i.i.d. = independ and identically distribuet

TR: $S_x = \sigma$

$odds(A^C) = \frac{1}{odds(A)}$

$logodds(A^C) = -logodds(A)$

Wahrscheinlichkeit (Wa., bsp. fairer Würfel)

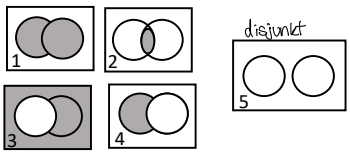
Ω	Grundraum	{1,2,3,4,5,6}
ω	Elementarereignis	{6}
A	Ereignis (A∈Ω)	{2,4,6}
ϕ	Leere Menge	{ }

Rechenoperationen

P(A)	≥0		Wa. von A
P(Ω)	=1		gesamte Wa.
P(A∪B)	P(A)+P(B)-P(A∩B)	[1]	A oder/und B
P(A∩B)	= P(A|B) P(B) (bed. Wa.)	[2]	A und B (geschnitten)
P(A^C)	1- P(A)	[3]	nicht A
P(A\B)	P(A)-P(A∩B)	[4]	A ohne B

Disjunkt [5]

P(A∩B)=0 (Schnittmenge gleich leere Menge)
P(A|B)=0
P(A∪B)=P(A)+P(B) (Wa. addieren)



Laplace

Jedes Ereignis ist gleich wahrscheinlich!

$$\mathbf{P(E)} = \frac{\mathbf{g}}{\mathbf{m}} \quad (\frac{\#\ \textbf{gesuchte Möglichkeiten}}{\#\ \textbf{alle Möglichkeiten}})$$

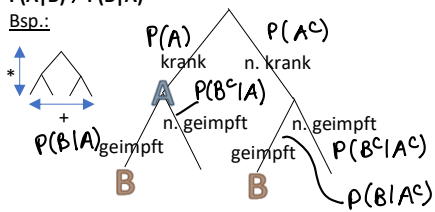
Bsp.: fairer Würfel
Wa. eine 6 zu würfeln: 1/6
Wa. eine gerade Zahl zu würfeln: 3/6

Unabhängigkeit

Die Wahrscheinlichkeiten beeinflussen sich gegenseitig nicht! (Gleichmässiger Baum)
Bsp.: Wa. zweimal hintereinander eine 6 zu würfeln (1/6 * 1/6)
P(A∩B) = P(A)*P(B)
P(A) = P(A|B)

Abhängigkeit/ bedingte Wahrscheinlichkeit

Die Wahrscheinlichkeiten beeinflussen sich gegenseitig! (Ungleichmässiger Baum)
Bsp.: Wa. krank zu werden ist abhängig von der Impfung.
P(A|B) = P(A∩B) / P(B)
Formel für Wa., dass A eintrifft, wenn B schon eingetroffen ist. (A gegeben B)
P(A|B) ≠ P(B|A)
Bsp.:



Wa. krank zu werden gegeben man ist geimpft
P(A|B) = P(A∩B) / P(B); A=krank, B=geimpft;
P(B)=P(krank)*P(geimpft)+P(n. krank)*P(geimpft)
P(A∩B)=P(krank)*P(geimpft)

Satz von Bayes

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

$P(A) = 0 \rightarrow P(A|B) = 0$

$P(A) > 0 \rightarrow P(A|B) = [0;1] \rightarrow P(A|B) \lessgtr P(A)$

Odds

Wie viel Mal wahrscheinlicher ist A als nicht A?
(Bsp.: Münzwurf 1:1 Wa. Kopf zu werfen; fairer Würfel 1:5 Wa. eine 6 zu würfeln) $P(A^C) = 1 - P(A)$

$$\mathbf{odds(A)} = \frac{\mathbf{P(A)}}{\mathbf{P(A^C)}} = \frac{\mathbf{P(A)}}{\mathbf{1 - P(A)}}$$

$$\mathbf{odds(A|B = 1)} = \frac{\mathbf{P(A|B = 1)}}{\mathbf{1 - P(A|B = 1)}}$$

Log-Odds

$$\mathbf{ln(odds(A))} = \mathbf{ln}(\frac{\mathbf{P(A)}}{\mathbf{P(A^C)}}) = \mathbf{ln}(\frac{\mathbf{P(A)}}{\mathbf{1 - P(A)}})$$

Odds-Ratio (OR)

↳ $A > A^C \Rightarrow \ln odds > 0$

Wie stark hängen zwei Merkmale zusammen?
(Für absolute Wa.s oder für bedingte Wa.s)

$$\mathbf{OR} = \frac{\mathbf{odds(A)}}{\mathbf{odds(B)}}; \mathbf{OR} = \frac{\mathbf{odds(A|B)}}{\mathbf{odds(A|B^C)}}$$

Bsp.: Wa. zu Erkranken sinkt um x wenn man geimpft ist. $\mathrm{OR} = x = \frac{\mathrm{odds(krank|nicht\ geimpft)}}{\mathrm{odds(krank|geimpft)}}$

Stichprobengröße:
$n \geq 4 \cdot \frac{\hat{\sigma}_x^2}{\delta^2}$
→ Breite 95%-VI = $2 \cdot \delta$

DISKRETE VERTEILUNG (abzählbar, mit „Punkt"-Wahrscheinlichkeiten rechnen möglich; P(X=x), PDF)
Kennzahlen einer Verteilung

X: Zufallsvariable, x: konkreter Wert von X

Erwartungswert ε(x): Beschreibt die mittlere Lage der Verteilung (Mittelwert bzw. Durchschnitt)

$$\varepsilon(x) = \sum_{x \in W} x * P(X = x); W = \text{Wertebereich von } x$$

Varianz Var(x): Beschreibt die mittlere quadratische Abweichung eines Zufallwertes von ihrem Erwartungswert.

$$\text{Var}(x) = \sum_{x \in W} (x - \varepsilon(x))^2 * P(X = x)$$

Standartabweichung σ(x): Beschreibt die Streuung der Verteilung bzw. die Breite

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

Zufallsvariablen

Rechenregeln

ε(X+Y)=ε(X)+ε(Y)	Varianz:
ε(a*X)=a*ε(X)	1. Konstanten=0
ε(a+bX)=a+bε(X)	2. Faktoren hoch 2
Var(X+Y)=Var(X)+Var(Y)	3. Minus wird Plus
Var(a*X)=a²*Var(X)	Var(2X+9-Y)= 4Var(X)+Var(Y)
Var(a+bX)= b²*Var(X)	

$\sigma(X \pm Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)}$ falls x und y unabhängig

$\sigma(X \pm Y) = \sqrt{\sigma^2(X) + \sigma^2(Y) \pm 2\text{Cov}(X,Y)}$ falls abhängig

Cov(X,Y)=ε((X-ε(X)(Y-ε(Y))

$g(a+b \cdot x) = a + b \cdot g(x)$

$\sigma(a+bx) = b \cdot \sigma(x)$

Berechnen von Vertrauensintervall für π

für Binomialverteilung durch Normalapproximation

$$I = \frac{x}{n} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) * \sqrt{\frac{x}{n}\left(1 - \frac{x}{n}\right)\frac{1}{n}}$$

$$\text{für } \alpha = 0.05: I = \frac{x}{n} \pm 1.96 * \sqrt{\frac{x}{n}\left(1 - \frac{x}{n}\right)\frac{1}{n}}$$

→ Breite VI abhängig von $\frac{1}{\sqrt{n}}$

VI für gesamte erwartete Anzahl von Erfolgen:

$$I = n\pi_0 \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) * \sqrt{n\pi_0(1 - \pi_0)}$$

$$\text{für } \alpha = 0.05: I = n\pi_0 \pm 1.96 * \sqrt{n\pi_0(1 - \pi_0)}$$

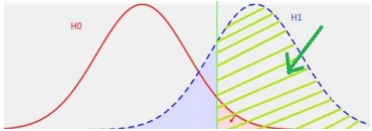
→ VI umfasst alle Werte π_0, für die ein (zweiseitiger) Binomialtest mit $H_0: \pi = \pi_0$ nicht verwirft

Macht: Wa., dass Beob. in Verwerfungsbereich fällt, obwohl H_1 wahr ist
→ Verwerfungsbereich {0; ...; 15} → Wahre Wa = P_A = Wa. H_0 zu verwerfen $P_{P_A}(x \leq 15)$

→ kann nur berechnet werden, wenn genaue Verteilung der Teststatistik unter H_A bekannt

Fehler

(hier: πA > π0; grüne Linie: Signifikanzlevel α, definiert Verwerfungsbereich K)



Rot: Fehler 1. Art ($\leq \alpha$); H_0 wird verworfen obwohl sie stimmen würde, P(X∈K)

Blau: Fehler 2. Art ($\leq \beta$); H_0 wird akzeptiert obwohl sie falsch ist, P(X∉K) → Wa., dass mit $p(H_1)$ Wert im VI liegt

Grün: Macht ($\leq 1 - \beta$); Wahrscheinlichkeit eine bestimmte Alternativhypothese zu erkennen, wenn diese stimmt, $P(X \in K)_{\pi A}$, kann man nur berechnen wenn π_A bekannt! Je grösser der Fehler 1. Art, desto grösser die Macht.

- Fehler 1. Art ↓ ⇒ Fehler 2. Art ↑ ⇒ Macht ↓
- höheres α = größerer Verwerfungsbereich

Auf 10% verwerfen ⇒ nicht unbedingt auf 1% usw.

→ größere Erfolgswa. unter H_0:
Verteilung verschiebt sich nach rechts
→ Verwerfungsbereich wird potentiell kleiner
→ Macht nimmt ab

Macht einseitig vs. zweiseitig:

$H_A: \pi < \pi_0$ ⇒ wird kleinere Erfolgswa. π_0 mit größerer Macht erkennen als zweiseitiger, für größere Erfolgswa. aber quasi keine Macht

$H_A: \pi > \pi_0$ ⇒ andersrum

⇒ da Macht (einseitig) > Macht (zweiseitig) hängt von konkreter Alternative ab

Grundfragen der Statistik

1. Plausibler Parameter π?

Punktschätzung für Erfolgswahrscheinlichkeit π

a) Momentenmethode

$$\pi = \frac{x}{n} = \frac{\varepsilon(x)}{n} = \frac{\#\text{beobachtet} \;\hat{=}\; \text{erwartet}}{\#\text{ unabhängige Versuche}}$$

b) Maximum likelyhood Methode

Gleichung aufstellen (P(A)*P(B)*...), logartithmieren falls hilfreich, ableiten und gleich null setzen, nach π auflösen; genauer, aufwendiger; nur wissen, dass es das gibt

2. Sind Werte von π mit den Daten vereinbar?

Vertrauensintervall VI

→ Die Werte von π_0 bei denen H_0 nicht verworfen wird bilden ein (1-α)-Vertrauensintervall.

→ Ein (1-α)-VI enthält den wahren Wert mit einer Wa. von (1-α).

Nur falls n gross ist, kann eine Normalapproximation gemacht werden! **Faustregel:** nπ >5 und n(1-π) >5

95% VI ist im 99% VI enthalten

P-Wert: Wa., unter gültiger H_0 das Ergebnis oder etwas noch extremeres (im Sinne der Alternative) zu erhalten

- Verwerfe H_0 falls $p \leq \alpha$
- Gewinnwa. π_0 unter Nullhypothese ist nicht im x VI für Gewinnwa. ⇒ $p \leq 1 - x$

Einseitiger Test:	**Zweiseitiger Test**:
eine Seite blind	sieht beide Seiten
man such nur grösser als oder kleiner als	man sucht grösser als und kleiner als
Grosse Macht	Kleine Macht

P-Wert

Die Wahrscheinlichkeit, die Beobachtung oder etwas noch Extremeres zu beobachten, falls H_0 wahr ist. Extrem: beide Seiten mit einbeziehen!! (p-wert: Summe der Wa. der Beobachtung und von den Wa.s von allen möglichen Beobachtungen mit noch kleineren Wahrscheinlichkeiten (sprich: Extremere Beobachtungen)

Gewinnwa. ist in 1-α-VI
→ auf α nicht verwerfen
p > α

$X \sim Bin(n,\pi)$ $Y \sim Bin(f,\pi) \Rightarrow$ unabhängig (+gleichwertig)
$\Rightarrow Y+X = Bin(f+n,\pi)$

Bernoulli nicht miteinander addieren $\Rightarrow n>1$ nicht mehr Ber

Summe von zwei unabhängigen Poissonvert. ist poissonvert.
$X \sim Pois(a)$ $Y \sim Pois(b)$ $Z = X+Y$ $Z \sim Pois(a+b)$

Binomialverteilung X~Bin(n,π)

Anzahl Erfolge (x) bei n unabhängigen Versuchen mit Erfolgswahrscheinlichkeit π.

$$\mathbf{P(X=x)} = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

ε(x)=nπ

σ²(x)=Var(x)=nπ(1-π) → am Größten, wenn π=0.5

Hypothesentest Binomialverteilung

1. Modell: X= # Erfolge von n Versuchen
X~Bin(n,π)

2. Nullhypothese
H_0: π_0 = ?
H_A: $\pi_A \neq \pi_0$, $\pi_A > \pi_0$, $\pi_A < \pi_0$
Einseitiger oder zweiseitiger Test?

3. Teststatistik
T= # Erfolge von n Versuchen unter der Annahme, dass H_0 stimmt.
T~Bin(n,π_0)

4. Signifikanzniveau
α = ? (meist 5% -> α = 0.05) → $P(x \leq k) \leq 0.05$

5. Verwerfungsbereich
$\pi_A \neq \pi_0$ $P(T \leq C_u) \leq \alpha/2$ und $P(T \geq C_o) \leq \alpha/2$
$\pi_A > \pi_0$ $P(T \geq C_o) \leq \alpha$
$\pi_A < \pi_0$ $P(T \leq C_u) \leq \alpha$

Normalapproximation von C_u/C_o:
Zweiseitig, α =0.05

$$C \approx n\pi_0 \pm 1.96\sqrt{n\pi_0(1-\pi_0)}$$

Einseitig, α =0.05: $\pi_A > \pi_0$: +1.64 → Obergrenze n
$\pi_A < \pi_0$: -1.64 → Untergrenze 0

6. Testentscheid
Liegt der beobachtete Wert im Verwerfungsbereich?
JA → H_0 wird verworfen
NEIN→ H_0 wird nicht verworfen **(aber auch nicht bewiesen)**

Bernoulliverteilung Y~Ber(π)

Erfolg oder Misserfolg bei nur einem Versuch? (Spez. Fall der Binom. mit n=1)

$$P(X=x) = \binom{1}{x} \pi^x (1-\pi)^{1-x}$$

$\mathbf{P(X=1)=\pi}$
$\mathbf{P(X=0)=1-\pi}$
ε(x)=π
σ²(x)=Var(x)=π(1-π)

Uniforme Verteilung Y~unif(n)

Alle Ereignisse sind gleich wahrscheinlich, n ist die Anzahl möglicher Ereignisse.

$$\mathbf{P(X=x)} = \frac{1}{n}$$

ε(x)= $\frac{(n+1)}{2}$

σ²(x)=Var(x)= $\frac{(n+1)(n-1)}{12}$

Hypergeometische Verteilung Y~hyper(n,N,m)

Modell: Urne mit N Kugeln, m sind markiert; es werden n gezogen und x davon sind markiert. Ohne zurücklegen!!

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

ε(x)= $\frac{n*m}{N}$

σ²(x)=Var(x)= $\frac{n*m*(N-n)(N-m)}{N^2*(N-1)}$

Poissonverteilung Y~Pios(λ)

Seltene Ereignisse und ihr vorkommen in einem bestimmten Zeitraum
λ: durchschnittliches Auftreten pro Zeit

$$\mathbf{P(X=x)} = e^{-\lambda} * \frac{\lambda^x}{x!}$$

ε(x)=λ
σ²(x)=Var(x)=λ

Summen von Poissonverteilten Zufallsvariablen:
Wenn X~Pois(λ_x) und Y~Pios(λ_y) unabhängig sind, dann ist X+Y~Pois($\lambda_x+\lambda_y$)

Prinzip „bis und mit"="1-ohne":
$P(X \leq x) = 1-P(X>x) = 1-P(X \geq x+1)$
$P(X \geq x) = 1-P(X<x) = 1-P(X \leq x-1)$

Für große n kann man Binomial mit Poisson annähern: $\lambda = n \cdot \pi$

Maximum likelyhood $\pi = \frac{x}{n}$

$H_A: p > p_0 \Rightarrow$ Für Verwerfungsbereich $P(X \geq k) \leq \alpha$

Verteilungen von diskreten Zufallsvariablen

In Versicherungen gängige Vert.
Ereignisse die beliebig oft eintreten können

	Binomialverteilung	Poissonverteilung	Allgemeine Formeln
Beschreibung	X = Anzahl „Erfolge/Treffer“ in n unabhängigen Versuchen (p = Erfolgswahrscheinlichkeit)	X = Anzahl „Erfolge“ pro Intervall (Zeit, Fläche usw.),	
Darstellung	$X \sim Bin(n,p)$	$X \sim Poisson(\lambda)$	
Werte, die X annehmen kann (diskret -> abzählbar)	• abzählbare, endliche Anzahl Ereignisse • 0, 1, 2, 3, 4 ..., n	• abzählbare, unendliche Anzahl Ereignisse • 0, 1, 2, 3, ...	• abzählbar • ...,-2, -1, 0, 1, 2, ...
Wahrscheinlichkeitsfunktion f(x)	$f(x) = P[X = x] = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$	$f(x) = P[X = x] = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$	Selber herausfinden, ohne Formel... z.B. bei Jasskarten: P[König]=P[X=4]=$\frac{4}{36} = \frac{1}{9}$
Kumulative Verteilungsfunktion F(x)	$F(x) = P[X \leq x] = \sum_{k=0}^{x} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$ **Wichtig: $F(n) = P[X \leq n] = 1$**	$F(x) = P[X \leq x] = \sum_{k=0}^{x} e^{-\lambda} \cdot \frac{\lambda^k}{k!}$	$F(x) = P[X \leq x] = \sum_{k \leq x} P[X = k]$
Erwartungswert E(X)	$E(X) = n \cdot p$	$E(X) = \lambda$	$E(X) = \sum_{i=1}^{n} k_i \cdot P[X = k_i]$
Varianz Var(X)	$Var(X) = n \cdot p \cdot (1-p)$	$Var(X) = \lambda$	$Var(X) = \sum_{i=1}^{n} (k_i - E(X))^2 \cdot P[X = k_i]$
Standardabweichung σ(X)	$\sigma(X) = \sqrt{Var(X)}$	$\sigma(X) = \sqrt{Var(X)}$	$\sigma(X) = \sqrt{Var(X)}$
Approximatives zweiseitiges 95% Vertrauensintervall	$I \approx \frac{x}{n} \pm 1.96 \cdot \sqrt{\frac{x}{n} \cdot \left(1 - \frac{x}{n}\right) \cdot \frac{1}{n}}$ Vertrauensintervall für die Erfolgswahr'keit p.	$I \approx x \pm 1.96\sqrt{x}$ Vertrauensintervall für λ.	
Approx. einseitiges 95% Vertrauensintervall	$\left]-\infty, \frac{x}{n} + 1.64\sqrt{\frac{x}{n} \cdot (1 - \frac{x}{n}) \cdot \frac{1}{n}}\right[$ für $H_A: p < p_0$ $\left]\frac{x}{n} - 1.64\sqrt{\frac{x}{n} \cdot \left(1 - \frac{x}{n}\right) \cdot \frac{1}{n}}, \infty\right[$ für $H_A: p > p_0$		

WICHTIG: Bernoulliverteilung $X \sim Bernoulli(p)$ ist ein Spezialfall der Binomialverteilung für n=1 (d.h. $X \sim Binom(1,p)$). Es wird nur ein einziger Versuch beschrieben, die Wahrscheinlichkeit für einen Erfolg ist gleich p und die Wahrscheinlichkeit für einen Misserfolg ist gleich 1-p.

WICHTIG: APPROXIMATIONEN: Für n = gross ist es mühsam, ohne Computer die kumulative Wahrscheinlichkeit $P[X \leq x]$ der Binomialverteilung zu berechnen. Deshalb wird meistens mit anderen Verteilungen approximiert:

1. **Normalapproximation (für n =gross, sehr häufig verwendet):** $X \approx Normal(\mu, \sigma^2); \mu = n \cdot p \; ; \sigma^2 = n \cdot p \cdot (1-p)$
   $P[X \leq x] \approx P\left[Z \leq \frac{x-\mu}{\sigma}\right] = \Phi\left(\frac{x-\mu}{\sigma}\right)$ → z-Tabelle
2. **Poissonapproximation (für n=gross und p=klein):** $X \approx Poisson(\lambda); \lambda = n \cdot p$

! Median ist nicht immer in Stichprobe

STETIGE VERTEILUNG (Wertebereich für x kontinuierlich, keine „Punkt"-Wahrscheinlichkeiten! P(X=x)=0 für alle x∈W_x, nur P(X≤x) möglich, Fläche unter der Kurve, Intervalle!!)

α-Quantil (empirisch)

Wert q_α, bei dem mindestens α *100% der Datenpunkte kleiner und (1-α) *100% grösser als q_α sind.
ACHTUNG! Daten der Reihenfolge nach sortieren!

$x_1 \le x_2 \le \cdots \le x_n$

$q_{0.25}$ = 1. Quartil
$q_{0.50}$ = Median
$q_{0.75}$ = 3. Quartil

$\alpha * n$

Keine ganze Zahl: $k = \alpha * n + \frac{1}{2}$, k runden, $q_\alpha = x_k$ (die k-te Zahl)

Ganze Zahl: $q_\alpha = \frac{x_{\alpha*n} + x_{\alpha*n+1}}{2}$

Inter-Quartile-Range: (~68% aller Werte)
IQR= $q_{0.75}$ - $q_{0.25}$
IQR und Median: robust gegen Ausreisser
P(X≤ q_α)=α

Kennzahlen

Für die Lage:
Arithmetisches Mittel = Durchschnitt (=μ)

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

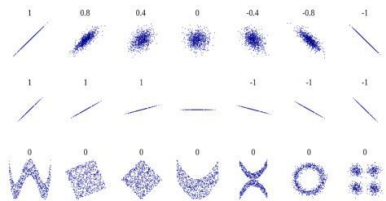
→ nicht robust gegen Ausreisser

Für die Streuung („Breite"):
Empirische Standartabweichung (=σ)

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

Korrelation: $y = a \cdot x$ → a ist Steigung im Streudiagramm zw. x und y
↳ misst wie stark Punkte auf einer Linie liegen
Gerade: Korrelation 1 | kein linearer Zusammenhang: Korrelation 0

Für linearen Zusammenhang: (zwischen X und Y, zwei stetigen Zufallsvariablen) Y=a+bX
E(Y)=a+bE(X); $Var(Y)=b^2Var(X)$; $q_{\bar{y}}=a+bq_{\bar{x}}$
wenn x normlverteilt, y auch; wenn nicht, keine Aussage möglich
Korrelation [-1,1]: (KEIN kausaler Zusammenhang!!!)



Stärke und Richtung von linearen Abhängigkeiten, sagt nichts über Steigung aus!
cor=0; kein linearer Zusammenhang, aber evt. Sonst
Empirische Korrelation:

$$r = \frac{S_{xy}}{S_x S_y} \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\frac{1}{n-1}$$

Gesetz der grossen Zahlen (GGZ)

$x_1, x_2 \dots x_n \sim$ F i.i.d (independent and identical distributed; F: irgendeine Verteilung)

$$E(\bar{x}_n) = \mu \qquad \sigma_{\bar{x}_n} = \frac{\sigma_x}{\sqrt{n}}$$

$Var(\bar{x}_n) \neq Var(x_i)$

Wurzel-n-Gesetz: Für eine halb so grosse Standardabweichung (Streuung) braucht man 4mal so viele Daten! (gilt für Std.abw. von Mittelwerten, nicht einzelne Messungen)

Zentraler Grenzwertsatz (ZGWS)

$x_1, x_2 \dots x_n \sim$ F i.i.d → aus GGZ: $\bar{x}_n \sim N(\mu, \frac{\sigma_x^2}{n})$

$S_n \sim N(n*\mu, n*\sigma_x{}^2)$

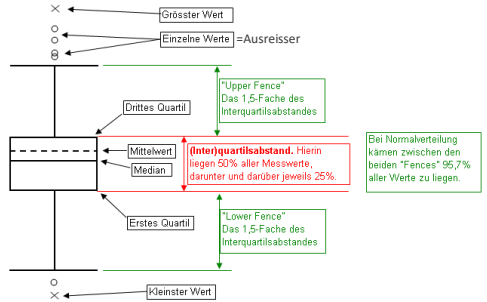
Bsp.: n=100 Spiele, E(x)=1/3, Var(x)=28.6
S_n=totale Gewinne (Standardisieren S.6)

$S_n \sim N(100 * {}^1/_3, 100 * 28.6) = N(33{,}2860)$

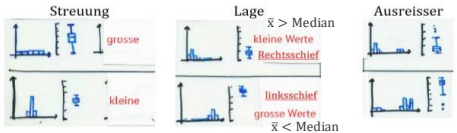
→ müssen unabhängig und gleich verteilt sein
→ Für große n ist Approximation besser

Graphische Methoden

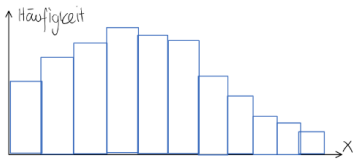
Boxplot



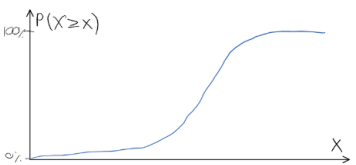
Hauptmerkmale zur Diskussion, bzw. zum Vergleich:



Histogramm (weit in Mitte)



Empirische kumulative Verteilungsfunktion (ECDF)



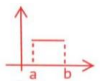
Kontinuierliche Verteilungsfunktionen

Wahrscheinlichkeitsdichte: **f(x)**
Kummulative Verteilungfnk.: **F(x)** (Integral von f(x))

Uniforme Verteilung X~Unif([a,b])

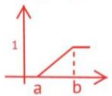
Jeder Wert ist gleich wahrscheinlich

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



$P(x \le X)$

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



$$\mathcal{E}(X) = \frac{a+b}{2},$$
$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}, \quad \sigma_X = \frac{b-a}{\sqrt{12}}.$$

Exponentialverteilung X~Exp(λ)

Warten ohne Gedächtnis (Bsp.: Radioaktiver Zerfall)
- Einfaches Modell für Wartezeiten auf Ausfälle
- Wenn die Zeit zwischen den Ausfällen Exponential-verteilt sind (λ), dann ist die Anzahl Ausfälle in einem Intervall t Poisson-verteilt (tλ).

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

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

$$\mathcal{E}(X) = \frac{1}{\lambda},$$
$$\mathrm{Var}(X) = \frac{1}{\lambda^2}, \quad \sigma_X = \frac{1}{\lambda}.$$

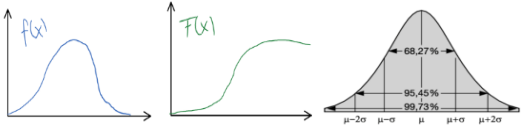
- $P(T > t) = 1 - P(T \le t) = e^{-\lambda t}$
- $P(T > t+s | T > s) = \frac{P(T>t+s \text{ und } T>s)}{P(T>s)} = \frac{P(T>t+s)}{P(T>s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} =$
  $= e^{-\lambda t} = P(T > t)$
  "Es spielt keine Rolle, ob man schon s Sekunden gewartet hat"

→ Menschl. Leben keine Ex-Vert., da Wa. zu Sterben im Alter zunimmt

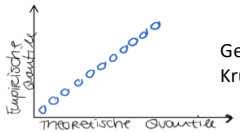
Normalverteilung X~N(μ,σ²) (σ(x) = √σ²)

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

$F(x) = ???$ → Standardnormalverteilung



Prüfen ob Normalverteilt: QQ-Plot (S.9)



Gerade: Normalverteilt
Krumm: NICHT n.v.

→ Wenn Steigung kleiner ⇒ kleinere Varianz

Standardnormalverteilung Z~N(0,1)

Mittelwert μ wird auf 0 gesetzt (Symmetrieachse=x-Achse)
Standardabweichung σ wird auf 1 gesetzt (x[-1,1])

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

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \varphi(x)\, dx$$

TABELLE!!

Von Normalverteilt zu Standardnormalverteilt:
Standardisieren (oder Normieren):

$$\mathbf{x_n = \frac{x - \mu}{\sigma}} \qquad \mathbf{disnormieren: x = x_n \sigma + \mu}$$

Bsp.: X~N(2,2²) → P(X≤5)=? (=0.93)

P(X≤5) = P(Z≤ $\frac{5-\mu}{\sigma}$) = P(Z≤ $\frac{5-2}{2}$) = P(Z≤ $\frac{3}{2}$)

P(Z≤ $\frac{3}{2}$) = Φ(1.5) =(Tabelle)= 0.93

Wichtig: Quadrat!! Wir rechnen mit σ, nicht mit σ²! Also zuerst Wurzelziehen!

Normalvert.: Median = Erwartungswert
↳ symmetrisch am E verteilt

Für z-Test: $\bar{X}_n = N(\mu, \sigma^2_{\bar{x}n}) \rightarrow N(\mu,1) = \frac{x_n - \mu}{\sigma_{\bar{x}n}}$
⇒ Für t-Test nicht weil σ geschätzt ist

Streuungswerte: Standardabweichung s.d.
Streuung arith. Mittel: " s.e.d.

Zur Erinnerung:
P(Z≤ - 0.21) = 1 - P(Z≤0.21)
P(Z≥0.21) = 1 - P(Z≤0.21)

Stichprobengröße: $n \geq 4 \cdot \frac{\hat{\sigma}_x^2}{\delta^2}$

$\delta = \frac{1}{2}$ Breite VI

Statistische Tests (für Erwartungswert μ)

Z-Test basierend auf mehreren Beobachtungen

σ_x: bekannt; $\sigma_{\overline{x_n}}$: berechnen

X_i: kontinuierlich, normalverteilt; Gepaarte Variablen!

1. Modell: $X_1,\dots,X_n \sim N(\mu,\sigma_x^2)$

2. Nullhypothese

H_0: $\mu_0 = ?$

H_A: $\mu_A \neq \mu_0$, $\mu_A > \mu_0$, $\mu_A < \mu_0$

Einseitiger oder zweiseitiger Test?

3. Teststatistik $Z \sim N(0,1)$

$$Z = \frac{\overline{X_n}-\mu_0}{\sigma_{\overline{x_n}}} = \frac{\sqrt{n}(\overline{X_n}-\mu_0)}{\sigma_x} \qquad \sigma_{\overline{x_n}} = \frac{\sigma_x}{\sqrt{n}}$$

$(= \frac{\text{beobachtet} - \text{erwartet}}{\text{Standartfehler}})$

4. Signifikanzniveau

α = ? (meist 5% -> α = 0.05)

5. Verwerfungsbereich →TABELLE!

$\mu_A \neq \mu_0$ K= $(-\infty, -\Phi^{-1}(1-\alpha/2)] \cup [\Phi^{-1}(1-\alpha/2), \infty)$

$\mu_A > \mu_0$ K= $[\Phi^{-1}(1-\alpha), \infty]$

$\mu_A < \mu_0$ K= $[-\infty, -\Phi^{-1}(1-\alpha)]$

6. Testentscheid

Liegt der beobachtete Wert (in Z einsetzen) im Verwerfungsbereich?

JA → H_0 wird verworfen

NEIN→ H_0 wird nicht verworfen **(aber auch nicht bewiesen)**

Vertrauensintervall für μ (α-V.I.)

$\mu_A \neq \mu_0$ $\left[\overline{x_n} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right) * \frac{\sigma_x}{\sqrt{n}}\right]$

$\mu_A > \mu_0$ $\left[\overline{x_n} - \Phi^{-1}(1-\alpha) * \frac{\sigma_x}{\sqrt{n}}, \infty\right)$

$\mu_A < \mu_0$ $\left(-\infty, \overline{x_n} + \Phi^{-1}(1-\alpha) * \frac{\sigma_x}{\sqrt{n}}\right]$

Schreibweise: $\Phi^{-1}\left(1-\frac{\alpha}{2}\right) = Z_{1-\alpha/2}$

T-Test besser als Z-Test

Bsp. $P(Z \leq 2) \approx 0.97 > P(T \leq 2) \approx 0.95$

Vertrauensintervall: enthält unplausible Werte für Teststatistik falls Nullhypothese stimmt und plausible Parameterwerte

10 Pers. Sprint 2x
1x mit Koffein
1x ohne
→ äußere Faktoren können variieren, intrinsische sind gleich

Gepaarter T-Test EINSTICHPROBENTEST

σ_x: unbekannt; $\widehat{\sigma_x}$: geschätzt

X_i: kontinuierlich, normalverteilt; Gepaarte Variablen!

SCHÄTZEN: $\widehat{\sigma_x}$ und $\widehat{\sigma_x}^2$ → geschätzte Varianz!

$$\widehat{\sigma_x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2 \qquad S_x \text{ in TR!}$$

1. Modell: $X_1,\dots,X_n \sim N(\mu,\sigma_x^2)$

2. Nullhypothese

H_0: $\mu_0 = ?$

H_A: $\mu_A \neq \mu_0$, $\mu_A > \mu_0$, $\mu_A < \mu_0$

Einseitiger oder zweiseitiger Test?

3. Teststatistik $T \sim t_{n-1}$

$$T = \frac{\overline{X_n}-\mu_0}{\sigma_{\overline{x_n}}} = \frac{\sqrt{n}(\overline{X_n}-\mu_0)}{\widehat{\sigma_x}} \qquad \sigma_{\overline{X_n}} = \frac{\widehat{\sigma_x}}{\sqrt{n}}$$

4. Signifikanzniveau

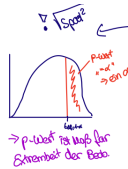
α = ? (meist 5% -> α = 0.05)

5. Verwerfungsbereich →TABELLE!

$\mu_A \neq \mu_0$ K= $(-\infty, -t_{n-1,\,1-\alpha/2}] \cup [t_{n-1,\,1-\alpha/2}, \infty)$ zweiseitig

$\mu_A > \mu_0$ K= $[t_{n-1,\,1-\alpha}, \infty)$ hier gepoolt

$\mu_A < \mu_0$ K= $(-\infty, -t_{n-1,\,1-\alpha}]$ einseitig



6. Testentscheid

Liegt der beobachtete Wert im Verwerfungsbereich?

Beob. Wert einsetzen: $t = \frac{\sqrt{n}\,(?-\mu_0)}{\widehat{\sigma_x}}$

JA → H_0 wird verworfen

NEIN→ H_0 wird nicht verworfen **(aber auch nicht bewiesen)**

Vertrauensintervall für μ (α-V.I.) → 99%-VI → α = 1%

$\mu_A \neq \mu_0$ $\left[\overline{x_n} \pm t_{n-1,1-\alpha/2} * \frac{\widehat{\sigma_x}}{\sqrt{n}}\right]$

$\mu_A > \mu_0$ $\left[\overline{x_n} - t_{n-1,1-\alpha} * \frac{\widehat{\sigma_x}}{\sqrt{n}}, \infty\right)$

$\mu_A < \mu_0$ $\left(-\infty, \overline{x_n} + t_{n-1,1-\alpha} * \frac{\widehat{\sigma_x}}{\sqrt{n}}\right]$

Anzahl Fehler bei Test bei α = a und x-Tests insg. = ax

Schreibweise: t_{df} → degrees of freedom

Generell: Je kleiner df, desto wahrscheinlicher sind Werte mit grossem Betrag

→ gepaarter 2 Stichproben T-test: gleiche Formel (Diff. zw. 2μ) 1μ

10 Pers. 1x Sprint
5 mit Koffein
5 ohne
⇒ nicht genau 1 kann 1 zugeordnet werden

Ungepaarter T-Test ZWEISTICHPROBENTEST

Annahme: $\sigma_x^2 = \sigma_y^2$ → gleiche Varianz!

$\overline{X_n} = \frac{1}{n}\sum x_i$ $\overline{Y_m} = \frac{1}{m}\sum y_i$ → arithmetisches Mittel

$$S_{pool}^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(x_i - \overline{x_n})^2 + \sum_{i=1}^{m}(y_i - \overline{y_m})^2\right)$$

$$= \frac{1}{n+m-2}\left((n-1)\widehat{\sigma_x}^2 + (m-1)\widehat{\sigma_y}^2\right)$$

1. Modell: $X_1,\dots,X_n \sim N(\mu,\sigma_x^2)$

2. Nullhypothese

H_0: $\mu_0 = ?$; $\mu_x = \mu_y$

H_A: $\mu_A \neq \mu_0$, $\mu_A > \mu_0$, $\mu_A < \mu_0$

Einseitiger oder zweiseitiger Test?

3. Teststatistik $T \sim t_{n+m-2}$

$$T = \frac{\overline{X_n}-\overline{Y_m}}{S_{pool}\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

beachte: oben S_{pool}^2, hier S_{pool}

4. Signifikanzniveau

α = ? (meist 5% -> α = 0.05)

5. Verwerfungsbereich →TABELLE!

$\mu_A \neq \mu_0$ K= $(-\infty, -t_{n+m-2,\,1-\alpha/2}] \cup [t_{n+m-2,\,1-\alpha/2}, \infty)$

$\mu_A > \mu_0$ K= $[t_{n+m-2,\,1-\alpha}, \infty)$

$\mu_A < \mu_0$ K= $(-\infty, -t_{n+m-2,\,1-\alpha}]$

6. Testentscheid

Liegt der beobachtete Wert im Verwerfungsbereich?

Beobachteten Wert einsetzen

JA → H_0 wird verworfen

NEIN→ H_0 wird nicht verworfen **(aber auch nicht bewiesen)**

GEPAART	UNGEPAART
Gleichgrosse Stichproben	Unterschiedliche Stichproben
Klare Zuordnung (rechts/links, vorher/nachher)	Keine Zuordnung
Differenz der Paare	
MEHR MACHT	WENIGER MACHT

ungepaarter T-test: andere Formel: keine Differenz → 2μ

p-wert in Tabelle: $t_{df,x} = T$

$p = 2 \cdot (1-x)$

↳ zweiseitig

Mann-Whitney U-Test (Zweistichproben W-Test)

Annahme: $X_i \sim F$ i.i.d.
$Y_i \sim G$ i.i.d.

Teste μ: H_0: F=G (X_i und Y_i gleich verteilt)
H_A: F=G+γ (X_i, Y_i gleiche Form aber verschoben)

Prinzip: H_0: F=G
1. Absolute Beträge der beob. Werte nehmen
2. Ränge nehmen
3. getrennt nach positiven und negativen Ausgangswerten Rangsummen bilden
4. mit PC U-Test machen

→ falls H_0 stimmt: + bzw. – Rangsummen etwa gleich, sonst stehen sie in einem bestimmten Verhältnis.

Welch-Test: setzt nicht voraus, dass σ beider Gruppen gleich sind.

Bonferroni Korrektur Multiples Testen

Fehler 1. Art (falsch positiv) höchstens = α

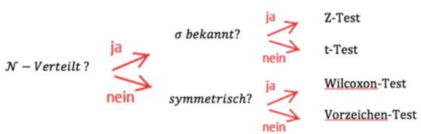
Bsp.: 1000 Tests mit α=0.05 (H_0 sei bei jedem wahr) Wegen Fehler 1. Art (H_0 verwerfen obwohl wahr) werden bei etwa 5% der Tests H_0 fälschlicherweise verworfen. ⇒ SG

Ziel: für alle 1000 Tests insgesamt eine Wa. von 5% zu haben, H_0 fälschlicherweise zu verwerfen!

Die Tests werden auf dem neuen Signifikanzniveau $\tilde{\alpha} = \frac{\alpha}{n}$ (n: Anzahl Tests) durchgeführt.

→ verringert Wa. für falsch positiv für alle Test zusammen auf ≤ 5 %

Welchen Test muss ich durchführen?



Stichprobengrösse (Faustregel)

$n \geq 4\frac{\widehat{\sigma_X}^2}{\delta^2}$ 2δ: Breite des 95%-VI

Vertrauens-/Vorhersage-Intervall

95% Vertrauensintervall < 95% Vorhersageintervall

Vertrauensintervall:	Vorhersageintervall:
Durchschnitt (viele Messungen) Gibt erwarteten Wert mit Wa. 1-α heraus	Man kann bei EINER Messung etwa so viel erwarten Grössere Streuung

Wa., dass Test bei wahrem H_0 kein sign. Ergebnis zeigt: $1-\alpha$
→ Wa., dass bei n Tests keiner eins zeigt: $(1-\alpha)^n$
→ Wa., dass mind. 1 Test fälschlicherweise verwirft: $1-(1-\alpha)^n$

→ Fehler folgt Normalverteilung mit Erwartungswert 0
→ für gegebene x folgt Y einer Normalverteilung mit $E(Y)=\beta_0+\beta_1 x$

Lineare Regression

R-Output:

Bsp.: Ein Bauer notiert den Ertrag von Kartoffelfeldern in kg pro m². Er stellt fest, dass sich durch Dünger den Ertrag erhöhen lässt, ist sich jedoch nicht sicher ob es wirklich am Dünger liegt. Er versucht deshalb, den Ertrag durch eine Lineare Regression zu beschreiben. Falls es am Dünger liegt, hängt der Ertag somit vom Grundertrag (β_0), vom Zusatzertrag durch Einsatz von x Einheiten Dünger (β_1*dünger) und von einem Standartfehler (ε_i) ab.

H_0: $\beta_1=0$ (Dünger hat keinen Einfluss)
H_A: β_1 hilft zur Erklärung des Modells (es kommt auf die Menge Dünger an)

$ertrag=\beta_0+\beta_1\cdot duenger+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma)$.

Der (unvollständige) Regressionsoutput sieht wie folgt aus:

Residuals:

Min	1Q	Median	3Q	Max
-86.00	-9.96	3.17	10.99	35.86

Coefficients: β σ(β) t(β) Teststatistik P-Wert

	Estimate	Std. Error	t value	Pr(>|t|)
β_0 (Intercept)	3.0437	6.5182	0.47	0.64
β_1 duenger	0.5471	0.0612	8.93	1.3e-11 ***

Achsenabschnitt, Steigung

---

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.6 on 46 degrees of freedom
Multiple R-squared: 0.634, Adjusted R-squared: 0.626

Anz. Datenpunkte: Anz. Felder = degrees of freedom + Anz. β's = 46+2= 48 = n (Meistens 2)
→ df=n-2

Koeffizienten: $\beta=\sigma(\beta)*t(\beta)$
Bsp.: $\beta_1=\sigma(\beta_1)*t(\beta_1)=0.0612*8.93=0.5471$

95%-VI: genau: $VI(\beta)=\beta\pm t_{df,0.975}*\sigma(\beta)$
Bsp.: $VI(\beta_1)=\beta_1\pm t_{46,0.975}*\sigma(\beta_1)=0.5471\pm2.01*0.0612=[0.42,0.67]$
→ da 0 nicht in VI, H_0 wird verworfen
Approx.: $VI(\beta)=\beta\pm2*\sigma(\beta)$

Verwerfungsbereich: $K(\beta)=(-\infty,-t_{df,1-\alpha/2}]\cup[t_{df,1-\alpha/2},\infty)$
Bsp.: für $\alpha=0.05$ $K(\beta_1)=(-\infty,-2.01]\cup[2.01,\infty)$
→ da $t(\beta_1)=8.93$ in K, H_0 wird verworfen

P-Wert: $t(\beta)=t_{df,1-P\text{-}Wert/2}$ → Tabelle: was ist P-Wert? (1-P-Wert/2)
Bsp.: $t(\beta_0)=t_{46,1-P\text{-}Wert/2}=0.47$ → Tabelle: (1-P-Wert/2)~ 0.68
P-wert = 2*(1-0.68)=0.64

→ Wenn t-Wert weit über größtem Wert von Tabelle in Zeile mit df ist, ist p-Wert null

→ P-Wert = α

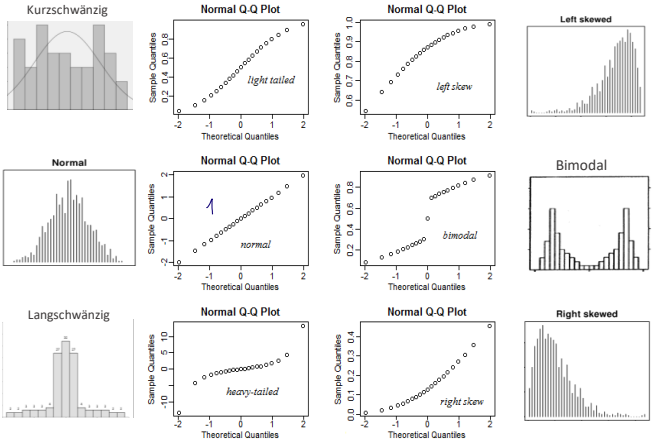
t-value = Estimate / Std. Error → Wert in Tabelle für t_n-Verteilung → bei zweiseitig: verdoppeln

Erwarteter Ertrag: Menge Dünger = x → Ertrag=$\beta_0+\beta_1$*x
Bsp.: Bei 1.5l/m² Dünger: Ertrag=3.0437+0.5471*1.5=3.8643

QQ-Plot und Histogramm (siehe auch S.5)

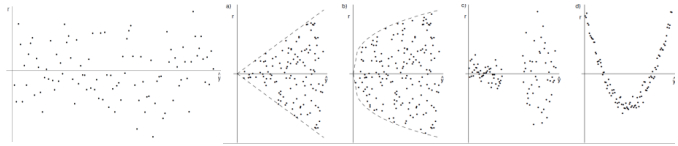
Ist es plausibel, dass QQ-Plot und Hitsogramm von denselben Daten stammen?
Welche Graphiken gehören zusammen?
Zur Erinnerung: nur bei Geraden im QQ-Plot ist die Normalitätsannahme bestätigt.

uniform



Tukey-Anscombe Plot

Ist die Fehlervarianz konstant? Passen die Daten zum Modell?
Ja: Nein:

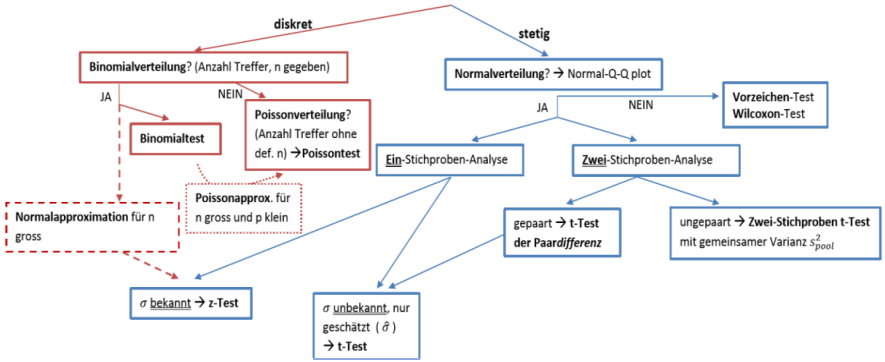


Fehlervarianz nicht konstant

Systematischer Fehler

Residuen: vertikaler Abstand zw. beobachtetem Punkt und gefitteter Gerade

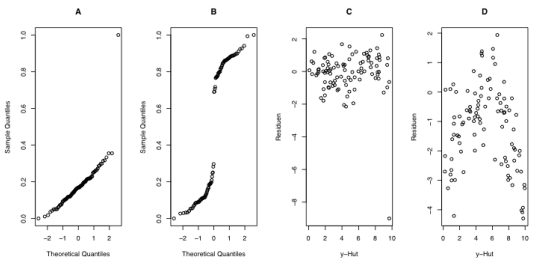
Man kann nicht alle Modellannahmen mit TA überprüfen



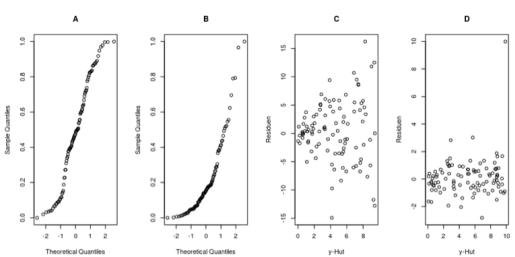
Zufallsvariablen Streudiagramme → arith. Mittel muss kleineren Stichprobenwert entsprechen

→ Median ($\alpha = 50\%$ - Quantil) nur wenn $\alpha \cdot n$ keine ganze Zahl ist

⇒ dann Zahl aus Stichprobe



- ✓ **a)** Die Residuen in 'A' könnten von einer Normalverteilung mit Aussreissern stammen.
- ✗ **b)** Die Residuen in 'B' könnten von einer Normalverteilung mit Aussreissern stammen.
- ✓ **c)** Plot 'C' zeigt, dass abgesehen von einigen Ausreissern die Linearitätsannahme erfüllt ist.
- ✓ **d)** Plot 'D' zeigt, dass die Linearitätsannahme des geschätzten Modells nicht angebracht ist.



- ✓ **a)** Betrachten Sie den QQ-Plot in Abbildung 'A'. Es ist plausibel, dass die Residuen von einer Uniformen Verteilung stammen.
- ✓ **b)** Betrachten Sie den QQ-Plot in Abbildung 'B'. Es ist plausibel, dass die Residuen von einer Rechtsschiefen Verteilung (d.h. die Dichtefunktion fällt bei grossen Zahlen langsamer ab als bei kleinen Zahlen) stammen.
- ✗ **c)** Betrachten Sie den TA-Plot in Abbildung 'C'. Der Plot zeigt konstante Varianz der Residuen.
- ✗ **d)** Betrachten Sie den TA-Plot in Abbildung 'D'. Der Plot zeigt keine auffallenden Abweichungen von den Modellannahmen.