Einführung in die Statistik

Wirtschaftsingenieurwesen, Bachelor

Sommer 2024

Prof. Dr. Christoph Becker

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Inhaltsverzeichnis

1. Deskriptive Statistik

1.1. Grundbegriffe

Ziele in der Statistik:

- Statistics is the science of using data to learn about the world around us.
- Testen von Theorien; Untersuche Charakteristiken von Verteilungen in Populationen.
- Prognose von Wirtschaftswachstum, Studienerfolg, Verbrechensraten...
- Kausale Beziehungen vs. Prognosen.

Grundbegriffe: Teilgebiete der Statistik

- **Deskriptive Statistik**: Übersichtliche Darstellung und Beschreibung von Daten (genauer: einer Stichprobe) mit Kennzahlen und Grafiken.
  - Mittelwerte, empirische Standardabweichung,...
  - Histogramm, Dichten, ...
- **Induktive Statistik**: Aussagen über **Grundgesamtheit**, Überprüfung von Hypothesen.
  - Schätzen von Parametern einer Verteilung (Punktschätzer).
  - Hypothesentest.
  - Konfidenzintervalle.

Grundbegriffe

Definition 1.1 (Grundgesamtheit). *Die Grundgesamtheit (Population) stellt eine Menge von räumlich, zeitlich und sachlich eindeutig definierten Objekten dar.*

Beispiele:

a) Alle aktuell im Studiengang Sozialwissenschaften an der HDA eingeschriebene Studierende,

b) Alle männlichen Personen in D., deren Blutdruck mindestens einmal täglich 180 mm Hg übersteigt.

Die Grundgesamtheit kann aufgrund der Größe oft nicht vollständig untersucht werden, denn zu aufwendig, zu teuer, oder unethisch. Stattdessen Analyse einer **Stichprobe**.

Deskriptive Statistik

Definition 1.2 (Stichprobe). *Eine Teilmenge von Objekten der Grundgesamtheit nennen wir Stichprobe. Die Anzahl der Objekte in der Stichprobe heißt Stichprobenumfang.*
Synonyme: Daten, Beobachtungen.

Stichprobe statt Grundgesamtheit zu untersuchen hat Vorteile:

- Billiger, schneller, Erhebung oft sorgfältiger möglich (Datenqualität).
- Manchmal unvermeidlich (Lebensdauertests).
- Ethischer (klinische Studien).
- Stichprobe wird oft zufällig aus Grundgesamtheit ausgewählt (eigene Wissenschaft!).

Ziel der Statistik: Stichprobe (Daten) $\Rightarrow$ Aussage über Grundgesamtheit.

Deskriptive Statistik

Einige Methoden zur Durchführung von Stichprobenuntersuchungen

- Zufallsstichprobe.
- Systematische Auswahl: Objektives Kriterium, z.B. jeder 100. Artikel.
- Quotenverfahren/Repräsentative Stichprobe. Die Stichprobe soll die Werte gewisser Merkmale mit den gleichen Quoten bzw. Anteilen, wie in der Grundgesamtheit enthalten.

Als Beispiel dafür, wie wichtig die Art der Erhebung einer Stichprobe sein kann siehe z.B. https://www.diw.de/documents/publikationen/73/diw_01.c.57345.de/diw_sp0019.pdf

1.2. Merkmalstypen

Definition 1.3 (Merkmal). *Beobachtbare bzw. messbare Eigenschaften von Objekten der Population nennen wir* ***Merkmale****.*

Bsp.: Körpergröße, Geschlecht, Raucher/Nichtraucher.

Merkmalsausprägungen *sind mögliche Werte, die ein Merkmal annehmen kann.*

Merkmalstypen

- Quantitativ vs. qualitativ
  - Quantitatives M.: Ausprägungen werden durch Messen oder Zählen erfasst (Bspl: Blutdruck, Kinderzahl)
  - Qualitatives M.: Alle anderen (z.B. Blutgruppe, Geschlecht)
- Diskret vs. stetig
  - Diskretes M.: Die Menge der Ausprägungen ist endlich (Bspl: Geschlecht, Kinderzahl)
  - Stetiges M.: Jeder Wert aus einem Intervall kann grundsätzlich angenommen werden (z.B. Blutdruck, Körpergröße, Prozentuale Veränderung des Preises einer Immobilie)

Skalenniveaus

- Metrisch: Ausprägung wird durch Zahlen erfasst und Differenzen (Abstände) sind sinnvoll interpretierbar (Bspl: Längen, Gewichte, Preise)
- Ordinal: Werte sind lediglich geordnet nach Größer-Kleiner-Relation (Bspl: Schulnoten)
- Nominal: Werte haben den Charakter von Namen oder Kategorien. Anordnung nicht möglich bzw. sinnvoll (Bspl: Haarfarbe, Städtenamen).

Hierarchie (von grob nach fein): nominal $\Rightarrow$ ordinal $\Rightarrow$ metrisch.
Statistische Methoden sollten Skalenniveau berücksichtigen.

Häufigkeiten

Definition 1.4 (Urliste, absolute und relative Häufigkeiten).

- *Urliste (Rohdaten): Ungeordnet aufgeschriebene Liste* $x_1, \dots, x_n$ *der Stichprobenwerte*
- *verschiedene Merkmalsausprägungen* $\{a_1, \dots, a_m\}$.
- $h_k = \#\{i | x_i = a_k\}$*: Anzahl der Beobachtungen, die die Ausprägung* a_k *besitzen.* h_k *heißt absolute Häufigkeit der Ausprägung* a_k.
- $f_k = \frac{h_k}{n}$ *heißt relative Häufigkeit der Ausprägung* a_k*, die Gesamtheit* $f_1, \dots, f_m$ *heißt Häufigkeitsverteilung.*

Es gilt $h_1 + \cdots + h_m = n$ bzw. $f_1 + \cdots + f_m = 1$.

1.3. Lagemaße

Oft wollen wir Daten nur anhand von wenigen (zwei) Kenngrößen beschreiben.

a) Lage: Wo liegen die Daten? Wo ist das 'Zentrum' der Daten?

b) Streuung: Wie weit weichen die Daten voneinander bzw. vom Zentrum ab?

1.3.1. Arithmetisches Mittel

Definition 1.5 (Arithmetisches Mittel (Mittelwert)). *Sei $x_1, \dots, x_n$ Stichprobe vom Umfang n (quantitatives Merkmal). Dann ist das* arithmetische Mittel *definiert durch*

$$\overline{x} = \overline{x}_n = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Alternative Bezeichnungen: Stichprobenmittel, Mittelwert.

- $\overline{x}$ nur bei quantitativen Merkmalen sinnvoll.
- $\overline{x}$ stark beeinflussbar durch Ausreißer.

1.3.2. Median

Definition 1.6 (Median). *Sei $x_1, \dots, x_n$ Stichprobe vom Umfang n (quantitatives oder ordinales Merkmal). Sei $x_{(1)} \leq \dots \leq x_{(n)}$ die zugehörige* geordnete Stichprobe. *Dann ist der* Median *der Stichprobe definiert durch*

$$med(x_1, \dots, x_n) = x_{(k+1)}, \qquad \text{falls } n = 2k+1 \text{ ungerade ist.}$$

Für $n = 2k$ gerade wird definiert

$$med(x_1, \dots, x_n) = \frac{1}{2}(x_{(k)} + x_{(k+1)}).$$

Alternative Schreibweise: $x_{0.5}$, $med(x)$, ...

Median

- Der Median beschreibt das Zentrum der Daten: Mindestens 50% der Daten sind $\geq$ und mindestens 50% der Daten sind $\leq$ als der Median ($\longrightarrow$ allgemeine Quantilsdefinition)
- Für n gerade ist Definition uneinheitlich. Jeder Wert zwischen $x_{(k)}$ und $x_{(k+1)}$ wäre möglich...
- Der Median wird weniger durch Ausreißer beeinflusst als $\overline{x}$.

1.4. Streuungsmaße

Definition 1.7 (Varianz und Standardabweichung). *Sei $x_1, \dots, x_n$ eine Stichprobe mit arithmetischem Mittel $\bar{x}$, dann heißt*

$$s^2 = s_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2$$

Varianz *der Stichprobe und die Wurzel*

$$s = s_n = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

Standardabweichung *der Stichprobe.*

Alternative Bezeichnungen: empirische Varianz, Stichprobenvarianz,...

Streuungsmaße

- Beide Maße, Varianz und Standardabweichung, messen die Streuung (quadratische Abweichung) der Stichprobe um den Mittelwert $\bar{x}$.
- Alternative Formel:

$$s^2 = \frac{1}{n-1}\left(\left(\sum_{k=1}^{n} x_k^2\right) - n\bar{x}^2\right)$$

- Manchmal auch mit $\frac{1}{n}$ statt $\frac{1}{n-1}$...

1.5. Grafische Darstellungen von Häufigkeiten

1.5.1. Das Stabdiagramm

- Geeignet für ordinale und nominale Merkmale, bei stetigen Merkmalen ist Histogramm geeigneter.
- *Stabdiagramm*:
  - Auf der x-Achse (Abszisse) des Stabdiagramms werden die beobachteten Merkmalsausprägungen abgetragen.
  - Auf der y-Achse (Ordinate) werden die absoluten oder relativen Häufigkeiten der jeweiligen Ausprägung in Form eines Stabes abgetragen. Die Länge des Stabes gibt die Häufigkeit der zugehörigen Ausprägung wider.
- *Säulen- bzw. Balkendiagramm*: Analog Stabdiagramm mit Säulen (Balken) statt Stäben. Die Säulen sollten die gleiche Breite besitzen
- Wird bei zu vielen Kategorien unübersichtlich

1.5.2. Die empirische Verteilungsfunktion

Annahme: Stichprobe $x_1, \ldots, x_n$ (metrisches Merkmal)

- Beispiel-Stichprobe:
  $110, 123, 130, 135, 138, 145, 100, 150, 155, 190$
- Anteil Beobachtungen ≤ 120?
- Anteil Beobachtungen ≤ 90?
- Anteil Beobachtungen ≤ 180?

Für beliebiges $x \in \mathbb{R}$ ergibt sich eine Funktion:

Definition 1.8 (Empirische Verteilungsfunktion).

$$\widehat{F}_n(x) = \textit{Anteil der Beobachtungen} \leq x = \frac{\#\{i | x_i \leq x\}}{n}$$

$\widehat{F}_n$ *heißt* empirische Verteilungsfunktion *(englisch: empirical (cumulative) distribution function).*

Die empirische Verteilungsfunktion

- $\widehat{F}_n(x)$ beschreibt die kumulierten relativen Häufigkeiten
- $\widehat{F}_n(x)$ hängt sowohl von x als auch von der Stichprobe ab!
- Die empirische Verteilungsfunktion enthält die gesamte Information über die Häufigkeitsverteilung der Daten.

Die empirische Verteilungsfunktion

1. Schritt: (Ordne Beobachtungen der Größe nach an und) trage Beobachtungen auf der x-Achse ein.

2. Schritt: $\widehat{F}_n$ ist eine monoton steigende Treppenfunktion mit folgenden Eigenschaften

   - Für $x < \min(x_1, \dots, x_n)$ gilt $\widehat{F}_n = 0$
   - An den Stellen $x_1, \dots, x_n$ springt $\widehat{F}_n$ um $\frac{1}{n}$ nach oben
   - Für $x \geq \max(x_1, \dots, x_n)$ gilt $\widehat{F}_n = 1$

Die empirische Verteilungsfunktion

Wichtige mathematische Eigenschaft: Es gilt der Satz von Glivenko-Cantelli:

$$\lim_{n\to\infty} \widehat{F}_n(x) = F(x),$$

wobei F die wahre aber unbekannte Verteilungsfunktion ist, die die Stichprobe generiert hat.

1.5.3. Das Histogramm

Betrachte Stichprobe $x_1, \ldots, x_n$ (metrisches Merkmal).

- Zerlege den Wertebereich der Beobachtungen in mehrere Intervalle
- Stelle relative oder absolute Häufigkeiten der entsprechenden Klassen grafisch dar.

Bemerkungen:

- Die Anzahl der Klassen kann frei gewählt werden.
- Die Klassenbreiten sollten gleich sein.

Das Histogramm

Konstruktion anhand der Beispiel-Daten

1. Teile Messbereich $[100, 200]$ in z.B. $k = 5$ gleich breite Klassen ein.
2. Ermittle absolute/relative Häufigkeiten pro Klasse
3. Grafische Darstellung: Stelle Häufigkeiten durch Rechtecke dar. Wenn alle Klassenbreiten identisch sind, kann die Höhe des Rechtecks proportional zu h_i bzw. f_i gewählt werden.

1.5.4. Quantile

Bekannt: Median = 50%-Quantil. Allgemeiner:

Definition 1.9 (p-Quantil). *Sei $x_1, \dots, x_n$ Stichprobe vom Umfang n (quantitatives oder ordinales Merkmal) und es sei $0 < p < 1$. Sei $x_{(1)} \leq \dots \leq x_{(n)}$ die zugehörige* geordnete Stichprobe. *Dann ist das* p-Quantil *der Stichprobe definiert durch*

$$\tilde{x}_p = x_{([n \cdot p]+1)}, \qquad \text{falls } n \cdot p \notin \mathbb{N}.$$

Für $n \cdot p \in \mathbb{N}$ wird definiert

$$\tilde{x}_p = \frac{1}{2}(x_{[np]} + x_{[np]+1}).$$

Dabei bezeichnet $[k]$ den ganzzahligen Anteil einer Zahl.

Quantile: Bemerkungen

- Es gilt $\text{med}(x) = \tilde{x}_{0.5}$
- Unteres bzw. oberes *Quartil*: $\tilde{x}_{0.25}$ bzw. $\tilde{x}_{0.75}$
- Anschaulich:
  - (Mindestens der) Anteil p der Daten $\leq \tilde{x}_p$
  - (Mindestens der) Anteil $1 - p$ der Daten $\geq \tilde{x}_p$

Beispiel-Daten: Berechnen Sie $\tilde{x}_{0.25}$ und $\tilde{x}_{0.75}$.

1.5.5. Box-Plots

Definition 1.10 (Fünf-Punkte-Zusammenfassung).
Die Fünf-Punkte-Zusammenfassung *(five point summary) eines Datensatzes besteht aus*

$$x_{(1)} \text{ (Minimum)}, \tilde{x}_{0.25}, \tilde{x}_{0.5}, \tilde{x}_{0.75}, x_{(n)} \text{ (Maximum)}$$

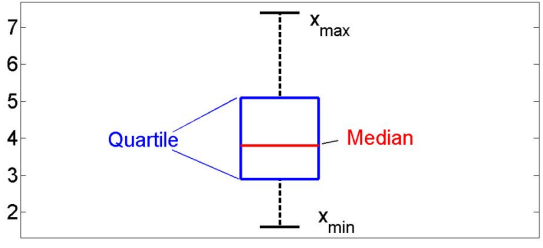


Abbildung 1: Visualisierung durch Box-Plot.

Box-Plots

Definition 1.11 (Box-Plot, Box-Whisker-Plot).

1. *Die Box (Schachtel) wird durch* $\tilde{x}_{0.25}$ *und* $\tilde{x}_{0.75}$ *begrenzt*
2. *Der Median wird durch einen Punkt oder einen Strich in der Box markiert.*
3. *Zwei Linien ('whiskers') außerhalb der Box gehen bis zu* $x_{(1)}$ *und* $x_{(n)}$

Box-Plots

Die Box repräsentiert den mittleren Teil der Daten. Die Länge der Box wird auch als

$$IQR = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

interquartile range bezeichnet.

Ein Box-Plot ermöglicht:

- Vergleich mehrerer Datensätze/Verteilungen
- Aussagen über Symmetrie
- Aussagen über Ausreisser.

Box-Plot für Beispieldaten.

1.6. How to make a bad graph

Positiv: Die Möglichkeiten zur Grafikerstellung in R sind hervorragend.

https://www.r-graph-gallery.com/

Aber oft ist weniger mehr:

- K. Broman: 'How to display data badly'

  http://ocw.jhsph.edu/courses/BiostatisticsLectureSeries05/PDFs/Broman.pdf

- E.R. Tufte, 'The visual display of quantitative information'

Beispiele: Grafiken

Abbildung 2: Quelle: Bosbach and Korff [2011]

Beispiele: Grafiken

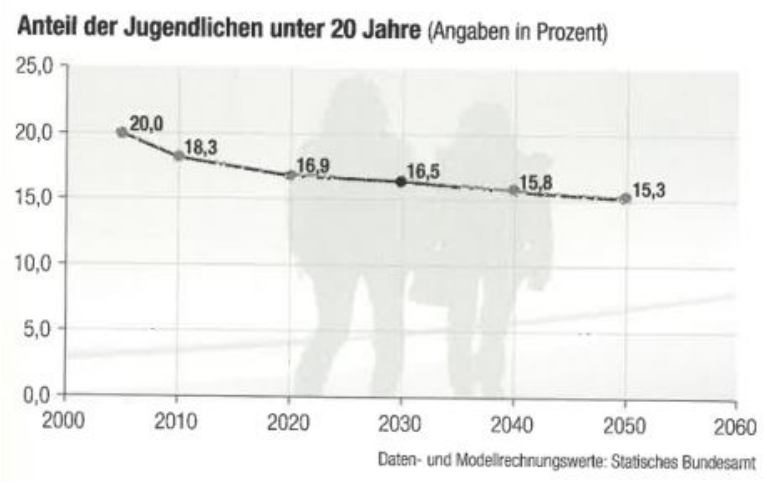


Abbildung 3: Quelle: Bosbach and Korff [2011]

Beispiele: Grafiken

Abbildung 4: Quelle: Bosbach and Korff [2011]

Beispiele: Grafiken

Abbildung 5: Quelle: Bosbach and Korff [2011]

Beispiele: Grafiken

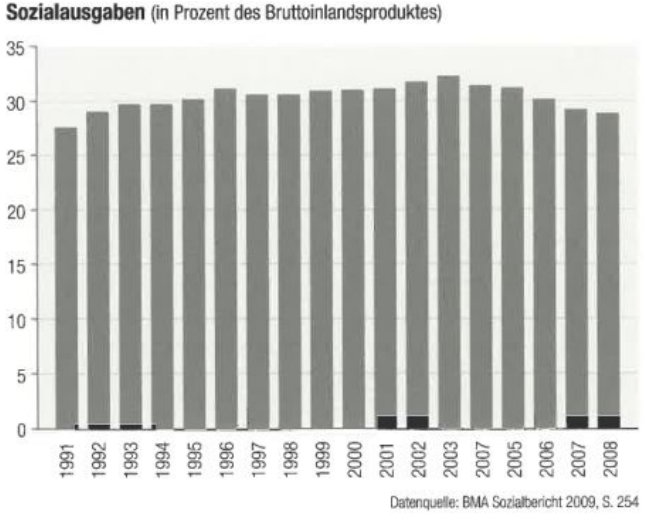


Abbildung 6: Quelle: Bosbach and Korff [2011]

2. Korrelation und Regression

2.1. Beschreibung von Zusammenhängen

Oft vermuten wir Zusammenhänge zwischen mehreren, gemeinsam beobachteten Merkmalen, z.B.

- Körpergröße $\longrightarrow$ Körpergewicht
- Risikofaktoren $\longrightarrow$ Krankheit
- Alter & BMI $\longrightarrow$ Blutdruck
- Körpergröße von Männern $\longrightarrow$ Körpergröße ihrer Söhne

Beobachtungen haben jetzt die Form

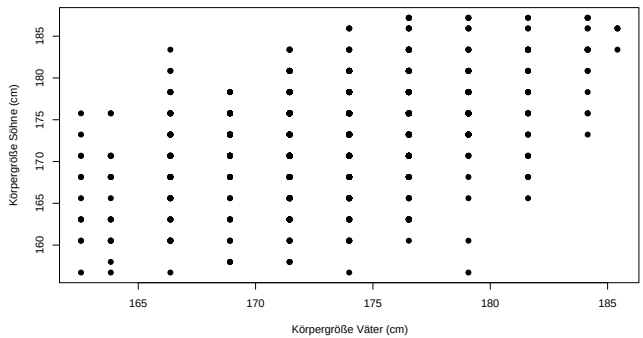
$$(x_1, y_1), \ldots, (x_n, y_n)$$

$\Rightarrow$ in jeder Beobachtung werden 2 Merkmale zusammengefasst.

Beschreibung von Zusammenhängen

Grafische Darstellung: *Punktwolke=Streudiagramm=Scatterplot*
Beispiel: Galton (1889) untersuchte die Größe von $n = 928$ Vätern und ihren Söhnen

- x_i: Größe des i-ten Vaters
- y_i: Größe des Sohnes vom i-ten Vater



Beschreibung von Zusammenhängen

Typische Fragen:

- Existiert ein **Zusammenhang** zwischen den x- und y-Werten?
- Wie **stark** ist ggf. der Zusammenhang?
- Welcher **Art** ist der Zusammenhang?

Weitergehend:

- **Vorhersage/Prognose**: z.B. Vater hat Körpergröße $x = 180$. Was erwarten wir für y, die Körpergröße seines Sohnes?
- **Kausalität**: z.B. Gibt es einen ursächlichen Zusammenhang zwischen BMI und Blutdruck?

2.2. Der Pearsonsche Korrelationskoeffizient

Gegeben: Beobachtungen $(x_1, y_1), \ldots, (x_n, y_n)$ von zwei Merkmalen X und Y.

Definition 2.1 (Kovarianz). *Die* Kovarianz s_{xy} *von* X *und* Y *ist gegeben durch*

$$\begin{aligned} s_{xy} &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y} \right). \end{aligned} \tag{1}$$

s_{xy} ist ein Maß für das gemeinsame „miteinander-variieren“ von X und Y.

s_{xy} kann beliebige positive und negative Werte annehmen.

Der Pearsonsche Korrelationskoeffizient

Gleichsinniger Zusammenhang wenn $s_{xy} > 0$. In der Summe (1) dominieren die Terme bei denen gleichzeitig entweder gilt

- $x_i > \overline{x}$ und $y_i > \overline{y}$, oder
- $x_i < \overline{x}$ und $y_i < \overline{y}$.

Gegensinniger Zusammenhang wenn $s_{xy} < 0$. In der Summe (1) dominieren die Terme bei denen gleichzeitig entweder gilt

- $x_i > \overline{x}$ und $y_i < \overline{y}$, oder
- $x_i < \overline{x}$ und $y_i > \overline{y}$.

Falls $X = Y \Rightarrow s_{xy} = s_x^2$ (Kovarianz=Varianz).

Der Pearsonsche Korrelationskoeffizient

Statt Kovarianz wird typischerweise (immer?) folgende normierte Version verwendet.

Definition 2.2 (Korrelationskoeffizient). *Der (Pearsonsche)* Korrelationskoeffizient *ist definiert durch*

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

Dabei ist s_{xy} die Kovarianz, s_x, s_y die Standardabweichungen von x und y.

r ist ein **Maß für den linearen Zusammenhang** zwischen X und Y.

- r nimmt Werte zwischen -1 und $+1$ an, $r_{xy} = r_{yx}$.
- r besitzt das gleiche Vorzeichen wie s_{xy}:

  i) $r > 0 \Rightarrow$ gleichsinniger Zusammenhang,

  ii) $r < 0 \Rightarrow$ gegensinniger Zusammenhang.

Der Pearsonsche Korrelationskoeffizient: Beispiele für Streudiagramme

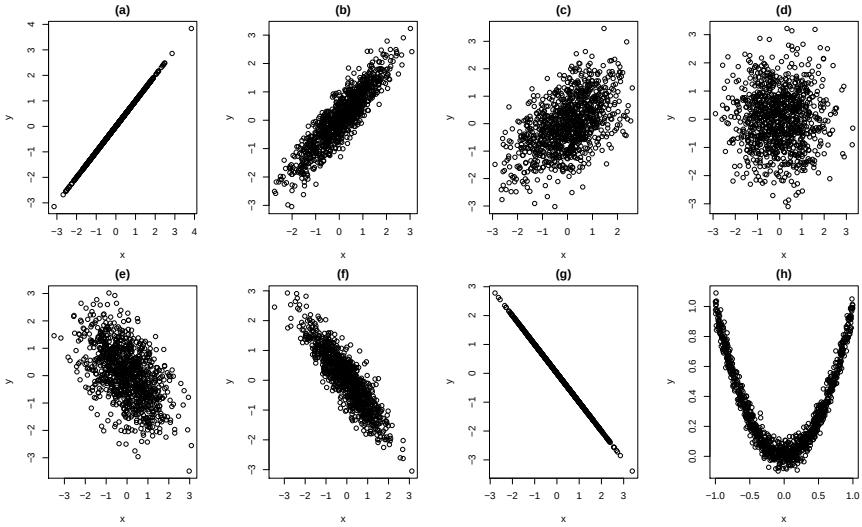


Abbildung 7: Streudiagramme mit jeweils 1000 Punkten und Korrelationskoeffizienten 1, 0.9, 0.5, 0, -0.5, -0.9, 0

Der Pearsonsche Korrelationskoeffizient

- Je näher $|r|$ bei 0, desto schwächer Zusammenhang und desto weiter streut Punktwolke.
- Je näher $|r|$ bei 1, desto stärker Zusammenhang und desto näher liegen die Punkte auf einer Geraden $y = ax + b$.

Es gibt noch weitere Typen von Korrelationskoeffizienten.

Korrelation ist ein Standardmaß zur Quantifizierung der Abhängigkeit zwischen zwei Größen.

Korrelation $\neq$ Kausalität. Gemeinsamkeitskorrelation/Scheinkorrelation ($\Rightarrow$ Scheinkausalität). https://www.tylervigen.com/spurious-correlations

2.3. Lineare Regression

- Beim Korrelationskoeffizienten werden Merkmale X und Y gleichberechtigt behandelt.
- Oft wollen wir jedoch eine der beiden Variablen - Y (*abhängige Variable, Regressand, Response*) - abhängig von der anderen Variablen X (*unabhängige Variable, Regressor, erklärende Variable*) modellieren.
- Bspl: x: Größe eine Mannes, y: Größe seines Sohnes

Ziel bei Regressionsanalysen ist die Schätzung des bedingten Erwartungswertes $\mathbb{E}(y|x)$, d.h. der erwartete (mittlere) Wert von y bei gegebenem x.

Lineare Regression

Bei der *linearen Regression* wird angenommen, dass $\mathbb{E}(y|x)$ eine lineare Funktion in x ist, d.h. es gibt $a, b \in \mathbb{R}$ mit

$$\mathbb{E}(y|x) = a + bx.$$

a und b heißen *Regressionsparameter*. Die Aussage ist äquivalent zu

$$y = a + bx + \epsilon,$$

wobei ϵ den Fehlerterm (oder das Residuum) bezeichnet.

Problem: Wahre Parameter a und b sind unbekannt und müssen geschätzt werden.

Lineare Regression

Definition 2.3. *Es seien Beobachtungen* $(x_1, y_1), \ldots, (x_n, y_n)$ *gegeben. Dann ist die* Regressionsgerade *gegeben durch*

$$y = \widehat{a} + \widehat{b} \cdot x.$$

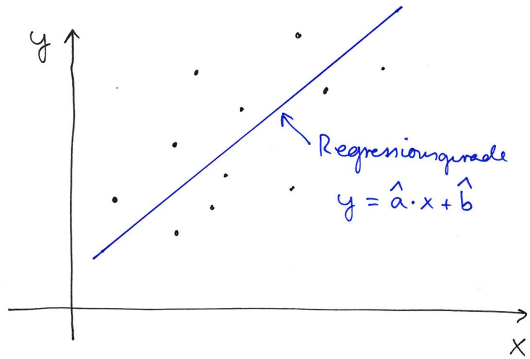
Dabei sind

$$\widehat{a} = \overline{y} - \widehat{b} \cdot \overline{x},$$
$$\widehat{b} = \frac{s_{xy}}{s_x^2}.$$

die sog. kleinste-Quadrate Schätzer *der Parameter.*

Lineare Regression

- a und b bzw. $\widehat{a}$ und $\widehat{b}$ werden als *(Regressions-)Koeffizienten* bezeichnet.
- Manchmal wird Regression nicht auf Originalwerten, sondern auf transformierten, z.B. $\ln(y)$, durchgeführt.
- Geometrische Interpretation



Lineare Regression

Die Regressionsgerade entsteht, indem der (quadratische) Abstand zu den Daten in der Punktwolke minimiert wird, d.h. $\widehat{a}$ und $\widehat{b}$ werden so gewählt, daß

$$\sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

minimal wird ($\Rightarrow$ „kleinste-Quadrate Schätzer“, „Ordinary Least Squares“ OLS)

Erstmals angewendet wird die Methode der kleinsten Quadrate (MkQ) bei der Fehlerausgleichsrechnung astronomischer Daten durch CARL FRIEDRICH GAUSS (1777 1855) im Jahr 1805. Ab 1807 war GAUSS Direktor der Sternwarte in Göttingen und Professor für Astronomie an der dortigen Universität. Die Methode der kleinsten Quadrate fand bald auch in anderen Bereichen der Wissenschaft Anwendung (Biologie, Psychologie, Soziologie, Wirtschaftswissenschaften. . .).

Lineare Regression

Ob die Regressionsgerade steigt oder fällt, hängt nur von der Kovarianz ab:

- $s_{xy} > 0 \Leftrightarrow$ Regressionsgerade steigend
- $s_{xy} < 0 \Leftrightarrow$ Regressionsgerade fallend
- $s_{xy} = 0 \Leftrightarrow$ Regressionsgerade horizontal

Lineare Regression

Beispiel (Forts.): Software-Output (in R) von Galtons Daten

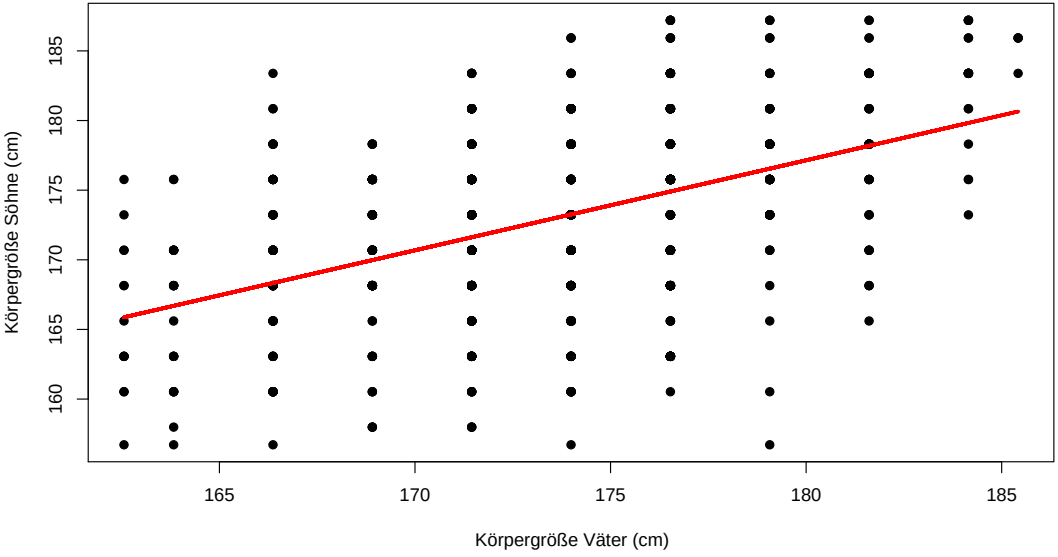
```
> lm.galton <- lm(child ~ parent, data=galton)
> lm.galton

Call:
lm(formula = child ~ parent, data = galton)

Coefficients:
(Intercept)       parent
    60.8115       0.6463
```

Erhalten $\widehat{a} = 60.8115$ und $\widehat{b} = 0.6463 \Rightarrow$ (geschätzte) Regressionsgerade:

$$y = 60.8115 + 0.6463 \cdot x$$

Galton.r

Lineare Regression



Lineare Regression

Für einen frei gewählten x-Wert kann der zugehörige y-Wert prognostiziert (geschätzt) werden. Wir schreiben dann $\widehat{y} = 60.811 + 0.6463 \cdot x$. Ergebnisse:

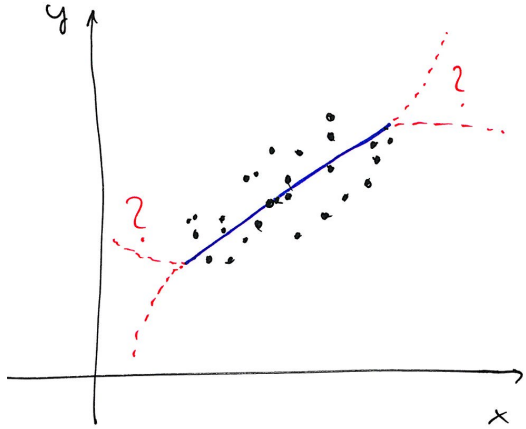
x	150	165	180
$\widehat{y}$	158	168	177

- Söhne extrem großer (bzw. kleiner) Väter sind im Mittel immer noch sehr groß (bzw. klein). ABER:
- Im Mittel sind Söhne extrem großer (bzw. kleiner) Väter kleiner (bzw. größer) als Ihre Väter!

Galton: „Regression towards mediocrity“ (Rückkehr zum Mittelmaß)
Ähnlich: Intelligenzquotient.

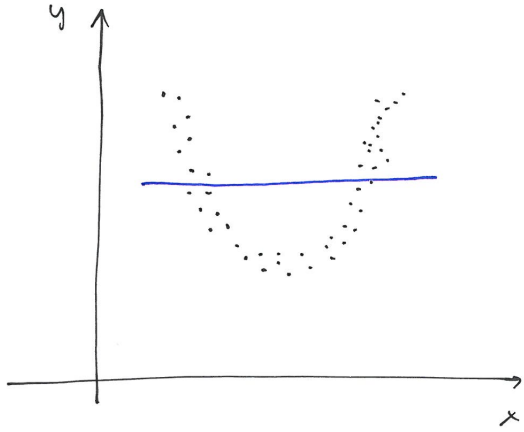
2.4. Interpretation von Korrelationen und Regression

Eine Extrapolation der Regressionsgeraden über den Bereich der Punktwolke hinaus ist i.A. keine gute Idee.



Interpretation von Korrelationen und Regression

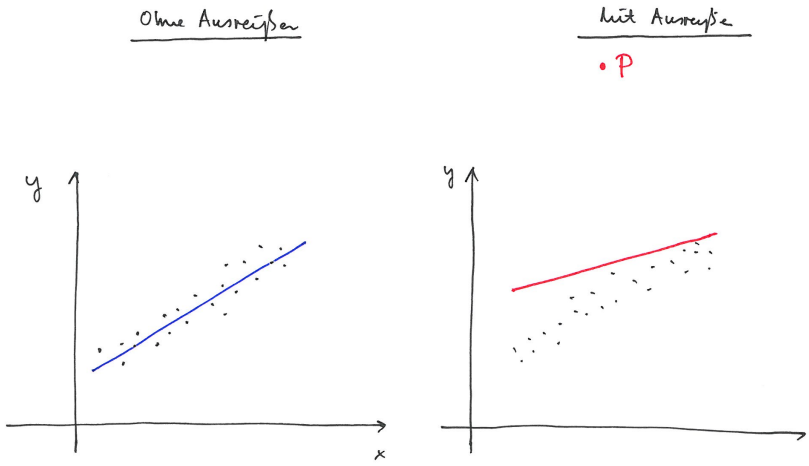
Fehlschluss: Aus $r_{xy} = 0 \not\Rightarrow$ Es gibt keinen Zusammenhang zwischen x und y. Erklä-



rung: Hier $r_{xy} = 0$. Zusammenhang nicht linear, sondern quadratisch.

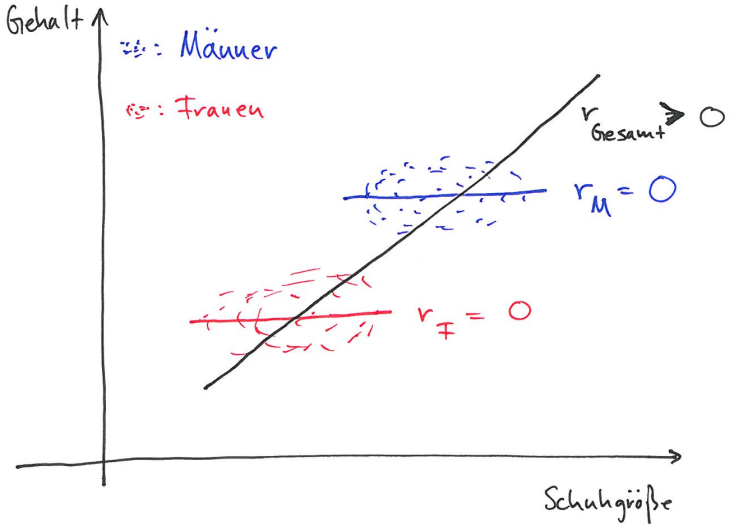
Interpretation von Korrelationen und Regression

Ausreißer: Einzelne extreme Werte können großen Einfluß auf die Regressionsgerade bzw. Korrelation haben.



Interpretation von Korrelationen und Regression

Inhomogenitätskorrelation: Gehalt und Schuhgröße korrellieren?



Interpretation von Korrelationen und Regression

Interpretation einer starken Korrelation ($|r_{xy}|$ groß):

- Höchstens Hinweis auf kausale Beziehung, kein Beweis.
- Keine Aussage darüber, ob
  - x von y oder umgekehrt beeinflußt wird
  - x und y wechselseitig aufeinander einwirken
  - ob x und y durch ein gemeinsames z beeinflußt werden.

Interpretation von Korrelationen und Regression

Abschließende Bemerkungen: Für zwei Merkmale X und Y gilt

- Der Korrelationskoeffizient beschreibt nur den Grad und die Richtung von *linearer* Abhängigkeit zwischen X und Y.
- Vollständige Beschreibung der Abhängigkeit i.A. nicht durch Korrelation möglich ($\Rightarrow$ Copulas)
- Sind X und Y unabhängig $\Rightarrow$ Korrelation $= 0$.
- Korrelation $= 0 \not\Rightarrow X$ und Y unabhängig! (Gilt nur in ganz speziellen Situationen, z.B. wenn X und Y multivariat normalverteilt sind).
- Wenn die Verteilungen von X und Y bekannt sind, kann der Korrelationskoeffizient i.A. nicht alle beliebigen Werte in $[-1, 1]$ annehmen.

2.5. Lineare Regression: Erweiterungen

Multiple lineare Regression mit einer abhängigen Variablen y aber mehreren erklärenden Variablen $x_1, \ldots, x_n$,

$$y = a + b_1 x_1 + \cdots + b_n x_n + \epsilon.$$

Schätzung der Parameter $a, b_1, \ldots, b_n$ mittels Kleinste-Quadrate-Schätzer.

Beispiel in R zur Illustration von

- Standard-Multipler Regression.
- Nichtlineare erklärende Variablen.
- Standardisierung von Variablen.

3. Wahrscheinlichkeitsrechnung

3.1. Zum Begriff der Wahrscheinlichkeit

Kennzeichen empirischer Forschung ist *Unsicherheit*. Ziel ist die Quantifizierung von Zufall / Unsicherheit $\Rightarrow$ 'Wahrscheinlichkeit'.

Das Konzept von Wahrscheinlichkeiten ist allgegenwärtig:

- Medizin: 'Mit einer Wahrscheinlichkeit von 0.05% ist Therapie A besser als Therapie B.'
- Wetter: Wahrscheinlichkeit für Regen morgen.
- Versicherung: Wahrscheinlichkeit, dass ein 35jähriger Kunde mindestens 65 Jahre alt wird.
- Business: Wie viele Mitarbeiter braucht ein Call-Center damit mit Wahrscheinlichkeit $< 1\%$ kein Kunde warten muss?

3.2. Rechnen mit Wahrscheinlichkeiten

Einfaches Beispiel: Faire Münze wird einmal geworfen. $P(\text{„Kopf“}) = P(\text{“Zahl“}) = \frac{1}{2}$. (Intuitive) Interpretation von P: „Langfristige Häufigkeit“: Wenn Experiment sehr oft wiederholt wird, ergibt sich als relative Häufigkeit vom Ereignis „Kopf“$\approx \frac{1}{2}$

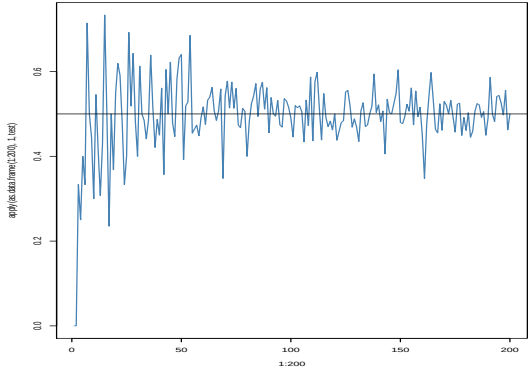


Abbildung 8: Relative Häufigkeit von 'Kopf' beim Münzwurf. `test=function(x)mean(sample(0:1, x, replace=TRUE, prob = c(0.5,0.5)))); plot(1:200,apply(as.data.frame(1:200),1, test), type="l", col="steelblue"); abline(a=0.5, b=0)`

Rechnen mit Wahrscheinlichkeiten

Weiteres Beispiel: Geschlechterverteilung bei Neugeborenen. In den meisten großen Populationen ist $P(\text{'Junge'}) > 0.5$ (England: $P(\text{'Junge'}) \approx 0.515$).

Mathematische Formalisierung 1930'er Jahre (Kolmogorov-Axiome)

Idee: Ordne jedem *Ereignis* eine *Wahrscheinlichkeit* (Zahl zwischen 0 und 1) zu.

Rechnen mit Wahrscheinlichkeiten

Die Menge aller möglichen Ergebnisse eines Zufallsexperiments wird **Ergebnisraum** Ω genannt.

Sei ω ein Element des Ergebnisraums Ω, dann heißt die einelementige Menge $\{\omega\}$ **Elementarereignis**. Geeignete Untermengen von $A \subset \Omega$ heißen **Ereignisse**. Insbesondere heißt das Ereignis $A = \Omega$ **sicheres Ereignis** und das Ereignis $A = \emptyset$ **unmögliches Ereignis**.

Ereignisse beschreiben Ergebnisse eines zufälligen Vorgangs (*Zufallsexperiment*). Ereignisse werden durch Aussagen bzw. Mengen ($A, B \subset \Omega$) beschrieben:

- Beispiel Ereignis A: „Morgen scheint die Sonne“.
- Ereignis B: „Morgen geht die Sonne im Osten auf.“

Ereignisse (Teilmengen von Ω) können miteinander verknüpft werden.

Rechnen mit Wahrscheinlichkeiten: Rechenregeln für Ereignisse

1. *Vereinigung* $A \cup B$: Ereignis A tritt ein, oder Ereignis B tritt ein, oder es treten beide Ereignisse gleichzeitig ein.
2. *Schnittmenge* $A \cap B$: Es treten beide Ereignisse gleichzeitig ein.
3. *Komplementärereignis* $\overline{A}$: Ereignis A tritt nicht ein.
4. A und B heißen *disjunkt*, wenn $A \cap B = \emptyset$ (leere Menge).

Bspl: Seien

- A: ' Patient besitzt Blutgruppe A'
- B: ' Patient besitzt positiven Rhesusfaktor'
- $\Rightarrow A \cup B$: Patient gehört zur Blutgruppe A oder besitzt positiven Rhesusfaktor oder beides, $A \cap B$=?, $\overline{A}$=?

Rechnen mit Wahrscheinlichkeiten: Rechenregeln für Ereignisse

Visualisierung durch Venn-Diagramme.

Praktische Bestimmung von Wahrscheinlichkeiten: Bei häufiger Durchführung des gleichen Experiments gilt

$$\text{relative Häufigkeit} = \text{Wahrscheinlichkeit des Ereignisses},$$

siehe Starkes Gesetz der großen Zahlen.

Kennt man einmal die Wahrscheinlichkeit gewisser 'Grundereignisse', so lassen sich die Wahrscheinlichkeiten für andere Ereignisse ableiten.

Rechnen mit Wahrscheinlichkeiten

Satz (Rechenregeln für Wahrscheinlichkeiten).

1. *Für jedes Ereignis A gilt $0 \leq P(A) \leq 1$.*
2. *Für das sichere Ereignis Ω gilt $P(\Omega) = 1$.*
3. *Additionsregel: Wenn Ereignisse A und B disjunkt sind, dann gilt*
$$P(A \cup B) = P(A) + P(B).$$
4. *Gegenereignis: Für jedes Ereignis A gilt $P(\overline{A}) = 1 - P(A)$.*
5. *Multiplikationsregel: Sind die Ereignisse A und B unabhängig, dann gilt $P(A \cap B) = P(A) \cdot P(B)$.*

Bspl. (Fortsetzung): $P(A) = 43\% = 0.43$, $P(B) = 0.81 \Rightarrow P(A \cap B) = 0.43 \times 0.81 = 0.3483$

3.3. Bedingte Wahrscheinlichkeiten

Wir suchen nun einen Begriff, der die **Wahrscheinlichkeit eines Ereignisses B** angibt, wenn wir schon **wissen, daß A eingetreten ist**.

Viele Krankheiten hängen vom Geschlecht der Person ab. Beispiel: W.keit an Diabetes zu erkranken: Für einen Mann ≈ 0.07, für eine Frau ≈ 0.02, Gesamtpopulation ≈ 0.045.

Welche W.keit ist (wenig) informativ/relevant?

Definition 3.1 (Bedingte Wahrscheinlichkeit). *Für zwei Ereignisse A und B definieren wir die* bedingte Wahrscheinlichkeit von A gegeben B *durch*

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

Aus der Definition der bedingten Wahrscheinlichkeit folgt direkt durch Umformen der **Multiplikationssatz**: Seien A und B Ereignisse mit einer Wahrscheinlichkeit ungleich Null. Dann gilt:

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

Bedingte Wahrscheinlichkeiten

Definiere Ereignisse

- Ereignis A='Person erkrankt an Diabetes'
- Ereignis B='Person ist männlich'
- Gesucht: $P(A|B)$

Beispiel: Ereignis A="Es regnet"; Ereignis B="Es ist bewölkt". Bekannt sei: $P(B) = 0.2$, $P(A \cap B) = 0.06$. Gesucht: $P(A|B)$

Lösung:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.06}{0.2} = 0.3$$

3.4. Satz von Bayes

Satz von Bayes

Wenn A, B unabhängige Ereignisse sind, gilt:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A).$$

Der **Satz von der totalen Wahrscheinlichkeit** besagt:
Sei $E_1, \ldots, E_n$ eine Partition eines Ereignisraumes Ω, d.h. $E_1 \cup \cdots \cup E_n = \Omega$ und $E_i \cap E_j = \emptyset$ für $i \neq j$. Sei $A \subset \Omega$ ein beliebiges Ereignis, dann gilt:

$$P(A) = \sum_{k=1}^{n} P(A \cap E_k) = \sum_{k=1}^{n} P(E_k) P(A|E_k).$$

Satz von Bayes

Satz. *Es seien A,B Ereignisse. Dann gilt*

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(\overline{A}) \cdot P(B|\overline{A})}.$$

Bemerkung: Ziel ist die Berechnung von $P(A|B)$ (sog. „a-posteriori Wahrscheinlichkeit“), wenn $P(A), P(B|A), P(B|\overline{A})$ bekannt sind.

Satz von Bayes

Beispiel: Screeningtest zur Erkennung von HIV-Infizierung.

Probanden sind entweder tatsächlich erkrankt K^+ oder gesund K^-,

Test beurteilt Probanden als testpositiv T^+ (i.e. „erkrankt“) oder testnegativ T^- („gesund“).

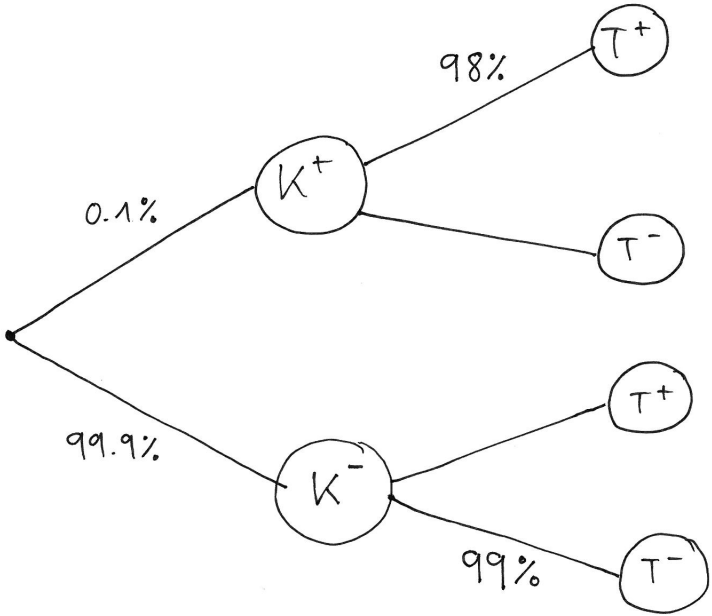
Möglicherweise können auch tatsächlich Gesunde (K^-) als testpositiv (T^+) bzw. tatsächlich Kranke (K^+) als testnegativ (T^-) beurteilt werden!

Bekannt sind:

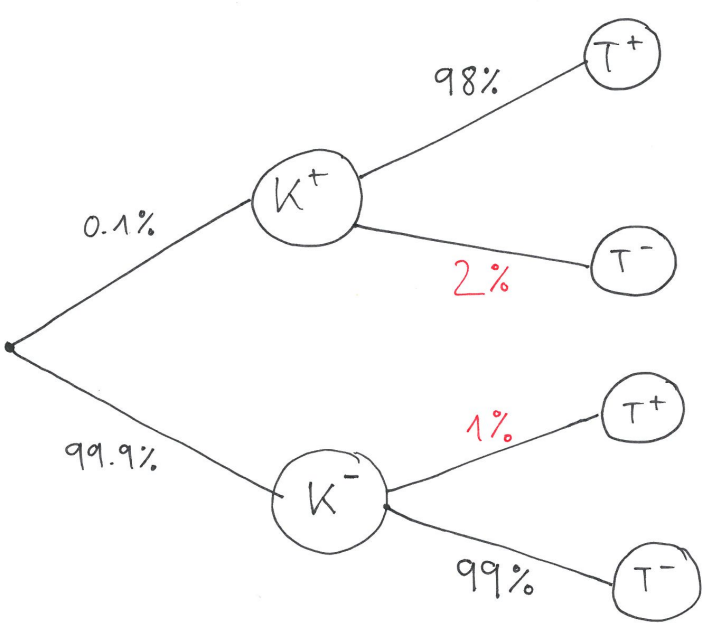
- $P(K^+) = 0.1\%$ (Prävalenz)
- $P(T^+|K^+) = 98\%$ (Sensitivität)
- $P(T^-|K^-) = 99\%$ (Spezifizität)

Angenommen positives Testergebnis liegt vor. Sollte man dann verzweifeln?
Gesucht: $P(K^+|T^+)$ (sog. positiver Vorhersagewert).

3.5. Visualisierung durch Wahrscheinlichkeitsbaum



Visualisierung durch Wahrscheinlichkeitsbaum



Visualisierung durch Wahrscheinlichkeitsbaum

Dann gilt

$$P(T^+ \cap K^+) = P(K^+) \cdot P(T^+|K^+) = 0.1\% \cdot 98\% = 0.00098,$$
$$P(T^+ \cap K^-) = P(K^-) \cdot P(T^+|K^-) = 99.9\% \cdot 1\% = 0.00999$$

und damit

$$P(T^+) = P(T^+ \cap K^+) + P(T^+ \cap K^-) = 0.01097.$$

Also gilt

$$P(K^+|T^+) = \frac{P(T^+ \cap K^+)}{P(T^+ \cap K^+) + P(T^+ \cap K^-)} = \frac{0.00098}{0.01097} = 0.08933$$

Allgemein als sog. *Formel (Satz) von Bayes*:

$$P(K^+|T^+) = \frac{P(K^+) \cdot P(T^+|K^+)}{P(K^+) \cdot P(T^+|K^+) + (1 - P(K^+)) \cdot (1 - P(T^-|K^-))}.$$

Der Satz von Bayes erlaubt die „Umkehrung“ von bedingten Wahrscheinlichkeiten. Standardanwendung: Gegeben ist $P(B|A)$, gesucht ist jedoch $P(A|B)$.

3.6. Zufallsvariablen

Definition 3.2 (Zufallsvariable). *Eine Größe X, deren Wert das Ergebnis eines Zufallsvorgangs (Zufallsexperiment) ist, heißt* Zufallsvariable *(oder Zufallsgröße). Genauer: Eine Zufallsvariable X zu einem Zufallsexperiment ist eine Funktion, die jedem Elementarereignis $\omega \in \Omega$ einen reellen Zahlenwert $X(\omega)$ zuordnet. Die Werte, die eine Zufallsvariable X annehmen kann, heißen Realisierungen der Zufallsvariable und werden mit Kleinbuchstaben x bezeichnet. Bezeichnungen: $X, Y, Z, \ldots$.*

Bei uns ordnen Zufallsvariablen jedem Ausgang des Zufallsvorgangs eine Zahl zu.

Beispiele:

- X= Anzahl von Ergebnis „Kopf“ beim 5-maligen Münzwurf.
- Y= Anzahl von Kreditausfällen in einem bestimmten Portfolio in einem bestimmten Zeitintervall.
- R= Rendite einer bestimmten Aktie in einem bestimmten Zeitintervall.

3.6.1. Stetig vs. diskret

Man unterscheidet *stetige* (stetig-verteilte) und *diskrete* (diskret-verteilte) Zufallsvariablen nach den möglichen Werten die sie annehmen können:

- Die möglichen Werte einer diskreten Zufallsvariablen lassen sich aufzählen, d.h. $X \in \{0, 1, 2, ...\}$ (Speziell: X kann nur endlich viele Werte annehmen.)
- Stetige Zufallsvariablen können prinzipiell alle Werte eines Intervalls $[a, b]$ oder alle Zahlen in $\mathbb{R}$ annehmen.

Beispiele:

- Y= Anzahl von Kreditausfällen in einem bestimmten Portfolio in einem bestimmten Zeitintervall.
- R= Rendite einer bestimmten Aktie in einem bestimmten Zeitintervall.
- X= Systolischer Blutdruck einer bestimmten Person zu einem bestimmten Zeitpunkt.

3.7. Dichtefunktion

Die Wahrscheinlichkeitsverteilungen von Zufallsvariablen lassen sich durch verschiedene Funktionen beschreiben.

Definition 3.3 (Dichtefunktion). *Die* Dichtefunktion *einer Zufallsvariablen X ist definiert durch*

a) eine Funktion $f = f_X$ mit $f(x) = f_X(x) = F'_X(x)$, falls X stetig,

b) eine Folge $p_0, p_1, p_2, \ldots$ mit $p_i = P(X = x_i)$, falls X diskret.

Alternative Bezeichnung: Wahrscheinlichkeitsfunktion, Wahrscheinlichkeitsdichte

Für diskrete Variablen gilt

$$0 \leq p_i \leq 1, \text{ für alle } i, \text{ und } \sum_i p_i = 1.$$

Die Dichtefunktion einer diskreten Zufallsvariablen lässt sich als Balkendiagramm darstellen.

Dichtefunktion

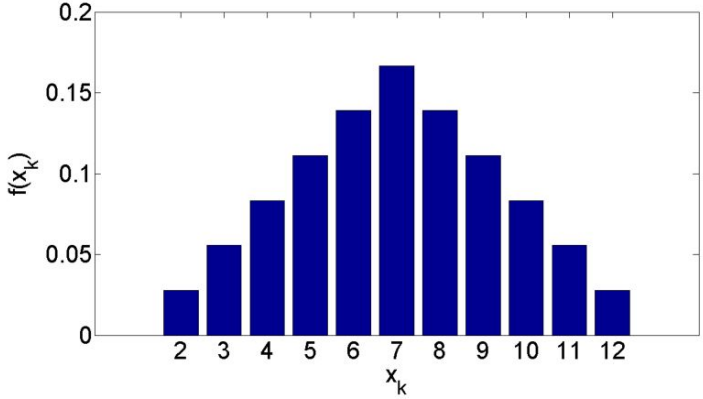


Abbildung 9: Augensummen beim Werfen zweier Würfel.

3.8. Verteilungsfunktionen

Bei einem Zufallsexperiment interessieren uns häufig nicht nur die Wahrscheinlichkeiten für eine Realisierung einer Zufallsvariablen $f(x_i) = P(X = x_i)$ sondern wir wollen wissen, wie wahrscheinlich es ist, dass X in einem bestimmten Bereich liegt: $P(a \leq X \leq b)$. Hierfür gibt es eine eigene Funktion, die Verteilungsfunktion.

Definition 3.4 (Verteilungsfunktion). *Die* Verteilungsfunktion *einer Zufallsvariablen X ist definiert durch*

$$F(x) = F_X(x) = P(X \leq x) \qquad (x \in \mathbb{R}).$$

Für eine diskrete Zufallsvariable gilt also

$$F(x) = P(X \leq x) = \sum_{i:x_i \leq x} f(x_i), \qquad x \in \mathbb{R}.$$

Verteilungsfunktionen

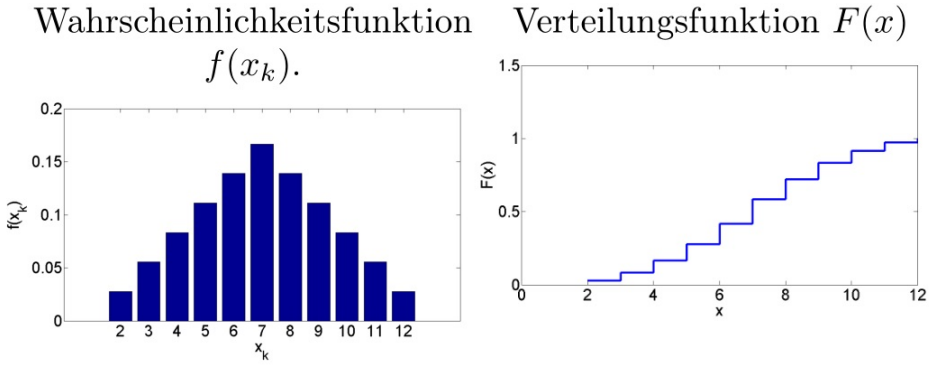


Abbildung 10: Wahrscheinlichkeits- und Verteilungsfunktion für die Augensumme zweier Würfel

Dichtefunktion

Verknüpfung zwischen Dichte- und Verteilungsfunktion:

a) falls X stetig gilt

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt.$$

b) falls X diskret gilt

$$F_X(x) = \sum_{i:i\leq x} p_i.$$

Verteilungs- und Dichtefunktionen enthalten die vollständige Information über die Verteilung einer Zufallsvariablen.
Oft ist man jedoch nur an gewissen Kenngrößen von Verteilungen interessiert...

Dichtefunktion

Für eine stetige Zufallsvariable X mit Dichte $f(x)$ und Verteilungsfunktion $F(x)$ gilt:

- $F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u$.
- $P(X \in [a, b]) = \int_{a}^{b} f(u)\mathrm{d}u = F(b) - F(a)$.
- $P(X = x) = 0$ für alle $x \in \mathbb{R}$.

3.9. Erwartungswert

Definition 3.5 (Erwartungswert). *a) Für eine stetige Zufallsvariable X ist der* Erwartungswert *definiert durch*

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) dx,$$

wobei f_X die zugehörige Dichtefunktion ist.

b) Für eine diskrete Zufallsvariable X ist der Erwartungswert *definiert durch*

$$\mathbb{E}(X) = \sum_{i=0}^{\infty} x_i \cdot p_i = \sum_{i=0}^{\infty} x_i \cdot P(X = x_i).$$

- $\mathbb{E}(X)$ gibt das „Zentrum“ (Schwerpunkt) der Verteilung an. $\mathbb{E}(X)$ entspricht dem arithmetischen Mittelwert einer Stichprobe (später mehr)
- Realisierungen x_i mit größerer Eintrittswahrscheinlichkeit $f(x_i)$ werden auch stärker berücksichtigt.

3.10. Varianz

Definition 3.6 (Varianz, Standardabweichung). *Die* Varianz *einer Zufallsvariable* X *ist gegeben durch*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2].$$

Für eine diskrete Zufallsvariable X *gilt*

$$\mathrm{Var}(X) = \sum_k (x_k - \mathbb{E}(X))^2 \, P(X = x_k).$$

Die Standardabweichung *ist gegeben durch* $\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$.

Die Varianz ist der mittlere (wahrscheinlichkeitsgewichtete) quadrierte Abstand der Realisierung der Zufallsvariablen von ihrem Erwartungswert.

$\mathrm{Var}(X)$ und $\mathrm{sd}(X)$ sind Maße für die Streuung/Variabilität der Zufallsvariablen X.

Erwartungswert und Varianz unter linearer Transformation

Sei X eine Zufallsvariable und $a, b \in \mathbb{R}$ beliebig. Dann gilt:

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

Standardisierung von Zufallsvariablen:

Definiere standardisierte/normierte Zufallsvariable

$$Z = \frac{X - \mathbb{E}(X)}{\mathrm{sd}(X)}.$$

Es gilt $\mathbb{E}(Z) = 0$, $\mathrm{Var}(Z) = 1$.

3.11. Wichtige Verteilungen

3.11.1. Binomialverteilung

Beispiel: Die Wahrscheinlichkeit, daß ein Junge geboren wird ist 0.51. Eine Familie habe vier Kinder. Wie groß ist die Wahrscheinlichkeit, daß alle 4 Kinder Mädchen sind?

- Sei $X = \#$ Mädchen bei 4 Geburten.
- Gesucht ist also $P(X = 4)$?

Lösung: Tafel

Binomialverteilung

Allgemeinere Situation:

- Es werden n unabhängige, identische Versuche durchgeführt.
- Bei jedem Versuch gibt es nur zwei mögliche Ausgänge: „Erfolg oder Mißerfolg“, „0 oder 1“, „ja oder nein“, ...
- Erfolgswahrscheinlichkeit für einzelnen Versuch: $p \in (0, 1)$
- Gesucht: Verteilung der Zufallsvariablen

$$X = \text{ Anzahl der Erfolge bei } n \text{ Versuchen.}$$

Binomialverteilung

Definition 3.7 (Binomialverteilung). *In der obigen Situation gilt: Für $k = 0, \dots, n$ ist*

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

In Zeichen: $X \sim \boldsymbol{Bin}(n, p)$
Sprechweise: „X ist binomialverteilt mit Parametern n und p“.
Dabei ist der Binomialkoeffizient („k aus n“) definiert durch

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

wobei

$$k! = k \cdot (k-1) \cdots 2 \cdot 1 \quad \textit{und} \quad 0! = 1.$$

Binomialverteilung

Eigenschaften: Für $X \sim \mathbf{Bin}(n, p)$ gilt

- X kann nur Werte $0, 1, \ldots, n$ annehmen, d.h. $\mathbf{Bin}(n, p)$ ist eine diskrete Verteilung.
- $\mathbb{E}(X) = n \cdot p$
- $\mathrm{Var}(X) = n \cdot p(1-p)$

Beispiel (Fortsetzung)

- $n = 4$ und $p = 1 - 0.51 = 0.49$ („Erfolg“ $\widehat{=}$ “Mädchen“)
- $X \sim \mathbf{Bin}(4, 0.49)$
- Lösung: $P(X = 4) = \binom{4}{4} 0.49^4 = 0.0576$
- $P(X = 2) = \binom{4}{2} 0.49^2 \cdot 0.51^2 = 0.3747$

Binomialverteilung

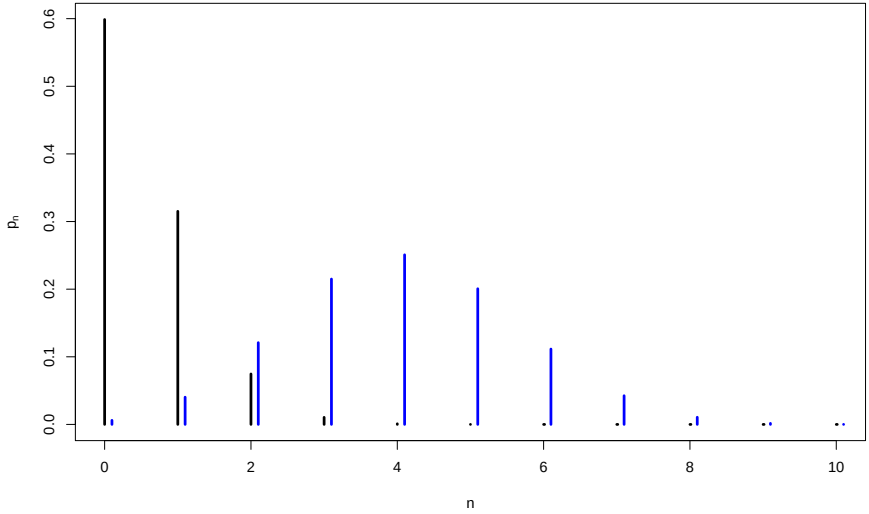


Abbildung 11: Dichtefunktion der Binomialverteilung: Dichten von **Bin**$(10, 0.05)$ (schwarz) und **Bin**$(10, 0.4)$ (blau).

3.11.2. Normalverteilung

Definition 3.8 (Normalverteilung). *Für $\mu \in \mathbb{R}, \sigma^2 > 0$ ist die Dichte der Normalverteilung $\boldsymbol{N}(\mu, \sigma^2)$ definiert durch*

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

Speziell für $\mu = 0$ und $\sigma = 1$ heißt $\boldsymbol{N}(0,1)$ Standardnormalverteilung. *Die Dichtefunktion lautet*

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

Normalverteilung

- $X \sim \mathsf{N}(\mu, \sigma^2) \Rightarrow X$ kann beliebige reelle Werte annehmen (ist also eine stetige Zufallsvariable)
- „Wichtigste“ Verteilung überhaupt, denn viele Daten lassen sich gut durch die Normalverteilung modellieren (Körpergrößen, Blutdruck,...).
- $\mathbb{E}(X) = \mu$ und $\mathsf{Var}(X) = \sigma^2$
- Sei X eine normalverteilte Zufallsvariable mit Parametern μ und σ, dann ist die **normierte Zufallsvariable** $Z = \frac{X-\mu}{\sigma}$ eine standardnormalverteilte Zufallsvariable mit Parametern $\mu = 0$ und $\sigma = 1$.
- Sei X eine normalverteilte Zufallsvariable mit Parametern μ und σ, dann gelten folgende Zusammenhänge zwischen der Verteilungsfunktion F und Dichte f von X und den entsprechenden Größen der Standardnormalverteilung Φ und ϕ:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \qquad f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

Normalverteilung

Dichtefunktion der Normalverteilung:

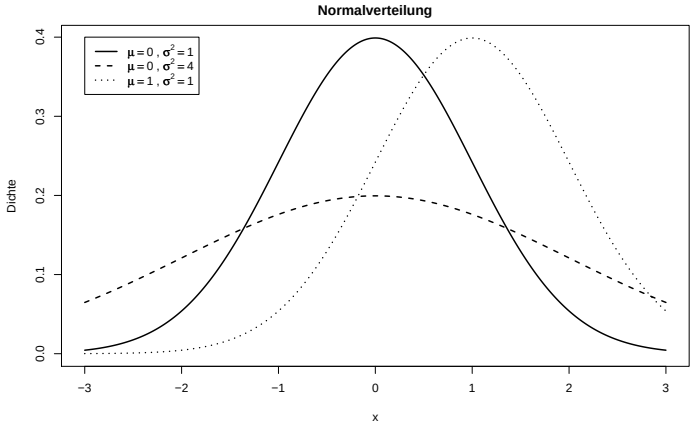


Abbildung 12: Die Dichtefunktion der Normalverteilung: die Gaußsche Glockenkurve. Hier: Dichten von $\mathbf{N}(0,1), \mathbf{N}(0,4)$ und $\mathbf{N}(1,1)$.

Normalverteilung

- Dichtefunktion symmetrisch um μ (dort maximal)
- μ beschreibt Lage der Verteilung
- σ beschreibt „Breite“ der Verteilung.

Warum taucht die Normalverteilung an so vielen Stellen auf?

Grund: Zentraler Grenzwertsatz (vage Formulierung):
„Unter sehr allgemeinen Bedingungen ist die Summe einer großen Anzahl von identisch verteilten Zufallsvariablen (die beliebige Verteilung besitzen dürfen!) normalverteilt“

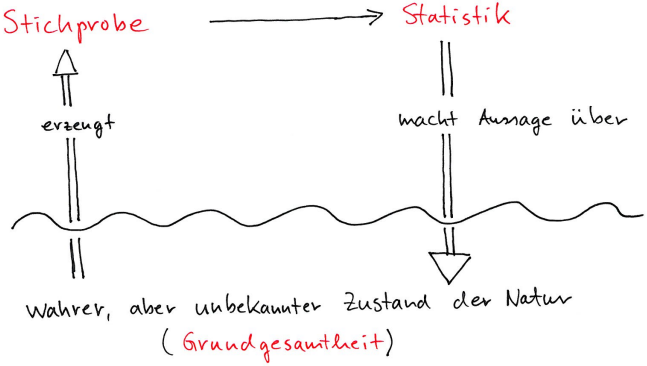
Normalverteilung

Sei X eine normalverteilte Zufallsvariable mit Parametern μ und σ, dann gilt:

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &\approx 68.3\% \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &\approx 95.5\% \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &\approx 99.7\% \end{aligned}$$

4. Parameterschätzung und Konfidenzintervalle

4.1. Einführung



Die Stichprobe soll Rückschlüsse auf die Grundgesamtheit liefern.

Einführung

Statistische Aussagen beziehen sich normalerweise auf einen (wahren, aber unbekannten) Parameter θ der Grundgesamtheit. Typische statistische Aussagen:

1. θ kann durch den Wert $\widehat{\theta} = 1.5$ geschätzt werden ($\Rightarrow$ **„Schätzer“** oder genauer **Punktschätzer**.).
2. Mit 95%-iger Wahrscheinlichkeit liegt θ zwischen 1 und 2 ($\Rightarrow$ **„Konfidenzintervall“**).
3. Mit einer Irrtumswahrscheinlichkeit von maximal 5% gilt $\theta > 0$ ($\Rightarrow$ **„Hypothesentest“**).

Einführung

Beispiel: Studie zur Wirksamkeit einer Therapie bei Bluthochdruck. Interessante Zielgröße ist die erreichte Blutdrucksenkung,

$$X \quad = \quad \text{Blutdruck vor Therapiebeginn - Blutdruck bei Therapieende}$$

Es werden $n = 10$ Patienten behandelt mit resultierender Stichprobe $x_1, \ldots, x_{10}$.

Fragestellung:

- Um wieviel reduziert die Therapie den Blutdruck?
- Genauer: Wie gross ist die *erwartete* Reduktion („mittlere Effektgrösse"), d.h. $\theta = \mathbb{E}(X)$?
- $\theta = \mathbb{E}(X)$: Wahre (aber unbekannte) Blutdruckreduktion *aller* Patienten

Einführung

Daten: $19, 7, 26, 25, -11, -8, 23, 30, 14, 16$.

- Problem: Gesucht ist $\theta = \mathbb{E}(X)$ aber wir haben nur die obigen Daten.
- Idee: Schätze $\theta = \mathbb{E}(X)$ durch das Stichprobenmittel $\overline{x} = \frac{1}{10}(x_1 + ... + x_{10})$.
- Sprechweise: „$\overline{x}$ ist ein Schätzer für $\mathbb{E}(X)$“.
- Hoffnung: $\overline{x}$ ist „guter“ Schätzer für $\mathbb{E}(X)$.
- Hier: $\overline{x} = 14.1$.

Was wissen wir jetzt über $\mathbb{E}(X)$? Ist tatsächlich $\mathbb{E}(X) = 14.1$?

- Sichere Aussage nicht möglich!
- Die beobachteten Daten sind zufällig, daher sind nur Aussagen möglich, die Unsicherheit berücksichtigen - aber wie?

$\Rightarrow$ Konfidenzintervall

Einführung

Wir können ein *Konfidenzintervall* angeben, d.h Zahlen $a < b$, so dass z.B. gilt

$$P(\text{wahrer Parameter } \mathbb{E}(X) \in [a, b]) = 95\%.$$

- Interpretation: Mit 95%-iger Sicherheit liegt die (wahre, aber unbekannte) Blutdruckreduktion zwischen den Werten a und b.
- In unserem Beispiel: 95%-Konfidenzintervall gegeben durch (fällt vom Himmel!)

  $$[7.902, 20.298]$$

  d.h. mit 95%-iger W.keit liegt die wahre (erwartete) Blutdruckreduktion zwischen 7.902 und 20.298.

4.2. Allgemeine Definition

Annahme: Es seien Daten (Stichprobe) $x_1, \dots, x_n$ gegeben.

Definition 4.1 (Schätzer). *Ein* Schätzer $\widehat{\theta}$ *(für einen Parameter* θ*) nimmt Daten* $x_1, \dots, x_n$ *und ordnet ihnen einen Wert* $\widehat{\theta} = \widehat{\theta}(x_1, \dots, x_n) \in \mathbb{R}$ *zu.*

Bemerkungen:

- Kurzschreibweise:

$$\text{Schätzer: Daten} \longmapsto \text{Wert}$$

- Da $x_1, \dots, x_n$ zufällig sind $\Rightarrow \widehat{\theta}$ ist eine Zufallsvariable!

Allgemeine Definition

Definition 4.2 (Konfidenzintervall). *Gesucht ist ein Parameter θ der Verteilung einer Zufallsvariablen X. Der Parameter soll aus einer Stichprobe $x_1, \dots, x_n$ geschätzt werden. Für eine vorgegebene (kleine) Wahrscheinlichkeit α ist $[a(x_1, \dots, x_n), b(x_1, \dots, x_n)]$ ein* Konfidenzintervall *zum* Konfidenzniveau $1 - \alpha$ *für den Parameter θ, wenn*

$$P(a(x_1, \dots, x_n) \leq \theta \leq b(x_1, \dots, x_n)) = 1 - \alpha.$$

Die Intervallgrenzen a und b hängen also von den Daten $x_1, \dots, x_n$ ab.

$1 - \alpha$ (grosse W.keit) heisst auch **Konfidenzniveau, Sicherheitsniveau, Überdeckungswahrscheinlichkeit**, α auch *Irrtumswahrscheinlichkeit*.

Allgemeine Definition: Konfidenzintervall

Da $x_1, \ldots, x_n$ zufällig sind $\Rightarrow$ Intervallgrenzen a und b sind Zufallsvariablen!

Interpretation: „Mit W.keit $1 - \alpha$ liegt der wahre aber unbekannte Parameter im Intervall $[a, b]$“.

Die Intervallgrenzen a und b hängen allgemein ab von:

- Sicherheitsniveau $1 - \alpha$,
- Stichprobenumfang n,
- Den beobachteten Daten $x_1, \ldots, x_n$.

4.3. Beispiel Normalverteilte Daten

Ist X eine normalverteilte Zufallsvariable mit Stichprobe $x_1, \ldots, x_n$, dann ist...

- Der Mittelwertschätzer $\overline{X} = \frac{1}{n} \sum_{j=1}^{n} x_j$ ein guter Schätzer für μ.
- Die empirische Varianz $s^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(x_j - \overline{X}\right)^2$ ein guter Schätzer für σ^2.

Guter Schätzer heißt streng genommen, dass der Schätzer erwartungstreu und konsistent ist.

Beispiel Normalverteilte Daten

Falls Varianz bekannt $\Rightarrow$ Konfidenzintervall:

Satz (Konfidenzintervall bei Normalverteilung, μ unbekannt σ^2 bekannt). *Seien $x_1, \dots, x_n \sim N(\mu, \sigma^2)$, μ unbekannt und σ^2 bekannt. Dann ist für α ein $1-\alpha$-KI gegeben durch*

$$\left[\overline{x} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right].$$

Dabei bezeichnet $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung.

Bemerkung: $z_{1-\alpha/2}$ wird aus Tabellen (oder vom Rechner) ermittelt (Bspl. Handout).

Beispiel Normalverteilte Daten

In unserem Beispiel: Es sei bekannt, dass $\sigma = 10$ ist. Dann berechnet sich das 95%-KI folgendermassen:

- $\overline{x} = 14.1$, $\sigma = 10$,
- $n = 10$,
- $\alpha = 5\%$,
- $z_{1-\alpha/2} = 1.95996$ (aus Tabelle).

Also KI $= [7.902, 20.298]$, d.h. mit 95%-iger Sicherheit, liegt die wahre aber unbekannte Blutdruckreduktion zwischen 7.902 und 20.298.

Beispiel Normalverteilte Daten

Satz (Konfidenzintervall bei Normalverteilung, μ und σ^2 unbekannt). *Seien $x_1, \dots, x_n \sim N(\mu, \sigma^2)$, μ und σ^2 unbekannt. Dann ist für α ein $(1-\alpha)$-KI gegeben durch*

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

Dabei bezeichnet $t_{n-1,1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der t_{n-1}-Verteilung.

Beispiel Normalverteilte Daten

Bemerkungen:

- Je grösser das Sicherheitsniveau, desto ??? das KI.
- Je grösser n, desto ??? das KI.
- Je grösser die Streuung σ, desto ??? das KI.
- Erstrebenswert: Möglichst ??? KI.
- Die Breite des KI ist ein Mass für die Präzision des Schätzers

Bspl: Tafel

Beispiel Normalverteilte Daten

Konsequenz:

Zur Bewertung eine Effekts (hier $\mathbb{E}(X)$) ist die Angabe eines Schätzers (hier $\overline{x}$) ohne die Angabe eines Konfidenzintervalls wenig informativ!

Regulatorische Relevanz (in der Medizin):

Bemerkungen:

ICH (E9):

- „Estimates of treatment effects should be accompanied by confidence intervals wherever possible“
- „... need to provide statistical estimates of treatment effects together with confidence intervals...“

Beispiel Normalverteilte Daten

Auch für viele andere Situationen kann man KI's berechnen:

- Daten normalverteilt und Varianz *unbekannt*
- Daten binomialverteilt (z.B. Therapieerfolg j/n)

5. Statistische Tests

5.1. Problemstellung und Grundbegriffe

Beispiel (Fortsetzung): Studie zur Wirksamkeit einer Therapie (z.b. Beta-Blocker) bei Bluthochdruck

Produzent (Pharma-Unternehmen) behauptet: „Medikament wirkt!“ Aber wie kann man das „beweisen“?

- Bekannt: Beobachtungen $x_1, \ldots, x_{10}$ (Blutdruckreduktion in 10 Probanden)
- Erinnerung: $\widehat{\mu} = 14.1 \Rightarrow$ deutet auf erfolgreiche Blutdrucksenkung hin, aber ist das sicher?
- Ziel: Wollen Aussage über (unbekanntes) $\mu = \mathbb{E}(X)$ treffen! Aber wie?

Problemstellung und Grundbegriffe

Übersetzung in ein „statistisches Entscheidungsproblem“.

- Übersetzung der abzusichernden Aussage:

$$H_1: \quad \mu > 0 \qquad (\text{„Alternative“})$$

- Gegenteil von Aussage H_1:

$$H_0: \quad \mu \leq 0 \qquad (\text{„Nullhypothese“})$$

- $\widehat{\mu} = \overline{x} = 14.1 \Rightarrow$ Entscheidung für H_0 oder H_1?

Problemstellung und Grundbegriffe

Statistischer Test: Daten $\longrightarrow$ Entscheidung für H_0 oder H_1.

Verfahren sollte

- transparent und reproduzierbar sein (Wissenschaftlichkeit!)
- möglichst wenig Fehler produzieren.

Problemstellung und Grundbegriffe

Mögliche Konstellationen beim Testen:

	Realität	
Testentscheidung	H_0 wahr	H_0 falsch
H_0 abgelehnt	Fehler 1. Art α-Fehler	Entscheidung richtig Power $(1-\beta)$
H_0 beibehalten	Entscheidung richtig Sensitivität $(1-\alpha)$	Fehler 2. Art β-Fehler

Fehler 1. und 2. Art können nicht gleichzeitig kontrolliert werden, deshalb wird immer der Fehler 1. Art kontrolliert. Das Signifikanzniveau entspricht einer oberen Grenze für den Fehler 1. Art. Auch bei kontrolliertem Fehler 1. Art versucht man mit möglichst hoher Power zu testen - dies erreicht man häufig durch einen größeren Stichprobenumfang.

Problemstellung und Grundbegriffe

Der Test entscheidet in zwei Situationen falsch:

1. H_0 wird abgelehnt (verworfen), obwohl H_0 wahr ist. Dies ist der *Fehler 1. Art* oder auch α-*Fehler*, in Formel

$$\text{Fehler 1. Art} = P(\text{Test lehnt } H_0 \text{ ab} | H_0)$$

2. H_0 wird nicht abgelehnt, obwohl H_0 falsch ist. Dies ist der *Fehler 2. Art* oder auch β-*Fehler*, in Formel

$$\text{Fehler 2. Art} = P(\text{Test lehnt } H_0 \text{ nicht ab} | H_1)$$

Problemstellung und Grundbegriffe

In unserem Beispiel (Übung):

- Fehler 1. Art = ???
- Fehler 2. Art = ???

Wichtig: Bei statistischen Tests kann nur der Fehler 1. Art vom Anwender vorgegeben („kontrolliert“) werden, Über den Fehler 2. Art weiß man i.A. nichts!

5.2. Ablauf eines statistischen Tests

1. Hypothesenformulierung: Aufstellen von H_0 (*Nullhypothese*) und H_1 (*Alternative*). Ziel: Verwerfung von H_0! (Im Beispiel: $H_0 : \mu \leq 0$ vs. $H_1 : \mu > 0$)

2. Lege *Signifikanzniveau* $\alpha \in (0, 1)$ fest. α: maximal tolerierbare W.keit für einen Fehler 1.Art. Konvention: 5%.

3. Berechne aus Stichprobe $x_1, \ldots, x_n$ den Wert der zugehörigen *Teststatistik* bzw. *Prüfgrösse* T. (Im Beispiel: $T = \sqrt{n}\frac{\bar{x}}{s} = \sqrt{10}\frac{14.1}{14.09846} = 3.1626$.)

4. Ermittle (Tabelle oder mit Software) *kritischen Wert* k_α ab dessen Über- bzw. Unterschreiten H_0 abgelehnt wird. (Im Beispiel: $k_\alpha = t_{n-1,1-\alpha} = \ldots = 1.8331$).

5. Testentscheidung: Lehne H_0 ab, wenn $T > k_\alpha$. (Im Beispiel: $T = 3.1626$ und $k_\alpha = 1.8331 \Rightarrow$ Lehne H_0 ab!)

Ablauf eines statistischen Tests

Anders formuliert: Aus einer Zufallsstichprobe $X_1, \ldots, X_n$ lässt sich eine Teststatistik $T(X_1, \ldots, X_n)$ ermitteln, deren Verteilung unter H_0 bekannt ist. Der Ablehnungsbereich A umfasst Werte für T, die wenn H_0 gilt ('unter H_0') sehr unwahrscheinlich sind, d.h. $P(T \in A|H_0) \leq \alpha$. Die Wahrscheinlichkeit α heißt das Signifikanzniveau des Tests und liegt üblicherweise bei $\alpha = 0.05$. Fällt das Testergebnis T in den Ablehnungsbereich, so wird H_0 zugunsten von H_1 abgelehnt. Ansonsten wird H_0 beibehalten.

5.3. Interpretation und Konsequenzen der Testentscheidung

H_0 wurde abgelehnt:

- Wenn H_0 abgelehnt wird, sprechen wir von einem *signifikantem* Testergebnis, d.h. das Ergebnis ist nicht mehr nur durch den Zufall erklärbar.

  Sprechweise, falls der Test H_0 ablehnt: Die Abweichung des Parameters von H_0 ist signifikant (statistisch gesichert) mit Wahrscheinlichkeit α.

- Genauer: Die W.keit, ein Ergebnis wie das realisierte (also $T = 3.1626$) unter der Nullhypothese zu beobachten ist „extrem“ klein - d.h. kleiner als α (=5%).

- Ziel war Verwerfung von $H_0 \Rightarrow$ „positives“ Testergebnis.

Interpretation und Konsequenzen der Testentscheidung

Der p-Wert:

- Oben: Nur Testentscheidung gegen H_0.
- Wie „stark“ ist die Evidenz gegen H_0?
- p-Wert: „W.keit, dass der gefundene Wert der Test-Statistik (also $T = 3.1626$ - oder noch größer) zustande kommt, wenn in Wirklichkeit die Nullhypothese wahr ist.“
- Salopp: „W.keit für einen Zufallsbefund.“
- Entscheidungsregel: Wenn p-Wert $\leq \alpha \Rightarrow$ Lehne H_0 ab!
- Berechnung des p-Werts durch Statistikprogramme. In unserem Beispiel: p-Wert $= 0.005751$.

Interpretation und Konsequenzen der Testentscheidung

Fehler 2. Art und *Power* des Tests:

- Power $= 1-$ Fehler 2. Art $= P($ Test lehnt H_0 ab $|H_1)$
- Power ist die W.keit, die richtige Entscheidung zu treffen, wenn die Alternative wahr ist.
- Die Power hängt i.A. ab von
  - Stichprobenumfang n
  - (unbekannte) Abweichung von der Nullhypothese
  - (unbekannte) Variabilität der Grundgesamtheit (üblicherweise $\mathrm{Var}(X)$ o.ä.)

Interpretation und Konsequenzen der Testentscheidung

p-Wert $\leq \alpha$ besagt lediglich, ob statistisch signifikanter Effekt vorliegt.

- Keine Information über die Größe des Effekts ($\Rightarrow$ Konfidenzintervall liefert das.)
- Keine Information über die praktische Relevanz des Effekts. Wenn die Stichprobe sehr groß ist, werden selbst kleinste (praktisch bedeutungslose) Abweichungen von H_0 als signifikant deklariert.

Statistische Signifikanz $\neq$ *Praktische Relevanz.*

Interpretation und Konsequenzen der Testentscheidung

Interpretation eines nicht-signifikanten Testergebnisses:

- Nicht „für“ Gültigkeit von H_0, sondern eher im Sinne von: Die Indizien sind nicht stark genug, um die Unschuldsvermutung (H_0) auszuschliessen. Es gibt zwei mögliche Erklärungen:
- H_0 ist wirklich wahr,
- H_1 ist wirklich wahr, aber die Power des Tests ist zu niedrig.

Interpretation und Konsequenzen der Testentscheidung

Software-Output (in R) für obiges Beispiel („Einseitiger Einstichproben t-Test“):

```
> blutdruck.reduktion
[1]  19   7  26  25 -11  -8  23  30  14  16
> t.test(blutdruck.reduktion,alternativ="greater")
One Sample t-test

data:  blutdruck.reduktion
t = 3.1626, df = 9, p-value = 0.005751
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
5.927386      Inf
sample estimates:
mean of x
14.1
```

5.4. Ausblick: Einteilung von Tests

Für fast jede Situation gibt es (mindestens) einen Test. Auswahl des Testverfahrens hängt von verschiedenen Faktoren ab:

a) Aufgabe des Tests:

   i. Vergleich von Erwartungswerten (unser Beispiel)

   ii. Vergleich von Varianzen

   iii. Vergleich von Häufigkeiten (Ausfallw.keiten)

   iv. ...

b) Anzahl Stichproben, die in den Test eingehen:

   i. Eine (unser Beispiel) (*Einstichprobentest*)

   ii. Zwei (*Zweistichprobentest*)

   iii. Mehr als zwei

c) Sind Sichproben voneinander unabhängig (*unverbunden*) oder abhängig (*verbunden*)? Bspl. klinische Studie

d) *Einseitige* Fragestellung, z.B.

$$H_0 : \mu \leq 0 \qquad \text{vs.} \qquad H_1 : \mu > 0$$

oder *zweiseitige* Fragestellung

$$H_0 : \mu = 0 \qquad \text{vs.} \qquad H_1 : \mu \neq 0.$$

e) *Parametrischer* Test (geht von bestimmten Verteilungsannahmen aus) vs. *nichtparametrischer* (verteilungsfreier) Test

Ausblick: Zweiseitiger Gauß-Test

Dies ist ein Test für den Erwartungswert bei bekannter Varianz unter Normalverteilungsannahme. Wir testen den Wert von μ eines Merkmals $X \sim \mathbb{N}(\mu, \sigma^2)$ bei bekanntem σ.

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

- Das Signifikanzniveau ist α.
- Gegeben sind Beobachtungen $x_1, \dots, x_n$.
- Unter H_0 gilt für die Teststatistik

  $$T(x_1, \dots, x_n) = \sqrt{n}\frac{\overline{x} - \mu_0}{\sigma} \sim \mathbb{N}(0.1)$$

  Bestimme das $1 - \frac{\alpha}{2}$-Quantil der Standardnormalverteilung $z_{1-\frac{\alpha}{2}}$.
- Entscheidungsregel: Lehne H_0 ab wenn $|T(x_1, \dots, x_n)| > z_{1-\frac{\alpha}{2}}$.

Ausblick: t-Test

Der t-Test ist ein Test für den Erwartungswert μ eines normalverteilten Merkmals $X \sim \mathbb{N}(\mu, \sigma^2)$ bei unbekanntem σ. Die Hypothesen sind

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

Das Signifikanzniveau sei α.

- Ziehe eine Stichprobe vom Umfang n, berechne den Mittelwert $\overline{x}$ und die empirische Standardardabweichung s.
- Bestimme das $1-\frac{\alpha}{2}$-Quantil $t_{n-1,1-\frac{\alpha}{2}}$ der t-Verteilung mit $n-1$ Freiheitsgraden.
- Berechne die Teststatistik
$$T(x_1, \dots, x_n) = \sqrt{n}\frac{\overline{x} - \mu_0}{s},$$
mit $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ und $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^n x_i^2 - (\overline{x})^2\right)$.
- Wenn H_0 stimmt, gilt $T(X_1, \dots, X_n) \sim t_{n-1}$. Lehne H_0 daher ab wenn
$$|T(x_1, \dots, x_n)| > t_{n-1,1-\frac{\alpha}{2}}.$$

Literatur

Gerd Bosbach and Jens Jürgen Korff. *Lügen mit Zahlen: Wie wir mit Statistiken manipuliert werden*. Heyne Verlag, 2011.

Prof. Dr. Christoph Becker
Fachbereich Mathematik und Naturwissenschaften
Hochschule Darmstadt
Haardtring 100
64295 Darmstadt
christoph.becker@h-da.de