# Statistik

Anton Klimovsky

29. Juni 2023

# Wozu brauchen wir Statistik?

- **Datenflut** in allen Wissenschaften.
- Technologie+Wissenschaft liefern (Un)Mengen von Daten.
- Daten kommen aus Experimenten, Messungen, Beobachtungen, Simulationen, usw.
- Die Daten sind oft **variabel** und **unübersichtlich**.
- Oft kennen wir die **Mechanismen**, die die Daten erzeugen nicht genau:
  - Unsicherheit, partielle Information, Messfehlern, Rauschen, usw.
  - Der Mechanismus kann sehr Wohl deterministisch sein, allerdings "kompliziert" (viele Parametern, die unbekannt sind).
- Wie gewinnt man das **Wissen** aus den Daten?

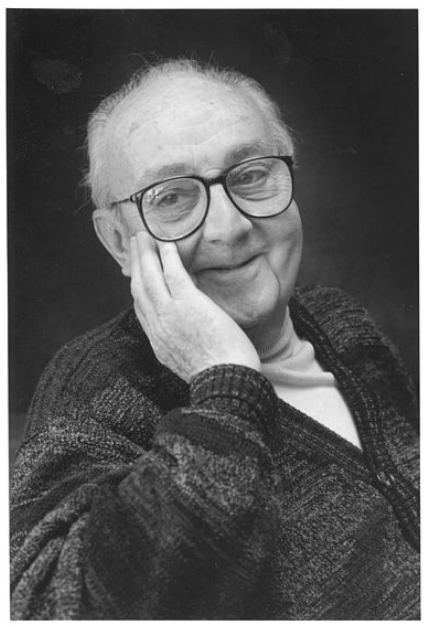
$$\boxed{\text{Daten}} + \boxed{\text{Modelle}} \overset{\textbf{Stochastik}}{\longleftrightarrow} \boxed{\text{Wissen}}.$$

# Ansatz der Statistik

- **Statistik**: modelliere Daten als **Realisierungen** von **ZVen**!
- $\rightsquigarrow$ wir modellieren die Variabilität (z.B. Erscheinung der Natur) durch Zufall (mathematische Abstraktion).
- Die ZV sind in einem **stochastischen Modell** zu spezifizieren.
- Die ZVen haben oft **Parameter** z.B.:
  - Die Erfolgswahrscheinlikeit $p$ bei der Binomial-Verteilung.
  - Der Erwartungswert $\mu$ und die Varianz $\sigma^2$ bei der Normalverteilung,
  - usw.
- Die Parameter sind oft nicht genau bekannt.
- **Statistik**: **präzisiere die Parameter anhand der Daten**.
- $\rightsquigarrow$ In Statistik macht man Rückschlüsse von den Daten auf die Parameter des Modells.

# (Stochastische) Modelle

- **Modell = ein Zufallsobjekt** (ZV, Zufallsvektor, Folge von ZVen, Zufallsmatrix, usw.).
- **Frage:** Welche Verteilung anzunehmen?
- **Vorsicht:** Die **Modellwahl** ist ein großes Problem. Man braucht **kritisches Denken**.
- **“Essentially, all models are wrong, but some are useful”**.



George E. P. Box (1919–2013)

# Was macht man mit Statistik?

- **Bsp.** Betrachten wir ein lineares Modell $Y = aX + W$, wobei $Y$, $X$ und $W$ ZVen sind.
  - $X$ ist das eigentliche Signal.
  - $Y$ ist der in einem Experiment messbare Wert.
  - $W$ ist das Rauschen (= der Messfehler).
  - $a$ ist ein Parameter des Modells.
- **Schätzung**: kenne $a$, schätze $X$ anhand von den Messungen von $Y$.
- **Modellbildung**: kenne $X$, vermesse $Y$, schätze $a$.

Speziell ist das Ziel:

- **Schätzung**: versuche den **kleinsten Schätzungsfehler** zu erreichen.
- **Hypothesentests** (als eine Form der Modellbildung): Eine unbekannte Variable (z.B. $a$) nimmt einen von den wenigen Werten an. Entscheide sich anhand von den Daten für einen Wert. Versuche dabei den Wert mit der **kleinsten Wahrscheinlichkeit der Fehlentscheidung** auszuwählen.

# Wie geht man mit den Daten um?

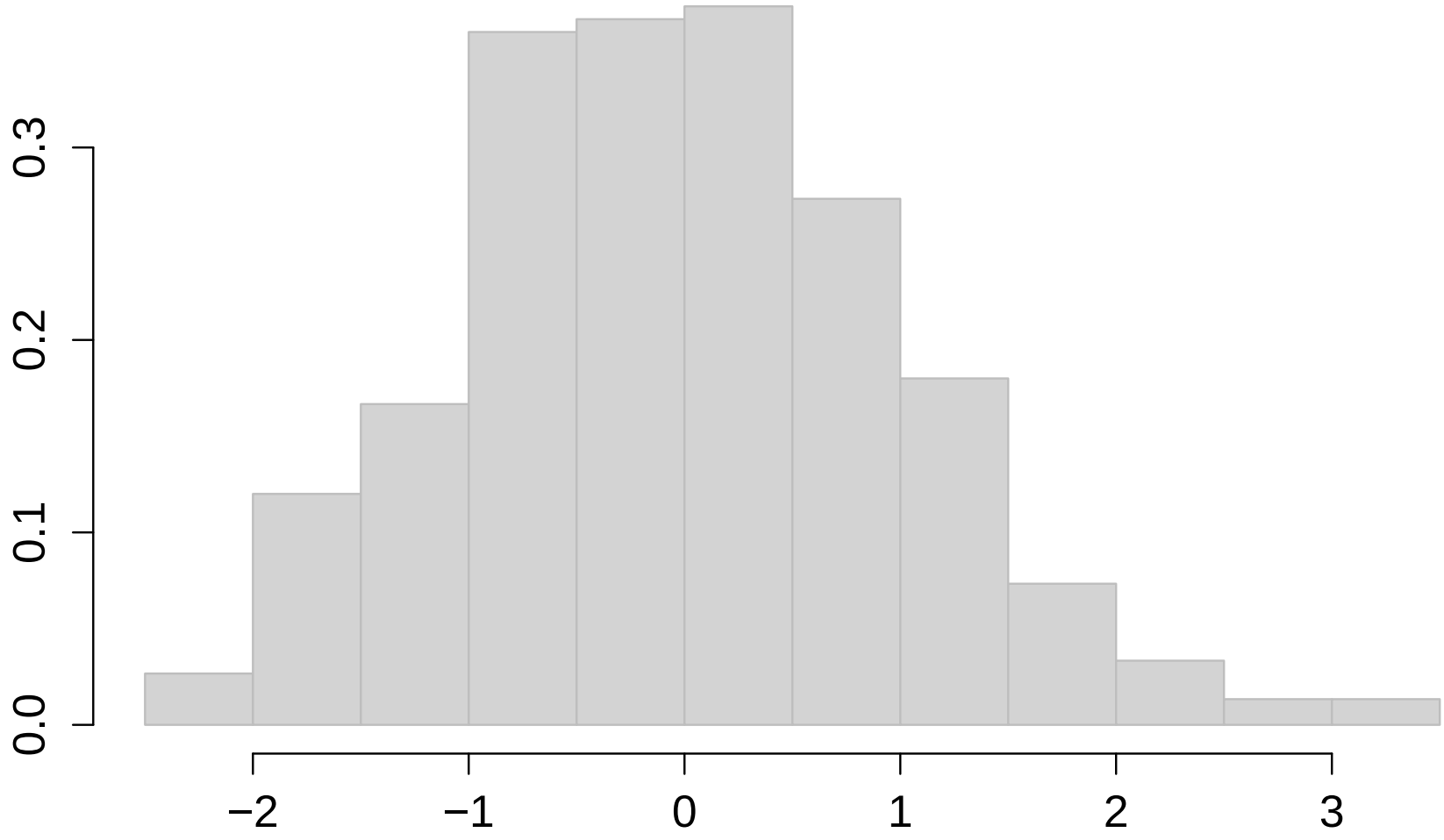
- (Schritt 0) **Deskriptive (= beschreibende) Statistik**. Verschaffe den ersten Eindruck von den Daten mittels graphischen Darstellungen und numerischen Kenngrößen. So kann man eventuell ein besseres Modell für die Daten wählen.
- (Schritt 1) **Induktive (= beurteilende) Statistik**: Das Schätzen und die Hypothesentests anhand von Modellen und Daten.
  Zwei Methodologien:
  - **Bayes'sche Statistik**: der unbekannte Parameter ($a$ im Besipiel) ist eine ZV. Starte mit einer "Vorabschätzung": **a priori Verteilung** von diesem Parameter. Anhand von den Daten verbessere die a priori Verteilung.
  - **Klassische (auch frequentistische) Statistik**: der Unbekannter Parameter ist eine feste Konstante.

# Deskriptive Statistik

Wie sehen die Daten aus?

## Histogramm

# Idee: die “empirische Dichte”.

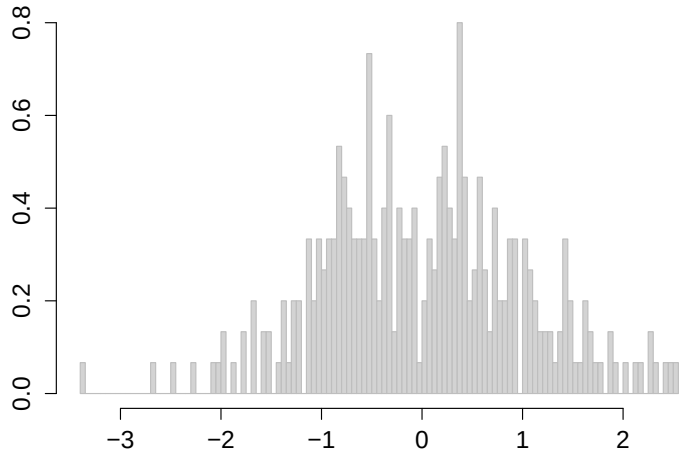


# Die “empirische Dichte”

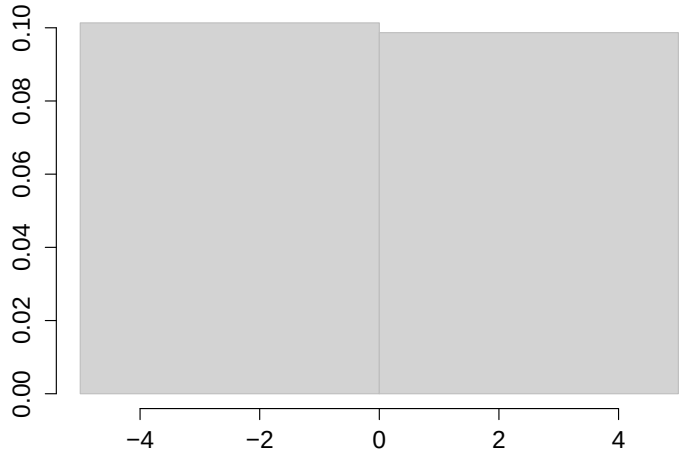
- Seien $x_1, x_2, \ldots, x_n \in \mathbb{R}$ die Daten.
- Sei $\mathbb{R} = \underbrace{(-\infty, a_1]}_{=I_0} \cup \bigcup_{=1}^{m} \underbrace{(a_k, a_{k+1}]}_{=I_k} \cup \underbrace{(a_m, +\infty)}_{=I_{k+1}}$ eine disjunkte **Zerlegung des Wertebereichs**, wobei $a_1 < a_2 < \ldots < a_m$.
- Diese Zerlegung induziert eine Gruppierung der Daten.
- Sei $n_k := |\{i : x_i \in I_k\}|$ die absolute und $f_k := n_k/n$ die **relative Häufigkeit** der Daten aus der Gruppe $k$.
- Die **Breite** des $k$-ten Rechtecks ist $a_{k+1} - a_k$.
- Die **Höhe** ist $f_k/(a_{k+1} - a_k)$.
- Somit ist die **Gesamtfläche** $= 1$.
- **Vorteil:** So kann man die Datensätze von verschiedenen Umfängen $n$ miteinander vergleichen.

# Wie klein/groß sollen die Zerlegungsintervalle sein?

- So **groß wie möglich** um das Rauschen zu unterdrücken:




- So **klein wie möglich** um die Struktur zu sehen:



# Boxplot

## Wie sind die Daten gestreut?

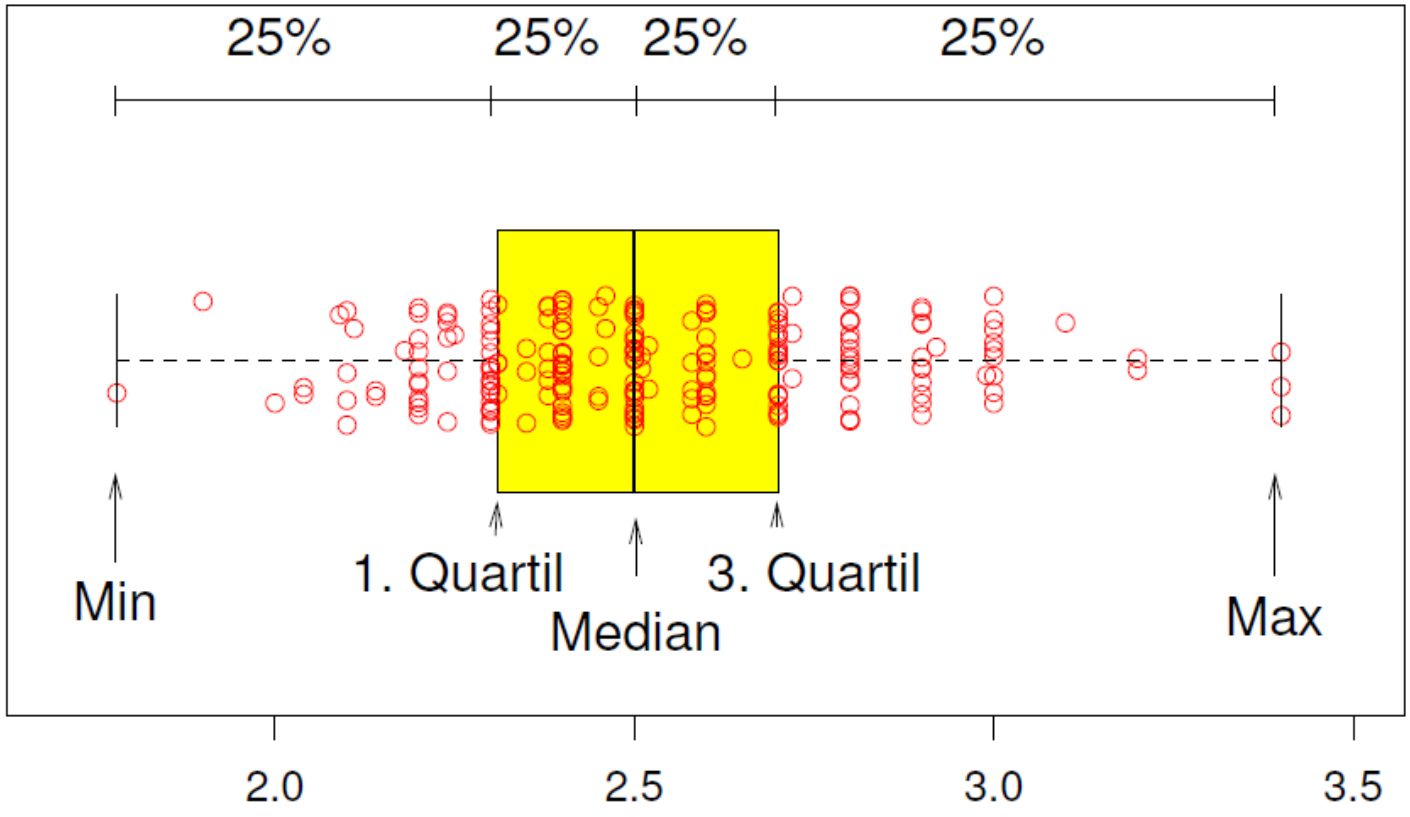



Bild ©Matthias Birkner

## Kenngrößen

- Wie groß? $\rightsquigarrow$ **Lageparameter**.
- Wie variabel?

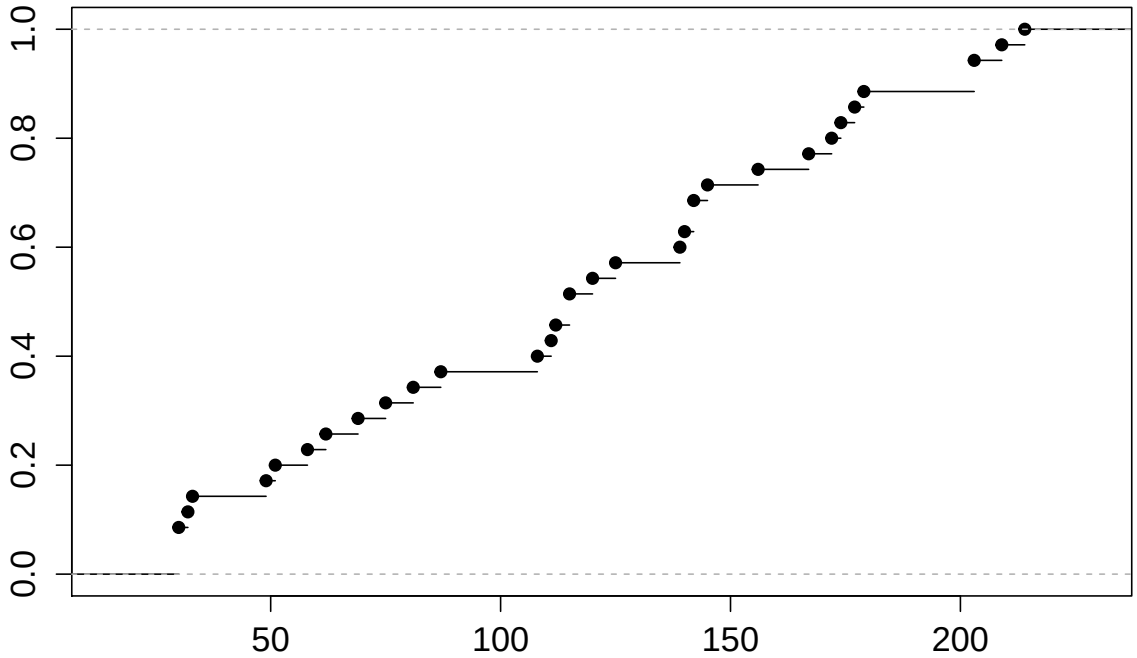
  $\rightsquigarrow$ **Steungsparameter**.

# Empirische Verteilung

- Sei $x_1, x_2, \ldots, x_n \in \mathbb{R}$ die **Strichprobe** (= die Daten).
- Die **empirische Verteilung** ist $P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \in \mathscr{M}_1(\mathbb{R})$, wobei

$$\delta_x(A) = \begin{cases} 1, & x \in A \\ 0, & \text{sonst.} \end{cases}$$

  die Dirac-Maß ist.
- Die **empirische Verteilungsfunktion** ist $F_n(x) := P_n(-\infty, x]$ für $x \in \mathbb{R}$.



# Lageparameter

- Der **empirische Mittelwert** ist $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} x_i = \int_{-\infty}^{\infty} x F_n(\mathrm{d}x)$.
- Der **Median** $m := F_n^{-1}(0.5)$.
- **N.B.** Der Median ist robuster als der empirische Mittelwert bzgl. der **Ausreißer**.
- Das $\alpha$**-Quantil** ist gleich $F_n^{-1}(\alpha)$ für $\alpha \in [0,1]$.
- $\rightsquigarrow$ Das $\alpha$-Quantil teilt den Datensatz in zwei Hälften:
  - $100\alpha\%$ der Daten sind kleiner als das $\alpha$-Quantil
  - und die restliche $100(1-\alpha)\%$ sind größer.
- $F_n^{-1}(3/4)$ heißt oberes Quartil.
- $F_n^{-1}(1/4)$ heißt unteres Quartil.

# Streuungsparametern

- Der **Quartilabstand** ist

$$\mathrm{IQR}_n := F_n^{-1}(3/4) - F_n^{-1}(1/4).$$

- Die **empirische Varianz** ist

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X}_n)^2 = \int_{-\infty}^{\infty}(x - \bar{X}_n)^2 F_n(\mathrm{d}x).$$

- Die (korrigierte) **Stichproben-Varianz** ist

$$s_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X}_n)^2 = \frac{n}{n-1}\sigma_n^2.$$

# Beurteilende Statistik

Rückschlüsse auf das Modell

# (Klassische) statistische Modelle

**Definition (Stichprobe)**

Sei $X\colon \Omega \to \mathscr{X}$ eine $\mathscr{X}$-wertige ZV. Jede Realisierung $x = X(\omega)$ heißt eine **Stichprobe** von $X$. $\mathscr{X}$ heißt der Beobachtungsraum.

**Bsp.** Oft ist $\mathscr{X} = \mathbb{R}^n$. Z.b. sind in der soziologischen Umfrage die Meinungen $(X_1, X_2, \ldots, X_n) \in \{0,1\}^n$ von den $n$ befragten Personen die Stichprobe.

**Definition (Statistisches Modell)**

Seien $X\colon \Omega \to \mathscr{X}$ die Daten. Ein (klassisches) **statistisches Modell** für $X$ ist eine Familie von W-Verteilungen auf $\mathscr{X}$

$$\mathscr{P} = \{P_\theta \in \mathscr{M}_1(\mathscr{X})\colon \theta \in \Theta\},$$

wobei $\Theta$ heißt der **Parameterraum** und $\theta \in \Theta$ ein **Parameter**.

**Bsp.** Bei der soziologischen Umfrage ist das Produkt von der Bernoulli-Verteilungen $\{\mathrm{Bernoulli}(p)^{\otimes n}\colon p \in [0,1]\}$ ein statistisches Modell. Die Erfolgswahrscheinlichkeit $p$ ist der unbekannte Parameter.

# Schätzer

**Definition (Schätzer)**

Eine messbare Abbildung $T\colon\ \mathscr{X} \to \Theta$ heißt der **Schätzer** (auch die **Schätzstatistik**).
Für ein gegebenes $x$ heißt $T(x)$ ein **Schätzwert** (auch **Punktschätzer**).

**Bsp.** In unserem Beispiel mit der soziologischen Umfrage ist der empirische Mittelwert $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ ein Schätzer.

**N.B.** Man missbraucht manchmal die Notation und betrachtet den Schätzer als $\Theta$-wertige ZV $T = T(X)$.

# Gewünschte Eigenschaften von den Schätzern

**Ansatz der klassischen Statistik:** Sei $X \sim P_\theta$.

Sei $\{T_n\}_{n=1}^{\infty}$ eine Familie von den Schätzern.

- $T_n$ heißt **erwartungstreu**, wenn $\mathbb{E}[T_n(X)] = \theta$.
- $T_n$ heißt **asymptotisch erwartungstreu**. wenn $\lim_{n\to\infty} \mathbb{E}[T_n(X)] = \theta$.

## Definition (Gewünschte Eigenschaften von den Schätzern)

- $\{T_n\}_{n=1}^{\infty}$ heißt **konsistent**, wenn $T_n(X) \xrightarrow[n\to\infty]{\mathbb{P}} \theta$.
- das **Risiko** (auch der **mittlere quadratische Fehler**) des Schätzers ist

$$R(T_n, \theta) := \mathbb{E}[(T_n(X) - \theta)^2] = \mathrm{Var}[T_n(X) - \theta] + (\underbrace{\mathbb{E}[T_n(X) - \theta]}_{=:\mathrm{Bias}[T_n(X)]})^2$$

$$= \mathrm{Var}[T_n(X)] + \mathrm{Bias}[T_n(X)]^2$$

(**Idee:** Das Risiko von einem guten Schätzer soll “möglichst klein” sein.)

**N.B.** Das kleine Risiko ist die wichtigste Eigenschaft von einem Schätzer. (Die Erwartungstreue ist überbewertet.)

**Bsp.**

- Der **empirische Mittelwert** $\bar{X}_n = \frac{1}{n}\sum_{n=1}^{\infty} X_i$ ist erwartungstreu (folgt aus der Linearität vom $\mathbb{E}$) und ist konsistent (folgt aus dem GGZ).
- Die **empirische Varianz** $\sigma_n^2 = \frac{1}{n}\sum_{n=1}^{\infty}(X_i - \bar{X}_n)^2$ ist nicht erwartungstreu, da $\mathbb{E}[\sigma_n^2] = \frac{n-1}{n}\mathrm{Var}[X_1]$. Die Varianz $\sigma_n^2$ ist konsistent (folgt aus GGZ).
- Die **korrigierte Stichproben-Varianz** $s_n^2 = \frac{n}{n-1}\sigma_n^2$ ist erwartungstreu und konsistent.

**Lemma (Risiko = Varianz + Bias$^2$)**

*Es gilt*

$$R(T,\theta) = \mathrm{Var}[T(X)] + (\mathrm{Bias}[T(X)])^2.$$

**Theorem (eine hinreichende Bedingung für die Konsistenz)**

*Sei* $R(T_n,\theta) \underset{n\to\infty}{\longrightarrow} 0$. *Dann ist* $T_n$ *konsistent.*

**Beweis.**

Folgt aus der Tschebyscheff-Ungleichung (oder Markov-Ungleichung). □

# Konstruktion von Schätzern

**Maximum-Likelihood-Methode:** wähle den geschätzten Parameter, so dass die Wahrscheinlichkeit von der Stichprobe maximal ist.

- Sei $X \sim P_\theta$.
- Sei $x = X(\omega)$ eine Realisierung von $X$.
- Wähle das $\theta$, das die **"Daten am wahrscheinlichsten macht"**:

$$\hat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta \in \Theta} P_\theta\{X \in [x, x + \mathrm{d}x]\}.$$

**Bsp.** Sei $X_1, X_2, \ldots, X_n$ u.i. $\mathrm{Exp}(\theta)$-verteilt.

- $\operatorname*{argmax}_\theta \left(\prod_{i=1}^n \theta \mathrm{e}^{-\theta x_i}\right) = \operatorname*{argmax}_\theta \left(n \log \theta - \theta \sum_{i=1}^n x_i\right)$
- $\Rightarrow \hat{\theta}_{\mathrm{ML}} = \frac{n}{x_1 + \ldots + x_n}$.

**N.B.** Man kann zeigen, dass die ML-Schätzern gute asymptotische ($n \to \infty$) Eigenschaften haben, wie z.B. Konsistenz.

**N.B.** Selten kann man den ML-Schätzer explizit analytisch finden. Man kann diesen aber approximativ numerisch ausrechnen.

# Bayes'sche Statistik

**Idee**: der unbekannte Parameter $\theta$ ist eine ZV.

## Definition (Bayes'sches Modell)

Das **Bayes'sche statistische Modell** ist gegeben durch die **bedingte Verteilung** von der Stichprobe

$$P_{\mathscr{X}|\Theta}(\cdot \mid \theta) \in \mathscr{M}_1(\mathscr{X})$$

gegeben $\theta$ und die **A-priori-Verteilung** $P_\Theta \in \mathscr{M}_1(\Theta)$. Somit ist

$$X \sim \int_\Theta P_{\mathscr{X}|\Theta}(\cdot \mid \theta) P_\Theta(\mathrm{d}\theta). \tag{1}$$

Die A-priori-Verteilung $P_\Theta$ repräsentiert unsere **ursprüngliche Vorstellung** über den unbekannten Parameter (aus Erfahrung, aus vorherigen Experimenten, usw.) und quantifiziert unsere **Unsicherheit** über den Wert vom unbekannten Parameter $\theta$.

**N.B.** Das Bayes'sche Modell ist keine Familie von Verteilungen, sondern eine Verteilung von der Bauart (1)!

Nun können wir unsere ursprüngliche Vorstellung über den unbekannten Parameter $\theta$ anhand von der Stichprobe $X$ korrigieren.

**Definition (A posteriori Verteilung)**

Nach der Bayes-Regel gilt

$$P_{\Theta|\mathscr{X}}(\mathrm{d}\theta \mid \mathrm{d}x) = \frac{P_{\mathscr{X}|\Theta}(\mathrm{d}x \mid \mathrm{d}\theta) P_{\Theta}(\mathrm{d}\theta)}{P_X(\mathrm{d}x)}, \tag{2}$$

wobei

$$P_X(\mathrm{d}x) := \int_{\Theta} P_{\mathscr{X}|\Theta}(\mathrm{d}x \mid \theta) P_{\Theta}(\mathrm{d}\theta). \tag{3}$$

Bedingte Verteilung $P_{\Theta|\mathscr{X}}$ heißt die **A-posteriori-Verteilung** von $\theta$ gegeben die Stichprobe $X$.

**N.B.** Analytisch kann man nur in wenigen Fällen die a posteriori Verteilung explizit ausrechnen, da (3) oft nicht explizit ausrechenbar ist.

**N.B.** Numerisch können wir sehr wohl den **Markov-Ketten-Monte-Carlo-Algorithmus** benutzen um die Stichproben aus (2) zu ziehen! Der Vorteil ist, dass wir (3) nicht explizit ausrechnen müssen.

# Bedingte Erwartung ist der beste Schätzer

- **Das Schätzen ohne Daten:** Sei $\theta$ eine ZV

$$c^* := \operatorname*{argmin}_{c \in \mathbb{R}} \mathbb{E}[(\theta - c)^2] = ?$$

Es gilt

$$\boxed{c^* = \mathbb{E}[\theta]}.$$

- **Das Schätzen mit Daten:** Sei $\theta$ und $X$ zwei ZVen.

$$\operatorname*{argmin}_{g\colon \mathscr{X} \to \Theta} \mathbb{E}[(\theta - g(X))^2 \mid X = x] = \mathbb{E}[\theta \mid X].$$

$\Rightarrow \mathbb{E}[\theta \mid X]$ **minimalisiert die quadratische Abweichung** $\mathbb{E}[(\theta - g(X))^2]$
**über alle Schätzern** $g\colon \mathscr{X} \to \Theta$**.**

**N.B.** Geometrisch ist also der bedingte Erwartungswert eine $L^2$**-Projektion** von $\theta$ auf dem Raum von allen Schätzern.

# Lineare Methode der kleinsten Quadraten

- Betrachte speziell die **lineare Schätzern**

$$\hat{\theta} = aX + b.$$

- **Frage:** $\operatorname{argmin}_{a,b\in\mathbb{R}} \mathbb{E}[(\theta - aX - b)^2] = (\hat{a}, \hat{b}) = ?$
- Sei $F(a,b) := \mathbb{E}[(\theta - aX - b)^2]$. Dann gilt

$$\begin{cases} \frac{\partial}{\partial a} F(\hat{a}, \hat{b}) = 0 \\ \frac{\partial}{\partial b} F(\hat{a}, \hat{b}) = 0. \end{cases}$$

- $\hat{a} = \frac{\mathrm{Cov}[X,\theta]}{\mathrm{Var}[X]}$.
- $\hat{b} = \mathbb{E}[\theta] - \hat{a}\mathbb{E}[X]$.
- Somit ist der **beste lineare Schätzer**

$$\hat{\theta}_{\mathrm{L}} := \mathbb{E}[\theta] + \frac{\mathrm{Cov}[X,\theta]}{\mathrm{Var}[X]}(X - \mathbb{E}[X]).$$

# Maximale a-posteriori-Wahscheinlichkeit Methode

Was tun wenn man nur einen Wert vom unbekannten Parameter berichten soll?

- **Maximale a-posteriori-Wahscheinlichkeit:**

$$\theta^* := \operatorname*{argmax}_{\theta \in \Theta} P_{\Theta|\mathscr{X}}(\theta \mid x).$$

- Alternativ kann man den **bedingten Erwartungswert angeben:**

$$\mathbb{E}[\theta \mid X = x] = \int \theta P_{\Theta|\mathscr{X}}(\mathrm{d}\theta \mid x).$$

**N.B.** Die Angabe von einem einzigen möglichen Wert ist nicht sehr informativ!

# Konfidenzintervale

- Sei $\{\mathbb{P}_\theta : \theta \in \Theta\}$ unser **statistisches Modell**
- Sei $\alpha \in (0,1)$ das vorgegebene **Niveau**.
- Sei $X$ unsere **Stichprobe**.
- Der **Schätzer** $T : \mathscr{X} \to \Theta$ liefert uns einen Schätzwert $\hat{\theta} = T(X)$, der in der Regel nicht mit dem wahren Parameter $\theta_0 \in \Theta$ übereinstimmt!
- Deswegen gibt man nicht nur den Schätzwert an, sondern gibt man noch zusätzlich ein **zufälliges Interval von den möglichen $\theta$-Werten** an, das mit der vorgegebenen Wahrscheinlichkeit $\alpha$ den unbekannten wahren Parameter $\theta_0$ enthält.
- Je breiter das Interval ist, desto weniger Präzision hat unsere Abschätzung $\hat{\theta} \approx \theta_0$.
- So erhält man mehr Information über die Präzision unseres Schätzverfahrens als Funktion vom Niveau
- $\rightsquigarrow$ **Fehlerbalken.**

# Konfidenzinterval

## Definition (Konfidenzinterval)

Sei $X \sim P_\theta$. Ein $(1-\alpha)$**-Konfindenzinteval** (auch Vertrauensinterval) für $\theta$ ist ein zufälliges Interval $[T_n^-, T_n^+]$, wobei $T_n^-, T_n^+ \colon \mathcal{X} \to \Theta$ zwei Statistiken sind, so dass

$$\mathbb{P}\{\theta \in [T_n^-(X), T_n^+(X)]\} \geq 1-\alpha, \quad \forall \theta \in \Theta.$$

Das $1-\alpha$ heißt das **Niveau** vom Konfidenzinterval.

**Idee:** Das $\alpha$ soll klein sein.

**Bsp.** $0.95$**-Konfidenzinterval für den unbekannten Erwartungswert.** Sei $X_1, X_2, \ldots, X_n$ u.i.v. mit $\mathbb{E}[X_1] = \theta$ (unbekannt) und $\mathrm{Var}[X_1] = \sigma^2$ (bekannt). Falls die Normalapproximation anwendbar ist (s. ZGS), gilt

$$\mathbb{P}\left\{\sqrt{n}\left|\frac{\bar{X}_n - \theta}{\sigma}\right| \leq 1.96\right\} \approx 0.95$$

$$\Leftrightarrow \mathbb{P}\left\{\bar{X}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \bar{X}_n + \frac{1.96\sigma}{\sqrt{n}}\right\} \approx 0.95,$$

da $\Phi\{1.96\} \approx 1 - 0.05/2$ (ein Tabellenwert).

# Strategie zur Konstruktion von Konfidenzintervalen

- Seien $\{\mathbb{P}_\theta : \theta \in \Theta\}$ unser statistisches Modell und die Irrtumswahrscheinlichkeit (das Niveau) $\alpha \in (0,1)$ gegeben.
- Ausgangspunkt ist ein Schätzer $T(X_1,...,X_n)$, der
  - den zu schätzenden Parameter $\theta$ enthält,
  - dessen Verteilung jedoch **nicht** von $\theta$ abhängt.
  - z.B. hängt (nach ZGS) die Verteilung von $\sqrt{n}\left|\frac{\bar{X}_n-\theta}{\sigma}\right| \overset{\mathrm{D}}{\approx} \mathcal{N}(0,1)$ von $\theta$ NICHT ab!
- Suche nun das $\frac{\alpha}{2}$-Quantil $q_{\frac{\alpha}{2}}$ und das $\frac{\alpha}{2}$-Quantil $q_{1-\frac{\alpha}{2}}$.
- Ein Konfidenzintervall zum Niveau $\alpha$ ist gegeben durch

$$\{\theta : T(X_1,...,X_n) \in [q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}]\}.$$

# Konfidenzintervale bei unbekannter Varianz

**Bsp.** Aus der **soziologischen Umfrage** erhalten wir eine Stichprobe der Meinungen $X_1, X_2, \ldots, X_n \in \{0,1\}$ aus der Gesamtpopulation, wobei es angenommen wird, dass $\{X_i\}_{i=1}^n$ u.i. $\mathrm{Bernoulli}(p)$-Verteilten ZVen sind. Hier ist $p \in [0,1]$ der unbekannte Parameter.

- Bis jetzt haben wir eine **ad hoc Abschätzung** benutzt: Da $X_1$ Bernoulli-verteilt sind, gilt $\sigma^2 = \mathrm{Var}[X_1] \leq 1/4$ immer.
- Alternativ können wir die **emprische Varianz als Schätzer** für $\sigma^2$ in das Konfidenzinterval einsetzen. Es gilt (nach dem GGZ):

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2.$$

- Somit gilt $\mathbb{P}\left\{\bar{X}_n - \frac{1.96\sigma_n}{\sqrt{n}} \leq \theta \leq \bar{X}_n + \frac{1.96\sigma_n}{\sqrt{n}}\right\} \approx 0.95$ für $n$ groß genug.

# Gauß'sche Stichproben

## Definition ($\xi^2$-Verteilung)

Sei $X_1, X_2, \ldots, X_n$ u.i. $\mathcal{N}(0,1)$-verteilte ZVen. Die Verteilung von der ZV $\sum_{i=1}^{n} X_i^2$ heißt die **$\chi_n^2$-Verteilung**.

## Definition ($t$-Verteilung)

Sei $X_0, X_1, X_2, \ldots, X_n$ u.i. $\mathcal{N}(0,1)$-verteilte ZVen. Die Verteilung von der ZV $X_0 / \sum_{i=1}^{n} X_i^2$ heißt die **$t_n^2$-Verteilung** (auch Student-Verteilung).

## Theorem (Verteilung von Gauß'schen Schätzern)

*Sei $X_1, X_2, \ldots, X_n$ u.i. $\mathcal{N}(\mu, \sigma^2)$-verteilte ZVen. Dann gilt*

- $\bar{X}_n$ *und* $s_n^2$ *sind* ***unabhängige ZVen****.*
- $\bar{X}_n$ *ist* $\mathcal{N}(\mu, \sigma^2)$***-verteilt****.*
- $(n-1)s_n^2/\sigma^2$ *ist* $\chi_{n-1}^2$***-verteilt****.*
- $T := \sqrt{n}\frac{\bar{X}_n - \mu}{s_n}$ *ist* $t_{n-1}$***-verteilt****.*

# Multivariate Gauß'sche Verteilung

## Theorem (Dichte der multivariaten Gauss'schen Verteilung)

*Seien $X_1, X_2, \ldots, X_n$ u.i.v. $\mathcal{N}(0,1)$. Sei $B \in \mathbb{R}^{n\times n}$ und $m \in \mathbb{R}^n$. Sei $C := BB^*$ eine invertierbare Matrix. Dann hat $Y := BX + m$ die Dichte*

$$f_{m,C}(y) := \frac{1}{(2\pi)^{n/2}|\det C|^{1/2}} \exp\left(-\frac{1}{2}\langle (y-m), C^{-1}(y-m)\rangle\right) \tag{4}$$

*und gilt $C_{i,j} = \mathrm{Cov}[Y_i, Y_j]$.*

## Definition (Multivariate Gauß'sche Verteilung)

Jede positiv definite symmetrische Matrix $C \in \mathbb{R}^{n\times n}$ und jeder Vektor $m \in \mathbb{R}^n$ definieren eine **multivariate Gauß'sche Verteilung** $\mathcal{N}(m, C)$ durch (4).

## Theorem (Lineare Transformationen von Gauß'schen Vektoren)

*Sei $Y \sim \mathcal{N}(m, C)$ ein Zufallsvektor in $\mathbb{R}^n$. Sei $A \in \mathbb{R}^{k\times n}$ eine Matrix mit dem maximalen Rang und $a \in \mathbb{R}^k$. Dann gilt $(AY + a) \sim \mathcal{N}(Am + a, ACA^*)$.*

# Konfidenzinteval für den Erwartungswert im Gauss'schen Model

- Sei $X \sim \mathcal{N}(\theta, \sigma^2)^{\otimes n}$.
- Der renormalisierte Schätzer für den Erwartungswert

$$T_n := \frac{\sqrt{n}(\bar{X}_n - \theta)}{s_n}.$$

- $T_n \sim t_{n-1}$!
- $\mathbb{P}\{\theta \in [\bar{X}_n - z\frac{s_n^2}{\sqrt{n}}, \bar{X}_n + z\frac{s_n^2}{\sqrt{n}}]\} \geq 0.95$, wobei $\alpha = 0.05$ und
- $z = F_{t_{n-1}}(1-\alpha/2)$ das $(1-\alpha/2)$**-Quantil** von der $t_{n-1}$-Verteilung ist.

**N.B.** $t_{n-1}$ konvergiert gegen $\mathcal{N}(0,1)$ als $n \to \infty$. (Dies folgt aus dem ZGS).

# Bootstrap-Verfahren

- **Bootstrap** (=Stiefelschlaufe). Dabei zieht man sich selbst an eigenen Stiefelschlaufen in die Höhe (=US Amerikanische Variante von **Baron Münchhausen**).
- Sei $X_1, X_2, \ldots, X_n$ ein Datensatz (= Sample), wobei $X_i$ unabhängig identisch Verteilt mit Verteilungsfunktion $F$.
- Sei $\theta = \theta(X_1, X_2, \ldots, X_n)$ ein Schätzer.
- **Idee:** approximiere die **Konfidenzintervale** und (sogar noch allgemeiner!) die **Verteilung** von $\theta$ durch wiederholtes ziehen mit Zurücklegen aus dem Datensatz:
  1. **Resample:** erzeuge $X_1^*, X_2^*, \ldots X_n^*$, wobei $X_i^* \sim \hat{F}_n$, wobei $\hat{F}_n$ die **empirische Verteilung** vom $X_1, X_2, \ldots, X_n$ ($\rightsquigarrow$ **ziehen mit Zurücklegen** aus $X_1, X_2, \ldots, X_n$.)
  2. Berechne das **Bootstraped Schätzer**: $\theta^* = \theta(X^*)$.
  3. Wiederhole Schritte 1. und 2. mehrfach (z.B. 1000 mal).
  4. $\rightsquigarrow \theta_1^*, \theta_2^*, \ldots, \theta_{1000}^*$ (Bootstrap-Kopien vom Schätzer) $\rightsquigarrow$ Bootstrap-Histogram.
  5. **Approximiere die Verteilung von $\theta$ mit dem Bootstrap-Histogram**.
- **Vorteil:** keine parametrische Modellannahmen! Beliebige Schätzern!

# Lineare Regression: Formulierung

- Sei $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ die **Stichprobe**.
- Sei

$$\boxed{y_i = \theta_0 + \theta_1 x_i + W_i,}$$

das **lineare Modell**, wobei z.B. $W_i \sim \mathcal{N}(0, \sigma^2)$ u.i.v. **Rauschen**.

- Die **Maximal-Likelihood-Methode** führt zur Aufgabe

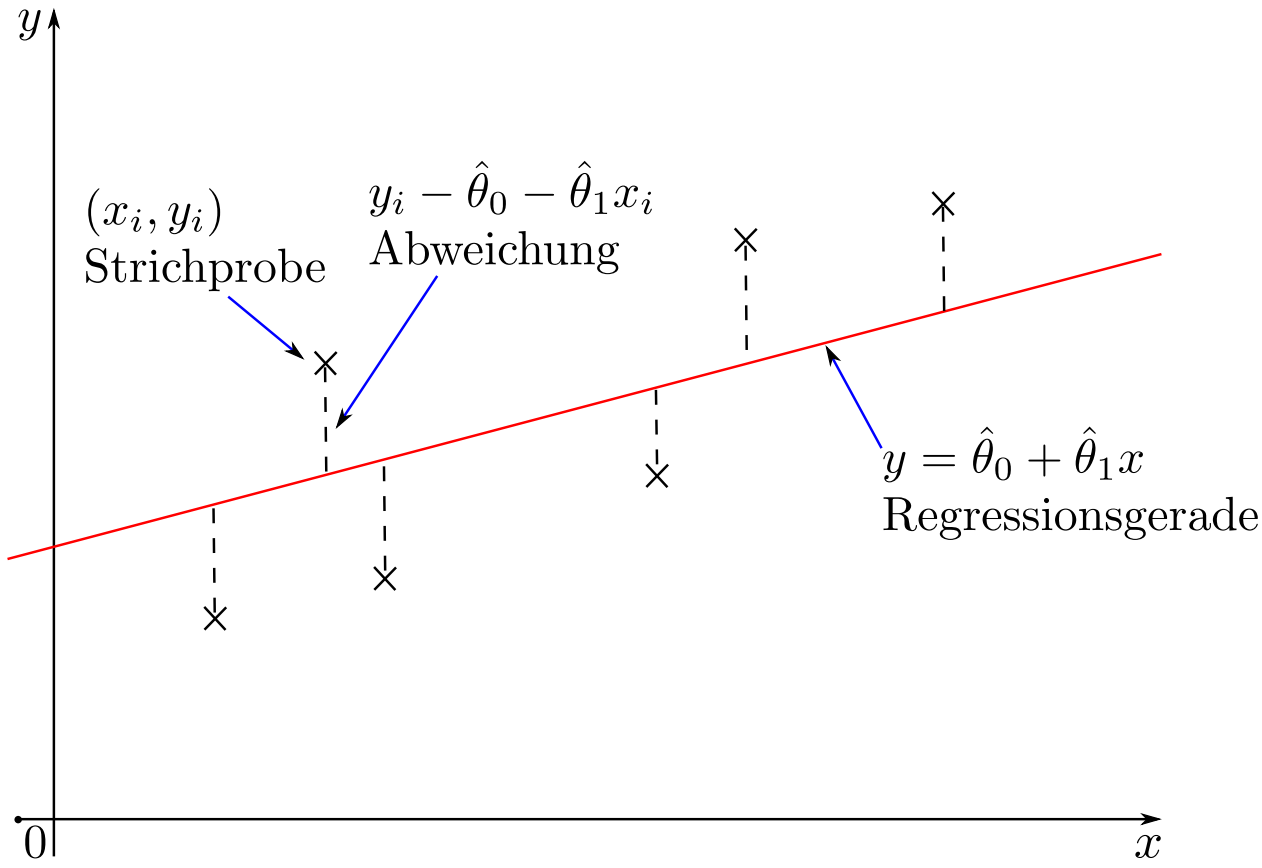
$$\text{argmax}_{\theta_0, \theta_1 \in \mathbb{R}} \left\{ c \cdot \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 \right) \right\} = \textbf{?}$$

- Dies ist Äquivalent zur **Methode der kleinsten-Quadraten**

$$\boxed{\text{argmin}_{\theta_0, \theta_1 \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 = \textbf{?}}$$

und heißt die **lineare Regression**.

# Lineare Regression: graphische Skizze



# Lineare Regression: analytische Lösung

$$\underset{\theta_0,\theta_1\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2. \tag{5}$$

- Die Lösung von (5) ist

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{6}$$

$$\hat{\theta}_2 = \bar{y} - \bar{\theta}_1 \bar{x}, \tag{7}$$

  wobei $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$ und $\bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n)$. (Diese bekommt man nach einer kurzen Rechnung aus der **Stationarität**.)
- **N.B.** Für das Modell $Y = \theta_0 + \theta_1 X + W$, wobei $X, Y, W$ ZVen mit $W$ unabhängig von $X$ und $\mathbb{E}[W] = 0$ sind, gilt

$$\theta_1 = \frac{\operatorname{Cov}[X,Y]}{\operatorname{Var}[X]}, \quad \theta_0 = \mathbb{E}[Y] - \theta_1 \mathbb{E}[X]. \tag{8}$$

**N.B.** Die Ausdrücke (6) und (7) sind die **empirische Version** von (8)!

# Hypothesentests: die Idee

- Der **Hypothesentest** ist ein Schätzverfahren, wobei der schätzende Parameter $\theta$ nur **wenige mögliche Werte** annehmen kann. Entscheide sich für den Wert, der die **Fehlerwahrscheinlichkeit minimier**t.
- Oft sieht die Hypothese folgendermaßen aus. Hypothese: $\theta \approx \theta_0$, wobei $\theta_0$ ein gegebener Wert ist.
- **Idee:**
  - Anhand von einer Stichprobe bilde die entsprechende **Konfidenzintervale** zum Niveau $\alpha$
  - Prüfe ob der **hypothetische Wert** $\theta_0$ im Konfindenzinterval liegt.
  - Falls ja, **verwerfe die Hypothese NICHT** (zum Niveau $\alpha$).

**N.B.** Die Wortwahl “verwerfe die Hypothese nicht” ist gewollt. Da wir dabei keine $100\%$ Sicherheit haben (lediglich $(1-\alpha)100\%$), können wir nicht sagen, dass wir die Hypothese annehmen.

# Hypothesentests: allgemeiner Ansatz

- Binäres $\theta \in \{H_0, H_1\}$:
  - **Nullhypothese** $H_0 \colon X \sim P_{H_0}$.
  - **Alternative** $H_1 \colon X \sim P_{H_1}$.
- Zerteile den Stichprobenraum in zwei Teile $\mathscr{X} = \mathscr{R} \cup (\mathscr{X} \setminus \mathscr{R})$. Sei $x \in \mathscr{X}$ die beobachtete Daten. Die **Entscheidungsregel**:

  verwerfe die Nullhypothese $H_0$, falls $x \in \mathscr{R}$.

  Der Menge $\mathscr{R}$ heißt Verwerfungsbereich
- Dabei sind folgende Fehlerarten möglich:
  - **Typ-I-Fehler** (= falsches Verwerfen, falscher Alarm): $H_0$ **ist richtig aber verworfen**.

    $$\alpha(\mathscr{R}) = P_{H_0}\{X \in \mathscr{R}\}.$$

  - **Typ-II-Fehler** (= falsche Akzeptanz): $H_0$ **ist falsch aber nicht verworfen**.

    $$\beta(\mathscr{R}) = P_{H_1}\{X \notin \mathscr{R}\}.$$

# Hypothesentests: Signifikanztests

**Bsp.** In einem **Münzwurfexperiment** haben wir $S = 472$ mal die Zahl in insgesamt $n = 1000$ Versuchen bekommen.

**Frage: Ist die Münze fair?**

**Formalisierung:** Gegeben sind die **Nullhypothese** $H_0 \colon p = 1/2$ und die **Alternative** $H_1 \colon p \neq 1/2$, wobei wir $(\mathrm{Bernoulli}(p))^{\otimes n}$ als das Modell benutzen.

- Wähle den Schätzer: $S/n$.
- Bilde den entsprechenden **Verwerfungsbereich**: $|S - n/2| \geq \xi$.
- Wähle das **Signifikanzniveau** $\alpha$, z.B. $\alpha = 0.05$.
- Wähle den **kritischen Wert** $\xi$, so dass

$$\mathbb{P}_{H_0}\{H_0 \text{ ist verworfen}\} = \mathbb{P}_{H_0}\{|S - n/2| \geq \xi\} \leq \alpha$$

- Die **Normalapproximation** impliziert $\xi = 31$. Demzufolge gilt

$$\mathbb{P}_{H_0}\{|S - 500| \leq 31\} \approx 0.95$$

- Somit ist im Beispiel $|S - 500| = |472 - 500| = 28 < \xi \rightsquigarrow H_0$ ist **nicht verworfen** beim Signifikanzniveau vom $5\%$.

# Methodologie der Signifikanztests

Die Hypothese $H_0\colon \theta = \theta_0$ ist zu testen anhand von einer Stichprobe $X$.

- Wir machen die folgende Schritte (bevor wir die Daten erhalten haben)
  - Wähle eine **Statistik** (auch eine **Prüfgröße**) $T\colon \mathscr{X} \to \mathbb{R}$.
  - Konstruiere eine **Familie von Verwerfungsbereichen** $\{\mathscr{R}_\xi\}_{\xi\in\mathbb{R}}$ und betrachte deren Abbildungen $T(\mathscr{R}_\xi)$. (In den einfachsten Fällen ist $T(\mathscr{R}_\xi)$ ein Interval.)
  - Wähle das **Signifikanzniveau**, d.h. die Wahrscheinlichkeit $\alpha$, dass wir $H_0$ irrtümlicherweise verwerfen.
  - Wähle den **kritischen Wert** $\xi_0$, so dass

$$\mathbb{P}\{H_0 \text{ ist irrtümlicherweise verworfen}\} = \mathbb{P}_{H_0}\{T(X) \in T(\mathscr{R}_{\xi_0})\} \approx \alpha.$$

- Sobald die Daten $x = X(\omega)$ (eine Realisierung von $X$) eingetroffen sind, machen wir die folgende Schritte:
  - Berechne $T(x)$.
  - Verwerfe $H_0$, falls $T(x) \in T(\mathscr{R}_{\xi_0})$.

# $p$-Wert

- **Idee:** Statt von einem festen Niveau $\alpha$ auszugehen und ein Konfidenzintervall für die Prüfgröße zu bestimmen, gehen wir vom beobachteten Wert der Prüfgröße $T(x)$ aus und bestimmen eine **kritische Grenze** für den Wert des Niveaus.

**Definition ($p$-Wert)**

Der **$p$-Wert** ist das kleinste Niveau, bei das die Nullhypothese gerade noch zu verwerfen ist, d.h.

$$p\text{-Wert} = \min\{\alpha : H_0 \text{ wird verworfen zum Signifikanzniveau } \alpha\}$$

**N.B.** Aus der Definition folgt: falls der $p$-Wert kleiner als das vorgegebene Niveau $\alpha_0$ ist, ist die $H_0$ zu verwerfen zum Niveau $\alpha_0$.
**Heuristik.** Falls der $p$-Wert nicht klein ist (z.B. $p$-Wert $> 0.1$), haben wir wenig Indizien um $H_0$ zu verwerfen.

# Ist mein Würfel fair?

- Die **Nullhypothese** $H_0$: $\mathbb{P}\{X = i\} = p_i = 1/6$ für $i \in [1,6] \cap N$.
- In einem Experiment aus insgesamt $n$ Versuchen erhalten wir $N_i$ mal die Zahl $i \in [1,6] \cap N$.
- Der **Verwerfungsbereich**: verwerfe $H_0$, falls

$$T(N_1,\ldots,N_n) := \sum_{i=1}^{6} \frac{(N_i - np_i)^2}{np_i} > \xi.$$

- Wähle $\xi$, so dass

$$\mathbb{P}_{H_0}\{\text{verwerfe } H_0\} = \alpha$$
$$\Leftrightarrow \mathbb{P}_{H_0}\{T(N_1,\ldots,N_6) > \xi\} = \alpha$$

- **Normalapproximation** (der ZGS, großes $n$) liefert: $T$ ist $\chi^2_6$-verteilt.

# $\chi^2$-Anpassungstest

**Frage: Passt das Modell zur Stichprobe?**

- Sei $X = (X_1, \ldots, X_n) \sim P^{\otimes n}$ eine **Stichprobe**.
- Zerteile den Stichprobenraum in $a \in \mathbb{N}$ disjunkte **Unterbereiche** $\mathscr{X}_i$. Also $\mathscr{X} = \bigcup_{i=1}^{a} \mathscr{X}_i$.
- $np_i$ ist die theoretische Anzahl von den Proben im Unterbereich $i$, wobei $p_i = P(\mathscr{X}_i)$.
- Sei $N_i := |\{k \colon X_k \in \mathscr{X}_i\}|$ die in Anzahl von den Realisierungen im Unterbereich $i$, die wir in einem Experiment erhalten haben.
- Die **Nullhypothese**: $H_0 \colon X \sim P^{\otimes N}$.
- Der **Verwerfungsbereich**: verwerfe $H_0$, falls

$$T(N_1, \ldots, N_n) := \sum_{i=1}^{a} \frac{(N_i - np_i)^2}{np_i} > \xi.$$

- Die Prüfgröße $T$ ist approximativ $\chi^2_{a-1}$**-verteilt**.

# Optischer Anpassungstest: QQ-Plot

**Frage: Passt das Modell zur Stichprobe?**

QQ-Plot (= Quantil-Quantil-Plot)