

1 Modelle für Zählraten

1.1 Wahrscheinlichkeitsmodelle

- Grundraum Ω mit Elementarereignissen ω_i (z.B. Augenzahl eines Würfels)
- Ereignisse A, B, C, \dots (Teilmenge von Ω) (z.B. Kombinationen von Augenzahlen)
- Wahrscheinlichkeit für jedes Ereignis $P(A), P(B), \dots$

1.2 Operatoren

- $A \cup B$ - ODER (inklusive, "und/oder")
- $A \cap B$ - UND (Konjunktion)
- A^c - NICHT (Negation)
- $A \setminus B = A \cap B^c$ - A UND NICHT B

1.3 Axiome der Wahrscheinlichkeitsrechnung

1. $P(A) \geq 0$ - Die Wahrscheinlichkeiten sind immer nicht-negativ
2. $P(\Omega) = 1$ - Das Ereignis Ω hat Wahrscheinlichkeit eins
3. $P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \emptyset$ (A und B sind disjunkt), d.h. für alle Ereignisse, die sich gegenseitig ausschliessen.

Daraus folgen:

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

1.4 Wahrscheinlichkeiten berechnen

Für diskrete Wahrscheinlichkeitsmodelle

1.4.1 Summe der Elementarereignisse (verschiedene $P(\omega_i)$)

$$P(A) = \sum_{\omega \in A} P(\{\omega\})$$

1.4.2 Laplace-Modell (gleiche $P(\omega_i)$)

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{günstig}}{\text{möglich}}$$

1.5 Unabhängigkeit

A und B sind stochastisch unabhängig, wenn gilt:

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

somit können wir dies annehmen, falls wir wissen, dass A und B nicht kausal voneinander abhängig sind

1.6 Bedingte Wahrscheinlichkeit (Abhängigkeit)

1.6.1 Satz von Bayes

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

somit ist $P(A|B)$ nicht unbedingt $P(B|A)^1$

1.6.2 Gesetz der totalen Wahrscheinlichkeit

$$P(B) = \sum_{i=1}^k P(B|A_k)P(A_k)$$

1.6.3 Odds

$$\text{odds}(E) = \frac{P(E)}{1 - P(E)} = \frac{P(E)}{P(E^c)}$$

(vgl. Abschnitt 1.4.2)

$$\text{odds}(E|A) = \frac{P(E|A)}{1 - P(E|A)}$$

¹ $P(A|B)$: $P(A)$ gegeben B

²Dabei ist $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ (TR: $nCr(n, x)$)

1.6.4 Odds-Ratio

$$\text{OR} = \frac{\text{odds}(E|A)}{\text{odds}(E|B)}$$

1.7 Zufallsvariable

$$X(\omega) = x$$

$$X: \Omega \rightarrow \mathbb{R} \\ \omega \rightarrow X(\omega)$$

Grossbuchstabe: Funktion, Kleinbuchstabe: Realisierung

$$P(X = x) = P(\{\omega; X(\omega) = x\}) = \sum_{\omega: X(\omega)=x} P(\omega)$$

So dass $\omega = x$, also einen gewünschten Wert (z.B. Jass: $P(\text{Koenig}) = P(\text{Schilten} - \text{Koenig}) + P(\text{Schellen} - \text{Koenig}) + \dots$)

1.8 Diskrete Verteilungen

1.8.1 Kennzahlen

Erwartungswert

$$\mathbb{E}(X) = \sum_{x \in \mathbb{W}_X} xP(X = x)$$

wobei \mathbb{W}_x der Wertebereich von X ist.

Varianz

$$\text{Var}(X) = \sum_{x \in \mathbb{W}_X} (x - \mathbb{E}(X))^2 P(X = x)$$

Standardabweichung

$$\sigma(X) = \sqrt{\text{Var}(X)}$$

1.8.2 Bernoulli-(π)-Verteilung

$$P(X = 1) = \pi, P(X = 0) = 1 - \pi, 0 \leq \pi \leq 1$$

Beschreibt das Eintreffen bzw. nicht-eintreffen eines bestimmten Ereignisses.

1.8.3 Binomialverteilung ²⁾

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, x \in \mathbb{N}_0$$

Dabei ist $0 \leq \pi \leq 1$ der Erfolgsparameter der Verteilung.

Notation: $X \sim \text{Bin}(n, \pi)$ (X folgt einer Binomialverteilung mit Parametern n und π)

Zusammenhänge:

- $\text{Bin}(1, \pi) = \text{Bernoulli}(\pi)$
- $X_1 \sim \text{Bin}(n_1, \pi); X_2 \sim \text{Bin}(n_2, \pi)$ unabhängig $\Rightarrow S := X_1 + X_2$, dann $S \sim \text{Bin}(n_1 + n_2, \pi)$
- $X_1 \sim \text{Bin}(n_1, \pi); X_2 \sim \text{Bin}(n_2, \pi)$ unabhängig $\Rightarrow P(X_1 = x_1 \cap X_2 = x_2) = P(X_1 = x_1) \cdot P(X_2 = x_2)$

Beispiel

Urne mit Zurücklegen

1.8.4 Poisson-(λ)-verteilung

$$P(X = x) = \exp(-\lambda) \frac{\lambda^x}{x!}, x \in \mathbb{N}_0$$

Dabei sind $\mathbb{E}(X) = \lambda, \text{Var}(X) = \lambda, \sigma(X) = \sqrt{\lambda}$
 Für zwei unabhängige Poisson-Verteilungen $X \sim \text{Poisson}(\lambda_x), Y \sim \text{Poisson}(\lambda_y)$
 ist $X + Y \sim \text{Poisson}(\lambda_x + \lambda_y)$
 Es gilt auch

$$P(X > n) = 1 - P(X \leq n) = 1 - (P(X = 0) + P(X = 1) + \dots + P(X = n))$$

1.8.5 Geometrische Verteilung

Sei $X \sim \text{Bernoulli}(\pi)$, dann ist

$$Y = P(X = n) = \pi(1 - \pi)^{n-1}$$

die Anzahl Fehlversuche bis zu einem erfolgreichen Versuch.

1.8.6 Poisson-Approximation der Binomial-Verteilung

$X \sim \text{Bin}(n, \pi)$ und $Y \sim \text{Poisson}(\lambda)$, für kleine π und grosse n gilt:

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \approx P(Y = x) = \exp(-\lambda) \frac{\lambda^x}{x!}, x \in \mathbb{N}_0$$

wobei $\lambda = n\pi$

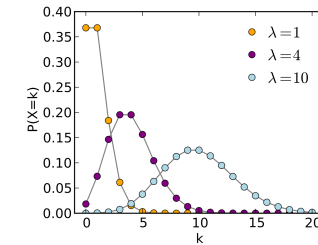


Abbildung 1: Poisson Approximation der Binomialverteilung

1.8.7 Diskrete Uniformverteilung

$$P(X = x_i) = \frac{1}{n}, i \in \mathbb{N}$$

$X \sim \text{Uniform}(x_i)$, alle n Ereignisse x sind gleich wahrscheinlich

1.8.8 Hypergeometrische Verteilung

Einfluss von entfernten Ereignissen auf Wahrscheinlichkeiten von neuen Ziehungen (ohne Zurücklegen).

$$P(X = x) = \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}}$$

$X \sim \text{Hyper}(N, n, m)$, dabei N die total möglichen Ereignisse, m die Gewinne und es wird n gezogen.

1.9 Kennwerte

1.9.1 Bernoulli-Verteilung

$$\mathbb{E}(X) = \pi \\ \text{Var}(X) = \pi(1 - \pi) \\ \sigma_X = \sqrt{\pi(1 - \pi)}$$

1.9.2 Binomialverteilung

$$\mathbb{E}(X) = n\pi \\ \text{Var}(X) = n\pi(1 - \pi) \\ \sigma_X = \sqrt{n\pi(1 - \pi)}$$

1.9.3 Poisson-Verteilung

$$\mathbb{E}(X) = \lambda \\ \text{Var}(X) = \lambda \\ \sigma_X = \sqrt{\lambda}$$

1.9.4 Geometrische Verteilung

$$\mathbb{E}(X) = \frac{1}{\pi} \\ \text{Var}(X) = \frac{1-\pi}{\pi^2} \\ \sigma_X = \frac{\sqrt{1-\pi}}{\pi}$$

1.9.5 Hypergeometrische Verteilung

$$\mathbb{E}(X) = \frac{nm}{M} \\ \text{Var}(X) = \frac{\frac{nm}{M}(N-m)(N-n)}{N^2(N-1)} \\ \sigma_X = \sqrt{\frac{nm(N-m)(N-n)}{N^2(N-1)}}$$

2 Statistik für Zählraten

- Grundfragestellung:** Welches ist der zu den Beobachtungen plausibelste Parameterwert? Die Antwort auf diese Frage heisst (Punkt-)Schätzung.
- Grundfragestellung:** Sind die Beobachtungen kompatibel (statistisch vereinbar) mit einem vorgegebenen Parameterwert? Die Antwort auf diese 2. Grundfrage heisst statistischer Test.
- Grundfragestellung:** Grundfragestellung: Welche Parameterwerte sind mit den Beobachtungen kompatibel (statistisch vereinbar)? Die Antwort auf diese 3. Grundfrage heisst Vertrauensintervall. Das Vertrauensintervall ist allgemeiner und informativer als ein statistischer Test.

2.1 Punktschätzung von Parametern

\hat{X} bezeichnet den Schätzwert von X

Bei **Binomialverteilung:**

2.1.1 Momentenmethode

Aus $\mathbb{E}(X) = n\pi \Leftrightarrow \pi = \frac{\mathbb{E}(X)}{n}$, daraus $\mathbb{E}(\hat{X}) = x$ und somit

$$\hat{\pi} = \frac{x}{n}$$

2.1.2 Maximum-Likelihood

Vorgehen:

- Funktion P der Wahrscheinlichkeit aufstellen
- $\log(P)$
- $\frac{dP}{d\pi} = 0$
- auflösen nach π

Dies ist für eine Binomialverteilung ebenfalls $\hat{\pi} = \frac{x}{n}$

2.2 Aufbau statistischer Test

$P(X \geq c)$ für verschiedene c

- Modell X erstellen
- Nullhypothese

$$H_0: \pi = \pi_0$$

und Alternativhypothese

$$H_A: \begin{cases} \pi \neq \pi_0 & (\text{zweiseitig}) \\ \pi > \pi_0 & (\text{einseitig nach oben}) \\ \pi < \pi_0 & (\text{einseitig nach unten}) \end{cases}$$

oft ist $H_0: \pi = 1/2$ (= reiner Zufall). Man testet also gegen Zufall.

- Teststatistik T (Anzahl treffer bei n Versuchen), Verteilung unter $H_0: T \sim \text{Bin}(n, \pi_0)$
- Festlegen von Signifikanzniveau α (meist $\alpha = 0.05$ oder $\alpha = 0.01$)
- Bestimmung Verwerfungsbereich

$$K = \begin{cases} [0, c_u] \cup [c_o, n] & H_A: \pi \neq \pi_0 \\ [c, n] & H_A: \pi > \pi_0 \\ [0, c] & H_A: \pi < \pi_0 \end{cases}$$

Wobei c der Wert ist bei dem noch $P(X \leq c) \leq \alpha$ für $H_A: \pi < \pi_0$, analog $P(X \geq c) \leq \alpha$ für $H_A: \pi > \pi_0$

- Testentscheid: Ist $t \in K$? Falls ja wird H_0 verworfen, falls nicht wird sie als korrekt angenommen³

Bsp. Berechnung von c

Es sei $X \sim \text{Bin}(150, 0.1)$ unter $H_A: \pi < 0.1$. Dann soll

$$P(X \leq c) \leq \alpha$$

Also berechne mit Tabelle (schaue wo $P(X = x) \leq \alpha$ für verschiedene x (kumulativ)) oder R.

³Achtung: Das heisst nicht, dass H_0 gültig ist! (Falsifizierbarkeit)

⁴Aber dies bedeutet nicht, dass falls $\text{Cor}(X, Y) = 0$, X und Y dann unabhängig sind!

2.2.1 Normalapproximation der Binomialverteilung

Gilt, wenn $n\pi > 5$ und $n(1 - \pi) > 5$ (Faustregel)

Für eine Verteilung $X \sim \text{Binom}(n, \pi)$ und $\alpha = 0.05$ gilt für einseitige Tests:

$$c \approx \begin{cases} n\pi_0 + 1.64\sqrt{n\pi_0(1 - \pi_0)} & \text{bei } H_0: \pi > \pi_0 \text{ (aufgerundet)} \\ n\pi_0 - 1.64\sqrt{n\pi_0(1 - \pi_0)} & \text{bei } H_0: \pi < \pi_0 \text{ (abgerundet)} \end{cases}$$

Für einen zweiseitigen Test ($\pi \neq \pi_0$) gilt:

$$c_o \approx n\pi_0 + 1.96\sqrt{n\pi_0(1 - \pi_0)} \text{ (aufgerundet)}$$

$$c_u \approx n\pi_0 - 1.96\sqrt{n\pi_0(1 - \pi_0)} \text{ (abgerundet)}$$

2.2.2 Fehler 1. und 2. Art

- Art: Fälschliches Verwerfen von H_0 , obwohl H_0 richtig ist.
- Art: Fälschliches Beibehalten von H_0 , obwohl H_A zutrifft.

$$P(\text{Fehler 1. Art}) = P_{H_0}(X \in K) \leq \alpha$$

Fehler 1. Art soll möglichst vermieden werden!

2.2.3 Macht (Power)

$$\text{Macht} := 1 - P(\text{Fehler 2. Art}) = P_{H_A}(X \in K) = P(X \geq c) \text{ z.B.}$$

Idee: Wie gross muss eine Stichprobe sein, damit mit einer bestimmten Macht $\beta = x$ eine Hypothese bewiesen werden kann auf Signifikanzniveau α ?

2.2.4 P-Wert

Gibt die Wahrscheinlichkeit an, dass die Beobachtung oder extremeres Ereignis eintritt unter H_0

$$P_{H_0}(T \geq t)$$

Es ist auch das kleinste Signifikanzniveau α , auf dem H_0 gerade noch verworfen wird. Also falls p -Wert $> \alpha$ wird H_0 beibehalten.

2.2.5 Vertrauensintervall (VI)

$$I := \{\pi_0; \text{Nullhypothese } H_0: \pi = \pi_0 \text{ wird beibehalten}\}$$

Für grosse n gilt

$$I \approx \frac{x}{n} \pm 1.96\sqrt{\frac{x}{n}\left(1 - \frac{x}{n}\right)\frac{1}{n}}$$

Die Werte von π_0 bei denen $H_0: \pi = \pi_0$ nicht verworfen wird, ist ein $(1 - \alpha)$ -VI.

$$P_\pi(\pi \in I(X)) \gtrsim 1 - \alpha$$

Ein $(1 - \alpha)$ -VI, enthält den wahren Parameter π mit einer Wahrscheinlichkeit von $(1 - \alpha)$

3 Modelle und Statistik für Zählraten

3.1 Deskriptive Statistik

3.1.1 Kennzahlen

Arithmetisches Mittel

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

Empirische Standardabweichung

$$s_x = \sqrt{\text{Var}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Quantile

α -Quantil

³Wert x bei dem $\alpha \cdot 100\%$ -Werte kleiner als x sind³

3.1.2 Kovarianz und Korrelation

Gemeinsame Verteilung von zwei Zufallsvariablen X und Y

Kovarianz

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_x)(Y - \mu_y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

es gilt somit auch

$$\text{Cov}(X, X) = \text{Var}(X)$$

Korrelation

$$\text{Cor}(X, Y) = \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

wobei $\rho_{XY} \in [-1, 1]$

Falls X, Y unabhängig $\text{Cor}(X, Y) = 0$.⁴

Empirische Korrelation

$$r = \frac{s_{xy}}{s_x s_y}$$

wobei $s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

3.1.3 Grafische Methoden

Histogramme

Einteilung in Klassen, auftragen der Beobachtungen je Klasse in Balkendiagramm

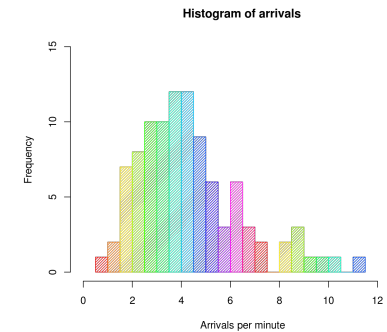


Abbildung 2: Histogramm

Boxplot

Rechteck, vom 75%- und 25%-Quantil begrenzt

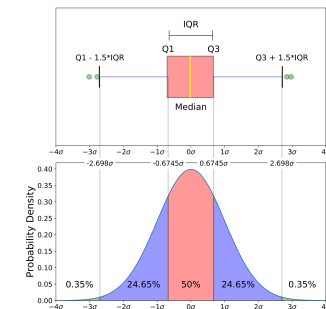


Abbildung 3: Beispiel Boxplot (IQR = Interquartile-Range)

Streudiagramm (Scatter-Plot)
Auftragen der Daten (x_n, y_n)

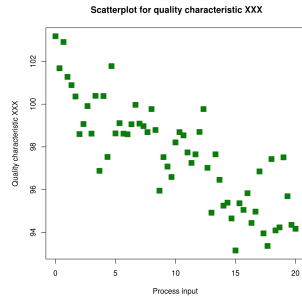


Abbildung 4: Streudiagramm

3.2 Stetige Zufallsvariablen und Wahrscheinlichkeitsverteilungen

Eine Zufallsvariable X heisst stetig, falls deren Wertebereich \mathbb{W}_X stetig ist Da Punktverteilung

$$P(X = x) = 0, \forall x \in \mathbb{W}_X,^5$$

benötigen wir

$$P(X \in (a, b]) = P(a < X \leq b)$$

Kumulative Verteilungsfunktion

$$F(x) = P(X \leq x)$$

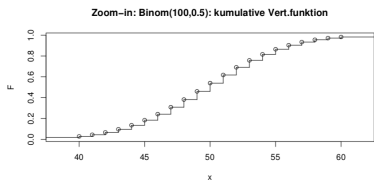
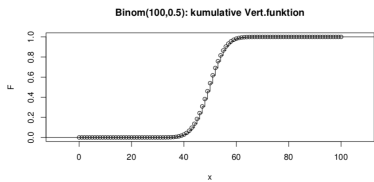


Abbildung 5: Kumulative Verteilungsfunktion

3.2.1 Wahrscheinlichkeits-Dichte

$$f(x) = \dot{F}(x) \iff F(x) = \int_{-\infty}^x f(y)dy$$

$$f(x) \geq 0, \forall x$$

3.3 Kennzahlen von stetigen Verteilungen

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x f(x) dx \\ \text{Var}(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x) dx \\ \sigma(X) &= \sqrt{\text{Var}(X)} \end{aligned}$$

⁵Da in jedem kontinuierlichen Intervall ∞ Werte sind

3.3.1 Quantile

$$P(X \leq q(\alpha)) = \alpha$$

$q(\alpha)$ ist der Punkt, an dem die Fläche unter der Dichtefunktion $f(x)$ von $-\infty$ bis $q(\alpha)$ gleich α ist. (z.B. beim Median ($\alpha = 50\%$) sind die Flächen darunter und darüber gleich gross)

3.4 Stetige Verteilungen

3.4.1 Uniforme Verteilung

$X \sim \text{Uniform}([a, b]), \mathbb{W}_X = [a, b]$

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{falls } a \leq x \leq b \\ 0, & \text{sonst} \end{cases}$$

somit ist die kumulative Verteilung

$$F(x) = \begin{cases} 0, & \text{falls } x < a \\ \frac{x-a}{b-a}, & \text{falls } a \leq x \leq b \\ 1, & \text{falls } x > b \end{cases}$$

Kennzahlen

$$\begin{aligned} \mathbb{E}(X) &= \frac{a+b}{2}x \\ \text{Var}(X) &= \frac{(b-a)^2}{12} \\ \sigma_X &= \frac{b-a}{\sqrt{12}} \end{aligned}$$

3.4.2 Exponential-Verteilung

$X \sim \text{Exp}(\lambda), \mathbb{W}_X = [0, \infty), \lambda \in \mathbb{R}^+$

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{falls } x \geq 0 \\ 0, & \text{sonst} \end{cases}$$

also

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{falls } x \geq 0 \\ 0, & \text{falls } x < 0 \end{cases}$$

Kennzahlen

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{\lambda} \\ \text{Var}(X) &= \frac{1}{\lambda^2} \\ \sigma_X &= \frac{1}{\lambda} \end{aligned}$$

3.4.3 Normalverteilung (Gauss'sche-Verteilung)

$X \sim \mathcal{N}(\mu, \sigma^2), \mathbb{W}_X = \mathbb{R}, \mu \in \mathbb{R}$ und $\sigma \in \mathbb{R}^+$

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

$F(x) \Rightarrow$ Tabelle!

Kennzahlen

$$\begin{aligned} \mathbb{E}(X) &= \mu \\ \text{Var}(X) &= \sigma^2 \\ \sigma_X &= \sigma \end{aligned}$$

Summe

Seien $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ i.i.d., $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ i.i.d. und $Y = X_1 + X_2$ dann ist

$$Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

3.4.4 Standard-Normalverteilung

$X \sim \mathcal{N}(0, 1), \mathbb{W}_X = \mathbb{R}, \mu = 0$ und $\sigma = 1$

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

$$\Phi(x) = \int_{-\infty}^x \varphi(y)dy$$

$$\Phi(-c) = P(X \leq -c) = P(X \geq c) = 1 - P(X \leq c) = 1 - \Phi(c)$$

3.5 Funktionen einer Zufallsvariable

Sei $g: \mathbb{R} \rightarrow \mathbb{R}$ und X eine Zufallsvariable, so ist

$$Y = g(X)$$

eine Transformation.

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

3.5.1 Lineare Transformation

Sei $X \sim \mathcal{N}(\sigma, \omega^2)$ und $Y = a + bX$ dann sind

$$\begin{aligned} \mathbb{E}(Y) &= a + b\mathbb{E}(X) \\ \text{Var}(Y) &= b^2 \cdot \text{Var}(X) \\ \sigma_Y &= |b| \cdot \sqrt{\text{Var}(X)} \\ q_Y(\alpha) &= a + b \cdot q_X(\alpha) \end{aligned}$$

3.5.2 Standardisieren einer Zufallsvariable

Überführen von X in eine *Standard-Normalverteilung* ($\mathbb{E} = 0, \sigma = 1$)

$$Z = g(X) = \frac{X - \mathbb{E}(X)}{\sigma_X} = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

3.5.3 Lognormal-Verteilung

Sei $Y \sim \mathcal{N}(\mu, \sigma^2)$ dann soll $X = \exp(Y)$ mit $\mu \in \mathbb{R}$ und $\sigma \in \mathbb{R}^+$

$$\mathbb{E}(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right) > \exp(\mathbb{E}(Y))$$

3.5.4 Berechnung von Momenten

Das k -te Moment ist gegeben als

$$m_k = \mathbb{E}(X^k)$$

also z.B.

$$m_2 = \mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$$

Verschiebungssatz für die Varianz:

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

3.6 Überprüfen der Normalverteilungs-Annahme

3.6.1 Q-Q Plot (Quantil-Quantil Plot)

Man plottet die empirischen Quantile gegen die theoretischen Quantile der Modell-Verteilung. Die Punkte sollten ungefähr auf der Winkelhalbierenden $y = f(x) = x$ liegen.

3.6.2 Normal-Plot

Für Klassen von Verteilungen, z.B. Klasse der Normalverteilungen mit verschiedenen μ, σ .

Sei $X \sim \mathcal{N}(\mu, \sigma^2)$, dann sind die Quantile von X

$$q(\alpha) = \mu + \sigma\Phi^{-1}(\alpha)$$

Ein *Q-Q Plot* bei dem die Modell-Verteilung gleich $\mathcal{N}(0, 1)$ ist, heisst Normal-Plot.

3.7 Funktionen von mehreren Zufallsvariablen

Statt einer Zufallsvariable X und deren n unabhängigen Realisierungen x_1, x_2, \dots, x_n , nimmt man oft X_1, X_2, \dots, X_n . Somit wird $y = g(x_1, x_2, \dots, x_n)$ zu einer Funktion von Zufallsvariablen

$$Y = g(X_1, X_2, \dots, X_n)$$

3.7.1 Unabhängigkeit und i.i.d. Annahme

Unabhängig heisst, dass es keine gemeinsamen Prozesse gibt, die den Ausgang beeinflussen.

Notation:

$$X_1, X_2, \dots, X_n \text{ i.i.d}$$

wobei *i.i.d* für *independent, identically distributed* steht.

Es gilt dann immer

$$\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$$

wenn X_1, X_2 unabhängig, auch

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2),$$

für nicht unabhängig

$$\text{Var}(aX_1 + bX_2) = a^2\text{Var}(X_1) + b^2\text{Var}(X_2) + 2ab\text{Cov}(X_1, X_2).$$

3.7.2 Gesetz der grossen Zahlen und \sqrt{n} -Gesetz

Sei X_1, X_2, \dots, X_n i.i.d \sim kumulative Verteilungsfunktion F , dann sind

$$\begin{aligned} \mathbb{E}(\bar{X}_n) &= \mu \\ \text{Var}(\bar{X}_n) &= \frac{\sigma_X^2}{n} \\ \sigma(\bar{X}_n) &= \frac{\sigma_X}{\sqrt{n}} \end{aligned}$$

Somit sind für eine doppelte Genauigkeit viermal so viele Messwerte nötig. Standardabweichung von X_n ist der *Standardfehler* des Arithmetischen Mittels.

$$\bar{X}_n \rightarrow \mu (n \rightarrow \infty)$$

3.7.3 Zentraler Grenzwertsatz

Sei X_1, X_2, \dots, X_n i.i.d, dann gilt

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right)$$

und daraus folgt für die Summe $\sum_{i=1}^n X_i$

$$S_X \approx \mathcal{N}(n\mu, n\sigma^2).$$

Aus

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma_X} \sim \mathcal{N}(0, 1)$$

folgt

$$\forall x : \lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x)$$

3.7.4 Verletzung der Unabhängigkeit

Sei X_1, X_2, \dots, X_n i.i.d

$$\begin{aligned} \mathbb{E}(\bar{X}_n) &= \mu \\ \text{Var}(\bar{X}_n) &= \frac{\sigma_X^2}{n} \left(1 + \frac{1}{n} \sum_{1 \leq i < j \leq n} \rho_{X_i X_j}\right) \end{aligned}$$

mit $\rho_{X_i X_j}$ die Korrelation zwischen X_i, X_j

Die Unabhängigkeit führt dazu, dass die Genauigkeit des arithmetischen Mittels beeinflusst wird!

3.8 Statistik für eine Stichprobe

	Annahmen				Macht
	σ_X bekannt	$X_i \sim \mathcal{N}$	sym. Verteilung	i.i.d.	
z-Test	•	•	•	•	***
t-Test		•	•	•	**
Wilcoxon			•	•	**
Vorzeichen				•	*

Abbildung 6: Übersicht der verschiedenen Tests für μ

3.8.1 Punktschätzung

Betrachtung von Daten x_1, x_2, \dots, x_n als Realisierungen von X_1, X_2, \dots, X_n i.i.d.

Wenn $\mathbb{E}(X_i) = \mu$ und $\text{Var}(X_i) = \sigma_X^2$ gesucht:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n \\ \hat{\sigma}_X^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \end{aligned}$$

3.8.2 z-Test (σ_X bekannt)

- Modell:** X_i ist eine kontinuierliche Messgrösse und Annahme X_1, X_2, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma_X^2)$
- Nullhypothese:**

$$H_0 : \mu = \mu_0$$

Alternativhypothese:

$$\begin{aligned} \text{oder } H_A : \mu &\neq \mu_0 && \text{zweiseitig} \\ \text{oder } H_A : \mu &> \mu_0 && \text{einseitig} \\ \text{oder } H_A : \mu &< \mu_0 && \text{einseitig} \end{aligned}$$

- Teststatistik:**

$$Z = \frac{(\bar{X}_n - \mu_0)}{\frac{\sigma_{X_n}}{\sigma_X}} = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma_X} = \frac{\text{beobachtet} - \text{erwartet}}{\text{Standardfehler}}$$

Verteilung der Teststatistik unter $H_0 : Z \sim \mathcal{N}(0, 1)$

- Signifikanzniveau:** α

- Verwerfungsbereich für die Teststatistik:**

$$K = \begin{cases} (-\infty, -\Phi^{-1}(1 - \frac{\alpha}{2})] \cup [\Phi^{-1}(1 - \frac{\alpha}{2}), \infty), & \text{bei } H_A : \mu \neq \mu_0 \\ (-\infty, -\Phi^{-1}(1 - \alpha)], & \text{bei } H_A : \mu < \mu_0 \\ [\Phi^{-1}(1 - \alpha), \infty), & \text{bei } H_A : \mu > \mu_0 \end{cases}$$

- Testentscheid:**

Überprüfen ob der beobachtete Wert der Teststatistik im Verwerfungsbereich K liegt.

3.8.3 Fehler 1./2. Art und Macht

Es gilt wie in Kapitel 2.2.2 und 2.2.3.

$$P_{\mu_0}(T \in K) = \alpha$$

$$P_{\mu}(T \in K) = \text{Macht}(\mu)$$

3.8.4 t-Test (σ_X unbekannt)

- Modell:** X_i ist eine kontinuierliche Messgrösse und Annahme X_1, X_2, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma_X^2)$
- Nullhypothese:**

$$H_0 : \mu = \mu_0$$

Alternativhypothese:

$$\begin{aligned} \text{oder } H_A : \mu &\neq \mu_0 && \text{zweiseitig} \\ \text{oder } H_A : \mu &> \mu_0 && \text{einseitig} \\ \text{oder } H_A : \mu &< \mu_0 && \text{einseitig} \end{aligned}$$

- Teststatistik:**

$$\hat{\sigma}_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\hat{\sigma}_X} = \frac{\text{beobachtet} - \text{erwartet}}{\text{geschätzter Standardfehler}}$$

Verteilung der Teststatistik unter $H_0 : T \sim t_{n-1}$

- Signifikanzniveau:** α

- Verwerfungsbereich für die Teststatistik:**

$$K = \begin{cases} (-\infty, -t_{n-1; 1-\frac{\alpha}{2}}] \cup [t_{n-1; 1-\frac{\alpha}{2}}, \infty), & \text{bei } H_A : \mu \neq \mu_0 \\ (-\infty, -t_{n-1; 1-\alpha}], & \text{bei } H_A : \mu < \mu_0 \\ [t_{n-1; 1-\alpha}, \infty), & \text{bei } H_A : \mu > \mu_0 \end{cases}$$

- Testentscheid:**

Überprüfen ob der beobachtete Wert der Teststatistik im Verwerfungsbereich K liegt.

3.8.5 P-Wert des t-Tests

$$P\text{-Wert} = P(|T| > |t|) = 2 \left(1 - F_{t_{n-1}}\left(\frac{\sqrt{n}|\bar{x}_n - \mu_0|}{\hat{\sigma}_X}\right)\right)$$

wobei $F_{t_{n-a}}$ die kumulative Verteilungsfunktion der t-Verteilung mit $n-1$ Freiheitsgraden ist ($F_{t_{n-1}}(t) = P(T \leq t), T \sim t_{n-1}$)

3.8.6 Vertrauensintervall für μ

Vgl. auch 2.2.5

Für einseitige Intervalle

$$\mu_0 \leq \bar{x}_n + \frac{\hat{\sigma}_X \cdot t_{n-1; 1-\alpha}}{\sqrt{n}} \quad \text{und} \quad \mu_0 \geq \bar{x}_n - \frac{\hat{\sigma}_X \cdot t_{n-1; 1-\alpha}}{\sqrt{n}}$$

und das zweiseitige Intervall

$$I = \left[\bar{x}_n - t_{n-1; 1-\frac{\alpha}{2}} \frac{\hat{\sigma}_X}{\sqrt{n}}, \bar{x}_n + t_{n-1; 1-\frac{\alpha}{2}} \frac{\hat{\sigma}_X}{\sqrt{n}} \right]$$

3.8.7 Vorzeichentest

- Modell:** X_1, X_2, \dots, X_n i.i.d. wobei X_i eine beliebige Verteilung hat
- Nullhypothese:**

$$H_0 : \mu = \mu_0 \quad (\mu \text{ ist der Median})$$

Alternativhypothese:

$$\begin{aligned} \text{oder } H_A : \mu &\neq \mu_0 && \text{zweiseitig} \\ \text{oder } H_A : \mu &> \mu_0 && \text{einseitig} \\ \text{oder } H_A : \mu &< \mu_0 && \text{einseitig} \end{aligned}$$

- Teststatistik:**

V : Anzahl X_i mit $X_i > \mu_0$

Verteilung der Teststatistik unter $H_0 : V \sim \text{Bin}(n, \pi_0)$, mit $\pi_0 = 0.5$

- Signifikanzniveau:** α

- Verwerfungsbereich für die Teststatistik:**

$$K = \begin{cases} [0, c_u] \cup [c_l, n], & \text{bei } H_A : \mu \neq \mu_0 \\ [0, c], & \text{bei } H_A : \mu < \mu_0 \\ [c, n], & \text{bei } H_A : \mu > \mu_0 \end{cases}$$

- Testentscheid:**

Überprüfen ob der beobachtete Wert der Teststatistik im Verwerfungsbereich K liegt.

3.8.8 Wilcoxon-Test

Voraussetzung: Realisierungen von X_1, X_2, \dots, X_n i.i.d., stetig und symmetrisch bezgl. $\mu = \mathbb{E}(X_i)$

Für Berechnung benutze R (5.7)

3.9 Statistik für zwei Stichproben

3.9.1 Gepaarte Stichprobe

Ligt vor falls:

- beide Versuchsbedingungen an derselben Versuchseinheit eingesetzt werden
- oder jeder Versuchseinheit aus der einen Gruppe genau eine Versuchseinheit aus der anderen Gruppe zugeordnet werden kann.

Die Daten entsprechen

$$x_1, \dots, x_n \text{ unter Versuchsbedingung 1}$$

$$y_1, \dots, y_n \text{ unter Versuchsbedingung 2}$$

wobei dasselbe n für beide nötig ist.

Gepoolte Varianz

$$S_{\text{pool}} = \sqrt{\frac{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2}{2}}$$

3.9.2 t-Test für gepaarte Stichproben

$$d_i = x_i - y_i, i \in \mathbb{N} \leq n$$

d_i seinen Realisierungen von D_1, \dots, D_n i.i.d. Somit vereinfacht sich die Betrachtung zu einer Variable auf welche wir den *t-Test* aus 3.8.4 anwenden können.

3.9.3 Ungepaarte Stichproben

Falls eine Paarung wie in 3.9.1 nicht möglich ist und die Daten

$$X_1, \dots, X_n \text{ i.i.d.}$$

$$Y_1, \dots, Y_m \text{ i.i.d.}$$

entsprechen, wobei $m \neq n$ nicht zwingend notwendig ist. Entscheidend ist, dass x_i und y_i zu verschiedenen Versuchseinheiten gehören und als unabhängig angenommen werden können.

3.9.4 t-Test für ungepaarte Stichproben

1. **Modell:**

$$X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu_X, \sigma^2)$$

$$Y_1, \dots, Y_m \text{ i.i.d. } \sim \mathcal{N}(\mu_Y, \sigma^2)$$

2. **Nullhypothese:**

$$H_0 : \mu_X = \mu_Y$$

Alternativhypothese:

$$\begin{aligned} \text{oder } H_A : \mu_X &\neq \mu_Y && \text{zweiseitig} \\ \text{oder } H_A : \mu_X &> \mu_Y && \text{einseitig} \\ \text{oder } H_A : \mu_X &< \mu_Y && \text{einseitig} \end{aligned}$$

3. **Teststatistik:**

$$T = \frac{\bar{X}_n - \bar{Y}_m}{S_{pool} \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

wobei

$$S_{pool} = \sqrt{\frac{1}{n+m-2} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{i=1}^m (Y_i - \bar{Y}_m)^2 \right)}$$

$$= \sqrt{\frac{(n-1)\hat{\sigma}_X^2 + (m-1)\hat{\sigma}_Y^2}{n+m-2}}$$

Verteilung der Teststatistik unter $H_0 : T \sim t_{n+m-2}$

4. **Signifikanzniveau:** α

5. **Verwerfungsbereich für die Teststatistik:**

$$K = \begin{cases} (-\infty, -t_{n+m-2; 1-\frac{\alpha}{2}}] \cup [t_{n+m-2; 1-\frac{\alpha}{2}}, \infty), & \text{bei } H_A : \mu_X \neq \mu_Y \\ (-\infty, -t_{n+m-2; 1-\alpha}], & \text{bei } H_A : \mu_X < \mu_Y \\ [t_{n+m-2; 1-\alpha}, \infty), & \text{bei } H_A : \mu_X > \mu_Y \end{cases}$$

6. **Testentscheid:**

Überprüfen ob der beobachtete Wert der Teststatistik im Verwerfungsbereich K liegt.

3.9.5 Zwei-Stichproben Wilcoxon-Test (Mann-Whitney Test)

Seien zwei Stichproben

$$X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu_X, \sigma^2)$$

$$Y_1, \dots, Y_m \text{ i.i.d. } \sim \mathcal{N}(\mu_Y, \sigma^2)$$

und F_X eine beliebige Verteilungsfunktion. Wir definieren nun

$$F_X(x) := F_X(x - \delta)$$

was einer verschobenen Funktion von F_X entspricht.

4 Regression

4.1 Einfache Lineare Regression

4.1.1 Modell

$$y_i = \beta_0 + \beta_1 x_i + E_i,$$

wobei $i \in \mathbb{N} \leq n$, $E_i \sim \mathcal{N}(0, \sigma^2)$, E_1, \dots, E_n i.i.d., $\mathbb{E}(E_i) = 0$ und $\text{Var}(E_i) = \sigma^2$
 Y bezeichnen wir als **Zielvariable (response variable)**, x als **erklärende Variable (explanatory/predictor variable)** oder **Co-Variable (covariate)** und E_i als Störfaktor (zufällig)

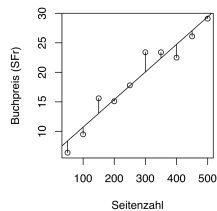


Abbildung 7: Einfache lineare Regression mit Residuen

4.1.2 Parameterschätzung

Das Modell aus 4.1.1 mit der *Methode der kleinsten Quadrate* liefert

$$\hat{\beta}_0, \hat{\beta}_1 \text{ Minimierung von } \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2,$$

daraus ergibt sich

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

und

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

dabei gilt

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \mathbb{E}(\hat{\beta}_1) = \beta_1$$

Für den **Standardfehler** gilt

$$s(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Die **Residuen**

$$R_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1)x_i, i \in \{1, 2, \dots, n\}$$

somit approximieren wir $E_i \approx R_i$ und daraus

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n R_i^2$$

4.2 Tests und Vertrauensintervalle der einfachen linearen Regression

4.2.1 t-Test in der Regression

1. **Modell:**

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

$$E_1, E_2, \dots, E_n \text{ i.i.d. } \mathcal{N}(0, \sigma_X^2)$$

2. **Nullhypothese:**

$$H_0 : \beta_1 = 0$$

Alternativhypothese:

$$H_A : \beta_1 \neq 0$$

3. **Teststatistik:**

$$T = \frac{\hat{\beta}_1 - 0}{\hat{s}(\hat{\beta}_1)} = \frac{\text{beobachtet} - \text{erwartet}}{\text{geschätzter Standardfehler}}$$

$$\text{Dabei ist } \hat{s} \text{ der geschätzte Standardfehler } \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}$$

Verteilung der Teststatistik unter $H_0 : T \sim t_{n-2}$

4. **Signifikanzniveau:** α

5. **Verwerfungsbereich für die Teststatistik:**

$$K = (-\infty, -t_{n-2; 1-\frac{\alpha}{2}}] \cup [t_{n-2; 1-\frac{\alpha}{2}}, \infty)$$

6. **Testentscheid:**

Überprüfen ob der beobachtete Wert der Teststatistik im Verwerfungsbereich K liegt.

Analog funktioniert auch ein *t-Test* für $H_0 : \beta_0 = 0, H_A : \beta_0 \neq 0$

4.2.2 t-Wert

$$\frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$$

4.2.3 P-Wert

Vgl. dazu 3.8.5, jedoch anstatt $n-1$ sind es hier $n-2$ Freiheitsgrade. Der P-Wert der Regression wird meist nicht von Hand berechnet (vgl. 5.9).

4.2.4 Vertrauensintervalle

Die zweiseitigen Vertrauensintervalle für $\beta_i (i = 0, 1)$ zum Niveau $1 - \alpha$ sind gegeben durch

$$[\hat{\beta}_i - \hat{s}(\hat{\beta}_i)t_{n-2; 1-\frac{\alpha}{2}}, \hat{\beta}_i + \hat{s}(\hat{\beta}_i)t_{n-2; 1-\frac{\alpha}{2}}]$$

Für grosse n approximieren wir $t_{n-2; 1-\frac{\alpha}{2}}$ mit $\Phi^{-1}(1 - \frac{\alpha}{2})$, somit für 95%-Vertrauensintervalle

$$[\hat{\beta}_i - 2\hat{s}(\hat{\beta}_i), \hat{\beta}_i + 2\hat{s}(\hat{\beta}_i)]$$

4.2.5 Bestimmtheitsmass R^2

Sei $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ der Wert auf der Regressionsgerade am Punkt x_i , dann gilt

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SS_Y} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SS_E} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SS_R}$$

wobei

- SS_Y : die totale Variation der Zielvariablen (ohne Einfluss der erklärenden Variablen x)
- SS_E : die Variation des Fehlers (Residuen-Quadratsumme)
- SS_R : die Variation, welche durch die Regression erklärt wird (Einfluss der erklärenden Variablen x).

Wir definieren

$$R^2 := \frac{SS_R}{SS_Y}, R^2 \in [0, 1]$$

als Mass für den Anteil der totalen Variation, welche durch die Regression erklärt wird.

Wenn R^2 gegen 1 geht ist es eine "gute" Regression.

$$R^2 = \hat{\rho}_{Y\hat{Y}}^2$$

4.2.6 Vorgehen bei einfacher linearer Regression

1. Plotten von Y und x in einem Streudiagramm. Überprüfen, ob eine lineare Regression überhaupt sinnvoll ist.
2. Anpassen der Regressionsgeraden; d.h. Berechnung der Punktschätzer $\hat{\beta}_0, \hat{\beta}_1$
3. Testen ob erklärende Variable x einen Einfluss auf die Zielvariable Y hat mittels *t-Test* für $H_0 : \beta_1 = 0$ und $H_A : \beta_1 \neq 0$. Falls dieser Test ein nicht-signifikantes Ergebnis liefert, so hat die erklärende Variable keinen signifikanten Einfluss auf die Zielvariable.
4. Testen ob Regression durch Nullpunkt geht mit *t-Test* für $H_0 : \beta_1 = 0$ und $H_A : \beta_1 \neq 0$. Falls dieser Test ein nicht-signifikantes Ergebnis liefert, so kann man das kleinere Modell mit Regression durch Nullpunkt benutzen (ohne Achsenabschnitt β_0).
5. Bei Interesse: Angabe von Vertrauensintervallen für $\hat{\beta}_0$ und $\hat{\beta}_1$.
6. Angabe des Bestimmtheitsmass R^2 . Dies ist in gewissem Sinne eine informellere (und zusätzliche) Quantifizierung als der statistische Test in Punkt 3.
7. Überprüfen der Modell-Voraussetzungen mittels Residuenanalyse (vgl. 4.3).

4.3 Residuenanalyse

Annahmen und deren Überprüfung:

1. $\mathbb{E}(E_i) = 0$ (*Tukey-Anscombe Plot*, vgl. 4.3.1)
Es gilt $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i + \mathbb{E}(E_i) = \beta_0 + \beta_1 x_i$, sodass keine systematischen Fehler auftreten können. Dennoch können Abweichungen auftreten (z.B. komplizierte quadr. Verteilung)
2. E_1, E_2, \dots, E_n i.i.d. (Plot bzgl. *serieller Korrelation*, *Tukey-Anscombe*)
Die Fehler müssen unabhängig voneinander sein, insbesondere sind $\text{Cor}(E_i, E_j) = 0$ für $i \neq j$, was bedeutet, dass keine *serielle Korrelation* auftritt. Da die Fehler gleich verteilt sein müssen, ist die Varianz der Fehler auch gleich.
3. E_1, E_2, \dots, E_n i.i.d. $\mathcal{N}(0, \sigma^2)$
Es wird angenommen, dass die Fehler normalverteilt sind. Überprüfung mit Normalplot der Residuen.

4.3.1 Tukey-Anscombe Plot

Plotten der Residuen R_i gegen die angepassten Werte \hat{y}_i . Idealerweise sind die Punkte gleichmässig um 0 gestreut. Bei verletzten Modellannahmen können auftreten:

- Kegelförmiges anwachsen von \hat{y}_i . Falls $\hat{y}_i > 0$ versuche $\log(Y_i) = \beta_0 + \beta_1 x_i + E_i$

- Ausreisser (Versuche robuste Regression)
- Unregelmässige Struktur (möglicherweise kein linearer Zusammenhang)

4.3.2 Serielle Korrelation

Überprüfung der Unabhängigkeitsannahme der E_1, E_2, \dots, E_n : Plotten von r_i gegen i .

Dabei sollte eine gleichmässige Verteilung um 0 entstehen.

4.3.3 Normaleplot

Wie in 3.6.2 erwarten wir möglichst eine Gerade, falls die Fehler normalverteilt sind.

4.4 Multiple lineare Regression

Oft sind erklärende Variablen $x_{i,1}, \dots, x_{i,p-1}$; ($p > 2$)

4.4.1 Modell

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{i,j} + E_i, i \in \mathbb{N} \leq n$$

$$E_1, E_2, \dots, E_i \text{ i.i.d., } \mathbb{E}(E_i) = 0, \text{Var}(E_i) = \sigma^2$$

In Matrixschreibweise:

$$\underbrace{Y}_{n \times 1} = \underbrace{X}_{n \times p} \times \underbrace{\beta}_{p \times 1} + \underbrace{E}_{n \times 1}$$

wobei:

- $Y = (Y_1, \dots, Y_n)^T$
- $X : (n \times p)$ -Matrix mit Spaltenvektoren $(1, 1, \dots, 1)^T, (x_{1,1}, x_{2,1}, \dots, x_{n,1})^T, \dots, (x_{1,p-1}, x_{2,p-1}, \dots, x_{n,p-1})^T$
- $\beta = (\beta_0, \dots, \beta_{p-1})$, der Parametervektor
- $E = (E_1, \dots, E_n)^T$, der Fehlervektor

Somit ist eine **einfache lineare Regression**

$$p = 2, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}^T$$

Analog dazu für **lineare Regression mit mehreren erklärenden Variablen**

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + E_i, i \in \mathbb{N} \leq n$$

$$p = 3, \quad X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}^T$$

ebenfalls für **lineare Regression mit quadratisch erklärenden Variablen**

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i, i \in \mathbb{N} \leq n$$

$$p = 3, \quad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}^T$$

und schlussendlich für eine **Regression mit transformierten erklärenden Variablen**

$$Y_i = \beta_0 + \beta_1 \log(x_{i,2}) + \beta_2 \sin(\pi x_{i,3}) + E_i, i \in \mathbb{N} \leq n$$

$$p = 3, \quad X = \begin{pmatrix} 1 & \log(x_{1,2}) & \sin(\pi x_{1,3}) \\ 1 & \log(x_{2,2}) & \sin(\pi x_{2,3}) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & \log(x_{n,2}) & \sin(\pi x_{n,3}) \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}^T$$

4.4.2 Interpretation

- Bei **einfacher linearer Regression** ist β_1 die erwartete Zunahme der Zielgrösse bei Erhöhung von x_1 um eine Einheit.
- Bei **multipler linearer Regression** ist β_i die erwartete Zunahme der Zielgrösse bei Erhöhung von x_i um eine Einheit - bei **Fixierung der anderen Variablen**.

4.4.3 Parameterschätzung und t-Test

Auch hier benutzen wir die *Methode der kleinsten Quadrate*.

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1} \text{ Minimierung von } \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1}))^2,$$

falls $p < n$

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Für die Fehlervarianz

$$\hat{\sigma} = \frac{1}{n-p} \sum_{i=1}^n R_i^2, R_i = Y_i - \left(\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{i,j} \right)$$

Den *t-Test* können wir analog zur *einfachen Regression* mit

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_A &: \beta_j \neq 0 \end{aligned}, j \in \mathbb{N} \leq p-1$$

durchführen. Dabei misst β_i den linearen Effekt der i -ten erklärenden Variable auf Zielvariable Y **nach Elimination** der linearen Effekte auf Y aller anderen Variablen. Es ist nicht möglich, durch direkte einfach lineare Regression von Y zur j -ten erklärenden Variable β_j zu erhalten!

4.4.4 F-Test

Prüft, ob es mindestens eine erklärende Variable gibt, die einen signifikanten Effekt auf die Zielvariable hat.

$$\begin{aligned} H_0 &: \beta_1 = \dots = \beta_{p-1} = 0 \\ H_A &: \text{mindestens ein } \beta_j \neq 0, \quad j \in \mathbb{N} \leq p-1 \end{aligned}$$

Hier können einzelne Variablen signifikant sein und andere nicht. Bei starker Korrelation zwischen zwei kann man eine weglassen, da keine neue Information.

4.4.5 Bestimmtheitsmass R^2

Es gilt wie in 4.2.5

$$R^2 = \hat{\rho}_{Y\hat{Y}}^2$$

5 R

5.1 diskrete Verteilungen

```
1 # d... berechnet P(X = x)
2 # p... berechnet P(X <= x)
3 # q... berechnet Quantile der Verteilung
4 # r... zieht eine bestimmte Anzahl Realisierungen der gewählten Verteilung
```

5.1.1 Binomialverteilung

```
1 dbinom(x = 5, size = 10, prob = 0.5) # Berechnet P(X = 5) fuer X ~ Binomial(10, 0.5)
2 pbinom(q = 5, size = 10, prob = 0.5) # Berechnet P(X <= 5) fuer X ~ Binomial(10, 0.5)
3 qbinom(p = 0.2, size = 10, prob = 0.5) # Berechnet das 20%-Quantil fuer X ~ Binomial(10, 0.5)
4 rbinom(n = 100, size = 10, prob = 0.5) # Zieht zufaellig n=100 Realisierungen von X ~ Binomial
  (10, 0.5)
5 # (fuehren Sie den oberen Befehl rbinom 2x aus, Sie erhalten andere Werte)
```

5.1.2 Poissonverteilung

```
1 dpois(x = 5, lambda = 2) # Berechnet P(X = 5) fuer X ~ Poisson(2)
2 ppois(q = 5, lambda = 2) # Berechnet P(X <= 5) fuer X ~ Poisson(2)
3 qpois(p = 0.2, lambda = 2) # Berechnet das 20%-Quantil fuer X ~ Poisson(2)
4 rpois(n = 100, lambda = 2) # Zieht n=100 Realisierungen von X ~ Poisson(2)
```

5.1.3 Binomialtest

```
1 ## Der Binomialtest kann in R mit dem Befehl binom.test(...) durchgefuehrt werden.
2 ## Die Argumente der Funktion sind:
3 ## - x: Der beobachtete Wert der Teststatistik
4 ## - n, p: Die Parameter der Verteilung der Teststatistik (Binomial(n,p)) unter der
  Nullhypothese
5 ## - alternative:
6 ## Die Wahl der Alternativhypothese. Moegliche Optionen sind:
7 ## - "less" fuer H_A: pi < pi_0
8 ## - "greater" fuer H_A: pi > pi_0
9 ## - "two.sided" fuer H_A: pi ungleich pi_0
10 ## - conf.level:
11 ## Das Konfidenzniveau fuer das Vertrauensintervall. Entspricht (1 - Signifikanzniveau).
12
13 ## Beispiel:
14 ## Wir vermuten, dass ein Wuerfel zu viele 6er wuerfelt.
15 ## Wir haben in 50 Wuerfen 13 mal eine sechs gewuerfelt.
16 ## Wir moechten auf dem 1%-Signifikanzniveau testen, ob der Wuerfel zu viele 6er wuerfelt.
17 binom.test(x = 13, n = 50, p = 1/6, alternative = "greater", conf.level = 0.99)
```

5.2 Kennzahlen

```
1 ## Wir haben folgende Daten beobachtet / gemessen
2 x <- c(1.1, 2.3, -2.4, 3.9, 5.1, -1.7, 2.0, -1.1, 3.4, 0.7)
3 y <- c(0.8, 2.1, -1.3, 1.0, 0.4, -3.2, 3.1, -0.1, 5.1, 4.3)
4
5 mean(x) # arithmetisches Mittel
6 var(x) # Varianz
7 sd(x) # Standardabweichung
8
9 max(x) # Maximum
10 min(x) # Minimum
11
12 median(x) # Median
13 quantile(x, probs = 0.25) # empirisches 25%-Quantil
14
15 summary(x) # Gibt Ueberblick ueber einige Kennzahlen
16
17 cor(x,y) # Empirische Korrelatin von x und y
```

5.3 Grafische Methoden

```
1 plot(x, y) # Streudiagramm von x und y
2 hist(x) # Histogramm Typ "Frequency" (siehe VL 8)
3 hist(x, freq = FALSE) # Histogramm Typ "Density" (siehe VL 8)
4 hist(x, breaks = 10) # mit breaks = ... kann die Anzahl Balken gesteuert werden, siehe Serie 8)
5 plot(ecdf(x)) # Empirische kumulative Verteilungsfunktion
6 boxplot(x) # Boxplot
7
8 z <- rnorm(n = 100, mean = 2, sd = 1)
9 qqnorm(z) # QQ-Plot, welcher mit den theoretischen Quantilen der N(0,1)-Verteilung vergleicht.
```

5.4 Stetige Verteilungen

5.4.1 Uniformverteilung

```
1 dunif(x = 2.5, min = 1, max = 3) # Wert der Dichte f(x) von X ~ Uniform([1,3]) an der Stelle x=2.5
2 punif(q = 2.5, min = 1, max = 3) # Wert der kum. Verteilungsfkt. F(x) von X ~ Uniform([1,3]) an
  der Stelle x=2.5
3 qunif(p = 0.2, min = 1, max = 3) # 20%-Quantil von X ~ Uniform([1,3])
4 runif(n = 100, min = 1, max = 3) # Zieht zufaellig 100 Realisierungen von X ~ Uniform([1,3])
```

5.4.2 Exponentialverteilung

```
1 dexp(x = 2, rate = 1) # Wert der Dichte f(x) von X ~ Exp(1) an der Stelle x = 2
2 pexp(q = 2, rate = 1) # Wert der kum. Verteilungsfunktion F(x) von X ~ Exp(1) an der Stelle x
  = 2
3 qexp(p = 0.2, rate = 1) # 20%-Quantil von X ~ Exp(1)
4 rexp(n = 100, rate = 1) # Zieht zufaellig 100 Realisierungen von X ~ Exp(1)
```

5.4.3 Normalverteilung

```
1 dnorm(x = 3, mean = 1, sd = sqrt(2)) # Wert der Dichte f(x) von X ~ N(1,2) an der Stelle x = 3
2 pnorm(q = 3, mean = 1, sd = sqrt(2)) # Wert der kum. VF F(x) von X ~ N(1,2) an der Stelle x =
  3
3 qnorm(p = 0.2, mean = 1, sd = sqrt(2)) # 20%-Quantil von X ~ N(1,2)
4 rnorm(n = 100, mean = 1, sd = sqrt(2)) # Zieht zufaellig 100 Realisierungen von X ~ N(1,2)
```

5.4.4 Standardnormalverteilung

```
1 dnorm(x = 3) # Wenn man meam = ..., sd = ... nicht angibt, wird eine N(0,1)-Verteilung
  angenommen.
2 pnorm(q = 3)
3 qnorm(p = 0.2) # entspricht Phi^{-1}(0.2)
4 rnorm(n = 100)
```

5.5 Ein-Stichproben t-Test (gepaart)

```
1 ## Der Ein-Stichproben t-Test kann in R mit dem Befehl t.test(...) durchgefuehrt werden.
2 ## Die benoetigten Argumente der Funktion sind:
3 ## - x: Der Vektor mit den beobachteten Werten
4 ## - mu: Der Wert mu_0 der Nullhypothese
5 ## - alternative:
6 ## Die Wahl der Alternativhypothese. Moegliche Optionen sind:
7 ## - "less" fuer H_A: mu < mu_0
8 ## - "greater" fuer H_A: mu > mu_0
9 ## - "two.sided" fuer H_A: mu ungleich mu_0
10 ## - conf.level:
11 ## Das Konfidenzniveau fuer das Vertrauensintervall. Entspricht (1 - Signifikanzniveau).
12
13 t.test(x = x, y = y, alternative = "greater", mu = 0, paired = TRUE, conf.level = 0.95)
```

5.6 Zwei-Stichproben t-Test (ungepaart)

```
1 ## Um in R einen ungepaarten Zwei-Stichproben t-Test durchzufuehren, verwenden
2 ## Sie ebenfalls die Funktion t.test(...) mit den Argumenten
3
4 ## - x: Der Vektor mit den beobachteten Werten der ersten Stichprobe
5 ## - y: Der Vektor mit den beobachteten Werten der zweiten Stichprobe
6 ## - mu: Der Wert mu_0 der Nullhypothese (typischerweise = 0, da Nullhypothese: "Es gibt
  keinen Unterschied")
7 ## - alternative:
8 ## Die Wahl der Alternativhypothese. Moegliche Optionen sind:
9 ## - "less" fuer H_A: mu_X - mu_Y < mu_0
10 ## - "greater" fuer H_A: mu_X - mu_Y > mu_0
11 ## - "two.sided" fuer H_A: mu_X - mu_Y ungleich mu_0
12 ## - paired = FALSE (ungepaarter Test)
13 ## - var.equal = TRUE (standardmaessig ist var.equal = FALSE, dann wird ein Welch-Test
  durchgefuehrt)
14 ## - conf.level:
15 ## Das Konfidenzniveau fuer das Vertrauensintervall. Entspricht (1 - Signifikanzniveau).
16
17 t.test(x = x, y = y, alternative = "two.sided", mu = 80.00, paired = FALSE, conf.level = 0.95)
```

5.7 Wilcoxon-Test

```
1 ## Ein- und Zwei-Stichproben Wilcoxon Tests stehen in R unter dem Befehl wilcox.test(...) zur
  Verfuegung.
2 ## Die Argumente der Funktion sind analog zu denjenigen der t-Tests.
3 wilcox.test(x = x, alternative = "greater", mu = 80)
```

5.8 Verteilungen

`pt` für kumulative Verteilungsfunktion
`qt` für Quantile

5.9 Regression

5.9.1 Einfache Lineare Regression

```
1 ## Um in R ein einfaches lineares Regressionsmodell anzupassen, verwendet man den R-Befehl lm
  (...).
2
3 ## Wir betrachten das Beispiel mit Buchpreis und Seitenzahl aus dem Vorlesungsskript.
4 x <- c(50, 100, 150, 200, 250, 300, 350, 400, 450, 500) ## Seitenzahl, erklärende Variable.
5 y <- c(9.9, 10.7, 13.3, 15.2, 16.4, 23.6, 23.5, 21.1, 28.9, 29.1) ## Buchpreis, Zielvariable.
```

Eigentliche Regression:

```
1 ## Um das lineare Regressionsmodell  $Y_i = \beta_0 + \beta_1 x_i + E_i$  zu fitten, benutzt man
2 fit <- lm(y ~ x) #("y gegen x")
3
4 ## Man kann Achsenabschnitt und Steigung sehen, wenn man sich das Objekt 'fit' anschaut:
5 fit
```

oder

```
1 fit <- lm(y ~ x)
2 summary(fit)
```

liefert den Output

```
1 Residuals:
2   Min       1Q   Median       3Q      Max
3 -3.6958 -0.5944 -0.2203  0.9300  3.3048
4
5 Coefficients:
6             Estimate Std. Error t value Pr(>|t|)
7 (Intercept)  6.793333   1.391060   4.884  0.00122 **
8 x            0.045006   0.004484  10.037  8.25e-06 ***
9 ---
10 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
11
12 Residual standard error: 2.036 on 8 degrees of freedom
13 Multiple R-squared:  0.9264, Adjusted R-squared:  0.9172
14 F-statistic: 100.8 on 1 and 8 DF, p-value: 8.254e-06
```

somit $Y_i = 6.793333 + 0.045006x_i$

Weitere Plots

```
1 ## Residuenplots erhaelt man einfach mittels
2 plot(fit) # man muss in der "Console" mehrmals die Eingabetaste dr cken, um die Plots zu sehen.
3
4 ## oder:
5 plot(fit$fitted, fit$resid) ## Tukey-Anscombe plot
6 qqnorm(fit$residuals) ## qq-Plot der Residuen
7
8 ## 95%-Vertrauensintervalle f r Koeffizienten (siehe VL 14, Slide 8)
9 confint(fit)
10
11 ## 95%-Vertrauens-/Vorhersageintervalle (siehe VL 14, Slides 9 und 10)
12 nd <- data.frame(x=1, y=NA)
13 predict(fit, newdata = nd, interval = "confidence") ## Vertrauensintervall
14 predict(fit, newdata = nd, interval = "prediction") ## Vorhersageintervall
15
16 ## Nehmen wir an, die Daten befinden sich in einem data.frame (anstelle von zwei Vektoren).
17 Daten_Buch <- data.frame(Seitenzahl = x, Buchpreis = y)
18 Daten_Buch
19
20 ## Dieselbe Regression wie oben kann man nun berechnen, indem man entweder schreibt:
21 fit2 <- lm(Daten_Buch$Buchpreis ~ Daten_Buch$Seitenzahl)
22 summary(fit2)
23
```

```
24 ## oder alternativ:
25 fit3 <- lm(Buchpreis ~ Seitenzahl, data = Daten_Buch)
26 summary(fit3)
27
28 ## Alle 3 Varianten (fit, fit2, fit3) liefern exakt dasselbe Resultat.
```

5.9.2 Multiple lineare Regression

```
1 ## Um in R ein multiples lineares Regressionsmodell anzupassen, verwendet man ebenfalls den R-Befehl
  lm(...).
2
3 ## Wir betrachten das Beispiel mit Buchpreis und Seitenzahl aus dem Vorlesungsskript, moechten nun
  jedoch
4 ## als zweite erklärende Variable noch das Erscheinungsjahr des Buches ins Modell aufnehmen.
5
6 x1 <- c(50, 100, 150, 200, 250, 300, 350, 400, 450, 500) ## Seitenzahl, erklärende Variable 1.
7 x2 <- c(2017, 1999, 2013, 2004, 2001, 1979, 2018, 2008, 2015, 2002) ## Erscheinungsjahr, erklärende
  Variable 2.
8 y <- c(9.9, 10.7, 13.3, 15.2, 16.4, 23.6, 23.5, 21.1, 28.9, 29.1) ## Buchpreis, Zielvariable.
9
10 ## Das multiple lineare Regressionsmodell  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + E_i$  berechnet man
11 ## mit dem Befehl:
12 fit <- lm(y ~ x1 + x2)
13
14 ## Die restlichen Befehle sind analog zur einfachen linearen Regression.
```

6 Anhang

Referenzen

1. Skript "Vorlesungsskript Mathematik IV für Agrarwissenschaften, Erdwissenschaften, Lebensmittelwissenschaften und Umweltnaturwissenschaften", Dr. Jan Ernest, HS19
2. Statistik_MatheIV.pdf, scmelina, HS18
3. ZF_Statistik_ClemenceBoutry.pdf ,clboutry, FS16

Bildquellen

- Abb. 1: Skbkakas, https://upload.wikimedia.org/wikipedia/commons/1/16/Poisson_pmf.svg
- Abb. 2: DanielPenfield, https://upload.wikimedia.org/wikipedia/commons/c/c3/Histogram_of_arrivals_per_minute.svg
- Abb. 3: towardsdatascience.com, <https://towardsdatascience.com/understanding-boxplots-5e2df7bcdb5>
- Abb. 4: DanielPenfield, https://upload.wikimedia.org/wikipedia/commons/a/af/Scatter_diagram_for_quality_characteristic_XXX.svg
- Abb. 5: Skript
- Abb. 7: Skript



Dieses Dokument ist unter (CC BY-SA 4.0) freigegeben

📄 <https://n.ethz.ch/~jannisip>
git <https://git.thisfro.ch/thisfro/statistik-zf>
Jannis Portmann, HS19