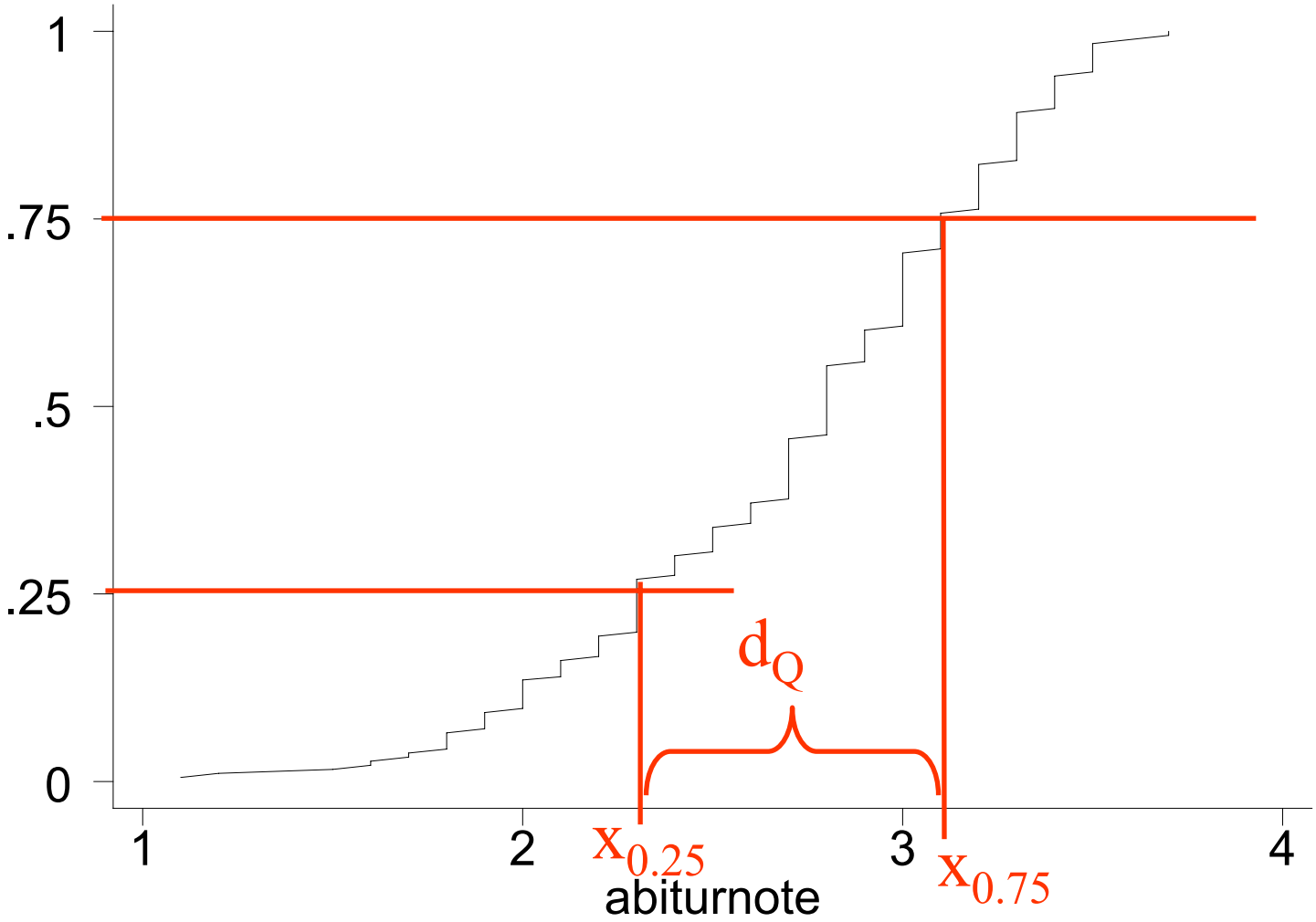
  $d_Q = x_{0.75} - x_{0.25}$

Beispiel: Abiturnote

$d_Q = 3.1\text{-}2.3 = 0.8$

abinote	Freq.	Percent	Cum.
1.1	1	0.54	0.54
1.2	1	0.54	1.08
1.5	1	0.54	1.61
1.6	2	1.08	2.69
1.7	2	1.08	3.76
1.8	5	2.69	6.45
1.9	5	2.69	9.14
2	8	4.30	13.44
2.1	5	2.69	16.13
2.2	6	3.23	19.35
2.3	14	7.53	26.88
2.4	6	3.23	30.11
2.5	7	3.76	33.87
2.6	6	3.23	37.10
2.7	16	8.60	45.70
2.8	18	9.68	55.38
2.9	9	4.84	60.22
3	19	10.22	70.43
3.1	10	5.38	75.81
3.2	12	6.45	82.26
3.3	13	6.99	89.25
3.4	9	4.84	94.09
3.5	8	4.30	98.39
3.6	1	0.54	98.92
3.7	2	1.08	100.00
Total	186	100.00	

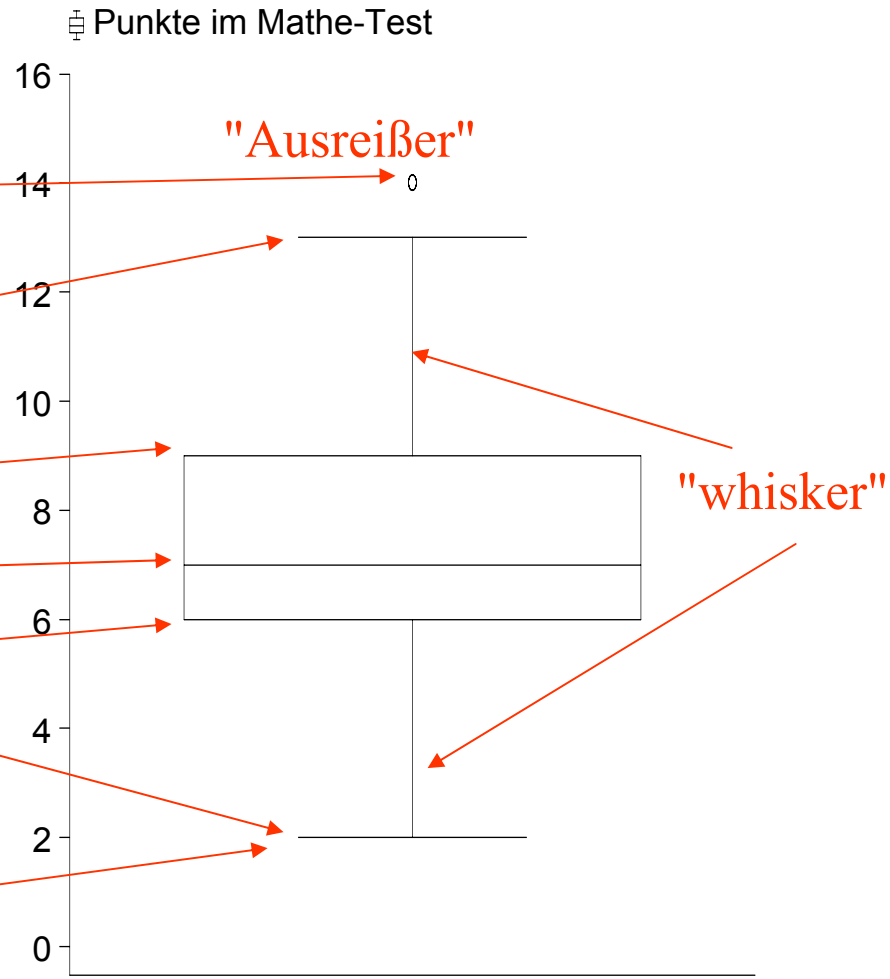
Graphische Bestimmung der Quartile und des Interquartilsabstandes



Exkurs: Der (modifizierte) Box-Plot

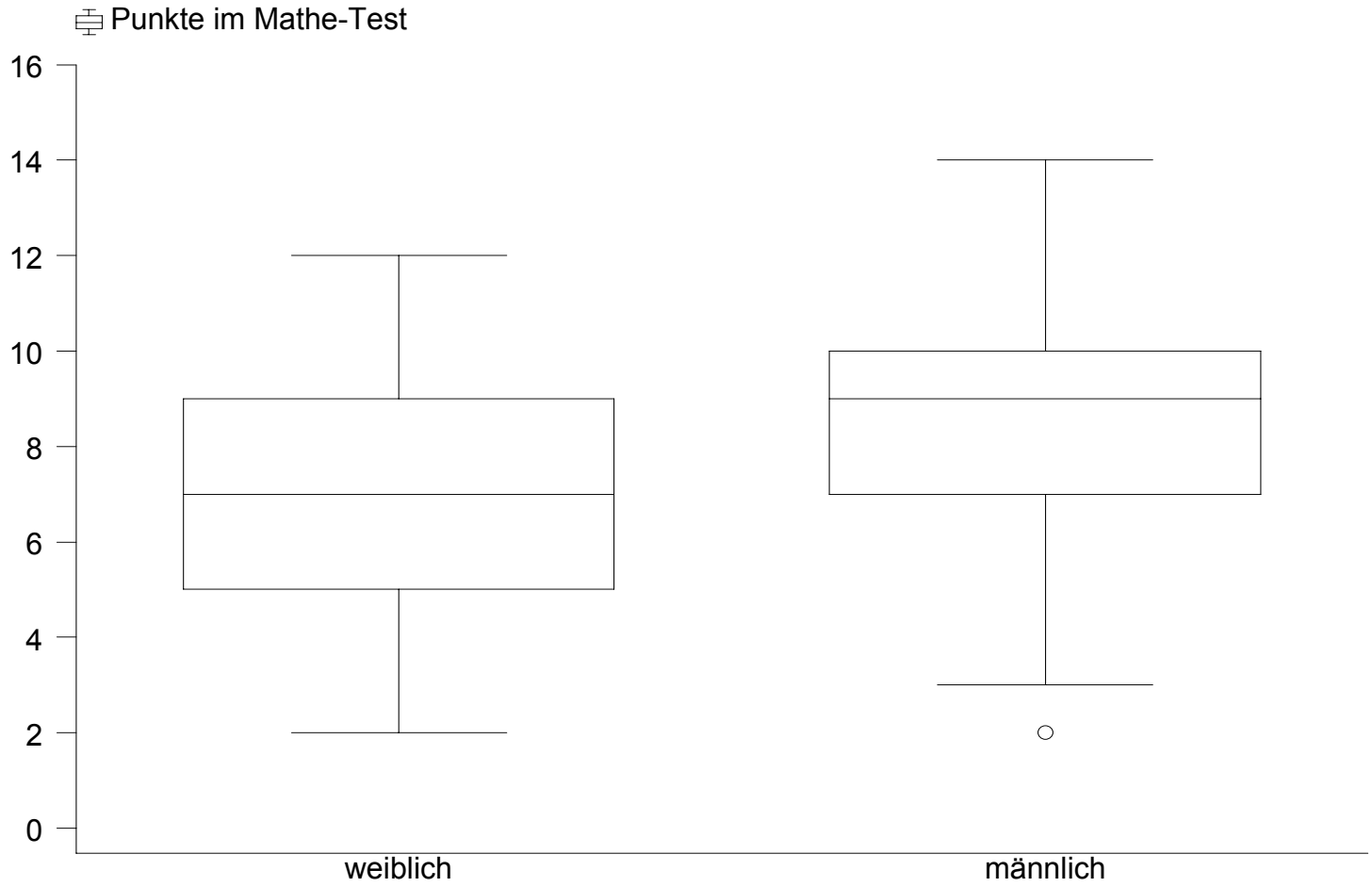
Fasst die Information aus folgenden Kennziffern zusammen:

- Maximum, x_{max}
- größter Wert unterhalb "oberer Zaungrenze"
  $z_o = x_{0.75} + 1.5d_Q$
- oberes Quartil, $x_{0.75}$
- Median, x_{MED}
- unteres Quartil, $x_{0,25}$
- kleinster Wert oberhalb "unterer Zaungrenze"
  $z_u = x_{0.25} - 1.5d_Q$
- Minimum, x_{min}
- Range, R
- Interquartilsabstand d_Q



Beispiel: Box-Plots

Box-Plots eignen sich gut für den Vergleich von Teilgruppen. Hier Männer – Frauen (dies ist bereits ein Bsp. für eine bivariate Zusammenhangsanalyse).



Empirische Varianz und Standardabweichung

- Die Varianz einer Verteilung ist definiert als:

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

- Die Varianz ist also die Summe der quadrierten Abweichungen vom arithmetischen Mittel geteilt durch die Anzahl der Werte

- Die Standardabweichung ergibt sich als Wurzel der Varianz, d.h.

$$\tilde{s} = \sqrt{\tilde{s}^2}$$

Stichprobenvarianz

- ACHTUNG!
- Von der empirischen Varianz wird in der Statistik auch noch eine Stichprobenvarianz unterschieden, die vor allem in der induktiven Statistik benötigt wird
- Die meisten Statistikprogramme geben diese Stichprobenvarianz standardmäßig aus
- Die Stichprobenvarianz ist definiert als:

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

Berechnung Varianz aus Rohdaten

Daten von Bsp. 1

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	2	-1.3	1.69
2	4	0.7	0.49
3	5	1.7	2.89
4	4	0.7	0.49
5	3	-0.3	0.09
6	1	-2.3	5.29
7	3	-0.3	0.09
8	4	0.7	0.49
9	2	-1.3	1.69
10	5	1.7	2.89
Σ	33	0	16.1

$$\Rightarrow \tilde{s}^2 = \frac{1}{10} \cdot 16.1 = 1.61$$

$$\Rightarrow \tilde{s} = \sqrt{1.61} \approx 1.27$$

Berechnung Varianz aus Häufigkeitstabelle

Formel:

$$\tilde{s}^2 = \sum_{j=1}^{k} (a_j - \bar{x})^2 \, f_j$$

Beispiel: Mittelwert $x_{MW} = 3.38$

mathenote oberstufe	Freq.	Perc.	$a_j - x_{MW}$	$(a_j - x_{MW})^2$	$f_j(a_j - x_{MW})^2$
1	4	2.17	-2.38	5.6644	0.1231
2	30	16.30	-1.38	1.9044	0.3105
3	66	35.87	-0.38	0.1444	0.0518
4	60	32.61	0.62	0.3844	0.1253
5	24	13.04	1.62	2.6244	0.3423
Total	184	100.00			0.953

Std.abw. = 0.976

Der Verschiebungssatz

- Eine Rechenerleichterung ergibt sich durch den sogenannten Verschiebungssatz; für jedes c gilt:

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

- Für den Spezialfall c=0 ergibt sich:

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2$$

Beispiel 1

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2$$

x_i	x_i^2
2	4
4	16
5	25
4	16
3	9
1	1
3	9
4	16
2	4
5	25
Σ	125

$$\Rightarrow \tilde{s}^2 = \frac{125}{10} - 3.3^2 = 12.5 - 10.89 = 1.61$$

Varianzzerlegung

- Ist die Stichprobe in r Schichten unterteilt, so gilt:

$$\tilde{s}^2 = \frac{1}{n}\sum_{j=1}^{r} n_j \tilde{s}_j^{\,2} + \frac{1}{n}\sum_{j=1}^{r} n_j (\bar{x}_j - \bar{x})^2$$

Dabei ist:

n_j *die Anzahl der Fälle in Schicht* j, $j \in \{1,...,r\}$

$\bar{x}_j$ *der Mittelwert innerhalb der Schicht* j

$\bar{x}$ *der Gesamtmittelwert*

$\tilde{s}_j$ *die Varianz innerhalb der Schicht* j

Die Gesamtstreuung lässt sich also in eine Streuung innerhalb der Schichten und eine Streuung zwischen den Schichten zerlegen

Beispiel 1:
Zerlegt in 2 Schichten

Schicht (j)	x_i	x_i^2
1	2	4
1	4	16
1	5	25
1	4	16
1	3	9
Σ	18	70
	MW: 3.6	
2	1	1
2	3	9
2	4	16
2	2	4
2	5	25
Σ	15	55
	MW: 3.0	

$$\tilde{s}_1^2 = \frac{70}{5} - 3.6^2 = 14 - 12.96 = 1.04$$

$$\tilde{s}_2^2 = \frac{55}{5} - 3.0^2 = 11 - 9 = 2.0$$

$$\tilde{s}^2 = \frac{1}{10}(5 \cdot 1.04 + 5 \cdot 2.0) + \frac{1}{10}\left(5 \cdot (3.6 - 3.3)^2 + 5 \cdot (3.0 - 3.3)^2\right) = 1.52 + 0.09 = 1.61$$

Der Variationskoeffizient

- Für Merkmale mit nichtnegativen Ausprägungen ist der Variationskoeffizient definiert als:

$$v = \frac{\tilde{s}}{\bar{x}}$$

$$für \ \bar{x} > 0$$

- Im Gegensatz zur Varianz und Standardabweichung ist dieser Streuungskoeffizient maßstabsunabhängig und deshalb zum allgemeineren Vergleich von Verteilungen geeignet (Beispiel: Einkommensvarianz in D und USA)

Transformationsregeln

LineareTransformation

$Y = aX + b,$

d.h. $y_i = ax_i + b$ für $i = 1,...,n$

Transformationsregel für das Mittel:

$\bar{y} = a\bar{x} + b$

Transformationsregel für die Varianz:

$\tilde{s}_y^{\,2} = a^2 \tilde{s}_x^{\,2}$ und $\tilde{s}_y = |a| \tilde{s}_x$

Standardisierung

Ist X eine (metrische) Variable, so besitzt die transformierte Variable

$$Y = \frac{X - \bar{x}}{\tilde{s}_x}$$

den Mittelwert 0 und die Standardabweichung 1.
Solche Variablen heißen „standardisierte Variablen“

<u>Beispiel:</u> Vergleich psychometrischer Testscores

$$\left.\begin{array}{ll} X: & \bar{x} = 30,\ \tilde{s}_x = 10;\quad x_i = 35 \\ Y: & \bar{y} = 40,\ \tilde{s}_y = 15;\quad y_i = 45 \end{array}\right\} \text{wo war Person i besser?}$$

$$\left.\begin{array}{l} \dfrac{35-30}{10} = 0.5 \\ \dfrac{45-40}{15} = 0.3\overline{3} \end{array}\right\} \text{in Test X!}$$

Kapitel III
Wahrscheinlichkeitstheorie

Prof. Dr. Josef Brüderl
Universität Mannheim

Frühjahrssemester 2007

Induktive Statistik

- In den Sozialwissenschaften ist es meist unmöglich Daten über alle Objekte zu sammeln, über die man Aussagen treffen will (Grundgesamtheit).
- Deshalb zieht man in der Regel Stichproben.
- Grundsätzliche Frage
  - Wie sicher ist es, dass die Schlüsse, die man aufgrund der Stichprobe zieht, auch wirklich für die Grundgesamtheit gelten?
- Induktion: Schluss vom Speziellen auf das Allgemeine
- Beispiel: Test eines neuen Medikaments zur Blutdrucksenkung
  - In einem Experiment mit 50 Versuchspersonen, stellt man fest, dass gegenüber der Kontrollgruppe (Placebo) in der Experimentalgruppe der mittlere Blutdruck niedriger ausfällt. Ist dieser Unterschied zufällig (durch die Stichprobenziehung verursacht) oder findet er sich auch in der Grundgesamtheit?

Grundbegriffe

- Die Wahrscheinlichkeitstheorie beschäftigt sich mit so genannten Zufallsvorgängen (Zufallsexperimenten)
  - Darunter versteht man Vorgänge, die zu verschiedenen Ergebnissen (Ereignissen) führen können, die sich gegenseitig ausschließen. Vor dem Vorgang ist ungewiss, welches der möglichen Ergebnisse eintreten wird
- Es wird nun versucht, den möglichen Ergebnissen Wahrscheinlichkeiten P(A) zuzuordnen
- Frequentistischer Wahrscheinlichkeitsbegriff
  - Die Wahrscheinlichkeiten lassen sich als Anteil verstehen, den ein Ergebnis hätte, wenn man den Vorgang sehr oft wiederholen würde

Beispiel: Würfeln

Ergebnis:	Wahrscheinlichkeit:
"1"	1/6
"4"	1/6
"gerade Zahl"	1/2
"keine 6"	5/6
"Zahl größer vier"	1/3
"1" oder "4"	1/3
"gerade und kleiner als fünf"	1/3
"ungerade und größer als fünf"	0

Mengen

Es ist nützlich, die Ergebnisse von Zufallsvorgängen als Mengen zu kennzeichnen:

- Eine *Menge* ist eine Zusammenfassung verschiedener Objekte. Die einzelnen Objekte werden *Elemente* genannt.
- Beispiele:
  - {1,2,3,4,5,6}
  - {"a", "b", "c",..., "z"}
  - {x: x ist eine natürliche Zahl mit $1 \leq x \leq 5$}
  - {Kopf, Zahl}

Standardmengen

- N : Menge der natürlichen Zahlen
- N_0: Menge der natürlichen Zahlen inklusive 0
- R: Menge der reellen Zahlen
- Ø: leere Menge
- Ω: Gesamtmenge aller Ereignisse (Ergebnisraum, Ergebnismenge)
  - Teilmengen von Ω heißen (Zufalls-) Ereignisse
  - Die einelementigen Teilmengen von Ω heißen Elementarereignisse

Grundlegende Definitionen

$x \in A$ x ist ein Element der Menge A

$x \notin A$ x ist kein Element der Menge A

$A \subset B$ A ist eine Teilmenge von B

für alle $x \in A$ gilt $x \in B$

$A \cap B$ Schnittmenge

$A \cap B = \{x: x \in A$ und $x \in B\}$

$A \cup B$ Vereinigungsmenge

$A \cup B = \{x: x \in A$ oder $x \in B\}$

$A \setminus B$ Differenzmenge

$A \setminus B = \{x: x \in A$ und $x \notin B\}$

$\neg A$ Komplementärmenge, "Nicht-A„

$\neg A = \Omega \setminus A = \{x: x \notin A\}$

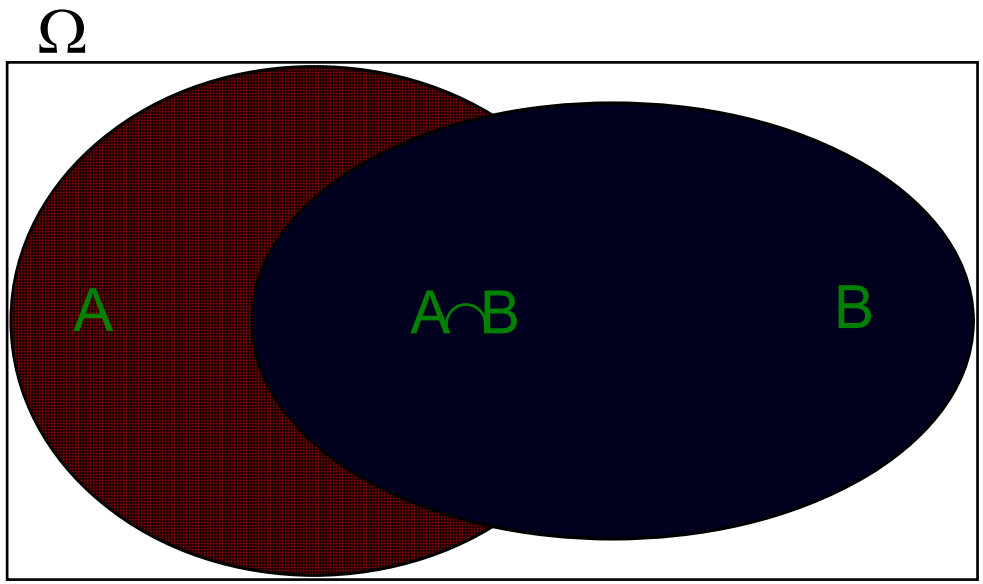
$\wp(A)$ Potenzmenge

die Menge aller Teilmengen von A; $\wp(A) = \{M: M \subset A\}$

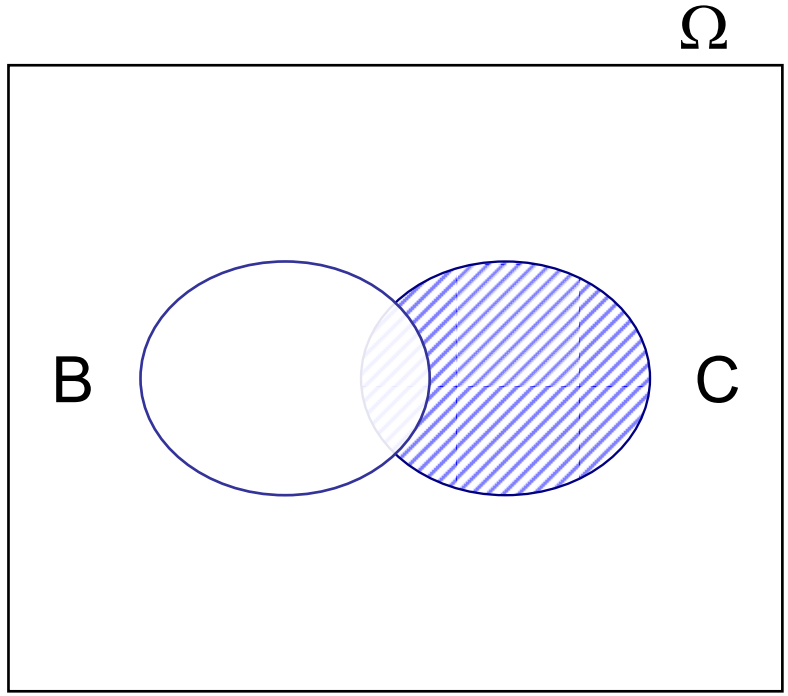
$|A|$ Mächtigkeit

Anzahl der Elemente, die in A enthalten sind

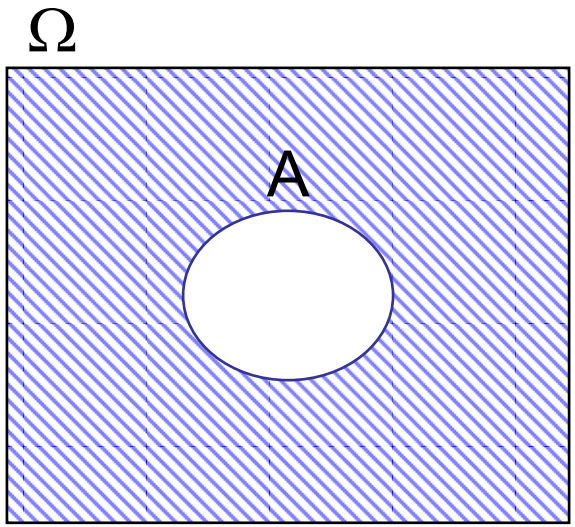
Das Venn-Diagramm



Schnittmenge



C \ B



$\neg A = \Omega \setminus A$

Disjunkte Mengen: $A \cap B = \emptyset$

Wahrscheinlichkeiten

Die Wahrscheinlichkeit P von Ereignissen lässt sich als eine Abbildung auffassen. Jedem Ereignis A wird eine Zahl P(A) zwischen 0 und 1 zugeordnet.

$$P: \{A: A \in \Omega\} \rightarrow [0,1]$$

Dabei müssen die drei *Axiome von Kolmogoroff* gelten:

K1: $P(A) \geq 0$ für alle $A \subset \Omega$

K2: $P(\Omega) = 1$

K3: Wenn: $A \cap B = \emptyset$ gilt $P(A \cup B) = P(A) + P(B)$

Abgeleitete Rechenregeln

Aus den drei *Axiomen von Kolmogoroff* lassen sich unter anderem die folgenden weiteren Regeln ableiten:

1. $0 \leq P(A) \leq 1$ für alle $A \subset \Omega$
2. $P(\emptyset) = 0$
3. $P(A) \leq P(B)$ falls $A \subset B$
4. $P(\neg A) = 1 - P(A)$
5. $P(A_1 \cup A_2 \cup ... \cup A_k) = P(A_1) + P(A_2) + ... + P(A_k)$
   falls $A_1, A_2, ..., A_k$ paarweise disjunkt
6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   (Additionssatz)

Empirische Wahrscheinlichkeitsinterpretation

- Frequentistisch
  - Ist $f_n(A)$ die relative Häufigkeit von A bei n Wiederholungen eines Zufallsexperiments, so gilt:
    $$f_n(A) \to P(A), \quad \text{für } n \to \infty$$
  - Die Wahrscheinlichkeit lässt sich also als Grenzwert der relativen Häufigkeit interpretieren, wenn man das Zufallsexperiment unendlich oft wiederholen würde
  - Die relativen Häufigkeiten, die aus endlichen Wiederholungen bestimmt werden, können zur Schätzung der unbekannten Wahrscheinlichkeit herangezogen werden.
- Subjektive Einschätzung
  - Wahrscheinlichkeiten können auch auf persönlichen Einschätzungen des Betrachters beruhen.
  - Beispiele: Sportergebnisse, Konjunkturentwicklung
  - Die subjektiven Wahrscheinlichkeit können mit "Wettquotienten" in Verbindung gebracht werden

Wie bestimmt man Wahrscheinlichkeiten?

- Laplace Wahrscheinlichkeit
  - Sind alle Elementarereignisse eines Zufallsexperiments gleich wahrscheinlich, so lässt sich die Wahrscheinlichkeit eines bestimmten Ergebnisses A durch folgende Abzählregel bestimmen:

$$P(A) = \frac{|A|}{|\Omega|}$$

  - Man teilt also die Anzahl der "für A günstigen Elementarereignisse" durch die Anzahl der "möglichen Elementarereignisse"
  - Beispiel Tombola: 25 Gewinnlose, 75 Nieten

$$P(\text{Gewinn}) = \frac{\text{Anzahl günstiger Ergebnisse}}{\text{Anzahl möglicher Ergebnisse}} = \frac{25}{100} = 0{,}25$$

$$P(\text{Niete}) = \frac{75}{100} = 0{,}75 = P(\overline{\text{G}}) = 1 - P(\text{G})$$

Kombinatorik

- Zur Berechnung der Laplace-Wahrscheinlichkeit benötigt man die Anzahl der günstigen und möglichen Ergebnisse. Dafür benötigt man häufig kombinatorische Überlegungen.
- Permutationen
  - Wieviele Möglichkeiten gibt es N unterscheidbare Objekte anzuordnen (wieviele Permutationen)? Antwort: N! (N Fakultät).

$$N! = N \cdot (N-1) \cdot (N-2) \cdot \ldots \cdot 2 \cdot 1$$

  - Beispiel 4 Asse: Es gibt 4! = 24 verschiedene Kombinationen Damit ist P({Herz, Karo, Kreuz, Pik}) = 1/24
- Wie viele verschiedene Stichproben vom Umfang n kann man aus einer Grundgesamtheit vom Umfang N ziehen?
  - Mit oder ohne Berücksichtigung der Reihenfolge
  - Mit oder ohne Zurücklegen

Kombinatorik

- Mit Reihenfolge / mit Zurücklegen: N^n
  - N=10, n=3: $|\Omega| = 10 \cdot 10 \cdot 10 = 10^3$
- Mit Reihenfolge / ohne Zurücklegen: $N! / (N-n)!$
  - N=10, n=3: $|\Omega| = 10 \cdot 9 \cdot 8 = 10! / 7!$
- Ohne Reihenfolge / mit Zurücklegen: $\binom{N+n-1}{n}$
  - Lotto (6 aus 49): $|\Omega| = \binom{49+6-1}{6} = 54! / (48! \cdot 6!) = 25.827.165$ (mit Zurücklegen)
- Ohne Reihenfolge / ohne Zurücklegen: $\binom{N}{n}$

$$Dabei\ ist\ \binom{N}{n} = \frac{N!}{(N-n)!\,n!}\ \ der\ sogenannte\ Binomialkoeffizient$$

  - Lotto (6 aus 49): $|\Omega| = \binom{49}{6} = 49! / (43! \cdot 6!) = 13.983.816$

Kombinatorik

Beispiel: n=2 aus {a,b,c}	ohne Zurückl.	mit Zurückl.
mit Reihen-folge	ab ac bc ba ca cb = 3! / 1!	ab ac bc ba ca cb aa bb cc $= 3^2$
ohne Reihen-folge	ab ac bc $= \binom{3}{2}$	ab ac bc aa bb cc $= \binom{4}{2}$

Bedingte Wahrscheinlichkeiten

Sind A und B zwei Ereignisse und gilt P(B) > 0, so ist die bedingte Wahrscheinlichkeit von A unter B definiert als:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Für bedingte Whs. gelten ebenfalls die Axiome von Kolmogoroff
- Die bedingte Whs. sagt uns, wie wahrscheinlich Ereignis A eintritt, wenn Ereignis B bereits eingetreten ist

- Sind A und B zwei Ereignisse und gilt P(B) > 0, so gilt der sogenannte "Produktsatz":

$$P(A \cap B) = P(A \mid B) \cdot P(B)$$

Stochastische Unabhängigkeit

- Zwei Ereignisse A und B heißen stochastisch unabhängig, wenn gilt:

$$P(A \mid B) = P(A) \qquad \text{mit } P(B) > 0$$

bzw.

$$P(B \mid A) = P(B) \qquad \text{mit } P(A) > 0$$

  - Das Eintreten des Ereignisses B führt zu keiner Neubewertung der Chance für das Eintreten von A

- Im Falle stochastischer Unabhängigkeit gilt der Produktsatz:

$$P(A \cap B) = P(A) \cdot P(B)$$

  - Diese Beziehung ist von großer praktischer Bedeutung, da sie besagt, dass die Whs. des gemeinsamen Auftretens zweier unabhängiger Ereignisse gleich dem Produkt der Einzelwahrscheinlichkeiten ist.

Unabhängigkeit: Beispiele

- Zweimaliges Würfeln
  - Wie wahrscheinlich ist es, zwei Einsen zu werfen?
    A: 1 beim 1. Wurf, B: 1 beim 2. Wurf; $P(A) = P(B) = 1/6$
    $P(A \cap B) = 1/36$ (da 36 mögliche Ergebnisse)
    Also: $P(A \cap B) = P(A) \cdot P(B)$
  - Gilt allgemein beim Würfeln: $P(6,6,6,6,6,6) = (1/6)^6 = 0{,}0000214$

- 3 Mädchen
  - $P(M,M,M) = (1/2)^3 = 0{,}125$

- Roulette (r := rot)
  - $P(r,r,r,r,r,r) = (1/2)^6 = 0{,}016$
  - $P(r \mid (r,r,r,r,r)) = 1/2$

Unabhängigkeit im Urnenmodell

- Urne: {1,2,3,4}, zweimaliges Ziehen
  - A = {1 beim 1. Zug}, B = {2 beim 2. Zug}
- Mit Zurücklegen
  - Ω = {(1,1)(1,2) … (4,4)}, | Ω| = 16, |A| = |B| = 4
  - P(A) = 4/16 = ¼
  - P(B) = 4/16 = ¼
  - P(A ∩ B) = P(1,2) = 1/16 = P(A) · P(B)
  - Es liegt somit Unabhängigkeit vor!
- Ohne Zurücklegen
  - Ω = {(1,2) … (4,3)}, | Ω| = 12, |A| = |B| = 3
  - P(A) = 3/12 = ¼
  - P(B) = 3/12 = ¼
  - P(A ∩ B) = P(1,2) = 1/12 ≠ P(A) · P(B)
  - Es liegt somit keine Unabhängigkeit vor!

Zufallsvariablen

- Eine Variable X heißt Zufallsvariable (ZV), falls ihre Werte bzw. Ausprägungen das Ergebnis eines Zufallsvorgangs sind
- Der Wert x, den X bei der Durchführung des Experiments annimmt, wird auch Realisierung genannt
- Wir unterscheiden wieder:
  - diskrete vs. stetige Zufallsvariablen
- Die Menge aller Wahrscheinlichkeiten (im Wertebereich) nennt man die *Wahrscheinlichkeitsverteilung* von X

Diskrete Zufallsvariablen

- Eine Zufallsvariable X ist *diskret*, falls sie nur endlich oder abzählbar unendlich viele Werte x_1, x_2, x_3,... annehmen kann
- Die Wertemenge von X wird als *Träger* T = {x_1, x_2, x_3,...} bezeichnet
- Die *Wahrscheinlichkeitsverteilung* von diskreten ZV ist durch die Whs. $P(X = x_i) = p_i$ für i = 1,2,3,... gegeben
- Reicht ein endlicher Wertebereich zur Beschreibung eines Zufallvorgangs aus, so ist die Wahrscheinlichkeitsverteilung $p_1,p_2,\ldots,p_k$ das wahrscheinlichkeitstheoretische Analogon zur relativen Häufigkeitsverteilung $f_1,f_2,\ldots,f_k$
- Die *Wahrscheinlichkeitsfunktion* f(x) ist definiert durch:

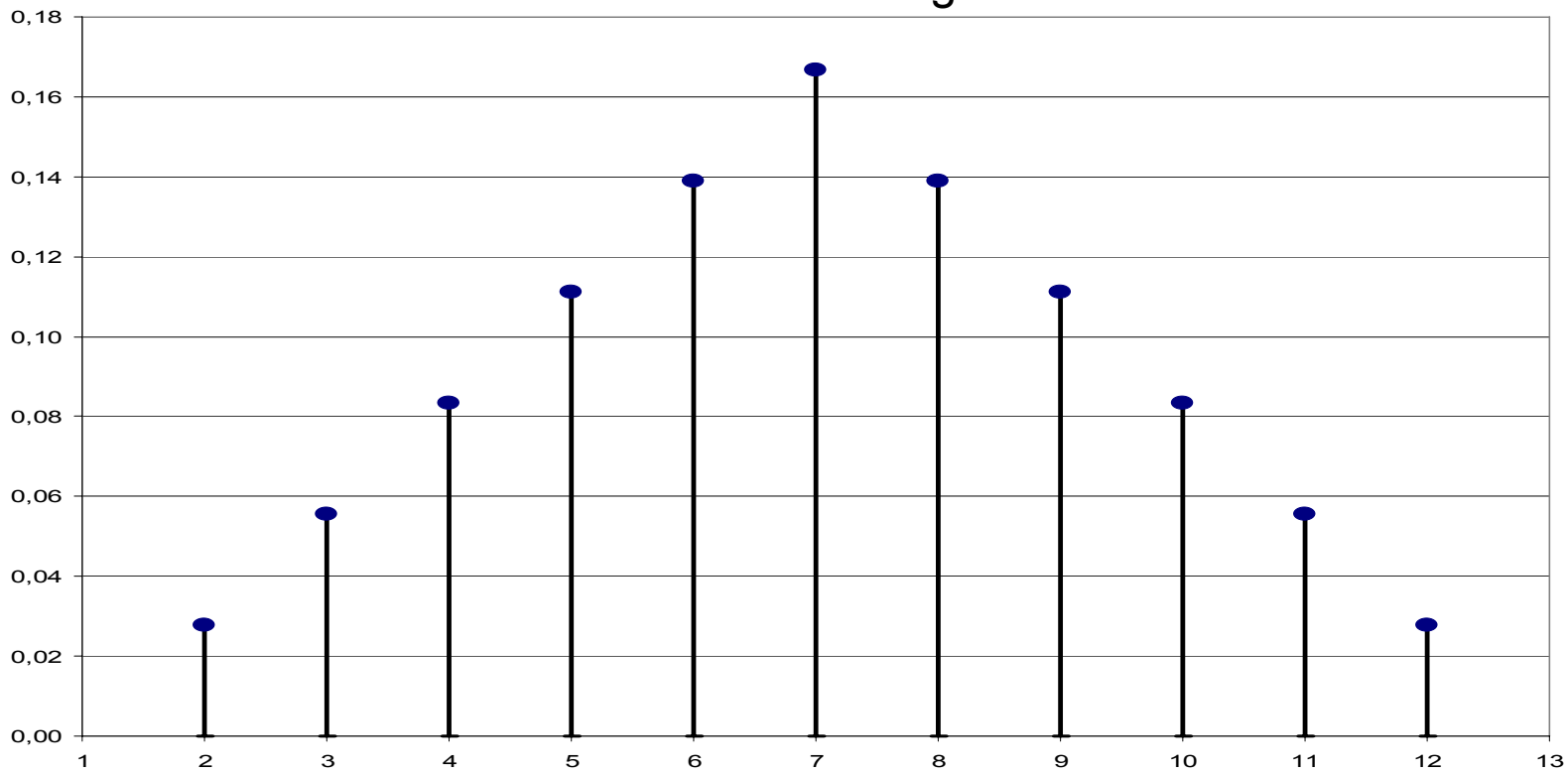
$$f(x) = \begin{cases} P(X = x_i) = p_i, \ \textit{für} \ x \in \mathrm{T} = \{x_1, x_2, \ldots\} \\ 0, \ \textit{für} \ x \notin \mathrm{T} \end{cases}$$

Beispiel: zweimaliges Würfeln

X = Summe der Augenzahlen, T = {2,3,4,...,12}

x_i	2	3	4	5	6	7	8	9	10	11	12
p_i	1/36	2/36	3/36	4/35	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Wahrscheinlichkeitsfunktion als Stabdiagramm



Beispiel: Bernoulli-Experiment

- Ist die ZV binär, so spricht man auch von einer *Bernoulli-Variablen* bzw. von der Durchführung eines *Bernoulli-Experimentes*
- Beispiel: Ereignis A tritt ein, oder nicht ein ($\neg$A)
  Die Zufallsvariable X (Indikator)
  X = 1 falls A eintritt
  X = 0 falls A nicht eintritt
  ist dann eine Bernoulli-Variable
- Mit P(A) = π ist

$$\left.\begin{aligned} P(X=1) &= \pi \\ P(X=0) &= 1-\pi \end{aligned}\right\} \text{Bernoulli} - \text{Verteilung}$$

Weitere Beispiele

- Gleichverteilung

  Eine diskrete Zufallsvariable auf dem Träger $T=\{x_1,...,x_k\}$ ist *gleichverteilt*, wenn für alle i = 1,...k:

$$P(X = x_i) = \frac{1}{k}$$

- Geometrische Verteilung

  Wird ein Bernoulli-Experiment mit $P(A) = \pi$ solange wiederholt, bis zum ersten Mal A eintritt, dann ist die Zufallsvariable X = „Anzahl der Versuche, bis zum ersten Mal A eintritt“ geometrisch verteilt mit Parameter π. Man schreibt $X \sim G(\pi)$. Es ist dann T = {1,2,3,...} und es gilt:

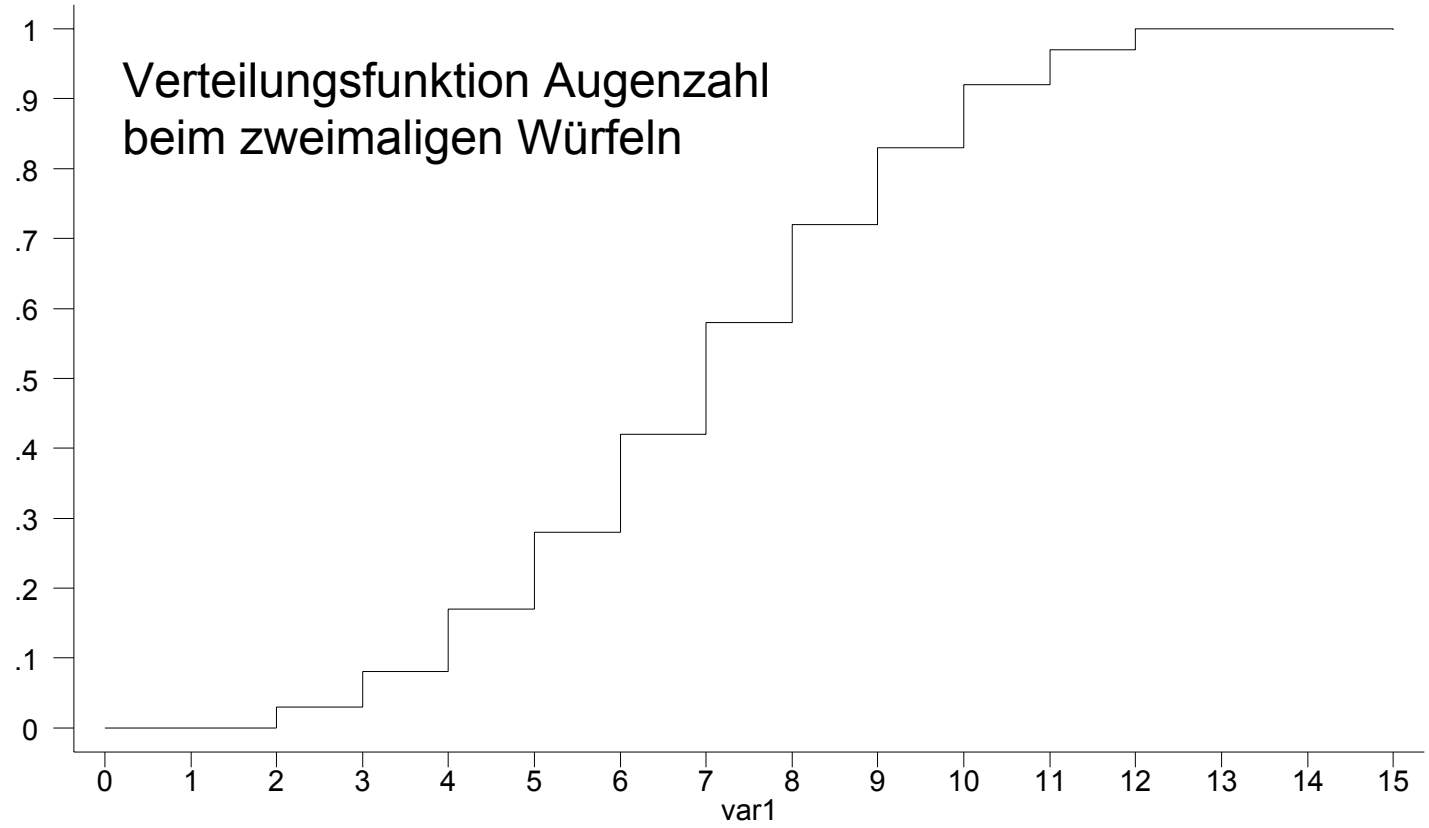
$$P(X = x_i) = (1-\pi)^{x_i - 1}\pi$$

  Bsp. Würfeln bis zum ersten Mal 6 auftritt:
  $P(X=1) = 1/6$, $P(X=2) = 5/6 \cdot 1/6$, $P(X=3) = (5/6)^2 \cdot 1/6$, …

Verteilungsfunktion

Die Verteilungsfunktion F(x) einer diskreten Zufallsvariable ist für jedes reelle x definiert durch

$$F(x) = P(X \leq x) = \sum_{i:x_i \leq x} p_i$$



Erwartungswert einer diskreten ZV

- Eine Maßzahl für das Zentrum einer Verteilung ist der Erwartungswert (μ). Analog zum arithmetischen Mittel:

$$\mu = E(X) = x_1 p_1 + x_2 p_2 + ... = \sum_{i \geq 1} x_i p_i$$

- Alternativ kann er auch errechnet werden nach:

$$E(X) = \sum_{i \geq 1} x_i f(x_i)$$

- Beispiele:
  - X = Augenzahl beim einmaligen Würfeln
    E(X) = (1 + 2 + 3 + 4 + 5 + 6) · 1/6 = 3,5
  - Bernoulli-Variable
    E(X) = 1·π + 0·(1 - π) = π
  - Glücksspiele im Casino
    E(X) < 0 !

Rechenregeln für Erwartungswerte

- Transformationsregel: Für $Y = aX + b$ gilt

  $E(Y) = aE(X) + b$

- Erwartungswert der Summe von ZV

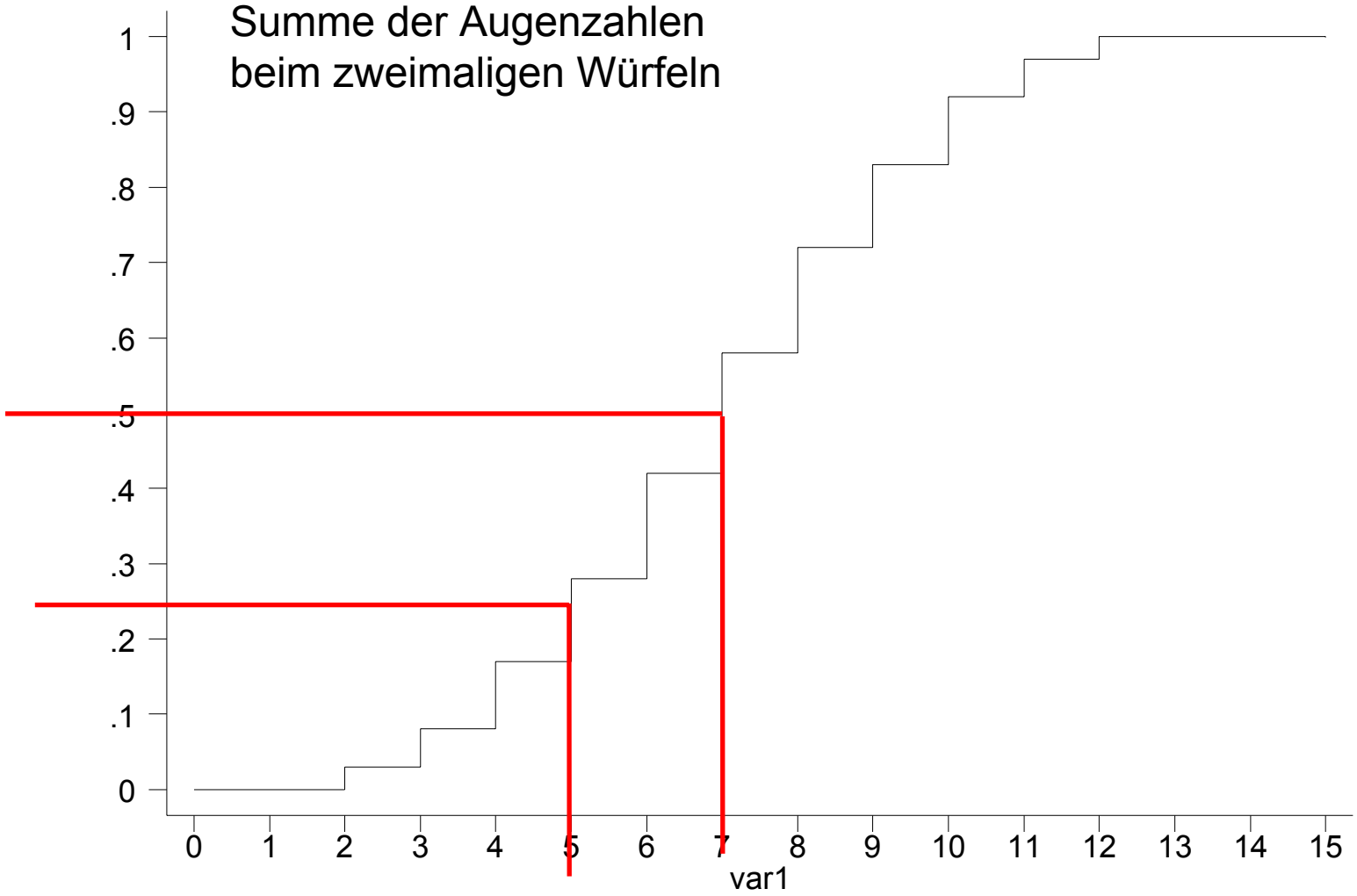
  $E(X + Y) = E(X) + E(Y)$

- Produktregel für unabhängige (!) ZV

  $E(X \cdot Y) = E(X) \cdot E(Y)$

- Beispiele:
  - Summe der Augenzahlen beim zweimaligen Würfeln
    $E(X_1 + X_2) = E(X_1) + E(X_2) = 3,5 + 3,5 = 7$
  - Produkt der Augenzahl beim zweimaligen Würfeln
    $E(X_1 \cdot X_2) = E(X_1) \cdot E(X_2) = 3,5 \cdot 3,5 = 12,25$

Quantile diskreter Zufallsvariablen



Die Varianz einer diskreten Zufallsvariable

- Die Varianz einer diskreten ZV X ist:

$$\sigma^2 = Var(X) = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + ...$$

$$= \sum_{i \geq 1} (x_i - \mu)^2 p_i = \sum_{i \geq 1} (x_i - \mu)^2 f(x_i)$$

- Die Standardabweichung ist gegeben durch:

$$\sigma = +\sqrt{\sigma^2}$$

- Verschiebungsregel

  $Var(X) = E(X^2) - \mu^2$

- Transformationsregel: Für Y = aX + b ist

  $Var(Y) = a^2 Var(X)$

- Für unabhängige (!) ZV gilt:

  Var(X + Y) = Var(X) + Var(Y)

Beispiele Varianz

- Bernoulli-Verteilung

  $\mu = \pi$
  $E(X^2) = 0^2 \cdot (1-\pi) + 1^2 \cdot \pi = \pi$
  $Var(X) = \pi - \pi^2 = \pi \cdot (1-\pi)$

- Einmaliges Würfeln

  $\mu = 3{,}5 = 7/2$
  $E(X^2) = (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) \cdot 1/6 = 91/6$
  $Var(X) = 91/6 – 49/4 = 2{,}92$

- Summe der Augenzahlen beim zweimaligen Würfeln

  $E(X^2)$ mühsam, aber
  $Var(X) = Var(X_1) + Var(X_2) = 2{,}92 + 2{,}92 = 5{,}83$

Die Binomialverteilung

- Wiederholt man ein Bernoulli-Experiment n-mal, dann ist die ZV X = "Anzahl der Versuche, bei denen A eintritt (Trefferzahl)" binomialverteilt mit T={0,1,…,n}. Notieren wir mit $P(A_i) = \pi$ die Whs., dass beim i-ten Experiment A eintritt, so lautet die Wahrscheinlichkeitsverteilung:

$$P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

- Notation: $X \sim B(n, \pi)$

  wobei n und π die Parameter der Binomialverteilung sind

- Es gilt:

  $E(X) = n \cdot \pi$, und $Var(X) = n \cdot \pi (1 - \pi)$

Beispiel: Lotterie (n=3, π=0,2)

Herleitung

Ergebnisse	x	P(Ergebnis)
(0,0,0)	0	0,8·0,8·0,8=0,512
(1,0,0)	1	0,2·0,8·0,8=0,128
(0,1,0)	1	0,8·0,2·0,8=0,128
(0,0,1)	1	0,8·0,8·0,2=0,128
(0,1,1)	2	0,8·0,2·0,2=0,032
(1,0,1)	2	0,2·0,8·0,2=0,032
(1,1,0)	2	0,2·0,2·0,8=0,032
(1,1,1)	3	0,2·0,2·0,2=0,008

Damit ist:

P(X=0) = 0,512
P(X=1) = 0,384
P(X=2) = 0,096
P(X=3) = 0,008

Berechnung mit Formel

$$P(X=0)=\binom{3}{0}0.2^0 0.8^3 = 1\cdot 1\cdot 0.512 = 0.512$$

$$P(X=1)=\binom{3}{1}0.2^1 0.8^2 = 3\cdot 0.2\cdot 0.64 = 0.384$$

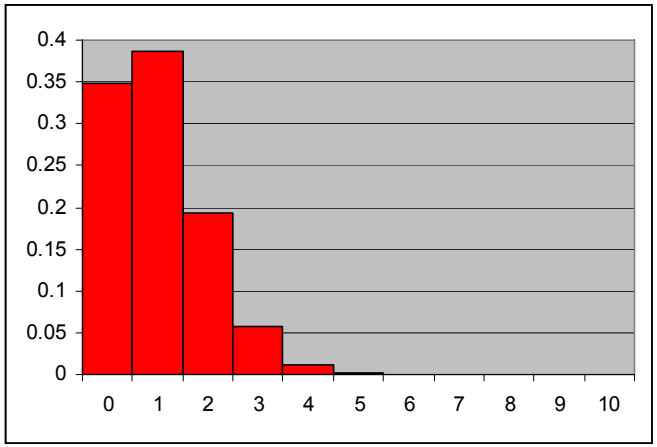
$$P(X=2)=\binom{3}{2}0.2^2 0.8^1 = 3\cdot 0.04\cdot 0.8 = 0.096$$

$$P(X=3)=\binom{3}{3}0.2^3 0.8^0 = 1\cdot 0.008\cdot 1 = 0.008$$

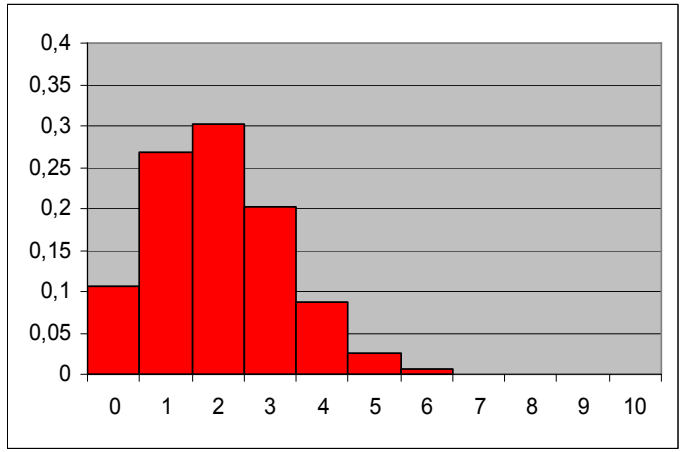
Whs. für mindestens 1 Treffer?

$P(X\geq 1) = 1 - P(X<1) = 1 - P(X=0) =$
$1 - 0.512 = 0.488$

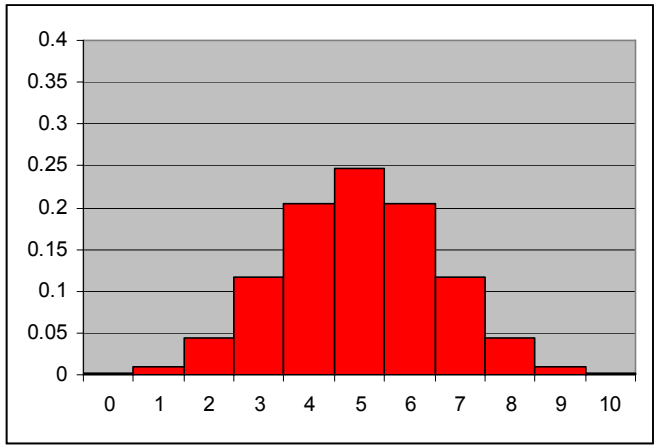
Beispiele n=10



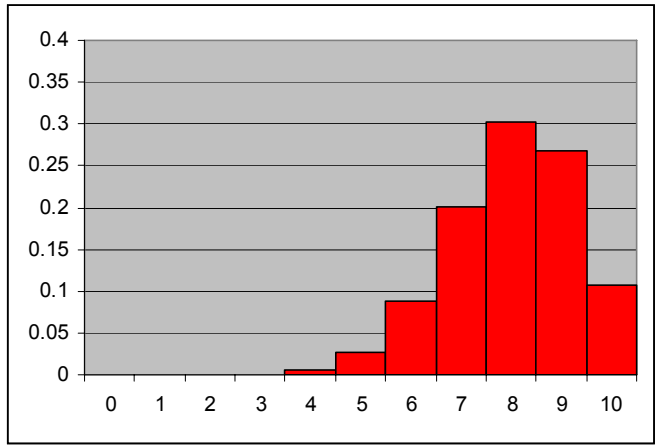
π=0.1



π=0.2



π=0.5



π=0.8

Stetige Zufallsvariablen

- Ein Merkmal X heißt stetig, falls für zwei mögliche Werte a und b auch jeder Wert x zwischen a und b (a < x < b) angenommen werden kann
  - Beispiel: reelle Zahlen mit beliebig vielen Nachkommastellen
- Problem: Die Wahrscheinlichkeit, dass in einem Zufallsvorgang ein bestimmter Wert erreicht wird, ist für alle Werte letztendlich Null

$$P(X = x_i) = 0$$

- Wahrscheinlichkeiten können deshalb sinnvoll nur für Intervalle angegeben werden

$$P(a < X \leq b) = P(a \leq X \leq b) =$$

$$P(a \leq X < b) = P(a < X < b)$$

Dichtefunktion f(x) und Verteilungsfunktion F(x)

- Die (Wahrscheinlichkeits-) *Dichte* (-funktion) f(x) beschreibt die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariable X, falls

$$P(a \le X \le b) = \int_a^b f(x)dx$$

- Es muss dann gelten: $f(x) \ge 0$ für alle x
- Außerdem muss die *Normierungseigenschaft* gelten:

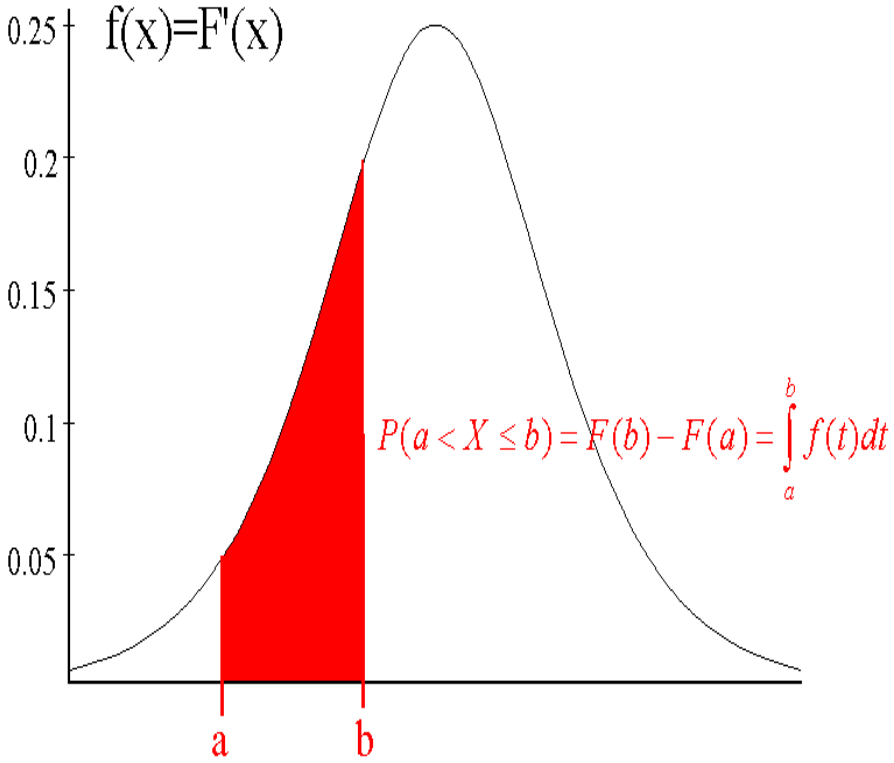
$$P(-\infty \le X \le +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$

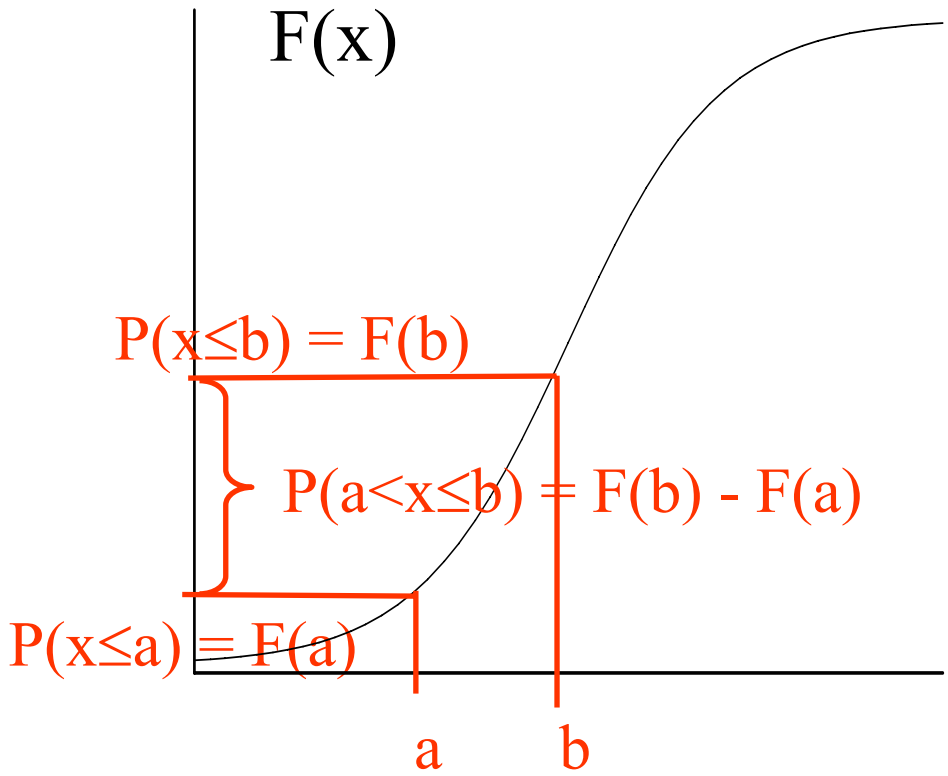
- Die *Verteilungsfunktion* lautet dann:

$$P(X \le x) = F(x) = \int_{-\infty}^{x} f(t)dt$$

Dichtefunktion f(x) und Verteilungsfunktion F(x)

- Es gilt:
  - F‘(x) = f(x)
  - Und $$P(a \le X \le b) = \int_a^b f(x)dx = F(b) - F(a)$$





Lage- und Streuungsparameter

- Zwei stetige ZV X und Y sind unabhängig, falls gilt

$$P(X \le x, Y \le y) = P(X \le x) \cdot P(Y \le y) = F_X(x) \cdot F_Y(y)$$

- Der Erwartungswert ist definiert als

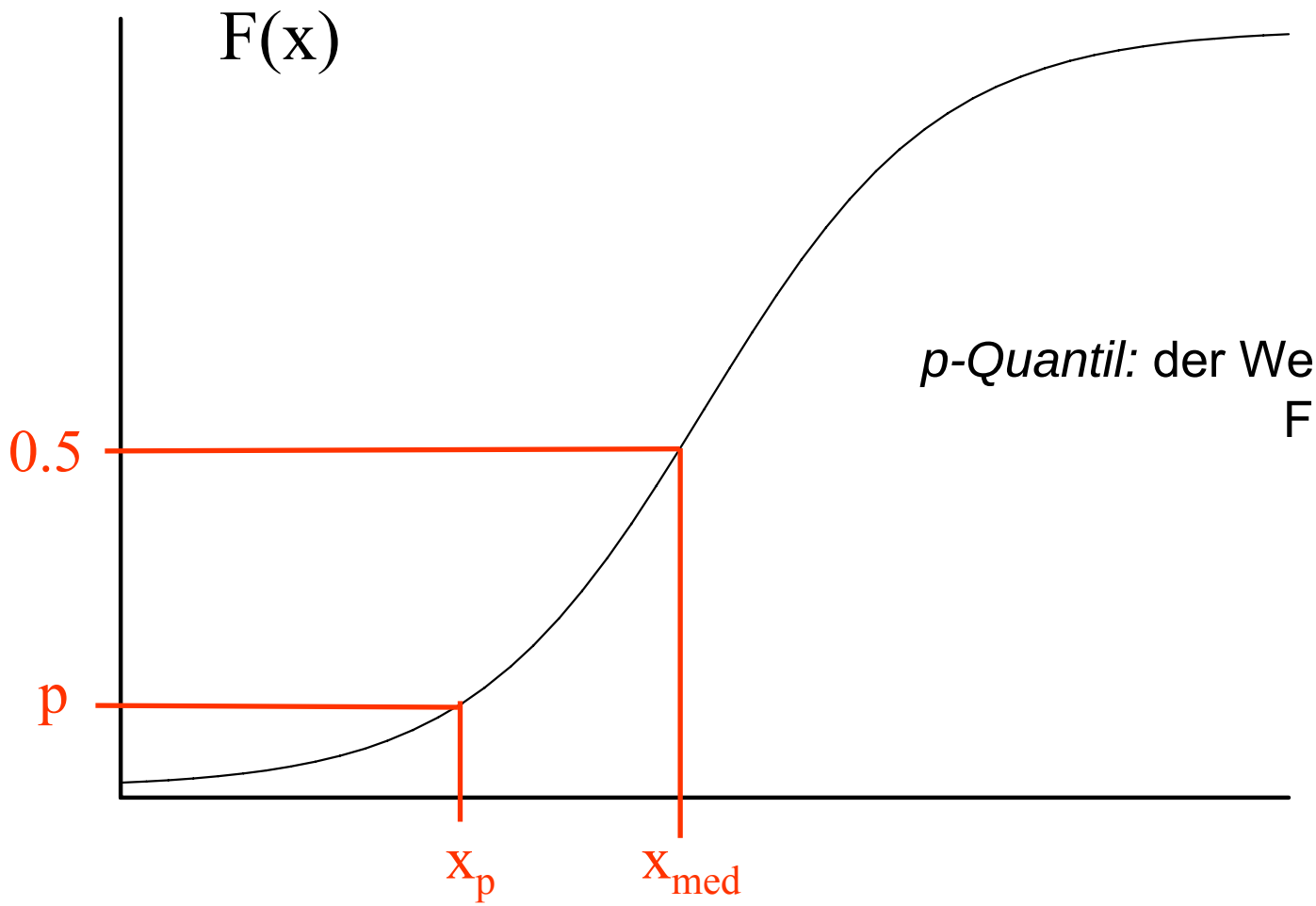
$$\mu = E(x) = \int_{-\infty}^{+\infty} x f(x) dx$$

- Die Varianz ist definiert als

$$\sigma^2 = Var(x) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

- Eigenschaften von Erwartungswert und Varianz analog wie bei diskreten ZV

Quantile



p-Quantil: der Wert x_p für den $F(x_p) = p$

Beispiel: stetige Gleichverteilung

- Die Dichtefunktion ist $$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \le x \le b \\ 0 & \text{sonst} \end{cases}$$

Beispiel: X : Wartezeit auf S-Bahn
(S-Bahn fährt alle 20 Minuten, zufälliges Eintreffen)

$$0 \le X \le 20 \Rightarrow f(x) = \frac{1}{20}, \; F(x) = \int_0^x \frac{1}{20} dt = \frac{x}{20}$$

$$E(X) = \int_0^{20} x \frac{1}{20} dx = \left[\frac{x^2}{40} \right]_0^{20} = \frac{400}{40} - 0 = 10$$

$$Var(X) = \int_0^{20} (x-10)^2 \frac{1}{20} dx = \left[\frac{(x-10)^3}{60} \right]_0^{20} = \frac{1000}{60} + \frac{1000}{60} = 33,\overline{3}$$

Die Normalverteilung

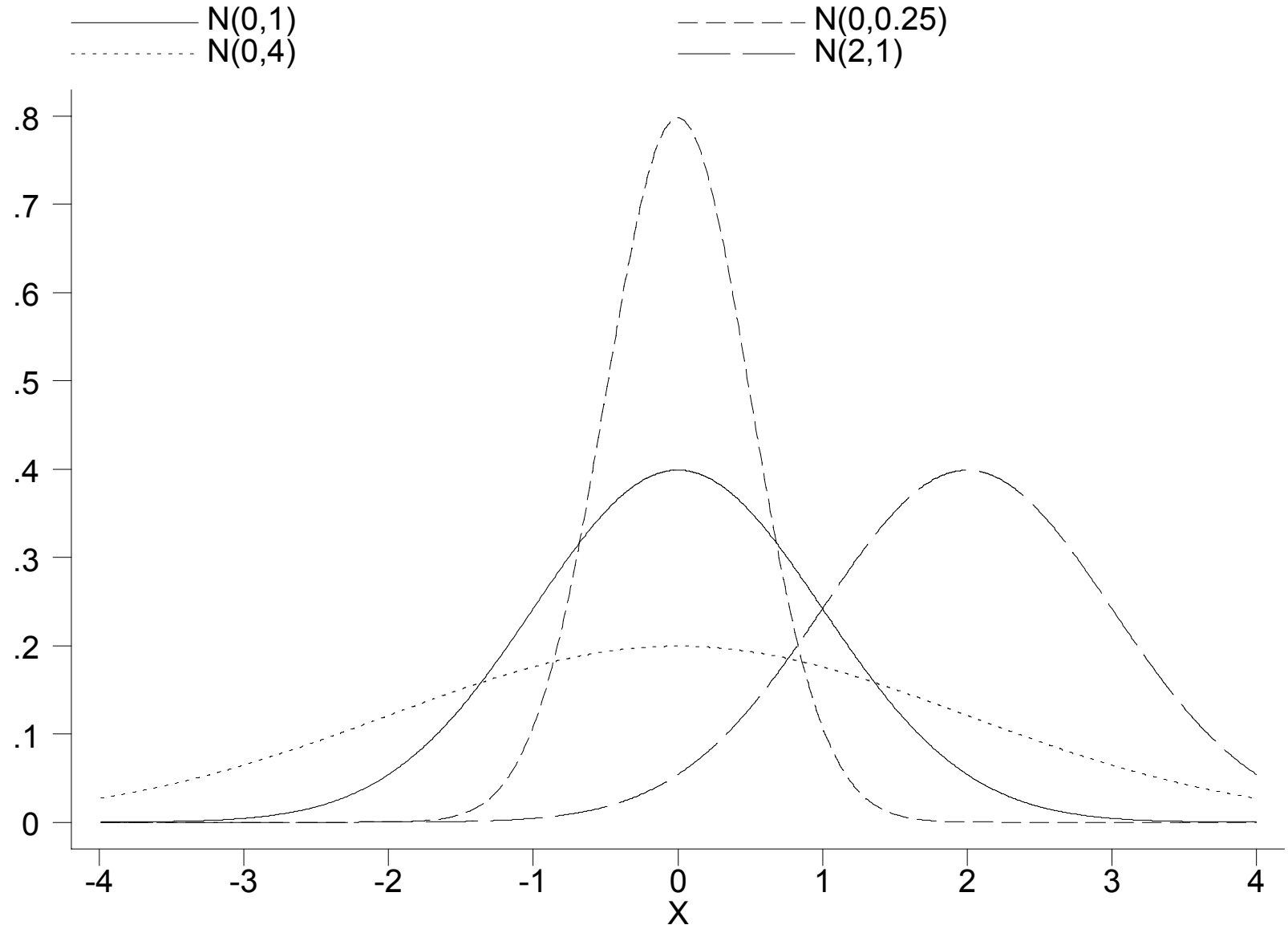
- Die *Normalverteilung* mit den Parametern μ und σ^2 besitzt die Dichte:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

- Ist X normalverteilt mit den Parametern μ und σ^2, so schreibt man auch $X \sim N(\mu,\sigma^2)$
- Für $X \sim N(\mu,\sigma^2)$ gilt: $E(X) = \mu$ und $Var(X) = \sigma^2$
- Die Normalverteilung wird auch als "Gauß-Verteilung" oder "Gaußsche (Glocken-)Kurve" bezeichnet
- N(0,1) ist die *Standardnormalverteilung*, für deren Dichte meist das Symbol $\phi(x)$ verwendet wird

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right)$$

Beispiele für Normalverteilungen



Eigenschaften der Normalverteilung

- Die Normalverteilung besitzt eine große praktische Bedeutung, die insbesondere durch den zentralen Grenzwertsatz ($\rightarrow$ s.u.) begründet ist
- Insbesondere wenn viele zufällige Einflüsse zusammenwirken liegt in der Regel annäherungsweise eine Normalverteilung vor
- Die Normalverteilung $N(\mu,\sigma^2)$ ist symmetrisch um μ
- f(x) ist unimodal und nimmt das Maximum in μ an
- Die Kurve besitzt in $\mu+\sigma$ und $\mu-\sigma$ jeweils einen Wendepunkt
- f(x) nähert sich der x-Geraden für $x\rightarrow+\infty$ und $x\rightarrow-\infty$ asymptotisch an

Verteilungsfunktion der Normalverteilung

- Die *Verteilungsfunktion* der Normalverteilung mit den Parametern μ und σ^2 ist per definitionem:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2} \right) dt$$

- Für die Verteilungsfunktion der Standardnormalverteilung schreibt man auch:

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$

- Die Werte von $\Phi(x)$ lassen sich nicht analytisch berechnen, sondern müssen numerisch approximiert werden; sie sind in Tabellen aufgeführt
- Es gilt (Symmetrieeigenschaft): $\Phi(-x) = 1 - \Phi(x)$

Z-Standardisierung

- Wird eine Zufallsvariable X ~ N(μ,σ^2) transformiert zu:

$$Z = \frac{X - \mu}{\sigma}$$

- so ist Z standardnormalverteilt, d.h. Z ~ N(0,1).

- Insbesondere gilt:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \Phi(z)$$

- Dies ist für Berechnungen praktisch, denn man braucht deshalb nur eine Tabelle für Φ(z)

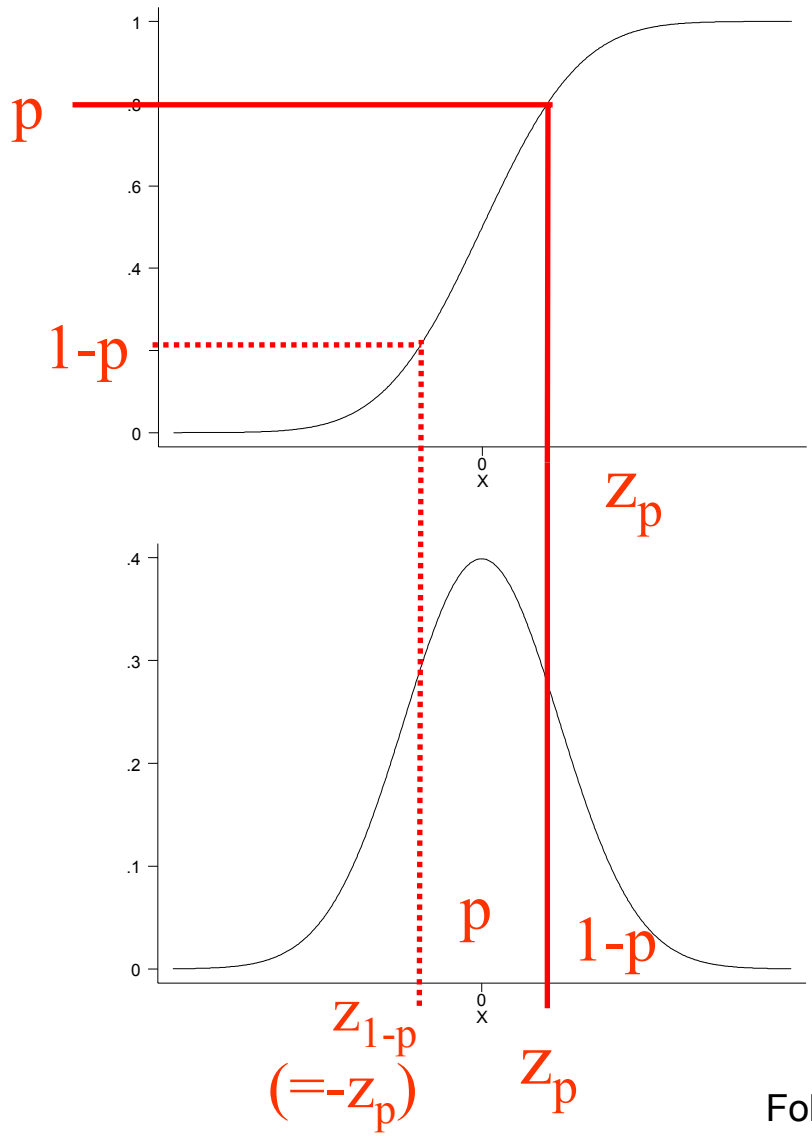
Quantile der Standardnormalverteilung

- Die p-Quantile z_p der Standardnormalvert. sind bestimmt durch:

  $$\Phi(z_p) = p$$

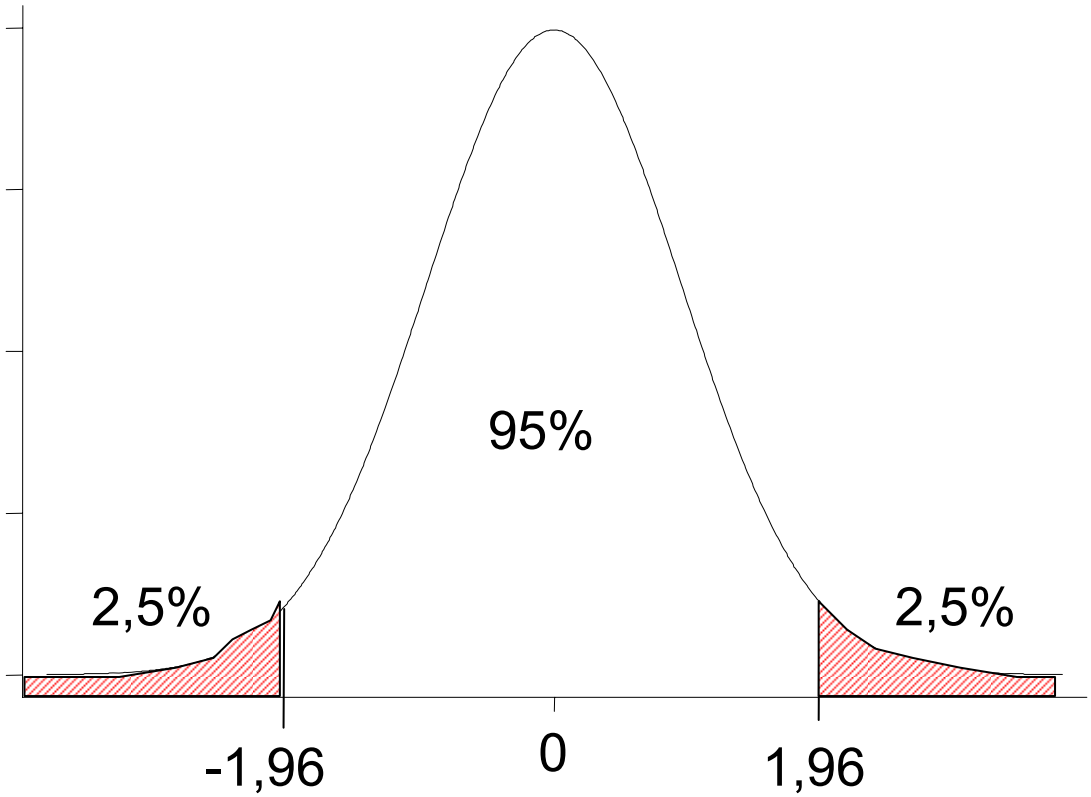
- Das p-Quantil teilt die Fläche unter der Dichte in einen Teil der Größe p (links) und einen Teil der Größe 1-p (rechts)
- Es gilt: $z_p = -z_{1-p}$
- Die p-Quantile x_p von $N(\mu,\sigma^2)$ ergeben sich durch:

  $$x_p = \sigma z_p + \mu$$



Wichtige Quantile der Standardnormalverteilung

p	z_p
.50	0
.75	0.67
.90	1.28
.95	1.64
.975	1.96
.99	2.33
.995	2.58



Beispiel: X~N(1,4)

$x_{0.975} = 1 + 2 \cdot 1.96 \quad = 4.92$

$x_{0.025} = 1 + 2 \cdot (-1.96) \quad = -2.92$

Arbeiten mit der Z-Tabelle

Tabelle A–3.1: Quantile der Normalverteilung

α	z_α	α	z_α	α	z_α	α	z_α	α	z_α
0.000	$-\infty$	0.200	−0.842	0.400	−0.253	0.600	0.253	0.800	0.842
0.005	−2.576	0.205	−0.824	0.405	−0.240	0.605	0.266	0.805	0.860
0.010	−2.326	0.210	−0.806	0.410	−0.228	0.610	0.279	0.810	0.878
0.015	−2.170	0.215	−0.789	0.415	−0.215	0.615	0.292	0.815	0.896
0.020	−2.054	0.220	−0.772	0.420	−0.202	0.620	0.305	0.820	0.915
0.025	−1.960	0.225	−0.755	0.425	−0.189	0.625	0.319	0.825	0.935
0.030	−1.881	0.230	−0.739	0.430	−0.176	0.630	0.332	0.830	0.954
0.035	−1.812	0.235	−0.722	0.435	−0.164	0.635	0.345	0.835	0.974
0.040	−1.751	0.240	−0.706	0.440	−0.151	0.640	0.358	0.840	0.994
0.045	−1.695	0.245	−0.690	0.445	−0.138	0.645	0.372	0.845	1.015
0.050	−1.645	0.250	−0.674	0.450	−0.126	0.650	0.385	0.850	1.036
0.055	−1.598	0.255	−0.659	0.455	−0.113	0.655	0.399	0.855	1.058
0.060	−1.555	0.260	−0.643	0.460	−0.100	0.660	0.412	0.860	1.080
0.065	−1.514	0.265	−0.628	0.465	−0.088	0.665	0.426	0.865	1.103
0.070	−1.476	0.270	−0.613	0.470	−0.075	0.670	0.440	0.870	1.126
0.075	−1.440	0.275	−0.598	0.475	−0.063	0.675	0.454	0.875	1.150
0.080	−1.405	0.280	−0.583	0.480	−0.050	0.680	0.468	0.880	1.175
0.085	−1.372	0.285	−0.568	0.485	−0.038	0.685	0.482	0.885	1.200
0.090	−1.341	0.290	−0.553	0.490	−0.025	0.690	0.496	0.890	1.227
0.095	−1.311	0.295	−0.539	0.495	−0.013	0.695	0.510	0.895	1.254

$Z_{0.85} = 1.036$

$X \sim N(0,16)$

$x_{0.885} = 0+4 \cdot 1.2=4.8$

$\Phi(1.15) = 0.875$

$X \sim N(2,4)$

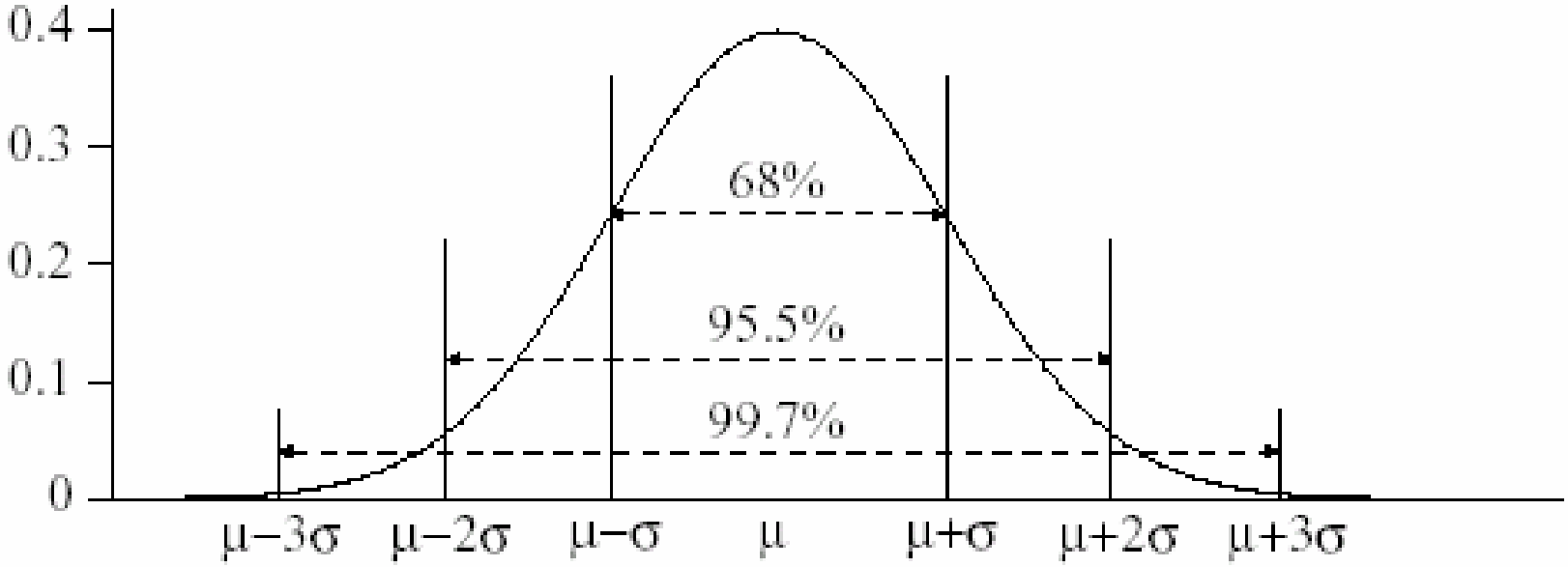
$F(3) = \Phi((3 - 2) / 2)$

$= \Phi(0.5)$

$= 0.69$

Zentrale Schwankungsbreiten

- 68.3% der Beobachtungen liegen zwischen μ-σ und μ+σ
- 95.5% der Beobachtungen liegen zwischen μ-2σ und μ+2σ
- 99.7% der Beobachtungen liegen zwischen μ-3σ und μ+3σ



Einführung in die Statistik, Ben Jann

χ^2-Verteilung

- Sind $X_1,..., X_n$ unabhängige Zufallsvariablen, die alle standardnormalverteilt sind, d.h. $X_i \sim N(0,1)$ für i = 1,...,n.

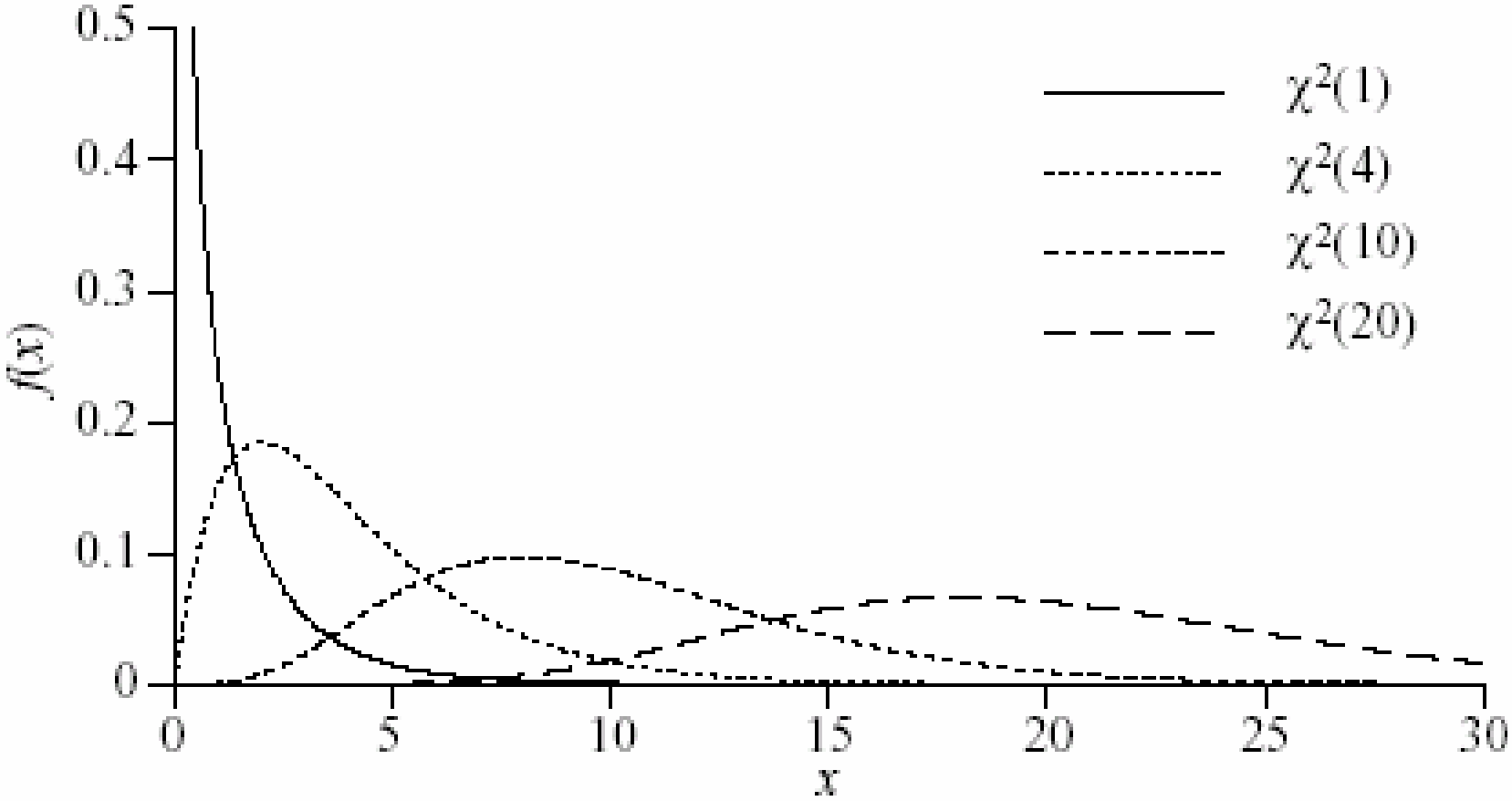
- Dann ist die Zufallsvariable Z

$$Z = X_1^2 + ... + X_n^2$$

Chiquadrat-verteilt mit *Freiheitsgrad n*
(degrees of freedom, df)

- Man schreibt auch $Z \sim \chi^2(n)$
- Es gilt: E(Z) = n und Var(Z) = 2n
- Für n>30 verwendet man die NV als Approximation

χ^2-Verteilung



Einführung in die Statistik, Ben Jann

Arbeiten mit der χ^2-Tabelle

Tabelle A–3.2: Ausgewählte Quantile der Chiquadrat–Verteilung

df	0.5%	1.0%	2.5%	5.0%	10.0%	90.0%	95.0%	97.5%	99.0%	99.5%
1	<.001	<.001	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.60
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09	16.75
6	0.676	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	0.989	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22	28.30
13	3.565	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	29.82
14	4.075	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58	32.80

$\chi^2_{0.95}(1) = 3.84$

$\chi^2_{0.99}(11) = 24.72$

t-Verteilung

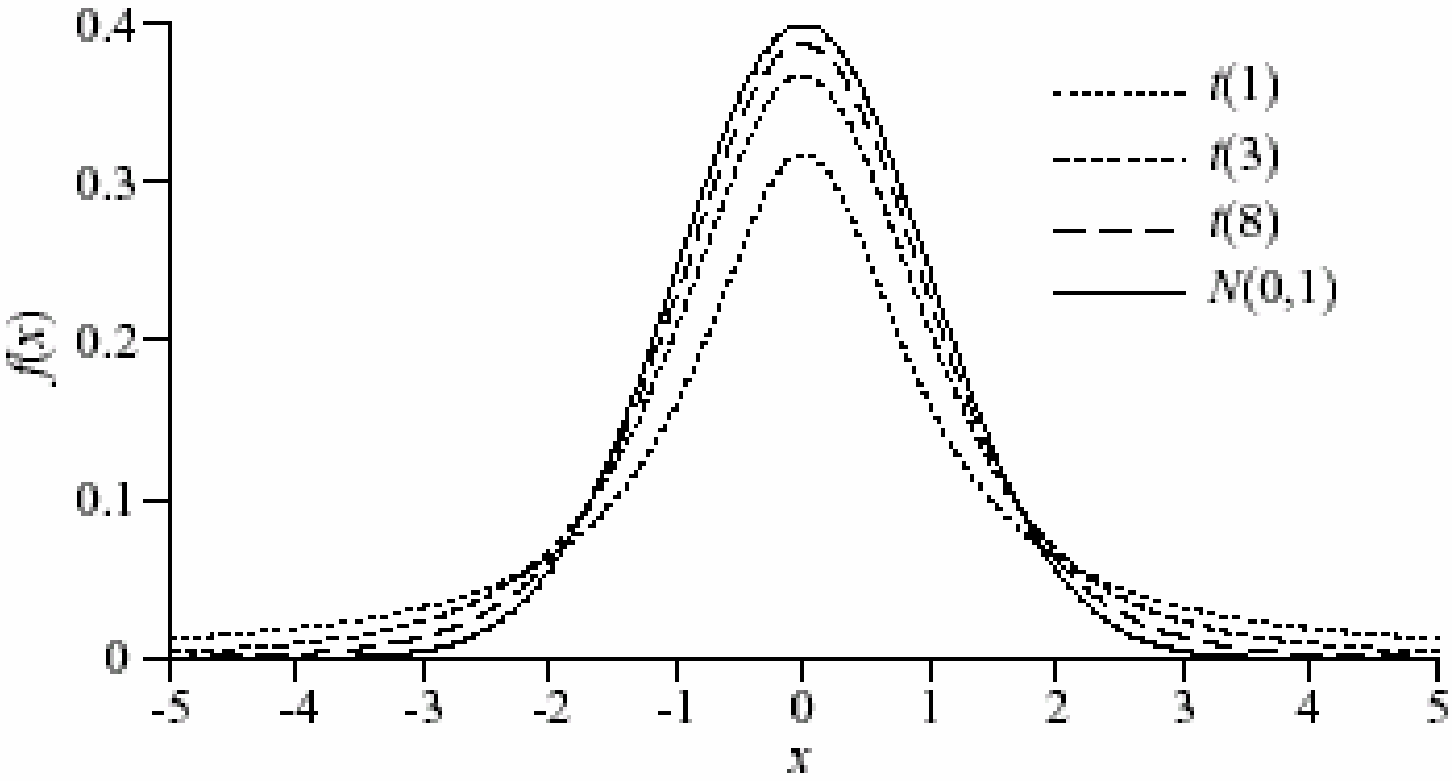
- Sind X und Z unabhängige Zufallsvariablen mit X ~ N(0,1) und Z ~ χ^2(n), so ist die Zufallsvariable

$$T = \frac{X}{\sqrt{\frac{Z}{n}}}$$

t-verteilt bzw. *Student-verteilt mit n Freiheitsgraden*

- Man schreibt auch T ~ t(n)
- Es gilt: E(T) = 0 und Var(T) = n/(n-2), für $n \geq 3$
- Für n>30 verwendet man die NV als Approximation

t-Verteilung



Einführung in die Statistik, Ben Jann

Arbeiten mit der t-Tabelle

Tabelle A–3.3: Ausgewählte Quantile der t–Verteilung

df	75.0%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%	99.95%
1	1.000	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	0.816	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.765	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

$t_{0.975}(1) = 12.71$

$t_{0.975}(30) = 2.04$

$t_{0.975}(\infty) = 1.96$

$t_{0.025}(1) = -t_{1-0.025}(1) = -12.71$

Grenzwertsätze

- Sei X eine Zufallsvariable mit einer Verteilungsfunktion F, Erwartungswert μ und Varianz σ^2
- Der zu X gehörende Zufallsvorgang wird nun n-mal unabhängig voneinander wiederholt, wobei X_i die Zufallsvariable für die i-te Wiederholung ist
- Dann bezeichnet

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  eine neue Zufallsvariable, die den durchschnittlichen Wert von X bei n Wiederholungen angibt
- Es gilt:

$$E(\overline{X}_n) = \mu \quad und \quad Var(\overline{X}_n) = \frac{\sigma^2}{n}$$

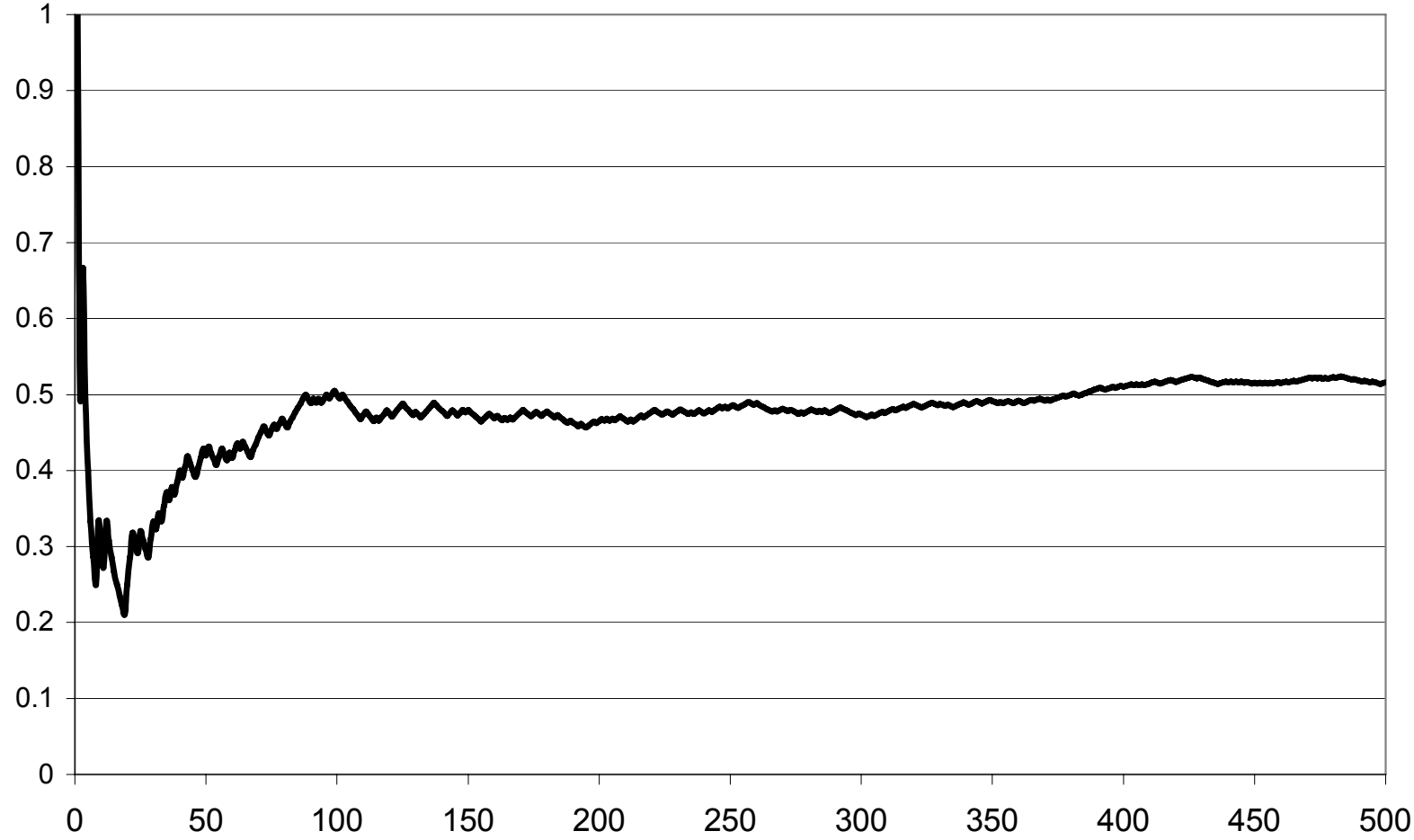
Das Gesetz der großen Zahlen

Für großes n wird somit die Varianz von $\overline{X}_n$ sehr klein, d.h. die Verteilung von $\overline{X}_n$ ist stark um μ konzentriert. Man kann dies so ausdrücken:

$$P(\left|\overline{X}_n - \mu\right| \leq \varepsilon) \to 1 \quad \text{für} \quad n \to \infty$$

- Dabei ist ε ein beliebig kleiner positiver Wert
- d.h.: Die Wahrscheinlichkeit, dass die Variable höchstens um ε vom Erwartungswert abweicht, konvergiert gegen 1, wenn n unendlich groß wird,
- bzw. durch Wahl eines genügend großen n kann der Erwartungswert beliebig nahe approximiert werden.

Beispiel: Wiederholter Münzwurf



Der zentrale Grenzwertsatz

- Betrachten wir nun die unstandardisierte Summe $X_1+\ldots+X_n$. Es gilt

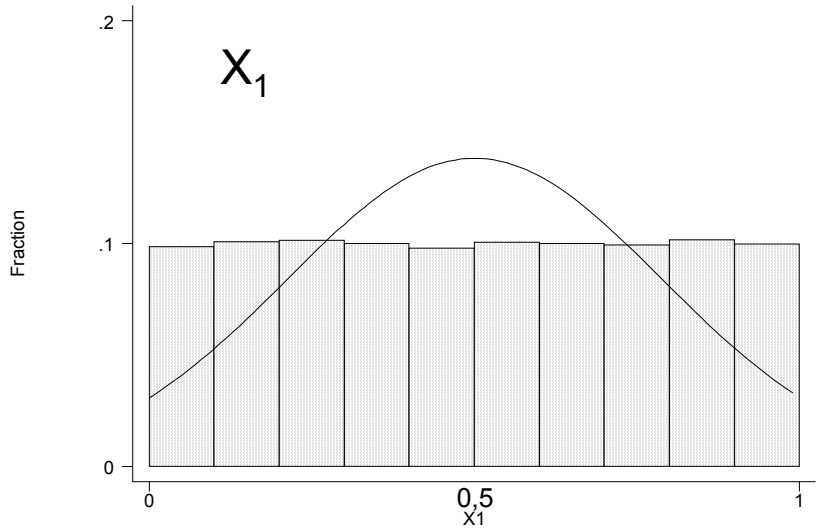
$$E(X_1+\ldots+X_n)=n\mu \quad Var(X_1+\ldots+X_n)=n\sigma^2$$

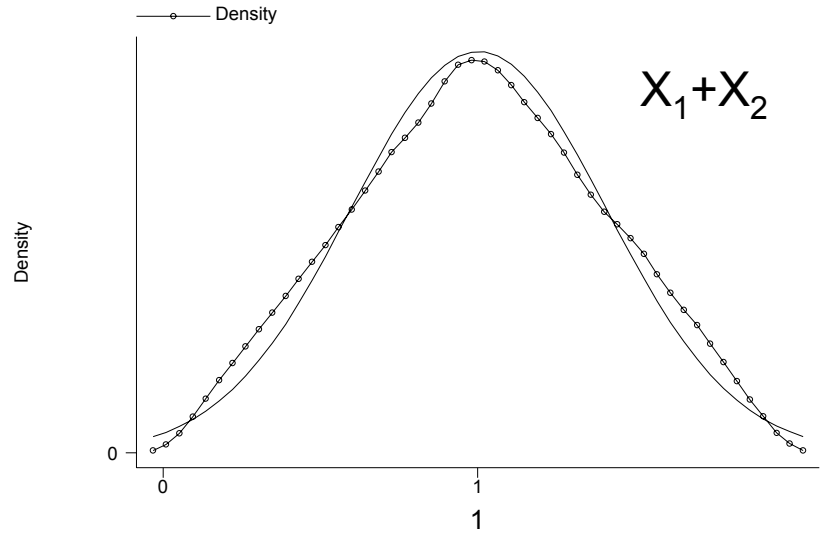
- Der zentrale Grenzwertsatz besagt nun, dass die Summe von unabhängig, identisch verteilten ZV asymptotisch normalverteilt ist:

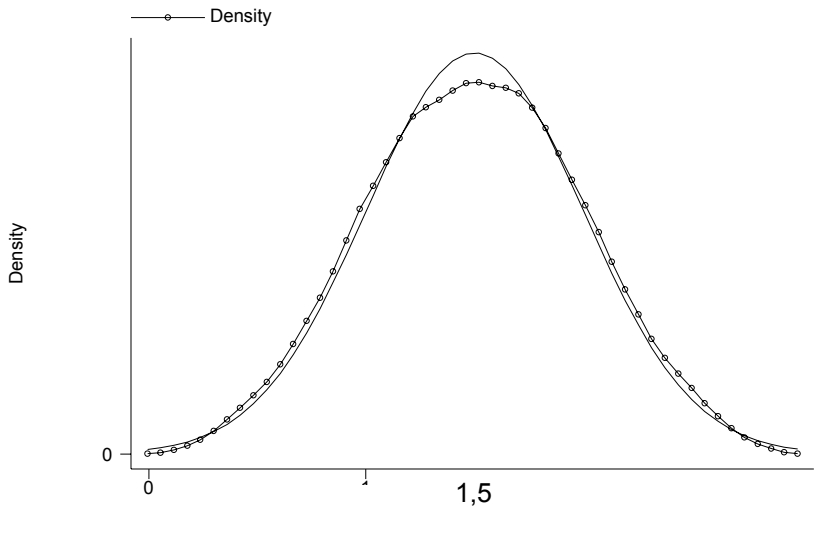
$$X_1+\ldots+X_n \overset{a}{\sim} N(n\mu, n\sigma^2)$$

- Dieser Satz begründet die zentrale Bedeutung der Normalverteilung in der Statistik
  - denn unabhängig von der Verteilung von X, konvergiert die Summe gegen die NV. Die Normalverteilung ist deshalb ein gutes Modell für alle ZV, die durch das Zusammenwirken von vielen kleinen zufälligen Effekten entstanden sind

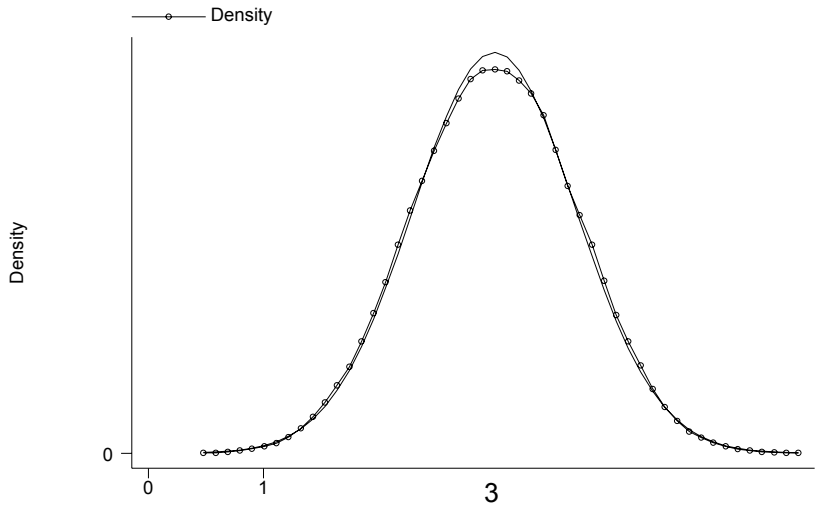
Beispiel: Summe gleichverteilter Variablen







$X_1+X_2+X_3$



$X_1+X_2+X_3+X_4+X_5+X_6$

Kapitel IV
Schätztheorie

Prof. Dr. Josef Brüderl
Universität Mannheim

Frühjahrssemester 2007

Schätzen

- Die interessierende Kenngröße (Parameter) der Grundgesamtheit (GG) ist unbekannt
- Wir ziehen eine Stichprobe (unabhängiges n-maliges Ziehen)
- Anhand dieser Daten können wir den Parameter schätzen
  - Punktschätzung: Mit welcher Formel können wir den Parameter schätzen?
  - Intervallschätzung: In welchem Bereich liegt der GG-Parameter mit hoher Wahrscheinlichkeit?
- Beispiele für zu schätzende Parameter:
  - Anteilswert π
  - Mittelwert μ
  - Varianz σ^2

Punktschätzung

- Mit welcher Formel berechnet man Schätzer für Punktschätzungen?
- θ sei der unbekannte GG-Parameter. Wir ziehen eine Stichprobe vom Unfang n. Als Ergebnis erhält man die Daten $x_1,x_2,...,x_n$, wobei jedes Datum eine Realisation der ZV X_i ist. Dann ist die Schätzfunktion (Schätzstatistik) definiert durch

$$T = g(X_1,X_2,...,X_n).$$

- Den konkreten Schätzer berechnet man aus den Daten durch

$$\hat{\theta} = g(x_1, x_2,..., x_n).$$

Beispiele für Schätzfunktionen

- Für den Mittelwert

$$E(X): \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Für die Varianz (nicht erwartungstreu)

$$Var(X): \widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Für die Varianz (erwartungstreu)

$$Var(X): S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Stichprobenverteilung

- Ein *Schätzwert* ist also die Realisierung eines Zufallsvorgangs, da die Daten aus einer zufälligen Stichprobenziehung resultieren.
- Damit ist ein Schätzer eine *Zufallsvariable* mit einer Wahrscheinlichkeitsverteilung, die man als *Stichprobenverteilung* (Prüfverteilung) bezeichnet.
- Die Stichprobenverteilung hat den Erwertungswert E(T) und die Varianz Var(T).
- Die Standardabweichung

$$\sigma_T = \sqrt{\mathrm{Var}(T)}$$

bezeichnet man auch als den *Standardfehler* des Schätzers.

Beispiel: Schätzung von μ durch $\overline{X}$

- Den Erwartungswert einer Zufallsvariable X in der GG schätzt man mit

$$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

  wobei die Zufallsvariable X_i das Ergebnis des i-ten Zugs für die Stichprobe repräsentiert
- Ist X normalverteilt, so ist die Stichprobenverteilung

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- Sonst gilt laut zentralem Grenzwertsatz

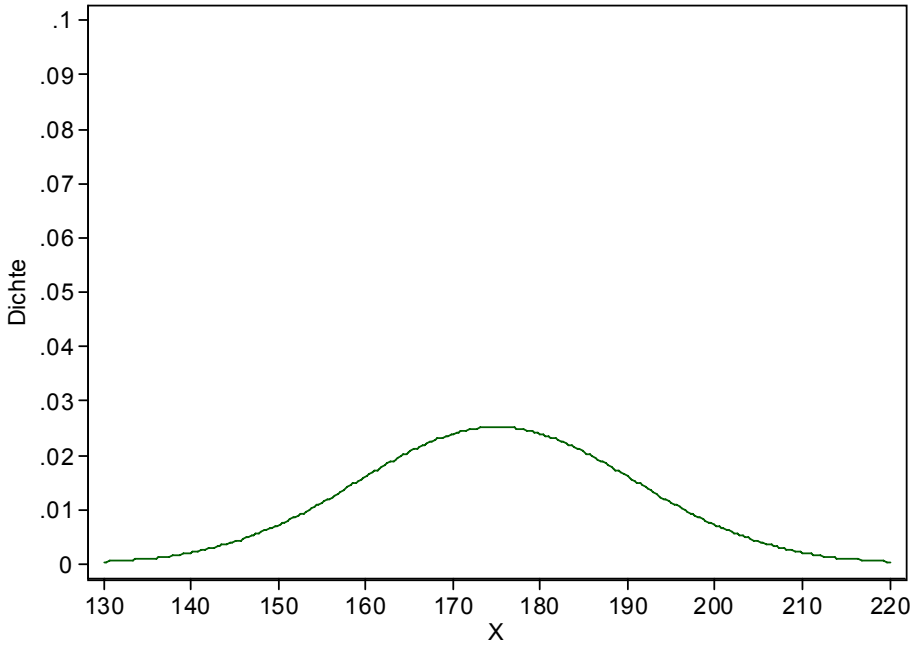
$$\overline{X} \sim^a N(\mu, \frac{\sigma^2}{n})$$

- Ab n>30 ist diese Approximation bereits brauchbar. Die Stichprobenverteilung sagt uns, mit welcher Wahrscheinlichkeit bestimmte Werte von $\overline{X}$ auftreten.

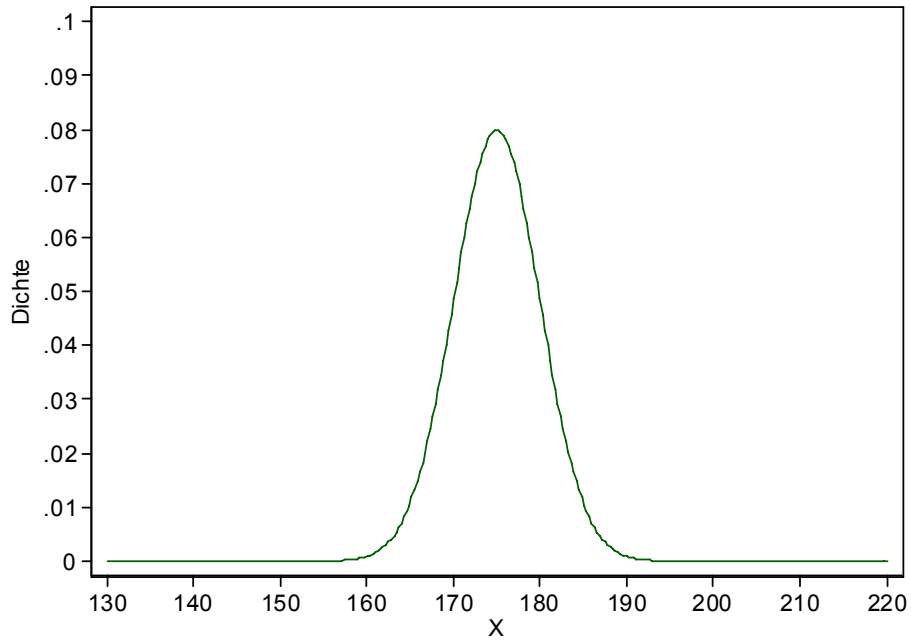
Beispiel: Schätzung der mittleren Körpergröße bei bekannter GG

- Die Körpergröße X in der GG sei bekannt. Weiterhin sei X normalverteilt mit μ = 175 (cm) und $\sigma^2 = 250$
- Damit ist die Stichprobenverteilung $\overline{X} \sim N(175, \frac{250}{n})$
- „Simulation“ für verschiedene n

n=1: $\overline{X} \sim N(175, 250)$



n=10: $\overline{X} \sim N(175, 25)$



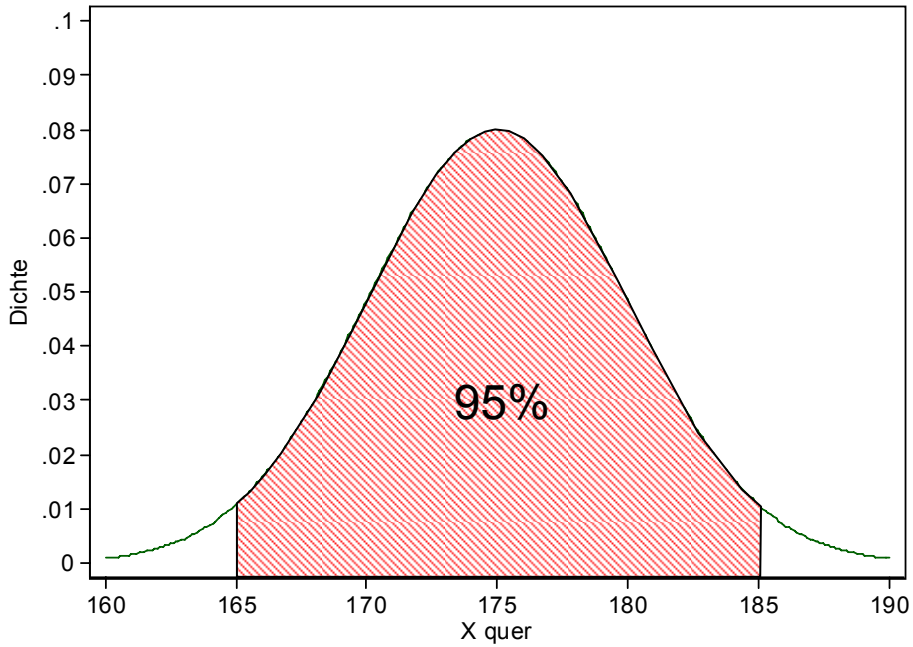
Beispiel: Schätzung der mittleren Körpergröße bei bekannter GG

- 95%-Konfidenzintervall (n=10)
  - In welches (um μ zentrierte) Intervall fallen 95% der Schätzer? Gesucht sind die 2,5% und 97,5% Quantile der Stichprobenverteilung

$$P(z_{0.025} \leq \frac{\overline{X} - 175}{5} \leq z_{0.975}) = 0.95$$

$$\Leftrightarrow P(175 + z_{0.025} \cdot 5 \leq \overline{X} \leq 175 + z_{0.975} \cdot 5) = 0.95$$

  - $-z_{0.025}=z_{0.975}=1.96$.
  - Damit ist KI = [165,2;184,8]



Gütekriterien

- Es sind viele Schätzfunktionen denkbar. Man wählt die "beste" Schätzung anhand von einigen wünschenswerten Eigenschaften aus:
  - Erwartungstreue: $E(T) = \theta$
    „Im Mittel trifft der Schätzer den wahren Wert“
  - Konsistenz: $Var(T) \rightarrow 0$ für $n \rightarrow \infty$
    „Eine größere Stichprobe lohnt sich“
  - Effizienz: Var(T) nimmt den kleinstmöglichen Wert an
    „Nimm den Schätzer mit dem kleineren Fehler“

- Konstruktionsprinzipien für Schätzfunktionen
  Bestimmte Schätzverfahren, die in verschiedensten Schätzsituationen Schätzformeln liefern. Z.B.:
  - Maximum-Likelihood
  - Ordinary-Least-Squares (OLS)

Intervallschätzung

- Der Punktschätzer $\hat{\theta}$ wird den wahren Wert θ im Regelfall nicht exakt treffen. Deshalb ist es sinnvoll, die Präzision des Schätzverfahrens anzugeben
- Eine Möglichkeit dies zu tun bietet der Standardfehler des Schätzers
- Eine andere Möglichkeit ist die Intervallschätzung, bei der man ein *Konfidenzintervall* bestimmt
  - Der wahre Wert des Parameters liegt mit der Wahrscheinlichkeit 1-α (Überdeckungswahrscheinlichkeit) innerhalb des Konfidenzintervalls
  - α gibt hierbei die Irrtumswahrscheinlichkeit an. Übliche Werte für α sind 0,05 und 0,01

Bestimmung des Konfidenzintervalls

- Anders als oben kennen wir jetzt $\hat{\theta}$ aber nicht θ. Wir müssen also aus den Daten eine untere Intervallgrenze G_u und eine obere Grenze G_o schätzen, so dass gilt:

$$P(G_u \leq \theta \leq G_o) = 1 - \alpha.$$

- $[G_u,G_o]$ ist dann das 1 - α-Konfidenzintervall. Dies ist eine Zufallsvariable.
  - Anhand der Daten einer Stichprobe schätzen wir dann das konkrete Konfidenzintervall $[g_u,g_o]$.
  - 1-α ist dann keine Wahrscheinlichkeit, denn der wahre Wert liegt entweder im Konfidenzintervall oder nicht. Man sagt deshalb häufig, "der wahre Wert ist mit der Sicherheit 1-α in $[g_u,g_o]$ enthalten".

Konfidenzintervall für μ ($X \sim N(\mu,\sigma^2)$, σ^2 bekannt)

- $X_1,...,X_n$ seien unabhängige Wiederholungen von $X \sim N(\mu,\sigma^2)$
- Die Schätzfunktion ist $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- Mit der Stichprobenverteilung $$\overline{X} = N(\mu, \frac{\sigma^2}{n}).$$
- Für das 1 – α-Konfidenzintervall müssen wir $G_u = -\overline{x}_{1-\alpha/2}$ und $G_o = \overline{x}_{1-\alpha/2}$ bestimmen.
- Diese Werte sind nicht tabelliert, deshalb verwenden wir die Z-Transformation $$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Konfidenzintervall für μ ($X \sim N(\mu,\sigma^2)$, σ^2 bekannt)

- Damit erhalten wir die Grenzen aus

$$1-\alpha = P(-z_{1-\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2})$$

$$= P(\overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$$

- Damit ist das 1 – α-Konfidenzintervall

$$\overline{X} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

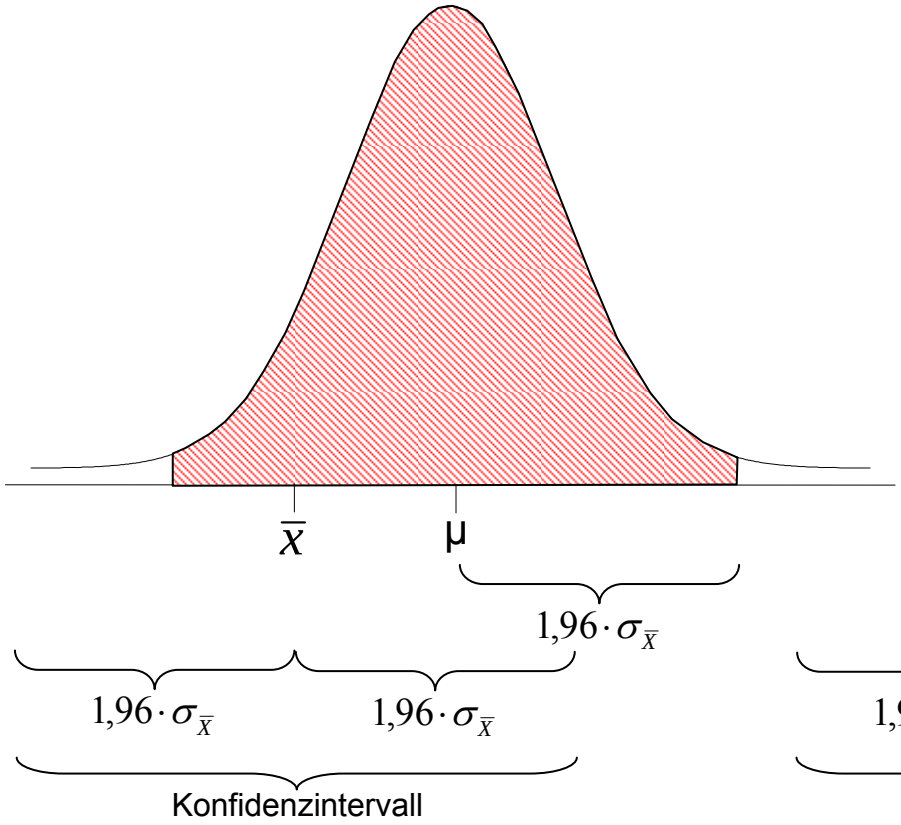
- Die Breite des Konfidenzintervalls nimmt ab
  - mit zunehmendem Stichprobenumfang n
  - mit abnehmender Standardabweichung σ
  - mit zunehmender Irrtumswahrscheinlichkeit α

Konfidenzintervall-“Logik“ (α=5%)

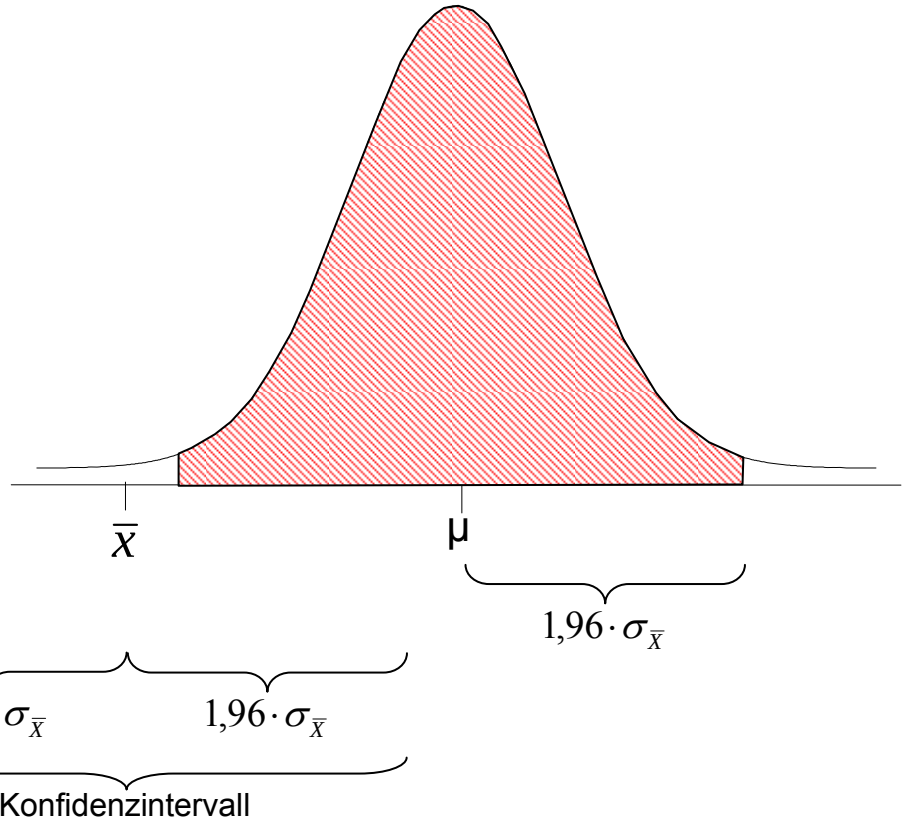
- Das 95%-Konfidenzintervall ist gegeben durch

$$\bar{x} \pm 1{,}96 \cdot \sigma_{\bar{X}} \quad \text{wobei} \quad \sigma_{\bar{X}} = \sigma / \sqrt{n}$$

In 95% der Stichproben



In 5% der Stichproben



Konfidenzintervall für μ ($X \sim N(\mu,\sigma^2)$, σ^2 unbekannt)

- Schätzfunktion und Stichprobenverteilung wie oben
- σ^2 ist nun aber unbekannt. Wir müssen es aus der Stichprobe schätzen. Die Schätzfunktion ist

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

- Es gilt

$$\frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$$

- Damit ist

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

Konfidenzintervall für μ (X~N(μ,σ²), σ² unbekannt)

- Somit finden wir das 1 – α-Konfidenzintervall aus

$$1-\alpha = P(-t_{1-\alpha/2}(n-1) \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t_{1-\alpha/2}(n-1)$$

$$= P(\overline{X} - t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}})$$

- Damit ist das 1-α Konfidenzintervall

$$\overline{X} \pm t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}$$

- Für n>30 kann man die Normalverteilungsapproximation verwenden, d.h. die t-Quantile durch z-Quantile ersetzen.
- Weiterer Fall: X beliebig verteilt
  - Ab n > 30 können wir ebenfalls mit obigen Formeln approximative Konfidenzintervalle berechnen.

Beispiel: Schätzung der mittleren Mathenote

- Erstsemester 2001/02
  - Wir ziehen eine Zufallsstichprobe von n=10 aus der Erstsemesterstudie. Geschätzt werden soll die mittlere Mathenote.
  - Dies sind die Daten: {4,5,4,4,3,4,4,3,3,3}
  - Bestimmen Sie das 95%-Konfidenzintervall.
  - Aus den Daten erhält man:

$$\overline{X} = \frac{1}{10}37 = 3.7, \qquad S^2 = 0.46, \qquad S = 0.67.$$

  - Aus der Tabelle der t-Verteilung kann man ablesen: $t_{0.975}(9) = 2.26$
  - Damit ist

$$2.26\frac{0.67}{\sqrt{10}} = 0.48$$

  - Damit ist das 95%-Konfidenzintervall

    [3.7 - 0.48, 3.7 + 0.48] = [3.22, 4.18]

  - Anmerkung: der wahre Wert ist μ = 3.4

Konfidenzintervall für π

- X sei nun eine Bernoulli-Variable. Wir wollen P(X = 1) = π schätzen. Die Schätzfunktion ist $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Wir wissen, dass $\sum_{i=1}^{n} X_i \sim B(n, \pi).$
- Für n>30 können wir durch die Normalverteilung approximieren. Deshalb lautet die Stichprobenverteilung

$$\overline{X} \sim N(\pi, \frac{\pi(1-\pi)}{n}).$$

- Damit ergibt sich das (approximative) Konfidenzintervall aus $(\overline{X} = \hat{\pi})$

$$1-\alpha \approx P\left(-z_{1-\alpha/2} \le \frac{\hat{\pi}-\pi}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \le z_{1-\alpha/2} \right)$$

$$\approx P\left(\hat{\pi} - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \le \pi \le \hat{\pi} + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right)$$

Beispiel: Wahlprognose

- Welche Partei gewinnt die nächste Wahl? Sie ermitteln in einer Umfrage 52% $(\hat{\pi})$ für Partei A und 48% $(1-\hat{\pi})$ für Partei B. Bestimmt werden soll das 95%-Konfidenzintervall für n=100 und n=2500.
- Bekannt ist: $\hat{\pi} = 0.52, \quad (1-\hat{\pi}) = 0.48, \quad -z_{0.025} = z_{0.975} = 1.96$
- n=100

$$1.96\sqrt{\frac{0.52 \cdot 0.48}{100}} = 0.1$$

  - 95%-KI = [0.52 - 0.1, 0.52 + 0.1] = [0.42, 0.62]
- n=2500

$$1.96\sqrt{\frac{0.52 \cdot 0.48}{2500}} = 0.0196$$

  - 95%-KI = [0.52 - 0.0196, 0.52 + 0.0196] = [0.5004, 0.5396]
- D. h. erst ab n=2500 kann mit 95% Sicherheit ein Wahlsieg der Partei A vorhergesagt werden.

Kapitel V
Testtheorie

Prof. Dr. Josef Brüderl
Universität Mannheim

Frühjahrssemester 2007

Hypothesenprüfung

- Statistik dient nicht nur der Schätzung bestimmter Parameter, sondern auch der Prüfung von Hypothesen
- Diese Hypothesen können oft als Vermutungen über bestimmte statistische Parameter geäußert werden
- Beispiele:
  - $\pi = 0.5$, $\mu = 2.7$, $\pi < 0.4$, $\mu \geq 0$
  - $\pi_1 > \pi_2$, $\mu_1 = \mu_2$, $\mu_1 \neq \mu_2$
- Die Frage ist nun, wie man aufgrund von Stichproben beurteilt, ob diese Hypothesen zutreffen oder nicht
- Da die mit den Stichprobendaten geschätzten Parameter Zufallsvariablen sind, lässt sich keine vollkommene Sicherheit im Hinblick auf die Gültigkeit solcher Hypothesen erzielen
- Man versucht deshalb zu ermitteln, mit welcher „Wahrscheinlichkeit“ eine aufgestellte Hypothese (H_0) gültig ist bzw. mit welcher Wahrscheinlichkeit eine Gegenhypothese (H_1) vorzuziehen wäre

Beispiel: Qualitätsprüfung

- Sie wollen nur Lieferungen mit maximal 10% Ausschuss akzeptieren.
- Sie ziehen eine 1000er Stichprobe und finden 108 Ausschussteile.
- Ist dieses Ergebnis mit ihrer Hypothese $\pi \leq 0{,}1$ vereinbar?
- Ein statistischer Test (approximativer Binomialtest) gibt die Antwort.
- Zuerst muss man die 2 Hypothesen formulieren, über die man eine Entscheidung herbeiführen will

  Nullhypothese H_0: $\pi = 0{,}1$

  Alternativhypothese H_1: $\pi > 0{,}1$

- Dann benötigt man eine Prüfgröße (Teststatistik), die aus den Daten einen Kennwert errechnet, der als Grundlage für den Test dient. In unserem Beispiel ist das Ergebnis jedes Stichprobenzuges die Bernoulli-ZV X_i. Es liegt deshalb nahe, die Summe dieser ZV als Prüfgröße zu verwenden

$$X = \sum_{i=1}^{1000} X_i = 108$$

Beispiel: Qualitätsprüfung

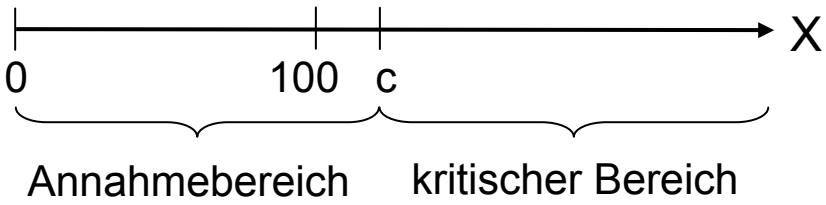
- Diese Prüfgröße ist binomialverteilt X~B(n,π). Da n sehr groß ist, können wir die Normalverteilungs Approximation verwenden

$$X \overset{a}{\sim} N(n\pi, n\pi(1-\pi))$$

- bzw. wenn die H_0 gilt (Prüfverteilung der Teststatistik)

$$X \overset{a}{\sim} N(100,90)$$

- Die Grundidee ist nun die folgende: Werte von X, die zu groß sind (weit über 100 liegen), sprechen eher für H_1. Dazu müssen wir einen kritischen Wert (c) bestimmen, ab dem wir H_1 akzeptieren. Für $x \leq c$ akzeptieren wir dann H_0 (Annahmebereich), für $x > c$ akzeptieren wir H_1 (Ablehnungsbereich, kritischer Bereich).

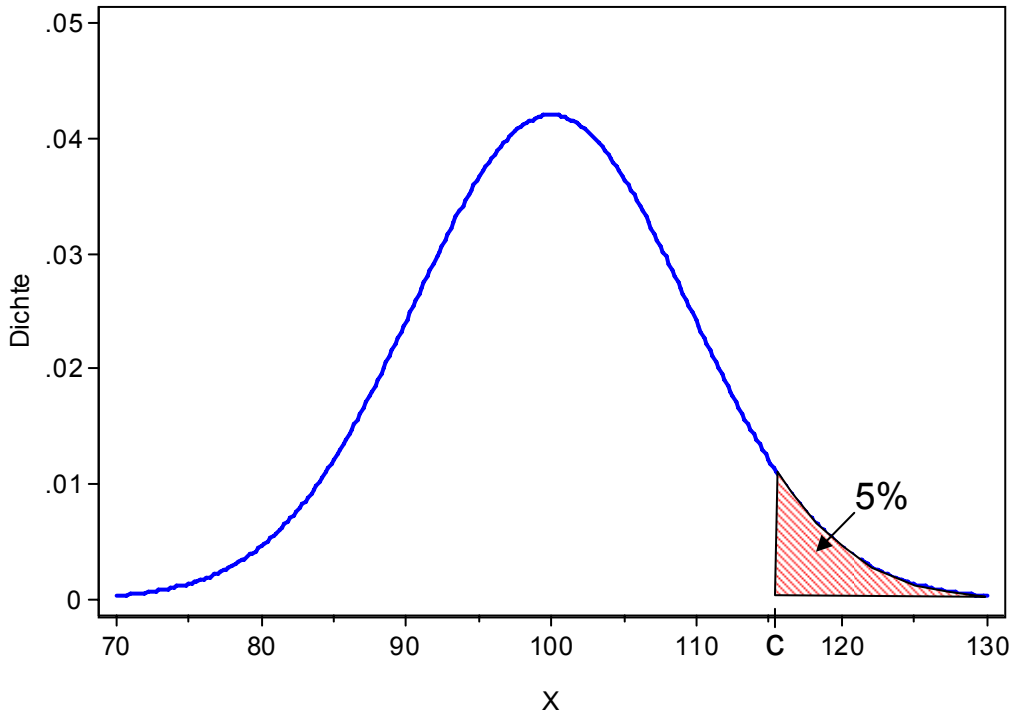


Beispiel: Qualitätsprüfung

- Dabei kann man einen Fehler machen: H_0 ist richtig, aber X liegt zufällig über c (also im kritischen Bereich). Die Whs. dieses Fehlers bezeichnet man als Irrtumswahrscheinlichkeit (Signifikanzniveau α). Wählen wir wie üblich α=0,05, so ergibt sich der kritische Wert aus

$$P_{H_0}(X > c) \leq 0.05$$

- c ist also das $x_{1-\alpha} = x_{0.95}$ Quantil.



Beispiel: Qualitätsprüfung

- Um das 95%-Quantil zu errechnen, standardisieren wir X

$$Z = \frac{X - 100}{\sqrt{90}}.$$

- Das $z_{0,95}$ Quantil ist 1,64. Daraus ergibt sich

$$c = 100 + \sqrt{90} \cdot 1.64 = 115.56.$$

- Der kritische Bereich ist also

$$C = \{x : x > 115\}.$$

- Wir haben aus unserer Stichprobe x = 108 erhalten. Die Prüfgröße liegt also unter dem kritischen Wert, also im Annahmebereich.
- Deshalb akzeptieren wir H_0 und nehmen die Lieferung an.

Ablauf statistischer Tests

1. Formulierung des Modells
2. Festlegung von H_0 und H_1
3. Festlegung des Signifikanzniveaus α
4. Auswahl einer geeigneten Prüfgröße, deren Verteilung unter der Nullhypothese bestimmt werden kann (Prüfverteilung)
5. Bestimmung des kritischen Bereichs
6. Errechnen der Prüfgröße anhand der Daten
7. Entscheidung über die Hypothesen:
   a. Fällt die Prüfgröße in den kritischen Bereich, so wird die Nullhypothese verworfen.
   b. Ansonsten behält man sie bei.

Festlegung von H_0 und H_1

- Statistische Tests „bevorzugen“ die H_0
  - Bei kleinem α müssen die Daten deutlich gegen die H_0 sprechen, bevor wir die H_1 akzeptieren
  - Übliche Werte für α sind 0,05 und 0,01
  - Je kleiner α, desto stärker die „Bevorzugung“ der H_0
- Es hängt vom „Weltbild“ ab, welche Hypothese H_0 wird
  - Qualitätsprüfung: Man vertraut seinen Geschäftspartnern. Deshalb ist „der Lieferant liefert gute Qualität“ die Nullhypothese.
  - Sozialforschung: Meist interessiert man sich für Zusammenhangshypothesen (s. Kap. VI). „Konservativ“ geht man dabei davon aus, dass kein Zusammenhang vorliegt (H_0). Dies verhindert die „leichtfertige“ Annahme einer Zusammenhangshypothese.

    H_0: „es besteht kein Zusammenhang“

    H_1: „es liegt ein Zusammenhang vor“

Entscheidung über die Hypothesen

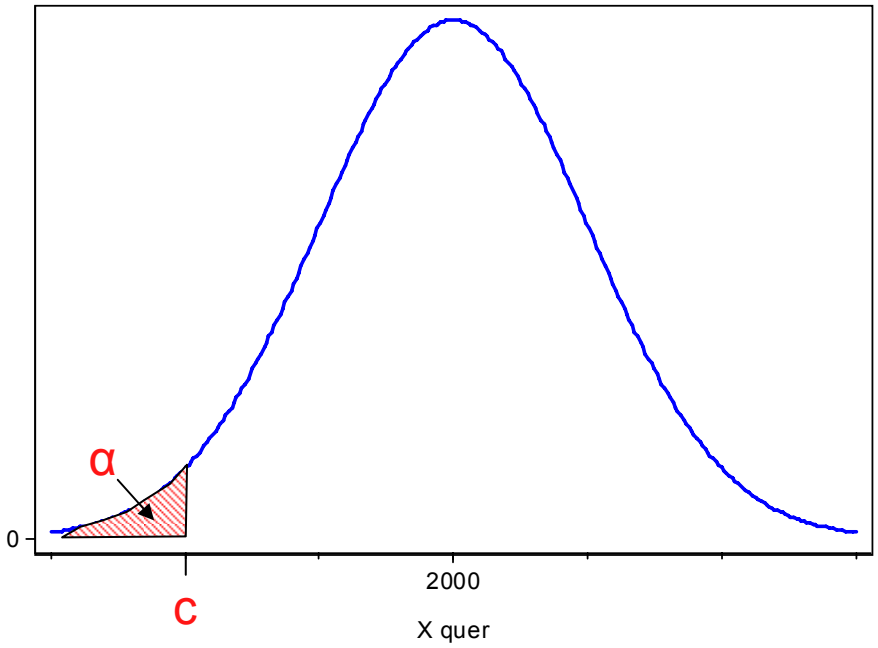
- α ist unbedingt vor Durchführung des Tests festzulegen.
  - Es ist statistisch gesehen Unsinn, nach Errechnung der Prüfgröße α so festzulegen, dass die Prüfgröße in den kritischen Bereich fällt und damit die Zusammenhangshypothese akzeptiert werden kann
- Fällt die Prüfgröße in den Annahmebereich, so sagt man:
  - "Die Nullhypothese kann nicht verworfen werden"
  - Damit geht man als "konservativer" Forscher davon aus, dass kein Zusammenhang besteht
- Fällt die Prüfgröße in den kritischen Bereich, so sagt man:
  - "Die Nullhypothese kann auf dem 5%-Signifikanzniveau verworfen werden". Damit akzeptieren wir (vorläufig) die Zusammenhangshypothese.
  - Man sagt auch "der Zusammenhang ist auf dem 5%-Niveau (statistisch) signifikant".
- Statistische Signifikanz ist nicht gleichbedeutend mit inhaltlicher Relevanz.

Einseitige und zweiseitige Tests

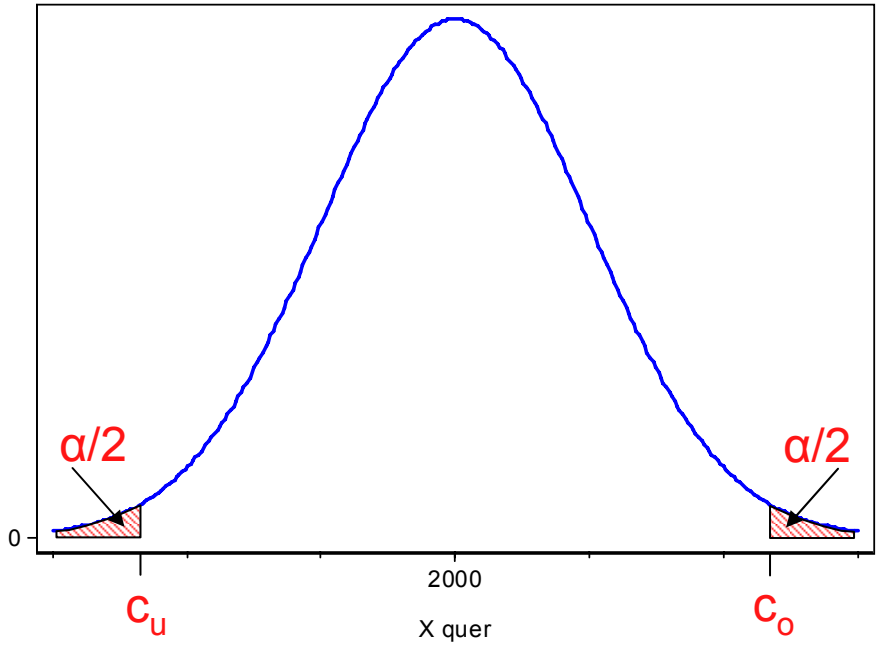
- Sind nur Abweichungen in eine Richtung von Interesse, so spricht man von einem einseitigen Test
  - Bsp. Mittleres Einkommen
    H_0 : „die Mannheimer verdienen im Schnitt mehr als 2000 €“.
    $H_0 : \mu \geq 2000$ gegen $H_1 : \mu < 2000$
    Der kritische Bereich ist damit ein Intervall, das bis c geht, so dass
    $$P_{H_0}(\overline{X} < c) \leq \alpha$$
- Bei zweiseitigen Tests dagegen interessieren Abweichungen nach oben und unten
  - Bsp. Mittleres Einkommen
    H_0 : „die Mannheimer verdienen im Schnitt 2000 €“.
    $H_0 : \mu = 2000$ gegen $H_1 : \mu \neq 2000$
    Der kritische Bereich umfasst damit „zu kleine“ und „zu große“ Werte. Es gibt also einen unteren kritischen Wert (c_u) und einen oberen (c_o), so dass gilt
    $$P_{H_0}(\overline{X} < c_u) = P_{H_0}(\overline{X} > c_o) \leq \alpha / 2$$

Einseitige und zweiseitige Tests

Einseitiger Test



Zweiseitiger Test



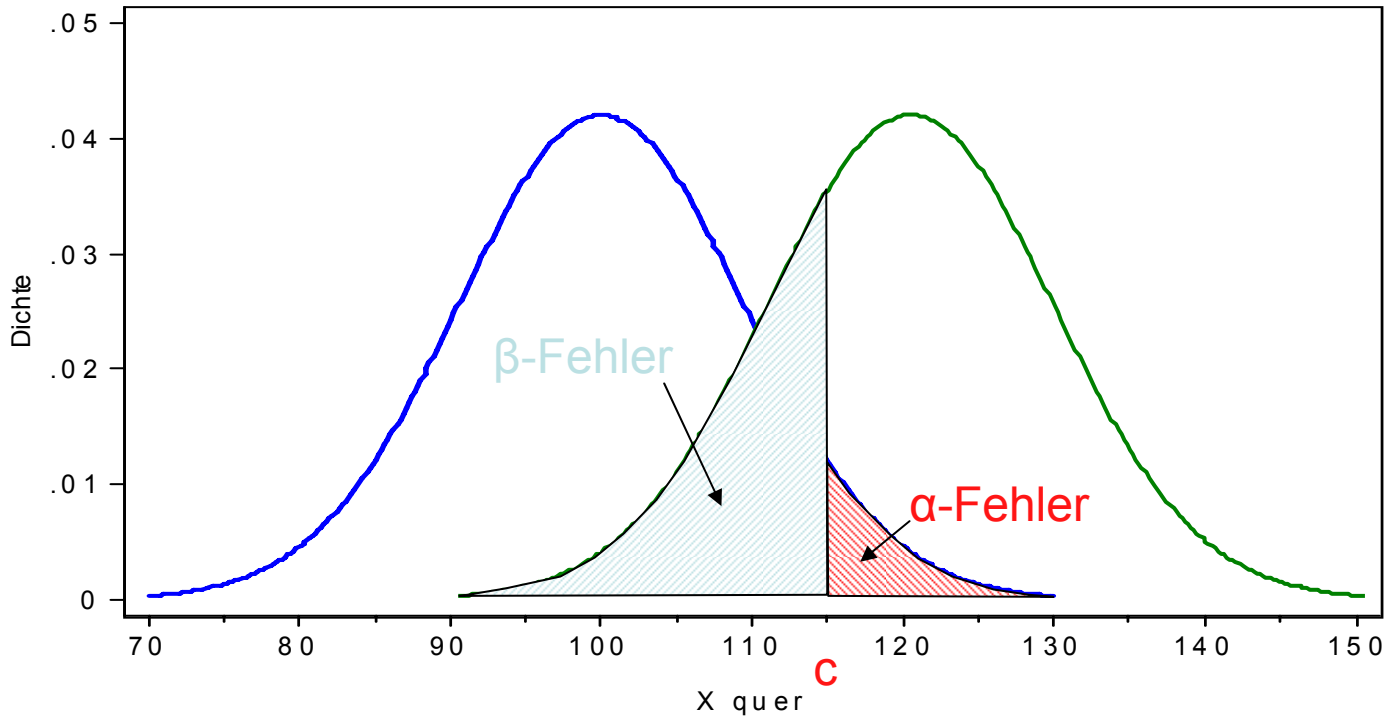
Fehlentscheidungen

- α lässt sich interpretieren, als die Wahrscheinlichkeit, „irrtümlicherweise“ die Nullhypothese H_0 abzulehnen und die Gegenthese H_1 zu akzeptieren
  - Dies wird auch als „Fehler 1. Art“ bzw. „α-Fehler“ bezeichnet
- Daneben gibt es noch einen „Fehler 2. Art“ bzw. „β-Fehler“
  - Irrtümliche Akzeptanz von H_0 (zuungunsten von H_1)

	Entscheidung	
Realität:	für H_0	für H_1
H_0 wahr	richtig	falsch (α-Fehler)
H_1 wahr	falsch (β-Fehler)	richtig

α- und β-Fehler

- Qualitätsprüfungsbeispiel von oben
  - α = 0.05 → c =115
  - Tatsächlich gelte in der GG: π = 0.12 (12% Ausschuss)
  - Wie hoch ist die Whs., dass wir die H_0 trotzdem beibehalten (β-Fehler)? (dass die Teststatistik in den Annahmebereich fällt?)

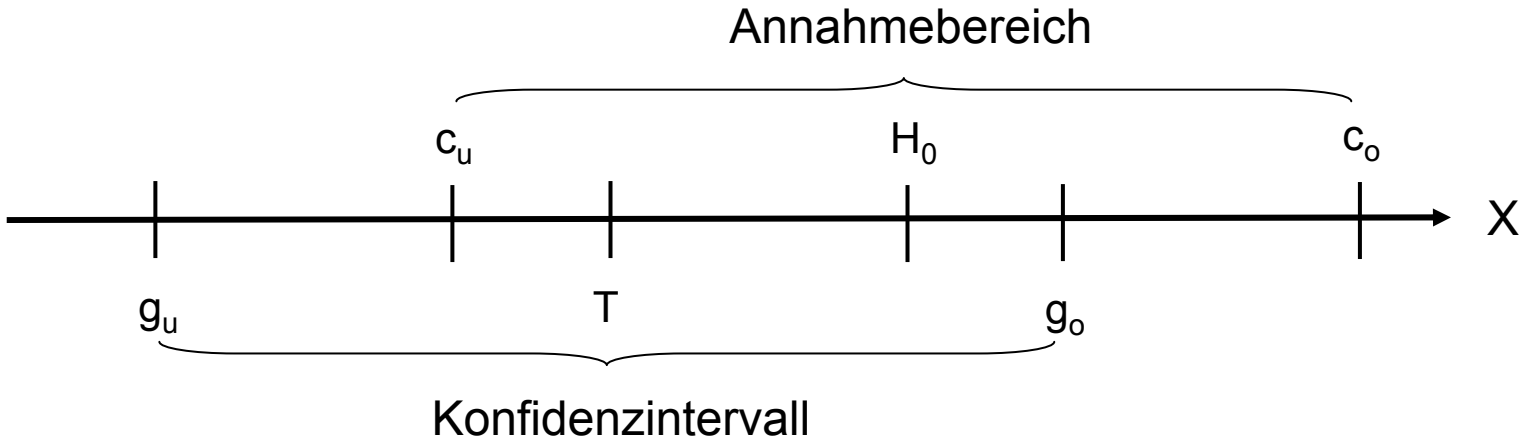


Fehlentscheidungen

- Der Fehler 1. Art wird als wichtiger angesehen, weil er die Annahme einer falschen Zusammenhangshypothese impliziert
- Deshalb werden Tests zu einem vorgegebenem α durchgeführt, womit die Irrtumswahrscheinlichkeit unter Kontrolle ist
- Nicht unter Kontrolle ist die Whs. für den Fehler 2. Art (β)
  - Allerdings verwendet man unter den möglichen Tests jenen, der β minimiert
- Anschauliche Interpretation von α
  - Angenommen in der GG gilt die H_0
  - Würden wir nun 100 Tests der H_0 durchführen, so würden wir in etwa $\alpha \cdot 100$ der Fälle irrtümlicherweise die H_0 verwerfen

Tests und Konfidenzintervalle

- Bei einem zweiseitigen Test sind der Annahmebereich und das 1-α-Konfidenzintervall gleich breit
- Man kann deshalb um den berechneten Wert der Prüfgröße das 1-α-Konfidenzintervall legen
- Wenn dieses Konfidenzintervall den Wert der Nullhypothese enthält wird diese nicht abgelehnt



Test über Mittelwerthypothesen

- Wir haben eine Hypothese über den unbekannten Erwartungswert einer Zufallsvariable X, die normalverteilt ist. Wir erwarten, dass er gleich μ_0 ist, d.h.

$$H_0: \mu = \mu_0 \qquad H_1: \mu \neq \mu_0$$

- Fall 1: σ^2 bekannt: (Gauß-Test)
  - Es liegt nahe, als Teststatistik $\overline{X}$ zu verwenden. Unter der Nullhypothese gilt $\overline{X} \sim N(\mu_0, \frac{\sigma^2}{n})$
  - Da diese Verteilung nicht tabelliert ist, verwenden wir die standardisierte Teststatistik $Z = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$
  - welche standardnormalverteilt ist. Zum Signifikanzniveau α ergibt sich damit der kritische Bereich als

$$C = \{Z: Z < z_{\alpha/2}\} \cup \{Z: Z > z_{1-\alpha/2}\}$$

kurz: $|Z| > z_{1-\alpha/2}$

Test über Mittelwerthypothesen

- Fall 2: σ^2 unbekannt (t-Test)
  - Die Teststatistik Z kommt nicht in Frage. Es liegt nahe, die unbekannte Varianz durch die Stichprobenvarianz S^2 zu schätzen
  - Als Teststatistik ergibt sich:

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n} \sim t(n-1)$$

  - Der kritische Bereich ergibt sich damit als:

$$|T| > t_{1-\alpha/2}(n-1)$$

  - Die H_0 wird also abgelehnt, falls T zu große, oder zu kleine Werte annimmt
  - Für n > 30 kann man wieder die Approximation durch die Normalverteilung verwenden, d.h. die t-Quantile durch die z-Quantile ersetzen

Beispiel: Mathenote der Erstsemester

- Wir wollen zum 5%-Signifikanzniveau die Hypothese testen, dass die mittlere Mathenote 3,2 beträgt:

$$H_0 : \mu = 3{,}2 \qquad H_1 : \mu \neq 3{,}2$$

- Wir ziehen eine Zufallsstichprobe von n=10 aus der Erstsemester-studie (SS 1999). Das sind die Daten: {3, 1, 3, 3, 3, 3, 3, 4, 3, 2}. Aus den Daten erhält man

$$\overline{X} = \frac{1}{10}28 = 2{,}8 \quad S^2 = \frac{1}{9}5{,}6 \quad S = 0{,}79$$

- Damit können wir die Teststatistik errechnen

$$T = \frac{2{,}8 - 3{,}2}{0{,}79}\sqrt{10} = -1{,}60$$

- Aus der Tabelle der t-Verteilung kann man ablesen $t_{0,975}(9) = 2{,}26$
- Da |-1,60| < 2,26 können wir die Nullhypothese nicht ablehnen, die Abweichung des Stichprobenmittels von 3,2 ist "zu klein".
- Wir gehen also weiterhin davon aus, dass die mittlere Mathenote in der GG 3,2 beträgt.

Beispiel: Mathenote der Erstsemester

- Was wäre, wenn alles gleich bleibt, aber n=100 ist?
- Dann wäre

$$T = \frac{2{,}8 - 3{,}2}{0{,}79}\sqrt{100} = -5{,}06$$

- Das z-Quantil lesen wir ab als $z_{0,975} = 1{,}96$
- Da |-5,06| > 1,96 können wir nun die Nullhypothese verwerfen
- Obwohl die Abweichung auch hier 0,4 beträgt, ist die nun "groß" genug, um die Nullhypothese verwerfen zu können. Grund hierfür ist der höhere Stichprobenumfang. Bei n=10 ist eine Abweichung von 0.4 noch mit einer Whs. größer α möglich. Bei n=100 ist eine solche Abweichung dagegen sehr unwahrscheinlich

Vergleich zweier Mittelwerte (t-Test)

- Wir vergleichen im Folgenden die Verteilung eines Merkmals in zwei Gruppen (X,Y)
- Entscheidend ist dabei, dass die beiden Stichproben unabhängig sind
- Die Stichprobenumfänge können verschieden sein, etwa n für X und m für Y
- Hypothesen: $H_0 : \mu_x = \mu_y$ $H_1 : \mu_x \neq \mu_y$
- Für den Test verwendet man die Differenz D der beiden Stichprobenmittel
- Unter der H_0 erwartet man dann, dass D gleich 0 ist
- Diese These über D kann dann prinzipiell wie im Ein-Stichprobenfall geprüft werden, allerdings muss man dazu noch die Varianz kennen

Vergleich zweier Mittelwerte (t-Test)

- Bei Unabhängigkeit von X und Y gilt:

$$Var(D) = Var(\overline{X} - \overline{Y}) = Var(\overline{X}) + Var(\overline{Y}) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$$

- 1. Fall: σ^2_x und σ^2_y bekannt
  - Hier verwendet man als Teststatistik

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0,1)$$

  - Z ist also unter der Nullhypothese standardnormalverteilt
  - Große Abweichungen von 0 führen somit zur Ablehnung der H_0
  - Der genaue Ablehnungsbereich ergibt sich durch:

$$|Z| > z_{1-\alpha/2}$$

Vergleich zweier Mittelwerte (t-Test)

- 2. Fall: σ^2_x und σ^2_y unbekannt, aber $\sigma^2_x = \sigma^2_y$
  - Man ersetzt die unbekannten Varianzen durch S^2_X und S^2_Y
  - Die Teststatistik ist dann:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{(n-1)S_x^{\,2} + (m-1)S_y^{\,2}}{n+m-2}}} \sim t(n+m-2)$$

  - Große Abweichungen von 0 führen wieder zur Ablehnung der H_0
  - Der genaue Ablehnungsbereich ergibt sich durch:

$$|T| > t_{1-\alpha/2}(n+m-2)$$

- 3. Fall: σ^2_x und σ^2_y unbekannt, aber $\sigma^2_x \neq \sigma^2_y$
  - Die Teststatistik ist hier $T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^{\,2}}{n} + \frac{S_y^{\,2}}{m}}} \sim t(k)$
  - Die Zahl der Freiheitsgrade k ergibt sich aus einer komplizierten Formel (Satterthwaite Formel)
  - Deshalb nicht klausurrelevant!

Beispiel: Vergleich mittlere Punktzahl im Mathetest

- Wir vergleichen die Punktezahlen von n=10 Sowis und m=8 Erzw.
  Sowi (X): 9 8 7 5 7 5 6 7 15 16
  Erzw. (Y): 6 7 5 7 6 5 6 12
- Aus diesen Daten erhalten wir

$$\overline{X} = 8{,}50 \quad S_X^2 = 15{,}13$$

$$\overline{Y} = 6{,}75 \quad S_Y^2 = 5{,}06$$

- Daraus errechnen wir die Teststatistik unter der Annahme, dass die Varianzen gleich sind

$$T = \frac{8{,}50 - 6{,}75}{\sqrt{\left(\frac{1}{10} + \frac{1}{8}\right)\frac{9 \cdot 15{,}13 + 7 \cdot 5{,}06}{16}}} = \frac{1{,}75}{1{,}55} = 1{,}13$$

- Aus der t-Tabelle lesen wir ab $t_{0,975}(16) = 2{,}12$
- Somit können wir die H_0 nicht verwerfen
- Wir müssen weiterhin davon ausgehen, dass Sowis und Erziehungswissenschaftler die gleiche mittlere Punktzahl im Mathetest haben

Kapitel VI
Bivariate Datenanalyse

Prof. Dr. Josef Brüderl
Universität Mannheim

Frühjahrssemester 2007

Bivariate Datenanalyse

- Besteht zwischen zwei Variablen ein Zusammenhang?
  - Zusammenhangsanalyse, oder „anspruchsvoller“
  - Kausalanalyse: X → Y
    - X: unabhängige Variable (uV), Y: abhängige Variable (aV)
    - Kausalanalyse ist das vorrangige Ziel jeder Wissenschaft
- Dazu betrachtet man bivariate (zweidimensionale) Verteilungen und berechnet Kennzahlen
- Fragestellungen:
  - Wie groß ist ein Zusammenhang (Beziehung, Kontingenz, Assoziation)?
  - In welche "Richtung" geht er (mind. ordinalskalierte Merkmale)
- Die Zusammenhangsmaße unterscheiden sich für nominalskalierte, ordinalskalierte und metrische Variablen

Die Kontingenztabelle (Kreuztabelle)

Beispiel: Sonntagsfrage in Abhängigkeit vom Wohnort (ALLBUS 1994)
(absolute Häufigkeiten)

Wahlabsicht	alte Bundesländer	neue Bundesländer	Total
Union	533	159	692
SPD	596	260	856
FDP	135	65	200
Bündnis 90/Grüne	225	91	316
PDS	4	116	120
andere Partei	83	31	114
Total	1576	722	2298

Kreuztabelle: bedingte Häufigkeiten (Spaltenprozente)

Wahlabsicht	alte Bundesländer	neue Bundesländer	Total
Union	33.82 %	22.02 %	30.11 %
SPD	37.82 %	36.01 %	37.25 %
FDP	8.57 %	9.00 %	8.70 %
Bündnis 90/Grüne	14.28 %	12.60 %	13.75 %
PDS	0.25 %	16.07 %	5.22 %
andere Partei	5.27 %	4.29 %	4.96 %
Total	100.00 %	100.00 %	100.00 %

Kreuztabelle: Notation

n	die Anzahl der Untersuchungseinheiten
Y, X	zwei Merkmale (Variablen)
$(y_1,x_1), (y_2,x_2), ..., (y_n,x_n)$	die 'Urliste', d.h. die gemeinsamen Ausprägungen von X und Y für jeden Fall
$a_1, a_2, ..., a_k$	die in der Urliste vorkommenden Ausprägungen von Y ($k \leq n$)
$b_1, b_2, ..., b_m$	die in der Urliste vorkommenden Ausprägungen von X ($m \leq n$)

Kreuztabelle: absolute Häufigkeiten

	b_1	...	b_m	
a_1	h_{11}	...	h_{1m}	$h_{1\bullet}$
a_2	h_{21}	...	h_{2m}	$h_{2\bullet}$
...	...	...	...	...
a_k	h_{k1}	...	h_{km}	$h_{k\bullet}$
	$h_{\bullet 1}$	...	$h_{\bullet m}$	n

wobei:

$h_{ij} = h(a_i,b_j)$ die absolute Häufigkeit der Kombination (a_i,b_j) für alle i=1,...,k und alle j=1,...,m

$h_{i\bullet}=h_{i1}+h_{i2}+...+h_{im}$ die Randhäufigkeit (Zeilensumme) der Ausprägung a_i für i=1,...,k

$h_{\bullet j}=h_{1j}+h_{2j}+...+h_{kj}$ die Randhäufigkeit (Spaltensumme) der Ausprägung b_j für j=1,...,m

Kreuztabelle: Spaltenprozente

	b_1	...	b_m	
a_1	$f_Y(a_1\|b_1)$	...	$f_Y(a_1\|b_m)$	$f_{1\bullet}$
a_2	$f_Y(a_2\|b_1)$	...	$f_Y(a_2\|b_m)$	$f_{2\bullet}$
...	...	...	...	...
a_k	$f_Y(a_k\|b_1)$	...	$f_Y(a_k\|b_m)$	$f_{k\bullet}$
	1	...	1	1

wobei:

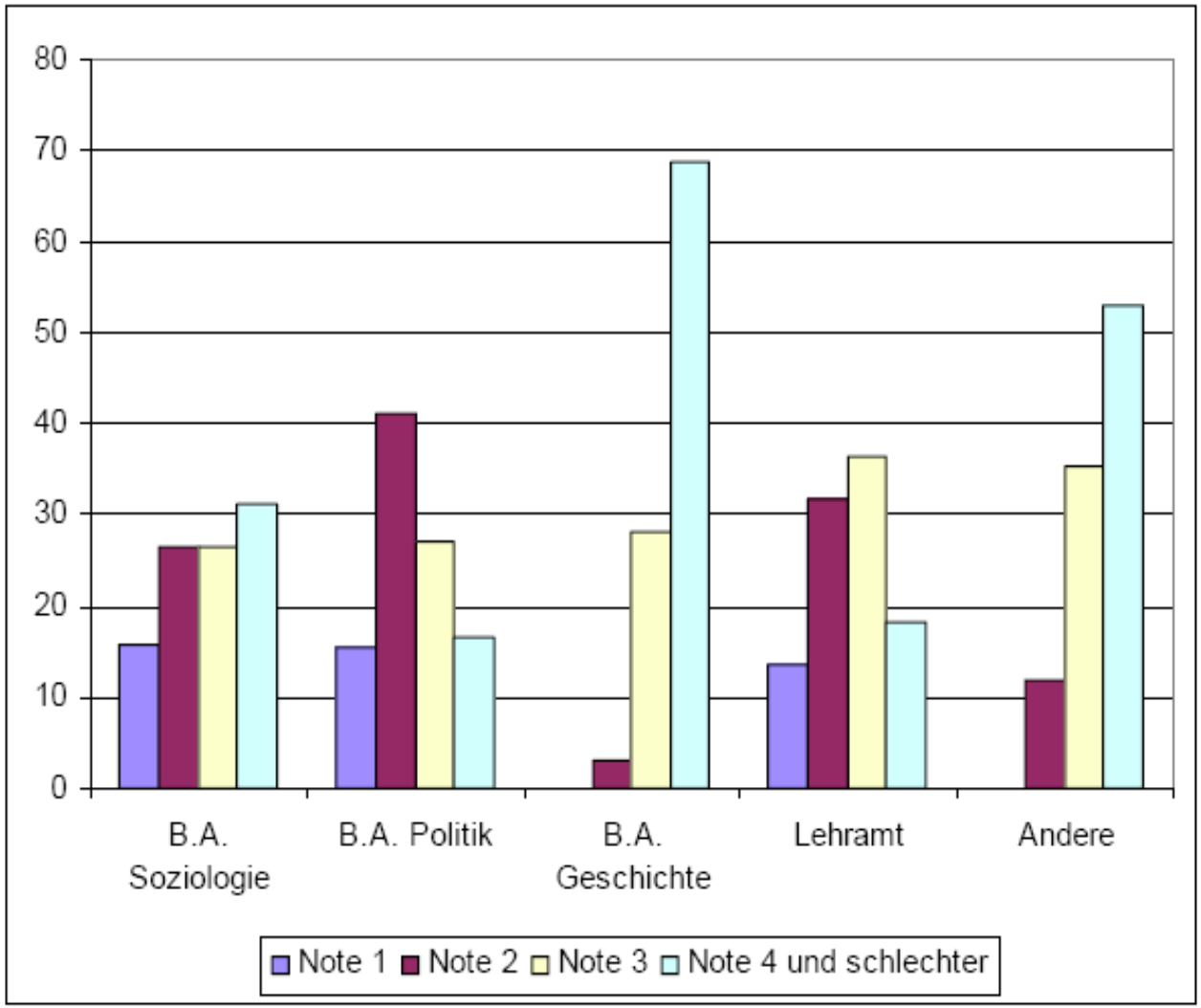
$f_Y(a_i|b_j) = h_{ij}/h_{\bullet j}$ die bedingte Häufigkeit von a_i unter der Bedingung $X=b_j$ für alle i=1,...,k und alle j=1,...,m

$f_{i\bullet}=f_{i1}+f_{i2}+...+f_{im}$ die relative Randhäufigkeit der Ausprägung a_i für i=1,...,k

Konventionen für die Konstruktion einer Kreuztabelle

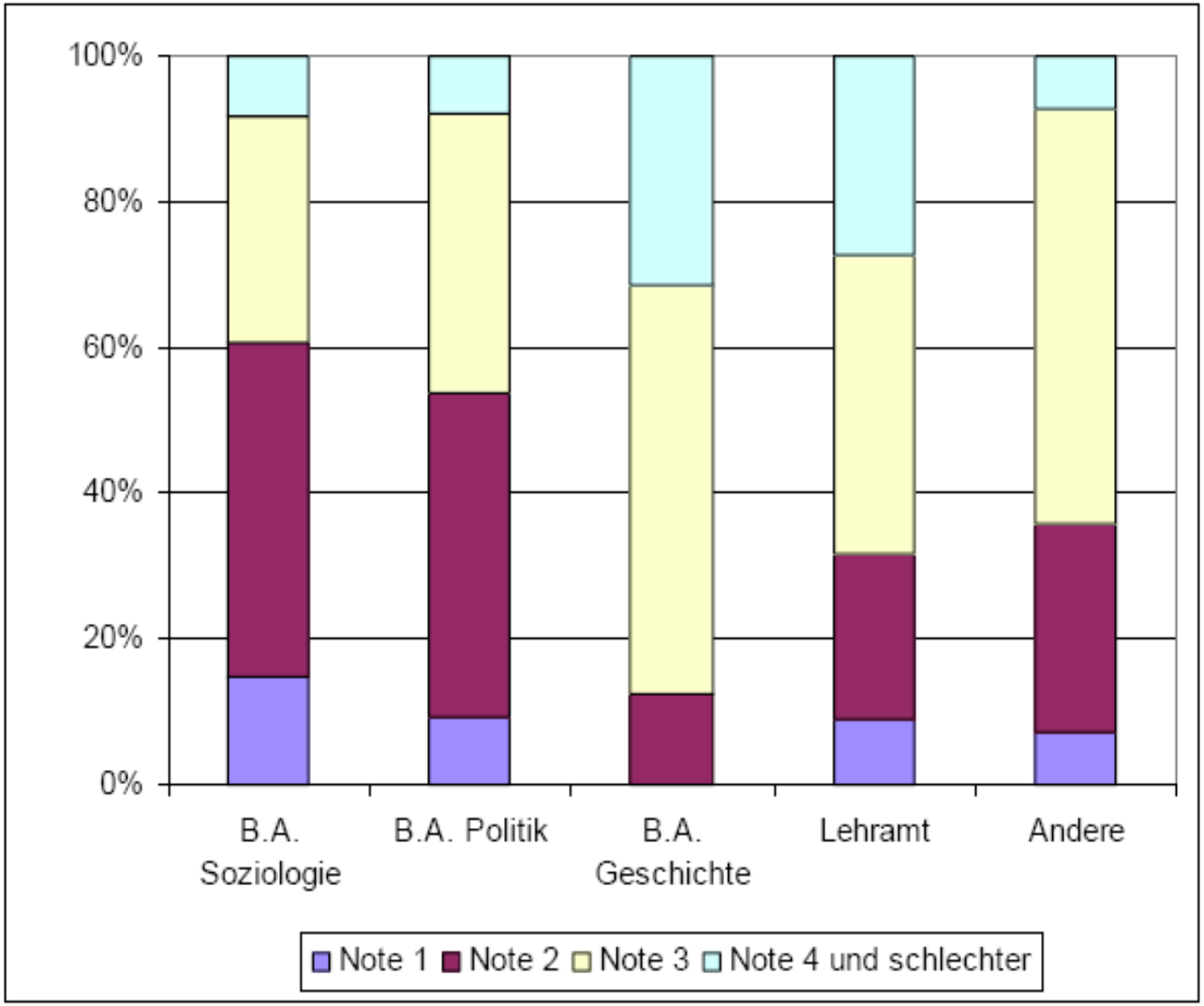
- Die Variable, von der man annimmt, dass sie die andere "verursacht", ist die unabhängige Variable (uV, X). Die andere Variable ist die abhängige Variable (aV, Y)
- Die Entscheidung zwischen uV und aV muss aus der Theorie, mit der man arbeitet, begründet werden
- Regel 1
  - Platziere die uV in den Spalten und die aV in den Zeilen
- Regel 2
  - Berichte die absoluten und die bedingten Häufigkeiten in Richtung der Spalten
- Regel 3
  - Zur Interpretation vergleiche die Prozentsätze in jeder Zeile. Wenn sich die Prozentsätze unterscheiden, besteht ein Zusammenhang zwischen uV und aV

Graphische Darstellung: Gruppiertes Balkendiagramm



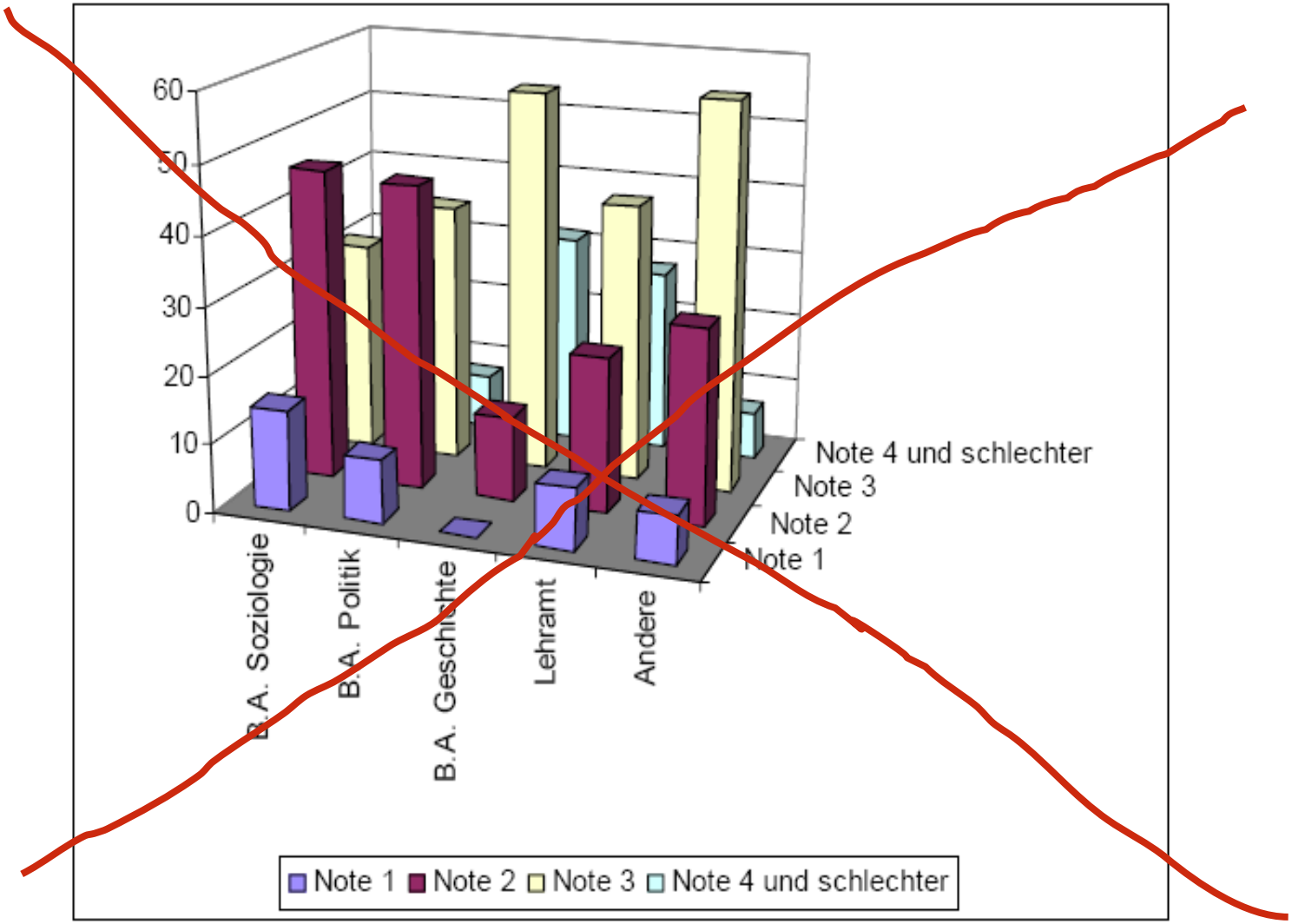
Mathenote im Abitur nach Studiengang
VL „Datenauswertung“ SS 2006

Graphische Darstellung: Gestapeltes Säulendiagramm



Erwartete Note in der Klausur nach Studiengang
VL „Datenauswertung“ SS 2006

Graphische Darstellung: Chart-Junk



Erwartete Note in der Klausur nach Studiengang
VL „Datenauswertung“ SS 2006

Das Konzept der statistischen Abhängigkeit

- Ein "statistischer Zusammenhang" bzw. "statistische Abhängigkeit" zwischen zwei Variabeln X und Y besteht, wenn die bedingten Verteilungen f_Y der einen Variable (Y) für verschiedene Werte der anderen Variable (X) unterschiedlich sind
- Sind die bedingten Verteilungen gleich, so liegt kein Zusammenhang bzw. statistische Unabhängigkeit vor

Unabhängigkeit

		X	
		1	2
Y	1	40	40
	2	10	10
		50	50

sehr starker Zusammenhang

		X	
		1	2
Y	1	40	10
	2	10	40
		50	50

perfekter Zusammenhang

		X	
		1	2
Y	1	50	0
	2	0	50
		50	50

Zusammenhangsmaße für (2x2)-Tabellen

		X		
		1	2	
Y	1	h_{11}	h_{12}	$h_{1\bullet}$
	2	h_{21}	h_{22}	$h_{2\bullet}$
		$h_{\bullet 1}$	$h_{\bullet 2}$	n

(2x2)-Tabelle
(Vierfeldertafel)

Die Prozentsatzdifferenz:

$$\%D = \frac{h_{11}}{h_{\bullet 1}} - \frac{h_{12}}{h_{\bullet 2}}$$

Das Odds-ratio
(Kreuzproduktverhältnis):

$$OR = \frac{\frac{h_{11}}{h_{21}}}{\frac{h_{12}}{h_{22}}} = \frac{h_{11}}{h_{21}} \cdot \frac{h_{22}}{h_{12}}$$

Das Odds-Ratio

- Relative Chance (Verhältnis der Odds)
  - OR=1: Chancen in beiden Gruppen gleich
  - OR>1: Chance in Gruppe X=1 höher
  - OR<1: Chance in Gruppe X=2 höher

Kommt aus dem Glücksspiel
(Verhältnis der Gewinnchancen)

	Spiel 1	Spiel 2
Gewinn	20	10
Verlust	100	100
Odds	20/100=0,2	10/100=0,1
Odds-ratio	0,2/0,1 = 2	

Auch in den Medien häufig berichtet
Beispiel PISA 2003 (fiktive Daten, OR real)

	Akademiker kinder	Arbeiter kinder
Abitur	60	33
Kein Abitur	40	67
Odds	60/40 = 1,5	33/67 = 0,5
Odds-ratio	1,5/0,5 = 3	

Odds sind keine Wahrscheinlichkeiten! Insofern sind OR keine Wahrscheinlichkeits-verhältnisse. Z.B.: die Chance der Akademikerkinder ist 3mal so hoch.

Beispiele

Unabhängigkeit

		X = 1	X = 2
Y	1	40	40
	2	10	10
		50	50

%D = 0
OR = 1

schwacher Zusammenhang

		X = 1	X = 2
Y	1	40	30
	2	10	20
		50	50

%D = 0,2
OR = 2,67

sehr starker Zusammenhang

		X = 1	X = 2
Y	1	40	10
	2	10	40
		50	50

%D = 0,6
OR = 16

perfekter Zusammenhang

		X = 1	X = 2
Y	1	50	0
	2	0	50
		50	50

%D = 1
(OR = ∞)

Beispiel: Erstsemesterstudie WS 2002/03

Mathe als Prüfungsfach?	weiblich	männlich	Total
ja	95	40	135
	67.86 %	*76.92 %*	*70.31 %*
nein	45	12	57
	32.14 %	*23.08 %*	*29.69*
Total	140	52	192
	100.00 %	*100.00 %*	*100.00%*

%D = 95/140 - 40/52 = 67.86% - 76.92% = -9.06 Prozentpunkte

OR = (95/45)/(40/12) = 0.63

χ^2 basierte Maße

- Allgemeine Idee hinter dem Wert χ^2:
  - wenn man die Randhäufigkeiten beider Variablen kennt:
  - wie müssten die absoluten Häufigkeiten der Kombinationen verteilt sein, damit kein Zusammenhang vorliegt
  - Abweichung von diesen "erwarteten" Häufigkeiten wird als Maßzahl für den Zusammenhang genommen

Welche Zellbesetzungen sind bei Unabhängigkeit zu erwarten?

Y \ X	1	2	
1	?	?	120
2	?	?	30
	50	100	150

Unabhängigkeit =

Bedingte Häufigkeiten unterscheiden sich nicht

Y \ X	1	2	
1	40	80	120
2	10	20	30
	50	100	150

Y \ X	1	2	
1	0,8	0,8	0,8
2	0,2	0,2	0,2
	1	1	1

Definition von χ^2

	b_1	...	b_m	
a_1	?	...	?	$h_{1\bullet}$
a_2	...	...	...	$h_{2\bullet}$
...	...	$\tilde{h}_{ij}$	...	...
a_k	?	...	?	$h_{k\bullet}$
	$h_{\bullet 1}$	...	$h_{\bullet m}$	n

Bedingte Häufigkeit = Bedingte Randhäufigkeit

$$\frac{\tilde{h}_{ij}}{h_{i\bullet}} = \frac{h_{\bullet j}}{n} \Rightarrow \tilde{h}_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n}$$

- Das Maß χ^2 misst nun die Abweichung der tatsächlich beobachteten Häufigkeiten von den erwarteten Häufigkeiten
- Dabei werden die Abweichungen pro Zelle quadriert, normiert, und über alle Zellen der Matrix aufsummiert

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

Beispiel: Erstsemesterstudie WS 2002/03

Mathe als Prüfungsfach? Beobachtete Häufigkeiten:			
	weibl.	männl.	Σ
ja	95	40	135
nein	45	12	57
Σ	140	52	192

Beobachtete Kreuztabelle

Erwartete Häufigkeiten:			
	weibl.	männl.	Σ
ja	(135*140)/192 = 98,44	(135*52)/192 = 36,56	135
nein	(57*140)/192 = 41,56	(57*52)/192 = 15,44	57
Σ	140	52	192

Indifferenztabelle

$$\chi^2 = \frac{(95-98.44)^2}{98.44} + \frac{(40-36.56)^2}{36.56} + \frac{(45-41.56)^2}{41.56} + \frac{(12-15.44)^2}{15.44}$$

$$= \frac{11.8336}{98.44} + \frac{11.8336}{36.56} + \frac{11.8336}{41.56} + \frac{11.8336}{15.44} = 0.1202 + 0.3237 + 0.2847 + 0.7664$$

$$= 1.495$$

Eigenschaften von χ^2

Für (2×2)-Kreuztabellen gilt: $\chi^2 = \frac{n(h_{11}h_{22} - h_{21}h_{12})^2}{h_{\bullet 1}h_{\bullet 2}h_{1\bullet}h_{2\bullet}}$

Im Beispiel: $\chi^2 = \frac{192(95 \cdot 12 - 40 \cdot 45)^2}{140 \cdot 52 \cdot 135 \cdot 57} = 1.493$

- χ^2 ist symmetrisch (d.h. man kann X und Y vertauschen)
- χ^2 ist invariant gegenüber Vertauschung von Zeilen oder Spalten (Anordnung ist bei Nominalskala willkürlich)
- Bei Unabhängigkeit gilt $\chi^2 = 0$
- χ^2 wächst mit n, deshalb gibt es keine generelle (!) Obergrenze: n'=10•n => $\chi'^2 = 10 \bullet \chi^2$
- Bei (2×2)-Tabellen: $\chi^2 = n$ bei perfektem Zusammenhang

Aus χ^2 abgeleitete Korrelationskoeffizienten

- Man normiert χ^2 auf sein Maximum, damit man Assoziationskoeffizienten im Intervall [0,1] erhält
- Phi (nur bei (2 × 2)-Tabellen)

$$\phi = \sqrt{\frac{\chi^2}{n}} = \frac{h_{11}h_{22} - h_{21}h_{12}}{\sqrt{h_{\bullet 1}h_{\bullet 2}h_{1\bullet}h_{2\bullet}}}$$

- Cramers V (bei (k × m)-Tabellen)

$$V = \sqrt{\frac{\chi^2}{n \bullet \min(m-1, k-1)}}$$

- Phi ist offensichtlich ein Spezialfall von Cramers V für k = m = 2
- Beispiel „Mathe als Prüfungsfach“

$$\phi = \sqrt{\frac{1.493}{192}} = 0.09$$

Beispiele

Unabhängigkeit

Y \ X	1	2	
1	40	40	80
2	10	10	20
	50	50	

$\Phi = 0$

schwacher Zusammenhang

Y \ X	1	2	
1	40	30	70
2	10	20	30
	50	50	

$\Phi = 0{,}22$

sehr starker Zusammenhang

Y \ X	1	2	
1	40	10	50
2	10	40	50
	50	50	

$\Phi = 0{,}60$

perfekter Zusammenhang

Y \ X	1	2	
1	50	0	50
2	0	50	50
	50	50	

$\Phi = 1$

χ^2-Unabhängigkeitstest

- Der χ^2-Unabhängigkeitstest testet, ob in einer Kreuztabelle ein signifikanter Zusammenhang vorliegt
- Die Hypothesen lauten:
  - H_0 : „X und Y sind unabhängig“
  - H_1 : „X und Y sind abhängig“
- Die unter der H_0 zu erwartenden Zellbesetzungen sind

$$\tilde{h}_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n}$$

- Weichen die beobachteten h_{ij} deutlich von den "erwarteten" Zellenhäufigkeiten ab, so spricht das gegen die H_0
- Als Teststatistik bietet sich an

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} \sim \chi^2((k-1)\cdot(m-1))$$

- Die H_0 wird abgelehnt, falls: $\chi^2 > \chi^2_{1-\alpha}((k-1)\cdot(m-1))$

Beispiele

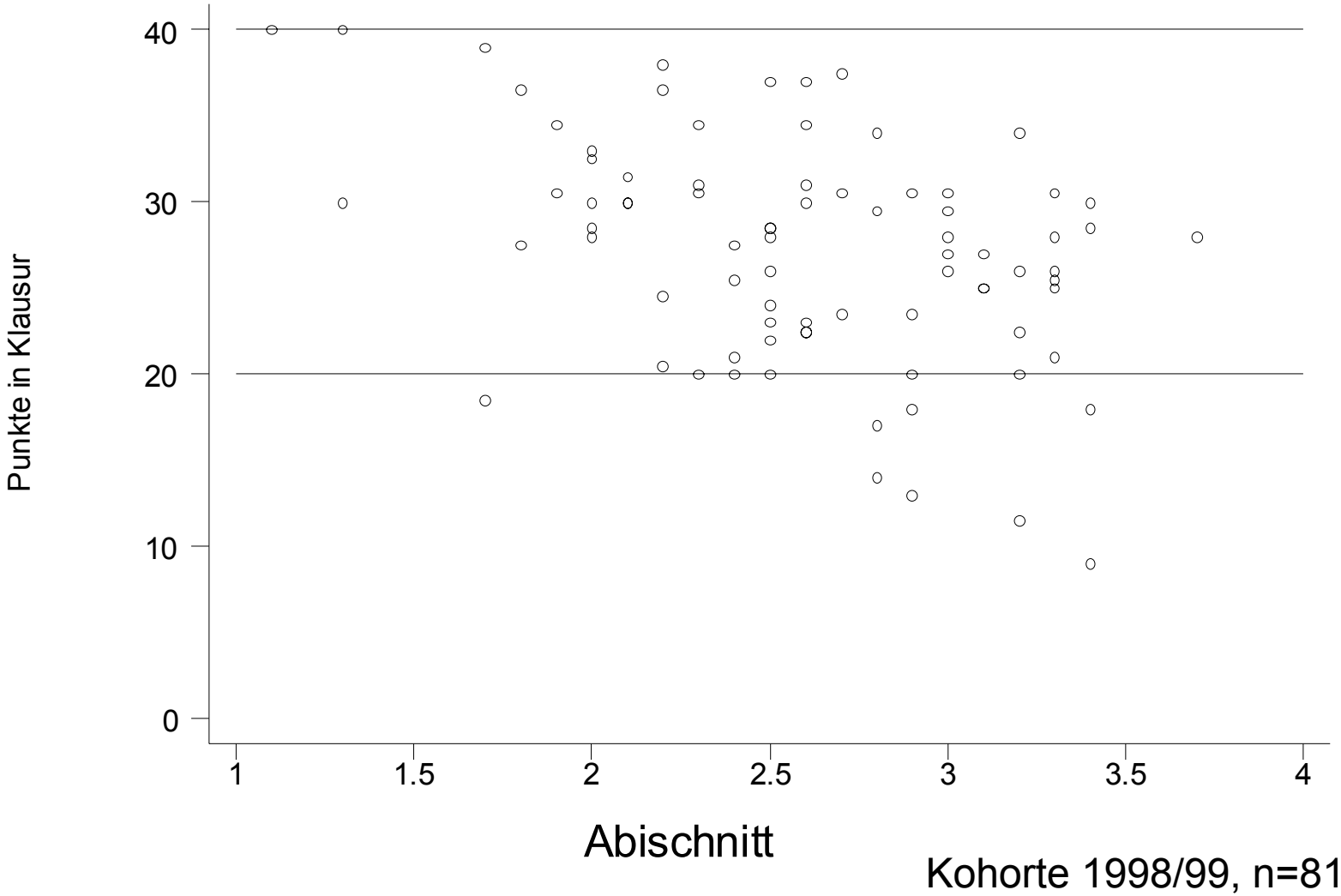
- Mathe als Prüfungsfach (α=0,05)
  - $\chi^2 = 1{,}493$ (s.o.)
  - $\chi^2_{0,95}$ (1·1) = 3,84 (s. Tabelle)
  - Damit können wir die H_0 nicht verwerfen. Es besteht kein signifikanter Zusammenhang zwischen Geschlecht und Mathe als Prüfungsfach
- Sonntagsfrage in Abhängigkeit vom Wohnort (α=0,05)
  - $\chi^2 = 262{,}5$ (mit Stata berechnet)
  - $\chi^2_{0,95}$ ((6-1)·(2-1)) = 11,07 (s. Tabelle)
  - Damit können wir die H_0 verwerfen. Es besteht ein signifikanter Zusammenhang zwischen Wohnort und Parteipräferenz
  - $$V = \sqrt{\frac{262{,}5}{2298 \bullet \min(6-1{,}2-1)}} = 0{,}34$$

Metrische Korrelationskoeffizienten

- Ab Ordinalskalenniveau ist auch die "Richtung" eines Zusammenhangs interessant. Korrelationskoeffizienten erhalten deshalb ein Vorzeichen
  - Wertebereich: [-1,1]
  - Positiver Zusammenhang: größere X gehen mit größeren Y einher
  - Negativer Zusammenhang: größere X gehen mit kleineren Y einher

- Das Streudiagramm (Scatterplot)
  - Ein Streudiagramm stellt die Kombinationen (y_1,x_1), ..., (y_n,x_n) graphisch in einem X-Y-Koordinatensystem dar
  - Regel: uV = X-Achse und aV = Y-Achse
  - Die Visualisierung liefert Hinweise auf einen evtl. Zusammenhang

Beispiel: Abinote und Klausurerfolg



Kohorte 1998/99, n=81

Korrelationskoeffizient nach Bravais-Pearson

- Auch Pearsonscher Produkt-Moment Korrelationskoeffizient genannt

Im Folgenden:
Pearsons r

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\tilde{s}_{XY}}{\tilde{s}_X \tilde{s}_Y}$$

wobei $\tilde{s}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ die empirische "Kovarianz„ ist

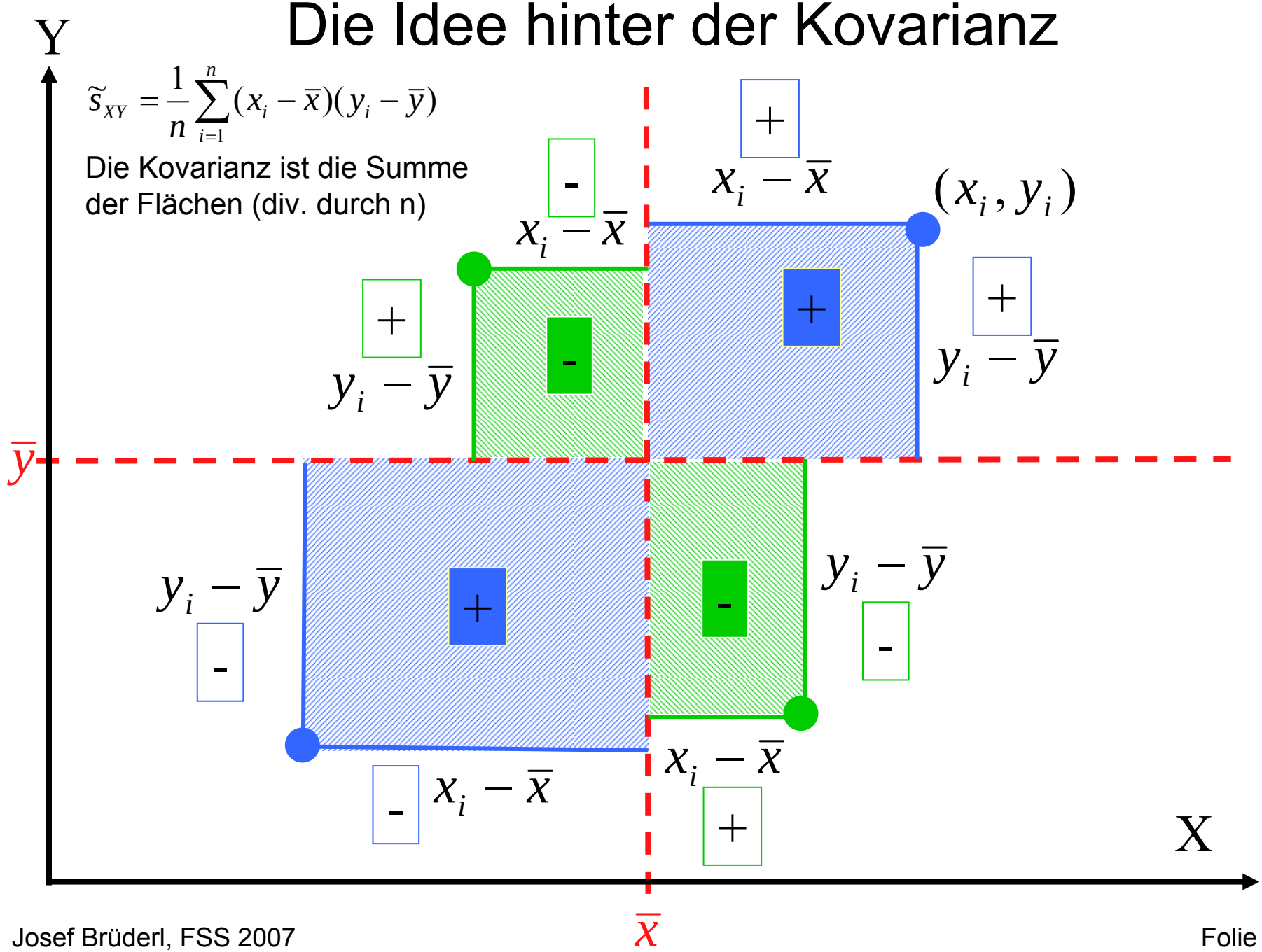
und $\tilde{s}_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ bzw. $\tilde{s}_Y$ die Standardabweichungen sind

- r hat das gleiche Vorzeichen wie die Kovarianz

Die Idee hinter der Kovarianz

$$\tilde{s}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Die Kovarianz ist die Summe der Flächen (div. durch n)



Beispiel

x_i	y_i	x_i-x_q	y_i-y_q	$(x_i-x_q)^2(y_i-y_q)^2$	$(x_i-x_q)(y_i-y_q)$
1.5	3	-1	0	1	0	0
3.5	4	1	1	1	1	1
2.5	2	0	-1	0	1	0
3	3	0.5	0	0.25	0	0
1	4	-1.5	1	2.25	1	-1.5
4	2	1.5	-1	2.25	1	-1.5
2	3	-0.5	0	0.25	0	0
17.5	21	(Σ)		7	4	-2
2.5	3	(Σ/n)				

$$r = \frac{-2}{\sqrt{7 \cdot 4}} = -0.378$$

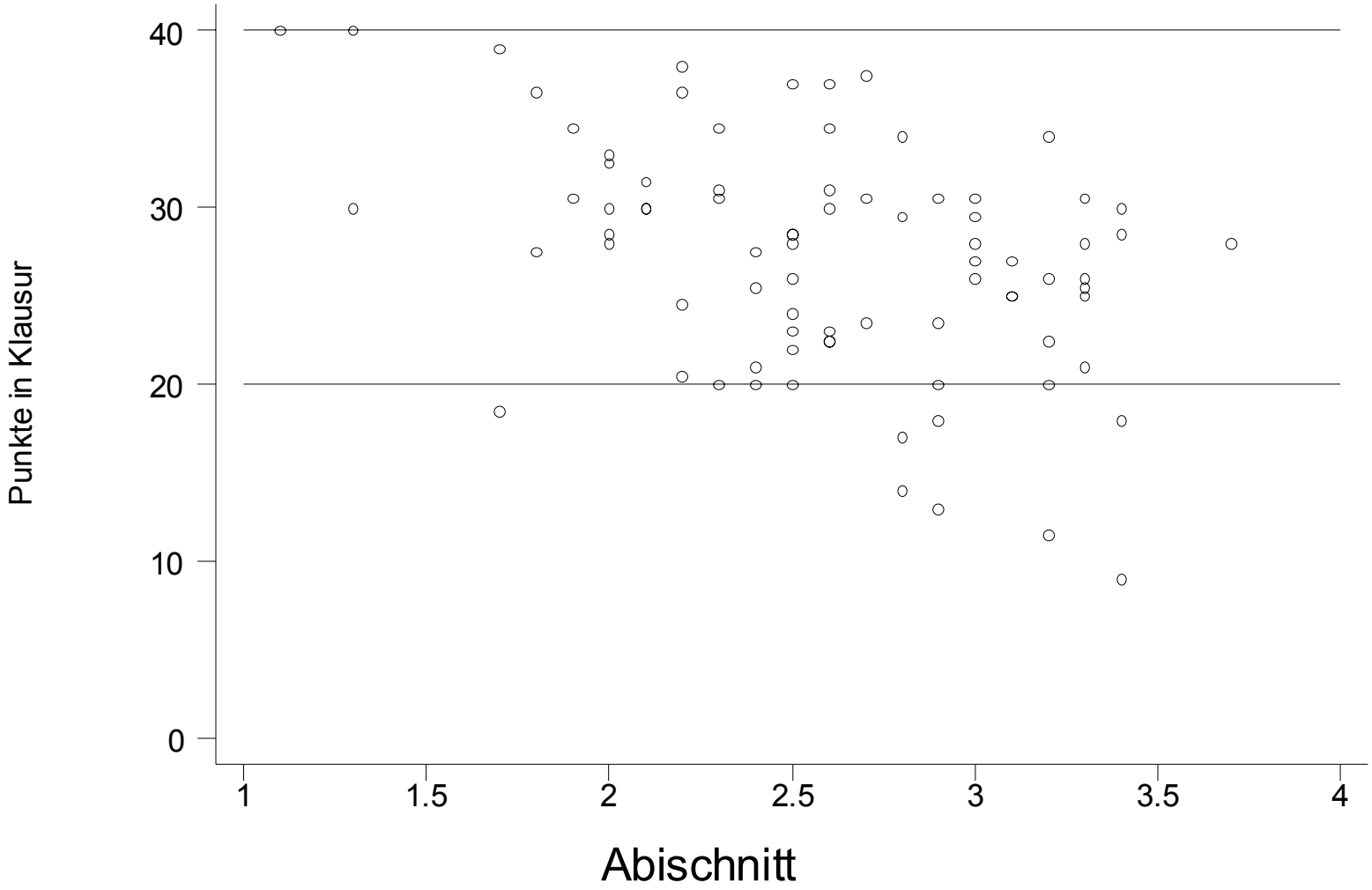
Beispiel: rechengünstige Formel

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1.5	3	2.25	9	4.5
3.5	4	12.25	16	14
2.5	2	6.25	4	5
3	3	9	9	9
1	4	1	16	4
4	2	16	4	8
2	3	4	9	6
17.5	21 (Σ)	50.75	67	50.5
2.5	3 (Σ/n)			

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\left(\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2\right)}}$$

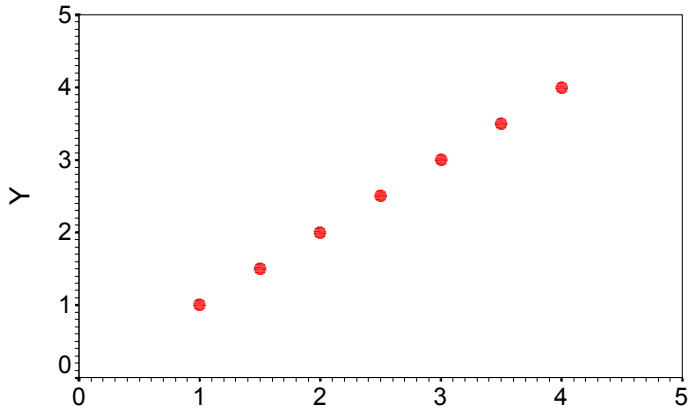
$$r = \frac{50.5 - 7 \cdot 2.5 \cdot 3}{\sqrt{(50.75 - 7 \cdot 2.5 \cdot 2.5) \cdot (67 - 7 \cdot 3 \cdot 3)}} = \frac{-2}{\sqrt{7 \cdot 4}} = -0.378$$

Beispiel: Abinote und Klausurerfolg

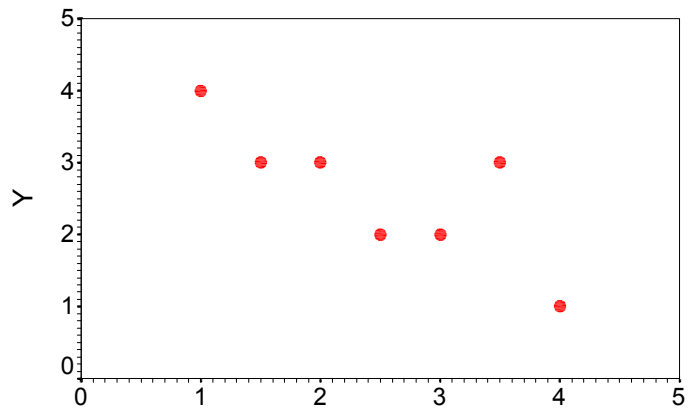


r = -0,43

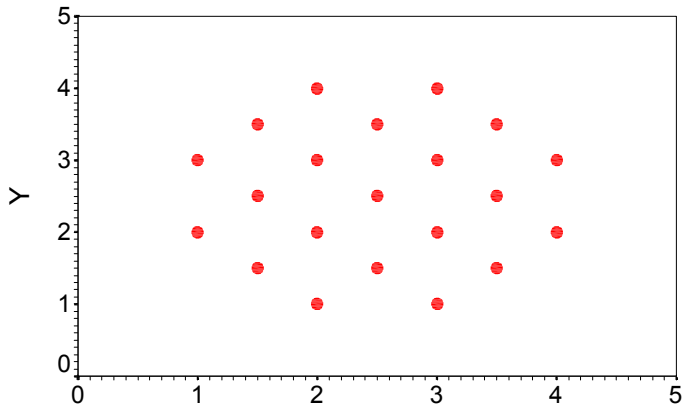
r misst die Stärke eines linearen Zusammenhangs



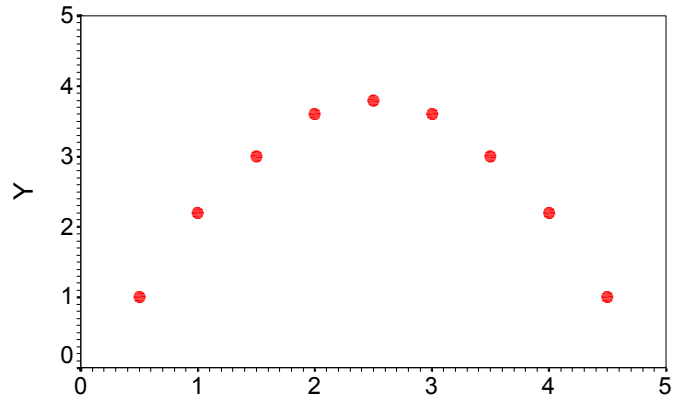
r = 1



r = -0.79



r = 0



r = 0

Rangkorrelation: Spearmans ρ

- Ordne die x_i der Größe nach und weise jeder Beobachtung den "Rang von x_i" zu
- Der Rang von x_i wird mit $rg(x_i)$ bezeichnet und ist die Platzzahl von x_i in der geordneten Liste
- Bei Bindungen wird allen betroffenen Fällen ihr Durchschnittsrang zugeordnet
- Verfahre analog mit Y, d.h. bilde die $rg(y_i)$
- Berechne ρ als Korrelation r dieser Ränge, d.h:

$$\rho = \frac{\sum_{i=1}^{n}(rg(x_i) - \overline{rg}_X)(rg(y_i) - \overline{rg}_Y)}{\sqrt{\sum_{i=1}^{n}(rg(x_i) - \overline{rg}_X)^2 \sum_{i=1}^{n}(rg(y_i) - \overline{rg}_Y)^2}}$$

$$Es\ gilt: \quad \overline{rg}_X = \overline{rg}_Y = \frac{n+1}{2}$$

- Auch hier gibt es eine rechengünstige Formel

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(rg(x_i) - rg(y_i))^2}{(n^2-1)n}$$

Exakt nur, falls keine Bindungen vorliegen

Zum Umgang mit Bindungen

x (geordnet)	mittlerer Rang	Rang von x
2	$\frac{1+2}{2}=1.5$	1.5
2		1.5
3	$\frac{3+4+5+6+7}{5}=$	5
3		5
3		5
3		5
3		5
4	$\frac{8+9+10}{3}=9$	9
4		9
4		9

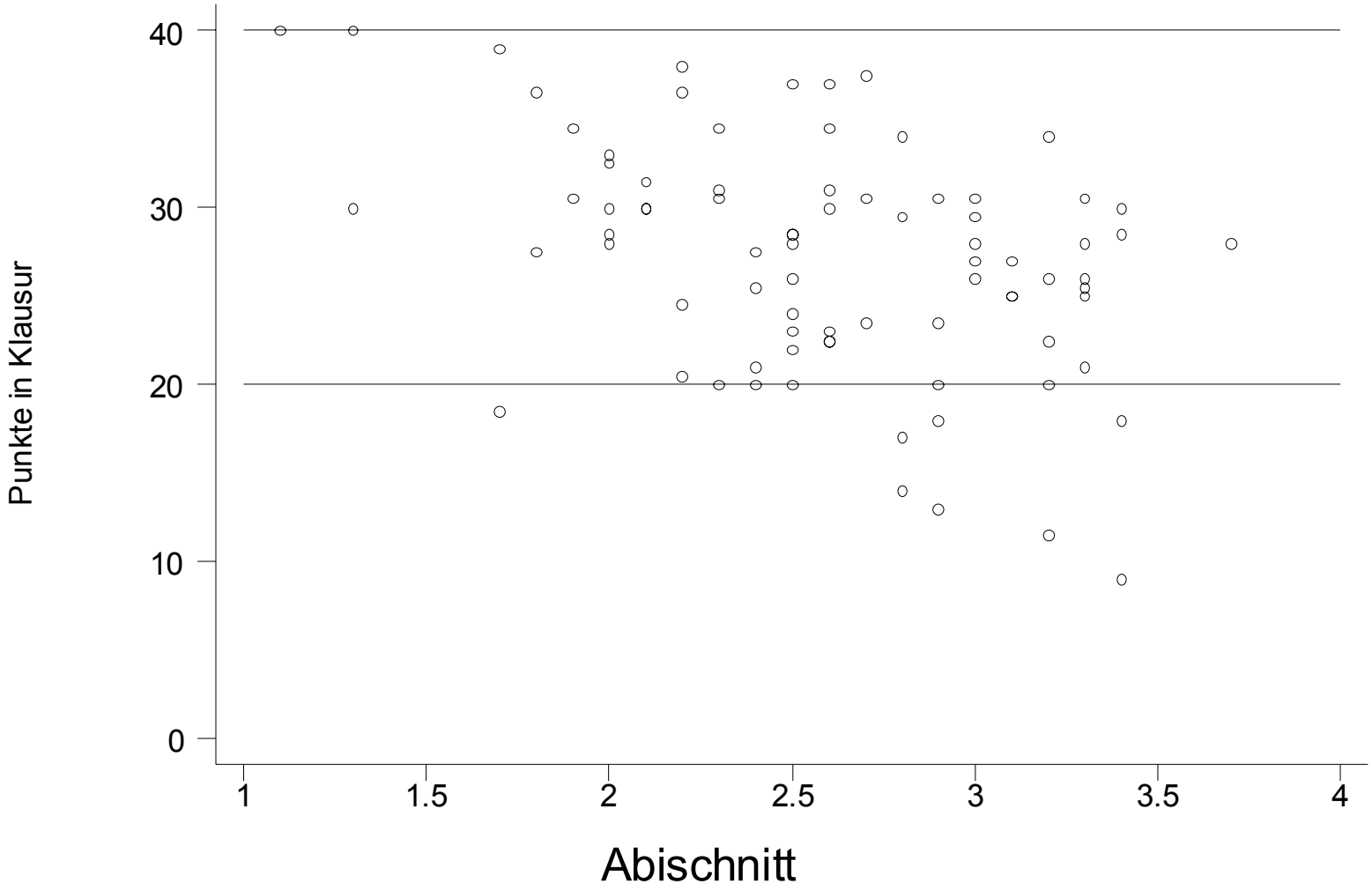
Beispiel: rechengünstige Formel

x_i	y_i	$rg(x_i)$	$rg(y_i)$	$(rg(x_i)-rg(y_i))^2$
1.5	3	2	4	4
3.5	4	6	6.5	0.25
2.5	2	4	1.5	6.25
3	3	5	4	1
1	4	1	6.5	30.25
4	2	7	1.5	30.25
2	3	3	4	1
			Σ	73

$$\rho = 1 - \frac{6 \cdot 73}{(49-1) \cdot 7} = -0.30$$

Hier wegen der vielen Bindungen großer Fehler.
Exakt ist ρ = -0.378

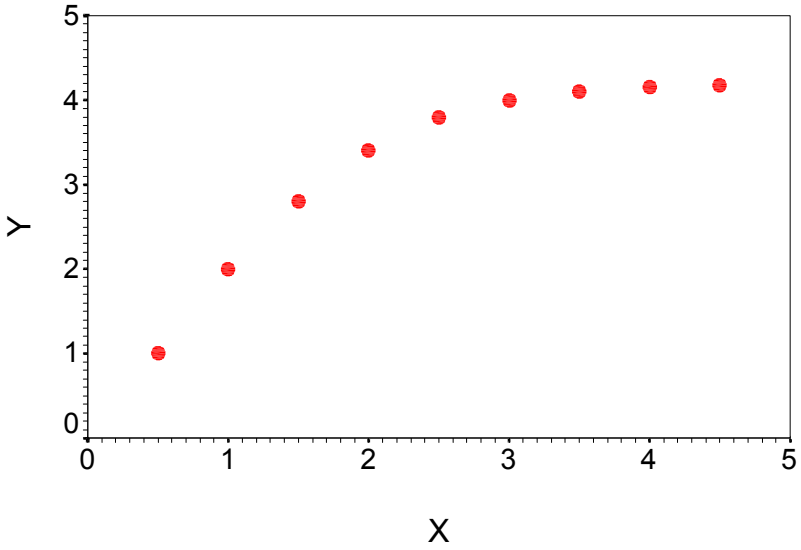
Beispiel: Abinote und Klausurerfolg



$r = -0{,}43$

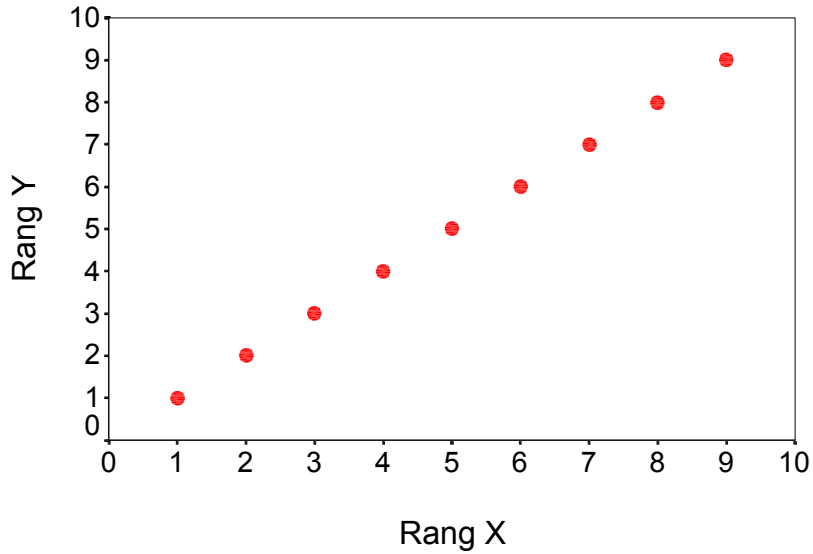
$\rho = -0{,}38$

ρ misst die Stärke eines monotonen Zusammenhangs



$\rho = 1$

$r = 0{,}91$



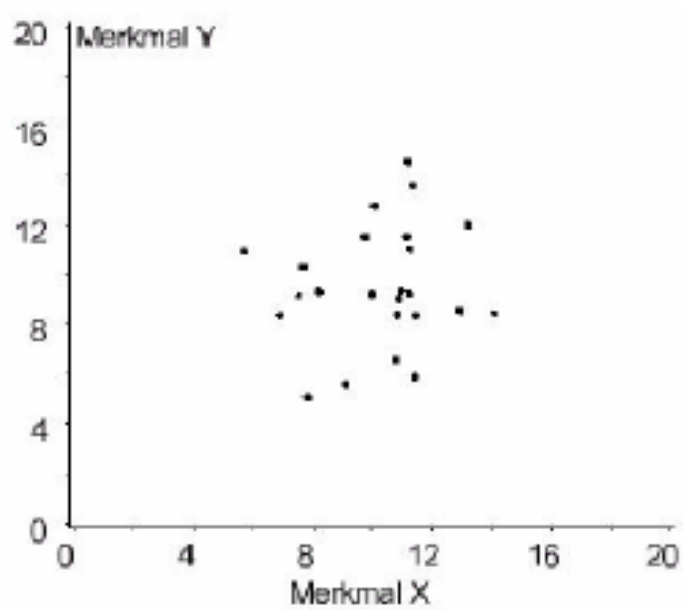
Grund ist, dass die Ränge hier einem linearen Zusammenhang folgen

Eigenschaften von r und ρ

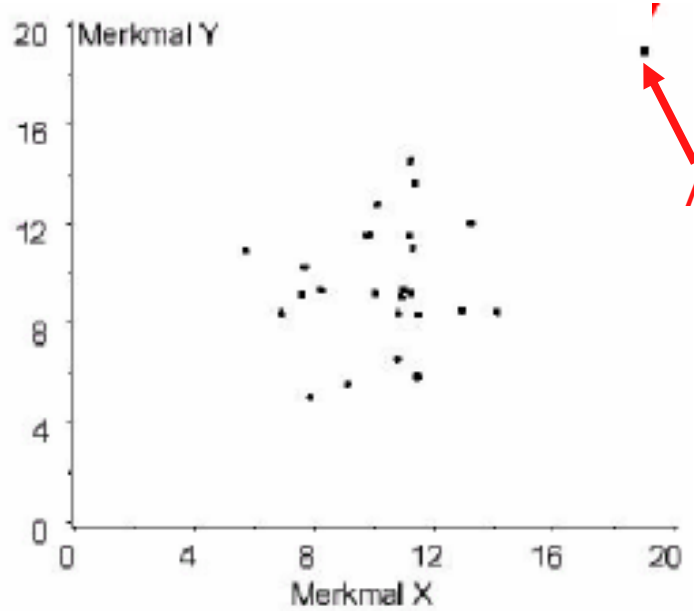
- Sowohl r als auch ρ liegen im Intervall [-1,1]
- Sowohl r als auch ρ sind symmetrisch
- r ist invariant gegenüber linearen Transformationen
- ρ ist invariant gegenüber monotonen Transformationen
- Da ρ nur von den Rängen abhängt, ist der Koeffizient bereits ab Ordinalskalenniveau anwendbar
- Wendet man r auf zwei Merkmale X und Y an, die nur die Ausprägungen 0 und 1 annehmen ('dichotome' bzw. 'binäre' Merkmale), so gilt:

$$|r| = \phi$$

Korrelationskoeffizienten und Ausreißer



Ohne Ausreißer
r = 0, 12
ρ = 0,05



Mit Ausreißer
r = 0, 48
ρ = 0,14

- r ist sehr anfällig für Ausreißer
  - Grund ist, dass die Kovarianz auf den Abstandsflächen beruht
- ρ dagegen ist deutlich weniger anfällig für Ausreißer
  - weil hier nur die Ränge in die Berechnung eingehen

Lineare Regressionsanalyse

- Die Regressionsanalyse ist das zentrale Analyseverfahren in den Sozialwissenschaften
- Im einfachsten Fall (einfache lineare Regression) handelt es sich um ein Verfahren zur bivariaten Zusammenhangsanalyse metrischer Variablen
- Ihre besondere Bedeutung erlangt die Methode
  - einerseits dadurch, dass sie einen bequemen Übergang zur multivariaten Datenanalyse erlaubt (multiple Regression),
  - andererseits dadurch, dass sie sich durchaus auch auf kategoriale Variablen erweitern lässt (uV: Dummy-Variablen; aV: logistische Regression, ordinales Logitmodell, multinomiales Logitmodell u.a.)

Einfache lineare Regression: Grundidee

- Es seien zwei metrische Variablen X und Y gegeben. X wird als uV betrachtet, Y als aV

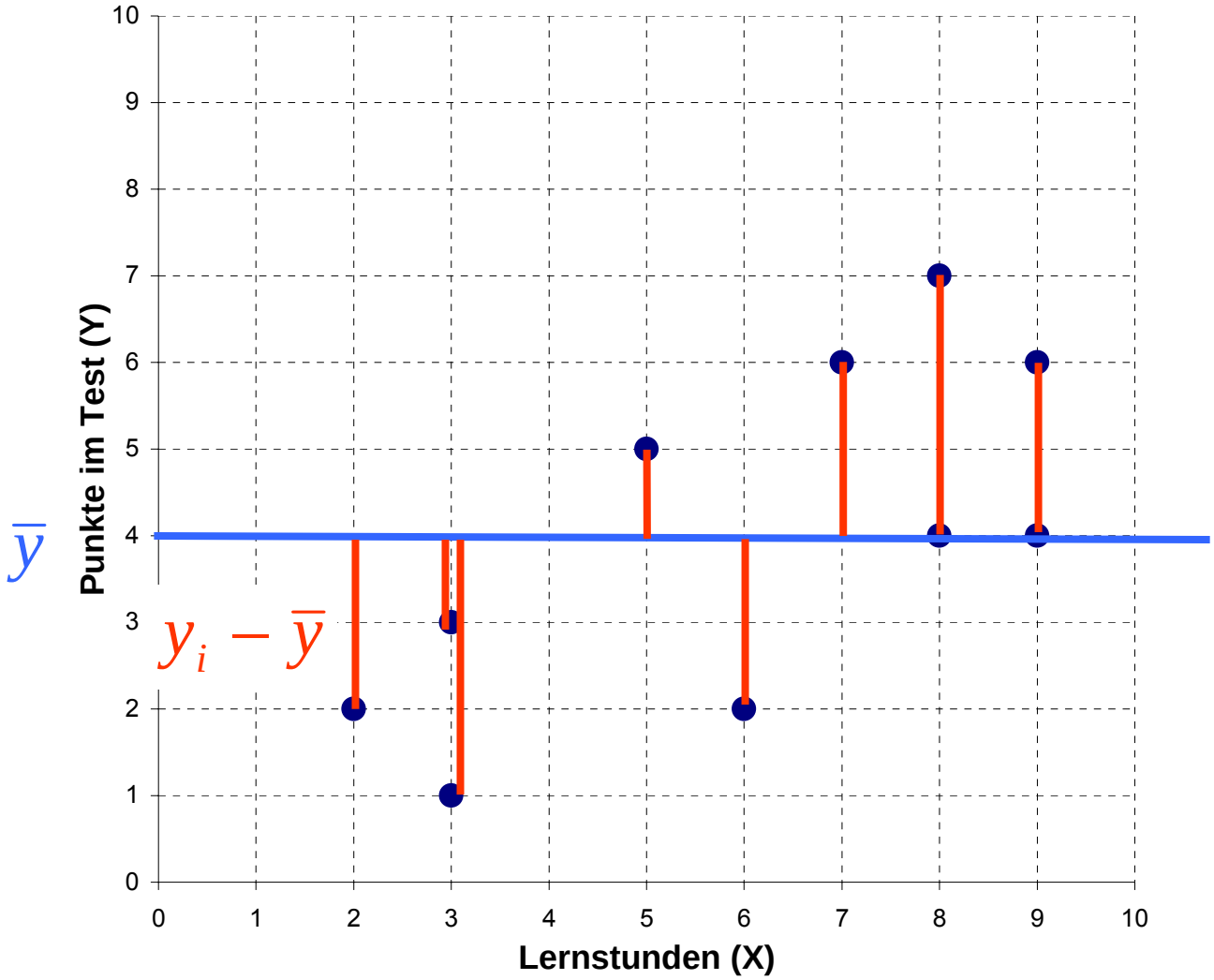
Lernstunden (X)	2	3	3	5	6	7	8	8	9	9
Punkte im Test (Y)	2	1	3	5	2	6	4	7	6	4

- Wie kann man nun eine Prognose über Y abgeben?
- Nehmen wir an, man hat keine Information über X: dann ist das arithmetische Mittel die Vorhersage, die die Summe der quadratischen Abweichungen minimiert (vgl. Kap. II)
- Im Beispielfall ist:

$$\bar{y} = 4$$

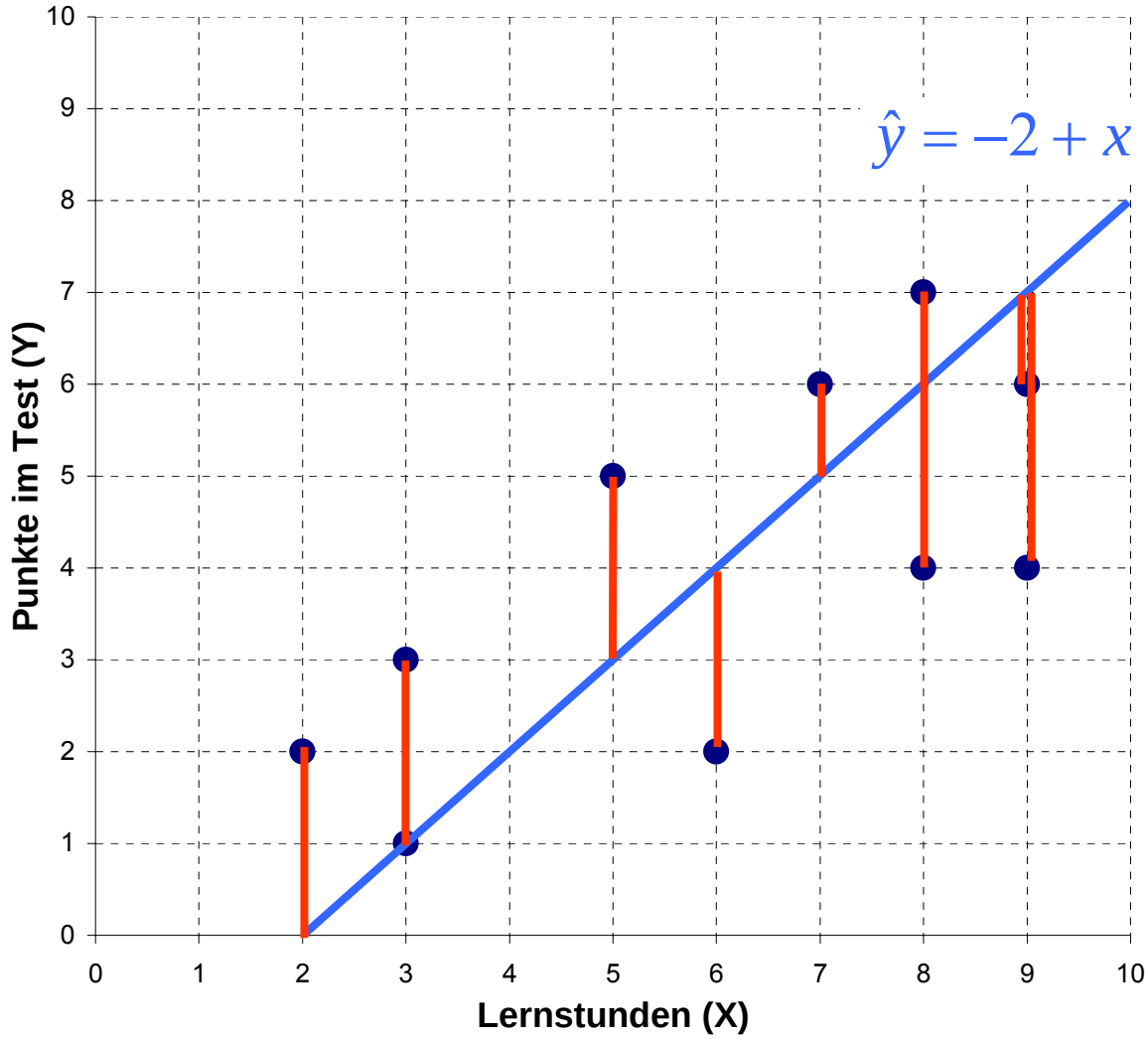
$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = 36$$

Graphische Darstellung im Streudiagramm



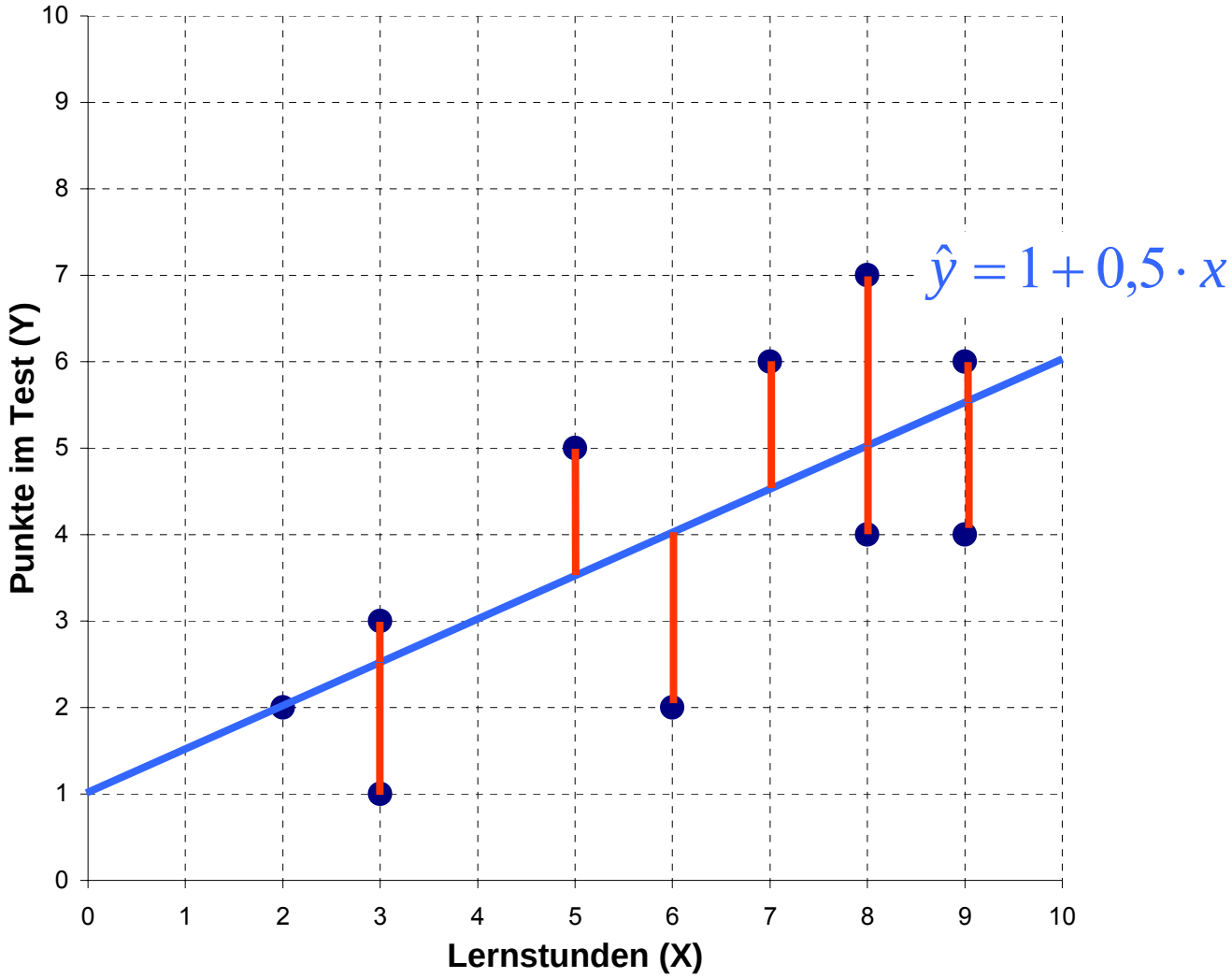
y_i	$(y_i-4)^2$
2	4
1	9
3	1
5	1
2	4
6	4
4	0
7	9
6	4
4	0
	36

Verbesserung der Prognose: Berücksichtigung der Lernstunden (X)



y_i	$(y_i-\hat{y})^2$
2	4
1	0
3	4
5	4
2	4
6	1
4	4
7	1
6	1
4	9
	32

Geht es noch besser?



y_i	$(y_i-\hat{y})^2$
2	0
1	2,25
3	0,25
5	2,25
2	4
6	2,25
4	1
7	4
6	0,25
4	2,25
	18,5

Welche Gerade ist die Beste?

- Prinzipiell lassen sich beliebig viele Geraden durch die Punkte legen
- Jede dieser Geraden ist gekennzeichnet durch

$$\hat{y} = \alpha + \beta x$$

  also durch die beiden Parameter α und β bestimmt
- Gesucht: Diejenige Gerade, d.h. die Werte α und β, für die die Summe der quadrierten Abweichungen minimal wird
- Lösung: Methode der kleinsten Quadrate
  (ordinary least squares, OLS)

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \hat{y})^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

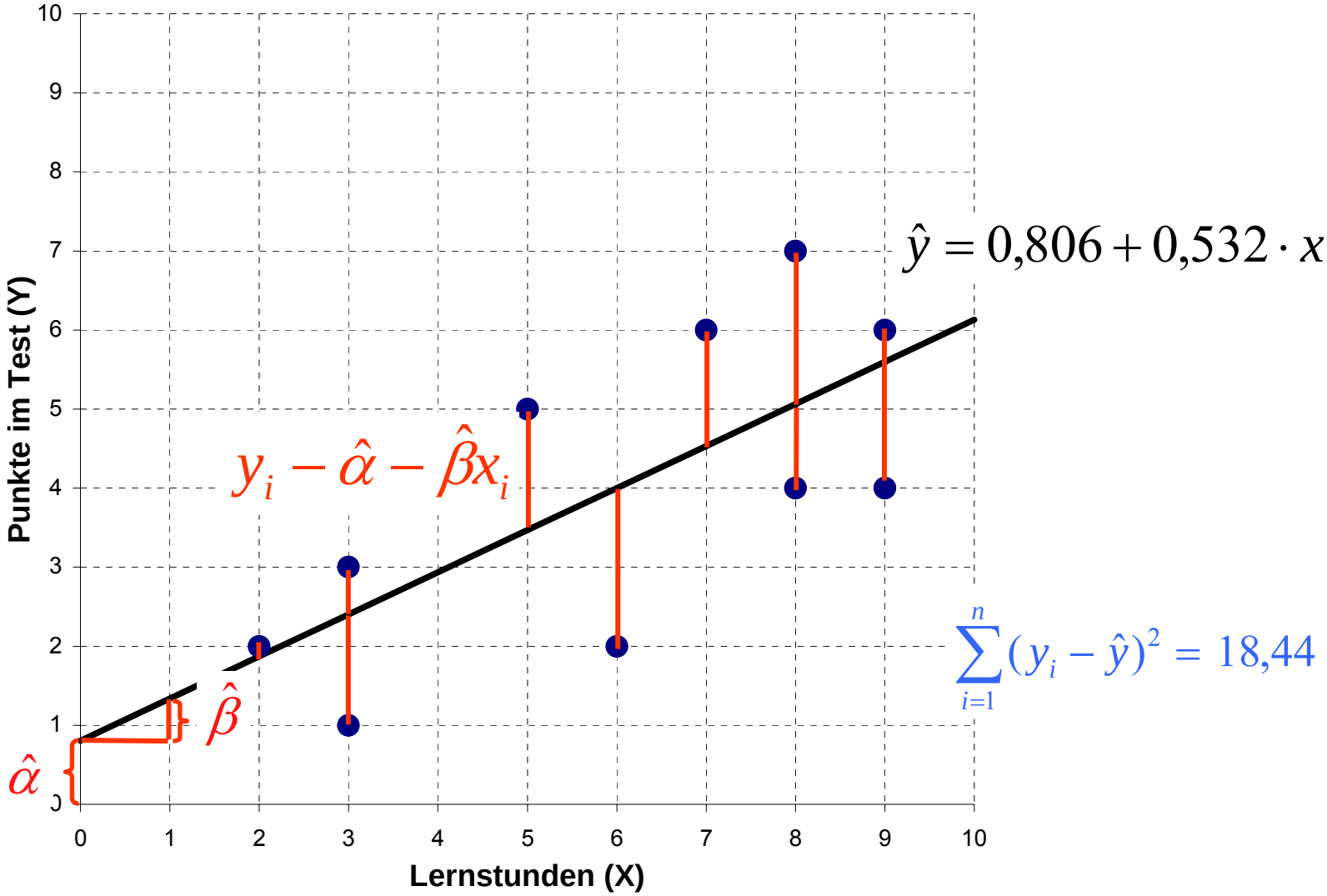
$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\tilde{s}_{XY}}{\tilde{s}_X{}^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i{}^2 - n\bar{x}^2}$$

Beispiel: Berechnung der OLS-Schätzer

x_i	y_i	$x_i - 6$	$y_i - 4$	$(x_i - 6)^2$	$(x_i-6)\cdot(y_i-4)$
2	2	-4	-2	16	8
3	1	-3	-3	9	9
3	3	-3	-1	9	3
5	5	-1	1	1	-1
6	2	0	-2	0	0
7	6	1	2	1	2
8	4	2	0	4	0
8	7	2	3	4	6
9	6	3	2	9	6
9	4	3	0	9	0
60	40			62	33

$$\hat{\beta} = \frac{33}{62} = 0{,}532 \qquad \hat{\alpha} = 4 - 0{,}532 \cdot 6 = 0{,}806$$

Beispiel: Die OLS-Regressionsgerade



Die Regression als Modell des bivariaten Zusammenhangs

- Sind aV (Regressand) und uV (Regressor) beide metrisch, so kann man zur Zusammenhangsanalyse das lineare Regressionsmodell einsetzen.
- Man formuliert folgendes lineare Modell des Zusammenhangs:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

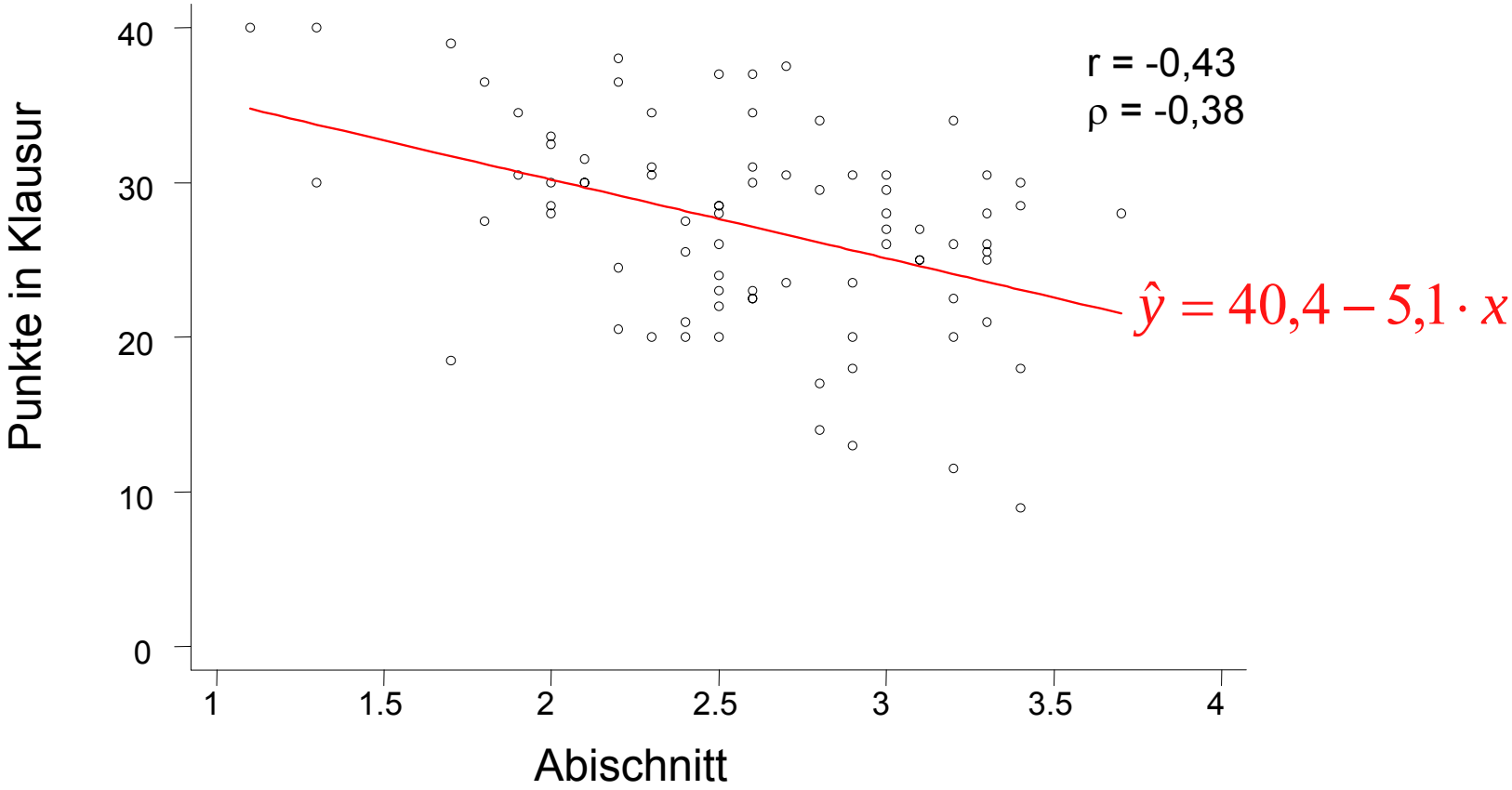
- α und β sind die Regressionskoeffizienten
  - α: Achsenabschnitt, β: Steigung
  - β: um wie viel Einheiten ändert sich Y, wenn X um eine Einheit steigt
  - β ist damit ein einfach zu interpretierendes Maß des Zusammenhangs
- ε_i ist der Fehlerterm (Abweichung der Daten von der Regressionsgerade)
- Man schätzt die Regressionskoeffizienten, indem man die Fehlerquadratsumme minimiert (OLS)

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \varepsilon_i^{\,2} = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Die OLS-Schätzer (Formeln s.o.) notieren wir mit $\hat{\alpha}$ und $\hat{\beta}$
- Die vom Regressionsmodell vorhergesagten Werte sind $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$
- Die geschätzten Fehler (Residuen) sind damit $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$

Beispiele

Abinote und Klausurerfolg



Lernstunden und Punkte im Test

$$\hat{y} = 0{,}81 + 0{,}53 \cdot x$$

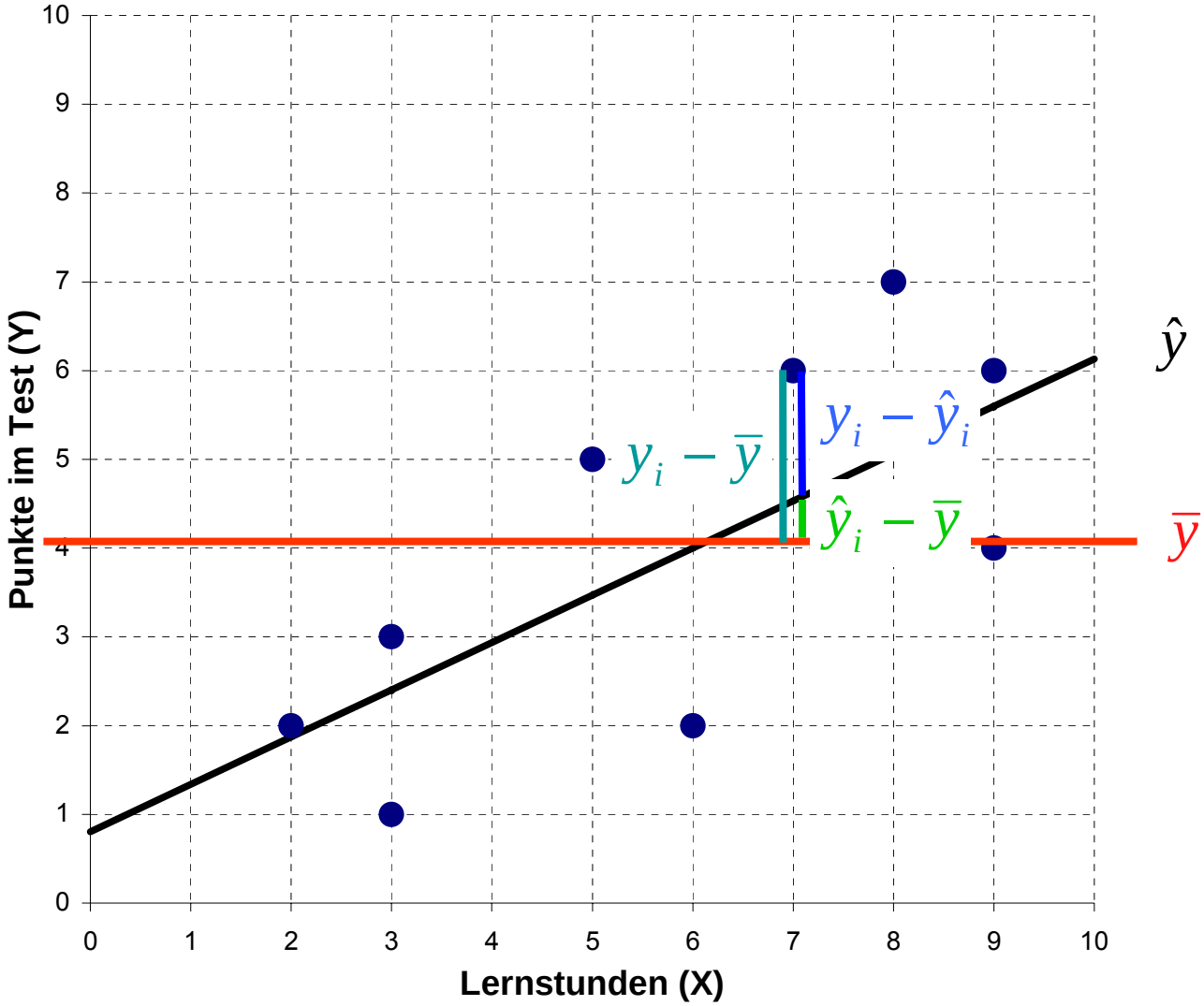
Das Bestimmtheitsmaß R^2

- Wie gut passt das Regressionsmodell auf die Daten?
- Die Grundidee ist: Welcher Anteil der Streuung von Y wird durch das Regressionsmodell „erklärt“?
- Streuungszerlegung
  - Total sum of squares (TSS): $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$
  - Model sum of squares (MSS): $MSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
  - Residual sum of squares (RSS): $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
  - Die gesamte Streuung kann damit in zwei Teile zerlegt werden

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = MSS + RSS$$

Graphische Interpretation der Streuungszerlegung



Das Bestimmtheitsmaß R^2

- Das Bestimmtheitsmaß ist nun definiert als

$$R^2 = \frac{\text{erklärte Streuung}}{\text{gesamte Streuung}} = \frac{MSS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Es gilt: $0 \leq R^2 \leq 1$
- R^2 lässt sich interpretieren als der Anteil der Varianz, der durch die Regressionsgerade (und damit durch X) erklärt wird
- Es gilt: $R^2 = r^2$
- Berechnung: am einfachsten durch Quadrierung von r

Beispiele

Lernstunden und Punkte im Test

x_i	y_i	$\hat{y}_i$	$(y_i - 4)^2$	$(\hat{y}_i - 4)^2$
2	2	1.870968	4	4.532777
3	1	2.403226	9	2.549687
3	3	2.403226	1	2.549687
5	5	3.467742	1	0.283299
6	2	4	4	0.000000
7	6	4.532258	4	0.283299
8	4	5.064516	0	1.133194
8	7	5.064516	9	1.133194
9	6	5.596774	4	2.549687
9	4	5.596774	0	2.549687
	40		36	17.564512

$\hat{\alpha}$ = 0.806452

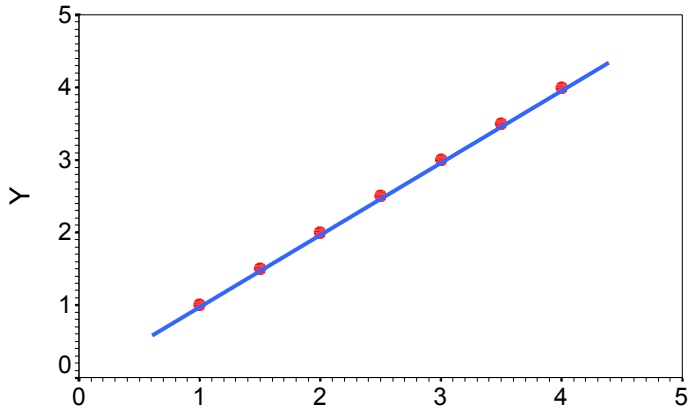
$\hat{\beta}$ = 0.532258

R^2 = 17.56 / 36
= .4879

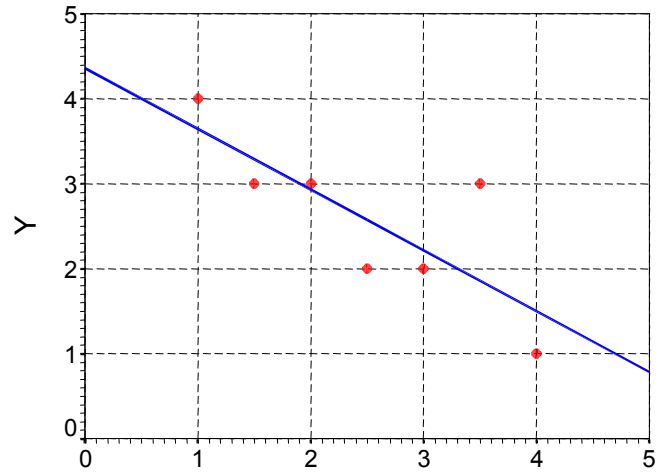
Abinote und Klausurerfolg

$r = -0{,}43 \rightarrow R^2 = 0{,}18$

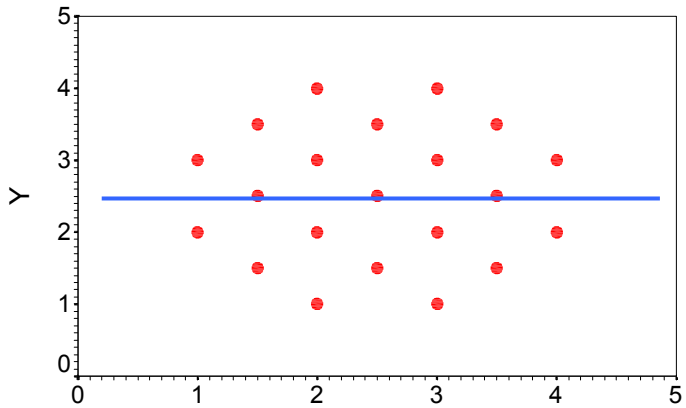
r und R^2



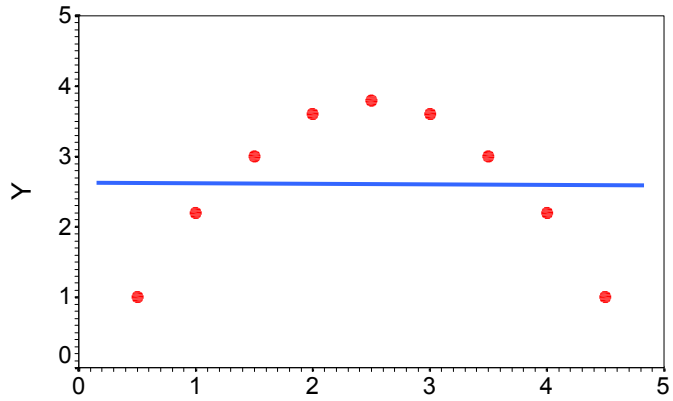
$r = 1, R^2 = 1$



$r = -0.79 , R^2 = 0,62$



$r = 0 , R^2 = 0$



$r = 0 , R^2 = 0$

Signifikanztest für $\hat{\beta}$

- $\hat{\beta}$ ist ein Schätzer
  - Mit einer Stichprobenverteilung
  - Und einem Standardfehler $\hat{\sigma}_{\hat{\beta}}$
- Damit kann man auch ein Konfidenzintervall schätzen
- Ebenso kann man einen Signifikanztest durchführen
  - Nullhypothese: X hat keinen Einfluss auf Y (kein Zusammenhang)
    $H_0: \beta = 0$
  - Die Teststatistik (t-Wert) ist

$$T = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \sim t(n-2)$$

  - Die H_0 wird abgelehnt, falls $|T| > t_{1-\alpha/2}(n-2)$
    - Ab n>30 das $z_{1-\alpha/2}$ Quantil (Faustregel für α=5%: |T| > 2)
  - Können wir die H_0 verwerfen, so spricht man davon, dass X einen signifikanten Einfluss auf Y hat

Abinote und Klausurerfolg: eine Regression mit STATA

```
. regress klpunkte abinote

  Source |       SS       df       MS      Number of obs =      81 [n]
---------+------------------------------   F(  1,    79) =   17.45
   Model |  620.99 [MSS]   1  620.991439   Prob > F      =  0.0001
Residual |  2811.3 [RSS]  79    35.58645   R-squared     =  0.1809 [R²]
---------+------------------------------   Adj R-squared =  0.1706
   Total |  3432.3 [TSS]  80  42.9040123   Root MSE      =  5.9654

------------------------------------------------------------------------------
klpunkte |      Coef.   Std. Err.      t    [95% Conf. Interv.]
---------+--------------------------------------------------------------------
 abinote | [beta]   -5.085771   1.217464    -4.18    -7.509   -2.662
   _cons | [alpha]   40.35849   3.229647    12.50     33.93    46.78
------------------------------------------------------------------------------
```