13.10.2012
Jörg Mäder

Ultrabrutalkurzskript Statistik

www.toolcase.org

Notation							
	X Datenreihe	$\bar{X}, \mu$	Mittelwert emp./theor.	H_0	Nullhypothese	P(X)	Wahrscheinlichkeit
	$\underline{X}$ Matrix	s^2, σ^2	Varianz emp./theor.	α	Signifikanzlevel	f(x)	Dichtefunktion
	n Anzahl Werte	s, σ	Standardabweichung	v	Freiheitsgrade	F(x)	kumul. Verteilungsfkt.
	x_i Einzelwert	^	geschätzter Wert	T	Testgrösse	ε	Zufallsfehler

Begriffe

a-priori & a-posteriori: basierend auf theoretischen respektive empirischen Überlegungen.

bias: systematische Abweichung/Fehler. Eine Schätzmethode mit *bias* gilt als *nicht erwartungstreu.*

diskret & stetig: Im diskreten Fall gibt es endlich, im stetigen unendlich viele Möglichkeiten.

Fehler (Residuen): Abweichung zwischen (reellen) Daten und Modell (Schätzwert, fit).

Freiheitsgrad: Beschreibt wie stark ein System (mathematisch) überdefiniert ist. Viele Daten ergeben einen hohen Freiheitsgrad, wodurch die Parameter besser bestimmt werden können.

Gesetz der grossen Zahlen: Die Abweichung zwischen empirischen Werten (Bsp Mittelwert) und theoretischen (Bsp Erwartungswert) nimmt mit steigender Anzahl Wiederholung ab.

independent identical distributed (iid): Die (meisten) Tests funktionieren nur dann korrekt, wenn die Daten die gleiche Verteilung haben und voneinander unabhängig sind.

Macht & Robustheit: Eine Methode ist mächtig, wenn sie mit wenigen Daten kleine Unterschiede signifikant feststellen kann. Eine Methode ist robust, wenn die Datenqualität (Ausreisser) wenig Einfluss hat. Meist sind robuste Methoden weniger mächtig und umgekehrt. Anhand der Datenqualität muss man beide Kriterien gegeneinander abwägen.

nominal, ordinal & skalar: Nominale Werte sind eine Art Beschriftung (Bsp: Automarke). Meist gibt es nur eine begrenzte Anzahl möglicher Werte. Nominalen Werte können als Zahlen ausgedrückt werden (Bsp: Postleitzahl), mathematische Operationen sind aber sinnlos. Besteht eine natürliche Ordnung (Bsp: Monate) spricht man von ordinal. Skalare Grössen können (prinzipiell) beliebig viele Werte annehmen, oft sind sie zähl- oder messbar (metrisch).

Nullhypothese (H_0): Vermutung, die getestet wird. H_0 kann verworfen oder beibehalten (bestätigt) werden, nicht aber bewiesen. Deshalb wird für H_0 die unerwünschte Vermutung verwendet.

Signifikanz: Ein Ergebnis in der Statistik ist nie 100%. Sobald aber die Wahrscheinlichkeit einer Falschaussage klein wird (typisch 5%), spricht man von signifikanten Ergebnissen.

Stichprobe & repräsentativ: Eine Stichprobe ist ein Ausschnitt aus der Grundgesamtheit. Bildet die Stichprobe deren Eigenschaften getreu ab, gilt sie als repräsentativ.

Unabhängigkeit & Wechselwirkung: Zwei Datenreihen (oder Ähnliches) sind unabhängig voneinander, wenn sie keinerlei gegenseitigen Einfluss aufeinander haben, weder direkt noch indirekt. Der gegenseitige Einfluss bei Abhängigkeit wird als Wechselwirkungen bezeichnet.

Weisses Rauschen: Normalverteilte Zufallsgrösse mit Mittelwert 0, frei von jeglicher Struktur. Idealfall für den Fehlerterm ε.

zentraler Grenzwertsatz: Die meisten Verteilungen nähern sich bei häufigem Wiederholen einer Normalverteilung an.

First-Aid Transformationen

Durch Transformation nähert sich die Verteilung von Daten oft der Normalverteilung an, was oft für Tests & Modelle zwingend ist.

Transformation	Anwendungen (Beispiele)	Daten
$y' = \log(y)$	Konzentrationen, Beträge, Gewichte	y > 0
$y' = \sqrt{y}$	Zähldaten, diskrete Werte	y ≥ 0
$y' = \arcsin\sqrt{y}$	Anteile, Prozentzahlen	1 ≥ y ≥ 0

Wahrscheinlichkeit

Ω	Grundgesamtheit, Ereignisraum, Menge aller möglichen Ereignisse	$P(\Omega) = 1$
{}, Ø	Leere Menge, unmögliches Ereignis	$P(Ø) = 0$
ω	Elementarereignis, nicht zusammengesetzt	$\omega \in \Omega$
A, B, ...	Mengen von Elementarereignissen	$A \subset \Omega$
A^C, $\overline{A}$, \A	Komplementär-Ereignis	$\{A \cup A^C\} = \Omega \quad \{A \cap A^C\} = Ø$
	Gegenwahrscheinlichkeit	$P(A^C) = 1 - P(A)$
$\cap$	Schnittmenge (*oder*)	$A^C \cup B^C = (B \cap A)^C$
$\cup$	Vereinigungsmenge (*und*)	$A^C \cap B^C = (B \cup A)^C$
P(X=x)	Wahrscheinlichkeit, dass ein Versuch X den Wert x ergibt.	diskret: $\mathbf{0 \le P(X=x) \le 1}$
		stetig: $\mathbf{P(X=x) = 0} \quad \mathbf{0 \le P(X \le x) \le 1}$
P(A)	Wahrscheinlichkeit, dass ein Versuch X ein Element aus Teilmenge A ergibt.	$P(A)=P(X=x, x \in A)=\sum_{x}^{A} P(X=x)$
	falls P(X=x) für alle x identisch ist, gilt:	$P(A)=\frac{n_A}{n_{\%Omega}}=\frac{\text{Anzahl Treffer}}{\text{Anzahl Möglichkeiten}}$
	falls A ein Wertebereich ($x_u \le x \le x_o$, $x \in A$):	$P(x_u \le X \le x_o) = P(X \le x_o) - P(X < x_u)$

Bei leerer Schnittmenge zweier Teilmengen gilt: $P(A \cup B) = P(A) + P(B)$

Ansonsten ($P(A \cap B) \neq \{\}$) muss um dies korrigiert werden: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Bei **Unabhängigkeit** gilt: $P(A \cap B) = P(A)\,P(B)$. Daten die diese Gleichung erfüllen sind aber nicht zwingend unabhängig, da diese Bedingung notwendig aber nicht hinreichend ist.
Beispiel: Ein Mann (M) hat eher Schuhnummer 45 als eine Frau: $P(M \cap 45) \neq P(M)\,P(45)$
Die Chance im 3. Stock zu wohnen wird hingegen nicht beeinflusst: $P(M \cap 3St) = P(M)\,P(3St)$

Bei **bedingten** Wahrscheinlichkeiten P(A|B) (sprich *P von A gegeben B*), werden nur Ereignisse der Teilmenge B beachtet. Der Satz von **Bayes** erlaubt die Umkehrung von A und B zu P(B|A).

$$P(A|B)=\frac{P(A \cap B)}{P(B)} \qquad P(B|A)=\frac{P(B)P(A|B)}{P(B)P(A|B)+P(B^c)P(A|B^c)} \qquad P(B_x|A)=\frac{P(B_x)P(A|B_x)}{\sum P(B_i)P(A|B_i)}$$

Tipp: Oft ist die Gegenwahrscheinlichkeit einfacher zu berechnen (Komplementär-Ereignis).
Beispiel: Ausfallwahrscheinlichkeit einer aus mehreren Komponenten bestehenden Maschine.
Tipp: Eine grafische Aufzeichnung als Mengen hilft beim Entschlüsseln der Verhältnisse.
Tipp: In komplizierten Fällen ist eine vollständige Auflistung oft einfacher als der formale Weg.

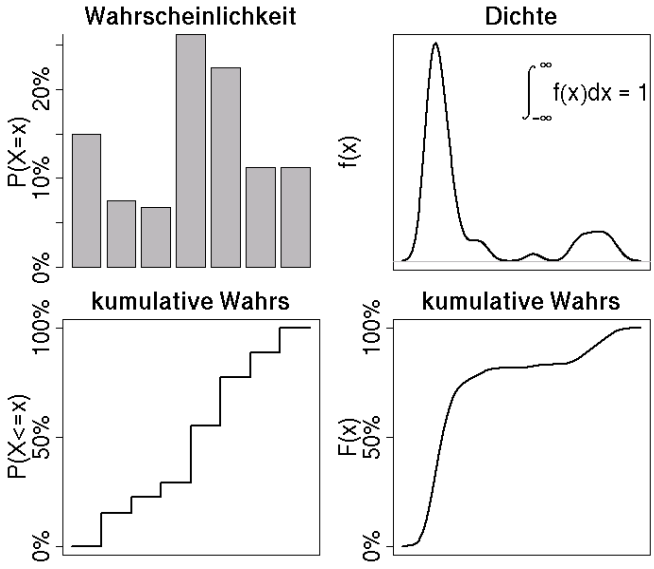
Dichtefunktion und kumulative Verteilungsfunktion

Im diskreten Fall wird P(X=x) meist als Balkendiagramm dargestellt. Im stetigen Fall verwendet man die **Dichtefunktion** f(x), da P(X=x) = 0 ist. Eine Teilfläche unter f(x), begrenzt von x_u und x_o, entspricht der Wahrscheinlichkeiten des **Wertebereiches** $P(x_u \le X \le x_o)$.

Die **kumulative Verteilungsfunktion F(x)** ist die Summe der einzelnen P(X=x) respektive das Integral von f(x). Im diskreten Fall ist F(x) treppenartig, im stetigen Fall „kontinuierlich“. Quantile lassen sich direkt an der Y-Achse von F(x) ablesen, da gilt.

$1 \ge P(X=x) \ge 0$ **$F(-\infty) = 0$**

F ist monoton steigend **$F(\infty) = 1$**



Quantile

1. Quartil = 25%Quantil — 2. Quartil = Median = 50%Quantil
3. Quartil = 75%Quantil — Quartilsdifferenz = 3. Quartil – 1. Quartil
1. Terzil = 33.3% Quantil — 2. Terzil = 66.6% Quantil

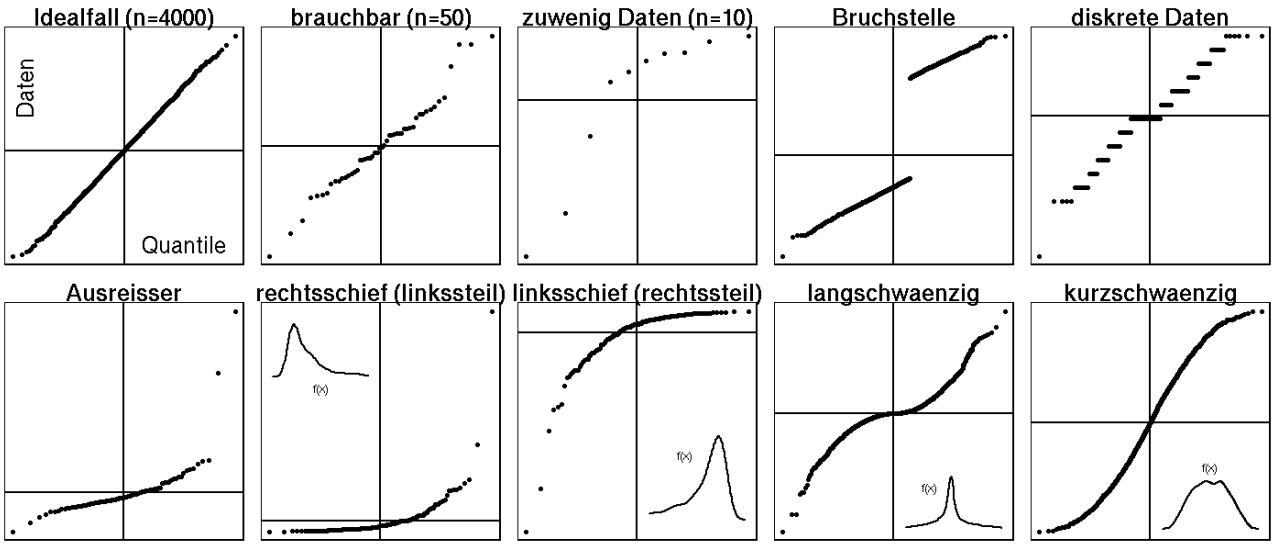
Ein Quantil (q_α) unterteilt eine (theoretische) Verteilung respektive einen (empirischen) Datensatz in zwei Teile $x_i < q_\alpha$ & $x_i > q_\alpha$. Beim Median ($q_{0.5} = q_{50\%}$) enthalten diese gleich viele Werte, das 20% Quantil ($q_{0.2}$) teilt im Verhältnis 1:4. Im sortierten Datensatz ($\dot{x}$) entspricht q_α dem Wert mit Index $\mathbf{j_\alpha = \alpha\, n + ½}$. Ist j_α nicht ganzzahlig, werden die beiden benachbarten Werte gemittelt (je nach Definition gewichtet oder nicht). Bei theoretischen Verteilungen gilt $q_\alpha = F^{-1}(\alpha)$.

Quantile sind **robust** gegenüber Ausreissern, da sie primär die Reihenfolge der Werte betrachten und nicht den Wert selbst. Im Gegenzug wird aber Information ignoriert.

Bsp: A: 1.2, 2.4, 3.6, 4.7, *6.0* → $q_{50} = 3.6$ $\bar{x} = 3.58$ B: 1.2, 2.4, 3.6, 4.7, *60* → $q_{50} = 3.6$ $\bar{x} = 13.98$

QQ-Plot

Im Quantil-Quantil-Plot kann man graphisch prüfen, ob ein Datensatz (x_i, i = 1...n) der erwarteten Verteilung entspricht. Im Idealfall entsprechen die sortierten Daten ($\dot{x}$) den theoretischen Quantilen (q_α) der Verteilung. Abweichungen ergeben im Plot ($\dot{x}_i \sim q_{\alpha i}$ $\alpha_i = (i - 0.5) / n$) Verzerrungen der Idealgeraden (y = x). Der QQ-Plot ist monoton steigend. Ausreisser befinden sich an den beiden Enden. Meist wird er für die Normalverteilung verwendet, ist aber allgemein anwendbar.

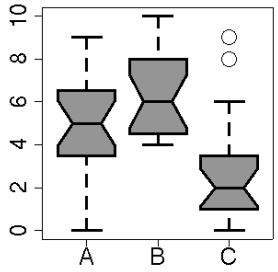


Boxplot

Boxplots erlauben einen optischen Vergleich mehrerer Datensätze. Eine typische Definition ist:

Linie innerhalb der Box: Median
Box: Quartilsdifferenz (QD)
1. bis 3. Quartil, 50% der Daten.
Kerbe: $Median \pm 1.58\, QD / \sqrt{n}$
Nicht überlappende Kerben deuten einen signifikanten Unterschied an.
Bsp.: A & B ↔ C. → Tests

Vertikale Linie: umfasst alle Punkte, die maximal 1.5 QD von der Box entfernt sind.
Punkte ober- und unterhalb: potentielle Ausreisser (C).
Schiefe Verteilungen ergeben asymmetrische Plots (B & C).



Diagnostik

Ein Testergebnis (T^+ & T^-) entspricht nicht immer der reellen Situation (R^+ & R^-). Speziell in der Medizin ist es aber wichtig, das Risiko einer Fehldiagnose zu kennen (seltene Krankheiten).

Sensitivität	$P(T^+\|R^+)$	falschpositiv	$P(T^+\|R^-)$	pos. Korrektheit	$P(R^+\|T^+)$
Spezifität	$P(T^-\|R^-)$	falschnegativ	$P(T^-\|R^+)$	neg. Korrektheit	$P(R^-\|T^-)$
diagn. Relevanz	$P(T^+\|R^+) / P(T^+\|R^-)$	Effizienz	$P(T^+\|R^+)P(R^+) + P(T^-\|R^-)P(R^-)$		

Kennzahlen

	Erwartungswert E(X)	Varianz Var(X)
	Lagemass, im Mittel zu erwartender Wert	Mass für die Streuung (Präzision)
diskret	$\mu = \sum P(X = x_i)\, x_i$	$\sigma^2 = \sum (P(X = x_i)(x_i - E(X))^2)$
stetig	$\mu = \int f(x)\, x\, dx$	$\sigma^2 = \int (f(x)(x - E(X))^2)$
empirisch	$\bar{x} = \frac{1}{n} \sum x_i$ Mittelwert als Näherung für E(X)	$\bar{x} = \frac{1}{n-1} \sum (x_i - \bar{x})^2$ Achtung! n – 1 nicht n

Standardabweichung $v = \frac{s}{\bar{x}}$

Variationskoeffizient, relativer Fehler
relative Streuung, oft in %, sinnvoll falls alle $x_i > 0$ $s = \sqrt{s^2}$ $\sigma = \sqrt{\sigma^2}$

Standardisierung, Normierung
lineare Transformation ($\rightarrow$ $\bar{x}' = 0$ $s' = 1$) $x'_i = \frac{x_i - \bar{x}}{s}$

Median ($Q_{50\%} \rightarrow$ *Quantile*) mittlerer Wert (sortiertes X)

Modus (Häufigster Wert) maximales P(X=x) resp. f(x)

Beispiel: 100 Würfe mit einem idealen 12 seitigen Würfel
(Augenzahlen: 5×1, 4×4 & 3×5) ergab 36×1, 36×4 & 28×5

$P(X)_{theoretisch}$	$P_1 = 5/12$ $P_4 = 4/12$ $P_5 = 3/12$
$P(X)_{empirisch}$	$P_1 = 0.36$ $P_4 = 0.36$ $P_5 = 0.28$
$E(X)_{theor}$	$\mu = (5 \cdot 1 + 4 \cdot 4 + 3 \cdot 5) / 12 = 3$
$E(X)_{empir.}$	$= (36 \cdot 1 + 36 \cdot 4 + 28 \cdot 5) / 100 = 3.2$
$Median_{theor}$	$P_{X \le 1} < 0.5$ & $P_{X \le 4} > 0.5 \rightarrow$ **4**
$Median_{empir.}$	$(\dot{x}_{50} + \dot{x}_{51}) / 2 =$ **4**
$Var(X)_{theor}$	$\sigma^2 = (5(1-3)^2 + 4(4-3)^2 + 3(5-3)^2) / 12 = 3$
$Var(X)_{empir.}$	$s^2 = (36(1-3.2)^2 + 36(4-3.2)^2 + 28(5-3.2)^2) / 99 \approx 2.9$
$Modus_{theor}$	P(X) maximal bei X = 1
$Modus_{empir.}$	Nicht eindeutig (1 oder 2)

Kovarianz & Korrelation

Kovarianz und Korrelation beschreiben die Ähnlichkeit zweier Reihen, indem die (skalierten) Differenzen zu den Mittelwerten multipliziert und aufsummiert werden. Beide sind nur für lineare Zusammenhänge geeignet und anfällig auf Ausreisser. Die grafische Darstellung Y versus X ist viel aussagekräftiger. **Achtung**: Eine Wert von (nahezu) 0 bedeutet nicht zwingend Unabhängigkeit!

$$Cor(X,Y) = \frac{1}{n-1} \sum \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

$$Cov(X,Y) = \frac{1}{n-1} \sum ((x_i - \bar{x})(y_i - \bar{y}))$$

$$-1 \le Cor(X,Y) \le +1$$

$$Cov(X,X) = Var(X)$$

Verlauf	Cov(X,Y)	Cor(X,Y)
synchron	stark positiv	nahe +1
entgegengesetzt	stark negativ	nahe –1
~ unabhängig	nahe 0	nahe 0

Die (Ko-)Varianzmatrix enthält von mehreren Vektoren (Ko-)Varianzen:

$$Var(\underline{X})_{ii} = Cov(X_i, Y_i)$$

Die Signifikanz einer Korrelation kann mittels t-Test berechnet werden (Freiheitsgrade: v = n – 2), das zugehörige Intervall über die Zwischengrösse $z_\pm$.

$$T = Cor(X,Y) \frac{\sqrt{n-2}}{\sqrt{1 - Cor(X,Y)^2}} \qquad z_\pm = \frac{1}{2} \ln\left(\frac{1 + Cor(X,Y)}{1 - Cor(X,Y)}\right) \pm \frac{t_{av}}{\sqrt{n-3}} \qquad Cor(X,Y)_\pm = \frac{e^{2z_\pm} - 1}{e^{2z_\pm} + 1}$$

Rechenregeln

X, Y, Z: Zufallsvariablen (Vektoren) a, b, c, d: feste Zahlen

$$E(X \pm Y) = E(X) \pm E(Y) \qquad E(a + bx) = a + b\,E(X)$$

$$Var(X \pm Y) = Var(X) + Var(Y) + 2\,Cov(X,Y) \qquad Var(a + bX) = b^2\,Var(X)$$

$$Cov(X \pm Y, Z) = Cov(X,Z) + Cov(Y,Z) \qquad Var(X) = E(X^2) - (E(X))^2$$

Gausssche Fehlerfortpflanzung

Bei der Verrechnung zweier unabhängiger Grössen (A & B) müssen auch deren Unsicherheiten (σ_A & σ_B) mit ins Ergebnis eingehen. Bei Addition & Subtraktion verwendet man die Summe der Varianzen, bei Multiplikation & Division das Quadrat der Variationskoeffizienten (relative Fehler, $v = \sigma/x$). Ansonsten muss man die Funktion an der gesuchten Stelle linearisieren.

$$\sigma_{A+B} = \sigma_{A-B} = \sqrt{\sigma_A^2 + \sigma_B^2}$$

$$\frac{\sigma_{A \cdot B}}{AB} = \frac{\sigma_{A/B}}{A/B} = \sqrt{\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2}$$

Diskrete Verteilungen

Binomial (Bernoulli) π, n

Anzahl Erfolge in n unabhängigen Versuchen (Bernoulli: n=1). Ein Experiment, das mit Wahrscheinlichkeit π gelingt, wird **n-mal** durchgeführt. Falls n gross und π klein näherungsweise Poisson verteilt ($\lambda \approx \pi n$) symmetrisch bei π = 0.5

Ω = {0, 1, 2 ... n} $P(X=x)=\binom{n}{x}\pi^x(1-\pi)^{n-x}$ $E(X)=n\pi$ $\sigma^2=n\pi(1-\pi)$

Poisson λ

Zähldaten. In einer (genähert) **unendlichen** Menge, befinden sich λ gesuchte Elemente. Für n Wiederholungen gilt $X'=\Sigma X_j$ und $\lambda'=n\lambda$. rechtsschief

Ω = {0, 1, 2 ... } $P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda}$ $E(X)=\lambda$ $\sigma^2=\lambda$

Geometrisch π

Anzahl Erfolge bis (und mit) zum ersten Misserfolg (oder umgekehrt)
Etwas mit Wahrscheinlichkeit π so oft probieren, bis es klappt (Trial and Error).

Ω = {0, 1, 2 ... } $P(X=x)=\pi^x(1-\pi)$ $E(X)=\frac{\pi}{1-\pi}$ $\sigma^2=\frac{\pi}{(1-\pi)^2}$

Stetige Verteilungen

Hinweis: σ wird auch als Parametername verwendet

Gamma-Funktion $\Gamma(x)=\int_0^\infty e^{-t}t^{x-1}\,dt \approx x!\,\text{kontinueierlich}$

Uniform α, β

konstante Dichtefunktion Rundungsfehler, selten symmetrisch **α ≤ x ≤ β**

$E(X)=\frac{\alpha+\beta}{2}$ $\sigma^2=\frac{(\beta-\alpha)^2}{12}$ $f(x)=\frac{1}{\beta-\alpha}$ $F(x)=\frac{X-\alpha}{\beta-\alpha}$

Normal μ, σ

Gaussverteilung, Glockenkurve Messungen Näherung für andere Verteilungen symmetrisch **Standard-Normalverteilung** bedeutet μ = 0 σ = 1

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

Logistisch μ, σ

langschwänziger als Normalverteilung z = (x - μ)/σ

$E(X)=\mu$ $\sigma^2=\sigma^2\frac{\pi^2}{3}$ $f(x)=\left(e^{z/2}+e^{-z/2}\right)^{-2}$ $F(x)=(1+e^{-z})^{-1}$

t-Verteilung v

Schliessende Statistik v: Freiheitsgrade (v>0) symmetrisch

$E(X)=0$ $\sigma^2=\frac{v}{v-2}$ $f(x)=\frac{\Gamma((v+1)/2)}{\Gamma(v/2)\sqrt{v\pi}}\left(1+\frac{x^2}{v}\right)^{\frac{n+1}{-2}}$ $F(x)=\int f(x)$

Lognormal μ, σ

logarithmierte Normalverteilung Grössen, Gewichte rechtsschief **x > 0**
Oft werden die Daten logarithmiert und mit Normalverteilung weitergearbeitet.

$E(X)=e^{\mu+\sigma^2/2}$ $f(x)=\frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x}\exp\left(-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right)$

$\sigma^2=e^{\sigma^2+2\mu}(e^{\sigma^2}-1)$ $F(x)=\int f(x)$

Exponential λ

verwandt mit Poisson Wartezeiten, radioaktiver Zerfall rechtsschief **x ≥ 0**

$E(X)=\frac{1}{\lambda}$ $\sigma^2=\frac{1}{\lambda^2}$ $f(x)=\lambda e^{-\lambda x}$ $F(x)=1-e^{-\lambda x}$

Chiquadrat υ

Schliessende Statistik rechtsschief **x ≥ 0**

$E(X)=\upsilon$ $\sigma^2=2\upsilon$ $f(x)=\left(2^{\upsilon/2}\Gamma(\upsilon/2)\right)^{-1}x^{\upsilon/2-1}e^{-x/2}$ $F(x)=\int f(x)$

Gamma η, σ

Beträge **x ≥ 0**

$E(X)=\eta\sigma$ $\sigma^2=\eta\sigma^2$ $f(x)=\left(\sigma\Gamma(\eta)\right)^{-1}(x/\sigma)^{\eta-1}e^{-x/\sigma}$ $F(x)=\int f(x)$

Weibull α, β

Materialermüdung **x ≥ 0**

$E(X)=\alpha^{-1/\beta}\Gamma(1/\beta+1)$ $f(x)=\alpha\beta x^{\beta-1}e^{\alpha x^\beta}$ $F(x)=1-e^{\alpha x^\beta}$

$\sigma^2=\alpha^{-2/\beta}\Gamma(2/\beta+1)-\left(\Gamma(1/\beta+1)\right)^2$

Kombinatorik

Sorten bieten einen unendlichen Vorrat an gleichartigen Elementen an.
Dies wird oft auch via ‚zurücklegen' der Elemente erreicht.

Permutation ohne Wiederholung	$n! = 1 \cdot 2 \cdot 3 \cdot ... \cdot n$	n verschiedene Elemente anordnen
Permutation mit Wiederholung	$\frac{n!}{k_1! \cdot k_2! \dots k_m!}$	n Elemente anordnen, wobei jeweils k_i identische Elemente existieren
Variation ohne Wiederholung	$\frac{n!}{(n-k)!}$	aus n verschiedenen Elementen k Elemente aussuchen und anordnen
Variation mit Wiederholung	n^k	aus n verschiedenen Sorten k Elemente aussuchen und anordnen
Kombination ohne Wiederholung	$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$	aus n verschiedenen Elementen k aussuchen (Anordnung beliebig)
Kombination mit Wiederholung	$\binom{n+k-1}{k} = \frac{(n+k-1)!}{(n-1)!\,k!}$	aus n verschiedenen Sorten k aussuchen (Anordnung beliebig)

Entscheidungsstrategie

	Frage	ja	nein
1	Sind alle Elemente unterscheidbar?	→ (4)	→ (2)
2	Ist die Anordnung der Elemente wichtig?	→ (3)	Kombination mit W
3	Müssen alle Elemente verwendet werden?	Permutation mit W	Variation mit W
4	Ist die Anordnung der Elemente wichtig?	→ (5)	Kombination ohne W
5	Müssen alle Elemente verwendet werden?	Permutation ohne W	Variation ohne W

Hauptkomponentenanalyse (EOF)

Durch lineare Transformation $\underline{Z}=B(\underline{X}-\underline{\mu})$ eines mehrdimensionalen Datensatz ($X_1 \dots X_n$) wird erreicht, dass die Komponenten von Z orthogonal zueinander stehen $Cor(Z_i,Z_j) = 0$ und mit absteigender Streuung sortiert sind. (Z_1 am meisten, Z_2 am zweit meisten ...). Ihre Interpretation ist oft heikel, da sie keine direkte Bedeutung mehr haben. Häufig werden nur die ersten Komponenten weiterverwendet (Dimensionsreduktion, Projektion). Die Transformationsmatrix B wird durch die Eigenvektoren der Kovarianzmatrix $Var(\underline{X})_{ij}$ definiert.

Simulationsstrategien

Die Genauigkeit von Resultaten kann man durch Simulation von Pseudo-Daten abschätzen (statt kompliziert berechnen). Dazu werden deren Kennzahlen und ihre Genauigkeiten (statistisch) ermittelt und als Ersatz verwendet. Eine Pseudo-Stichprobe wird aus dem ursprünglichen Datensatz erzeugt.

Bootstrap: Aus den Originaldaten wird zufällig eine gleich grosse Pseudo-Stichprobe gezogen (mit zurücklegen). Dieser Vorgang wird hundert bis tausendfach durchgeführt.

Jackknife: Die Pseudodaten entsprechen den Originaldaten minus eines Werts. Entsprechend werden n Pseudo-Stichproben erzeugt. Einfacher aber weniger allgemein anwendbar als Bootstrap.

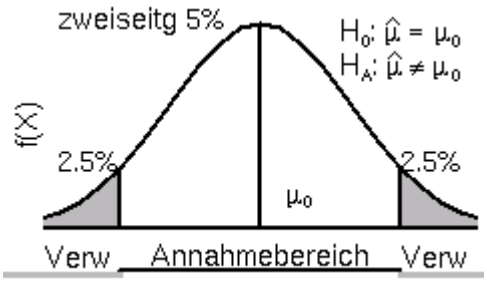
Monte-Carlo-Simulation: Ist ein zu prüfender Stichprobenraum zu gross, zieht man zufällig Stichproben und analysiert nur diese. Beim Ziehen die Wahrscheinlichkeiten berücksichtigen (falls nicht alle gleich).

Tests

Mit statistischen Tests werden Daten gegen eine Hypothese geprüft oder zwei Datensätze miteinander vergleichen. Tests können **nichts beweisen**, weshalb das Vorgehen umgekehrt wird. Als Nullhypothese (H_0) verwendet man die unerwünschte Aussage. Ergibt der Test eine kleine Wahrscheinlichkeit für H_0 (unterhalb Signifikanzlevel α), wird H_0 **verworfen**, andernfalls **beibehalten** (ist aber nicht bewiesen). Die Alternativhypothese (H_A) ist das Gegenstück zu H_0.

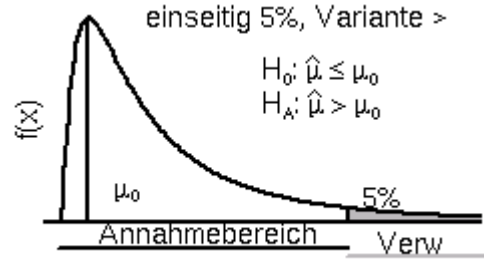
Signifikanzlevel - einseitig oder zweiseitig

Das Signifikanzlevel (α, häufig 5%) definiert das Risiko einer Falschaussage für einen Test (Fehler 1. Art). Das Level unterteilt die Verteilung in Annahme- und Verwerfungsbereich (extreme Werte). Liegt der Testwert im Verwerfungsbereich, wird H_0 verworfen.



Zweiseitig: Je die Hälfte des Verwerfungsbereich befindet sich an den beiden Rändern der Verteilung
Bsp.: Präzisionskontrollen, Parameterabschätzungen

Einseitig: Der Verwerfungsbereich beschränkt sich auf eine Seite. Bsp.: Grenzwertkontrollen.

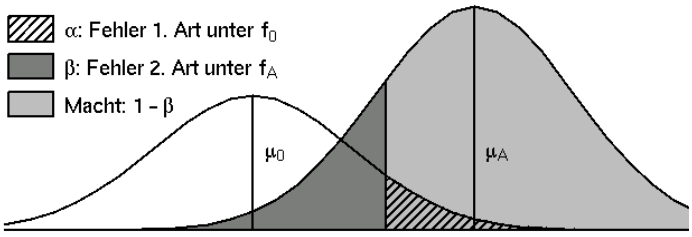


Ein zweiseitiger Test entspricht immer auch zwei einzelnen einseitigen Tests mit je $\alpha' = \alpha/2$.

Fehler und Macht

Fehler 1. Art (α, typisch 5%):
Die Nullhypothese (H_0: f_0 mit μ_0) wird verworfen, obwohl sie wahr ist.

Fehler 2. Art (β, typisch 20%):
Die Alternativhypothese (H_A: f_A mit μ_A) wird verworfen, obwohl sie wahr ist.



Macht ($1 - \beta$,): Gegenstück zu β. Wird α zu klein gewählt, schrumpft die Macht übermässig stark

p-Wert

Statt der Entscheidung *Annahmebereich ja/nein*, kann man den p-Wert der Nullhypothese berechnen. Dieser gibt an, bis auf welches Niveau (α) hinunter H_0 verworfen werden könnte. Je kleiner p, desto signifikanter das Resultat. Der p-Wert wird auch Übertretungs- oder Irrtumswahrscheinlichkeit genannt. Statistikprogramme geben meist den p-Wert direkt mit an.

Ein p-Wert von 0.05 bedeutet, dass die Testgrösse exakt dem kritischen Wert des 5% Test entspricht. Mit 0.001 wäre sie weit im Verwerfungsbereich, mit 0.34 mitten im Annahmebereich.

Approximation / Näherungen

Komplizierte (oder unbekannte) Verteilungen lassen sich durch andere approximieren, sofern Parameter und Anzahl Wiederholungen gewissen (Faust-)Regeln genügen. Meistens verwendet man die Normalverteilung (und s werden von der ursprünglichen Verteilung übernommen).

Nähert man diskrete Verteilungen durch Stetige, muss man Rundungsregeln beachten, da die Stetige beliebige Werte annehmen kann (**Stetigkeitskorrektur**). Beispiel Binomial $P(40 \leq X \leq 50)$ nähern mit Normalverteilung $P(39.5 < X < 50.5)$. Auch < und ≤ sind unterschiedlich zu behandeln!

Ein- / Zweistichprobentests (gepaart, ungepaart)

Bei einem Datensatz (Stichprobe X) wird dieser (deren Mittelwert) gegen eine vorgängig bekannte (festgelegte) Hypothese getestet (Grenzwert, kein Effekt ...). Bei zwei Datensätze (X & Y) ist die Signifikanz des Unterschiedes (der Mittelwerte) wichtig. Bei 2 **gepaarten Stichproben** ist jedem Element einer Reihe ein Element der anderen **eindeutig zugeordnet**. Die Differenzen $x_j - y_j$ werden **als eine Stichprobe aufgefasst** ($\mu_0 = 0$). Diese Differenzbildung schaltet externe Faktoren weitgehend aus und erhöht damit die Qualität der Aussage stark (Gewinn an Macht).

Beispiele gepaart	**Beispiele ungepaart**
2 Hautcremes an je einem Arm (l/r) bei allen 10 Personen	2 Blutdruckpräparate an je 10 Personen
2 Schlafmittel im Abstand von 30 Tagen an allen 10 Personen	2 Dünger auf 10 resp. 11 Weizenfeldern

Standardtests: Bei allen Tests muss die Nullhypothese (H_0 mit μ_0) im Voraus festgelegt sein! Die Daten müssen unabhängig und identisch verteilt sein (**iid**). Tests mit grosser Macht können mit weniger Daten kleinere Unterschiede signifikant erkennen. Es kann aber auch sein, dass ein Vorzeichentest Signifikanz zeigt, der z-Test aber nicht. Oft sind dann die Annahmen nicht korrekt.

Testgrösse	1 Stichprobe	2 Stichproben ungepaart	Macht	Robustheit
z-Test	$Z=\frac{\lvert\bar{x}-\mu_0\rvert}{\sqrt{\sigma^2/n}}$	$Z=\frac{\lvert\bar{x}-\bar{y}\rvert}{\sigma\sqrt{1/n_x+1/n_y}}$	**sehr gross**	sehr klein
t-Test	$T=\frac{\lvert\bar{x}-\mu_0\rvert}{\sqrt{s^2/n}}$	$Z=\frac{\lvert\bar{x}-\bar{y}\rvert}{s\sqrt{1/n_x+1/n_y}}$	gross	klein
Wilcox u-Test	R=min(R+;R-) *min: kleinere der beiden*	$T=min\left(R_x-\frac{n_x(n_x+1)}{2};R_y-\frac{n_y(n_y+1)}{2}\right)$	**gross**	**gross**
Vorzeichentest	A = Anzahl($x_i>\mu_0$)	*geht nicht*	klein	**sehr gross**

	kritische Werte	weitere Annahmen (neben *iid*)
z-Test	Tabelle Standardnormalverteilung	Daten sind **normalverteilt** σ^2 ist im **Voraus bekannt**!
t-Test	Tabelle t-Test Freiheitsgrad $v = n - 1$ resp. $v = n_1 + n_2 - 2$	Daten sind **normalverteilt** s^2 aus den Daten geschätzt
Wilcox u-Test	Tabellen Wilcox verschiedene Tabellen für 1 oder 2-Stichproben!	Verteilung der Daten symmetrisch wenige mehrfache Werte
Vorzeichentest	Tabelle Vorzeichentest oder Binomial(n, ½)	**keine**

Der **Wilcox-Test** arbeitet mit der Rangierung der Daten. Bei 2 ungepaarten Stichproben werden alle Werte zusammen rangiert (kleinster Wert 1 Rangpunkt etc), pro Stichprobe aufsummiert (R_X, R_Y) und anschliessend mit n_i korrigiert. Bei einer resp. 2 gepaarten werden die Beträge rangiert ($|x_i - \mu_0|$ resp. $|x_i - y_j|$) und nach dem Vorzeichen aufsummiert. Mehrfach vorkommende Werte (Bindungen) erhalten den mittleren Rang (Bsp: Daten 2, 3, 4, 4, 4, 4, 11 → Ränge 1, 2, 4.5, 4.5, 4.5, 4.5, 7)
Es gilt: $R_+ + R_- = n(n+1)/2$ bei 2 Stichproben: $T_X + T_Y = n_1 n_2$

Bemerkungen

- Der **Wilcox-Test** ist oft die beste Wahl (sehr robust bei minim kleinerer Macht als der t-Test).
- Der Wilcox-Test entspricht dem Kruskal-Wallis-Test mit 2 Gruppen.
- Beim **z-Test** aber auch beim **t-Test** müssen die Annahmen sehr genau eingehalten werden.
- Der **Vorzeichentest** zählt die Anzahl der x_i ober- / unterhalb von μ_0. (Werte $x = \mu_0$ entfernen)
- Bei 2 ungepaarten Stichproben mit verschiedenen Varianzen gilt: $s_{xy}^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_X+n_y-2}$
- Werden mehrere Gruppen gegeneinander getestet häufen sich zufällige Resultate. Die Korrektur des Signifikanzlevels um die Anzahl Tests (m) $\alpha' = \alpha/m$ verhindert dies (**Bonferroni**). Wird eine der m H_0's verworfen, verwendet man $\alpha' = \alpha/(m-1)$ für die übrigen (**Holm**). Dieses Vorgehen setzt Unabhängigkeit **nicht** voraus, bei hohem m wird die Macht aber klein. Es ist also nicht sinnvoll, möglichst viele Tests durchzuführen. →Varianzanalyse

Nominale Klassen und Kontingenztafeln (Chiquadrat)

Bei in **nominalen Klassen** aufgeteilten Zähldaten (Bsp: Berufe), kann mittels Chiquadrattest (χ^2) geprüft werden, ob die Anzahl Werte pro Klasse $S^{(j)}$ der erwarteten Anzahl $n\,\pi_o^{(j)}$ entsprechen.

T	Testgrösse (χ^2-verteilt, → Tabellen)	m	Anzahl Klassen
$\pi_0^{(j)}$	Wahrscheinlichkeit der Klasse j	n	Gesamtanzahl an Beobachtungen
v	Freiheitsgrade: v = m – p – 1	p	Anzahl geschätzter $\pi^{(j)}$; oft: p = 0

$$T=\sum_{j=1}^{m}\frac{\left(S^{(j)}-n\,\pi_0^{(j)}\right)^2}{n\,\pi_0^{(j)}}$$

Unwahrscheinliche Klassen sollten (zu Restklassen) zusammengefasst werden, ansonsten wird der Test unbrauchbar (Freiheitsgrade anpassen!). Faustregel: $n\,\pi_0^{(j)}\geq 4$ für 80% der Klassen, sonst $n\,\pi_0^{(j)}\geq 1$. Zähldaten direkt verwenden (keine Mittelwerte). Bei 2 Klassen ist Binomial besser.

Mit **Kontingenztafeln** kann man überprüfen, ob zwei nominalen Einteilungen (A und B mit n resp. m Klassen) sich gegenseitig beeinflussen (Abhängigkeit; Bsp: Beruf versus Automarke).

	A_1	...	A_i	...	A_{mA}	Σ
B_1	N_{11}	...	N_{i1}	...	N_{mA1}	N_{*1}
...	...	...	...	...	...	...
B_j	N_{1j}	...	N_{ij}	...	N_{nj}	N_{*j}
...	...	...	...	...	...	...
B_m	N_{1m}	...	N_{ij}	...	N_{nm}	N_{*m}
Σ	N_{1*}	...	N_{i*}	...	N_{n*}	N

N Gesamtanzahl
N_{i*} Anzahl Beobachtungen in A_i
N_{*j} Anzahl Beobachtungen in B_j
N_{ij} Anzahl Beobachtungen gleichzeitig in A_i und B_j.
Freiheitsgrade: v = (n – 1)(m – 1)

$$T=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(N_{ij}-N_{i*}N_{*j}/n\right)^2}{N_{i*}N_{*j}/n}$$

$$T=\sum\frac{(\text{beobachtet}-\text{erwartet})^2}{erwartet}$$

Mit Chiquadrat kann man auch **allgemein Verteilungen testen**, indem man Wertbereiche als Klassen definiert (hier oft: p ≥1).

Varianzanalyse

Der **Globaltest** (hat die Gruppierung einen Einfluss?) wird via **Varianzanalysetabelle** gemacht.

	Quadratsumme	Freiheitsgr.	mittleres Quadrat	Testgrösse (F-verteilt): T = MS_G / MS_E	
Gruppe	$SS_G=\sum_i n_j(\overline{Y}_{i.}-\overline{Y}_{..})^2$	DF_G = g – 1	MS_G = SS_G / DF_G	g: Anzahl Gruppen	n: Anzahl Daten
Fehler	$SS_E=\sum_{ij}(Y_{ij}-\overline{Y}_{i.})^2$	DF_E = n – g	MS_E = SS_E / DF_E	$\overline{Y}_{i.}$: Gruppenmittel	$\overline{Y}_{..}$: Gesamtmittel
Total	$SS_T=\sum_{ij}(Y_{ij}-\overline{Y})^2$	DF_T = n – 1		SS_G + SS_E = SS_T	DF_G + DF_E = DF_T

Computerprogramme berechnen direkt den p-Wert von T (Funktion *aov* in R). Die Varianzanalyse kann auch via lineare Regression durchgeführt werden (Gruppeneinteilung als Faktorvariable). Bei Mehrweg-Varianzanalysen existiert pro Einteilung je eine Zeile *Gruppe* und eine Testgrösse T_k.

Vertrauens- / Konfidenzintervall

Die Unsicherheit (Präzision) einer Zahl wird meist mit einem Intervall, welches die wahre Zahl mit einer bestimmten Wahrscheinlichkeit enthält, angegeben. Typische Intervalle der Normalverteilung sind $\mu_0 \pm \sigma$ (~66%) $\pm 2\sigma$ (~95%) und $\pm 3\sigma$ (~99.7%). Meist werden sie auch als Näherung für andere (auch unbekannte) Verteilungen genommen. Wird σ geschätzt (s), sind die Koeffizienten der t-Verteilung zu verwenden (→ Tabelle). Intervalle können auch asymmetrisch sein.

Intervalle von Mittelwerten werden durch die Wiederholung kleiner.

$$\overline{x}_{\pm}=\overline{x}\pm t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$

$t_{\alpha,n-1}$: Wert der t-Verteilung n: Anzahl Werte

Tipp für Poisson mit n Wiederholungen: Werte für nehmen und diese durch n dividieren.

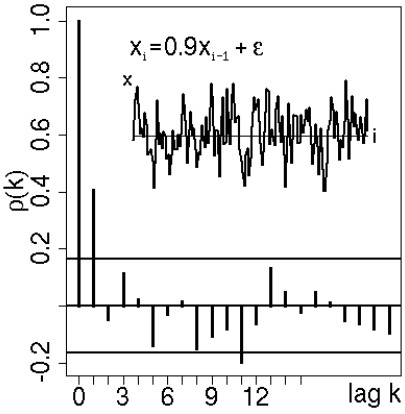
Vertrauensintervall versus Annahmebereich

Liegen die Kennzahlen eines Datensatzes im **Annahmebereich** gelten sie als verträglich im Bezug auf die vorher festgelegte Verteilung (inkl. Parameter). Die Umkehrung ist das **Vertrauens-** oder **Konfidenzintervall**. Es umfasst für eine vorher festgelegte Verteilung die Parameter, die mit dem aktuellen Datensatz verträglich sind. Meist zweiseitig verwendet, häufig beim Modellieren.

Autokorrelation

Werte einer Datenreihe beeinflussen sich oft gegenseitig. Diese sequentielle Struktur verfälscht Intervalle und Tests. Sie wird mittels Autokorrelation (AK) erfasst. Die Residuen eines Modells sollten keine AK aufweisen (Y und X_i hingegen dürfen). AK ist bei im Vergleich zum Messintervall langsamen Prozessen häufig.

Die Grafik zeigt einen simulierten AR(1)-Prozesses. Einzig ρ(1) ist signifikant (ausserhalb von $\pm\rho_{krit}$), ρ(0) ist immer 1 und ρ(11) ist zufällig signifikant. Ausserhalb sollte ρ(k) nicht benutzt werden. Um die korrekten Intervalle zu berechnen, korrigiert man die Daten (Bsp: Cochrane-Orcutt-Verfahren).



k: Distanz (lag) $x_i = f(x_{i-k,\ldots.})$ $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$ $\gamma(k) = \frac{1}{n}\sum_{s=1}^{n-k}(x_{k+s}-\bar{x})(x_s-\bar{x})$ $\gamma_{krit} = \frac{\pm 2}{\sqrt{n}}$

Kritische Werte

Die Grenzen k zwischen Annahme- und Verwerfungsbereich lassen sich aus kumulativer Verteilung F(x) und Signifikanzlevel α berechnen, respektive an der Y-Achse der Grafik ablesen.

Diskret Die kritischen Werte werden direkt aus der Definition der Verteilung berechnet: $F(k) = \alpha$
Je nach Fragestellung gehören die Grenzwerte zum Annahme oder Verwerfungsbereich.

Stetig Meist entspricht der kritische Wert nicht genau einem möglichen Wert. Entsprechend liegt je einer der benachbarten Punkte im Verwerfungs- resp. Annahmebereich.
Durch sukzessives Addieren kann man die Grenzwerte ebenfalls ermitteln.
Es gilt: $P(X \leq k) < \alpha$ und $P(X \leq k+1) > \alpha$ respektive $P(X > k) < 100\% - \alpha$
Tipp: Bei symmetrischen Verteilungen dies ausnutzen: $P(X = i) = P(X = n - i)$
Tipp: Teilweise ist der Rechenaufwand für $P(X > k) < 100\% - \alpha$ kleiner.

Aus historischen Gründen (Mangel an Computerpower) und für einen schnellen Überblick sind für viele Verteilungen die kritischen Werte in **Nomogrammen** aufgezeichnet oder **tabelliert**.

Aus den Tabellenwerten lassen sich die entsprechenden Intervalle für die Kennzahlen herleiten, indem man die Gleichungen für die Testgrösse entsprechend umformt. Beim Vorzeichentest lassen sich nur ungefähre Schranken ermitteln (Zwischen dem c und c+1 kleinsten Wert, mit c als Schranke aus der Tabelle. Die obere Grenze wird via Symmetrie hergeleitet).

Bei grossen Datensätzen kann man auch mit den Quantilen arbeiten.

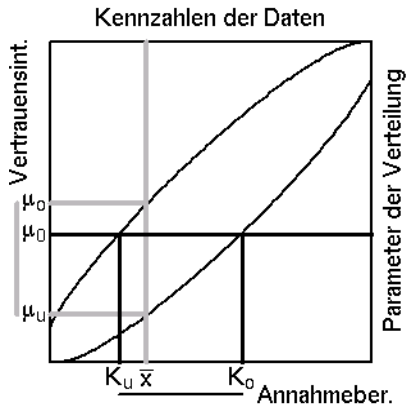
Nomogramm (Binomialverteilung)

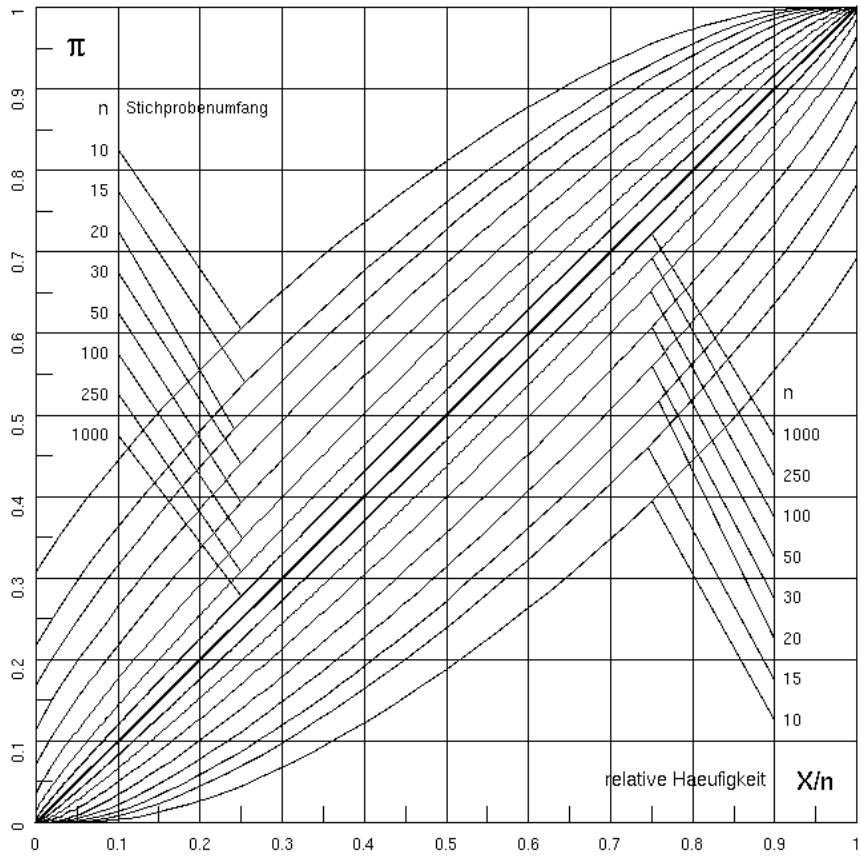
5% zweiseitig Verwerfen falls ausserhalb
X/n = relative Häufigkeit

Hinweis: Die Kurven entsprechen dem Stichprobenumfang und wurden zugunsten der Lesbarkeit interpoliert (*kontinuierlich*).

Hypothese → Annahmebereich
1. waagerechte Linie durch μ_0
2. Schnittpunkte ergeben kritische Werte K_u und K_o

Daten → Vertrauensintervall
1. senkrechte Linie durch
2. Schnittpunkte ergeben kritische Werte μ_u und μ_o





Tabellen

„x↓→“ bedeutet, dass x die Summe aus Spalten- und Zeilenbeschriftungen ist.
Ein zweiseitiger Test entspricht immer auch einem Einseitigen mit halbem Level.

Standard-Normal-Verteilung

$P(X \geq z) = P(X \leq -z) = 1 - F(z) = F(-z)$ für $z > 0$

z↓→	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	***0.0505***	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	***0.0250***	0.0244	0.0239	0.0233

Werte × 0.001

z↓→	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2.0	22.750	17.864	13.903	10.724	8.198	6.210	4.661	3.467	2.555	1.866
3.0	1.350	0.968	0.687	0.483	0.337	0.233	0.159	0.108	0.072	0.048
4.0	0.032	0.021	0.013	0.009	0.005	0.003	0.002	0.001	0.001	0.000

z-Test

5% zweiseitig: verwerfen falls $|Z| >$ **1.96** 5% einseitig: verwerfen falls $|Z| >$ **1.64**
Die Werte gehören zur Standard-Normalverteilung (siehe markierte Felder in der Tabelle).

t-Test

5% und 1%, zwei- und einseitig Verwerfen falls $|T| > q$ v = Anzahl Freiheitsgrade
bei Regressionen gilt: v = Anzahl Daten – Anzahl geschätzter Parameter

	q 1-seitig		q 2-seitig	
v	5%	1%	5%	1%
1	6.31	31.82	12.71	63.66
2	2.92	6.96	4.30	9.92
3	2.35	4.54	3.18	5.84
4	2.13	3.75	2.78	4.60
5	2.02	3.36	2.57	4.03
6	1.94	3.14	2.45	3.71
7	1.89	3.00	2.36	3.50

	q 1-seitig		q 2-seitig	
v	5%	1%	5%	1%
8	1.86	2.90	2.31	3.36
9	1.83	2.82	2.26	3.25
10	1.81	2.76	2.23	3.17
11	1.80	2.72	2.20	3.11
12	1.78	2.68	2.18	3.05
13	1.77	2.65	2.16	3.01
14	1.76	2.62	2.14	2.98

	q 1-seitig		q 2-seitig	
v	5%	1%	5%	1%
15	1.75	2.60	2.13	2.95
20	1.72	2.53	2.09	2.85
30	1.70	2.46	2.04	2.75
40	1.68	2.42	2.02	2.70
50	1.68	2.40	2.01	2.68
100	1.66	2.36	1.98	2.63
∞	1.64	2.33	1.96	2.58

Vorzeichentest

5% zweiseitig entspricht Binomial(n,½) Verwerfen falls $A \leq c$ oder $A \geq n - c$
symmetrisch Näherung (z aus z-Test):

n↓→	0	1	2	3	4	5	6	7	8	9
0	-	-	-	-	-	-	0	0	0	1
10	1	1	2	2	2	3	3	4	4	4
20	5	5	5	6	6	7	7	7	8	8
30	9	9	9	10	10	11	11	12	12	12
40	13	13	14	14	15	15	15	16	16	17

n	50	60	70	80	90	100	200	300	400	500
c	17	21	26	30	35	39	85	132	179	227

Chiquadrat

5% einseitig Verwerfen falls T > c v = Anzahl Freiheitsgrade

Näherung (z aus z-Test): $c \approx v + z\sqrt{2v}$

v↓→	0	1	2	3	4	5	6	7	8	9
0		3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92
10	18.31	19.68	21.03	22.36	23.69	25.00	26.30	27.59	28.87	30.15
20	31.41	32.67	33.93	35.18	36.42	37.66	38.89	40.12	41.34	42.56
30	43.78	44.99	46.20	47.40	48.61	49.81	41.00	52.20	53.39	54.58
40	55.76	56.95	58.13	59.31	60.49	61.6	62.83	64.01	65.18	66.34

v	50	60	70	80	90	100	200	300	500	1000
c	67.51	79.09	90.54	101.88	113.15	124.35	234.00	341.40	553.14	1074.70

Poisson

5% zweiseitig verwerfen falls T ausserhalb (λ_u, λ_o) x = Anzahl *Erfolge*.

Näherung (z aus z-Test) $\lambda \approx x + 2 \pm \sqrt{x+1}$ grobe Näherung $\lambda \approx x \pm z\sqrt{x}$

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13
λ_u	0	0.03	0.24	0.62	1.09	1.62	2.20	2.81	3.45	4.12	4.80	5.49	6.20	6.92
λ_o	3.69	5.57	7.22	8.77	10.24	11.67	13.06	14.42	15.76	17.08	18.39	19.68	20.96	22.23

x	14	15	16	17	18	19	20	25	30	35	40	45	50	60
λ_u	7.65	8.40	9.15	9.90	10.67	11.44	12.22	16.18	20.24	24.38	28.58	32.82	37.11	45.79
λ_o	23.49	24.74	25.98	27.22	28.45	29.67	30.89	36.90	42.83	48.68	54.47	60.21	65.92	77.23

Wilcox

Verwerfen falls T ≤ U

Einstichprobentest oder
2 gepaarte Stichproben

Näherung (z aus z-Test): $U = \frac{n(n+1)}{4} - z\sqrt{\frac{n(n+1)(2n+1)}{24}}$

5% einseitig

n↓→	0	1	2	3	4	5	6	7	8	9
0	-	-	-	-	-	0	2	3	5	8
10	10	13	17	21	25	30	35	41	47	53
20	60	67	75	83	91	100	110	119	130	140

5% zweiseitig

n↓→	0	1	2	3	4	5	6	7	8	9
0	-	-	-	-	-	-	0	2	3	5
10	8	10	13	17	21	25	29	34	40	46
20	53	58	65	73	81	89	98	107	113	126

2 ungepaart Stichproben
5% zweiseitig

Näherung (z aus z-Test): $U = \frac{n_1 n_2}{2} - z\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$

n_1/n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	0	0	0	0	1	1	1	1	1	2	2	2	2
3	-	-	-	-	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	-	-	-	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
5	-	-	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	-	-	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	-	-	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	-	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	-	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	-	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	-	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	-	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	-	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	-	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	69	74	78	83
15	-	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	-	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	-	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87	93	99	105
18	-	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	-	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	-	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

n_1/n_2	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
2	3	3	3	3	3	4	4	4	4	5	5	5	5	5	6	6	6	6	7	7
3	8	9	9	10	10	11	11	12	13	13	14	14	15	15	16	16	17	17	18	18
4	15	16	17	17	18	19	20	21	22	23	24	24	25	26	27	28	29	30	31	31

Multiple lineare Regression

Mittels Regression wird einer zu erklärenden Variable (Y) eine (X) oder mehrere erklärende Variablen (X_1, X_2, ... X_m) gegenübergestellt (einfache, respektive multiple lineare Regression).

Bei einfacher Regression verwendet man eine Geradengleichung mit Y-Achsenabschnitt (α), Steigung (β) und einem Fehlerterm (ε). Bei der multiplen existiert für jede X-Variable ein β_i ($\alpha = \beta_0$). In der Notation mit Matrizen enthält eine Spalte mit Einsen für den Achsenabschnitt.

$$y_i = \alpha + \sum \beta_i x_{ij} + \epsilon_j$$
$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\epsilon}$$

Die Parameter werden durch Minimierung der Summe der quadrierten Fehler (RSS = SS_E =) bestimmt (**Methode der kleinsten Quadrate**). Der Fehler r_j (**Residuum**) ist die Differenz aus Mess- und Modellwert (**fit**). Diese Methode ist nicht robust.

$$\hat{\beta} = \frac{\sum (y_i - \bar{y})(x_j - \bar{x})}{\sum (x_j - \bar{x})^2}$$
$$\hat{\beta} (X^T X)^{-1} X^T Y$$
$$\hat{y}_j = \hat{\alpha} + \sum \hat{\beta}_i x_{ij}$$
$$r_j = y_j - \hat{y}_j$$

Standardfehler $\hat{s}_{\beta_i}$ für $\hat{\beta}$ mit n: Anzahl Daten und p: Anzahl geschätzter Parameter (inkl. α)

$$\hat{s}_{\beta i} = \sqrt{\frac{1}{n-p} \frac{\sum (r_j - \bar{r})^2}{\sum (x_{ij} - \bar{x}_i)^2}}$$

t-Test (T) gegen einen vorgegebenen (theoretischen) Wert (β_i)

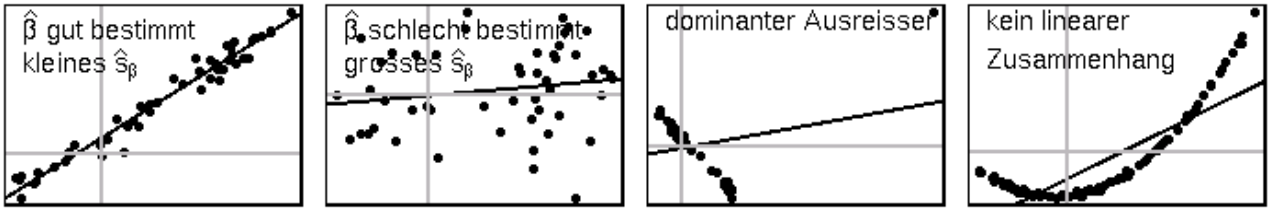
$$T_{i,v} = \left| \frac{\hat{\beta}_i - \beta_i}{\hat{s}_{\beta_i}} \right|$$

Vertrauensintervall für $\hat{\beta}$ (t aus t-Verteilung mit Signifikanzlevel α) Freiheitsgrade (Degree of freedom DF) v = n – p

$$\hat{\beta}_i \pm t_{\alpha v} \hat{s}_{\beta_i}$$

R^2 (erklärte Varianz): Anteil der Varianz der Daten (Y), der durch das Modell (R) erklärt wird. Die Residuen (E) enthalten den Rest. Nützlich beim Vergleich mehrerer Modelle (hohes R^2 besser). Jede zusätzliche Variable erhöht das R^2, weshalb man zwischen dessen Wert und der Anzahl Variablen optimieren sollte. Die Variante R^2_{adj} enthält einen Strafterm für *grosse* Modelle (kleine werden bevorzugt).

$$R^2 = \frac{SS_R}{SS_Y} = \frac{\sum (\hat{y}_j - \bar{\hat{y}})^2}{\sum (y_j - \bar{y})^2}$$
$$SS_Y = SS_R + SS_E$$
$$R^2_{adj} = 1 - \frac{n-1}{n-p}(1 - R^2)$$



normierte Koeffizienten $\hat{\beta}'$ sind Einheits-/Dimensionslos. Sie erlauben Vergleiche der Koeffizienten sowohl innerhalb als auch zwischen Modellen. Bsp: Bei einem Wert von –0.5 sinkt Y um ein halb s_Y (Standardabweichung von Y), wenn X um ein s_X steigt.

$$\hat{\beta}_i' = \hat{\beta} \frac{s_{X_i}}{s_Y}$$

Anmerkung: Wenn von der Theorie her sinnvoll, kann man ohne Achsenabschnitt modellieren.
Anmerkung: Beim Aufbau des Modells immer wieder die Qualität (mittels Grafiken) prüfen.
Anmerkung: Modelle wie $\log(y) = \alpha + \beta_i x_i^2 + \beta_2 \cos(x_2) + \beta_3 (x_1 + x_2) + \epsilon$ sind linear, da mit $y' = \log(y)$, $x'_1 = x_1^2$, $x'_2 = \cos(x_2)$, $x'_3 = x_1 + x_2$ gilt $y' = \alpha + \beta_1 x'_1 + \beta_2 x'_2 + \beta_3 x'_3 + \epsilon$

lineare Regression in R (Funktion *lm*)

Für Regression verwendet man in R den Befehl *lm*. Das Modell wird ohne Koeffizienten β_i und Fehlerterm ε notiert: *lm(y ~ u + v)* für $y = \beta_0 + \beta_1 u + \beta_2 v + \varepsilon$. Bei $\beta_0 = 0$ schreibt man *y ~ 0 + u + v*. Sind die Daten in einem *dataframe* gespeichert, kann man dessen Spaltennamen verwenden: *lm(y ~ u + v, data=messdaten)*. *coef(mod)* zeigt die Koeffizienten, *fitted(mod)* die gefitteten Werte *residuals(mod)* die Residuen des in *mod* gespeicherten Ergebnisses von *lm*. *summary(mod)* zeigt unter anderem die Freiheitsgrade, das R^2 sowie eine Tabelle, mit je einer Zeile für Achsenabschnitt α (*Intercept*) und jedes β_i der erklärenden Variablen. Spalte 3 und 4 enthalten T- und p-Wert zum Test H_0: $\beta_i = 0$.

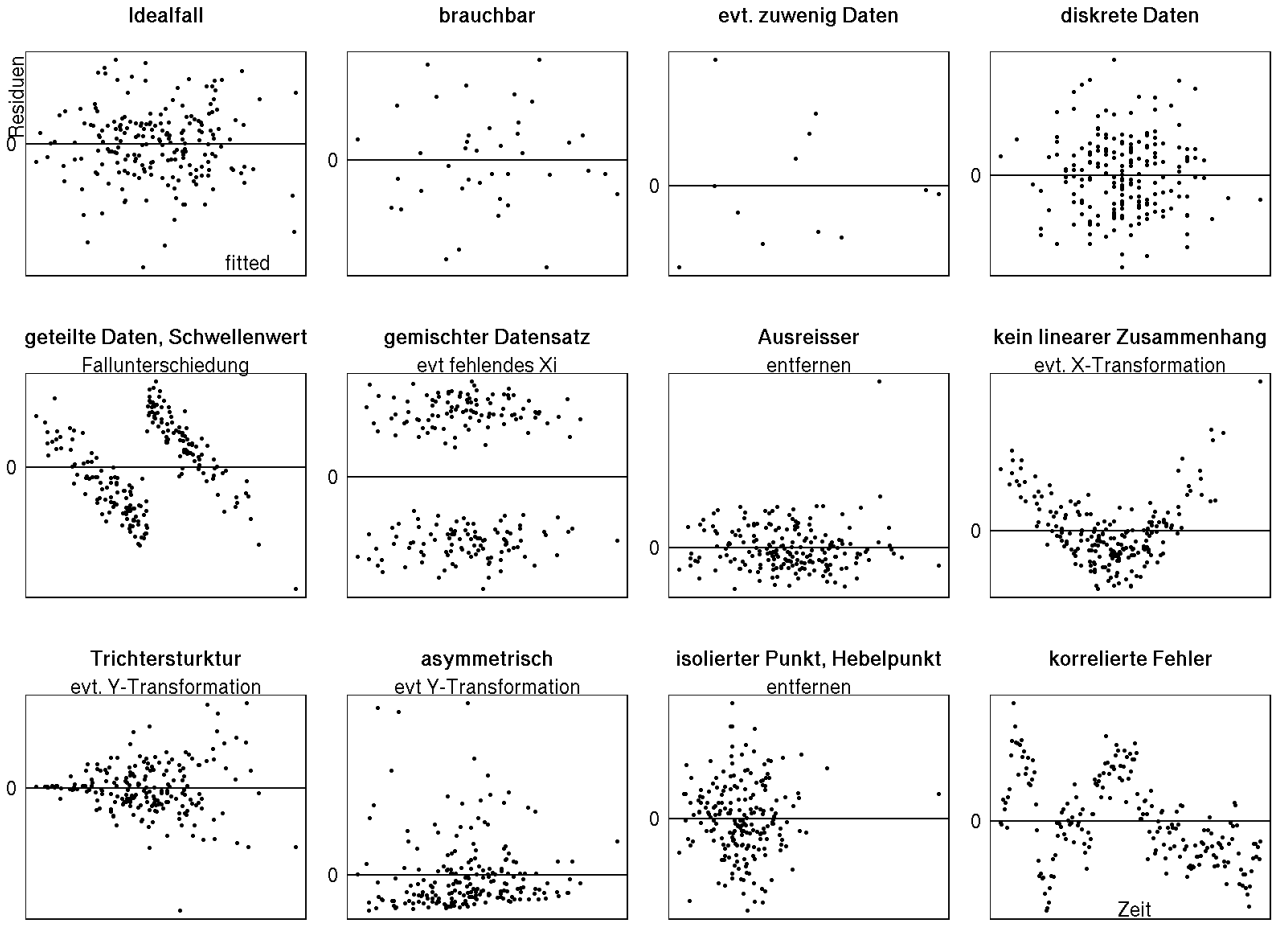
	Estimated	Std. Error	t Value	Pr(>|t|)
Intercept	$\hat{\alpha}$ \ $\hat{\beta}_0$	$\hat{s}_\alpha$	T_α	p_α
u	$\hat{\beta}_u$	$\hat{s}_{\beta_u}$	$T_{\beta u}$	$p_{\beta u}$
v	$\hat{\beta}_v$	$\hat{s}_{\beta_v}$	$T_{\beta v}$	$p_{\beta v}$

Interpretation von Regressionsmodellen

Die Interpretation einer Regression besteht vor allem aus Betrachtung von Grafiken (speziell der Residuen) und Analyse der geschätzten Parameter (inklusive **p-Wert** und **t-Test**).

Die Residuen (r_i) sollten normalverteilt und strukturlos sein (*iid*). Ersteres wird im (→) **QQ-Plot**, Letzteres meist im **Tukey-Anscombe-Plot** (r_i versus $\hat{y}_i$) geprüft. Im Idealfall entsteht eine waagerechte strukturlose Punktwolke um 0, vertikal der Normalverteilung entsprechend auslaufend. Um mögliche Strukturen (Hinweise auf Zusammenhänge) zu erkennen plotet man die Residuen versus (potentielle) erklärende Variablen.

Bei (zeitlich) geordneten Daten sind die Fehler häufig voneinander abhängig (serielle Fehler, → Autokorrelation), wodurch sich die Signifikanzen (p-Werte) verfälschen. Häufig sind „träge“ Fehler (r_i nahe bei r_{i-1}). Die Reduktion der (zeitlichen) Auflösung kann korrelierte Fehler beseitigen (Mittelwertbildung, Datenreduktion!). Schnelltest: Im ungestörten Fall wechseln die Residuen (sortiert nach der Reihenfolge) rund jedes zweite Mal das Vorzeichen.



Ein Modell sollte immer auch optimiert werden. Speziell Parameter die nicht signifikant von 0 (oder von einem theoretischen Wert) verschieden sind (p-Wert < 0.1), sollten entfernt (resp. ersetzt) werden. Folgendes einfaches **Optimierungs**verfahren bewährt sich (*backward elimination*).

1. Modell mit allen verfügbaren erklärenden Variablen aufsetzen und p-Werte berechnen.
2. Die Variable mit dem schlechtesten (höchsten) p-Wert entfernen.
3. Schritt 1 & 2 so oft wiederhohlen, bis alle Variablen signifikant sind. Verlauf von R^2 beachten.

Achtung: Immer nur eine Variable pro Durchgang entfernen. Seiteneffekte können bewirken, dass durch die Entfernung einer Variablen eine andere massiv an Signifikanz zulegt.

Anmerkung: Interpretierbarkeit und Erhebungskosten von erklärenden Variablen sind wichtige Kriterien bei der Modellauswahl/Optimierung (zum Teil stärker als deren Signifikanz).

Anmerkung: Korrelierte erklärende Variablen führen zu Unsicherheiten in den Schätzungen und sollten daher vermieden werden (→ EOF), was sich aber nicht immer erreichen lässt.

Verteilungen