

Statistik für Umweltingenieurwesen

npikall

06.03.2025

Inhaltsverzeichnis

1 Zufallsvariablen und Verteilungen	3
1.1 Wahrscheinlichkeiten	3
1.1.1 Kolmogorov-Axiome	3
1.1.2 Eigenschaften von Wahrscheinlichkeiten	3
1.1.3 Bedingte Wahrscheinlichkeit	3
1.1.4 Satz von Bayes	3
1.2 Zufallsvariablen	4
1.3 Verteilungsfunktionen	4
1.4 Quantile	5
1.5 Erwartungswert	5
1.6 Wichtige Verteilungen	6
2 Schätzungen	6
2.1 Stichprobe und Grundgesamtheit	6
2.2 Simulation von Zufallsvariablen	6
2.3 Schätzer für Mittelwert und Varianz	6
2.3.1 Eigenschaften von Schätzern	6
2.3.2 Schätzer für Mittelwert und Varianz	7
2.4 Momentenmethode	7
2.5 Maximum Likelihood	7
2.6 Gesetz der großen Zahlen	7
2.7 Zentraler Grenzwertsatz	8
2.8 Konfidenzintervalle	8
3 Hypothesentest	8
3.1 Parametrische Hypothesen	8
3.2 Hypothesentests	9
3.3 Teststatistik	9
3.4 Typ I und Typ II Fehler	9
3.5 p-Wert	9
3.6 Hypothesentests und Konfidenzintervalle	9
3.7 Wichtige Hypothesentests	10
4 Multivariate Zufallsvariablen und Methoden	10
4.1 Bivariate Zufallsvariablen und Verteilungen	10
4.1.1 Unabhängigkeit	10
4.2 Multivariate Zufallsvariablen und Verteilungen	10
4.2.1 Kovarianz	11
4.3 Korrelation	11
4.3.1 Definition	11

4.3.2 Scheinkorrelation	11
4.4 Lineare Regression	11
4.5 ANOVA - Analysis of Variance	12
4.6 Wichtige Hypothesentests für multivariate Statistik	12
5 Nichtparametrische Methoden	12
5.1 Parametrische und nichtparametrische statistische Modelle	12
6 Experimental design	12
6.1 Grundlegende Aspekte der statistischen Versuchsplanung	12
6.2 Varianzquellen in statistischen Experimenten	13
6.3 Prinzipien der statistischen Versuchsplanung	13
6.3.1 Arten von Störgrößen	13
6.3.2 Randomisierung	13
6.3.3 Blockbildung	13
6.3.4 Statistische Kontrolle von Störfaktoren	13
6.3.5 Weitere Prinzipien	13
6.4 Typen von Stichproben	13
6.4.1 Einfache Zufallsstichprobe	13
6.4.2 Geschichtete Stichproben	13
6.4.3 Klumpenstichproben	14
6.5 Inverse Probleme der Stichprobengröße	14

1 Zufallsvariablen und Verteilungen

1.1 Wahrscheinlichkeiten

1.1.1 Kolmogorov-Axiome

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(\Omega) = 1$
3. Für eine Folge von Ereignissen A_1, A_2, \dots mit $A_i \cap A_j = \emptyset$ für $i \neq j$ gilt $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$. Die Vereinigung von Ereignissen (die sich nicht überschneiden) entspricht ihrer Summe.

1.1.2 Eigenschaften von Wahrscheinlichkeiten

- Ein Ereignis kann eintreten oder nicht:
 - $A^c = \frac{\Omega}{A}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
 - $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$
- Ein Ereignis kann nicht gleichzeitig eintreten und nicht eintreten.
 - $\mathbb{P}(\emptyset) = \mathbb{P}(A \cup A^c) = 0$
- Monotonie der Wahrscheinlichkeit für Teilereignisse:
 - $B \subseteq A \rightarrow \mathbb{P}(B) \leq \mathbb{P}(A)$
- Wahrscheinlichkeit für A oder B
 - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

1.1.3 Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit von A gegeben B beschreibt die Wahrscheinlichkeit, dass das Ereignis A eintritt, gegeben, dass B eingetreten ist (angenommen $\mathbb{P}(B) > 0$).

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad (1.1)$$

Unabhängige Ereignisse wenn gilt: $\mathbb{P}(A|B) = \mathbb{P}(A) \rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$

1.1.4 Satz von Bayes

Der Satz von Bayes ist ein wichtiges Theorem der Wahrscheinlichkeitstheorie und beschäftigt sich mit bedingten Wahrscheinlichkeiten. Er folgt aus der Definition der bedingten Wahrscheinlichkeit. Das ist mathematisch einfach, philosophisch jedoch nicht.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}$$
$$\mathbb{P}(E_i|B) = \frac{\mathbb{P}(B|E_i) \cdot \mathbb{P}(E_i)}{\sum_{i=1}^{\infty} \mathbb{P}(B|E_i) \cdot \mathbb{P}(E_i)} \quad (1.2)$$

Beispiel Medizinischer Test:

Es sei +− ein positives/negatives Testergebnis und K bzw. $\neg K$ ob eine Person tatsächlich krank ist.

- Test hat mit 90% ein positives Ergebnis bei kranken Personen
 - $\mathbb{P}(+|K) = 0.9$ (True Positive Rate **TP**)
 - ↳ $\mathbb{P}(-|K) = 0.1$ (False Negative Rate **FN**)
- Test zu 95% richtig (negatives Ergebnis) bei nicht kranken Personen
 - $\mathbb{P}(-|\neg K) = 0.95$ (True Negative Rate **TN**)
 - ↳ $\mathbb{P}(+|\neg K) = 0.05$ (False Positive Rate **FP**)
- Die Krankheit tritt bei 1% aller Menschen auf.
 - $\mathbb{P}(K) = 0.01$ (**Prior**)
 - ↳ $\mathbb{P}(\neg K) = 0.99$

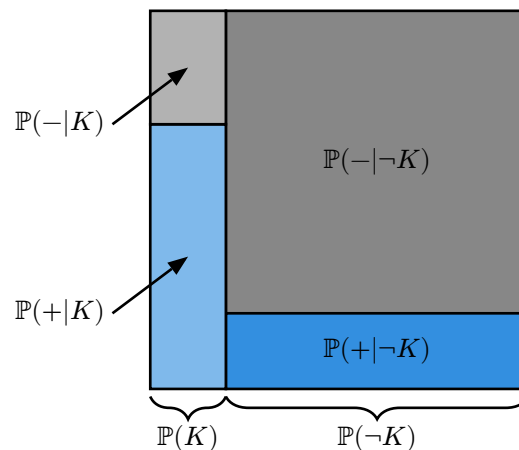
Wie wahrscheinlich ist es, dass eine Person mit positivem Test wirklich krank ist? ($\mathbb{P}(K|+)$ wird **Posterior** genannt)

$$\mathbb{P}(K|+) = \frac{\mathbb{P}(+|K) \cdot \mathbb{P}(K)}{\mathbb{P}(+)} \quad (1.3)$$

$$\begin{aligned} \mathbb{P}(+) &= \mathbb{P}(+|K) \cdot \mathbb{P}(K) + \mathbb{P}(+|\neg K) \cdot \mathbb{P}(\neg K) \\ &= 0.9 \cdot 0.01 + 0.05 \cdot 0.99 = 0.0585 \end{aligned} \quad (1.4)$$

$$\mathbb{P}(K|+) = \left(\frac{\text{TP}}{\text{TP} + \text{FP}} \right) = \frac{0.9 \cdot 0.01}{0.0585} \approx 0.154 \quad (1.5)$$

Die Wahrscheinlichkeit ist somit von 1% auf 15.4% aktualisiert worden.



Grafik 1: Darstellung des Wahrscheinlichkeitsraumes. Der linke Balken stellt die Wahrscheinlichkeit dar wenn man krank ist und ein richtiges Testergebnis bekommt. Der rechte Balken zeigt die Wahrscheinlichkeit ein Falsch Positiv zu bekommen.

Gute Erklärungen sind in folgenden Videos zu finden:

- 3B1B The medical Test Paradox
- 3B1B The quick proof of Bayes Theorem
- 3B1B Bayes-Theorem, the Geometry of changing Beliefs
- Veritasium The Bayesian Trap

1.2 Zufallsvariablen

Zufallsvariablen X ordnen Ereignissen eine Zahl zu (z.B. Kopf = 1 und Zahl = 0 oder bei Regen 20 mm).

1.3 Verteilungsfunktionen

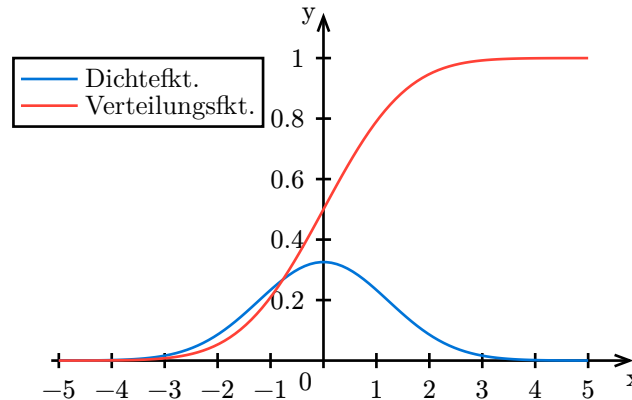
X bezeichnet immer eine Zufallsvariable und x einen möglichen Wert, den diese annimmt.

Definition: $F_X : \mathbb{R} \rightarrow [0, 1]$ ist die Verteilungsfunktion der Zufallsvariable X definiert durch

$$F_X(x) = \mathbb{P}(X \leq x) \quad (1.6)$$

Somit ist die Verteilungsfunktion, ausgewertet an x die **Unterschreitungswahrscheinlichkeit** der Zufallsvariablen X für diesen Wert. Es gilt:

$$\begin{aligned} \lim_{x \rightarrow -\infty} F_X(x) &= 0 \\ \lim_{x \rightarrow \infty} F_X(x) &= 1 \end{aligned} \quad (1.7)$$



Grafik 2: Dichte- und Verteilungsfunktion von einer kontinuierlichen Normalverteilung

- Diskrete Zufallsvariable (cdf und pmf)
- Kontinuierliche Zufallsvariable (cdf und pdf)

PDF Probability Density Function

PMF Probability Mass Function

CDF Cumulative Distribution Function

1.4 Quantile

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt
 ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aequaleam animo,
 cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infi-
 nitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

1.5 Erwartungswert

Zufallsvariablen werden durch Verteilungs- bzw. Dichtefunktionen charakterisiert, welche durch Parameter beschreiben sind. Diese Parameter hängen wiederum von den sog. Momenten (Mittelwert, Varianz) ab.

Erwartungswert:

$$\begin{aligned}\mathbb{E}(X) &= \int_{\mathcal{S}} x f_X(x) \, dx \\ \mathbb{E}(X) &= \sum_{x \in \mathcal{S}} x p_X(x)\end{aligned}\tag{1.8}$$

Daraus folgt:

$$\begin{aligned}\mathbb{E}(X + Y) &= \mathbb{E}(X) + \mathbb{E}(Y) \\ \mathbb{E}(\lambda \cdot X) &= \lambda \cdot \mathbb{E}(X)\end{aligned}\tag{1.9}$$

Varianz:

$$\text{Var}(X) = \sigma^2 = \mathbb{E}(X - \mu_X)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (1.10)$$

Variationskoeffizient:

$$\text{CV} = \frac{\sigma_X}{\mu_X} \quad (1.11)$$

Standardisierte Variable:

Sie hat einen Mittelwert von 0 und einen Varianz von 1.

$$Y = \frac{X - \mu_X}{\sigma_X} \quad (1.12)$$

1.6 Wichtige Verteilungen

- Exponentialverteilung $\text{Exp}(\lambda) = F(X) = 1 - e^{-\lambda x}$
- Normalverteilung
- Log-Normalverteilung
- Students t-Verteilung
- χ^2 - Verteilung
- F-Verteilung
- Gumbel-Verteilung
- Stetige Uniform Verteilung
- Diskrete Uniform Verteilung
- Bernoulli Verteilung
- Binomial Verteilung
- Poisson Verteilung


2 Schätzungen**2.1 Stichprobe und Grundgesamtheit**

Eine **Menge** von n Zufallsvariablen $\{X_1, \dots, X_n\}$ mit **derselben Verteilung** F_X ist eine Zufallsstichprobe der Größe n der Verteilung F_X . Das statistische Modell, welches die Stichprobe erzeugt hat, nennt man auch Grundgesamtheit, in diesem Fall einfach die Verteilung F_X .

Man kann an eine Stichprobe ein statistisches Modell anpassen. Dafür müssen wir ein Modell auswählen und die Eigenschaften des Modells aus der Stichprobe schätzen.

2.2 Simulation von Zufallsvariablen

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 mu, sigma = 0, 0.1
5 x = np.random.normal(mu, sigma, size=1000)
6
7 fig, ax = plt.subplots()
8 ax.hist(x, bins=30)
9 plt.show()
```

 Python
2.3 Schätzer für Mittelwert und Varianz

Statistische Schätzer verwenden Information aus einer Stichprobe, um auf Eigenschaften des dahinterliegenden Modells, der **Grundgesamtheit**, **zu schließen**. Das Verhalten von Schätzern ist abhängig vom zugrundeliegenden statistischen Modell der Zufallsstichprobe und dem Stichprobenumfang n . Im Allgemeinen: Je größer der Stichprobenumfang n , desto kleiner ist die mit dem Schätzverfahren verbundene Unsicherheit.

2.3.1 Eigenschaften von Schätzern

- Konsistenz (Schätzer haben eine Unsicherheit, die aber gegen 0 geht bei steigendem Stichprobenumfang).

$$\forall \varepsilon > 0 : \lim_{n \rightarrow \infty} \mathbb{P}(\|\Gamma_n - \Theta\| > \varepsilon) = 0 \quad (2.1)$$

- (asymptotische) Erwartungstreue

$$\left(\lim_{n \rightarrow \infty} \right) \mathbb{E}(\Gamma_n) = \Theta \quad (2.2)$$

2.3.2 Schätzer für Mittelwert und Varianz

Die wohl bekanntesten statistischen Schätzer sind die Schätzer für den Mittelwert \bar{x} und die Varianz σ^2 .

$$\bar{X} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.3)$$

$$S^2 = \sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.4)$$

Beide sind konsistent und erwartungstreu. Also:

$$\begin{aligned} \mathbb{E}(\bar{X}) &= \mu_x \\ \mathbb{E}(S^2) &= \text{Var}(X) \end{aligned} \quad (2.5)$$

2.4 Momentenmethode

Die Schätzung von Verteilungsparametern mit Hilfe von Schätzern für die Momente nennt man Momentenmethode. Die Intuition hinter dieser Schätzmethode ist, dass Stichprobenmomente mit den Momenten der Grundgesamtheit gleichgesetzt werden.

$$(\bar{x}, \sigma^2) \rightarrow (\xi, \alpha) \rightarrow \text{Dichtefunktion} \quad (2.6)$$

Die Momentenmethode führt zu konsistenten Schätzern für Parameter, allerdings im Allgemeinen nicht zu erwartungstreuen Schätzern (allerdings asymptotisch erwartungstreu).

2.5 Maximum Likelihood

Maximum-Likelihood-Schätzung ist wie die Momentenmethode ein Schätzverfahren für die Parameter einer Verteilung. Die Intuition ist, dass für eine gegebene Dichte (Verteilungsfamilie) die Parameter gesucht werden, welche eine Stichprobe am besten beschreiben.

$$L(\Theta, x) = \prod_{i=1}^n f_X(x_i, \Theta) \quad (2.7)$$

Jener Wert für Θ welcher die Likelihood für eine Stichprobe maximiert wird Maximum-Likelihood Schätzer $\hat{\Theta}_{\text{ML}}$ genannt. Dabei leitet man zuerst die Verteilung ab und setzt sie null.

2.6 Gesetz der großen Zahlen

Das Gesetz der großen Zahlen ist eines der wichtigsten Theoreme der Statistik.

Theorem

Sei X_1, X_2, \dots eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit $\mathbb{E}(X_i) = \mu_X < \infty$ und $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ dann gilt:

$$\bar{X}_n \rightarrow \mu_X \quad (2.8)$$

Das Gesetz der großen Zahlen besagt, dass der Durchschnitt einer großen Anzahl unabhängiger, identisch verteilter Zufallsvariablen mit hoher Wahrscheinlichkeit dem Erwartungswert der zugrunde liegenden Verteilung nahekommt.

2.7 Zentraler Grenzwertsatz

Der zentrale Grenzwertsatz ist eines der wichtigsten Theoreme der Statistik.

Theorem

Sei X_1, X_2, \dots eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit $\mathbb{E}(X_i) = \mu_X$, $\text{Var}(X_i) = \sigma_X^2 < \infty$ und $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Dann gilt:

$$\sqrt{n}(\bar{X}_n - \mu_X) \rightarrow \mathcal{N}(0, \sigma_X^2) \quad (2.9)$$

Der zentrale Grenzwertsatz besagt, dass die Summe oder der Durchschnitt vieler unabhängiger, identisch verteilter Zufallsvariablen – unabhängig von der ursprünglichen Verteilung – näherungsweise einer Normalverteilung (Glockenkurve) folgt, wenn die Anzahl der Variablen groß genug ist.

2.8 Konfidenzintervalle

Approximative Konfidenzintervalle sind eine unmittelbare Anwendung des zentralen Grenzwertsatzes. Ein Konfidenzintervall sind Funktionen $T_1(X_1, \dots, X_n)$, $T_2(X_1, \dots, X_n)$ der Zufallsstichprobe, sodass für den Parameter Θ zum Konfidenzniveau $1 - \alpha$ gilt:

$$\mathbb{P}(T_1 < \Theta < T_2) \geq 1 - \alpha \quad (2.10)$$

Konfidenzintervalle schließen den zu schätzenden Parameter zwischen zwei zufälligen Werten (Intervall) ein.

q	0.75	0.8	0.9	0.95	0.975	0.99	0.995
z_q	0.674	0.841	1.281	1.644	1.959	2.326	2.575

Beispiel

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

3 Hypothesentest

Hypothesentests bieten einen formalen Rahmen zur Entscheidungsfindung bei statistischen Fragestellungen.

3.1 Parametrische Hypothesen

Parametrische Hypothesen sind Annahmen über unbekannte Parameter Θ einer Verteilung F_Θ . Dabei wird eine Nullhypothese $\mathcal{H}_0 : \Theta \in \Theta_0$ gegen eine Alternativhypothese $\mathcal{H}_1 : \Theta \in \Theta_1$ getestet, wobei Θ_0 und Θ_1 disjunkt sind.

Üblicherweise spiegelt \mathcal{H}_0 den *Status quo* bzw. *Gleichheit* wider, und \mathcal{H}_1 repräsentiert *Veränderung*. Enthält Θ_0 nur einen Punkt, so nennt man die Hypothese *einfach*, sonst *zusammengesetzt*. Weiters unterscheidet man zwischen einseitigen und zweiseitigen Hypothesen. Ein einfaches Beispiel:

$$\begin{aligned}
\mathcal{H}_0 : \Theta &= \Theta_0 \\
\mathcal{H}_1 : \Theta &< \Theta_0 \quad \text{einseitig} \\
&\text{oder} \\
\mathcal{H}_1 : \Theta &\neq \Theta_0 \quad \text{zweiseitig}
\end{aligned} \tag{3.1}$$

3.2 Hypothesentests

Gegeben sei eine Stichprobe x , die aus einer Verteilung F_Θ mit unbekanntem Parameter Θ stammt. Basierend auf den Hypothesen \mathcal{H}_0 und \mathcal{H}_1 soll entschieden werden, ob \mathcal{H}_0 angelehnt oder beibehalten wird. Dazu wird eine kritische region C definiert, die den Ergebnisraum in zwei disjunkte Mengen unterteilt:

- $x \in C \rightarrow$ Ablehnung von \mathcal{H}_0
- $x \in C^c \rightarrow$ Beibehaltung von \mathcal{H}_0

Diese Entscheidung erfolgt mithilfe einer Teststatistik.

3.3 Teststatistik

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aequi doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

3.4 Typ I und Typ II Fehler

Entscheidung	\mathcal{H}_0 richtig	\mathcal{H}_1 richtig
\mathcal{H}_0 ablehnen	Typ I Fehler	korrekt
\mathcal{H}_0 nicht ablehnen	korrekt	Typ II Fehler

3.5 p-Wert

Der p-Wert ist das kleinste Signifikanzniveau, bei welchem die Nullhypothese für die gegebenen Daten abgelehnt werden würde.

Der p-Wert für einen festgelegten Hypothesentest und eine gegebene Stichprobe entspricht der Wahrscheinlichkeit einen Wert für die Teststatistik unter der Nullhypothese zu erhalten, der genauso extrem oder extremer ist, als die gegebene Realisation t von T

Liegt der p-Wert unter dem vorgegebenen Signifikanzniveau α , so wird die \mathcal{H}_0 abgelehnt.



Beispiel

Eine Münze wird 100 mal geworfen, es erscheint 80 mal Kopf. Ist die Münze fair?

$$\mathbb{P}\left(T \geq 80 \mid \mathcal{H}_0 : p = \frac{1}{2}\right) = 1 - \sum_{i=1}^{80} \binom{n}{i} p_i (1-p)^{n-i} \approx 5.6 \cdot 10^{-10} \tag{3.2}$$

Ein Ausgang mit 80 mal Kopf ist extrem unwahrscheinlich ($5.6 \cdot 10^{-10} \ll \alpha = 0.05$) bei einer fairen Münze und \mathcal{H}_0 wird abgelehnt.

3.6 Hypothesentests und Konfidenzintervalle

Hypothesentests und Konfidenzintervalle sind eng miteinander verknüpft. Ein Konfidenzintervall gibt den Bereich an, in dem ein unbekannter Parameter mit einer bestimmten Wahrscheinlichkeit liegt. Ein Hypothesentest prüft, ob ein spezifischer Parameterwert plausibel ist.

Ein zentraler Zusammenhang besteht darin, dass ein Parameterwert Θ_0 genau dann nicht abgelehnt wird, wenn er im entsprechenden $(1 - \alpha)$ -Konfidenzintervall liegt. Ist Θ_0 außerhalb des Intervalls, wird die Nullhypothese verworfen. Dies ermöglicht eine alternative Interpretation von Hypothesentests und unterstützt die statistische Entscheidungsfindung.

! Merke

Die Nullhypothese wird **nie** angenommen. Die korrekte Formulierung ist:

„Die Nullhypothese konnte nicht abgelehnt werden.“

Nur weil sie nicht abgelehnt wird, heißt das nicht, dass sie richtig ist.

3.7 Wichtige Hypothesentests

- ... für Mittelwert bei unbekannter Varianz: $T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$
- ... für Varianz einer Normalverteilung: $T = \frac{(n-1)S^2}{\sigma_0^2}$
- ... für vers. Mittelwerte bei gleicher unbekannter Varianz Δ
- ... für unterschiedliche Mittelwerte bei unbekannter Varianz
- ... für unterschiedliche Mittelwerte bei gepaarten Stichproben
- ... für Verhältnis der Varianzen $T = \frac{S_X^2}{S_Y^2}$
- χ^2 -Test für kategoriale Daten

4 Multivariate Zufallsvariablen und Methoden

4.1 Bivariate Zufallsvariablen und Verteilungen

Eine **bivariate Zufallsvariable** (X_1, X_2) ordnet Ereignissen Wertepaare reeller Zahlen zu und wird auch als **Zufallsvektor** der Dimension 2 bezeichnet. Dabei steht X_1 für eine Zufallsvariable und x_1 für eine ihrer Realisationen.

Wir betrachten Wahrscheinlichkeiten von Ereignissen der Form $\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\}$.

Die **Verteilungsfunktion** einer bivariaten Zufallsvariable $\{(X_1, X_2)\}$ ist gegeben durch:

$$F_{X_1, X_2}(x_1, x_2) = \mathbb{P}(\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\}) \quad (4.1)$$

Sie beschreibt die Wahrscheinlichkeit, dass beide Zufallsvariablen jeweils einen Wert kleiner oder gleich (x_1, x_2) annehmen.

4.1.1 Unabhängigkeit

Für **abhängige** Zufallsvariablen hängt die bedingte Verteilung von X_2 von der bedingten Information ab (z.B.: wenn X_1 groß ist, dann auch X_2 wahrscheinlich groß).

Für **unabhängige** Zufallsvariablen hat die Verteilung von X_1 keinen Einfluss auf die Verteilung von X_2 und umgekehrt.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

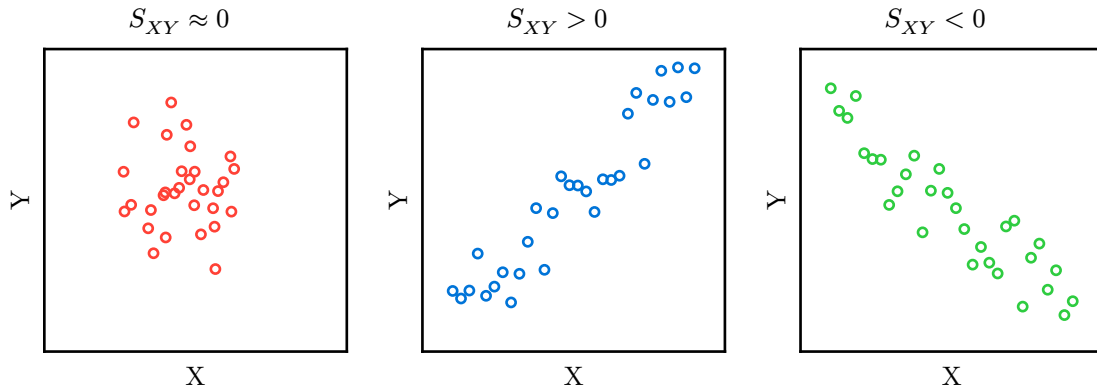
4.2 Multivariate Zufallsvariablen und Verteilungen

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

4.2.1 Kovarianz

Eines der wichtigsten, bivariaten Abhängigkeitsmaße ist die Kovarianz. Sie ist ein Maß für die lineare Abhängigkeit zwischen X und Y. Die Kovarianz kann folgendermaßen geschätzt werden:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}) \cdot (Y_i - \bar{Y}) \quad (4.2)$$



4.3 Korrelation

4.3.1 Definition

Die Kovarianz liefert nur eingeschränkt Information zu linearen Abhängigkeit zwischen X und Y, das sie von von X und Y abhängt. Die Korrelation ist definiert als **normierte Kovarianz**:

$$\begin{aligned} \rho_{XY} &= \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \\ R_{XY} &= \frac{S_{XY}}{S_X \cdot S_Y} \end{aligned} \quad (4.3)$$

Die Korrelation erfasst nur lineare Zusammenhänge (keine Kreise oder so). Kausalität bezeichnet die Beziehung zwischen Ursache und Wirkung. Korrelation bezeichnet einen symmetrischen statistischen Zusammenhang. Korrelation kann dabei helfen Kausalzusammenhänge aufzudecken, allerdings kann **Kausalität nie** aus einer **statistischen Analyse** **gefolgert** werden.

4.3.2 Scheinkorrelation

- Reverse Causality (zB: je schneller Windräder desto schneller der Wind)
- Third factor / confounding variable (X und Y ähnlich aufgrund von Z)
- Spurious relationship (reiner Zufall)

4.4 Lineare Regression

Lineare Regression setzt Variablen in eine lineare Beziehung zueinander. Das Ziel von linearer Regression ist üblicherweise die Prognose und die Evaluation von statistischen Abhängigkeiten.

$$Y = \beta_0 + X\beta_1 + \varepsilon \quad (4.4)$$

$$Y = X\beta + \varepsilon = \underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{n \times 1} = \underbrace{\begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}}_{n \times 2} \cdot \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{2 \times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{n \times 1} \quad (4.5)$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_n \end{pmatrix} \quad (4.6)$$

4.5 ANOVA - Analysis of Variance

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua quaerat voluptatem. Ut enim aequale doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

4.6 Wichtige Hypothesentests für multivariate Statistik

Um zu überprüfen, ob eine beobachtete Stichprobenkorrelation statistisch signifikant ist kann die folgende Teststatistik verwendet werden.

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \quad (4.7)$$

R bezeichnet hier den Schätzer für die Korrelation und n die Anzahl der Beobachtungen.

5 Nichtparametrische Methoden

5.1 Parametrische und nichtparametrische statistische Modelle

Ein parametrisches statistisches Modell für eine eindimensionale Zufallsvariable ist eine Menge an Verteilungen derselben Verteilungsfamilie, indiziert durch eine Parametermenge

$$\dots \quad (5.1)$$

Ein nichtparametrisches statistisches Modell für eine eindimensionale Zufallsvariable ist eine Menge an Verteilungen, die entweder diskret oder kontinuierlich sind

$$\dots \quad (5.2)$$

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua quaerat voluptatem. Ut enim aequale doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

6 Experimental design

6.1 Grundlegende Aspekte der statistischen Versuchsplanung

Statistische Versuchsplanung ist die Formalisierung von Arbeitsschritten zur Untersuchung einer Forschungsfrage.

- Reproduzierbarkeit gewährleisten
- passende Daten sammeln und passende statistische Methoden verwenden
- den individuellen Charakter der Frage berücksichtigen
- ökonomische Aspekte berücksichtigen und Kosten wenn möglich minimieren

Dazu wird wie folgt vorgegangen:

1. Klarifizieren bzw Ableiten einer Fragestellung
2. Übersetzung der Frage in ein fachspezifisches Modell und anschließend ein passendes statistisches Modell wählen
3. Detaillierte Festlegung des Versuchsplans
4. Festlegung der Datenbeschaffung oder Stichprobenerhebung
5. Detaillierte Festlegung der statistischen Methoden
6. Ökonomische Optimierung der Versuchsplanung (zB optimaler Stichprobenumfang)

6.2 Varianzquellen in statistischen Experimenten

Primärvarianz: Anteil der Variabilität in Daten, der durch die Variation der experimentellen Bedingungen erklärt werden kann

Sekundärvarianz: Anteil der Variabilität in Daten, der durch die Wirkung von Störvariablen verursacht wird (Systematischer Fehler)

Fehlervarianz: Anteil der Variabilität in Daten, der durch zufällige Unterschiede der Messungen oder durch zufällige, unsystematische Einflüsse verursacht wird (Zufälliger Fehler)

Das Ziel der statistischen Versuchsplanung ist es, unter anderem, die Primärvarianz zu maximieren, die Sekundärvarianz zu kontrollieren und die Fehlervarianz zu minimieren.

6.3 Prinzipien der statistischen Versuchsplanung

6.3.1 Arten von Störgrößen

- Eliminierbare Störgrößen
- Kontrollierbare Störgrößen
- Nicht kontrollierbare Störgrößen

6.3.2 Randomisierung

Randomisierung reduziert den Einfluss von Störgrößen in statistischen Experimenten, indem sie Gruppen und Bedingungen zufällig zuweist. Dadurch werden Störvariablen ausgeglichen, was statistische Zusammenhänge zuverlässiger macht. Sie ist besonders in der medizinischen Statistik verbreitet, funktioniert aber nur bei **großen Stichproben** gut.

6.3.3 Blockbildung

Blockbildung ist eine Alternative zur Randomisierung bei **kleinen Stichproben**, um den Einfluss von Störgrößen zu kontrollieren. Dabei wird die Stichprobe in Blöcke mit ähnlichen Eigenschaften eingeteilt, sodass Heterogenitäten berücksichtigt werden. Die Blöcke können in der Analyse separat betrachtet oder als erklärende Variable einbezogen werden. Eine Kombination mit Randomisierung innerhalb der Blöcke ist möglich.

6.3.4 Statistische Kontrolle von Störfaktoren

Die statistische Kontrolle von Störfaktoren integriert diese direkt in die Modellierung, um ihren Einfluss zu berücksichtigen. Sie hängt von der verwendeten Analysemethode ab und ist oft unvermeidbar, wenn Störgrößen nicht ausgeschlossen oder gesteuert werden können, z. B. in ökonomischen Studien.

6.3.5 Weitere Prinzipien

- Wiederholungen
- Kontrollgruppen
- Ökonomie

6.4 Typen von Stichproben

6.4.1 Einfache Zufallsstichprobe

Eine einfache Zufallsstichprobe ist eine zufällige Auswahl von Elementen aus einer Population, bei der jedes Element die gleiche Chance hat, ausgewählt zu werden. Dabei werden Störfaktoren nicht berücksichtigt, und es wird angenommen, dass die Population homogen ist, sodass die Stichprobenwerte unabhängig und identisch verteilt sind.

6.4.2 Geschichtete Stichproben

Bei geschichteten Stichproben wird die Population in homogene Schichten unterteilt, um den Einfluss bestimmter Merkmale zu berücksichtigen. Jede Schicht wird separat beprobt und kann einzeln oder

gemeinsam analysiert werden. Die Stichprobengröße sollte proportional zur Schichtgröße sein, um Verzerrungen zu vermeiden.

6.4.3 Klumpenstichproben

Bei einer Klumpenstichprobe wird die Population in natürliche Gruppen (Klumpen) unterteilt, aus denen zufällig einige ausgewählt oder vollständig analysiert werden. Sie ist kosteneffizient, berücksichtigt jedoch Störfaktoren nicht explizit. Im Gegensatz zur geschichteten Stichprobe wird angenommen, dass die Heterogenität innerhalb der Klumpen der Gesamtpopulation entspricht.

6.5 Inverse Probleme der Stichprobengröße

Der minimale Stichprobenumfang, sodass ein Effekt mit einer vorgegebener Wahrscheinlichkeit durch den statistischen Versuch nachgewiesen werden kann, beziehungsweise die statistische Schätzung einer Merkmalausprägung einer vorgegebenen Genauigkeit genügt.

Einstichproben t-Test: Beim Einstichproben-t-Test wird überprüft ob der Mittelwert einer normalverteilten Zufallsvariablen signifikant von einem vorgegebenen Wert abweicht, wobei die Varianz σ_x^2 nicht bekannt ist.

$$T = \frac{\sqrt{n} \cdot (\bar{X} - \mu_0)}{S} \quad (6.1)$$

Zweistichproben t-Test: Beim einem Zweistichproben-t-Test wird überprüft ob die Differenz des Mittelwerts von zwei normalverteilten Zufallsvariablen signifikant von einem vorgegebenen Wert (meistens 0) abweicht. Die Varianzen werden nicht als bekannt vorausgesetzt.