# Einführung in Web- und Data-Science

Prof. Dr. Ralf Möller

**Universität zu Lübeck**

**Institut für Informationssysteme**

Tanya Braun (Übungen)

# Warenkorbanalyse: Kombinatorische Explosion

- Operationen über Potenzmengenverband
  - Verbesserung durch:
    - Berechnung von häufigen Artikelmengen mit Beschneidung des Suchraums (Pruning) …
    - … und anschließender Bestimmung von Assoziationsregeln durch Betrachtungen aller binären Partitionierungen der häufigen Artikelmengen
  - Aber: Join für jede Ebene → Jeder mit jedem
    - Quadratischer Aufwand!
    - Ist das wirklich handhabbar?
    - Nicht wenn viele verschiedene Artikel vorkommen!
- Können wir eventuell nur eine Teilmenge der Daten analysieren?

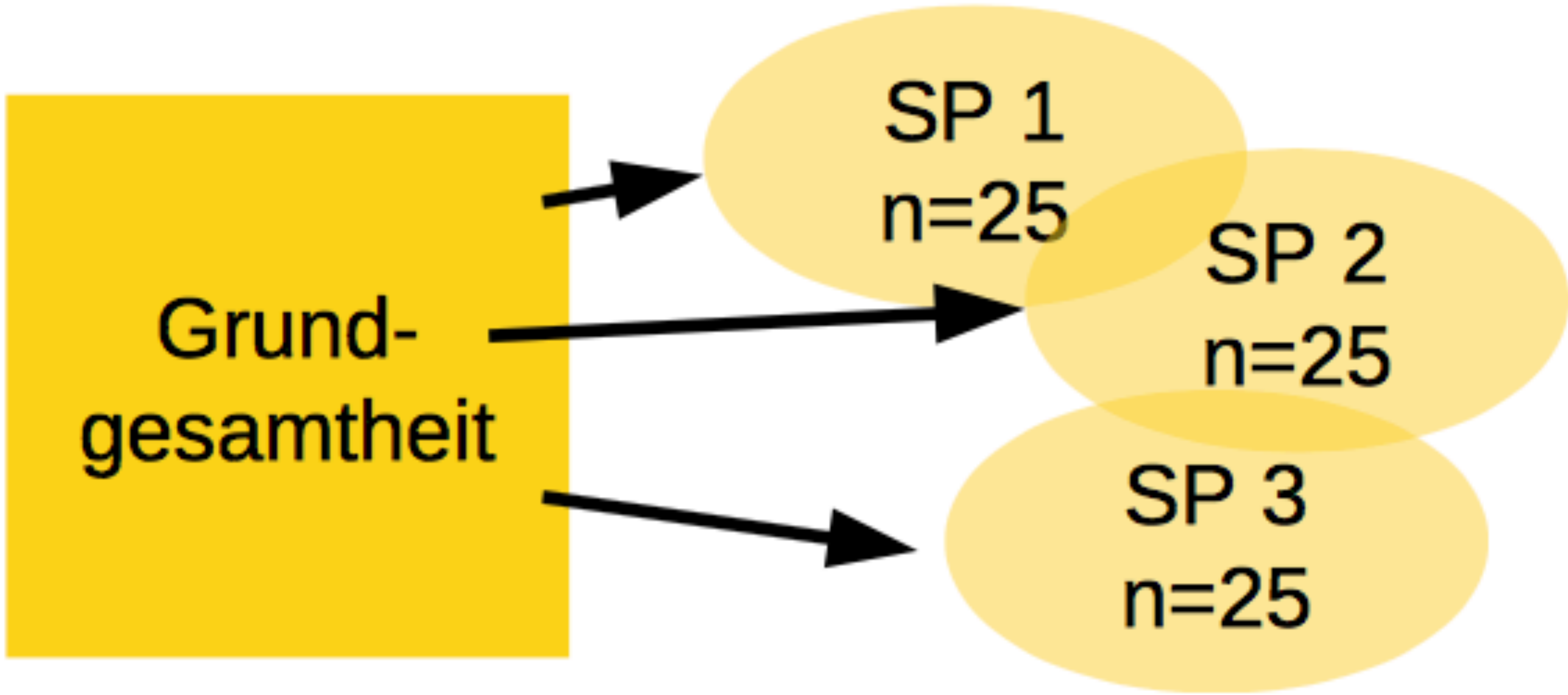
# Welche Probleme können vereinfacht werden?

- Bei allgemeiner Warenkorbanalyse erstmal kaum welche
  - Anzahl der Warenkorbtypen zählt (verschiedene Artikelmengen)
- Nehmen wir an, die Anzahl der interessierenden Warenkorbtypen sei vorgegeben (und gar nicht so groß)
  - Support- und Konfidenzberechnung immer noch aufwendig: Anzahl der Warenkörbe zählt
- Teilmenge der Warenkörbe betrachten?
  - Support- und Konfidenzwerte wie bei Gesamtmenge?
  - Wie groß muss die Teilmenge sein, um Aussagen treffen zu können?
  - Welche Teilmenge(n) auswählen?

# Betrachtung einer Teilmenge der Daten

Daten, auch Grundgesamtheit oder Population genannt

Erhebung einer Teilmenge der Daten auch Stichprobe (SP) genannt



# Betrachtung einer Teilmenge der Daten

Daten, auch Grundgesamtheit oder Population genannt

Teilmenge der Daten, auch Stichprobe (SP) genannt

| Definition | Population | Stichprobe |
|---|---|---|
| | Grundgesamtheit | Teilmenge einer Grundgesamtheit |
| Symbole | griechisch | latein |
| Mittel | $\mu$ | $\overline{x}$ |
| Standardab-weichung | $\sigma$ | s |

# Begriff der statistischen Variable

- Statische Variable ordnet einem Attribut (Merkmal) einer Erhebungseinheit (Merkmalsträger, Objekt) einen Wert zu (Merkmalsausprägung)
- Beispiele
  - Grundgesamtheit: *Einwohner der Stadt Lübeck*
    - Merkmalsträger: *ein Einwohner*
    - Merkmal: *Geschlecht*
    - Merkmalsausprägung: *männlich*
  - Grundgesamtheit: *Tage eines Untersuchungszeitraums*
    - Merkmalsträger: *ein Tag*
    - Merkmal: *Niederschlagsmenge in Deutschland*
    - Merkmalsausprägung: *1,5 Kubikkilometer*

# Statistik

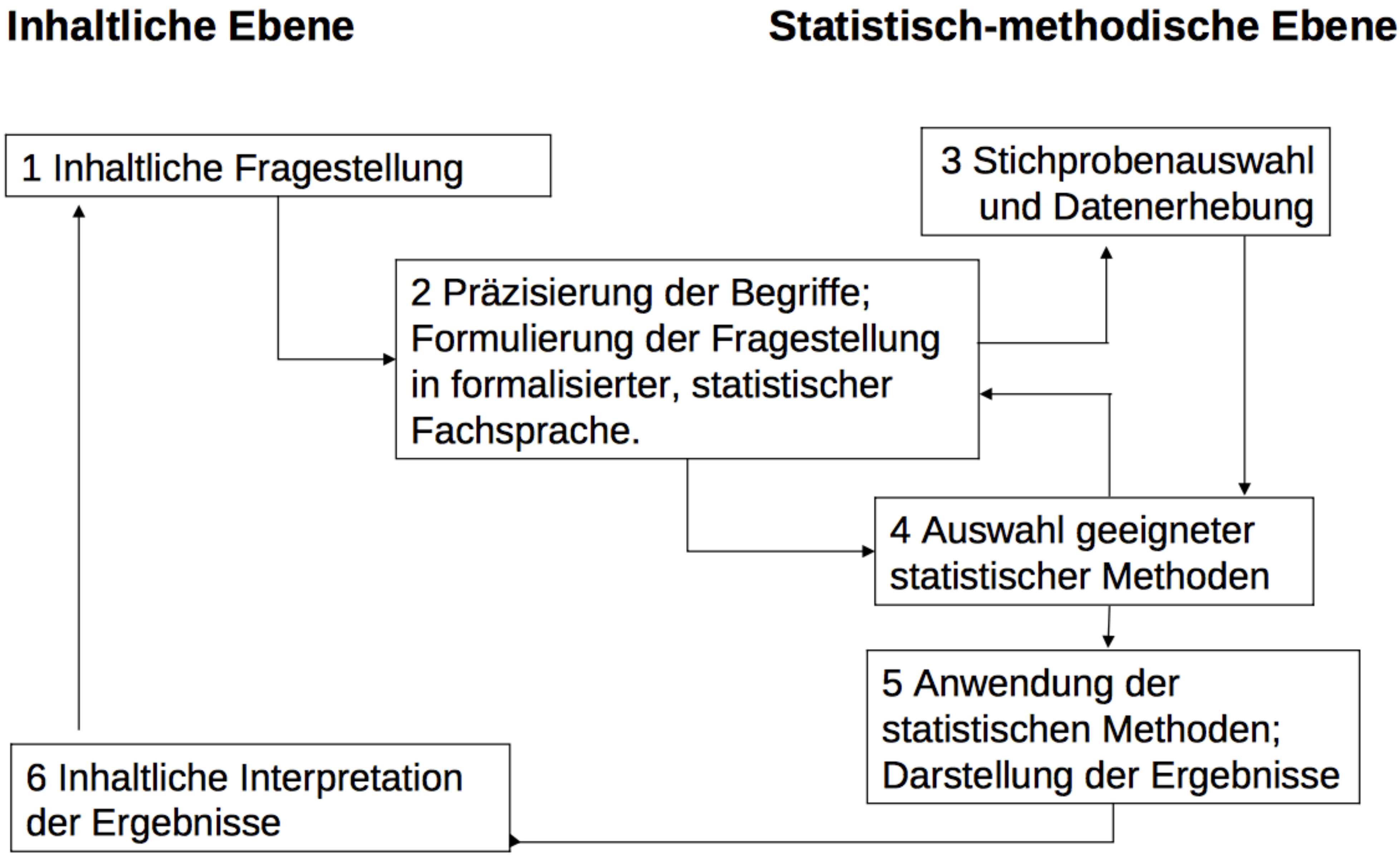
- Deskriptive Statistik
  - Beschreiben von Daten (auch: Teilmenge von Daten)
    - Beispiele: Mittelwert, Varianz, Regression, Korrelation, ...
  - Suchen nach Trends / Mustern
    - Beispiele: Häufige Artikelmengen, Assoziationsregeln
- Induktive Statistik
  - Ziel: Verallgemeinerung der Beschreibung einer Teilmenge der Daten auf Grundgesamtheit
  - Rückschlüsse auf Grundgesamtheit/Population durch Erhebung einer "repräsentativen" Stichprobe

# "Repräsentativ"

- Durch Aussagen über Stichprobe kann auf Eigenschaften der Grundgesamtheit geschlossen werden
- Elemente zufällig aus Grundgesamtheit nehmen?
- Größe der Stichprobe sollte ausreichend sein
  - Wir kommen später darauf zurück

- Zunächst: Kein formal definiertes Konzept, basiert je nach Anwendung in vielen Fällen eher auf plausiblen Argumenten

# Ablauf systematischer Untersuchungen



# Planung von Auswerte-Untersuchungen

- Welche Stichprobeneinheit soll verwendet werden?
  - Welche Skalierung/Normalisierung der Daten?
- Welches räumliche Probennahmemuster verwenden?
  - Welche Aufteilung einer Fläche zur Beprobung?
- Welches zeitliche Probennahmemuster verwenden?
  - Was sind angemessene Intervalle?

# Erhebung von Stichproben

- Verbundene Stichproben
  - z.B. wiederholte Messungen am gleichen Versuchsobjekt
  - Stichprobe zu einem Zeitpunkt kann Einfluss auf Stichprobe eines anderen Zeitpunkts haben
- Unverbundene Stichproben
  - Stichproben haben keinen Einfluss aufeinander
  - z.B. unterschiedliche Populationen, Vergleich unterschiedlicher Objekte

# Merkmale / Variablen

| | | |
|---|---|---|
| **Objekt/Merkmalsträger** | Individuum, an dem interessierende Größen erhoben werden. | Person, die einen Fragebogen ausfüllt |
| **Merkmal** | Die interessierende Eigenschaft des Objektes | Qualitativ: Geschlecht, Religionszugehörigkeit<br>Quantitativ: Alter, Einkommen |
| **Merkmalsausprägung** | Mögliche Werte eines Merkmals | Kategorien; Zahlenwerte |

Experimente werden normalerweise gestaltet, um den Einfluss eines oder mehrerer Faktoren auf eine Variable zu untersuchen

# Systematischer Fehler/Trend (Bias)

- Auftretender, meist störender systematischer Effekt mit einer Grundtendenz, so dass Werte von den wahren Ergebnissen abweichen
- Beispiele
  - Schätzung von Fischpopulationen mit Netzen einer bestimmten Maschenweite: kleine Fische können immer entkommen
  - Fangen von Säugetieren: manche Individuen sind "trap happy", manche sind "trap shy"

# Lagemaße - Mittelwerte

- Arithmetisches Mittel

$$\bar{x}_{\text{arithm}} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Geometrisches Mittel

$$\bar{x}_{\text{geom}} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Das geometrische Mittel zweier Zahlen a und b liefert die Seitenlänge eines Quadrates, das den gleichen Flächeninhalt hat wie das Rechteck mit den Seitenlängen a und b

$$\log_a \bar{x}_{\text{geom}} = \frac{1}{n}\sum_{i=1}^{n} \log_a x_i \,,$$

Relevant u.a. für logarithmierte Daten, z.B. Populationswachstum

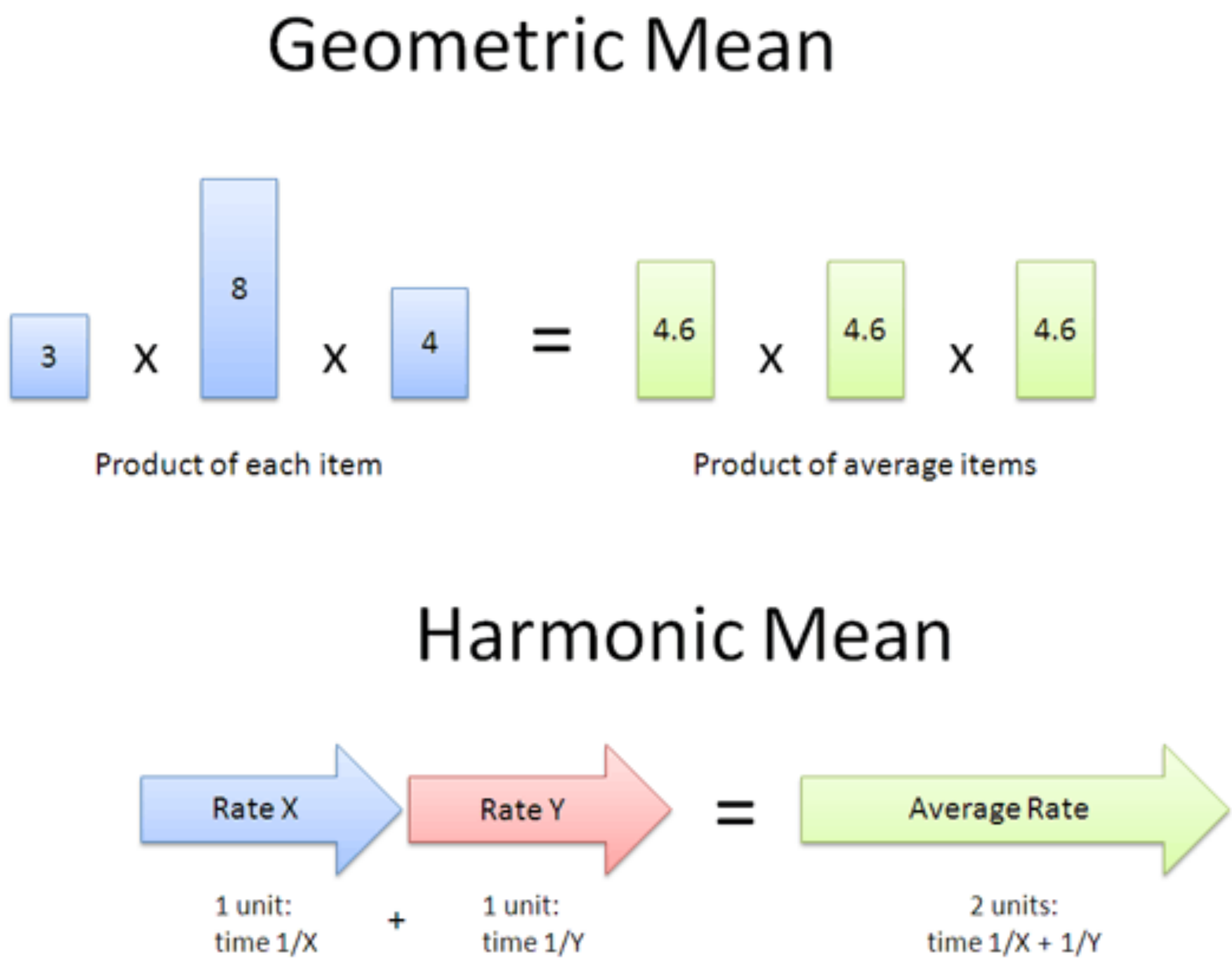
- Harmonisches Mittel

$$\bar{x}_{\text{harm}} = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

$$\frac{1}{\bar{x}_{\text{harm}}} = \frac{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}{n}$$

- $\min(x_1, \ldots, x_n) \leq \bar{x}_{\text{harm}} \leq \bar{x}_{\text{geom}} \leq \bar{x}_{\text{arithm}} \leq \max(x_1, \ldots, x_n).$

# Visualisierung



# Weitere Lagemaße

- Median (der Wert, der bei einer Auflistung von Zahlenwerten in der Mitte steht)

  4, 1, 37, 2, 1 → Median = 2 (1, 1, 2, 4, 37)

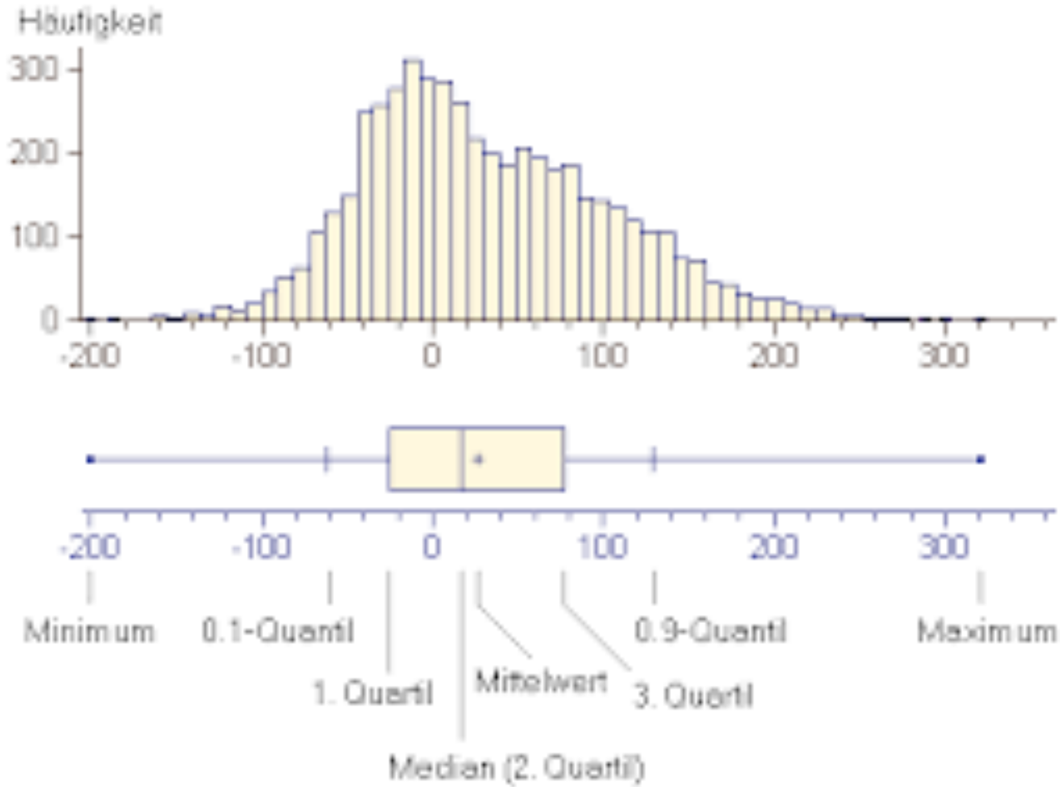
- Modalwert

  2, 2, 3, 5, 5, 5, 9, 9, 15

- Quantil, Quartil

  Geordnete Reihe der Merkmalsausprägungen wird in gleichgroße Teile zerlegt

  Kumulierte Häufigkeiten

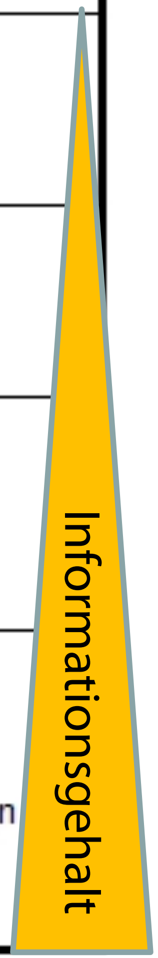


# Anwendungen

| Name & Meaning | Formula / Example | Used for |
|---|---|---|
| Arithmetic Mean<br>[average] | $\frac{sum}{size} = \frac{a+b+c}{3}$ | Most situations<br>("average item") |
| Median<br>[middle value] | Middle of sorted list<br>(2 middles? Average 'em) | Wildly varying samples<br>(houses, incomes) |
| Mode<br>[most popular] | Most popular value | No compromises<br>(winner takes all) |
| Geometric Mean<br>[average factor] | $\sqrt[3]{abc}$ | Investments, growth, area, volume |
| Harmonic Mean<br>[average rate] | $\frac{3}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}}$ | Speed, production, cost |

# Datentypen

| *Skalenniveaus* | *mögliche Aussagen* | *mögliche Methoden (Beispiel Lage)* | *Beispiele* |
|---|---|---|---|
| **Nominal**<br>(keine Ordnung der Daten möglich) | **1. Gleichheit und Ungleichheit** (=, #) können festgestellt werden) | Häufigkeiten, relative Häufigkeiten,<br>Modalwert | Lieblingszeitungen<br>Geschlecht<br>Studienrichtungen |
| **Ordinal**<br>(größenmäßige Ordnung möglich, aber Abstände ohne Aussagekraft) | 1 Gleichheit und Ungleichheit<br>**2. Rangreihung (<, >, =)** | dazu: z.B. kumulierte Häufigkeiten, Median | Beliebtheitsrangliste<br>soziale Schichten |
| **Intervall**<br>(Abstände können interpretiert werden, nicht aber das Verhältnis von Größen) | 1. Gleichheit und Ungleichheit<br>2. Rangreihung<br>**3. Gleichheit der Unterschiede** | dazu: u.a.<br>arithmetisches Mittel | Intelligenzquotient<br>Temperatur |
| **Verhältnis**<br>(die Ausprägungen haben einen absoluten Nullpunkt; das Verhältnis kann interpretiert werden) | 1. Gleichheit und Ungleichheit<br>2. Rangreihung<br>3. Gleichheit der Unterschiede<br>**4. Proportionalität**<br>**x11=3*x12** | dazu: u.a.<br>geometrisches Mittel | Alter<br>Preis<br>Größe<br>Nahrungswert in Kalorien<br>Inflation |



# Metrische Variablen

- Intervall- und Verhältnisskala werden zur sog. Kardinalskala zusammengefasst.
- Merkmale auf dieser Skala werden dann als metrisch bezeichnet

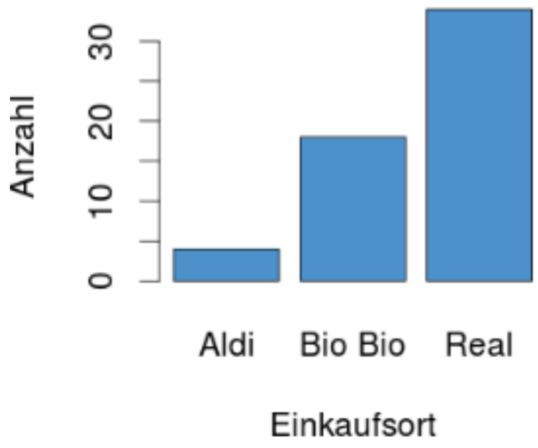
# Kategoriale Variablen

- nominalskalierte Variablen
- ordinalskalierte Variablen
- metrische Variablen, die nur wenige Ausprägungen haben (nicht von allen Autoren unterstützt)
- Variablen, die durch Kategorisierung aus ordinalskalierten oder metrischen Variablen entstanden sind (Beispiel: Variable „Einkommen“ mit den Kategorien „500-999 €“, „1000-1499 €“ usw.)

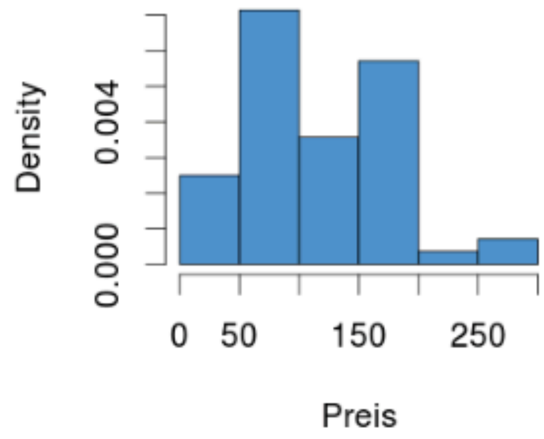
# Streuungsmaße

- Spannweite
  - Maximale Differenz zwischen zugrunde liegenden Daten
  - Mindestens Ordinaldaten notwendig
- Varianz
  - Mittlere quadratische Abweichung der einzelnen Datenwerte vom arithmetischen Mittelwert
  - Einheiten quadriert
- Standardabweichung
  - Als Standardabweichung bezeichnet man die Wurzel aus der Varianz
  - Streuungsmaß besitzt dieselbe Einheit wie die Daten und der Mittelwert

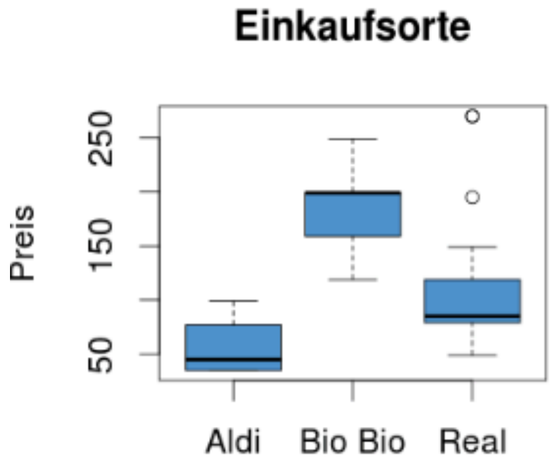
# Darstellung von Dateneigenschaften



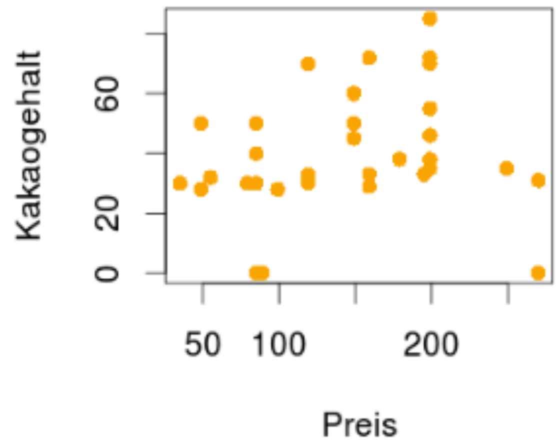

Säulendiagramm




Histogramm




Boxplot




Scatterplot

# Darstellung von Daten

Barplot/Säulendiagramm/Balkendiagramm

- Nominale und ordinalskalierte Variablen: Anzahl

Histogramm

- Ordinalskalierte oder metrische Variablen

Scatterplot

- Für 2 Variablen
- Normalerweise metrische Variablen

Boxplot

- Metrische Variablen, die verschiedenen Kategorien angehören können.

# Relative Häufigkeiten

- Histogramm: Zähler für Anzahl von Ausprägungen
  - Häufigkeitsverteilung
- Normierung der Anzahlen auf [0, 1] (Skalierung) ergibt relative Häufigkeiten
- Verteilung meist in Bezug auf relative Häufigkeiten betrachtet

# Verteilungen

- Einige Verteilungen, die natürlich vorkommen
  - Exponentialverteiltung (hatten wir schon)
    - Städte (nominal) Anzahl Einwohner (metrisch)
    - Über Einwohner wird Städte sortierbar
  - Binomialverteilung
  - Normalverteilung

- Beschreibung durch Funktion

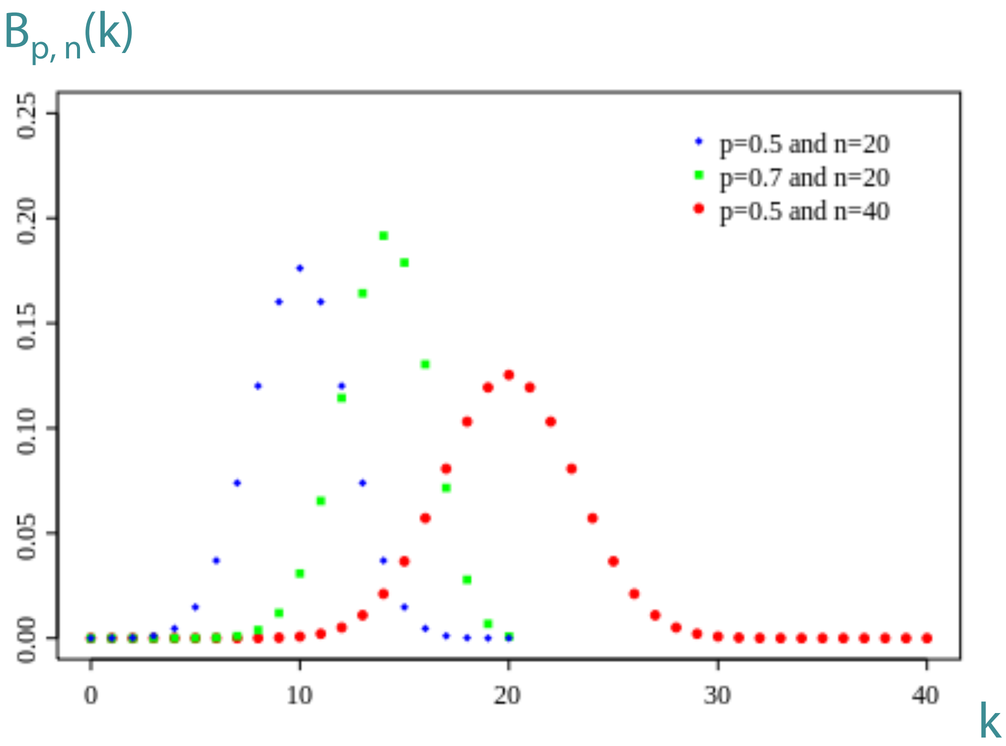
$$f: \text{Grundmenge} \longrightarrow [0, 1]$$

# Binomialverteilung

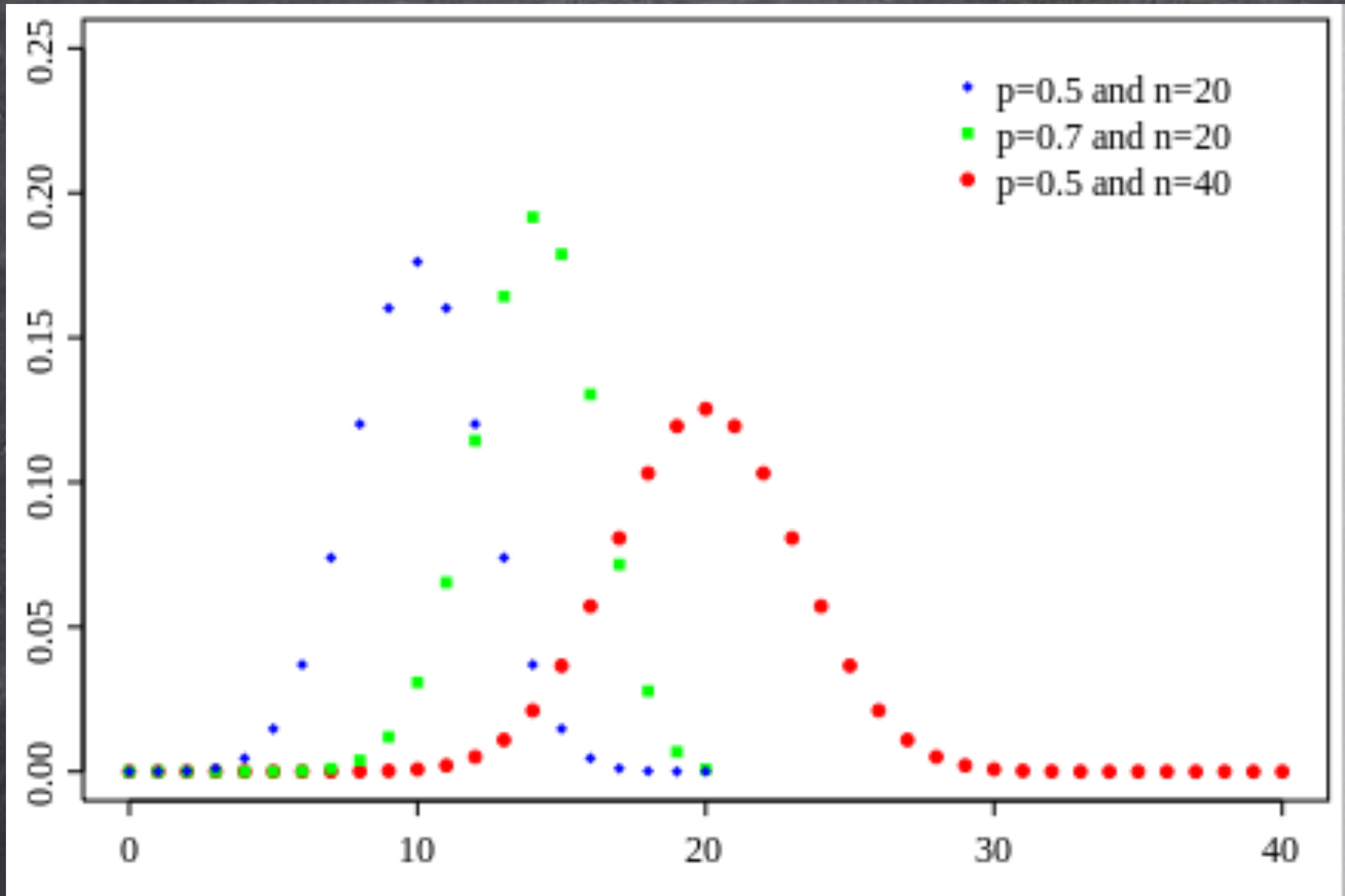
- Beschreibt Anzahl der Erfolge in einer Serie von gleichartigen und unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben: „Erfolg“ oder „Misserfolg“

- $n$ = #Versuche
  $p$ = #erfolgr. Vers. / #Versuche

- Beschreibung der relativen Häufigkeit, genau $k$ Erfolge zu erzielen, als Funktion $B_{p,n}(k)$



# Aufgabe




Wie bestimmen wir die relative Häufigkeit, bis zu k Erfolge zu erzielen?

$\sum_{i=0}^{k} B_{p,n}(i)$

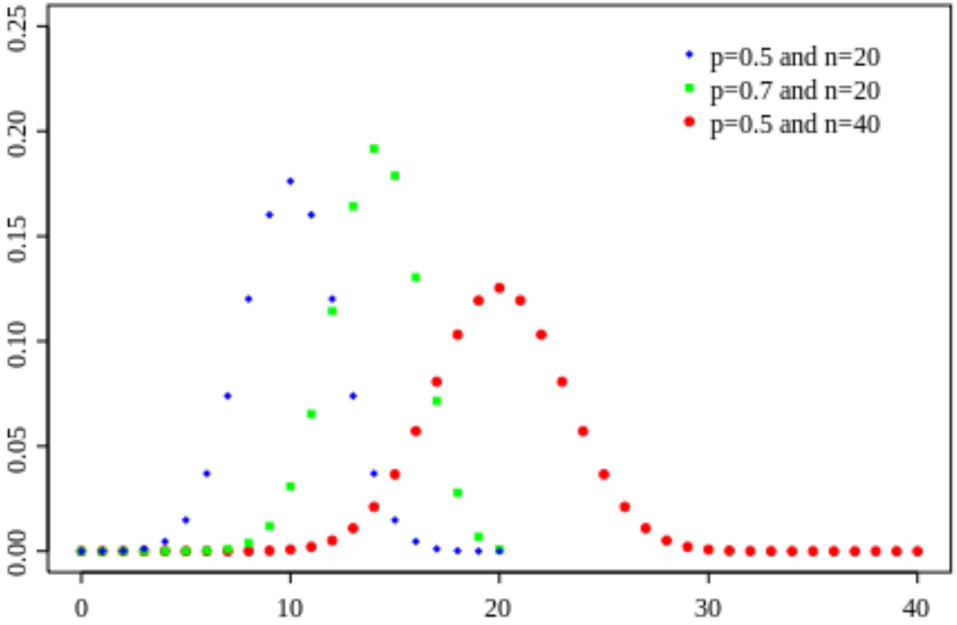
# Binomialverteilung

- Beschreibt Anzahl der Erfolge in einer Serie von gleichartigen und unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben:
  „Erfolg“ oder „Misserfolg“

- $n$ = #Versuche
  $p$ = #erfolgr. Vers. / #Versuche
- Beschreibung der relativen Häufigkeit, genau $k$ Erfolge zu erzielen, als Funktion $B_{p,\,n}(k)$

- Es gilt: $\Sigma_{i=0}^{n} B_{p,n}(i) = 1$

$B_{p,\,n}(k)$



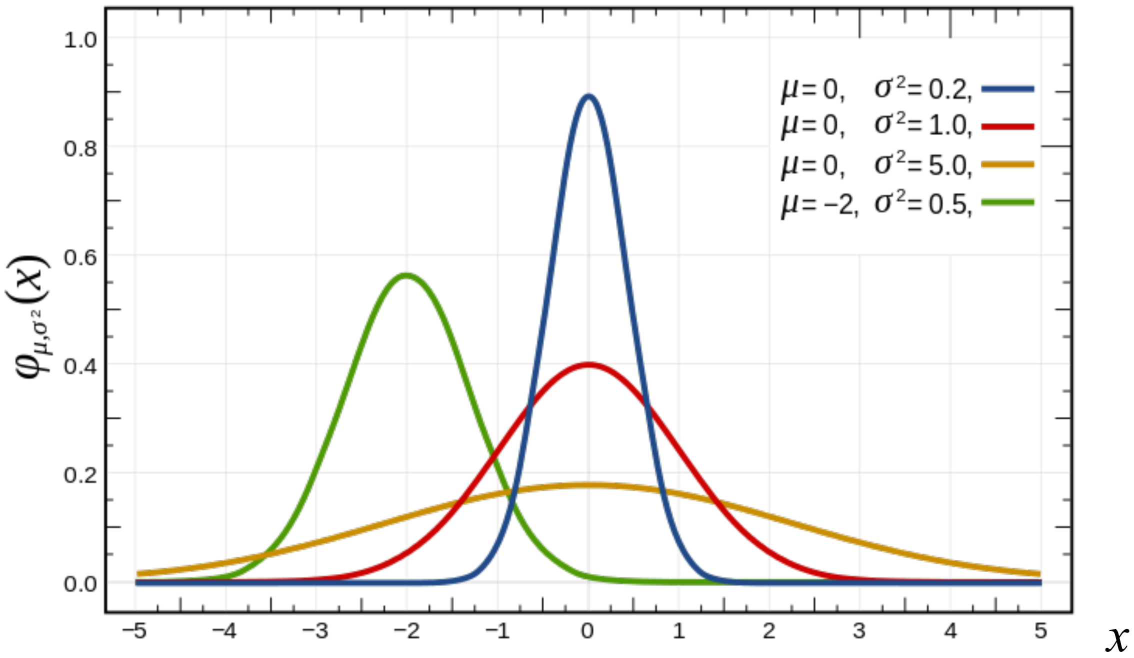

k

# Normalverteilung

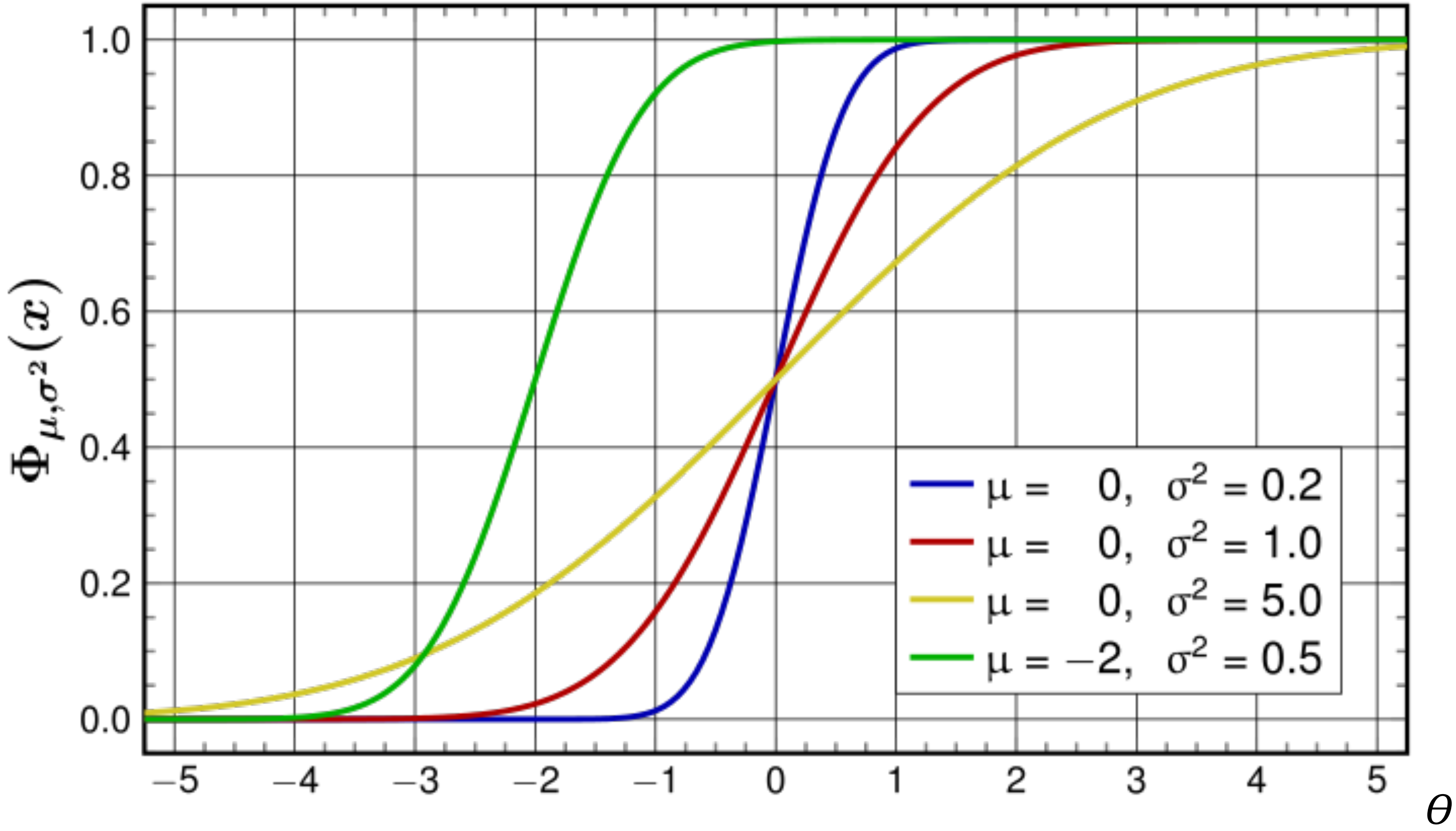
- Grundmenge: $\mathbb{R}$
- Lagemaß: Mittelwert $\mu$
- Streuungsmaß: Varianz $\sigma^2$
- Funktion für Häufigkeitsverteilung wird im kontinuierlichen Fall Dichtefunktion genannt:

Bei einer Normalverteilung sind Mittelwert und Median gleich

$$\int_{-\infty}^{\infty} \varphi_{\mu\sigma^2}(x)\, dx = 1$$



# Verteilung für relative Häufigkeit von $\varphi_{\mu,\sigma^2}(x) \leq \theta$




$\Phi_{\mu,\sigma^2}(x)$ = Fläche unter der Häufigkeitsverteilung $\varphi_{\mu,\sigma^2}(x)$ von -∞ bis $\theta$

→ sog. Verteilungsfunktion

# Von relativen Häufigkeiten zu Wahrscheinlichkeiten

- Übergang von relativen Häufigkeiten auf sog. Wahrscheinlichkeiten als Eigenschaften des Daten erzeugenden Prozesses
  - Johann Bernoulli (1667-1748) und Pierre Laplace (1749-1822)
  - Beispiel: Wahrscheinlichkeit,
    männlich zu sein, wenn man ≥400,000 Euro verdient
  - Aber: Auch bei großen Datenmengen wird die Wahrscheinlichkeit für eine Eigenschaft des die Daten generierenden Prozesses offensichtlich durch
    #günstige Fälle / #mögliche Fälle nur sehr grob geschätzt
- Betrachtung des Grenzfalles: #mögliche Fälle $\longrightarrow \infty$
  - Richard von Mises (ca. 1883-1953)
- Weitere Entwicklung ab 1930 durch Andrei Kolmogorov

# Wahrscheinlichkeits- vs. Dichtefunktion

- Wahrscheinlichkeitsfunktion
  - Wahrscheinlichkeit für jede Merkmalsausprägung
- Geht nicht bei dichter Grundmenge
  - Wahrscheinlichkeit für jeden einzelnen Wert: 0
- Daher in diesem Fall: Dichtefunktion
- Verwendung der Dichte in Verteilungsfunktion
  - Bestimmung der Wahrscheinlichkeit, dass ein gewisses Ereignis höchstens x mal auftritt
  - Verteilungsfunktionen für die Normalverteilung

$$\Phi_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \varphi_{\mu\sigma^2}(t)\, dt$$

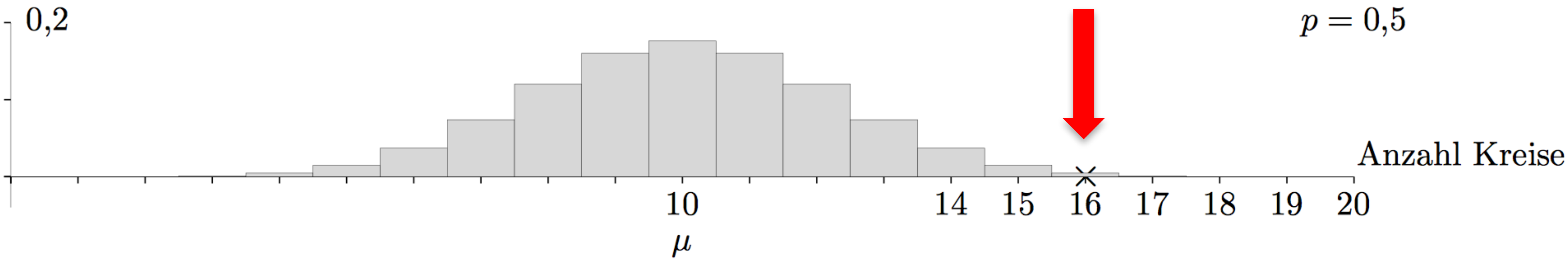
  - Geht für $x \longrightarrow \infty$ gegen 1

# Hypothesentest

- Vermutung: Küken können Körner schon erkennen und müssen die Form des Futters nicht erst lernen
- Experiment:
  - Kreise und Dreiecke je zur Hälfte zum Picken vorgegeben (sagen wir 20 Objekte insgesamt)
  - Wenn Vermutung wahr, sollte $p_{Kreis} >> 0.5$ gelten
- Hypothese $H_0$:
  - Küken unterscheiden nicht zwischen Kreisen und Dreiecken, $p_{Kreis} = 0.5$, Mittelwert des Experiments sollte 10 sein, Varianz sei 2
- Hypothese $H_1$:
  - Küken unterscheiden zwischen Kreis und Dreick, sie picken häufiger in einen Kreis

# Experiment unter Normalverteilungsannahme

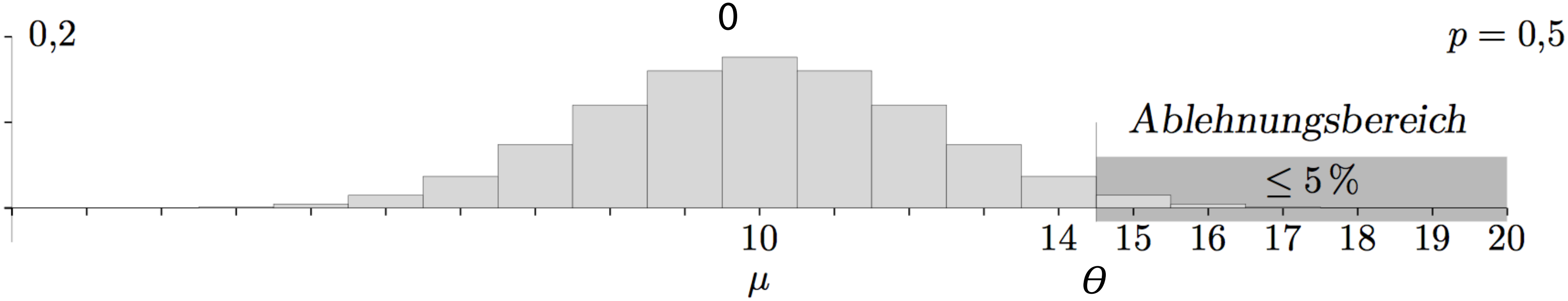
- Wenn Vermutung falsch, (also $H_0$ wahr), dann $p_{Kreis}$=0.5, Mittelwert von $\mu$=10, $\sigma^2$=2 (empirisch)




- Ausgang des Experiments:
  - Ausgang: Küken pickt im Mittel 16 mal auf Kreis
- Annahme: Wir wollen die Wahrscheinlichkeit minimieren, $H_0$ abzulehnen, obwohl sie wahr ist.

# Ablehnungsbereich

- Ziel: Wahrscheinlichkeit für Fehler (Ablehnung von $H_0$, obwohl wahr) klein halten
- Setze Irrtumswahrscheinlichkeit $\alpha$ auf 0.05
- Bestimme $\theta$ , so dass $$\int_0^{\theta} \varphi_{\mu\sigma^2}(x)\, dx = 0.95$$




- Fällt Test in Ablehnungsbereich, liegt signifikante Abweichung vor
- Wir sprechen von einem Test mit Signifikanzniveau $\alpha$

# Auswertung des Experiments: Fehleranalyse

- Das Experiment fällt in den Ablehnungsbereich für $H_0$
- Also: Annahme der Vermutung als wahr
- Anzahl der Ausgänge mit Kreis sogar 16
- Bestimme

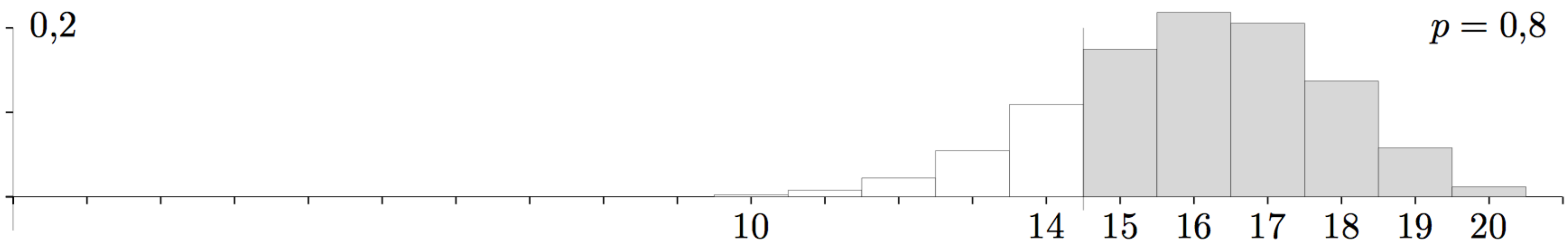
$$\int_0^{16} \varphi_{\mu\sigma^2}(x)\, dx = 0.979$$

- Irrtumswahrscheinlichkeit sogar nur 0.021
- Wir sagen $\alpha = 0.021$ (oder 2.1%)
und nennen das Fehler 1. Art

# Weitere Fragestellung

- Nehmen wir an, wir kennen die Verteilung N($\mu\sigma^2$) für den Falls, dass Küken eine angeborende Körnererkennungsbegabung haben.
- Mit welcher Wahrscheinlichkeit würde die Begabung der Küken nicht erkannt?
- Würden Küken Kreise mit Wahrscheinlichkeit p=0.8 bevorzugen, ergäbe sich:

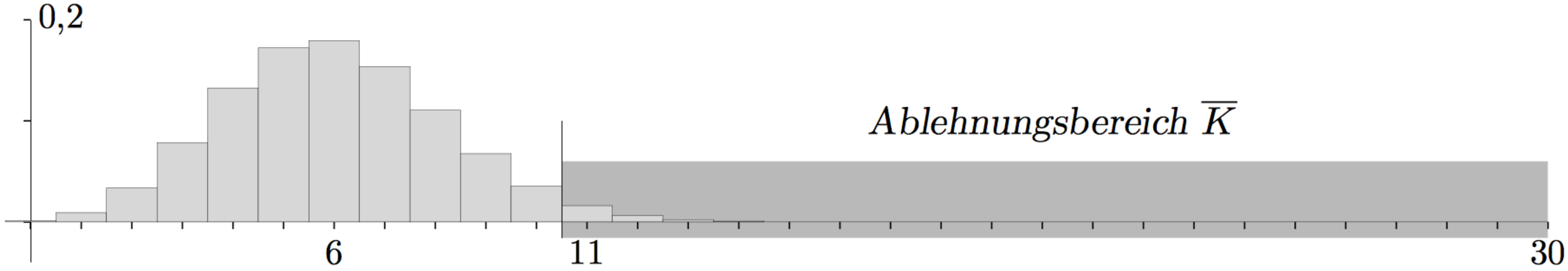
$\beta = \boldsymbol{\Phi}\mu\sigma^2(14) = 0{,}196$




- Je mehr sich p dem Wert 0.5 nähert, umso größer wird der Fehler 2. Art

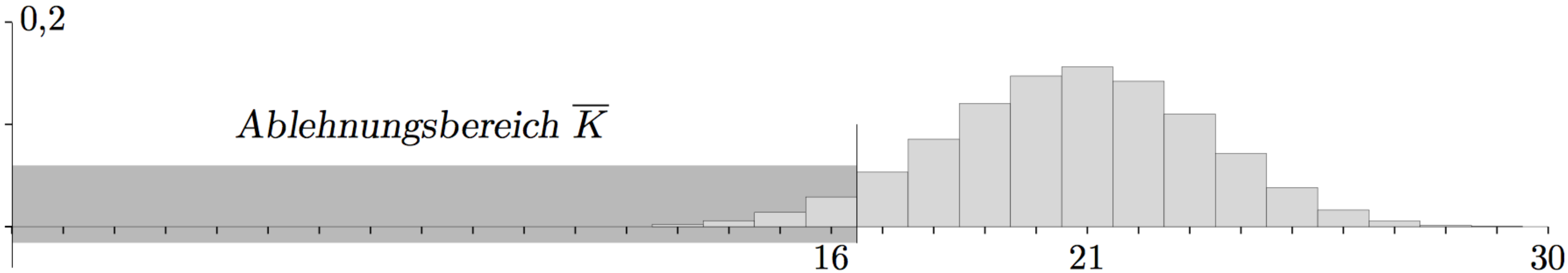
# Ablehnungsbereichs rechts

- Behauptung: Ein bestimmtes Medikament verursacht höchstens bei 20 % der Patienten Nebenwirkungen. Wir bezweifeln dies und testen die Nullhypothese auf dem 5%-Niveau. Die Stichprobenlänge sei n = 30
- $H_0$: $p \leq p_0$ $H_1 > p_0$



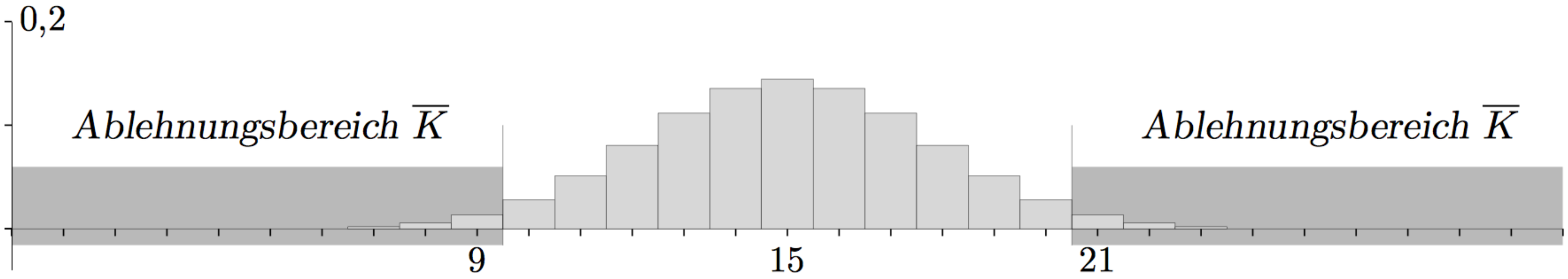
# Ablehnungsbereichs links

- Behauptung: Mindestens 70 % der gelieferten Gurken erfüllen die europäsche Krümmungsnorm. Wir vermuten das Gegenteil und testen auf dem 5%-Niveau.
- $H_0$: $p \geq p_0$ $H_1 < p_0$



# Ablehnungsbereichs beidseitig

- Bei der zufälligen Farbgebung sollen 50 % der Serienprodukte eine helle Tönung besitzen.
  Wir wollen Abweichungen aufdecken.
- $H_0$: $p = p_0$ $H_1 \neq p_0$



# Typ-1- und Typ-2-Fehler

- Typ 1: Wir lehnen $H_0$ ab, obwohl sie wahr ist
  - Wenn $\alpha$=0,05 , dann lehnen wir $H_0$ in 5% der Fälle ab
  - Wahrscheinlichkeit $\alpha$, mit der wir $H_0$ ablehnen, also einen Typ-1-Fehler zu machen

- Typ 2: Wir akzeptieren $H_0$ obwohl sie falsch ist
  - Die Wahrscheinlichkeit einen Typ-2-Fehler zu machen, ist $\beta$
  - 1-$\beta$ ist dann die Wahrscheinlichkeit $H_0$ (richtigerweise) NICHT zu akzeptieren

- Es werden aber unterschiedliche Verteilungen zugrunde gelegt: $\alpha \neq 1-\beta$

HYPOTHESIS TESTING OUTCOMES

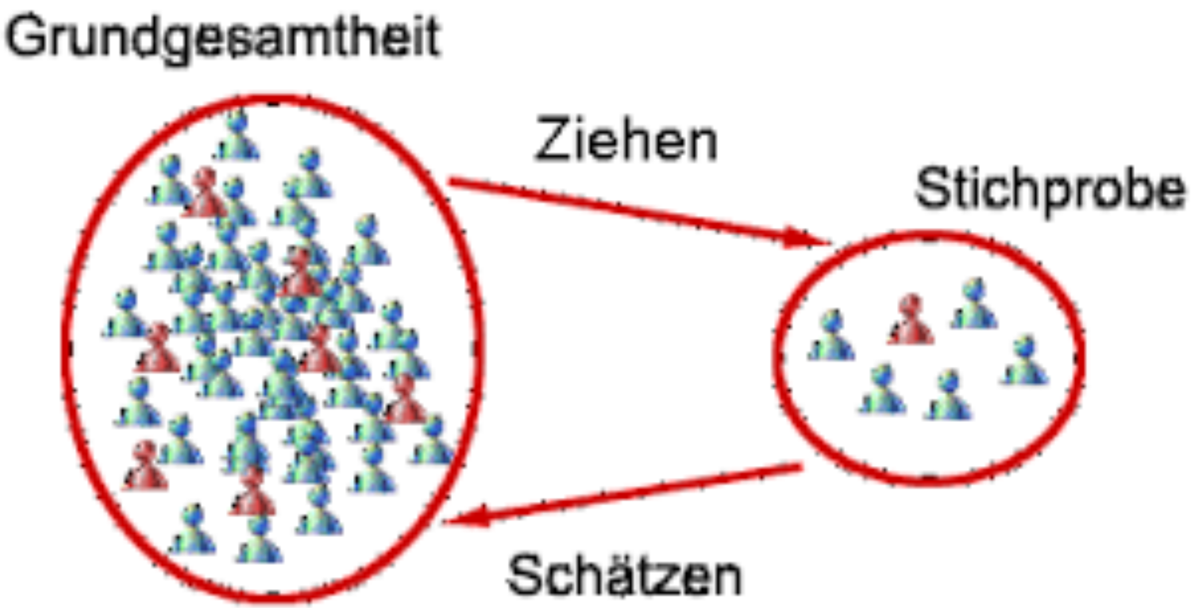
| | | Reality | |
|---|---|---|---|
| | | The Null Hypothesis Is True | The Alternative Hypothesis is True |
| Research | The Null Hypothesis Is True | Accurate $1 - \alpha$ | Type II Error $\beta$ |
| | The Alternative Hypothesis is True | Type I Error $\alpha$ | Accurate $1 - \beta$ |

# Zusammenfassung: Hypothesentest

- Um eine Hypothese zu beweisen zeigt man,
dass die Gegenhypothese wegen eines Testergebnisses äußerst unwahrscheinlich ist.
- Welche Hypothese als Nullhypothese getestet wird, hängt von Zielsetzung ab
- Wichtig: Verteilungsannahme der Nullhypothese muss gerechtfertigt sein
- Parameter der jeweils angenommenen Verteilung müssen sinnvoll bestimmt werden
- Wie groß sollte die Stichprobe sein?
- Wieviele Daten benötigen wir, um gewissen Aussagen machen zu können?

# Schätzung von Parametern

- Auswertung der Daten einer Stichprobe
- Rückschlüsse auf Eigenschaften der Grundgesamtheit
- Wir betrachten zunächst einmal die Normalverteilung
  - Aus Stichprobe Parameter bestimmen



# Experimente, Zufallsvariablen, Verteilungen

- Durchführung von Experimenten / Auswertung von Daten
  - Merkmalsausprägungen bestimmen
  - Werte von statistischen Variablen
  - Im Sinne des Ziehens aus GG: Zufallsvariable
- Beispiel: Zufallsvariable X normalverteilt
- Wir schreiben: X~N($\mu$, $\sigma^2$)

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

- Standardnormalverteilung:
  $\mu$ =0 und $\sigma$=1:

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



GG = Grundgesamtheit

# Erwartungen formal

- Erwartungswert von Zufallsvariable X:
  - Wert, den X im Mittel einnimmt
  - Diskret:

$$\mathbf{E}(X) = \sum_{i \in I} x_i p_i$$

wobei $p_i$ die relative Häufigkeit des Auftretens des Wertes $x_i$ ist

  - Kontinuierlich:

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

$f$ ist die Wahrscheinlichkeits-verteilung von X

  - Notation manchmal auch: E[X]

- E[X], wenn X~N($\mu$, $\sigma^2$) ?

# Varianz formal

- Varianz von Zufallsvariable X:
  - Wert, den X im Mittel einnimmt
  - Definition:

$$\operatorname{Var}(X) := \operatorname{E}\left((X-\mu)^2\right)$$

  - Notation manchmal auch: Var[X]

- Var[X], wenn X~N($\mu$, $\sigma^2$) ?

$$\operatorname{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$$

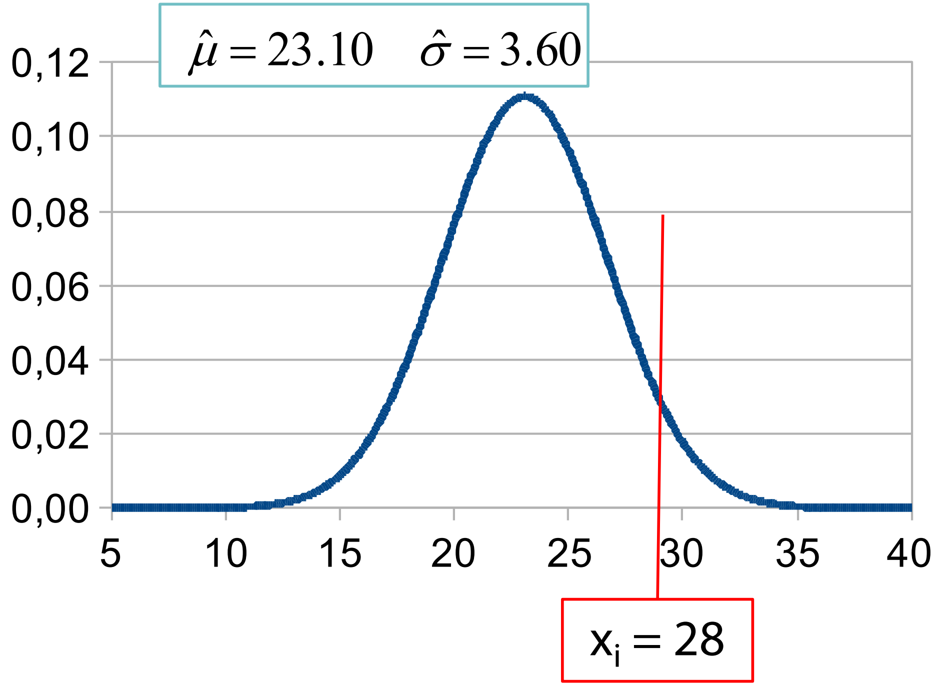
wobei

$$\mu = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x.$$

# Interpretation eines Messwertes

## Beispiel $x_i$=28

- Interpretierbar nur bei gegebener Verteilung
- $x_i$ liegt über dem arithm. Mittel
- Genauer: $x_i$ liegt mehr als eine Standardabweichung über dem arithm. Mittel
- Genauer: Wie viel Prozent der Gesamtheit geben Werte unter / über 28 an?
- Um diese Frage zu beantworten, hilft die z-Standardisierung

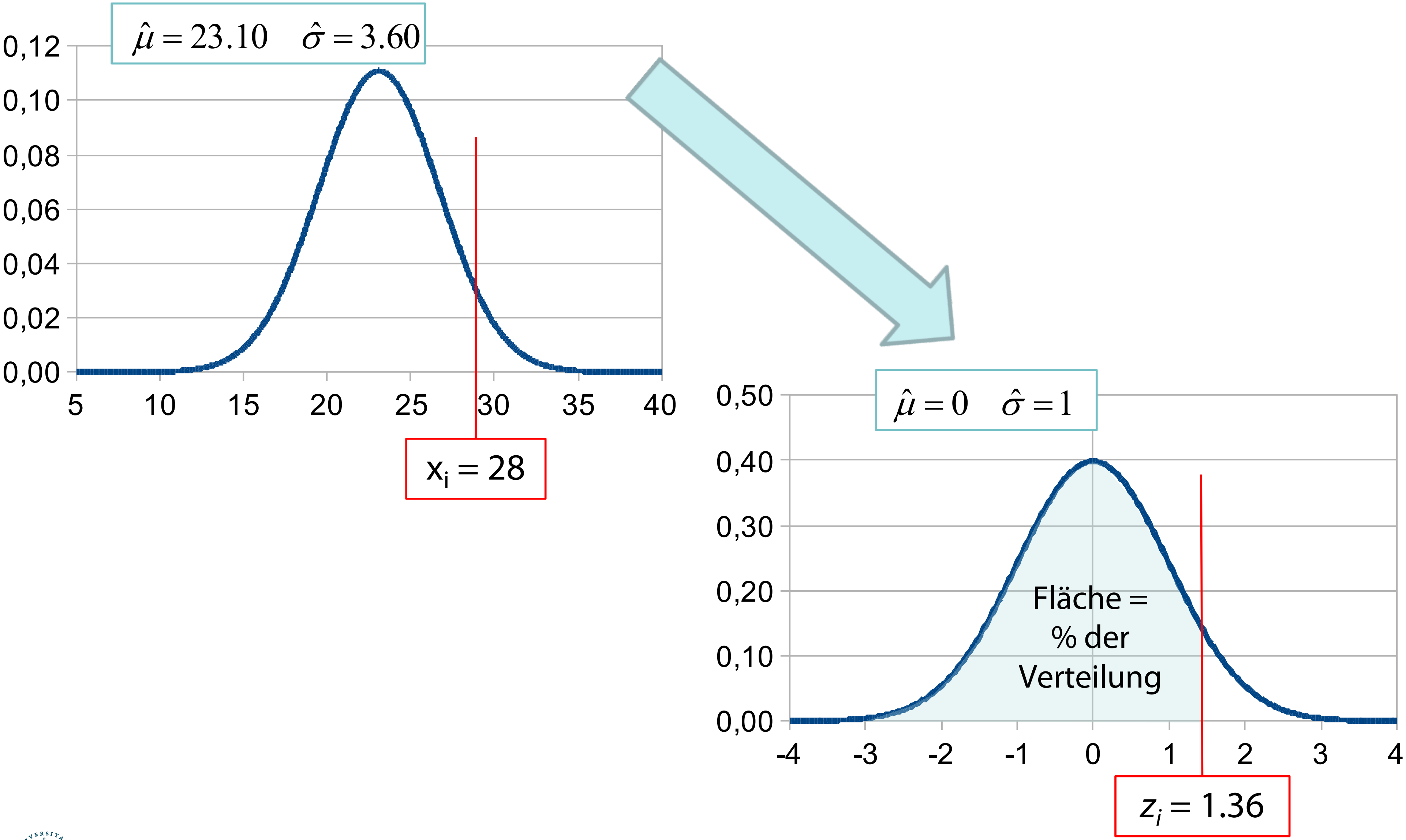


# *z*-Standardisierung

- Mit der ***z-Standardisierung*** wird eine Normalverteilung in eine Standardnormalverteilung umgewandelt.
- Die z-Standardisierung erfolgt in zwei Schritten:
  (1) Zunächst wird von jedem Messwert der *Mittelwert* subtrahiert.
  (2) Dann wird das Ergebnis durch die *Standardabweichung* geteilt.

$$z_i = \frac{x_i - \bar{x}}{\hat{\sigma}}$$

# $z$-Standardisierung



# $z$-Standardisierung

- ***z Werte*** können mit Hilfe einer $z$-Tabelle einfach interpretiert werden.
- In Tabellen zur Standardnormalverteilung ist immer angegeben, wie groß die Fläche unter der Kurve links von einem $z$-Wert ist.
- Die Fläche gibt den Anteil der Verteilung an, deren Werte kleiner oder gleich des „kritischen" z-Werts ist.
- Beispiel:
  - $x_i = 28$
  - $z_i = 1.36$
  - Fläche($z_i$) = $\boldsymbol{\Phi}(z_i) = 0.91$
  - Anteil der z-Werte $\leq 1.36 \Rightarrow 0.91$
  - 91% der Population haben z-Werte kleiner oder gleich 1.36
  - 91% der Population haben x-Werte von 28 oder darunter
  - Nur 9% der Population haben x-Werte größer als *i*.

# *z*-Standardisierung

## *Die z-Tabelle (Standardnormalverteilung)*

| z | Fläche | z | Fläche | z | Fläche | z | Fläche |
|---|---|---|---|---|---|---|---|
| -3.00 | 0.00 | -1.50 | 0.07 | 0.00 | 0.50 | 1.50 | 0.93 |
| -2.90 | 0.00 | -1.40 | 0.08 | 0.10 | 0.54 | 1.60 | 0.95 |
| -2.80 | 0.00 | -1.30 | 0.10 | 0.20 | 0.58 | 1.70 | 0.96 |
| -2.70 | 0.00 | -1.20 | 0.12 | 0.30 | 0.62 | 1.80 | 0.96 |
| -2.60 | 0.00 | -1.10 | 0.14 | 0.40 | 0.66 | 1.90 | 0.97 |
| -2.50 | 0.01 | -1.00 | 0.16 | 0.50 | 0.69 | 2.00 | 0.98 |
| -2.40 | 0.01 | -0.90 | 0.18 | 0.60 | 0.73 | 2.10 | 0.98 |
| -2.30 | 0.01 | -0.80 | 0.21 | 0.70 | 0.76 | 2.20 | 0.99 |
| -2.20 | 0.01 | -0.70 | 0.24 | 0.80 | 0.79 | 2.30 | 0.99 |
| -2.10 | 0.02 | -0.60 | 0.27 | 0.90 | 0.82 | 2.40 | 0.99 |
| -2.00 | 0.02 | -0.50 | 0.31 | 1.00 | 0.84 | 2.50 | 0.99 |
| -1.90 | 0.03 | -0.40 | 0.34 | 1.10 | 0.86 | 2.60 | 1.00 |
| -1.80 | 0.04 | -0.30 | 0.38 | 1.20 | 0.88 | 2.70 | 1.00 |
| -1.70 | 0.04 | -0.20 | 0.42 | 1.30 | 0.90 | 2.80 | 1.00 |
| -1.60 | 0.05 | -0.10 | 0.46 | 1.40 | 0.92 | 2.90 | 1.00 |

# $z$-Standardisierung

***Interpretation der Ausprägung eines normalverteilten Merkmals***

- Erhebung einer Stichprobe
  - Berechnung von Mittelwert und Standardabweichung
- Erhebung des Merkmals bei der Person i
- Berechnung des $z$-Werts
- Nachschlagen der Größe der Fläche unterhalb der $z$-Verteilung, die links von $z_i$ liegt
- Die Fläche $f(z_i)$ gibt an, wie viel Prozent der Population Werte kleiner oder gleich $z_i$ bzw. $x_i$ haben.
- 1 - $f(z_i)$ gibt an, wie viel Prozent der Population Werte größer $z_i$ bzw. $x_i$ haben.

# Prozentränge

- Ein ***Prozentrang*** (PR) gibt an, wie viel Prozent der Population Werte *kleiner oder gleich* einem kritischen Wert haben.

# Aufgabe: IQ-Wert-Analyse

Annahme: Normalverteilung
mit $\mu$=100; $\sigma$=15
Welchem Prozentrang entspricht
ein IQ-Wert von
(a) 130; (b) 92.5; (c) 85; (d) 100; (e) 115?

| z | Fläche | z | Fläche | z | Fläche | z | Fläche |
|---|---|---|---|---|---|---|---|
| -3.00 | 0.00 | -1.50 | 0.07 | 0.00 | 0.50 | 1.50 | 0.93 |
| -2.90 | 0.00 | -1.40 | 0.08 | 0.10 | 0.54 | 1.60 | 0.95 |
| -2.80 | 0.00 | -1.30 | 0.10 | 0.20 | 0.58 | 1.70 | 0.96 |
| -2.70 | 0.00 | -1.20 | 0.12 | 0.30 | 0.62 | 1.80 | 0.96 |
| -2.60 | 0.00 | -1.10 | 0.14 | 0.40 | 0.66 | 1.90 | 0.97 |
| -2.50 | 0.01 | -1.00 | 0.16 | 0.50 | 0.69 | 2.00 | 0.98 |
| -2.40 | 0.01 | -0.90 | 0.18 | 0.60 | 0.73 | 2.10 | 0.98 |
| -2.30 | 0.01 | -0.80 | 0.21 | 0.70 | 0.76 | 2.20 | 0.99 |
| -2.20 | 0.01 | -0.70 | 0.24 | 0.80 | 0.79 | 2.30 | 0.99 |
| -2.10 | 0.02 | -0.60 | 0.27 | 0.90 | 0.82 | 2.40 | 0.99 |
| -2.00 | 0.02 | -0.50 | 0.31 | 1.00 | 0.84 | 2.50 | 0.99 |
| -1.90 | 0.03 | -0.40 | 0.34 | 1.10 | 0.86 | 2.60 | 1.00 |
| -1.80 | 0.04 | -0.30 | 0.38 | 1.20 | 0.88 | 2.70 | 1.00 |
| -1.70 | 0.04 | -0.20 | 0.42 | 1.30 | 0.90 | 2.80 | 1.00 |
| -1.60 | 0.05 | -0.10 | 0.46 | 1.40 | 0.92 | 2.90 | 1.00 |

| IQ | z(IQ) | PR |
|---|---|---|
| 130 | 2.0 | 98 |
| 92.5 | -0.5 | 31 |
| 85 | -1.0 | 16 |
| 100 | 0.0 | 50 |
| 115 | 1.0 | 84 |

# Prozentränge

- Ein ***Prozentrang*** (PR) gibt an, wie viel Prozent der Population Werte *kleiner oder gleich* einem kritischen Wert haben.
- Damit entspricht der Prozentrang der Wahrscheinlichkeit des *z*-Werts

# Wahrscheinlichkeiten

- Die z-Tabelle ermöglicht es auch, ***Wahrscheinlichkeitsaussagen*** für bestimmte Intervalle zu machen.
- Wie groß ist die Wahrscheinlichkeit für einen IQ-Wert (a) von 85 bis 115; (b) von 70 bis 130; (c) von 0 bis 70; (d) von über 100

| IQ | $z(IQ_1)$ | $z(IQ_2)$ | $p(z_1)$ | $p(z_2)$ | $\Delta p$ |
|---|---|---|---|---|---|
| 85 bis 115 | -1.0 | 1.0 | .16 | .84 | .68 |
| 70 bis 130 | -2.0 | 2.0 | .02 | .98 | .96 |
| 0 bis 70 | -6.7 | -2.0 | .00 | .02 | .02 |
| > 100 | 0 | ∞ | .50 | 1.00 | .50 |

# Wahrscheinlichkeiten

Generell gilt für normalverteilte Merkmale:

- ***68.26%*** der Werte liegen im Bereich:

$$\mu - 1.0 \cdot \sigma < x_i < \mu + 1.0 \cdot \sigma$$

bzw.

$$-1.0 < z_i < +1.0$$

- ***95.44%*** der Werte liegen im Bereich:

$$\mu - 2.0 \cdot \sigma < x_i < \mu + 2.0 \cdot \sigma$$

bzw.

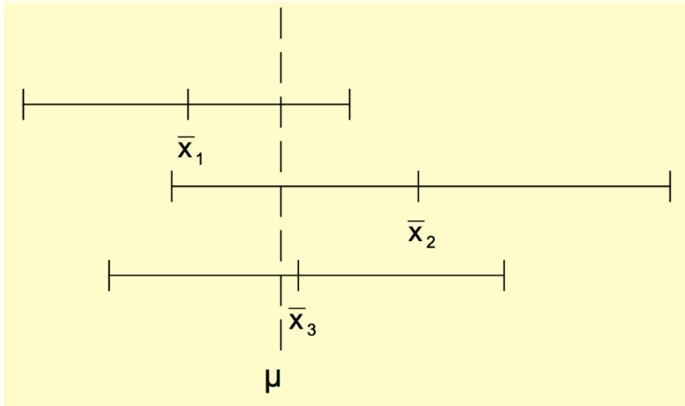
$$-2.0 < z_i < +2.0$$

# Stichprobenkennwerteverteilungen

- Wir haben verschiedene Stichprobenkennwerte kennengelernt: z.B. Mittelwert, Median, Varianz ("Punktschätzer")
- Meist interessieren nicht die Werte für die konkrete Stichprobe, sondern für die zugrundeliegenden Population
- Die Kennwerte aus einer Stichprobe werden daher als Schätzer für die entsprechenden Populationskennwerte verwendet
- Wir erwarten: Je größer eine (repräsentative) Stichprobe, desto genauer ist die Schätzung

# Stichprobenkennwerteverteilungen

- Wenn man aus der gleichen Population immer wieder Stichproben zieht, ergibt sich für jede Stichprobe ein neuer Mittelwert
- Wenn man sehr viele Stichproben erhebt, erhält man auch viele Mittelwerte
- Nun kann man die Verteilung der resultierenden Mittelwerte betrachten
- Diese Verteilung heißt Stichprobenkennwerteverteilung des Mittelwerts



# Standardfehler

- Diese „Verteilung der der Mittelwerte" ist selbst wieder normalverteilt (wenn das Merkmal normalverteilt ist)
- Der Mittelwert der Stichprobenkennwerteverteilung entspricht dem Mittelwert in der Population
- Die Streuung der Stichprobenkennwerteverteilung wird als Standardfehler (des Mittelwerts) bezeichnet.
  - Der Standardfehler gibt an, wie nah ein empirischer Stichprobenmittelwert am wahren Populationsmittelwert liegt
  - Dieser Standardfehler des Mittelwertes kann auch aus einer einzigen Stichprobe geschätzt werden:

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}_x^{\,2}}{N}} = \frac{\hat{\sigma}_x}{\sqrt{N}} \qquad \sigma^2 = \mathrm{Var} = 1/n \cdot \Sigma_{i=1}^{n}(x-x_i)^2$$

# Standardfehler

Beispiel: Unter den Mitarbeiter einer großen Firma soll die Leistungsmotivation bestimmt werden. Es werden 10 Mitarbeiter zufällig ausgewählt und getestet.

- Es ergibt sich Mittelwert von 60 bei einer geschätzten Populationsvarianz von 90.

- Wie groß ist der Standardfehler dieses Mittelwerts?

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{90}{10}} = \sqrt{9} = 3$$

- Wie groß wäre der Standardfehler bei $\sigma^2=250$ und $N=10$?

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{250}{10}} = \sqrt{25} = 5$$

- Wie groß wäre der Standardfehler bei $\sigma^2=90$ und $N=90$?

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{90}{90}} = \sqrt{1} = 1$$

# Konfidenzintervalle

- Der Standardfehler ist die Standardabweichung der Stichprobenkennwerteverteilung
- Da die Stichprobenkennwerteverteilung normalverteilt ist, kann die Wahrscheinlichkeit dafür berechnet werden, dass der Mittelwert in einem bestimmten Intervall liegt
- Mit *p*=0.68 ist der Populationsmittelwert höchstens einen Standardfehler vom Stichprobenmittelwert entfernt
- Beispiel:
  - Wenn $\bar{x} = 60$ und $\hat{\sigma}_{\bar{x}} = 3$, dann gilt mit *p*=0.68 für den Populationsmittelwert: $57 < \mu < 63$

- Notation: P( *Bedingung* ) = *p* mit $p \in [0, 1]$
- Beispiel: P( $57 < \mu < 63$ ) = 0.68

# Konfidenzintervalle

- Ein Konfidenzintervall ist ein symmetrischer Bereich um den Stichprobenmittelwert, in welchem der Populationsmittelwert mit einer bestimmten Wahrscheinlichkeit liegt.

$$\mathrm{P}\,(\bar{x} - 1.00 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 1.00 \cdot \hat{\sigma}_{\bar{x}}) = .682$$

$$\mathrm{P}\,(\bar{x} - 2.00 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 2.00 \cdot \hat{\sigma}_{\bar{x}}) = .954$$

$$\mathrm{P}\,(\bar{x} - 1.96 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 1.96 \cdot \hat{\sigma}_{\bar{x}}) = .95$$

$$\mathrm{P}\,(\bar{x} - 2.57 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 2.57 \cdot \hat{\sigma}_{\bar{x}}) = .99$$

# Standardfehler für weitere Kennwerte

| ***Kennwert*** | ***Standardfehler*** |
| --- | --- |
| Relative Häufigkeit (p) | $\hat{\sigma}_p = \sqrt{\frac{p \cdot (1-p)}{N}}$ |
| Median | $\hat{\sigma}_{Md} = \frac{1.253 \cdot \hat{\sigma}_x}{\sqrt{N}}$ |
| Arithmetisches Mittel | $\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{N}}$ |
| Standardabweichung | $\hat{\sigma}_s = \frac{\hat{\sigma}}{\sqrt{2 \cdot N}}$ |

# Standardfehler der relativen Häufigkeit

- Wie groß ist der ***Standardfehler der relativen Häufigkeit*** von Frauen unter Psychologiestudierenden ($p$=.76)?

$$\hat{\sigma}_p = \sqrt{\frac{p \cdot (1-p)}{N}}$$

$$\hat{\sigma}_p = \sqrt{\frac{.76 \cdot (1-.76)}{96}} = \sqrt{\frac{.76 \cdot .24}{96}} = \sqrt{.002} = .04$$

- Wie groß das 95% Konfidenzintervall?

$$.76 - 1.96 \cdot .04 < \pi < .76 + 1.96 \cdot .04$$

$$.68 < \pi < .84$$

# Standardfehler des Medians

- Wie groß ist der ***Standardfehler des Medians*** der Statistik-vorkenntnisse?

$$\hat{\sigma}_{Md} = \frac{1.253 \cdot \hat{\sigma}_x}{\sqrt{N}}$$

$$\hat{\sigma}_{Md} = \frac{1.253 \cdot 20}{\sqrt{98}} = \frac{25.06}{9.90} = 2.53$$

- Wie groß das 95% Konfidenzintervall?

$$20 - 1.96 \cdot 2.53 < Md < 20 + 1.96 \cdot 2.53$$

$$15.04 < Md < 24.96$$

# Standardfehler der Standardabweichung

- Wie groß ist der ***Standardfehler der Standardabweichung*** der Statistikvorkenntnisse?

$$\hat{\sigma}_s = \frac{\hat{\sigma}}{\sqrt{2 \cdot N}}$$

$$\hat{\sigma}_s = \frac{19.90}{\sqrt{2 \cdot 98}} = \frac{19.90}{14} = 1.42$$

- Wie groß das 95% Konfidenzintervall?

$$19.90 - 1.96 \cdot 1.42 < \sigma < 19.90 + 1.96 \cdot 1.42$$

$$17.12 < \sigma < 22.68$$

# Konfidenzintervall

- Die Lage und Breite des Konfidenzintervalls ist abhängig von den zufälligen Konfidenzgrenzen
- Diese hängen ab von:
  - dem Stichprobenumfang
  - der Schätzfunktion und deren Verteilung und
  - dem Konfidenzniveau
- Breite des Konfidenzintervalls ist Ausdruck für die Genauigkeit der Parameterschätzung!
  - Ein höheres Konfidenzniveau (kleineres $\alpha$) führt zu einer Verbreiterung des Konfidenzintervalls und …
  - ... ein größerer Stichprobenumfang führt zu einer Verkleinerung des Konfidenzintervalls

Sei $X \sim N(\mu, \sigma^2)$ eine normalverteilte ZG und $(X_1, ..., X_n)$ eine mathematische Stichprobe aus der GG X.

**1. Fall:** Die Varianz $\sigma^2$ der normalverteilten GG sei bekannt
Für den unbekannten Parameter $\mu$ ist eine Konfidenzschätzung anzugeben.

Als Punktschätzer für $\mu$ wählen wir das arithmetische Mittel

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{mit} \quad \overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Die Wahrscheinlichkeit, daß der Betrag des Schätzfehlers kleiner als die Schranke d ist, wird mit $(1-\alpha)$ vorgegeben, d.h.

$$\boxed{P(\left|\overline{X} - \mu\right| \leq d) = 1 - \alpha}$$

$\overline{X} - \mu$ ist der Schätzfehler

$$= P(\overline{X} - d \leq \mu \leq \overline{X} + d)$$

(Symmetrie der NV- Dichtefunktion)

ZG= Zufallsgröße = Zufallsvariable GG=Grundgesamtheit NV = Normalverteilung

Zur Bestimmung der Größe d standardisieren wir die ZG $\overline{X}$ :

$$Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1) \quad \text{, da} \quad \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\Rightarrow \quad P\left(\left|\frac{\overline{X}-\mu}{\sigma}\cdot\sqrt{n}\right| \le \frac{d}{\sigma}\cdot\sqrt{n}\right) = 1-\alpha \qquad z_{1-\frac{\alpha}{2}} = \frac{d}{\sigma}\cdot\sqrt{n}$$

$$= P\left(|Z| \le z_{1-\frac{\alpha}{2}}\right) \qquad \rightarrow \quad \boxed{d = \frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}}$$

$$= P\left(z_{\frac{\alpha}{2}} \le Z \le z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \qquad z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$$

$$\Rightarrow \boxed{= P\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}\right) = 1-\alpha}$$

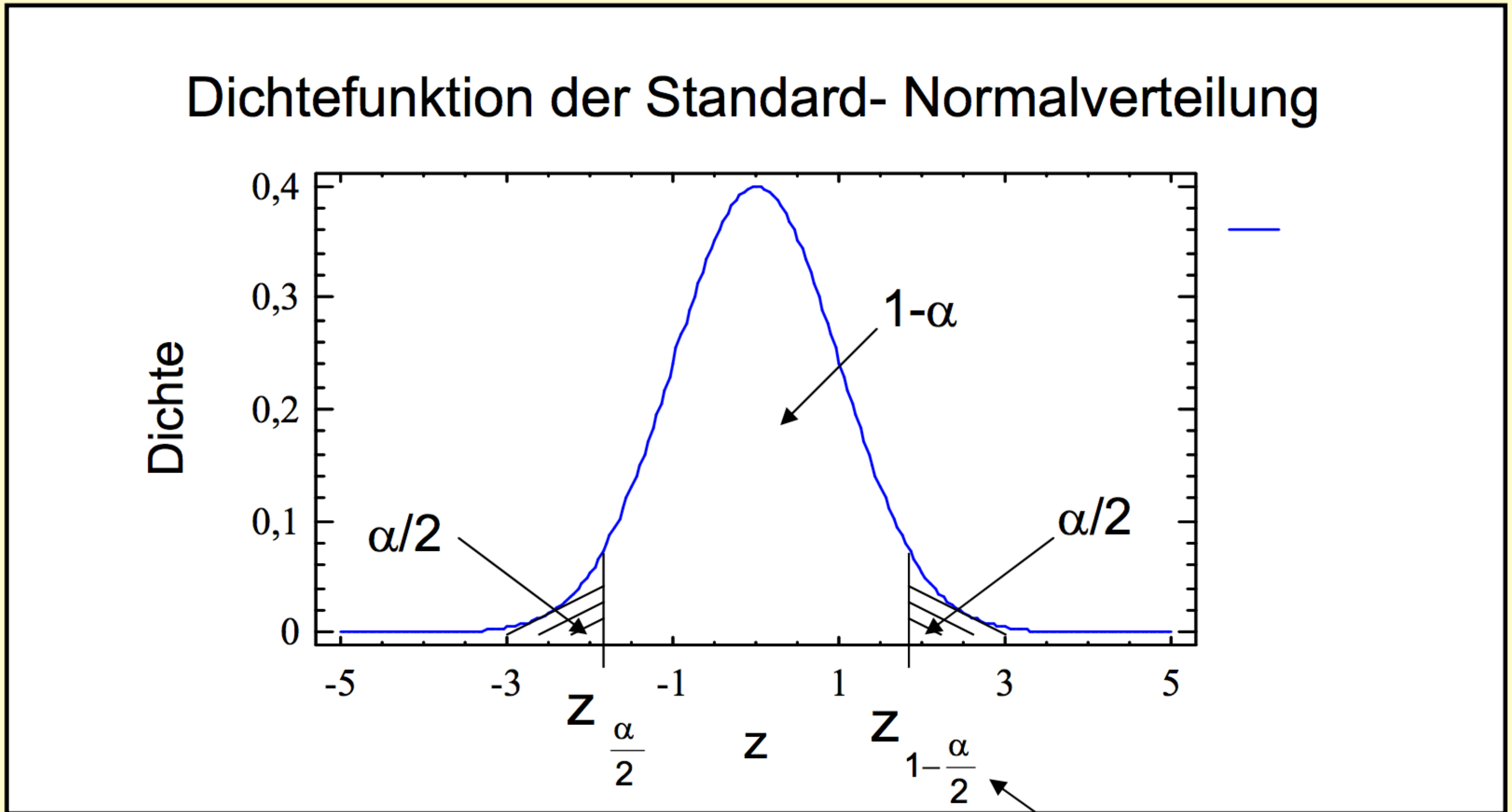


U. Römisch

# Das Konfidenzintervall

$$\left[\overline{X}-\frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}};\overline{X}+\frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}\right]$$

überdeckt also den wahren Parameter μ mit der Wahrscheinlichkeit (1-α).




$= -z_{1-\frac{\alpha}{2}}$

$1-\frac{\alpha}{2}$ -Quantil der Standard- NV

Jede konkrete Stichprobe liefert uns eine Realisierung der ZG $\overline{X}$ und damit ein realisiertes Konfidenzintervall:

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} ; \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right]$$

Einige typische $z_{1-\frac{\alpha}{2}}$ - Werte (2-seitige Fragestellung) und $z_{1-\alpha}$- Werte (1- seitige Fragestellung) enthält die Tabelle:

| $1-\alpha$ | $\alpha$ | $z_{1-\frac{\alpha}{2}}$ | $z_{1-\alpha}$ |
|---|---|---|---|
| 0,95 | 0,05 | 1,96 | 1,64 |
| 0,99 | 0,01 | 2,58 | 2,33 |
| 0,999 | 0,001 | 3,29 | 3,09 |

$$\Phi\left(z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

$$\Phi\left(z_{1-\alpha}\right) = 1 - \alpha$$

**Bem.:** Die Breite dieses Konfidenzintervalls für den Erwartungswert $\mu$ beträgt 2d und ist von $\alpha$, n, $\sigma$ und der Verteilung der zugehörigen Schätzfunktion abhängig.

$$2\,d = 2\frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$

Je größer $\alpha$ (bei festem n!)

$\longrightarrow$ desto kleiner das Konfidenzintervall

Je größer n $\longrightarrow$ desto kleiner das Konfidenzintervall

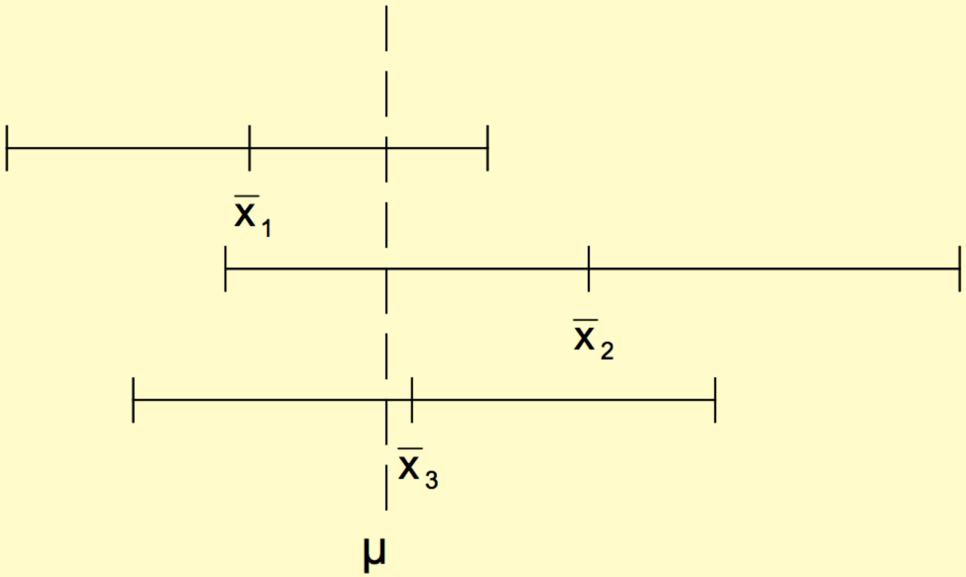
Die Breite des Konfidenzintervalls ist hier ein Maß für die Genauigkeit der Schätzung von μ und die Irrtumswahrscheinlichkeit $\alpha$ ein Maß für das Risiko.

$\Rightarrow$ **Planung des Stichprobenumfangs:**

geg.: halbe Breite des Konf.intervalls d,
Varianz $\sigma^2$,
Konfidenzniveau (1-$\alpha$)

$$\rightarrow \boxed{n = \frac{\sigma^2}{d^2} \cdot z^2_{1-\frac{\alpha}{2}}}$$

**Bem.:** Die Lage des konkreten Konfidenzintervalls wird durch die konkrete Stichprobe bestimmt.




**Bem.:** Wählen wir ein Konfidenzniveau $(1-\alpha) = 0{,}95$.
So heißt das:
In 95% aller Fälle enthält der Vertrauensbereich den unbekannten Parameter der GG und in 5% der Fälle nicht.
D.h.: Behauptet man $k$ mal, der unbekannte Parameter liege im Vertrauensbereich, so hat man im Mittel $\alpha \cdot k$ Fehlschlüsse zu erwarten.

# Schätzung der Varianz

- Ähnliche Überlegungen
- Auch hierfür Herleitung der erforderlichen Stichprobengröße möglich

**2. Fall:** Die Varianz $\sigma^2$ der normalverteilten GG sei nicht bekannt und muß geschätzt werden.
Für den unbekannten Parameter μ ist eine Konfidenzschätzung anzugeben.

Wir wählen als Punktschätzer:

für den Erwartungswert μ: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ und

für die Varianz $\sigma^2$ $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Wir wählen nun zur Bestimmung der Größe d aus

$$P(|\overline{X} - \mu| < d) = 1 - \alpha$$

die Stichprobenfunktion $t = \frac{\overline{X} - \mu}{S} \cdot \sqrt{n} \sim t(m)$ aus,

die einer t-Verteilung mit m = n – 1 Freiheitsgraden genügt.

Aufgrund der Symmetrie der Dichtefunktion der t-Verteilung gilt wieder:

$$P\left(|t| < t_{1-\frac{\alpha}{2};m}\right) = 1 - \alpha$$

und man erhält analog zum 1. Fall durch Umformungen das Konfidnzintervall für den gesuchten Parameter μ:

$$\Rightarrow \quad \boxed{P\left(\overline{X} - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};m} < \mu < \overline{X} + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};m}\right) = (1-\alpha)}$$

$$\boxed{d = \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};m}}$$

Veranschaulichung analog wie beim 1. Fall!

Das Konfidenzintervall

$$\left[\overline{X} - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};m} ; \overline{X} + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};m}\right]$$

überdeckt also den wahren Parameter μ mit der Wahrscheinlichkeit (1-$\alpha$).

Jede konkrete Stichprobe liefert uns wieder ein realisiertes Konfidenzintervall:

$$\left[\overline{x} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};m} ; \overline{x} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};m}\right]$$

**Bem.:** - Für gleichen Stichprobenumfang n und gleiche Irrtumswahrscheinlichkeit $\alpha$ ist das realisierte Konfidenzintervall im 2. Fall größer als das im 1. Fall.

- Für hinreichend große m = n –1 (m>500)verwendet man als Näherung für die Quantile $t_{1-\frac{\alpha}{2};m} = t_{1-\frac{\alpha}{2};\infty} \approx z_{1-\frac{\alpha}{2}}$

- Die Quantile der t-Verteilung können aus Tabellen abgelesen werden (in Abhängigkeit von $\alpha$ und m)!

## $\Rightarrow$ Planung des Stichprobenumfangs:

<u>geg.:</u> halbe Breite des Konf.intervalls d,
geschätzte Varianz $s^2$,
Konfidenzniveau $(1-\alpha)$

**Iteration:**

$$\boxed{n_{i+1} = \frac{s^2}{d^2} \cdot t^2_{1-\frac{\alpha}{2};n_i-1}}$$

da m = n -1

1. Schritt: $n_0 = \infty \quad \rightarrow$

$$\boxed{n_1 = \frac{s^2}{d^2} \cdot t^2_{1-\frac{\alpha}{2};\infty}}$$

$z^2_{1-\frac{\alpha}{2}}$

$n_1$ wird auf die nächste ganze Z. aufgerundet!

2. Schritt: $n_1 \quad \rightarrow$

$$\boxed{n_2 = \frac{s^2}{d^2} \cdot t^2_{1-\frac{\alpha}{2};n_1-1}}$$

…

*Die Iteration bricht ab, wenn die aufgerundeten Stichproben-umfänge zweier aufeinander folgender Schritte gleich sind, bzw. nach Erreichen einer vorgegebenen Anzahl von Schritten.*

**Bsp.:** Der mittlere fettfreie Trockenmassegehalt (TM) von Kondensmilch ist auf 0,5 % genau zu schätzen. Aus früheren Bestimmungen des TM- Gehaltes lag ein Schätzwert der Standardabweichung von $s = 0{,}996$ % vor. Als Konfidenzniveau wurde $(1-\alpha) = 0{,}95$ festgelegt.

1. Berechnung des mindestens notwendigen Stichprobenumfangs:

   Iteration: 1. Schritt: $n_0 = \infty \rightarrow n_1 = 15{,}24 \sim 16$

   ...

   4. Schritt: $n_3 = 18 \rightarrow \mathbf{n_4} = 17{,}66 \sim \mathbf{18}$

2. Ermittlung der Punktschätzung $\bar{x}$ aus einer Stichprobe vom Umfang n = 18.

3. Das 95%- ige geschätzte Konfidenzintervall lautet dann:

$$[\bar{x} - d; \bar{x} + d] = [20{,}2 ; 21{,}2] \quad [\%]$$

# Begriff der Erwartungstreue

- Ein Schätzer heißt erwartungstreu, wenn sein Erwartungswert gleich dem wahren Wert des zu schätzenden Parameters ist

**Beispiel Stichprobenmittel** [Wikipedia]

Zur Schätzung des Erwartungswertes $\gamma = \mu$ der Grundgesamtheit wird üblicherweise das Stichprobenmittel

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

verwendet. Werden alle Stichprobenvariablen $X_i$ zufällig aus der Grundgesamtheit gezogen, so haben alle den Erwartungswert $\operatorname{E}(X_i) = \mu$. Damit berechnet sich der Erwartungswert des Stichprobenmittels zu

$$\operatorname{E}(\overline{X}_n) = \operatorname{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\operatorname{E}(X_i) = \frac{1}{n}\cdot n\cdot\mu = \mu\,.$$

Das Stichprobenmittel ist also ein erwartungstreuer Schätzer des unbekannten Verteilungsparameters $\mu$.

# Korrigierte Stichprobenvarianz

Die korrigierte Stichprobenvarianz der Stichprobenwerte $x_1, \ldots, x_n$ ist definiert als

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

Dabei ist $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ der empirische Mittelwert, also das arithmetische Mittel der Stichprobe.

Der Nenner $n-1$ in der korrigierten Stichprobenvarianz erklärt sich folgendermaßen: Sind $x_1, \ldots, x_n$ die Ausprägungen der unabhängig identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit Varianz $\sigma^2$ und ist der Mittelwert $\mu$ der Grundgesamtheit bekannt, so ist

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

eine erwartungstreue Schätzfunktion für die Varianz $\sigma^2$ der Grundgesamtheit[2] und damit auch

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2,$$

eine erwartungstreue Schätzung für die Varianz. Es gilt namentlich:

$$\mathrm{E}(S_0^2) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\big((X_i - \mu)^2\big) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n}n\sigma^2 = \sigma^2.$$

Üblicherweise kennt man aber den Mittelwert $\mu$ der Grundgesamtheit nicht und schätzt ihn daher durch den Stichprobenmittelwert

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

Setzt man diesen Schätzwert unbekümmert in obige Formel ein, so erhält man für die Varianz $\sigma^2$ der Grundgesamtheit die Schätzung

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Um zu entscheiden, ob dieser Schätzer erwartungstreu ist, betrachtet man ihn als Ausprägung der Schätzfunktion

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 .$$

und berechnet wie folgt den Erwartungswert:

$$
\begin{aligned}
\mathrm{E}(S_1^2) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big((X_i-\overline{X})^2\Big) = \frac{1}{n}\,\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu+\mu-\overline{X})^2\right)\\
&= \frac{1}{n}\,\mathrm{E}\left(\sum_{i=1}^{n}\Big((X_i-\mu)^2-2(X_i-\mu)(\overline{X}-\mu)+(\overline{X}-\mu)^2\Big)\right)\\
&= \frac{1}{n}\,\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2-2\sum_{i=1}^{n}(X_i-\mu)(\overline{X}-\mu)+n(\overline{X}-\mu)^2\right)\\
&= \frac{1}{n}\,\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2-2n(\overline{X}-\mu)(\overline{X}-\mu)+n(\overline{X}-\mu)^2\right)\\
&= \frac{1}{n}\,\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}-\mu)^2\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\mathrm{E}\big((X_i-\mu)^2\big)-n\,\mathrm{E}\Big((\overline{X}-\mu)^2\Big)\right)\\
&= \frac{1}{n}\Big(n\,\mathrm{Var}(X)-n\,\mathrm{Var}(\overline{X})\Big)\\
&= \mathrm{Var}(X)-\mathrm{Var}(\overline{X}) = \sigma^2-\frac{\sigma^2}{n} = \frac{n-1}{n}\,\sigma^2,
\end{aligned}
$$

wobei beim vorletzten Gleichheitszeichen die Definition der Varianz und die Formel zur Berechnung des Standardfehlers aus der Varianz der Grundgesamtheit und dem Stichprobenumfang benutzt wurden. Daraus ergibt sich, dass die Schätzfunktion $S_1^2$ nicht erwartungstreu ist, und dass man einen erwartungstreuen Schätzer für die Varianz erhält, wenn man $s_1^2$ mit dem Faktor $\frac{n}{n-1}$ multipliziert. So gelangt man zur korrigierten Stichprobenvarianz

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

Für diese gilt nun unabhängig von der genauen Verteilung der $x_i$

$$\mathrm{E}(S^2) = \sigma^2.$$

Der Erwartungswert der korrigierten Stichprobenvarianz ist also gleich der Varianz der Grundgesamtheit. Die korrigierte Stichprobenvarianz ist somit eine erwartungstreue Schätzung für die Varianz.[2]

# Verteilungsannahmen gerechtfertigt?

- Wie kann man prüfen, ob die Verteilungsannahme gerechtfertigt ist?
- Eigenschaften von Schätzern bzgl. Verteilungen betrachten und mit konkreter Stichprobe prüfen
- Sind die Eigenschaften offensichlich in einer Stichprobe verletzt, liegt andere als die angenommene Verteilung vor
- $\chi^2$-Test
  - Funktionsanpassungstest (z.B. Verteilungstest)
  - Unabhängigkeitstest
  - Homogenitätstest

F. R. Helmert. In: Zeitschrift fuer Math. und Physik 21, S. 102-219, **1876**
Karl Pearson: On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it Can Reasonably Be Supposed to have Arisen from Random Sampling.
In: Philosophical Magazine 5, Band 50, S. 157-175 , 1900
Zitiert nach L. Schmetterer: Mathematische Statistik. Springer, Wien 1966, S. 93

# Zusammenfassung

- Konzept der Stichprobe
- Relative Häufigkeiten
- Verteilungen
- Beschreibungsmaße
  - Mittelwert, Varianz (Streuung), …
- Wahrscheinlichkeiten
- Hypothesentest, Signifikanzniveau
- Zufallsvariable,
- Normalverteilung, Standardnormalverteilung
- Standardfehler
- Konfidenzintervall, Stichprobenumfang
- Erwartungstreue