EST WS2024/2025

Zusammenfassung

January 24, 2025

Contents

1 Grundlagen

1.1 Grundlegende Konzepte

um eine statistische Frage wohldefiniert zu stellen benötigen wir:

- Stichprobenraum Ω
- Elementarereignisse $\omega_i \in \Omega$
- Ereignisse $A \subseteq \Omega$

1.1.1 Mengenoperationen

Wir nennen

- $A^c = \{\omega \in \Omega : \omega \notin A\}$
  - das Komplement von A
  - Nicht-A
- $A \cup B = \{\omega \in \Omega : \omega \in A \text{ oder } \omega \in B \text{ oder } (\omega \in A \text{ und } \omega \in B)\}$
  - (= A ODER B)
  - union
- $A \cap B = \{\omega \in \Omega : \omega \in A \text{ und } \omega \in B\}$
  - den Durchschnitt von A und B
  - (= A UND B)
  - intersection
- $A - B = \{\omega \in \Omega : \omega \in A, \omega \notin B\}$
  - die Differenz von A und B
- Wenn jedes $\omega \in A$ auch in B ist, dann schreiben wir $A \subseteq B$
  - A ist eine Teilmenge von B
- Die Anzahl der Elemente in A schreiben wir als $|A|$
- Wir nennen zwei Ereignisse disjunkt, wenn $A \cap B = \emptyset$

1.1.2 Definition

Eine Funktion P die jedem Ereignis A eine reellwertige Zahl $P(A)$ zuweist heißt Wahrscheinlichkeitsmaß wenn es die folgenden (Kolmogorov-)Axiome erfüllt:

Kolmogorov Axiome

1. $P(A) \geq 0$, für jedes A
   - Wahrscheinlichkeiten sind nicht negativ
2. $P(\Omega) = 1$
   - Ein Ereignis, das dem kompletten Stichprobenraum umfasst, hat die Wahrscheinlichkeit 1
3. Wenn $A_1, A_2, \ldots$ disjunkt sind dann gilt $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$
   - Die Wahrscheinlichkeit der Vereinigung von Ereignissen ist die Summe der Wahrscheinlichkeiten

Folgerungen Aus den Axiomen

- $P(\emptyset) = 0$
- $A \subset B \implies P(A) \leq P(B)$
- $P(A^c) = 1 - P(A)$
- $A \cap B = \emptyset \implies P(A \cup B) = P(A) + P(B)$
  - generell gilt (siehe Venn Diagramm):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Bedingte Wahrscheinlichkeit (A gegeben B) (siehe Venn Diagramm):

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

- anders herum:

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

- $\implies P(A, B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$
  - **Bayes Regel**

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

Beispiel Corona Schnelltest (ST) soll validiert werden, indem wir mit einem PCR-Test vergleichen.

	# PCR positiv	# PCR negativ
# ST positiv	99	50
# ST negativ	1	9850

- Sensitivität (true positive rate): $P(ST+|PCR+) = \frac{99}{100} = 0.99$
- Spezifizität (true negative rate): $P(ST-|PCR-) = \frac{9850}{9900} \approx 0.995$
- Frage: $P(PCR+|ST+) =?$

$$\frac{P(ST+|PCR+) \cdot P(PCR+)}{P(ST+)} = \frac{0.99 \cdot \frac{100}{10000}}{\frac{149}{10000}} = \frac{0.0099}{0.0149} \approx 0.664$$

1.2 Kombinatorik

1. Wie viele Möglichkeiten gibt es n Objekte k mal zu kombinieren? (Ziehen mit zurücklegen)

$$n^k$$

2. Wie viele Möglichkeiten gibt es n Objekte anzuordnen (Ziehen ohne Zurücklegen)?

$$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 1 = n!$$

3. Wie viele Möglichkeiten gibt es k Objekte aus n Objekten auszuwählen? (n choose k)

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)}{k!} = \frac{(n \cdot (n-1) \cdot \ldots \cdot 1)}{((n-k) \cdot (n-k-1) \cdot \ldots \cdot 1) \cdot k!} = \frac{n!}{k! \cdot (n-k)!}$$

2 Zufall

2.1 Zufallsvariable

2.1.1 Definition

- Eine Zufallsvariable ist eine Funktion $X : \Omega \to \mathbb{R}$ die jedem Elementarereignis $\omega \in \Omega$ eine reellwertige Zahl $X(\omega)$ zuweist.
- Wir beschreiben das Verhalten einer Zufallsvariable durch deren kumulative Verteilungsfunktion $F_X(x) = P(X \leq x)$

2.2 Verteilungsfunktion

- Eine Zufallsvariable kann diskret oder stetig/kontinuierlich sein

2.2.1 Diskret

- **diskret**: wenn die ZV abzählbar viele Werte annehmen kann
  - In diesem Fall (diskret) definieren wir die sog. *Wahrscheinlichkeitsfunktion*: $f_X(x) = P(X = x)$, wir fordern, $\sum f_X(x) = 1 \quad x \in \chi$ (oder $x \in X(\omega)$)
- Wir schreiben $X \sim f_X$: "ZV X folgt der Verteilung f_X"
  - Beobachtungen von X werden gemäß der Wahrscheinlichkeiten in f_X generiert
- Wie erhalten wir f_X?
  1. Aus Beobachtungen schätzen
  2. Aus Vorwissen konstruieren, z.B. Lotto
  3. Beobachtungen und Vorwissen vereinen: Aus Beobachtungen die Parameter einer angenommenen Verteilung schätzen

diskrete Verteilungen

Punktmasseverteilung

- $X \sim \delta_a,\ P(X = a) = 1$
- kumulierte Verteilungsfunktion: $F(x) = \begin{cases} 0 & x < a \\ 1 & x \geq a \end{cases}$
- Wahrscheinlichkeitsfunktion: $f(x) = \begin{cases} 1 & x = a \\ 0 & \text{sonst} \end{cases}$

Gleichverteilung:

- Für $x \in K = \{x_1, \ldots, x_k\}, k \in \mathbb{N}$
- $F(x) = \frac{|\{z \in K; z \leq x\}|}{|K|}$
- $f(x) = \begin{cases} \frac{1}{k} & x \leq k \\ 0 & \text{sonst} \end{cases}$

Bernoulli-Verteilung

- $x \in \{0, 1\}, P(X = 1) = p, P(X = 0) = 1 - p$
- $f_X(x) = p^x(1 - p)^{1-x}$
  - $f(1) = p^1(1 - p)^0 = 0$
  - $f(0) = p^0(1 - p)^1 = |1 - p|$

Binomial-Verteilung

- Anzahl der Erfolge wiederholter (unabhängiger) Bernoulli Experimente
  - Jeder Versuch hat dieselbe Wahrscheinlichkeit für Erfolg/Misserfolg
- Sei $X \sim \text{Bernoulli}(p)$
- Ziehen sie n Beobachtungen von X und nennen sie y die Summe dieser Beobachtungen
- $y \sim \text{Binomial}(n, p)$
  - $f_Y(k) = \binom{n}{k} p^k(1 - p)^{n-k}$
  - $F_Y(k) = \sum_{i=0}^{k} \binom{n}{i} p^i(1 - p)^{n-i}$
  - $\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$ Binomialkoeffizient
- $p \cdot p \cdot \cdots \cdot p = p^k$: Wahrscheinlichkeit für k Erfolge
- $(1 - p) \cdot (1 - p) \cdot \cdots \cdot (1 - p) = (1 - p)^{n-k}$: Wahrscheinlichkeiten für $n - k$ Misserfolge

Geometrische Verteilung

- $X \sim \text{Geo}(p)$
- X : Anzahl von Bernoulli(p)-Experimenten bis zum ersten Erfolg
- $P(X = k) = f_X(k) = (1 - p)^{k-1} \cdot p$

Poisson Verteilung

- $x \in \{0, 1, 2, 3, \ldots\}$ kann unendlich viele diskrete Werte annehmen
- $X \ldots$ Anzahl der Erfolge/Ereignissen in einer Zeitperiode
- $X \sim Pois(\lambda)$
- $\lambda \ldots$ Zeit zwischen 2 Erfolgen/Ereignissen
- $P(X = k) = f_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$

2.2.2 Kontinuierlich

- Im kontinuierlichen Fall werden immer Intervalle betrachtet, da $P(X = c) = 0$ (einzelne Werte haben Wahrscheinlichkeit 0)
- **kontinuierlich**: Wir nennen eine Zufallsvariable X kontinuierlich, wenn es eine *Wahrscheinlichkeitsdichtefunktion (PDF)* f_X mit den folgenden Eigenschaften gibt:
  1. $f_X(x) \geq 0$

2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. $P(a < x < b) = \int_a^b f_X(x)dx$ (mit $a \leq b$))

- Anmerkung: $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ bzw. $f_X = F'(x)$ für alle Punkte an denen F_X diff. ist.
- Generell definieren wir den Erwartungswert einer Zufallsvariable X:

$$\mu_X := E(x) = \int_{-\infty}^{\infty} x \cdot dF(x) = \begin{cases} \sum x \cdot f_X(x) & \text{für X diskret} \\ \int_{-\infty}^{\infty} x \cdot f_X(x)dx & \text{für X kont. wenn F diff'bar} \end{cases}$$

- $E(X)$ ist linear: $E(aX) = aE(X); E(a + bX) = a + bE(X)$
- und die Varianz

$$\sigma^2 := E((X - \mu)^2) = \int_{-\infty}^{\infty} (X - \mu)^2 df(x)$$

- $Var(x) = E(X^2) - E(X)^2$
  - $E(X^2) = \sum x^2 f_X(x)$

kontinuierliche Verteilungen

Gauss Verteilung

- Wir nennen X normalverteilt (oder Gauss-verteilt) mit Parametern μ (Mittelwert) und σ (Standardabweichung) (bzw σ^2, Varianz) wenn

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

- In diesem Fall schreiben wir $X \sim N(\mu, \sigma^2)$
- Wenn $X \sim N(\mu, \sigma^2)$ nennen wir $Z = \frac{X-\mu}{\sigma} \sim N(0, 1) =: \emptyset(x)$ eine Zufallsvariable mit standardisierter Normalverteilung
- Was macht eine Normalverteilung "normal"?
  - Beispiel
    - Betrachten wir einen Menge U_i von statistisch unabhängigen, diskreten Zufallsvariablen mit $u_i \in \{-1, 0, 1\}$ und $P(U_i = .) = \frac{1}{3}$
    - Definiere $U^k = \sum_{i=1}^{k} U_i$, z.b. $U^2 = U_1 + U_2$ Berechne $P(U^2)$
      - $P(U^2)$ kann die Werte $\{-2, -1, 0, 1, 2\}$ annehmen
      - $P(U^2 = -2) = P(U_1 = -1, U_2 = -1) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$
      - gleiches für $P(U^2 = 2)$
      - $P(U^2 = -1) = P(U^2 = 1) = \frac{2}{9}$
      - $P(U^2 = 0) = \frac{3}{9}$
    - Je mehr Zufallsvariablen addiert werden, desto breiter und höher wird die Verteilung.
    - Übergang zu den kontinuierlichen Zufallsvariablen $U_i \sim$ Gleichverteilt$(-\frac{a}{2}, \frac{a}{2})$
      - $U^2 = U_1 + U_2$ kleinster Wert: $-a$, größter Wert: a
      - Geometrisch: Eine der Verteilungen "flippen" und übereinanderlegen
      - Ergebniss: Trapezverteilung
  - Generell gilt für $Z = X + Y$:

$$f_Z = f_X * f_Y = \int_{-\infty}^{\infty} f_X(t) \cdot f_Y(z - t)dt$$

    - "Faltung"
  - Frage: "Welche Verteilung ergibt mit sich selbst gefaltet wieder sich selbst"
    - Normalverteilung

* Wenn Zufallsverteilungen oft miteinander gefaltet werden kommt irgendwann eine Normalverteilung raus.
  · Bsp. Körpergrößen
* Wähle $f_X(x) = N(0, \sigma^2)$:

$$f_{X+X}(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2} t^2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}(z-t)^2 dt = \cdots = \frac{1}{\sqrt{2\pi 2\sigma^2}} \exp -\frac{1}{2\sigma^2} z^2 = N(0, 2\sigma^2)$$

* Die Normalverteilung ist ein stabiler Punkt der Operation $f_X * f_X$.

- Zentrale Grenzwertsatz (ZGS) - "central limit theorem"
  - Für eine Sequenz von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots . X_n$ mit Erwartungswert μ und Varianz σ^2 konvergiert die Verteilung von

$$\sqrt{n} \cdot (\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu))$$

  gegen $N(0, \sigma^2)$ für $n \to \infty$.
    * Anmerkung: Die Variante des ZGS von Lyapunov generalisiert den ZGS für beliebige (unterschiedliche) Verteilungen von $X_1, \ldots, X_n$ (unter bestimmten Annahmen über deren Erwartungswerte und Varianzen).

Exponential Verteilung

- $X \sim Exp(\lambda)$
- Analog zur Poisson Verteilung im diskreten Fall
- bsp. $X \ldots$ Lebensdauer eines Bauteils in Stunden
- $\lambda \ldots$ Anzahl der Erfolge innerhalb einer Zeitperiode
  - bsp. $X \sim Exp(\frac{1}{500})$: in einer Stunde passieren $\frac{1}{500}$ Ereignisse ($\frac{1}{500}$ Bauteile gehen kaputt)
- $f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$
- $E(X) = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$

Multivariate Verteilungen Die wichtigste Mulivariate Verteilung ist die multivariate Normalverteilung

$$N(\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

mit den Parametern $\mu = [\mu_1, \ldots, \mu_k]^T$ und $\Sigma \in \mathbb{R}^{k \times k}$ mit den Einträgen

$$\sigma_{ij} = E((x_i - \mu_i) \cdot (x_j - \mu_j)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) dF(x_i, x_j)$$

$w^T \Sigma w \geq 0 \; \forall w \in \mathbb{R}^k$ (Σ muss positiv semidefinit sein)

- Σ und Σ^{-1} sind symmetrisch
- lineare Transformationen von multivariaten normalverteilten ZV sind multivariat normalverteilt.
- Die Korrelation von X und Y ist gegeben durch

$$\rho(X, Y) = \frac{\sigma xy}{\sigma x \sigma y} \in [-1, 1]$$

- Die Elemente der inversen Kovarianzmatrix Σ repräsentieren die partielle Kovarianz $\sigma xy|\chi$ $\{x, y\}$

3 Schätzen

Wie schätzen wir Parameter ϕ einer Verteilung aus Beobachtungen $S = \{x_1, \ldots, x_m\}$ einer ZV X?

$$\hat{\phi}^* = \max_{\phi} P(\phi|S)$$

- Bayes Regel:
$$P(\phi|S) = \frac{P(S|\phi) \cdot P(\phi)}{P(S)}$$
  - $P(\phi)$... subjektiver *Prior* (wenn bestimmte Werte für ϕ bevorzugt werden; a priori)
  - $P(S|\phi)$... Unter Annahme einer Verteilung berechenbar!
  - $P(S)$... Wahrscheinlichkeit von S --> unbekannt!

3.1 Maximum-Likelihood Estimation

- Für Beobachtungen $x_1, \ldots, x_m$ die unabhängig aus der gleichen Verteilung f_X gezogen sind (i.i.d Annahme) wählen wir die Parameter, die die gemeinsame Wahrscheinlichkeit aller Beobachtungen maximieren:
$$\underline{\hat{\phi}} = \max_{\underline{\phi}} \prod_{i=1}^{m} f_X(x_i; \underline{\phi})$$

3.1.1 Beispiel

$X \sim N(\mu, \sigma^2)$; $S = \{x_1, \ldots, x_m\}; \hat{\mu} =?, \hat{\sigma}^2 =?$ $\hat{\mu} = \max_\mu \prod_{i=1}^m N(x_i; \mu, \sigma^2) = \max_\mu \prod_{i=1}^m \frac{1}{2\pi\sigma^2} \exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\}$

- Ableiten, da Maximum genommen werden soll: Um Produkt wegzubekommen --> Logarithmus nehmen
  $\max_\mu \sum_{i=1}^{m} \log \frac{1}{2\pi\sigma^2} - \sum_{i=1}^{m} \frac{1}{2\sigma^2}(x_i - \mu)^2 \ \frac{\delta}{\delta\mu} g(\mu)$
  $= -\frac{1}{2\sigma^2} \sum_{i=1}^{m} 2(x_i - \mu)^2 \cdot (-1) = 0$
  $\Leftrightarrow \sum_{i=1}^{m} x_i - \sum_{i=1}^{m} \mu = 0$
  $\Leftrightarrow \hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x_i$
- Das gleiche für $\hat{\sigma}$:
  $\hat{\sigma} = \max_\sigma \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\}$
  $= \max_\sigma -\frac{m}{2} \log 2\pi - \frac{m}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \mu)^2 \ \frac{\delta}{\delta\sigma} g'(\sigma)$
  $= \frac{-m}{2} \frac{1}{\sigma^2} 2\sigma + \frac{1}{\sigma^3} \sum_{i=1}^{m} (x_i - \mu)^2 = 0$
  $\Leftrightarrow -m \cdot \sigma^2 + \sum_{i=1}^{m} (x_i \mu)^2 = 0$
  $\Leftrightarrow \sigma = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2}$
- Stichprobenvarianz
  - Ziel 1: Der Erwartungswert eines Schätzers soll der zu schätzende Parameter sein!
    $E(\hat{\mu}) = \int_{-\infty}^{\infty} \frac{1}{m} \sum_{i=1}^{m} x_i \cdot f_X(x) dx$
    $= \frac{1}{m} \sum_{i=1}^{m} \int_{-\infty}^{\infty} x_i \cdot f_X(x) dx$
    $= \frac{1}{m} \sum_{i=1}^{m} \mu = \frac{1}{m} \cdot m \cdot \mu = \mu$ ✓
    $E(\hat{\sigma}^2) = \int_{-\infty}^{\infty} \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2 \cdot f_X(x) dx$
    $= \frac{1}{m} \sum_{i=1}^{m} \int_{-\infty}^{\infty} (x_i - \mu)^2 f_X(x) dx$
    $= \frac{1}{m} \sum_{i=1}^{m} \sigma^2 = \sigma^2$ ✓
    $\Rightarrow \hat{\mu}$ und $\hat{\sigma}$ sind unverzerrte Schätzer!

∗ unverzerrter Schätzer: Bessel Korrektur (Kann mit geschätzem Erwartungswert $\hat{\mu}$ statt tatsächlichem μ berechnet werden)

$$\hat{\sigma}'^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \hat{\mu})^2$$

– Ziel 2: Der Schätzer soll eine geringe Varianz haben Für die Abschätzung der Varianz nutzen wir die Markov-Ungleichung Nun wählen wir $Z = |\hat{\mu_x} - \mu_x| = |\frac{1}{m}\sum_{i=1}^{n} X_i - \mu_x|$:

$$P(|\hat{\mu} - \mu| \geq \epsilon) = P((\hat{\mu} - \mu)^2 \geq \epsilon^2) = \frac{E((\hat{\mu} - \mu)^2)}{\epsilon^2}$$

$$= \frac{1}{\epsilon^2} E\left(\left(\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right)^2\right)$$

$$= \frac{1}{\epsilon^2} E\left(\left(\frac{1}{m}\sum_{i=1}^{m} (X_i - \mu)\right)^2\right)$$

$$= \frac{1}{\epsilon^2 m^2} E\left(\left[(X_1 - \mu) + (X_2 - \mu) + \cdots + (X_m - \mu)\right]^2\right)$$

$$= \frac{1}{\epsilon^2 m^2} E\left([(X_1 - \mu)(X_1 - \mu) + (X_1 - \mu)(X_2 - \mu) + \cdots + (X_m - \mu)(X_m - \mu)]\right)$$

Alle nicht quadratischen Terme sind kovariant und gleichen sich gegenseitig aus.

$$= \frac{1}{\epsilon^2 m^2} E\left(\left[(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_m - \mu)^2\right]\right)$$

$$= \frac{1}{\epsilon^2 m^2} E\left(\sum_{i=1}^{m}(X_i - \mu)^2\right) = \frac{1}{\epsilon^2 m^2}\sum_{i=1}^{m} E\left((X_i - \mu)^2\right) = \frac{\sigma^2 m}{\epsilon^2 m^2}$$

$$\Leftrightarrow P(|\hat{\mu}_x - \mu_x| \geq \epsilon) \leq \frac{\sigma_x^2}{\epsilon^2 m}$$

Beachte: Aufgrund des ZGW Zentraler Grenzwertsatz ist $\hat{\mu}$ annährend normalverteilt! Schätzer für spezifische Parameter einer Verteilung oder Eigenschaften einer ZV nennt man **Punktschätzer**

3.1.2 i.i.d Annahme

- independent and identically distributed
- Wann immer man Datenpunkte hat, nimmt man an, dass diese aus der Gleichen Verteilung stammen.
  – Sonst ist das schätzen nicht möglich
- Man nimmt ebenso an, dass die Beobachtungen unabhängig sind.
- Achtung: Regression to the mean - Extremwerte sind unwahrscheinlich, aber nicht weil vorher schon Extremwerte beobachtet wurden sind (Annahme, alle Beobachtungen sind unabhängig)

3.1.3 Markov-Ungleichung

Sei Z eine ZV mit x eine Realisierung von Z und $x \geq 0$. Dann gilt

$$E(x) = \int_0^\infty x \cdot f_Z(x)dx = \int_0^\epsilon x \cdot f_Z(x)dx + \int_\epsilon^\infty x \cdot f_Z(x)dx \geq \int_\epsilon^\infty x \cdot f_Z(x)dx$$

$$\geq \int_\epsilon^\infty \epsilon f_Z(x)dx = \epsilon \int_\epsilon^\infty f_Z(x)dx = \epsilon \cdot P(Z \geq \epsilon)$$

$$\Leftrightarrow P(Z \geq \epsilon) \leq \frac{E(Z)}{\epsilon}$$

- Der erste Teil $\int_0^\epsilon x \cdot f_Z(x)dx$ ist ≥ 0, daher ist die Summe $\geq$ dem zweiten Teil der Summe

3.2 Lineare Regression

Betrachte ZV Y und X mit der folgenden Beziehung:

$$Y = \beta_0 + \beta_1 X + \epsilon;\ \epsilon \sim N(0, \sigma^2)$$

ϵ ist ein Störfaktor, da die Gerade nicht immer alle Punkte umfassen kann. Schätze β_0, β_1 aus $S = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$

$$\epsilon = (Y - \beta_0 - \beta_1 X) \sim N(0, \sigma^2)$$

$$\beta_0^*, \beta_1^* = \max_{\beta_0, \beta_1} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 \cdot X_i)^2\right\}$$

Schätzen mit Maximum Likelidhood Estimation

- Mit Logarithmus vereinfachen

$$\max_{\beta_0, \beta_1} \log \frac{m}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{m} \frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 \cdot X_i)^2$$

- konstante Terme können vernachlässigt werden, da sie die Funktion lediglich skalieren und nicht die Extrempunkt verändern
- Der negative Term wird umgedreht und es wird das Minimum gesucht

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{m} (Y_i - \beta_0 - \beta_1 \cdot X_1)^2$$

- Wir wollen den quadratischen Fehler minimieren

$$\underline{\beta} := \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix};\quad \underline{x_i} =^l \begin{bmatrix}1 \\ x_i\end{bmatrix}$$

$$\underline{\beta^*} = \min_{\underline{\beta}} \left(\sum_{i=1}^{m} (Y_i - \underline{\beta^T}\ \underline{x_i})^2 = \sum_{i=1}^{m} (Y_i - [\beta_0, \beta_1] \begin{bmatrix}1 \\ x_i\end{bmatrix})^2 \right)$$

$$\frac{\delta}{\delta \underline{\beta}} \sum_{i=1}^{m} (Y_i - \underline{\beta}^T\ \underline{x_i}) = \underline{0}$$

$$\sum_{i=1}^{m} 2(Y_i - \underline{\beta}^T\ \underline{x_i}) \cdot \underline{x_i^T} = \underline{0}$$

$$\sum_{i=1}^{m} Y_i \cdot \underline{x_i^T} - \underline{\beta}^T \sum_{i=1}^{m} \underline{x_i} \cdot \underline{x_i^T} = \underline{0}$$

$$m\underline{\hat{\sigma}}(Y, \underline{X}) - \underline{\beta}^T m \cdot \underline{\hat{\sigma}}(\underline{X}, \underline{X}) = 0$$

$$\underline{\beta} = \underline{\hat{\sigma}}(\underline{X}, \underline{X})^{-1} \cdot \underline{\hat{\sigma}}(Y, \underline{X})$$

- Mit

$$\sigma(\underline{X}, \underline{X}) = \begin{bmatrix}1 & \hat{\mu}x \\ \hat{\mu}_x & \sigma^2(x)\end{bmatrix} \text{ und } \underline{\hat{\sigma}}(Y, X) = \begin{bmatrix}\hat{\mu}_y \\ \hat{\sigma}(y, x)\end{bmatrix}$$

- Dieser Term entspricht der Stichproben-Kovarianz wenn $\mu x = \mu y = 0$

4 Hypothesentests

4.1 Hypothesen

Wir haben zwei Beobachtungen $S_X = \{X_1, \ldots, X_m\}$ mit $X_i \sim f_X$ und $S_Y = \{Y_1, \ldots, Y_n\}$ mit $Y_i \sim f_Y$. Wir möchten untersuchten, ob sich die ZV X und Y bzgl. eines Merkmals unterscheiden (z.B. Ein Impfstoff wirkt besser als ein anderer) welches wir durch die Teststatistik T abbilden, z.B. $T(S_X, S_Y) = |\hat{\mu}_X - \hat{\mu}_Y|$

Wir formulieren unsere sog. **Nullhypothese**:

$$H_0 = T(.) = 0$$

Das strikte Gegenteil der Nullhypothese ist die **Alternativhypothese**:

$$H_1 = T(.) \neq 0$$

Da Abweichungen in beide Seiten gleich betrachtet werden, handelt es sich um einen Zweiseitigen Test.

- Man könnte auch einen Einseitigen Test konstruieren, indem man den absoluten Wert weglassen würde, und anstatt $T(.) = 0$ für die Nullhypothese etwas wie $T(.) < 0$ nehmen
- Mit diesem Test kann also nur bestimmt werden ob die Ergebnisse gleich oder ungleich sind, nicht welcher von beiden *bessser* ist

Sei $T(S_X, S_Y) = c$. Was ist $P(H_0|T(.) = c)$?

$$P(H_0|T) = \frac{P(T|H_0) \cdot P(H_0)}{P(T)}$$

- wir verwenden Bayes
- $P(H_0)$ und $P(T)$ sind unbekannt, es ist also prinzipiell nicht berechenbar
  - $P(H_0)$ ist die a priori Wahrscheinlichkeit
  - $P(T)$ ist die Wahrscheinlichkeit, das die Daten die Gesamtmenge genau repräsentieren
- $P(T|H_0)$ kann berechnet werden

Welche Fehlerarten gibt es beim Hypothesen testen?

	H0 akzeptiert	H0 verworfen
H0 wahr	+	alpha / Typ I Fehler
H0 falsch	beta / Typ II Fehler	+

Wir nennen $P(T(.) \geq c|H_0)$ den p-Wert. Für das Signifikanzniveau α (typischerweise $\alpha = 0.05$) verwerfen wir die Nullhypothese wenn $p \leq \alpha$

4.2 Nullverteilung

Wie berechnen wir $P(T(.) \geq c|H_0)$?

- Hierzu benötigen wir die Nullverteilung $f_{T|H_0}$
- In der Klassischen Statistik wird diese unter Annahmen über den daten-generierenden Prozess hergeleitet (-> Wald-Test, t-Test, χ^2-Test, etc.)
- In der Praxis benutzen wir meist **Permutations-Methoden**

4.2.1 Permutationstests

Idee: verschiedene Gruppen durch Permutation der Daten bilden, unter der Annahme der Nullhypothese (kein unterschied in den Daten) Den Prozess wiederholt man viele Male mit zufälligen Permutationen. Alle Ergebnisse sind Punkte in der Nullverteilung. Durch Zählen kann eine Annahme getroffen werden.

Mittelwert Sei $S_z = \{z_1, \ldots, z_{m+n}\} = \{x_1, \ldots, x_n, y_1, \ldots, y_m\}$

1. Berechne $T(z) = |\frac{1}{m} \sum_{i=1}^{m} z_i - \frac{1}{n} \sum_{i=m+1}^{m+n} z_i| = c$
2. Generiere eine zufällige Permutation von S_z, z.b.

$$S_{z'} = \{z_{10}, z_3, z_1, z_m, z_{23}, \ldots\}$$

3. Berechne $T_{z'}$ (stellt einen Wert aus $f_{T|H_0}$ dar)
4. Wiederhole 2. und 3. k-mal, $T_{z'}^{(1)}, \ldots, T_{z'}^{(k)}$.
5. Schätze den p-Wert als

$$\hat{p} = \frac{1}{k} \sum_{i=1}^{k} (T_{z'}^{i} \geq c)$$

Der Test auf Mittelwerte, wie hier beschrieben wird in der Literatur als Student's t-test bezeichnet.

ANOVA Neben dem t-test auf Mittelwerte wird in der Analyse empirischer Daten häufig die *ANOVA - Analysis of Variance* eingesetzt:

Wir betrachten eine Zufalssvariable X die wir unter $c \in \{1, \dots, k\}$ unterschiedlichen Bedingungen beobachten:

- H_0: $\mu_{X|c=1} = \cdots = \mu_{X|c=k}$
- H_1: $\exists i, j \in \{1, \dots, k\} : \mu_{X|c=i} \neq \mu_{X|c=j}$ Als Teststatistik wählen wir die f-Teststatistik

$$f := \frac{\sum_{i=1}^{k} \left(\hat{\mu}_{X|c=i} - \sum_{j=1}^{k} \hat{\mu}_{X|c=j} \right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{i|c=i} - \hat{\mu}_{X|c=j} \right)^2}$$

wobei n_i die Anzahl der beobachteten Werte von X für die Klasse $c = i$ darstellt
(Im Zähler steht die Varianz der Mittelwerte aller Bedingungen, und im Nenner steht die Varianz aller Werte)

5 Kausalität

- Frage: Wie entstehen statistische Abhängigkeiten durch Ursache-Wirkung Beziehungen
  - Grundlage für Untersuchungen welche Effekte Handlungen, auch Interventionen gennannt, haben
- Überlegungen gehen bis auf Aristoteles zurück
- in der modernen Statistik haben sich zwei Konzepte durchgesetzt:
  - Potential Outcomes Framework
  - Kausale Strukturgleichungen

5.1 Reichenbach Prinzip

- Notwendigkeit einer kausalen Betrachtung von statistischen Abhängigkeiten
- Eine statistische Abhängigkeit zwischen zwei Zufallsvariablen X und Y ($P(X,Y) \neq P(X) \cdot P(Y) \Leftrightarrow X \not\perp Y$ kann durch drei Ursache-Wirkung Beziehungen entstehen:
  1. X ist Ursache von Y ($X \rightarrow Y$)
  2. Y ist Ursache von X ($Y \rightarrow X$)
  3. X und Y haben eine gemeinsame Ursache Z ($X \leftarrow Z \rightarrow Y$)

5.1.1 Simpson Paradoxon

Beispiel: wir vergleichen zwei Behandlungen (OP-A und OP-B) für die Entfernung von großen oder kleinen Nierensteinen

Große Nierensteine

OP	#	#erfolgreich	%
A	24	12	50%
B	52	28	53,8%

$\Longrightarrow$ B > A!

Kleine Nierensteine

OP	#	#erfolgreich	%
A	76	56	73,7%
B	48	38	79,2%

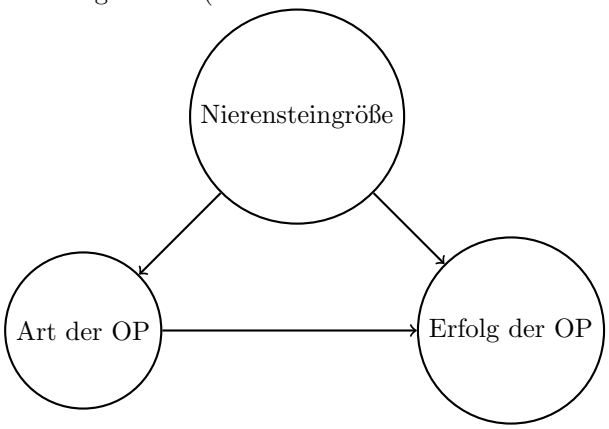
$\Longrightarrow$ B > A!

Alle Nierensteine			
OP	#	#erfolgreich	%
A	100	68	68%
B	100	66	66%
$\Longrightarrow$ A > B!			

Wie kann es sein, dass das Ergebnis unterschiedlich zu den Untergruppen ist?

- OP-A wurde öfter bei kleinen Nierensteinen angewendet (kleine Nierensteine sind generell leichter zu operieren)

Nierensteingröße → Erfolg # Der Erfolg hängt von der Größe des Steins ab Art der OP → Erfolg # Der Erfolg hängt von der Art der OP ab Nierensteingröße → Art der OP # Die Art der OP hängt auch von der Nierensteingröße ab (Ärzte wählen andere OP für andere Größen)



5.2 Frameworks

5.2.1 Potential Outcomes Framework

- Besteht aus
  - einer Menge von Einheiten ("units") $u \in U$ (z.b. Patienten)
  - einer Zuweisungsfunktion jeder Einheit zu euner Behandlungs- oder Kontrollgruppe ("treatment" - t vs. "control" - c) $S : U \to \{t, c\}$ z.b. neues Medikament vs. Placebo
  - einem Endpunkt ("outcome") $Y : U \times \{t, c\} \to \mathbb{R}$ z.B. Überlebensdauer
- Das grundlegende Problem der kausalen Inferenz
  - Um eine optimale Entscheidung für jede Einheit treffen zu können, wüssten wir gerne den Behandlungseffekt ("treatment effect") für jede Einheit

$$TE(u_i) = Y(u_i, t) - Y(u_i, c)$$

  - Können wir nicht beobachten, da wir nicht für dieselbe unit verschiedene Zuweisungsfunktionen verwenden können (Patient kann nicht gleichzeitig ein Medikament und ein Placebo erhalten; units werden zerstört oder verändert)
  - Um dieses Problem zu bewältigen, gibt es verschiedene Ansätze:
    - Ansatz: Wir begnügen uns damit den Erwartungswert des Behandlungseffekts zu schätzen

$$E(TE) = E(Y(u, t)) - E(Y(u, c))$$

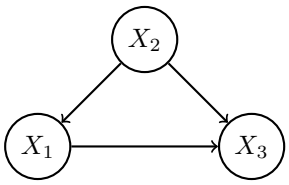
    - Wir können aber nur $E(Y(u, t)|S(u) = t)$ und $E(Y(U, c)|S(u) = c)$ messen, also die Endpunkte für die Einheiten die wir in Behandlungs- bzw. der Kontrollgruppe zuweisen (führt einen Bias ein, Simpson Paradoxon)
      - Wenn die Zuweisungsfunktion S unabhängig vom Endpunkt Y ist, dann ist $E(Y(u, t)|S(u) = t) = E(Y(u, t))$
      - Die Zuweisungsfunktion wird daher randomisiert

5.2.2 Kausale Strukturgleichung

- Eine kausale Strukturgleichung ("structural causal model"-SCM) wird beschrieben durch
  - eine Menge $\mathcal{X} = \{X_i\}_{i=1}^N$ endogener Zufallsvariablen,
  - eine Menge $\mathcal{E} = \{\epsilon_i\}_{i=1}^N$ exogener Rauschterme
  - eine Menge von Funktionen $\mathcal{F} = \{f_i\}_{i=1}^N$ mit $x_i = f_i(pa(X_i), \epsilon_i)$ und $pa(X_i) \subset \mathcal{X} \setminus X_i$ den direkten Ursachen (Eltern) von X_i relativ zu $\mathcal{X}$, wobei keine Variable eine Vorfahre von sich selbst sein darf
- Die kausalen Beziehungen von einer Strukturgleichung können in einem gerichteten azyklischen Graphen dargestellt werden
  - Graphen dürfen keine Kreise beinhalten (Gerichteter Azyklischer Graph), da sonst eine Zufallsvariable von sich selber abhängig wäre

Beispiel:
$X_1 = X_2 + \epsilon_1$
$X_2 = \epsilon_2$
$X_3 = X_1 \cdot X_2 + \epsilon_3$

$pa(X_1) = \{X_2\}$ $pa(X_2) = \{\}$
$pa(X_3) = \{X_1, X_2\}$



- Dieser Graph gibt eine kausale Faktorisierung der gemeinsamen Wahrscheinlichkeitsdichte-Funktion:

$$P(X_1, X_2, X_3) = P(X_3|X_1, X_2) \cdot P(X_1|X_2) \cdot P(X_2)$$

- Diese Faktorisierung kann für kausale Aussagen genutzt werden:

$$P(X_1, X_2, X_3|do\{X_2 = x'\}) = P(X_3|X_1, X_2 = x') \cdot P(X_1|X_2 = x')$$

  - Intervention $do\{X_2 = x'\}$: Setze X_2 zu x'