Berufsverband
Deutscher Markt- und
Sozialforscher e.V.

# Grundlagen der Statistik

01. und 02. Oktober 2018 in Berlin

Referent:

**Dr. Johannes Gladitz**

*(Statistik-Service, Berlin)*

# Statistik - umgangssprachlich

## Notlüge - Lüge - Statistik

"Ich glaube nur einer Statistik, die ich selbst gefälscht habe!"

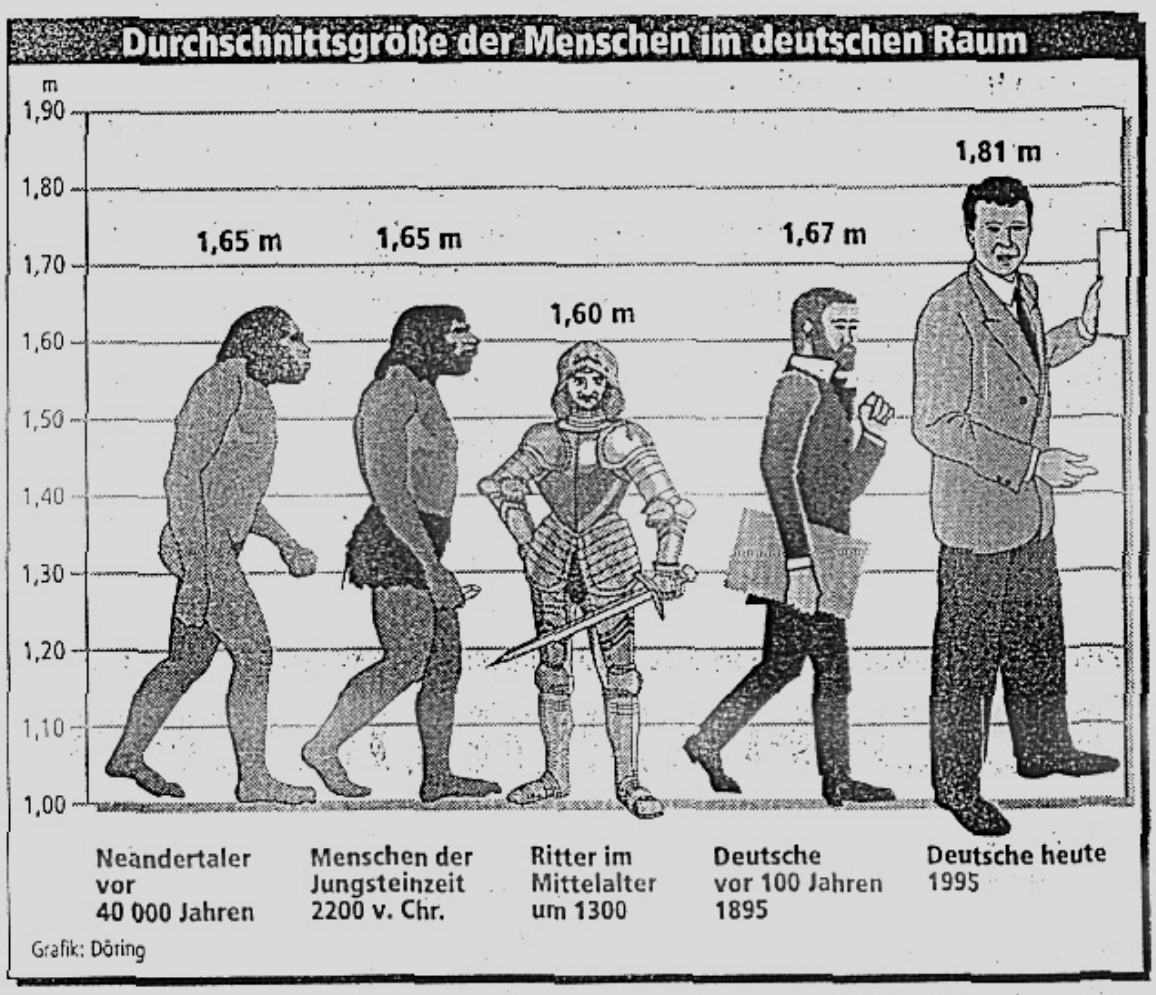
Die Statistik ist in Verruf geraten durch

- ⇨ die unkritische Zahlengläubigkeit der Menschen
- ⇨ mangelnden Einblick in statistische Methoden
- ⇨ Scharlatanerie (dadurch begünstigt)
- ⇨ statistische Manipulationen

zu empfehlen:

Walter Krämer (2011): "So lügt man mit Statistik"
(Piper Verlag GmbH, München, 4. Auflage)
ISBN: 3492264131

# Lügen mit Statistik




Über Jahrtausende hat sich die Körperhöhe kaum verändert. Erst seit gut 100 Jahren ist ein Wachsstumsschub zu beobachten. Grafik: Döring

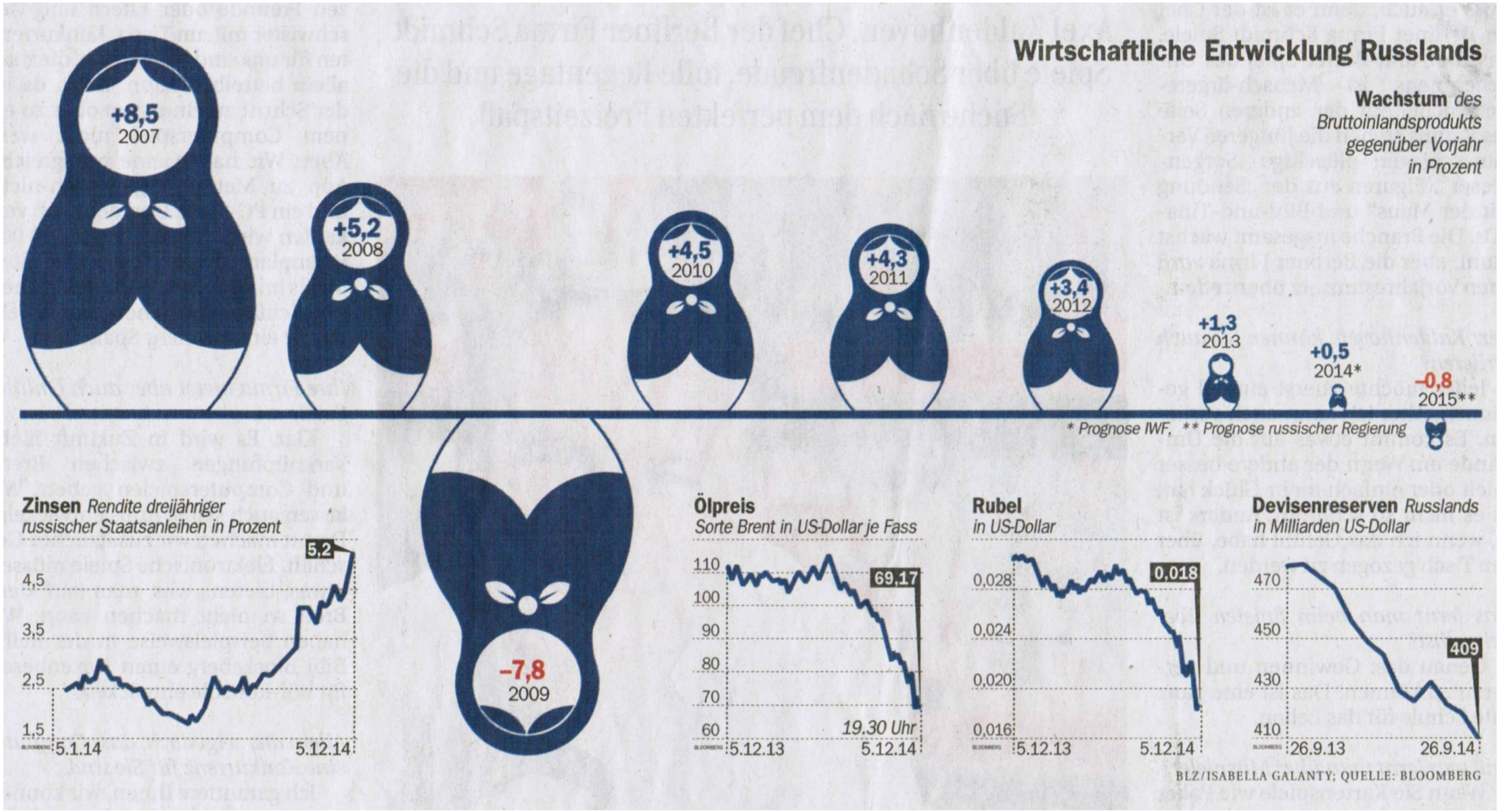
So lügt man mit Statistik!

Wo wurde hier manipuliert?

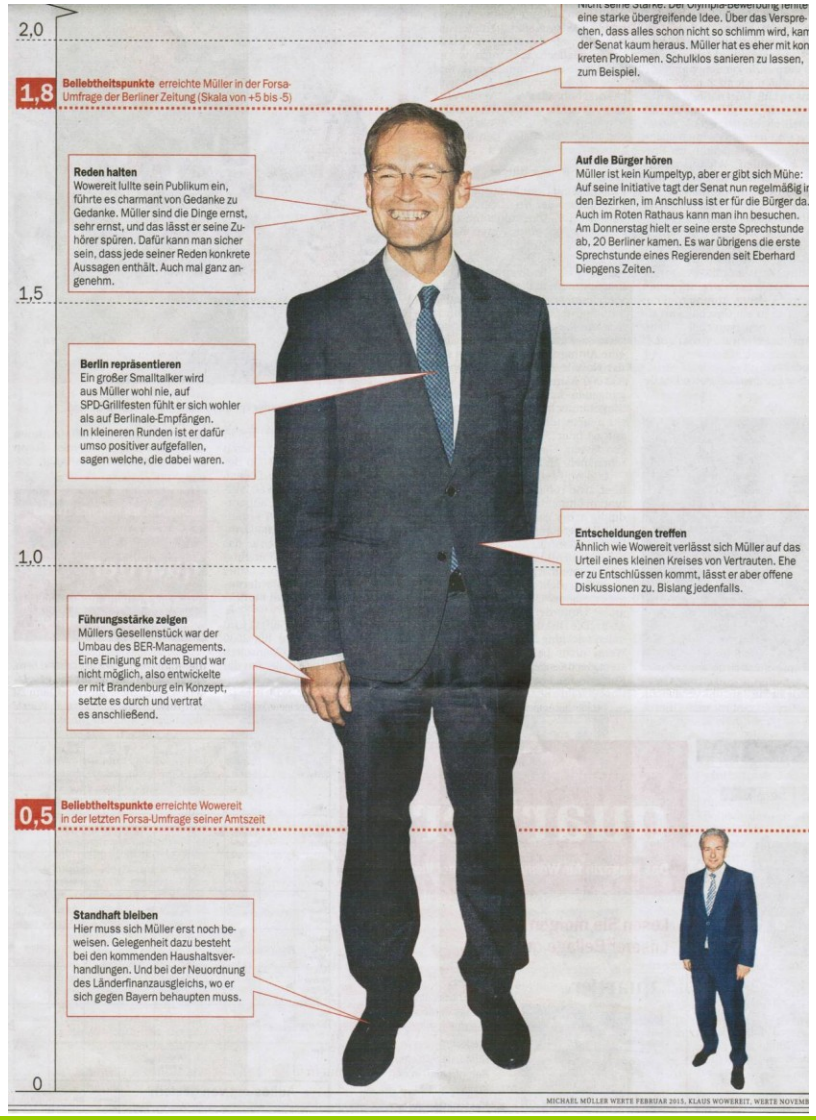
Aus der Berliner Zeitung

# Lügen mit Statistik

Ein weiteres Beispiel: So lügt man mit Statistik!



# Lügen mit Statistik




Aus der Berliner Zeitung vom 20.03.2015

Artikel von Jan Thomsen und Frederik Bombosch

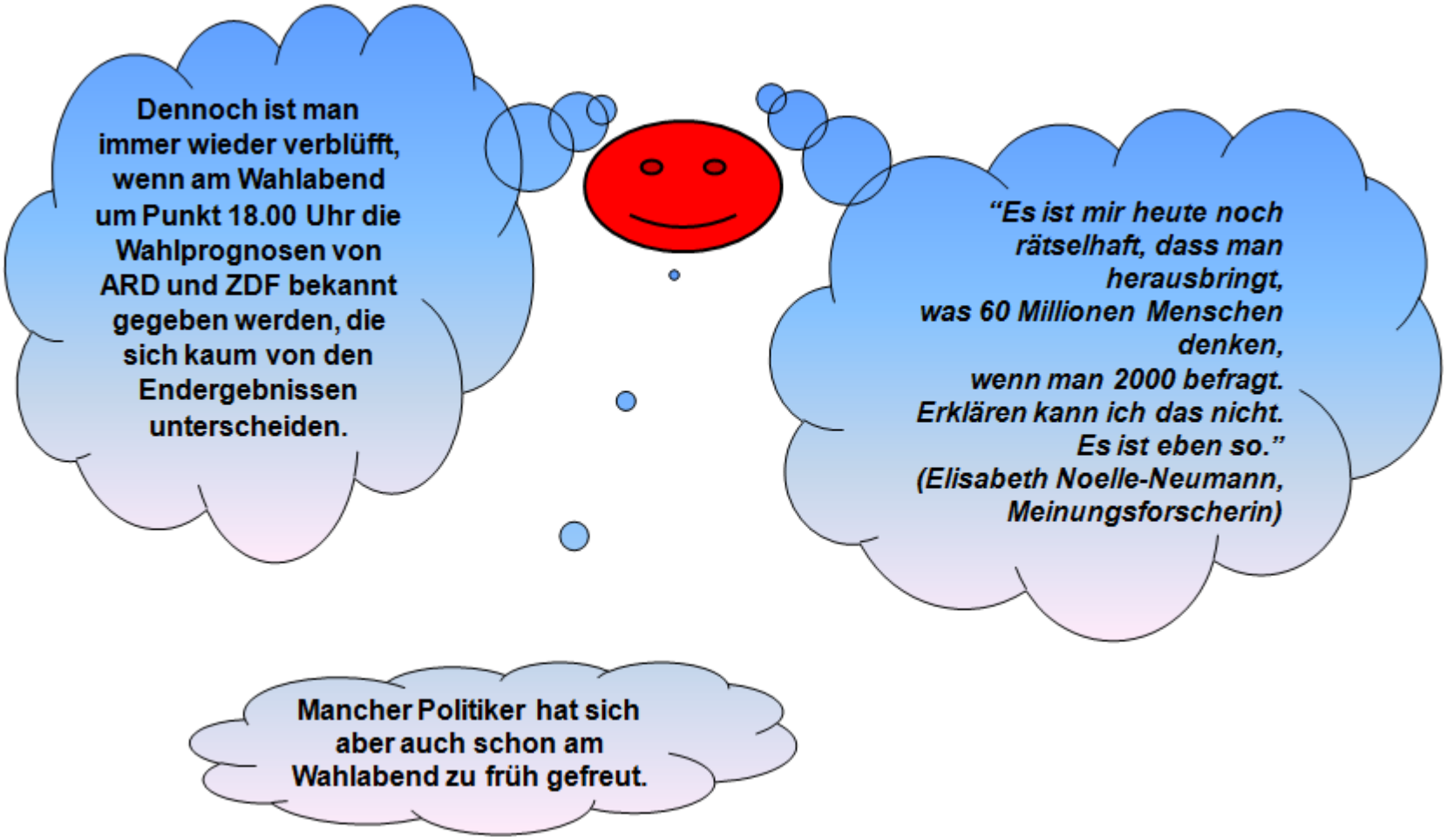
Die Befragten konnten auf einer Skala von -5 bis +5 antworten.

Das ist keine Verhältnisskala! Es gibt keine natürliche Null.

Hinzu kommt, dass eine Flächendarstellung verwendet wurde, also mit der Höhe wächst die Fläche quadratisch.

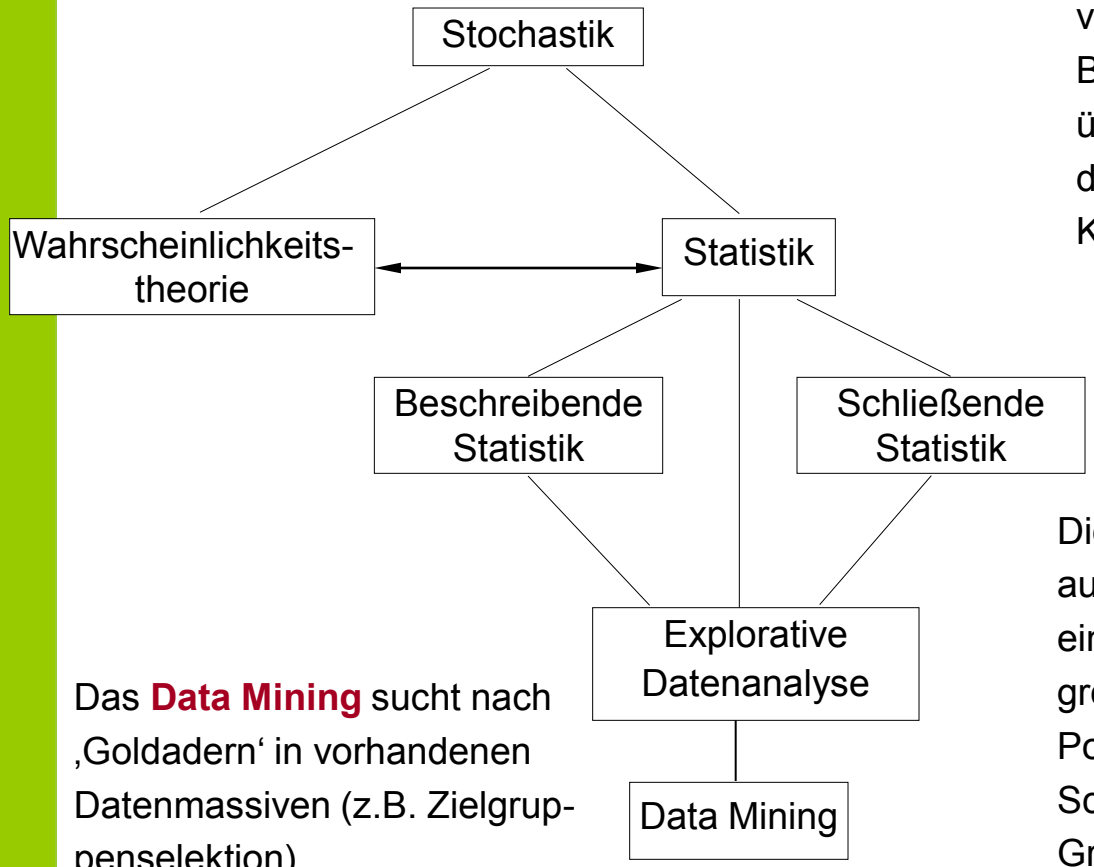
„So lügt man mit Statistik“

# Statistik - Wahlabend



# Statistik als Wissenschaftsdisziplin

**Überblick**




Die **beschreibende Statistik** versucht, vorhandene Beobachtungen oder Daten übersichtlich zusammenzufassen, d.h. einen Informationsgewinn durch Komplexitätsreduktion zu erreichen.

Die **schließende Statistik** versucht, aus der Beobachtung und Analyse einer Stichprobe, die zufällig einer größeren Grundgesamtheit oder Population entnommen wurde, Schlüsse über die gesamte Grundgesamtheit zu ziehen.

Das **Data Mining** sucht nach ‚Goldadern' in vorhandenen Datenmassiven (z.B. Zielgruppenselektion).

# Statistik als Wissenschaftsdisziplin

## **Big Data** – Was versteht man darunter?

Gartner IT Glossary: „Big data is **high-volume**, **high-velocity** and **high-variety** information assets that demand cost- effective, innovative forms of information processing for enhanced insight and decision making“. Abgerufen am 10. Oktober 2016 von: http://www.gartner.com/it-glossary/big-data

**Herausforderungen**:

- Verarbeitung vieler Datensätze
- Verarbeitung vieler Spalten innerhalb eines Datensatzes
- Schneller Import großer Datenmengen
- Sofortige Abfrage importierter Daten (Realtime Processing)
- Kurze Antwortzeiten auch bei komplexen Abfragen
- Möglichkeit zur Verarbeitung vieler gleichzeitiger Abfragen (Concurrent Queries)
- Analyse verschiedenartiger Informationstypen (Zahlen, Texte, Bilder, …)

**Statistische Algorithmen** auch hier verwendet, aber auch aus Künstlicher Intelligenz, maschinellem Lernen, Linguistik, Bildverarbeitung, Spracherkennung …

# Beschreibende und schließende Statistik

## Beschreibende Statistik = Deskriptive Statistik

Beschreibung von meist unübersichtlichen Erhebungsdaten, indem

- Prozente
- Mittelwerte
- Streuparameter
- Zusammenhangsmaße etc.

tabellarisch und grafisch dargestellt werden.

## Schließende Statistik = Inferenzstatistik

Auf der Grundlage einer Stichprobe

- Prognosen
- Schätzungen und Konfidenzintervalle
- Statistische Signifikanztests etc.

über Verteilungsparameter in der Grundgesamtheit

# Was macht schließende Statistik?

Ziel der schließenden Statistik ist
die Gewinnung von Informationen
über **Wahrscheinlichkeitsverteilungen**
in der Grundgesamtheit.

Schließende Statistik ist die
**Quantifizierung der Ungewissheit**
bei zufallsbehafteten Erscheinungen.

J. Gladitz

# Aufgaben und Erkenntniswert

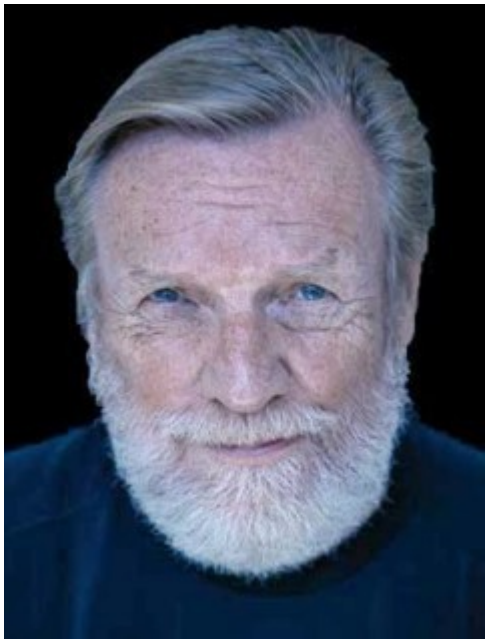
## Beispiele:

- **Wahlprognosen** vor der Wahl und am Wahlabend (schließend)
- Monatliche Veröffentlichung der **Arbeitslosenzahlen** (deskriptiv)
- **Quotenermittlung** bei Fernsehsendern (schließend)
- **Mietspiegel** für eine Großstadt (deskriptiv und schließend)
- Ermittlung der **Zufriedenheit** der Passagiere im ÖPNV mittels einer Umfrage, Bonus-Malus-relevant für Subventionen (schließend)
- Internationales Tracking von **Markenbekanntheit** und -**stärke** mittels Umfragen zur Steuerung der Marketinginvestitionen (schließend)
- Finden von marketing-relevanten **Zielgruppen** im stationären und Distanzhandel (Business Analytics; schließend)
- Prognose der **Konjunkturentwicklung** in Deutschland (demagogisch, weil meistens falsch; es geht nicht, weil zu komplex)

# Aufgaben und Erkenntniswert

**John Naisbitt: * 1929, amer. Zukunftsforscher:**

**(Wien & Tianjin, China)**



**„Wir ertrinken in Informationen und wir dürsten nach Wissen."**

**Methoden der Statistik fördern aus Informationen Wissen, eine allgemein unterschätzte Wissenschaft!**

# Aufgaben und Erkenntniswert

## Erkenntniswert:

- Durch eine sinnvolle **Datenaggregation** werden Informationen erst wahrnehmbar und Botschaften transportiert.
- Die **moderne Kommunikation** ist voll von statistischen Begriffen, die heute zur Allgemeinbildung gehören sollten. Die Medien sind überfrachtet mit Prozenten, Durchschnitten, Indizes, Summen usw. Die allgemeine Schulbildung hat sich darauf ungenügend eingestellt.
- Fast alle **Entscheidungen** müssen **unter Ungewissheit** über Teilaspekte gefällt werden. Die Statistik liefert Instrumente, um Möglichkeiten und Wahrscheinlichkeiten quantifiziert bei komplexen Entscheidungen zu berücksichtigen. Manager sind ungenügend darauf vorbereitet.
- Die Medizin, die Psychologie, die Sozialwissenschaften, die Wirtschaftwissenschaften, alle Wissenschaften, die mit empirischen Daten arbeiten, sind mit dem Zufall konfrontiert. **Wesentliche Einflüsse sind von zufälligen zu trennen**. **Dabei hilft die Statistik**.
- … und es gibt weitere unzählige Beispiele.

# Statistik im Dienste anderer Wissenschaften

- Die Markt- und Sozialforschung lebt von der Statistik
- Psychologie (Faktorenanalyse)
- Pharmaforschung: Klinische Studien (Dosis-Wirkung)
- Landwirtschaft: Feldversuchswesen (Varianzanalyse)
- Kriminologie: Wurzeln der Kriminalität
- Banken und Versicherungen (Credit Scoring, Data Mining)
- Börsenmakler (Chart Analysen)
- Verarbeitende Industrie (Statistische Qualitätskontrolle)
- Eich- und Meßwesen
- Unternehmensberatung (Strategiebildung)
- Ökonometrie (Konjunkturprognosen)
- Zoo- und Wildtierforschung (Verhaltensanalysen)
- Fischereiwesen (Analyse von Gewässerproben)
- Human- und Veterinärmedizin (Epidemiologie)
- Zuverlässigkeitstheorie bei komplexen Konstruktionen
- .....

# Wichtige Grundbegriffe

## Grundbegriffe: Grundgesamtheit, Population

Alle Objekte, auf die sich die Untersuchung bezieht, für welche die Ergebnisse der Untersuchung später gelten sollen

**Beispiele**

- alle Kunden und potenziellen Kunden eines Unternehmens
- alle Wahlberechtigten in Deutschland
- alle Zulieferer eines Autoherstellers
- alle Adressen eines Versandhauses
- alle Autos in einem Wohngebiet
- alle Mitarbeiter eines Unternehmens …

**Vor einer Untersuchung muss klar sein, für welche Grundgesamtheit die Aussagen später zutreffen sollen.**

# Wichtige Grundbegriffe

## Grundbegriffe: Erhebung

Wie werden die Informationen erhoben.

| Art | Beschreibung |
|---|---|
| **Vollerhebung** | **Über alle Elemente der Grundgesamtheit werden Daten erhoben.** |
| **Teilerhebung** | **Nur über einen Teil aller Element der Grundgesamtheit werden Daten erhoben.** |
| **Primärerhebung** | **Die Daten werden direkt an den Elementen der Grundgesamtheit erfasst.** |
| **Sekundärerhebung** | **Es wird auf schon erfasste Daten einer früheren Erhebung zurückgegriffen.** |

# Wichtige Grundbegriffe

## Grundbegriffe: Repräsentativität

bezieht sich immer auf eine Grundgesamtheit, die klar definiert sein muss.

- Repräsentativ ist eine Stichprobe dann, wenn sie in Bezug auf die Verteilung der erhobenen Merkmale ein wirklichkeitsgetreues Abbild der Grundgesamtheit darstellt.
- Repräsentativität ist praktisch schwer nachprüfbar.
- Repräsentativ wird eine Stichprobe dann, wenn die Auswahl der Elemente der Grundgesamtheit zufällig erfolgt (jedes Element der GG die gleiche Chance hatte, Mitglied der Stichprobe zu werden).
- Das ist meist nicht zu realisieren. Deshalb wurden allerhand Verfahren zur Stichprobenplanung entwickelt, die sich diesem Ziel nähern.

# Wichtige Grundbegriffe

## Methoden der Stichprobenziehung

- Reine Zufallsauswahl (bei großen Stichproben)
- Einstufig oder mehrstufig (z.B. Sample Points)
- Schichtung mit Quotenvorgabe (bei kleineren Stichproben)
- Klumpenstichprobe
- Bewusste Auswahl (z.B. für Gruppengespräche)

Lothar Kreienbrock (2004). Einführung in die Stichprobenverfahren, Neuauflage, Oldenbourg-Verlag. ISBN: 978-3-486-25998-8

# Wichtige Grundbegriffe

## Grundbegriffe: Merkmal (Variable)

Eigenschaften, Charakteristika, die untersucht werden sollen.

- Alter, Geschlecht, Bundesland, Zufriedenheit mit Leben, Umsatz für Lebensmittel pro Woche

## Grundbegriffe: Merkmalsausprägung

Werte, die ein Merkmal annehmen kann.

- Jahre, m/w, Bayern, unzufrieden (Note 5), 200 €,

## Grundbegriffe: Fall

Basiseinheit der Analysen, für die Merkmale beobachtet und Merkmalsausprägungen gemessen werden.

- Personen (jede Person ist ein Fall), Firmen, Staaten, bei Zeitreihen jeder Zeitpunkt oder Zeitraum ein Fall

# Wichtige Grundbegriffe

## Skalenniveaus von Merkmalen

### Nominal

nur verbal beschreibbar, keine Rangordnung, Zuordnung von Zahlen (Codes) zu Inhalten willkürlich.

Beispiele: Bundesland, Beruf, Geschlecht, PLZ, Partei

### Ordinal

nur verbal beschreibbar, aber Rangordnung möglich, Abstand zwischen Werten nicht klar definiert.

Beispiele: Schulnoten, Zufriedenheit (1=sehr zufrieden…5=sehr unzufried.), Zustimmungsgrade (1=bin voll dafür bis 4=bin voll dagegen)

### Metrisch

natürliche Skalen, Abstand macht Sinn, Unterscheidung zwischen verhältnisskalierten und intervallskalierten Daten.

Beispiele: Einkommen (€), Umsätze (€), Anzahl Bestellungen
gemessene Temperatur in °C (nur intervallskaliert)

# Wichtige Grundbegriffe

## Beispiel einer Datenmatrix

Merkmale / Variablen

Fälle

| | idnrneu | f1 | f2 | f3 | f4 | f5 | f6 | f7 | gesamt | bindg | alter | sex |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 74 | . | 4 | . | . | . | . | . | 5 | Nein | 41-50 | männlic |
| 2 | 79 | sehr | sehr | sehr | 2 | 2 | 2 | sehr | 2 | Nein | über 5 | weiblich |
| 3 | 35 | 3 | 5 | 5 | 5 | 4 | 6 | 6 | 5 | Nein | über 5 | weiblich |
| 4 | 24 | 3 | 4 | 3 | 4 | 5 | 4 | 2 | 3 | Nein | über 5 | weiblich |
| 5 | 32 | 3 | 5 | 4 | 5 | 3 | 2 | 2 | 4 | Ja | 21-30 | männlic |
| 6 | 3 | 4 | sehr | 6 | 5 | 6 | sehr | 3 | 4 | Nein | 31-40 | männlic |
| 7 | 85 | 4 | 6 | 5 | 5 | 6 | 5 | 6 | 6 | Ja | über 5 | weiblich |
| 8 | 30 | 4 | 5 | 5 | 5 | 5 | 5 | 3 | 6 | Ja | 41-50 | weiblich |
| 9 | 55 | 4 | 4 | 6 | 4 | 2 | 5 | sehr | 3 | Nein | 41-50 | weiblich |
| 10 | 53 | 4 | 5 | 5 | 6 | 5 | 4 | 3 | 5 | Nein | 21-30 | weiblich |
| 11 | 86 | 4 | 4 | 6 | 5 | 5 | 4 | sehr | 5 | Nein | 21-30 | männlic |
| 12 | 80 | 4 | 3 | 3 | 3 | 4 | 4 | sehr | 3 | Nein | 21-30 | weiblich |
| 13 | 95 | 4 | 2 | 5 | 3 | 3 | 3 | 4 | 4 | Ja | 41-50 | weiblich |
| 14 | 89 | 4 | 5 | 6 | 5 | 5 | 2 | 2 | 6 | Nein | 21-30 | weiblich |
| 15 | 41 | 4 | 4 | 5 | 4 | 4 | 2 | 5 | 4 | Nein | 41-50 | weiblich |
| 16 | 44 | 4 | 3 | 6 | 6 | 2 | 2 | 3 | 5 | Ja | 41-50 | weiblich |
| 17 | 90 | 4 | sehr | 5 | 4 | sehr | 2 | 2 | 3 | Nein | 31-40 | männlic |
| 18 | 72 | 5 | 5 | sehr | sehr | 6 | sehr | 4 | sehr zuf | Ja | 41-50 | weiblich |

# Wichtige Grundbegriffe

## Verteilungen in der Statistik

- Um eine Grundgesamtheit exakt zu beschreiben, könnte man sämtliche Daten über alle Mitglieder der Grundgesamtheit auflisten.
- Man kann sich oft ein ausreichendes Bild von der Grundgesamtheit verschaffen, wenn man für bestimmte wichtige (interessierende) Merkmale weiß, *welche Werte wie häufig auftreten*.
- Das führt zum Begriff der Verteilung, auch Wahrscheinlichkeitsverteilung genannt.
- *Eine Verteilung gibt an, welche Werte wie häufig auftreten*.
- Solche Verteilungen kann man für ein Merkmal angeben: z.B. Altersverteilung
- oder auch für mehrere: Altersverteilung für Frauen und Männer
- Monovariate (eindimensionale), Bivariate, Multivariate Verteilungen

# Wichtige Grundbegriffe

## Grundbegriffe: Wahrscheinlichkeit P

- ‚mit dem Schein der Wahrheit‘
- ‚So die Wahrscheinlichkeit etwas ist, zwischen dem Wahren und dem Falschen gleichsam mitten inne ist.‘ (Definition von Immanuel Kant)
- Die Wahrscheinlichkeit P ergibt sich als logische Wahrscheinlichkeit bei gleichmöglichen Ereignissen, z.B. Würfeln, z.B. Lotto spielen, Münze werfen.

- Wahrscheinlichkeit = Anzahl der günstigen / Anzahl der möglichen Ereignisse.
- P (Augenzahl gerade) = 3 / 6 = 1/2
- P (Augenzahl = 6) = 1/6

# Häufigkeit und Wahrscheinlichkeit

## Häufigkeit, relative Häufigkeit

- Hat man keine gleichwahrscheinlichen Ereignisse, definiert man die Wahrscheinlichkeit als Grenzwert der relativen Häufigkeit:
- n Versuche, bei denen H mal das Ereignis A eintritt
- H heißt dann absolute Häufigkeit, n ist der Stichprobenumfang
- die relative Häufigkeit:

$$h(A) = H(A) / n \xrightarrow[\text{n gegen Unendlich}]{} P(A)$$

## Gesetz der großen Zahlen

- Bei wachsender Anzahl von Versuchswiederholungen strebt die relative Häufigkeit gegen die Wahrscheinlichkeit.

# Häufigkeit und Wahrscheinlichkeit

## Würfeln: Ergebnis einer Simulation

| | n=10 | | n=100 | | n=1.000 | | n=10.000 | | n=30.000 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Punkte | Häufigkeit | Anteil | Häufigkeit | Anteil | Häufigkeit | Anteil | Häufigkeit | Anteil | Häufigkeit | Anteil |
| 1 | 1 | 10,0% | 18 | 18,0% | 177 | 17,7% | 1.641 | 16,4% | 4.898 | 16,3% |
| 2 | 1 | 10,0% | 13 | 13,0% | 147 | 14,7% | 1.640 | 16,4% | 4.970 | 16,6% |
| 3 | 2 | 20,0% | 18 | 18,0% | 158 | 15,8% | 1.640 | 16,4% | 4.979 | 16,6% |
| 4 | 2 | 20,0% | 19 | 19,0% | 183 | 18,3% | 1.720 | 17,2% | 5.077 | 16,9% |
| 5 | 3 | 30,0% | 17 | 17,0% | 176 | 17,6% | 1.667 | 16,7% | 4.991 | 16,6% |
| 6 | 1 | 10,0% | 15 | 15,0% | 159 | 15,9% | 1.692 | 16,9% | 5.085 | 17,0% |
| | 10 | 100,0% | 100 | 100,0% | 1.000 | 100,0% | 10.000 | 100,0% | 30.000 | 100,0% |

Zunächst werden die ersten 10 Würfe ausgezählt, dann 100, dann 1.000, 10.000 und 30.000. Jedes Mal werden die relativen Häufigkeiten berechnet.

**Die relative Häufigkeit nähert sich mit größer werdender Fallzahl der Wahrscheinlichkeit an. Diese ist 1/6 = 16,667%.**

# Einige Mathematische Symbole

| | |
|---|---|
| (groß)P | Symbol für Wahrscheinlichkeit |
| (klein)p | oft für Signifikanz verwendet |
| $X_i$ | Merkmal (zufällige Variable) |
| $X_i=a$ | Das Merkmal $X_i$ nimmt die Ausprägung a an |
| n | Stichprobenumfang |
| $\Sigma$ | Summenzeichen Beispiel: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ |
| $\bar{X}$ | arithmetisches Mittel |
| $S^2$ | Varianz |
| S | Standardabweichung |
| R | Korrelation nach Pearson |
| S.E. | Standard Error |

# Diskrete und stetige Verteilungen

**Diskrete Verteilung (kategoriale Verteilung)**

Man kann die Ausprägungen eines Merkmals zählen.

**Beispiele**:

Verteilung vom Männern und Frauen in der GG

Verteilung der Gesamtzufriedenheit in der GG

Verteilung der Augenzahl beim Würfeln mit 2 Würfeln

**Stetige Verteilung (kontinuierliche Verteilung)**

Es gibt theoretisch unendlich viele denkbare Werte, die ein Merkmal annehmen kann.

**Beispiele**:

Verteilung der Einkommen, Umsätze, Telefonierdauer

Kaufwahrscheinlichkeit, Kündigungswahrscheinlichkeit

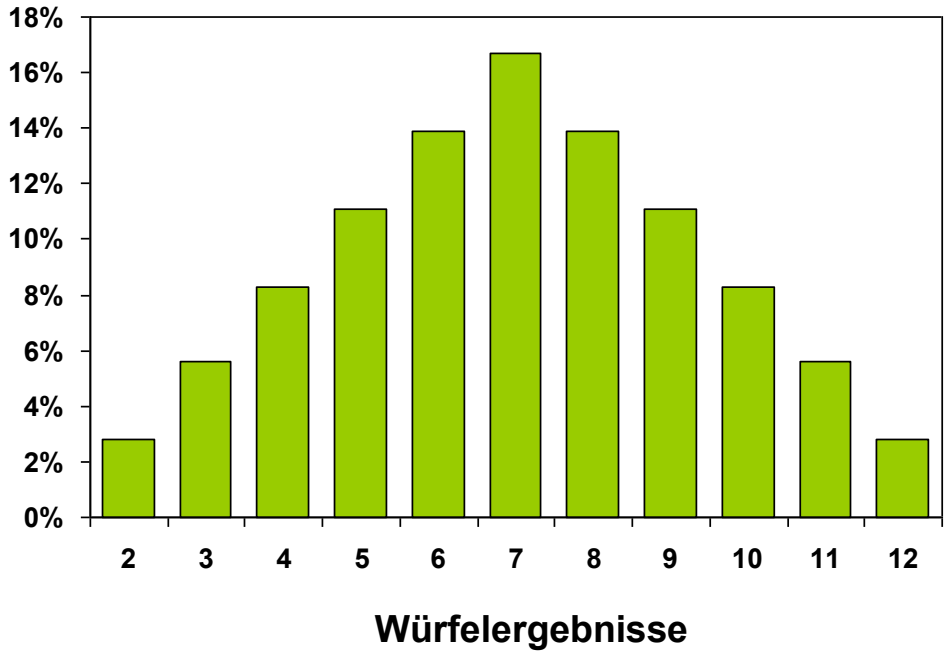
# Diskrete Verteilungen: Beispiele

## Würfeln mit zwei Würfeln

| Werte | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P | 2,8% | 5,6% | 8,3% | 11,1% | 13,9% | 16,7% | 13,9% | 11,1% | 8,3% | 5,6% | 2,8% |

Die Verteilung wird angegeben, indem man die möglichen Werte und die zugehörigen Wahrscheinlichkeiten (relativen Häufigkeiten) aufschreibt.

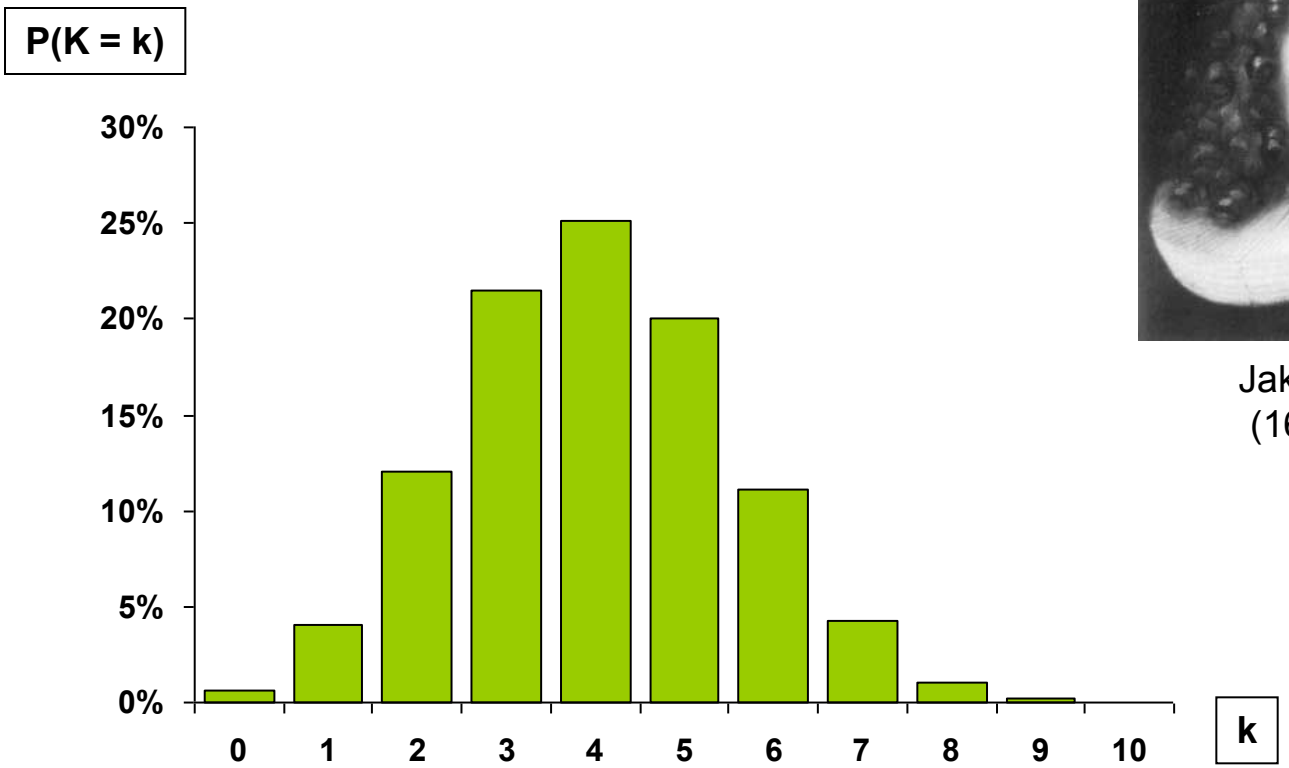
Darstellung erfolgt als Balken- oder Tortendiagramm (hier sind Balken besser).



# Diskrete Verteilungen: Beispiele

## Beispiel Binominalverteilung:

n = 10; p = 0,4




Jakob Bernoulli
(1655 – 1705)

# Diskrete Verteilungen: Beispiele

## Die Binomialverteilung

### Das Bernoullische Versuchsschema:

Bei n durchgeführten Versuchen tritt k mal das zufällige Ereignis A ein. Wie groß ist die Wahrscheinlichkeit dafür, wenn P(A) = p in der Grundgesamtheit?

1,1,1,1,1,....1,1,1,0,0,0,0,.....0,0,0,0,0
k mal die 1 (n-k) mal die 0

Die Versuche sind unabhängig voneinander. Also ist die Wahrscheinlichkeit für die angegebene Realisierung des Versuches

$$p^k (1-p)^{n-k}$$

Es gibt aber $\binom{n}{k}$ Möglichkeiten, die k Einsen in dieser Folge zu platzieren.

Daraus folgt

$$P_n(K=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

# Diskrete Verteilungen: Beispiele

**Formel: n über k**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$n! = n*(n-1)*(n-2)*....*2*1$$

5! = 5 * 4 * 3 * 2 * 1 = 120
10! = 3.628.800

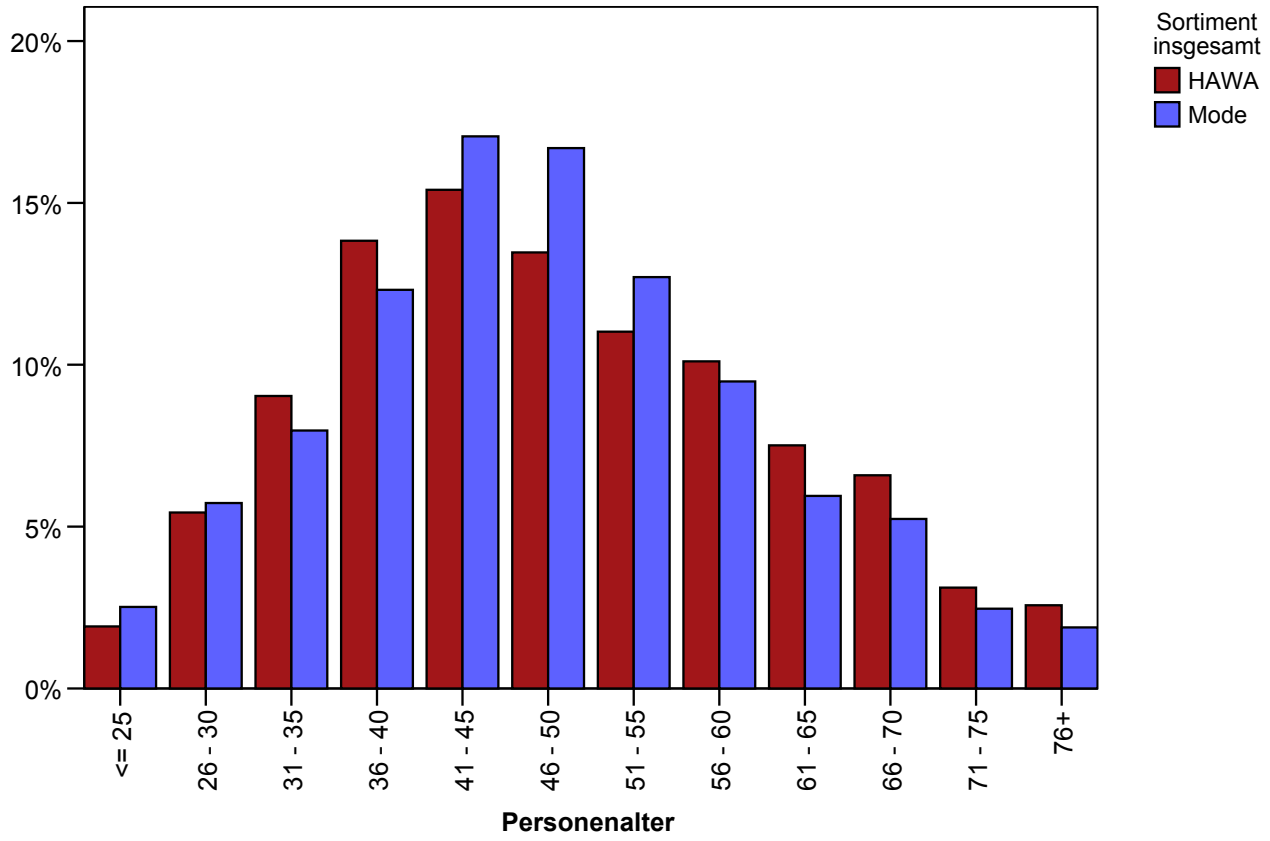
**Formel nützlich bei Glücksspielen: 6 aus 49 z.B.**

$$\binom{49}{6} = \frac{49*48*47*46*45*44}{6 * 5 * 4 * 3 * 2 * 1} = 13.983.816$$

So viele Möglichkeiten gibt es, 6 aus 49 Zahlen auszuwählen.

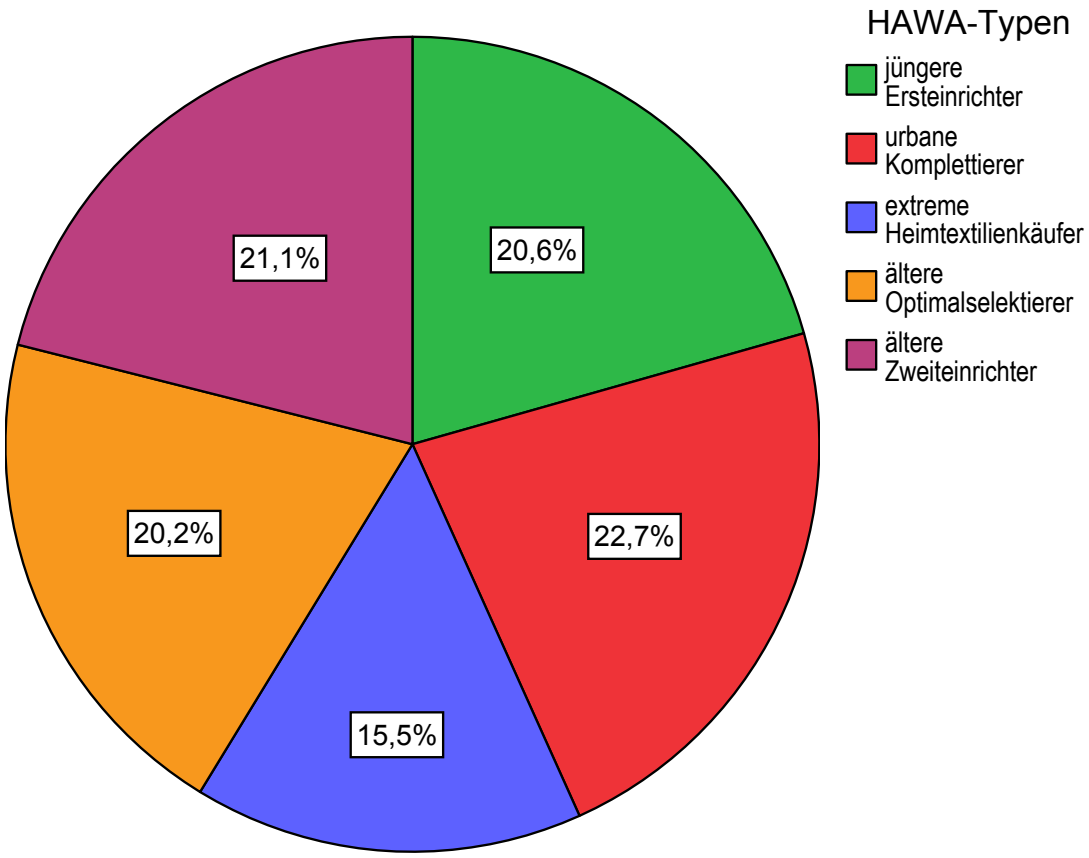
# Diskrete Verteilungen: Beispiele

## Beispiel Altersgruppenverteilung:



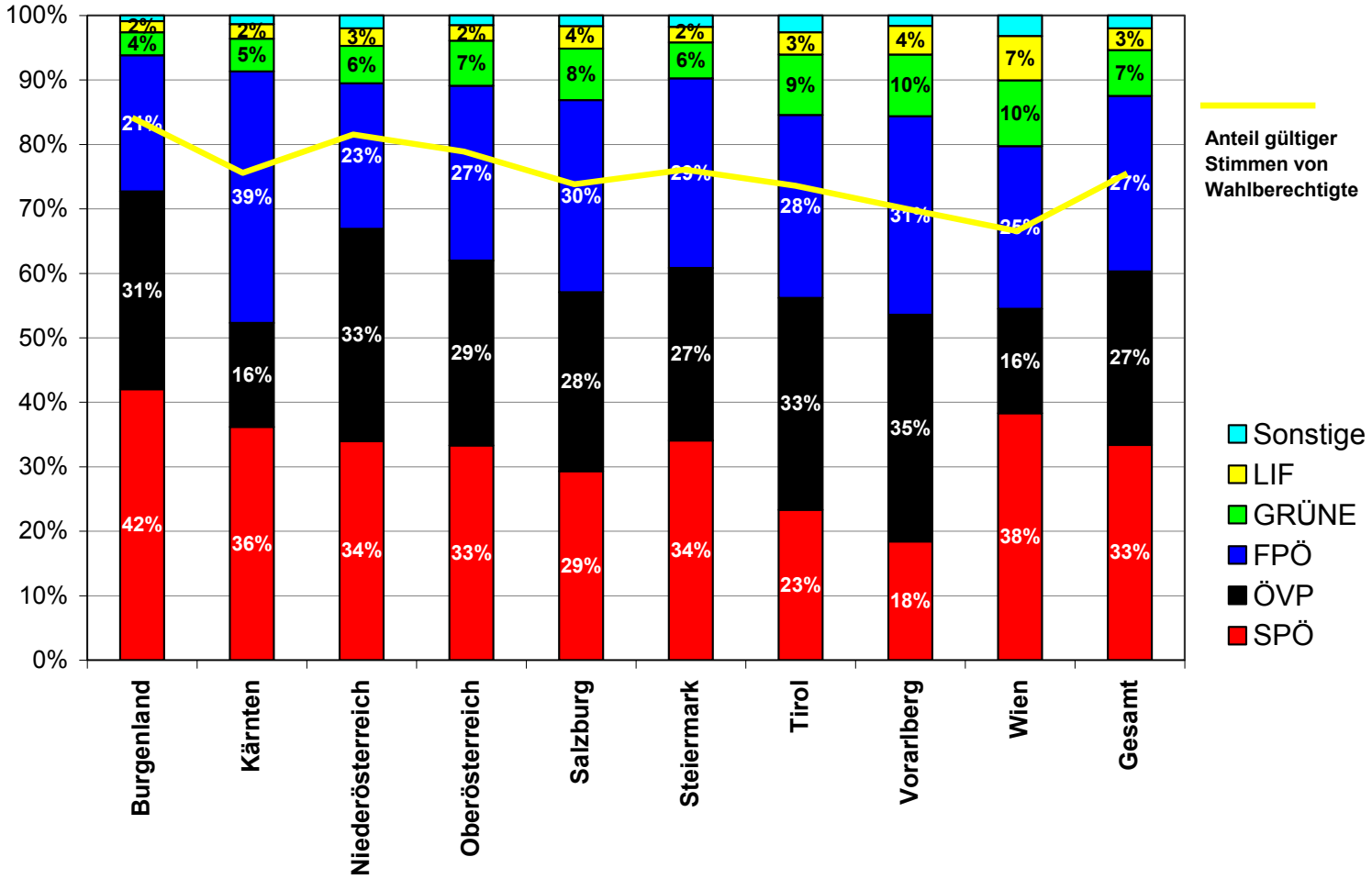
# Diskrete Verteilungen: Beispiele

## Beispiel Verteilung von Kundentypen:



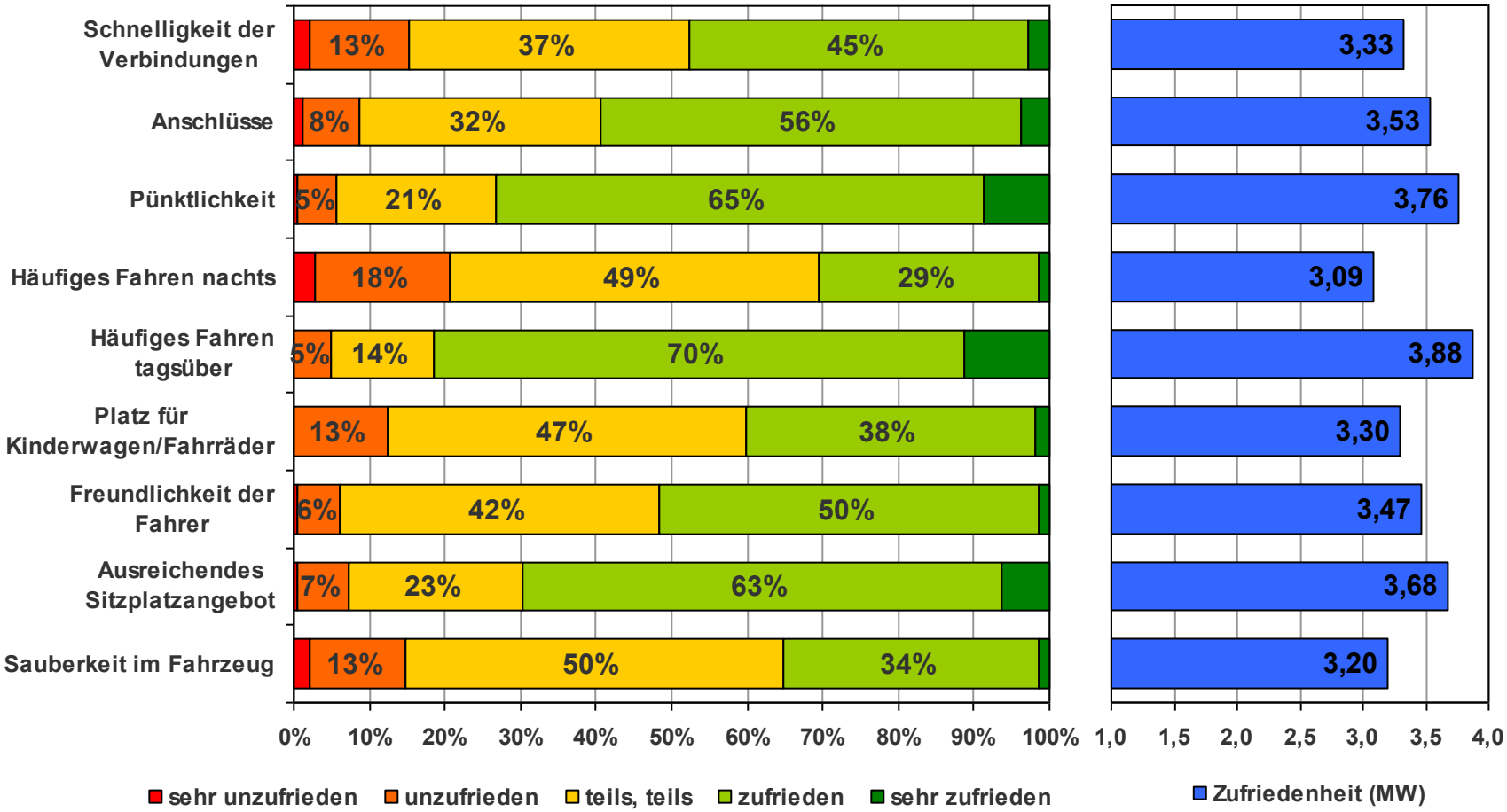
# Diskrete Verteilungen: Beispiele

## Beispiel Stimmenverteilung bei der NRW 1999:



# Diskrete Verteilungen: Beispiele

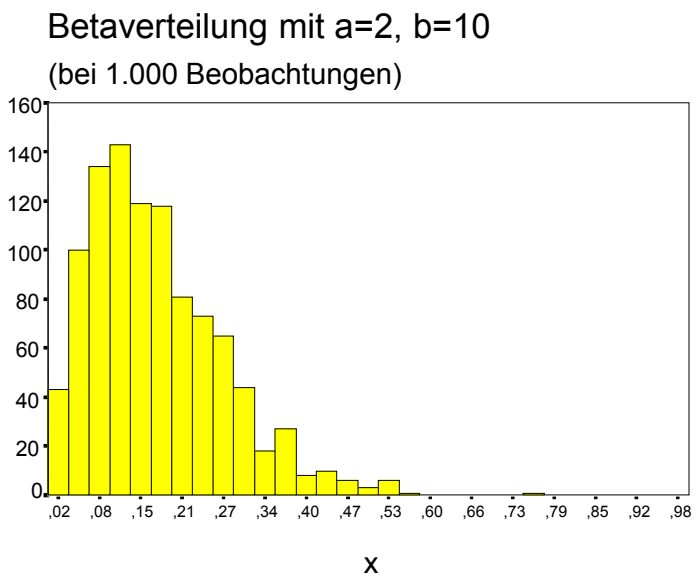
## Beispiel Ergebnis Kundenzufriedenheitsanalyse:

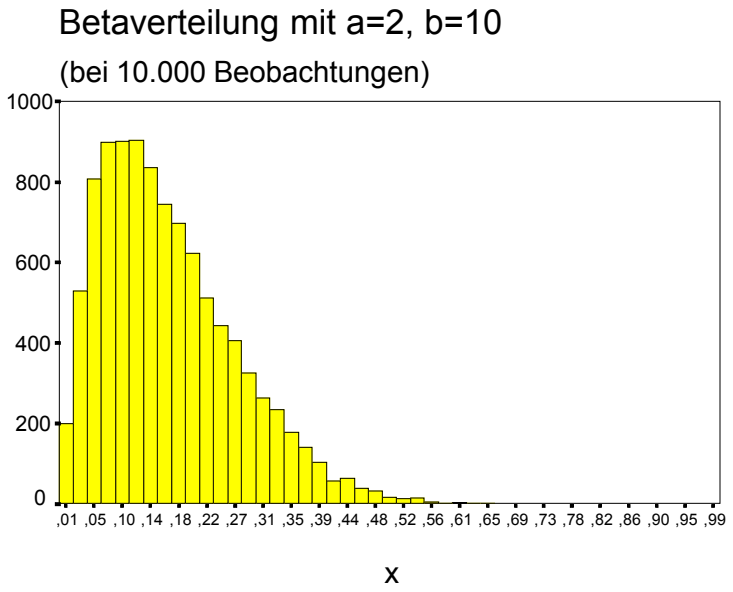


# Stetige Verteilungen: Histogramm

## Stetige Verteilungen

Bei stetigen Verteilungen kann man nicht alle Werte und die zugehörigen Wahrscheinlichkeiten abtragen. Es sind zu viele Werte und P für jeden einzelnen Wert ist sehr, sehr klein. Deshalb wählt man Klasseneinteilung und zählt, wie häufig die Klassen auftreten. Diese Darstellung nennt man ***Histogramm.***

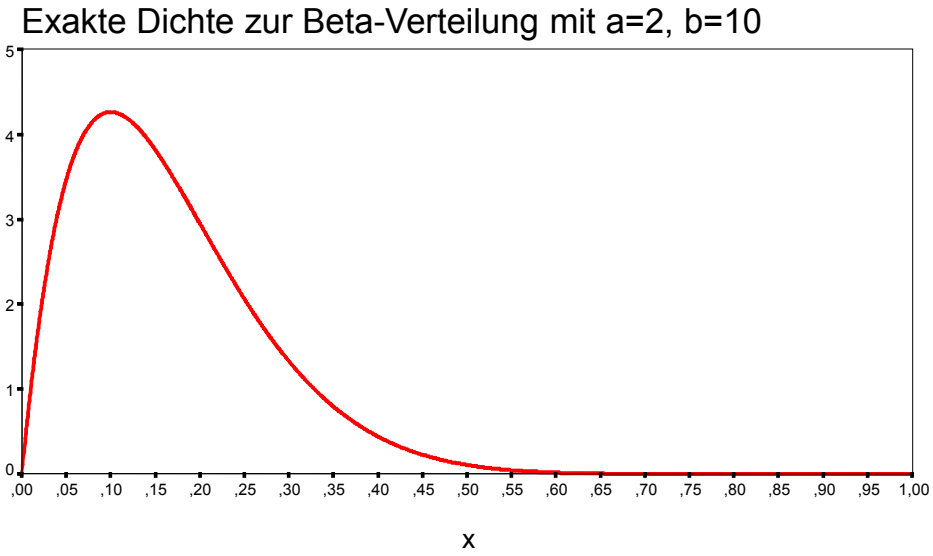


# Stetige Verteilungen: Dichtefunktion

## Stetige Verteilungen

Wird die Fallzahl immer größer und die Klassen immer feiner, geht das Histogramm gegen eine glatte Kurve, die ***Dichtefunktion f(x)***.




**Eigenschaften der Dichte**:

$f(x) \geq 0$

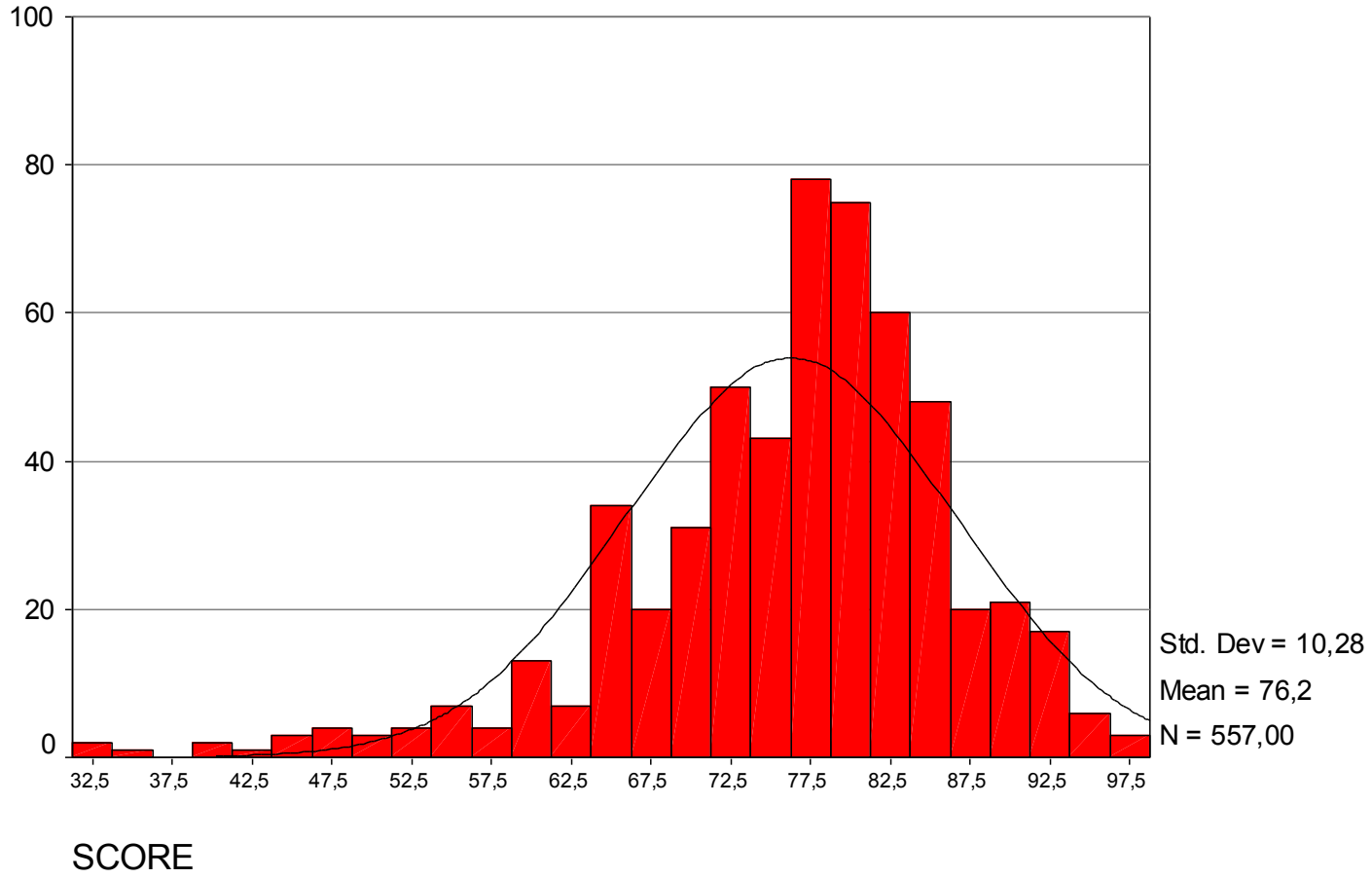
und Fläche unter der Kurve = 1

(Normierung)

Es ist eine glatte Kurve.

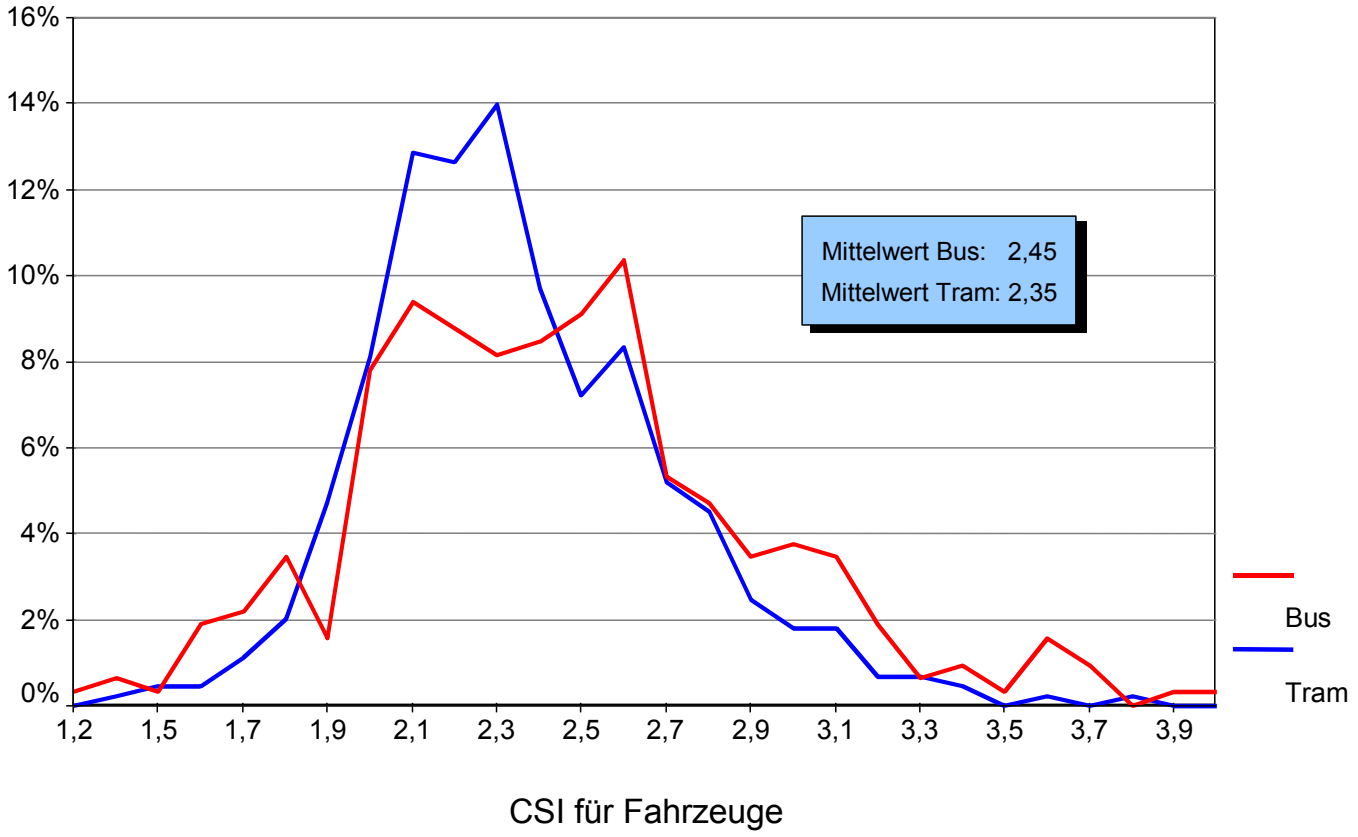
# Stetige Verteilungen: Beispiele

## Beispiel Histogramm: Kundenzufriedenheit



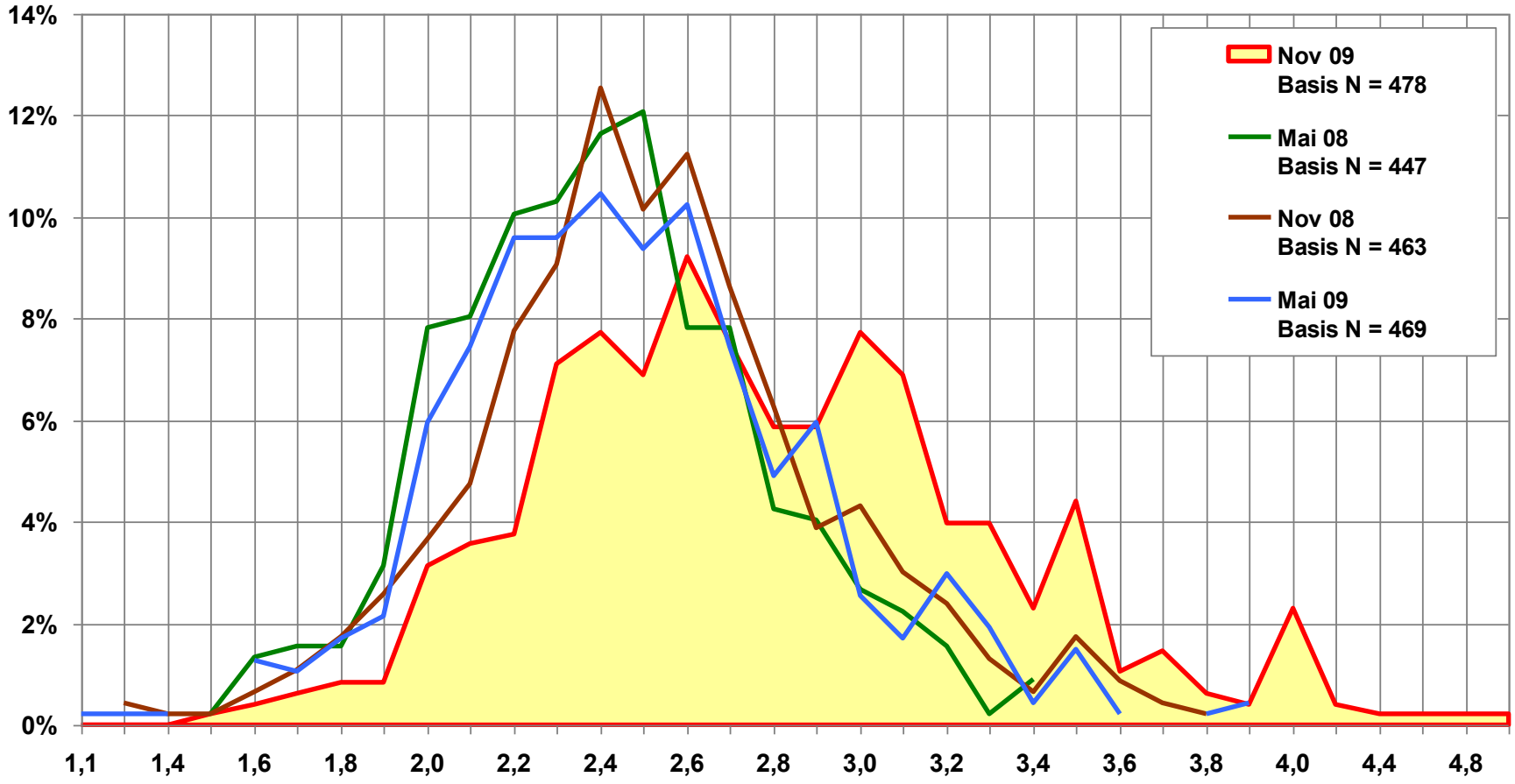
# Stetige Verteilungen: Beispiele

## Beispiel Histogramm als Häufigkeitspolygon: Kundenzufriedenheit



# Stetige Verteilungen: Beispiele

## Beispiel Histogramm als Häufigkeitspolygon: Kundenzufriedenheit

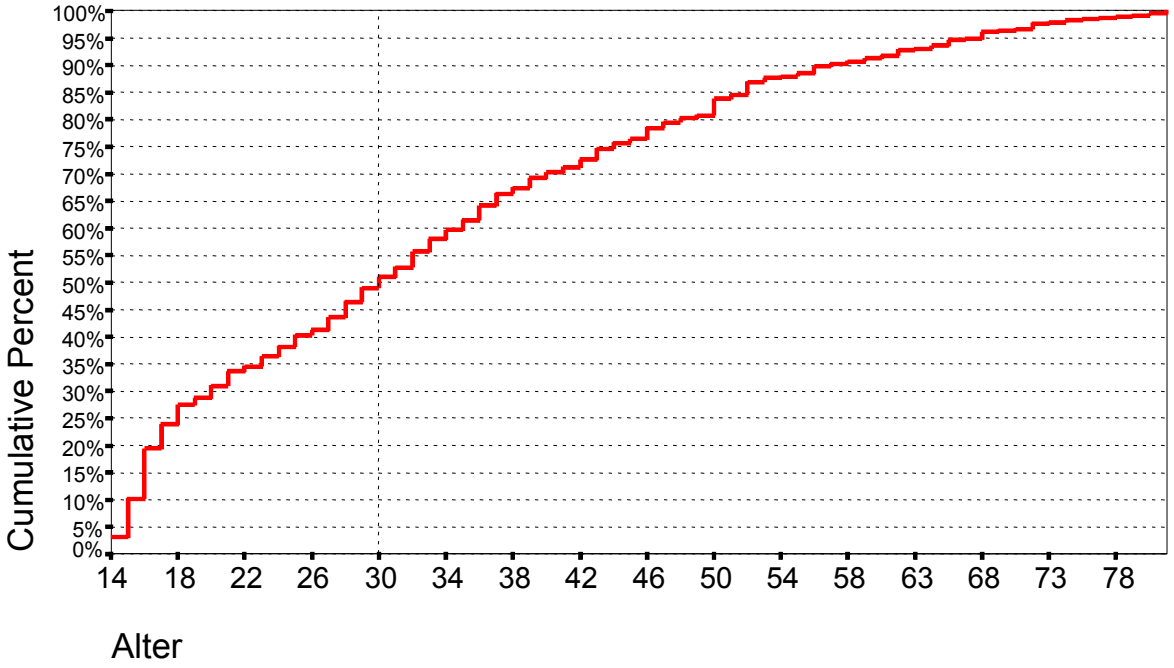


# Stetige Verteilungen: Verteilungsfunktion

## Stetige Verteilungen

Eine andere Form der Darstellung stetiger Verteilungen ist die kumulierte Summe.
Nennt man auch ***empirische Verteilungsfunktion***.

### Kumulative Summenfunktion

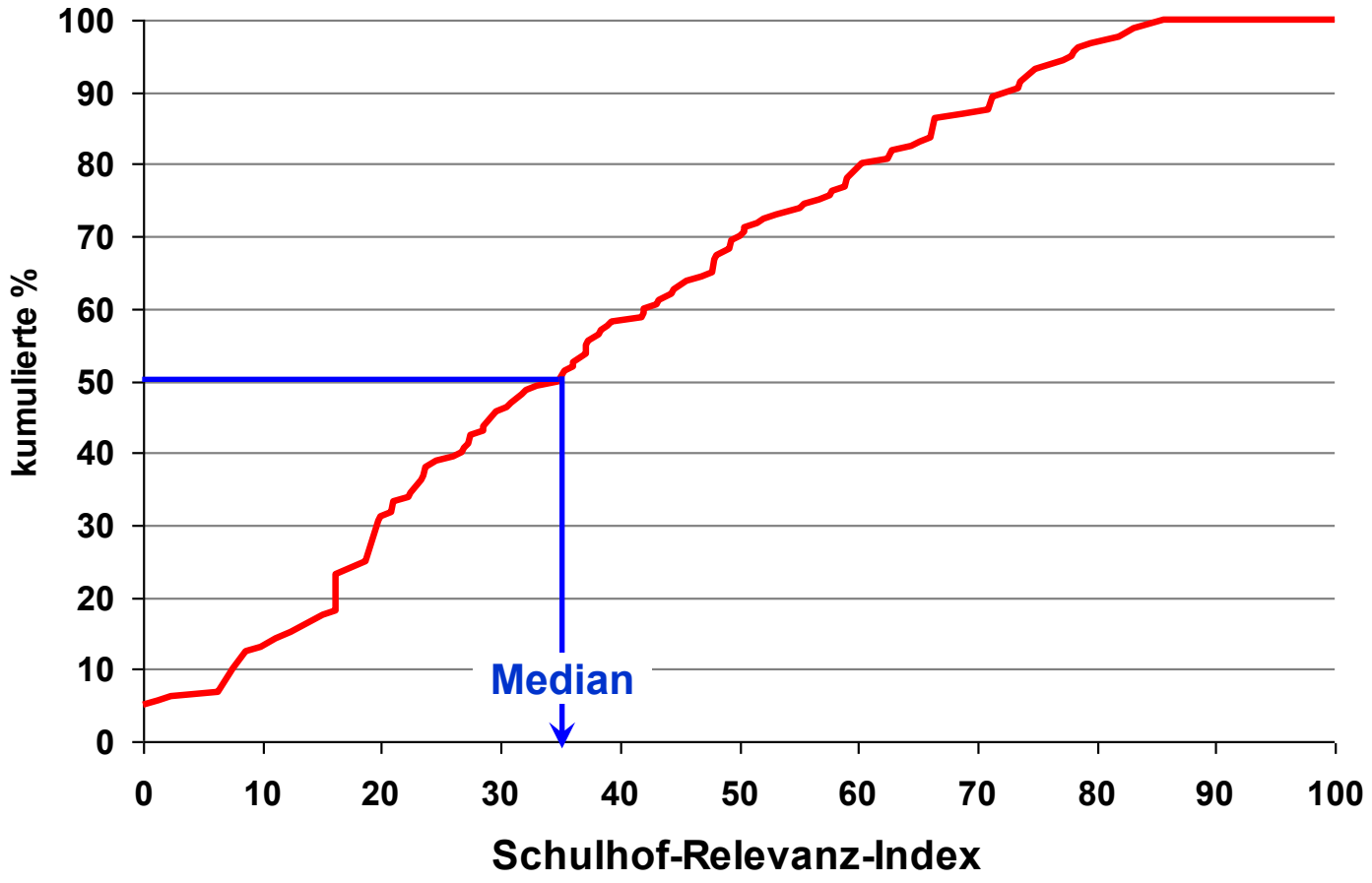



**Hauptsatz der mathematischen Statistik:**

Für $n \to \infty$ konvergiert die empirische Verteilungsfunktion gegen die wahre Verteilungsfunktion der Grundgesamtheit.

# Kumulierte Summe: Beispiele

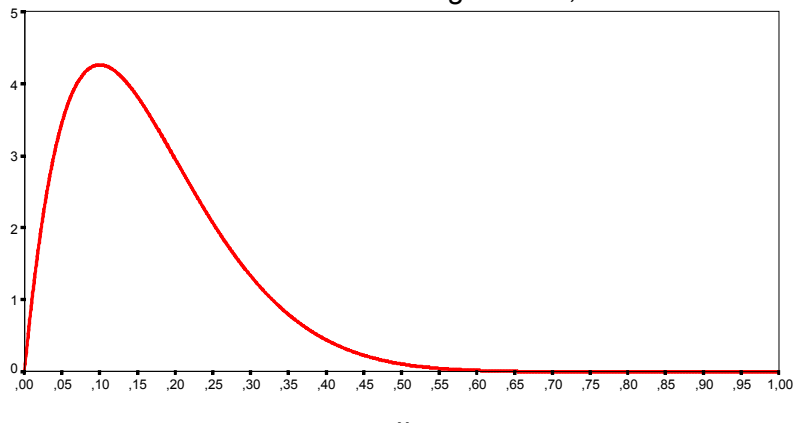
## Beispiel Kumulierte Summenfunktion: SRI



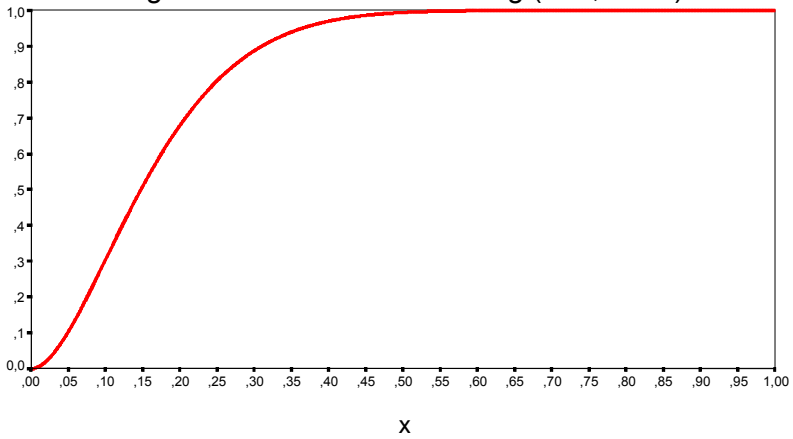
# Dichte und Verteilungsfunktion

## Stetige Verteilungen

Exakte Dichte zur Beta-Verteilung mit a=2, b=10




Verteilungsfunktion zur Betaverteilung (a=2, b=10)




Die Dichte f(x) ist der Anstieg der Verteilungsfunktion F(x) im Punkte x.

Die Verteilungsfunktion ist die Fläche unter der Dichte von -∞ bis x, d.h.

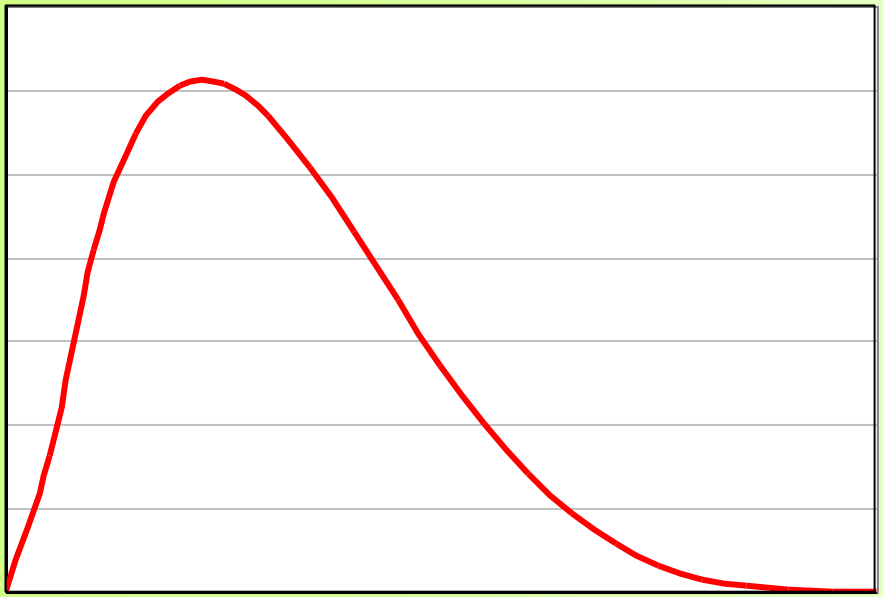
$$f(x) = \frac{dF(x)}{dx}$$

$$F(x) = \int_{-\infty}^{x} f(z)dz$$

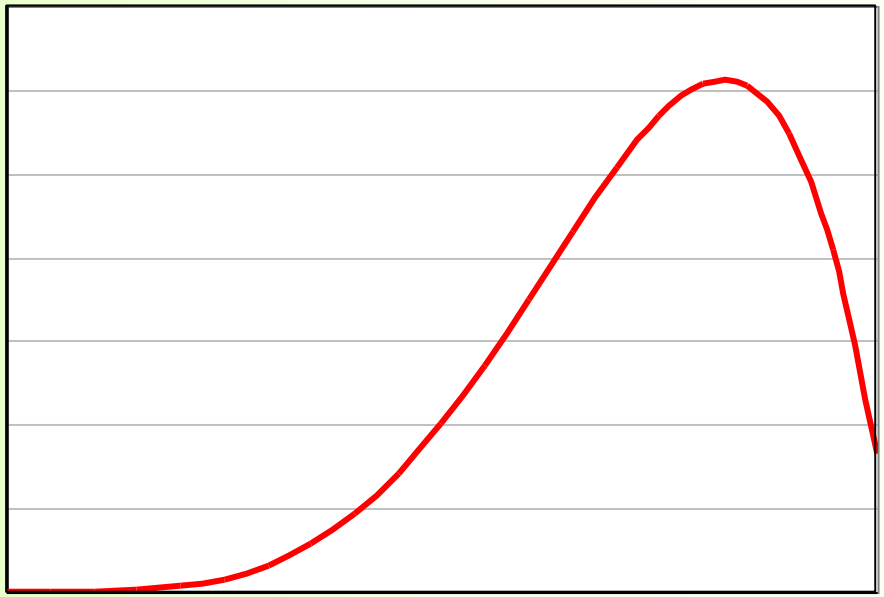
$$F(x) = P(X < x)$$

# Verteilungsformen

**Rechtsschiefe Verteilung: Schwanz nach rechts**



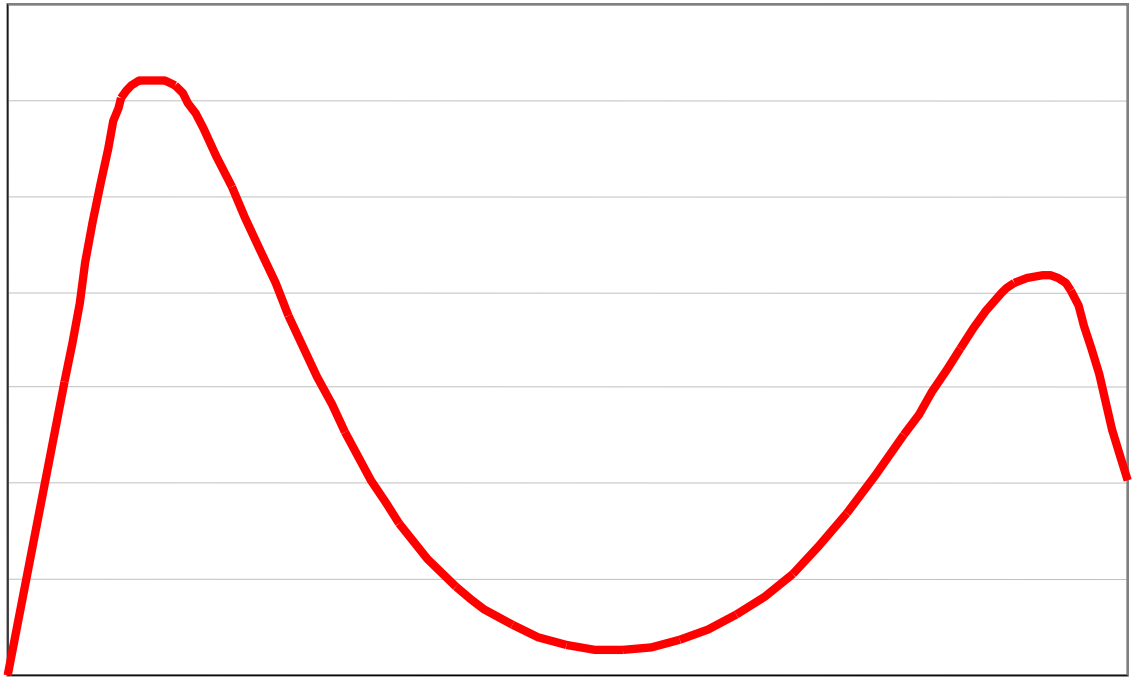
**Linksschiefe Verteilung: Schwanz nach links**



***Beide Verteilungen sind unimodal (eingipflig): Sie haben nur einen Gipfel.***

# Verteilungsformen

**Bimodale (zweigipflige) Verteilung**



Oft Mischung aus zwei unimodalen Verteilungen.

# Parameter von Verteilungen

Oftmals reicht es aus, einige **charakteristische Merkmale** der Verteilung zu kennen, um sich eine etwaige Vorstellung von einer Verteilung zu machen.

Solche Merkmale bezeichnet man als ***Parameter***.

Man unterscheidet:

- Lageparameter: z. B. arithmetisches Mittel, Median
- Streuparameter: z. B. Spannweite, Varianz
- Gestaltsparameter: z. B. Schiefe, Exzess (Kurtosis)

# Lageparameter

**arithmetisches Mittel**

Bekanntester Lage- oder Lokationsparameter:

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

[ $x_1,...x_n$ Beobachtungen, n Stichprobenumfang]

Es hat eine Reihe von angenehmen Eigenschaften, auf die wir noch eingehen werden, allerdings auch unangenehme: Es ist **nicht unabhängig von Ausreißern** am Rande der Verteilung.

# Lageparameter

**5% gestutztes Mittel**

Der arithmetische Mittelwert, in dessen Berechnung die 5% größten und die 5% kleinsten Werte nicht eingehen. Er wird empfohlen, wenn Werte auftreten, die viel größer oder viel kleiner als die restlichen Werte sind. Das 5% gestutzte Mittel ist **robust gegen** derartige **Ausreißer** und führt in solchen Fällen zu besseren Schätzungen der zentralen Tendenz als das gewöhnliche arithmetische Mittel.

# Lageparameter

**Median $x_{0.5}$**

Ein Maß für die zentrale Tendenz. Der Wert über und unter dem jeweils 50% der Beobachtungen liegen. Für fünf Beobachtungen wäre der Median der drittgrößte (oder drittkleinste) Wert der Beobachtungen. Für eine gerade Anzahl an Beobachtungen ergibt sich der Median als Durchschnitt der beiden "mittleren" Beobachtungen. Der Median ist **robust gegen Ausreißer** an den Enden der Verteilung.

**Mode, Modalwert, Modus**

Der am häufigsten auftretende Wert (oder Werte). Eine unimodale Verteilung besitzt einen Modalwert (bei empirischen Verteilungen wenig sinnvoll).

Nützlich vor allem als Verteilungsparameter **für nominale Variablen**.

# Lageparameter

**Perzentile (Prozentpunkte)**

Ähnlich dem Median (der 50%-Punkt) kann man für jeden anderen Prozentwert den zugehörigen Merkmalswert bestimmen.

Das 10. Perzentil $\mathbf{x_{0.1}}$ z.B. ist ein Wert der Verteilung, so dass 10% aller Werte des Merkmals kleiner als $\mathbf{x_{0.1}}$ sind.

Wichtig sind: das 25. Perzentil (unteres Quartil)
das 50. Perzentil (Median)
das 75. Perzentil (oberes Quartil)

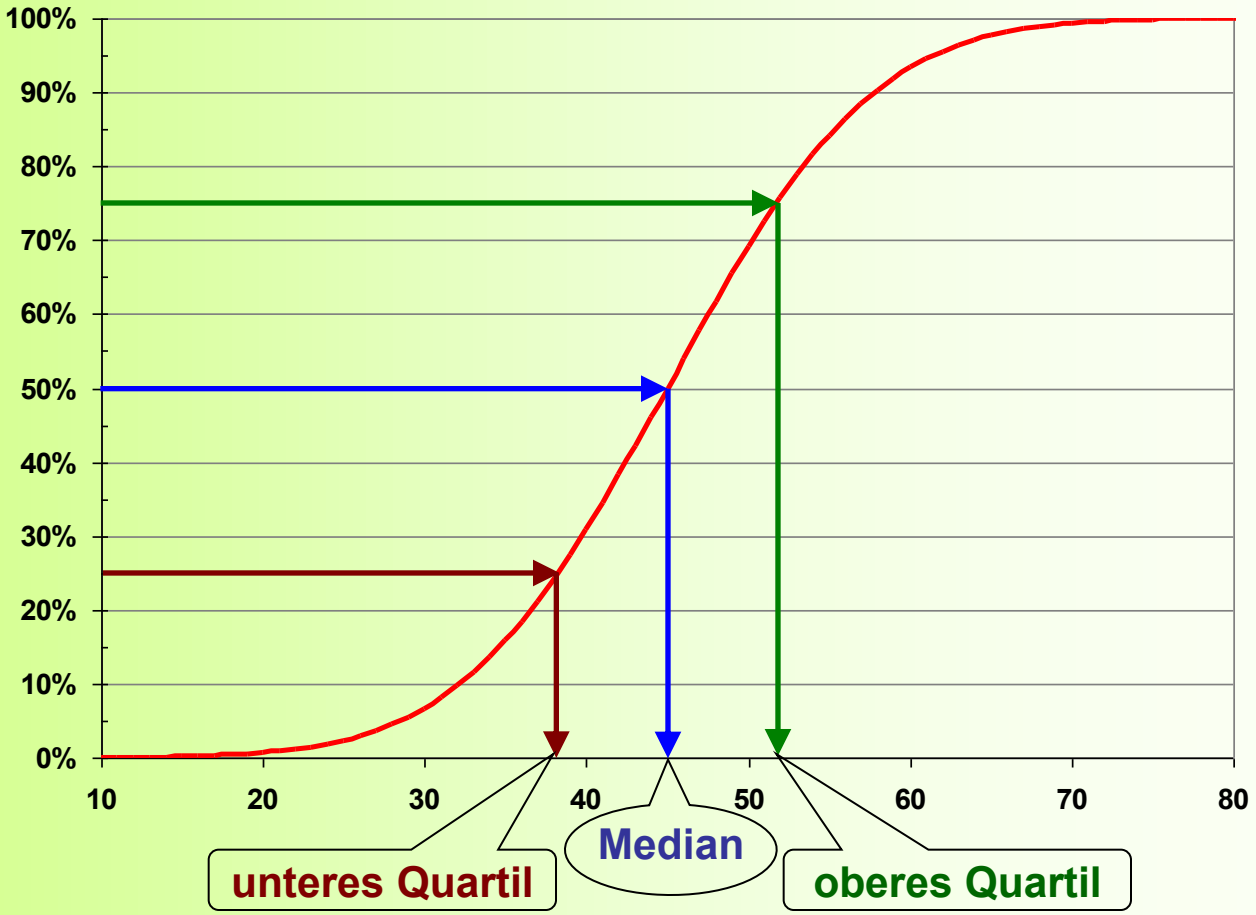
(später für Boxplots verwendet)

Zwischen dem unteren und dem oberen Quartil liegt die Hälfte der mittleren Werte des betrachten Merkmals.

# Lageparameter

## Perzentile anschaulich

**Perzentile ergeben sich als Spiegelung an der kumulierten Summenfunktion**



# Streuparameter (Skalenparameter)

**Bereich, Wertebereich, Spannweite**
(Range)

Der Abstand zwischen dem größten und dem kleinsten Wert einer Variablen. Er ist extrem Ausreißer-abhängig.

**Interquartils-Abstand**
(interquartile range)

Ein Ausreißer-robustes Maß für die Streuung der Daten. Der Abstand zwischen dem Wert des oberen Quartils (75ter Prozentpunkt) und dem Wert für das untere Quartil (25ter Prozentpunkt).

# Streuparameter (Skalenparameter)

**Varianz $s^2$**

Ein Maß für die Streuung der Werte um ihren Mittelwert. Sie wird berechnet als Summe der quadrierten Abweichungen vom Mittelwert, dividiert durch die um 1 verminderte Anzahl an gültigen Beobachtungen.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad , \quad \bar{x} \text{ arithmetisches Mittel}$$

Wie das arithmetische Mittel, ist auch die Varianz **extrem Ausreißer-abhängig**.

**Standardabweichung s**

(Standard Deviation)

Die Quadratwurzel der Varianz. Die Standardabweichung ist ein Streuungsmaß, das in der **gleichen Maßeinheit wie die Beobachtungen** vorliegt.

# Die Z-Transformation

Die **Z-Transformation** wird zur **Standardisierung von zufälligen Merkmalen** *x* verwendet:

$$z = \frac{x - \bar{x}}{s}$$ , $\bar{x}$ = arithmetisches Mittel, $s$ = Standardabweichung

Die Z-Werte sind **skalenunabhängig und dimensionslos**. Z-transformierte Werte haben immer den Mittelwert 0 und die Standardabweichung 1.

Damit werden Merkmale mit unterschiedlichen Skalen besser vergleichbar. Die Null ist dann immer der Merkmalsdurchschnitt, und die Eins bedeutet dann immer eine Standardabweichung.

Die Z-Transformation ist eine **lineare Transformation**.

**Beispiel**: Die Faktorscores nach einer Faktorenanalyse werden als Z-transformierte Werte berechnet.

# Gestaltsparameter

Sie sagen etwas über die Form der Verteilung aus. Die wichtigsten Gestaltsparameter sind Schiefe und Exzeß, auch Kurtosis genannt:

**Schiefe**

(Skewness)

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

Ein Index für die Schiefe, also den Grad, zu dem eine unimodale Verteilung nicht symmetrisch ist, so dass ein Ende der Verteilung stärker nach links oder rechts ausgeprägt ist.

Für eine **Normalverteilung**, die symmetrisch ist, ist die **Schiefestatistik=0**. Eine Verteilung mit einer signifikant positiven Schiefe hat ein langes rechtes Ende (rechtsschief). Eine Verteilung mit einer signifikant negativen Schiefe besitzt ein langes linkes Ende (linksschief).

Die Schiefe wird zusammen mit der Kurtosis benutzt, um eine Variable auf Normalverteilung zu prüfen.

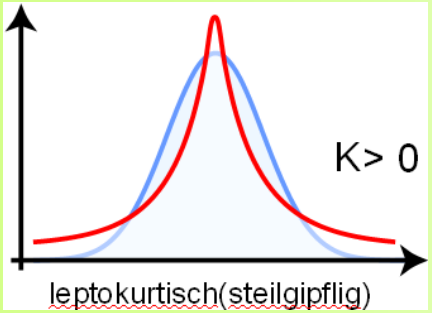
# Gestaltsparameter

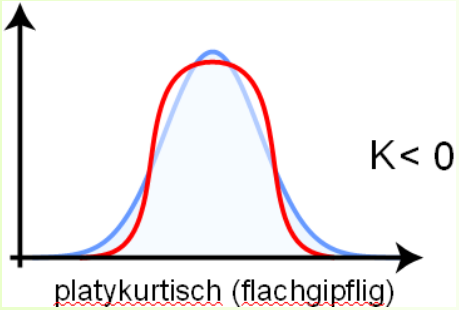
**Kurtosis, Exzeß, Wölbung**

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

Ein Maß für den Anteil der Beobachtungen, die in den Enden (Schwänzen) einer Verteilung liegen, bzw. für das absolute Maximum der Häufigkeitsverteilung im Vergleich zur Dichtefunktion einer Normalverteilung.

Für eine theoretische **Normalverteilung** hat **die Kurtosis einen Wert von 0** (mesokurtisch). Eine Variable mit negativer Kurtosis (platykurtisch) hat schwächere Enden als eine Normalverteilung, eine positive Kurtosis (leptokurtisch) zeigt an, daß mehr Beobachtungen in die Enden der Verteilung fallen als bei einer Normalverteilung. Zusammen mit der Schiefe läßt sich abschätzen, ob eine Variable normalverteilt ist.



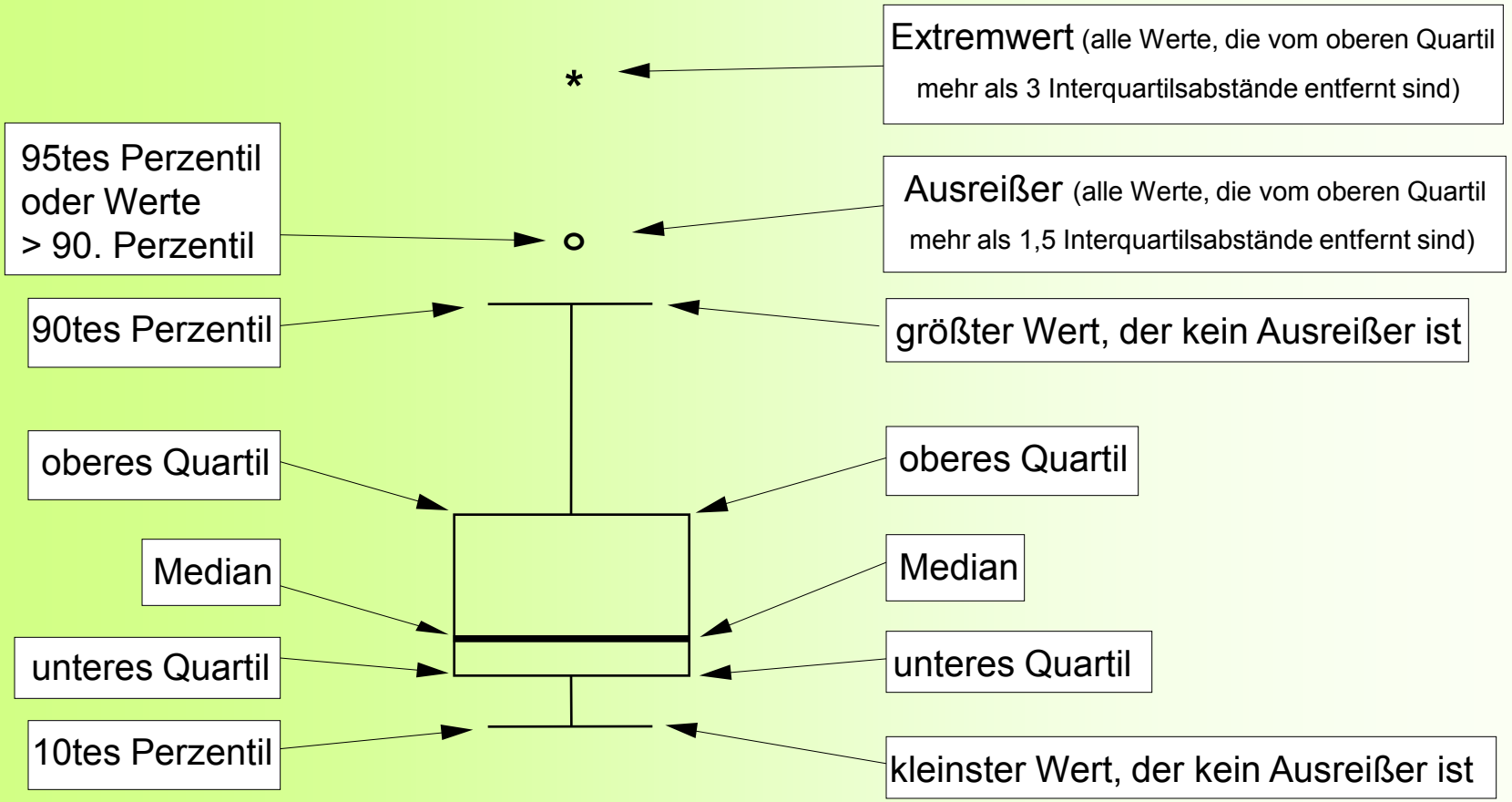



# Boxplots zur Darstellung von Verteilungen

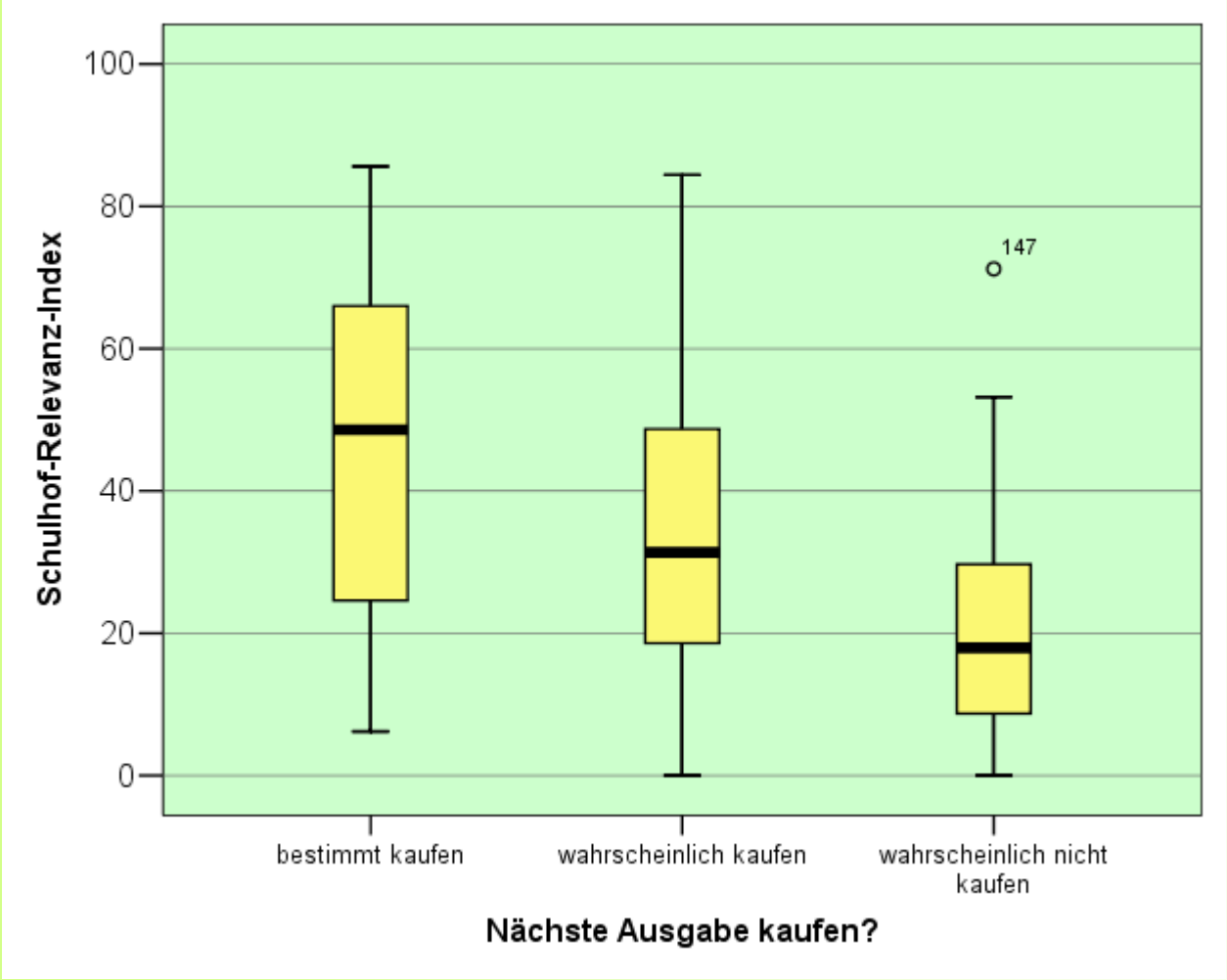
## Boxplots nicht einheitlich definiert

### SigmaStat / SigmaPlot

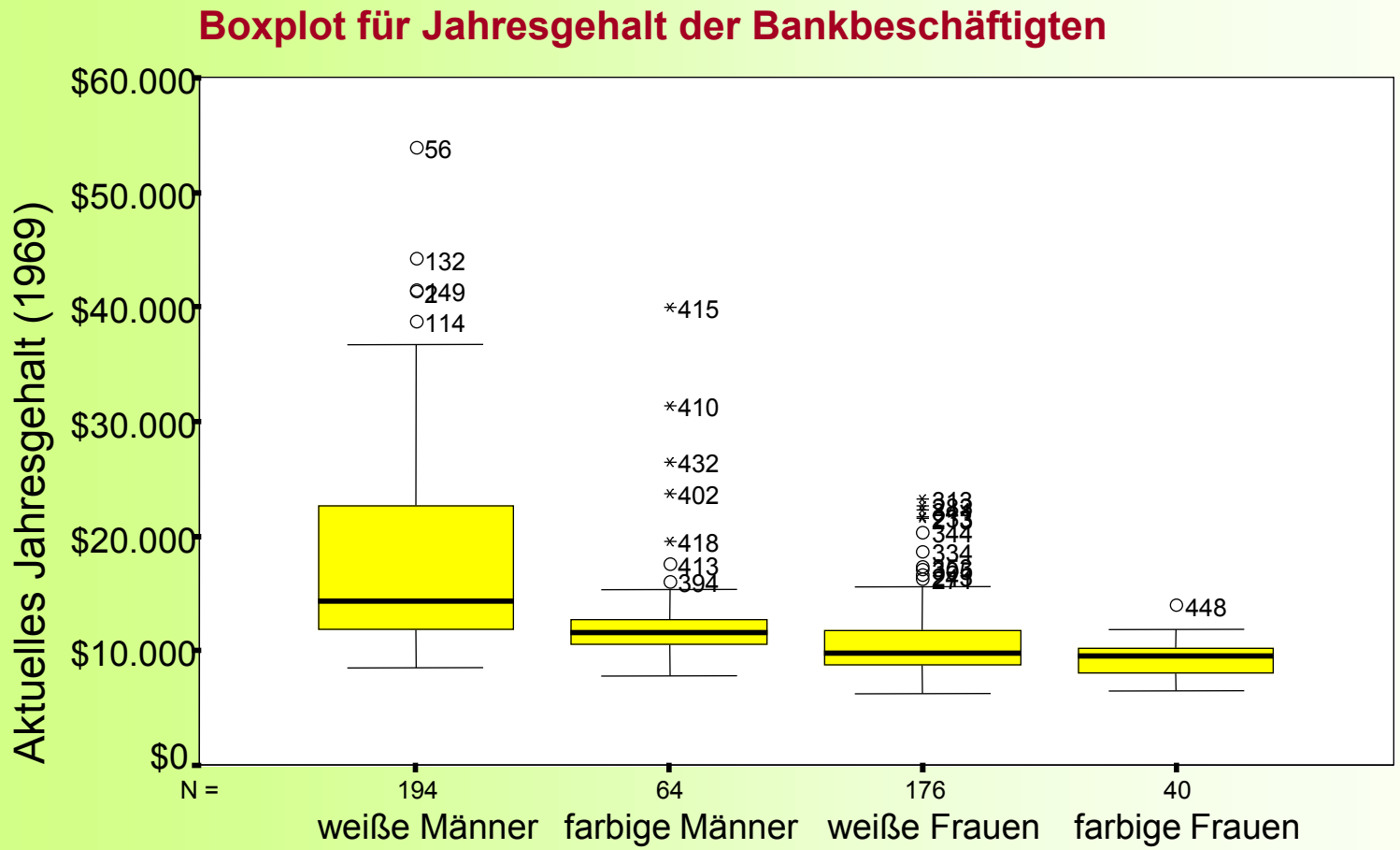
### SPSS



# Boxplots: Beispiele (SPSS)



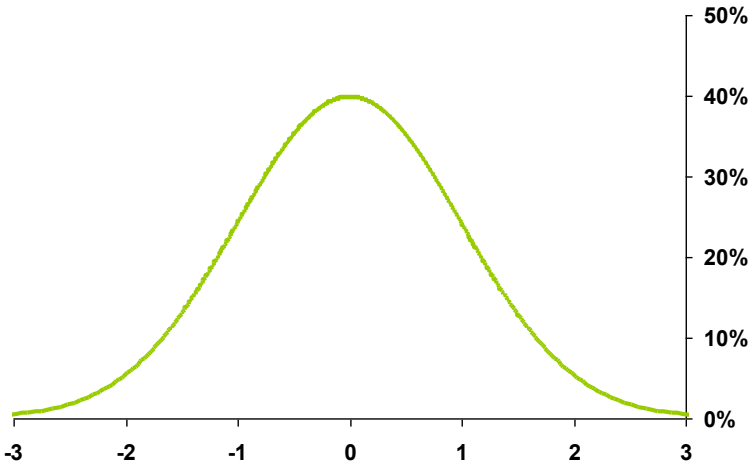
# Boxplots: Beispiele (SPSS)



# Die Normalverteilung

Eine stetige Zufallsgröße X heißt normalverteilt mit dem Mittelwert μ und der Varianz $\sigma^2$, wenn sie folgende Dichte- bzw. Verteilungsfunktion hat:

**Dichtefunktion f(x)**




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

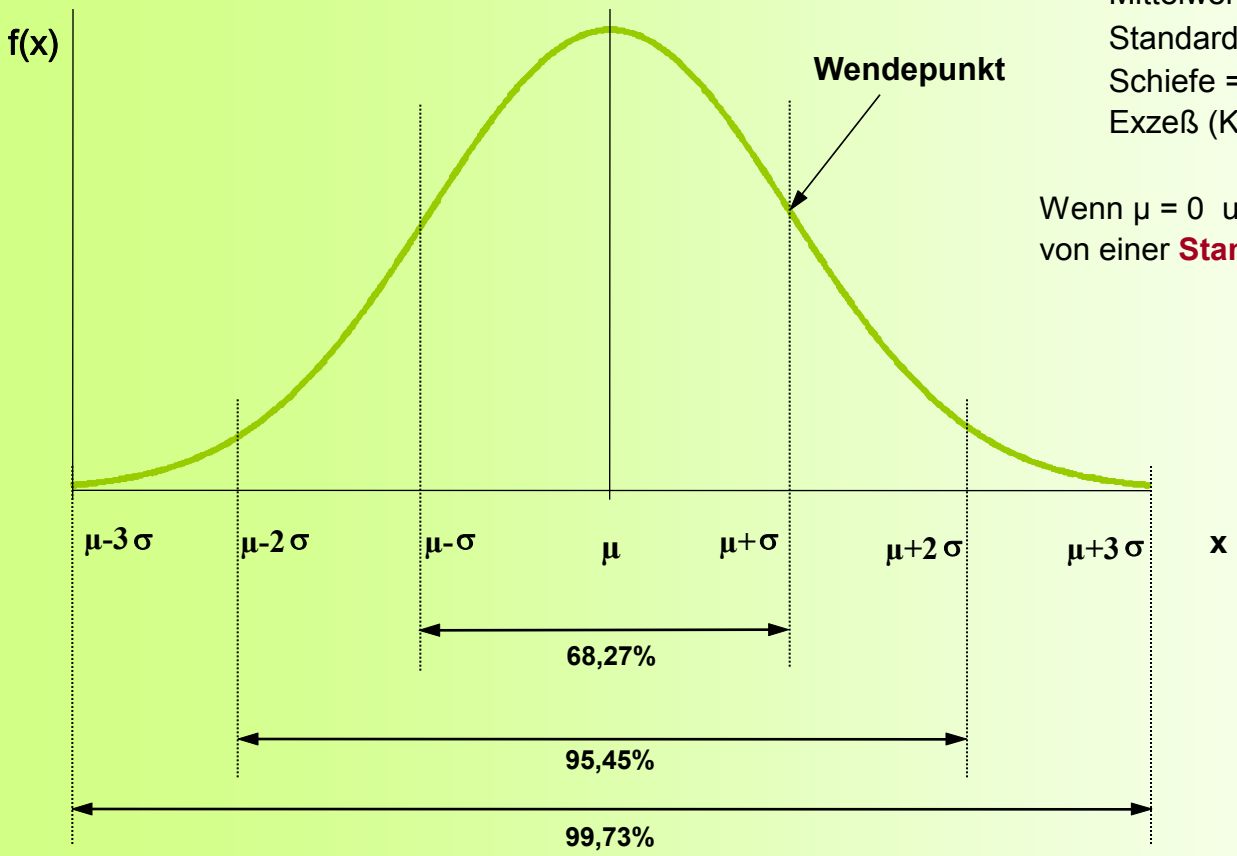
**Verteilungsfunktion F(x)**


Die Dichtefunktion der Normalverteilung wurde von Gauß entdeckt und wird auch ***Gaußsche Glockenkurve*** genannt.

# Die Normalverteilung

## Die Dichte der Normalverteilung

***Symmetrisch um den Mittelwert***




Hier sind

Mittelwert = Median = Modalwert =μ
Standardabweichung = σ
Schiefe = 0
Exzeß (Kurtosis) = 0

Wenn μ = 0 und σ = 1, so spricht man von einer **Standardnormalverteilung**.

Carl Friedrich Gauß
(1777 – 1855)

# Standardfehler vs. Standardabweichung

Geschätzte Parameterwerte treffen bei einer Stichprobenerhebung naturgemäß nicht die wahren Werte der Grundgesamtheit (GG).

Z.B. ist das arithmetische Mittel als Schätzung des wahren Mittels der GG selbst eine Zufallsgröße mit einer Wahrscheinlichkeitsverteilung.

Bortz spricht von *Stichprobenkennwerteverteilungen*.

**Die Standardabweichung eines geschätzten Parameters nennt man Standardfehler (Standard Error, S.E.)**

Er sagt etwas aus über die Genauigkeit einer Schätzung. Damit kann man einschätzen, wie weit der wahre Wert der GG vom geschätzten Wert noch entfernt sein kann.

S.E. hängt ab von der Fallzahl n und der Standardabweichung s.

**Wird verwendet zum Bau von Konfidenzintervallen.**

# Standardfehler

## Standardfehler für den Mittelwert

Für das arithmetische Mittel kann man den Standardfehler leicht berechnen. Er ergibt sich zu:

$$\text{Standardfehler}(\bar{x}) = \frac{s}{\sqrt{n}}$$

$s$ ist hier die Standardabweichung des Merkmals (berechnet aus der Stichprobe) und $n$ ist der Stichprobenumfang.

Für wachsende $n$ verringert sich der Standardfehler.

Die Standardabweichung der Stichprobe kann man nicht beeinflussen, bestenfalls durch die Elimination von schlimmen Ausreißern.

**Die Schätzgenauigkeit kann man also über die Fallzahl steuern.**

# Konfidenzintervalle

**Ausgangspunkt** ist die Frage:

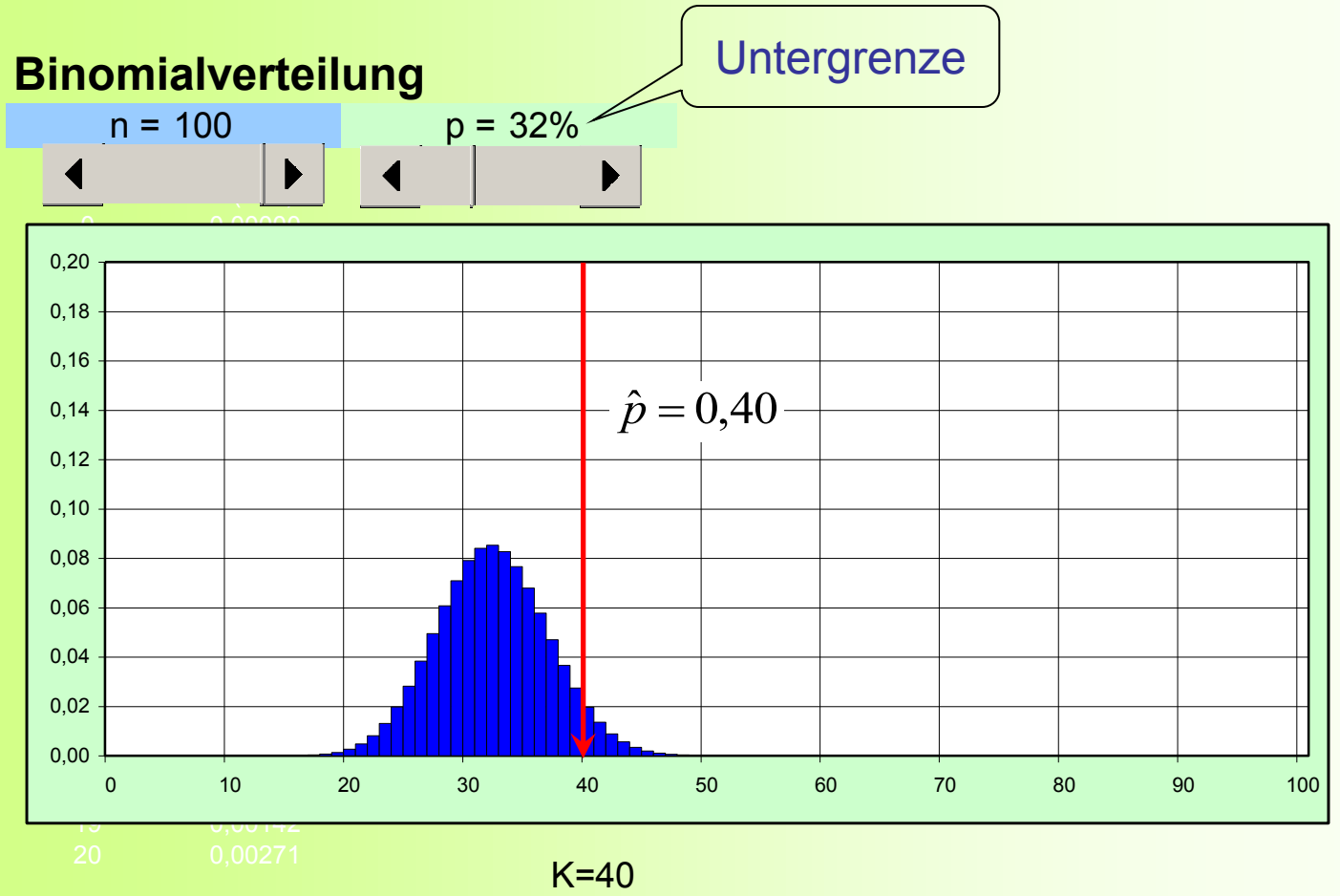
**Wie weit kann der unbekannte Parameter der Verteilung der Grundgesamtheit (GG) vom geschätzten Parameterwert entfernt sein?**

Dazu betrachtet man alle möglichen Werte, die der interessierende Parameter für die GG annehmen kann und wählt diejenigen aus, für die der berechnete Schätzwert plausibel ist.

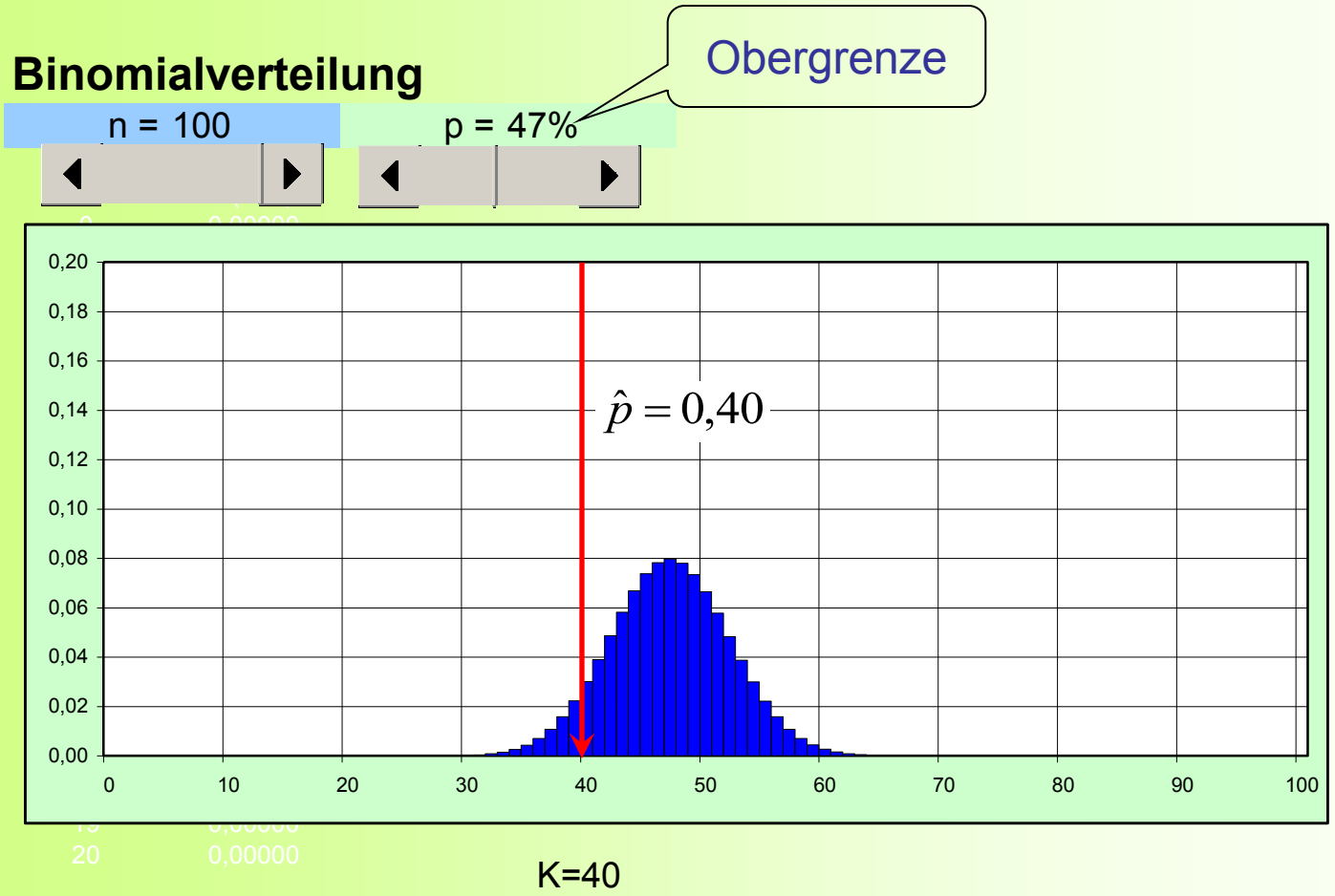
Die Verteilung aus der GG des betrachteten Merkmals generiert eine **Stichprobenkennwerteverteilung** für die Parameterschätzung. Plausibel sind solche Parameter für die Verteilung in der GG, die den berechneten Schätzwert nicht ganz unwahrscheinlich erscheinen lassen.

‚Nicht ganz unwahrscheinlich' wird über das Konfidenzniveau (95%) oder die Irrtumswahrscheinlichkeit (5%) gesteuert.

# Konfidenzintervalle: Beispiel



# Konfidenzintervalle: Beispiel



# Konfidenzintervalle: Berechnung

## Konfidenzintervall für Mittelwerte (n ≥ 50)

Man kann zeigen, dass das arithmetische Mittel als Schätzung um den wahren unbekannten Mittelwert streut, mit einer Standardabweichung von

$$S.E.(Mean) = \frac{s}{\sqrt{n}}$$ ← Standardfehler des MW

Für wachsende Fallzahl n wird der Standardfehler kleiner, also die Schätzung genauer.

### Zentraler Grenzwertsatz der Statistik

Das arithmetische Mittel ist mit wachsendem Stichprobenumfang n näherungsweise normalverteilt, unabhängig von der Originalverteilung der Daten.

Für n ≥ 50 ist diese Näherung (Approximation) meist schon hinreichend gut.

# Konfidenzintervalle: Berechnung

Für Normalverteilung oder n > 50 gilt damit (Eigenschaft der NV):

$$P(\mu - 1{,}96 * \frac{s}{\sqrt{n}} \leq \bar{x} \leq \mu + 1{,}96 * \frac{s}{\sqrt{n}}) = 95\%$$

$$P(\bar{x} - 1{,}96 * \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + 1{,}96 * \frac{s}{\sqrt{n}}) = 95\%$$

Dieses sagt aus, dass *das* ***Konfidenzintervall mit seinen zufälligen Grenzen mit einer Wahrscheinlichkeit von 95% den wahren unbekannten MW überdeckt*.**

**Allgemein gilt für n > 50**:

Sei ĉ eine Schätzung eines Parameters c, und S.E.(ĉ) sei der Standardfehler von ĉ.

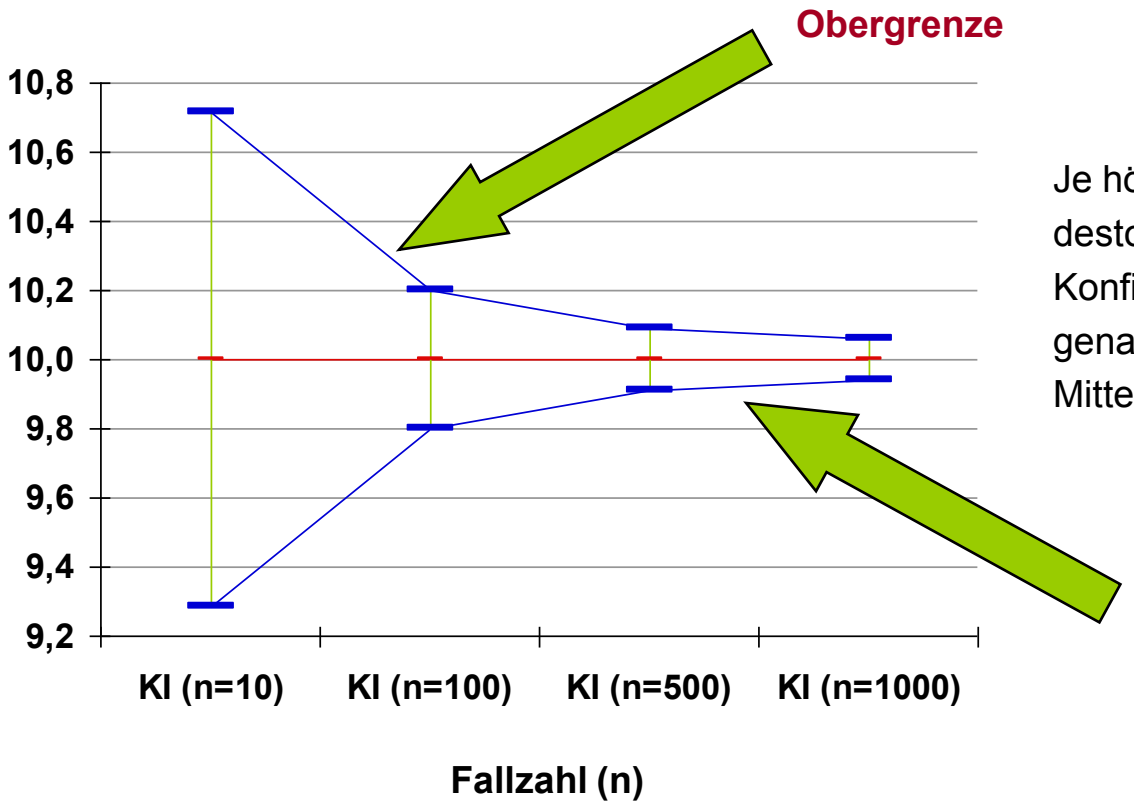
Dann ergibt sich durch **[ĉ ± 2*S.E.(ĉ)]** näherungsweise ein 95% KI für c.

Die Standardfehler werden bei der Berechnung von Schätzungen oft mit angegeben.

# Konfidenzintervalle: Einflüsse auf die Breite

## Konfidenzintervall und Fallzahl

MW = 10; STDEV = 1; Sicherheit = 95%

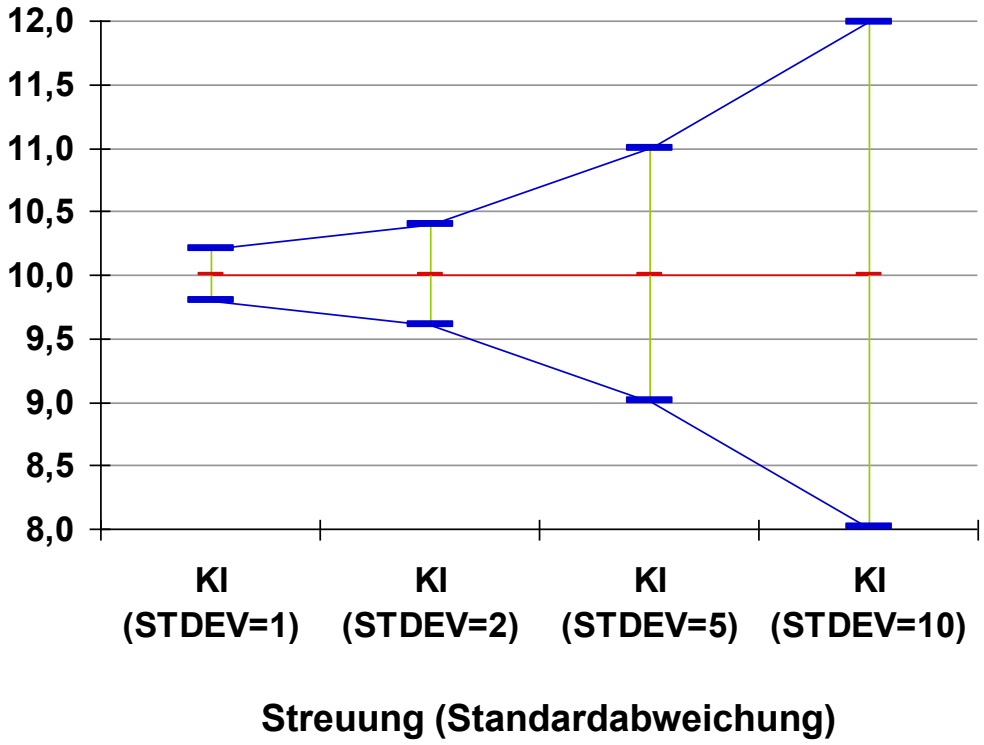



Je höher die Fallzahl (n), desto schmaler das Konfidenzintervall, desto genauer ist der geschätzte Mittelwert.

# Konfidenzintervalle: Einflüsse auf die Breite

## Konfidenzintervall und Streuung

MW = 10; n = 100; Sicherheit = 95%

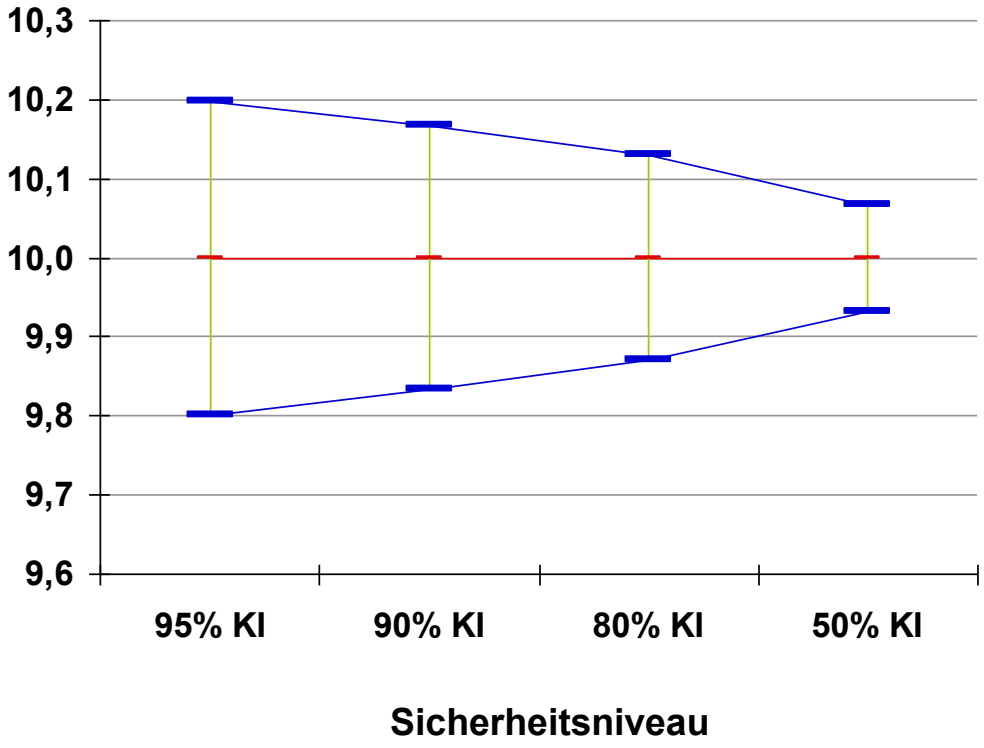



Je höher die Streuung, desto größer das Konfidenzintervall, desto ungenauer ist der geschätzte Mittelwert.

# Konfidenzintervalle: Einflüsse auf die Breite

## Konfidenzintervall und Sicherheitsniveau

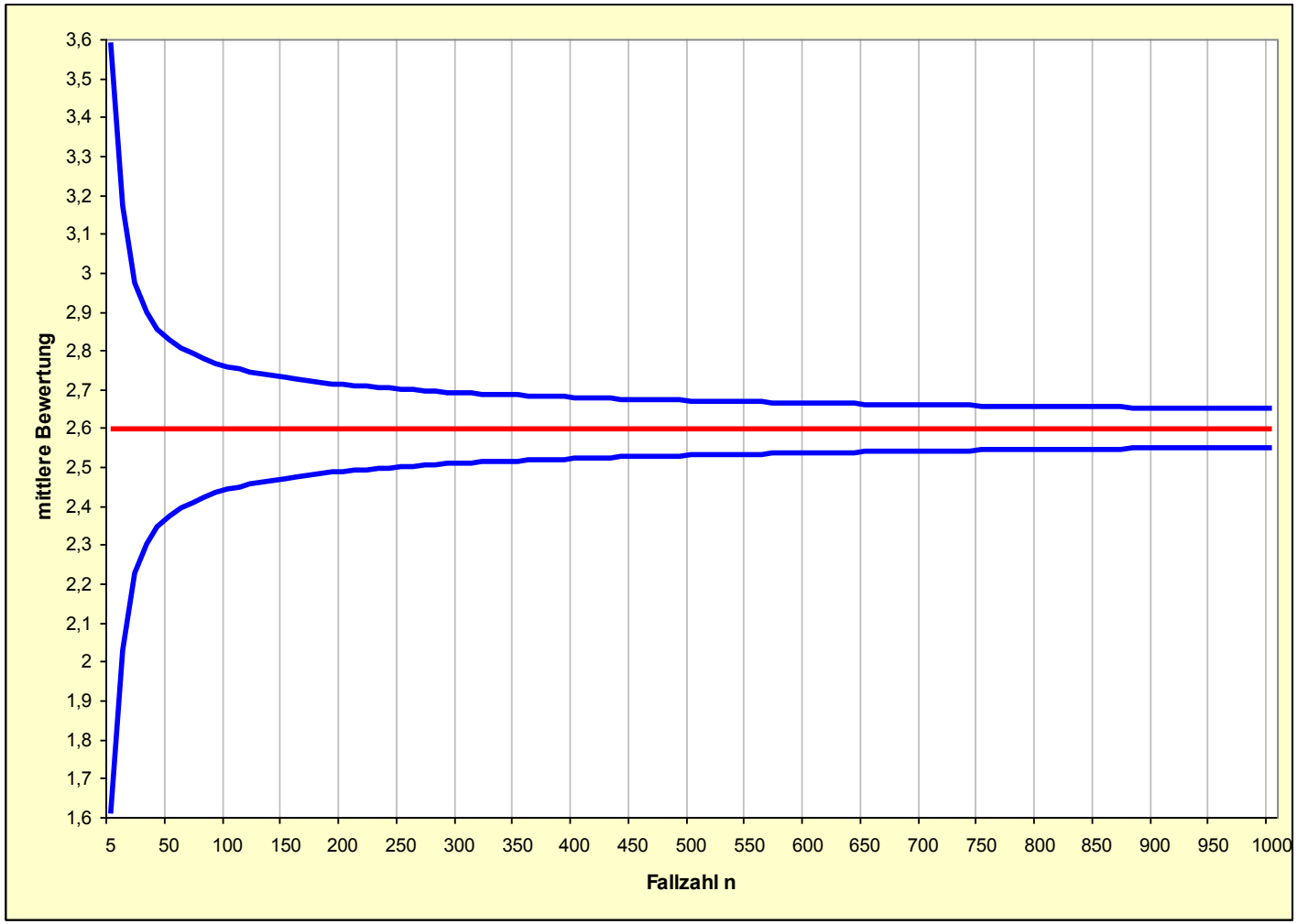
MW = 10; STDEV = 1; n = 100




Je größer die gewünschte Sicherheit, desto breiter ist das Konfidenzintervall.

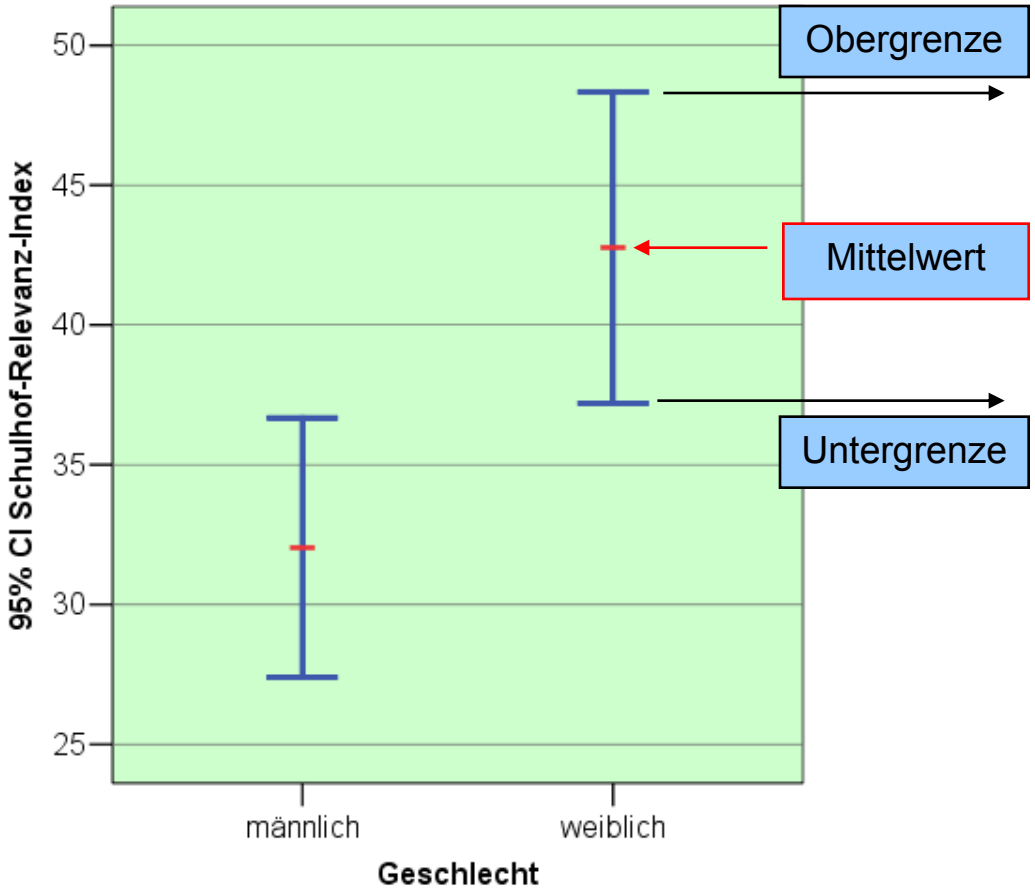
# Konfidenzintervalle: Einflüsse auf die Breite

## 95% Konfidenzintervall und Fallzahl: MW=2,6 Stdev=0,8



# Konfidenzintervalle: Beispiele

## Fehlerbalkendiagramm




Überschneiden sich die Konfidenzintervalle von 2 Mittelwerten **nicht**, dann gibt es signifikante Unterschiede.

Auch bei Überschneidungen können die Unterschiede noch signifikant sein. Hier sollte zur Sicherheit zusätzlich ein statistischer Test durchgeführt werden.

Wenn ein Mittelwert vom Konfidenzintervall eines anderen Mittelwertes überdeckt wird, sind die Unterschiede definitiv nicht signifikant.

# Statistische Signifikanztests

Zur Konstruktion von statistischen Tests formuliert man **2 Hypothesen**.

Z. B. für den t-Test für 2 unabhängige Gruppen:

| | |
|---|---|
| **Nullhypothese:** | Der Mittelwertunterschied ist zufällig. |
| **Alternative:** | Der Mittelwertunterschied ist nicht zufällig. |

Die Prüfstatistik hat Verfahren entwickelt, mit denen aus den gegebenen Stichprobenkennwerten (z. B. Mittelwert, Varianz, Fallzahl) nach bestimmten Formeln so genannte Prüfgrößen berechnet werden.

Die Irrtumswahrscheinlichkeit α ist die Wahrscheinlichkeit sich zu irren, wenn man sich für ‚Mittelwertunterschiede vorhanden' entscheidet, obwohl in der Grundgesamtheit keine Unterschiede vorliegen, wenn man sich also **für die Alternative entscheidet, obwohl die Nullhypothese richtig** ist.

# Statistische Signifikanztests

**Mögliche Fehler:**

Nullhypothese wird verworfen, obwohl sie richtig ist (Fehler 1. Art = Alpha).

Nullhypothese wird beibehalten, obwohl sie falsch ist (Fehler 2. Art = Beta).

| | Nullhypothese wahr | Nullhypothese falsch |
|---|---|---|
| Nullhypothese abgelehnt | Fehler 1. Art | Richtige Entscheidung |
| Nullhypothese beibehalten | Richtige Entscheidung | Fehler 2. Art |

**Der Fehler 1. Art ist die Irrtumswahrscheinlichkeit Alpha.**

Bei nichtsignifikanten Testergebnissen ist die Gefahr, die Nullhypothese fälschlicherweise beizubehalten desto kleiner, je größer die im Test berechnete Irrtumswahrscheinlichkeit ist.

# Statistische Signifikanztests

## Signifikanz:

Irrtumswahrscheinlichkeit p soll nicht größer als 0,05 (5%) sein, um von signifikanten Unterschieden zu sprechen.

In der Statistik hat sich aber **folgender Sprachgebrauch** durchgesetzt:

| Irrtumswahrscheinlichkeit | Bedeutung | Symbol |
|---|---|---|
| p > 0,05 | nicht signifikant | kein Stern |
| p <= 0,05 | signifikant | ☆ |
| p <= 0,01 | hoch signifikant | ☆☆ |
| p <= 0,001 | höchst signifikant | ☆☆☆ |

# Statistische Signifikanztests

Statistikprogramme geben das **beobachtete Signifikanzniveau (p-Wert)** an. Es dient der Entscheidung, ob die Nullhypothese abzulehnen ist oder nicht. Es gibt die Wahrscheinlichkeit an, dass ein statistisches Ergebnis (z. B. der Wert der Teststatistik), mindestens so extrem wie das beobachtete, herauskommen würde, wenn die Nullhypothese richtig wäre. **Wenn das beobachtete Signifikanzniveau klein genug ist (0,05 (5%) in der Markforschung verwendet), wird die Nullhypothese abgelehnt**.

Dabei ist die **Irrtumswahrscheinlichkeit** höchstens so groß, wie das beobachtete Signifikanzniveau. Dieser berechnete Signifikanzwert muss dann mit der vorgegebenen Irrtumswahrscheinlichkeit (i. d. R. 0,05) verglichen werden. Liegt der Wert über 0,05 (5%), dann spricht man von **nicht signifikant**.

Die **Nullhypothese** kann durch Stichprobenbefunde **nur widerlegt, nicht aber bewiesen werden**. Wird die Nullhypothese beibehalten, so heißt das lediglich, dass sie der Alternativhypothese vorgezogen wird, bis sie evtl. durch eine weitere Prüfung mit größeren Stichprobenumfängen widerlegt werden kann.

# Statistische Signifikanztests

**Achtung:**

**Je kleiner p (= Alpha), desto höher ist die Signifikanz.**

**Je dichter p bei 1 liegt, desto weniger wahrscheinlich sind Unterschiede zwischen den Mittelwerten.**

**Für p=1 sind die Mittelwerte identisch.**

# Statistische Signifikanztests: Beispiel

## Beispiel: Mittelwertvergleich für 2 Gruppen

Stichprobe 1: $x_{11}, x_{12}, x_{13},\ldots,x_{1n_1}$ vom Umfang n1 mit MW $\mu_1$

Stichprobe 2: $x_{21}, x_{22}, x_{23},\ldots,x_{2n_2}$ vom Umfang n2 mit MW $\mu_2$

Beide Stichproben sind näherungsweise normalverteilt mit derselben unbekannten Varianz $\sigma^2$.

Man will prüfen, ob die Mittelwerte in beiden Gruppen gleich oder verschieden sind:

$H_0: \mu_1=\mu_2$ gegen $H_1: \mu_1 \neq \mu_2$

Man konstruiert jetzt eine sinnvolle Prüfgröße. Das könnte die Differenz der Stichprobenmittelwerte sein. Um diese Differenz unabhängig von der Skala des Merkmals zu machen, dividiert man die Differenz durch den Standardfehler dieser Differenz.

# Statistische Signifikanztests: Beispiel

## Beispiel: Mittelwertvergleich für 2 Gruppen

Prüfgröße für den t-Test für 2 unabhängige Gruppen:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S.E.(\bar{x}_1 - \bar{x}_2)}$$

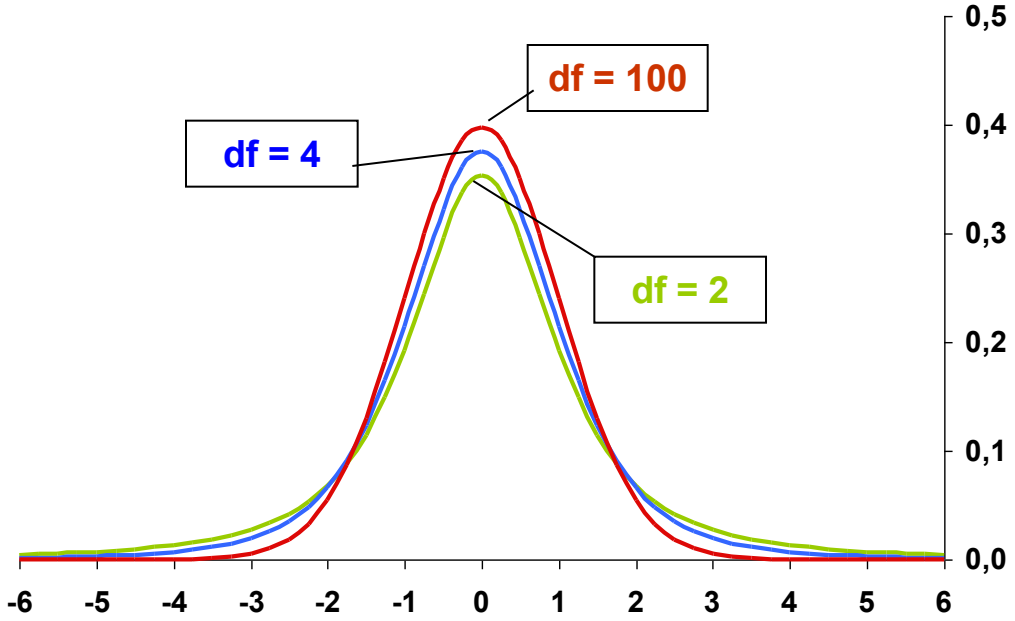
S.E. = Standard Error

Unter der Annahme, dass $\mu_1=\mu_2$ hat die Prüfgröße t eine t-Verteilung mit $n_1+n_2-2$ Freiheitsgraden (df = degrees of freedom).

Nachdem man die Prüfgröße berechnet hat, prüft man, ob der sich ergebende Wert zur hypothetischen Verteilung passt. Etwa für t-Werte zwischen -2 und +2 kann man das annehmen. Weicht t weiter von der Null ab (hypothetischer Wert der Differenz), lehnt man die Nullhypothese ab. Der Grenzwert ±2 heißt hier kritischer Wert. Er ergibt sich aus der geforderten Grenze für die Irrtumswahrscheinlichkeit, die hier mit 5% angenommen wurde.

# Die t-Verteilung

## Die t-Verteilung mit df Freiheitsgraden:




William Sealy Gosset
alias ‚Student'
(1876 – 1937)

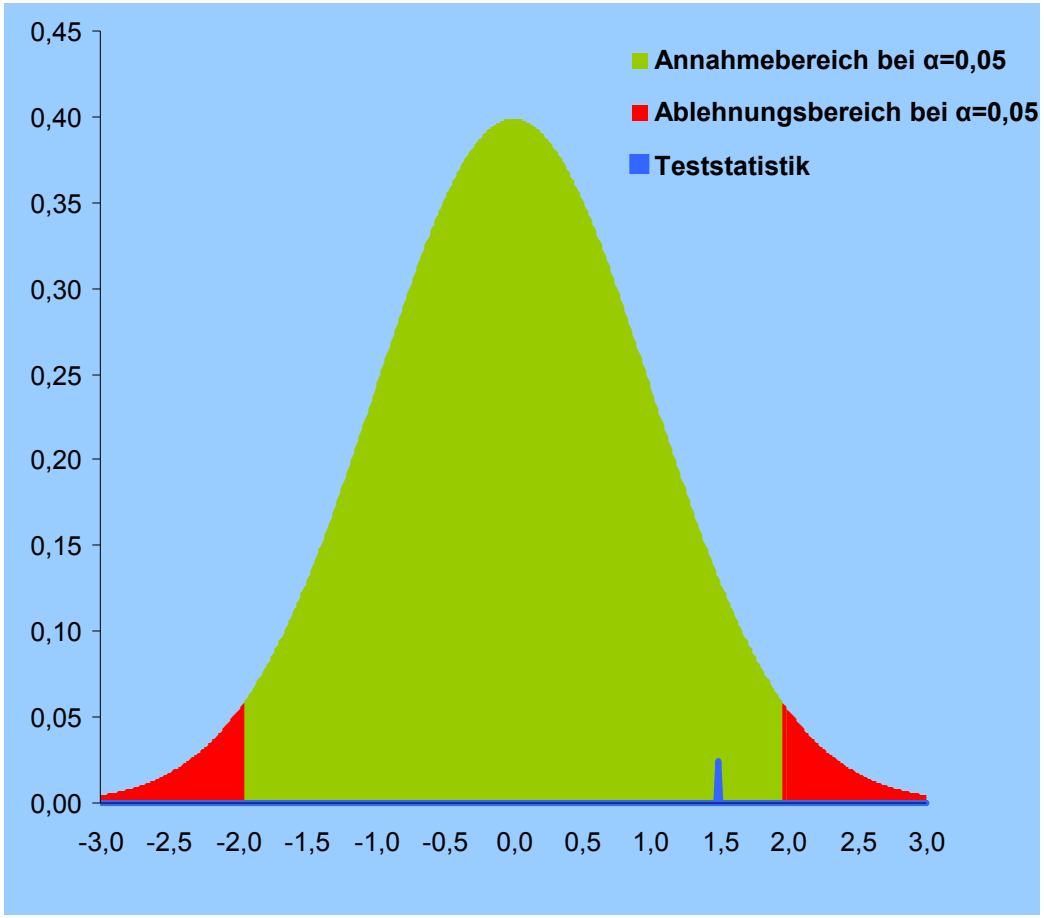
Ist ähnlich zur Normalverteilung (NV), für df > 100 fast identisch mit NV.

# Statistische Signifikanztests: Beispiel

## Beispiel: Mittelwertvergleich für 2 Gruppen

Die kritische Grenze $t_\alpha$ berechnet man so, dass

$P(|t| \geq t_\alpha) = 0{,}05$



# T-Test für unabhängige Gruppen

## Simulation des t-Tests für 2 unabhängige Gruppen

| | Mittelwert | Std-Abw. | Fallzahl | Alpha / Power |
|---|---|---|---|---|
| Gruppe 1 | 28 | 12 | 26 | 5,0% |
| Gruppe 2 | 20 | 12 | 26 | 65,3% |

*p = 0,0200*

*t = 2,404*

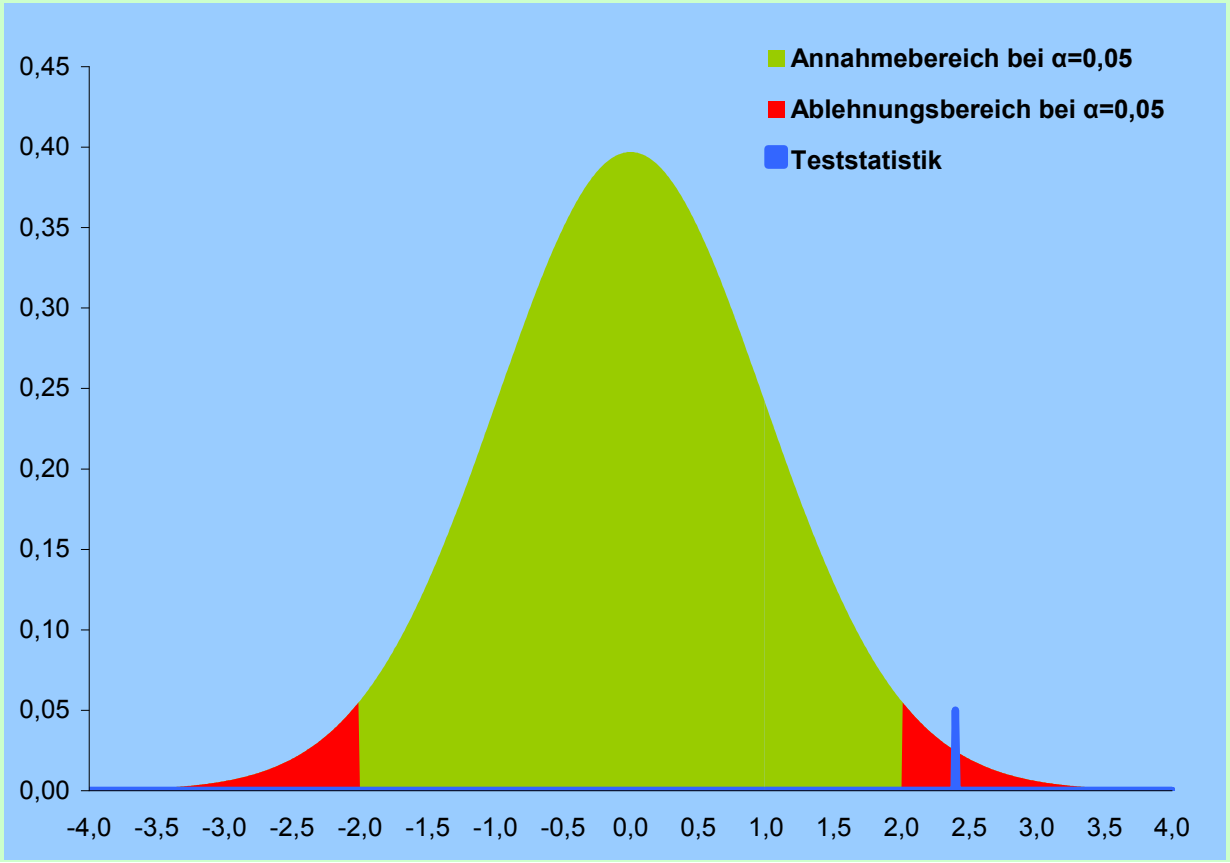
Die blauen Felder bitte nicht ändern

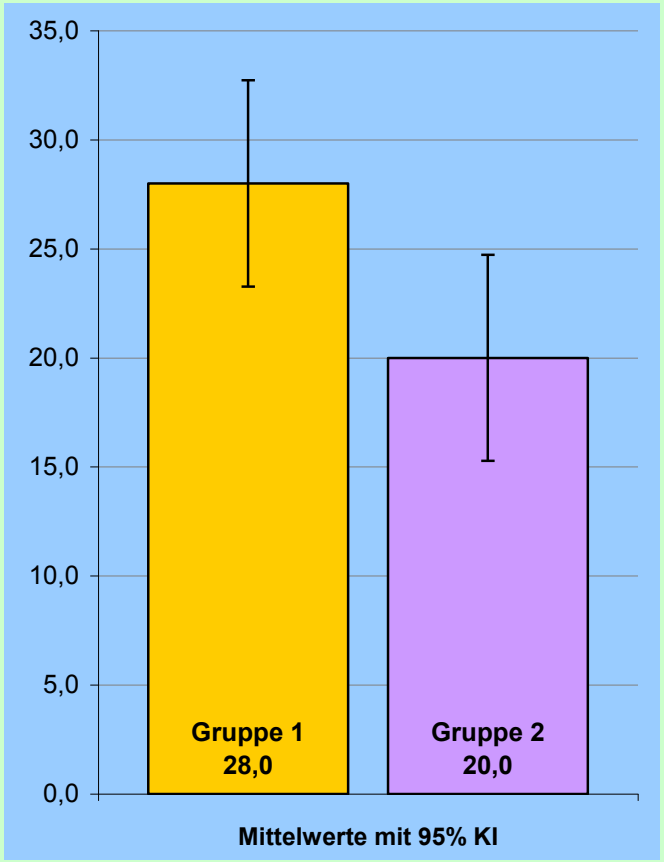
Im roten Feld Änderungen per Hand

Vollbildmodus

H0: Die Mittelwerte sind gleich

H1: Die Mittelwerte sind ungleich



# Statistische Signifikanztests: Beispiel

## Beispiel: Mittelwertvergleich für 2 Gruppen

Praktisch berechnet man heute nicht mehr die kritischen Grenzen, sondern die Wahrscheinlichkeit, für t einen Wert zu erhalten mindestens so extrem gelegen wie der berechnete unter der Annahme, dass die Nullhypothese gilt, also es keinen Mittelwert-unterschied zwischen beiden Gruppen gibt.

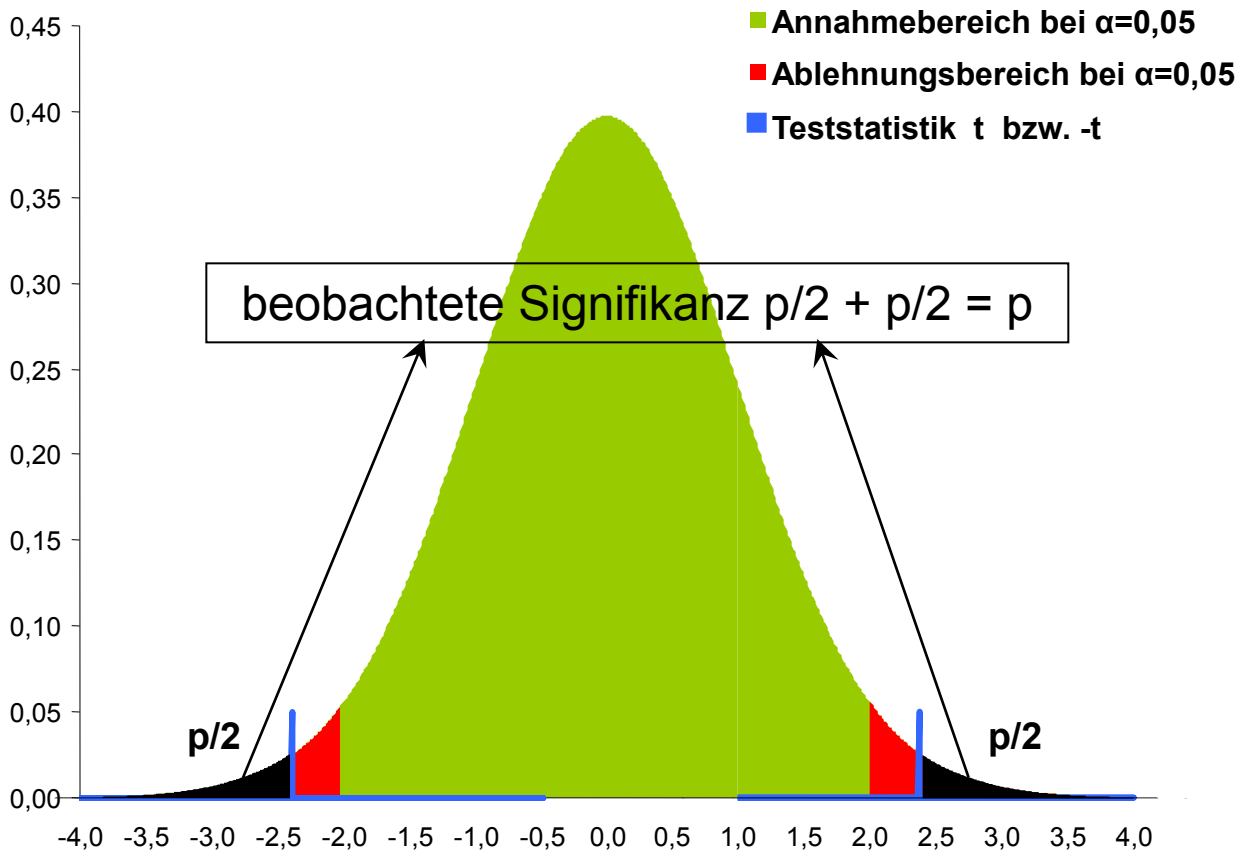
**p = beobachtete Signifikanz**

Dieser Wert wird von Statistikprogrammen mit p oder Signifikanz bezeichnet. Je kleiner dieser Wert ist, desto ausgeprägter ist der Mittelwertunterschied.

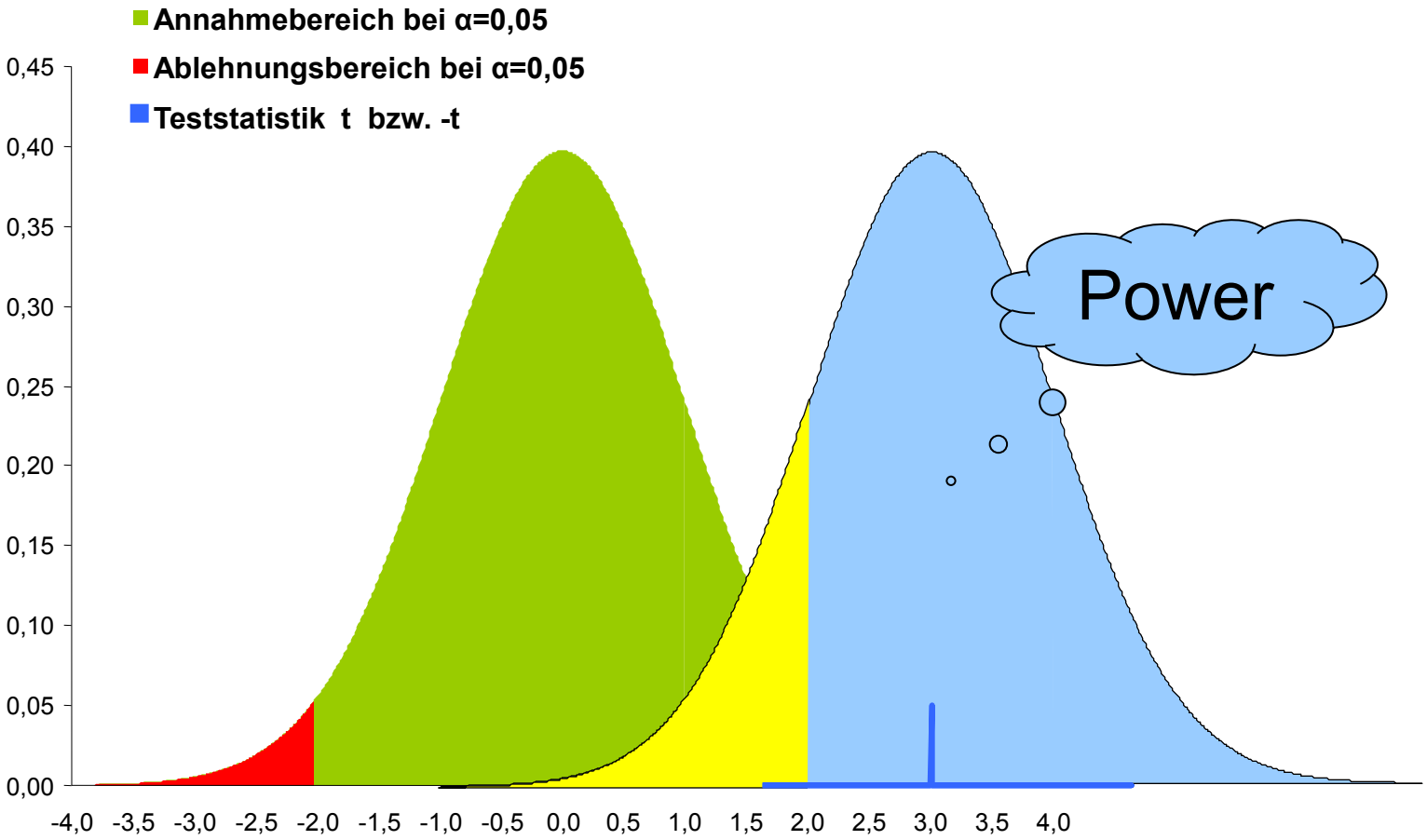
Gedankenbrücke: p ist die Wahrscheinlichkeit für die Gültigkeit der Nullhypothese

# Beobachtete Signifikanz von Tests

## Beobachtete Signifikanz



# Beobachtete Power von Tests



# Power von Signifikanztests

## Beobachtete Power =

Wahrscheinlichkeit, dass bei Versuchswiederholung ein signifikanter Mittelwertunterschied herauskommt, wenn der Unterschied in der Population genau so groß ist, wie hier beobachtet.

**Faustregel:**

Die beobachtete Power sollte ≥ 0,8 sein.

In der Marktforschung bisher kaum verwendet. In der klinischen Forschung unverzichtbar.

# T-Test für unabhängige Gruppen mit SPSS

## Beispiel: Mittelwertvergleich für 2 Gruppen

**Gruppenstatistiken**

| | | | N | Mittelwert | Standardabweichung | Standardfehler des Mittelwertes |
|---|---|---|---|---|---|---|
| Schulhof-Relevanz-Index | Geschlecht | männlich | 80 | 32,035 | 20,8572 | 2,3319 |
| | | weiblich | 80 | 42,773 | 25,0256 | 2,7979 |

**Test bei unabhängigen Stichproben**

| | | Levene-Test der Varianzgleichheit | | T-Test für die Mittelwertgleichheit | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Signifikanz | T | df | Sig. (2-seitig) | Mittlere Differenz | Standard-fehler der Differenz | 95% Konfidenzintervall der Differenz | |
| | | | | | | | | | Untere | Obere |
| Schulhof-Relevanz-Index | Varianzen sind gleich | 6,989 | ,009 | -2,948 | 158 | ,004 | -10,7375 | 3,6423 | -17,9314 | -3,5436 |
| | Varianzen sind nicht gleich | | | -2,948 | 153,03 | ,004 | -10,7375 | 3,6423 | -17,9332 | -3,5418 |

Schulhof-Relevanz-Index (Werte von 0 bis 100).

Unterschiede zwischen den Geschlechtern sind statistisch signifikant.

# Prüfung auf Normalverteilung

Der Standardtest zur Prüfung auf Normalverteilung ist der **Test von KOLMOGOROV-SMIRNOV**.

Die entscheidende Prüfgröße bei diesem Test ist die maximale Abweichung zwischen gegebener und theoretischer Verteilungsfunktion.

**Es gibt aber auch visuelle Prüfmöglichkeiten.**

- Histogramm
- Boxplot
- P-P-Diagramm
- Q-Q-Diagramm

Andrej N. Kolmogorov
(1903 – 1987)

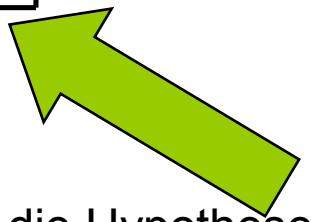
# Prüfung auf Normalverteilung

**Tests auf Normalverteilung**

Signifikanz

| | nein nicht normalverteilt | ja normalverteilt |
|---|---|---|
| Kolmogorov-Smirnov[a] | ,000 | ,177 |
| Shapiro-Wilk | ,000 | ,241 |

a. Signifikanzkorrektur nach Lilliefors

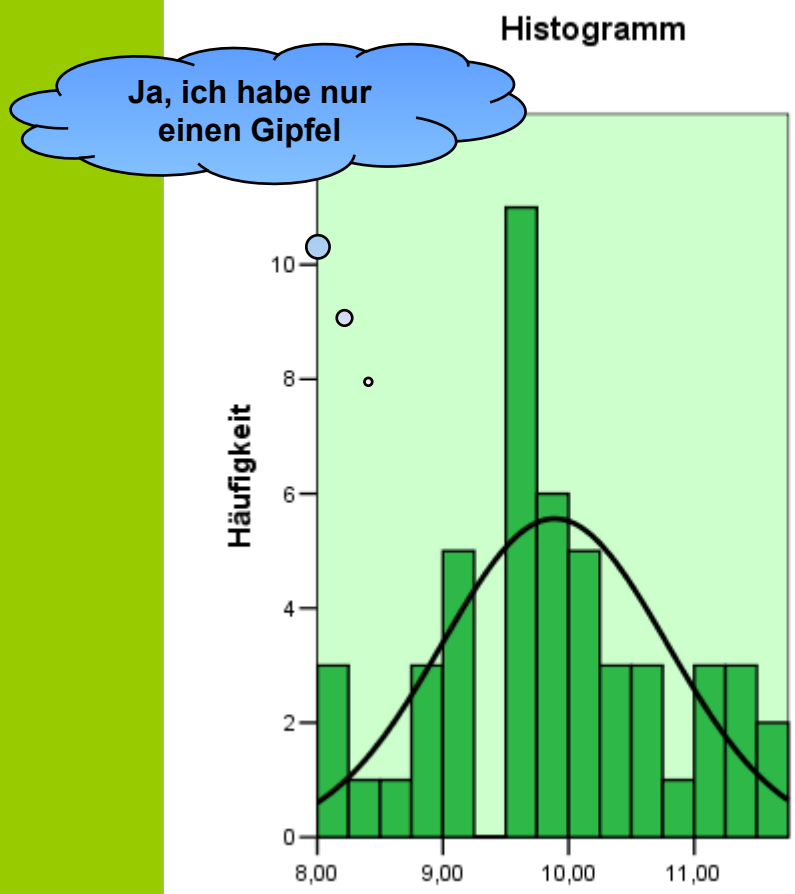


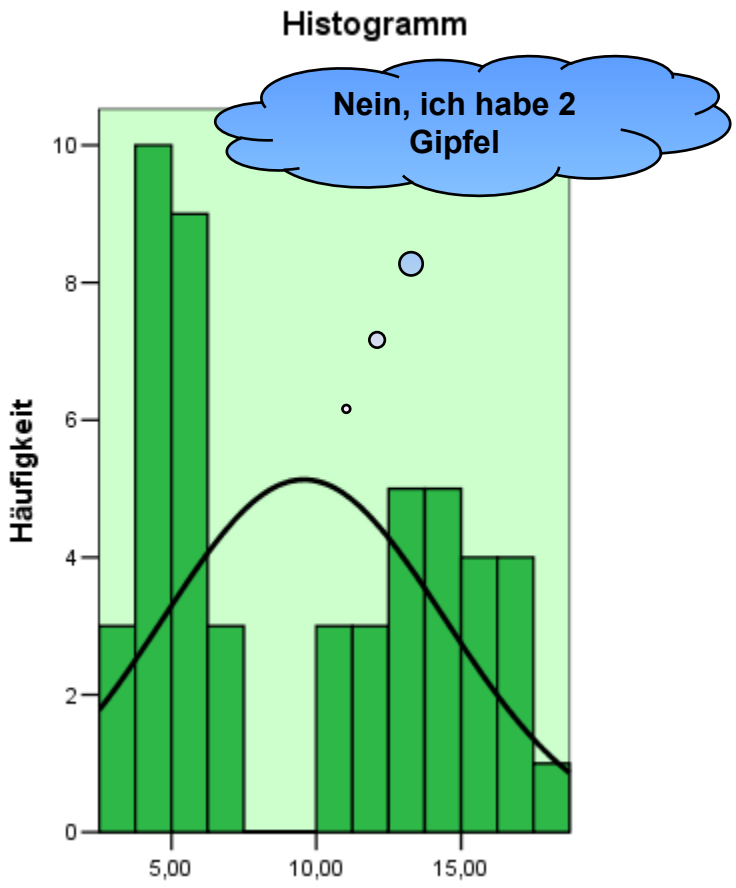
Ist der Signifikanzwert kleiner als 0,05, dann muss die Hypothese der Normalverteilung signifikant abgelehnt werden.

Ist der Wert größer als 0,05, dann kann die Hypothese der Normalverteilung nicht signifikant abgelehnt werden. Das heißt aber nicht, dass man das Vorliegen der Normalverteilung nachweisen konnte.

# Prüfung auf Normalverteilung

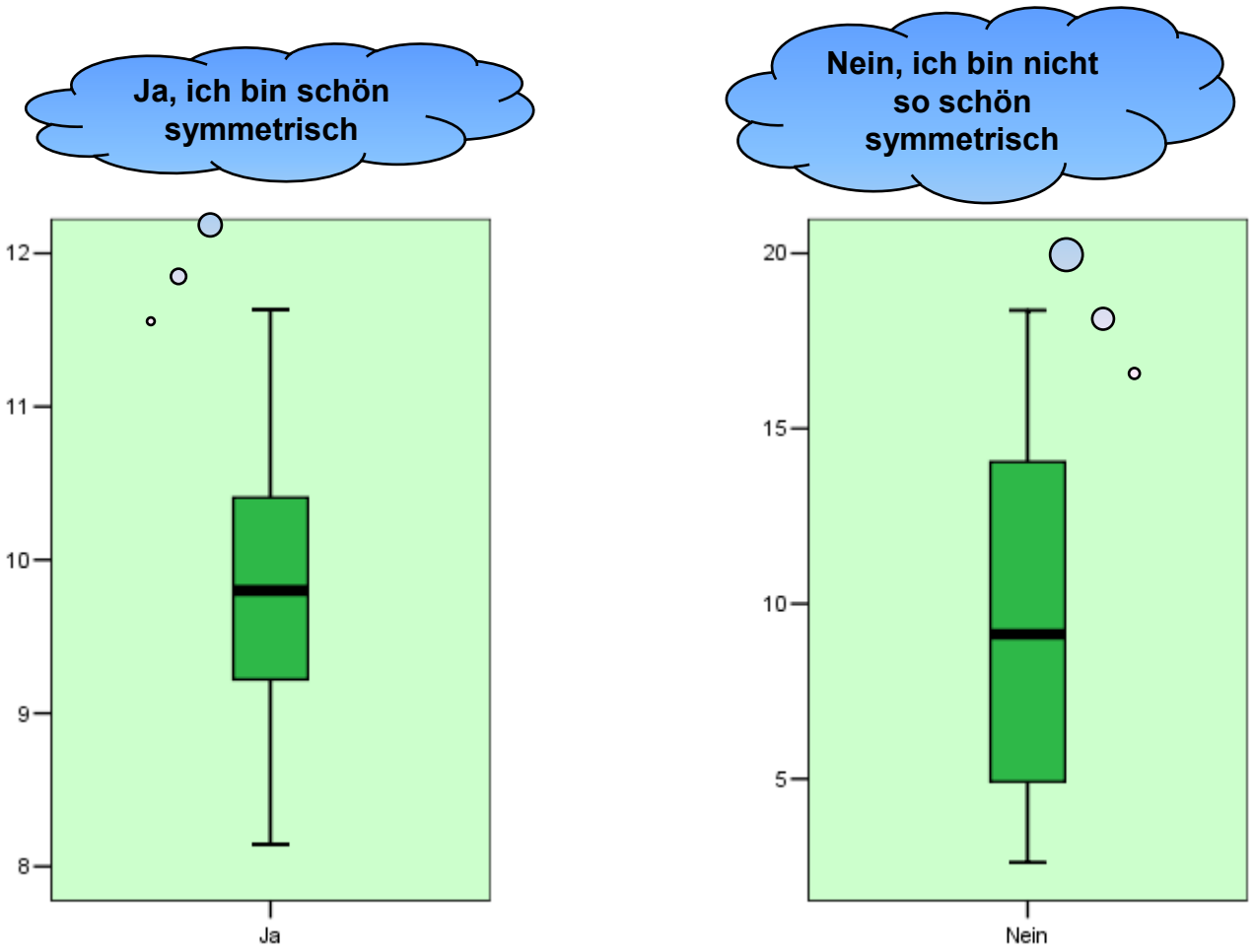
## Bin ich normalverteilt?



# Prüfung auf Normalverteilung

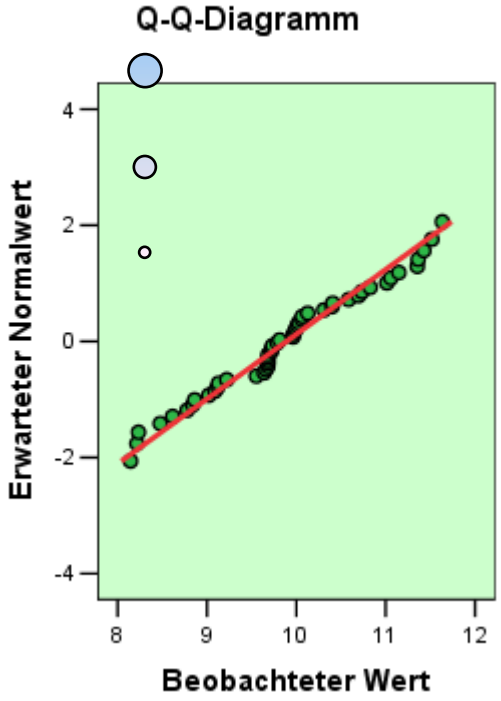
## Bin ich normalverteilt?



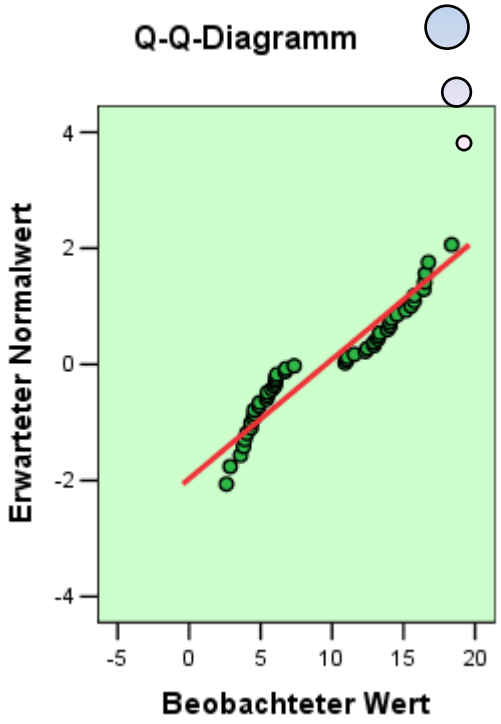
# Prüfung auf Normalverteilung

## Bin ich normalverteilt?

Ja, meine Punkte liegen dicht an der roten Linie




Nein, meine Punkte liegen nicht so schön um die rote Linie

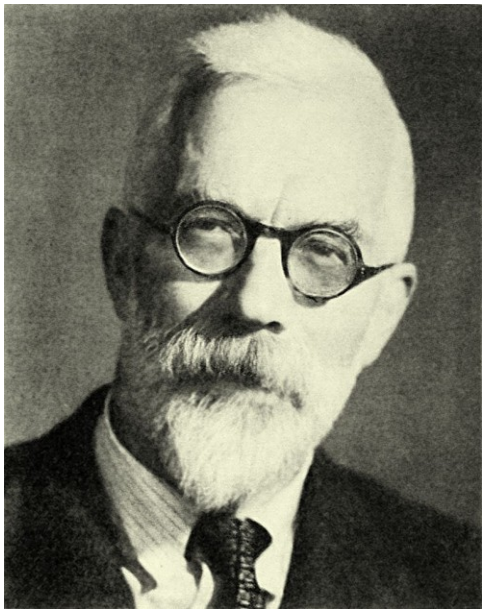


# Übersicht über Mittelwertvergleiche

Die am häufigsten verwendeten Test sind **Mittelwertvergleiche zwischen zwei oder mehr als zwei (k) Gruppen**.

SPSS und SigmaStat enthalten Ratgeber, die über das Hilfemenü aufgerufen werden. Diese führen über einige Abfragen zum richtigen Test.

Unabhängig davon wollen wir hier auf den nächsten Seiten eine Orientierungshilfe geben.



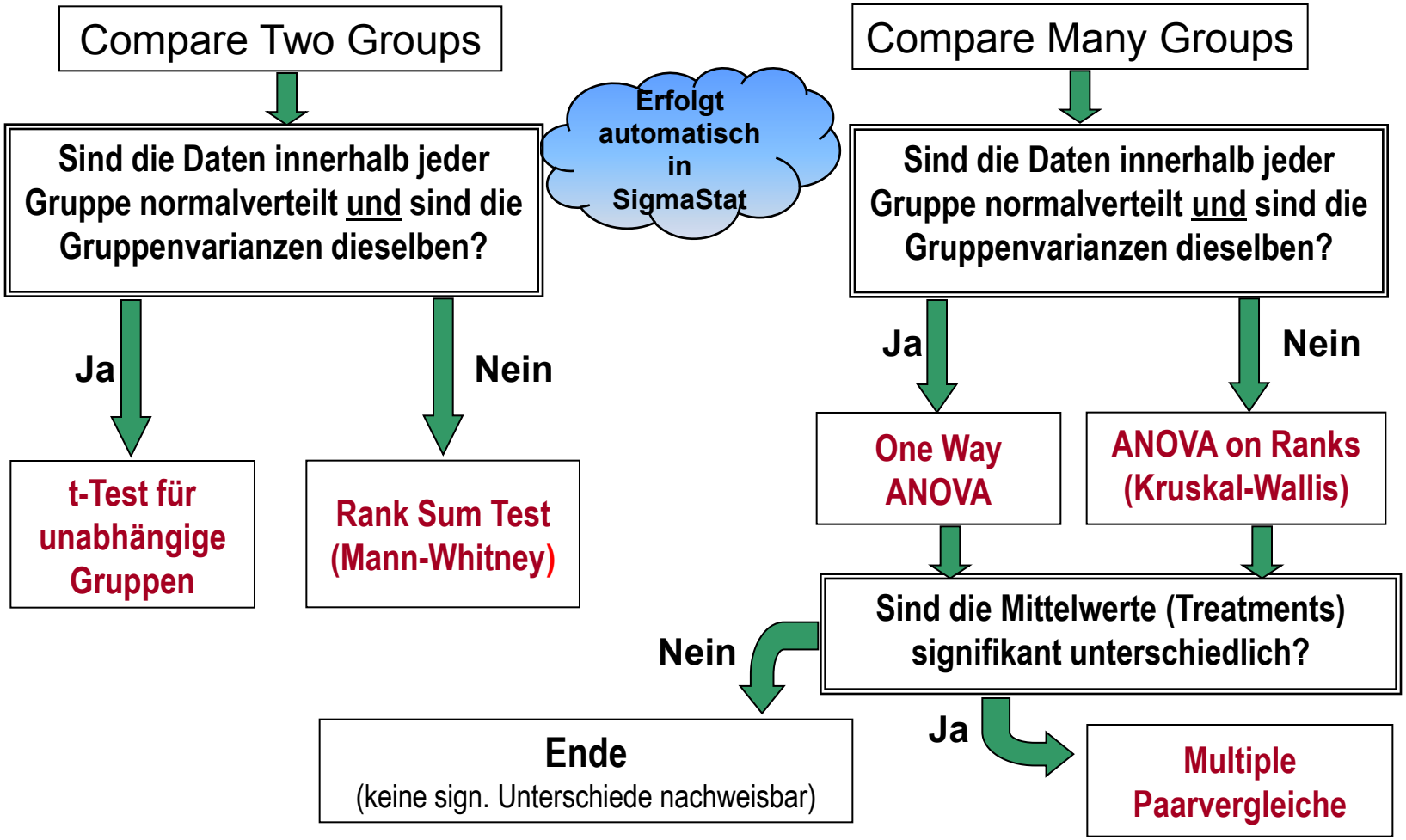
Ronald Aylmer Fisher
(1890 – 1962)

Die erste entscheidende Frage ist die nach der Abhängigkeit der zu vergleichenden Messreihen. Abhängig sind die Messreihen dann, wenn dieselben Personen oder Objekte in verschiedenen Messreihen vorkommen (z.B. Vorher-Nachher oder Messung aller 3 Tage über einen gewissen Zeitraum).

Als unabhängig gelten die Messreihen dann, wenn diese mit jeweils anderen Personen oder Objekten gewonnen wurden.

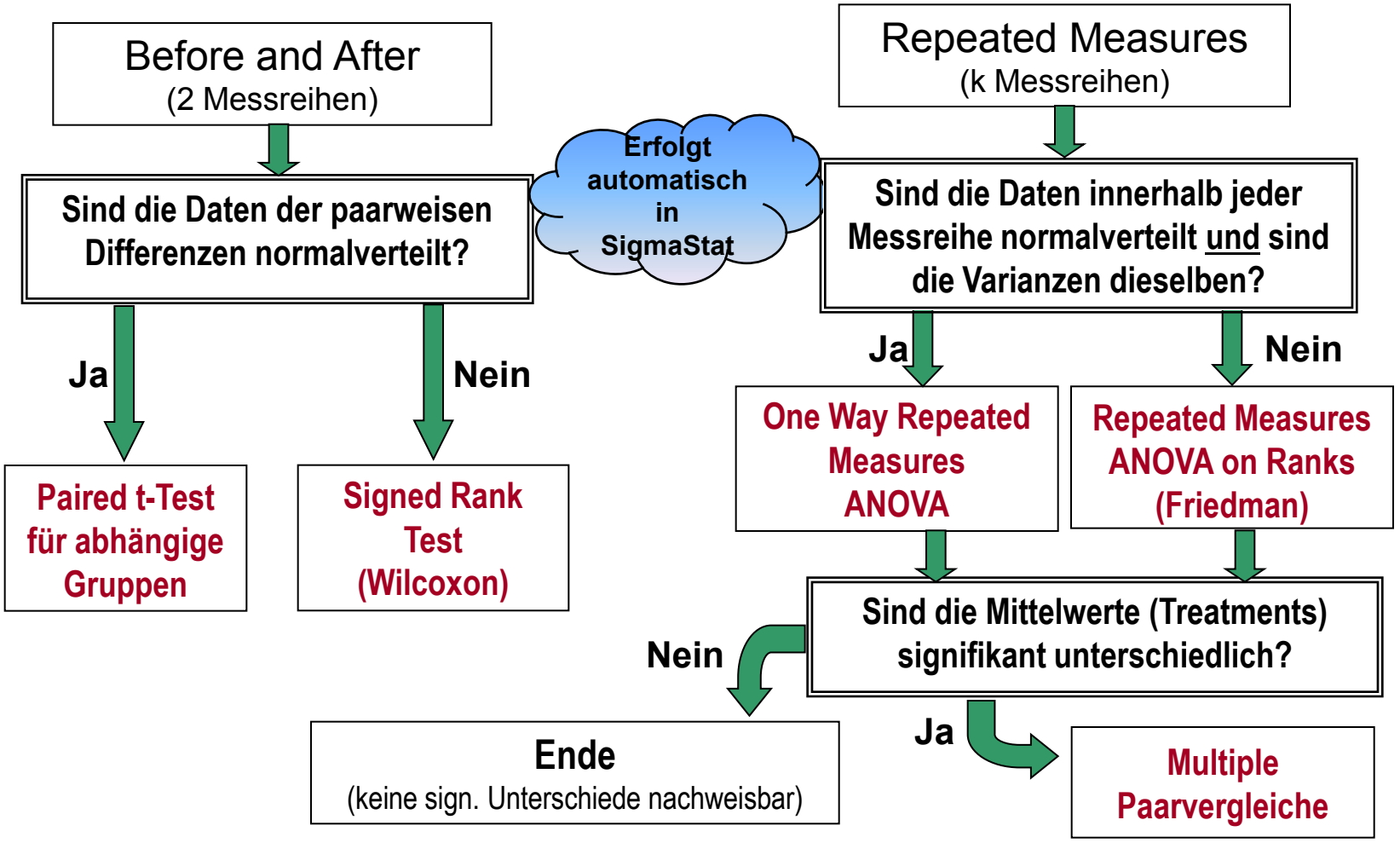
# Übersicht über Mittelwertvergleiche

## Unabhängige Stichproben



# Übersicht über Mittelwertvergleiche

## Abhängige Stichproben



# Zusammenhangsmaße

## Korrelation

Karl Pearson
(1857 – 1936)

**Maßkorrelation nach Pearson**

- Normiertes Maß für die Stärke eines linearen statistischen Zusammenhangs zwischen zwei metrischen Merkmalen. Wertebereich: [-1, 1]

**Partieller Korrelationskoeffizient**

- Normiertes Maß für die Stärke eines linearen statistischen Zusammenhangs zwischen zwei metrischen Merkmalen, wenn der Einfluss eines dritten metrischen Merkmals aus beiden Größen herausgerechnet wurde. Wertebereich: [-1, 1]

**Rangkorrelation nach Spearman**

- Normiertes Maß für die Stärke eines linearen statistischen Zusammenhangs zwischen zwei ordinalen Merkmalen. Wertebereich: [-1, 1]
  Spearman‘s Rho ist die Pearsonkorrelation der Ränge.

## Korrelation

### Gebräuchliche Beschreibungen des Zusammenhangs

| Wert des Zusammenhangs-maßes (W) | Beschreibung |
|---|---|
| **0,0 < \|W\| < 0,2** | **Sehr schwacher Zusammenhang** |
| **0,2 < \|W\| < 0,5** | **Schwacher Zusammenhang** |
| **0,5 < \|W\| < 0,7** | **Mittelstarker Zusammenhang** |
| **0,7 < \|W\| < 0,9** | **Starker Zusammenhang** |
| **0,9 < \|W\| < 1,0** | **Sehr starker Zusammenhang** |

Diese Einteilung hängt aber auch stark vom inhaltlichen Kontext ab.

# Regressionsanalysen

## Multiple Lineare Regression:

- Eine abhängige metrische Variable (Zielgröße).
- Mehrere unabhängige metrische Variablen.
- Die zu einem bestimmten „Objekt“ gehörende Ausprägung einer (metrischen) Zielgröße wird auf Grund der entsprechenden Werte einer oder mehrerer Prädiktorvariablen vorhergesagt.
- Die Vorhersage-Funktion („Regressionsgleichung“) ist eine Linear-kombination der Prädiktoren (unabhängige Merkmale).
- Gesucht ist diejenige Funktion, die die Ausprägungen der Zielgröße möglichst gut wiedergibt.
- Versucht Zusammenhänge zwischen der Zielgröße und den Prädiktorvariablen zu erkennen und zu erklären.
- Möglichkeit der schrittweisen Auswahl der Prädiktoren.
- Hat man nur eine Prädiktorvariable, dann spricht man von einer einfachen linearen Regression, sonst von einer multiplen linearen Regression.

# Regressionsanalysen

## Lineare Regression: Anwendungen in der Marktforschung:

- **Umsatzprognose einer neu zu eröffnenden Supermarkt-Filiale**
- **Wie viele Schulbücher können an eine Schule in Abhängigkeit von Größe, Schultyp und Standort durch einen Schulbuchverlag verkauft werden?**
- **Was sind die wichtigsten Treiber für die Kundenzufriedenheit? Kann man die Wichtigkeit einzelner Teilzufriedenheiten für die Gesamtzufriedenheit indirekt aus Regressionsanalysen ohne direkte Abfrage bestimmen?**
- **Wovon hängt die Dauer des Besuchs einer Internetseite ab?**
- **Wie kann ich den Gebrauchswert einer Immobilie bestimmen? Wovon hängt dieser vor allem ab, was sind die wichtigsten Einflussfaktoren?**
- **Wie kann ich die Nettokaltmiete/m² für eine Wohnung berechnen, was sind die wichtigsten Preistreiber (Mietspiegel)?**
- **Wovon hängt die Akzeptanz eines neuen Produktes ab? Wie kann man diese erhöhen?**
- **Was bestimmt den Wert von gebrauchten PKWs?**

# Regressionsanalysen

## Lineare Regression: Technische Voraussetzungen:

- Metrische Zielgröße und metrische Prädiktorvariablen
- Auch kategoriale Prädiktorvariablen sind möglich – Dummycodierung
- Stichprobe sollte genügend groß sein (>5 x so viele Fälle wie Merkmale)
- Fehler (Residuen) müssen normalverteilt sein
  - Kolmogoroff-Smirnov-Test, q-q-Diagramm, p-p-Diagramm oder weitere Residuenstatistiken und Diagramme
- Fehler müssen eine konstante Varianz aufweisen (Varianzhomogenität)
  - Residuenplots und andere Diagramme
- Die Fehler müssen unabhängig sein
  - Durbin-Watson-Test für Autokorrelationen
- Zwischen Zielgröße und Prädiktorvariable(n) wird ein linearer oder quasilinearer Zusammenhang vorausgesetzt.
- Keine starken Abhängigkeiten zwischen den Prädiktoren (Kollinearitäten), geschätzte Koeffizienten sind sonst unzuverlässig, bzw. ändern sich stark
  - Prüfung mit Korrelationskoeffizienten, Toleranzmaß, Faktorenanalyse

# Regressionsanalysen

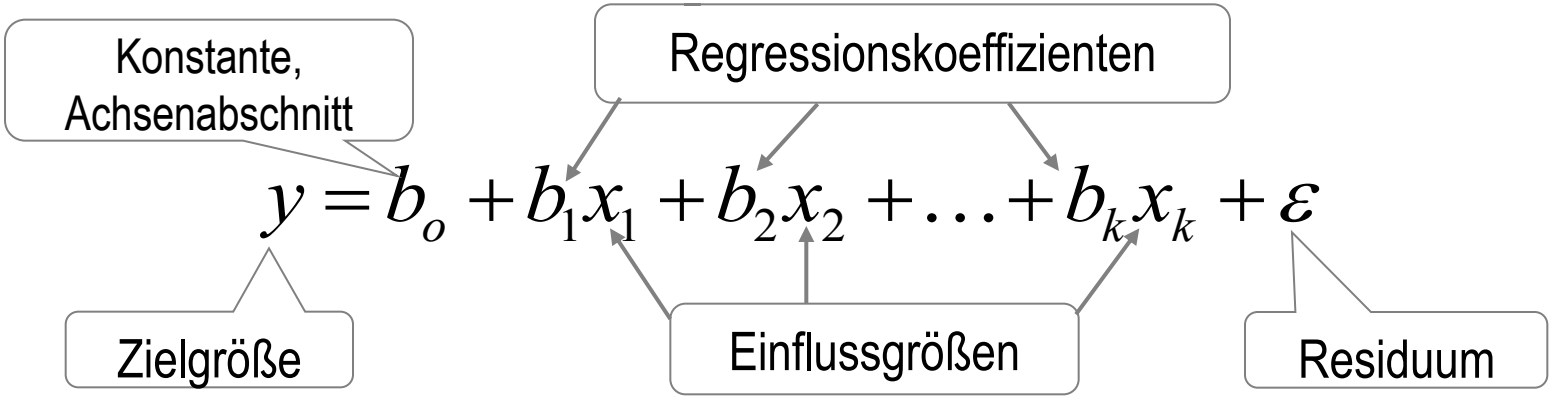
## Lineare Regression: Schätzung der Koeffizienten:

- Die Koeffizienten (Modellparameter) werden mit Hilfe der Methode der kleinsten Quadrate geschätzt (oder Maximum-Likelihood-Methode).
- Nur signifikante unabhängige Variablen ins Modell aufnehmen.
- Je mehr Koeffizienten für unabhängige Variablen geschätzt werden müssen, desto stärker hängt das Modell von der speziellen Stichprobe ab, desto schwieriger ist die Verallgemeinerbarkeit.
- Modellgüte:
  - Bestimmtheitsmaß ($R^2$): Anteil der erklärten Streuung der Zielgröße bezogen auf die Gesamtstreuung, sollte größer als 0,4 sein, besser > 0,6.
  - Korrigierte Bestimmtheitsmaß: Bestimmtheitsmaß korrigiert um die Anzahl der geschätzten Parameter.
  - F-Test: Güte der geschätzten Werte (Erklärungskraft der Regression insge-samt), Signifikanz sollte kleiner als 0,05 sein; trotz eines Bestimmtheitsmaßes von größer als 0,4 kann der F-Test einen nichtsignifikanten Wert ausgeben.
  - T-Test zur Prüfung der Regressionskoeffizienten, Signifikanz sollte kleiner als 0,05 sein.
  - Konfidenzintervalle für die Regressionskoeffizienten, das Konfidenzintervall darf den Wert Null nicht einschließen.

# Regressionsanalysen

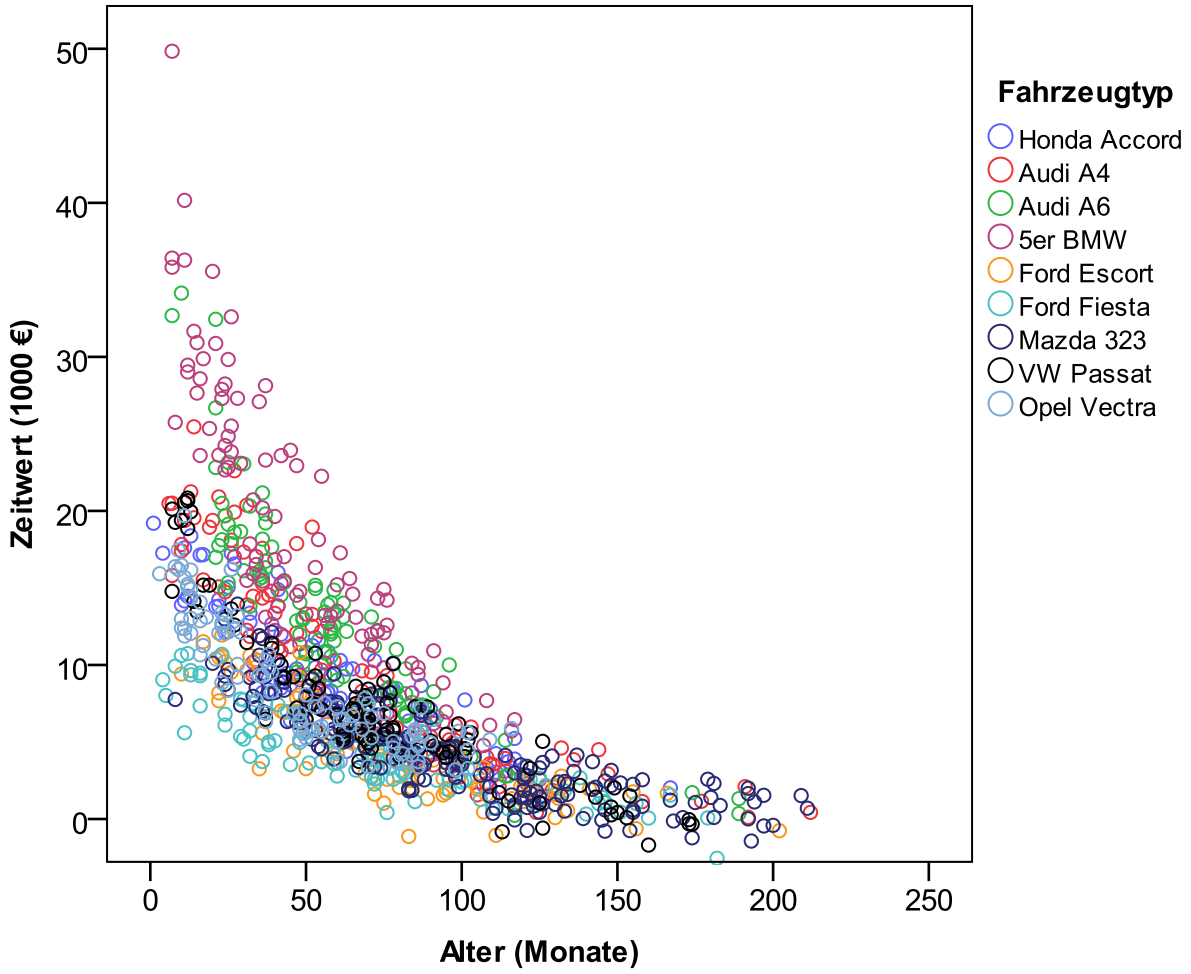
## Lineare Regression: Ergebnis:

- Das Ergebnis wird qualitativ oder quantitativ (oder beides) verwendet.
- Qualitativ in dem Sinne, dass man sich nicht wirklich für die Prognose interessiert, sondern nur dafür, welche Einflussgrößen wirken in welcher Stärke auf die Zielgröße. Was sind die erklärungsstärksten Prädiktoren.
- Quantitativ, wenn man Prognosen für bestimmte Konstellationen der Einflüsse berechnen möchte, eventuell mit Prädiktionsintervallen.
- Modellgleichung:



# Regressionsanalyse: Beispiel

## Lineare Regression: Streudiagramm Wert von Gebrauchtwagen



# Kontingenzanalyse und Chi-Quadrat

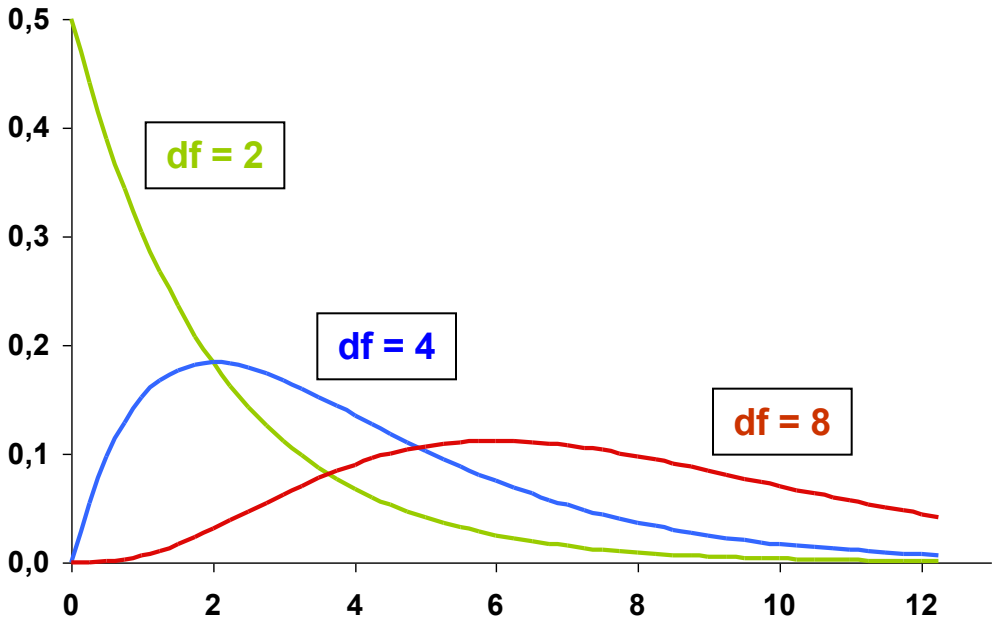
Die Grundlage einiger wichtiger Zusammenhangsmaße für nominale Merkmale ist der **Chi-Quadrat-Test**. Der Chi-Quadrat-Test basiert auf den einzelnen Zellen einer **Kreuztabelle**. Für jede Zelle wird geprüft, wie stark die beobachtete Häufigkeit von der bei Unabhängigkeit zu erwartenden Häufigkeit abweicht. Ist die Summe dieser Abweichungen groß genug, dann gibt es zwischen den beiden Merkmalen eine Abhängigkeit, d. h. zwischen diesen beiden Merkmalen besteht ein Zusammenhang. Die zu erwartenden Häufigkeiten bei Unabhängigkeit ergeben sich aus den Randverteilungen der beiden Merkmale.

$$\chi^2 = \sum_{i,j} \frac{(Beobachtet_{ij} - Erwartet_{ij})^2}{Erwartet_{ij}}$$

Bei Unabhängigkeit der beiden Merkmale ist diese Größe Chi-Quadrat-verteilt mit (r-1)*(s-1) Freiheitsgraden (r = Anzahl Zeilen; s = Anzahl Spalten)

# Die Chi-Quadrat-Verteilung

## Die Chi-Quadrat-Verteilung mit df Freiheitsgraden:




Friedrich Robert Helmert
(1843 – 1917)

# Kontingenzanalyse und Chi-Quadrat

Auf dem Chi-Quadrat-Maß basieren die **Zusammenhangsmaße** Kontingenzkoeffizient, Phi und Cramers-V. Man spricht von so genannten chi-quadrat-basierten Kontingenzmaßen.

**Cramers-V** ist ein normiertes Kontingenzmaß, das sowohl auf quadratische als auch auf rechteckige Kreuztabellen angewandt werden kann. Cramers-V = 0 bedeutet, es gibt keinen statistischen Zusammenhang, Cramers-V = 1 bedeutet, zwischen den beiden Merkmalen liegt totale Abhängigkeit vor. Cramers-V gilt als das beste chi-quadrat-basierte Kontingenzmaß.

**Cramèrs-V:**

$$V = \sqrt{\frac{\chi^2}{n^*(\min(r;\ s) - 1)}}$$

**Beispiele für Kreuztabellen:** siehe Beispiele S. 22 ff.

# Literaturempfehlungen


- Peter Eckstein (2014). Statistik für Wirtschaftswissenschaftler, 4. Auflage, Gabler-Verlag.
- Andy Field (2013). Discovering Statistics using IBM SPSS Statistics. 4th Edition, SAGE Publications Ltd. London.
- Achim Bühl (2011). SPSS 20: Einführung in die moderne Datenanalyse, 13. Auflage, Pearson Studium - Scientific Tools.
- Jürgen Bortz & Christof Schuster (2010). Statistik für Human- und Sozialwissenschaftler. 7. Auflage, Springer-Verlag.
- Rainer Schnell, Paul B. Hill, Elke Esser (2008). Methoden der empirischen Sozialforschung, 8. Auflage, Oldenbourg-Verlag.
- Jürgen Bortz & Nicola Döring (2006). Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler. 4. Auflage, Springer-Verlag.
- Lothar Kreienbrock (2004). Einführung in die Stichprobenverfahren, Oldenbourg-Verlag.
- Achim Bühl & Peter Zöfel (2000). SPSS Methoden für die Markt- und Meinungsforschung, Addison-Wesley.

# Danke, dass Sie da waren!

**Statistik-Service**
**Dr. Johannes Gladitz**
**Zionskirchstraße 27**
**10119 Berlin**
**Tel.: +49 [0] 30/44342668**
**Mobil: +49 [0] 151/67507838**

**www.statistik-service.de**
**j.gladitz@statistik-service.de**