Statistische Methoden

Marianne Müller

25. September 2003

Inhaltsverzeichnis

1. Deskriptive Statistik

Die deskriptive Statistik (beschreibende Statistik, explorative Datenanalyse) versucht, das Wesentliche eines Zahlenhaufens zu beschreiben, um die Daten zu verstehen oder präsentieren zu können. Daten bestehen aus Messungen oder Beobachtungen an mehreren Versuchseinheiten, z. B. Personen oder Tiere. Die Grössen oder Merkmale, die erfasst werden, heissen Variablen (sie variieren zwischen den Versuchseinheiten).
Zu allererst interessiert die Verteilung einer Variablen: welche Werte werden wie oft angenommen. Dazu macht man eine Häufigkeitstabelle und stellt diese graphisch dar mit einem Kuchen- oder Balkendiagramm, bzw. mit einem Histogramm. Andere Darstellungsmöglichkeiten sind Stem-and-Leaf-Plot und Boxplot. Die Form der Verteilung sollte in diesen Graphiken erkennbar werden: symmetrisch, links- oder rechtsschief, bimodal.
Um die mittlere Lage und die Streuung einer Variablen zu beschreiben, werden statistische Kennzahlen berechnet, wie z. B. arithmetisches Mittel, Median, Quartilsdifferenz und Standardabweichung. Kennzahlen werden auch Statistiken genannt.

- **Lagemasse:**
  - arithmet. Mittel $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$
  - Median : 50% aller Beobachtungen sind kleiner gleich diesem Wert.
  - Modus: der häufigste Wert.
  - Quartile, Perzentile: Das 1. Quartil Q_1 teilt die Daten im Verhältnis 25 : 75. Das 2. Quartil ist der Median und das 3. Quartil Q_3 wird so bestimmt, dass 75% aller Werte kleiner und 25% grösser als Q_3 sind. Das k. Perzentil unterteilt die Daten im Verhältnis k zu 100-k.
- **Streuungsmasse**:
  - Spannweite = Maximum − Minimum
  - Varianz, Standardabweichung:

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 \qquad s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

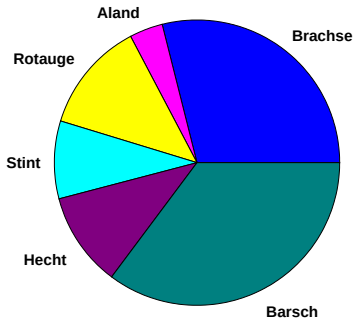
  - Quartilsdifferenz: $IQR = Q_3 - Q_1$

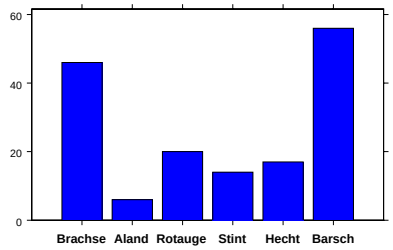
Bei symmetrisch verteilten Daten hat die Standardabweichung zusammen mit $\bar{x}$ eine spezielle Bedeutung:

> Etwa 95% aller Einzelmessungen liegen innerhalb von $\pm 2 \cdot s$ um $\bar{x}$ herum; 2.5% der Werte sind kleiner als $\bar{x} - 2 \cdot s$ und 2.5% der Werte sind grösser als $\bar{x} + 2 \cdot s$.

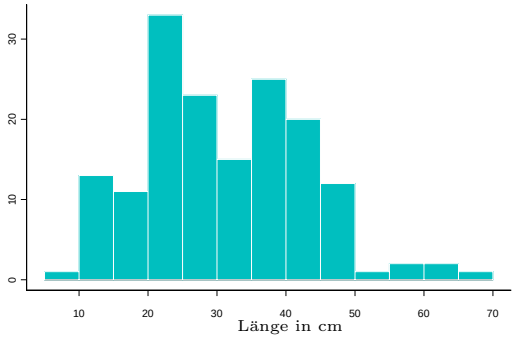
Welche numerische oder graphische Beschreibung den Daten angemessen ist, hängt vor allem vom Typ der Daten ab: Nominal- oder Ordinaldaten, bzw. diskrete oder stetige Daten.

Fischgattungen: Pie- und Barchart





Histogramm



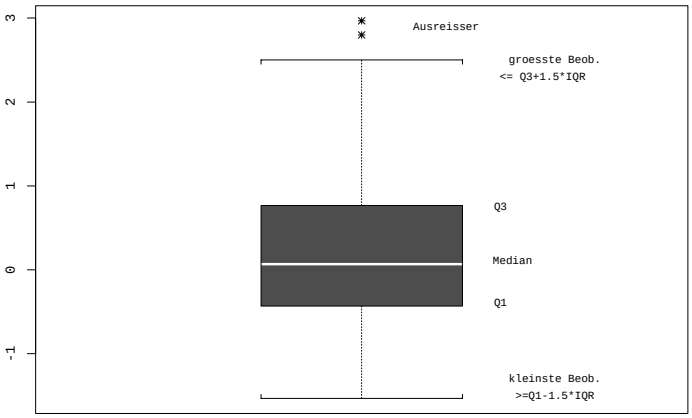

```
> stem(fish$length3)

N = 159   Median = 29.4 Quartiles = 23.1, 39.7

Decimal point is 1 place to the right of the colon

  0 : 9
  1 : 122223333334
  1 : 55666677899
  2 : 000111122222333333333334444444
  2 : 55555556667777888999999999
  3 : 001111234444
  3 : 555556666677778888999999999
  4 : 0000011111111222223344
  4 : 5555556666789
  5 : 1
  5 : 5
  6 : 044
  6 : 8
```

Definition des Boxplots



2. Zufallsvariablen

2.1. Zusammenhang: Wirklichkeit–Modell

Wirklichkeit	Modell
Stichprobe	Population
Daten	Zufallsvariable
diskret	diskret
stetig	stetig
rel. Häufigkeit	Wahrscheinlichkeit
Häufigkeitstabelle	Wahrscheinlichkeitsverteilung
Stabdiagramm	Stabdiagramm
Histogramm	Dichte
empir. Verteilungsfunktion	Verteilungsfunktion
empir. Kennzahlen	theoret. Kennzahlen
Mittelwert $\bar{x}$	Erwartungswert $E(X)$
Varianz s^2	Varianz $Var(X)$

Eine *Zufallsvariable* (random variable) ist eine quantitative Variable, deren Wert durch das zufälllige Ergebnis von Experimenten oder Beobachtungen bestimmt wird. Zufallsvariablen bilden ein Modell für die beobachteten Grössen, die Daten.
Es gibt *diskrete* (discrete) und *stetige* (continuous) Zufallsvariablen. Diskrete Zufallsvariablen haben nur eine endliche oder abzählbare Anzahl möglicher Werte, stetige Zufallsvariablen können alle Werte innerhalb eines Intervalls der reellen Zahlen annehmen.

2.2. Diskrete Zufallsvariablen

Die *Wahrscheinlichkeitsverteilung* p (probability function) ist definiert durch:

X	x_1	x_2	...	x_k
p	$p(x_1)$	$p(x_2)$	...	$p(x_k)$

$$p(x_i) := P(X = x_i), \quad \sum p(x_i) = 1.$$

Die *kumulative Verteilungsfunktion* F (cumulative distribution function) ist definiert durch:

$$F(x) = P(X \leq x), \qquad -\infty < x < \infty.$$

F ist monoton wachsend, $\lim_{x\to-\infty} F(x) = 0$ und $\lim_{x\to\infty} F(x) = 1$.

Uniforme Verteilung

Eine Verteilung, bei der alle Werte die gleiche Wahrscheinlichkeit haben, heisst *uniform*.

X	x_1	x_2	...	x_n
p	$1/n$	$1/n$	...	$1/n$

.

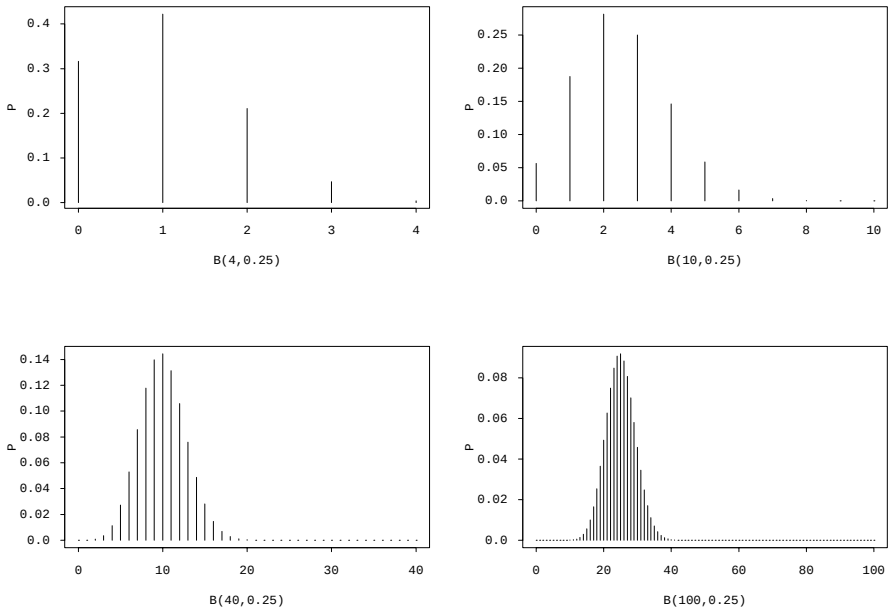


Abbildung 2.1.: Binomialverteilungen

Bernoulli Verteilung

Ein Experiment habe zwei mögliche Ausgänge, "Erfolg" und "Misserfolg", mit den Wahrscheinlichkeiten p und $1-p$.

$$X = \begin{cases} 0 & : \quad \text{"Misserfolg"} \\ 1 & : \quad \text{"Erfolg"} \end{cases}$$

X	0	1
p(x)	1-p	p

Binomialverteilung

Es werden n voneinander unabhängige Versuche gemacht. Jeder einzelne Versuch hat zwei mögliche Ausgänge, „Erfolg“ und „Misserfolg“. Die Wahrscheinlichkeit p für einen Erfolg ist konstant. Die Anzahl Erfolge X hat dann eine Binomialverteilung $\mathcal{B}(n,p)$ und die Wahrscheinlichkeit für k Erfolge ist gegeben durch:

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{für } k = 0, 1, \ldots, n.$$

Geometrische Verteilung

Es werden unabhängige Versuche durchgeführt. Jeder einzelne Versuch hat zwei mögliche Ausgänge, „Erfolg“ und „Misserfolg“. Die Wahrscheinlichkeit p für einen Erfolg ist konstant. X ist die Anzahl Versuche bis und mit dem ersten Erfolg.

$$P(X=k) = (1-p)^{k-1} p \qquad \text{für } k = 1, 2, 3, \ldots.$$

Negative Binomialverteilung

Es werden unabhängige Versuche durchgeführt. Jeder einzelne Versuch hat zwei mögliche Ausgänge, „Erfolg“ und „Misserfolg“. Die Wahrscheinlichkeit p für einen Erfolg ist konstant.

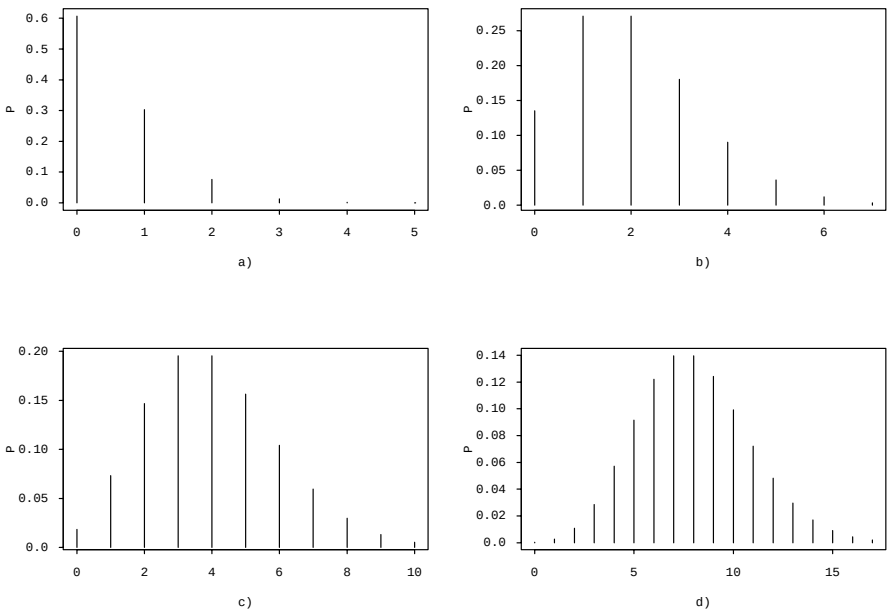


Abbildung 2.2.: Poissonverteilungen mit a) $\lambda = 0.5$, b) $\lambda = 2$, c) $\lambda = 4$, d) $\lambda = 8$

X ist die Anzahl Misserfolge bis r Erfolge eingetreten sind.

$$P(X = k) = \binom{k+r-1}{k} p^r (1-p)^k \qquad \text{für } k = 0, 1, 2, \ldots.$$

Poissonverteilung
Betrachte die absolute Häufigkeit, mit der ein bestimmtes Ereignis eintritt. Wenn die Ereignisse unabhängig voneinander mit einer konstanten Rate λ passieren, dann hat die Anzahl Ereignisse X eine Poissonverteilung $\mathcal{P}(\lambda)$. Die Wahrscheinlichkeit für k Ereignisse pro Zeiteinheit ist:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k = 0, 1, 2, \ldots$$

Für n gross und p klein ist die Poissonverteilung eine Näherung für die Binomialverteilung mit $\lambda = np$.
Ein Prozess, der innerhalb eines festen zeitlichen oder räumlichen Intervalls eine Anzahl Ereignisse erzeugt, die einer Poissonverteilung folgt, heisst *Poissonprozess*.

2.3. Stetige Zufallsvariablen

Die *Dichte* f (density) ist eine stückweise stetige Funktion mit $f(x) \geq 0$ und $\int_{-\infty}^{\infty} f(x)\,dx$.
Wenn X eine stetige Zufallsvariable mit Dichte f ist, dann gilt:

$$P(a < X < b) = \int_a^b f(x)\,dx \quad \text{für } a < b.$$

Die *kumulative Verteilungsfunktion* F (cumulative distribution function) ist definiert durch:

$$F(x) = P(X \leq x), \qquad -\infty < x < \infty.$$

Es gilt:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

Das α–*Quantil* x_α ist definiert durch: $F(x_\alpha) = \alpha$. Für $\alpha = 1/2$ erhält man den *Median*, für $\alpha = 1/4$ und $\alpha = 3/4$ das 1. und das 3. *Quartil*.

Uniforme Verteilung
Die Dichte einer uniformverteilten Zufallsvariablen ist:

$$f(x) = \frac{1}{b-a} \quad \text{für } a \leq x \leq b.$$

Die Verteilungsfunktion ist:

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

Exponentialverteilung
Modell für Warte- oder Ueberlebenszeiten. Wird in einem Poissonprozess mit Paramter λ statt der Anzahl Ereignisse in einem bestimmten Zeitintervall die Dauer bis zum Eintreten des nächsten Ereignisses betrachtet, so ist diese Dauer exponentialverteilt, $Exp(\lambda)$. Die Exponentialverteilung ist *gedächtnislos* (memoryless).
Die Dichte einer exponentialverteilten Zufallsvariablen ist:

$$f(x) = \lambda e^{-\lambda x} \quad \text{für } x \geq 0, \qquad \lambda > 0.$$

Die Verteilungsfunktion ist:

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

Gammaverteilung
Die Dichte einer gammaverteilten Zufallsvariablen ist:

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \quad \text{für } x \geq 0, \qquad \alpha > 0, \lambda > 0.$$

Die Gammafunktion $\Gamma(x)$ ist definiert durch: $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du \quad x > 0$.
Es gilt: $\Gamma(1) = 1, \quad \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ für $\alpha > 1, \quad \Gamma(\alpha) = (\alpha-1)!$, wenn α eine ganze Zahl grösser als 1 ist.

Normalverteilung
Die Normalverteilung ist das weitaus häufigste Modell für Messdaten. Entwickelt wurde sie als Modell für Messfehler, sie passt aber oft auch in andern Situationen recht gut. Das

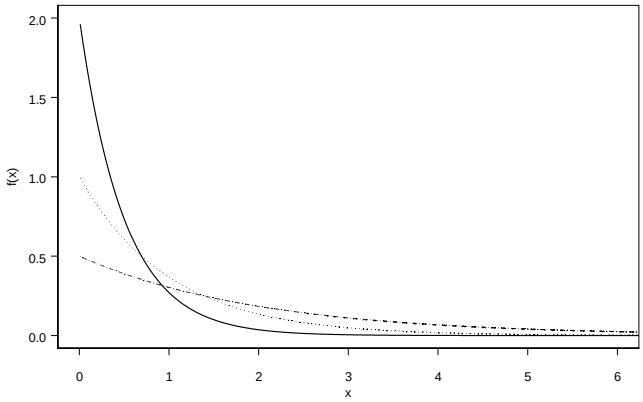


Abbildung 2.3.: Exponentialverteilungen mit $\lambda = 0.5$ $(- \cdot - \cdot)$, $\lambda = 1$ $(\cdots\cdots)$, $\lambda = 2$ (—)

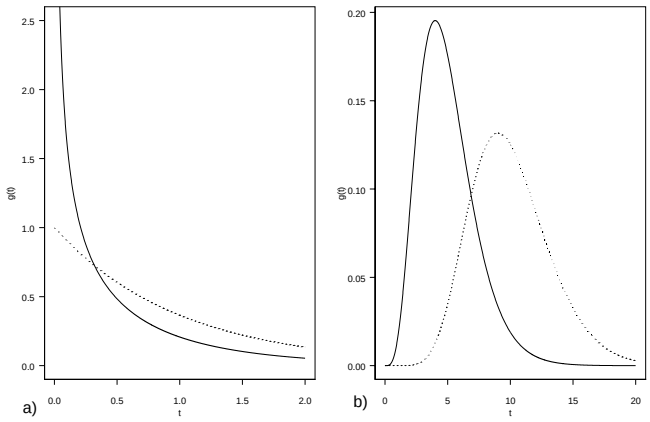


Abbildung 2.4.: Gammaverteilungen mit $\lambda = 1$ und a) $\alpha = 0.5$ (—) und $\alpha = 1$ $(\cdots)$,
b) $\alpha = 5$ (—) und $\alpha = 10$ $(\cdots)$

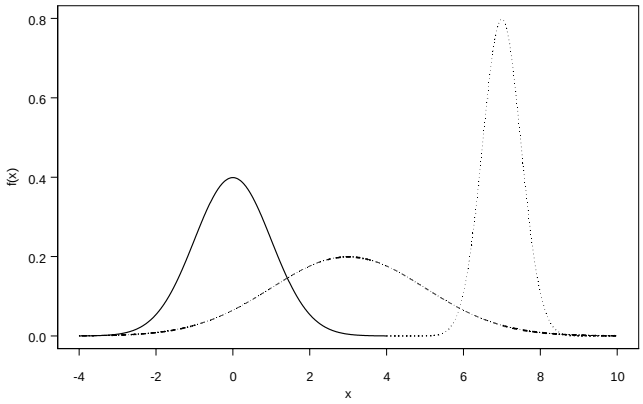


Abbildung 2.5.: Dichtekurven von $\mathcal{N}(0,1)$, $\mathcal{N}(3,4)$ und $\mathcal{N}(7,1/4)$

hat sich empirisch gezeigt und ein mathematisches Resultat, der *Zentrale Grenzwertsatz*, bestätigt das. Ein grosser Teil der statistischen Methoden setzt Normalverteilung voraus. Die Dichte einer Zufallsvariablen mit Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ ist gegeben durch:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad -\infty < x < +\infty, \qquad -\infty < \mu < +\infty, \quad \sigma > 0.$$

Die spezielle Normalverteilung $\mathcal{N}(0,1)$ heisst *Standardnormalverteilung*. Für Dichte und kumulative Verteilungsfunktion verwendet man in diesem Fall die Bezeichnungen ϕ und Φ.

Chiquadrat-Verteilung

Seien $Z_1, \ldots, Z_n \sim \mathcal{N}(0,1)$, iid. Dann hat $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ eine Chiquadrat-Verteilung mit n Freiheitsgraden, $X \sim \chi_n^2$.

t-Verteilung

Seien $Z \sim \mathcal{N}(0,1)$ und $X \sim \chi_n^2$ unabhängige Zufallsvariablen. Dann hat $T = \dfrac{Z}{\sqrt{X/n}}$ eine t-Verteilung mit n Freiheitsgraden, $T \sim t_n$.

F-Verteilung

Seien $X_1 \sim \chi_n^2$ und $X_2 \sim \chi_m^2$ unabhängige Zufallsvariablen. Dann hat $F = \dfrac{X_1/n}{X_2/m}$ eine F-Verteilung mit n und m Freiheitsgraden, $F \sim F_{n,m}$.

2.4. Erwartungswert und Varianz

Sei X eine diskrete Zufallsvariable mit Wahrscheinlichkeitsfunktion p. Der *Erwartungswert* von X (expected value, mean) ist definiert als:

$$E(X) = \sum_i x_i p(x_i),$$

falls die Summe existiert. Oft wird der Erwartungswert mit μ bezeichnet.

Sei X eine stetige Zufallsvariable mit Dichte f. Der *Erwartungswert* $E(X)$ von X ist definiert als:

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

falls das Integral existiert.

Sei X eine Zufallsvariable mit Erwartungswert $E(X)$. Dann ist die *Varianz* von X (variance) gegeben durch:

$$Var(X) = E\{[X - E(X)]^2\},$$

falls der Erwartungswert existiert. Die *Standardabweichung* von X (standard deviation) ist die Wurzel aus der Varianz. Oft wird die Varianz mit σ^2 und die Standardabweichung mit σ bezeichnet.
Wenn X diskret ist, gilt mit $E(X) = \mu$:

$$Var(X) = \sum_i (x_i - \mu)^2 p(x_i),$$

für X stetig:

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx.$$

Rechenregeln

Seien X_1 und X_2 Zufallsvariablen und a, b konstante Zahlen. Dann gilt:

$$\begin{aligned} E(X_1 + X_2) &= E(X_1) + E(X_2) \\ E(a + bX_1) &= a + bE(X_1) \\ Var(X_1 + X_2) &= Var(X_1) + Var(X_2) \text{ , falls } X_1 \text{ und } X_2 \text{ unabhängig sind} \\ Var(a + bX_1) &= b^2 Var(X_1). \end{aligned}$$

Zusammenstellung der wichtigsten Verteilungen

Verteilung	$\boldsymbol{P(X = k)}$ **bzw.** $\boldsymbol{f(x)}$	Wertebereich	$\boldsymbol{E(X)}$	$\boldsymbol{Var(X)}$
Binomial	$\binom{n}{k} p^k (1-p)^{n-k}$	$k = 0, 1, \ldots, n$	np	$np(1-p)$
Neg. Binomial	$\binom{k+r-1}{k} p^r (1-p)^k$	$k = 0, 1, \ldots$	$r\frac{1-p}{p}$	$r\frac{1-p}{p^2}$
Geometrisch	$(1-p)^{k-1} p$	$k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$\frac{\lambda^k e^{-\lambda}}{k!}$	$k = 0, 1, \ldots$	λ	λ
Uniform	$\frac{1}{b-a}$	$a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$	$-\infty < x < \infty$	μ	σ^2
Gamma	$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$	$x > 0$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Exponential	$\lambda e^{-\lambda x}$	$x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Chiquadrat	$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\, x^{\frac{n}{2}-1} e^{-x/2}$	$x > 0$	n	$2n$
t	$f(x) = \frac{\Gamma[\frac{n+1}{2}]}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$	$-\infty < x < \infty$	0	$\frac{n}{n-2}$
F	$f(x) = \frac{\Gamma[\frac{n+m}{2}]}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} (\frac{n}{m})^{\frac{n}{2}} x^{\frac{n}{2}-1} \left(1 + \frac{n}{m}x\right)^{-\frac{n+m}{2}}$	$x > 0$	$\frac{n}{n-2}$	

2.5. Kovarianz und Korrelation

Seien X und Y gemeinsam verteilte Zufallsvariablen mit Erwartungswerten μ_X und μ_Y. Dann ist die *Kovarianz* von X und Y (covariance) gegeben durch:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)],$$

falls der Erwartungswert existiert.

Seien X_1, X_2, Y_1 und Y_2 Zufallsvariablen und a, b, c und d konstante Zahlen, dann gilt:

$$\begin{aligned} Cov(X_1+X_2, Y_1+Y_2) &= Cov(X_1,Y_1)+Cov(X_1,Y_2)+Cov(X_2,Y_1)+Cov(X_2,Y_2) \\ Cov(a+bX_1, c+dY_1) &= bdCov(X_1,Y_1). \end{aligned}$$

Daraus folgt:

$$Var(aX_1+bX_2) = a^2Var(X_1)+b^2Var(X_2)+2abCov(X_1,X_2).$$

Seien X und Y gemeinsam verteilte Zufallsvariablen mit Varianzen verschieden von Null. Dann ist die *Korrelation* von X und Y (correlation) gegeben durch:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Es gilt: $-1 \leq \rho \leq 1$ und $\rho = \pm 1$ für $Y = a+bX$.

2.6. Anwendungsbereiche von verschiedenen Wahrscheinlichkeitsmodellen

Verteilung	**Beispiele**
Diskret Uniform	Würfeln, einstellige Zufallszahlen
Bernoulli	Spezialfall der Binomial ($n = 1$)
Binomial	Anzahl Uebertragungsfehler in einer Sequenz fester Länge, Anzahl defekte Stücke unter n Einheiten, Anzahl Bluter unter n Knaben, Anzahl Mädchen unter n Kindern, Anzahl von Objekten, die regulär verteilt sind (zeitlich oder räumlich)
Geometrisch	Anzahl gekaufte Lose bis zu einem Gewinn
Hypergeometrisch	„Ziehen ohne Zurücklegen“, Capture-Recapture
Negative Binomial	Anzahl von Objekten, die geklumpt verteilt sind (zeitlich oder räumlich), Anzahl Corn-Flakes-Käufe
Poisson	Anzahl Todesfälle pro Jahr, Bakterien in 10ml Lösung, Fahrzeuge vor einer Ampel, Telephonanrufe pro 5 Min.
Stetig Uniform	„Zufallszahl zwischen 0 und 1“
Exponential	Warte- oder Überlebenszeiten zwischen Ereignissen, die poissonverteilt sind („ohne Gedächtnis“), Aufnahmezeiten von neurologischen Rezeptoren
Gamma	Warte- oder Ueberlebenszeiten „mit Klumpung“, Zeit zwischen zwei Erdbeben
Normal	Messfehler, Längenmessungen, Mittelwerte (ZGS)

3. Schätzungen

Statistiken werden benutzt, um entsprechende Parameter der Population zu schätzen. Eine Schätzung, die aus einer einzelnen Zahl besteht, heisst Punktschätzung.
Wegen der Stichprobenvariabilität ergibt sich je nach Stichprobe ein anderer Wert. Es ist wichtig zu wissen, wie stark diese Schwankungen sein können, um die Genauigkeit der Schätzung beurteilen zu können.

3.1. Grenzwertsätze

Gegeben sei eine Population mit Mittelwert μ und Standardabweichung σ. Der Mittelwert einer Zufallsstichprobe $X_1, \ldots, X_n$ vom Umfang n aus dieser Population kann als Zufallsvariable $\bar{X}$ betrachtet werden. Für die Wahrscheinlichkeitsverteilung von $\bar{X}$ gilt dann:

$$\mu_{\bar{x}} = \mu \quad \text{und } \sigma_{\bar{x}} = \sigma/\sqrt{n} \tag{3.1}$$

Im Mittel liegt die Schätzung richtig, sie ist erwartungstreu, und die Streuung nimmt mit zunehmendem n ab.

Mit Hilfe des Gesetzes der grossen Zahl und dem Zentralen Grenzwertsatz kann diese Aussage noch präzisiert werden.

Satz 1 (Gesetz der grossen Zahl) *Seien $X_1, X_2, \ldots, X_n$ unabhängige Zufallsvariablen mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$. Sei $\bar{X}_n = 1/n \sum_{i=1}^{n} X_i$. Dann gilt für ein beliebiges $\epsilon > 0$:*

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0 \qquad \textit{für } n \to \infty.$$

Für genügend grosses n liegt also $\bar{X}$ nahe bei μ.

Satz 2 (Zentraler Grenzwertsatz) *Seien $X_1, X_2, \ldots, X_n$ unabhängige Zufallsvariablen mit $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ und gemeinsamer Verteilungsfunktion F. Dann nähert sich die Verteilung des Mittelwerts $\bar{X}_n$ der X_i für wachsendes n der Normalverteilung an:*

$$\lim_{n\to\infty} P(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x) = \Phi(x) \qquad -\infty < x < \infty.$$

$\bar{X}$ ist genähert normalverteilt für grosses n.

3.2. Vertrauensintervalle

Vertrauensintervall für μ

Wegen des Zentralen Grenzwertsatzes gilt mindestens genähert :

$$P(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95,$$

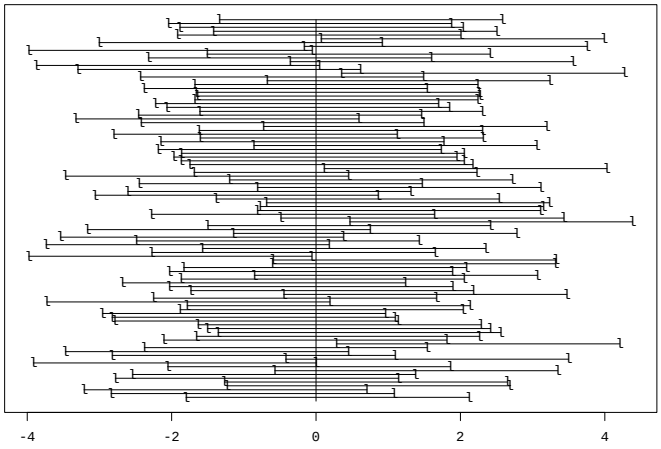


Abbildung 3.1.: Simulation von 100 Vertrauensintervallen um $\mu = 0$

d. h.

$$P(\mu - 1.96\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma/\sqrt{n}) = 0.95.$$

Nun ist

$$\mu - 1.96\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma/\sqrt{n}$$

aequivalent zu

$$\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}.$$

Wenn σ bekannt ist, kann man also eine Stichprobe ziehen, $\bar{x}$ berechnen, zusammen mit σ in der obigen Ungleichung einsetzen und dann erhält man ein 95%-*Vertrauensintervall* oder *Konfidenzintervall* (confidence interval) für den unbekannten Parameter μ:

$$\bar{x} \pm 1.96\sigma/\sqrt{n} \tag{3.2}$$

Man kann **nicht** sagen, dass μ mit einer Wahrscheinlichkeit von 95% im berechneten Intervall liegt. Der Parameter μ ist eine Konstante und nicht zufällig. Aus einer Stichprobe berechnet man den Mittelwert $\bar{x}$ und daraus die Intervallgrenzen. Auch die Intervallgrenzen sind also fest und nicht zufällig. Entweder liegt μ im Intervall oder nicht.
Zufällig ist der Wert, den die Zufallsvariable $\bar{X}$ annimmt. Jede Stichprobe liefert ein etwas anderes $\bar{x}$ und damit ein etwas anderes Intervall $\bar{x} \pm 1.96\sigma/\sqrt{n}$. 95% der so erzeugten Intervalle enthalten tatsächlich μ, die restlichen 5% nicht.

Vertrauensintervall für Proportion p

Sei p der Anteil der Stimmbevölkerung, der für Kandidat A ist und $\hat{p}$ = der entsprechende Anteil in einer „repräsentativen“ Stichprobe vom Umfang n, also eine Schätzung für p.
Gegeben seien Zufallsvariablen $X_1, \ldots, X_n$ mit

$$X_i = \begin{cases} 1 & \text{Person i ist für A} \\ 0 & \text{sonst} \end{cases}$$

$$\sum X_i \sim \mathcal{B}(n, p),$$
$$\hat{p} = \frac{\sum X_i}{n}$$

Wegen dem Zentralen Grenzwertsatz ist $\hat{p} \overset{as}{\approx} \mathcal{N}(p, \frac{p(1-p)}{n}), \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Deshalb gilt: $P\left(p - 1.96\sqrt{\frac{p(1-p)}{n}} < \hat{p} < p + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95$

$P\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95$

Für genügend grosses n ist $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ein 95%-Vertrauensintervall für p.

Vertrauensintervall für μ, σ unbekannt

Satz 3 *Seien $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid und $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ die Stichprobenvarianz. Dann hat*

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

eine t-Verteilung mit $n-1$ Freiheitsgraden.

Nun gilt also

$$P(-t_{0.975} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{0.975}) = 0.95$$
$$\Longleftrightarrow$$
$$P(\mu - t_{0.975}\ s/\sqrt{n} \leq \bar{X} \leq \mu + t_{0.975}\ s/\sqrt{n}) = 0.95$$

Aber:

$$\mu - t_{0.975}\ s/\sqrt{n} \leq \bar{X} \leq \mu + t_{0.975}\ s/\sqrt{n} \quad \Longleftrightarrow \quad \bar{X} - t_{0.975}\ s/\sqrt{n} \leq \mu \leq \bar{X} + t_{0.975}\ s/\sqrt{n}$$

Ein 95% Vertrauensintervall für den unbekannten Parameter μ ist also

$$\bar{x} \pm t_{0.975}\ s/\sqrt{n}$$

4. Statistische Tests

4.1. Begriffe und Vorgehensweise

Allgemein **Beispiel**

1. Problem formulieren:
   Nullhypothese H_0 festlegen
2. Alternativen bestimmen:
   Alternativhypothese H_A
3. zu beobachtende Grösse festlegen:
   Teststatistik T
4. Wahrscheinlichkeitsverteilung von T unter H_0 bestimmen
5. Menge aller "extremen" Beobachtungen definieren:
   Verwerfungsbereich K mit *Signifikanzniveau* α
6. Daten erheben, Wert von T berechnen:
   $T = t$
7. *P-Wert* berechnen
8. Entscheidung:
   t im Verwerfungbereich: verwerfe H_0
   t im Annahmebereich: behalte H_0 bei

Erläuterungen:

1. Wir nehmen an, dass kein Effekt oder Unterschied vorhanden ist und versuchen Evidenz gegen diese Annahme, die *Nullhypothese* H_0, zu finden. Die Nullhypothese ist das zu überprüfende Modell und besteht meistens aus einer Verteilungsannahme und einer Aussage über einen Parameter.

2. Weil man Evidenz **gegen** und nicht für etwas sucht, entspricht die Alternativhypothese der *Arbeitshypothese*. Die Nullhypothese möchte man möglichst widerlegen.

3. Die Teststatistik basiert meistens auf einer Schätzung des Parameters, der in Null- und Alternativhypothese auftritt.

5. Das Signifikanzniveau α (significance level) ist gleich der Wahrscheinlichkeit, ein „extremes" Resultat zu erhalten, unter der Annahme, dass H_0 stimmt. Je grösser also α gewählt wird, desto grösser ist der Verwerfungsbereich und umgekehrt. Üblich sind $\alpha = 5\%$ oder $\alpha = 1\%$. Das Komplement zum Verwerfungsbereich heisst *Annahmebereich*.

7. Der P-Wert (p-value) ist die Wahrscheinlichkeit, dass in einem neuen Versuch ein mindestens so extremes Resultat herauskommt, unter der Annahme, dass H_0 richtig

ist. Der P-Wert ist **nicht** die Wahrscheinlichkeit, dass H_0 richtig ist. Einer Hypothese kann gar keine Wahrscheinlichkeit zugeordnet werden, sie ist entweder richtig oder falsch.

8. Der Wert von T ist genau dann im Verwerfungsbereich, wenn der P-Wert$\leq \alpha$. In diesem Fall wird die Nullhypothese abgelehnt oder verworfen. H_0 wird als statistisch widerlegt betrachtet. Der Test ist (statistisch) signifikant.

   Wenn der P-Wert$> \alpha$ ist, dann liegt der Wert von T im Annnahmebereich und die Daten sprechen zu wenig gegen H_0. H_0 kann nicht verworfen werden. Der Test ist nicht signifikant.

Fehler 1. Art: H_0 wird verworfen, obschon H_0 richtig wäre. Die Wahrscheinlichkeit eines Fehlers 1. Art ist α und wird auf 5% oder 1% festgelegt.

Fehler 2. Art: H_0 wird beibehalten, obschon H_A stimmt. Die Wahrscheinlichkeit eines Fehlers 2. Art wird meist mit β bezeichnet.

Macht: $1 - \beta$ ist die Wahrscheinlichkeit, eine wahre Alternativhypothese zu erkennen und heisst die *Macht* des Tests (power).

Beispiel: Vergleich von zwei Betriebssystemen

Braucht die Installation von Linux mehr Zeit? 15 InformatikerInnen richteten je zwei Netzwerke ein (Betriebssystem, Applikationen, Peripheriegeräte).
Benötigte Zeit in Minuten:

Inf. Nr.	Linux	WinNT	Diff. d
1	154	145	9
2	164	162	2
3	198	156	42
4	168	152	16
5	180	168	12
6	172	157	15
7	142	155	-13
8	165	140	25
9	172	145	27
10	158	160	-2
11	170	165	5
12	148	174	-26
13	188	138	50
14	145	142	3
15	146	149	-3

4.2. Tests für Lageparameter

z-Test

Seien $X_1, \ldots, X_n$ unabhängig normalverteilt mit Erwartungswert μ und bekannter Varianz σ^2. Betrachte die folgenden Hypothesen:

H_0: $\mu = \mu_0$
H_A: $\mu \neq \mu_0$.

Die Teststatistik des z-Tests ist:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \quad \text{unter } H_0 \text{ standardnormalverteilt.}$$

t-Test für eine Stichprobe

Seien $X_1, \ldots, X_n$ unabhängig normalverteilt mit Erwartungswert μ und unbekannter Varianz. Betrachte die folgenden Hypothesen:

H_0: $\mu = \mu_0$
H_A: $\mu \neq \mu_0$.

Die Teststatistik des t-Tests ist:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}, \quad \text{unter } H_0 \text{ t-verteilt mit } n-1 \text{ Freiheitsgraden (degrees of freedom).}$$

Vorzeichentest (sign test)

Seien $X_1, \ldots, X_n$ iid mit Median m.

H_0: $m = \mu_0$
H_A: $m \neq \mu_0$.

Zur Berechnung der Teststatistik T wird von jedem Wert μ_0 subtrahiert. Dann ist T die Anzahl positiver (oder negativer) Beobachtungen; $T \sim \mathcal{B}(n, 0.5)$.

Wilcoxontest (Rangsummentest, signed rank test)

Seien $X_1, \ldots, X_n$ iid stetige, symmetrisch verteilte Zufallsvariablen mit Median m.

H_0: $m = \mu_0$
H_A: $m \neq \mu_0$.

Zuerst wird von allen Beobachtungen μ_0 subtrahiert. Die Werte werden dann dem Absolutbetrag nach geordnet und die zugehörigen Ränge bestimmt. Sind mehrere Werte gleich gross, werden die Ränge gemittelt.
Die Teststatistik ist die Rangsumme aller positiven (oder negativen) Werte T^+ (oder T^-).
Es gibt Tabellen mit den kritischen Werten; für $n > 30$ Normalapproximation.

4.3. Welcher Test soll benutzt werden?

Im Betriebssystembeispiel wird von den drei Tests auf Lageparameter einer signifikant und zwei nicht. Was nun? Nehmen wir das für uns günstigste Ergebnis und vergessen den Rest? Die verschiedenen Tests haben ganz unterschiedliche Voraussetzungen:

z-Test:	Daten sind normalverteilt, σ^2 ist bekannt, unabh. Beobachtungen
t-Test:	Daten sind normalverteilt, unabh. Beobachtungen
Wilcoxontest:	Daten sind stetig und symmetrisch verteilt, unabh. Beobachtungen
Vorzeichentest:	unabh. Beobachtungen

Die Null- und Alternativhypothese sind entsprechend:
H_0: die obigen Voraussetzungen und $\mu = \mu_0$ oder $m = \mu_0$
H_A: die obigen Voraussetzungen und $\mu \neq \mu_0$ oder $m \neq \mu_0$

Der z-Test ist praktisch kaum je benutzbar, weil σ fast immer unbekannt ist. Der t-Test reagiert auf Ausreisser sehr empfindlich, weil $\bar{x}$ und s empfindlich sind. Er ist hingegen robust (unempfindlich) gegen kleinere Abweichungen von der Normalverteilung, insbesondere für grösseres n. Die Robustheit bezieht sich aber nur auf das Signifikanzniveau α. Die Macht des

t-Tests kann ziemlich drastisch abnehmen. Der Wilcoxontest hat unter exakter Normalverteilung eine etwas kleinere Macht als der t-Test, bei Abweichungen von der Normalverteilung ist er aber deutlich effizienter (mächtiger) als der t-Test. Der Vorzeichentest ist eher ineffizient.

Empfehlung:
Falls die Daten stetig und symmetrisch sind den Wilcoxontest benutzen, sonst den Vorzeichentest. Falls der t-Test benutzt wird, den Normalplot anschauen.

t-Test für zwei unabhängige Stichproben

Seien $X_1, \ldots, X_n$ unabhängig normalverteilt mit Erwartungswert μ_X und Varianz σ^2 und $Y_1, \ldots, Y_m$ unabhängig normalverteilt mit Erwartungswert μ_Y und Varianz σ^2. Die Y_i seien unabhängig von den X_i.

H_0: $\mu_X = \mu_Y$
H_A: $\mu_X \neq \mu_Y$.

Die Teststatistik des t-Tests ist:

$$t = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \text{mit } S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

Unter H_0 ist t t-verteilt mit $n + m - 2$ Freiheitsgraden.

Mann-Whitney-Test (Rangsummentest, rank sum test)

Seien $X_1, \ldots, X_n$ iid stetige Zufallsvariablen mit Erwartungswert μ_X und $Y_1, \ldots, Y_m$ iid stetige Zufallsvariablen mit Erwartungswert μ_Y. Die Y_i seien unabhängig von den X_i.

H_0: $\mu_X = \mu_Y$
H_A: $\mu_X \neq \mu_Y$.

Bestimme die Rangsummen $T^{(1)}$ der X_i und $T^{(2)}$ der Y_i in der „gemeinsamen“ Stichprobe. Setze

$$U^{(1)} = T^{(1)} - \frac{n(n+1)}{2}, \quad U^{(2)} = T^{(2)} - \frac{m(m+1)}{2} \quad \text{und } U = \min(U^{(1)}, U^{(2)}).$$

Die kritischen Werte für U sind tabelliert.

4.4. Dualität zwischen Tests und Vertrauensintervallen

Bei einem Test lautet die Frage: „Welche Beobachtungen sind vereinbar mit H_0, bzw. einem Parameterwert μ_0?“ Der Annahmebereich liefert die Antwort.
Umgekehrt wird bei einem Vertrauensintervall gefragt: „Welche Parameter sind vereinbar mit den Beobachtungen?“ Alle diese Parameterwerte bilden dann das Vertrauensintervall.

Satz 4 (Dualitätssatz) *Ein Test mit Signifikanzniveau α verwirft $H_0 : \mu = \mu_0$ genau dann nicht, wenn μ_0 innerhalb des $(1-\alpha)100\%$-Vertrauensintervalls liegt.*

5. Zusammenhang zwischen zwei kategoriellen Variablen

5.1. 2×2-Kreuztabelle, Kontingenztafel

Beispiel: Desinfektion
Joseph Lister, brit. Arzt im 19. Jh., experimentierte mit Karbolsäure, um die hohe Todesrate durch postoperative Infektionen nach Amputationen zu senken.
Eine *2×2-Kreuztabelle*, oder *Kontingenztafel*, zeigt die 75 Operationen:

	überlebt		
Desinfektion	ja	nein	Total
ja	34	6	40
nein	19	16	35
Total	53	22	75

	Y		
X	1	2	Total
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Verschiedene relative Häufigkeiten:

Gemeinsame Häufigkeit: der Anteil der Beobachtungen mit $X = 1$ und $Y = 1$ an der Gesamtzahl Beobachtungen.

Randhäufigkeit, marginale Häufigkeit: der Anteil der Beobachtungen mit $X = 1$ an der Gesamtzahl Beobachtungen.

Bedingte Häufigkeit: der Anteil der Beobachtungen mit $Y = 1$ unter den Beobachtungen mit $X = 1$.

Besteht ein Zusammenhang zwischen X und Y? Vergleiche die bedingten Häufigkeiten miteinander.

5.2. Chiquadrat-Test auf Unabhängigkeit

Gemeinsame Verteilung von X und Y: $p_{ij} = P(X = i, Y = j)$ mit $i = 1, 2, j = 1, 2$.
Randverteilungen: $p_{i.} = P(X = i) = \sum_j p_{ij}$ und $p_{.j} = P(Y = j) = \sum_i p_{ij}$.

	Y		
X	1	2	Total
1	p_{11}	p_{12}	$p_{1.}$
2	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	1

Wenn X und Y unabhängig sind, so gilt:
$P(X = i, Y = j) = P(X = i)P(Y = j)$, d. h. $p_{ij} = p_{i.} \cdot p_{.j}$

Null- und Alternativhypothese:

$H_0 : p_{ij} = p_{i.} \cdot p_{.j}$ kein Zusammenhang zwischen X und Y
$H_A : p_{ij} \neq p_{i.} \cdot p_{.j}$ es gibt einen Zusammenhang zwischen X und Y

Erwartete Häufigkeiten unter H_0:

	Y		
X	1	2	Total
1	$\frac{n_{1.}n_{.1}}{n_{..}}$	$\frac{n_{1.}n_{.2}}{n_{..}}$	$n_{1.}$
2	$\frac{n_{2.}n_{.1}}{n_{..}}$	$\frac{n_{2.}n_{.2}}{n_{..}}$	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Die Testgrösse basiert auf den Abweichungen zwischen beobachteten (O_k) und erwarteten Häufigkeiten (E_k) in den einzelnen Zellen k: $X^2 = \sum_{k=1}^{4} \frac{(O_k - E_k)^2}{E_k} = \sum_{ij} \frac{(n_{ij} - n_{i.}n_{.j}/n_{..})^2}{n_{i.}n_{.j}/n_{..}}$.
X^2 ist unter H_0 genähert χ^2-verteilt mit einem Freiheitsgrad. Näherung ist gut, wenn $n_{..} \geq 30$ und alle **erwarteten** Häufigkeiten ≥ 5.
Der χ^2-Test ist ein einseitiger Test! Mit $\alpha = 5\%$ wird H_0 verworfen, wenn $X^2 > 3.84$.

5.3. $r \times s$ -Kontingenztafeln

Die kategoriellen Variablen X und Y haben r und s Ausprägungen.
Nullhypothese H_0: kein Zusammenhang zwischen X und Y
Testgrösse basiert wieder auf den Abweichungen zwischen beobachteten (O_k) und erwarteten Häufigkeiten (E_k): $X^2 = \sum_{k=1}^{rs} \frac{(O_k - E_k)^2}{E_k}$.
X^2 hat unter H_0 genähert eine χ^2-Verteilung mit $(r-1)(s-1)$ Freiheitsgraden.
Näherung ist gut, falls $n_{..} \geq 30$, die meisten erwarteten Häufigkeiten ≥ 4 und höchstens 20% aller erwarteten Häufigkeiten zwischen 1 und 4 sind.

Beispiel: Umfrage zum Umweltschutz

	Beeinträchtigung				
Bildung	1	2	3	4	Total
1	212	85	38	20	355
2	434	245	85	35	799
3	169	146	74	30	419
4	79	93	56	21	249
5	45	69	48	20	182
Total	939	638	301	126	2004

H_0 : „Es gibt keinen Zusammenhang zwischen Beeinträchtigung und Bildung“.
Die unter H_0 erwarteten Häufigkeiten sind:

	Beeinträchtigung			
Bildung	1	2	3	4
1	166.3	113.0	53.3	22.3
2	374.4	254.4	120.0	50.2
3	196.3	133.4	62.9	26.3
4	116.7	79.3	37.4	15.7
5	85.3	57.9	27.3	11.4

5.4. Bemerkungen zum Chiquadrat–Test

Immer Absolutzahlen verwenden, keine Prozentzahlen oder sonst irgendwie standardisierte Zahlen nehmen.

Beobachtungen müssen unabhängig sein. Bei z. B. gepaarten Daten ist *McNemar's Test* statt dem χ^2- Test durchzuführen.

Für Kategorien mit einer Rangordnung gibt es den Chiquadrattest auf Trend.

Bei zu kleinen Anzahlen ist *Fisher's exakter Test* durchzuführen.

5.5. McNemar's Test für gepaarte Daten

	1982		
1980	RaucherIn	NichtraucherIn	Total
RaucherIn	620	97	717
NichtraucherIn	76	1317	1393
Total	696	1414	2110

H_0: „kein Zusammenhang zwischen Jahr und Rauchverhalten“ oder „Zwischen 1980 und 1982 begannen gleichviele Personen neu zu rauchen wie RaucherInnen aufhörten“.
H_A: „Es gibt eine Änderungstendenz in eine Richtung“.

Betrachte nur diejenigen Personen, die sich verändern. Es gibt $r = 97$ Raucherinnen, die aufhörten und $s = 76$ NichtraucherInnen, die mit Rauchen anfingen. Die Teststatistik ist:

$$X^2 = \frac{(|r-s|-1)^2}{r+s} \text{ und ist } \chi^2\text{-verteilt mit 1 fg}$$

Wir haben also $X^2 = 2.31 < 3.84$, H_0 kann nicht verworfen werden.

5.6. Chiquadrat–Anpassungstest

Stimmt eine beobachtete Häufigkeitsverteilung mit einer theoretischen Verteilung überein?

Die Testgrösse $X^2 = \sum_{k=1}^{r} \frac{(O_k - E_k)^2}{E_k}$ ist χ^2 -verteilt mit ν Freiheitsgraden, wobei
$\nu =$ Anz. Klassen $-1-$ Anz. geschätzter Parameter.

Beispiel: Kreuzungsversuch:

Nachkommen mit drei Phänotypen mit W'keiten $\frac{1}{4}, \frac{1}{2}$ und $\frac{1}{4}$. Sind die beobachteten Häufigkeiten unter 150 Nachkommen mit dem Erbgesetz vereinbar?

Phänotyp	beobachtet (O_k)	erwartet (E_k)
I	29	37.5
II	77	75
III	44	37.5
Total	150	150

Beispiel: Poissonverteilung für die Anzahl Bakterien

Anz. Bakterien pro Fläche	Anzahl Flächen beobachtet	erwartet
0	34	32.8
1	68	82.1
2	112	102.6
3	94	85.5
4	55	53.4
5	21	26.7
6	12	11.1
7+	4	5.7
Total	400	399.9

6. Einfache lineare Regression

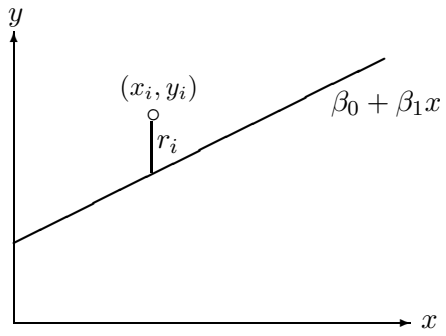
6.1. Das Modell

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

y_i ist die Zielvariable der i-ten Beobachtung.
x_i ist die erklärende Variable der i-ten Beobachtung. x_i ist eine feste, nicht zufällige Grösse.
β_0, β_1 sind unbekannte *Parameter*, die sog. Regressionskoeffizienten. Diese sollen mit Hilfe der vorhandenen Daten geschätzt werden.
ϵ_i ist der *zufällige Rest* oder *Fehler*, d. h. die zufällige Abweichung von y_i von der Geraden. Es wird vorausgesetzt, dass $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ ist und dass $Cov(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$.

6.2. Methode der Kleinsten Quadrate

Residuen: $r_i = y_i - (\beta_0 + \beta_1 x_i)$
Minimiere $Q(\beta_0, \beta_1) = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$



Normalgleichungen:

$$\begin{aligned} n\beta_0 + \beta_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i \end{aligned}$$

Least Squares-Lösung:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Regressionsgerade:
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Eigenschaften der LS-Schätzer

a) erwartungstreu $E(\hat{\beta}_0) = \beta_0$ und $E(\hat{\beta}_1) = \beta_1$
b) **BLUE** Best Linear Unbiased Estimator

$$Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}) \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Schätzung für σ^2 :

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}$$

6.3. Tests und Vertrauensintervalle

Vor: ϵ_i normalverteilt und unabhängig.
Das Modell kann nun so geschrieben werden:

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, y_i und y_j unabhängig für $i \neq j$

t-Test für $H_0 : \beta_1 = \beta$ gegen $H_A : \beta_1 \neq \beta$:

$$t^* = \frac{\hat{\beta}_1 - \beta}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta}{\sqrt{\hat{\sigma}^2 / \sum(x_i - \bar{x})^2}}$$

t^* hat eine t-Verteilung mit $n-2$ Freiheitsgraden. Die Grösse $se(\hat{\beta}_1)$ heisst *Standardfehler (standard error)* von $\hat{\beta}_1$. Verwerfe H_0, wenn $|t^*| > t_{97.5\%,n-2}$.

Ein 95%-Vertrauensintervall für β_1 ist:

$$\hat{\beta}_1 \pm t_{97.5\%,n-2} \cdot \sqrt{\hat{\sigma}^2 / \sum(x_i - \bar{x})^2}$$

Ein 95%-Vertrauensintervall für $\beta_0 + \beta_1 x_0$ ist

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{97.5\%,n-2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Ein 95%-*Prognoseintervall* für y_0 ist

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{97.5\%,n-2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

6.4. Varianzanalyse

Zerlegung der Quadratsummen:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SST &= SSR \quad + \quad SSE \\ \text{total sum of squares} &= \text{regression sum of squares} + \text{error sum of squares} \end{aligned}$$

Mean square of $\ldots = \dfrac{\text{sum of squares of } \ldots}{\text{Freiheitsgrade}}$

F-Test für $H_0 : \beta_1 = 0$ mit der Teststatistik:

$$F^* = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

F^* hat unter H_0 eine F-Verteilung mit 1 und $n-2$ Freiheitsgraden. Verwerfe H_0, wenn $F^* > F_{95\%,1,n-2}$

Anova-Tabelle:

Source of Variation	Sum of squares	Degrees of Freedom	Mean square	F^*
Regression	SSR	1	MSR	MSR/MSE
Residual	SSE	$n-2$	MSE	
Total	SST	$n-1$		

Bestimmtheitsmass R^2: Anteil an der Gesamtvariabilität, der „durch die Regression erklärt wird“:

$$R^2 = 1 - \frac{SSE}{SST}$$

Es gilt $R^2 = r^2$, wobei r die Korrelation zwischen x und y ist.

6.5. Residuenanalyse

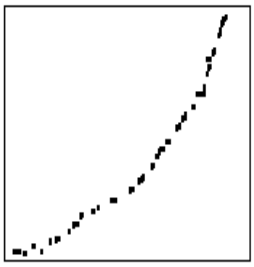
Modellannahmen:

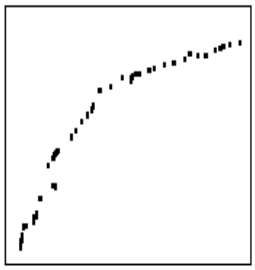
a) Zusammenhang zwischen y und x genähert linear.
b) ϵ_i haben Erwartungswert 0.
c) ϵ_i haben Varianz σ^2.
d) ϵ_i sind unkorreliert.
e) ϵ_i sind normalverteilt.

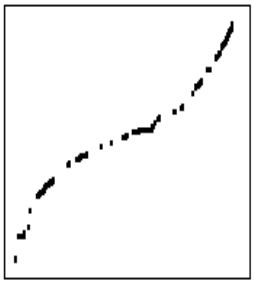
Residuenplots:

a) Normalplot
b) Plot von r_i gegen $\hat{y}_i$
c) Plot von r_i gegen x_i
d) Plot von r_i gegen i

Normalplots

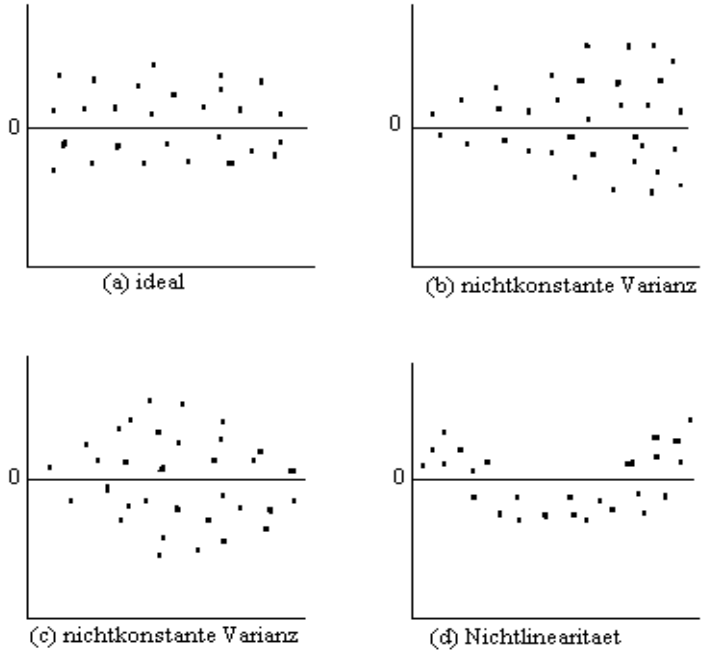

rechtsschief


linksschief


dickschwaenzig


duennschwaenzig

Plot von r_i gegen $\hat{y}_i$ oder x_i



7. Multiple lineare Regression

Beispiel: Luftverschmutzung und Mortalität
Um den Einfluss der Luftverschmutzung auf die allgemeine Mortalität zu untersuchen, wurden in einer US-Studie (finanziert von General Motors) Daten aus 60 verschiedenen Regionen zusammengetragen. Neben der altersstandardisierten Mortalität und der Belastung durch CO, NOx und SO_2 wurden verschiedene demographische und meteorologische Variablen erfasst.
Eine einfache lineare Regression von Mortalität auf SO_2 zeigt, dass mit zunehmender SO_2-Konzentration die allgemeine Sterblichkeit signifikant ansteigt. Aber auch der Zusammenhang zwischen Mortalität und allgemeinem Bildungsstand, Bevölkerungsdichte, %-Nichtweisse, Einkommen, Niederschlagsmenge usw. ist jeweils signifikant.
Statt viele einzelne einfache Regressionen zu rechnen, ist es besser, den Zusammenhang mit mehreren erklärenden Variablen gleichzeitig zu untersuchen.

7.1. Das Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \qquad i = 1, \ldots, n$$

y_i : Zielvariable
$x_{i1}, \ldots, x_{ip}$: erklärende Variablen, fest.
$\beta_0, \ldots, \beta_p$: unbekannte Parameter, Regressionskoeffizienten.
ϵ_i : zufälliger Rest oder Fehler. $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ und $Cov(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$.
Für ϵ_i normalverteilt gilt $y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \sigma^2)$ und $Cov(y_i, y_j) = 0$ für $i \neq j$.

in Matrixschreibweise:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{y}$: Zielvariablenvektor der Länge n.
$\mathbf{X}$: Designmatrix der Dimension $n \times (p+1)$.
$\boldsymbol{\beta}$: Parametervektor der Länge $p+1$.
$\boldsymbol{\epsilon}$: Fehlervektor, $E(\boldsymbol{\epsilon}) = \mathbf{0}$ und $Cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.
Für ϵ_i normalverteilt gilt $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Die Regressionsgleichungen von drei einfachen linearen Regressionen sind

$$\begin{aligned}
\hat{y} &= 886.85 + 16.73 \cdot \log SO_2 \\
\hat{y} &= 887.06 + 4.49 \cdot \%\text{-Nichtweisse} \\
\hat{y} &= 849.53 + 2.37 \cdot \text{Niederschlag} \\
\hat{y} &= 776.22 + 16.9 \cdot \log SO_2 + 3.66 \cdot \%\text{-Nichtweisse} + 1.73 \cdot \text{Niederschlag}
\end{aligned}$$

Interpretation der Regressionskoeffizienten?

$\hat{\beta}_j$ gibt die Veränderung in y bei einem Anstieg von x_j um eine Einheit an, vorausgesetzt alle andern Variablen bleiben konstant.

Methode der kleinsten Quadrate
Gesucht sind $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ so, dass
$Q = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}))^2$ minimal wird.

Normalgleichungen:

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - (\beta_0 + \ldots + \beta_p x_{ip})) = 0 \\ \frac{\partial Q}{\partial \beta_1} &= -2\sum_{i=1}^{n}(y_i - (\beta_0 + \ldots + \beta_p x_{ip}))x_{i1} = 0 \\ \vdots &\quad \vdots \\ \frac{\partial Q}{\partial \beta_p} &= -2\sum_{i=1}^{n}(y_i - (\beta_0 + \ldots + \beta_p x_{ip}))x_{ip} = 0 \end{aligned}$$

$$\mathbf{X}^t(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

oder

$$\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{y}$$

Least-Squares-Schätzungen:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ und $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$
mit Normalverteilung: $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1})$

Schätzung für σ^2 :

$$SSE = \sum r_i^2 = (\mathbf{y} - \hat{\mathbf{y}})^t(\mathbf{y} - \hat{\mathbf{y}}) \qquad \hat{\sigma}^2 = \frac{SSE}{n - p - 1} = MSE$$

Geschätzte Werte und Residuen:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{H}\mathbf{y} \quad \mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} \qquad \mathbf{H} \text{ heisst } \mathbf{Hat\text{-}Matrix}$$

7.2. Tests und Vertrauensintervalle

ANOVA-Tabelle:

Source	Sum of squares	df	Mean square	F^*
Regres	$SSR = \sum(\hat{y}_i - \bar{y})^2$	p	MSR	MSR/MSE
Resid	$SSE = \sum(y_i - \hat{y}_i)^2$	$n - 1 - p$	MSE	
Total	$SST = \sum(y_i - \bar{y})^2$	$n - 1$		

Globaler F-Test:
$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ gegen $H_A :$ mindestens ein $\beta_j \neq 0$

Testgrösse $F^* = MSR/MSE$ hat unter H_0 eine F-Verteilung mit p und $n - p - 1$ Freiheitsgraden. Verwerfe H_0, wenn $F^* > F_{95\%,p,n-p-1}$.

Multiples Bestimmtheitsmass:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad adjR^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\frac{SSE}{SST}$$

Tests von individuellen Parametern:

$H_0 : \beta_j = 0$ gegen $H_A : \beta_j \neq 0$

Teststatistik

$$t^* = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\mathbf{X}^t\mathbf{X})^{-1}_{jj}}}$$

hat unter H_0 eine t-Verteilung mit n–p–1 Freiheitsgraden. Verwerfe H_0, wenn $|t^*| > t_{97.5\%,n-p-1}$.

95%-Vertrauensintervall für β_j : $\hat{\beta}_j \pm t_{97.5\%,n-p-1} \cdot \hat{\sigma}\sqrt{(\mathbf{X}^t\mathbf{X})^{-1}_{jj}}$

95%-Vertrauensintervall für $E(y_0)$: $\hat{y_0} \pm t_{97.5\%,n-p-1} \cdot \hat{\sigma}\sqrt{\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}$ wobei $\mathbf{x}_0^t = (1 x_{01} x_{02} \cdots x_{0p})^t$

95%-Prognoseintervall für eine zukünftige Beobachtung: $\hat{y_0} \pm t_{97.5\%,n-p-1} \cdot \hat{\sigma}\sqrt{1 + \mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}$

Beispiel: Luftverschmutzungsstudie

```
Call: lm(formula = mort ~ log(so2) + nonwhite + rain, data = smsa)

Residuals:
     Min       1Q   Median       3Q      Max
-75.9671 -25.0296   0.5792  20.7507 128.3527

Coefficients:
            Estimate Std.Error tvalue Pr(>|t|)
(Intercept) 776.225  21.248  36.532  < 2e-16 ***
log(so2)     16.949   3.348   5.060 4.84e-06 ***
nonwhite      3.665   0.586   6.248 5.98e-08 ***
rain          1.732   0.456   3.796 0.000363 ***
---

Residual standard error: 38.17 on 56 df Multiple R-Squared: 0.6428,
Adjusted R-squared: 0.6237 F-statistic:  33.6 on 3 and 56 df, p-value: 1.48e-012
```

Modell mit allen sozialen und meteorolog. Variablen

jantemp	Mittlere Januar-Temperatur in Fahrenheit
julytemp	Mittlere Juli-Temperatur in Fahrenheit
relhum	Mittlere relative Luftfeuchtigkeit um 13 Uhr
rain	Mittlere jährl. Niederschlagsmenge in inches
educ	Median der absolvierten Schuljahre aller über 25jährigen
dens	Bevölkerungsdichte pro Quadratmeile

nonwhite	Anteil der nichtweissen Bevölkerung in %
wc	Anteil "white-collar worker" in %
pop	Bevölkerung
house	Mittlere Anzahl Personen pro Haushalt
income	Median des Einkommens

```
Call: lm(formula = mort ~ educ + jantemp +
julytemp + relhum + rain + dens + nonwhite +
wc + pop + house + income + log(so2),
data = smsa, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-70.915 -20.941  -2.773  18.859 105.931

Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept) 1.16e+03  2.94e+02  3.96  0.00026 ***
educ       -1.11e+01  9.45e+00 -1.17  0.24698
jantemp    -1.67e+00  7.93e-01 -2.10  0.04079 *
julytemp   -1.17e+00  1.94e+00 -0.60  0.55021
relhum      7.02e-01  1.11e+00  0.63  0.52864
rain        1.22e+00  5.49e-01  2.23  0.03074 *
dens        5.62e-03  4.48e-03  1.25  0.21594
nonwhite    5.08e+00  1.01e+00  5.02  8.3e-06 ***
wc         -1.93e+00  1.26e+00 -1.52  0.13462
pop         2.07e-06  4.05e-06  0.51  0.61180
house      -2.22e+01  4.04e+01 -0.55  0.58607
income      2.43e-04  1.33e-03  0.18  0.85562
log(so2)    6.83e+00  5.43e+00  1.26  0.21426
---

Residual standard error: 36.2 on 46 df
Multiple R-Squared: 0.733,
Adjusted R-squared: 0.664
F-statistic: 10.5 on 12 and 46 df,
p-value: 1.42e-009
```

Multicollinearität:

Sind x_1 und x_2 korreliert, dann ändern sich die geschätzten Koeffizienten, je nachdem welche Variablen im Modell sind. Es ist möglich, dass der globale F-Test signifikant ist und alle einzelnen t-Tests sind nicht signifikant.

Partielle F-Tests:
Effekt von $p-q$ Variablen gemeinsam testen. Partitioniere Parametervektor und Designmatrix :

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \\ \beta_{q+1} \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{und} \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix},$$

Modell: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$
Teste auf $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ gegen $H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$
Testgrösse

$$F^* = \frac{(SSR_{H_1} - SSR_{H_0})/(p-q)}{SSE_{H_1}/(n-p-1)}$$

hat unter H_0 eine F-Verteilung. Verwerfe H_0, wenn $F^* > F_{95\%,p-q,n-p-1}$.

7.3. Modelldiagnostik

Residuenplot:
Normalplot der Residuen r_i
Residuen r_i gegen geschätzte y-Werte $\hat{y}_i$
Residuen r_i gegen eine erklärende Variable x_i des Modells
Residuen r_i gegen eine neue Variable x_i', die nicht im Modell ist
Residuen r_i gegen den Index i

Ausreisser und einflussreiche Beobachtungen:
Ausreisser: Beobachtung mit grossem Residuum
einflussreiche Beobachtung: Beobachtung mit grossem Einfluss auf Parameterschätzungen
Hebelpunkt: Beobachtung mit extremen x-Werten

Leverages: Diagonalelemente h_{ii} der Hat-Matrix $\mathbf{H}$. Messen, wie extrem Beobachtungen bezüglich der x-Variablen sind. Es gilt $0 \leq h_{ii} \leq 1$. Hebelpunkte: $h_{ii} > 2(p+1)/n$.
Gefährlich, wenn r_i und h_{ii} gross. Betrachte Plot r_i gegen h_{ii} .

Cook's Distanz:

$$D_i = \frac{\sum(\hat{y}_j - y_{\hat{j(i)}})^2}{(p+1)\hat{\sigma}^2} = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{r_i^{*2}}{p+1}$$

$$r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \qquad \text{heisst studentisiertes Residuum}$$

Punkte mit $D_i > 1$ sollten genauer untersucht werden.

7.4. Modellwahl

Es gibt verschiedene Strategien, das „beste“ Modell zu finden und verschiedene Kriterien, was das „beste“ Modell ist. Meistens gibt es allerdings, gleich nach welchem Kriterium, nicht ein „bestes“ Modell, sondern mehrere gleich „gute“.

Strategien

Rückwärts-Elimination

Bei der Rückwärts-Elimination (backward elimination) beginnt man mit dem vollständigen Modell, d. h. mit allen zur Verfügung stehenden, erklärenden Variablen. Man eliminiert diejenige Variable mit dem kleinsten F-Wert, sofern dieser kleiner als eine vorgegebene Schranke von z. B. $F_{OUT} = 3$ ist. Dann berechnet man eine neue Regression und eliminiert die nächst unwichtigste Variable im Modell, bis keine Variable mehr einen F-Wert unterhalb der Schranke besitzt. Diese Strategie ist nur durchführbar, wenn die Anzahl vorhandener erklärender Variablen deutlich kleiner ist als die Anzahl Beobachtungen.

Vorwärts-Selektion

Bei der Vorwärts-Selektion (forward selection) beginnt man mit dem „leeren“ Modell (keine erklärende Variablen) und nimmt schrittweise jeweils die wichtigste, zusätzliche Variable in das Modell auf, solange diese eine vorgegebene Schranke von z. B. $F_{IN} = 3$ überschreitet. Die Vorwärtsselektion braucht viel weniger Rechenaufwand als die Rückwärtsmethode.

Schrittweise Regression

Die schrittweise Regression (stepwise regression) ist eine Kombination von Vorwärts- und Rückwärtsstrategie. Man beginnt vorwärts, überprüft nach jeder Aufnahme einer neuen Variable aber die F-Werte der anderen Variablen. Es ist also möglich, dass einmal aufgenommene Variablen wieder eliminiert werden, oder dass eliminierte Variablen später wieder aufgenommen werden.

„Alle Gleichungen“

Bei diesem Verfahren wird unter allen Regressionsmodellen mit den vorhandenen erklärenden Variablen das „beste“ gesucht (all subsets). Die Zahl der Modelle wächst mit der Anzahl Variablen allerdings ziemlich schnell an. Bsp: mit 10 Variablen gibt es $2^{10} = 1024$ Modelle. Intelligente Algorithmen ersparen es sich, alle Modelle durchzurechnen.

Gütekriterien

In Frage kommen die folgenden Grössen:

1. Maximales Korrigiertes Bestimmtheitsmass :

$$adjR^2 = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right) = 1 - \frac{MSE}{MST}$$

2. Minimaler Mean Square Error:

$$MSE = \frac{SSE}{n-p-1}$$

3. Maximaler Wert der Teststatistik des globalen F-Tests:

$$F^* = \frac{MSR}{MSE}$$

4. Minimale **PRESS-Statistik** : Dieses Kriterium misst die Güte des Modells an seinem Vorhersagewert. Man rechnet das Regressionsmodell jeweils ohne die i-te Beobachtung und vergleicht dann den geschätzten Wert für die i-te Beobachtung mit dem tatsächlichen y-Wert.

   Das i-te PRESS-Residuum ist definiert als

$$r_{i,-i} = y_i - \hat{y}_{i,-i}$$

   mit dem geschätzten y-Wert für $\mathbf{x}_i^t$:

$$\hat{y}_{i,-i} = \mathbf{x}_i^t \boldsymbol{\beta}_{-i}.$$

   $\boldsymbol{\beta}_{-i}$ bezeichnet die LS-Lösung ohne die i-te Beobachtung.

   Führt man diese Berechnungen für jede Beobachtung aus (also n mal) und summiert die quadrierten PRESS-Residuen auf, so erhält man die PRESS-Statistik

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{n} r_{i,-i}^2$$

   Glücklicherweise ist es nicht nötig, alle n Regressionen durchzuführen, denn es gilt:

$$r_{i,-i} = \frac{r_i}{1 - h_{ii}}$$

   Somit erhält man

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{r_i}{1 - h_{ii}} \right)^2$$

   Die PRESS-Statistik lässt sich also aus den gewöhnlichen Residuen $r_i$und den Leverages h_{ii} (Diagonalelementen der Hat-Matrix) berechnen.

5. Mallows C_q:

   Die Statistik von Mallows ist gegeben durch

$$C_q = \frac{SSE_q}{\hat{\sigma}_f^2} - n + 2q$$

   Dabei ist SSE_q das Fehlersummenquadrat des betrachteten Modells, q ist die Anzahl der Parameter des entsprechenden Modells ($q = p+1$), und $\hat{\sigma}_f^2$ ist die geschätzte Varianz unter dem vollen Modell mit allen erklärenden Variablen. Modelle mit C_q nahe bei q sind gute Kandidaten für die Modellwahl.

Für eine feste Anzahl von Variablen im Modell führen all diese Kriterien zum gleichen Ergebnis. Wenn hingegen Modelle miteinander verglichen werden mit einer unterschiedlichen Anzahl Variablen, dann können verschiedene „beste“ Modelle herauskommen. Es gibt eben auch nicht ein „bestes“ oder gar „richtiges“ Modell und alles andere ist schlechter oder falsch. Neben der automatisierten Suche aufgrund eines objektiven Kriteriums braucht es immer auch viel Fachwissen und subjektive Entscheidungen, um zu einem oder mehreren „geeigneten“ Modellen zu gelangen.

8. Versuchsplanung

Stufen einer Studie

1. Problem formulieren, in überprüfbare Hypothesen übersetzen
2. Information sammeln: Literatur, existierende Daten
3. Versuchsplanung
4. Daten sammeln und kontrollieren
5. Statistische Analyse
6. Interpretation

Versuchsplanung

- Wahl des Studientyps: beobachtend/experimentell
- Protokoll
  - Variablen (erklärende und Zielvariable)
  - Messinstrumente
  - Zielpopulation, Stichprobe
  - genauer Ablauf, Verantwortlichkeiten
  - Vorgehen bei Protokollabweichungen
- Pilotstudie
  - Reliabilität
  - Management
- Design
  - Block Design, Faktorieller Versuchsplan, Split-plot Design
  - Studiengrösse

Studientypen

- experimentell: Versuchseinheiten, -personen werden verschiedenen Behandlungen (treatments) zugeteilt. "Behandlung" umfasst alles, was kontrolliert werden kann, die gesamten Versuchsbedingungen. Bsp: Interventionsstudie, klinische Studie.
- beobachtend: Kein Einfluss auf die Gruppenzuteilung, keine Kontrolle der erklärenden Variablen. Bsp: epidemiologische Studie.
  - Umfrage (survey)
  - Fall-Kontroll-Studie (case-controll study): Vergleich der Häufigkeiten des Auftretens eines Risikofaktors bei kranken und bei gesunden Personen.
  - Kohortenstudie (cohort study): Vergleich der Häufigkeiten des Auftretens von Krankheiten bei Personen mit/ohne Risikofaktor.

Randomisierte, kontrollierte Studien

Goldstandard für experimentelle Studien sind randomisierte, kontrollierte Studien (RCT): die Versuchseinheiten werden zufällig der Behandlungs- oder der Kontrollgruppe zugeteilt und die Studie ist, wo möglich, doppelblind, d. h. weder Versuchspersonen noch ExperimentatorInnen wissen, wer in welcher Gruppe ist.

9. Varianzanalyse

Beispiel: Postoperative Zahnschmerzen
Eine Anaesthesistin untersuchte in einer Doppel-Blind-Studie die Wirksamkeit von vier verschiedenen Behandlungen gegen Zahnschmerzen nach einem chirurgischen Eingriff. Die vier Behandlungen waren:

I: Codein und Akupunktur
II: nur Akupunktur (und Zuckerkapsel)
III: nur Codein (und „Scheinakupunktur“)
IV: Placebo (Zuckerkapsel und „Scheinakupunktur“)

40 Patienten, alles Männer im Alter zwischen 18 und 30 Jahren, wurden mittels Randomisierung den vier Gruppen zugeteilt. Für jede Person wurde ein Schmerz-Intensitäts-Index erhoben, sowohl vor der Zahnbehandlung als auch 2 Stunden danach. Zielvariable war die Differenz zwischen „vorher“ und „nachher“. Die Daten sind:

Behandlung	Einzelwerte									
I	3.1	2.7	1.6	1.3	1.0	0.8	2.5	1.7	1.1	0.9
II	0.6	1.3	0.3	1.5	0.0	0.4	2.0	1.3	0.2	0.8
III	2.1	1.6	0.8	0.5	0.3	0.0	1.0	1.4	−0.2	0.6
IV	1.8	0.0	−1.8	−1.6	1.7	−1.6	1.2	0.3	0.0	−0.7

Idee der Varianzanalyse

$$\boxed{\text{Gesamtvariabilität der Beobachtungen}} = \boxed{\text{Variations-ursache 1}} + \boxed{\text{Variations-ursache 2}} + \boxed{\text{Variations-ursache 3}} + \dots$$

Begriffe
`Faktor:` diskrete, erklärende Variable
`Levels:` Werte, die der Faktor annimmt
`Ein-Weg-Varianzanalyse:` Einfluss eines Faktors wird untersucht
`Zwei-Weg-Varianzanalyse:` Einfluss von zwei Faktoren wird untersucht
`Treatment:` Faktorkombination
`Plot, experimentelle Einheit:` kleinste Einheit, die einem Treatment zugeteilt werden kann.

9.1. Ein-Weg-Varianzanalyse

Modell:

$$y_{ij} = \mu + A_i + \epsilon_{ij}, \quad \text{wobei } \sum A_i = 0$$

y_{ij} ist die Messung für Person j mit Treatment i, $j = 1, \ldots, J; i = 1, \ldots, I$,
μ ist das Gesamtmittel,
A_i ist der Effekt des i-ten Levels von Faktor A, bzw. die Abweichung der Gruppe i vom Gesamtmittel und
ϵ_{ij} ist ein zufälliger „Fehler“ oder Rest, über den folgendes vorausgesetzt wird:

1) $E(\epsilon_{ij}) = 0$ für alle i und j
2) die ϵ_{ij} sind unabhängig und haben alle die gleiche Varianz σ^2
3) die ϵ_{ij} sind normalverteilt

Varianzanalyse-Grundgleichung

$$\underbrace{\sum_i \sum_j (y_{ij} - y_{..})^2}_{\text{Gesamtvariabilität}} = \underbrace{\sum_i \sum_j (y_{i.} - y_{..})^2}_{\text{Variabilität zwischen den Gruppen}} + \underbrace{\sum_i \sum_j (y_{ij} - yi.)^2}_{\text{Variabilität innerhalb der Gruppen}}$$

total sum of squares	=	treatment sum of squares	+	residual sum of squares
SS_{tot}	=	SS_{treat}	+	SS_{res}
N–1	=	I–1	+	N–I
df_{tot}	=	df_{treat}	+	df_{res}

Total mean square: $MS_{tot} = SS_{tot}/(N-1)$
Residual mean square: $MS_{res} = SS_{res}/I(J-1) = \hat{\sigma}^2, \quad E(MS_{res}) = \sigma^2$
Treatment mean square: $MS_{treat} = SS_{treat}/(I-1), \quad E(MS_{treat}) = \sigma^2 + \sum JA_i^2/(I-1)$

Anova-Tabelle

Source	SS	df	MS	F	P-Value
Treatment	$\sum_i \sum_j (y_{i.} - y_{..})^2$	I–1	MS_{treat}	$MS_{treat}/MSres$	$P_{H_0}(F > F^*)$
Residuals	$\sum_i \sum_j (y_{ij} - yi.)^2$	N–I	MS_{res}		
Total	$\sum_i \sum_j (y_{ij} - y_{..})^2$	N–1			

Tests und Parameterschätzungen

`F-Test`
$H_0:$ alle $A_i = 0, \quad H_A:$ mindestens ein $A_i \neq 0$
Wenn die ϵ_{ij} normalverteilt sind, dann ist $F = MS_{treat}/MS_{res}$ unter H_0 F-verteilt mit $I-1$ und $N-I$ Freiheitsgraden. Verwerfe H_0, falls $F > F_{95\%,I-1,N-I}$.

`Effekt Modell`: $y_{ij} = \mu + A_i + \epsilon_{ij}, \qquad \sum A_i = 0$
Schätzungen: $\hat{\mu} = y_{..} \quad \widehat{\mu + A_i} = y_{i.} \quad \hat{A}_i = y_{i.} - y_{..}$
Voraussage: $\hat{y}_{ij} = \hat{\mu} + \hat{A}_i = yi.$ Residuum: $\hat{\epsilon}_{ij} = y_{ij} - y_{i.}$

`Mean Modell`: $y_{ij} = \mu_i + \epsilon_{ij} \qquad \hat{\mu}_i = y_{i.}$

`Effekt Modell mit anderer Nebenbedingung`: $y_{ij} = \mu + A_i + \epsilon_{ij}$, $A_1 = 0$
$\hat{\mu} = y_{1.}, \quad \hat{A}_i = y_{i.} - y_{1.}$

`Treatmentvergleiche`
Treatmentunterschied $A_i - A_{i'}$ wird geschätzt durch $y_{i.} - y_{i'.}$, mit Standardfehler $\sqrt{\sigma^2(1/J + 1/J)} = \sqrt{2\sigma^2/J}$, geschätzt durch $\sqrt{2MS_{res}/J}$.

Multiple Vergleiche

Ein `Kontrast` C ist eine Linearkombination von Effekten: $C = \sum_{i=1}^{I} \lambda_i A_i$ mit $\sum \lambda_i = 0$ und wird durch $\hat{C} = \sum \lambda_i y_{i.}$ geschätzt.
Zwei Kontraste $C_1 = \sum \lambda_i A_i$ und $C_2 = \sum \lambda_i' A_i$ heissen `orthogonal`, wenn $\sum \lambda_i \lambda_i' = 0$. Die entsprechenden Schätzungen sind dann unkorreliert. Es gibt I–1 orthogonale Kontraste.

n geplante, orthog. Kontraste ($n \leq I-1$)	`Bonferroni` (-Holm) Signifikanzniveau α/n
alle Paarvergleiche	`Tukey`: krit. Werte für die Verteilung von $\max \lvert y_{i.} - y_{i'.} \rvert$
komplexe nichtorthogonale oder komplexe ungeplante Vergleiche	`Scheffé`: krit. Wert $\sqrt{(I-1)F_{I-1,N-I,95\%}}$

9.2. Vollständiges Blockdesign

Jedes Treatment kommt in jedem Block gleich oft vor.
Modell: $y_{ij} = \mu + A_i + b_j + \epsilon_{ij}$, b_j: Effekt des Blocks j.

`Fixed-Effects Model / Modell I`	$\sum A_i = 0,\ \sum b_j = 0,\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$
`Mixed Model / Modell III`	$\sum A_i = 0,\ b_j \sim \mathcal{N}(0, \sigma_b^2),\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$, alle b_j und ϵ_{ij} unabhängig
`Random-Effects Model / Modell II`	alle Faktoren haben zufällige Effekte

Anova-Tabelle:

Source	SS	df	MS	F
Blocks	...	$J-1$	...	
Treatments	...	$I-1$	...	...
Residual	...	$(I-1)(J-1)$	...	
Total	...	$N-1$		

9.3. Multi–Faktor–Experimente

2–Weg–Varianzanalyse

Modell:

$$y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K.$$

mit den Nebenbedingungen: $\sum A_i = 0,\ \sum B_j = 0,\ \sum_i (AB)_{ij} = 0,\ \sum_j (AB)_{ij} = 0.$

y_{ijk} ist die k-te Beobachtung mit Faktor A auf Level i und Faktor B auf Level j, μ ist das Gesamtmittel, A_i ist der Haupteffekt des i-ten Levels des Faktors A, B_j ist der Haupteffekt des j-ten Levels des Faktors B, $(AB)_{ij}$ ist die Interaktion des i-ten Levels von A mit dem j-ten Level von B und ϵ_{ijk} ist der zufällige Rest mit $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

Zerlegung der Gesamtvariabilität: $SS_{tot} = SS_A + SS_B + SS_{AB} + SS_{res}$

$SS_{tot} = \sum\sum\sum (y_{ijk} - y_{...})^2 \quad SS_A = \sum\sum\sum (y_{i..} - y_{...})^2 \quad SS_B = \sum\sum\sum (y_{.j.} - y_{...})^2$
$SS_{AB} = \sum\sum\sum (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 \quad SS_{res} =$ «Differenz»

Ein Faktor mit I Levels hat $I-1$, eine Interaktion zwischen zwei Faktoren mit I und J Levels hat $(I-1)(J-1)$ Freiheitsgrade.

Anova-Tabelle

Source	SS	df	MS	F	P-Wert
A		$I-1$		MS_A/MS_{res}	
B		$J-1$		MS_B/MS_{res}	
AB		$(I-1)(J-1)$		MS_{AB}/MS_{res}	
Residual		«Differenz»			
Total		$IJK-1$			

Modell mit zufälligen Effekten (Modell II)

$$y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}, \qquad i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K.$$

μ ist das Gesamtmittel,
a_i ist der zufällige Effekt von Faktor a, $a_i \sim \mathcal{N}(0, \sigma_a^2)$,
b_j ist der zufällige Effekt von Faktor b, $b_j \sim \mathcal{N}(0, \sigma_b^2)$,
ϵ_{ijk} ist der zufällige Rest, $\epsilon_i \sim \mathcal{N}(0, \sigma_e^2)$, die a_i, b_j und ϵ_{ijk} sind alle unabhängig.

Schätzungen:
$\hat{\sigma}_e^2 = MS_{res}$
$\hat{\sigma}_a^2 = (MS_a - MS_{res})/JK$
$\hat{\sigma}_b^2 = (MS_b - MS_{res})/IK$

A. Matrizen und Vektoren

A.1. Definition

Eine Matrix ist eine rechteckige Anordnung von Zahlen in Zeilen und Spalten.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}$$

Die Dimension von A ist dim A = $n \times m$ (Anzahl Zeilen $\times$ Anzahl Spalten)
a_{ij} : Element in der i-ten Zeile und j-ten Spalte von $\mathbf{A}$

Transponierte Matrix von $\mathbf{A}$:

$$\mathbf{A^t} = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1m} & a_{2m} & \dots & a_{nm} \end{pmatrix}$$

Spezialfälle:
Eine Matrix der Dimension 1×1 ist eine Zahl (Skalar). Eine quadratische Matrix hat gleichviele Spalten wie Zeilen.
Eine Matrix, die nur aus einer Spalte besteht, ist ein Vektor.
Wenn $\mathbf{A} = \mathbf{A^t}$, so heisst $\mathbf{A}$ symmetrisch.
Eine Diagonalmatrix ist symmetrisch und alle Elemente ausserhalb der Diagonalen sind 0.
Die Einheitsmatrix $\mathbf{I}$ ist diagonal mit lauter Einsen in der Diagonale.

A.2. Rechnen mit Matrizen

Addition und Subtraktion von zwei Matrizen und die Multiplikation einer Matrix mit einem Skalar geschieht elementweise.

Multiplikation zweier Matrizen:

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 1 & 4 & 6 \\ 1 & 5 & 8 \end{pmatrix}$$

$$\left(\begin{array}{|cc|} \hline 2 & 3 \\ \hline 4 & 1 \end{array} \right) \cdot \left(\begin{array}{|c|cc} \hline 1 & 4 & 6 \\ \hline 1 & 5 & 8 \end{array} \right) = \begin{pmatrix} 5 & . & . \\ . & . & . \end{pmatrix}$$

$$2 \cdot 1 + 3 \cdot 1 = 5$$

Einfaches lineares Regressionsmodell $\mathbf{y = X\beta + \epsilon}$

mit

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{und} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

A.3. Lineare Unabhängigkeit und inverse Matrizen

Eine Menge von Vektoren heisst **linear unabhängig**, wenn keine der Spalten als Linearkombination der übrigen geschrieben werden kann.

Das **Inverse $\mathbf{A^{-1}}$** einer Matrix $\mathbf{A}$ ist :

$$\mathbf{A^{-1} \cdot A = A \cdot A^{-1} = I}$$

Nicht jede Matrix ist invertierbar.
Ein Inverses existiert genau dann, wenn alle Spalten, resp. Zeilen linear unabhängig sind.

A.4. Zufallsvektoren und Kovarianzmatrizen

Ein Zufallsvektor ist ein Vektor aus Zufallsvariablen $Y_1, Y_2, \ldots, Y_n$:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Erwartungswert

$$E(\mathbf{Y}) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix}$$

Kovarianzmatrix

$$Cov(\mathbf{Y}) = \begin{pmatrix} VarY_1 & Cov(Y_1,Y_2) & Cov(Y_1,Y_3) & \cdots & Cov(Y_1,Y_n) \\ Cov(Y_1,Y_2) & VarY_2 & Cov(Y_2,Y_3) & \cdots & Cov(Y_2,Y_n) \\ Cov(Y_1,Y_3) & Cov(Y_2,Y_3) & Var(Y_3) & \cdots & Cov(Y_3,Y_n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Cov(Y_1,Y_n) & Cov(Y_2,Y_n) & Cov(Y_3,Y_n) & \cdots & VarY_n \end{pmatrix}$$

Rechenregeln:

$$\begin{aligned} E(\mathbf{a} + \mathbf{B} \cdot \mathbf{Y}) &= \mathbf{a} + \mathbf{B} \cdot E(\mathbf{Y}) \\ Cov(\mathbf{a} + \mathbf{B} \cdot \mathbf{Y}) &= \mathbf{B} \cdot Cov(\mathbf{Y}) \cdot \mathbf{B}^t \end{aligned}$$

Mehrdimensionale Verteilungen

Die Wahrscheinlichkeitsverteilung eines Zufallsvektors ist die gemeinsame Verteilung der einzelnen Variablen. Am häufigsten benutzt wird die multivariate Normalverteilung. Was

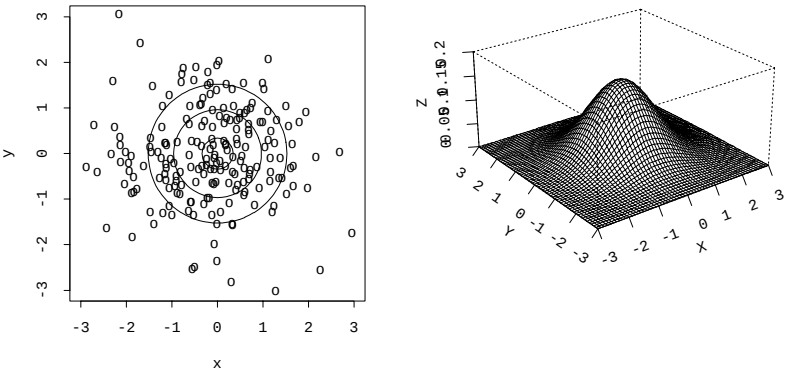


Abbildung A.1.: Bivariate Normalverteilungen mit $\rho = 0$

man sich unter einer zweidimensionalen, sogenannte bivariaten Normalverteilung vorstellen soll, zeigen die Abbildungen A.1 und A.2.
X und Y, wie auch U und V haben univariate Normalverteilungen. Daneben bestimmt die Korrelation zwischen den Variablen die genaue Form der gemeinsamen Verteilung. Je grösser die Korrelation ρ, desto enger werden die elliptischen Kontourlinien (Punkte mit gleicher Dichte).
Eine bivariate Normalverteilung wird demnach durch fünf Parameter festgelegt: $\mu_x, \sigma_x^2, \mu_y, \sigma_y^2, \rho_{xy}$.
Für eine 3-dimensionale Normalverteilung braucht es schon 9 Parameter und die Liste wird länger und länger mit wachsender Dimension. Benutzt man die Matrixnotation, so genügt die Angabe des Vektors der Erwartungswerte und der Kovarianzmatrix.
Also zum Beispiel:

$$\mathbf{Z} \sim N(\boldsymbol{\mu}, Cov(\mathbf{Z}))$$

mit

$$\mathbf{Z} = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \text{und} \quad Cov(\mathbf{Z}) = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

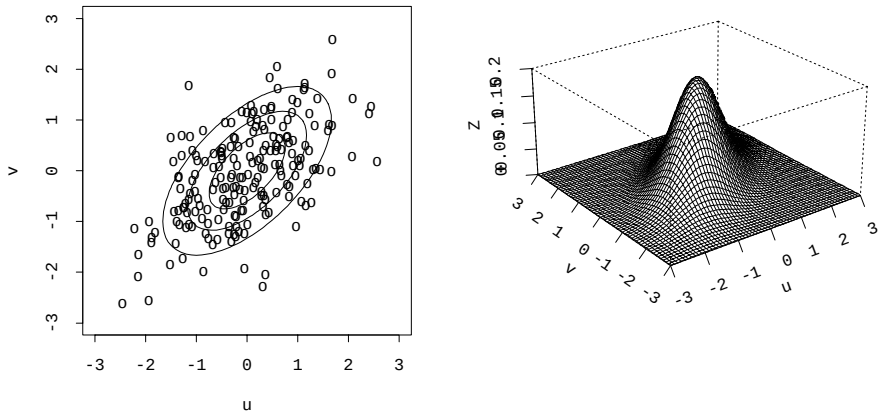


Abbildung A.2.: Bivariate Normalverteilungen mit $\rho = 0.6$