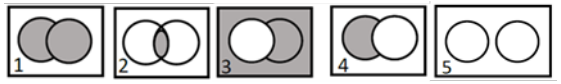
1. MODELLE FÜR ZÄHLDATEN

1.1 WAHRSCHEINLICHKEITSMODELLE

Ω	Grundraum	{1,2,3,4,5,6}
ω	Elementarereignis	{6}
A	Ereignis ($A \in Q$)	{2,4,6}
$\emptyset$	Leere Menge	{ }

$P(A)$	≥ 0	Wa. von A
$P(\Omega)$	= 1	gesamte Wa.

$P(A \cup B)$	$P(A) + P(B) - P(A \cap B)$	A oder B	[1]
$P(A \cap B)$		A und B	[2]
$P(A^c) \mid P(\bar{A})$	$1 - P(A)$	nicht A	[3]
$P(A \backslash B)$	$P(A) - P(A \cap B)$	A ohne B	[4]



- Disjunkt = getrennte Elemente [5]
  $P(A \cap B) = 0 = \{\ \} \mid P(A|B) = 0 \mid P(A \cup B) = P(A) + P(B)$

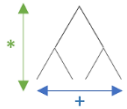
1.2 WAHRSCHEINLICHKEITEN BERECHNEN

- Laplace-Modell = alle Elementarereignisse sind gleich wahrsch
  $P(A) = \frac{Anzahl\ günstige\ \omega}{Anzahl\ mögliche\ \omega}$
  Bsp. Fairer Würfel:
  Wa. eine 6 zu Würfeln: 1/6
  Wa. eine gerade Zahl zu Würfeln 3/6 = 1/2
- mit Gegenereignis rechnen → z.T. weniger Rechenaufwand

1.3 UNABHÄNGIGKEIT

- Auftreten von A beeinflusst Wa. von B nicht → gleicher Baum
  $P(A \cap B) = P(A) * P(B)$ $P(A) = P(A|B)$
  Bsp. Wa. zweimal hintereinander eine 6 zu Würfeln $\frac{1}{6} * \frac{1}{6}$

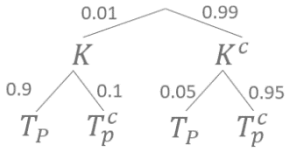
1.4 ABHÄNGIGKEIT / BEDINGTE WAHRSCHEINLICHKEIT



- Wa. beeinflussen sich → ungleicher Baum
- Bedingte Wa. von A wen B eingetreten: $P(A|B)$
  $P(A|B) \neq P(B|A)$
- Satz von Bayes: $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) * P(A)}{P(B)}$
  → richtiger Weg & korrektes Ergebnis dividiert durch alle Wege, welche das gleiche Ergebnis hervorrufen
- Satz der totalen Wa. (zwei Ereignisse)
  $P(B) = P(B|A) * P(A) + P(B|A^c) * P(A^c)$
  bei mehr Ereignissen: $P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$

Bsp. Wa., dass wenn Test positiv (T_p) auch tatsächlich die Krankheit (K) vorhanden ist -> Gesucht: $P(D|T)$
Gegeben: $P(K) = 0.01 \rightarrow P(K^c) = 0.99$
$P(T_P|K) = 0.9 \rightarrow P(T_p^c|K) = 0.1$
$P(T_p^c|K^c) = 0.95 \rightarrow P(T_p|K^c) = 0.05$

Satz von Bayes: $P(K|T_p) = \frac{P(T_p|K) * P(K)}{P(T_p)} \rightarrow P(T_p)$ gesucht
Satz tot. Wa. $P(T_p) = P(T_p|K) * P(K) + P(T_p|K^c) * P(K^c)$
$= 0.9 * 0.01\ 1\ 0.05 * 0.99 = 0.0585 \rightarrow$ pos. Vorhersagewert
$P(K|T_p) = \frac{0.9*0.01}{0.0585} = 0.1538 \approx 15.4\%$



1.5 ODDS, LOG-ODDS, ODDS-RATIO

- Odds = wie viel Mal wahrsch. dass A eintritt statt A^c
  $odds\ (A) = \frac{P(A)}{1-P(A)} = \frac{P(A)}{P(A^c)}$ wenn odds ↑ dann auch log-odds ↑
  $odds(A^c) = \frac{1}{odds(A)}$
- Log-Odds $\ln(Odds(A)) = \ln\left(\frac{P(A)}{P(A^c)}\right)$
  $\ln(Odds(A^c)) = -\ln(Odds(A))$

Krankheit \ Risikofaktor	Ja	Nein
Ja	a	b
Nein	c	d

- Odds Ratio (OR) = Chancenverhältnis, Zahl, welche etw. Über die Stärke eines Zusammenhangs aussagt
  $OR = \frac{\text{odds}(A|B)}{\text{odds}\ (A|B^c)} = \frac{a*d}{c*b}$
  $P(A) = \frac{a}{a+c} \rightarrow$ Wa. zu erkranken mit Risikofaktor
  $P(B) = \frac{b}{b+c} \rightarrow$ Wa. zu erkranken ohne Risikofaktor
- Risk difference (RD) = Wirksamkeit eines bestimmten Faktors
  $RD = P(A|B) - P(A|B^c)$
- Risk Ratio (RR) = Wa. durch einen bestimmten Faktor vergrössert / verkleinert wird
  $RR = \frac{P(A|B)}{P(A|B^c)}$

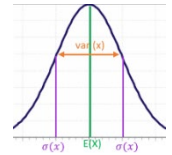
Bsp. Wie hoch ist das Chancenverhältnis an einer Krankheit (K) zu erkranken, wenn unter Einfluss eines Risikofaktors (RF) steht? (z.B. Lungenkrebs, Raucher) → Gesucht: OR
Gegeben: $P(K|RF) = 0.54$ $P(K|RF^c) = 0.18$
$RD = P(K|RF) - P(K|RF^c) = 0.54 - 0.18 = 0.36$
$RR = \frac{P(K|RF)}{P(K|RF^c)} = \frac{0.54}{0.18} = 3$
$odds(K|RF) = \frac{P(K|RF)}{P(K^c|RF)} = \frac{0.54}{1-0.54} = 1.17$
$odds(K|RF^c) = \frac{P(K|RF^c)}{P(K^c|RF^c)} = \frac{0.18}{1-0.18} = 0.22$
$OR = \frac{odds(K|RF)}{odds(K|RF^c)} = \frac{1.17}{0.22} = 5.33$
→ Wa. Krank zu werden ist mit diesem RF ~5-mal grösser

2. DISKRETE WAHRSCHEINLICHKEITSVERTEILUNGEN

2.1 ZUFALLSVARIABLE

- Diskrete Wa.verteilung → nur für bestimmte Werte definiert x nimmt endliche, abzählbare Werte an
- Kumulative Verteilungsfkt. → Summe der Wa. welche ≤ Wert x sind → stetig steigend, höchster Wert ist 1 → springt an den Stellen, die zum Wertebereich gehören
- Zufallsvariable = Zufallsexperiment = als Ergebnis eine Zahl hat
- Grossbuchstabe = Funktion || Kleinbuchstabe = konkreter Wert

2.2 KENNZAHLEN EINER VERTEILUNG



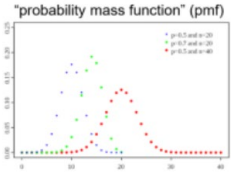
- Erwartungswert = mittlere Lage einer Verteilung
  $E(X) = \sum x * P(X = x)$
  → Zahl$_1$ * Wa$_1$ + Zahl$_2$ * Wa$_2$ + …
- Varianz = Abweichung vom Erwartungswert
  $Var(X) = \sum(x - E(X))^2 * P(X = x)$
- Standardabweichung = Streuung der Verteilung
  $\sigma(X) = \sqrt{Var(X)}$

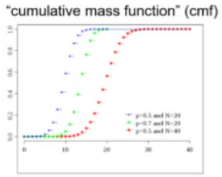
2.3 BINOMIALVERTEILUNG

- Beschreibung des Eintreffens / Nicht-Eintreffens eines bestimmten Ereignisses
- Binomialkoeffizient (TR: nCr(n, x), menu;5;3) = auf wie viele Arten man x Dinge auf n Plätzen anordnen kann (Lottozahlen)
  $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ $\binom{n}{2} = \frac{n(n-1)}{2} \rightarrow$ # Händeschütteln
- Binomialverteilung: Anz. Erfolge in einer Serie von gleichartigen, unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben (= Bernoulli-Prozesse)
  $X \sim Bin(n, \pi)$ $P(X = x) = \binom{n}{x} * \pi^x * (1-\pi)^{n-x}$ (binompdf)
  x = # Erfolge | n = # Versuche | π = # Wa. für Erfolg
  TR: menu;5;5;A → pdf || menu;5;5;B → cdf
  $P[X = x] = binompdf(n, \pi, x)$ «genau x Treffer»
  $P[X \leq x] = binomcdf(n, \pi, x)$ «höchstens x»
  $P[X < x] = binomcdf(n, \pi, x-1)$
  $P[X \geq x] = 1 - binomcdf(n, \pi, x-1)$ «mindestens x»
  $P[X > x] = 1 - binomcdf(n, \pi, x)$

$E(X) = n * \pi$ $Var(X) = n\pi(1-\pi)$
$\sigma(X) = \sqrt{Var(X)} = \sqrt{n * \pi * (1 - \pi)}$

- Streuung Var(x) wächst mit n → aber immer langsamer
- Festes n ist $Var_{max}(x)$ bei $\pi = 0.5$
- P(X=x) (pmf = probability mass fkt) ist maximal wenn $x \approx n * \pi$
- Wenn n gross, sind Wa. nur um $n * \pi$ herum gross, sonst klein
- Wenn $n * \pi * (1 - \pi)$ nicht zu klein → Verteilung symmetrisch





2.4 BERNOULLIVERTEILUNG

- Erfolg oder Misserfolg bei nur einem Versuch

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

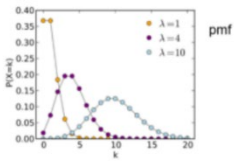
$$E(X) = \pi \qquad Var(X) = \pi(1 - \pi) = \sigma^2$$

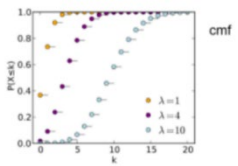
- Unabhängige Zufallsvariable
  $X \sim Bin(n, \pi) \mid Y \sim Ber(\pi) \rightarrow X + Y \; bin(n + 1, \pi)$

2.5 POISSONVERTEILUNG

- Seltene Ereignisse werden in einem vorgegebenen Zeitraum gezählt (Bspw. Unfallraten, radioaktiver Zerfall)
  → keine klare Oberschranke vorhanden
  $X \sim Pois(\lambda) \quad P(X = x) = \frac{\lambda^x}{x!} * e^{-\lambda} \; (x \in \{0,1, \dots \infty\})$
  λ = durchschnittliches auftreten pro Zeit
  ∞ = einfach eine grosse Zahl im TR eintippen
- Poissonverteilungen addierbar, wenn unabhängig voneinander → $E((X + Y)$ || $E(a * X) = a * E(X)$
- Poisson-Approximation: für grosse n und kleine π gilt: Bin(n,π) ≈ Pois(λ)
- TR: poissCdf (λ, untere Grenze, obere Grenze) (menu;5;5; k)

$$E(X) = \lambda \qquad Var(X) = \lambda \qquad \sigma(X) = \sqrt{\lambda}$$





2.6 UNIFORME VERTEILUNG

- Alle Ereignisse sind gleich wahrsch.
  $X \sim Unif(n) \quad P(X = x) = \frac{1}{n} \; (x \in \{0,1, \dots, n\})$

$$E(X) = \frac{n+1}{2} \qquad Var(X) = \frac{(n+1)(n-1)}{12} = \sigma^2$$

2.7 HYPERGEOMETRISCHE VERTEILUNG

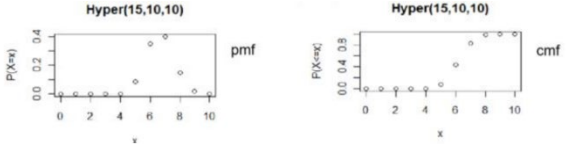
- Gewinwa. nicht konst., da abhängig was davor gezogen wurde → vernachlässigbar, wenn sehr grosse Lostrommel und wenige Ziehungen

$$X \sim hyper(N, n, m) \qquad P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} = \frac{\text{«günstige»}}{\text{«mögliche»}}$$

TR (gespeicherte Funktion): hyp(x,N, m, n)
Bezogen auf Kugelziehen aus Urne: N = # Anz
m = #markierte n = gezogen || x = markiert von gezogenen

$$E(X) = \frac{n * m}{N}$$

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



3. GRUNDFRAGESTELLUNGEN DER STATISTIK

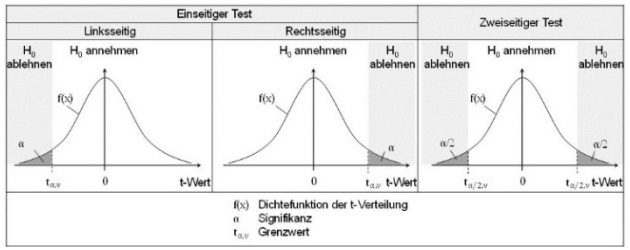
3.1 WELCHES IST DER ZU DEN BEOBACHTUNGEN PLAUSIBELSTE PARAMETERWERT π ?

- Antwort: Punkt-Schätzung → Schätzwert mit Dach: $\hat{\pi}$
- Momenten-Methode (MM): Beobachtung = E(X)
- Maximum-Likelihood-Methode (MLE):
  - Welches ist die höchste Wa. basierend auf Messdaten?
  $\frac{dP(X=x)}{d\pi} = \frac{d}{\pi}\left(\binom{n}{x}\pi^x * (1 - \pi)^{n-x}\right)$
  - Nullstellen der Ableitung suchen → Extrema
  - Für eine Beobachtung: $\hat{\pi} = \frac{x}{n}$
  - π soll so gewählt werden, dass am besten zur Beobachtung passt
  Bsp. Münze wird 3-mal geworfen Ergebnis: ZKK, Würfe sind unabhängig voneinander, p = Wa. für K, MLE(p) = 0.33 -> 1/3

3.2 SIND DIE BEOBACHTUNGEN KOMPATIBEL (STATISTISCH VEREINBAR) MIT VORGEGEBENEM PARAMETER π?

- Antwort: statistischer Test

1) Modell → passende Wa.-Verteilung und X festlegen
2) H_0 (Nullhypothese) & H_A (Alternativhypothese) aufstellen
   $H_0: \pi = \pi_0$ kein Effekt / Veränderung
   $H_A: \pi \neq \pi_0$ zweiseitig
   $H_A: \pi > \pi_0$ einseitig nach oben
   $H_A: \pi < \pi_0$ einseitig nach unten
3) Teststatistik (T) = # Erfolge bei n Versuchen
   - unter Annahme, dass H_0 stimmt
   - T ist eine Funktion von X
4) Signifikanzniveau α festlegen (meist: 0.05 || 0.01)
= Stichprobendaten, welche von der Annahme abweichen, dass Annahme verworfen wird
- wenn α grösser wird vergrössert sich Verwerfungsbereich d.h. Macht nimmt tendenziell zu
5) Verwerfungsbereich K = enthält alle Ereignisse, welche «abnormal» genug sind
→ bei Binomialtest mit binomcdf
- Gesucht: kleinste Zahl c, für die gilt: $P(X \geq c) \leq \alpha$



$H_A: \pi \neq \pi_0$	$K = [0, c_u] \cup [c_o, n]$	zweiseitig → $\frac{\alpha}{2}$
$H_A: \pi > \pi_0$	$K = [c, n]$	einseitig
$H_A: \pi < \pi_0$	$K = [0, c]$	einseitig

- Einseitiger Test = nur Abweichungen in eine Richtung von H_0 detektiert → nicht grosse Abweichung nötig, damit detektiert
  - Macht gross
  - unsicher, dass die Abweichung, in die eine Richtung von der H_0 relevant ist→ einseitiger Test bevorzugt
- Zweiseitiger Test = Abweichungen in beide Richtungen von H_0
  - müssen aber gross sein, damit detektiert → Macht ist klein
  - zweiseitiger Verwerfungsbereich mit ½ Signifikanzniveau α
  → Immer eine untere und eine obere Grenze vorhanden
  - Spezifität nicht sehr hoch → eignet sich, wenn man grosse/kleine Gewinn-Wa. erkennen will

- Normalapproximation (nur für Überschlagsrechnungen)
  - Falls n gross und π ≈ 0.5
  - gut falls $n * \pi > 5$ und $n * (1 - \pi) > 5$
  - je grösser # Versuche → bessere Normalapproximation

$\pi \neq \pi_0$ K = [0, c_u] ∪ [c_o, n] $c = n\pi_o \pm 1.96\sqrt{n\pi_o(1-\pi_o)}$

TR (gesp. Fkt.): napp_1o(n, π_o); napp_1u $\frac{\alpha}{2}$ pro Seite

→ o = aufrunden, u = abrunden

$\pi > \pi_0$ K = [c, n] $c = n\pi_o + 1.64\sqrt{n\pi_o(1-\pi_o)}$

TR (gesp. Fkt.): napp_2(n, π_o); aufrunden

$\pi < \pi_0$ K = [0, c] $c = n\pi_o - 1.64\sqrt{n\pi_o(1-\pi_o)}$

TR (gesp. Fkt.): napp_3(n, π_o); abrunden

6) Testentscheid: Beobachteter Wert im Verwerfungsbereich?
- Ja = H_0 verwerfen
- Nein = H_0 nicht verworfen aber auch nicht bewiesen

- Fehlertypen
  1. Art = α-Fehler = H_0 verwerfen, obwohl H_0 korrekt
  2. Art = β-Fehler = H_0 beibehalten, obwohl H_A korrekt
  -α beschränkt Mögl. für Fehler 2. Art → grösser falls α kleiner
  - wenn Fehler 1. Art kleiner wird, wird Fehler 2. Art grösser
  - wenn Fehler 2. Art grösser wird, wird Fehler 1. Art kleiner

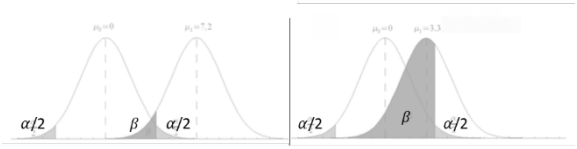
α = 0.05	H_0 wahr	H_0 falsch
H_0 keep	Spezifität → 1 - α	Fehler 2. Art → β
H_0 discard	Fehler 1. Art → ≤ α	Sensitivität / Macht → 1 - β

- Macht = Wa. H_A zu entdecken, falls H_A richtig ist

$1 - P(Fehler\ 2.Art)$

  - Je grösser Fehler 1. Art (α), desto grösser die Macht
  - wird bei zweiseitigem Test meist kleiner (ausser einseitiger detektiert auf falsche Seite → Macht = 0)

$H_0 \uparrow | H_A \downarrow | Macht \downarrow$ || $\alpha \downarrow | \beta \uparrow | Macht \downarrow$



Bsp. Berechnung Macht

→ $[0,8] \cup [18,19] | p_A = 0.85 | n = 19$

Behauptung: Macht = 0.199

binomcdf für beide Verwerfungsbereiche (=0.00002 & =0.198)

Macht = Addieren der Werte = 0.199 → Behauptung richtig

- P-Wert = Wa. unter Gültigkeit von H_0 das beobachtete Ergebnis oder ein extremeres zu erhalten
  = das kleinste α, bei dem H_0 gerade noch verworfen wird
  - Verwerfe H_0 falls P-Wert ≤ α
  - Belasse H_0 falls P-Wert > α
  - je kleiner der P-Wert, desto signifikanter Berechnung → Ergebnis spricht gegen H_0
  - wenn x grösser wird, dann wird p-Wert kleiner (Rest konst.)

Bsp. Einseitiger Binomialtest (p-Wert bestimmen)

→ $\pi = 0.55 | H_A: \pi > \pi_0 | n = 10 | x = 3 |$

Behauptung: p=0.204

- zuerst binompdf (für genau 3 Treffer) = 0.0746 = Wa. dafür
- p-Wert → extremere Ereignisse, d.h. $x \geq 3$ →

binomcdf = 0.97 → d.h. Behauptung falsch

3.3 WELCHE PARAMETERWERTE π SIND MIT DEN BEOBACHTUNGEN KOMPATIBEL?

- Antwort: Konfidenzintervall / Vertrauensintervall (VI) = Bereich, der mit einer gewissen Wa. den Parameter eine Verteilung einer Zufallsvariablen einschliesst
- Schliesst wahren Parameter mit $P(1-\alpha)$ ein
- Werte von π_0 bei denen H_0 nicht verworfen wird
- Ein (1-α)-Vertrauensintervall enthält den wahren Parameter mit Wahrscheinlichkeit 1-α
- VI = {π_0 ; Nullhypothese (H_0: π = π_0) wird belassen}

Bei α = 0.05: $95\%\text{-}VI = \frac{x}{n} \pm 1.96 * \sqrt{\frac{x}{n} * \left(1 - \frac{x}{n}\right) * \frac{1}{n}}$

TR (gesp. Fkt.): na_95o(n, x) oder na_95u(n, x)

- zweiseitiges 95%-VI umfasst bei einer Binomvert. alle Werte p_0, für die ein zweiseitiger Binomialtest mit H_0: p=p_0 nicht verwirft → mit 95%-Wa. ist korrekt

- Normalapproximation: für 95%-VI (falls n gross) → Faustregel
  - Parameter in VI, wenn nein → verwerfen
  - $\sqrt{n}$ – Gesetz = mit n-mal so vielen Beobachtungen wird das VI um den Faktor $\sqrt{n}$ kleiner (weniger breit)
  → für eine halb so grosse Standardabweichung (Streuung) braucht man 4-mal so viele Daten Bsp. $n = 4, d.h. \frac{VI}{\sqrt{4}}$
  - doppelte Genauigkeit bei viermal so vielen Daten

3.4 VERGLEICHE

- Hypothesentest (gut für internationale Feststellungen)
  ⊕ klares Prozedere, klare Aussage über Fehler 1. & 2.Art
  ⊖ wie deutlich verworfen? wie gross wahrer Parameter?
- P-Wert
  ⊕ klar ob und wie deutlich verworfen wird
  ⊖ keine Aussage Fehler 1&2 , wie gross wahrer Parameter?
- 95%-VI (Unsicherheit vermindern → beste Art)
  ⊕ klar ob und wie deutlich verworfen, klar wie gross der Parameter etwa ist
  ⊖ keine klare Aussage über Fehler 1. & 2. Art

Bsp. Einseitiger Binomialtest (Verwerfungsbereich bestimmen)

→ $H_0: \pi = 0.6 | H_A: \pi > 0.6 | n = 26\ \alpha = 0.05$

Behauptung: K[21,26] → K = [c, n] (aus Theorie) , weil H_A: π > π_0

binomcdf(26, 0.6, 21, 26) = 0.021 < 0.05 → d.h. K richtig gewählt

3. Zahl → untere Grenze (hier c) , 4. Zahl → obere Grenze (hier n)

4. DESKRIPTIVE STATISTIK

4.1 KENNZAHLEN

- Kennzahlen charakterisieren einen Datensatz
- Kennzahlen für Lage
  - Arithmetisches Mittel $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Median $q_{0.5}$
- Kennzahlen für Streuung
  - Streuung → empirische Standardabweichung

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

  - IQR (=Inter-Quartile-Range)

- α-Quantil (empirisch, α = %)
  - q_α = Wert aus Daten → bei dem mindestens $\alpha * 100\%$ der Datenpunkte kleiner und $(1-\alpha) * 100\%$ grösser als q_α sind.
  - Daten müssen der Reihenfolge nach sortiert sein
  - $\alpha * n$ berechnen ($q_{0.25}$ → 1. Quartil, $q_{0.75}$ → 3.Quartil)
    - ganze Zahl: $q_\alpha = \frac{1}{2}(x_{\alpha*n} + x_{\alpha*n+1})$
    - keine ganze Zahl: $k = \alpha * n + \frac{1}{2}$ → Runden zu Zahl $q_\alpha = x_k$ (k-te Zahl)
  - empirische Median = $q_{0.5}$
    - mittlere Beobachtung, robust gegen Ausreisser
    - Zahl aus Stichprobe, wenn $\alpha * n$ keine ganze Zahl
    - wenn ganze Zahl → Mittelwert aus zwei Beobachtungen aus der Stichprobe
  - Quartilsdifferenz (IQR)
    - Streuungsmass für Daten, robust gegen Ausreisser

$IQR = q_{0.75} - q_{0.25}$

- Kovarianz = Mass für gegenseitige Abhängigkeit → misst ob Zahlen in die gleiche Richtung gehen

$Cov(X,Y) = E[(X - \mu x)(Y - \mu y)]$

x, y = stetige Zufallsvariablen, μ = Erwartungswert

- Korrelation = Stärke einer statistischen Beziehung von zwei Variablen $Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$
  - dimensionslose normierte Zahl ∈ $[-1, 1]$

$Corr(X,Y) = 1 \rightarrow Y = a + bX$ b>0

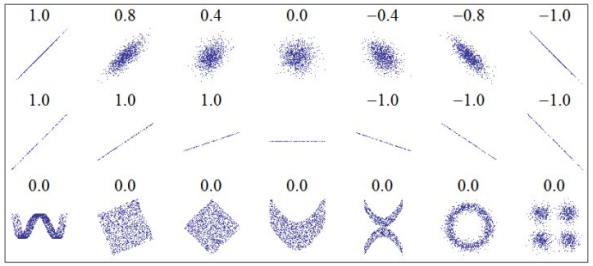
$Corr(X,Y) = -1 \rightarrow Y = a - bX$ b<0

$Corr(X,Y) = 0$ → X und Y unabhängig voneinander

- Empirische Korrelation

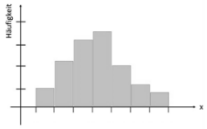
$\hat{\rho} = r = \frac{s_{xy}}{s_x s_y}$ || $s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

- dimensionslose Zahl $\in [-1, 1]$
- Vorzeichen gibt Richtung an, Betrag gibt die Stärke des linearen Zusammenhangs zwischen zwei Variablen an
- Korrelationskoeffizient = Grad des lin. Zusammenhangs

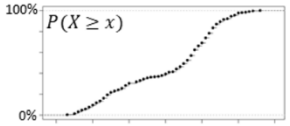


4.2 GRAPHISCHE METHODEN

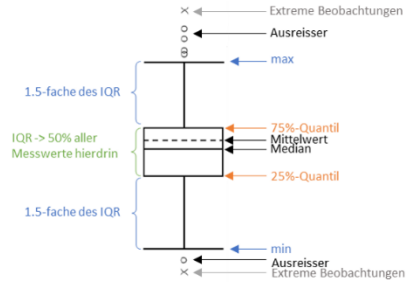
[1] Histogramm



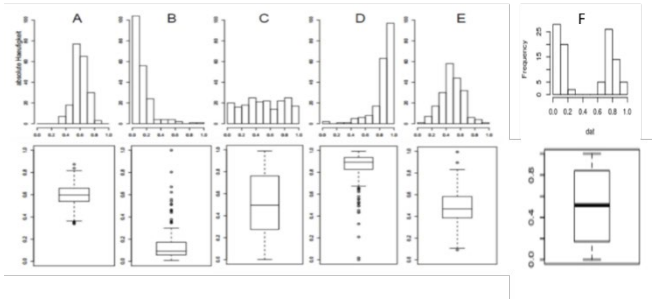
[2] Kumulative Verteilungsfkt.(ECDF)



[3] Boxplot



- Zusammenhang Histogramm – Boxplot



- Hypothese
  - konfirmatorisch: langsam, formell, «bestätigen»
  - explorativ: schnell, informell

4.3 ZUFALLSVARABLEN (X=Zufallsvariable, x =Wert)

- $\sum P(X = x_i) = 1$ → wenn Elemente mehrfach gezählt werden > 1
  P(X=x): Wa., dass ZV genau den Wert x produziert
  P(X≥x): Wa., dass ZV mindestens Wert x oder höher produziert
  P(X≤x): Wa., dass ZV mindestens Wert x oder tiefer produziert
- Funktion einer Zufallsvariable
  - Lineare Transformation $g(x) = a + b * x$ für $Y = a + b * X$

$$E(Y) = a + b * E(X) \mid Var(Y) = b^2 * Var(X)$$

$$\sigma(Y) = |b| * \sigma(X) = \sqrt{b^2 * Var(X)}$$

- mehrere x vorhanden:

$E(Y) = a + b_1 * E(X_1) + b_2 * E(X_2)$

$Var(Y) = (b_1)^2 * Var(X_1) + (b_2)^2 * Var(X_2)$

- Quantil: $q_y^{\alpha} = a * b * q_x^{\alpha}$ falls b>0

- i.i.d = independent, identically, distributet → $X_1, \dots, X_n$ $i.i.d$
  Zufallsvariablen unabhängig und haben dieselbe Verteilung
- Funktion mehrerer Zufallsvariablen $y = g(x_1, x_2, \dots x_n)$
  $E(X_1 + X_2) = E(X_1) + E(X_2)$ → gilt immer
  $Var(X_1 + X_2) = Var(X_1) + Var(X_2)$ wenn X_1, X_2 unabhängig
- Gesetz der Grossen Zahlen (GGZ) = relative beobachtete Häufigkeit nähert sich mit steigendem Stichprobenumfang seinem Erwartungswert → Einpendeln
  - je grösser die Stichprobe, desto genauer die Schätzung
  - Std.abweichung von $\overline{X}_n$ = Standard-fehler des arith. Mittels

$E(\overline{X}_n) = \mu \quad | Var(\overline{X}_n) = \frac{\sigma_x^2}{n} \quad | \sigma(\overline{X}_n) = \frac{\sigma x}{\sqrt{n}} \overline{X}_n \to \mu(n \to \infty)$

- Zentraler Grenzwertsatz (ZGS) = Mittelwert einer beliebigen Verteilung nähert sich mit zunehmendem Stichprobenumfang der Normalverteilung an
  - je grösser Stichprobe, desto näher an Normalverteilung

$(\overline{X}_n) \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right) \quad | \quad S_n = \sum_{i=1}^{n} x_n \quad | \quad S_n \sim N(n * \mu_x, n * \sigma_x^2)$

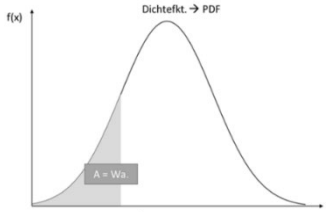
5. STETIGE VERTEILUNGEN

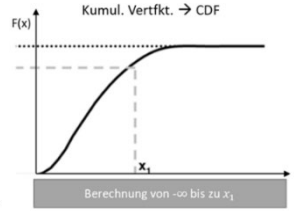
- Zufallsvariable X ist stetig, der Wertebereich kontinuierlich
  Bsp. Grösse messen → unendlich viele Nachkommastellen
- Pkt.-Wa. $P(X = x) = 0$ | $P(X \in (a, b]) = P(a < x \leq b)$
- Kumulative Verteilungsfunktion (CDF) = Wa. dass das Ergebnis X kleiner oder gleich dem Wert x
  $F(x) = P(X \leq x)$ weil $P(a < x \leq b) = F(b) - F(a)$
  → «kleiner als», Monotoner Anstieg von 0 bis 1

- Wahrscheinlichkeitsdichte $f(x) = F'(x)$ (PDF)
  $f(x) \geq 0$ für alle x (weil F monoton steigend)
  → keine einzelnen Punkte vorhanden, sondern über unendlich viele Werte ein Integral → Wa. der Dichtefunktion
  $P(a < x \leq b) = F(b) - F(a) = \int_b^a f(x)dx$
  $\int_{-\infty}^{\infty} f(x)dx = 1$
- Nullverteilung (null distribution) = Stichprobenverteilung der Teststatistik gemäss H_0
- Quantile q
  $P(x \leq q(\alpha)) = \alpha \mid F(q(\alpha)) = \alpha \mid q(\alpha) = F^{-1}(\alpha)$

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

$$Var(x) = \int_{-\infty}^{\infty} (x - E(x))^2 * f(x)dx = E(X^2) - (E(X))^2 = \sigma^2$$





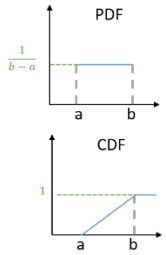
5.1 UNIFORME VERTEILUNG

- Völlige Ignoranz → Dichte konstant auf dem Intervall → auf ganzem Wertebereich gleiche Wa. → $X \sim Unif([a, b])$
- PDF → $P(X = x)$

$$f(x) = \begin{cases} \frac{1}{b-a} für\ a \leq x \leq b \\ 0\ falls\ nicht \end{cases}$$

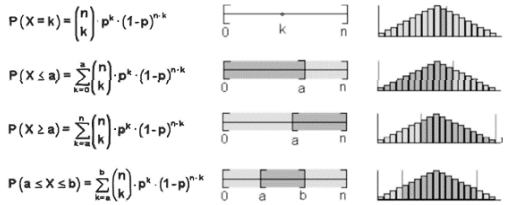
- CDF → $P(X \geq x) = 1 - P(X \leq x)$

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



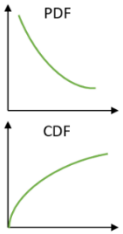
$$E(X) = \frac{a+b}{2} \qquad Var(x) = \frac{(b-a)^2}{12} = \sigma^2$$

$$Median\ m(x) \to 0.5 = \frac{m-a}{b-a}$$



5.2 EXPONENTIALVERTEILUNG

- Zeitdauer für seltene Ereignisse → Warten ohne Gedächtnis
  Wa. verändert sich nie (Bsp. Wartezeiten auf Ausfälle)
  $X \sim Exp(\lambda) \qquad P(X = x) = \frac{\lambda^x}{x!} * e^{-\lambda}$
- Pois(x) hat eine bestimmte Anzahl an Events, bei Exp(x) ist es die Dauer zwischen zwei Events
  - wenn die Zeiten zwischen den Ausfällen eines Systems exponential-verteilt sind, dann ist die # Ausfälle im Intervall der Länge t Poisson-verteilt
- PDF
  $f(x) = \begin{cases} \lambda * e^{-\lambda * x} \; falls \; x \geq 0 \\ 0 \; falls \; x < 0 \end{cases}$
- CDF
  $F(x) = \begin{cases} 1 - e^{-\lambda * x} \; falls \; x \geq 0 \\ 0 \; falls \; x < 0 \end{cases}$



$E(x) = \frac{1}{\lambda} \qquad Var(x) = \frac{1}{\lambda^2} \qquad m(x) \rightarrow \; 0.5 * 1 - e^{-\lambda * m}$

5.3 NORMALVERTEILUNG / GAUSSVERTEILUNG

- Zufällige Messfehler, Summe von unabhängigen, gleichverteilten Zufallsvariablen $X \sim N(\mu, \sigma^2)$
- PDF
  $f(x) = \frac{1}{\sigma * \sqrt{2\pi}} * exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- CDF: keine konkrete Angabe → TABELLE

$E(x) = \mu$	$Var(x) = \sigma^2$
Kurvenhöhepunkt	Breite

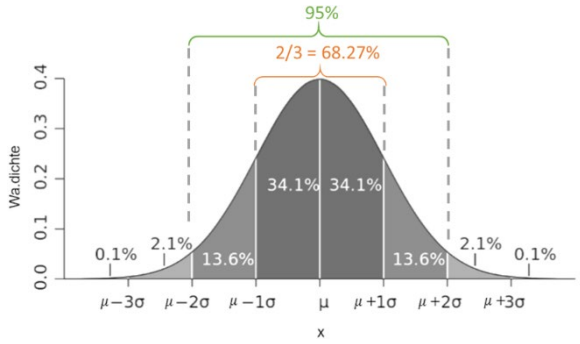
5.4 STANDARDNORMALVERTEILUNG

- Standardnormalverteilung $Z \sim N(0,1)$
- PDF → $\varphi(x) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{x^2}{2}\right)$
- CDF → $\Phi(x) = \int_{-\infty}^{x} \varphi(y) dy$ (Symmetrieachse = x-Achse)
- Standardisierung = Transformation einer Zufallsvariablen, so dass $E(x) = 0 \mid Var(x) = 1 | \sigma(x) = 1$ gilt → Erreichung durch linear transformierte Variablen
  $g(x) = \frac{x - E(X)}{\sigma_X} = -\frac{E(X)}{\sigma_X} + \frac{1}{\sigma_X} x = a + bx$
  $Z = g(X) = \frac{X - E(X)}{\sigma_X}$
- Standardisieren: $P(X \leq x) = P\left(z \leq \frac{x-\mu}{\sigma}\right) = \Phi$

1.) Gegeben: Stelle(x) → cdf gesucht, nur für Wa. kleiner als
  $P(X \geq \mu)$ → Ausgangslage
  $P(X \leq -x) = 1 - P(X \leq x) = P(X \geq x)$
  $z = \frac{x-\mu}{\sigma}$ → in TABELLE nachschauen
2.) Gegeben: Wa. P → Stelle(x) gesucht, Wa. in TABELLE nachschauen = z → $x = z * \sigma + \mu$

Bsp. Gegeben: $X \sim N(\mu = -85, \sigma^2 = 0.09)$
Behauptung: $P(X \leq -85.608) = 0.0091$
$\sigma = \sqrt{0.09} = 0.3 \mid z = \frac{x-\mu}{\sigma} = \frac{(-85.608)-(-85)}{0.3} = -2.86$ → Tabelle Wert: 0.9909, weil $P(X \geq \mu)$ sein soll → $1 - 0.9909 = 0.0091$

- Prozentzahlen: wie viele Messwerte darin enthalten
- Normalverteilung erkennbar, wenn QQ-Plot = Gerade ergibt



5.5 LOGNORMAL-VERTEILUNG

- Für positive Zufallsvariablen verwendet → wenn Transformation auf eine Normalverteilung führt = X-Lognormal-verteilt
  $Y \sim N(\mu, \sigma^2), X = \exp(Y) \rightarrow \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^2$
- Lognormal-Verteilung nicht mehr symmetrisch
  $E(X) = \exp\left(\mu, \frac{\sigma^2}{2}\right) > \exp(E(Y))$

6. STATISTIK FÜR STICHPROBEN

~~Vorzeichen-Test, Wilcoxon-Test, Mann-Whitney Test~~

- (Punkt-)Schätzungen
  $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \mid \hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$

- T-Test besser als Z-Test

Test	Annahmen			
	σ_x bekannt	$x_i \sim N$	Symmetrisch	i.i.d.
z	X	X	X	X
t		X	X	X

6.1 Z-TEST

1.) Basierend auf mehreren Beobachtungen $Z \sim N(\mu, \sigma_x^2)$
  σ_X = bekannt | $\sigma_{\overline{X_n}}$ = berechnen | X_i = kontinuierlich
2.) Hypothese betrachten
  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
  $Z = \frac{\overline{x}_n - \mu_0}{\sigma_{\overline{x}_n}} = \frac{\sqrt{n} * (\overline{x}_n - \mu_0)}{\sigma_x} = \frac{beobachtet - erwartet}{Standardfehler} \; \mu \rightarrow meist \; 0$
4.) Signifikanzniveau & Verwerfungsbereich → TABELLE
  $H_A: \mu_A \neq \mu_0 \qquad K = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}, \infty\right)$
  $H_A: \mu_A < \mu_0 \qquad K = (-\infty, -z_{1-\alpha}]$
  $H_A: \mu_A > \mu_0 \qquad K = [z_{1-\alpha}, \infty)$
  Bsp. $Z_{0.985}$, 0.985 in Tabelle suchen → Spalte und Zeile = Grenze
5.) Testentscheid & Vertrauensintervall
  Beobachteter Wert (in Z einsetzen) im Verwerfungsbereich?
  Ja → H_0 verwerfen | Nein → H_0 nicht verworfen
  $H_A: \mu_A \neq \mu_0 \qquad VI = \left[\overline{X}_n \pm z_{1-\frac{\alpha}{2}} * \frac{\sigma_x}{\sqrt{n}}\right]$
  $H_A: \mu_A < \mu_0 \qquad VI = \left(-\infty, \overline{X}_n - z_{1-\alpha} * \frac{\sigma_x}{\sqrt{n}}\right]$
  $H_A: \mu_A > \mu_0 \qquad VI = \left[\overline{X}_n - z_{1-\alpha} * \frac{\sigma_x}{\sqrt{n}}, \infty\right)$

6.2 T-TEST

1.) $T \sim N(\mu, \sigma_x^2)$
σ_x= unbekannt, $\sigma_{\hat{x}}$ = geschätzt, X_i = kontinuierlich

$$E(T) = 0 \mid Var(T) = \frac{n}{n-2} \mid \hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

2.) Hypothese betrachten
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 $T \sim t_{n-1}$ → Freiheitsgrade

$T = \frac{\overline{X}_n - \mu_0}{\hat{\sigma}_{\overline{x}_n}} = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\hat{\sigma}_x} = \frac{beobachtet - erwartet}{Standardfehler} \mid \sigma_{\overline{X}_n} = \frac{\hat{\sigma}_x}{\sqrt{n}}$

μ_0 → meist 0

P-Wert (bei H_A: $\mu \neq \mu_0$) $P(|T| > |t|) = 2 * (1 - F_{t_{n-1}}(\frac{\sqrt{n} * |\overline{x}_n - \mu_0|}{\hat{\sigma}_x}))$

4.) Signifikanzniveau & Verwerfungsbereich → TABELLE
Quantile in Tabelle: $T \sim t_{13}, P(T \leq t) = 0.95$ → Z13, Sp 0.95
$T \sim t_{13}, P(X \leq 2.659)$ → Z13, 2.659 suchen & Spalte ablesen

$H_A: \mu_A \neq \mu_0 \quad K = \left(-\infty, -t_{n-1;\,1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1;\,1-\frac{\alpha}{2}}, \infty\right)$

$H_A: \mu_A < \mu_0 \quad K = \left(-\infty;\ -t_{n-1;\,1-\frac{\alpha}{2}}\right]$

$H_A: \mu_A > \mu_0 \quad K = \left[t_{n-1;\,1-\frac{\alpha}{2}};\ \infty\right)$

- Je kleiner degrees of freedom (df) bei der Verteilung t_{df}, desto wahrscheinlicher sind Werte mit grossem Absolutbetrag

5.) Testentscheid & Vertrauensintervall
- Beobachteter Wert (in T einsetzen) im Verwerfungsbereich?
  Ja → H_0 verwerfen
  Nein → H_0 nicht verwerfen: falls $\left|\frac{\sqrt{n} * |\overline{x}_n - \mu_0|}{\hat{\sigma}_x}\right| < t_{n-1;\,1-\frac{\alpha}{2}}$

$H_A: \mu_A \neq \mu_0 \quad VI = \left[\overline{X}_n \pm t_{n-1;\,1-\frac{\alpha}{2}} * \frac{\hat{\sigma}_x}{\sqrt{n}}\right)$

$H_A: \mu_A < \mu_0 \quad VI = \left(-\infty, \overline{X}_n + t_{n-1;\,1-\frac{\alpha}{2}} * \frac{\hat{\sigma}_x}{\sqrt{n}}\right]$

$H_A: \mu_A > \mu_0 \quad VI = \left[\overline{X}_n + t_{n-1;\,1-\frac{\alpha}{2}} * \frac{\hat{\sigma}_x}{\sqrt{n}}, \infty\right)$

- Faustregel für minimale Stichprobengrösse n

$n \geq 4 * \frac{\hat{\sigma}_x^2}{\delta^2} \mid \delta = \frac{1}{2}$ von der totalen Breite des VI

6.3 ZWEI STICHPROBEN: GEPAART / UNGEPAART

- Gepaarte Stichprobe:
  Direkte Zuordnung → Stichprobengrösse dieselbe
  Bsp. Zwillinge, Messung vorher/nachher
  1.) Differenz bilden $u_i = x_i - y_i \rightarrow (i = 1, \dots, n)$ x_i, y_i sind abhängig
  2.) Teststatistik: $T_u = \frac{\sqrt{n} * (u_i - 0)}{\hat{\sigma}_u}$
  Falls Daten normalverteilt → T-Test

- Ungepaarte Stichprobe
  Paarung nicht möglich
  x_i, y_i unabhängig
  $m \neq n$ → oft, nicht immer

Gepaart	Ungepaart
Gleichgrosse Stichproben	Unterschiedlich
Klare Zuordnung Differenz der Paare	Keine Zuordnung
Mehr Macht	Weniger Macht

Bsp. Messungen mit Methode A oder B, aber nicht mit beiden gleichzeitig → kein eindeutiger Zusammenhang

6.4 ZWEI STICHPROBEN T-TEST MIT GLEICHEN VARIANZEN

- Annahme: $\sigma_X^2 = \sigma_Y^2$ → gleiche Varianz
  Arithmetisches Mittel: $\overline{X}_n = \frac{1}{n}\sum X_i \mid \overline{Y}_m = \frac{1}{m}\sum Y_i$
  Empirische Varianz: $\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

1.) $x_1, \dots x_n \ i.i.d. \sim N(\mu_x, \sigma^2) \mid y_1, \dots y_m \ i.i.d. \sim N(\mu_Y, \sigma^2)$

2.) Hypothese betrachten
H_0: $\mu_x = \mu_y$ | H_A = $\mu_x \neq \mu_y, \mu_x > \mu_y, \mu_x < \mu_y$

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

$S_{Pool}^2 = \frac{1}{n+m-2}\left((n-1)\hat{\sigma}_x^2 + (m-1)\hat{\sigma}_y^2\right) = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2\right)$

$T = \frac{(\overline{X}_n - \overline{Y}_m) - (\mu_x - \mu_y)}{S_{pool} * \sqrt{\frac{1}{n} + \frac{1}{m}}}$ μ_x, μ_y →meist 0, $\sqrt{S_{pool}{}^2}$!

4.) Signifikanzniveau & Verwerfungsbereich → TABELLE

$H_A: \mu_x \neq \mu_y \quad K = \left(-\infty; -t_{n+m-2;1-\frac{\alpha}{2}}\right] \cup \left[t_{n+m-2;1-\frac{\alpha}{2}}; \infty\right)$

$H_A: \mu_x < \mu_y \quad K = \left(-\infty; -t_{n+m-2;1-\alpha}\right]$

$H_A: \mu_x > \mu_y \quad K = \left[t_{n+m-2;1-\alpha}; \infty\right)$

5.) Testentscheid: Beobachteter Wert in Verwertungsbereich?
Ja → H_0 verwerfen | Nein → H_0 nicht verworfen

- $(\mu_x - \mu_y)$ ist häufig unter H_0 gleich 0 → wird weggelassen
- $S_{pool}^2 = \hat{\sigma}^2$
- Bei gepaarten Stichproben auch ungepaarter Test durchführbar → umgekehrt nicht möglich

6.5 WELCH-TEST; ZWEI STICHPR. T-TEST UNGLEICHE VARIANZEN

- Varianzen in beiden Gruppen müssen nicht gleich sein → aber kleinere Macht → an Prüfung gleiche Varianzen annehmen

6.6 MULTIPLES TESTEN – BONFERRONI-KORREKTUR

- Liste von wirklich "wichtigen" Resultaten generieren
- Anpassung der Signifikanzniveaus für mehrere zusammenhängende Tests

$\alpha_{adj} = \frac{\alpha}{j} \mid korr.\,\alpha - Niveau\ für\ jeden\ Test = \frac{gewünschtes\ \alpha}{\#\ Test}$

- Wenn mehrere Tests durchgeführt werden bei verschiedenen Gruppen gibt es eine Alpha-Fehler-Kumulierung
  Alpha-Fehler-Risiko: $1 - (1 - \alpha)^j$
  Bsp. 4 Gruppen Mittelwerte vergleichen
  alle $\alpha = 5\%$ → $1 - (1 - 0.05)^6 = 0.265$
  (6 Tests, weil alle miteinander vergleichen)
  d.h auch wenn kein signifikanter Unterschied im Mittelwert würde zu 26.5% ein Test fälschlicherweise signifikant werden
  Daher Bonferroni-Korrektur anwenden, damit das Gesamt α für alle Tests zusammen bei max. 5% liegt
  $\alpha_{adj} = \frac{0.05}{6} = 0.0083$ → für jeden Test $\alpha = 0.0083$ einsetzen
- Nachteil: korrigiert α-Niveau stark nach unten → Risiko für Fehler 2. Art steigt

6.7 HYPOTHESENTEST VERGLEICHE

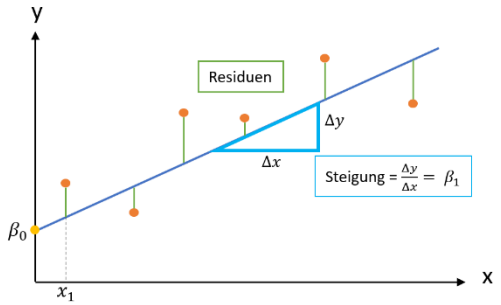
Test	Was wird getestet	Bsp.
Binomialtest	Anteil in Gruppe gleich π_0?	Wirksamkeit von Medikament
t-Test, eine Stichprobe	Erwartungswert in Gruppe gleich μ_0?	Füllmenge in Getränkeflaschen
t-Test, 2 gepaarte Stichproben	Erwartungswert in beiden Gruppen gleich?	Reaktionszeit von Haupt- und Nebenhand
t-Test, 2 ungepaarte Stichproben	Erwartungswert in beiden Gruppen gleich?	Aktivität von Gen XY bei Gesunden und Kranken

7. REGRESSION

~~Multiple Lineare Regression~~

- Voraussagen basierend auf Messdaten machen / gerade fitten
- $Y = \beta_0 + \beta_1 * x + \varepsilon \mid \varepsilon \sim N(0, \sigma^2)$
- ε = zufälliger Fehler / Fehler-Variable / Rausch-Terme
  Zusammenhang zwischen der erklärenden und der Ziel-Variablen nicht exakt
- X_i = erklärende Variable, deterministisch, wurde bestimmt
- Y_i = Zielvariable, Zufallsvariable von ε abhängig
- Residuen = Vertikaler Abstand zwischen Punkt und Geraden
  $R_i = Y_i - (\beta_0 + \beta_1 * x_i)\ (i = 1, \dots, n)$
- Einfache lineare Regression: nur eine x-Variable vorkommend
- Linear = wenn Parameter β in der Ableitung verschwindet

$$\varepsilon_i \sim N(0, \sigma^2) \quad | \quad E(\varepsilon_i) = 0 \quad | \quad Var(\varepsilon_i) = \sigma^2$$



7.1 PARAMETERSCHÄTZUNGEN
METHODE DER KLEINSTEN QUADRATE

- Schätzungen haben keinen systematischen Fehler, sondern streuen nur um den wahren Parameter herum
- Methode der kleinsten Quadrate und Maximum Likelihood Methode sind äquivalent
- β_0 und β_1 sind Minimierer von $\sum_{i=1}^{n}(Y_i-(\beta_0+\beta_1*x_i))^2$
- Approximation $R_i \approx \varepsilon_i$
- Parameterschätzung
  $\widehat{\beta_1}=\frac{\sum_{i=1}^{n}(Y_i-\bar{Y}_n)(x_i-\bar{x}_n)}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}$ | $\widehat{\beta_0}=\bar{y}_n-\widehat{\beta_1}*\bar{x}_n$
- Varianzschätzung $\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}R_i^2$
- Standardfehler $\widehat{s.e.}(\widehat{\beta_1})=\sqrt{\widehat{Var}(\widehat{\beta_1})}=\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}}$
- Teststatistik $T=\frac{\widehat{\beta_1}-0}{\widehat{s.e.}(\widehat{\beta_1})}=\frac{beobachtet-erwartet}{geschätzter\ Standardfehler}$
- Verwerfungsbereich $K=\left(-\infty,-t_{n-2;1-\frac{\alpha}{2}}\right]\cup\left[t_{n-2;1-\frac{\alpha}{2}},\infty\right)$
- Vertrauensintervall
  $VI=\left[\widehat{\beta_1}-\widehat{s.e.}(\widehat{\beta_1})t_{n-2;\,1-\frac{\alpha}{2}};\widehat{\beta_1}+\widehat{s.e.}(\widehat{\beta_1})t_{n-2;1-\frac{\alpha}{2}}\right]$
- Faustregel 95%-VI $\left[\widehat{\beta_1}-2*\widehat{s.e.}(\widehat{\beta_1});\ \widehat{\beta_1}+2*\widehat{s.e.}(\widehat{\beta_1})\right]$
- Bestimmtheitsmass = wie gut stimmt das Regressionsmodell mit den tatsächlichen Beobachtungen überein
  $R^2=\frac{SS_R}{SS_Y}=\hat{\rho}^2_{Y\hat{Y}}$ → je näher R^2 an 1 desto besser Modell

1.) Regressionsgerade anpassen → β_0 und β_1 berechnen
2.) T-Test → hat x einen Einfluss auf Y? → H_0: $\beta_1=0$
wenn nicht signifikant → x hat keinen Einfluss auf Y
3.) T-Test → Regression durch Nullpunkt? → H_0: $\beta_0\neq 0$
wenn nicht signifikant → kleineres Modell mit Regression durch Nullpunkt benutzen
4.) Vertrauensintervall β_0 und β_1 & Bestimmtheitsmass R^2
5.) Modell-Voraussetzungen mit Residuenanalyse überprüfen

7.2 GENERALISED LINEAR MODELS (GLMS)

- Zusammenhang zwischen erklärenden Variablen und Parametern einer Verteilung feststellen
- Störgrösse bei klassischen linearen Modellen normalverteilt bei GLMs nicht notwendig
- Binomialregression / logistische Regression $Y\sim Bin\left(n,p(x)\right)$
  $P(x)=\frac{e^{\beta_0+\beta_1*x}}{1+e^{\beta_0+\beta_1*x}}$ | $\log\left(\frac{p(x)}{1-p(x)}\right)=\beta_0+\beta_1*x$
  Bsp. Bei welcher Dosis ist Genesungswahrscheinlichkeit 80%?
- Poissonregression $Y\sim Pois\left(\lambda(x)\right)$
  $\lambda(x)=e^{\beta_0+\beta_1*x}$ | $log\left(\lambda(x)\right)=\beta_0+\beta_1*x$
  Bsp. Morgen -5°C, 95%-Quantil der Unfälle morgen ?

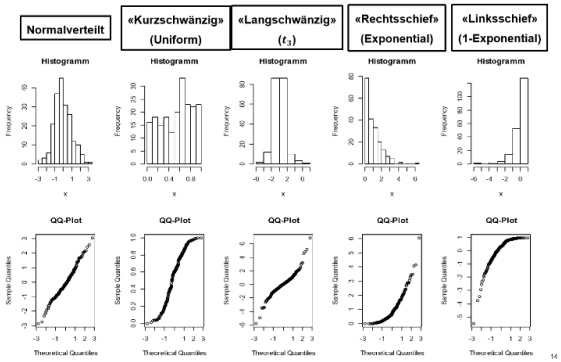
- Lineare Regression $Y\sim N(\mu(x),\sigma^2)$
  Einfache Regression: $\mu(x)=\beta_0+\beta_1x+\beta_2x$
  Multiple Regression: $\mu(x)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots$

8. QUALITÄTSTEST - RESIDUENANALYSE

- Für die Überprüfung der Modell-Voraussetzungen für die einfache lineare Regression
- Residuen messen den vertikalen Abstand zwischen beobachtetem Punkt und gefitteter Gerade
- $\mathcal{E}(E_i)=0\rightarrow\ \mathcal{E}(Y_i)=\beta_0+\beta_1x_1$ → kein systematischer Fehler im Modell → Überprüfung im Tukey-Abscombe Plot
- $E_1,\dots,E_n\ i.i.d.\ |\ Corr(E_i,E_j)=0\ (i\neq j)$ → Fehler müssen voneinander unabhängig sein, keine seriellen Korrelationen vorhanden
  → Überprüfen mit QQ-Plot und Tukey-Anscombe (Fehler müssen ebenfalls gleichverteilt sein)
- $E_1,\dots,E_n\ i.i.d.\ N(0,\sigma^2)$ → Fehler sind normalverteilt
  → Überprüfen mit Normalplot
- Übergang zwischen Verletzen und Einhalten des Modells ist fliessend → leichte Verletzungen weniger schlimm und Ergebnisse doch noch brauchbar

8.1 QQ-PLOT

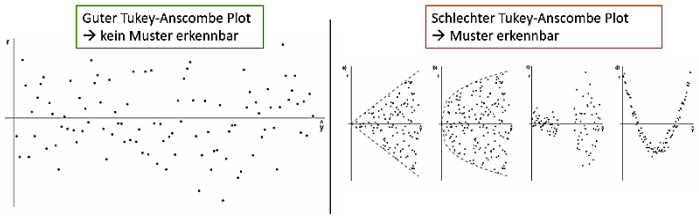
- Histogramm transformieren → Hilfe, für Vergleich ob Modell und Verteilung funktionieren
- Q-Q-Plot (Quantil-Quantil Plot) = empirische Quantile werden gegen die theoretischen Quantile der Modell-Verteilung aufgetragen → wenn Beobachtungen gemäss Modell, dann Punkte auf Winkelhalbierenden von $y=x$
  $\alpha=\frac{0.5}{n},\frac{1.5}{n},\dots,\frac{(n-0.5)}{n}$
- Normal-Plot = Überprüfung ob Normalverteilung vorliegt
  $q(\alpha)=\mu+\sigma\Phi^{-1}(\alpha)$



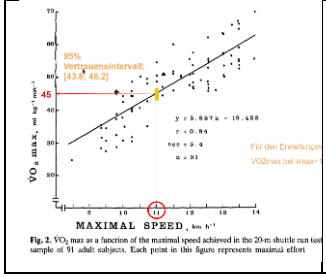
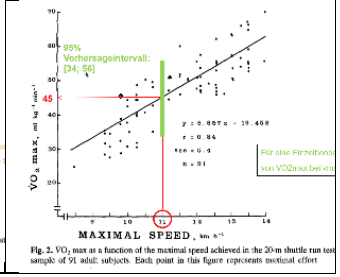
- Punkte sollen auf der Geraden mit Achsenabschnitt μ und Steigung σ liegen = dann gut

8.2 TUKEY ANSCOMBE PLOT

- Plot der Residuen r_i (Y-Achse) gegen die angepassten Werte $\hat{y}_i$ (X-Achse)
- Idealfall: gleichmässige Streuung der Punkte um Null → konstante Breite, uniforme Streuung innerhalb des Bandes Abweichungen: kegelförmiges Anwachsen der Streuung mit $\hat{y}_i$
- Möglicherweise kann man Zielvariable logarithmieren
  $\log(Y_i)=\ \beta_0+\beta_1x_i+\varepsilon_i$



8.3 VERTRAUENSINTERVALL UND VORHERSAGEINTERVALL VON Y

Vertrauensintervall VTI	Vorhersageintervall VHI
plausibler Bereich für E[Y_i] gegeben X_i	Plausibler Bereich für Y_i gegeben X_i
Unsicherheit wegen β's	Unsicherheit β's und ε_i
	

8.4 R-OUTPUT

Modell $Y_i = \beta_0 + \beta_1 x_1 + E_i \,; E_i \sim N(0, \sigma^2)\ i.i.d.$

$Y_i = -19.5 + 5.9 x_1 + E_i \,; E_i \sim N(0, 5.4^2)\ i.i.d$

Residuals:	Min	1Q	Median	3Q	Max
	-10.2	-4.5	-0.2	4.7	12.0

Coefficients:		Estimate	Std. Error	t value	Pr(>|t|)
β_0	Intercept (Achsenabschnitt)	-19.5	4.7	-4.1	8.5e-05
β_1	x	5.9	0.4	14.3	<2e-16

Signif. Codes: 0 0.001 0.01 0.05 0.1 1

Residual standard error: 5.4 on 89 degrees of freedom
Multiple R-squared: 0.6981, Adjusted R-squared: 0.8948
F-statistic: 205.8 on 1 and 89 DF, p-Value <2.2e-16

- erwarteter Achsenabschnitt
- Falls man x um eine Einheit erhöht, sagt das Modell voraus, dass sich y um den Wert 5.9 (Steigung) erhöht
- Wenn es um α geht → Estimate miteinander vergleichen
- Standardfehler von $\widehat{\beta_1}$ $(= \sigma_{\widehat{\beta_1}})$
  Approx. 95%-VI: $5.9 \pm 2 * 0.4$
  Exaktes, 95%-VI: $5.9 \pm 1.99 * 0.4$
  → $t_{89,0.975} = 1.99$ → TABELLE
- Freiheitsgrade: $df = n - 2$
- t Value: Zwischen ± 2, wenn ↑, dann Fläche klein
  → d.h. P-Wert astronomisch klein
- Beobachtete Teststatistik
  Im Test $H_0: \beta_0 = 0\ vs.\ H_1: \beta_1 \neq 0$
  $\frac{5.9}{0.4} = 14.3$
- P-Wert: Angenommen $\beta_1 = 0$, wie Wa. ist es diese Beobachtung oder etwas Extremeres zu erhalten.