

Andreas Quatember:

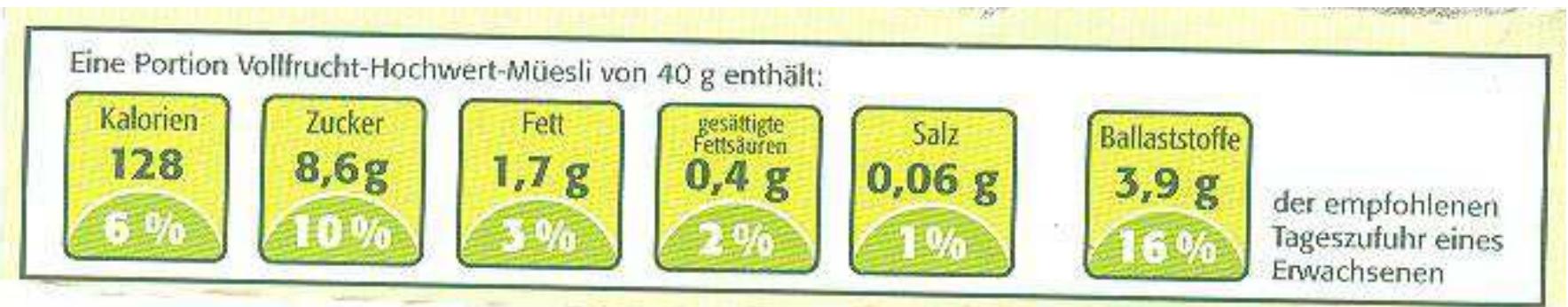
Statistik ohne *Angst* vor Formeln

**Eine verständnisorientierte Einführung in
die Grundlagen der Statistik**

Bedeutung des Faches Statistik



Statistik ist Alltag!



Beispiele:

- Analysen des Finanzmarktes, High-Frequency-Trading, Stopp-Loss-Automatik
- Big Data, Kundendatenanalysen im Web (Amazon, iTunes, Facebook, Google)
- Statistische Analysen im Sport: zB Matchstatistiken im Fußball

UEFA.com

LIVE | PULS

BAYERN	1 - 1	ARSENAL
64%	BALL POSSESSION	36%
14	TOTAL ATTEMPTS	8
9	ATTEMPTS ON TARGET	5
4	BLOCKS AND SAVES	8
6	CORNERS	5
4	OFFSIDES	3
119.62 km	DISTANCE COVERED	114.51 km
631 (83%)	PASSES COMPLETED	269 (67%)
14	FOULS COMMITTED	14
2 / 0	YELLOW/RED CARDS	3 / 0

Image des Faches

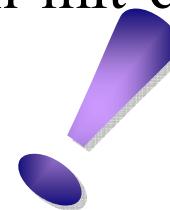


„,... und jetzt noch etwas für die Statistiker unter unseren Zusehern“

„Mit Statistik lässt sich alles beweisen!“

„Glaube keiner Statistik, die Du nicht selber gefälscht hast!“

Verwechslung der Qualität der statistischen Methoden mit der Qualität ihrer Anwendung



Was versteht man unter **Statistik**? → Methoden der Datenanalyse zum Zweck der Informationsbündelung



Gliederung in 3 Teile:

1 Beschreibende Statistik

Es liegen vollständige Daten über eine Gesamtheit vor

2 Wahrscheinlichkeitstheorie

Kombiniert 1 mit 3

3 Schließende Statistik

Es liegen nur Daten über einen ausgewählten Teil der Gesamtheit vor

1 Beschreibende Statistik

1.1 Grundbegriffe

Erhebungseinheiten: Objekte, über die Daten erhoben werden

Grundgesamtheit: Gesamtheit aller Erhebungseinheiten

Merkmal: Eine interessierende Eigenschaft

Merkmalsausprägungen: Die einzelnen möglichen Werte eines Merkmals

Wertebereich: Alle Merkmalsausprägungen

Beispiel 1: Grundbegriffe einer statistischen Erhebung

Erhebung der Punkteverteilung bei der Statistikklausur:

Grundgesamtheit:	alle Prüflinge
Merkmal:	Punkte
Merkmalsausprägungen:	0, 1, 2, ...

Erhebung der Zufriedenheit von Kunden:

Grundgesamtheit:	alle Kunden
Merkmal:	Zufriedenheit mit der Beratung
Merkmalsausprägungen:	sehr zufrieden, eher zufrieden, teils-teils, eher unzufrieden, sehr unzufrieden

Erhebung des besten Kinofilms:

Grundgesamtheit:	alle teilnahmewilligen Leser und -innen
Merkmal:	bester Film
Merkmalsausprägungen:	Film 1, Film 2, ...

■ Merkmalstypen (1. Unterscheidung): **nominal** - **ordinal** - **metrisch**

nominal: Unterscheidung der Merkmalsausprägungen dem Namen nach
(Prototyp: Geschlecht)

ordinal: Merkmalsausprägungen besitzen eine natürliche Reihenfolge
(Prototyp: Schulnoten)

metrisch: Merkmalsausprägungen lassen sich reihen **und** sind Vielfache
einer Einheit (Prototyp: Körpergröße)

■ Merkmalstypen (2. Unterscheidung): **diskret** - **stetig**

diskret: Wertebereich umfasst nur bestimmte Merkmalsausprägungen
(Prototyp: Schulnoten)

stetig: ... umfasst alle reellen Werte eines Intervalls (Körpergröße)

Kodierung der Merkmalsausprägungen

1. Geschlecht: weiblich (=1) männlich (=2)

2. Alter (in vollendeten Lebensjahren): Jahre

3. Wie schätzen Sie die didaktisch-methodische Qualität der LVA ein?

1 (=sehr gut) 2 3 4 5 (=sehr schlecht)

4. Waren die angegebenen Lernunterlagen hilfreich?

1 2 3 4 5 (1=sehr hilfreich, ... , 5=überhaupt nicht hilfreich)

Dateneingabe für die elektronische Verarbeitung (z.B. in Excel):

	A	B	C	D	E	F	G
1.	Erhebungseinheit:	2	21	1	3		
2.	Erhebungseinheit:	1	38	2	2		
3.	Erhebungseinheit:	3					
	4						

Antwort auf 1. Frage

Antwort auf 2. Frage

Beispiel 2: Merkmalstypen

Merkmal	Merkmalsausprägungen	n / o / m	d / s
Familienstand	ledig (=1), verheiratet (=2), geschieden (=3), verwitwet (=4)	nominal	diskret
100-m-Zeiten	11,21 sec., 11,24 sec., ...	metrisch	stetig
Preis eines Sportartikels	29,90 €, 34,90 €, ...	metrisch	diskret
Platzierungen in einem 100m-Lauf	1., 2., 3., ...	ordinal	diskret
Weitsprungleistung (in ganzen cm)	516 cm, 492 cm, ...	metrisch	stetig

1.2 Tabellarische und graf. Darstellung von Häufigkeitsverteilungen

1.2.1 Häufigkeitsverteilungen einzelner Merkmale

Tabellarische Darstellung

Beispiel 3: Tabellarische Darstellung einer Häufigkeitsverteilung

Punktezahlen	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenhäufigkeit
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1

Häufigkeiten (h): Erster Überblick

Relative Häufigkeiten oder Anteile (p) einer Merkmalsausprägung i:

$$p_i = h_i / N$$

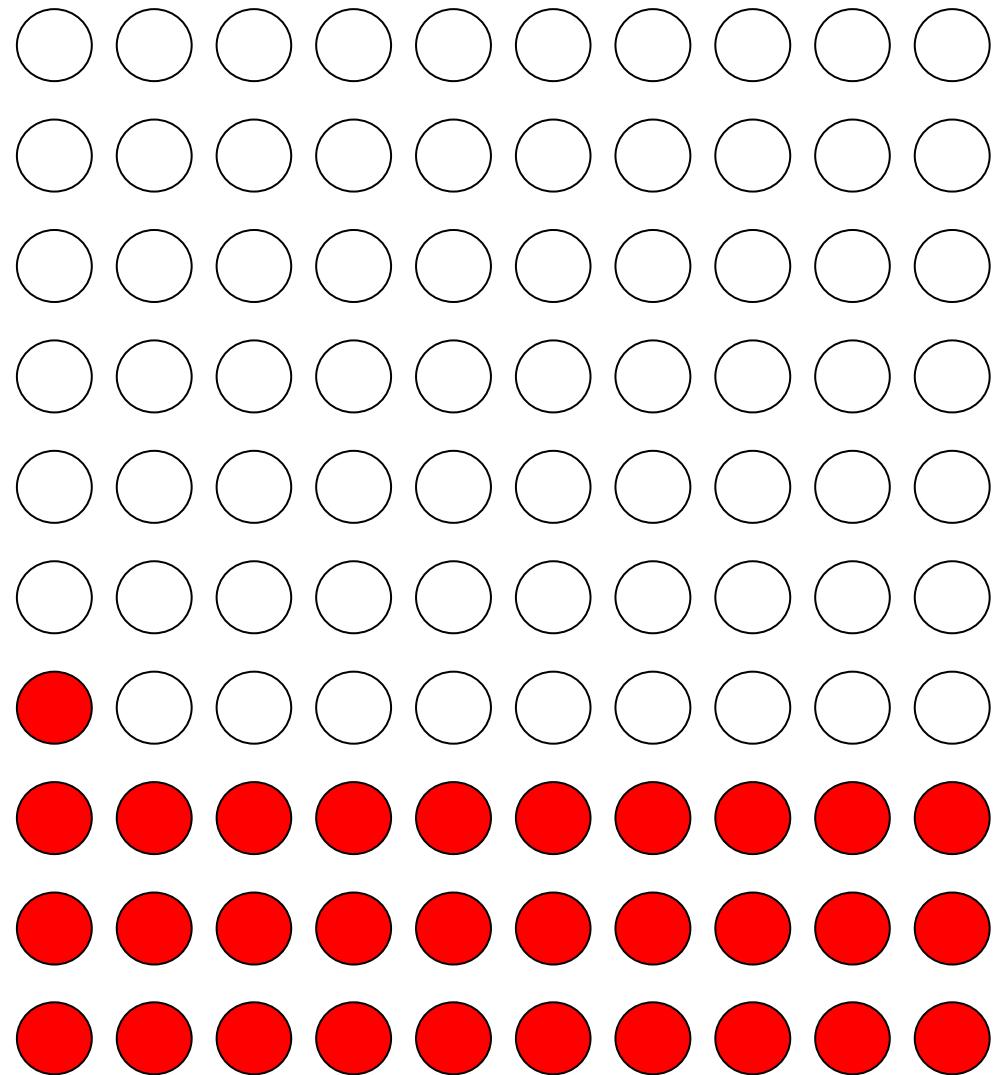
(Formel 1)

Punktzahlen	Häufigkeit h	Relative Häufigkeit p	Prozent	Relative Summenh.
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1

$$N=142$$

Prozentzahlen: $p_i \cdot 100$

Abbildung 1: Die Prozentzahlen (am Beispiel von 31,0 %)



Relative Summenhäufigkeit (oder **empirische Verteilungsfunktion**):
 Summe der relativen Häufigkeiten einer Merkmalsausprägung und aller kleineren Merkmalsausprägungen

z.B. die relative Summenhäufigkeit zur Merkmalsausprägung 3:

Punktezahlen	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenh.
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1

Nur sinnvoll bei metrischen oder ordinalen Merkmalen!

Zusammenfassung von Merkmalsausprägungen zu Intervallen:

Beispiel 4: Tabellarische Darstellung einer Häufigkeitsverteilung

Altersklasse	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenhäufigkeit
0 – unter 15	1.317.707	0,160	16,0	0,160
15 – unter 30	1.526.909	0,185	18,5	0,345
30 – unter 45	1.984.501	0,241	24,1	0,586
45 – unter 60	1.596.849	0,194	19,4	0,780
60 – unter 75	1.173.166	0,142	14,2	0,922
75 und mehr	634.174	0,077	7,7	1
N=8.233.306				

Häufigkeiten, relative Häufigkeiten und relative Summenhäufigkeiten beziehen sich auf ganze Intervalle von Merkmalsausprägungen

<http://www.ifas.jku.at>

Institut für Angewandte Statistik

Gast [Login](#) Personen JKU.at Suchen Schnellzugriff

[JKU](#) | [IFAS](#) | [Unsinn in den Medien](#)

 The IFAS logo features the letters "IFAS" in a bold, blue, sans-serif font. Above the "I", there is a red stylized shape resembling a dome or a graph curve.

Unsinn in den Medien – vom sorglosen Umgang mit Daten

Irren ist menschlich! "Journalistischer Irrtum" in Zusammenhang mit statistischen Daten ist insofern gefährlich, als solche Irrtümer – verbreitet in Zeitungen, Magazinen oder im TV – Bestandteil des "Wissens" der Bevölkerung in Diskussionen im Freundeskreis oder am Stammtisch oder in der Politik werden. Wie häufig selbst einfachste Kennzahlen falsch interpretiert werden, dokumentiert diese kommentierte Seite, die ständig aktualisiert wird (den Artikeln des aktuellen Jahres wird die Jahreszahl hinzugefügt). Diese Kommentare sind bewusst launig verfasst. Dabei sollen aber jedenfalls die Fehler und keinesfalls die Verfasser im Mittelpunkt der Betrachtung stehen. Denn es gilt: siehe oben.

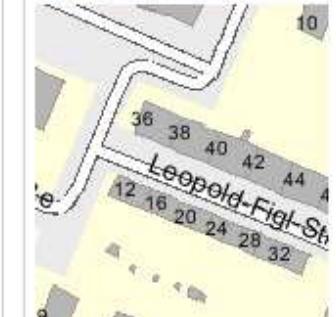
Für den Inhalt verantwortlich: [Andreas Quatember](#)

[Übersicht](#) [Unsinn in den Medien – NEU!](#)

News

- [Neuer Institutsvorstand](#)
Prof. Müller übernimmt mit 1. Jänner 2011 die Leitung des IFAS ... [mehr](#)
- [Mitherausgeberschaft](#)
Professor Müller zum neuen Mitherausgeber der "Statistical Papers" bestellt ... [mehr](#)
- [Neues Web-Layout](#)
Das IFAS zeigt sich in neuem Gewand! ... [mehr](#)

Lageplan

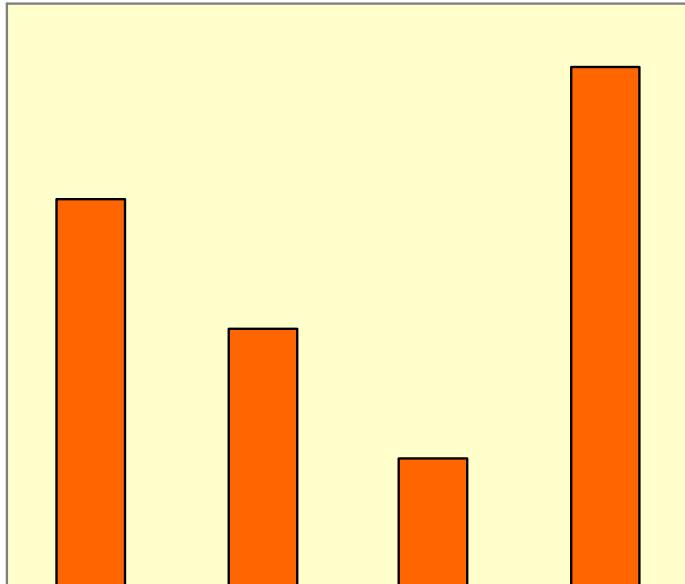


A small map showing the location of the IFAS office within a larger building complex. The map includes street names like "Leopold-Figl-Str." and house numbers 10, 36, 38, 40, 42, 44, 12, 16, 20, 24, 28, 32, 9. The IFAS logo is overlaid on the map at its approximate location.

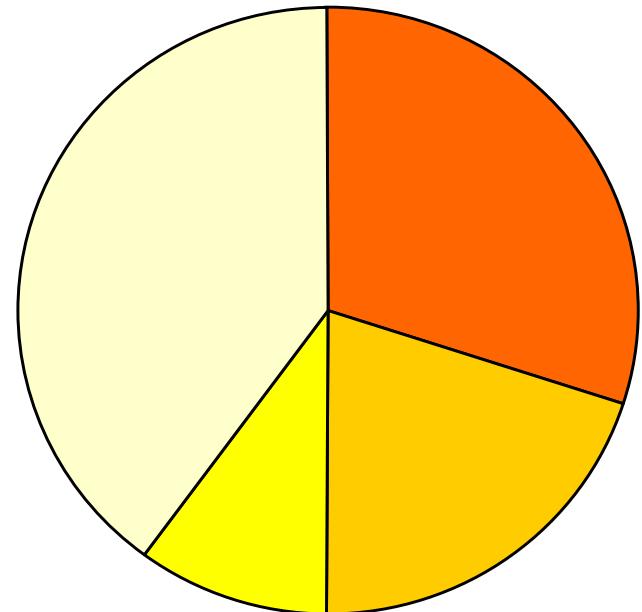
Grafische Darstellung

Aufgabe: Die wesentlichsten Informationen „auf einen Blick“ erfassbar machen

Säulendiagramm:
(Balken-, Stabdiagramm)



Kreisdiagramm:
(Kuchen-, Tortendiagramm)



Beispiel 5: Tabellarische Darstellung der Häufigkeitsverteilung

Note	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenh.
1	16	0,113	11,3	0,113
2	20	0,141	14,1	0,254
3	44	0,310	31,0	0,564
4	32	0,225	22,5	0,789
5	30	0,211	21,1	1

Abbildung 2: Ein Säulendiagramm

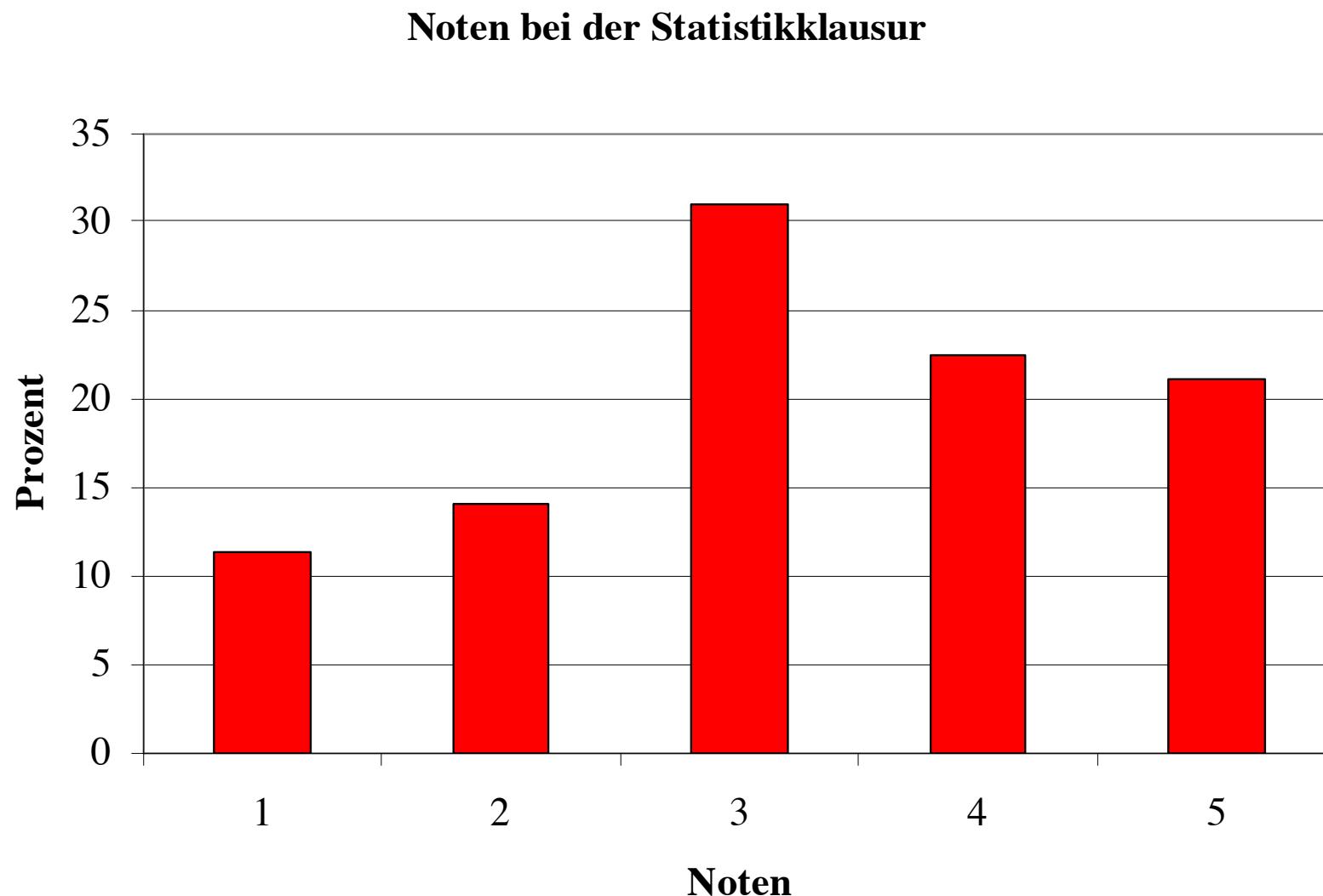
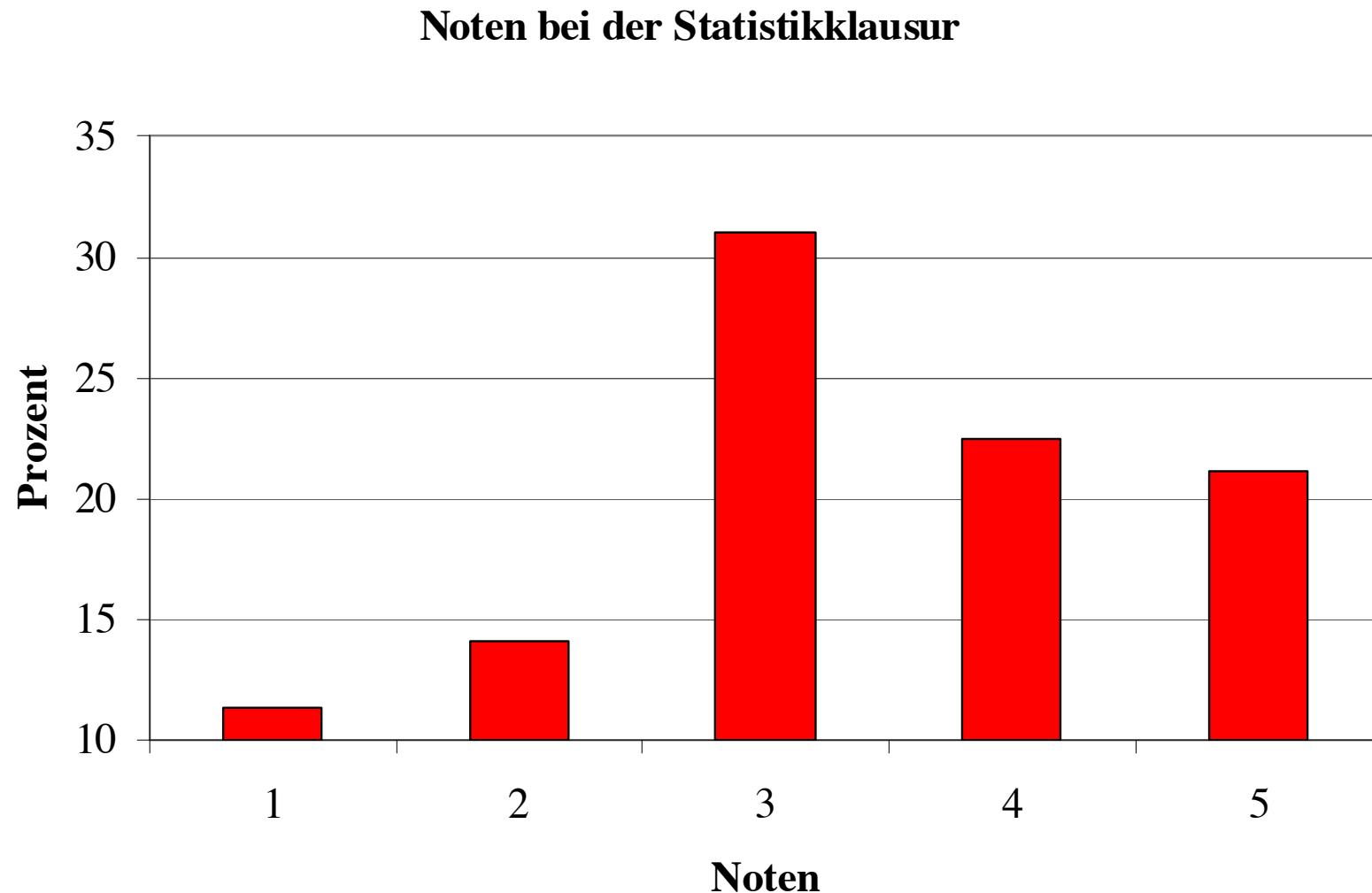


Abbildung 3: Säulendiagramm mit verschobenem Nullpunkt auf der y-Achse



Unsinn in
den Medien

The chart displays 'Gelenkschmiere in mg/ml' on the y-axis (ranging from 0 to 50) against time on the x-axis. It shows a baseline level at 'vorher' (before) and a much higher level at 'nachher' (after), with a red arrow indicating a 25% increase in joint lubrication.

294 018

B.DREXEL GELENK-VITAL SET 3-tlg. Presslinge, Tinktur & 1 gratis Messbecher

HSE24 Preis **€ 79,99**
~~CHF 137,95~~

+Versand: €5,95/CHF7,95

AdT-Tiefpreis
€ 29,99
CHF 51,95

277 430 BDE Duo Karde Vital Massage Sprays € 19,99

Angebot des Tages

0800 29 888 88
EASy 0800 29 888 29

HSE24

Noch 3942 Stk

Abbildung 4: Säulendiagramm mit umgeordneten Merkmalsausprägungen

Noten bei der Statistikklausur

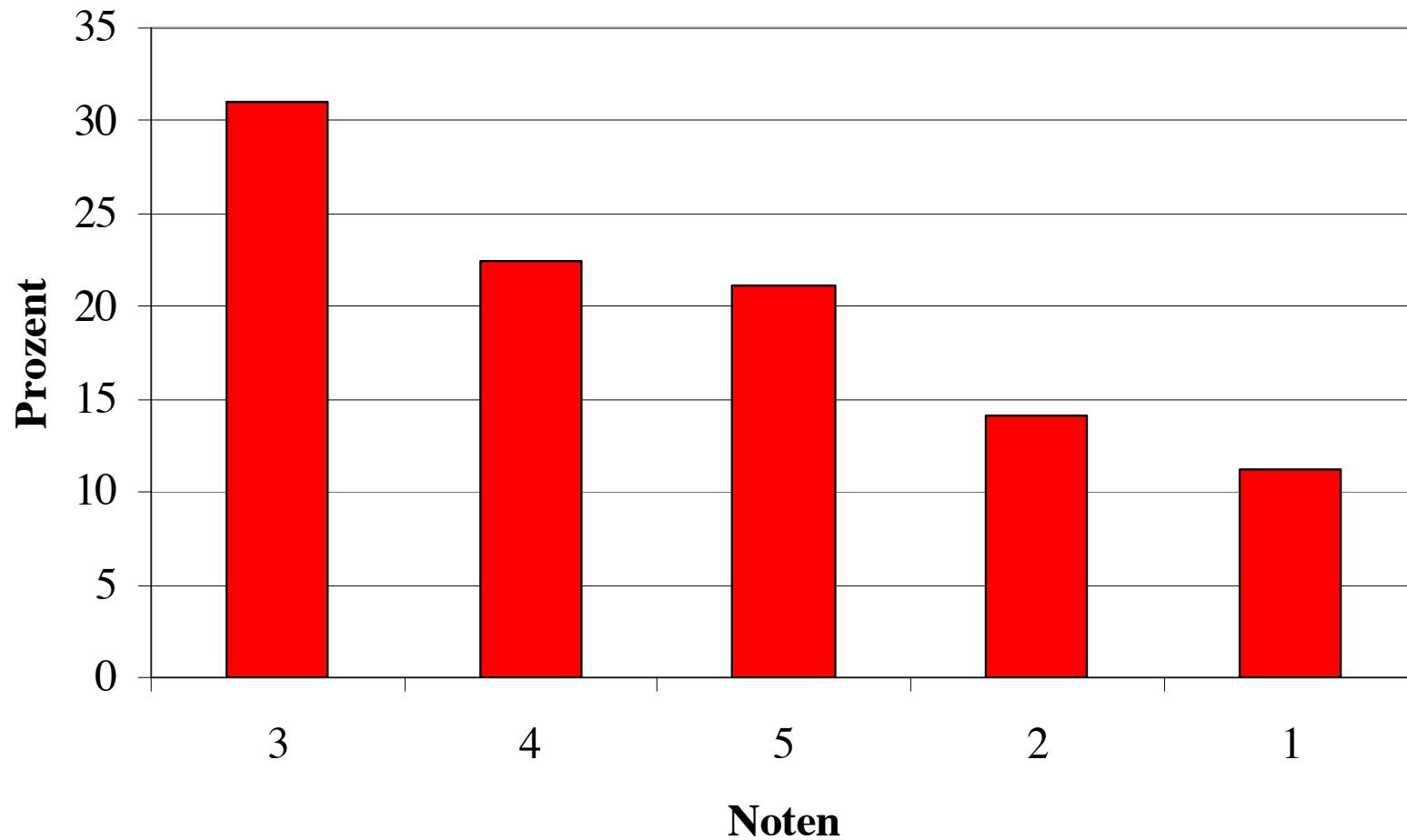


Abbildung 5: Säulendiagramm mit 3-D-Darstellung

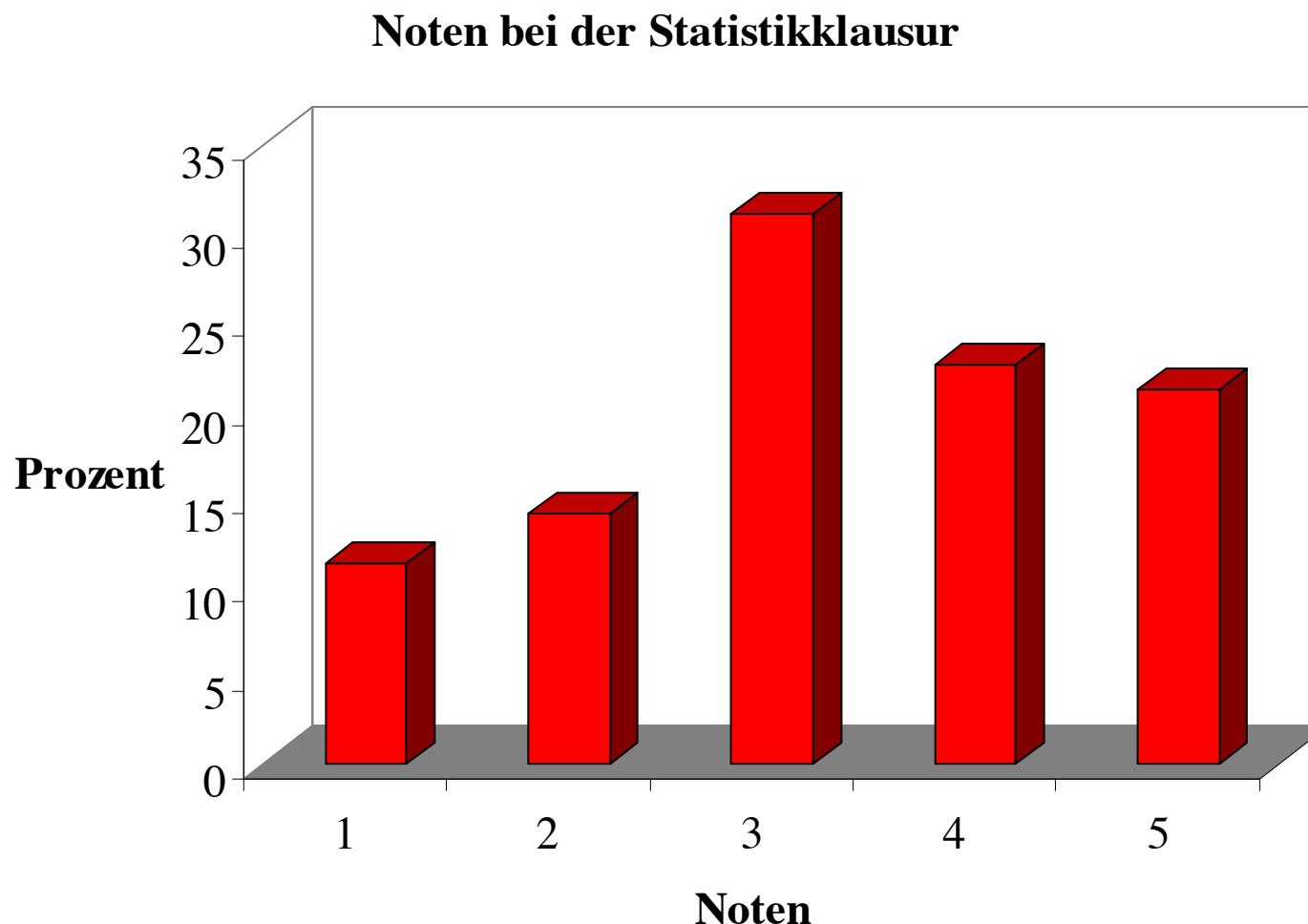
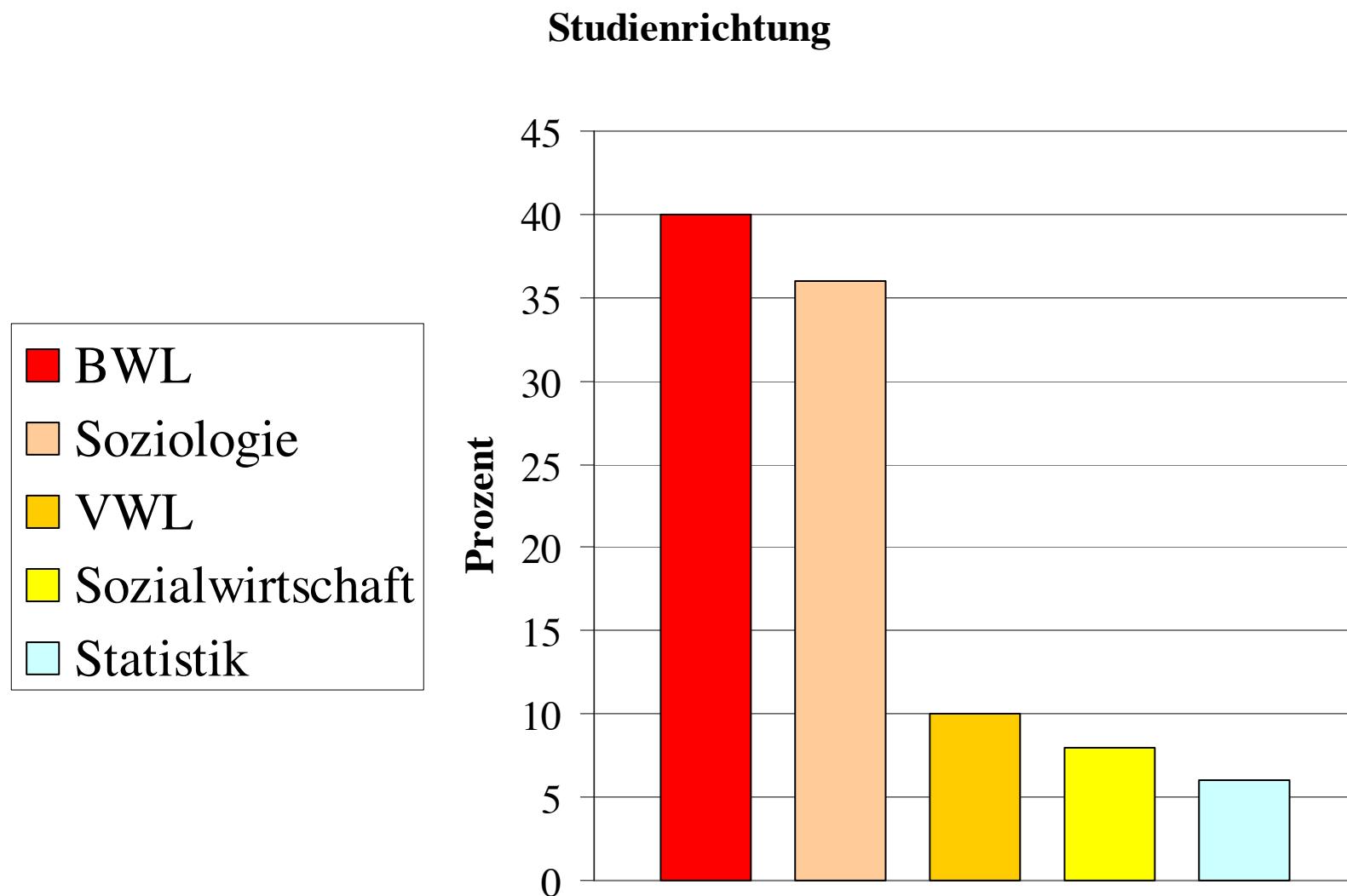
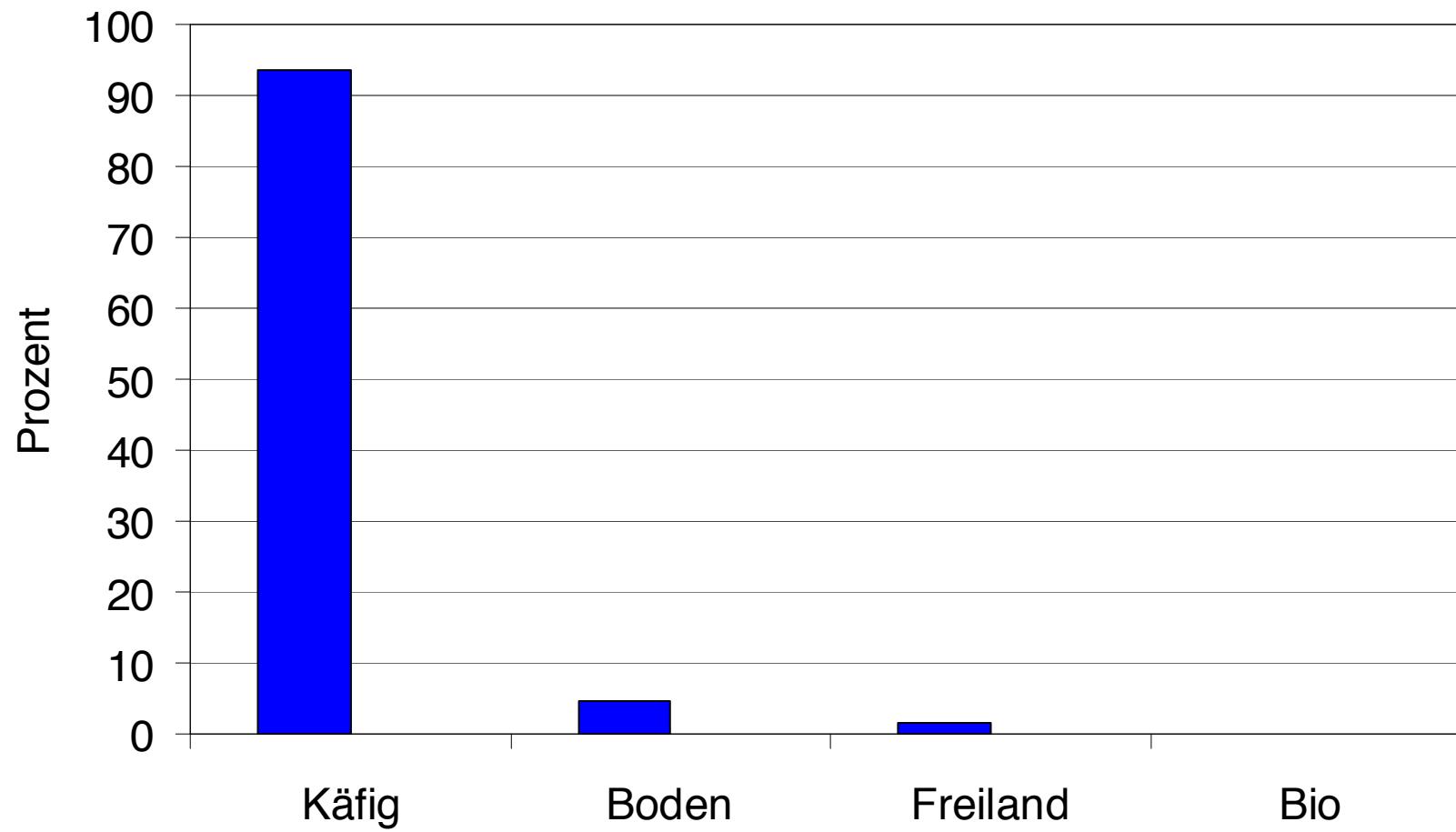


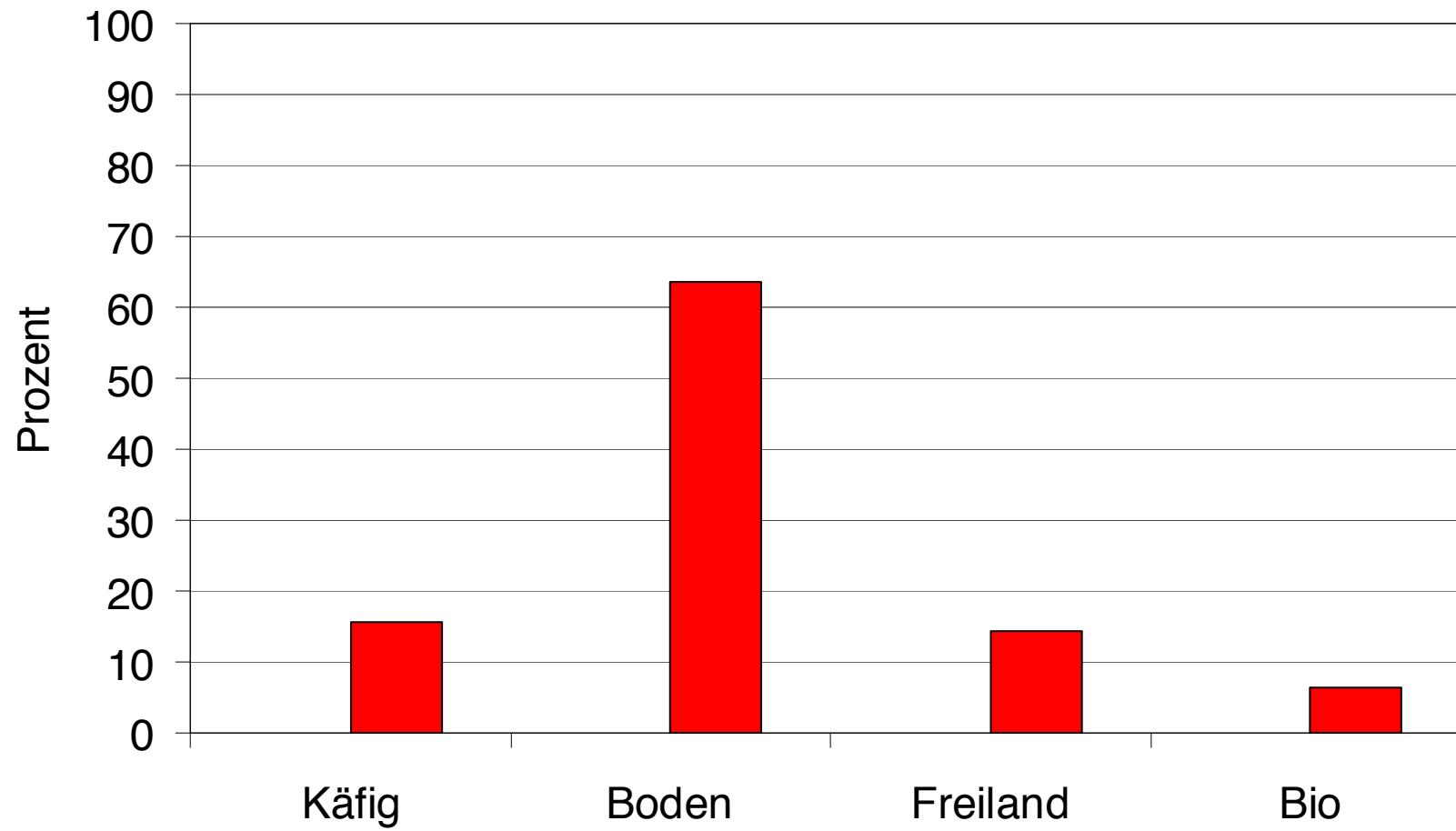
Abbildung 6: Säulendiagramm mit Legende



Hühnerhaltungsformen: 1995



Hühnerhaltungsformen: 2010



Hühnerhaltungsformen: Vergleich 1995 und 2010

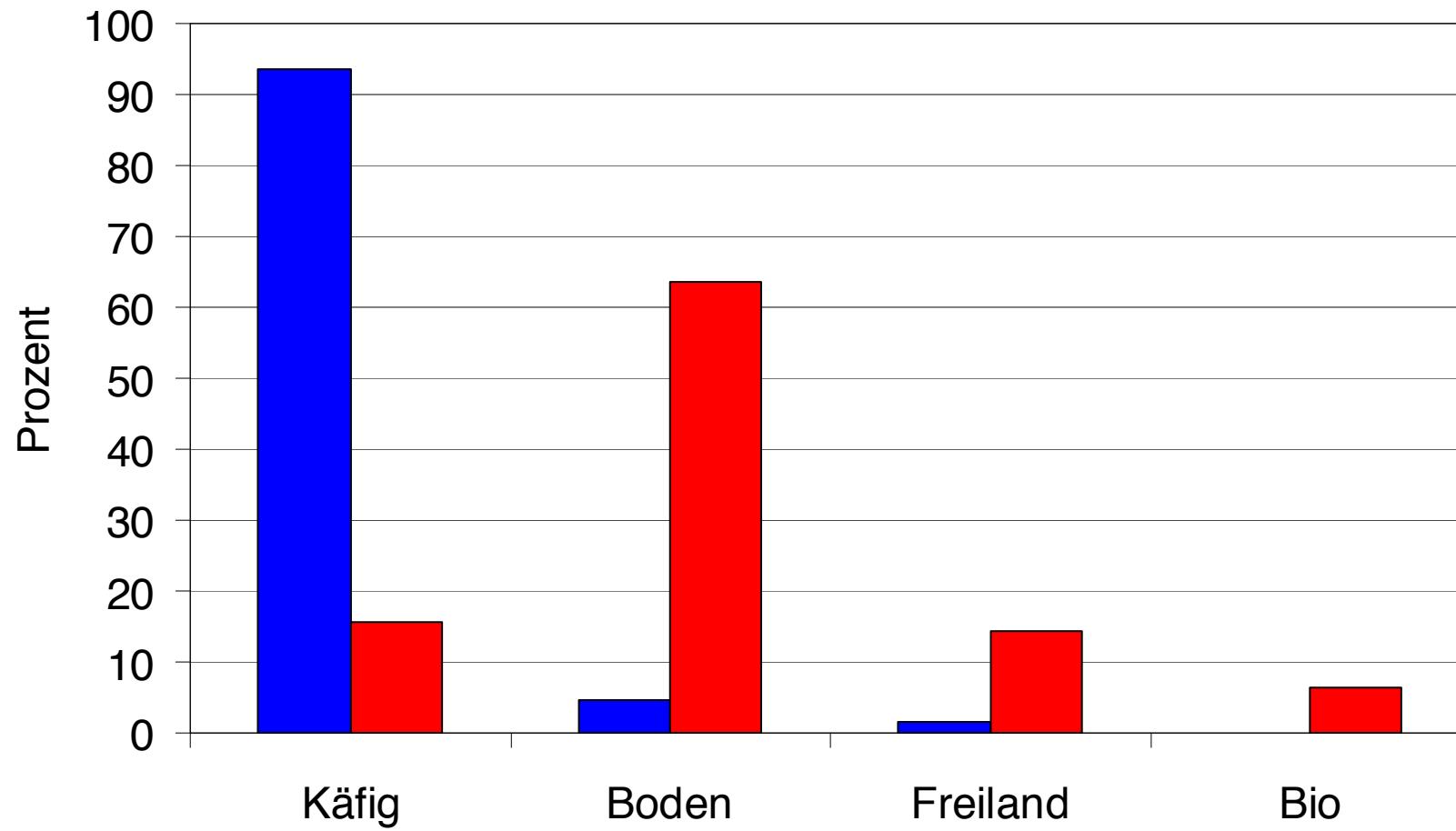


Abbildung 7: Ein Säulendiagramm einer Zeitreihe

Entwicklung der Jahresumsätze
in den Jahren 2004 - 2008

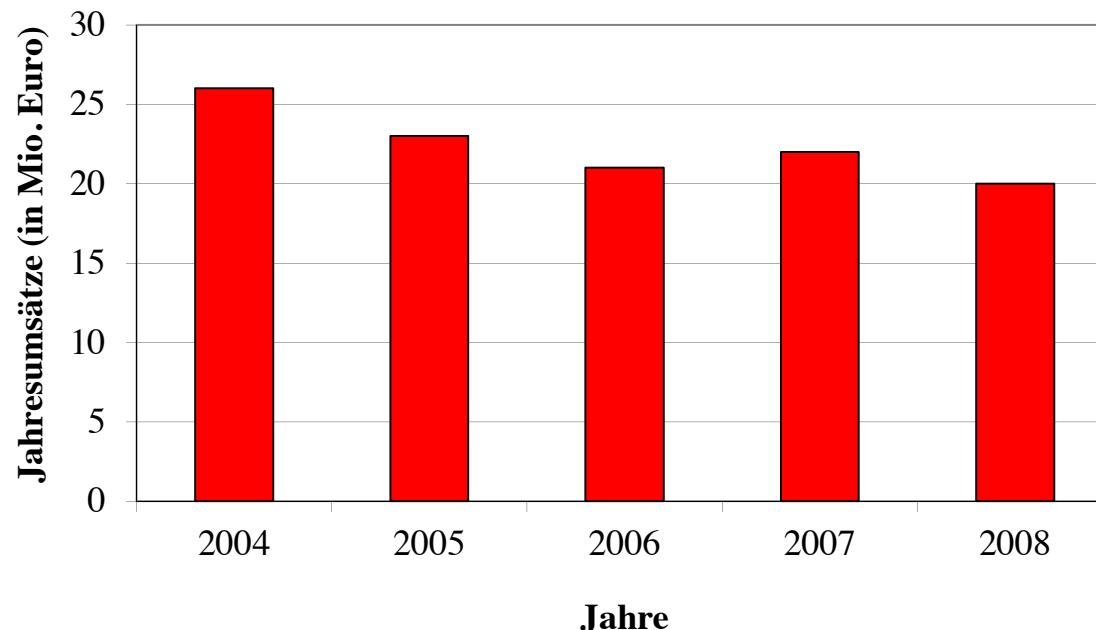
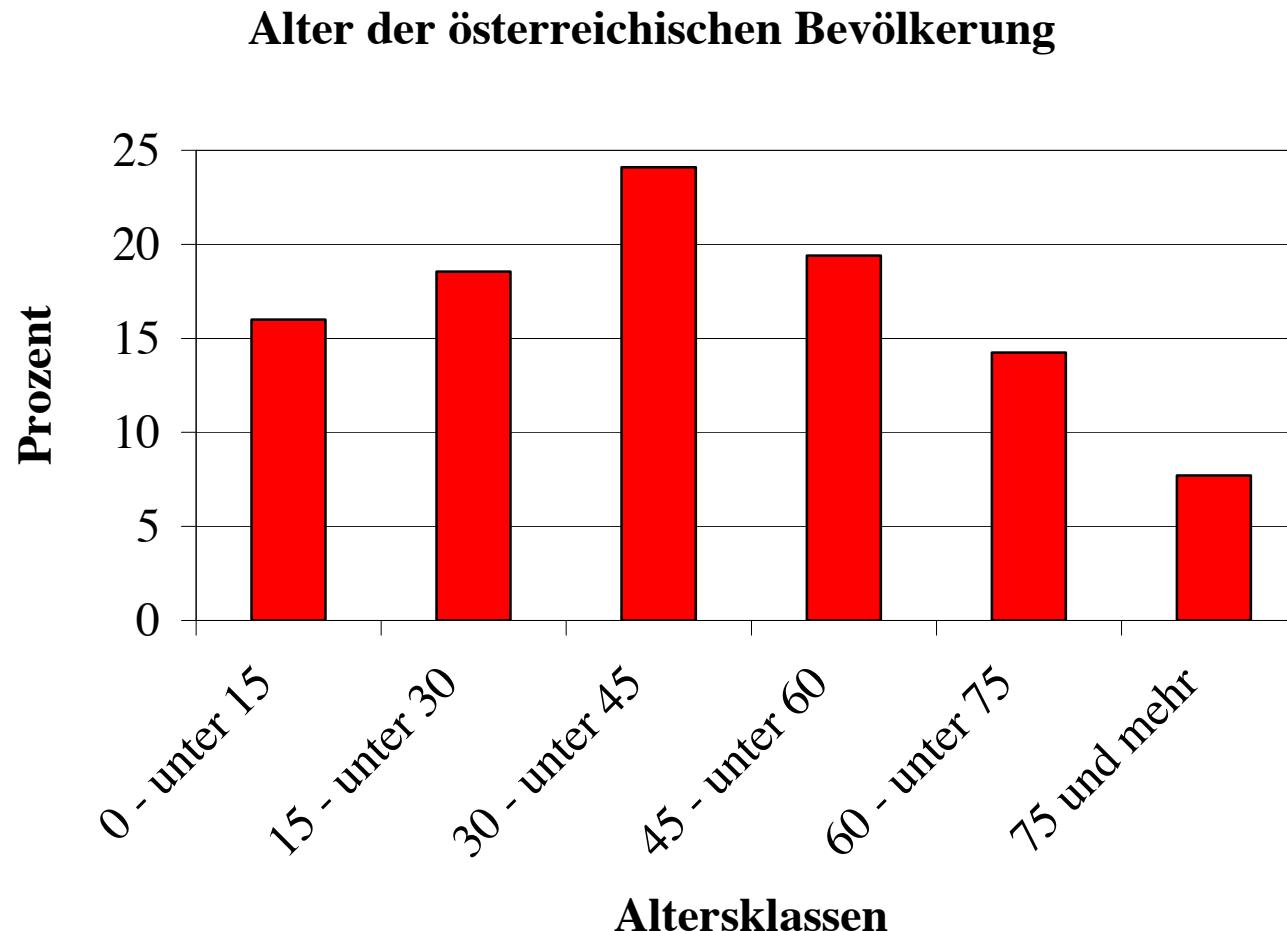
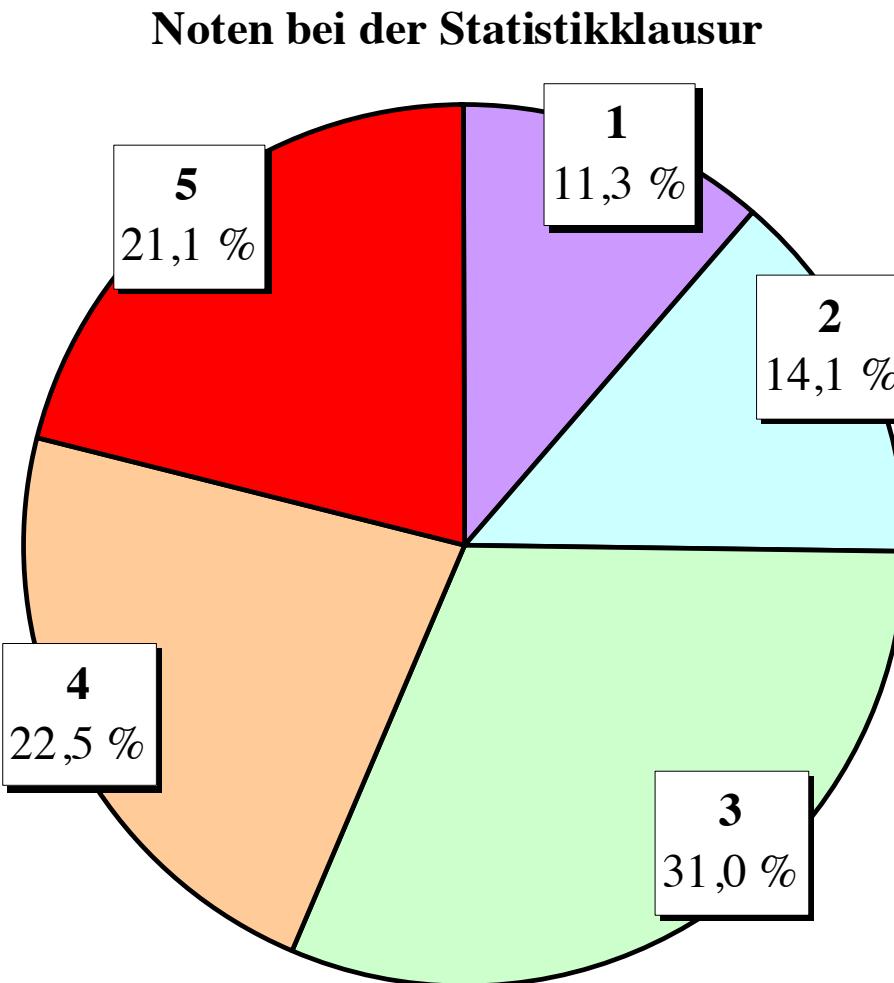


Abbildung 8: Säulendiagramm für ein in Intervalle zerlegtes Merkmal



Mitberücksichtigung verschiedener Intervallbreiten: Histogramm

Abbildung 9: Kreisdiagramm



Auch relative Summenhäufigkeiten sind im Kreisdiagramm ablesbar!

Regeln für die grafische Darstellung:

Säulendiagramme:

- Beschriftungen der x- und y-Achse sind unbedingt anzuführen
- Nullpunkt der Prozentzahlen auf der y-Achse sollte am Schnittpunkt zur x-Achse liegen

Säulen- und Kreisdiagramme:

- Überschriften und sind unbedingt anzuführen
- Ordnung innerhalb der Merkmalsausprägungen beibehalten
- 3-D-Darstellungen vermeiden
- Direkte Beschriftungen sind Legenden vorzuziehen



Die einfachste Grafik ist zumeist auch die Beste!

1.2.2 Gemeinsame Häufigkeitsverteilungen zweier Merkmale

Beispiel 6: Tabellarische Darstellung einer gemeinsamen Häufigkeitsverteilung zweier Merkmale

Häufigkeiten h:

		Studienrichtung					Summe
		BWL	Soz	VWL	SoWi	Stat	
Geschlecht	weiblich	//		/			
	männlich	/			/		
	Summe						

Häufigkeiten h:

		Studienrichtung					Summe	
Geschlecht		BWL	Soz	VWL	SoWi	Stat		
		weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	20	200
	Summe	200	180	50	40	30	500	

Relative Häufigkeiten p:

		Studienrichtung					Summe	
Geschlecht		BWL	Soz	VWL	SoWi	Stat		
		weiblich	0,22	0,24	0,04	0,06	0,04	0,60
	männlich	0,18	0,12	0,06	0,02	0,02	0,40	
	Summe	0,40	0,36	0,10	0,08	0,06	1	

Vergleich: *Häufigkeitsverteilung der Studienrichtung unter den Frauen und unter den Männern*

Studienrichtung

Geschlecht		BWL	Soz	VWL	SoWi	Stat	Summe
		weiblich	110	120	20	30	20
	männlich	90	60	30	10	10	200
	Summe	200	180	50	40	30	500

Zerlegung der Grundgesamtheit in die Teilgesamtheiten der F und M:

Beispiel 7: Tabellarische Darstellung einer **bedingten Verteilung**

Geschlecht		Studienrichtung					Summe
		BWL	Soz	VWL	SoWi	Stat	
	weiblich	0,367	0,400	0,067	0,100	0,067	1
	männlich	0,450	0,300	0,150	0,050	0,050	1

Beispiel 7: Tabellarische Darstellung einer **bedingten** Verteilung

		Studienrichtung					
		BWL	Soz	VWL	SoWi	Stat	Summe
Geschlecht	weiblich	0,367	0,400	0,067	0,100	0,067	1
	männlich	0,450	0,300	0,150	0,050	0,050	1

Unter den Frauen studieren 36,7 % BWL, 40,0 % Soz ...



Unter den Männern studieren ...



Korrekte Angabe der jeweiligen Grundgesamtheit, auf die sich die Prozentzahlen beziehen:

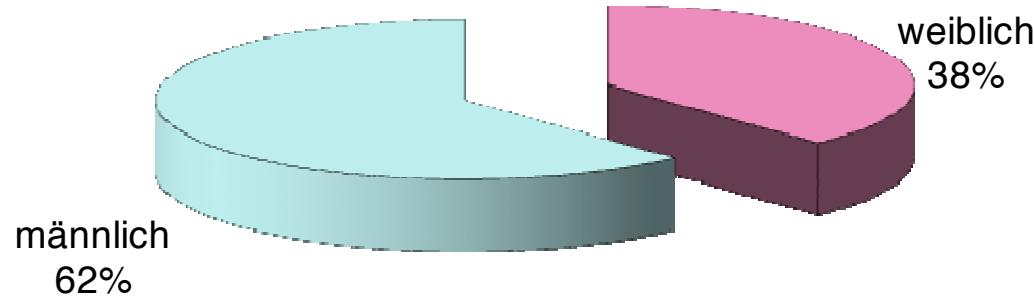
Richtig: „Zwei Drittel aller gefälschten Euro-Banknoten sind Fünfziger“

Falsch: „Zwei Drittel aller Fünfziger sind gefälschte Banknoten!“

Abbildung 10: Fehlinterpretation (DER STANDARD, 8.5. 1992)

*Unsinn in
den medien*

Durchgefallene nach Geschlecht



„Die Mädchen – oft zahlenmäßig überlegen – stellen nur etwas mehr als ein Drittel der *Sitzenbleiber*. Bei den Burschen dagegen erreichen 62% ihr Klassenziel nicht.“

1.3 Kennzahlen statistischer Verteilungen

Informationsbündelung auf einen einzigen Repräsentanten der Verteilung

1.3.1 Kennzahlen der Lage

Mittelwert (oder Durchschnitt):

Idee: Stellvertreter für alle Daten ist jener Wert, der sich bei gleichmäßiger Aufteilung der Summe aller aufgetretenen Daten (=Merkmalssumme) auf die Erhebungseinheiten ergeben würde

Beispiel:

Einkommen von fünf Personen: 1.000, 3.000, 4.000, 1.000 und 1.000 €

Merkmalssumme: $1.000 + 3.000 + 4.000 + 1.000 + 1.000 = 10.000 \text{ €}$

Gleichmäßige Aufteilung: $10.000 : 5 = 2.000 \text{ €}$

Formale Umsetzung der Idee des Mittelwerts:

Zeichen für den Mittelwert: \bar{x} (sprich: „x quer“)

N ... Anzahl der Erhebungseinheiten

x_1 ... Merkmalsausprägung der 1. Erhebungseinheit,

x_2 ... Merkmalsausprägung der 2. Erhebungseinheit und so fort

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} \quad (2)$$

Merkmalssumme (2. Variante): $1.000 \cdot 3 + 3.000 \cdot 1 + 4.000 \cdot 1 = 10.000 \text{ €}$



Merkmalsausprägungen · Häufigkeiten

k ... Anzahl der verschiedenen Merkmalsausprägungen

h_1 ... Häufigkeit der 1. Merkmalsausprägung,

h_2 ... Häufigkeit der 2. Merkmalsausprägung und so fort

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot h_i}{N} \quad (2a)$$

Auch mit den relativen Häufigkeiten p :

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot h_i}{N} = \sum_{i=1}^k x_i \cdot \frac{h_i}{N} = \sum_{i=1}^k x_i \cdot p_i \quad (2b)$$

Beispiel 8: Berechnung des Mittelwerts (Fortsetzung von Beispiel 3)

Punktezahlen Häufigkeit Relative Häufigkeit

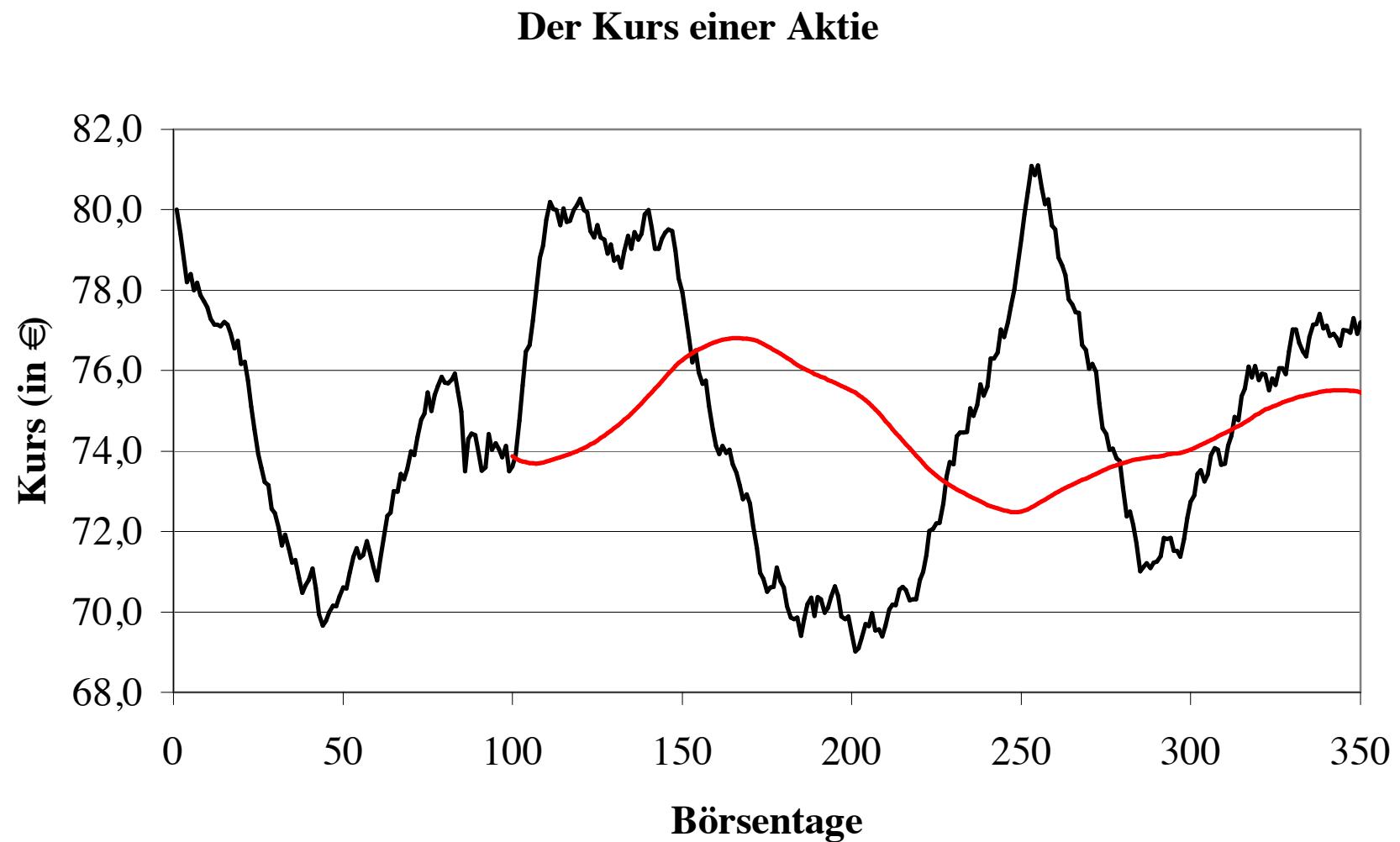
0	1	0,007
1	3	0,021
2	10	0,070
3	16	0,113
4	32	0,225
5	44	0,310
6	20	0,141
7	16	0,113

$$(2a): \bar{x} = \frac{\sum_{i=1}^k x_i \cdot h_i}{N} = \frac{(0 \cdot 1 + 1 \cdot 3 + \dots + 7 \cdot 16)}{142} = \frac{651}{142} = 4,58$$

oder mit (2b): $\bar{x} = \sum_{i=1}^k x_i \cdot p_i = 0 \cdot 0,007 + 1 \cdot 0,021 + \dots + 7 \cdot 0,113 = 4,58$

Beispiel für ein Anwendungsgebiet: [Zeitreihenanalyse](#)

Abbildung 11: Gleitende Mittelwerte in Zeitreihen



Der Mittelwert eignet sich nur für metrische Merkmale (und auch da nicht immer →)

Beispiel 9: Der Mittelwert von Wachstumsfaktoren

Vor drei Jahren: Umsatz von 20 Millionen €. In den drei Jahren seither jährliche Umsatzzuwächse von 10, 90 und 50 %.

Um wie viel Prozent ist der Umsatz pro Jahr durchschnittlich gestiegen?

Mittelwert: $(10+90+50):3 = 50\%$.

Verlauf des Umsatzes (in Mio. €):

1. Jahr: $20 \cdot 1,10 = 22 \text{ €}$ (1,10 ist der **Wachstumsfaktor**)

2. Jahr: $22 \cdot 1,90 = 41,8 \text{ €}$

3. Jahr: $41,8 \cdot 1,50 = 62,7 \text{ €}$

Mit dem Mittelwert der Prozentzahlen:

1. Jahr: $20 \cdot 1,5 = 30 \text{ €}$

2. Jahr: $30 \cdot 1,5 = 45 \text{ €}$

3. Jahr: $45 \cdot 1,5 = 67,5 \text{ €}$! ?

Welcher konstante Wachstumsfaktor würde also 62,7 ergeben?

$$20 \cdot g^3 = 62,7?$$

Aus $g^3 = \frac{62,7}{20} = 3,135$ folgt $g = \sqrt[3]{3,135} = \underline{\underline{1,464}}$

Auch aus Wachstumsfaktoren: $g = \sqrt[3]{1,1 \cdot 1,9 \cdot 1,5} = \sqrt[3]{3,135} = \underline{\underline{1,464}}$

... geometrischer Mittelwert der Wachstumsfaktoren

→ das durchschnittliche jährliche prozentuelle Wachstum ist 46,4 %.

Häufiges Anwendungsgebiet des geometrischen Mittelwerts: prozentuelles Wachstum von **Indizes** (z.B. Preisindex für die Lebenshaltung, Aktienindizes ...)

Gegenstand eines Indexes: Preisliche Entwicklung eines **Warenkorbs**

Inflationsrate: Quotient des aktuellen Werts des Preisindexes für die Lebenshaltung und des Werts vor genau einem Jahr

Probleme: Relevanz des Warenkorbs für den Einzelnen, Veralterung des Warenkorbs



Eine negative Eigenschaft von Mittelwert und geometrischem Mittelwert: Empfindlichkeit gegenüber untypischen Merkmalsausprägungen („**Ausreißern**“)

Möglicherweise besser: Berechnung des Mittelwertes ohne Ausreißer

Wichtig: Korrekte Beschreibung der Grundgesamtheit, auf die sich ein Mittelwert bezieht

Andreas Quatember:

Statistik ohne *Angst* vor Formeln

**Eine verständnisorientierte Einführung in
die Grundlagen der Statistik**

Bedeutung des Faches Statistik



Statistik ist Alltag!



Beispiele:

- Analysen des Finanzmarktes, High-Frequency-Trading, Stopp-Loss-Automatik
- Big Data, Kundendatenanalysen im Web (Amazon, iTunes, Facebook, Google)
- Statistische Analysen im Sport: zB Matchstatistiken im Fußball

UEFA.com

LIVE | PULS

BAYERN	1 - 1	ARSENAL
64%	BALL POSSESSION	36%
14	TOTAL ATTEMPTS	8
9	ATTEMPTS ON TARGET	5
4	BLOCKS AND SAVES	8
6	CORNERS	5
4	OFFSIDES	3
119.62 km	DISTANCE COVERED	114.51 km
631 (83%)	PASSES COMPLETED	269 (67%)
14	FOULS COMMITTED	14
2 / 0	YELLOW/RED CARDS	3 / 0

Image des Faches

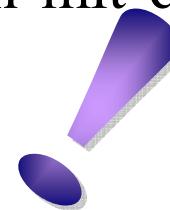


„,... und jetzt noch etwas für die Statistiker unter unseren Zusehern“

„Mit Statistik lässt sich alles beweisen!“

„Glaube keiner Statistik, die Du nicht selber gefälscht hast!“

Verwechslung der Qualität der statistischen Methoden mit der Qualität ihrer Anwendung



Was versteht man unter **Statistik**? → Methoden der Datenanalyse zum Zweck der Informationsbündelung



Gliederung in 3 Teile:

1 Beschreibende Statistik

Es liegen vollständige Daten über eine Gesamtheit vor

2 Wahrscheinlichkeitstheorie

Kombiniert 1 mit 3

3 Schließende Statistik

Es liegen nur Daten über einen ausgewählten Teil der Gesamtheit vor

1 Beschreibende Statistik

1.1 Grundbegriffe

Erhebungseinheiten: Objekte, über die Daten erhoben werden

Grundgesamtheit: Gesamtheit aller Erhebungseinheiten

Merkmal: Eine interessierende Eigenschaft

Merkmalsausprägungen: Die einzelnen möglichen Werte eines Merkmals

Wertebereich: Alle Merkmalsausprägungen

Beispiel 1: Grundbegriffe einer statistischen Erhebung

Erhebung der Punkteverteilung bei der Statistikklausur:

Grundgesamtheit:	alle Prüflinge
Merkmal:	Punkte
Merkmalsausprägungen:	0, 1, 2, ...

Erhebung der Zufriedenheit von Kunden:

Grundgesamtheit:	alle Kunden
Merkmal:	Zufriedenheit mit der Beratung
Merkmalsausprägungen:	sehr zufrieden, eher zufrieden, teils-teils, eher unzufrieden, sehr unzufrieden

Erhebung des besten Kinofilms:

Grundgesamtheit:	alle teilnahmewilligen Leser und -innen
Merkmal:	bester Film
Merkmalsausprägungen:	Film 1, Film 2, ...

■ Merkmalstypen (1. Unterscheidung): **nominal** - **ordinal** - **metrisch**

nominal: Unterscheidung der Merkmalsausprägungen dem Namen nach
(Prototyp: Geschlecht)

ordinal: Merkmalsausprägungen besitzen eine natürliche Reihenfolge
(Prototyp: Schulnoten)

metrisch: Merkmalsausprägungen lassen sich reihen **und** sind Vielfache
einer Einheit (Prototyp: Körpergröße)

■ Merkmalstypen (2. Unterscheidung): **diskret** - **stetig**

diskret: Wertebereich umfasst nur bestimmte Merkmalsausprägungen
(Prototyp: Schulnoten)

stetig: ... umfasst alle reellen Werte eines Intervalls (Körpergröße)

Kodierung der Merkmalsausprägungen

1. Geschlecht: weiblich (=1) männlich (=2)

2. Alter (in vollendeten Lebensjahren): Jahre

3. Wie schätzen Sie die didaktisch-methodische Qualität der LVA ein?

1 (=sehr gut) 2 3 4 5 (=sehr schlecht)

4. Waren die angegebenen Lernunterlagen hilfreich?

1 2 3 4 5 (1=sehr hilfreich, ... , 5=überhaupt nicht hilfreich)

Dateneingabe für die elektronische Verarbeitung (z.B. in Excel):

	A	B	C	D	E	F	G
1.	Erhebungseinheit:	2	21	1	3		
2.	Erhebungseinheit:	1	38	2	2		
3.	Erhebungseinheit:	3					
	4						

Antwort auf 1. Frage

Antwort auf 2. Frage

Beispiel 2: Merkmalstypen

Merkmal	Merkmalsausprägungen	n / o / m	d / s
Familienstand	ledig (=1), verheiratet (=2), geschieden (=3), verwitwet (=4)	nominal	diskret
100-m-Zeiten	11,21 sec., 11,24 sec., ...	metrisch	stetig
Preis eines Sportartikels	29,90 €, 34,90 €, ...	metrisch	diskret
Platzierungen in einem 100m-Lauf	1., 2., 3., ...	ordinal	diskret
Weitsprungleistung (in ganzen cm)	516 cm, 492 cm, ...	metrisch	stetig

1.2 Tabellarische und graf. Darstellung von Häufigkeitsverteilungen

1.2.1 Häufigkeitsverteilungen einzelner Merkmale

Tabellarische Darstellung

Beispiel 3: Tabellarische Darstellung einer Häufigkeitsverteilung

Punktezahlen	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenhäufigkeit
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1

Häufigkeiten (h): Erster Überblick

Relative Häufigkeiten oder Anteile (p) einer Merkmalsausprägung i:

$$p_i = h_i / N$$

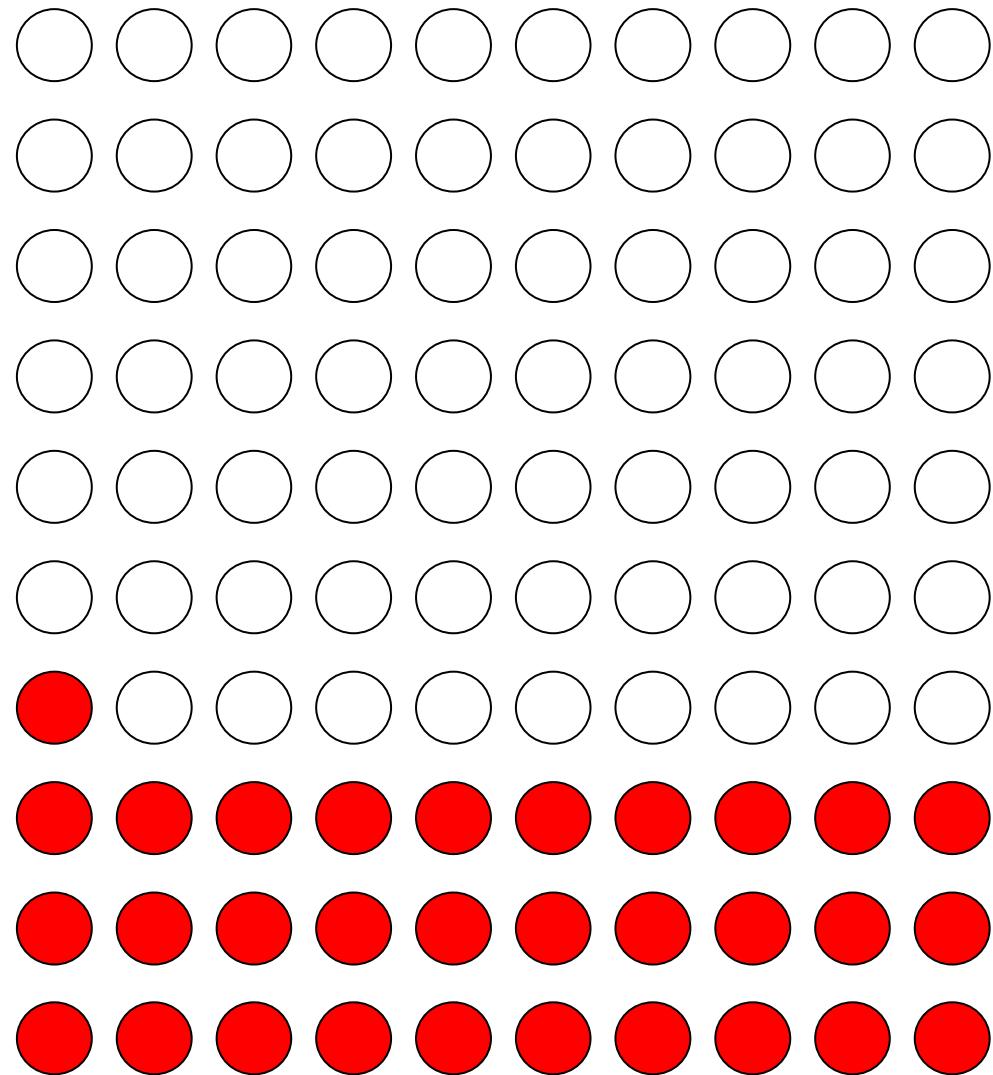
(Formel 1)

Punktzahlen	Häufigkeit h	Relative Häufigkeit p	Prozent	Relative Summenh.
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1

$$N=142$$

Prozentzahlen: $p_i \cdot 100$

Abbildung 1: Die Prozentzahlen (am Beispiel von 31,0 %)



Relative Summenhäufigkeit (oder **empirische Verteilungsfunktion**):
 Summe der relativen Häufigkeiten einer Merkmalsausprägung und aller kleineren Merkmalsausprägungen

z.B. die relative Summenhäufigkeit zur Merkmalsausprägung 3:

Punktezahlen	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenh.
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1

Nur sinnvoll bei metrischen oder ordinalen Merkmalen!

Zusammenfassung von Merkmalsausprägungen zu Intervallen:

Beispiel 4: Tabellarische Darstellung einer Häufigkeitsverteilung

Altersklasse	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenhäufigkeit
0 – unter 15	1.317.707	0,160	16,0	0,160
15 – unter 30	1.526.909	0,185	18,5	0,345
30 – unter 45	1.984.501	0,241	24,1	0,586
45 – unter 60	1.596.849	0,194	19,4	0,780
60 – unter 75	1.173.166	0,142	14,2	0,922
75 und mehr	634.174	0,077	7,7	1
N=8.233.306				

Häufigkeiten, relative Häufigkeiten und relative Summenhäufigkeiten beziehen sich auf ganze Intervalle von Merkmalsausprägungen

<http://www.ifas.jku.at>

Institut für Angewandte Statistik

Gast [Login](#) Personen JKU.at Suchen Schnellzugriff

[JKU](#) | [IFAS](#) | [Unsinn in den Medien](#)

 The IFAS logo features the letters "IFAS" in a bold, blue, sans-serif font. Above the "I", there is a red stylized shape resembling a dome or a graph curve.

Unsinn in den Medien – vom sorglosen Umgang mit Daten

Irren ist menschlich! "Journalistischer Irrtum" in Zusammenhang mit statistischen Daten ist insofern gefährlich, als solche Irrtümer – verbreitet in Zeitungen, Magazinen oder im TV – Bestandteil des "Wissens" der Bevölkerung in Diskussionen im Freundeskreis oder am Stammtisch oder in der Politik werden. Wie häufig selbst einfachste Kennzahlen falsch interpretiert werden, dokumentiert diese kommentierte Seite, die ständig aktualisiert wird (den Artikeln des aktuellen Jahres wird die Jahreszahl hinzugefügt). Diese Kommentare sind bewusst launig verfasst. Dabei sollen aber jedenfalls die Fehler und keinesfalls die Verfasser im Mittelpunkt der Betrachtung stehen. Denn es gilt: siehe oben.

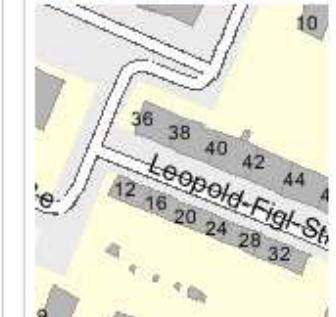
Für den Inhalt verantwortlich: [Andreas Quatember](#)

[Übersicht](#) [Unsinn in den Medien – NEU!](#)

News

- [Neuer Institutsvorstand](#)
Prof. Müller übernimmt mit 1. Jänner 2011 die Leitung des IFAS ... [mehr](#)
- [Mitherausgeberschaft](#)
Professor Müller zum neuen Mitherausgeber der "Statistical Papers" bestellt ... [mehr](#)
- [Neues Web-Layout](#)
Das IFAS zeigt sich in neuem Gewand! ... [mehr](#)

Lageplan

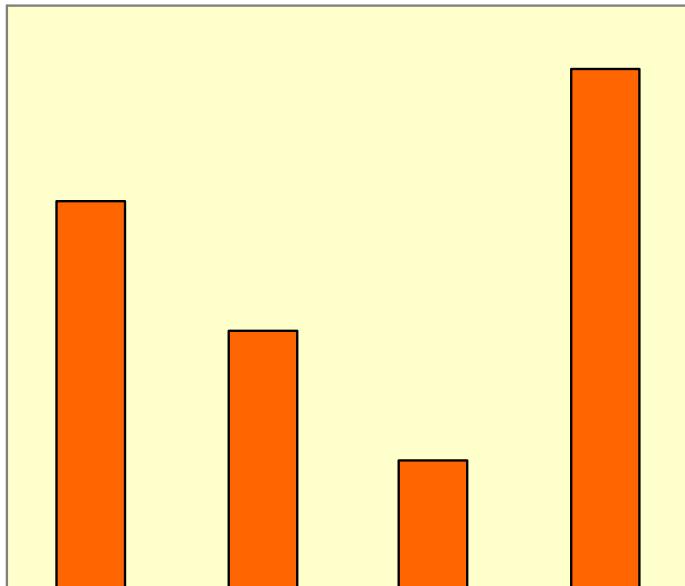


A small map showing the location of the IFAS office within a larger building complex. The map includes street names like "Leopold-Figl-Str." and house numbers 10, 36, 38, 40, 42, 44, 12, 16, 20, 24, 28, 32, 9. The IFAS logo is overlaid on the map at its approximate location.

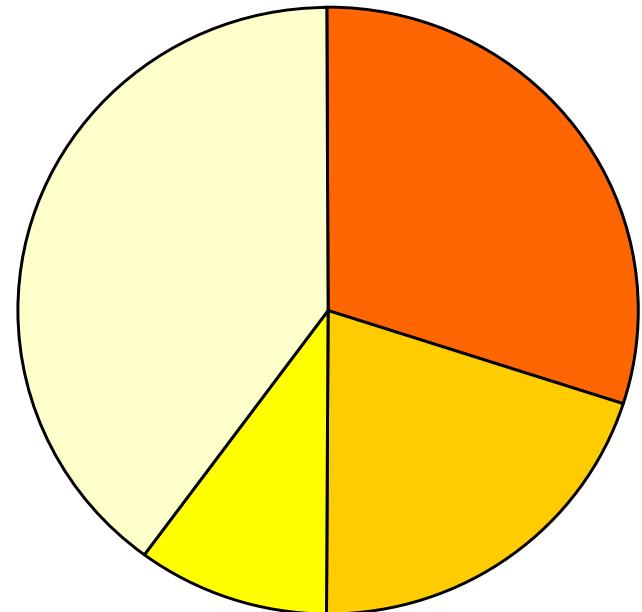
Grafische Darstellung

Aufgabe: Die wesentlichsten Informationen „auf einen Blick“ erfassbar machen

Säulendiagramm:
(Balken-, Stabdiagramm)



Kreisdiagramm:
(Kuchen-, Tortendiagramm)



Beispiel 5: Tabellarische Darstellung der Häufigkeitsverteilung

Note	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenh.
1	16	0,113	11,3	0,113
2	20	0,141	14,1	0,254
3	44	0,310	31,0	0,564
4	32	0,225	22,5	0,789
5	30	0,211	21,1	1

Abbildung 2: Ein Säulendiagramm

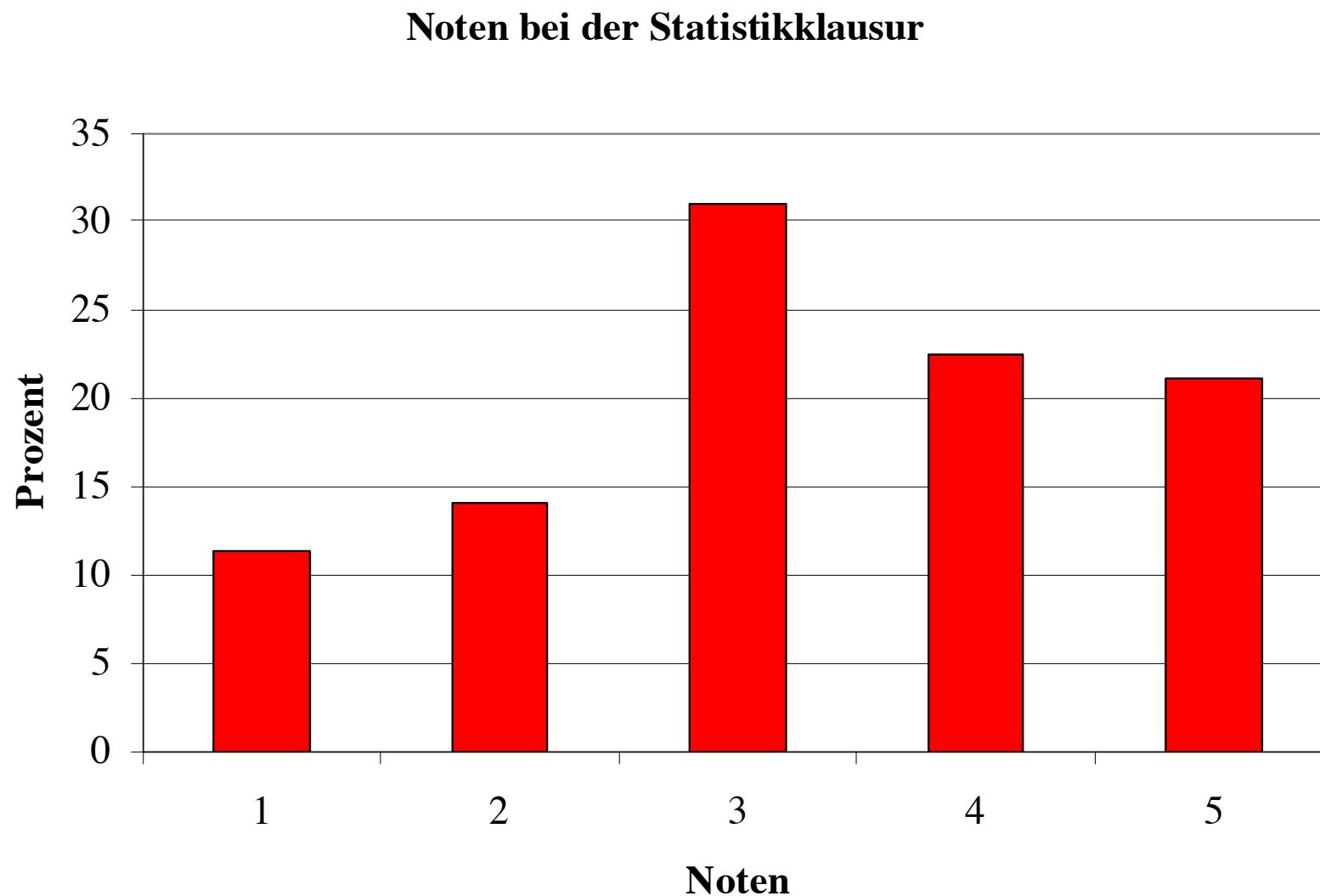
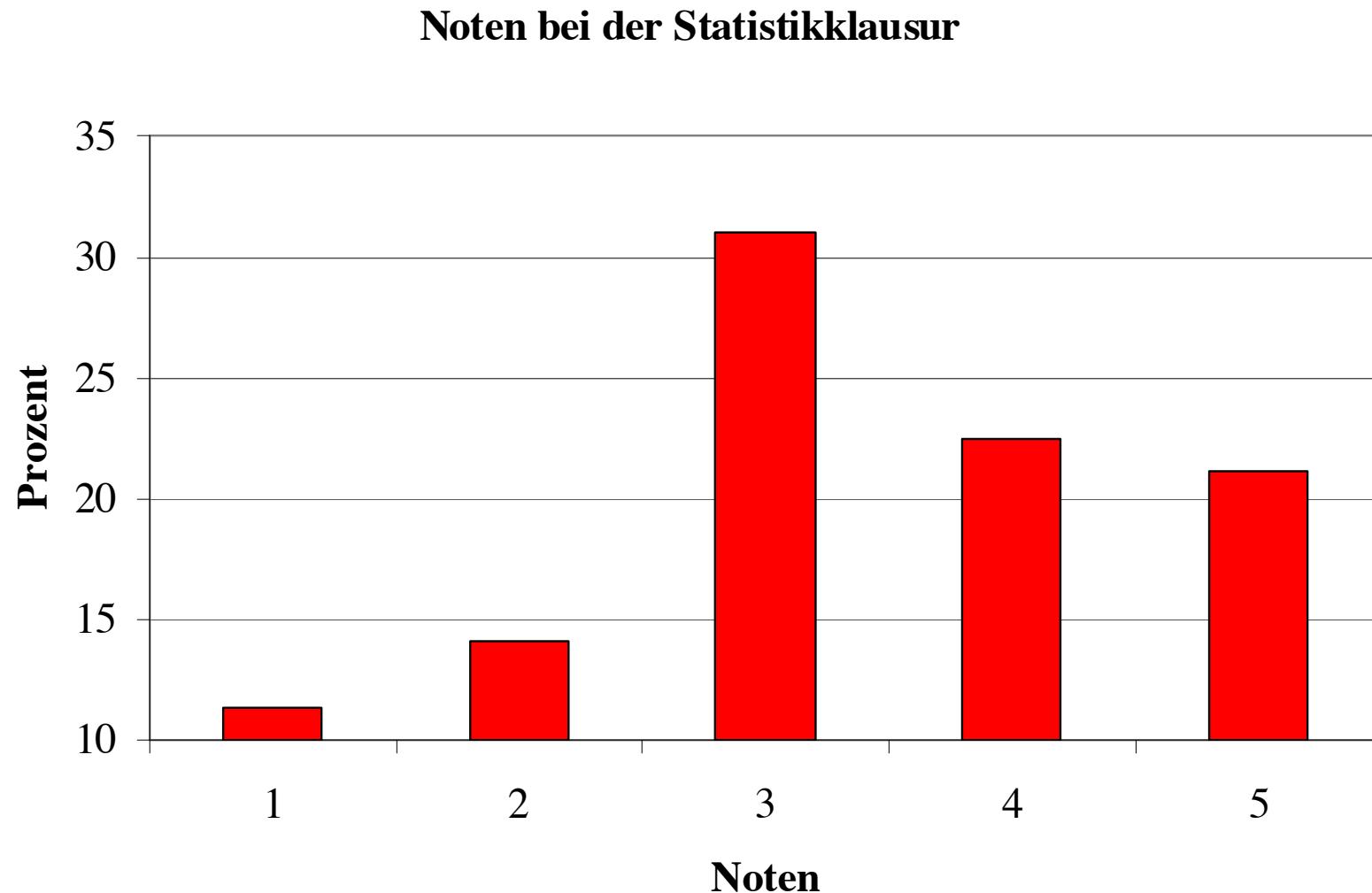


Abbildung 3: Säulendiagramm mit verschobenem Nullpunkt auf der y-Achse



Unsinn in
den Medien

The chart displays 'Gelenkschmiere in mg/ml' on the y-axis (ranging from 0 to 50) against time on the x-axis. It shows a baseline at approximately 40 mg/ml, followed by a sharp drop to about 10 mg/ml, and then a large green bar reaching nearly 50 mg/ml, labeled 'nachher' (after). A red arrow points upwards from the baseline to the green bar, with the text '25 % mehr Gelenkschmiere' (25% more joint lubrication).

294 018

1 B.DREXEL
GELENK-VITAL
SET 3-tlg. Press-
linge, Tinktur &
1 gratis Messbecher

HSE24 Preis
€ **79,99**
CHF 137,95

+Versand: €5,95/CHF7,95

AdT-Tiefpreis
€ **29,99**
CHF 51,95

277 430 BDE Duo Karde Vital Massage Sprays € 19,99

Noch
3942
Stück

Angebot des Tages

0800 29 888 88
EASy 0800 29 888 29

HSE24

Abbildung 4: Säulendiagramm mit umgeordneten Merkmalsausprägungen

Noten bei der Statistikklausur

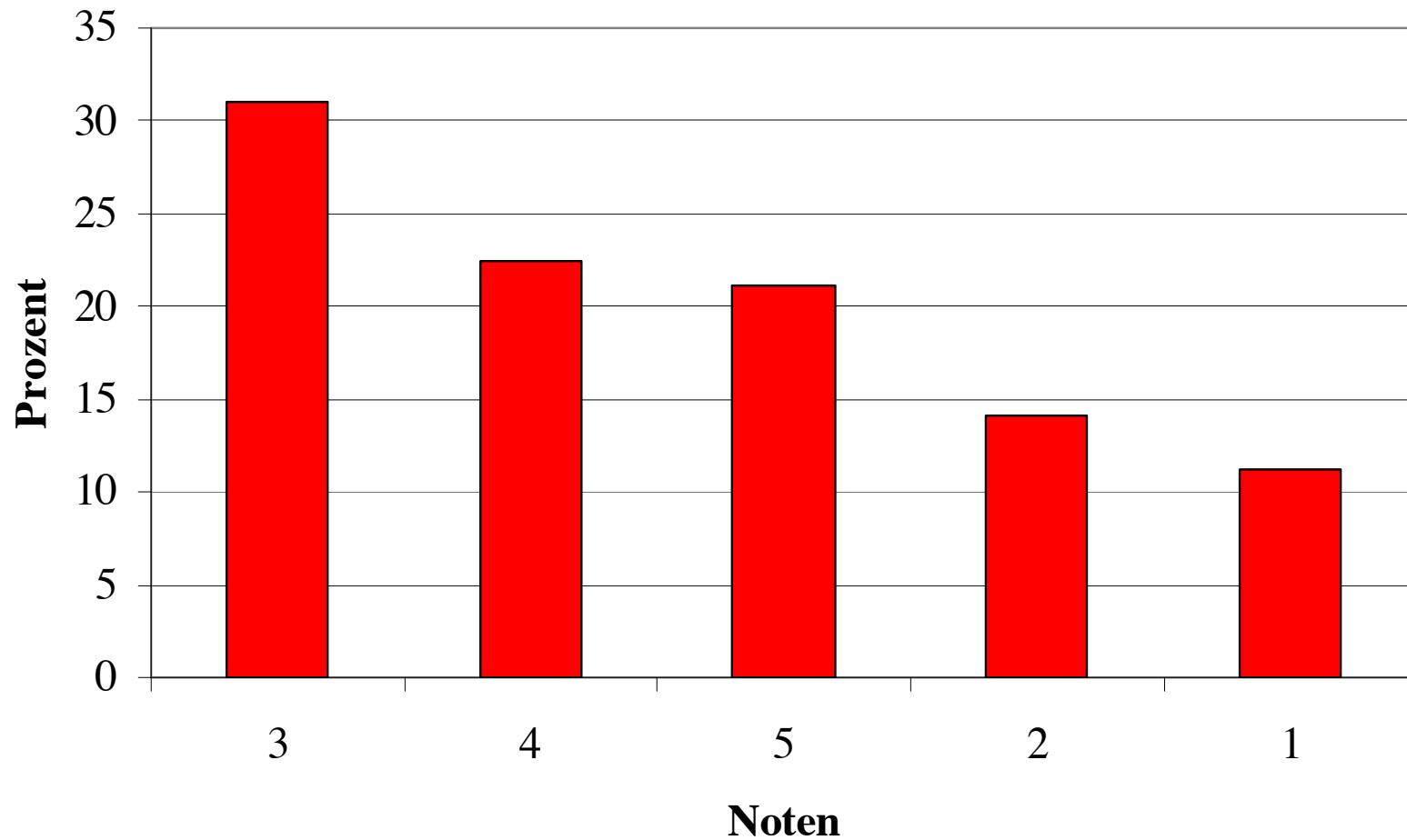


Abbildung 5: Säulendiagramm mit 3-D-Darstellung

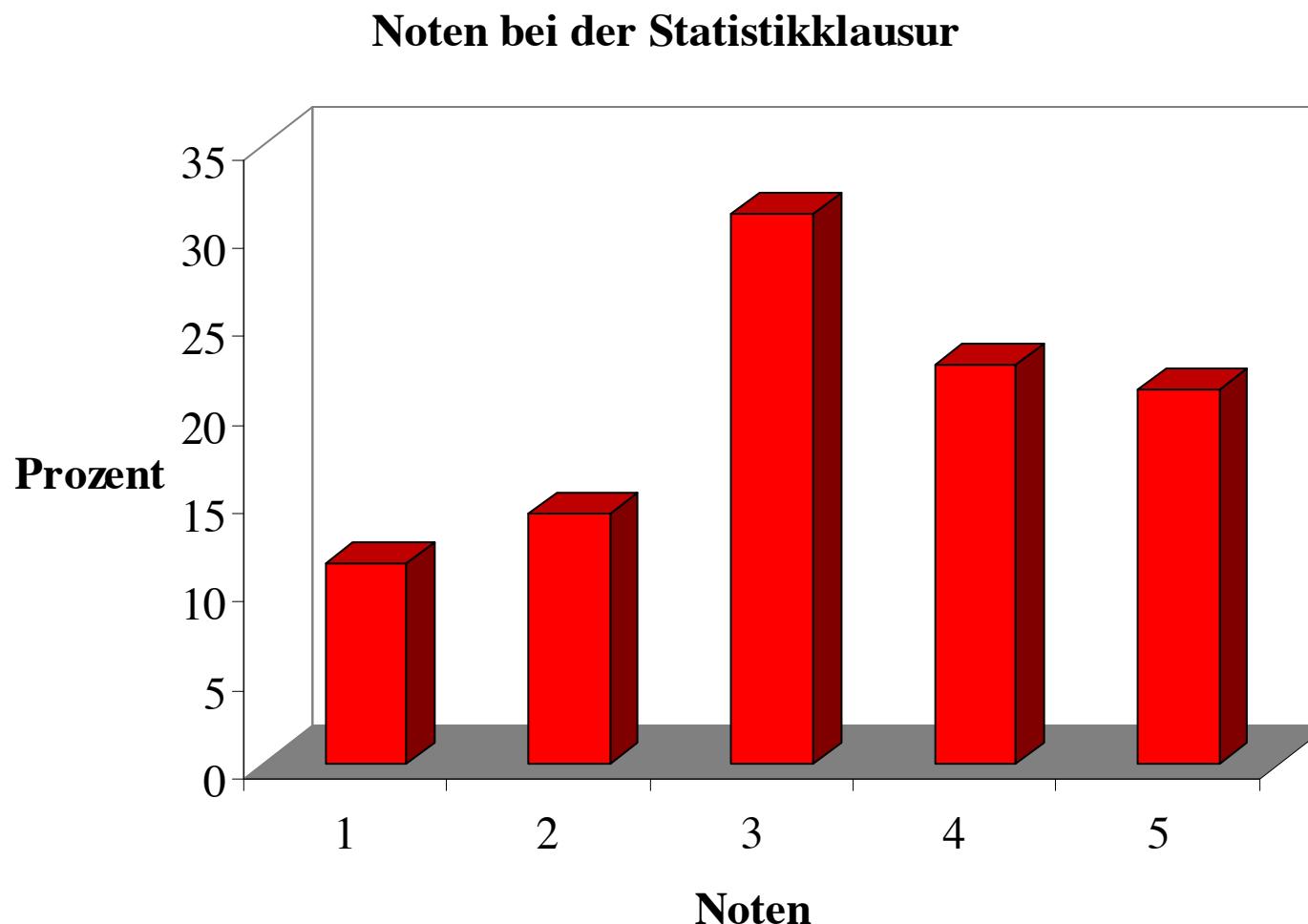
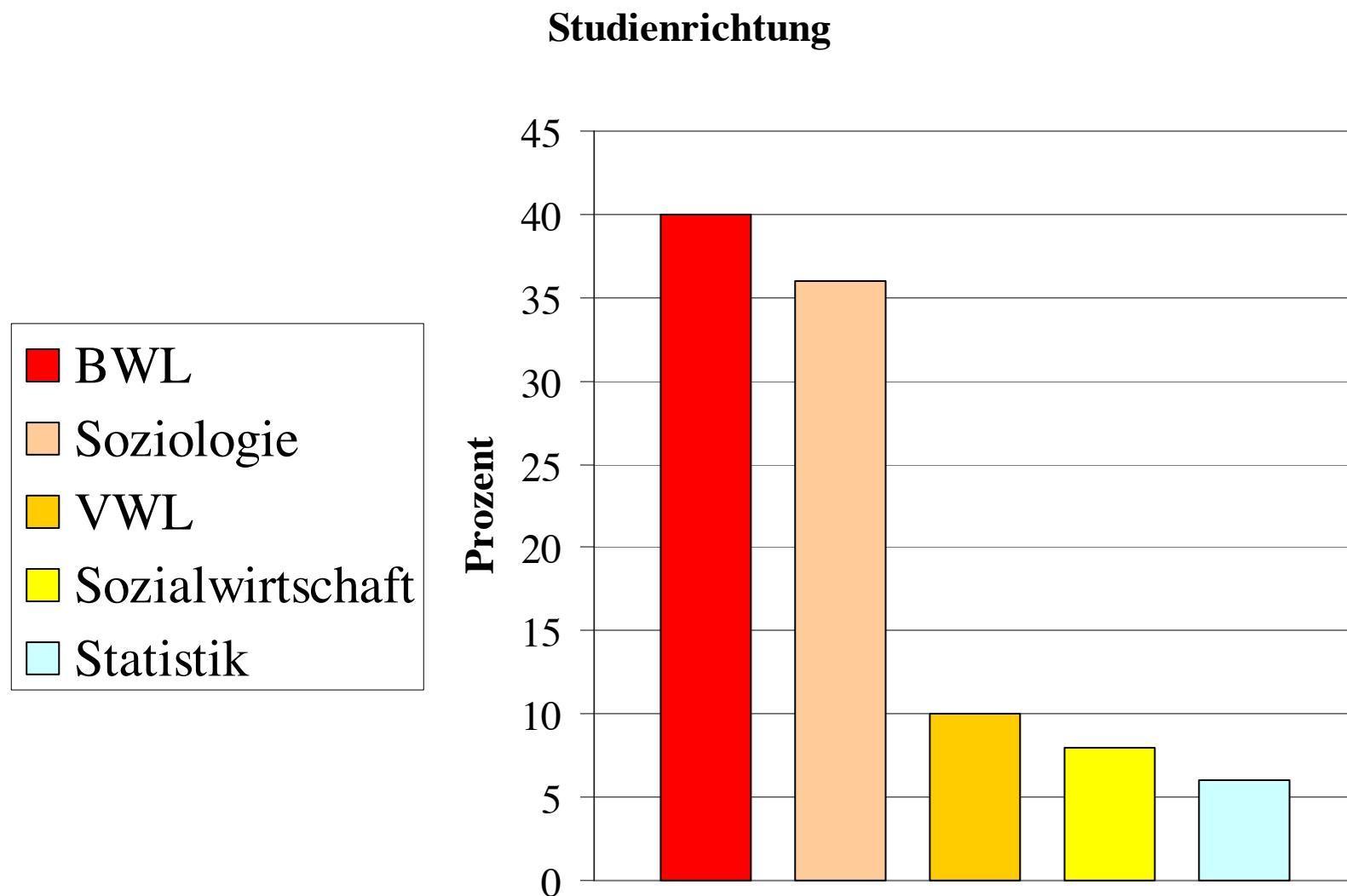
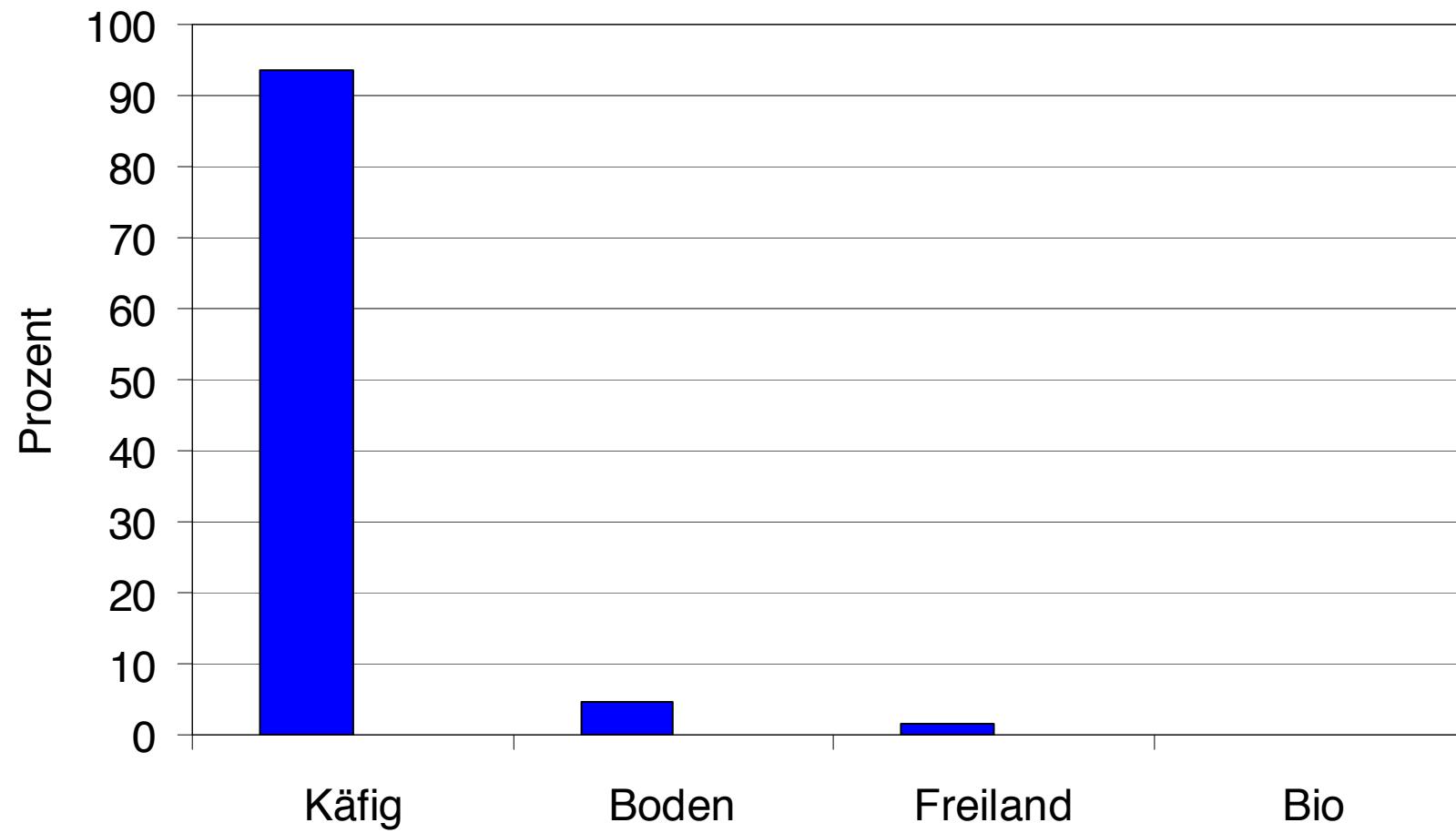


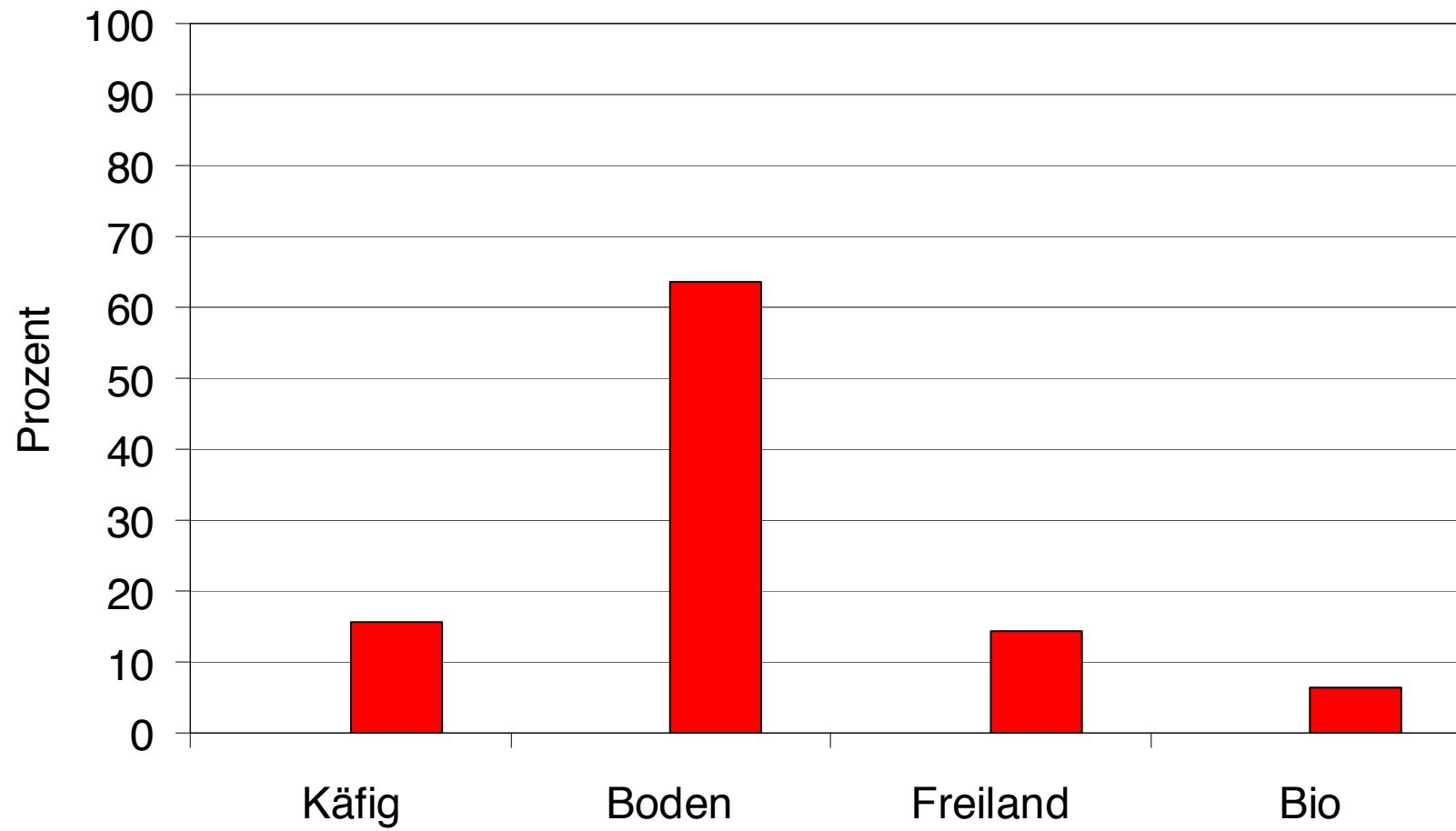
Abbildung 6: Säulendiagramm mit Legende



Hühnerhaltungsformen: 1995



Hühnerhaltungsformen: 2010



Hühnerhaltungsformen: Vergleich 1995 und 2010

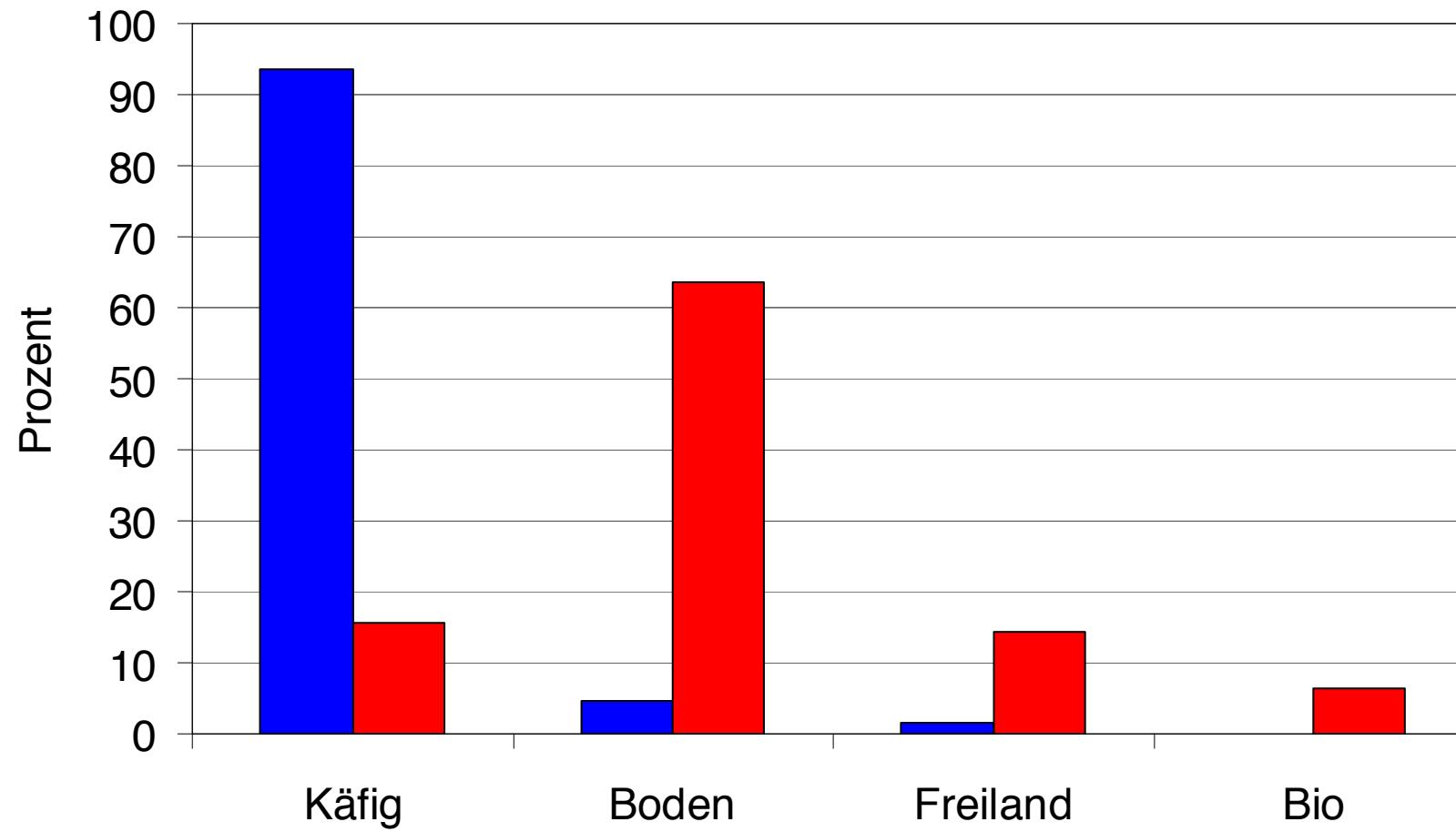


Abbildung 7: Ein Säulendiagramm einer Zeitreihe

Entwicklung der Jahresumsätze
in den Jahren 2004 - 2008

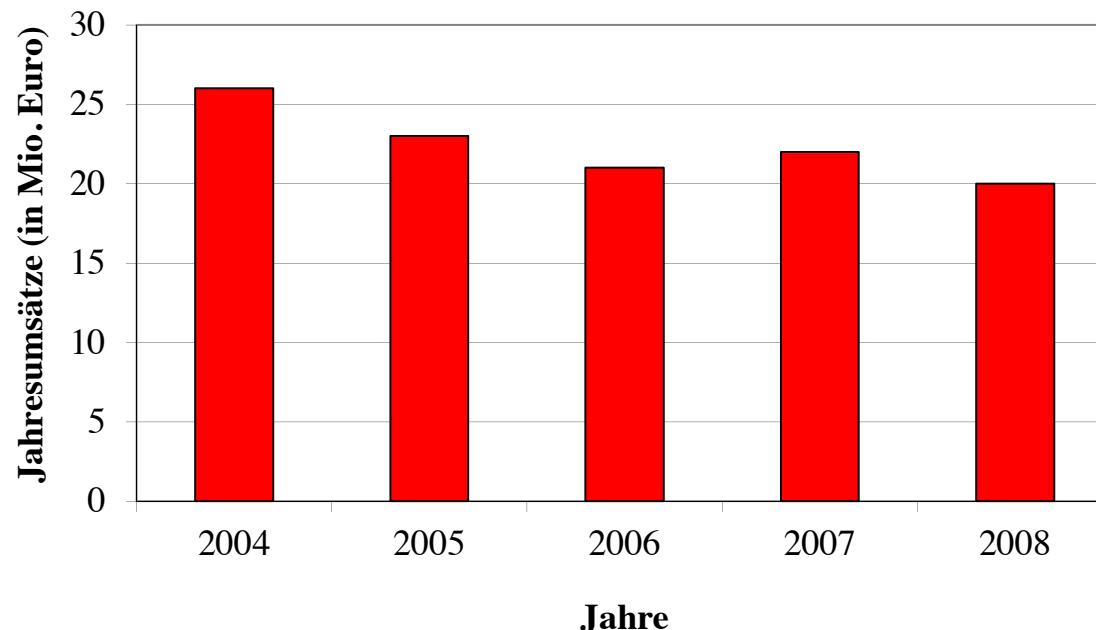
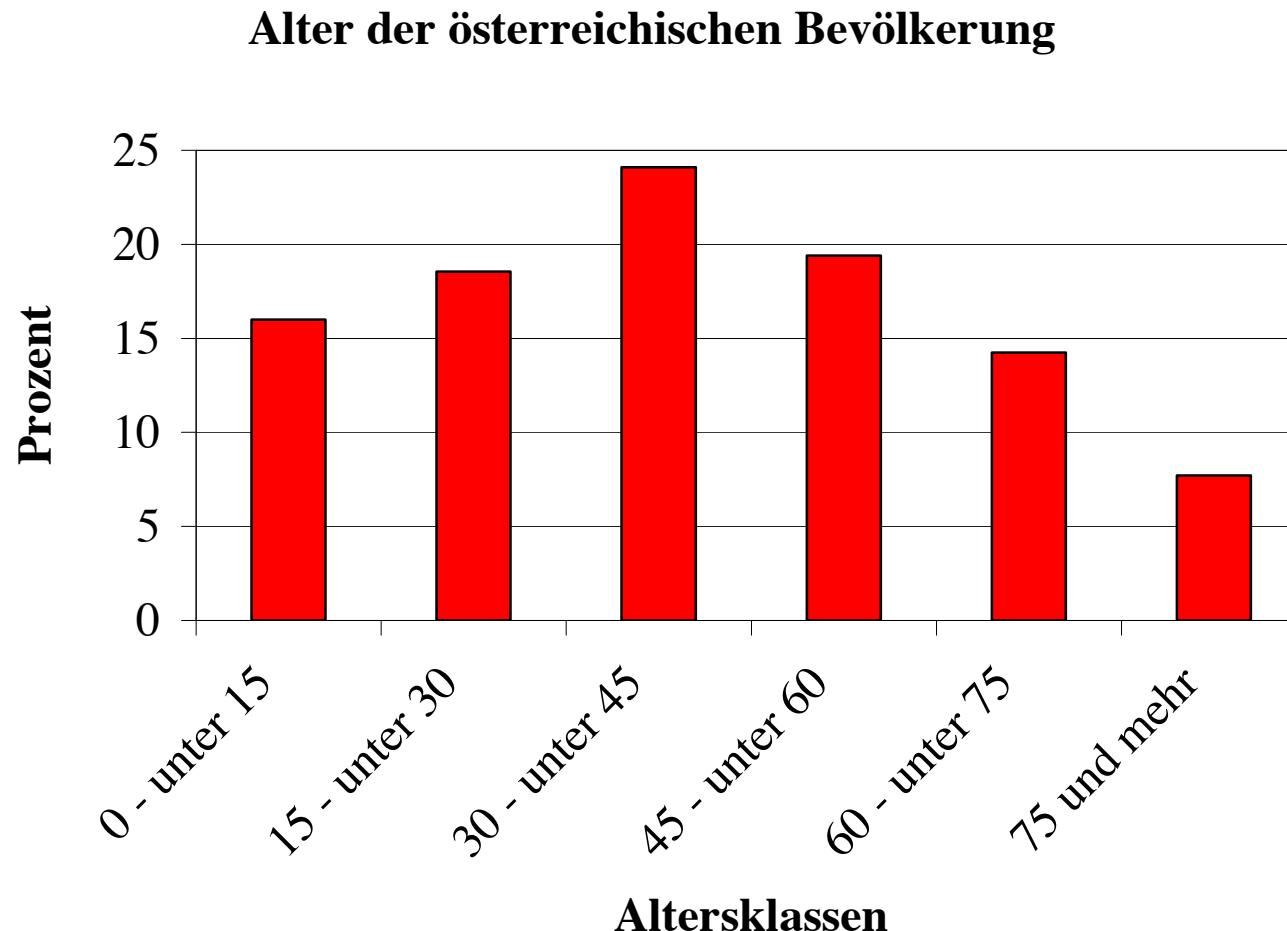
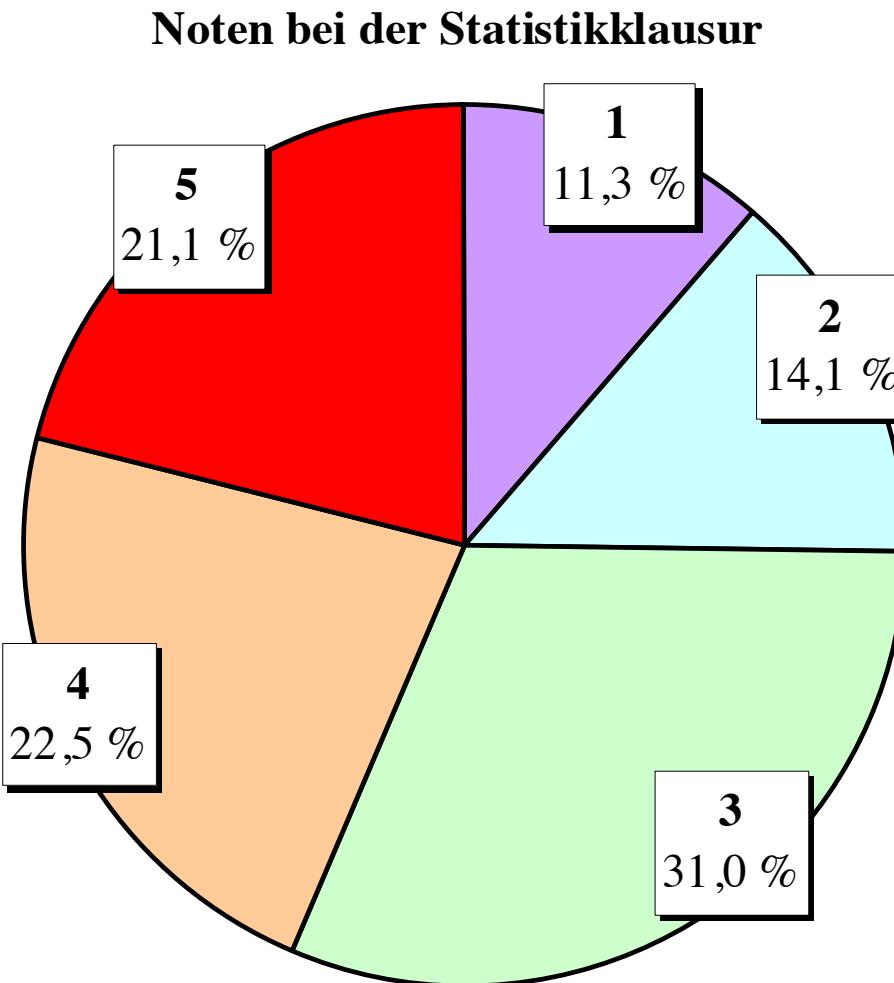


Abbildung 8: Säulendiagramm für ein in Intervalle zerlegtes Merkmal



Mitberücksichtigung verschiedener Intervallbreiten: Histogramm

Abbildung 9: Kreisdiagramm



Auch relative Summenhäufigkeiten sind im Kreisdiagramm ablesbar!

Regeln für die grafische Darstellung:

Säulendiagramme:

- Beschriftungen der x- und y-Achse sind unbedingt anzuführen
- Nullpunkt der Prozentzahlen auf der y-Achse sollte am Schnittpunkt zur x-Achse liegen

Säulen- und Kreisdiagramme:

- Überschriften und sind unbedingt anzuführen
- Ordnung innerhalb der Merkmalsausprägungen beibehalten
- 3-D-Darstellungen vermeiden
- Direkte Beschriftungen sind Legenden vorzuziehen



Die einfachste Grafik ist zumeist auch die Beste!

1.2.2 Gemeinsame Häufigkeitsverteilungen zweier Merkmale

Beispiel 6: Tabellarische Darstellung einer gemeinsamen Häufigkeitsverteilung zweier Merkmale

Häufigkeiten h:

		Studienrichtung					Summe
		BWL	Soz	VWL	SoWi	Stat	
Geschlecht	weiblich	//		/			
	männlich	/			/		
	Summe						

Häufigkeiten h:

		Studienrichtung					Summe
		BWL	Soz	VWL	SoWi	Stat	
Geschlecht	weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	200
	Summe	200	180	50	40	30	500

Relative Häufigkeiten p:

		Studienrichtung					Summe
		BWL	Soz	VWL	SoWi	Stat	
Geschlecht	weiblich	0,22	0,24	0,04	0,06	0,04	0,60
	männlich	0,18	0,12	0,06	0,02	0,02	0,40
	Summe	0,40	0,36	0,10	0,08	0,06	1

Vergleich: *Häufigkeitsverteilung der Studienrichtung unter den Frauen und unter den Männern*

Studienrichtung

		BWL	Soz	VWL	SoWi	Stat	Summe
Geschlecht	weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	200
	Summe	200	180	50	40	30	500

Zerlegung der Grundgesamtheit in die Teilgesamtheiten der F und M:

Beispiel 7: Tabellarische Darstellung einer **bedingten Verteilung**

		BWL	Soz	VWL	SoWi	Stat	Summe
Geschlecht	weiblich	0,367	0,400	0,067	0,100	0,067	1
	männlich	0,450	0,300	0,150	0,050	0,050	1

Beispiel 7: Tabellarische Darstellung einer **bedingten** Verteilung

		Studienrichtung					
		BWL	Soz	VWL	SoWi	Stat	Summe
Geschlecht	weiblich	0,367	0,400	0,067	0,100	0,067	1
	männlich	0,450	0,300	0,150	0,050	0,050	1

Unter den Frauen studieren 36,7 % BWL, 40,0 % Soz ...



Unter den Männern studieren ...



Korrekte Angabe der jeweiligen Grundgesamtheit, auf die sich die Prozentzahlen beziehen:

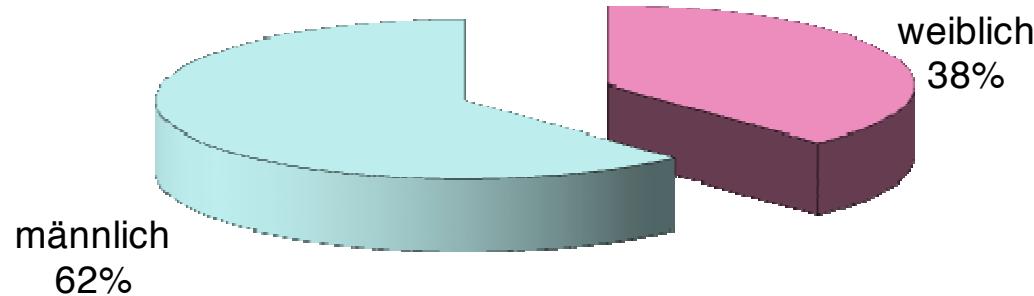
Richtig: „Zwei Drittel aller gefälschten Euro-Banknoten sind Fünfziger“

Falsch: „Zwei Drittel aller Fünfziger sind gefälschte Banknoten!“

Abbildung 10: Fehlinterpretation (DER STANDARD, 8.5. 1992)

*Unsinn in
den medien*

Durchgefallene nach Geschlecht



„Die Mädchen – oft zahlenmäßig überlegen – stellen nur etwas mehr als ein Drittel der *Sitzenbleiber*. Bei den Burschen dagegen erreichen 62% ihr Klassenziel nicht.“

1.3 Kennzahlen statistischer Verteilungen

Informationsbündelung auf einen einzigen Repräsentanten der Verteilung

1.3.1 Kennzahlen der Lage

Mittelwert (oder Durchschnitt):

Idee: Stellvertreter für alle Daten ist jener Wert, der sich bei gleichmäßiger Aufteilung der Summe aller aufgetretenen Daten (=Merkmalssumme) auf die Erhebungseinheiten ergeben würde

Beispiel:

Einkommen von fünf Personen: 1.000, 3.000, 4.000, 1.000 und 1.000 €

Merkmalssumme: $1.000 + 3.000 + 4.000 + 1.000 + 1.000 = 10.000 \text{ €}$

Gleichmäßige Aufteilung: $10.000 : 5 = 2.000 \text{ €}$

Formale Umsetzung der Idee des Mittelwerts:

Zeichen für den Mittelwert: \bar{x} (sprich: „x quer“)

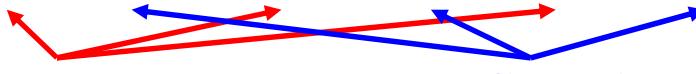
N ... Anzahl der Erhebungseinheiten

x_1 ... Merkmalsausprägung der 1. Erhebungseinheit,

x_2 ... Merkmalsausprägung der 2. Erhebungseinheit und so fort

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} \quad (2)$$

Merkmalssumme (2. Variante): $1.000 \cdot 3 + 3.000 \cdot 1 + 4.000 \cdot 1 = 10.000 \text{ €}$



Merkmalsausprägungen · Häufigkeiten

k ... Anzahl der verschiedenen Merkmalsausprägungen

h_1 ... Häufigkeit der 1. Merkmalsausprägung,

h_2 ... Häufigkeit der 2. Merkmalsausprägung und so fort

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot h_i}{N} \quad (2a)$$

Auch mit den relativen Häufigkeiten p :

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot h_i}{N} = \sum_{i=1}^k x_i \cdot \frac{h_i}{N} = \sum_{i=1}^k x_i \cdot p_i \quad (2b)$$

Beispiel 8: Berechnung des Mittelwerts (Fortsetzung von Beispiel 3)

Punktezahlen Häufigkeit Relative Häufigkeit

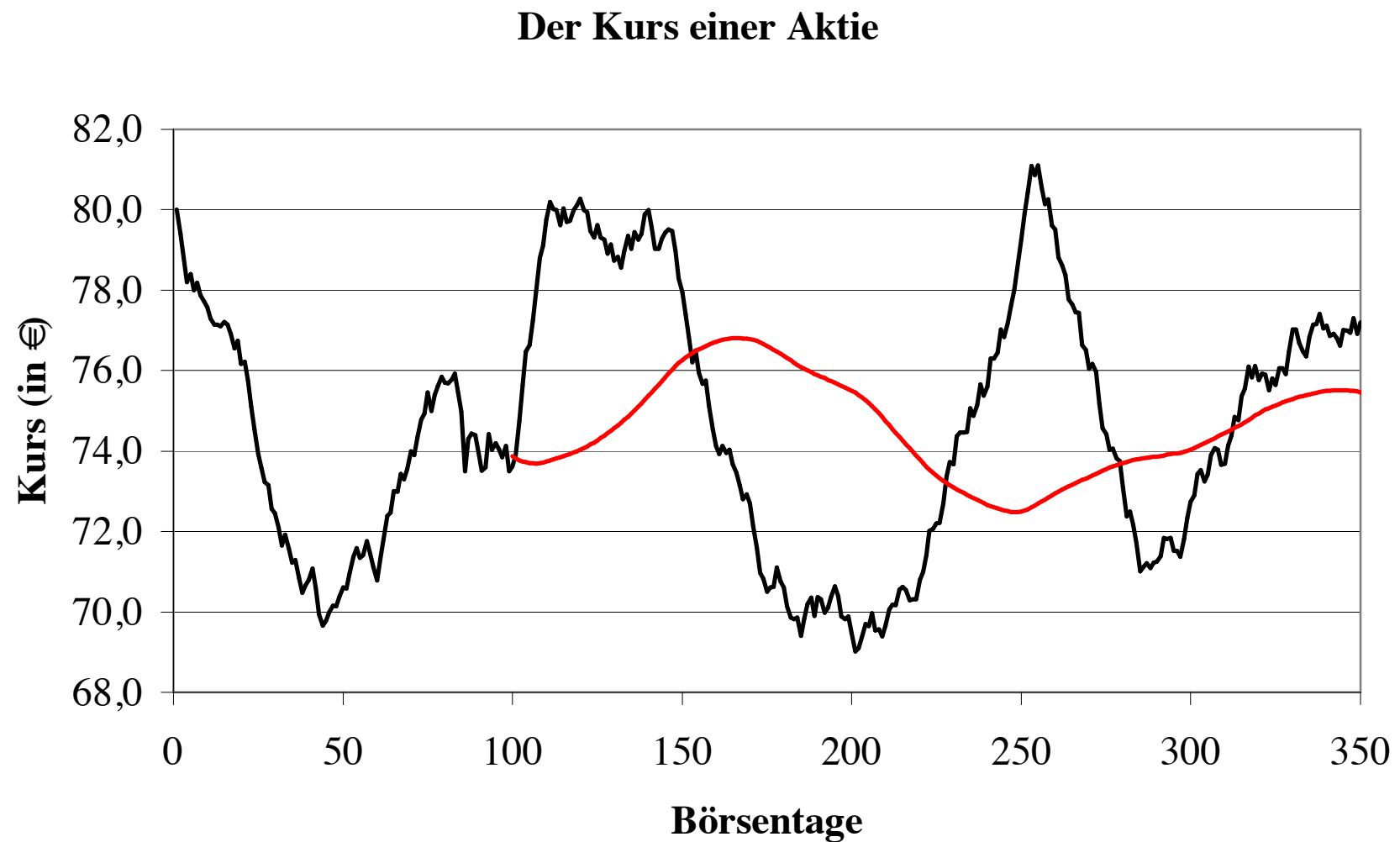
0	1	0,007
1	3	0,021
2	10	0,070
3	16	0,113
4	32	0,225
5	44	0,310
6	20	0,141
7	16	0,113

$$(2a): \bar{x} = \frac{\sum_{i=1}^k x_i \cdot h_i}{N} = \frac{(0 \cdot 1 + 1 \cdot 3 + \dots + 7 \cdot 16)}{142} = \frac{651}{142} = 4,58$$

oder mit (2b): $\bar{x} = \sum_{i=1}^k x_i \cdot p_i = 0 \cdot 0,007 + 1 \cdot 0,021 + \dots + 7 \cdot 0,113 = 4,58$

Beispiel für ein Anwendungsgebiet: [Zeitreihenanalyse](#)

Abbildung 11: Gleitende Mittelwerte in Zeitreihen



Der Mittelwert eignet sich nur für metrische Merkmale (und auch da nicht immer →)

Beispiel 9: Der Mittelwert von Wachstumsfaktoren

Vor drei Jahren: Umsatz von 20 Millionen €. In den drei Jahren seither jährliche Umsatzzuwächse von 10, 90 und 50 %.

Um wie viel Prozent ist der Umsatz pro Jahr durchschnittlich gestiegen?

Mittelwert: $(10+90+50):3 = 50\%$.

Verlauf des Umsatzes (in Mio. €):

1. Jahr: $20 \cdot 1,10 = 22 \text{ €}$ (1,10 ist der **Wachstumsfaktor**)

2. Jahr: $22 \cdot 1,90 = 41,8 \text{ €}$

3. Jahr: $41,8 \cdot 1,50 = 62,7 \text{ €}$

Mit dem Mittelwert der Prozentzahlen:

1. Jahr: $20 \cdot 1,5 = 30 \text{ €}$

2. Jahr: $30 \cdot 1,5 = 45 \text{ €}$

3. Jahr: $45 \cdot 1,5 = 67,5 \text{ €}$! ?

Welcher konstante Wachstumsfaktor würde also 62,7 ergeben?

$$20 \cdot g^3 = 62,7?$$

Aus $g^3 = \frac{62,7}{20} = 3,135$ folgt $g = \sqrt[3]{3,135} = \underline{\underline{1,464}}$

Auch aus Wachstumsfaktoren: $g = \sqrt[3]{1,1 \cdot 1,9 \cdot 1,5} = \sqrt[3]{3,135} = \underline{\underline{1,464}}$

... geometrischer Mittelwert der Wachstumsfaktoren

→ das durchschnittliche jährliche prozentuelle Wachstum ist 46,4 %.

Häufiges Anwendungsgebiet des geometrischen Mittelwerts: prozentuelles Wachstum von **Indizes** (z.B. Preisindex für die Lebenshaltung, Aktienindizes ...)

Gegenstand eines Indexes: Preisliche Entwicklung eines **Warenkorbs**

Inflationsrate: Quotient des aktuellen Werts des Preisindexes für die Lebenshaltung und des Werts vor genau einem Jahr

Probleme: Relevanz des Warenkorbs für den Einzelnen, Veralterung des Warenkorbs



Eine negative Eigenschaft von Mittelwert und geometrischem Mittelwert: Empfindlichkeit gegenüber untypischen Merkmalsausprägungen („**Ausreißern**“)

Möglicherweise besser: Berechnung des Mittelwertes ohne Ausreißer

Wichtig: Korrekte Beschreibung der Grundgesamtheit, auf die sich ein Mittelwert bezieht

Median \tilde{x} (sprich: „x Welle“) (auch: das 50%-Perzentil):

Idee: Als Stellvertreter für alle Daten gilt jener Wert, der – bei Sortierung der Daten aller N Erhebungseinheiten nach der Größe – in der Mitte steht.

Körpergrößen von fünf Erhebungseinheiten (*ungerade* Anzahl):

148, 158, 148, 160, 155

Sortierung: 148 148 **155** 158 160 → $\tilde{x} = 155$

Bei sechs Erhebungseinheiten (*gerade* Anzahl):

148, 158, 148, 160, 155, 157

Sortierung: 148 148 **155 157** 158 160 → $\tilde{x} = \frac{155 + 157}{2} = \underline{\underline{156}}$

Beispiel 10: Median eines diskreten Merkmals

Note	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenh.
1	16	0,113	11,3	0,113
2	20	0,141	14,1	0,254
3	44	0,310	31,0	0,564
4	32	0,225	22,5	0,789
5	30	0,211	21,1	1

142 Erhebungseinheiten → 71. und 72. stehen in der Mitte

Ergebnis: $\tilde{x} = 3$

Sortierbarkeit der Merkmalsausprägungen → nur bei metrischen und ordinalen Merkmalen

Eigenschaft des Medians: Unempfindlich (*robust*) gegen „Ausreißer“!

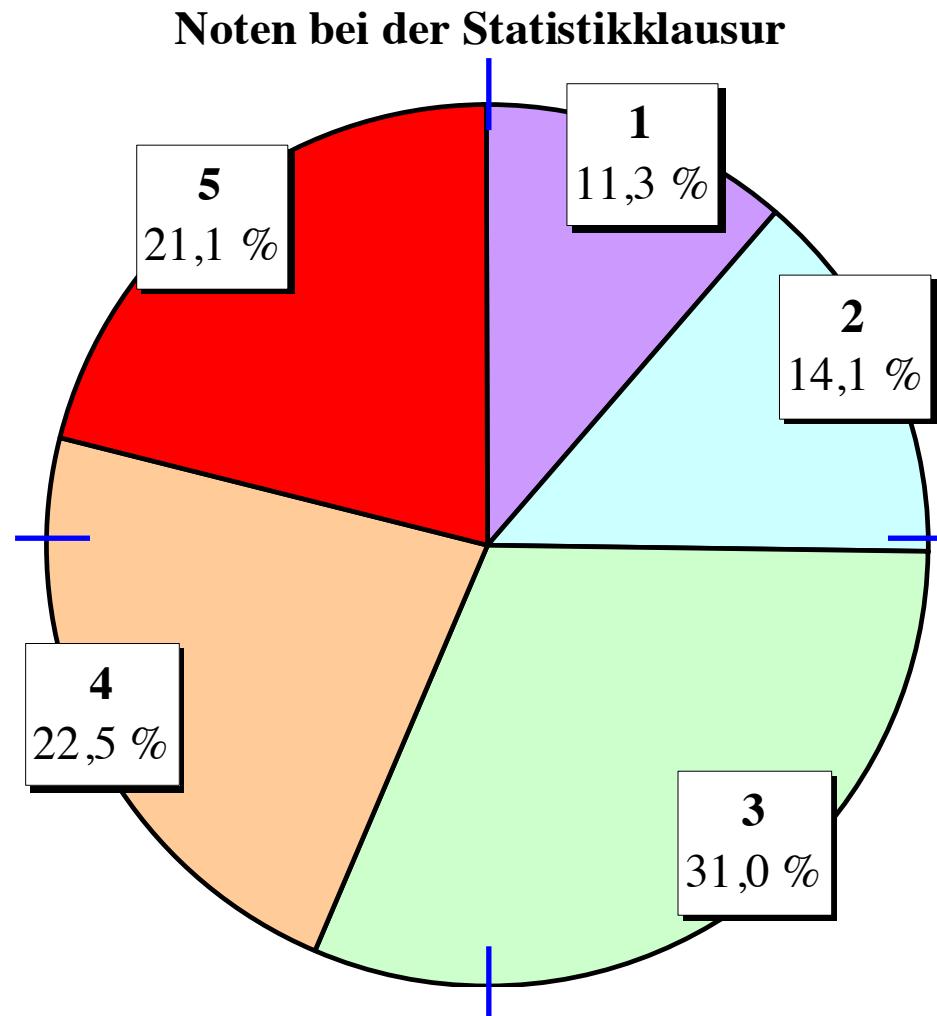
Bestimmung des Medians (50 Prozent-Perzentil) aus den relativen Summenhäufigkeiten

25 Prozent- (unteres Quartil), 75 Prozent-Perzentil (oberes Quartil)

Note	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenh.
1	16	0,113	11,3	0,113
2	20	0,141	14,1	0,254
3	44	0,310	31,0	0,564
4	32	0,225	22,5	0,789
5	30	0,211	21,1	1

Solche Perzentile lassen sich daher leicht an Kreisdiagrammen ablesen

Abbildung 12: Kreisdiagramm



Weitere grafische Veranschaulichung einer Häufigkeitsverteilung mit Hilfe der Perzentile: **Boxplot** (oder **Box-Whisker-Plot** oder **Kastendiagramm**)

Dazu werden benötigt:

Minimum

Unteres Quartil

Median

Oberes Quartil

Interquartilsdistanz: Abstand zwischen unterem und oberem Quartil

Maximum

Ausreißer

Beispiel 3 (Rückblick): Tabellarische Darstellung einer Häufigkeitsverteilung

Punktezahlen	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenhäufigkeit
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1

Abbildung 14: Ein Säulendiagramm

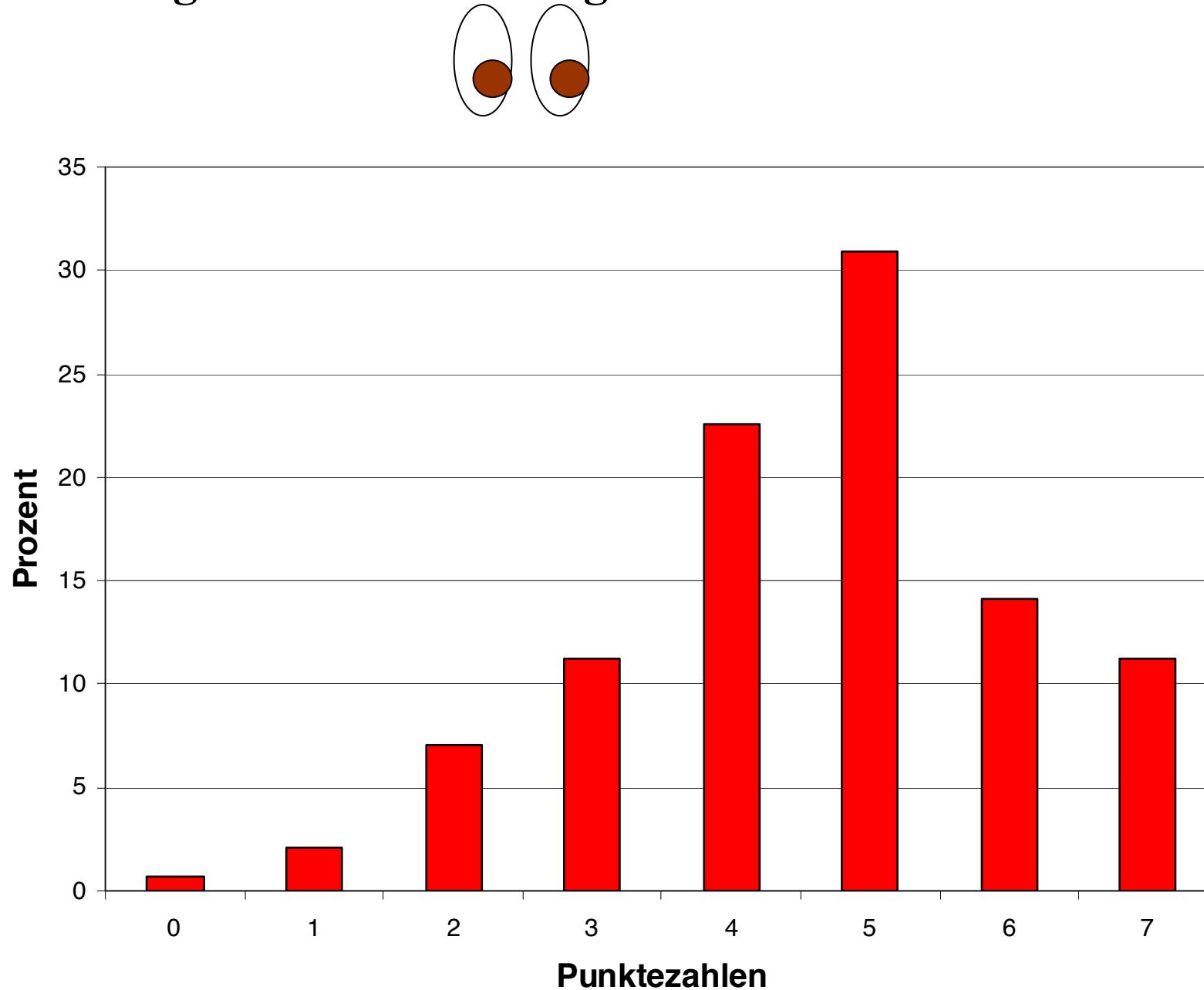
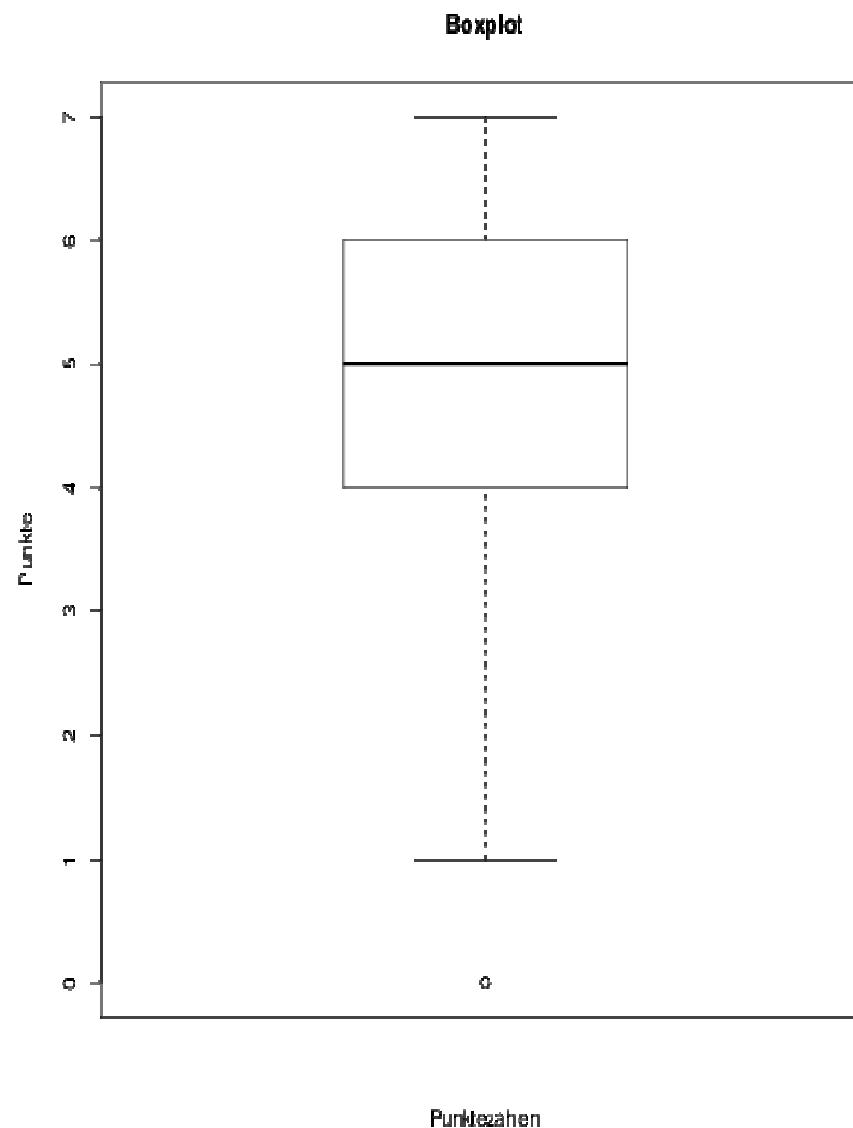
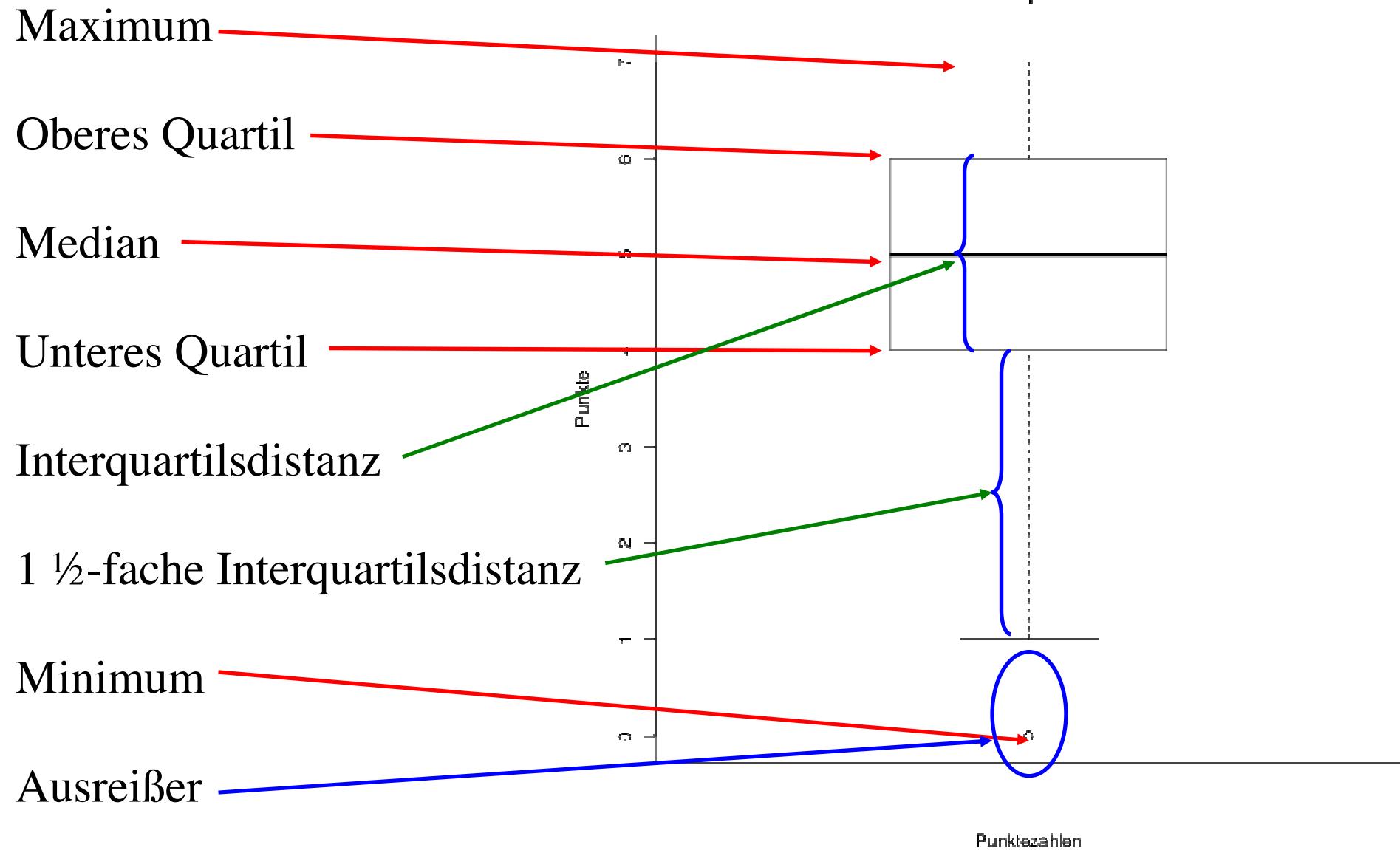


Abbildung 13: Ein Boxplot

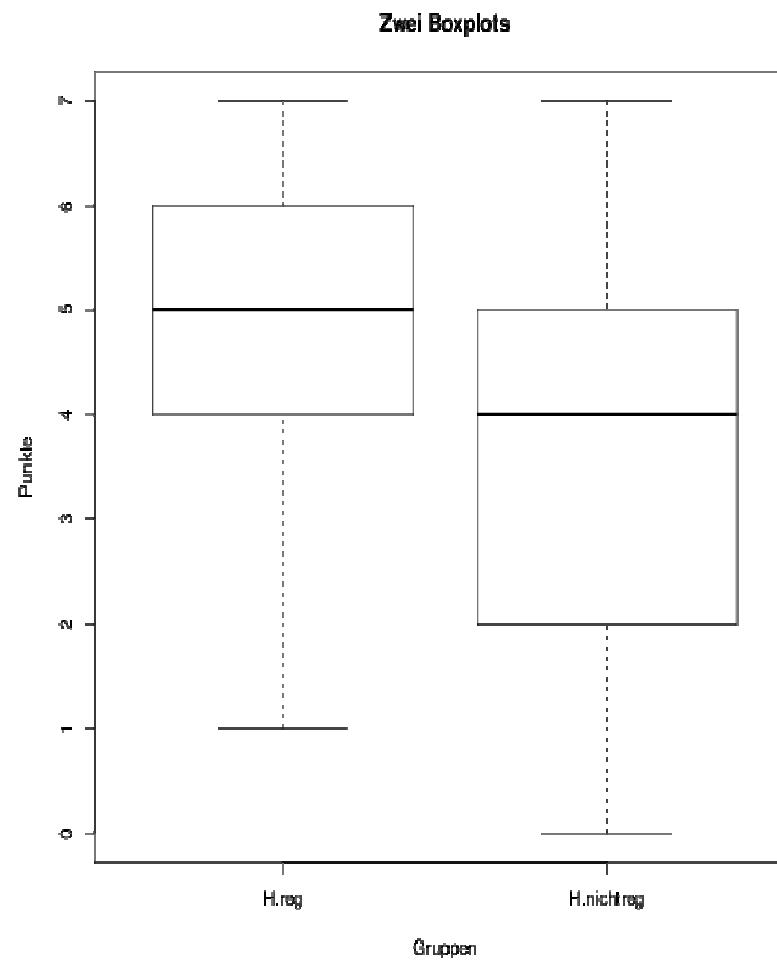


Boxplot



Boxplots eignen sich besonders zum grafischen Vergleich von bedingten Häufigkeitsverteilungen, da man sie nebeneinanderstellen kann.

Abbildung 15: Zwei Boxplots



Modus:

Idee: Stellvertreter ist jene Merkmalsausprägung, die am Häufigsten vor kommt

Abbildung 16: Der Modus einer Häufigkeitsverteilung



1.3.2. Kennzahlen der Streuung

Einkommen von fünf Personen: 1.000, 3.000, 4.000, 1.000 und 1.000 €

Und fünf andere Personen: 1.800, 2.200, 2.400, 1.800, 1.800

In beiden Gruppen: $\bar{x} = 2.000$

Unterschiedliche Streuung des Merkmals

Wie lässt sich das messen?

Beispiel: Alter der Personen in diesem Raum

Idee: Quadrierte Abweichungen der Merkmalsausprägungen aller Erhebungseinheiten vom Mittelwert (nur für metrische Merkmale) bestimmen und davon den Mittelwert berechnen

Beispiel:

Einkommen: 1.000, 3.000, 4.000, 1.000 und 1.000 €

Mittelwert $\bar{x} = 2.000 \text{ €}$

Quadrierte Abweichungen: $(1.000 - 2.000)^2, (3.000 - 2.000)^2, (4.000 - 2.000)^2,$
 $(1.000 - 2.000)^2$ und $(1.000 - 2.000)^2$

Davon den Mittelwert berechnen: 8 Millionen durch 5 = 1,6 Millionen

Man nennt diese Kennzahl: **Varianz**

Varianz s^2 (sprich: “s Quadrat”):

$$s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N} \quad (3)$$

Berechnung mit den Häufigkeiten:

$$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot h_i}{N} \quad (3a)$$

Berechnung mit den relativen Häufigkeiten:

$$s^2 = \sum_{i=1}^k (x_i - \bar{x})^2 \cdot p_i \quad (3b)$$

Parameter der Normalverteilung

Beispiel 11: Berechnung der Varianz

Punkte x_i	$(x_i - \bar{x})^2$	Häufigkeit h_i
0	$(0-4,58)^2 = 20,98$	1
1	$(1-4,58)^2 = 12,82$	3
2	$(2-4,58)^2 = 6,66$	10
3	$(3-4,58)^2 = 2,50$	16
4	$(4-4,58)^2 = 0,34$	32
5	$(5-4,58)^2 = 0,18$	44
6	$(6-4,58)^2 = 2,02$	20
7	$(7-4,58)^2 = 5,86$	16

Mit (3a):

$$s^2 = \frac{20,98 \cdot 1 + 12,82 \cdot 3 + \dots + 5,86 \cdot 16}{142} = 2,24 \text{ (Punkte-Quadrat)}$$

Die Standardabweichung s:

$$s = \sqrt{s^2} \quad (4)$$

in Beispiel 11: $s = \sqrt{2,243} = 1,498$ (Punkte)

Vergleich der Streuungen von verschiedenen Häufigkeitsverteilungen:

Der Variationskoeffizient v:

$$v = \frac{s}{x} \quad (5)$$

Weitere Charakteristika von Häufigkeitsverteilungen haben untergeordnete Bedeutung

1.3.3 Eine Kennzahl der Konzentration

- Einkommen von 5 Personen:

1.000, 3.000, 4.000, 1.000 und nochmals 1.000 €

$$s^2 = \frac{(1.000 - 2.000)^2 \cdot 3 + (3.000 - 2.000)^2 \cdot 1 + (4.000 - 2.000)^2 \cdot 1}{5} = 1,600.000$$

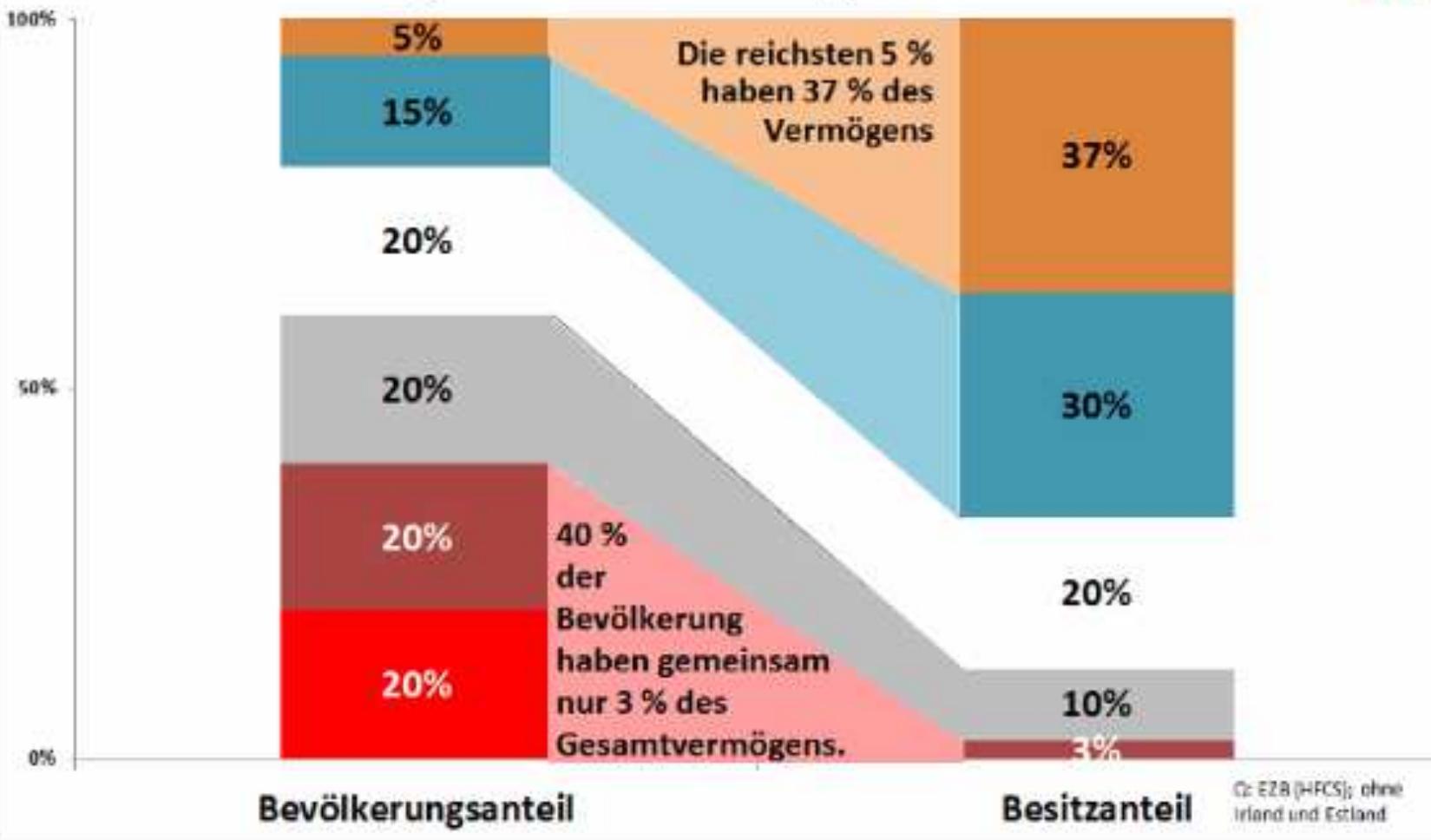
- Jede Person erhält nun 10.000 € als Prämie zusätzlich:

11.000, 13.000, 14.000, 11.000 und nochmals 11.000 €

Neuer Mittelwert: $\bar{x} = 12.000$; Varianz bleibt gleich.

Wie gleichmäßig konzentriert sich die Merkmalssumme auf die einzelnen Erhebungseinheiten? →

Verteilung Netto-Privat-Vermögen Euroraum 2010



5 Personen nach der Größe ihrer Merkmalsausprägungen sortieren

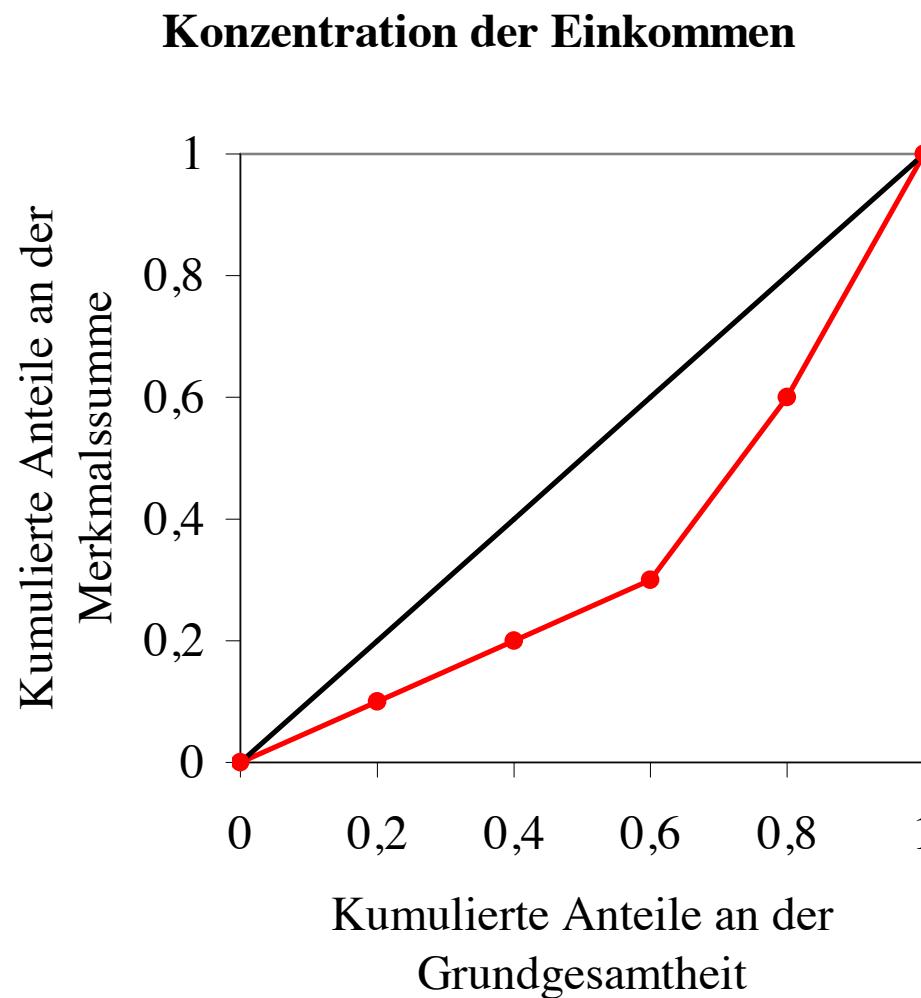
Beispiel 12: Messung der Konzentration einer Merkmalssumme auf die Erhebungseinheiten

Person	Anteile an der Grund- gesamtheit	Kumulierte Anteile an der Grund- gesamtheit	Einkom- men	Anteile am Gesamtein- kommen	Kumulierte Anteile am Gesamtein- kommen
A	0,2	0,2	1.000	0,1	0,1
D	0,2	0,4	1.000	0,1	0,2
E	0,2	0,6	1.000	0,1	0,3
B	0,2	0,8	3.000	0,3	0,6
C	0,2	1	4.000	0,4	1

Aussagen über die **Konzentration** einer Merkmalssumme auf die Erhebungseinheiten

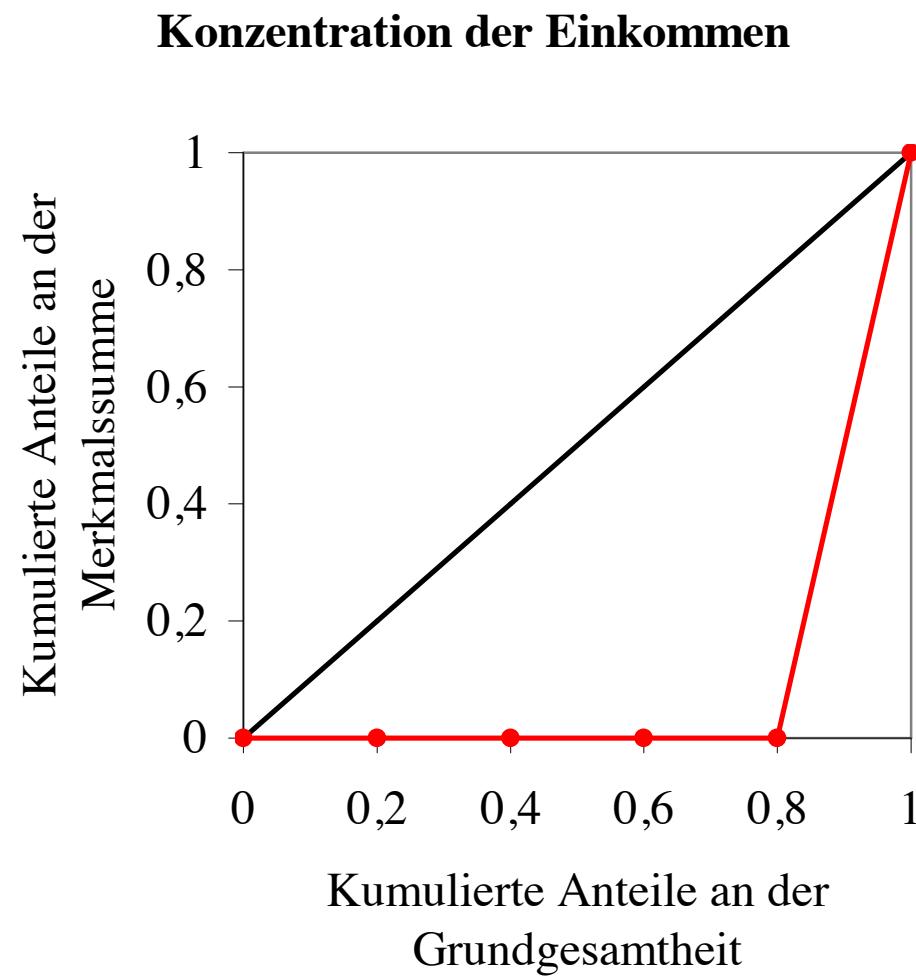
Grafische Veranschaulichung: **Lorenzkurve**

Abbildung 17: Die Lorenzkurve der Konzentration



Nullkonzentration – Maximalkonzentration

Abbildung 18: Maximalkonzentration



Fläche zwischen Lorenzkurve und Diagonale:

Bei Nullkonzentration ist diese 0, bei Maximalkonzentration ist sie:

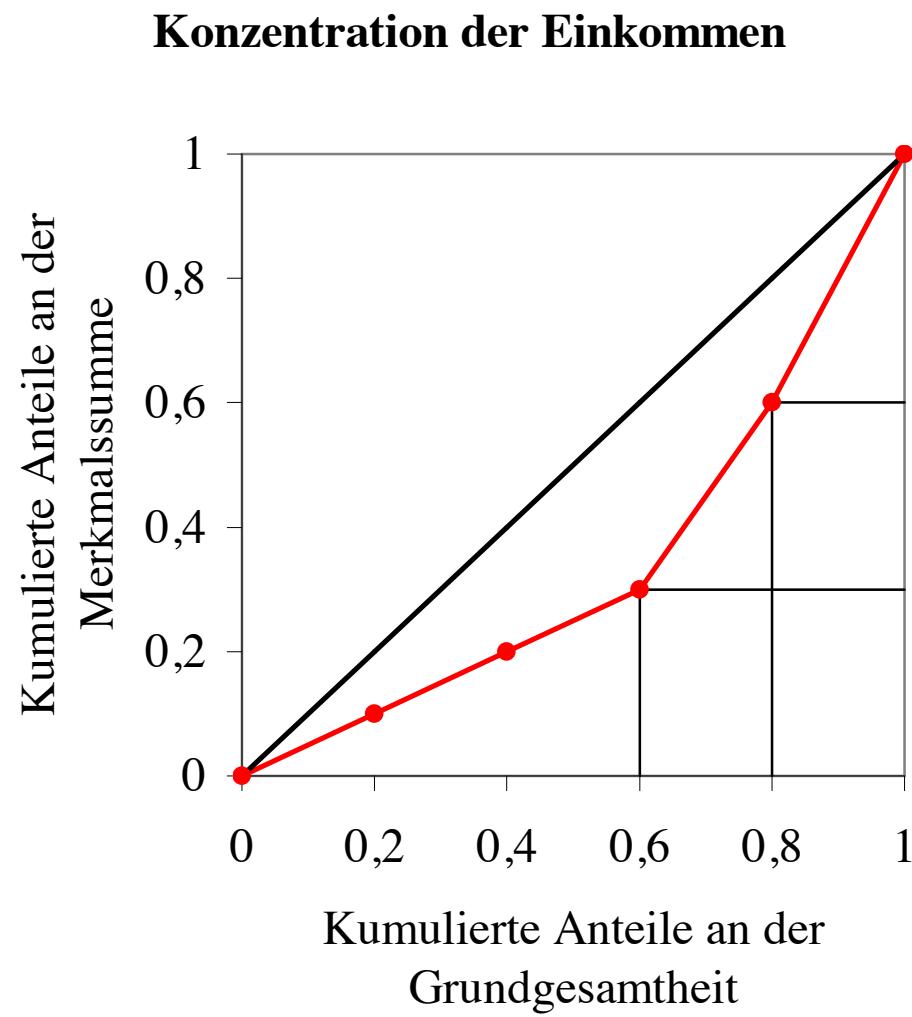
$$\frac{1}{2} - \frac{1}{2 \cdot N} = \frac{1}{2} \cdot \left(1 - \frac{1}{N} \right)$$

Normierter **Ginikoeffizient**:

Fläche zwischen der Lorenzkurve und der Diagonalen dividiert durch maximale Fläche zwischen der Lorenzkurve und Diagonale.

0 bei Nullkonzentration, 1 bei Maximalkonzentration

Abbildung 19: Die Berechnung des Ginikoeffizienten



Die Fläche zwischen Lorenzkurve und Diagonale ist 0,16

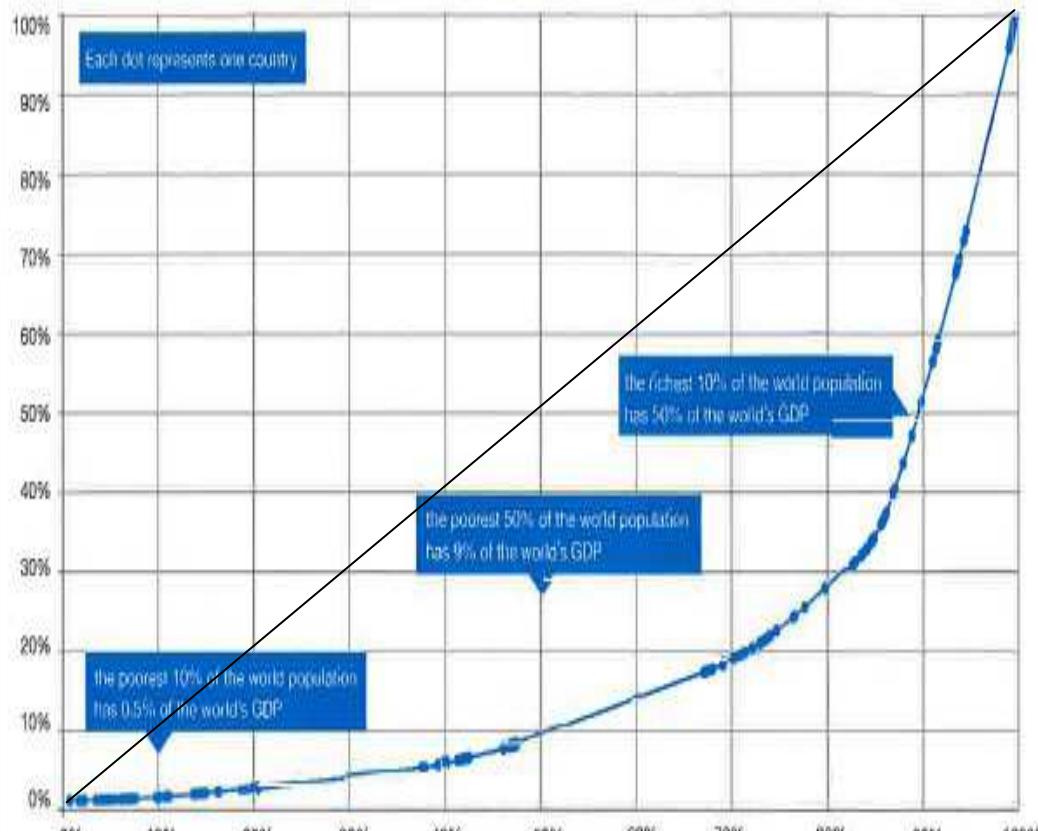
Die Fläche bei Maximalkonzentration ist $\frac{1}{2} \cdot \left(1 - \frac{1}{N}\right) = \frac{1}{2} \cdot \left(1 - \frac{1}{5}\right) = 0,4$

Der normierte Ginikoeffizient ist somit: $\frac{0,16}{0,4} = 0,4$

Die Fläche zwischen Lorenzkurve und Diagonale beträgt 40 % der Fläche, die bei Maximalkonzentration aufgetreten wäre.

» World income is still not equally distributed

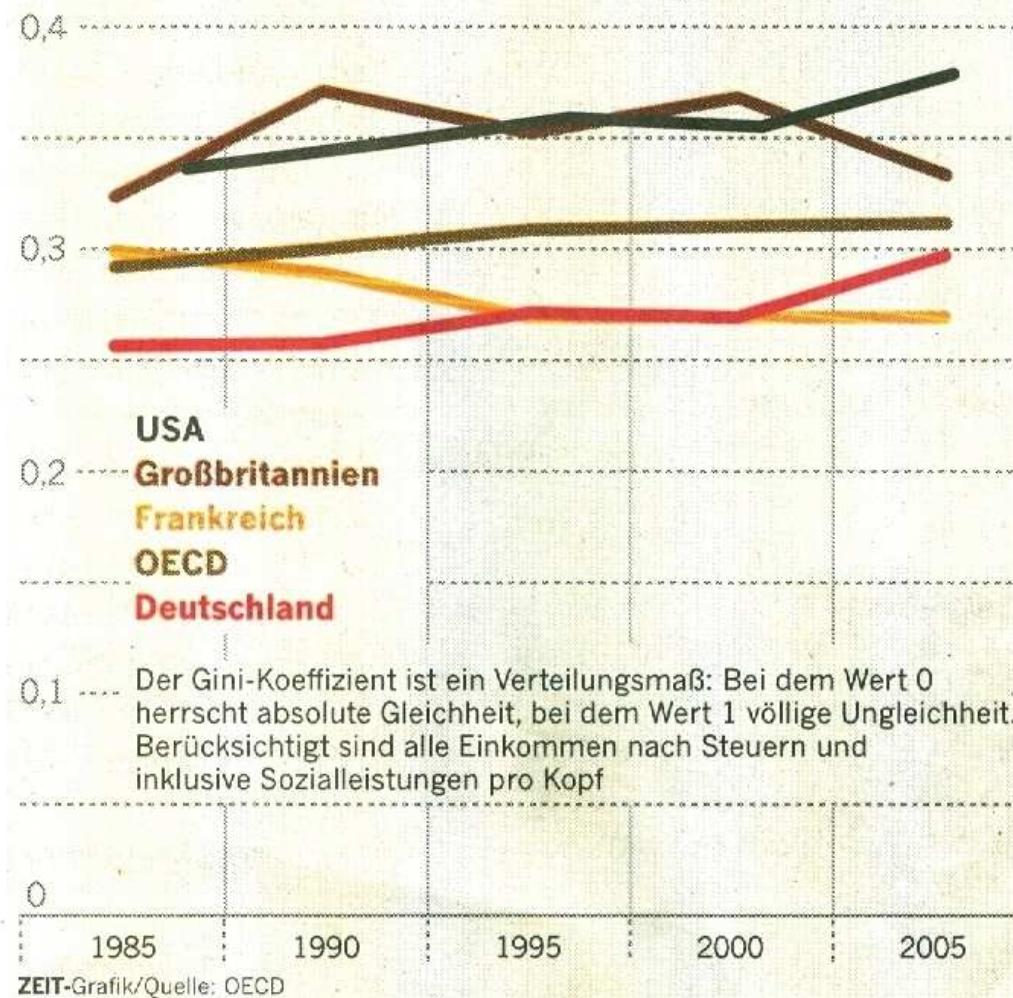
Cumulated GDP in function of cumulated population, 2010 or latest available year



Source: OECD national accounts, World Bank national accounts

Lange wuchs die Ungleichheit

In Deutschland sind die Einkommen aber noch immer gleichmäßiger verteilt als in vielen anderen Ländern



1.3.4 Kennzahlen des statistischen Zusammenhangs

Gemeinsame Häufigkeitsverteilung zweier Merkmale

Statistischer vs. kausaler Zusammenhang



Verschiedene Merkmalstypen Verschiedene Kennzahlen

Unterscheidung der Methoden: ... für zwei metrische, nominale, ordinale Merkmale

Bei zwei Merkmalen unterschiedlicher Merkmalstypen: Die Kennzahl des „niedrigeren“ Merkmalstyps wählen!

Hierarchie der Merkmale: metrisch – ordinal – nominal

metrisch – ordinal: wie 2 ordinarle Merkmale

metrisch – nominal: wie 2 nominale Merkmale

ordinal – nominal: wie 2 nominale Merkmale

Nominale Merkmale

Beispiel 13: Messung des Zusammenhangs von nominalen Merkmalen
Häufigkeiten h:

		Studienrichtung					Summe
		BWL	Soz	VWL	SoWi	Stat	
Geschlecht	weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	200
	Summe	200	180	50	40	30	500

Beobachtete relative Häufigkeiten p_{ij}^b :

		Studienrichtung					Summe
		BWL	Soz	VWL	SoWi	Stat	
Geschlecht	weiblich	0,22	0,24	0,04	0,06	0,04	0,60
	männlich	0,18	0,12	0,06	0,02	0,02	0,40
	Summe	0,40	0,36	0,10	0,08	0,06	1

Wie geht das bei Geschlecht und Einkommen zum Beispiel ?

Wenn *kein statistischer Zusammenhang* zwischen Geschlecht und Studienrichtung vorliegt → gleiche bedingte Häufigkeitsverteilungen der Studienrichtung unter den Frauen und unter den Männern

		BWL	Soz	VWL	SoWi	Stat	Summe
Geschlecht	weiblich	120	108	30	24	18	300
	männlich	80	72	20	16	12	200
	Summe	200	180	50	40	30	500
	Rel. Häufigk.	0,40	0,36	0,10	0,08	0,06	

... bei Fehlen eines Zusammenhangs **erwartete Häufigkeiten**

Bei Fehlen eines Zusammenhangs **erwartete relative Häufigkeiten p_{ij}^e** :

		BWL	Soz	VWL	SoWi	Stat	Summe
Geschlecht	weiblich	0,24	0,216	0,06	0,048	0,036	0,60
	männlich	0,15	0,144	0,04	0,032	0,024	0,40
	Summe	0,40	0,36	0,10	0,08	0,06	1

Idee: Verwendung der Differenzen der beobachteten und der bei Fehlen eines Zusammenhangs erwarteten (relativen) Häufigkeiten

Beobachtete relative Häufigkeiten p_{ij}^b :

		Studienrichtung					Summe
Geschlecht		BWL	Soz	VWL	SoWi	Stat	
	weiblich	0,22	0,24	0,04	0,06	0,04	0,60
	männlich	0,18	0,12	0,06	0,02	0,02	0,40
	Summe	0,40	0,36	0,10	0,08	0,06	1

Bei Fehlen eines Zusammenhangs erwartete relative Häufigkeiten p_{ij}^e :

		Studienrichtung					Summe
Geschlecht		BWL	Soz	VWL	SoWi	Stat	
	weiblich	0,24	0,216	0,06	0,048	0,036	0,60
	männlich	0,16	0,144	0,04	0,032	0,024	0,40
	Summe	0,40	0,36	0,10	0,08	0,06	1

Das Zusammenhangsmaß **Chiquadrat** χ^2 :

$$\chi^2 = N \cdot \sum \frac{(p_{ij}^b - p_{ij}^e)^2}{p_{ij}^e} \quad (6)$$

Wenn die Merkmale nicht statistisch zusammenhängen: $\chi^2 = 0$.

Beispiel 13:

$$\chi^2 = 500 \cdot \left[\frac{(0,22 - 0,24)^2}{0,24} + \frac{(0,24 - 0,216)^2}{0,216} + \dots \right] = 18,06$$

Normierung nötig!

Das **Cramersche Zusammenhangsmaß V** (= **Cramers V**) liegt zwischen 0 und 1:

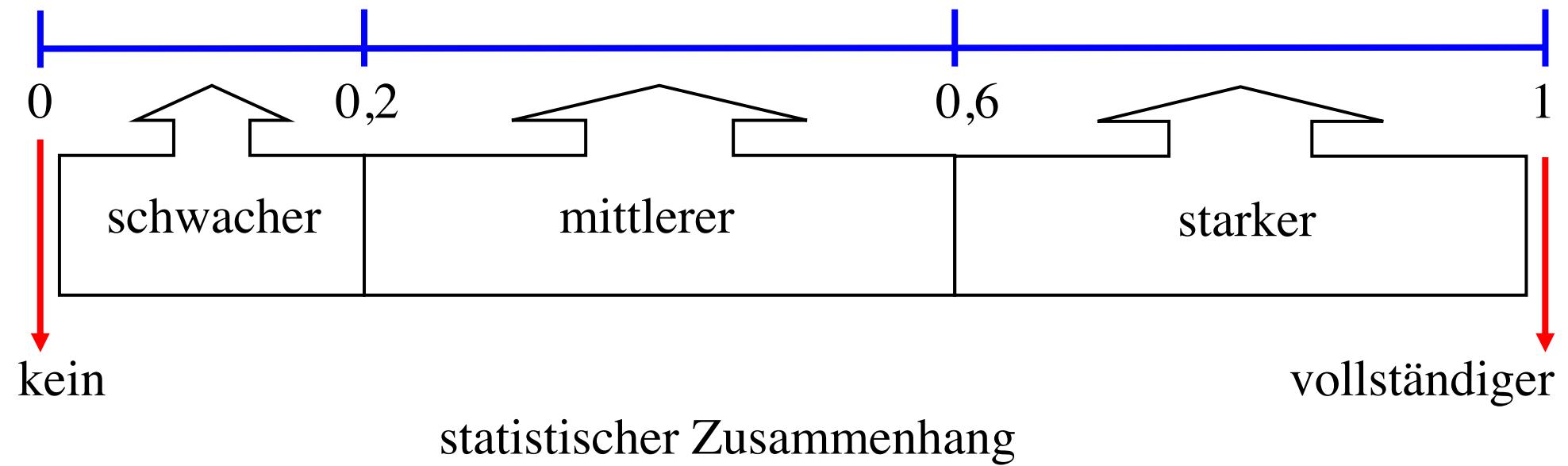
$$V = \sqrt[+]{\frac{\chi^2}{N \cdot (\min(s, t) - 1)}} \quad (7)$$

($s, t \dots$ die Anzahlen der Merkmalsausprägungen der beiden Merkmale;
 $\min(s, t) \dots$ die kleinere der beiden Anzahlen).

Beispiel 13:

$$V = \sqrt[+]{\frac{18,06}{500 \cdot (2 - 1)}} = 0,19$$

Abbildung 20: Die Interpretation von Cramers V (Faustregeln)



Metrische Merkmale

Beispiel 14: Erhebung von zwei metrischen Merkmalen

Person	A	B	C	D	E
Alter	21	46	55	35	28
Einkommen	1.850	2.500	2.560	2.230	1.800

Abbildung 21: Streudiagramm zweier metrischer Merkmale

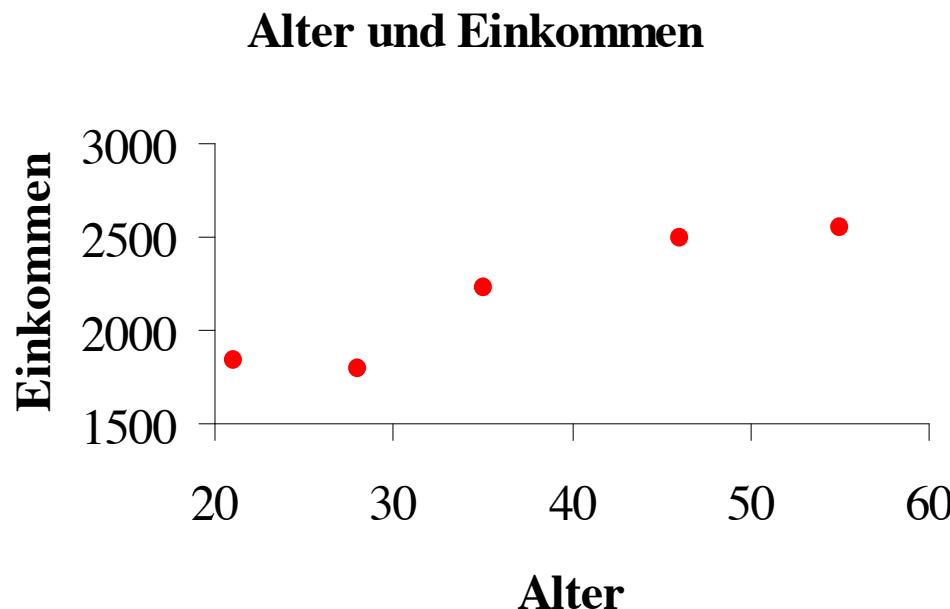
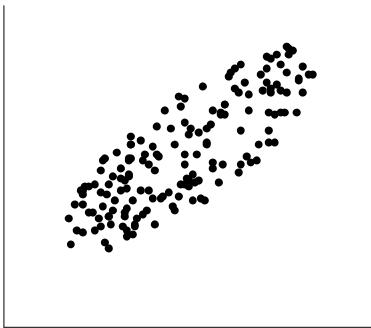
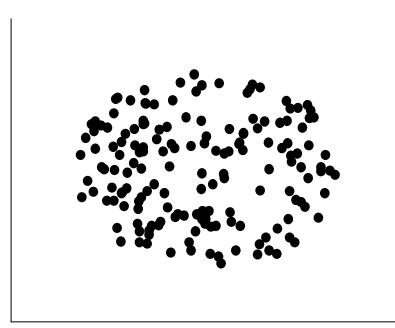


Abbildung 22: Drei Streudiagramme für beliebige Merkmale x und y
Richtung des statistischen Zusammenhangs



gleichsinnig

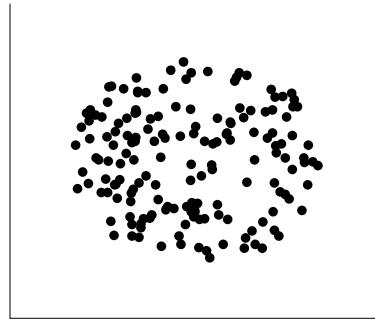


keine Richtung

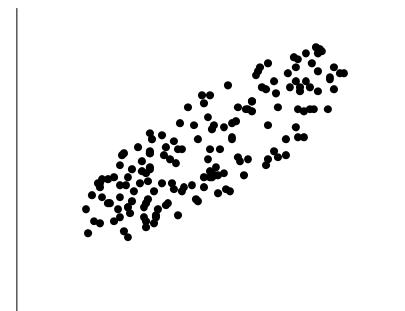


gegensinnig

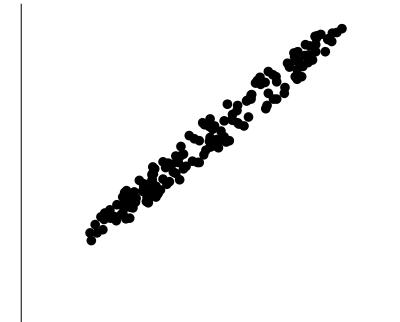
Abbildung 23: Drei Streudiagramme für beliebige Merkmale x und y
Stärke des statistischen Zusammenhangs



kein Zusammenh.



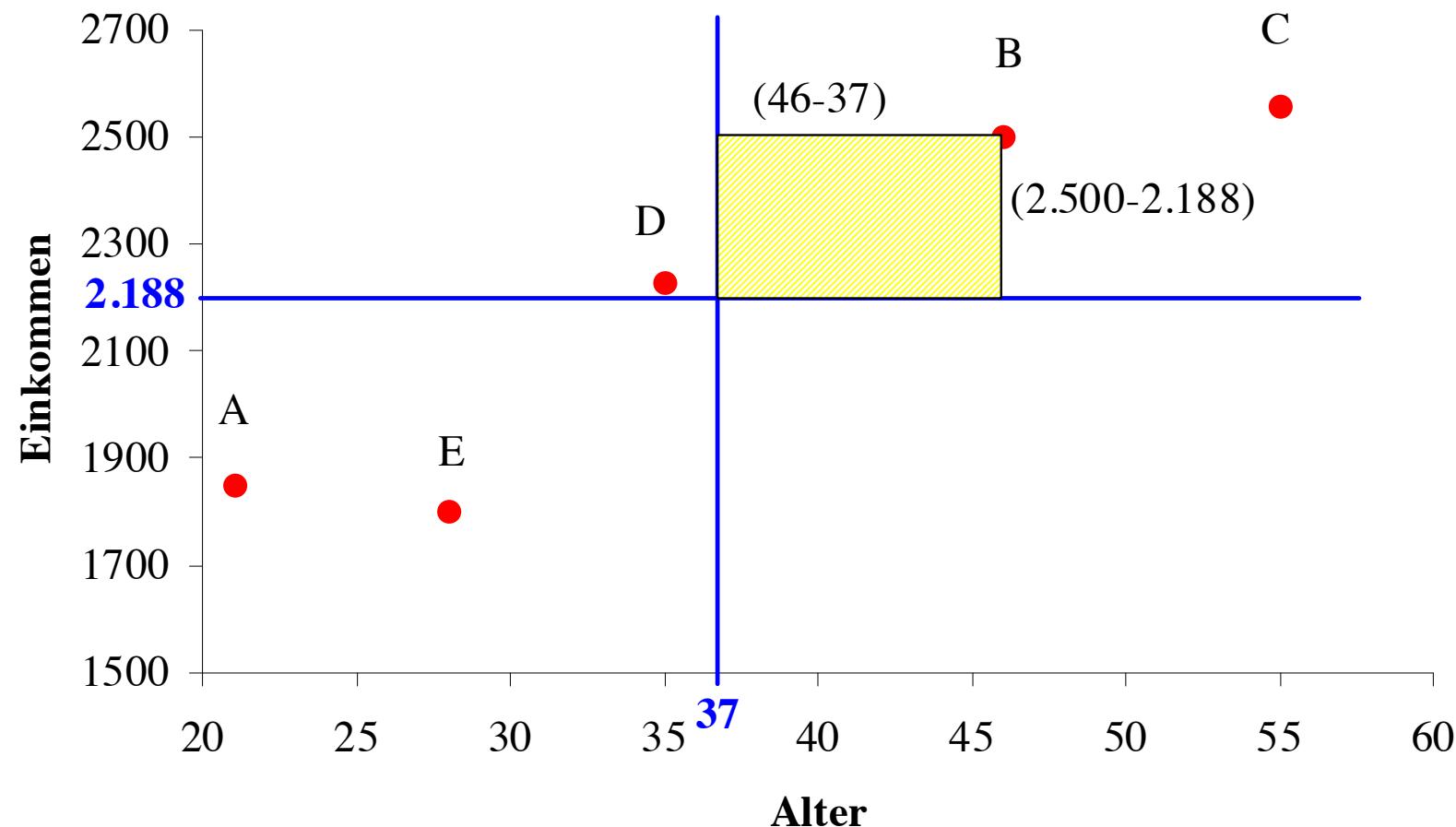
mittlerer Zusammenh.



starker Zusammenh.

Idee: Berechnung des folgenden Produktes für jede Erhebungseinheit:
 $(x_i - \bar{x}) \cdot (y_i - \bar{y})$

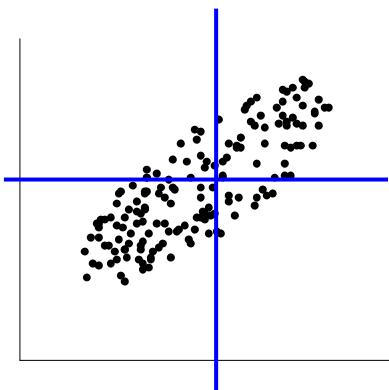
Abbildung 24: Grafische Darstellung der Idee



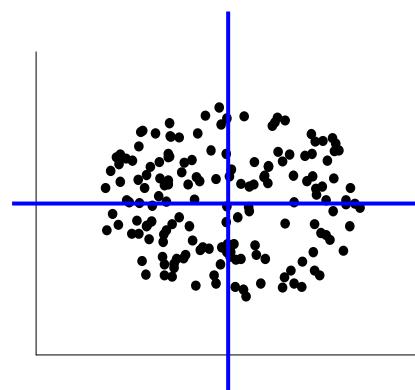
Die **Kovarianz** s_{xy} ist der Mittelwert der „gerichteten“ Rechtecksflächen:

$$s_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N} \quad (8)$$

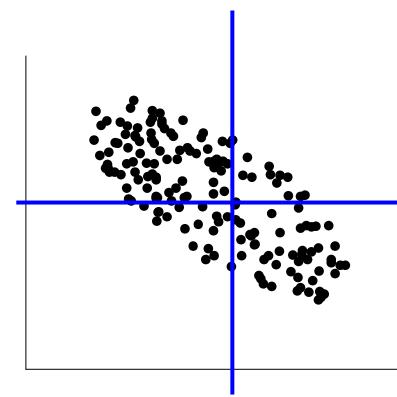
Vorzeichen der Kovarianzen:



gleichsinnig

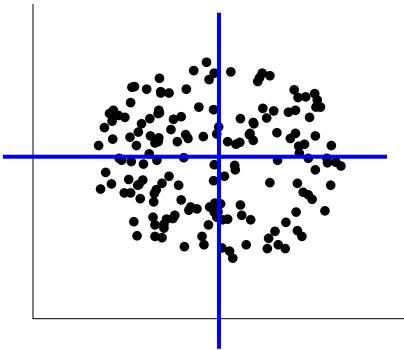


keine Richtung

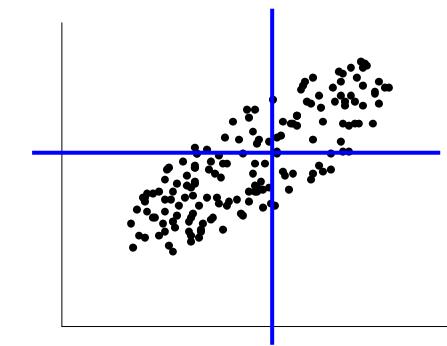


gegensinnig

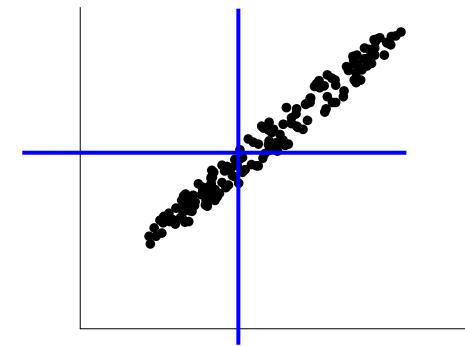
Ausmaß der Kovarianzen:



kein Zusammenh.



mittlerer Zusammenh.



starker Zusammenh.

In Beispiel 14:

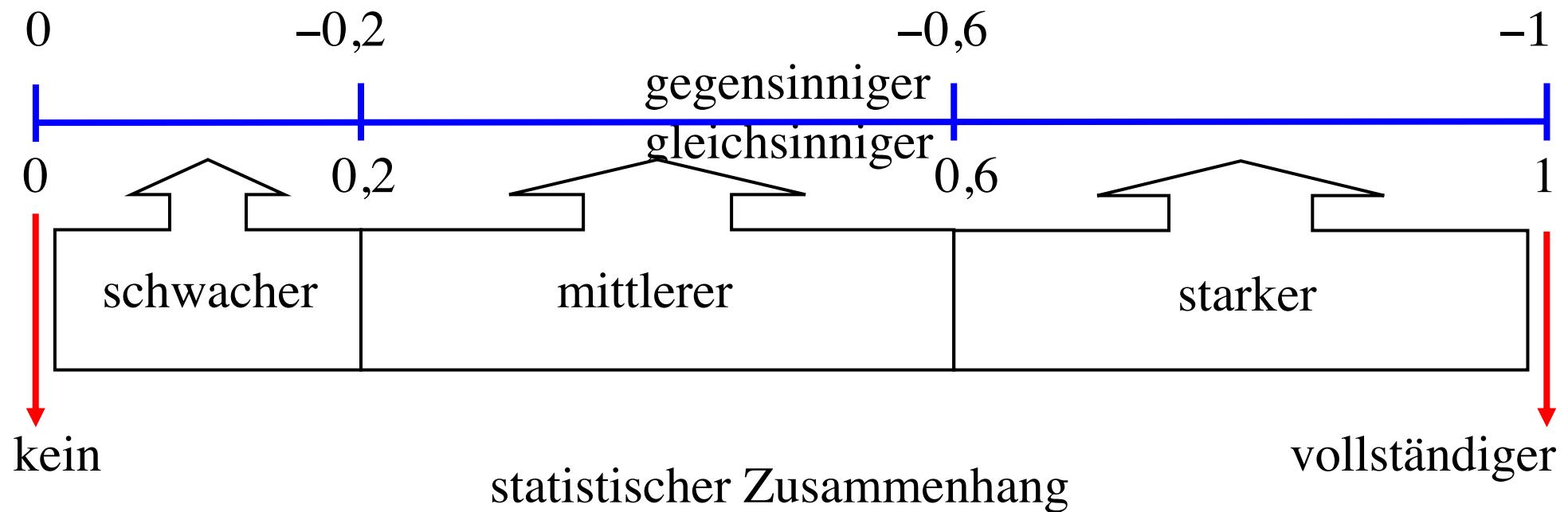
$$s_{xy} = \frac{(21 - 37) \cdot (1.850 - 2.188) + \dots + (28 - 37) \cdot (1.800 - 2.188)}{5} = 3.664$$

Normierung der Kovarianz nötig → **Korrelationskoeffizient r:**

$$r = \frac{s_{xy}}{s_x \cdot s_y} \quad (9)$$

Messung des *linearen* statistischen Zusammenhangs

Abbildung 25: Interpretation (Faustregeln)



in Beispiel 14:

$$s_{xy} = \frac{(21 - 37) \cdot (1.850 - 2.188) + \dots + (28 - 37) \cdot (1.800 - 2.188)}{5} = 3.664$$

$$r = \frac{3.664}{\sqrt{149,2} \cdot \sqrt{100.456}} = 0,946$$

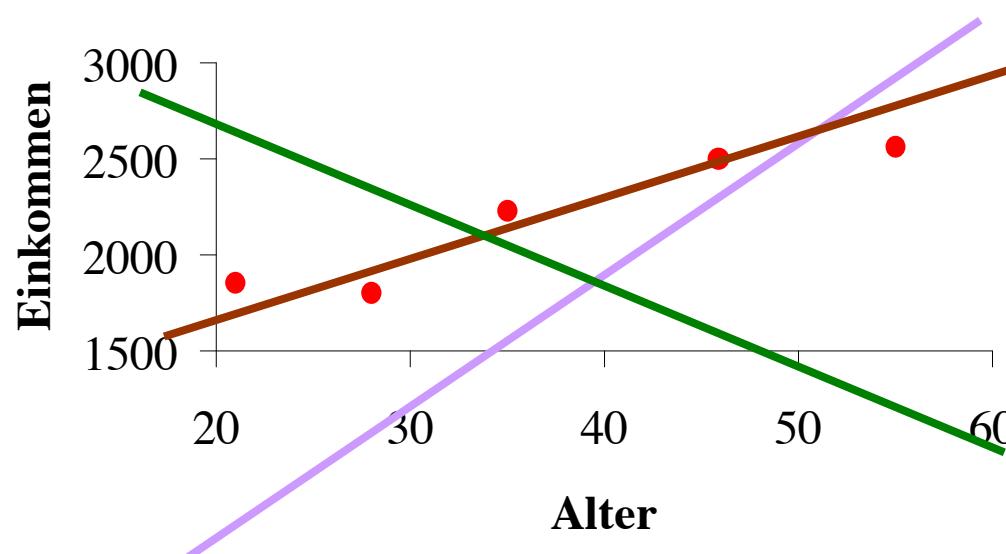
Regressionsrechnung

... beschäftigt sich mit dem *kausalen* Zusammenhang

Aus Werten der *Regressoren* Schätzung des Wertes des *Regressanden*
(Beispiel: Schätzung des Bremswegs)

Darstellung des linearen statistischen Zusammenhangs zweier metrischer Merkmale durch die **Regressionsgerade**

→ Die Gerade, die „am Nächsten zu den Punkten“ liegt.



Methode der kleinsten Quadrate (Extremwertaufgabe) →

Gleichung der Regressionsgeraden:

$$y = b_1 \cdot x + b_2 \quad (10)$$

mit den Regressionskoeffizienten

$$b_1 = \frac{s_{xy}}{s_x^2}$$

und

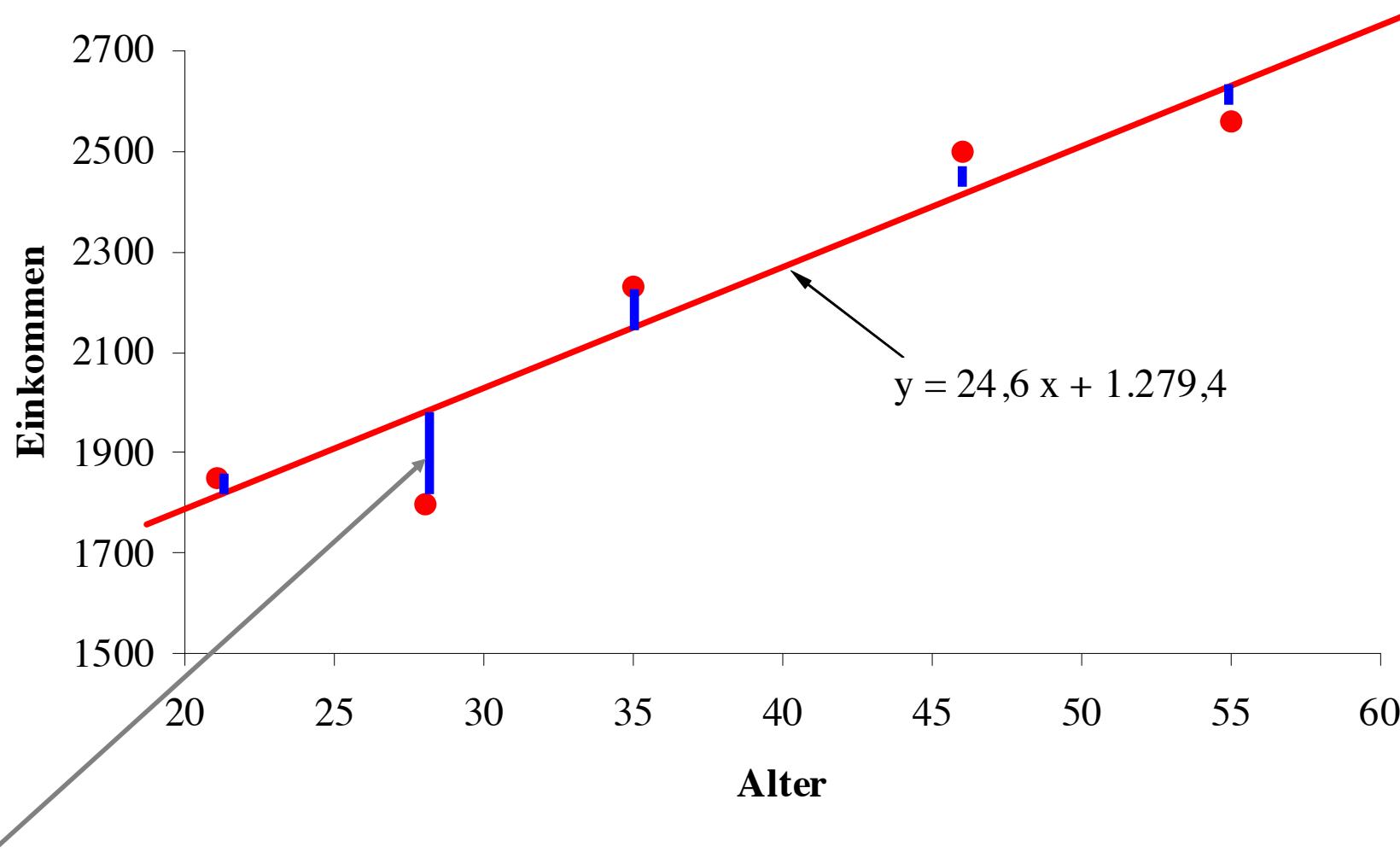
$$b_2 = \bar{y} - b_1 \cdot \bar{x}$$

Beispiel 15: Berechnung der Gleichung der Regressionsgeraden

$$b_1 = \frac{3.664}{149,2} = 24,6 \quad \text{und} \quad b_2 = 2.188 - 24,6 \cdot 37 = 1.279,4$$

Geradengleichung der Regressionsgeraden: $y = 24,6 \cdot x + 1.279,4$

Abbildung 26: Die Regressionsgerade



Residuen – beispielsweise bei Person A: $1.850 - 1.796$

Verwendung der Regressionsgeraden zur Schätzung fehlender Werte bzw. zur Prognose:

Beispielweise: $y = 24,6 \cdot x + 1.279,4$

Alter $x = 40 \rightarrow$ Einkommen $y = 24,6 \cdot 40 + 1.279,4 = 2.263,4 \text{ €}$

Vertrauen in die Schätzung:

Das **Bestimmtheitsmaß** B :

$$B = r^2 \quad (11)$$

In Beispiel 15: $B = r^2 = 0,947^2 = 0,897$

B gibt den Anteil der durch die Regression erklärten Varianz des Regressanden an („Erklärungsanteil“)

Faustregel: r sollte im Bereich des starken linearen Zusammenhangs liegen.

Ordinale Merkmale

... fast wie bei metrischen Merkmalen

Problematik:

Beispiel 16: Erhebung von zwei ordinalen Merkmalen

Studierender	A	B	C	D	E	F
Mathematiknote x	1	1	5	5	4	2
Statistiknote y	2	2	5	4	4	3

$$r = 0,96$$

Andere Kodierung der Noten: 1, 10, 100, 1.000 und 10.000

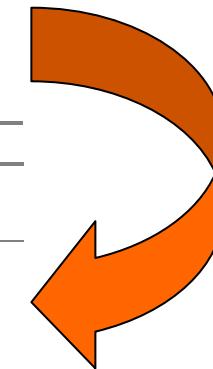
Studierender	A	B	C	D	E	F
Mathematiknote x	1	1	10.000	10.000	1.000	10
Statistiknote y	10	10	10.000	1.000	1.000	100

$$r = 0,68$$

→ Korrelationskoeffizient der Ränge !

Studierender	A	B	C	D	E	F
Mathematiknote x	1	1	5	5	4	2
Statistiknote y	2	2	5	4	4	3

Studierender	A	B	C	D	E	F
Mathematikrang u	1,5	1,5	5,5	5,5	4	3
Statistikrang v	1,5	1,5	6	4,5	4,5	3



Korrelationskoeffizient der Rangzahlen ist unabhängig von der gewählten Kodierung:

$$r = \frac{s_{uv}}{s_u \cdot s_v} = \frac{2,625}{\sqrt{2,75} \cdot \sqrt{2,75}} = 0,955$$

... Spearmanscher Korrelationskoeffizient der Rangzahlen

Interpretation wie beim Korrelationskoeffizienten

Kapitel 2: Wahrscheinlichkeitsrechnung

2.1 Grundbegriffe

Verbindung von **beschreibender** und **schließender Statistik** durch die
Wahrscheinlichkeitsrechnung: Zuordnung von Zahlen zu bestimmten Ereignissen (Auskunft über die Wahrscheinlichkeit ihres Eintreffens)



Zufallsexperiment (z.B. das Werfen eines Würfels)

Merkmal (*Zufallsvariable*; z.B. die Augenzahl)

Merkmalsausprägungen (*Elementarereignisse*; z.B. die ganzen Zahlen von 1 bis 6)

Ereignis (Teilmenge des Wertebereichs; z.B. die geraden Zahlen).

Unmögliche und sichere Ereignisse

Jedem Ereignis (E) wird die Wahrscheinlichkeit seines Eintreffens $\text{Pr}(E)$ zugeordnet.

Es gilt:

- $0 \leq \text{Pr}(E) \leq 1$
- $\text{Pr}(\text{unmögliches Ereignis}) = 0$
- $\text{Pr}(\text{sicheres Ereignis}) = 1$

Beispiel 17: Rechnen mit Wahrscheinlichkeiten



Anzahl pro Serie	„Gewinn“ in Euro
10 x	Eine Schatztruhe voller Gold
13 x	30.000
20 x	3.000
60 x	1.000
130 x	300
3.000 x	100
7.000 x	60
20.000 x	30
70.000 x	9
290.000 x	6
642.000 x	3
1.140.000 x	1,50

Alle Fälle sind gleich wahrscheinlich →

Abzählregel: „günstige durch mögliche“

Beispiel 17: Rechnen mit Wahrscheinlichkeiten

x ... Auszahlungsbetrag

Anzahl pro Serie	„Gewinn“ in Euro
10 x	Eine Schatztruhe voller Gold
13 x	30.000
20 x	3.000
60 x	1.000
130 x	300
3.000 x	100
7.000 x	60
20.000 x	30
70.000 x	9
290.000 x	6
642.000 x	3
1.140.000 x	1,50

$$\Pr(x = \text{„Schatztruhe“}) = \frac{10}{10.000.000} = \underline{\underline{0,000001}}$$

„In durchschnittlich nur einem von 1 Million Fällen wird das Ereignis eintreffen!“

$$\Pr(x = 1,50) = \frac{1.140.000}{10.000.000} = \underline{\underline{0,114}}$$

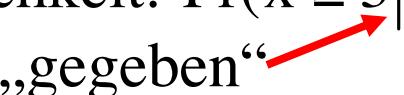
„In durchschnittlich elf von 100 Fällen ...“

$$\Pr(x \geq 3) = \frac{642.000 + 290.000 + \dots + 13 + 10}{10.000.000} = \frac{1.032.233}{10.000.000} = \underline{\underline{0,103}}$$

$$1 - 0,103 = \underline{\underline{0,897}} \dots \text{Gegenwahrscheinlichkeit } \rightarrow \Pr(x < 3)$$

Anzahl pro Serie	„Gewinn“ in Euro
10 x	Eine Schatztruhe voller Gold
13 x	30.000
20 x	3.000
60 x	1.000
130 x	300
3.000 x	100
7.000 x	60
20.000 x	30
70.000 x	9
290.000 x	6
642.000 x	3
1.140.000 x	1,50

Bedingte Wahrscheinlichkeit: $\Pr(x \geq 3 | x \neq 0) = \frac{1.032.233}{2.172.233} = \underline{\underline{0,475}}$

„gegeben“ 

Lotto 6 aus 49: Alle möglichen Zahlenkombinationen sind gleich wahrscheinlich → Abzählregel: „günstige durch mögliche“

mögliche: Anzahl der verschiedenen 6er Gruppen bei 49 Kugeln

$$\binom{49}{6} = \frac{49!}{6!(49-6)!} = 13.983.816 \text{ Möglichkeiten}$$



günstige: Anzahl der ausgefüllten Tippkolonnen

x ... Anzahl der “Richtigen” bei *einem* ausgefüllten Tipp

$$\Pr(x = 6) = \frac{1}{13.983.816} = 0,000000072$$

2.2 Wahrscheinlichkeitsverteilungen

Wahrscheinlichkeitsverteilung - Häufigkeitsverteilung

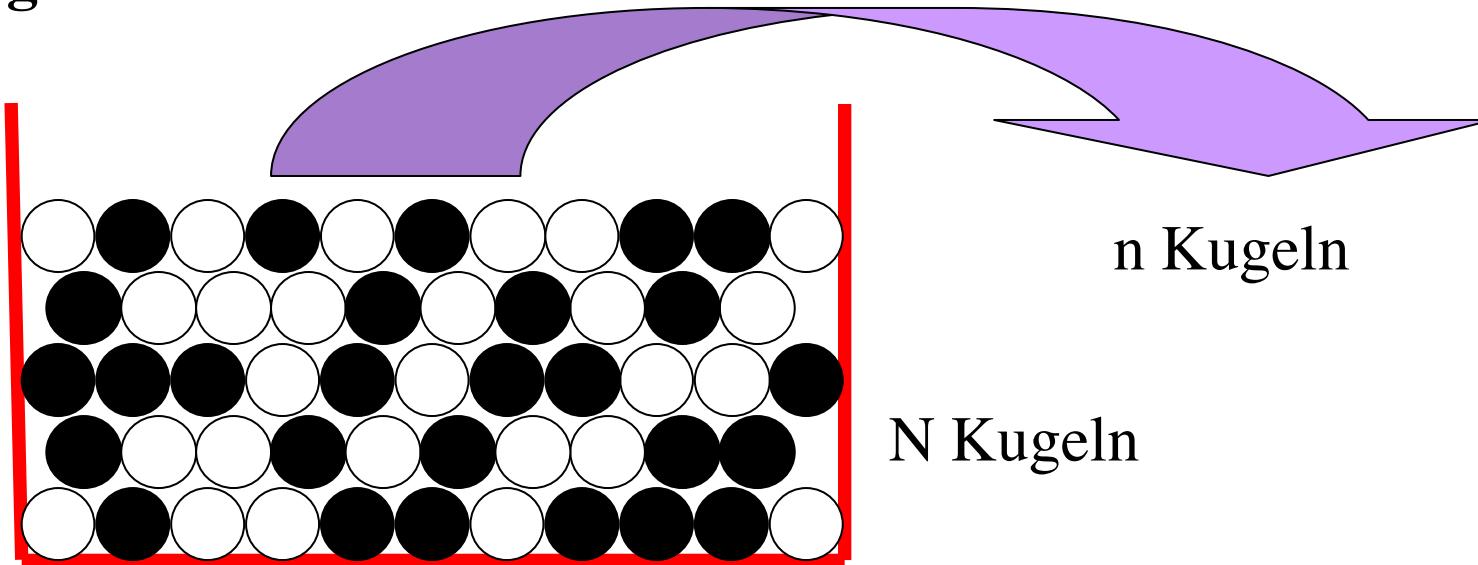
Tabellarisch und grafisch darstellbar

Theoretische Kennzahlen (theoretischer Mittelwert = Erwartungswert; theoretische Varianz etc.)

Diskrete und stetige Merkmale

2.2.1 Die hypergeometrische Verteilung

Abbildung 27: Das Urnenmodell



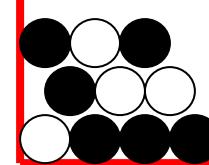
Zufälliges Ziehen ohne Zurücklegen:

- Lotto
- Art der Stichprobenziehung, die Rückschluss von Stichprobe auf Grundgesamtheit ermöglicht

Beispiel 18: Rechnen mit Wahrscheinlichkeiten (Urnenmodell ohne Z.)

$N = 10$ Kugeln, $A = 4$ weiße, $N-A = 6$ schwarze

$n = 3$ Kugeln werden entnommen



Wie wahrscheinlich ist es, dass unter den 3 gezogenen Kugeln genau eine weiße ist ($x \dots$ Anzahl der gezogenen weißen Kugeln)?

Alle Kombinationen der 10 Kugeln gleich wahrscheinlich → Abzählregel

Mögliche Kombinationen: $\binom{10}{3} = 120$

(Kugeln mit den Nummern 1,2,3; 1,2,4; 1,2,5 ...)

Günstige Kombinationen:

eine weiße Kugel mit zwei schwarzen Kugeln kombinieren

Kugel 1 bis 4 sind weiß, Kugeln 5 bis 10 schwarz

Für eine weiße Kugel unter den drei gezogenen Kugeln muss eine der vier weißen gezogen werden, für zwei schwarze unter den drei gezogenen Kugeln müssen zwei der sechs schwarzen gezogen werden:

Gesamtzahl der Möglichkeiten: $\binom{4}{1} \cdot \binom{6}{2}$



$$\rightarrow \Pr(x=1) = \frac{\binom{4}{1} \cdot \binom{6}{2}}{\binom{10}{3}} = \frac{4 \cdot 15}{120} = 0,5; \quad \Pr(x=2), \Pr(x=0) \quad ?$$

Verallgemeinerung: N Kugeln, A weiße, n werden zufällig gezogen
 x ... Anzahl der gezogenen weißen Kugeln

$$\Pr(x = a) = \frac{\binom{A}{a} \cdot \binom{N - A}{n - a}}{\binom{N}{n}} \dots \text{Hypergeometrische Verteilung} \quad (12)$$

Theoretischer Mittelwert μ an gezogenen weißen Kugeln:

$$\mu = n \cdot \frac{A}{N} \quad (\text{in Beispiel 18: } \mu = 3 \cdot \frac{4}{10} = 1,2)$$

Theoretische Varianz σ^2 :

$$\sigma^2 = n \cdot \frac{A}{N} \cdot \left(1 - \frac{A}{N}\right) \cdot \frac{N - n}{N - 1} \quad (\text{in 18: } \sigma^2 = 3 \cdot \frac{4}{10} \cdot \left(1 - \frac{4}{10}\right) \cdot \frac{10 - 3}{10 - 1} = 0,56)$$

2.2.2 Die Binomialverteilung

Urnenmodell, aber Ziehen *mit* Zurücklegen

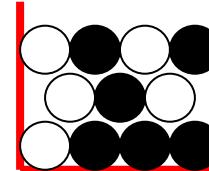
Im Gegensatz zum Ziehen *ohne* Zurücklegen: Vor der Ziehung einer neuen Kugel wieder derselbe Urneninhalt

Anwendung bei unabhängigen Wiederholungen ein und desselben Versuchs

Beispiel 19: Rechnen mit Wahrscheinlichkeiten (Urnenmodell mit Zurücklegen)

$N = 10$ Kugeln, $A = 4$ weiße, $N-A = 6$ schwarze

$n = 3$ Kugeln werden entnommen



Wie wahrscheinlich ist es, dass unter den 3 gezogenen Kugeln genau eine weiße ist ($x \dots$ Anzahl der gezogenen weißen Kugeln)?

Beispiel: $\bigcirc \bullet \bullet$ (die erste gezogene Kugel ist weiß, die danach gezogenen sind schwarz)

Wahrscheinlichkeit für die Reihenfolge $\bigcirc \bullet \bullet$:

$$0,4 \cdot 0,6 \cdot 0,6 = 0,4^1 \cdot 0,6^2 = 0,144.$$

Aber auch $\bullet \bigcirc \bullet$ enthält nur eine weiße Kugel:

$$0,6 \cdot 0,4 \cdot 0,6 = 0,4^1 \cdot 0,6^2 = 0,144.$$

Ebenso $\bullet \bullet \bigcirc$:

$$0,6 \cdot 0,6 \cdot 0,4 = 0,4^1 \cdot 0,6^2 = 0,144.$$

Die Gesamtwahrscheinlichkeit für eine gezogene weiße Kugel ist daher:

$$3 \cdot 0,144 = 0,432$$

oder einfach

$$\Pr(x = 1) = \binom{3}{1} \cdot 0,4^1 \cdot 0,6^2 = 0,432.$$

Verallgemeinerung: N Kugeln, A weiße, n werden zufällig gezogen

$$\pi = A/N \text{ und } 1-\pi = 1-A/N$$

x ... Anzahl der gezogenen weißen Kugeln

$$\Pr(x = a) = \binom{n}{a} \cdot \pi^a \cdot (1 - \pi)^{n-a} \dots \text{Binomialverteilung} \quad (13)$$

Theoretischer Mittelwert μ an gezogenen weißen Kugeln:

$$\mu = n \cdot \pi \quad (\text{in Beispiel 19: } \mu = 3 \cdot \frac{4}{10} = 1,2)$$

Theoretische Varianz σ^2 :

$$\sigma^2 = n \cdot \pi \cdot (1 - \pi)$$

Bei sehr kleinem π und großem n: **Poissonverteilung** (Grenzverteilung der Binomialverteilung)

2.2.3 Die Normalverteilung

Stetige Merkmale: Wahrscheinlichkeiten nur für Intervalle

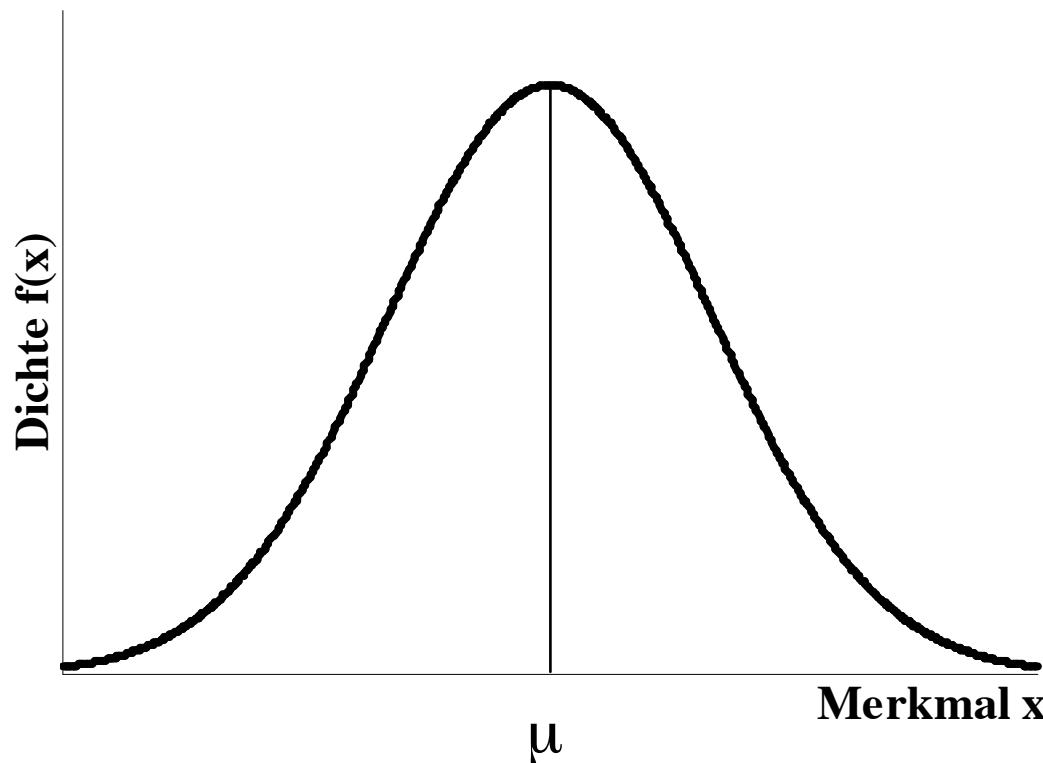
Grafische Darstellung: **Dichte** $f(x)$ statt Wahrscheinlichkeit nach oben auftragen

Eigenschaft der Dichte:

Die Fläche zwischen der Dichte und der x-Achse muss in jedem Intervall der Wahrscheinlichkeit dieses Intervalls entsprechen.

→ Berechnung von Wahrscheinlichkeiten = Berechnen von Flächen

Abbildung 29: Ein normalverteiltes Merkmal



Die Dichte der Normalverteilung ist mathematisch beschreibbar:

$$f(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}}$$

Beispiel 21: Funktionsdauer von Taschenrechnern

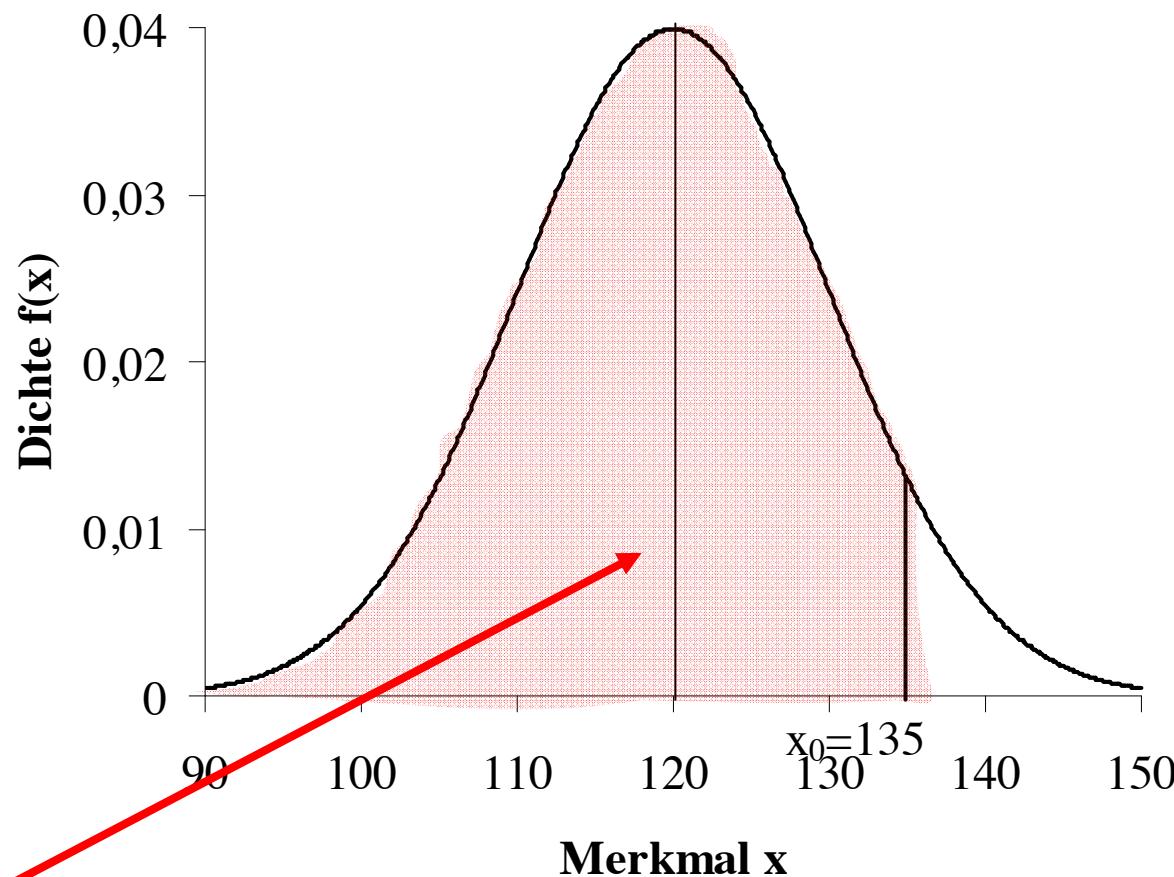
Die Funktionsdauer x ist normalverteilt mit Erwartungswert $\mu = 120$ h und theoretischer Varianz $\sigma^2 = 100$.

Wie wahrscheinlich ist es, dass die Funktionsdauer

- a) höchstens 135 h
- b) mehr als 135 h
- c) mehr als 105 h
- d) höchstens 105 h beträgt?

Berechnung von a)

Abbildung 30: Normalverteilung



$$\Pr(x \leq 135) = \int_{-\infty}^{135} f(x)dx = \int_{-\infty}^{135} \frac{1}{\sqrt{2 \cdot \pi} \cdot 10} \cdot e^{-\frac{1}{2} \cdot \frac{(x-120)^2}{100}} dx$$

Beispiel 22: Standardnormalverteilung

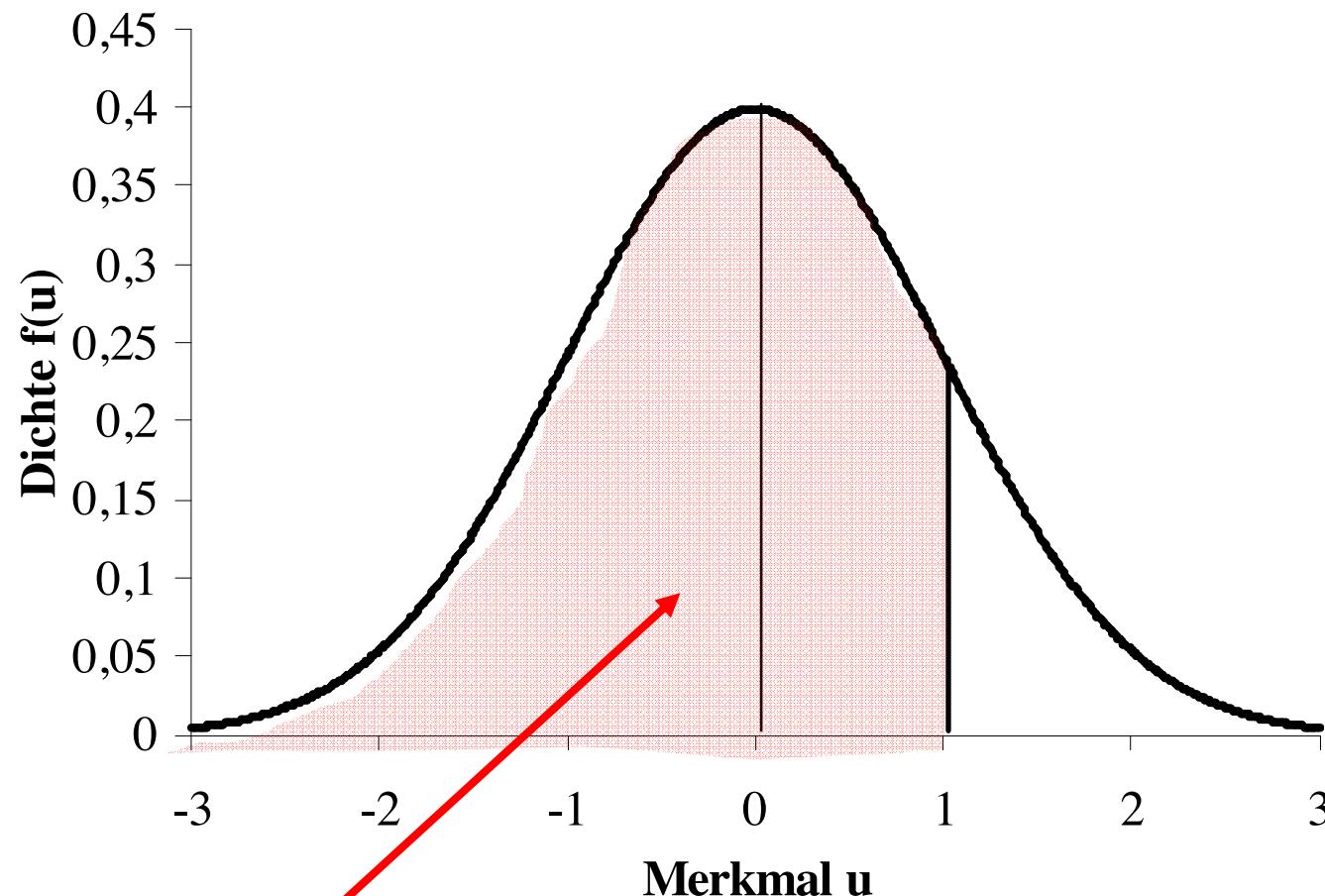
Ein stetiges Merkmal u ist normalverteilt mit Erwartungswert $\mu = 0$ und theoretischer Varianz $\sigma^2 = 1$.

Zu berechnen ist die Wahrscheinlichkeit dafür, dass ein Messwert

- a) höchstens 1
- b) größer als 1
- c) größer als -1
- d) höchstens -1 ist.

Berechnung von a)

Abbildung 31: Standardnormalverteilung



$$\Pr(u \leq 1) = \int_{-\infty}^1 \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{x^2}{2}} dx. \text{ Die Lösung des Integrals ist tabelliert!}$$



Tabelle A (Standardnormalverteilung):

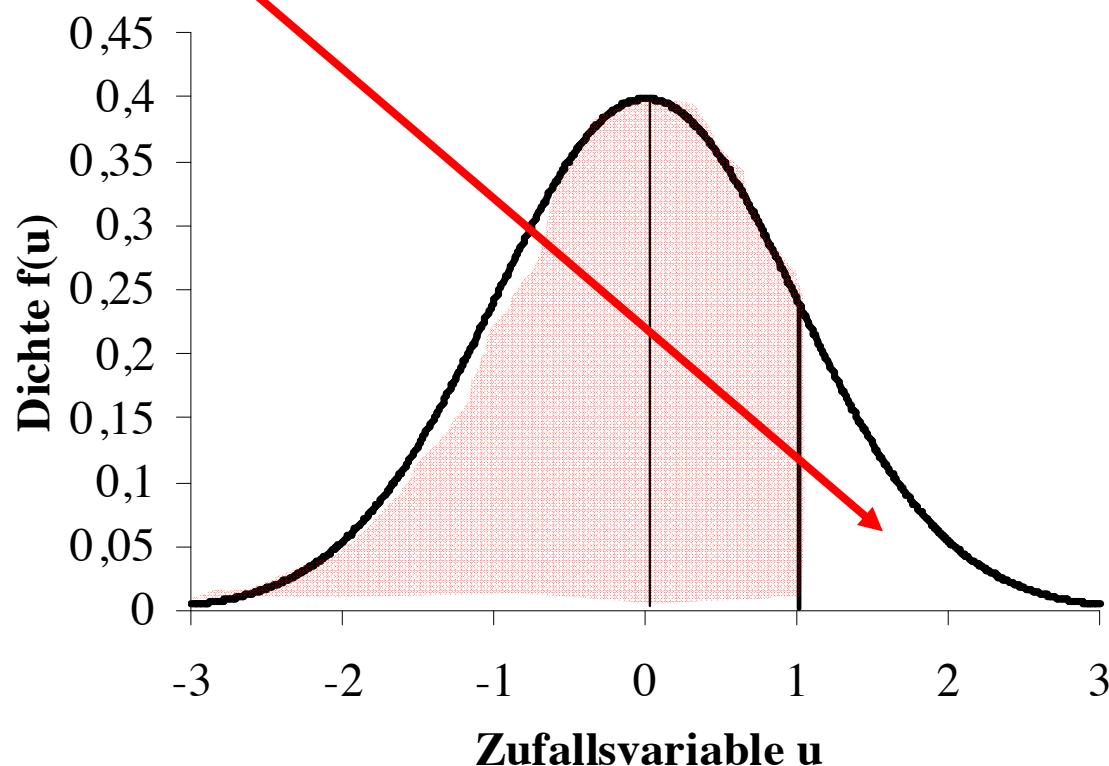
u_0	$\Pr(u \leq u_0)$										
0,01	0,5040	0,51	0,6950	1,01	0,8438	1,51	0,9345	2,01	0,9778	2,51	0,9940
0,02	0,5080	0,52	0,6985	1,02	0,8461	1,52	0,9357	2,02	0,9783	2,52	0,9941
0,03	0,5120	0,53	0,7019	1,03	0,8485	1,53	0,9370	2,03	0,9788	2,53	0,9943
0,04	0,5160	0,54	0,7054	1,04	0,8508	1,54	0,9382	2,04	0,9793	2,54	0,9945
0,05	0,5199	0,55	0,7088	1,05	0,8531	1,55	0,9394	2,05	0,9798	2,55	0,9946
0,06	0,5239	0,56	0,7123	1,06	0,8554	1,56	0,9406	2,06	0,9803	2,56	0,9948
0,07	0,5279	0,57	0,7157	1,07	0,8577	1,57	0,9418	2,07	0,9808	2,57	0,9949
0,08	0,5319	0,58	0,7190	1,08	0,8599	1,58	0,9429	2,08	0,9812	2,58	0,9951
0,09	0,5359	0,59	0,7224	1,09	0,8621	1,59	0,9441	2,09	0,9817	2,59	0,9952
0,1	0,5398	0,6	0,7257	1,1	0,8643	1,6	0,9452	2,1	0,9821	2,6	0,9953
...
0,41	0,6591	0,91	0,8186	1,41	0,9207	1,91	0,9719	2,41	0,9920	2,91	0,9982
0,42	0,6628	0,92	0,8212	1,42	0,9222	1,92	0,9726	2,42	0,9922	2,92	0,9982
0,43	0,6664	0,93	0,8238	1,43	0,9236	1,93	0,9732	2,43	0,9925	2,93	0,9983
0,44	0,6700	0,94	0,8264	1,44	0,9251	1,94	0,9738	2,44	0,9927	2,94	0,9984
0,45	0,6736	0,95	0,8289	1,45	0,9265	1,95	0,9744	2,45	0,9929	2,95	0,9984
0,46	0,6772	0,96	0,8315	1,46	0,9279	1,96	0,9750	2,46	0,9931	2,96	0,9985
0,47	0,6808	0,97	0,8340	1,47	0,9292	1,97	0,9756	2,47	0,9932	2,97	0,9985
0,48	0,6844	0,98	0,8365	1,48	0,9306	1,98	0,9761	2,48	0,9934	2,98	0,9986
0,49	0,6879	0,99	0,8389	1,49	0,9319	1,99	0,9767	2,49	0,9936	2,99	0,9986
0,5	0,6915	1	0,8413	1,5	0,9332	2	0,9772	2,5	0,9938	3	0,9987

a) $\Pr(u \leq 1) = 0,841$

[Anmerkung: Bei stetigen Merkmalen gilt:

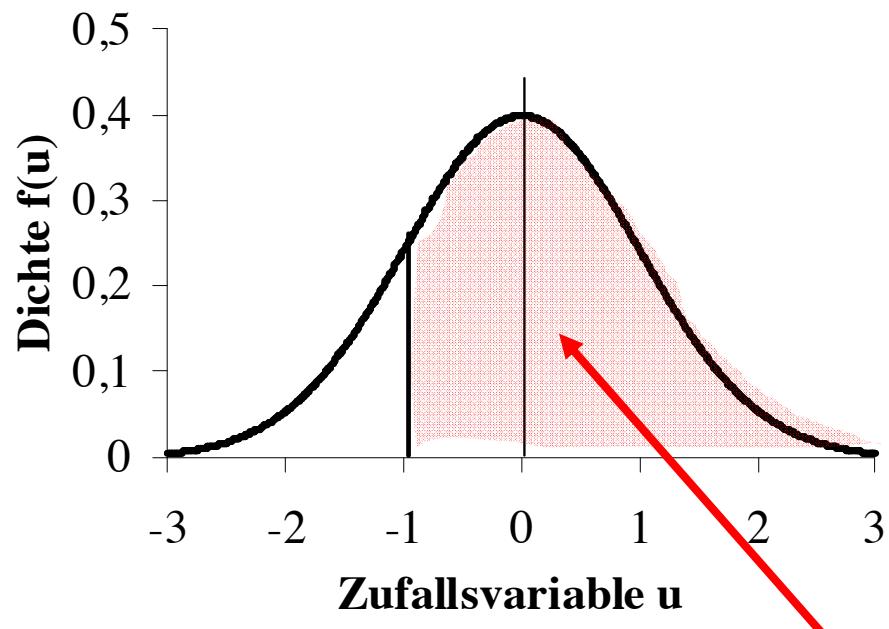
$$\Pr(u \leq 1) = \Pr(u < 1)]$$

b) $\Pr(u > 1)$:



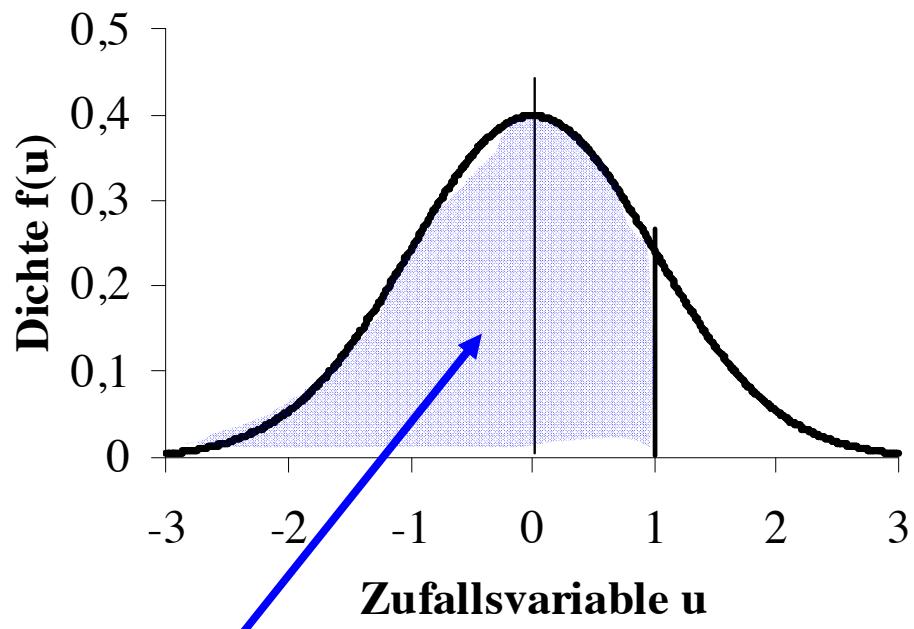
Gegenwahrscheinlichkeit: $1 - \Pr(u \leq 1) = 1 - 0,841 = \underline{\underline{0,159}}$

c) $\Pr(u > -1)$:

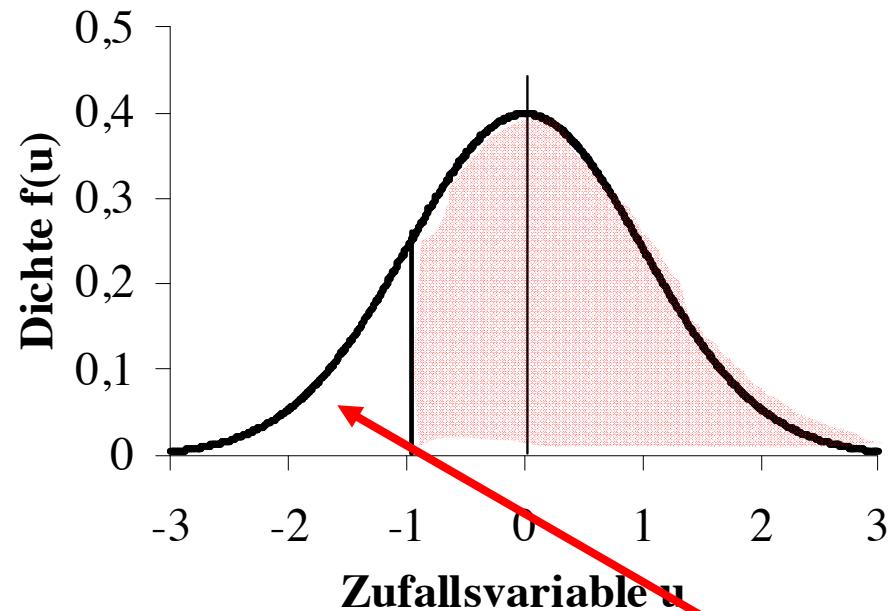


$$\Pr(u > -1) = \Pr(u < +1)$$

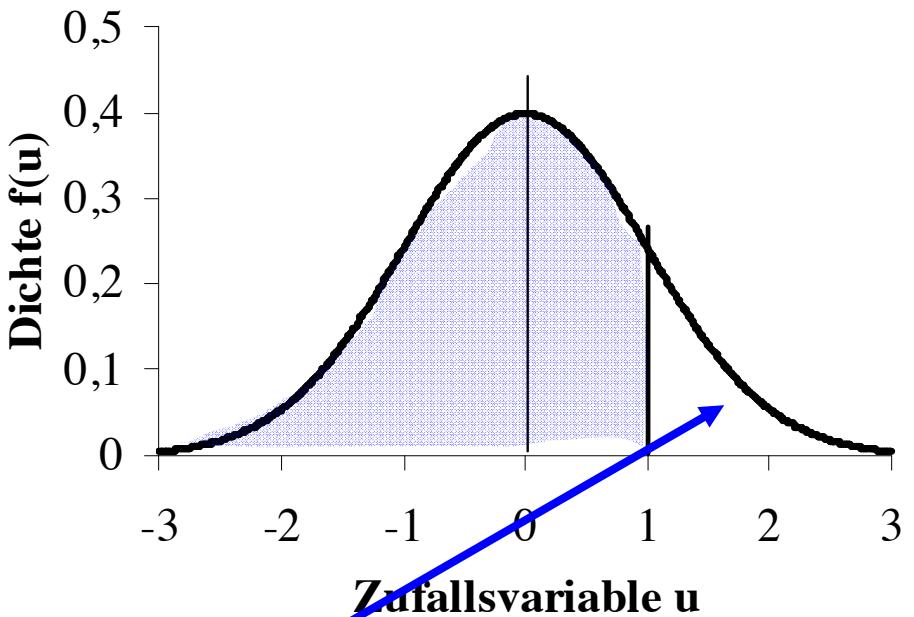
→ $\Pr(u > -1) = \underline{0,841}$



d) $\Pr(u \leq -1)$:



$$\Pr(u \leq -1) = \Pr(u \geq +1)$$

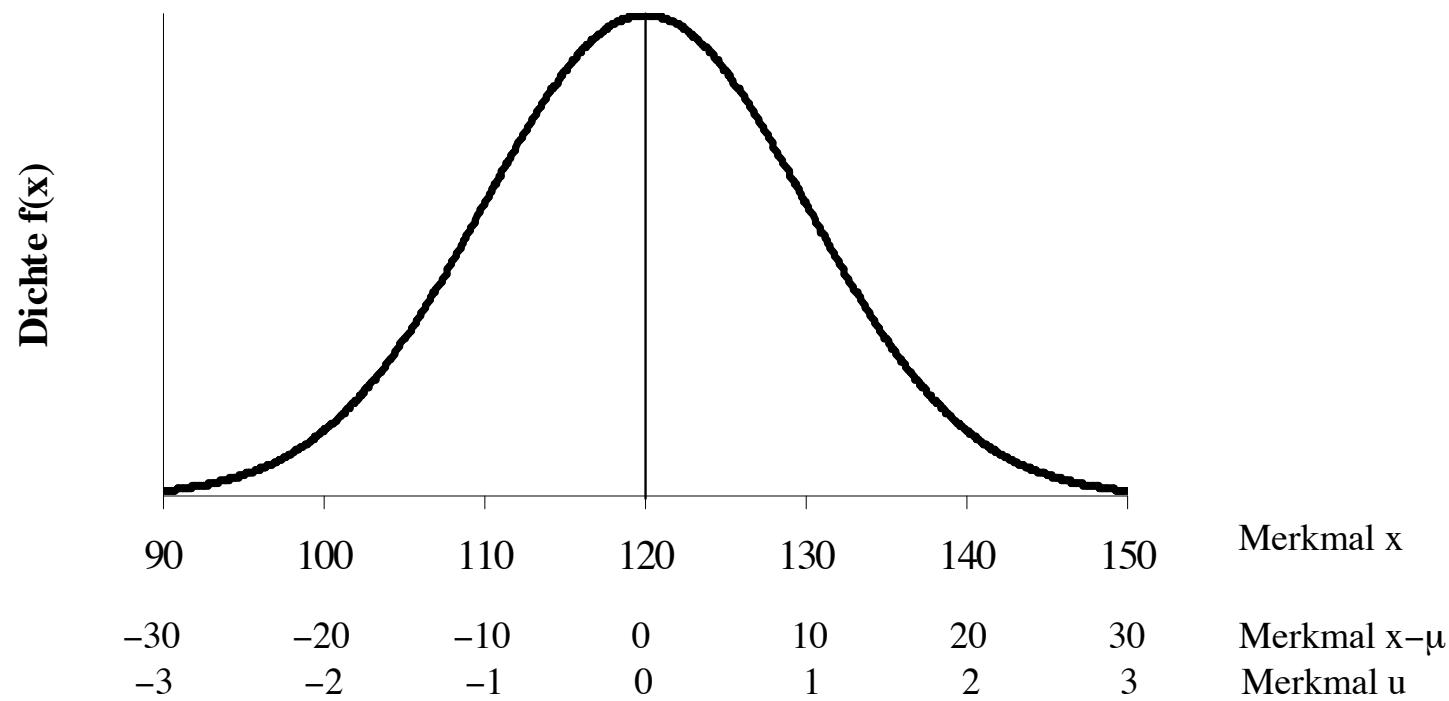


$$\rightarrow \Pr(u \leq -1) = 1 - 0,841 = \underline{\underline{0,159}}$$

Verallgemeinerung auf alle Normalverteilungen durch **Standardisierung**:

$$u_0 = \frac{x_0 - \mu}{\sigma} \quad (14)$$

Abbildung 33: Grafische Darstellung der Standardisierungsschritte



Beispiel 21: Funktionsdauer von Taschenrechnern

Die Funktionsdauer x ist normalverteilt mit Erwartungswert $\mu = 120$ h und theoretischer Varianz $\sigma^2 = 100$.

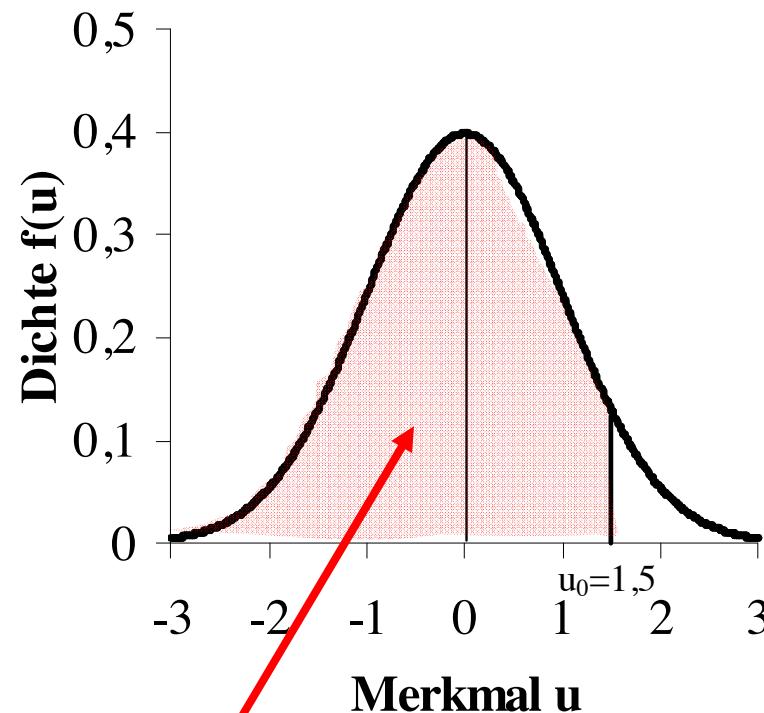
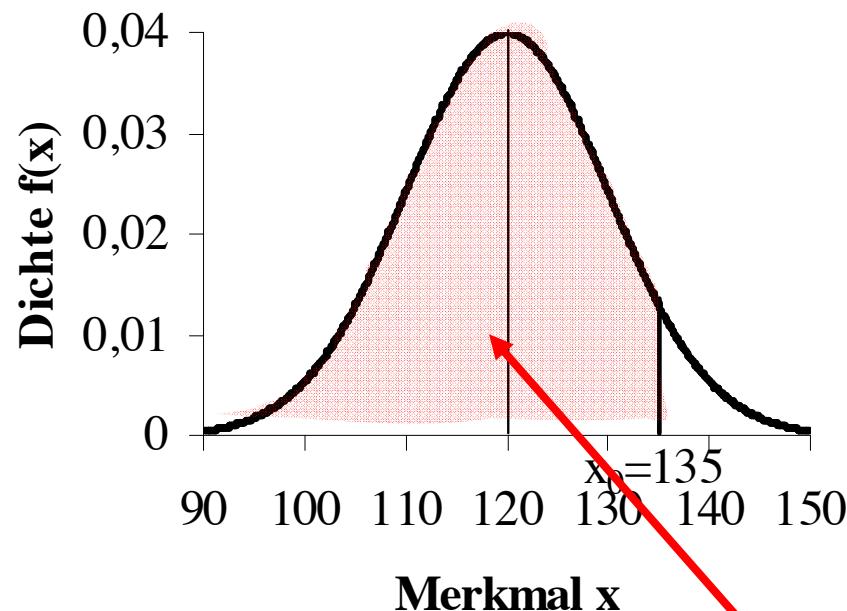
Wie wahrscheinlich ist es, dass die Funktionsdauer

- a) höchstens 135 h
- b) mehr als 135 h
- c) mehr als 105 h
- d) höchstens 105 h beträgt?

a) $\Pr(x \leq 135)$:

Der standardisierte Wert u_0 von $x_0 = 135$ ist:

$$u_0 = \frac{135 - 120}{10} = 1,5$$



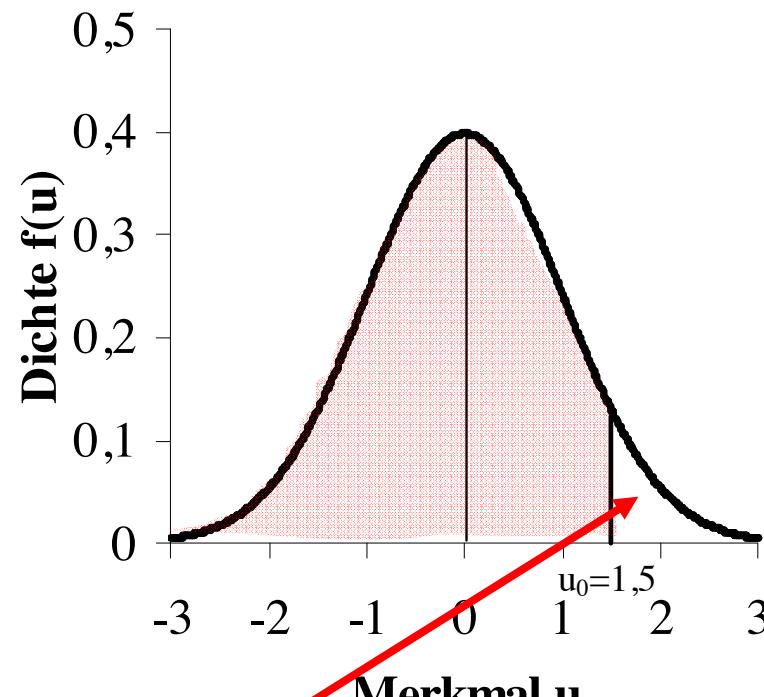
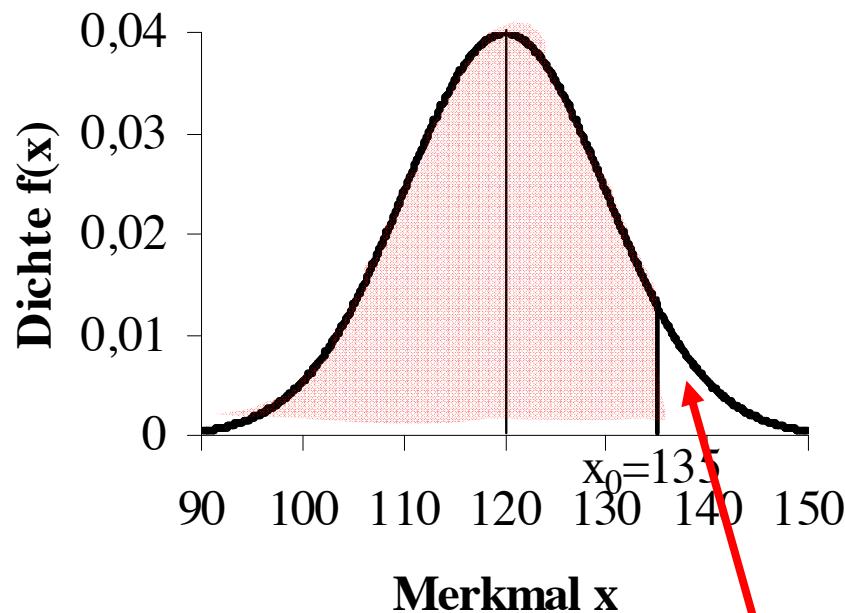
$$\Pr(x \leq 135) = \Pr(u \leq 1,5)$$

→ $\Pr(x \leq 135) = \underline{0,933}$

b) $\Pr(x > 135)$:

Der standardisierte Wert u_0 von $x_0 = 135$ ist:

$$u_0 = \frac{135 - 120}{10} = 1,5$$



$$\Pr(x > 135) = \Pr(u > 1,5)$$

$$\Rightarrow \Pr(x > 135) = 1 - 0,933 = \underline{\underline{0,067}}$$

c) $\Pr(x > 105)$:

Der zum Wert $x_0 = 105$ standardisierte Wert u_0 ist:

$$u_0 = \frac{105 - 120}{10} = -1,5$$

$$\Pr(x > 105) = \Pr(u > -1,5) = \Pr(u < +1,5) = \underline{\underline{0,933}}$$

d) $\Pr(x \leq 105)$:

Der zum Wert $x_0 = 105$ standardisierte Wert u_0 ist:

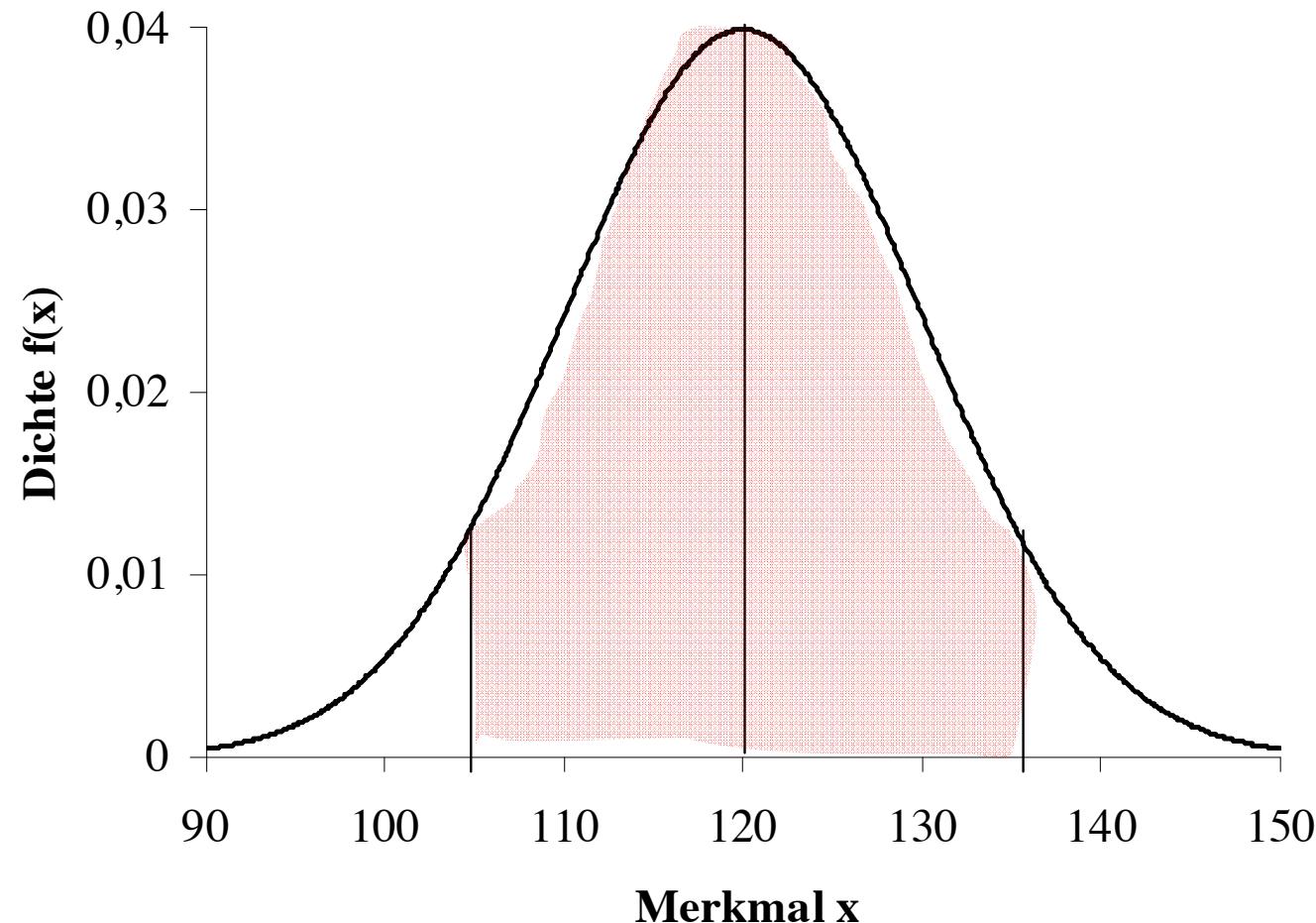
$$u_0 = \frac{105 - 120}{10} = -1,5$$

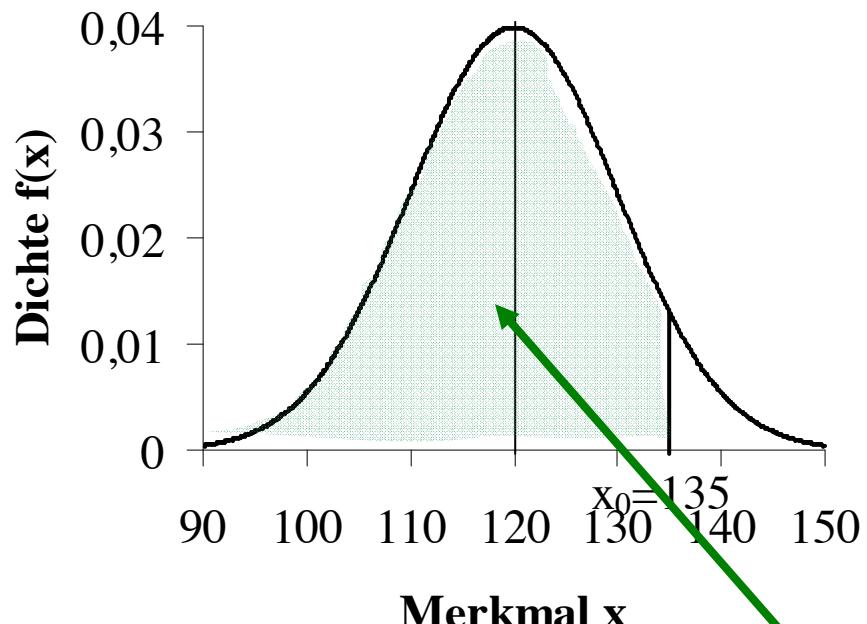
$$\Pr(x \leq 105) = \Pr(u \leq -1,5) = \Pr(u \geq +1,5) = 1 - 0,933 = \underline{\underline{0,067}}$$

Beispiel 23: Rechnen mit der Normalverteilung

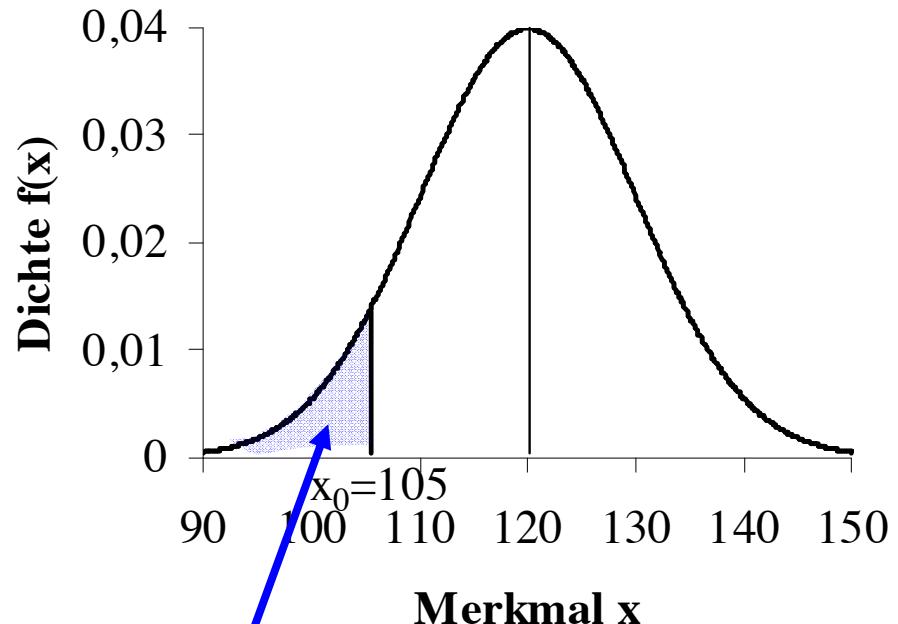
Wahrscheinlichkeit, dass x zwischen 105 und 135 Stunden liegt.

Abbildung 34: Normalverteilung





$$\Pr(105 \leq x \leq 135) = \Pr(x \leq 135) - \Pr(x < 105)$$



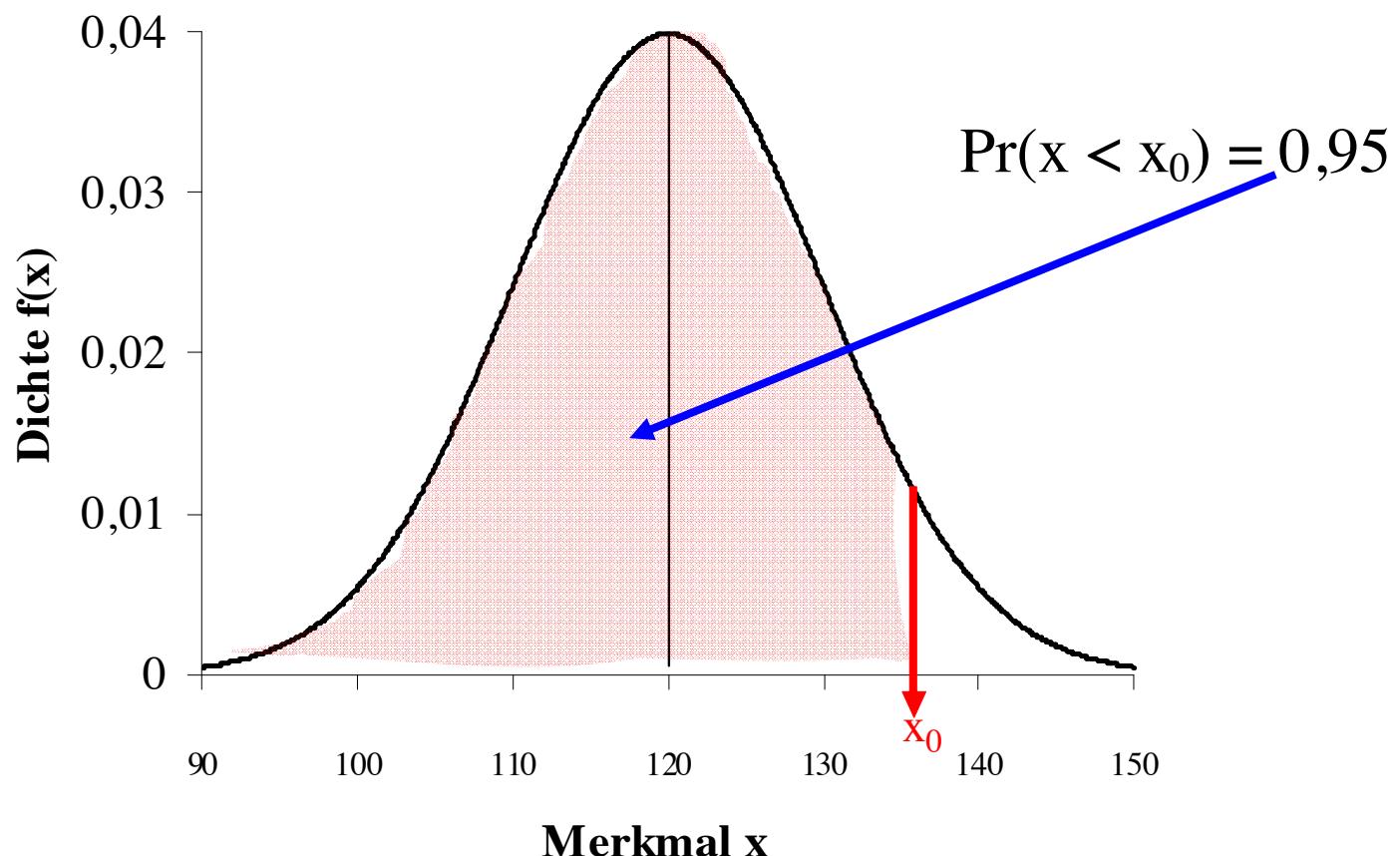
$$\text{Standardisierungen: } u_0 = \frac{135 - 120}{10} = 1,5 \quad \text{und} \quad u_0 = \frac{105 - 120}{10} = -1,5$$

$$\begin{aligned} \Pr(x \leq 135) - \Pr(x < 105) &= \Pr(u \leq 1,5) - \Pr(x < -1,5) = 0,933 - 0,067 = \\ &= 0,866 \end{aligned}$$

Beispiel 24: Rechnen mit der Normalverteilung

Funktionsdauer x_0 , die mit einer Wahrscheinlichkeit von 0,95 unterschritten wird.

Abbildung 35: Normalverteilung



$$(14): u_0 = \frac{x_0 - \mu}{\sigma}$$

$\Pr(x < x_0) = \Pr(u < u_0) = 0,95 \rightarrow u_0 = ? \rightarrow$ Tabelle A (Standardnormalverteilung):

$$1,65 = \frac{x_0 - 120}{10}$$

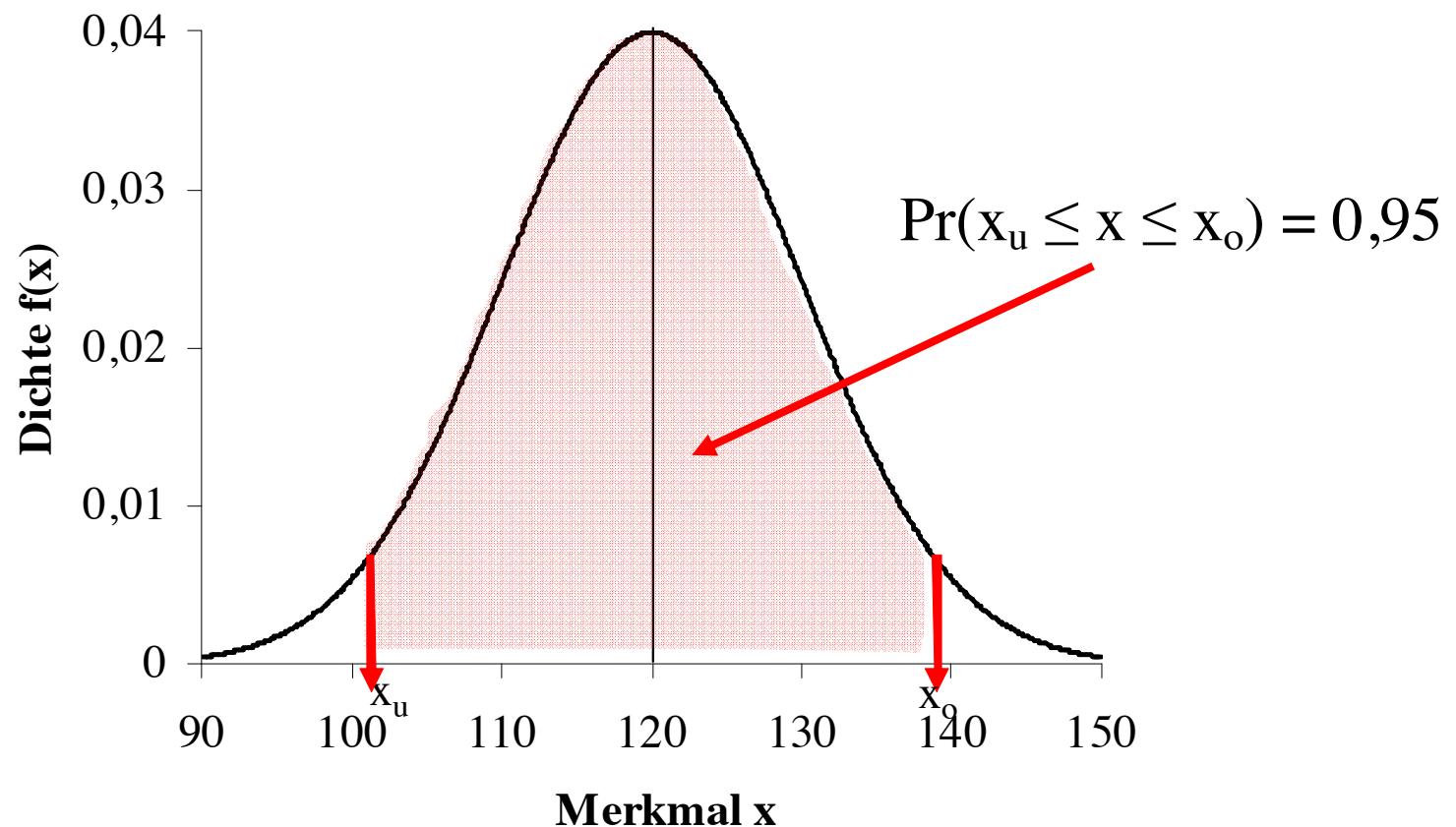
$$x_0 = 1,65 \cdot 10 + 120 = 136,5$$

u_0	$\Pr(u \leq u_0)$
1,51	0,9345
1,52	0,9357
1,53	0,9370
1,54	0,9382
1,55	0,9394
1,56	0,9406
1,57	0,9418
1,58	0,9429
1,59	0,9441
1,6	0,9452
1,61	0,9463
1,62	0,9474
1,63	0,9484
1,64	0,9495
1,65	0,9505
1,66	0,9515
1,67	0,9525
1,68	0,9535
1,69	0,9545
1,7	0,9554
...	...

Beispiel 25: Rechnen mit der Normalverteilung

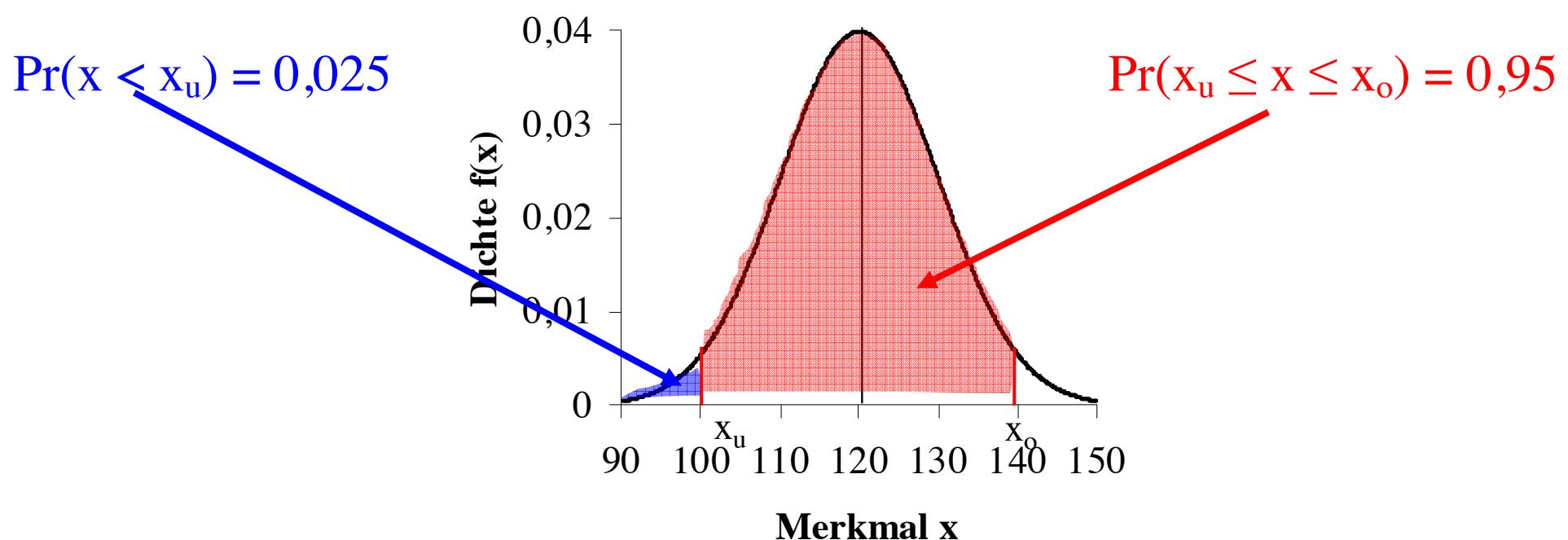
„Um den Erwartungswert symmetrisches Intervall“ (Untergrenze x_u und Obergrenze x_o), in dem die Wahrscheinlichkeit 0,95 beträgt.

Abbildung 36: Normalverteilung



$$(14): u_0 = \frac{x_0 - \mu}{\sigma}$$

$$\Pr(x_u \leq x \leq x_o) = 0,95 \rightarrow \Pr(x \leq x_o) = ?$$



$$\Pr(x \leq x_o) = \Pr(u \leq u_0) = 0,975 \rightarrow u_0 = ? \quad \rightarrow$$

$$1,96 = \frac{x_0 - 120}{10}$$

$$x_o = 1,96 \cdot 10 + 120 = 139,6$$

$$x_u = 120 - \underbrace{(139,6 - 120)}_{19,6} = 100,4$$

Tabelle A (Standardnormalverteilung):

1,9	0,9713
1,91	0,9719
1,92	0,9726
1,93	0,9732
1,94	0,9738
1,95	0,9744
1,96	0,9750
1,97	0,9756
1,98	0,9761
1,99	0,9767
2	0,9772

Beispiel 26: Annäherung der Hypergeometrischen Verteilung an die Normalverteilung

$N = 6,000,000$, $A = 2,400,000$ (Wähler einer bestimmten Partei in der Grundgesamtheit), $x \dots$ Anzahl von Wählern in der Stichprobe

Die Verteilung der Anzahlen von Wählern $n = 3$:

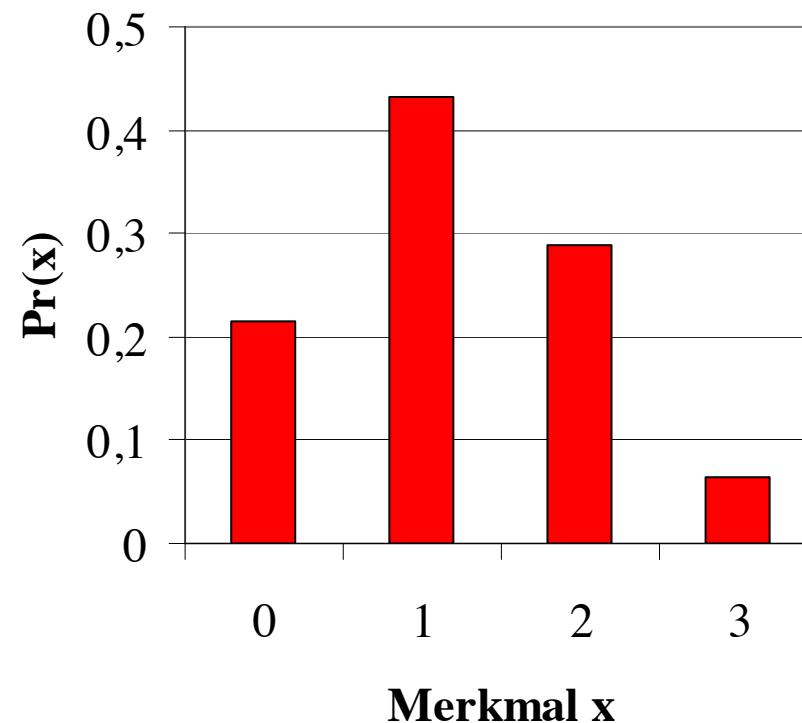


Abbildung 37: $n = 10$

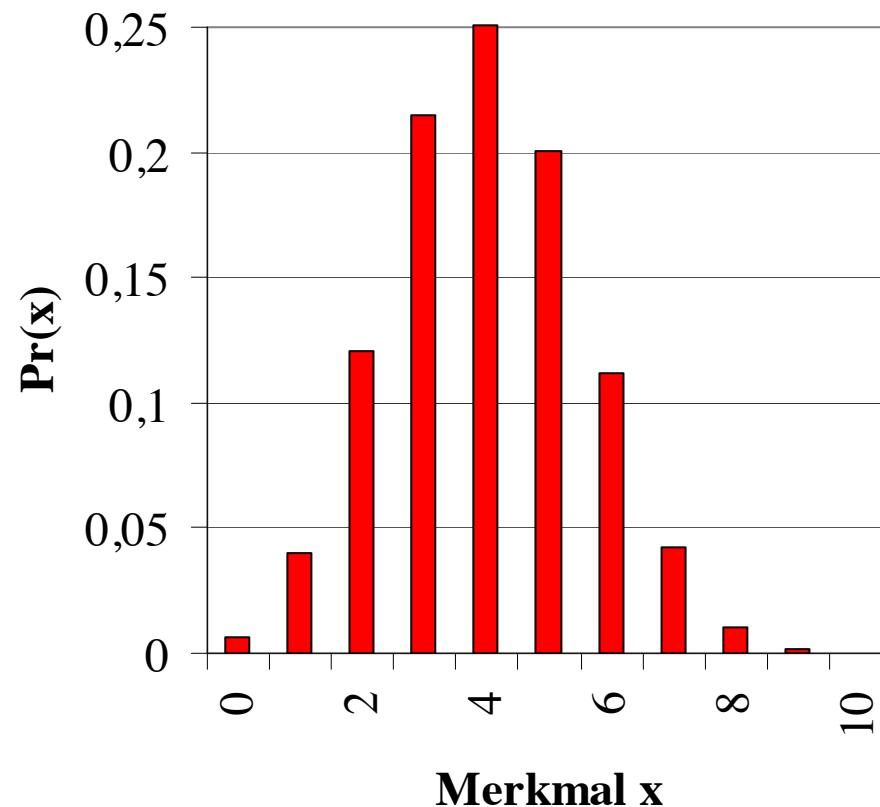


Abbildung 38: $n = 50$

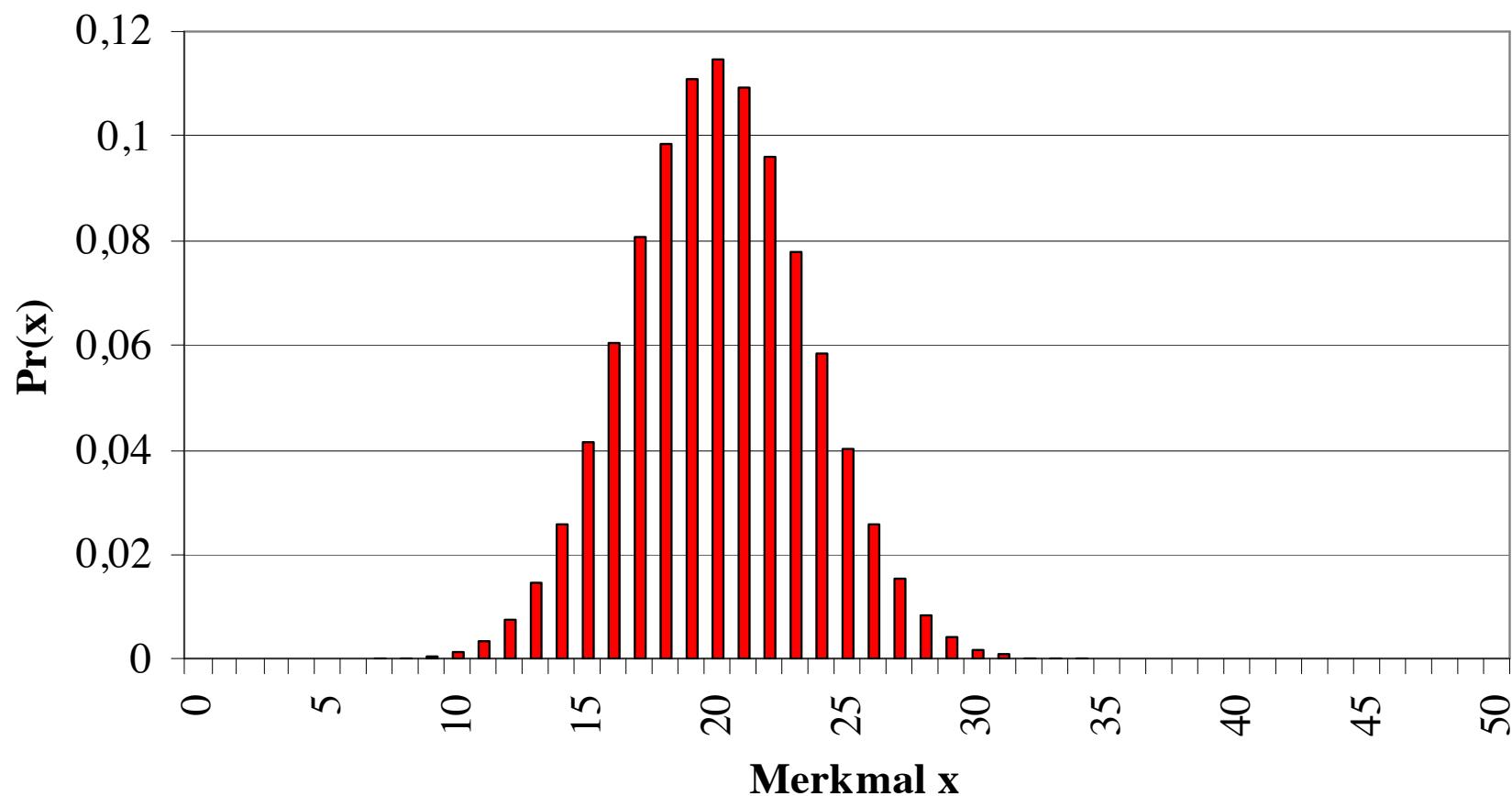
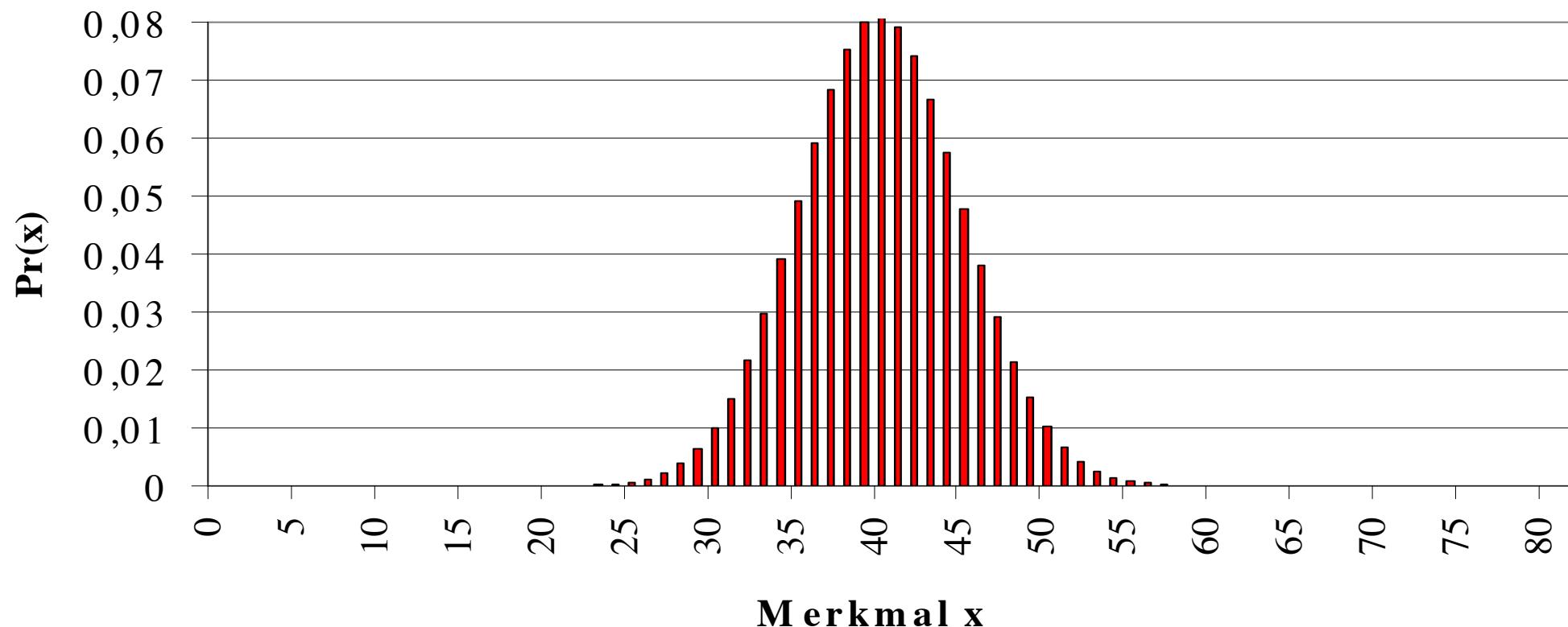


Abbildung 39: $n = 100$



Das ist der Inhalt des „Zentralen Grenzwertsatzes“ der Statistik

Erwartungswert und theoretische Varianz dieser Normalverteilung entsprechen jenen der hypergeometrischen Verteilung

Verallgemeinerung: N Kugeln, A weiße, n werden zufällig gezogen
 x ... Anzahl der gezogenen weißen Kugeln

$$\Pr(x = a) = \frac{\binom{A}{a} \cdot \binom{N - A}{n - a}}{\binom{N}{n}} \dots \text{Hypergeometrische Verteilung} \quad (12)$$

Theoretischer Mittelwert μ an gezogenen weißen Kugeln:

$$\mu = n \cdot \frac{A}{N} \quad (\text{in Beispiel 18: } \mu = 3 \cdot \frac{4}{10} = 1,2)$$

Theoretische Varianz σ^2 :

$$\sigma^2 = n \cdot \frac{A}{N} \cdot \left(1 - \frac{A}{N}\right) \cdot \frac{N - n}{N - 1}$$

Kapitel 3: Schließende Statistik

3.1 Grundbegriffe

Vollerhebung – Daten der Grundgesamtheit – beschreibende Statistik – Parameter

Stichprobenerhebung – Daten einer Stichprobe – schließende Statistik – Rückschluss auf Parameter

Aus: KRONEN-ZEITUNG <

11. Dezember 2005, S. 3

Wien. – Die drastische Trendumkehr erschreckt. Vor zehn Jahren stimmten 66 % der Österreicher für den EU-Beitritt. Heute dagegen finden 70 %, laut neuester „market“-Umfrage, der Beitritt hätte nichts oder nur wenig gebracht

...

Voraussetzung für den Rückschluss von den Stichprobenergebnissen auf die Parameter: **Repräsentativität** der Stichprobe für die Grundgesamtheit hinsichtlich der interessierenden Merkmale

Große Suggestivwirkung

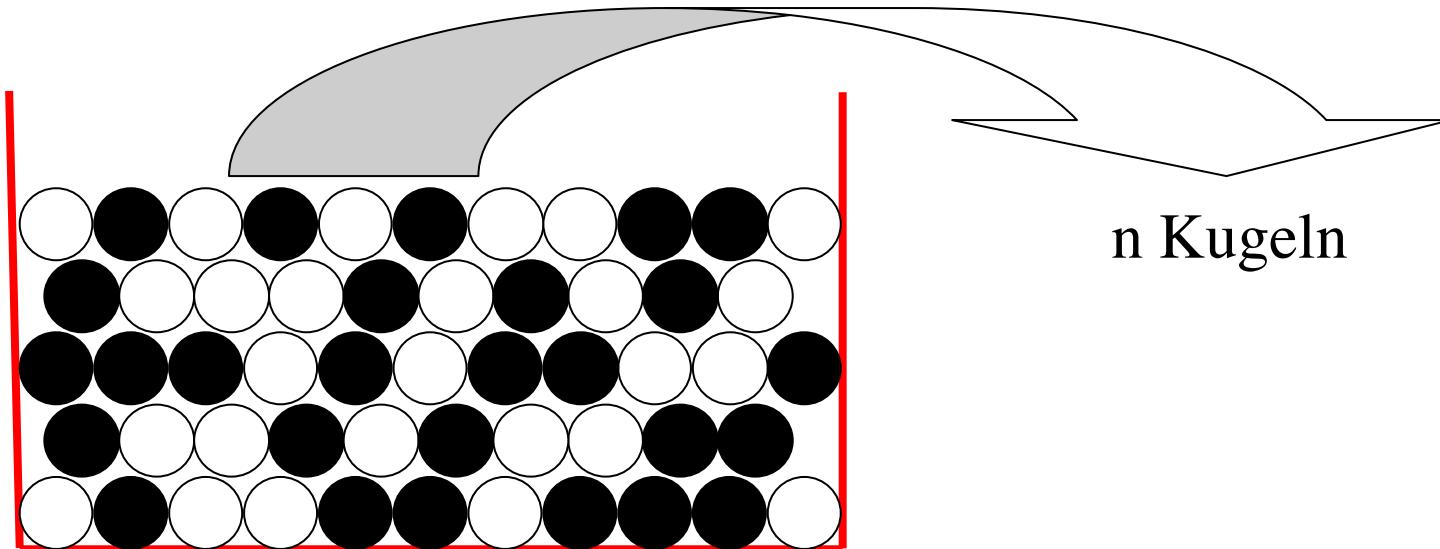
Repräsentativität der Stichprobe ~ Ähnlichkeit zur Grundgesamtheit

Entscheidend für Repräsentativität der Stichprobe: **Stichprobenverfahren** und **Stichprobenumfang**

Nicht jede Stichprobe muss repräsentativ sein: **Informative** Stichproben für den Erhebungszweck

3.2 Stichprobenverfahren

Lösung der Repräsentativitätsaufgabe durch Ziehung exakt nach dem Urnenmodell



... uneingeschränkte (einfache) Zufallsstichprobe

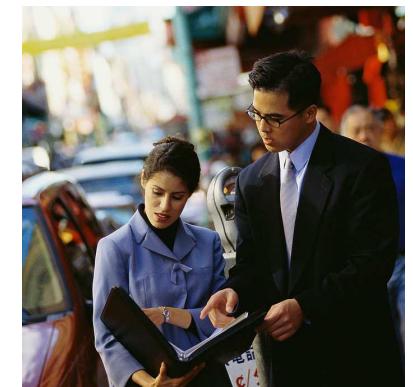
Praktische Umsetzung z.B. in Excel: Funktion ZUFALLSZAHL

Andere **Zufalls**stichprobenverfahren liefern ebenfalls repräsentative Stichproben

Basis ist immer eine Form der Zufallsauswahl:

- zufällige Ziehung der Erhebungseinheiten
- alle Erhebungseinheiten besitzen eine von null verschiedene und berechenbare Auswahlwahrscheinlichkeit

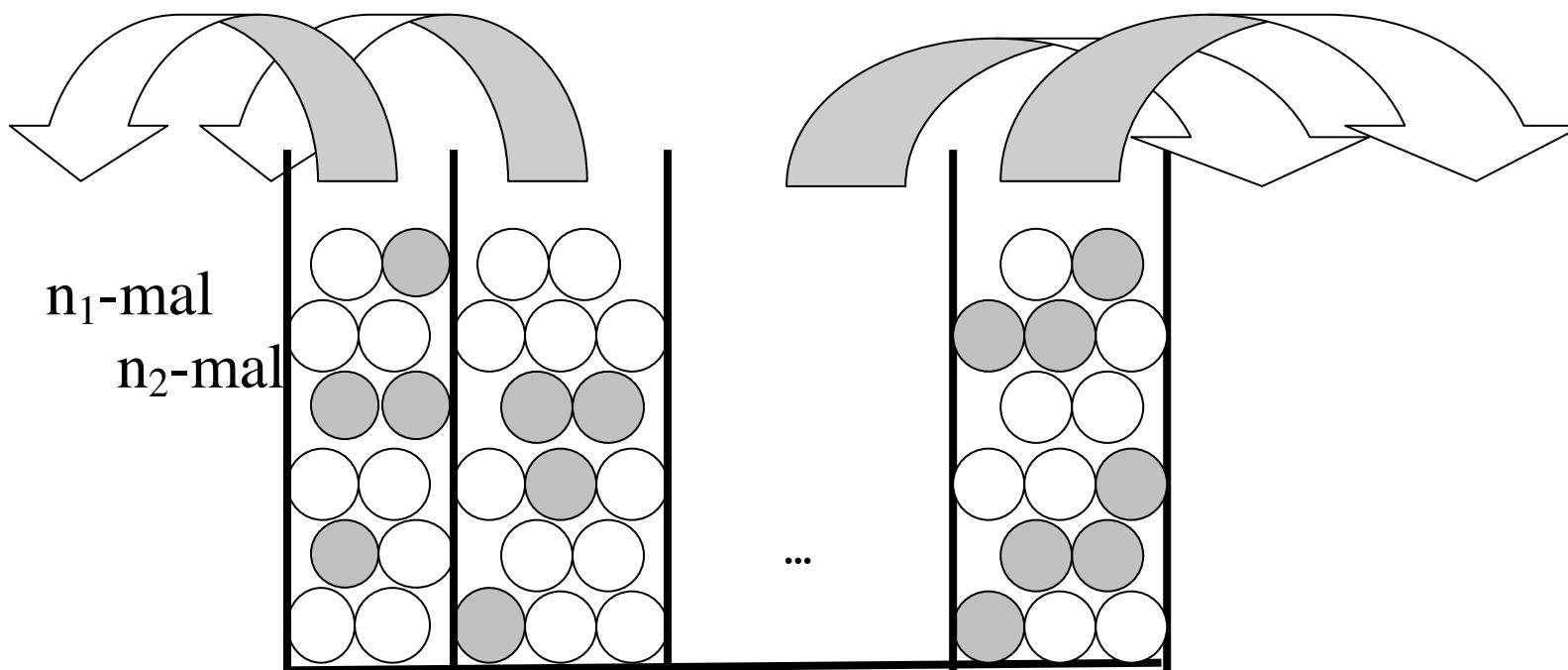
Die Befragung von auf einem Stadtplatz vorbei kommenden Passanten kann (auch bei Einhaltung bestimmter **Quoten** bzgl. Geschlecht, Alter etc.) keine repräsentativen Ergebnisse zum Beispiel für die Wohnbevölkerung eines Landes garantieren (**verzerrte** Stichprobe)



- Geschichtete uneingeschränkte Zufallsauswahl

Zerlegung der Grundgesamtheit in Schichten (z.B. der Männer und Frauen) und uneingeschränkt zufällige Ziehung aus jeder Schicht zum Zweck eines Genauigkeitsgewinns

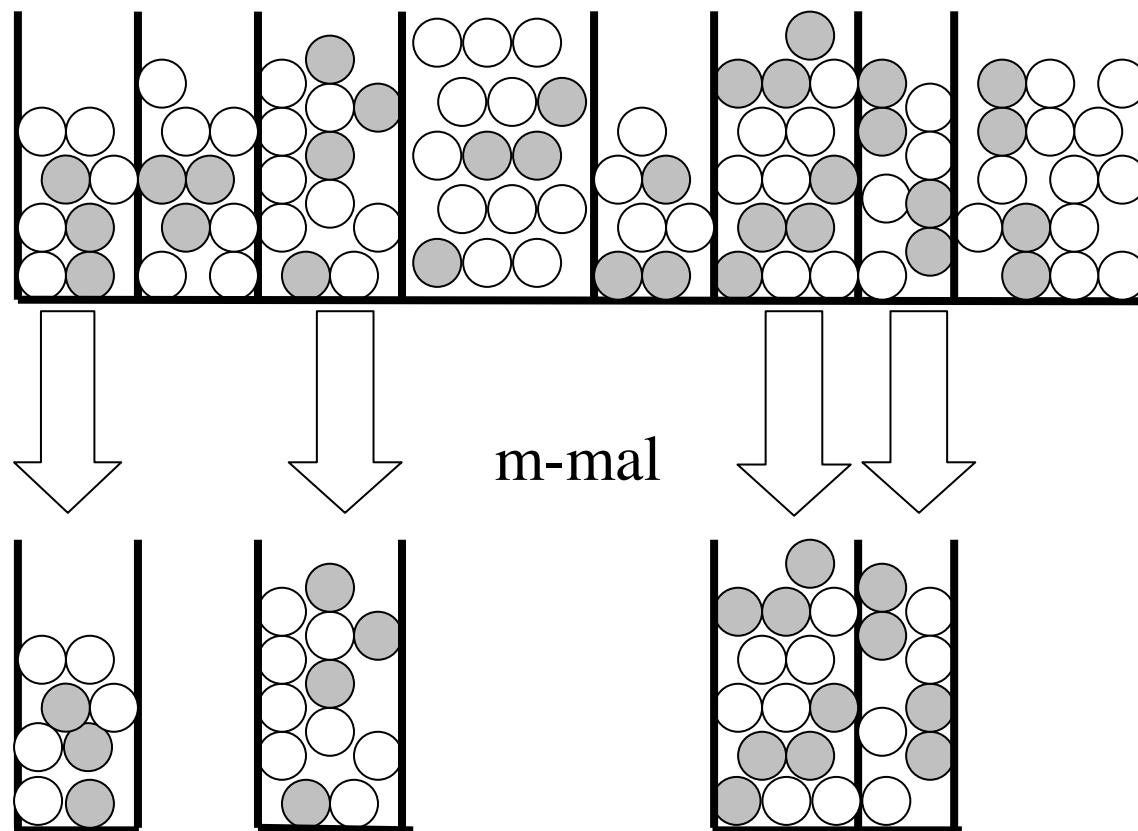
Abbildung 40: Das Urnenmodell bei einer geschichteten uneingeschränkten Zufallsauswahl



- Uneingeschränkte Klumpenauswahl

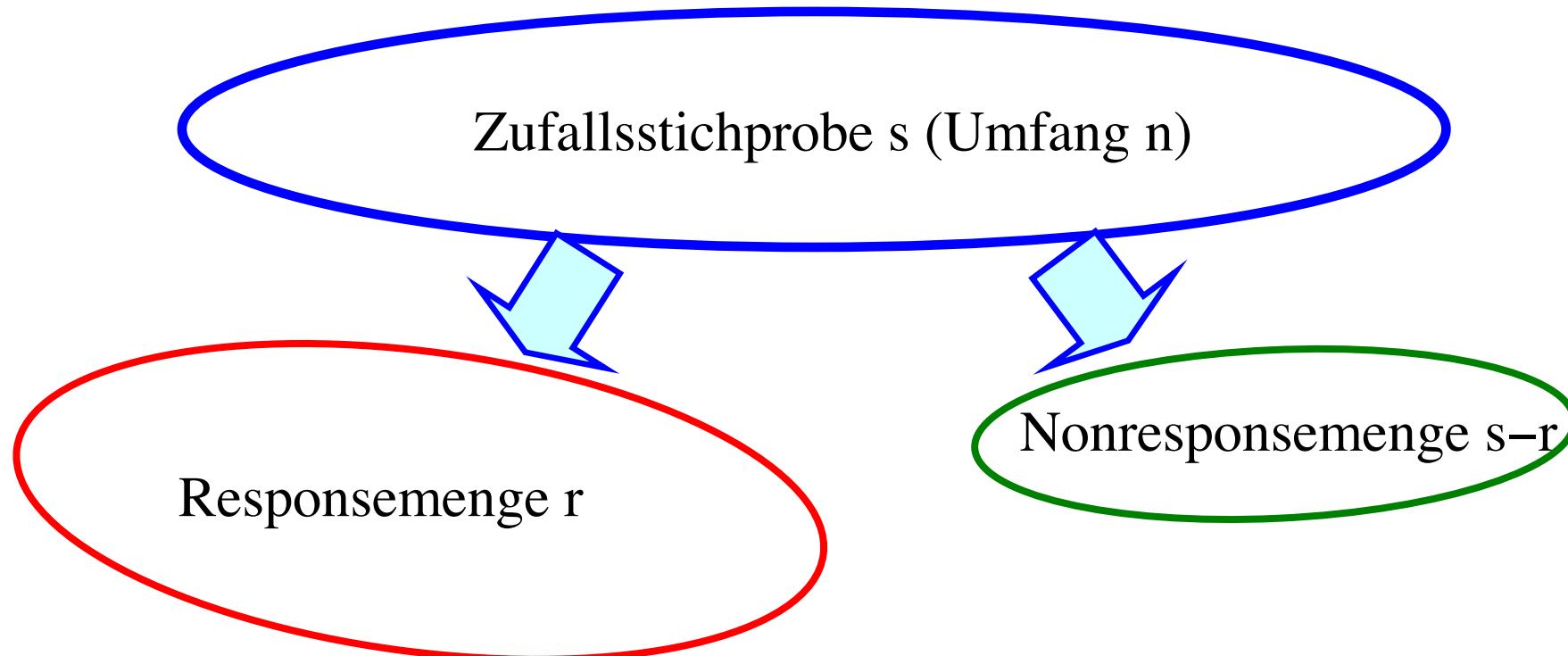
Zerlegung der Grundgesamtheit in Klumpen, die zufällig gezogen werden, zum Zweck der Kostenersparnis

Abbildung 41: Das Urnenmodell einer uneing. Klumpenauswahl



Weiterer Grund für die Verzerrung einer Stichprobe: **Nonresponse**

Abbildung 42: Die Zerlegung einer Stichprobe bei Vorliegen von Nonresponse



Menge r liefert **verzerrte** Stichprobe, wenn $s-r$ nicht sehr klein und sich r und $s-r$ beim interessierenden Merkmal unterscheiden

Nonresponserate so gering wie möglich halten (Anreize zur Teilnahme an der Stichprobenerhebung setzen)

Nachträgliche Kompensierung von Nonresponse nur durch aufwändige statistische Methoden möglich: **Gewichtungsanpassung** und **Datenimputation**

3.3 Die Handlungslogik der schließenden Statistik

In Zufallsstichproben: Stichprobenergebnisse sind **Punktschätzer** für die unbekannten Parameter der Grundgesamtheit

Punktschätzer (aus der Stichprobe) **Parameter (in der Grundgesamtheit)**

Relative Häufigkeit p

Mittelwert \bar{x}

Stichprobenvarianz s^2

Differenz zweier relativer Häufigkeiten (oder Mittelwerte) d

Chiquaret χ_{err}^2

Korrelationskoeffizient r

Regressionskoeffizienten b_1, b_2

Relative Häufigkeit π

Mittelwert μ

Varianz σ^2

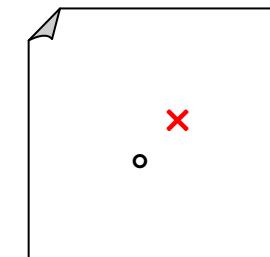
Differenz zweier relativer Häufigkeiten (oder Mittelwerte) δ

Chiquaret χ^2

Korrelationskoeffizient ρ

Regressionskoeffizienten β_1, β_2

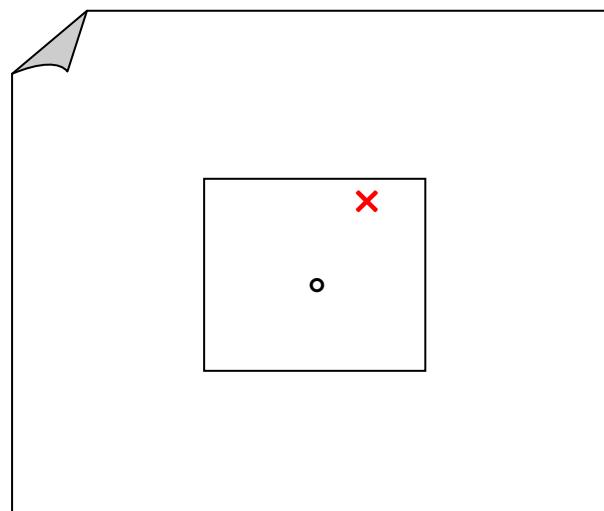
Punktschätzer (\circ) in der Nähe der Parameter (\times)?



Intervallschätzung:

Idee: Intervall, das mit einer Wahrscheinlichkeit $1-\alpha$ den unbekannten Parameter überdeckt.

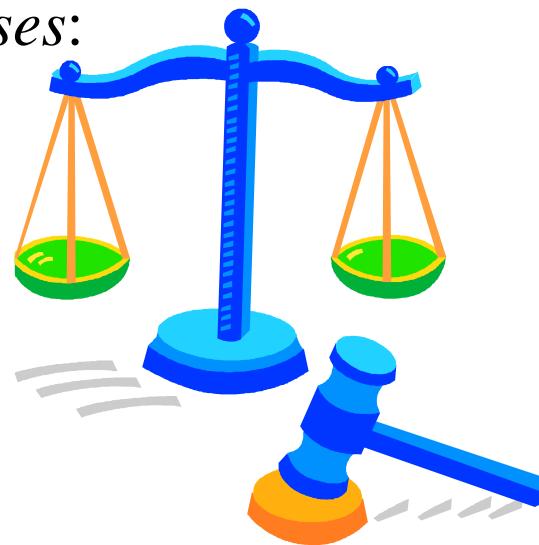
Überdeckungswahrscheinlichkeit $1-\alpha$



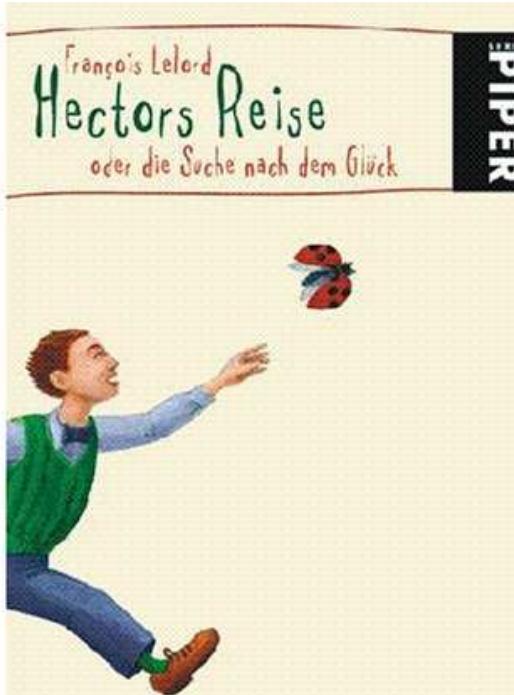
Testen von Hypothesen:

Aufgabe: Treffen einer fundierten Entscheidung zwischen zwei konkurrierenden Unterstellungen (Beispiel: Ist eine Mehrheit EU-skeptisch oder nicht?)

Handlungslogik eines *Indizienprozesses*:



	Indizienprozess	Statistisches Testen von Hypothesen
Zu prüfende Hypothese (Einshypothese):	Schuld	Forschungshypothese
Ausgangshypothese (Nullhypothese):	Unschuld	Gegenteil der zu überprüfenden Hypothese
Sammlung von Indizien gegen die Nullhypothese:	Zeugeneinvernahmen etc.	Stichprobenerhebung
Entscheidung:	Bei starken Indizien: Verurteilung, sonst Freispruch	Bei starken Indizien: Akzeptieren der Einshypothese , sonst Beibehaltung der Nullhypothese
Fehlermöglichkeiten	Justizirrtum, irrtümlicher Freispruch	α - Fehler (Signifikanzniveau), β -Fehler



„... so ist Wissenschaft eben: Es reicht nicht, wenn man irgendwas denkt, man muss versuchen nachzuprüfen, ob es auch stimmt. Wenn nicht, könnte ja alle Welt sonst was denken und behaupten, und wenn es Leute behaupteten, die gerade in Mode waren, würde man ihnen glauben.“ (Francois Lelord, *Hectors Reise oder die Suche nach dem Glück*, S.153)

Aufgabe der Statistik: Festlegung jener Schranken, die die schwachen von den starken Indizien gegen die Nullhypothese trennen

3.4 Schätzen und Testen einer relativen Häufigkeit

3.4.1 Schätzen einer relativen Häufigkeit

Aufgabe: Schätzung einer unbekannten relativen Häufigkeit π in der Grundgesamtheit

Punktschätzer für π : Die relative Häufigkeit p in einer uneingeschränkten Zufallsstichprobe.

Intervallschätzung: Konstruktion eines **Konfidenzintervalls**, das den Parameter π mit einer Wahrscheinlichkeit $1-\alpha$ (*zumeist 95 %*) überdeckt.

Relative Häufigkeit $\pi = h/N$

Häufigkeiten in der Stichprobe sind *hypergeometrisch* verteilt mit Erwartungswert und theoretischer Varianz

$$\mu = n \cdot \pi \quad \text{und} \quad \sigma^2 = n \cdot \pi \cdot (1 - \pi) \cdot \frac{N - n}{N - 1} \quad (12a) \text{ und } (12b)$$

Relative Häufigkeiten in der Stichprobe $p = \frac{h}{n}$ besitzen Erwartungswert und theoretische Varianz

$$\mu = \pi \quad \text{und} \quad \sigma^2 = \frac{\pi \cdot (1 - \pi)}{n} \cdot \frac{N - n}{N - 1}$$

Rolle der Größe N der Grundgesamtheit

$$\text{Große Grundgesamtheiten (}N \rightarrow \infty\text{): } \sigma^2 = \frac{\pi \cdot (1 - \pi)}{n} \cdot \frac{N - n}{\underbrace{N - 1}_{\approx 1}} \approx \frac{\pi \cdot (1 - \pi)}{n}$$

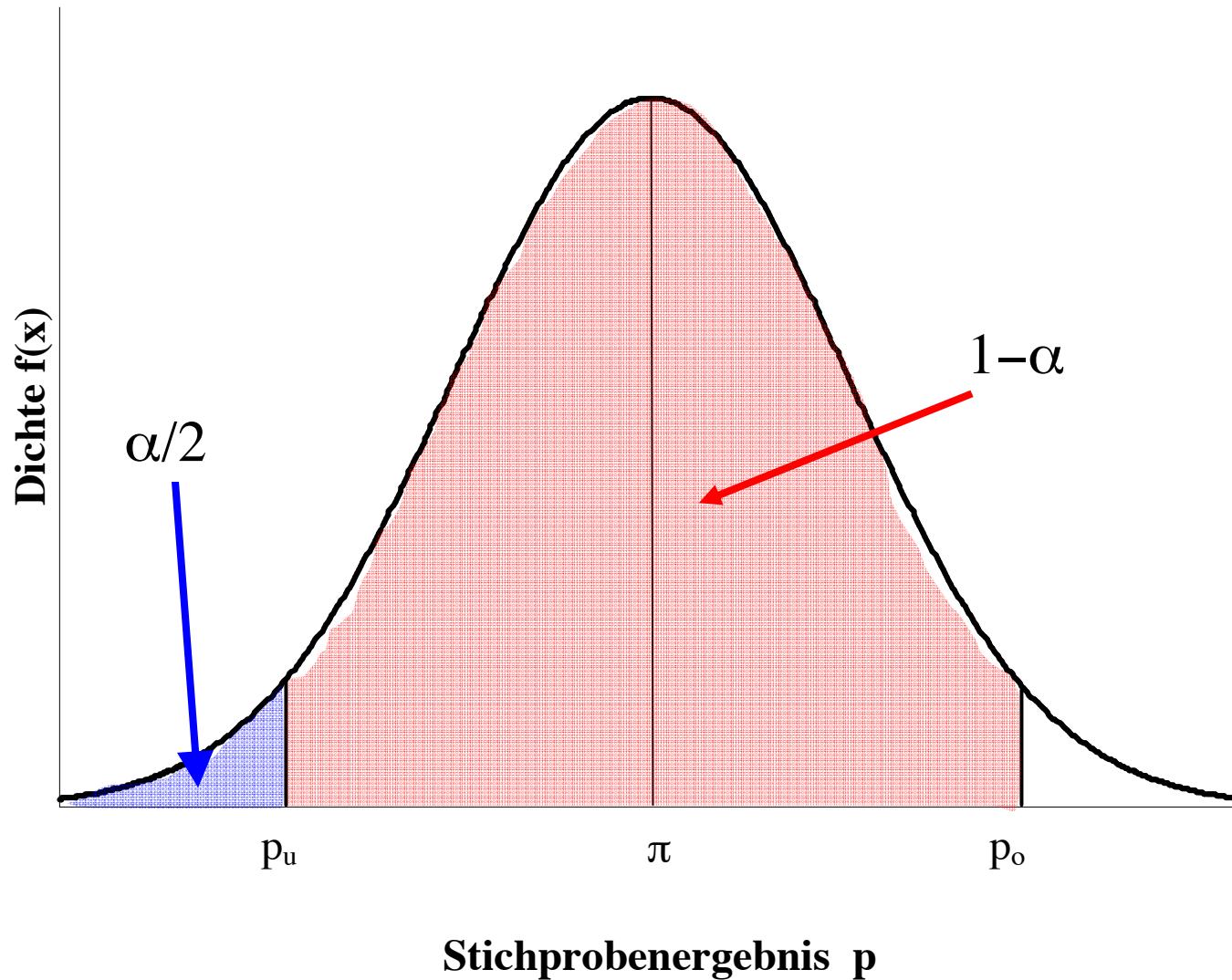
Zentraler Grenzwertsatz der Statistik:

Bei großem n (Faustregel: $n \geq 100$) sind relative Häufigkeiten p annähernd *normalverteilt*.

→ Rechnen mit der Normalverteilung:

Jenes symmetrische Intervall $[p_u; p_o]$, in dem die möglichen Stichprobenergebnisse p mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ liegen

Abbildung 43: Die (annähernde) Stichprobenverteilung relativer Häufigkeiten für große Stichproben und große Grundgesamtheiten



$$(14): u_0 = \frac{x_0 - \mu}{\sigma}$$

Obere Schranke p_o : $u_{1-\alpha/2} = \sqrt{\frac{\pi \cdot (1 - \pi)}{n} \cdot \frac{N - n}{N - 1}}$ in großen Stichproben aus kleinen Grundgesamtheiten

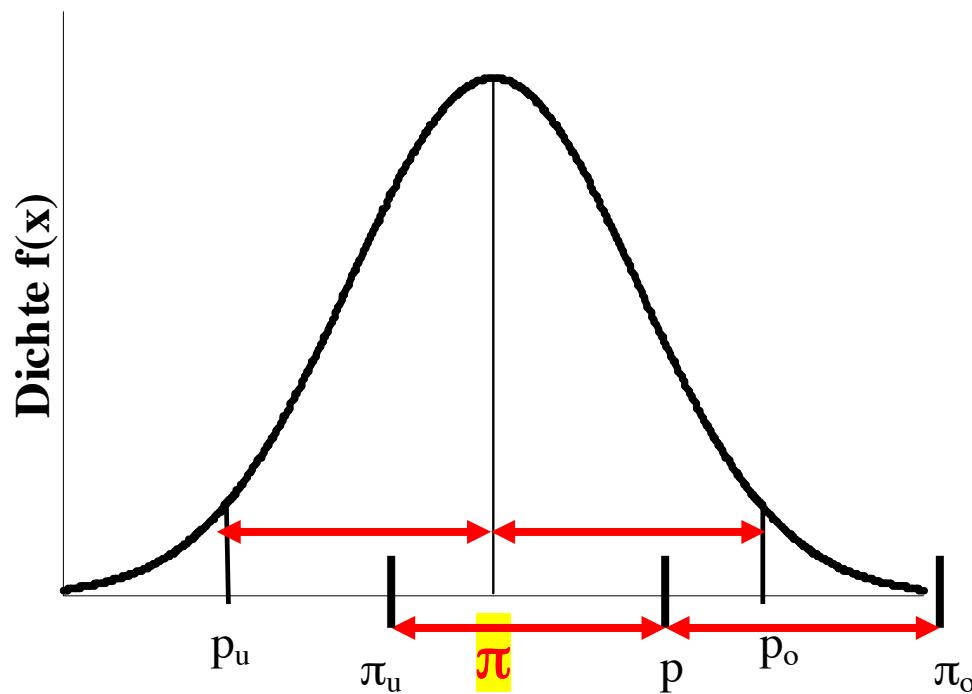
Große Grundgesamtheiten und große Stichproben (Faustregel : $n \geq 100$):

$$p_o = \pi + u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} \quad (15a)$$

$$p_u = \pi - u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} \quad (15b)$$

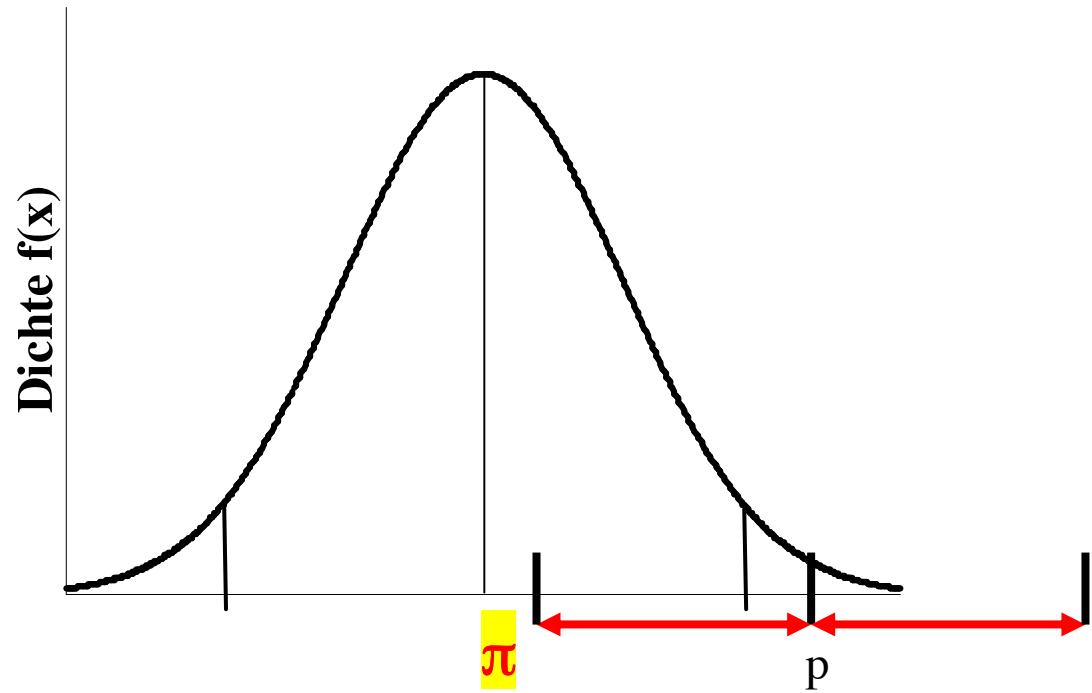
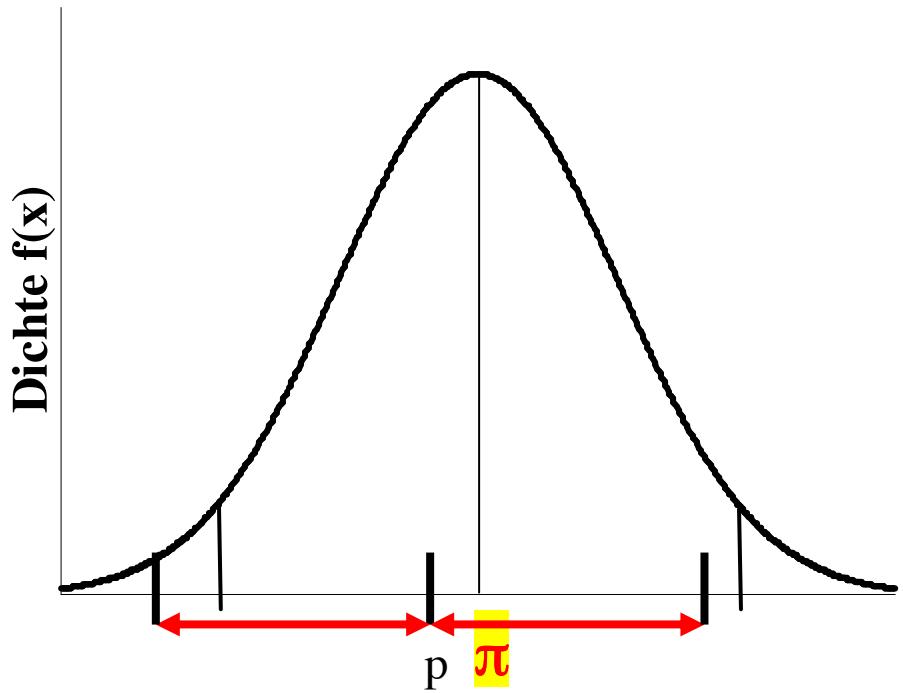
Aussage über Wahrscheinlichkeiten von *Stichprobenergebnissen*

Eigentliche Fragestellung: Aussage über den Parameter



$$p_o = \pi \pm u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}}$$

(15a,b)



$$p_o = \pi \pm u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} \quad (15a,b)$$

π wird durch p ersetzt:

$$\begin{aligned} \pi_o &= p + u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1 - p)}{n}} \\ \pi_u &= p - u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1 - p)}{n}} \end{aligned} \quad (16)$$

... Konfidenzintervall zur Sicherheit $1 - \alpha$ („Parameter π wird mit Wahrscheinlichkeit $1 - \alpha$ überdeckt“)

Beispiel 27: Konfidenzintervall für die relative Häufigkeit

Zufallsstichprobe: 23 % von 400 Befragten sehen mit Zuversicht der wirtschaftlichen Zukunft entgegen

Punktschätzer für π : $p = 0,23$.

Konfidenzintervall zur Sicherheit $1-\alpha = 0,95$:

$$\pi_o = p + u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = 0,23 + 1,96 \cdot \sqrt{\frac{0,23 \cdot (1-0,23)}{400}} = \underline{\underline{0,271}}$$

$$\pi_u = p - u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = 0,23 - 1,96 \cdot \sqrt{\frac{0,23 \cdot (1-0,23)}{400}} = \underline{\underline{0,189}}$$

„Die relative Häufigkeit an Zuversichtlichen in der Grundgesamtheit wird mit einer Wahrscheinlichkeit von 0,95 vom Intervall [0,189; 0,271] überdeckt.“

Beispiel 28: Konfidenzintervall für die relative Häufigkeit

Zufallsstichprobe: 23 % von 10.000 Befragten sehen mit Zuversicht der wirtschaftlichen Zukunft entgegen

Der *Punktschätzer* für π beträgt $p = 0,23$.

Konfidenzintervall zur Sicherheit $1-\alpha = 0,95$:

$$\pi_o = p + u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = 0,23 + 1,96 \cdot \sqrt{\frac{0,23 \cdot (1-0,23)}{10.000}} = 0,238$$

$$\pi_u = p - u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = 0,23 - 1,96 \cdot \sqrt{\frac{0,23 \cdot (1-0,23)}{10.000}} = 0,222$$



Erforderlicher Stichprobenumfang in großen Grundgesamtheiten:

(15a) und (15b): Aussage über Stichprobenergebnisse

$$p_o = \pi + u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} \quad (15a)$$

$$p_u = \pi - u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} \quad (15b)$$

Schwankungsbreite ε

$$\varepsilon = u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} \rightarrow$$

$$n = \frac{u^2}{\varepsilon^2} \cdot \pi \cdot (1 - \pi) \quad (17)$$

Erforderlicher Stichprobenumfang in allen Grundgesamtheiten:

$$p_o = \pi + u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n} \cdot \frac{N - n}{N - 1}}$$

$$p_u = \pi - u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n} \cdot \frac{N - n}{N - 1}}$$

Schwankungsbreite ε

$$\varepsilon = u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n} \cdot \frac{N - n}{N - 1}} \rightarrow$$

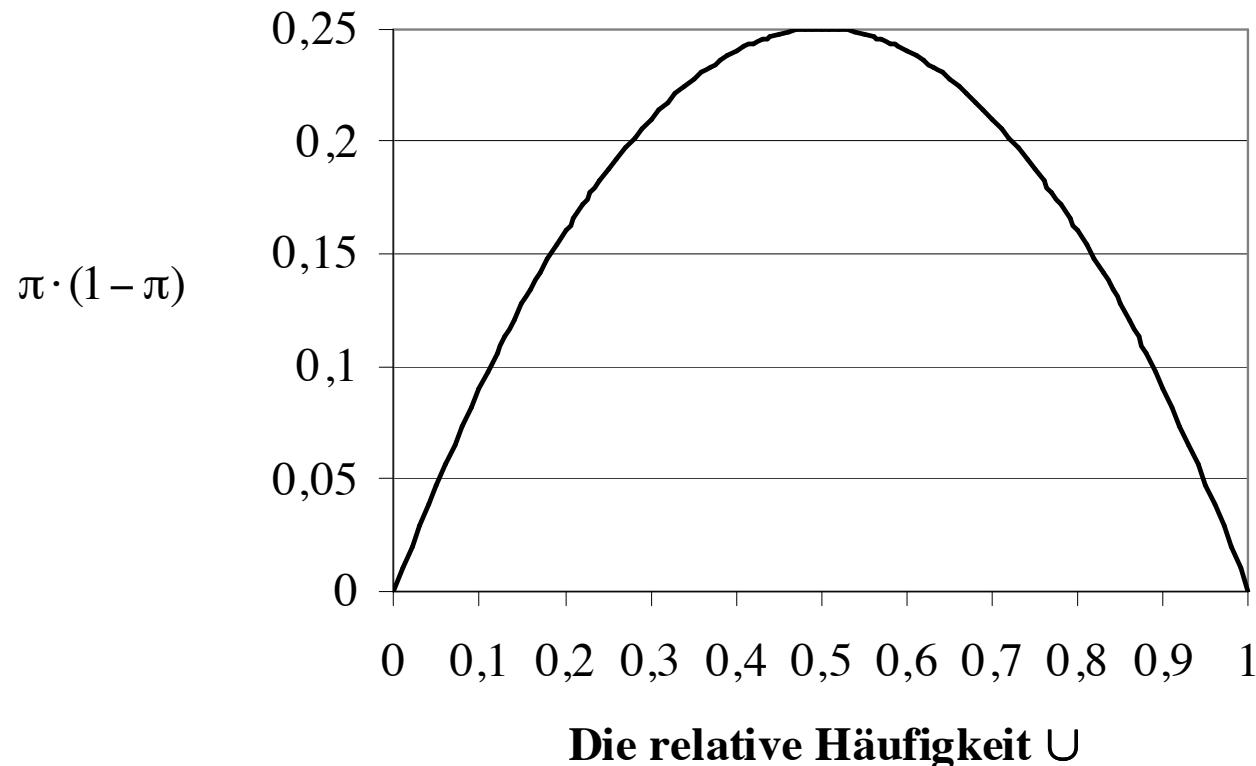
$$n = \frac{N \cdot u_{1-\alpha/2}^2 \cdot \pi \cdot (1 - \pi)}{(N - 1) \cdot \varepsilon^2 + u_{1-\alpha/2}^2 \cdot \pi \cdot (1 - \pi)}$$

Festzulegen:

- Überdeckungswahrscheinlichkeit $1-\alpha$
- Schwankungsbreite ε
- Parameter π

$$n = \frac{u^2}{\varepsilon^2} \cdot \pi \cdot (1 - \pi)$$

Abbildung 44: Der Verlauf der Funktion $\pi \cdot (1-\pi)$

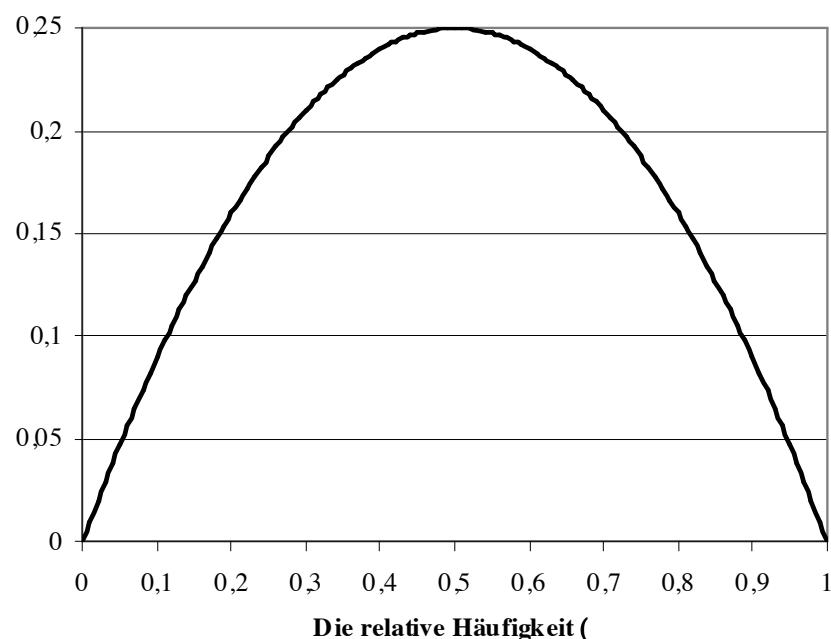


Beispiel 29: Berechnung des erforderlichen Stichprobenumfangs

Stichprobenumfang für Konfidenzintervall zur Sicherheit $1 - \alpha = 0,95$ mit Schwankungsbreite $\varepsilon = 0,02$ und

- a) Partei etwa 15 % der Stimmen
- b) Partei zwischen 15 und 25 % der Stimmen
- c) Partei zwischen 40 und 55 % der Stimmen
- d) keinerlei Abschätzung des Stimmenanteils

$$n = \frac{u^2}{\varepsilon^2} \cdot \pi \cdot (1 - \pi)$$



- a) $n = \frac{1,96^2}{0,02^2} \cdot 0,15 \cdot (1 - 0,15) = 1.225$
- b) $n = \frac{1,96^2}{0,02^2} \cdot 0,25 \cdot (1 - 0,25) = 1.801$
- c) $n = \frac{1,96^2}{0,02^2} \cdot 0,5 \cdot (1 - 0,5) = 2.401$
- d) wie c)

Beispiel: $N = 10.000$

Stichprobenumfang für Konfidenzintervall zur Sicherheit $1 - \alpha = 0,95$ mit Schwankungsbreite $\varepsilon = 0,02$ und Partei etwa 15 % der Stimmen

$$n = \frac{N \cdot u_{1-\alpha/2}^2 \cdot \pi \cdot (1 - \pi)}{(N - 1) \cdot \varepsilon^2 + u_{1-\alpha/2}^2 \cdot \pi \cdot (1 - \pi)}$$

$$n = \frac{10.000 \cdot 1,96^2 \cdot 0,15 \cdot (1 - 0,15)}{9.999 \cdot 0,02^2 + 1,96^2 \cdot 0,15 \cdot (1 - 0,15)} = 1.092$$

3.4.2 Testen von Hypothesen über eine relative Häufigkeit

Aufgabe: Treffen einer fundierten Entscheidung zwischen zwei konkurrierenden Unterstellungen (=Hypothesen) über eine relative Häufigkeit in der Grundgesamtheit

Allgemeine Handlungslogik für das statistische Testen von Hypothesen:

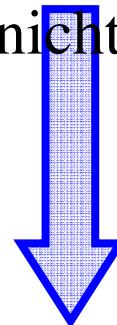
- Aufstellen von Eins- und Nullhypothese
- Sammlung von Indizien gegen Nullhypothese
- Entscheidung: Entweder Beibehaltung der Null- oder Akzeptierung der Einshypothese

Beispiel 30: Statistisches Testen von zweiseitigen Hypothesen

EU: 20 % Zuversichtliche

Auf dem Signifikanzniveau $\alpha = 0,05$ Überprüfung, ob in einem Land die relative Häufigkeit der Zuversichtlichen nicht mit dem EU-weiten Wert übereinstimmt.

Stichprobe: $n = 400$, $p = 0,23$



Hypothesenformulierung:

$$H_0: \pi = 0,2 \quad \text{und} \quad H_1: \pi \neq 0,2$$

Einshypothese H_1



... zweiseitige Fragestellung

Hat man mit $p = 0,23$ ein *starkes Indiz* gegen H_0 gefunden?

Wäre $p = 0,9; 0,6; 0,3; 0,25; 0,23; 0,21$ ein starkes Indiz gegen H_0 ?

$$H_0: \pi = 0,2 \quad \text{und} \quad H_1: \pi \neq 0,2$$

Bereich der schwachen Indizien gegen die Nullhypothese (=Beibehaltungsregion der Nullhypothese) nach (15a,b):

$$p_o = \pi + u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} = 0,2 + 1,96 \cdot \sqrt{\frac{0,2 \cdot (1 - 0,2)}{400}} = 0,239$$

$$p_u = \pi - u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} = 0,2 - 1,96 \cdot \sqrt{\frac{0,2 \cdot (1 - 0,2)}{400}} = 0,161$$

$p = 0,23 \in [0,161;0,239] \rightarrow$ Schwaches Indiz gegen $H_0 \rightarrow$ Beibehaltung der Nullhypothese; Testergebnis ist **nicht signifikant**

Die **zweiseitige Fragestellung** in allgemeiner Darstellung:

$$H_0: \pi = \pi_0 \quad \text{und} \quad H_1: \pi \neq \pi_0$$

Entscheidungsregel: Beibehaltung von H_0 , wenn gilt: $p \in [p_u; p_o]$

Abbildung 45: Beibehaltung der Nullhypothese beim zweiseitigen Test

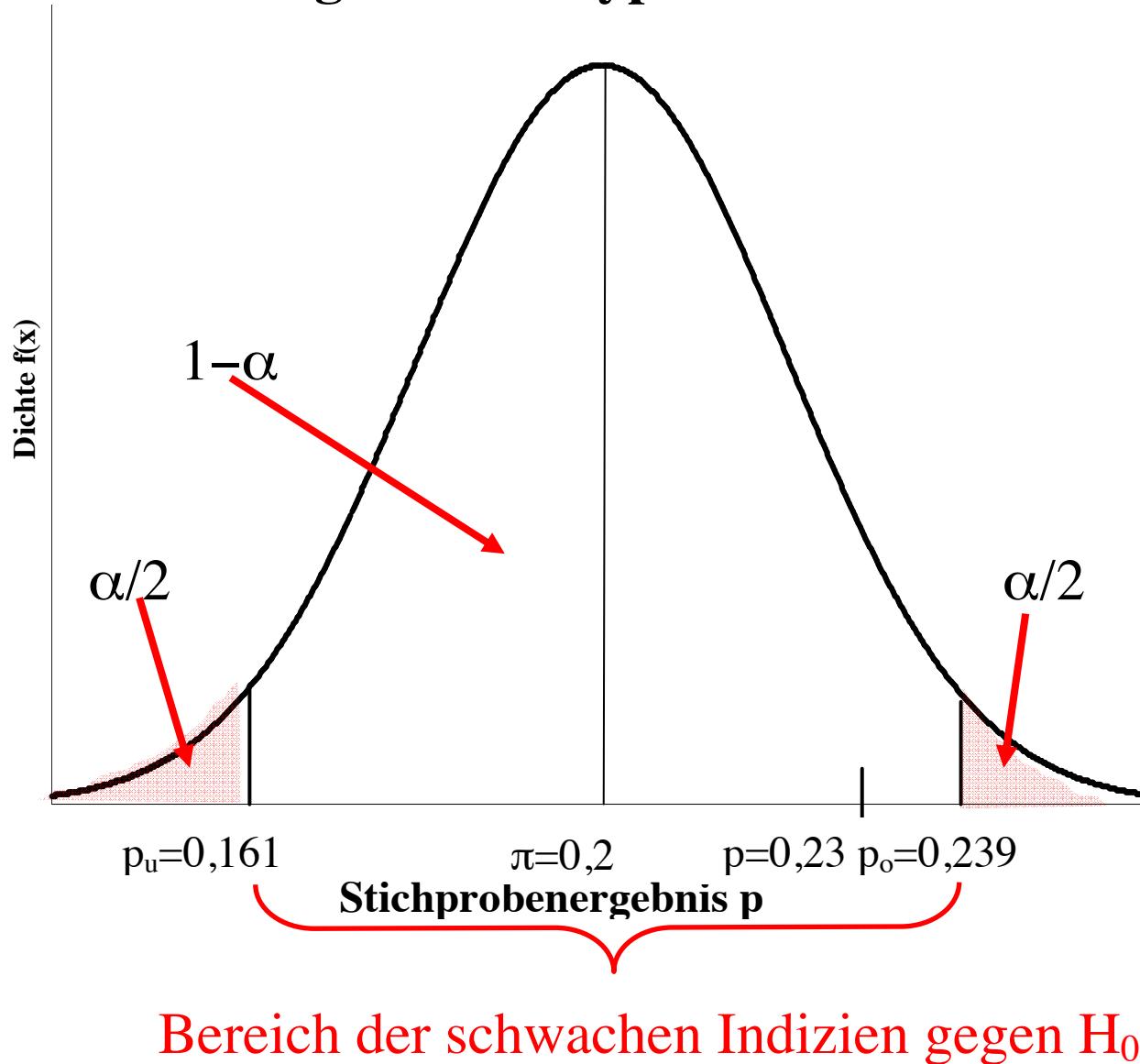
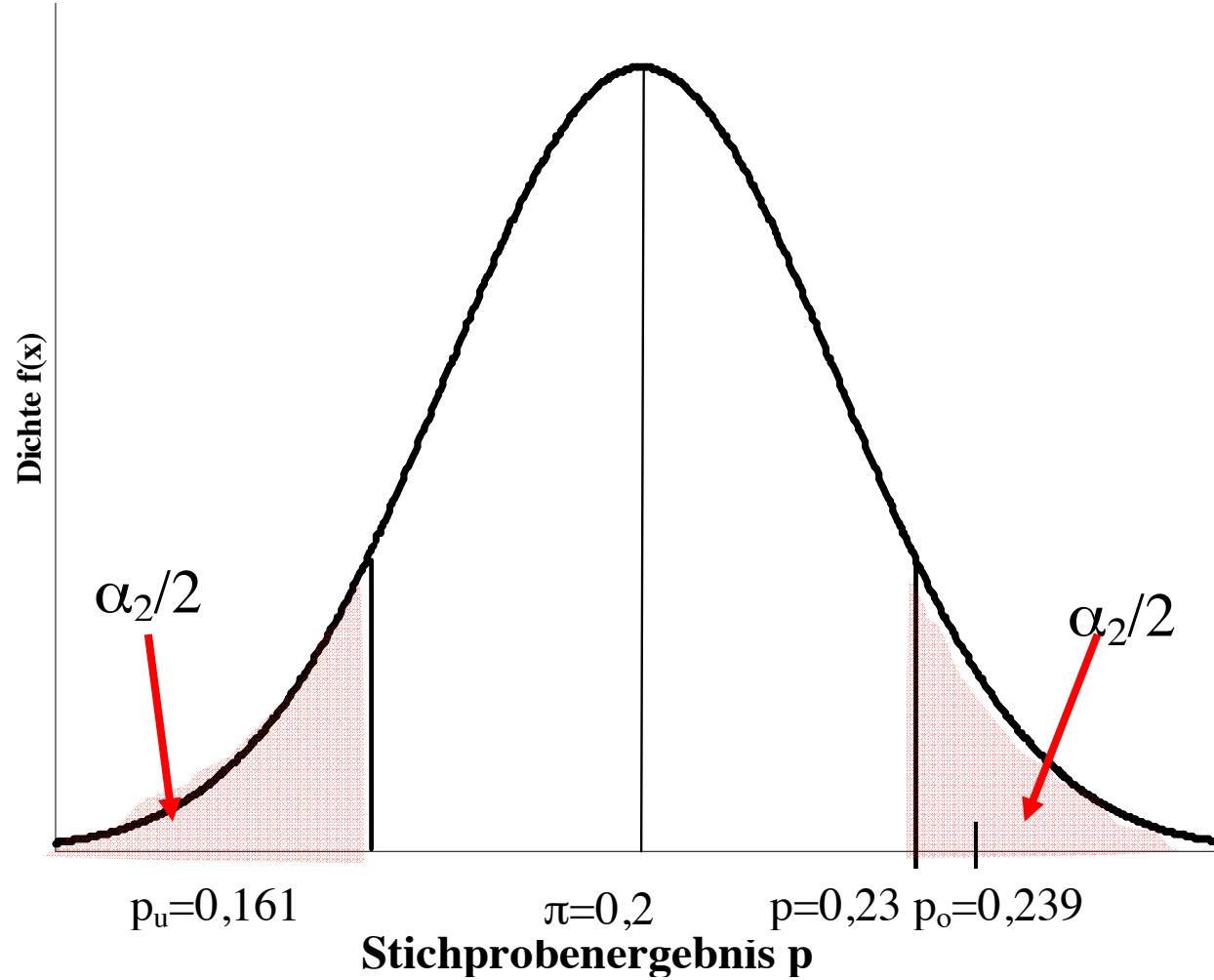


Abbildung 46: Entscheidung durch Verwendung des p-Wertes beim zweiseitigen Test



Beibehaltung der Nullhypothese: p -Wert $\alpha_2 >$ Signifikanzniveau α .

In Beispiel 30: $\alpha_2 = 0,134 > 0,05 \rightarrow$ Beibehaltung der Nullhypothese!

Signifikanzniveau α ist immer *vor* der Untersuchung festlegen

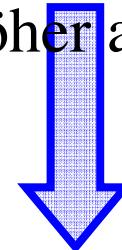
Konvention: zumeist $\alpha = 0,05$

Beispiel 31: Testen von einseitigen Hypothesen

EU: 20 % Zuversichtliche

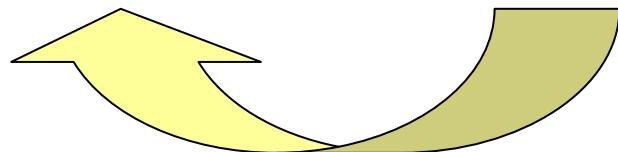
Auf dem Signifikanzniveau $\alpha = 0,05$ Überprüfung, ob in einem Land die relative Häufigkeit der Zuversichtlichen höher als der EU-weite Wert ist.

Stichprobe: $n = 400$, $p = 0,23$



Hypothesenformulierung:

$$H_0: \pi \leq 0,2 \quad \text{und} \quad H_1: \pi > 0,2 \quad \text{Einshypothese } H_1$$

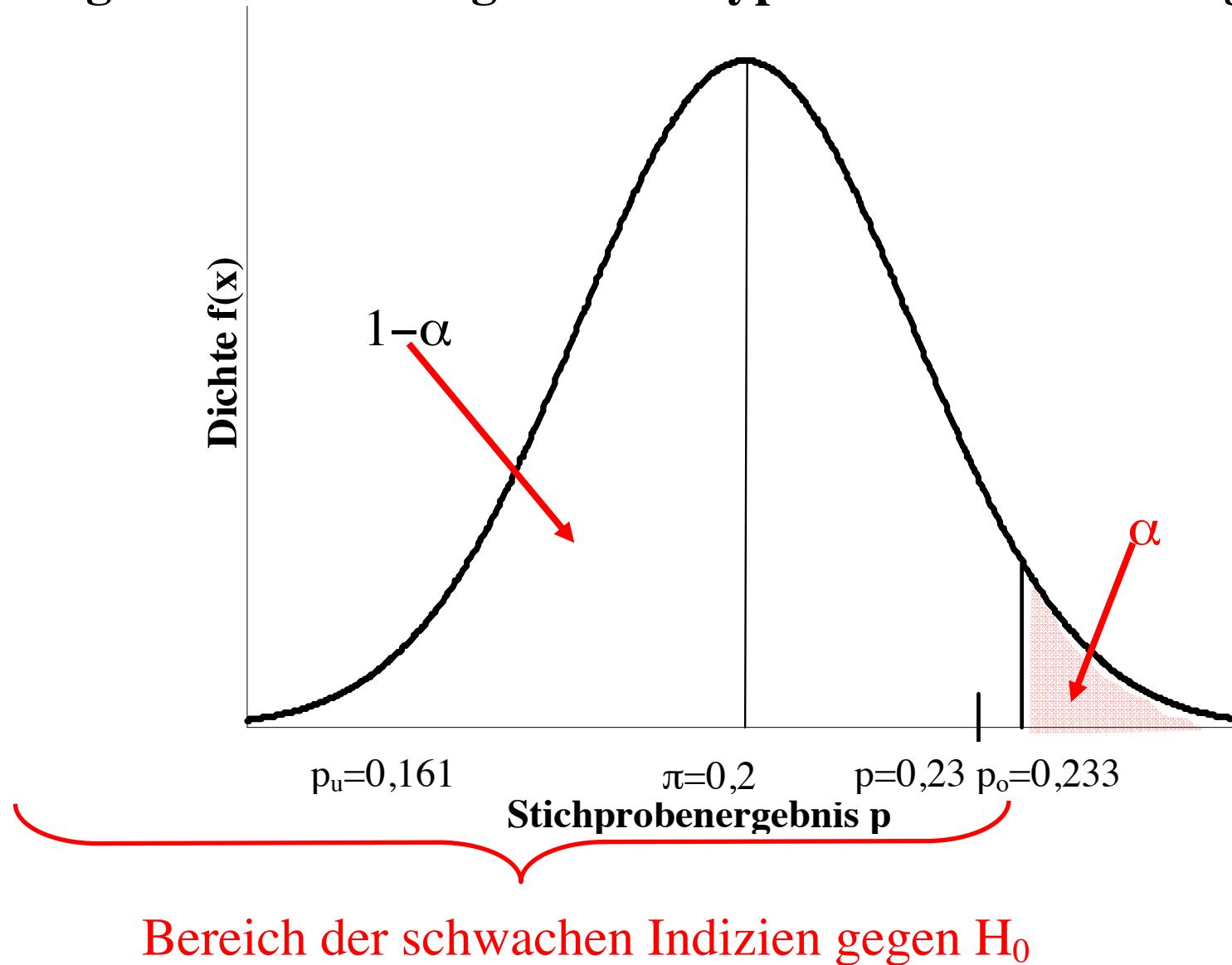


Hat man mit $p = 0,23$ ein starkes Indiz gegen H_0 gefunden?

Beispiele für relative Häufigkeiten p , die ein starkes Indiz gegen H_0 sind

Obere Schranke der schwachen Indizien gegen die Nullhypothese

Abbildung 47: Beibehaltung der Nullhypothese beim einseitigen Test



Obere Schranke der schwachen Indizien gegen die Nullhypothese:

$$p_o = \pi + u_{1-\alpha} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} = 0,2 + 1,65 \cdot \sqrt{\frac{0,2 \cdot (1 - 0,2)}{400}} = 0,233$$

$p = 0,23 \leq 0,233 \rightarrow$ Schwaches Indiz gegen $H_0 \rightarrow$ Beibehaltung der Nullhypothese

Testergebnis ist nicht signifikant

Abbildung 47: Beibehaltung der Nullhypothese beim einseitigen Test

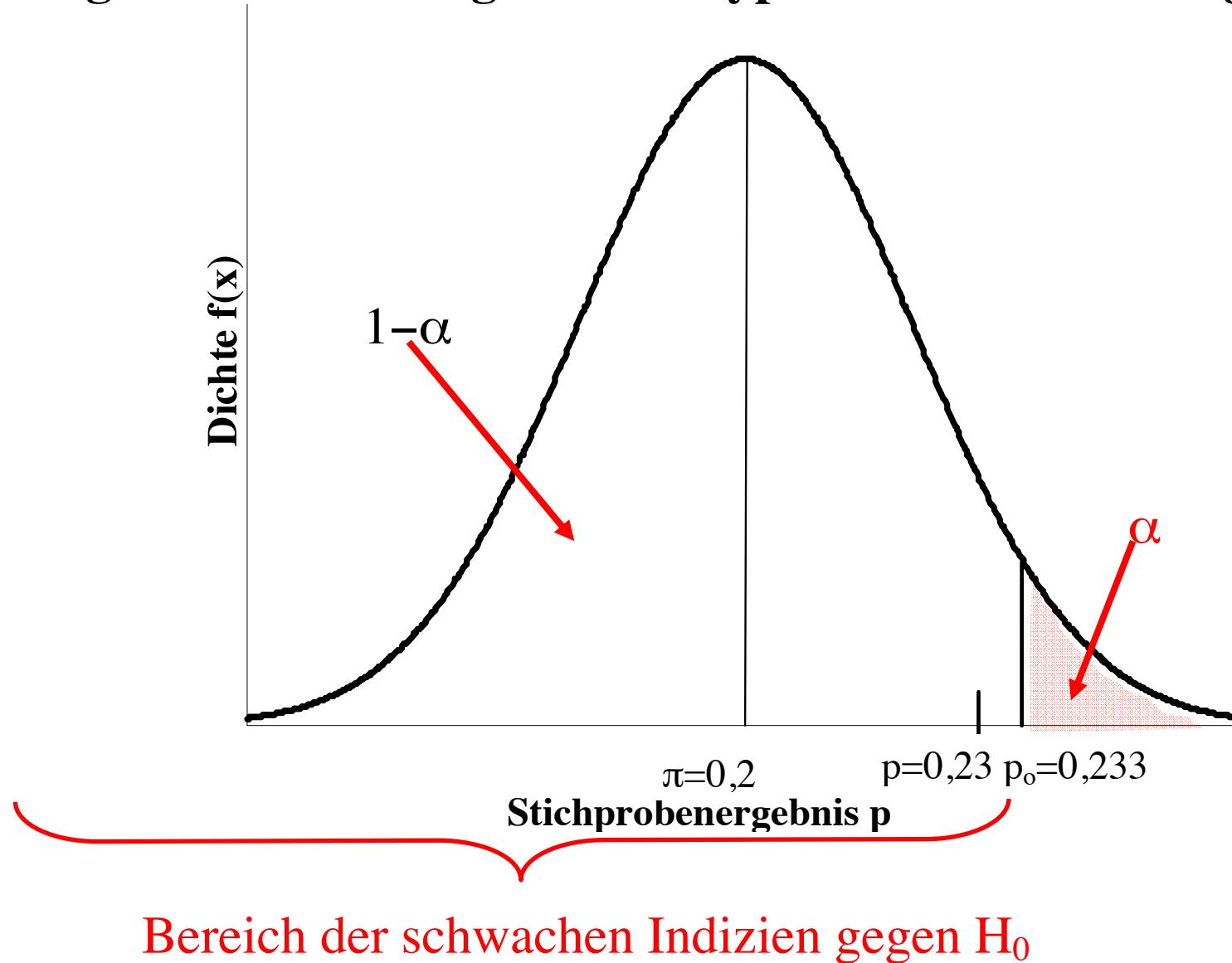
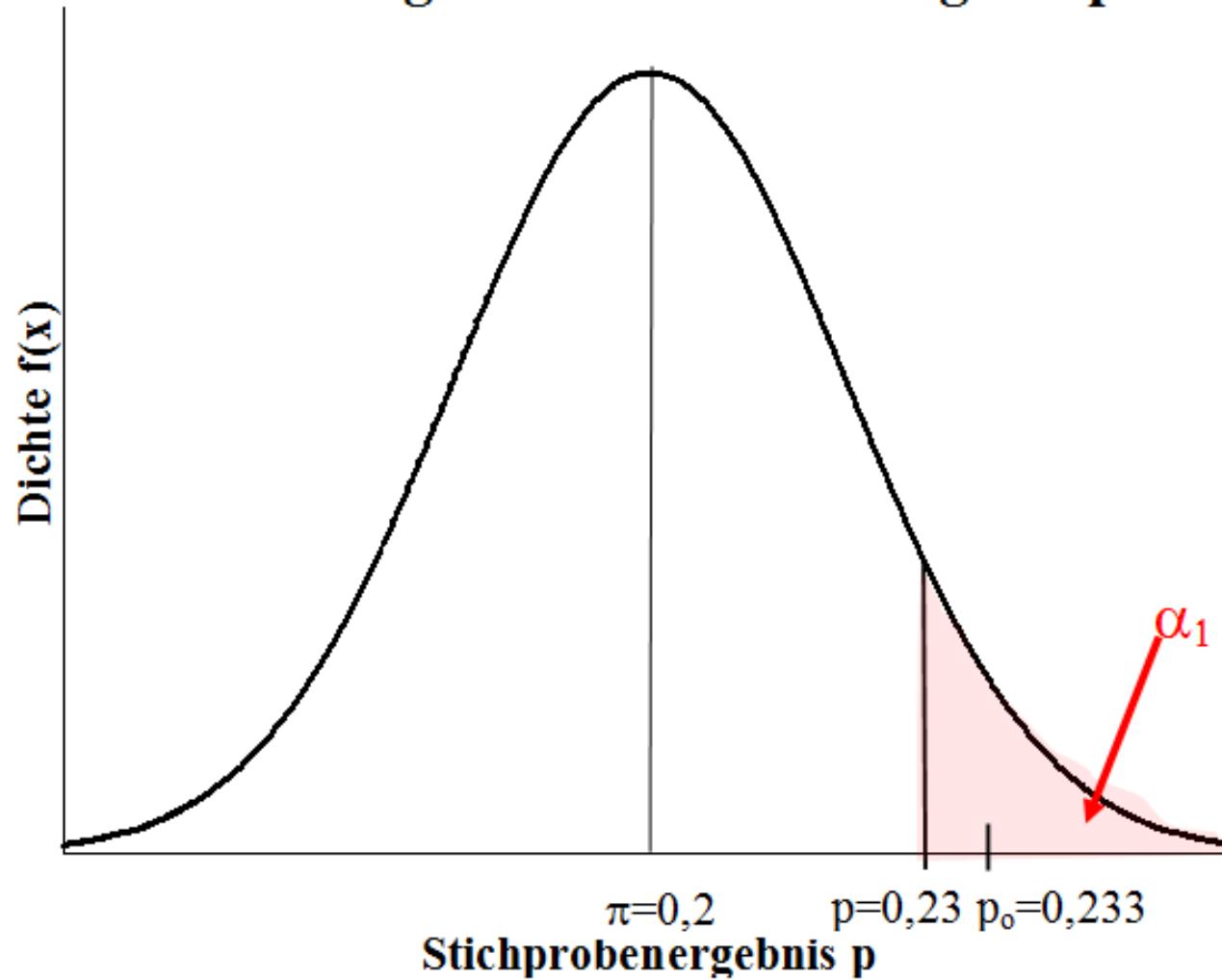


Abbildung 48: Entscheidung durch Verwendung des p-Wertes



Beibehaltung der Nullhypothese: p-Wert $\alpha_1 >$ Signifikanzniveau α .

Beispiel 31: $\alpha_1 = 0,067 > 0,05 \rightarrow$ Beibehaltung der Nullhypothese!

Beispiel für einen SPSS-Output:

Test auf Binomialverteilung					
	Kategorie	N	Beobachteter Anteil	Testanteil	Exakte Signifikanz (1-seitig)
Zuversichtlich	Gruppe 1 ja	92	,230	,2	,067
	Gruppe 2 nein	308	,770		
	Gesamt	400	1,000		

Beziehung der ein- und zweiseitigen p-Werte α_1 und α_2 :

$$\alpha_1 = \frac{\alpha_2}{2}$$

Beispiel 32: Testen von einseitigen Hypothesen

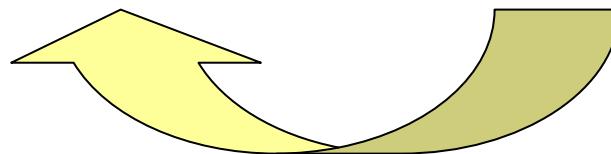
EU: 20 % Zuversichtliche

Auf dem Signifikanzniveau $\alpha = 0,05$ Überprüfung, ob in einem Land die relative Häufigkeit der Zuversichtlichen geringer als in der EU ist.



Hypothesenformulierung:

$$H_0: \pi \geq 0,2 \quad \text{und} \quad H_1: \pi < 0,2 \quad \text{Einhypothese } H_1$$



Untere Schranke der schwachen Indizien gegen H_0 :

$$p_u = \pi - u_{1-\alpha} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}}$$

Starkes Indiz gegen die Nullhypothese, wenn $p < p_u$

Die **einseitige Fragestellung** in allgemeiner Darstellung:

Es gibt zwei Arten von einseitigen Fragestellungen:

- a) $H_0: \pi \leq \pi_0$ und $H_1: \pi > \pi_0$
- b) $H_0: \pi \geq \pi_0$ und $H_1: \pi < \pi_0$

Entscheidungsregeln:

- a) H_0 wird beibehalten, wenn $p \leq p_o$.
- b) H_0 wird beibehalten, wenn $p \geq p_u$.

3.5 Schätzen und Testen eines Mittelwerts

3.5.1 Schätzen eines Mittelwerts

Aufgabe: Schätzung eines unbekannten Mittelwerts μ in der Grundgesamtheit

Punktschätzer für μ : Der Mittelwert \bar{x} in einer uneingeschränkten Zufallsstichprobe

Intervallschätzung: Konstruktion eines **Konfidenzintervalls**, das den Parameter μ mit einer Wahrscheinlichkeit $1-\alpha$ überdeckt.

Zentraler Grenzwertsatz der Statistik:

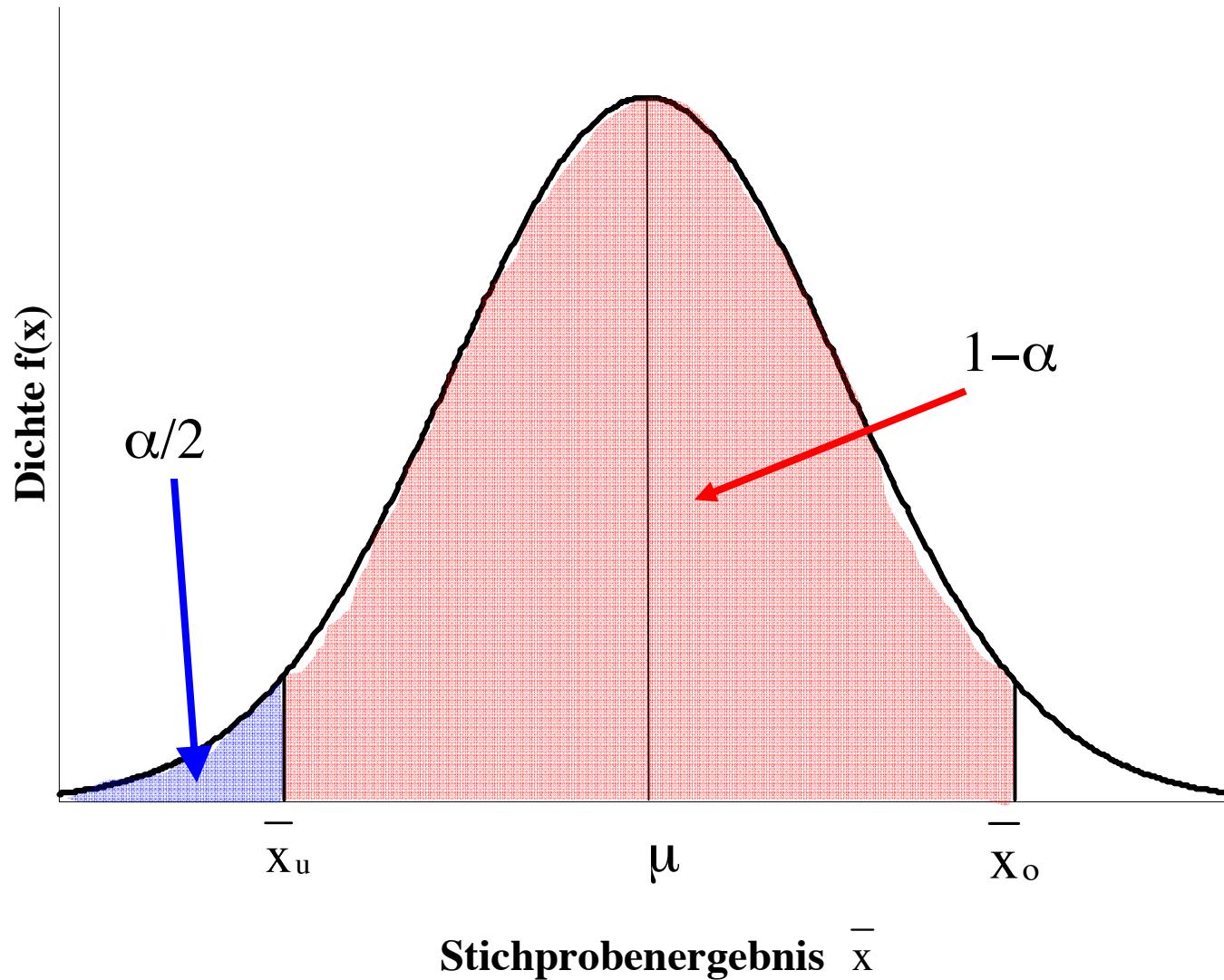
Bei großem n (Faustregel: $n \geq 100$) sind Mittelwerte \bar{x} annähernd *normalverteilt* mit dem Erwartungswert μ und der theoretischen Varianz $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$.

in großen Grundgesamtheiten ≈ 1

→ Rechnen mit der Normalverteilung:

Symmetrisches Intervall $[\bar{x}_u; \bar{x}_o]$, in dem die möglichen Stichprobenergebnisse \bar{x} mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ liegen

Abbildung 49: Die annähernde Stichprobenverteilung von Mittelwerten für große Stichprobenumfänge



(14): $u_0 = \frac{x_0 - \mu}{\sigma}$ beim Merkmal x ; beim Stichprobenmittelwert \bar{x} gilt für

die obere Schranke \bar{x}_o : $u_{1-\alpha/2} \approx \frac{\bar{x}_o - \mu}{\sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}}}$ in großen Stichproben aus kleinen Grundgesamtheiten

Große Grundgesamtheiten und *große* Stichproben (Faustregel : $n \geq 100$):

$$\begin{aligned}\bar{x}_o &= \mu + u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}} \\ \bar{x}_u &= \mu - u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}\end{aligned}\tag{18}$$

Aussage über Wahrscheinlichkeiten von *Stichprobenergebnissen*

Eigentliche Fragestellung: Aussage über den *Parameter*

Gleiche Überlegungen wie beim Konfidenzintervall für relative Häufigkeiten π nach (16):

$$\mu_o = \bar{x} + u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}$$

$$\mu_u = \bar{x} - u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}$$

Die Varianz σ^2 der Grundgesamtheit durch die Stichprobenvarianz s^2 ersetzen:

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 \quad (19)$$

s^2 schätzt σ^2 durchschnittlich richtig (*Unverzerrtheit*)



Konfidenzintervall zur Sicherheit $1-\alpha$ für μ :

$$\begin{aligned}\mu_o &= \bar{x} + u_{1-\alpha/2} \cdot \sqrt{\frac{s^2}{n}} \\ \mu_u &= \bar{x} - u_{1-\alpha/2} \cdot \sqrt{\frac{s^2}{n}}\end{aligned}\tag{20}$$

(„Parameter μ wird mit Wahrscheinlichkeit $1-\alpha$ überdeckt“)

Kleine Stichproben: Die standardisierten \bar{x} sind bei normalverteiltem Merkmal eigentlich **t-verteilt** (nicht normalverteilt)

Statt $u_{1-\alpha/2}$ nehme man den Wert der t-Verteilung bei $n-1$ „Freiheitsgraden“

t-Verteilung:

FG: n-1	$t_{0,95;n-1}$	$t_{0,975;n-1}$
10	1,812	2,228
20	1,725	2,086
30	1,697	2,042
40	1,684	2,021
50	1,676	2,009
100	1,660	1,984
150	1,655	1,976
200	1,653	1,972
300	1,650	1,968
400	1,649	1,966
500	1,648	1,965
1000	1,646	1,962
2000	1,646	1,961
5000	1,645	1,960
10000	1,645	1,960

Vergleiche:

$$u_{0,95}=1,645$$

$$u_{0,975}=1,96$$

Faustregel:

Für $n \geq 100$ darf
 $t \approx u$ gelten!

Beispiel 33: Konfidenzintervall für einen Mittelwert

Zufallsstichprobe: $n = 100$, $\bar{x} = 998$ und $s^2 = 2,56$

Punktschätzer für μ : $\bar{x} = 998$

Konfidenzintervall zur Sicherheit $1-\alpha = 0,95$ nach (20):

$$\mu_o = \bar{x} + u_{1-\alpha/2} \cdot \sqrt{\frac{s^2}{n}} = 998 + 1,96 \cdot \sqrt{\frac{2,56}{100}} = \underline{\underline{998,31}}$$

$$\mu_u = \bar{x} - u_{1-\alpha/2} \cdot \sqrt{\frac{s^2}{n}} = 998 - 1,96 \cdot \sqrt{\frac{2,56}{100}} = \underline{\underline{997,69}}$$

„Der Mittelwert μ des Merkmals x in der Grundgesamtheit wird mit einer Wahrscheinlichkeit von 0,95 von diesem Intervall überdeckt.“

3.5.2 Testen von Hypothesen über einen Mittelwert

Aufgabe: Treffen einer fundierten Entscheidung zwischen zwei konkurrierenden Unterstellungen über einen Mittelwert

Analoge Vorgehensweise zu den Überlegungen bei relativen Häufigkeiten:

Zweiseitige Fragestellung:

$$H_0: \mu = \mu_0 \quad \text{und} \quad H_1: \mu \neq \mu_0$$

Bereich der schwachen Indizien gegen die Nullhypothese (=Beibehaltungsregion der Nullhypothese) nach (18):

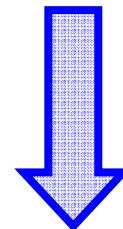
Entscheidungsregel: Beibehaltung von H_0 , wenn gilt: $\bar{x} \in [\bar{x}_u; \bar{x}_o]$

Beispiel 34: Testen von zweiseitigen Hypothesen

Normgewicht: $\mu = 1.000$ g

Überprüfung auf einem Signifikanzniveau $\alpha = 0,05$, ob das Durchschnittsgewicht der laufenden Produktion davon abweicht.

Stichprobe: $n = 100$; $\bar{x} = 998$; $s^2 = 2,56$



Einshypothese H_1

Hypothesenformulierung:

$$H_0: \mu = 1.000 \quad \text{und} \quad H_1: \mu \neq 1.000$$



Ist das Stichprobenergebnis ein starkes Indiz gegen H_0 ?

$$H_0: \mu = 1.000 \quad \text{und} \quad H_1: \mu \neq 1.000$$

$$s^2 = 2,56$$

Bereich der schwachen Indizien gegen die Nullhypothese (=Beibehaltungsregion der Nullhypothese) nach (18):

$$\bar{x}_o = \mu + u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}} = 1.000 + 1,96 \cdot \sqrt{\frac{2,56}{100}} = 1.000,31$$

$$\bar{x}_u = \mu - u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}} = 1.000 - 1,96 \cdot \sqrt{\frac{2,56}{100}} = 999,69$$

→ $\bar{x} = 998 \notin [999,69; 1.000,31]$ → Starkes Indiz gegen H_0 → Annahme der Einshypothese; Testergebnis ist signifikant

Kleine Stichproben: Standardisierte \bar{x} sind bei normalverteiltem x t-verteilt
→ „t-Test“

Alternative Entscheidungsregel: Beibehaltung der Nullhypothese, wenn der p-Wert $\alpha_2 >$ Signifikanzniveau α .

Einseitige Fragestellungen:

$$H_0: \mu \leq \mu_0 \quad \text{und} \quad H_1: \mu > \mu_0 \quad \text{bzw.} \quad H_0: \mu \geq \mu_0 \quad \text{und} \quad H_1: \mu < \mu_0$$

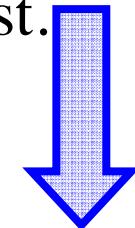
Eine Schranke der schwachen Indizien gegen H_0

Beispiel 35: Testen von einseitigen Hypothesen über einen Mittelwert

Normgewicht: $\mu = 1.000$ g

Überprüfung auf einem Signifikanzniveau $\alpha = 0,05$, ob das Durchschnittsgewicht der laufenden Produktion zu gering ist.

Stichprobe: $n = 100$; $\bar{x} = 999,62$; $s^2 = 14,44$



Einshypothese H_1

Hypothesenformulierung:

$$H_0: \mu \geq 1.000 \quad \text{und} \quad H_1: \mu < 1.000$$



Ist das Stichprobenergebnis ein starkes Indiz gegen H_0 ?

Untere Schranke der schwachen Indizien gegen die Nullhypothese:

$$\bar{x}_u \approx \mu - u_{1-\alpha} \cdot \sqrt{\frac{s^2}{n}} = 1.000 - 1,65 \cdot \sqrt{\frac{14,44}{100}} = 999,37$$

→ $\bar{x} = 999,62 \geq 999,37 \rightarrow$ schwaches Indiz gegen H_0 ($\bar{x} = 999,62$ liegt in Beibehaltungsregion der Nullhypothese) → H_0 wird beibehalten!

Das Testergebnis ist **nicht signifikant**.

Einseitiger Test in die andere Richtung: Obere Schranke \bar{x}_o

p-Wert: Beibehaltung der Nullhypothese, wenn $\alpha_1 > \alpha$

3.6 Testen von Hypothesen über zwei relative Häufigkeiten

Aufgabe: Fundierte Entscheidung zwischen zwei konkurrierenden Unterstellungen über den Vergleich zweier relativer Häufigkeiten zweier Grundgesamtheiten A und B

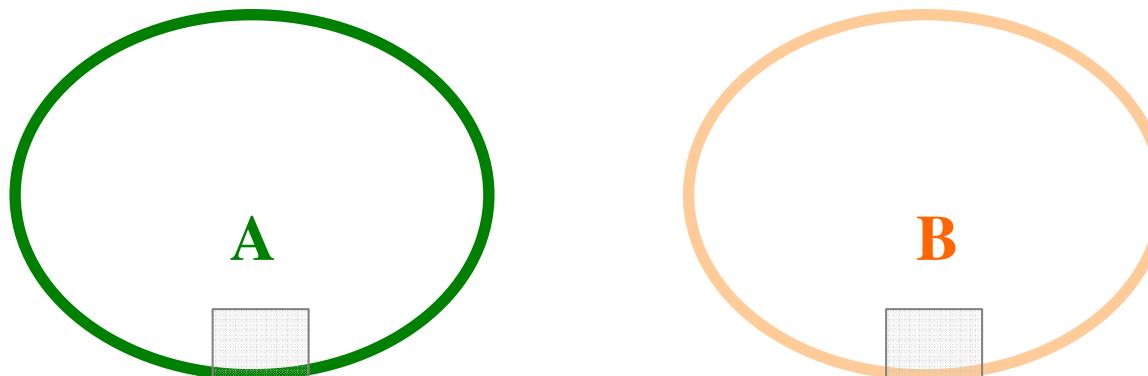
Beispiele:

Vergleich der relativen Häufigkeiten an Atomkraftgegnern in Österreich (A) und Deutschland (B)

Vergleich der relativen Häufigkeiten an Studienbeitragsbefürwortern an der JKU Linz (A) und der WU Wien (B)

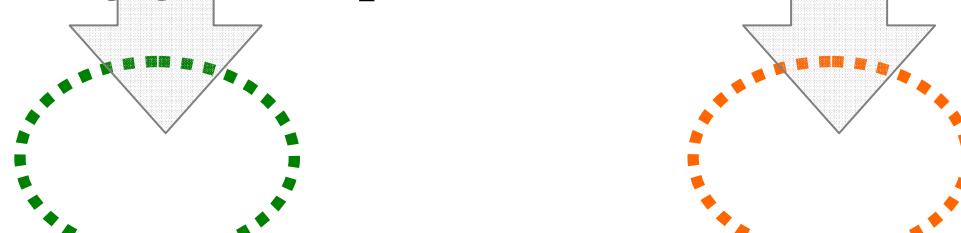
Vergleich der relativen Häufigkeiten an Anteilen einer Partei vor einem Monat (A) und heute (B)

Grundgesamtheiten A und B:



Parameter: Relative Häufigkeiten π_A und π_B und $\delta = \pi_A - \pi_B$

Unabhängige oder abhängige Stichproben aus A und B:



Stichprobenergebnisse: p_A und p_B und $d = p_A - p_B$ (Punktschätzer)

Bei zwei unabhängigen Stichproben:

Zentraler Grenzwertsatz der Statistik:

d ist in großen Stichproben aus großen Grundgesamtheiten näherungsweise normalverteilt mit dem Erwartungswert δ und der theoretischen Varianz

$$\frac{\pi_A \cdot (1 - \pi_A)}{n_A} + \frac{\pi_B \cdot (1 - \pi_B)}{n_B}.$$

→ Rechnen mit der Normalverteilung

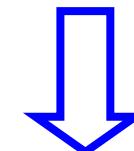
Wenn $\pi_A = \pi_B$: Schätzung von π_A und von π_B kann durch die gemeinsame relative Häufigkeit p über beide Stichproben erfolgen

$$p = \frac{h_A + h_B}{n_A + n_B} = \frac{n_A \cdot p_A + n_B \cdot p_B}{n_A + n_B}$$

Die Varianz von d wird dann geschätzt durch: $p \cdot (1 - p) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B} \right)$.

Zweiseitiger Test:

Überprüfung, ob $\pi_A \neq \pi_B$  Überprüfung, ob $\delta = \pi_A - \pi_B \neq 0$



$$H_0: \delta = 0 \quad \text{und} \quad H_1: \delta \neq 0$$

Einshypothese H_1

Bereich der schwachen Indizien gegen die Nullhypothese:

$$d_o = 0 + u_{1-\alpha/2} \cdot \sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B} \right)} \quad (21)$$

$$d_u = 0 - u_{1-\alpha/2} \cdot \sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}$$

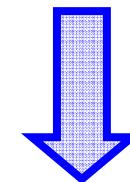
Entscheidungsregel: Beibehaltung von H_0 , wenn $d \in [d_u; d_o]$

Beispiel 36: Statistisches Testen von Hypothesen über die Differenz zweier relativer Häufigkeiten

Überprüfung auf einem Signifikanzniveau von $\alpha = 0,05$, ob sich der Stimmenanteil einer Partei innerhalb eines Monats verändert hat.

Hypothesenformulierung:

$$H_0: \delta = 0 \quad \text{und} \quad H_1: \delta \neq 0$$



Einshypothese H_1

Bezeichnungen A und B sind beliebig wählbar

$$n_A = 400; p_A = 0,343 (= 137 \text{ Personen})$$

$$n_B = 600; p_B = 0,368 (= 221 \text{ Personen})$$

$$\rightarrow p = \frac{0,343 \cdot 400 + 0,368 \cdot 600}{400 + 600} = \frac{358}{1.000} = 0,358$$

Bereich der schwachen Indizien gegen H_0 nach (21):

$$d_o = +1,96 \cdot \sqrt{0,358 \cdot (1 - 0,358) \cdot \left(\frac{1}{400} + \frac{1}{600} \right)} = \underline{\underline{0,061}}$$

$$d_u = -1,96 \cdot \sqrt{0,358 \cdot (1 - 0,358) \cdot \left(\frac{1}{400} + \frac{1}{600} \right)} = \underline{\underline{-0,061}}$$

→ $d = 0,343 - 0,368 = -0,025 \in [-0,061; 0,061]$ → schwaches Indiz gegen H_0 → Beibehaltung von H_0 !

Testergebnis ist **nicht signifikant**.

Einseitige Fragestellungen:

Mögliche Hypothesenformulierungen sind

$$H_0: \delta = \pi_A - \pi_B \leq 0 \quad \text{und} \quad H_1: \delta = \pi_A - \pi_B > 0$$

oder

$$H_0: \delta = \pi_A - \pi_B \geq 0 \quad \text{und} \quad H_1: \delta = \pi_A - \pi_B < 0$$

Einseitige Schranke d_o oder d_u der schwachen Indizien gegen H_0 durch (21) mit $u_{1-\alpha}$ an Stelle von $u_{1-\alpha/2}$

Entscheidungsregel: Beibehaltung von H_0 , wenn $d \leq d_o$ (bzw. wenn $d \geq d_u$)

Abwandlung von Beispiel 36:

Überprüfung auf einem Signifikanzniveau von $\alpha = 0,05$, ob sich der Stimmenanteil innerhalb eines Monats **vergrößert/verkleinert** hat.

Festlegung von A und B z.B.:

A ... Grundgesamtheit zum früheren Zeitpunkt

B ... Grundgesamtheit zum späteren Zeitpunkt

vergrößert: $H_1: \delta = \pi_A - \pi_B < 0$



verkleinert: $H_1: \delta = \pi_A - \pi_B > 0$

Entscheidungsregeln mit p-Wert:

Beibehaltung von H_0 bei zweiseitiger Fragestellung, wenn $\alpha_2 > \alpha$

Beibehaltung von H_0 bei einseitiger Fragestellung, wenn $\alpha_1 > \alpha$

3.7 Testen von Hypothesen über zwei Mittelwerte

Aufgabe: Treffen einer fundierten Entscheidung zwischen zwei konkurrierenden Unterstellungen über den Vergleich der Mittelwerte zweier Grundgesamtheiten A und B:

Parameter μ_A , μ_B und $\delta = \mu_A - \mu_B$

Stichprobenergebnisse \bar{x}_A , \bar{x}_B und $d = \bar{x}_A - \bar{x}_B$ (Punktschätzer)

Bei zwei unabhängigen Stichproben:

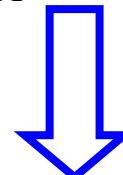
Zentraler Grenzwertsatz der Statistik:

d ist in großen Stichproben aus großen Grundgesamtheiten näherungsweise normalverteilt mit Erwartungswert δ und theoretischer Varianz $\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}$.

→ Rechnen mit der Normalverteilung (σ_A^2 und σ_B^2 durch Stichprobenvarianzen s_A^2 und s_B^2 ersetzen)

Zweiseitige Fragestellung:

Überprüfung der Verschiedenheit von μ_A und μ_B auf einem Signifikanzniveau $\alpha = 0,05 \rightarrow \delta = \mu_A - \mu_B \neq 0$



$$H_0: \delta = \mu_A - \mu_B = 0 \quad \text{und} \quad H_1: \delta = \mu_A - \mu_B \neq 0$$

Bereich der schwachen Indizien gegen die Nullhypothese

$$d_o = +u_{1-\alpha/2} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} \quad (22)$$

$$d_u = -u_{1-\alpha/2} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Entscheidungsregel: Beibehaltung von H_0 , wenn gilt: $d \in [d_u; d_o]$

Einseitige Fragestellungen:

$$H_0: \delta = \mu_A - \mu_B \leq 0 \quad \text{und} \quad H_1: \delta = \mu_A - \mu_B > 0$$

oder

$$H_0: \delta = \mu_A - \mu_B \geq 0 \quad \text{und} \quad H_1: \delta = \mu_A - \mu_B < 0$$

Einseitige Schranke d_o oder d_u der schwachen Indizien gegen H_0 durch (22) mit $u_{1-\alpha}$ an Stelle von $u_{1-\alpha/2}$

Entscheidungsregel: Beibehaltung von H_0 , wenn $d \leq d_o$ (bzw. wenn $d \geq d_u$)

Entscheidungsregeln bei p-Werten

Beispiel 37: Statistisches Testen von Hypothesen über die Differenz zweier Mittelwerte

Überprüfung auf einem Signifikanzniveau von $\alpha = 0,05$, ob Ergebnisse von Leistungstests in B besser als in A sind.

$$H_0: \delta \geq 0 \quad \text{und} \quad H_1: \delta < 0 \quad (\mu_B > \mu_A!)$$

$$n_A = 500; \bar{x}_A = 501,5; s_A = 25$$

$$n_B = 550; \bar{x}_B = 503,2; s_B = 30$$

Untere Schranke der schwachen Indizien gegen H_0 :

$$d_u = -1,65 \cdot \sqrt{\frac{625}{500} + \frac{900}{550}} = -2,803$$

→ $d = 501,5 - 503,2 = -1,7 \geq -2,803 \rightarrow$ schwaches Indiz gegen $H_0 \rightarrow$ Beibehaltung von H_0 !

Testergebnis ist nicht signifikant.

3.8 Testen einer Hypothese über einen statistischen Zusammenhang zweier nominaler Merkmale

Aufgabe: Fundierte Entscheidung zwischen zwei konkurrierenden Unterstellungen über den Zusammenhang zweier nominaler Merkmale

Kennzahl des statistischen Zusammenhangs: χ^2 („Chiquadrat“)

→ „Chiquadrattest“

Beispiel 38: Testen des Zusammenhangs zwischen zwei nominalen Merkmalen

Häufigkeiten h :

		Parteipräferenz					Summe
Geschlecht		A	B	C	D	E	
	weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	200
Summe		200	180	50	40	30	500

Beobachtete relative Häufigkeiten p_{ij}^b :

		Parteipräferenz					Summe
		A	B	C	D	E	
Geschlecht	weiblich	0,22	0,24	0,04	0,06	0,04	0,60
	männlich	0,18	0,12	0,06	0,02	0,02	0,40
	Summe	0,40	0,36	0,10	0,08	0,06	1

Bei Fehlen eines Zusammenhangs erwartete relative Häufigkeiten p_{ij}^e :

		Parteipräferenz					Summe
		A	B	C	D	E	
Geschlecht	weiblich	0,24	0,216	0,06	0,048	0,036	0,60
	männlich	0,16	0,144	0,04	0,032	0,024	0,40
	Summe	0,40	0,36	0,10	0,08	0,06	1

χ^2 ... Parameter; χ^2_{err} ... Stichprobenergebnis

$$\chi_{\text{err}}^2 = n \cdot \sum \frac{(p_{ij}^b - p_{ij}^e)^2}{p_{ij}^e} \quad (23)$$

Überprüfung auf einem Signifikanzniveau α , ob ein Zusammenhang in der Grundgesamtheit besteht.

$$H_0: \chi^2 = 0 \quad \text{und} \quad H_1: \chi^2 > 0$$

Einseitige Fragestellung

Gilt H_0 : Stichprobenergebnisse χ_{err}^2 haben eine **Chiquadratverteilung**

Voraussetzung: $n \cdot p_{ij}^e > 5$ für alle i, j

Einshypothese H_1

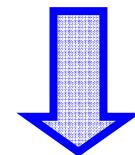
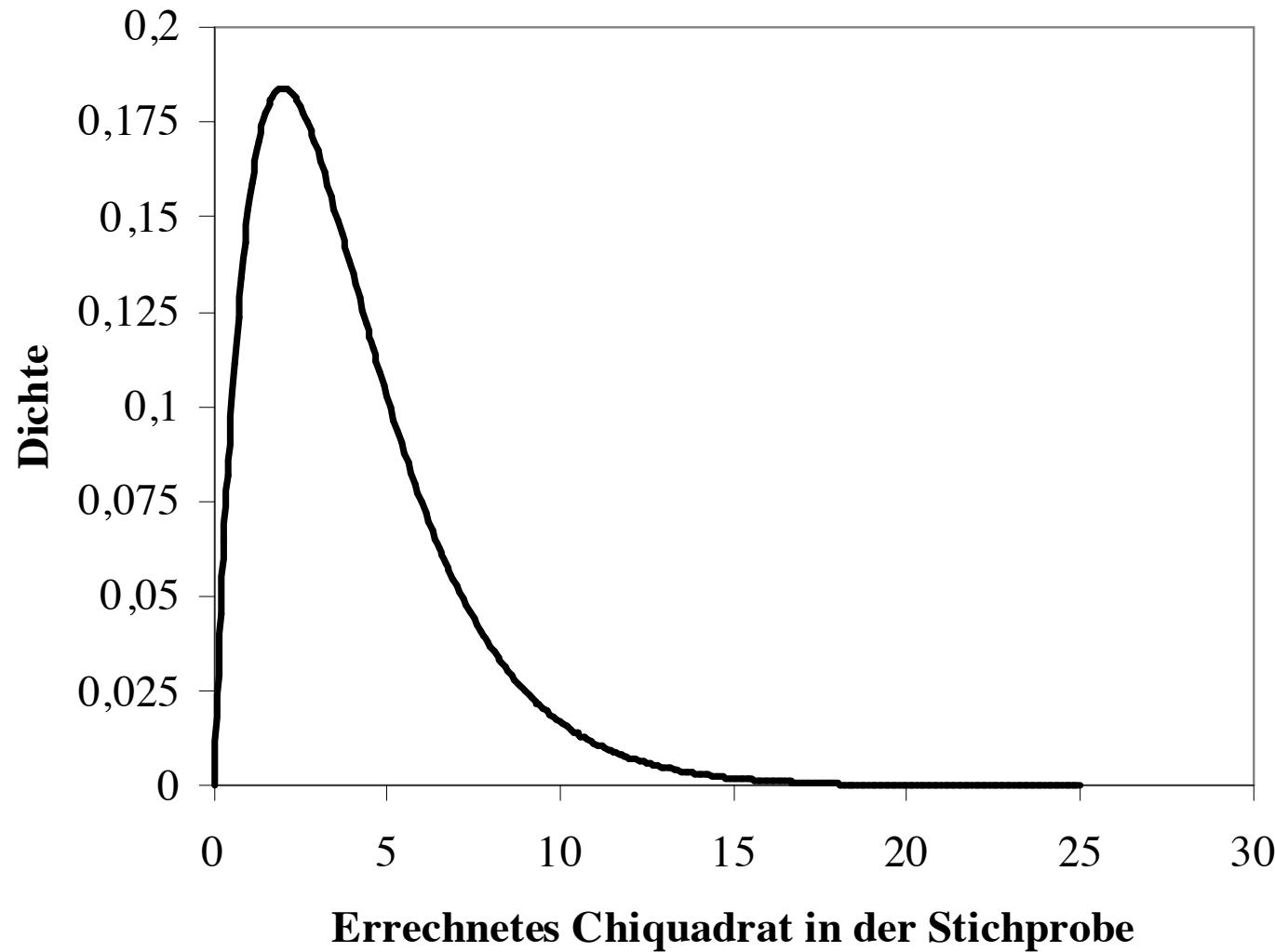


Abbildung 50: Eine chiquadratverteilte Zufallsvariable

Die Dichte der Zufallsvariablen χ_{err}^2 bei $\chi^2 = 0$ und 4 Freiheitsgraden



Bestimmung der oberen Schranke der schwachen Indizien gegen H_0 : Tabelle B im Anhang

- Signifikanzniveau α
- Freiheitsgrade: das Produkt der jeweils um 1 verminderten Anzahlen der Merkmalsausprägungen der beiden Merkmale



Erklärung der Freiheitsgrade: $(2-1) \cdot (5-1) = 4$

		Parteipräferenz					Summe
		A	B	C	D	E	
Geschlecht	weiblich	110	120	20	30		300
	männlich						200
	Summe	200	180	50	40	30	500

$$\chi^2_{\text{err}} = 18,06 \text{ (aus Beispiel 13)}$$

Obere Schranke der schwachen Indizien gegen H_0 : 9,49 (Tabelle B)

Freiheitsgrade	$\alpha=0,1$	$\alpha=0,05$	$\alpha=0,01$
1	2,71	3,84	6,63
2	4,61	5,99	9,21
3	6,25	7,81	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81
7	12,02	14,07	18,48
8	13,36	15,51	20,09
9	14,68	16,92	21,67
10	15,99	18,31	23,21

$$\chi^2_{\text{err}} = 18,06 > 9,49 \rightarrow \text{Entscheidung für die Einshypothese}$$

Testergebnis ist signifikant

Abbildung 50: Eine chiquadratverteilte Zufallsvariable

Die Dichte der Zufallsvariablen χ_{err}^2 bei $\chi^2 = 0$ und 4 Freiheitsgraden

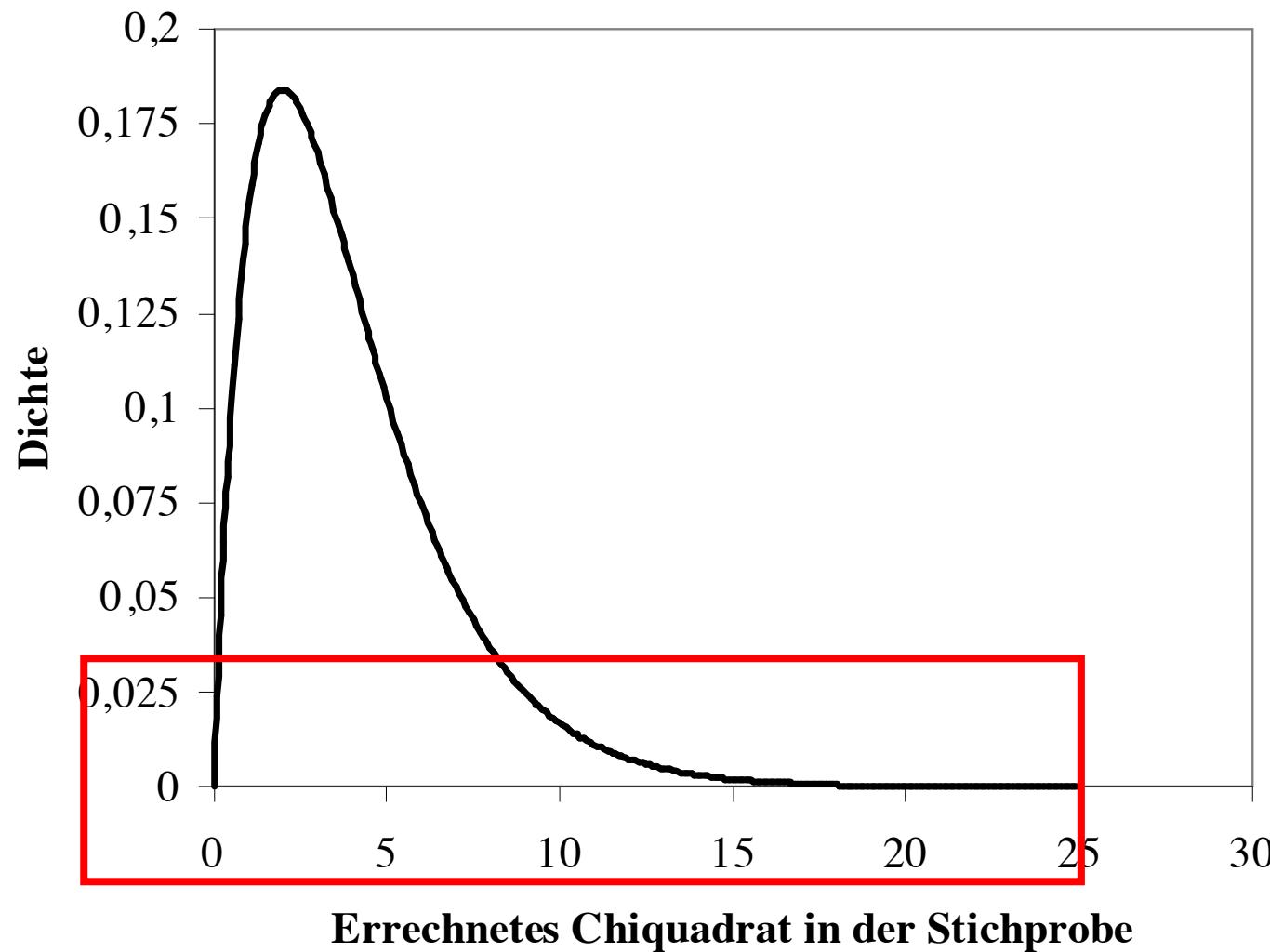
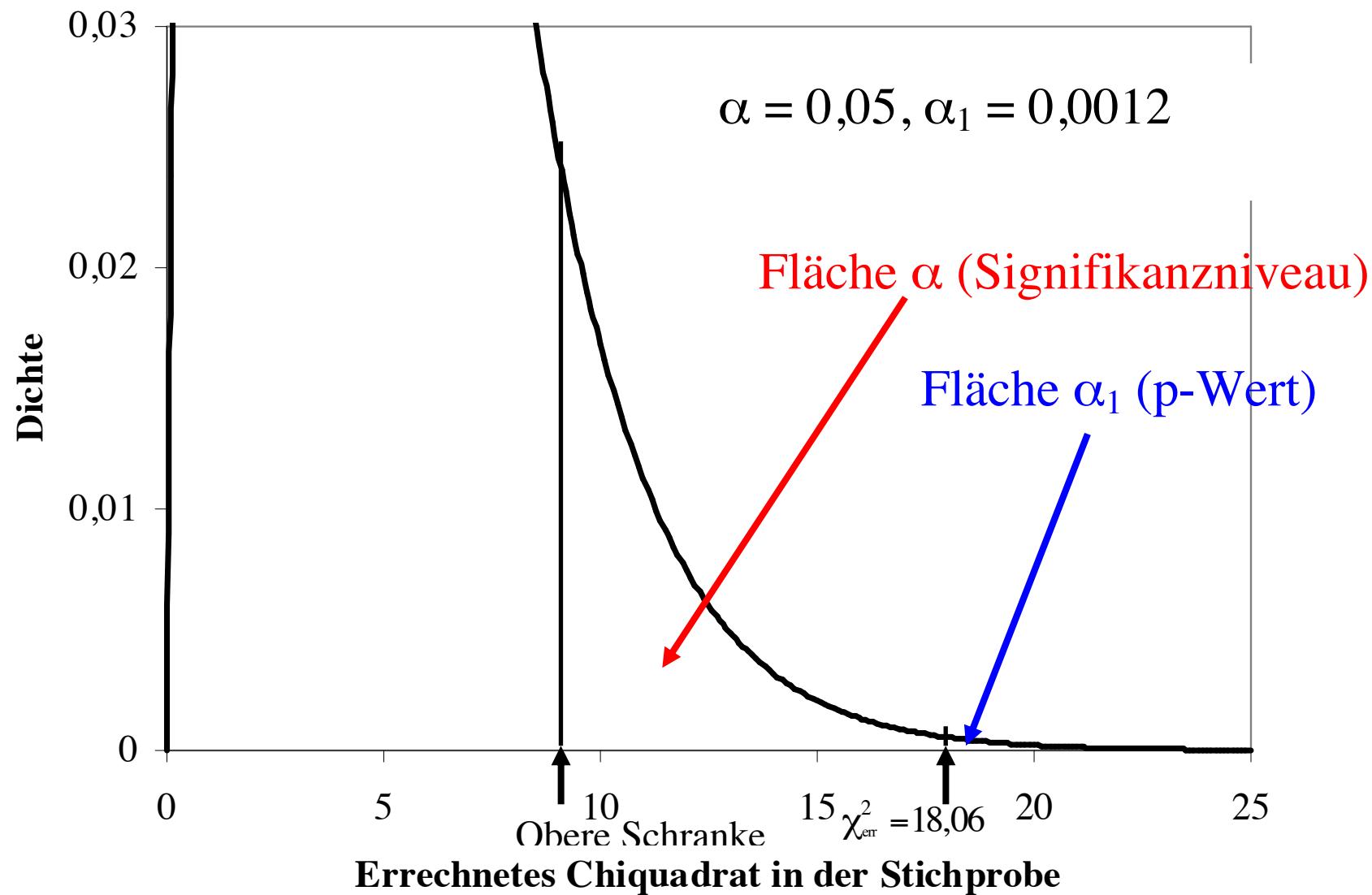


Abbildung 51: Chiquadratverteilte Zufallsvariable (Vergrößerung)



3.9 Testen von Hypothesen über eine Verteilungsform

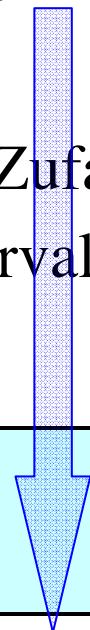
Aufgabe: Fundierte Entscheidung zwischen zwei konkurrierenden Unterstellungen über die Verteilungsform eines Merkmals

Normalverteilung der Daten ist Voraussetzung für die Anwendung einiger Methoden der schließenden Statistik (Abschnitte 3.10 bis 3.12)

Überprüft wird auf einem Signifikanzniveau α , ob die Normalverteilungsannahme nicht zutrifft

χ^2_{err} misst hier die Abweichungen zwischen in der Zufallsstichprobe beobachteten relativen Häufigkeiten p_i^b eines Intervalls und bei Normalverteilung erwarteten relativen Häufigkeiten p_i^e :

$$\chi^2_{\text{err}} = n \cdot \sum \frac{(p_i^b - p_i^e)^2}{p_i^e}$$



(24)

Einhypothese H_1

Hypothesenformulierung:

$$H_0: \chi^2 = 0 \quad \text{und} \quad H_1: \chi^2 > 0$$

Faustregel für die Anzahl der zu bildenden Intervalle: \sqrt{n}

Beispiel 39: Testen einer Verteilungsform

Überprüfung auf einem Signifikanzniveau $\alpha = 0,05$, ob das Gewicht von Zuckerpaketen nicht normalverteilt ist.

Stichprobe: $n = 100$, $\bar{x} = 999,93$, $s^2 = 0,25$

Beobachtete relative Häufigkeiten (zur Vereinfachung der Darstellung nur 4 Intervalle):

Intervall	relative Häufigkeit p_i^b
unter 999,5	0,2
[999,5; 1.000]	0,32
[1.000; 1.000,5]	0,34
über 1.000,5	0,14

Bei Normalverteilung erwartete relative Häufigkeiten (= die Wahrscheinlichkeiten der Intervalle):

$$\Pr(x > 1.000,5) = \Pr(u > 1,14) = 1 - \Pr(u \leq 1,14) = 1 - 0,873 = 0,127$$

$$\text{Nach (14)} : u = \frac{1.000,5 - 999,93}{0,5} = 1,14$$

und so fort ...

Intervall	relative Häufigkeit p_i^e	relative Häufigkeit p_i^b
unter 999,5	0,195	0,2
[999,5; 1.000]	0,361	0,32
[1.000; 1.000,5]	0,317	0,34
über 1.000,5	0,127	0,14

$$\chi_{\text{err}}^2 = n \cdot \sum \frac{(p_i^b - p_i^e)^2}{p_i^e} = 100 \cdot \left[\frac{(0,2 - 0,195)^2}{0,195} + \frac{(0,32 - 0,361)^2}{0,361} + \dots \right] = 0,78$$



χ^2 -Tabelle → obere Schranke der schwachen Indizien gegen H_0

- Signifikanzniveau α
- Freiheitsgrade in diesem Fall: Anzahl der Intervalle minus 3

$\alpha = 0,05$, Freiheitsgrade = $4 - 3 \rightarrow$ Obere Schranke = 3,84

$\chi_{\text{err}}^2 = 0,78 \leq 3,84 \rightarrow$ Testergebnis ist nicht signifikant.

H_0 (=Normalverteilungshypothese) wird beibehalten.

Entscheidung mittels p-Wert:

p-Wert $\alpha_1: 0,377$

Beibehaltung von H_0 , wenn $\alpha_1 > \alpha$

3.10 Testen von Hypothesen über einen statistischen Zusammenhang zweier metrischer Merkmale

Aufgabe: Fundierte Entscheidung zwischen zwei konkurrierenden Unterstellungen über den Zusammenhang zweier metrischer Merkmale

Messung des Zusammenhangs zweier metrischer Merkmale: **Korrelationskoeffizient**

Bezeichnungen:

ρ ... Korrelation in der Grundgesamtheit

r ... Korrelation in der Stichprobe

Voraussetzung:

Beide Merkmale normalverteilt → Chiquadrat-test aus Abschnitt 3.9! (Weitere Möglichkeit: Kolmogorov-Smirnov-Test)

Zweiseitige Fragestellung:

Überprüfung auf einem Signifikanzniveau $\alpha = 0,05$, ob zwischen den beiden Merkmalen ein statistischer Zusammenhang besteht (egal in welcher Richtung):

Hypothesenformulierung:

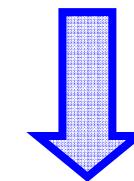
$$H_0: \rho = 0 \quad \text{und} \quad H_1: \rho \neq 0$$

Umweg:

$$u = \rho \cdot \sqrt{\frac{n - 2}{1 - \rho^2}}$$

Umformulierung der Hypothesen:

$$H_0: u = 0 \quad \text{und} \quad H_1: u \neq 0$$



Einshypothese H_1

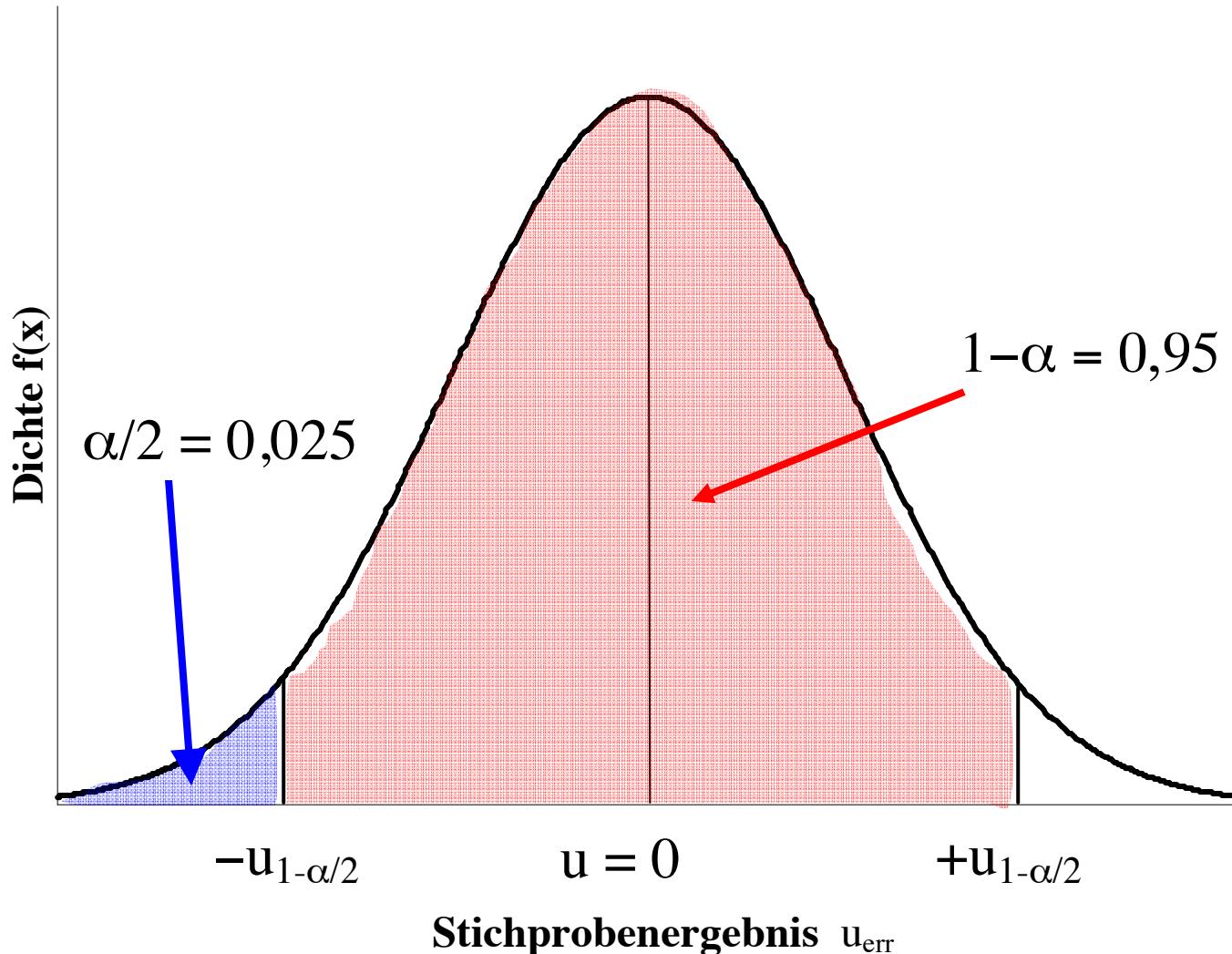
Korrelationskoeffizienten r nach (9) mit den Daten der Stichprobe:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Testgröße:

$$u_{\text{err}} = r \cdot \sqrt{\frac{n - 2}{1 - r^2}} \quad (25)$$

Wenn $u = 0$: u_{err} ist t-verteilt, in großen Stichproben annähernd *standard-normalverteilt* ➔



Bereich der schwachen Indizien gegen H_0 : $[-u_{1-\alpha/2}; +u_{1-\alpha/2}]$

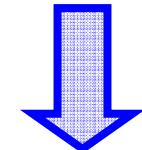
Entscheidungsregel: Beibehaltung von H_0 , wenn $u_{\text{err}} \in [-u_{1-\alpha/2}; +u_{1-\alpha/2}]$

Einseitige Fragestellungen:

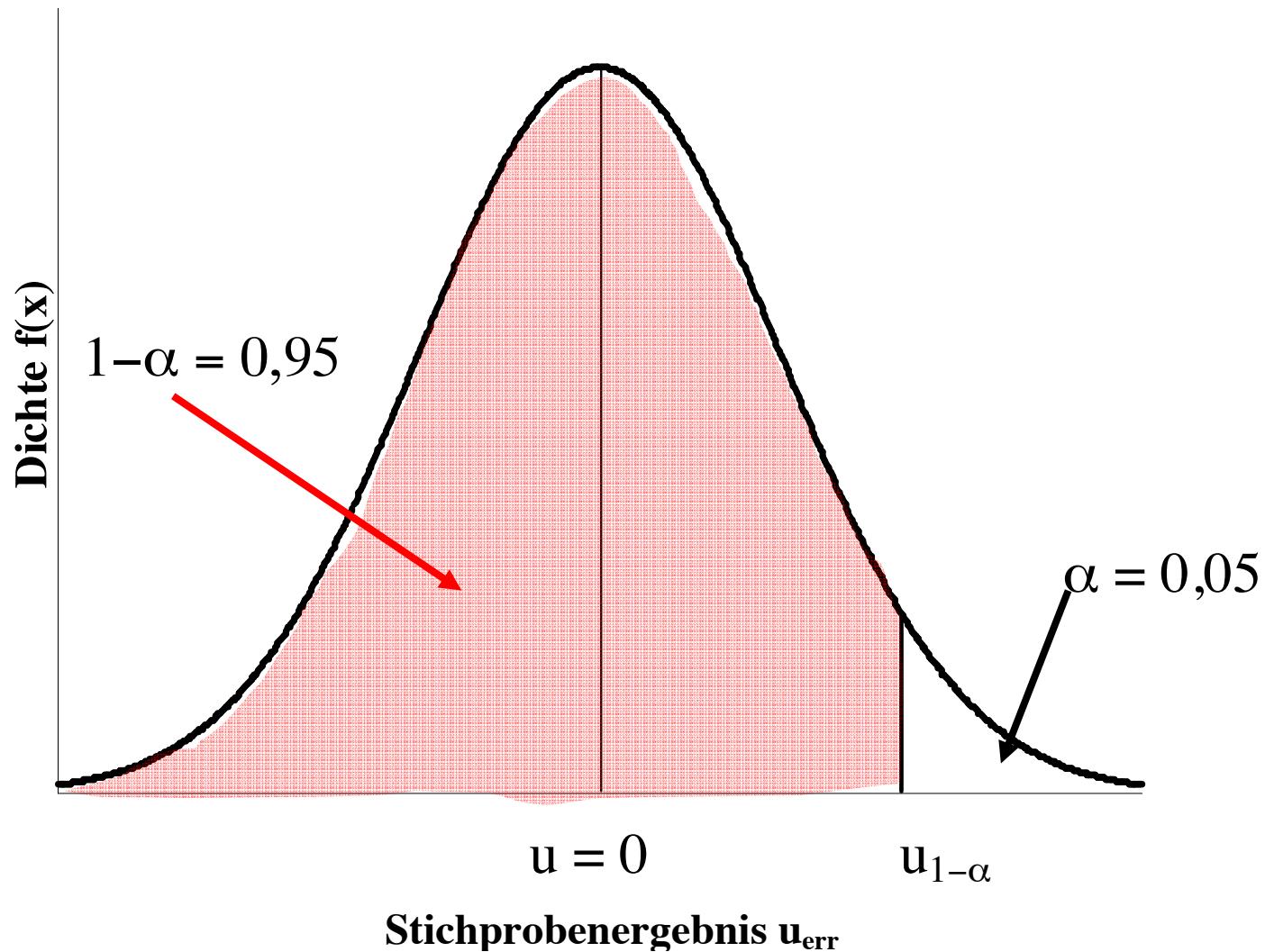
- Überprüfung auf einem Signifikanzniveau $\alpha = 0,05$, ob ein gleichsin-
niger Zusammenhang besteht:

Hypothesenformulierung:

$$H_0: \rho \leq 0 \quad \text{und} \quad H_1: \rho > 0$$



Einshypothese H_1



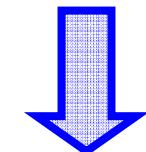
Oberschranke der schwachen Indizien gegen H_0 : $+u_{1-\alpha}$

Entscheidungsregel: Beibehaltung der Nullhypothese, wenn $u_{\text{err}} \leq +u_{1-\alpha}$.

- Überprüfung auf einem Signifikanzniveau $\alpha = 0,05$, ob ein gegensinniger Zusammenhang besteht:

Hypothesenformulierung:

$$H_0: \rho \geq 0 \quad \text{und} \quad H_1: \rho < 0$$



Einshypothese H_1

Untere Schranke der schwachen Indizien gegen H_0 : $-u_{1-\alpha}$

Entscheidungsregel: Beibehaltung der Nullhypothese, wenn $u_{\text{err}} \geq -u_{1-\alpha}$

Beispiel 40: Testen eines gleichsinnigen Zusammenhangs

Überprüfung eines gleichsinnigen Zusammenhangs auf einem Signifikanzniveau $\alpha = 0,05$

$$H_0: \rho \leq 0 \quad \text{und} \quad H_1: \rho > 0$$

Nach Beibehaltung der Nullhypotesen bei den Chiquadrattests auf Normalverteilung:

$$n = 600, r = 0,36$$

$$u_{\text{err}} = r \cdot \sqrt{\frac{n-2}{1-r^2}} = 0,36 \cdot \sqrt{\frac{600-2}{1-0,36^2}} = 9,44$$

Obere Schranke des Bereichs der schwachen Indizien gegen $H_0: + 1,65$

$9,44 > 1,65$: Die Einstypothese wird akzeptiert

Testergebnis ist signifikant.

Entscheidungsregeln mit p-Werten:

Beibehaltung von H_0 , wenn bei zweiseitiger Fragestellung gilt: $\alpha_2 > \alpha$

Beibehaltung von H_0 , wenn bei einseitiger Fragestellung $\alpha_1 > \alpha$ gilt

Beispiel für einen SPSS-Output:

Kolmogorov-Smirnov-Anpassungstest

		Hausübung	Klausur
N		155	155
Parameter der Normalverteilung ^{a,b}	Mittelwert	32,85	48,59
	Standardabweichung	10,136	17,947
Extremste Differenzen	Absolut	,081	,086
	Positiv	,067	,074
	Negativ	-,081	-,086
Kolmogorov-Smirnov-Z		1,006	1,069
Asymptotische Signifikanz (2-seitig)		,263	,203

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
Hausübung * Klausur	155	100,0%	0	0,0%	155	100,0%

Symmetrische Maße

	Wert	Asymptotischer Standardfehler ^a	Näherungsweises T ^b	Näherungsweise Signifikanz
Intervall- bzgl. Intervallmaß Pearson-R	,716	,039	12,697	,000
Ordinal- bzgl. Ordinalmaß Korrelation nach Spearman	,728	,039	13,118	,000
Anzahl der gültigen Fälle	155			

a. Die Null-Hypothese wird nicht angenommen.

b. Unter Annahme der Null-Hypothese wird der asymptotische Standardfehler verwendet.

c. Basierend auf normaler Näherung

3.11 Testen von Hypothesen über Regressionskoeffizienten (einfache Regressionsanalyse)

Korrelationsrechnung: *statistischer* Zusammenhang metrischer Merkmale

Regressionsrechnung: *Kausalität* des Zusammenhangs

Ökonometrie

Regressand Regressor

Regressionsgerade in der Grundgesamtheit: $\hat{y} = \beta_1 \cdot x + \beta_2$

Aufgabe: Fundierte Entscheidung zwischen Hypothesen über die Steigung β_1 oder den Achsenabschnitt β_2 der Regressionsgeraden

Schätzen und Testen der beiden Regressionskoeffizienten: einfache Regressionsanalyse

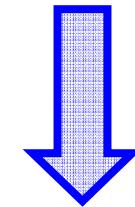
Schätzer b_1 und b_2 wie in Abschnitt 1.3.4 mit Daten aus einer uneingeschränkten Zufallsstichprobe

Testen von Hypothesen über β_1

- Überprüfung auf einem Signifikanzniveau α , ob der Regressor x einen Einfluss auf den Regressanden y ausübt

Hypothesenformulierung:

$$H_0: \beta_1 = 0 \quad \text{und} \quad H_1: \beta_1 \neq 0$$



Einhypothese H_1

Voraussetzung: Normalverteilung beider Merkmale (Chiquadrat-test über die Verteilungsform)

Für ausreichend große Stichproben (Faustregel: $n \geq 100$) sind b_{1o} und b_{1u} Schranken des Bereichs der schwachen Indizien gegen H_0 :

$$\begin{aligned} b_{1o} &= u_{1-\alpha/2} \cdot s_{b_1} \\ b_{1u} &= -u_{1-\alpha/2} \cdot s_{b_1} \end{aligned} \tag{26}$$

In kleinen Stichproben: t-Verteilung mit $n-2$ Freiheitsgraden

Stichprobenstandardabweichung s_{b_1} von b_1 aus $s_{b_1}^2 = \frac{1-r^2}{n-2} \cdot \frac{s_y^2}{s_x^2}$

Entscheidungsregel: Beibehaltung der Nullhypothese, wenn $b_1 \in [b_{1u}; b_{1o}]$

Fragestellung identisch mit dem Test von

$$H_0: \rho = 0 \quad \text{und} \quad H_1: \rho \neq 0$$

in der Korrelationsanalyse

Einseitige Hypothesenformulierungen:

$$H_0: \beta_1 \leq 0 \quad \text{und} \quad H_1: \beta_1 > 0$$

oder

$$H_0: \beta_1 \geq 0 \quad \text{und} \quad H_1: \beta_1 < 0 .$$

Obere Schranke b_{1o} oder untere Schranke b_{1u} der schwachen Indizien gegen H_0 aus (26) mit $u_{1-\alpha}$ statt $u_{1-\alpha/2}$

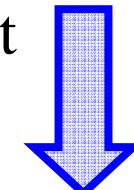
Entscheidungsregeln: Beibehaltung von H_0 , wenn $b_1 \leq b_{1o}$ ($b_1 \geq b_{1u}$)

Beispiel 41: Testen von einseitigen Hypothesen über die Steigung der Regressionsgeraden

Überprüfung auf einem Signifikanzniveau $\alpha = 0,05$, ob die Steigung der Regressionsgeraden in der Grundgesamtheit positiv ist

Hypothesenformulierung:

$$H_0: \beta_1 \leq 0 \quad \text{und} \quad H_1: \beta_1 > 0$$

 Eishypothese H_1

Voraussetzung: Normalverteilung

Stichprobe: $n=330$, $s_{xy} = 67,72$ nach (8), $s_x^2 = 219,63$ und $s_y^2 = 1.065,37$

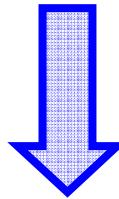
$$\text{nach (3), } r = 0,14, b_1 = \frac{67,72}{219,63} = 0,31.$$

Obere Schranke der schwachen Indizien gegen H_0 :

$$b_{10} = u_{1-\alpha} \cdot s_{b_1} = u_{1-\alpha} \cdot \sqrt{\frac{1-r^2}{n-2} \cdot \frac{s_y^2}{s_x^2}} = 1,65 \cdot \sqrt{\frac{1-0,14^2}{330-2} \cdot \frac{1.065,37}{219,63}} = 0,20$$

$0,31 > 0,20 \rightarrow$ Testergebnis ist signifikant \rightarrow Entscheidung für die Einst-hypothese

■ Überprüfung, ob β_1 anders ist als eine vorgegebene Steigung β_1^0



Hypothesenformulierung:

$$H_0: \beta_1 = \beta_1^0 \quad \text{und} \quad H_1: \beta_1 \neq \beta_1^0$$

Einshypothese H_1

Bereich der schwachen Indizien gegen diese Nullhypothese in großen Stichproben:

$$b_{1o} = \beta_1^0 + u_{1-\alpha/2} \cdot s_{b_1} \quad (27)$$

$$b_{1u} = \beta_1^0 - u_{1-\alpha/2} \cdot s_{b_1}$$

Beibehaltung von H_0 auf einem Signifikanzniveau α , wenn $b_1 \in [b_{1u}, b_{1o}]$.

Einseitige Fragestellungen: In (27) die obere beziehungsweise die untere Schranke mit $u_{1-\alpha}$ statt $u_{1-\alpha/2}$ berechnen

Test auf größere Steigung: $H_0 (\beta_1 \leq \beta_1^0)$ beibehalten, wenn $b_1 \leq b_{1o}$

Test auf kleinere Steigung: $H_0 (\beta_1 \geq \beta_1^0)$ beibehalten, wenn $b_1 \geq b_{1u}$

Beispiel 42: Testen von zweiseitigen Hypothesen über die Steigung der Regressionsgeraden

Überprüfung, ob die Steigung β_1 der Regressionsgeraden von $\beta_1^0 = 0,25$ abweicht

Hypothesenformulierung:

$$H_0: \beta_1 = 0,25 \quad \text{und} \quad H_1: \beta_1 \neq 0,25$$

Bereich der schwachen Indizien gegen die Nullhypothese nach (27):

$$b_{1o} = \beta_1^0 + u_{1-\alpha/2} \cdot s_{b_1} = 0,25 + 1,96 \cdot \sqrt{\frac{1 - 0,14^2}{330 - 2} \cdot \frac{1.065,37}{219,63}} = 0,49$$

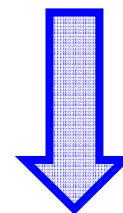
und

$$b_{1u} = \beta_1^0 - u_{1-\alpha/2} \cdot s_{b_1} = 0,01$$

Stichprobenergebnis $b_1 = 0,31 \in [0,01; 0,49] \rightarrow$ Nullhypothese beibehalten \rightarrow Testergebnis ist nicht signifikant

Testen von Hypothesen über β_2

- Überprüfung, ob der Achsenabschnitts β_2 anders ist als ein vorgegebener Aschenabschnitt β_2^0



Einshypothese H_1

Hypothesenformulierung:

$$H_0: \beta_2 = \beta_2^0 \quad \text{und} \quad H_1: \beta_2 \neq \beta_2^0 .$$

Bereich der schwachen Indizien gegen die Nullhypothese:

$$\begin{aligned} b_{2o} &= \beta_2^0 + u_{1-\alpha/2} \cdot s_{b_2} \\ b_{2u} &= \beta_2^0 - u_{1-\alpha/2} \cdot s_{b_2} \end{aligned} \tag{28}$$

$$\text{Varianz } s_{b_2}^2 = \frac{1-r^2}{n-2} \cdot s_y^2 \cdot \left(1 + \frac{\bar{x}^2}{s_x^2} \right)$$

s_x^2 und s_y^2 nach (3) und nicht nach (19): Sonst an Stelle der Zahl 1 in der Klammer den Quotient $(n-1)/n$ verwenden

Entscheidungsregel: Beibehaltung der Nullhypothese, wenn $b_2 \in [b_{2u}, b_{2o}]$

Einseitige Fragestellungen:

$$H_0: \beta_2 \leq \beta_2^0 \quad \text{und} \quad H_1: \beta_2 > \beta_2^0$$

beziehungsweise

$$H_0: \beta_2 \geq \beta_2^0 \quad \text{und} \quad H_1: \beta_2 < \beta_2^0$$

In (28) obere beziehungsweise untere Schranke mit $u_{1-\alpha}$ statt $u_{1-\alpha/2}$

Bei kleinen Stichproben: $u_{1-\alpha}$ (bei zweiseitigen Fragestellungen: $u_{1-\alpha/2}$) durch Wert der t-Verteilung bei $n-2$ Freiheitsgraden ersetzen

p-Werte: Keine Änderung der Entscheidungsregeln

Multiple lineare Regressionsanalysen

Nichtlineare Regressionsanalysen



3.12 Testen von Hypothesen über mehr als zwei Mittelwerte (Einfache Varianzanalyse)

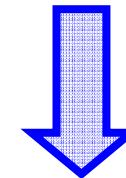
Aufgabe: Treffen einer fundierten Entscheidung zwischen zwei konkurrierenden Unterstellungen über den Vergleich der Mittelwerte (stellvertretend für die Verteilungen) eines metrischen Merkmals in *mehr als zwei* Grundgesamtheiten (= Gruppen)

2 Voraussetzungen:

- Normalverteilung des Merkmals in allen Grundgesamtheiten (Chiquadrat-test über die Verteilungsform)
- Gleiche Varianz σ^2 des Merkmals in allen Grundgesamtheiten

Einfache (oder einfaktorielle) Varianzanalyse:

Ein Merkmal, der sogenannte **Faktor**, erzeugt h verschiedene Grundgesamtheiten. Überprüfung, ob sich mindestens 2 der h Gruppenmittelwerte unterscheiden:



Hypothesenformulierung:

Einshypothese H_1

$$H_0: \mu_A = \mu_B = \dots = \mu_K \quad \text{und} \quad H_1: \mu_i \neq \mu_j \text{ (für mindestens 2 Gruppen)}$$

Idee: Betrachtung der Differenzen $\mu_A - \mu$, $\mu_B - \mu$ und so fort

Nach (3) gilt mit neuer Notation:

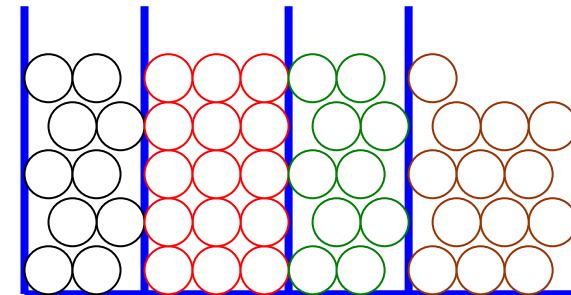
$$\sigma^2 = \frac{1}{N} \cdot \sum_{i=1}^N (x_i - \mu)^2 \quad (3)$$

Varianzzerlegung:

$$\sigma^2 = \sigma_I^2 + \sigma_Z^2$$

σ_I^2 ... Varianz innerhalb der Gruppen:

$$\sigma_I^2 = \frac{1}{N} \cdot [N_A \cdot \sigma_A^2 + N_B \cdot \sigma_B^2 + \dots + N_K \cdot \sigma_K^2]$$



mit den Innergruppenvarianzen

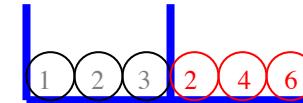
$$\sigma_A^2 = \frac{1}{N_A} \cdot \sum_{i=1}^{N_A} (x_i - \mu_A)^2 \text{ und so weiter}$$

σ_Z^2 ... Varianz zwischen den Gruppen:

$$\sigma_Z^2 = \frac{1}{N} \cdot [N_A \cdot (\mu_A - \mu)^2 + N_B \cdot (\mu_B - \mu)^2 + \dots + N_K \cdot (\mu_K - \mu)^2]$$

Beispiel 43: Varianzzerlegung

Gruppe A: 1, 2, 3 und Gruppe B: 2, 4, 6



$$\mu = \frac{1}{6} \cdot (1 + 2 + 3 + 2 + 4 + 6) = 3, \sigma^2 = \frac{1}{6} \cdot [(1-3)^2 + \dots + (6-3)^2] = 2,6$$

Varianzzerlegung: $\mu_A = 2, \mu_B = 4,$

$$\sigma_A^2 = \frac{1}{3} \cdot [(1-2)^2 + (2-2)^2 + (3-2)^2] = 0,6,$$

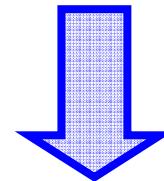
$$\sigma_B^2 = \frac{1}{3} \cdot [(2-4)^2 + (4-4)^2 + (6-4)^2] = 2,6$$

$$\sigma_I^2 + \sigma_Z^2 = \frac{1}{N} \cdot [N_A \cdot \sigma_A^2 + N_B \cdot \sigma_B^2] + \frac{1}{N} \cdot [N_A \cdot (\mu_A - \mu)^2 + N_B \cdot (\mu_B - \mu)^2] =$$

$$= \frac{1}{6} \cdot [3 \cdot 0,6 + 3 \cdot 2,6] + \frac{1}{6} \cdot [3 \cdot (2-3)^2 + 3 \cdot (4-3)^2] = 1,6 + 1 = 2,6$$

Gilt H_1 , dann ist $\sigma_z^2 \neq 0 \rightarrow$ Umformulierung der Hypothesen:

Überprüft wird, ob $\Phi = \sigma_z^2 / \sigma_I^2$ größer als null ist.



Hypothesenformulierung:

$$H_0: \Phi = 0 \quad \text{und} \quad H_1: \Phi > 0$$

Einshypothese H_1

Zufallsstichproben aus allen h Grundgesamtheiten (n_A, n_B, \dots, n_K):

$$F_{\text{err}} = \frac{\frac{1}{h-1} \cdot [n_A \cdot (\bar{x}_A - \bar{x})^2 + n_B \cdot (\bar{x}_B - \bar{x})^2 + \dots + n_K \cdot (\bar{x}_K - \bar{x})^2]}{\frac{1}{n-h} \cdot [(n_A - 1) \cdot s_A^2 + (n_B - 1) \cdot s_B^2 + \dots + (n_k - 1) \cdot s_K^2]} \quad (29)$$

$$\text{mit } s_A^2 = \frac{1}{n_A - 1} \cdot \sum_{i=1}^{n_A} (x_i - \bar{x}_A)^2 \text{ nach (19)}$$

F_{err} besitzt bei Gültigkeit von $H_0: \Phi = 0$ eine F-Verteilung

Einseitige Fragestellung:

Obere Schranke der schwachen Indizien gegen die Nullhypothese (eine Tabelle der F-Verteilung):

- Signifikanzniveau α
- „Freiheitsgrade“ $h-1$ des Ausdrucks im Zähler
- „Freiheitsgrade“ $n-h$ des Ausdrucks im Nenner von F_{err}

Beispielsweise bei 10 Grundgesamtheiten, gesamte Anzahl an Befragten $n = 5.000$, Signifikanzniveau $\alpha = 0,05$:

Obere Schranke: 1,882 ($\alpha=0,05$, Freiheitsgrade 9 und 4.990)

Beispiel 44: Einfache Varianzanalyse

Überprüfung, ob sich die Mathematikkenntnisse in drei Bundesländern auf einem Signifikanzniveau $\alpha = 0,05$ unterscheiden

$$H_0: \Phi = 0 \quad \text{und} \quad H_1: \Phi > 0$$

$$n_A = 6, n_B = 6, n_C = 3$$

Bundesland: Testergebnisse:

A	502	508	498	505	508	509
B	492	498	490	512	488	490
C	499	507	488			

Annahme: Normalverteilungen und Gleichheit der Varianzen

$$\bar{x}_A = \frac{1}{6} \cdot (502 + 508 + \dots + 509) = 505$$

$$\bar{x}_B = \frac{1}{6} \cdot (492 + 498 + \dots + 490) = 495$$

$$\bar{x}_C = \frac{1}{3} \cdot (499 + 507 + 488) = 498$$

$$\bar{x} = \frac{1}{15} \cdot (502 + 508 + \dots + 488) = 499,6$$

Zähler von F_{err} nach (29):

$$\begin{aligned} n_A \cdot (\bar{x}_A - \bar{x})^2 + n_B \cdot (\bar{x}_B - \bar{x})^2 + \dots + n_K \cdot (\bar{x}_K - \bar{x})^2 &= \\ = \frac{1}{2} \left[6 \cdot (505 - 499,6)^2 + 6 \cdot (495 - 499,6)^2 + 3 \cdot (498 - 499,6)^2 \right] &= 154,8 \end{aligned}$$

$$s_A^2 = \frac{1}{5} \cdot [(502 - 505)^2 + (508 - 505)^2 + \dots + (509 - 505)^2] = 18,4$$

$$s_B^2 = \frac{1}{5} \cdot [(492 - 495)^2 + (498 - 495)^2 + \dots + (490 - 495)^2] = 81,2$$

$$s_C^2 = \frac{1}{2} \cdot [(499 - 498)^2 + (507 - 498)^2 + (488 - 498)^2] = 91$$

Nenner von F_{err} nach (29):

$$\frac{1}{n-h} \cdot [(n_A - 1) \cdot s_A^2 + (n_B - 1) \cdot s_B^2 + \dots + (n_k - 1) \cdot s_K^2] =$$

$$\frac{1}{12} \cdot (5 \cdot 18,4 + 5 \cdot 81,2 + 2 \cdot 91) = 56,6$$

$$\rightarrow F_{\text{err}} = \frac{154,8}{56,6} = 2,732$$

Obere Schranke der schwachen Indizien gegen die Nullhypothese:

Z.B. in Office 2003 Excel-Funktion „FINV“ mit $\alpha=0,05$, Freiheitsgrade 2 und 12: **3,886**.

Da $2,732 < 3,886$: Beibehaltung der Nullhypothese

Testergebnis ist **nicht signifikant**.

Abbildung 52: F-verteilte Zufallsvariable

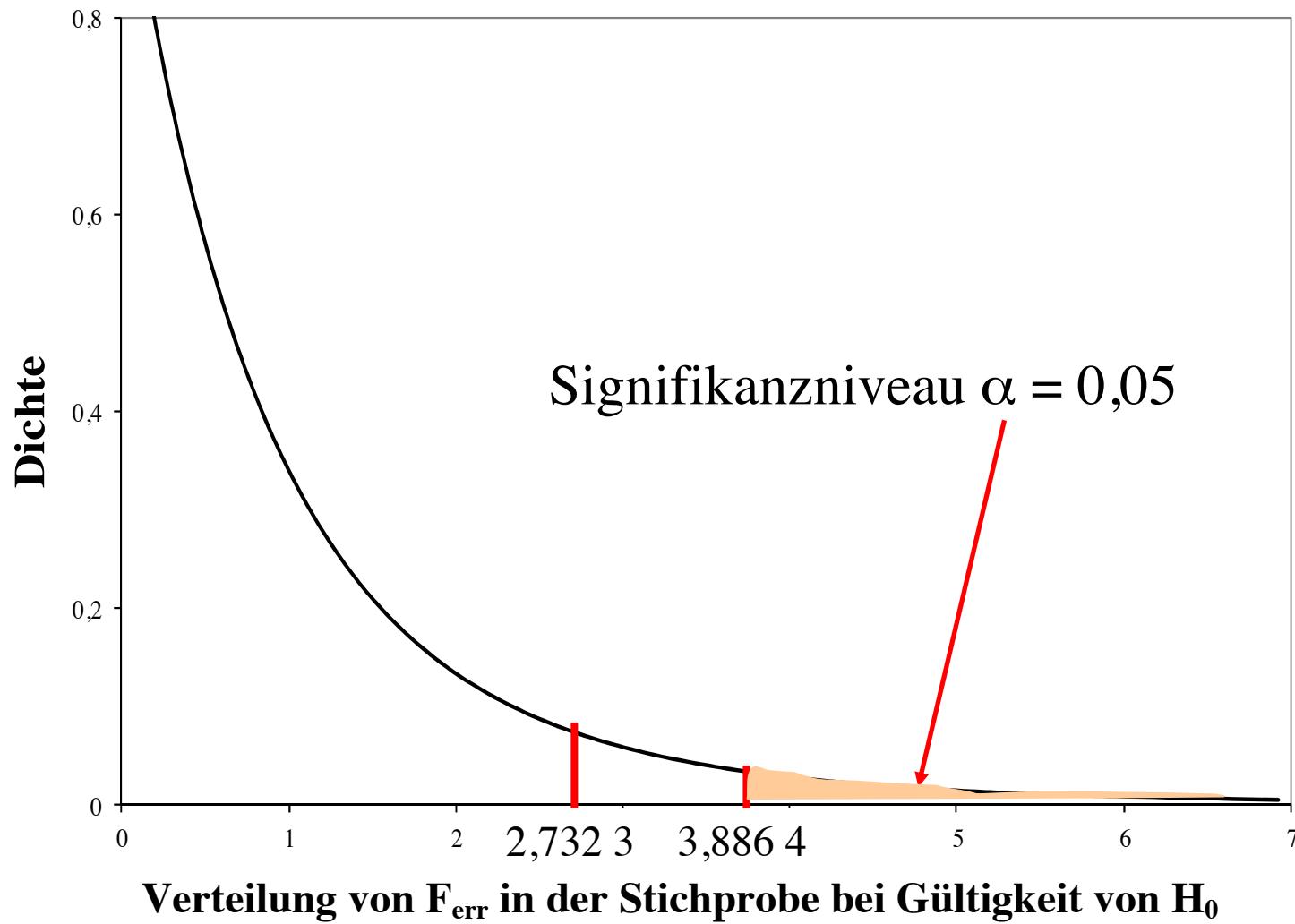
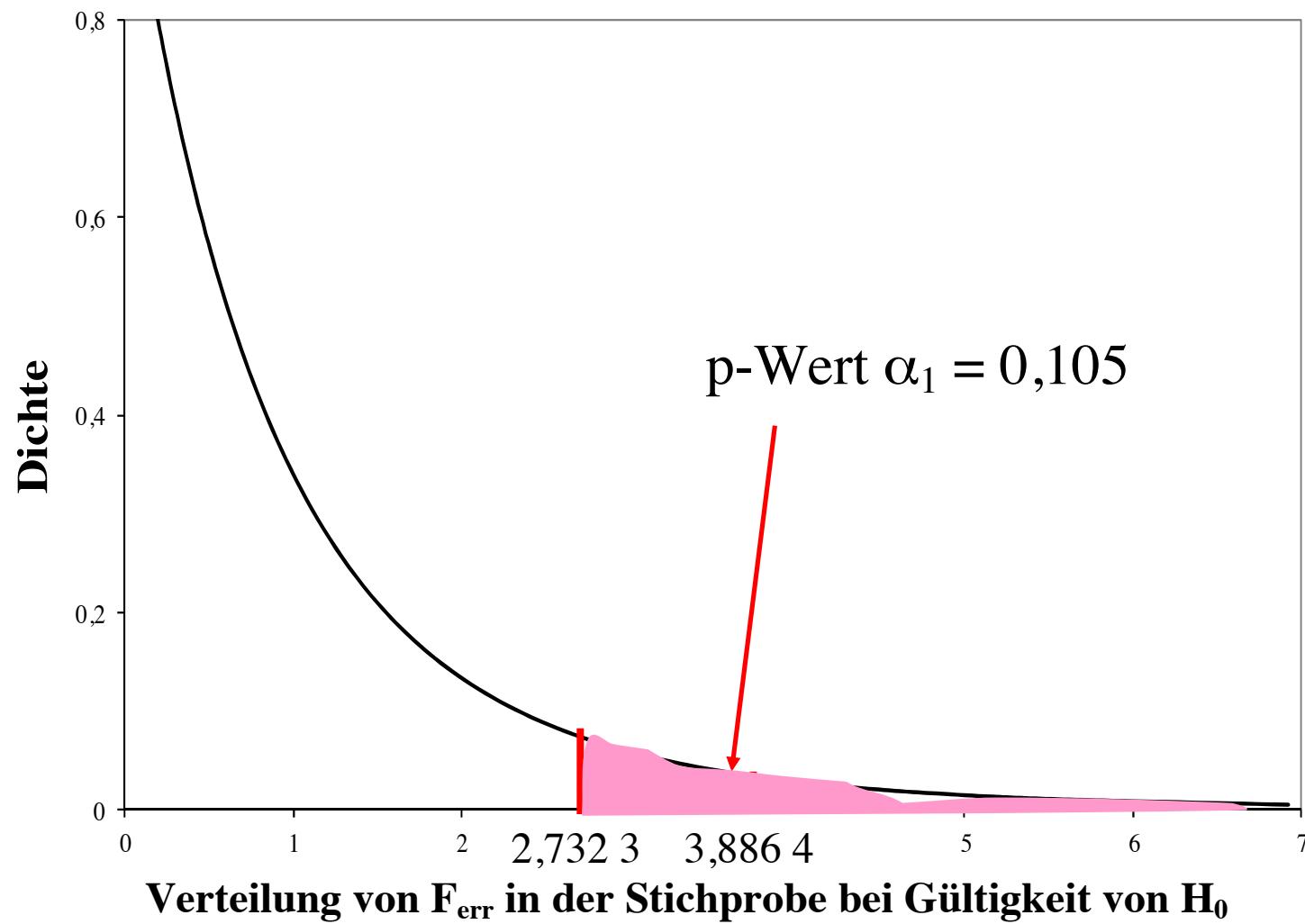


Abbildung 53: F-verteilte Zufallsvariable ($\alpha_1 = 0,105$)





Mehrfaktorielle Varianzanalysen

3.13 Probleme in der Anwendung statistischer Tests

1 Das Finden der geeigneten Teststrategie:

Einheitliche Handlungslogik - unterschiedliche mathematische Ansätze

Lösung: Experten der Statistik zu Rate zu ziehen



2 Das Signifikanz-Relevanz-Problem:

Zunehmende Stichprobenumfänge → immer mehr signifikante Stichprobenergebnisse

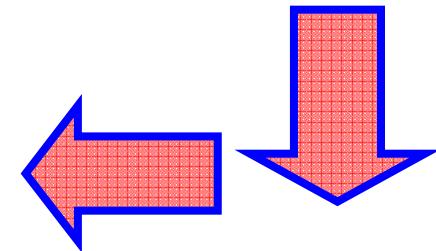
In Beispiel 40 ($H_0: \rho \leq 0$, $H_1: \rho > 0$; $r = 0,36$): Auch $r = 0,07$ wäre signifikant

$$u_{\text{err}} = 0,07 \cdot \sqrt{\frac{600 - 2}{1 - 0,07^2}} = 1,72 > 1,65$$

Praktische Relevanz?

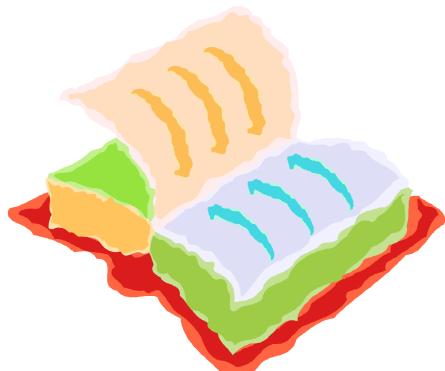
Lösung: Richtige Hypothesenformulierung

Beispiel: Überprüfung, ob ein praktisch bedeutsamer gleichsinniger Zusammenhang besteht und schätzt man nur Korrelationen größer als 0,4 als praktisch relevant ein → Hypothesenformulierung:



$$H_0: \rho \leq 0,4 \quad \text{und} \quad H_1: \rho > 0,4$$

Zunehmende Stichprobenumfänge → mehr richtige Entscheidungen!



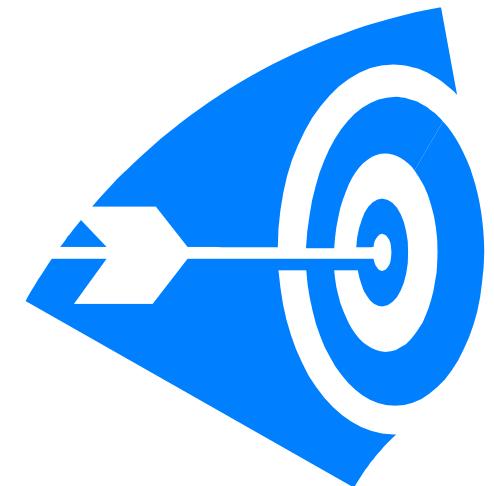
Literaturhinweis:

Quatember, A. (2005). *Das Signifikanz-Relevanz-Problem beim statistischen Testen von Hypothesen*. ZUMA-Nachrichten 57, S. 128-150.

3 Das Alles-mit-Allem-Testen:

Statistik-Programmpakete: Automatismus des Auswerfens von p-Werten
„Alles-mit-Allem-Testen“ ohne Forschungshypothesen („Ausquetschen“ der Daten)

„Es (gehört) zum Standard wissenschaftlicher Studien, dass *erst* das Untersuchungsziel und die Hypothese angegeben werden müssen und *dann* die Daten erhoben werden. Wer aber nach *irgendwelchen Mustern* in Datensammlungen sucht und *anschließend* seine Theorien bildet, schießt sozusagen auf die weiße Scheibe und malt danach die Kreise um das Einschussloch“ (von Randow (1994), S.94).



„Münzwurfexperiment“:

400 Studierende in einem Hörsaal

Vorhersage der Ausgänge von 5 Münzwürfen (Untersuchungsziel:
Auffinden von hellseherisch veranlagten Personen)

Für *einen* Ratenden:

$$\Pr(x = 5) = \left(\frac{1}{2}\right)^5 = 0,03125 \text{ Signifikantes Ergebnis!}$$

Bei *400* Ratenden:

Durchschnittlich $400 \cdot 0,5^5 = 12,5$ Personen (Forderung der statistischen Theorie!)

Staunende Öffentlichkeit: 12 Personen mit „statistisch nachgewiesenen“ telepathischen Fähigkeiten



Zeitungen:

„....Wissenschaftler haben auf Grund von Stichprobenuntersuchungen etwas festgestellt, wofür sie jedoch keine Erklärung anbieten können ...“

Wer ein Ergebnis nicht erklären kann, hat *nichts* gefunden!

Unsinn in
den Medien

GÖN 27.12.03

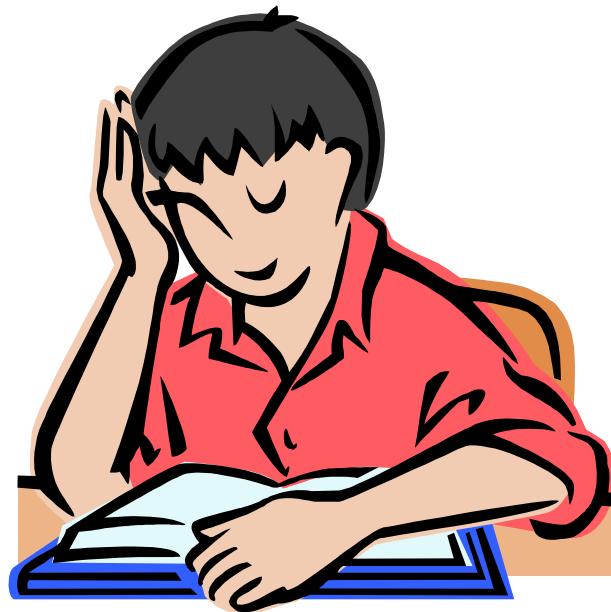
Durchfall stört Intelligenz

Häufiger Durchfall in Kleinkinderalter kann die Entwicklung der Intelligenz beeinträchtigen. Das ergab eine amerikanische Studie. Warum das so ist, konnten die Forscher allerdings nicht erklären.

Alles-mit-Allem-Testen: Theorien haben nie eine Chance zur Widerlegung

Lösung: Überprüfung *vorab* formulierter Forschungshypothesen

Alles-mit-Allem-Tests nur als Unterstützung des Nachdenkens des Anwenders (und danach: Theorienbildung und neue Untersuchung)

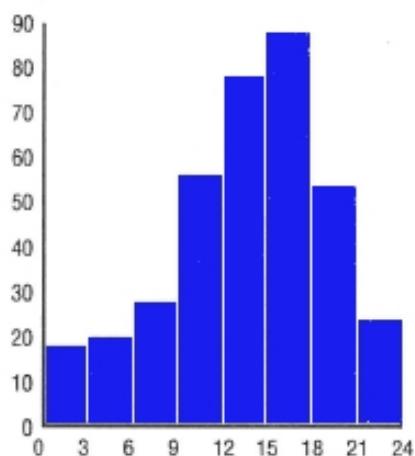


Fallstricke bei Histogrammen

Punkteverteilung Statistik-Prüfung

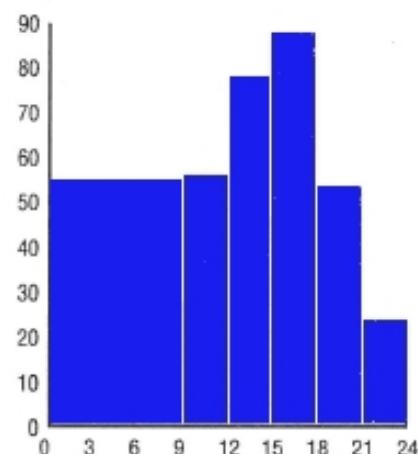
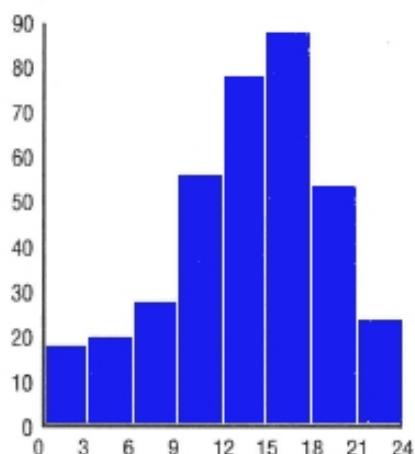
Fallstricke bei Histogrammen

Punkteverteilung Statistik-Prüfung



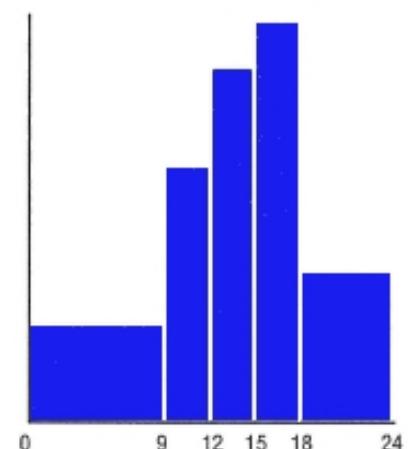
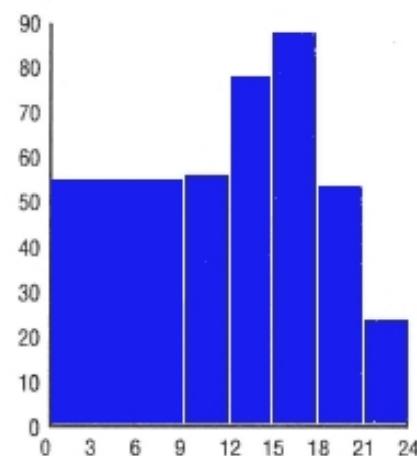
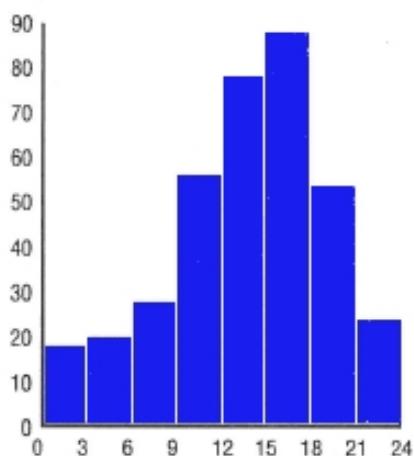
Fallstricke bei Histogrammen

Punkteverteilung Statistik-Prüfung



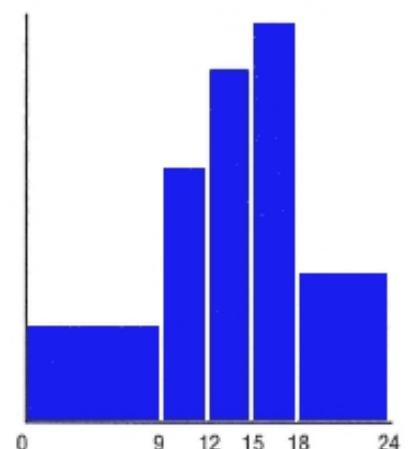
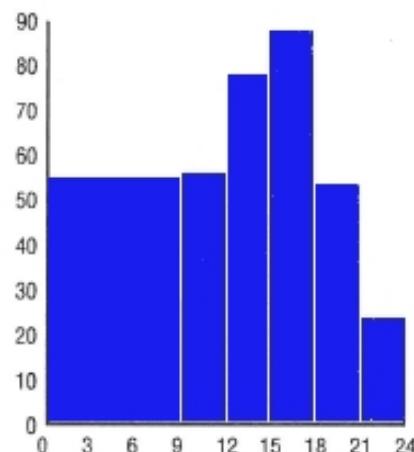
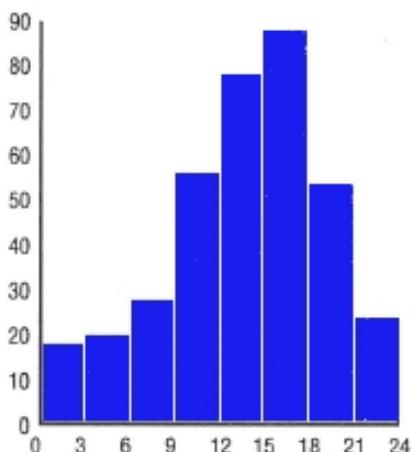
Fallstricke bei Histogrammen

Punkteverteilung Statistik-Prüfung



Fallstricke bei Histogrammen

Punkteverteilung Statistik-Prüfung



Das gleiche Histogramm mit verschiedenen breiten Intervallen, falsch und richtig.

Das Flächenprinzip

Falls unterschiedliche Klassenbreiten verwendet werden:

$$\text{Höhe} = \frac{\text{Besetzungszahl}}{\text{Breite}}$$

Theorem

Die Fläche ist proportional zur dargestellten Zahl.

Aufgabe:

Für eine Menge aus fünf natürlichen Zahlen gilt: der Mittelwert \bar{x} ist 4, der Median x_Z ist 5 und der Modus x_D ist 1. Bestimmen Sie die fünf Zahlen.

Aufgabe:

Für eine Menge aus fünf natürlichen Zahlen gilt: der Mittelwert \bar{x} ist 4, der Median x_Z ist 5 und der Modus x_D ist 1. Bestimmen Sie die fünf Zahlen.

Lösung:

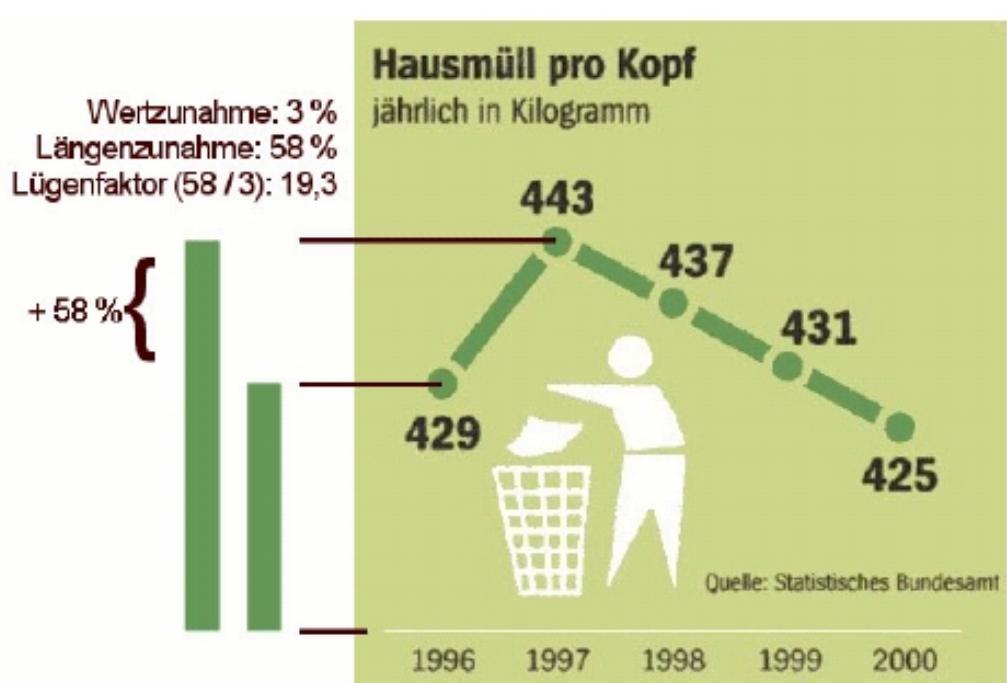
$$1; 1; 5; 6; 7$$

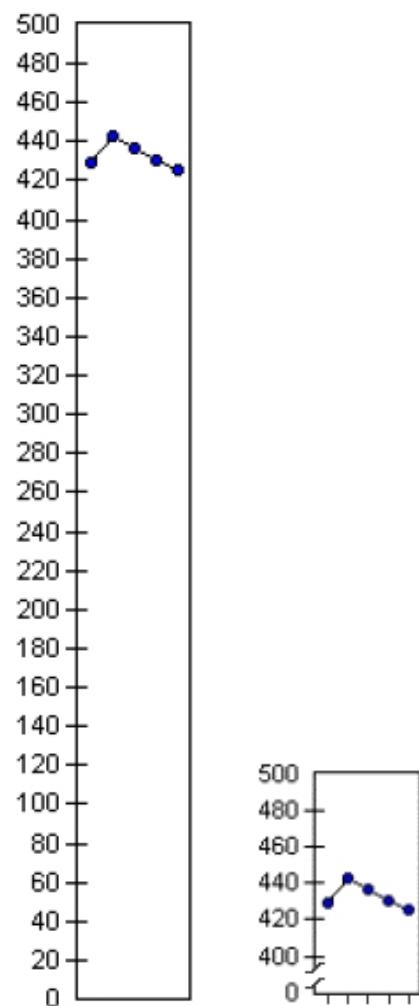
Lügen mit Grafiken

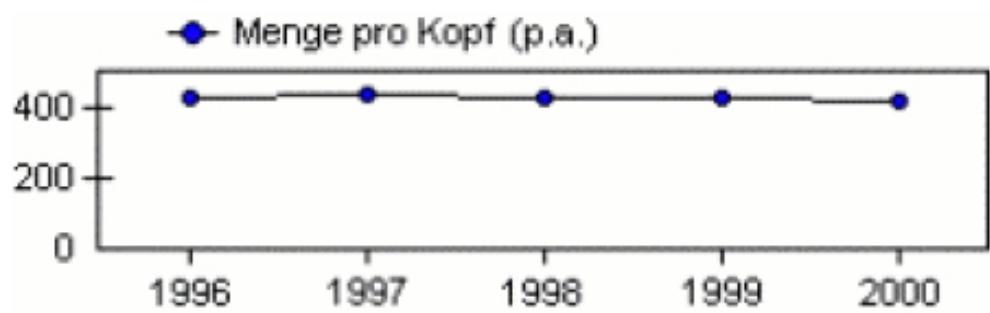


Strecken und Stauchen

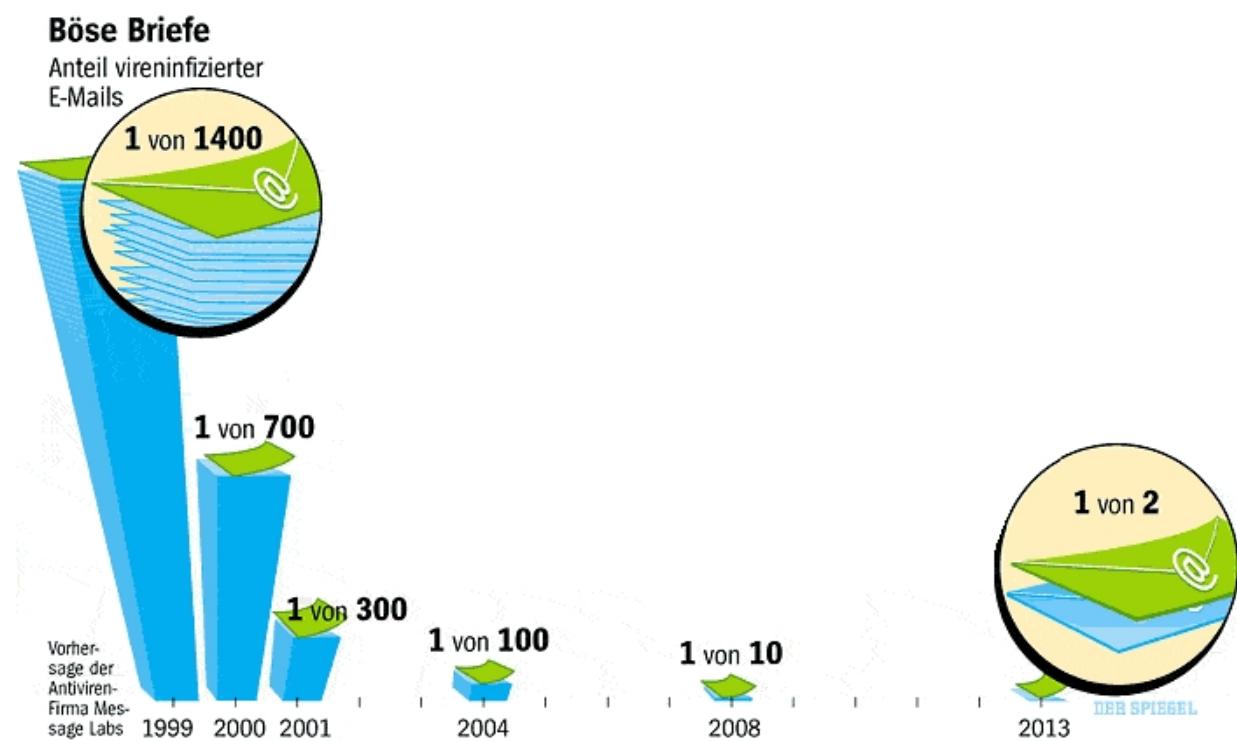
Jahr	Hausmüll pro Kopf in kg	Veränderung zum Vorjahr	
		abs.	in %
1996	429	-	-
1997	443	14	3%
1998	437	-6	-1%
1999	431	-6	-1%
2000	425	-6	-1%

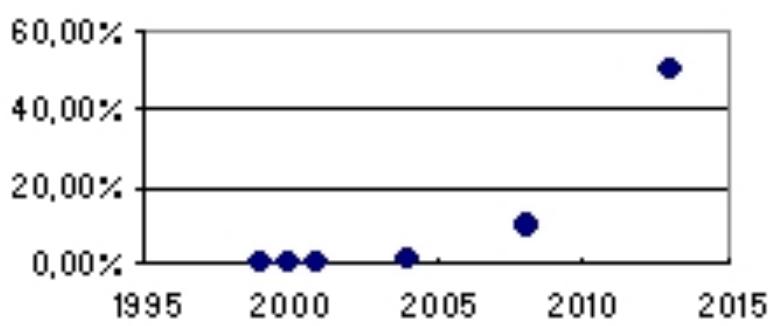






Verkehrte Proportionen

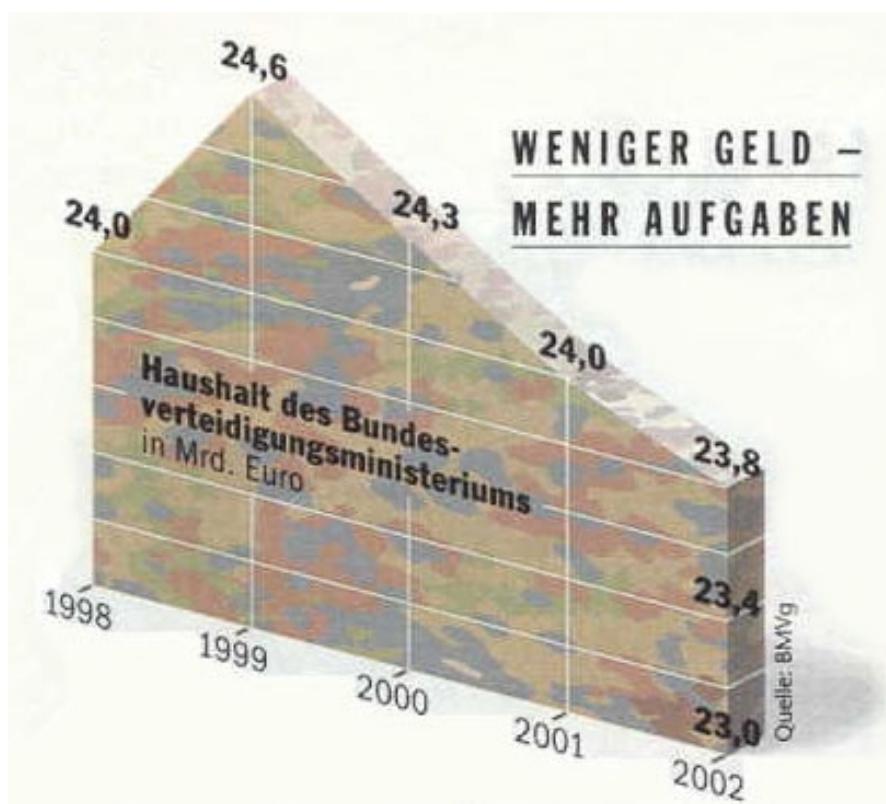




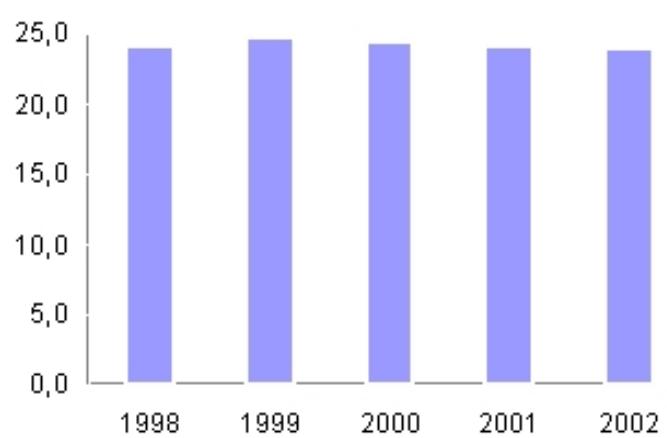
Seriöser Trend?

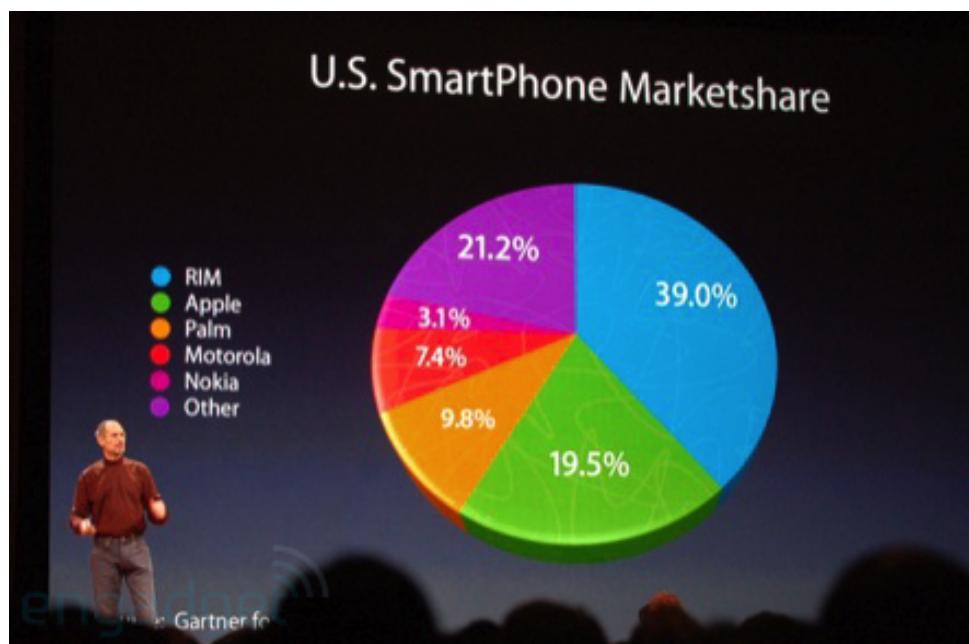
Jahr	Werte in %	Jahr	Werte in %
1999	0,07	2007	
2000	0,14	2008	10
2001	0,33	2009	
2002		2010	
2003		2011	
2004	1	2012	
2005		2013	50
2006			

3D-Übel



Jahr	Wehretat	Veränderung zum Vorjahr		Säule Wehretat	Veränderung zum Vorjahr	
	in Mrd. €	abs.	in %	mm	abs.	in %
1998	24,0			26		
1999	24,6	+0,6	+2,5	41	+15	+57,7
2000	24,3	-0,3	-1,2	33	-8	-19,5
2001	24,0	-0,3	-1,2	26	-7	-21,2
2002	23,8	-0,2	-0,8	21	-5	-19,2





POSITIVE ENTWICKLUNG

Im Straßenverkehr verunglückte Kinder
im Alter von unter 15 Jahren, in Deutschland



RÜCKGANG Die Zahl der verunglückten Kinder nimmt Jahr für Jahr ab – obwohl immer mehr Autos unterwegs sind

Quelle: Statistisches Bundesamt

Bedingte Wahrscheinlichkeiten

„Der Tod fährt mit! Vier von zehn tödlich verunglückten Autofahrern trugen keinen Sicherheitsgurt!“ (ADAC-Motorwelt)

Bedingte Wahrscheinlichkeiten

„Der Tod fährt mit! Vier von zehn tödlich verunglückten Autofahrern trugen keinen Sicherheitsgurt!“ (ADAC-Motorwelt)

Völlig uninteressant! Spannend wäre:

x von y angeschnallten Autofahrern verunglücken tödlich.

Bedingte Wahrscheinlichkeiten

„Der Tod fährt mit! Vier von zehn tödlich verunglückten Autofahrern trugen keinen Sicherheitsgurt!“ (ADAC-Motorwelt)

Völlig uninteressant! Spannend wäre:

x von y angeschnallten Autofahrern verunglücken tödlich.

Falsch interpretierte bedingte Wahrscheinlichkeiten finden sich leider oft in der Tagespresse.

Unfallstatistik:

	angeschnallt	nicht angeschnallt	Total
†	4	6	10
♡			
Total			

Unfallstatistik:

	angeschnallt	nicht angeschnallt	Total
†	4	6	10
♡	a	b	
Total	$a + 4$	$b + 6$	

Interessant wären die Zahlen a und b : Wie viele der bei einem Unfall angeschnallten (resp. nicht angeschnallten) Personen überleben den Unfall nicht?

- „Autofahrer fahren kurz vor dem Ziel sehr unvorsichtig!“

- „Autofahrer fahren kurz vor dem Ziel sehr unvorsichtig!“
- „Central Park ist sicherer als die eigenen Wohnung!“

- „Autofahrer fahren kurz vor dem Ziel sehr unvorsichtig!“
- „Central Park ist sicherer als die eigenen Wohnung!“
- „Frauen, hütet Euch vor Euren Ehemännern!“

- „Autofahrer fahren kurz vor dem Ziel sehr unvorsichtig!“
- „Central Park ist sicherer als die eigenen Wohnung!“
- „Frauen, hütet Euch vor Euren Ehemännern!“
- „Freizeit kann ein günstiger Nährboden für Kriminalität sein!“

- „Autofahrer fahren kurz vor dem Ziel sehr unvorsichtig!“
- „Central Park ist sicherer als die eigenen Wohnung!“
- „Frauen, hütet Euch vor Euren Ehemännern!“
- „Freizeit kann ein günstiger Nährboden für Kriminalität sein!“
- „Schnelles Autofahren ist sicherer als Bummeln.“

- „Autofahrer fahren kurz vor dem Ziel sehr unvorsichtig!“
- „Central Park ist sicherer als die eigenen Wohnung!“
- „Frauen, hütet Euch vor Euren Ehemännern!“
- „Freizeit kann ein günstiger Nährboden für Kriminalität sein!“
- „Schnelles Autofahren ist sicherer als Bummeln.“
- “Nachts fährts sich ungefährlicher.“

- „Autofahrer fahren kurz vor dem Ziel sehr unvorsichtig!“
- „Central Park ist sicherer als die eigenen Wohnung!“
- „Frauen, hütet Euch vor Euren Ehemännern!“
- „Freizeit kann ein günstiger Nährboden für Kriminalität sein!“
- „Schnelles Autofahren ist sicherer als Bummeln.“
- “Nachts fährts sich ungefährlicher.“
- „Ehefrau drängt Partner in den Alkoholismus!“

- „Autofahrer fahren kurz vor dem Ziel sehr unvorsichtig!“
- „Central Park ist sicherer als die eigenen Wohnung!“
- „Frauen, hütet Euch vor Euren Ehemännern!“
- „Freizeit kann ein günstiger Nährboden für Kriminalität sein!“
- „Schnelles Autofahren ist sicherer als Bummeln.“
- “Nachts fährts sich ungefährlicher.“
- „Ehefrau drängt Partner in den Alkoholismus!“
- „Haie lieben Männer!“

Persönliches Lieblingsbeispiel

„30% der Unfälle entstehen unter Einfluss von Alkohol.“

Persönliches Lieblingsbeispiel

„30% der Unfälle entstehen unter Einfluss von Alkohol.“

Na dann, Prost!

Persönliches Lieblingsbeispiel

„30% der Unfälle entstehen unter Einfluss von Alkohol.“

Na dann, Prost!

Persönliches Lieblingsbeispiel

„30% der Unfälle entstehen unter Einfluss von Alkohol.“

Na dann, Prost!



Jeder fünfte Unfall geschieht in der Romandie unter Alkoholeinfluss: Verkehrskontrolle in Genf.
Bild: Keystone

Ein Problem der Standardabweichung



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	0.95	0.85	1.1	0.9	1.0		

Ein Problem der Standardabweichung



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	0.95	0.85	1.1	0.9	1.0	0.96	

Ein Problem der Standardabweichung



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	0.95	0.85	1.1	0.9	1.0	0.96	0.0962

Ein Problem der Standardabweichung



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	0.95	0.85	1.1	0.9	1.0	0.96	0.0962



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	12790	11999	13490	12690	13390		

Ein Problem der Standardabweichung



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	0.95	0.85	1.1	0.9	1.0	0.96	0.0962



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	12790	11999	13490	12690	13390	12872	

Ein Problem der Standardabweichung



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	0.95	0.85	1.1	0.9	1.0	0.96	0.0962



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	12790	11999	13490	12690	13390	12872	603

Ein Problem der Standardabweichung



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	0.95	0.85	1.1	0.9	1.0	0.96	0.0962



Nr.	1	2	3	4	5	\bar{x}	SA
Preis	12790	11999	13490	12690	13390	12872	603

Gibt es mehr Variation in den Motorradpreisen?

Variationskoeffizient

Der **Variationskoeffizient** ist ein relatives Streumass. Er misst die Streuung relativ zum Mittelwert.

$$c_v = \frac{s}{\bar{y}}$$

Variationskoeffizient

Der **Variationskoeffizient** ist ein relatives Streumass. Er misst die Streuung relativ zum Mittelwert.

$$c_v = \frac{s}{\bar{y}}$$

- Wenn $c_v > 1$, dann haben die Daten beträchtliche Streuung.

Variationskoeffizient

Der **Variationskoeffizient** ist ein relatives Streumass. Er misst die Streuung relativ zum Mittelwert.

$$c_v = \frac{s}{\bar{y}}$$

- ▶ Wenn $c_v > 1$, dann haben die Daten beträchtliche Streuung.
- ▶ Wenn c_v nahe bei 0, dann haben die Daten relativ kleine Streuung.

Berechnen von c_v

Wir vergleichen einige Werte vom Variationskoeffizienten aus unserer Datensammlung DatenFS15.tns.

$$\frac{\text{stDevSamp}(\text{apfel})}{\text{mean}(\text{apfel})}$$

$$\frac{\text{stDevSamp}(\text{distanz})}{\text{mean}(\text{distanz})}$$

$$\frac{\text{stDevSamp}(\text{fussball})}{\text{mean}(\text{fussball})}$$

Binomialverteilung mit R

Beispiel: Die Wahrscheinlichkeit, dass man im Roulette bei einmaligem Setzen auf „rot“ gewinnt, ist $p = \frac{18}{37} = 0.486$. Definieren wir mit x jene Anzahl der Spiele, bei denen man bei fünfmaligem Setzen auf „rot“ gewinnt.

Binomialverteilung mit R

- Die Wahrscheinlichkeit, genau a Gewinne zu erzielen, ist

$$P(x = a) = \binom{n}{a} \cdot p^a \cdot (1 - p)^{n-a}$$

Binomialverteilung mit R

- Die Wahrscheinlichkeit, genau a Gewinne zu erzielen, ist

$$P(x = a) = \binom{n}{a} \cdot p^a \cdot (1 - p)^{n-a}$$

- Diese Formel ist in R abgebildet:

```
dbinom(a, size=n, prob=p)
```

Binomialverteilung mit R

- Die Wahrscheinlichkeit, genau a Gewinne zu erzielen, ist

$$P(x = a) = \binom{n}{a} \cdot p^a \cdot (1 - p)^{n-a}$$

- Diese Formel ist in R abgebildet:

```
dbinom(a, size=n, prob=p)
```

- Weitere Informationen finden wir unter `help(Binomial)`.

Binomiale Wahrscheinlichkeitsverteilung mit R

Bestimmen Sie mit R die Wahrscheinlichkeitsverteilung des Beispiels:

„rot“ gewinnt a mal	Wahrscheinlichkeitsverteilung
0	
1	
2	
3	
4	
5	

Binomiale Wahrscheinlichkeitsverteilung mit R

Bestimmen Sie mit R die Wahrscheinlichkeitsverteilung des Beispiels:

„rot“ gewinnt a mal	Wahrscheinlichkeitsverteilung
0	<code>dbinom(0, size=5, prob=18/37) = 0.036</code>
1	
2	
3	
4	
5	

Binomiale Wahrscheinlichkeitsverteilung mit R

Bestimmen Sie mit R die Wahrscheinlichkeitsverteilung des Beispiels:

„rot“ gewinnt a mal	Wahrscheinlichkeitsverteilung
0	<code>dbinom(0, size=5, prob=18/37) = 0.036</code>
1	<code>dbinom(1, size=5, prob=18/37) = 0.170</code>
2	
3	
4	
5	

Binomiale Wahrscheinlichkeitsverteilung mit R

Bestimmen Sie mit R die Wahrscheinlichkeitsverteilung des Beispiels:

„rot“ gewinnt a mal	Wahrscheinlichkeitsverteilung
0	<code>dbinom(0, size=5, prob=18/37) =0.036</code>
1	<code>dbinom(1, size=5, prob=18/37) =0.170</code>
2	<code>dbinom(2, size=5, prob=18/37) =0.320</code>
3	
4	
5	

Binomiale Wahrscheinlichkeitsverteilung mit R

Bestimmen Sie mit R die Wahrscheinlichkeitsverteilung des Beispiels:

„rot“ gewinnt a mal	Wahrscheinlichkeitsverteilung
0	<code>dbinom(0, size=5, prob=18/37) =0.036</code>
1	<code>dbinom(1, size=5, prob=18/37) =0.170</code>
2	<code>dbinom(2, size=5, prob=18/37) =0.320</code>
3	<code>dbinom(3, size=5, prob=18/37) =0.304</code>
4	
5	

Binomiale Wahrscheinlichkeitsverteilung mit R

Bestimmen Sie mit R die Wahrscheinlichkeitsverteilung des Beispiels:

„rot“ gewinnt a mal	Wahrscheinlichkeitsverteilung
0	<code>dbinom(0, size=5, prob=18/37) =0.036</code>
1	<code>dbinom(1, size=5, prob=18/37) =0.170</code>
2	<code>dbinom(2, size=5, prob=18/37) =0.320</code>
3	<code>dbinom(3, size=5, prob=18/37) =0.304</code>
4	<code>dbinom(4, size=5, prob=18/37) =0.144</code>
5	

Binomiale Wahrscheinlichkeitsverteilung mit R

Bestimmen Sie mit R die Wahrscheinlichkeitsverteilung des Beispiels:

„rot“ gewinnt a mal	Wahrscheinlichkeitsverteilung
0	<code>dbinom(0, size=5, prob=18/37) =0.036</code>
1	<code>dbinom(1, size=5, prob=18/37) =0.170</code>
2	<code>dbinom(2, size=5, prob=18/37) =0.320</code>
3	<code>dbinom(3, size=5, prob=18/37) =0.304</code>
4	<code>dbinom(4, size=5, prob=18/37) =0.144</code>
5	<code>dbinom(5, size=5, prob=18/37) =0.027</code>

Binomiale Wahrscheinlichkeitsverteilung mit R

```
> yprob<-dbinom(0:5, size=length(0:5)-1, prob = 17/38)
```

Binomiale Wahrscheinlichkeitsverteilung mit R

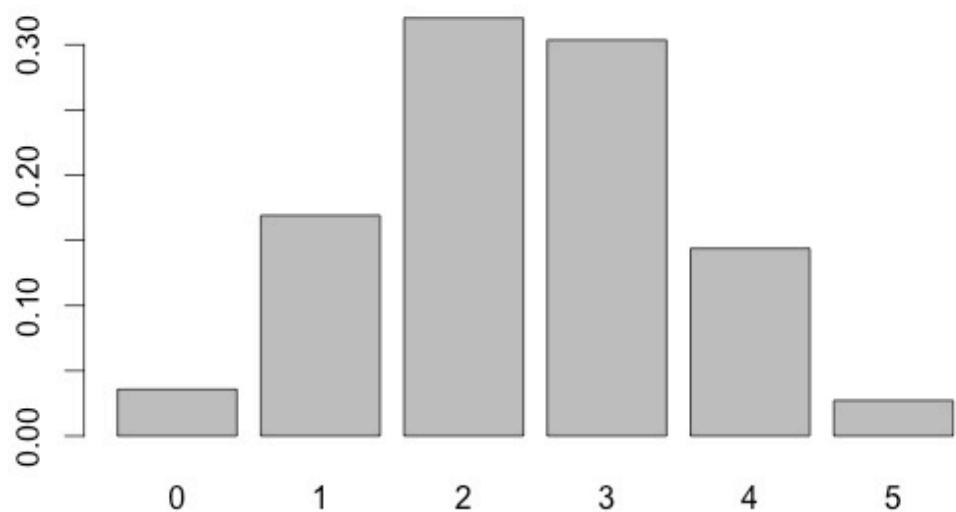
```
> yprob<-dbinom(0:5, size=length(0:5)-1, prob = 17/38)  
> names(yprob)<-0:5
```

Binomiale Wahrscheinlichkeitsverteilung mit R

```
> yprob<-dbinom(0:5, size=length(0:5)-1, prob = 17/38)
> names(yprob)<-0:5
> barplot(yprob)
```

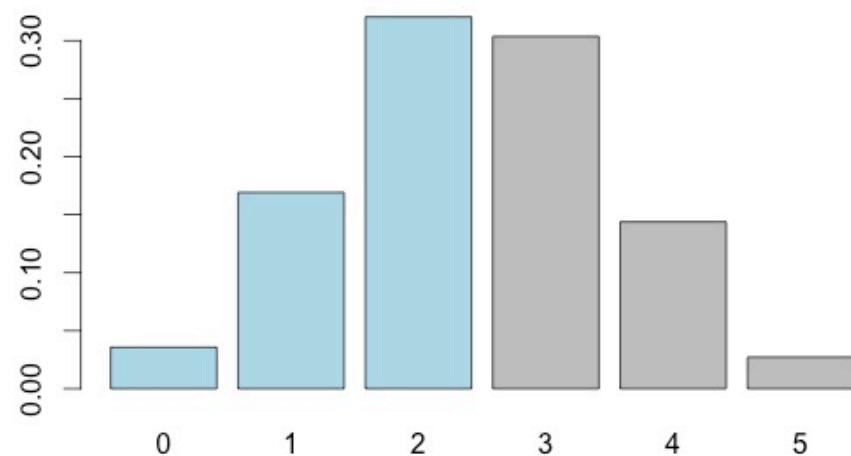
Binomiale Wahrscheinlichkeitsverteilung mit R

```
> yprob<-dbinom(0:5, size=length(0:5)-1, prob = 17/38)  
> names(yprob)<-0:5  
> barplot(yprob)
```



Kummulierte binomiale Wahrscheinlichkeit mit R

Beispiel: Wie gross ist bei fünfmaligem Setzen auf „rot“ die Wahrscheinlichkeit, dass man öfter verliert als gewinnt?



Kummulierte binomiale Wahrscheinlichkeit mit R

- Diese Wahrscheinlichkeit ist $P(x = 0) + P(x = 1) + P(x = 2)$.

Kummulierte binomiale Wahrscheinlichkeit mit R

- Diese Wahrscheinlichkeit ist $P(x = 0) + P(x = 1) + P(x = 2)$.
- Wir bestimmen diesen Wert mit

```
dbinom(0, size=5, prob=18/37) +  
dbinom(1, size=5, prob=18/37) +  
dbinom(2, size=5, prob=18/37)
```

Kummulierte binomiale Wahrscheinlichkeit mit R

- Diese Wahrscheinlichkeit ist $P(x = 0) + P(x = 1) + P(x = 2)$.
- Wir bestimmen diesen Wert mit

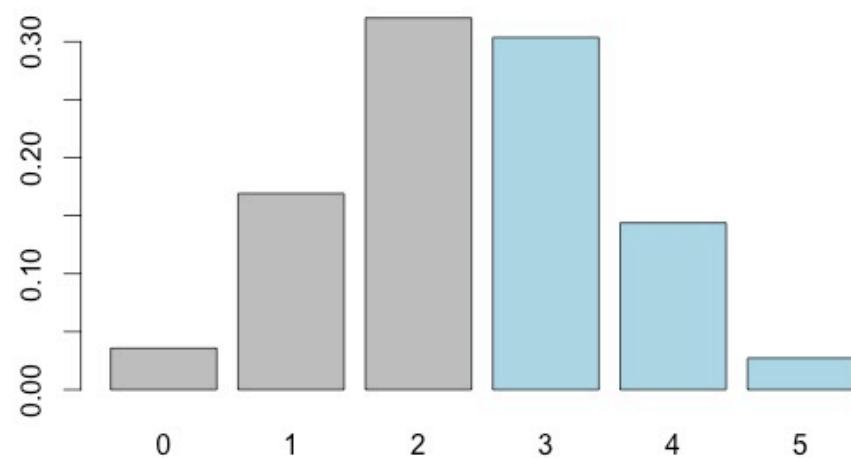
```
dbinom(0, size=5, prob=18/37) +  
dbinom(1, size=5, prob=18/37) +  
dbinom(2, size=5, prob=18/37)
```

- Kummulierte Wahrscheinlichkeiten bestimmen sich in R mit

```
pbisnom(2, size=5, prob=18/37)
```

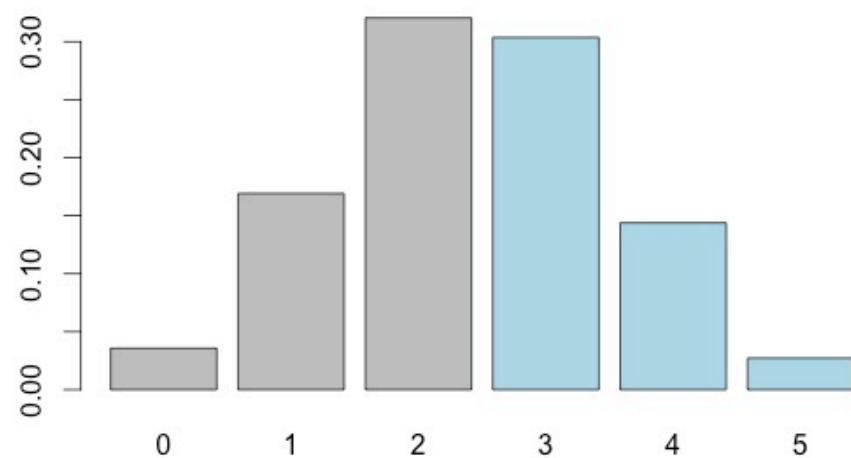
Kummulierte binomiale Wahrscheinlichkeit mit R

Beispiel: Wie gross ist bei fünfmaligem Setzen auf „rot“ die Wahrscheinlichkeit, dass man öfter gewinnt als verliert?



Kummulierte binomiale Wahrscheinlichkeit mit R

Beispiel: Wie gross ist bei fünfmaligem Setzen auf „rot“ die Wahrscheinlichkeit, dass man öfter gewinnt als verliert?

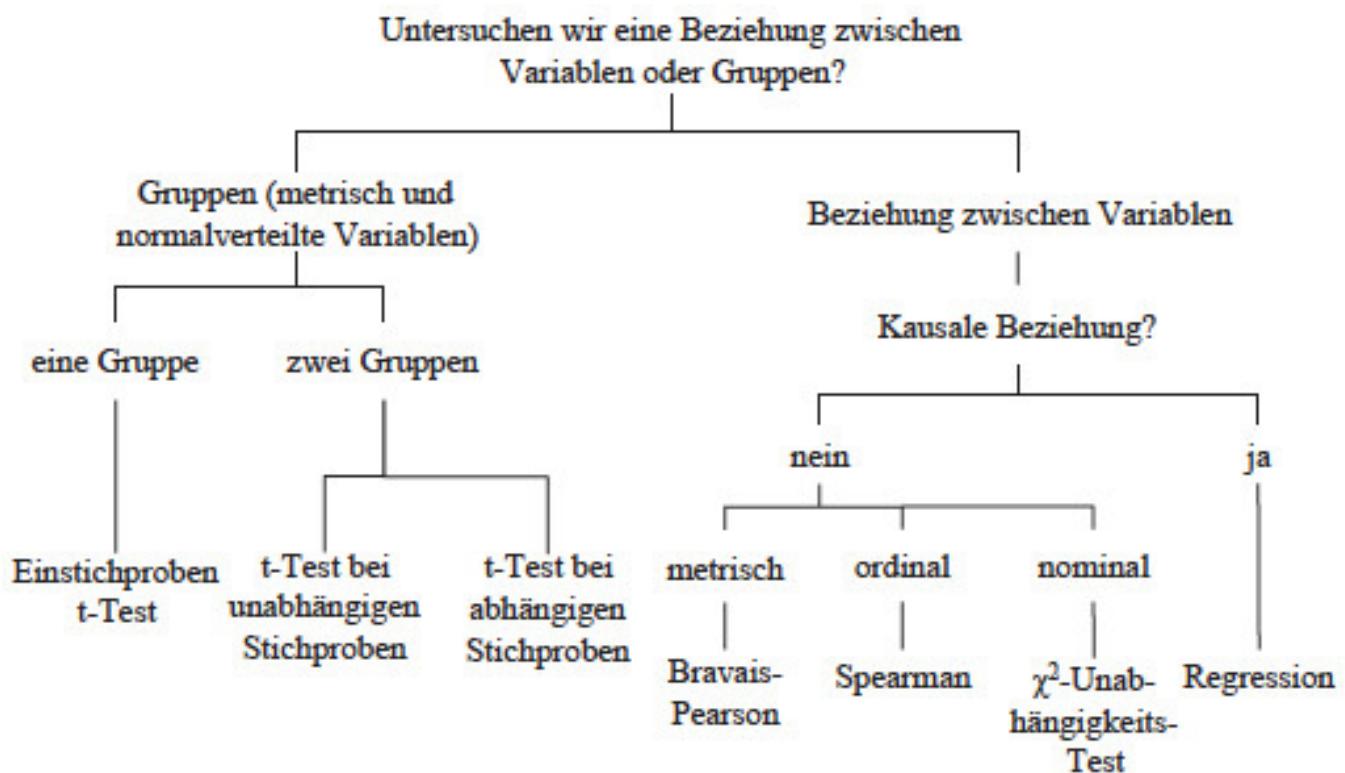


```
> pbinom(2, size=5, prob = 18/37, lower.tail = FALSE)
```

Inverse binomiale Wahrscheinlichkeit mit R

Beispiel: Welche Anzahl der Gewinne wird in 90% der Fälle höchstens erreicht?

Testverfahren



Testverfahren

Test	Wann der Test benutzt wird?	Beispiele
<i>Untersuchung einer Gruppe</i>		
Einstichproben t-Test	Untersuchung eines Mittelwertes bei metrischen, normalverteilten Variablen	Wie alt sind Unternehmensgründer im Durchschnitt?
Wilcoxon Test	Untersuchung eines Mittelwertes bei einer kleinen Stichprobe und metrisch nicht-normalverteilten Variablen oder bei ordinalen Variablen	Wie viel Branchenberufserfahrung haben Gründer im Durchschnitt?
χ^2 -Test	Vergleich der beobachteten Häufigkeiten mit den theoretisch erwarteten Häufigkeiten bei nominalen Variablen	Entspricht der Anteil der Gründerinnen an den gründenden Personen insgesamt dem Geschlechterverhältnis in der Gesellschaft?
<i>Untersuchung von zwei unabhängigen Gruppen</i>		
t-Test bei unab- hängigen Stichproben	Untersuchung auf einen Unterschied der Mittelwerte von Gruppen bei metrischen, normalverteilten Variablen	Gibt es bei der Gründung von Unternehmen einen Altersunterschied zwischen Männern und Frauen?
Mann-Whitney Test	Untersuchung auf einen Unterschied der Mittelwerte von Gruppen bei kleiner Stichprobe und metrisch nicht-normalverteilten Variablen oder bei ordinalen Variablen	Gibt es einen Unterschied in der Einschätzung der Branchenberufserfahrung zwischen Gründern und Gründerinnen?
χ^2 -Test	Untersuchung, ob sich zwei Gruppen hinsichtlich der Häufigkeitsanteile einer nominalen Variable unterscheiden	Unterscheiden sich Gründer von Dienstleistungsunternehmen und solche von Industrieunternehmen hinsichtlich ihrer Gründungsmotive?

Testverfahren

<i>Untersuchung von zwei abhängigen Gruppen</i>		
t-Test bei abhängigen Stichproben	Untersuchung, ob es in einer Gruppe vor und nach einer Maßnahme Unterschiede gibt, bei metrischen, normalverteilten Variablen	Hat eine Werbemaßnahme einen Einfluss auf das Image meines Produktes (metrisch gemessen)?
Wilcoxon Test bei abhängigen Stichproben	Untersuchung, ob es in einer Gruppe vor und nach einer Maßnahme Unterschiede gibt, bei kleiner Stichprobe und metrisch nicht-normalverteilten Variablen oder bei ordinalen Variablen	Hat eine Werbemaßnahme einen Einfluss auf das Image meines Produktes (ordinal gemessen)?
McNemar χ^2 -Test	Untersuchung, ob es in einer Gruppe vor und nach einer Maßnahme Unterschiede gibt, bei nominalen Variablen	Hat die Werbemaßnahme einen Einfluss darauf hat, ob Personen mein Produkt kaufen oder nicht?

Testverfahren

Test	Wann der Test benutzt wird?	Beispiele
<i>Untersuchung auf Korrelation zwischen zwei Variablen bzw. auf Unabhängigkeit</i>		
Bravais-Pearson	Untersuchung, ob zwischen zwei metrischen Variablen ein Zusammenhang besteht	Gibt es einen Zusammenhang zwischen dem Aufwand, den ein Jungunternehmer für Marketing betreibt und dem für Produktverbesserungen?
Spearman	Untersuchung, ob zwischen ordinalen Variablen ein Zusammenhang besteht	Gibt es einen Zusammenhang zwischen der Erwartung hinsichtlich der zukünftigen Unternehmensentwicklung und der Selbsteinschätzung der eigenen Unternehmensführungskenntnisse?
χ^2 -Unabhängigkeitstest	Untersuchung, ob zwischen nominalen Variablen ein Zusammenhang besteht	Gibt es einen Zusammenhang zwischen der Branche, in der gegründet wird und dem Gründungsmotiv

Inverse binomiale Wahrscheinlichkeit mit R

Beispiel: Welche Anzahl der Gewinne wird in 90% der Fälle höchstens erreicht?

- Wir kennen die Wahrscheinlichkeit und suchen den zugehörigen Wert auf der x -Achse.

Inverse binomiale Wahrscheinlichkeit mit R

Beispiel: Welche Anzahl der Gewinne wird in 90% der Fälle höchstens erreicht?

- Wir kennen die Wahrscheinlichkeit und suchen den zugehörigen Wert auf der x -Achse.
- Auch diese Berechnung ist in R abgespeichert:

```
qbinom(0.9, size=5, prob=18/37)
```

Hypergeometrische Verteilung mit R

Beispiel: Beim Schweizer Zahlenlotto sind 6 Zahlen aus 42 zu ziehen.

Wir bezeichnen mit x die Anzahl der richtig angekreuzten Zahlen.

Bestimmen Sie die Wahrscheinlichkeitsverteilung und stellen Sie diese grafisch dar.

Hypergeometrische Verteilung mit R

- R verwendet bei der hypergeometrischen Verteilung die Notation

$$P(x = a) = \frac{\binom{m}{x} \cdot \binom{n}{k-x}}{\binom{N}{k}}$$

Hypergeometrische Verteilung mit R

- R verwendet bei der hypergeometrischen Verteilung die Notation

$$P(x = a) = \frac{\binom{m}{x} \cdot \binom{n}{k-x}}{\binom{N}{k}}$$

- Wir finden somit:

Hypergeometrische Verteilung mit R

- R verwendet bei der hypergeometrischen Verteilung die Notation

$$P(x = a) = \frac{\binom{m}{x} \cdot \binom{n}{k-x}}{\binom{N}{k}}$$

- Wir finden somit:

```
> ylotto<-dhyper(0:6, m=6, n=39, k=6)
```

Hypergeometrische Verteilung mit R

- R verwendet bei der hypergeometrischen Verteilung die Notation

$$P(x = a) = \frac{\binom{m}{x} \cdot \binom{n}{k-x}}{\binom{N}{k}}$$

- Wir finden somit:

```
> ylotto<-dhyper(0:6, m=6, n=39, k=6)  
> names(ylotto)<-0:6
```

Hypergeometrische Verteilung mit R

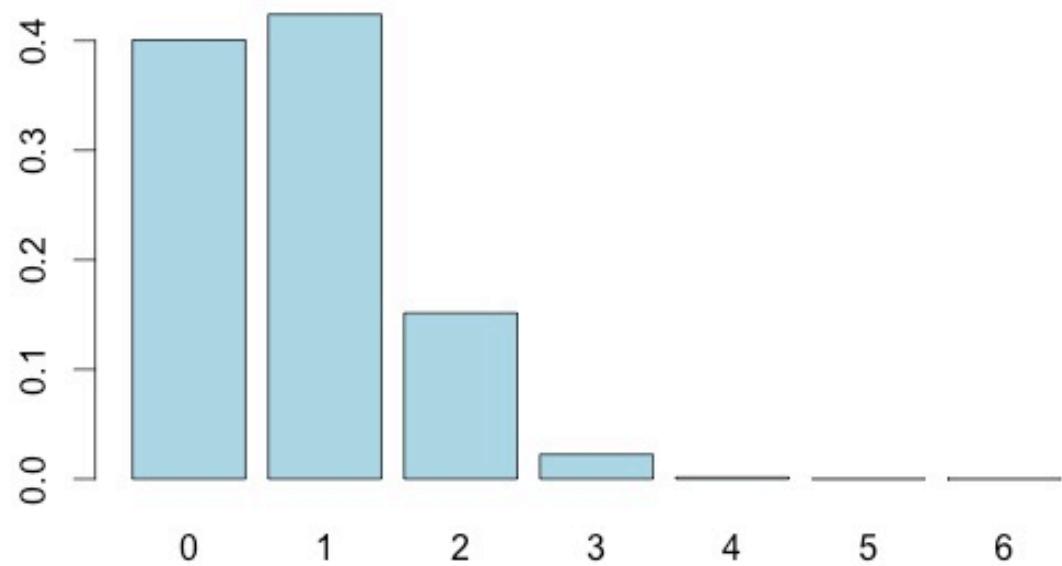
- R verwendet bei der hypergeometrischen Verteilung die Notation

$$P(x = a) = \frac{\binom{m}{x} \cdot \binom{n}{k-x}}{\binom{N}{k}}$$

- Wir finden somit:

```
> ylotto<-dhyper(0:6, m=6, n=39, k=6)
> names(ylotto)<-0:6
> barplot(ylotto)
```

Hypergeometrische Verteilung mit R



Übungen zur Hypergeometrischen Verteilung mit R

- Wie gross ist beim Schweizer Zahlenlotto die Wahrscheinlichkeit, keinen Gewinn zu erzielen?
- Wie gross ist der Anteil, tatsächlich einen Gewinn zu erzielen?
- Wie viele richtige Zahlen kreuzen die besten 1% der Spieler an?

Lösungen

- Wie gross ist beim Schweizer Zahlenlotto die Wahrscheinlichkeit, keinen Gewinn zu erzielen?
> `phyper(2, m=6, n=39, k=6)`
- Wie gross ist der Anteil, tatsächlich einen Gewinn zu erzielen?

Lösungen

- Wie gross ist beim Schweizer Zahlenlotto die Wahrscheinlichkeit, keinen Gewinn zu erzielen?
> `phyper(2, m=6, n=39, k=6)`
- Wie gross ist der Anteil, tatsächlich einen Gewinn zu erzielen?
> `phyper(2, m=6, n=39, k=6, lower.tail=FALSE)`
- Wie viele richtige Zahlen kreuzen die besten 1% der Spieler an?

Lösungen

- Wie gross ist beim Schweizer Zahlenlotto die Wahrscheinlichkeit, keinen Gewinn zu erzielen?

> `phyper(2, m=6, n=39, k=6)`

- Wie gross ist der Anteil, tatsächlich einen Gewinn zu erzielen?

> `phyper(2, m=6, n=39, k=6, lower.tail=FALSE)`

- Wie viele richtige Zahlen kreuzen die besten 1% der Spieler an?

> `qhyper(0.99, m=6, n=39, k=6)`

oder

> `qhyper(0.019, m=6, n=39, k=6, lower.tail=FALSE)`

Regressionsanalyse

- Versuchen wir die Wirkung einer unabhängigen Variable x auf eine abhängige Variable y zu modellieren, können wir diese Abhängigkeit z.B. mit einer linearen Regression beschreiben.

$$y = m \cdot x + b$$

Regressionsanalyse

- Versuchen wir die Wirkung einer unabhängigen Variable x auf eine abhängige Variable y zu modellieren, können wir diese Abhängigkeit z.B. mit einer linearen Regression beschreiben.

$$y = m \cdot x + b$$

- Dazu legen wir in die Punktwolke eine Gerade, die nach der Methode der kleinsten Quadrate bestimmt wird.

Regressionsanalyse

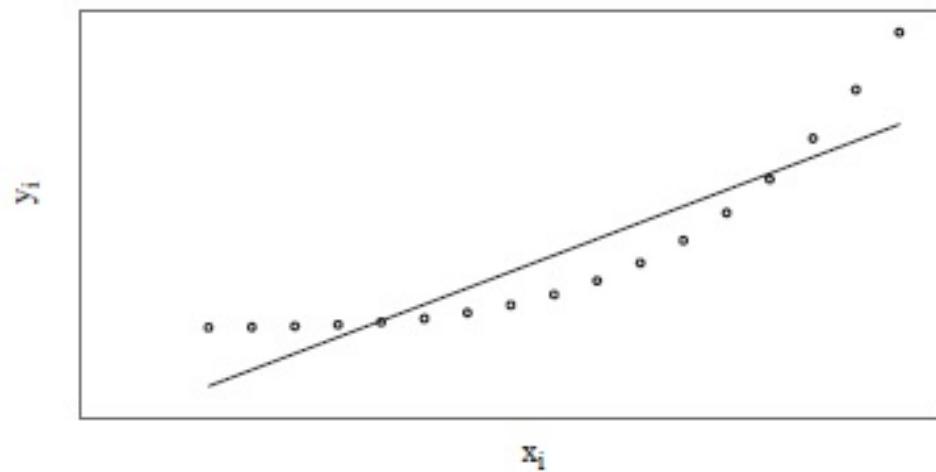
- Versuchen wir die Wirkung einer unabhängigen Variable x auf eine abhängige Variable y zu modellieren, können wir diese Abhängigkeit z.B. mit einer linearen Regression beschreiben.

$$y = m \cdot x + b$$

- Dazu legen wir in die Punktwolke eine Gerade, die nach der Methode der kleinsten Quadrate bestimmt wird.
- Die Regressionsanalyse muss theoretisch fundiert sein!

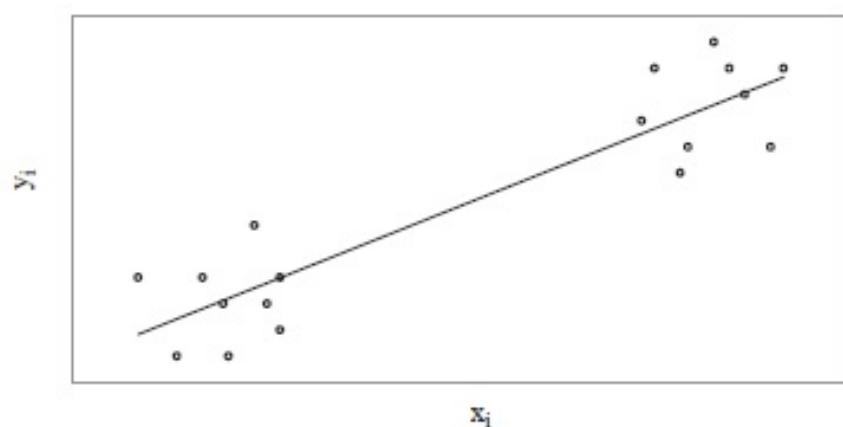
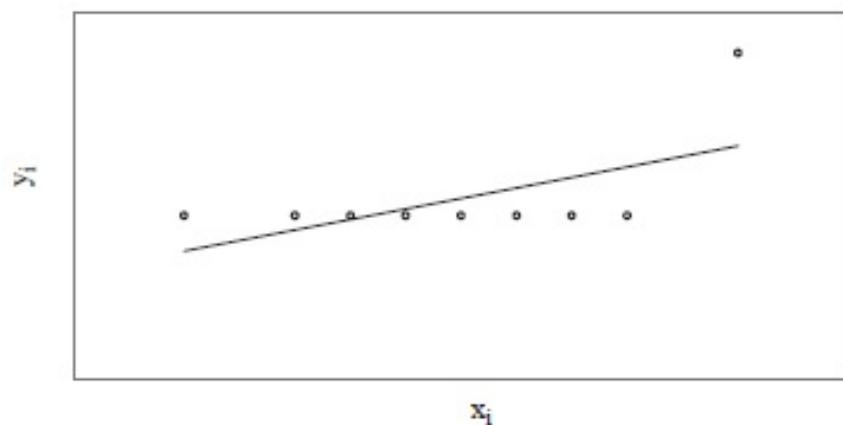
Vorsicht bei der Regression: lineare Wolke

- Die Punktewolke muss ein lineares Muster zeigen.
- Nicht-lineare Regressionen sind möglich, aber nicht Bestandteil dieses Kurses.



Vorsicht bei der Regression: Ausreißer

- Ausreißer können die Gerade beeinflussen.



Vorsicht bei der Regression: Out-of-Sample

- Machen Sie keine Out-of-Sample Vorhersage!

