Zusammenfassung Statistik 1 - FS17

v5

Gleb Ebert

10. August 2017

Vorwort

Diese Zusammenfassung soll den gesamten Stoff der Vorlesung Statistik 1 (Stand Frühjahrssemester 2017) in kompakter Form enthalten und soll an der Basisprüfung verwendet werden können. Ich kann leider weder Vollständigkeit noch die Abwesenheit von Fehlern garantieren. Für Fragen, Anregungen oder Verbesserungsvorschlägen kann ich unter **glebert@student.ethz.ch** erreicht werden.
Die Instruktionen für die TI-83/84 Taschenrechner befinden sich auf der letzten Seite und können somit einfach weggelassen werden.
An dieser Stelle möchte ich Jonathas Enders für die vielen Verbesserungsvorschläge danken.

1 Hypothesentest

1) Modell
2) Nullhypothese und Alternative
3) Teststatistik
4) Signifikanzniveau α
5) Verwerfungsbereich der Teststatistik K
6) Testentscheid

2 Modelle für Zähldaten

2.1 Wahrscheinlichkeitsmodelle

- Grundraum Ω, Elemantarereignisse w
- Ereignis: Teilmenge von Ω
- Wahrscheinlichkeit für jedes Ereignis

$A \cup B$ (oder); $A \cap B$ (und); $A^c, \overline{A}$ (nicht A)
Zwei Mengen sind disjunkt, wenn sie kein gemeinsames Element besitzen.
Axiome:

1) $P(A) \geq 0$ **2)** $P(\Omega) = 1$
3) $P(A \cup B) = P(A) + P(B) \Longleftarrow A \cap B = \{\} = \emptyset$

2.1.1 Wahrscheinlichkeit berechnen

1) Summe von Elementarereignissen

$$P(A) = \sum_{n=1}^{i} P(w_i) = 1$$

2) Laplace Modell: El.ereignisse gleich wa.

$$P(A) = \frac{\#\,günstiger\,El.ereignisse}{\#\,möglicher\,El.ereignisse}$$

3) Mengenoperationen / Venn-Diagramme:
z.B. Gegenereignis (A und A^C)

2.2 Unabhängigkeit

A, B sind unabhängig, wenn das Auftreten von A die Wa. von B nicht beeinflusst $\Longleftrightarrow P(A \cap B) = P(A) * P(B)$

2.3 Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit von Ereignis A wenn B eingetreten ist, wird mit $P(A|B)$ bezeichnet. Es gilt:

$$P(A^c|B) = 1 - P(A|B)$$

Satz von Bayes: $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) * P(A)}{P(B)}$$

⚠ $P(A|B) \neq P(B|A)$

Satz der totalen Wahrscheinlichkeiten:

$$P(X) = P(X|K) * P(K) + P(X|K^c) * P(K^c)$$

- $odds(E) = \frac{P(E)}{1-P(E)}$
- $log-odds(E) = ln(odds(E))$
- $odds-Ratio = \frac{odds(E|G=1)}{odds(E|G=2)}$
  mit Ereignisgruppen $G = 1$ und $G = 2$

2.4 Zufallsvariable

Funktion $\Omega \curvearrowright \Re\,;\; X : A \longrightarrow X(A) = x$

- Grossbuchstabe: X = Funktion
- Kleinbuchstabe: x = konkreter Wert

$P(X = x) = P(\{w|X(w) = x\})\,;\; \sum_{alle\ x} P(X = x) = 1$

2.5 Binomialverteilung

Allgemein gilt:
n = #Lose; x = #Gewinne; π = #Wa. Gewinn

- Binomialkoeffizient: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$
- Binomialverteilung: $\binom{n}{x} * \pi^x * (1-\pi)^{n-x}$
- Erwartungswert: $E(X) = n\pi$
- Varianz: $Var(X) = n\pi(1-\pi)$

2.6 Kennzahlen einer Verteilung

- Erwartungswert: $E(X) = \sum P(X = x) * x$
  Sind X, Y unabhängig so gilt:
  $E(aX + bY + c) = a * E(X) + b * E(Y) + c$
- Varianz: $Var(X) = E[(X - E(X))^2]$
  $= \sum P(X = x) * [x - E(x)]^2$

$$Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 2ab * Cov(X, Y)$$

- Standardabweichung: $\sigma_X = \sqrt{Var(X)}$

2.7 Diskrete Verteilung

- Binomialverteilung: $X \sim Bin(n, \pi)$ (→ siehe Kapitel 2.5)
- Uniforme Verteilung: alle Ereignisse gleiche Wa.
  - $X \sim Unif(n)$
  - $P(X = x) = \frac{1}{n}$, $\{1, 2, \cdots, n\}$
  - $E(X) = \frac{n+1}{2}$, $Var(X) = \frac{(n+1)(n-1)}{12}$
- Poissonverteilung:
  vergleichsweise seltene Ereignisse während eines bestimmten Zeitraums.
  - $X \sim Pois(x)$
  - $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ $(x = 0, 1, 2, \ldots)$
  - $E(X) = \lambda$, $Var(X) = \lambda$

  Die Summe zwei voneinander unabhängigen und poisson-verteilten Zufallsvariablen ist ebenfalls poisson-verteilt:
  $X \sim Poisson(\lambda_X), Y \sim Poisson(\lambda_Y)$
  $\rightarrow X + Y \sim Pois(\lambda_X + \lambda_Y)$
- Hypergeometrischer Verteilung: Urne, N Kugeln, m markiert, n ziehen ohne zurücklegen, wieviele markierte? Die Chance eine Markierte Kugel zu ziehen verändert sich nach jedem Zug. Bei sehr grossen N ist dies aber vernachlässigbar und die Binomialverteilung ist eine gute Approximation.
  - $X \sim Hyper(N, n, m)$
  - $P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$, $\{0, 1, \ldots, \min(n, m)\}$
  - $E(X) = \frac{nm}{N}$, $Var(X) = \frac{nm(N-m)(N-n)}{N^2(N-1)}$

3 Statistik für Zähldaten

3.1 Drei Grundfragen

- Bester Schätzwert für Parameter
  → Punktschätzung
- Sind Beobachtungen und gewisse Parameterwerte kompatibel? → Hypothesentest
- In welchem Bereich liegt Parameter?
  → Vertrauensintervall (VI)

3.2 Schätzung, Test und VI bei Binomialtest

3.2.1 Punktschätzung

- Momenthenmethode (MM)
- Maximum-Likelihood Methode / M-L Estimate (MLE)

MM, Bsp 1

100 Patienten bekommen neues Medikament. 54 davon werden gesund. Was ist die Wirkwahrscheinlichkeit des Medikaments?

X: gesund gewordene Patienten ($x = 54$)

$X \sim Bin(n = 100,\ \pi = ?)$

Momenthenmethode um π zu schätzen:

$E(X) = n * \pi$

$E(X) \approx x = 54 \rightarrow x \approx N * \pi \rightarrow \pi \approx \frac{x}{n} = 0.54$

MM, Bsp 2: Capture-Recapture

Gesucht: Grösse unbekannter Population. Lincoln-Peterson Methode:

- m zufällige Tiere fangen, markieren, freilassen
- n zufällige Tiere fangen
- ZV X: Anzahl markierter Tiere in 2. Fang

$X \sim Hyper(N, n, m)$ mit N als Populationsgrösse

x Markierte in 2. Fang

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

Ungenau aber richtige Grössenordnung

MLE, Bsp 1

$n = 600$ Personen erhalten Medikament; $x = 30$ haben Nebenwirkung. Wie oft treten diese auf?

X: Anzahl Personen mit Nebenwirkung

$X \sim Bin(n = 600, \pi)$; $P(X = 30) = \binom{600}{30}\pi^{30}(1-\pi)^{570}$

MLE $\hat{\pi}$ für π, ist der Wert, der $P(X = 30)$ maximiert.

$P(X = x) = \binom{n}{x}\pi^{x}(1-\pi)^{n-x} := f(\pi)$ 'likelihood'

$\hat{\pi} = \frac{x}{n} = \frac{30}{600} = 0.05$

3.2.2 Statistischer Test (Binomialtest)

1) Modell: X: Anzahl Erfolge bei n Versuchen
$X \sim Bin(n, \pi)$

2) Nullhypothese $\boldsymbol{H_0}$: $\pi = \pi_0$
Alternative $\boldsymbol{H_A}$:

- $\pi \neq \pi_0$ (zweiseitig)
- $\pi > \pi_0$ (einseitig nach oben)
- $\pi < \pi_0$ (einseitig nach unten)

3) Teststatistik $\boldsymbol{T}$: Anzahl Treffer bei n Versuchen
Verteilung von T falls H_0 stimmt: $T \sim Bin(n, \pi_0)$

4) Signifikanzniveau $\boldsymbol{\alpha}$: Konvention; meist $\alpha = 0.05$

5) Verwerfungsbereich von $\boldsymbol{T}$:
Form des Verwerfungsbereiches

- $\pi \neq \pi_0:\ K = [0, c_u] \cup [c_o, n]$
- $\pi > \pi_0:\ K = [c, n]$
- $\pi < \pi_0:\ K = [0, c]$

c kann dabei mit dem Taschenrechner berechnet werden. Dabei ist folgendes zu beachten:

$P(X \geq x) = 1 - P(X \leq x-1)$ bzw. $P(X > x) = P(X <= x)$

$\Rightarrow$ rechtsseitiger Test (analog für linkss.):

$P(X \geq c) \leq \alpha \Leftrightarrow 1 - P(X \leq c-1) \leq \alpha$

$\Leftrightarrow 1 - \alpha \leq P(X \leq c-1)$

Beim zweiseitigen Verwerfungsbereich benutzt man $\frac{\alpha}{2}$. Alternativ kann auch die Normalapproximation benutzt werden, wenn $n\pi_0 > 0$, $n(1-\pi_0) > 5$ und $\alpha = 0.05$:

$$c \approx n\pi_0 + z\sqrt{n\pi_0(1-\pi_0)}$$

- $\pi > \pi_0:\ z = 1.64$
  c aufgerundet auf die nächste ganze Zahl
- $\pi < \pi_0:\ z = -1.64$
  c abgerundet auf die nächste ganze Zahl
- $\pi \neq \pi_0:\ z = \pm 1.96$
  c analog gerundet

6) Testentscheid

Beispiel Panini-Bilder

1) Ziehen 500 aus 661 Bildern mit Zurücklegen

2) H_0: zufällig eingepackt
H_A: weniger Doppelte eingepackt

3) T: Anzahl einzigartiger Bilder
Verteilung wenn Nullhypothese stimmt:
Simulation

4) $\alpha = 1/1'000'000$

5) Verwerfungsbereich von T: Bei 1 Mio Simulationen nie mehr als 387 einzigartige Bilder
$\rightarrow K = \{388, 389, \ldots, 500\}$

6) Beobachteter Wert (477) liegt in K. H_0 wird auf α daher verworfen.

Fehler:

1. Art: Fälschliches Verwerfen von H_0, obwohl richtig

2. Art: Fälschliches Behalten, obwohl H_A stimmt

Per Definition ist die Wahrscheinlichkeit eines Fehler 1. Art höchstens α. Die Wahrscheinlichkeit eines Fehler 2. Art wird grösser mit kleinerem α. Da man primär Fehler 1. Art vermeiden will, wählt man α klein.

Statt der Wahrscheinlichkeit des Fehlers 2. Art wird oft die Macht angegeben. Macht $= 1 - P(\text{Fehler 2. Art}) = P_{H_A}(X \epsilon K)$ Sie gibt die Wahrscheinlichkeit an H_A zu bestätigen, falls diese richtig ist.

Der einseitige Test erkennt kleinere Abweichungen in eine Richtung von H_0. Seine Macht ist also gross.

Der zweiseitige Test erkennt nur grössere Abweichungen in beide Richtungen von H_0. Seine Macht ist also klein. Man rechnet mit zwei Verwerfungsbereichen an beiden Seiten des Spektrums. Dabei rechnet man jeweils mit dem halben Signifikanzniveau $\frac{\alpha}{2}$ und nimmt anschliessend die Vereinigung der beiden Verwerfungsbereiche.

P-Wert:

Def. 1: Kleinstes α bei dem H_0 gerade noch verworfen wird.

Def. 2: Wa. die Beobachtung oder einen extremeren Fall zu beobachten, falls H_0 wahr ist.

3.2.3 Vertrauensintervall (VI, engl. CI)

Def. 1: Die Werte von π_0 bei denen H_0 nicht verworfen wird auf α, sind $(1-\alpha)$−VI für π
Def. 2: Ein $(1-\alpha)$−VI enthält den wahren Parameter mit Wahrscheinlichkeit $1-\alpha$.

Für $\alpha = 0.05$ (95%-VI) kann die Normalapproximation benutzt werden. Dabei wird $z = 1.64$ für ein einseitiges VI und $z = 1.96$ für das zweiseitige VI eingesetzt:

$$I \approx \frac{x}{n} \pm z\sqrt{\frac{x}{n^2}\left(1-\frac{x}{n}\right)}$$

3.2.4 Vorhersage- und Vertrauensintervall

Das 95%-Vorhersageintervall für ein Ereignis ist in der Regel grösser als das 95%-Vertrauensintervall für das erwartete Ereignis. Ersteres gibt den Bereich für den wahren Wert bei einer Messung an, während letzteres bei vielen Wiederholungen einer Messung mit der Wahrscheinlichkeit $1-\alpha$ angibt, dass der Wert darin liegt.

4 Modelle und Statistik für Messdaten

4.1 Deskriptive Statistik

4.1.1 Kennzahlen

Das α-Quantil ist der Wert q_α, bei dem $\alpha * 100\%$ der Datenpunkte kleiner als q_α sind.
$q_{0.5}$ = “Median”, $q_{0.25}$ = “1. Quartil”, $q_{0.75}$ = “3. Quartil"
Besteht unser Datensatz aus geordneten Werten
$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, können empirische α-Quantile wie folgt berechnet werden:
$\frac{1}{2}\left(x_{(\alpha n)} + x_{(\alpha n+1)}\right)$ wenn $\alpha * n \in \mathbb{Z}$
$x_{(\alpha n+\frac{1}{2})}$ gerundet auf ganze Zahl wenn $\alpha * n \notin \mathbb{Z}$

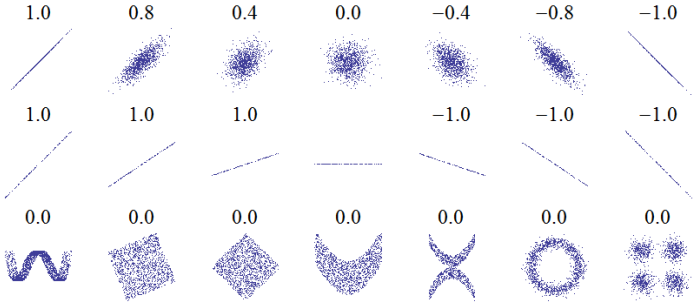
Kennzahlen für die Lage
arithmethische Mittel: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median: $q_{0.5}$ (robust)

Kennzahlen für die Streuung
empirische Standardabweichung:

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

Inter-Quartile Range (IQR):
$IQR = q_{0.75} - q_{0.25}$ (robust)

Kennzahlen für linearen Zusammenhang
$Var(X) = E((X-\mu_x)^2)$ wobei $\mu_x = E(X)$
Kovarianz: $Cov(X,Y) = E[(X-E(X)) * (Y-E(X))]$
$= E(X*Y) - E(X) * E(Y)$ mit $Cov(X,X) = Var(X)$
Korrelation = “skalierte Kovarianz”
Misst Stärke und Richtung von linearer Abhängigkeit. Korrelation $\in [-1,1]$
$\rho_{XY} = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x * \sigma_y}$
$Corr(X,Y) = 1 \longleftrightarrow Y = a + b*X,\ b>0$
$Corr(X,Y) = -1 \longleftrightarrow Y = a + b*X,\ b<0$
X, Y unabhängig $\longrightarrow Corr(X,Y) = 0$



empirische Korrelation:

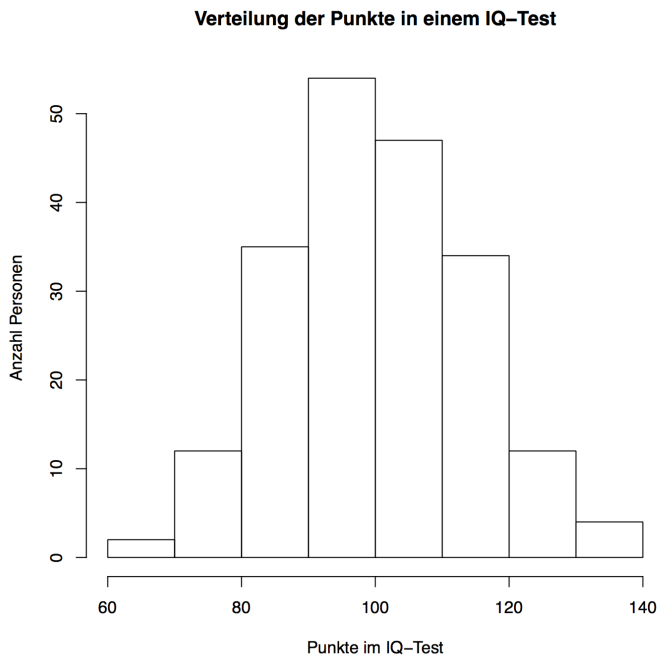
$$r_{XY} = \frac{s_{XY}}{s_X * s_Y},\ s_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x}) * (y_i-\bar{y})}{n-1}$$

Standardisierung: Ein Datensatz kann standardisiert werden, so dass arithmetisches Mittel gleich Null und Standardabweichung gleich 1 sind.

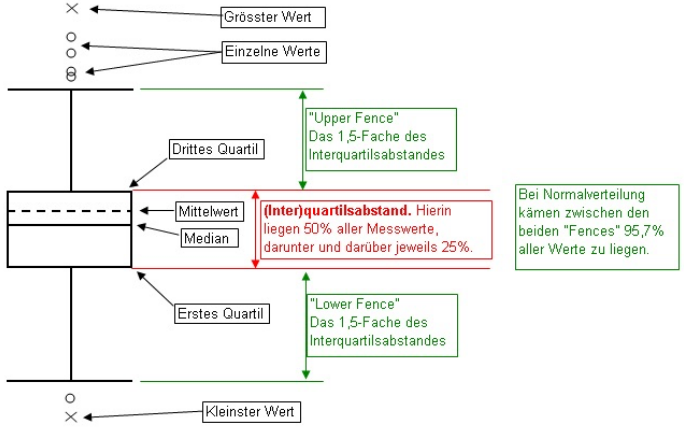
$$z_i = \frac{x_i - \bar{x}}{s_X},\ (i = 1,\ldots,n)$$

4.1.2 Grafische Methoden

Histogramm: Klassen konstanter Breite; Anzahl Beobachtungen pro Klasse; Balken proportional zur Anzahl Beobachtungen in der jeweiligen Klasse

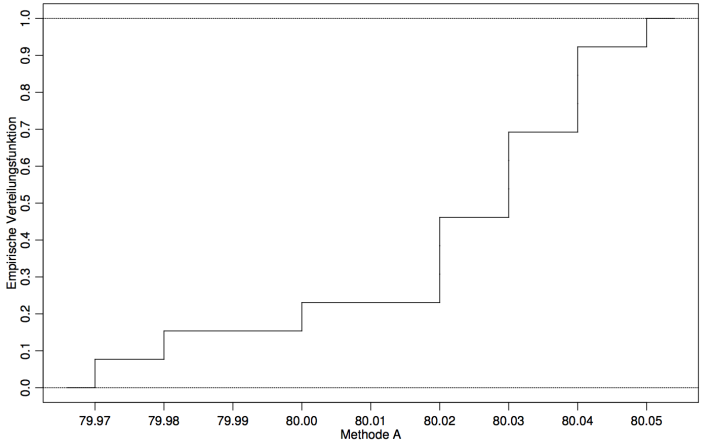


Boxplot: Rechteck, dass von den empirischen 25%- und 75%-Quantilen begrenzt wird; Linien, welche von dem Rechteck bis zum kleinsten bzw grössten Wert reichen, der höchstens 1.5 mal die Quartilsdifferenz von einem der beiden Quartile entfernt ist; Ausreisser sind als Sterne aufgeführt; ein Strich, welcher den Median anzeigt

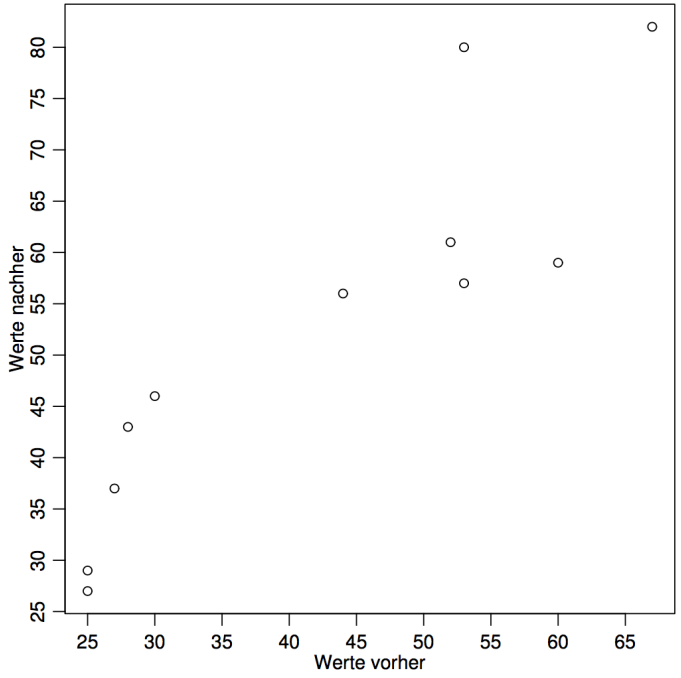


empirische kumulative Verteilungsfunktion $Fn(\cdot)$: Treppenfunktion, die bei jedem $x_{(i)}$ einen Sprung der Höhe $\frac{1}{n}$ oder eines Vielfachen bei mehrfachem Auftreten des jeweiligen Wertes macht

$$F_n(x) = \frac{1}{n}\text{Anzahl}\{i|x_i \geq x\}$$



Streudiagramm: Datenpunkte i mit Koordinaten (x_i, y_i) werden in einer Ebene dargestellt



4.2 Stetige Zufallsvariablen und Wahrscheinlichkeitsverteilung

Wertebereich stetig → $P(X = x) = 0$ für alle x ↯

4.2.1 Wahrscheinlichkeitsdichte

$P(X \leq x) =: F(x)$ kumulative Verteilungsfunktion
$f(x) = \frac{d}{dx}F(x)$ Wahrscheinlichkeitsdichte

$$\Rightarrow F(x) = \int_{-\infty}^{x} f(x')dx'$$

4.2.2 Kennzahlen

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx\,; \quad Var(X) = E((X - E(X))^2)$$

$$\sigma_x = \sqrt{Var(X)}\,; \quad Quantil: q_\alpha = F^{-1}(\alpha)$$

4.3 Wichtige stetige Verteilung

4.3.1 Uniforme Verteilung

$$X \sim Uniform([a,b])$$

Jeder Wert im Intervall [a,b] ist gleich wahrscheinlich.

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

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

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

4.3.2 Exponentialverteilung

$$X \sim Exp(\lambda)$$

"Wartezeit auf Ausfälle"
X mit Wertebereich $W_x = \Re^+ = [0,\infty)$ ist exponentiell verteilt mit Parameter $\lambda \in \Re^+$ ($X \sim e^\lambda$), falls

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

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

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

4.3.3 Normal- oder Gauss-Verteilung

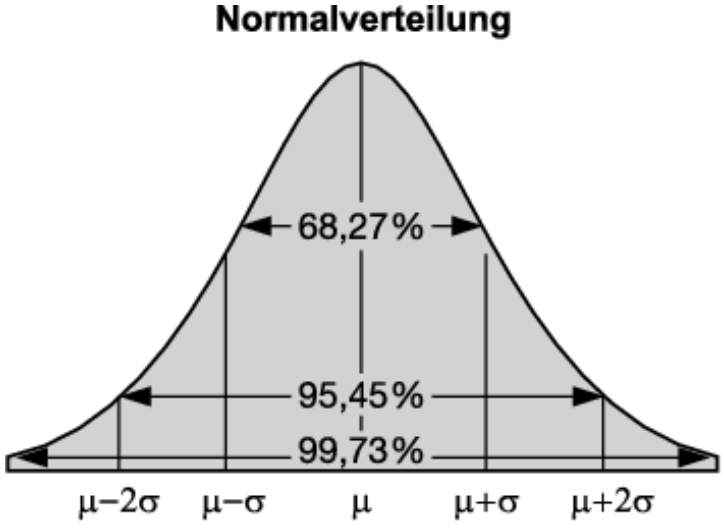
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Häufigste Verteilung für Messwerte
X mit Wertebereich $W_x = \Re$ ist normalverteilt mit Parameter $\mu \in \Re$ und $\sigma^2 \in \Re^+$ falls

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

Die kumulative Verteilungsfunktion ist nicht explizit darstellbar und wird deswegen tabelliert. Dabei reicht eine Tabelle für die Standard-Normalverteilung da jede Normalverteilung immer in eine Standard-Normalverteilung transformiert werden kann (siehe Standardisierung einer Zufallsvariablen weiter unten).

$$E(X) = \mu, Var(X) = \sigma^2, \sigma_X = \sigma$$



Standard-Normalverteilung Die Normalverteilung mit $\mu = 0$ und $\sigma^2 = 1$ heisst Standard-Normalverteilung. Dichte und kumulative Verteilungsfunktion sehen wie folgt aus:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad \Phi(x) = \int_{-\infty}^{x} \phi(y)dy$$

4.3.4 Funktion einer Zufallsvariable

Ist X eine Zufallsvariable, dann ist es $Y = g(X)$ ebenfalls. Es gilt stets

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

Lineare Transformation

Ist $g(x) = a + bx$ so gilt für $Y = a + bX$:

$$E(Y) = E(a + bX) = a + bE(X)$$

$$Var(Y) = Var(a + bX) = b^2 Var(x), \quad \sigma_Y = |b|\sigma_X$$

$$q_Y = a + bq_X$$

Standardisierung einer Zufallsvariablen

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

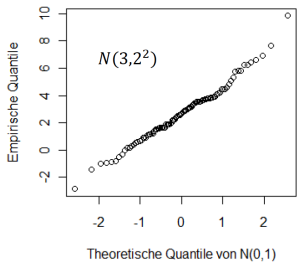
Bsp: $X \sim \mathcal{N}(\mu, \sigma^2)$ mit $\mu = 2$, $\sigma^2 = 4$. Berechne $P(X \geq 5)$

$$\begin{aligned} P(X \geq 5) &= P\left(\frac{X-\mu}{\sigma} \geq \frac{5-\mu}{\sigma}\right) = P\left(Z \geq \frac{5-2}{2}\right) \\ &= P(Z \geq 1.5) = 1 - P(Z \leq 1.5) = 1 - \Phi(1.5) \\ &= 1 - 0.933 = 0.067 \end{aligned}$$

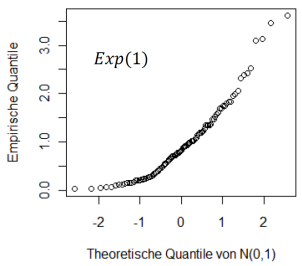
4.3.5 Überprüfen der Normalverteilungs-Annahme

Der Q-Q Plot vergleicht die empirischen mit den theoretischen Quantilen der Modell-Verteilung. Entsprechen die empirischen Quantile also den theoretischen, hat der Plot in etwa die Form der Winkelhalbierenden $y = x$.

Gerade:
Normalverteilung OK



Krümmung:
Keine Normalverteilung



4.4 Funktionen von mehreren Zufallsvar.

Haben wir mehrere Zufallsvariablen, treffen wir die Annahme $X_1, \ldots, X_n \sim F \quad iid$. Dies bedeutet, dass $X_1, \ldots, X_n$ unabängig voneinander und dabei gleich verteilt sind (iid steht für "independent & identically distributet"). Somit gilt folgendes (a sei eine Konstante):

$$E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n)$$

$$Var(X_1 + \ldots + X_n) = Var(X_1) + \ldots + Var(X_n)$$

$$Cov(a, X) = 0;\ Cov(X, Y) = Cov(Y, X)$$

$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$$

Gesetz der grossen Zahlen (GGZ)

$$E(\overline{X}_n) = \mu, \quad Var(\overline{X}_n) = \frac{\sigma_X^2}{n}, \quad \sigma(\overline{X}_n) = \frac{\sigma_X}{\sqrt{n}}$$

$\sqrt{n}$-Gesetz: n-mal mehr Beobachtungen
→ Vertrauensintervall um $\sqrt{n}$ kleiner
→ σ um $\sqrt{n}$ kleiner bzw. Genauigkeit um $\sqrt{n}$ grösser

Zentraler Grenzwertsatz (ZGS)

Sind $X_1, \ldots, X_n \quad iid$, dann gilt

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right), \quad S_n \approx \mathcal{N}(n\mu, n\sigma_X^2)$$

4.5 Statistik für eine Stichprobe

4.5.1 Punkt-Schätzung

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad \hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

4.5.2 Tests für μ

Die in diesem Teil aufgeführten Test sind wie folgt nach der Anzahl Annahmen und gleichzeitig der Macht einzuordnen.
z > t > Wilcoxon > VZ

z-Test: σ_X bekannt

1) **Modell**: X_i kontinuierliche Messgrösse, $X_1, \ldots, X_n \quad iid \quad \mathcal{N}(\mu, \sigma_X^2)$, σ_X bekannt.
2) **Nullhypothese** $H_0 : \mu = \mu_0$
   Alternative $H_A : \mu \neq \mu_0$ (oder < oder >)
3) **Teststatistik** T:
   $Z = \frac{(\overline{X}_n - \mu_0)}{\sigma_{\overline{X}_n}} = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma_X}$
   Verteilung der Teststatistik unter H_0:
   $Z \sim \mathcal{N}(0,1)$
4) **Signifikanzniveau**: α
5) **Verwerfungsbereich von** T:
   - $\mu \neq \mu_0$:
     $K = (-\infty, -\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)] \cup [\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \infty)$
   - $\mu < \mu_0$: $K = \left(-\infty, -\Phi^{-1}(1-\alpha)\right]$
   - $\mu > \mu_0$: $K = (\Phi^{-1}(1-\alpha), \infty)$

   (siehe Kapitel 6 für Φ)
6) **Testentscheid**: Überprüfen, ob Wert im Verwerfungsbereich liegt

t-Test: σ_X unbekannt

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

1) **Modell**: X_i kontinuierliche Messgrösse, $X_1, \ldots, X_n \quad iid \quad \mathcal{N}(\mu, \sigma_X^2)$, σ_X wird durch $\hat{\sigma}_X$ geschätzt.
2) **Nullhypothese** $H_0 : \mu = \mu_0$
   Alternative $H_A : \mu \neq \mu_0$ (oder < oder >)
3) **Teststatistik** T:

$$t = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\hat{\sigma}_X}$$

   Verteilung der Teststatistik unter $H_0 : T \sim t_{n-1}$
4) **Signifikanzniveau**: α
5) **Verwerfungsbereich von** T:
   - $\mu \neq \mu_0$: $K = (-\infty, -t_{n-1;1-\frac{\alpha}{2}}] \cup [t_{n-1;1-\frac{\alpha}{2}}, \infty)$
   - $\mu < \mu_0$: $K = (-\infty, -t_{n-1;1-\alpha}]$
   - $\mu > \mu_0$: $K = (t_{n-1;1-\alpha}, \infty)$
6) **Testentscheid**: Überprüfen, ob Wert im Verwerfungsbereich liegt

P-Wert: $p = P(T > t) = 1 - F_{t_{n-1}}(t)$
Beachte: $P(T > t) = 1 - P(T <= t)$, $P(T >= t) = 1 - P(T < t)$
$F_{t_{n-1}}$ kann aus der Tabelle ausgelesen werden.
Dazu sucht man in der Zeile $n-1$ nach der Bedingung (z.B. T > 2.228). Die dazugehörige Spalte ($t_{0.975}$) gibt einem den p-Wert an: $p = 1-0.975 = 0.025$. Bei einem zweiseitigen Test währe $p = 2(1-0.975) = 2 * 0.05 = 0.05$.

Vorzeichentest (Binomialtest)

1) **Modell**: $X_1, \ldots, X_n \quad iid$ wobei X_1 eine beliebige Verteilung hat.
2) **Nullhypothese** $H_0 : \mu = \mu_0$ (μ ist der Median)
   Verteilung unter H_0 $V \sim Bin(n, \pi_0)$ mit $\pi_0 = 0.5$
   Alternative $H_A : \mu \neq \mu_0$ (oder < oder >)
3) **Teststatistik** V : Anzahl X_i mit $X_i > \mu_0$
4) **Signifikanzniveau**: α
5) **Verwerfungsbereich von** T :
   $\mu \neq \mu_0 : \; K = [0, c_u] \cup [c_o, n]$
   c_u und c_o müssen mit der Binomialverteilung oder der Normalapproximation berechnet werden.
6) **Testentscheid**: Überprüfen, ob Wert im Verwerfungsbereich liegt

Wilcoxon-Test

- Kompromiss, der Normalverteilung nicht voraussetzt (t-Test) aber die Information der Daten besser ausnützt als der Vorzeichentest.
- Annahme: $X_i \sim F \quad iid$, F ist symmetrisch
- Teste Median μ : $H_0 : \; \mu = \mu_0$
- Intuition der Teststatistik
  - Rangiere $|x_i - \mu_0| \rightarrow r_i$
  - Gib Rängen ursprüngliches Vorzeichen von $(x_i - \mu_0)$ (“signed ranks”)
  - Teststatistik: Summe aller Ränge, bei denen $(x_i - \mu_0)$ positiv ist.
- Falls H_0 stimmt, sollte Summe weder zu gross noch zu klein sein.

4.5.3 Vertrauensintervall für μ

95%-VI werden nach dem folgenden Schema berechnet. Dabei

$$\begin{aligned} \mu \neq \mu_0 : &\quad [c_u, c_o] \\ \mu < \mu_0 : &\quad [-\infty, c_o] \\ \mu > \mu_0 : &\quad [c_u, \infty] \end{aligned}$$

⚠ In der folgenden Formel für zweiseitiges Vertrauensintervall $\alpha/2$ statt α verwenden.

$$c_{o/u} = \overline{x}_n \pm t_{(n-1,1-\alpha)} \frac{\hat{\sigma}_X}{\sqrt{n}} = \overline{x}_n \pm \Phi^{-1}_{(1-\alpha)} \frac{\sigma_X}{\sqrt{n}}$$

⚠ Die Formel mit t und $\hat{\sigma}_X$ gilt für den t-Test (geschätztes σ) und diejenige mit Φ^{-1} und σ_X für den z-Test (bekanntes σ). Für Φ^{-1} siehe Kapitel 6.

4.6 Test bei zwei Stichproben

4.6.1 Vergleich gepaarter und ungepaarter t-Test

Bei gepaarten Stichproben kann auch der ungepaarte t-Test verwendet werden.

gepaart	ungepaart
gleich grosse Stichproben	können, müssen aber nicht gleich gross sein
klare Zuordnung (rechts - links, vorher - nachher)	keine Zuordnung
mehr Macht	weniger Macht

4.6.2 Gepaarte Stichproben

Sind Daten gepaart (z.B. Messung vor und nach der Einnahme eines Medikamentes), arbeitet man mit den Differenzen innerhalb der Paare (Test für eine Stichprobe).

$$u_i = x_i - y_i \quad (i = 1, \ldots, n)$$

4.6.3 Ungepaarte Stichproben

Sind Daten ungepaart wendet man den **ungepaarten t-Test** an.

1) **Modell**:

$$\begin{aligned} X_1, \ldots, X_n \; iid &\quad \sim \mathcal{N}(\mu_X, \sigma^2) \\ Y_1, \ldots, Y_m \; iid &\quad \sim \mathcal{N}(\mu_Y, \sigma^2) \end{aligned}$$

2) **Nullhypothese** $H_0 : \mu_X = \mu_Y$
   Alternative:
   - $H_A : \; \mu_X \neq \mu_Y$ (zweiseitig)
   - $H_A : \; \mu_X > \mu_Y$ (einseitig)
   - $H_A : \; \mu_X < \mu_Y$ (einseitig)
3) **Teststatistik** T :

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

$$S^2_{pool} = \frac{1}{n+m-2}\left((n-1)\hat{\sigma}^2_x + (m-1)\hat{\sigma}^2_y\right)$$

   Verteilung der Teststatistik unter $H_0 : T \sim t_{n+m-2}$
4) **Signifikanzniveau**: α
5) **Verwerfungsbereich von** T:
   - $\mu_X \neq \mu_Y$:
     $K = (-\infty, -t_{n+m-2;1-\frac{\alpha}{2}}] \cup [t_{n+m-2;1-\frac{\alpha}{2}}, \infty)$
   - $\mu_X > \mu_Y : \; K = [-t_{n+m-2;1-\alpha}, \infty)$
   - $\mu_X < \mu_Y : \; K = (-\infty, t_{n+m-2;1-\alpha}]$
6) **Testentscheid**: Überprüfen, ob Wert im Verwerfungsbereich liegt

Zwei-Stichproben t-Test bei ungleichen Varianzen (Welch-Test)

In den meisten Fällen erhält man ähnliche P-Werte wie unter der Annahme von gleichen Varianzen.

$$\begin{aligned} X_1, \ldots, X_n \; iid &\quad \sim \mathcal{N}(\mu_X, \sigma^2_X) \\ Y_1, \ldots, Y_m \; iid &\quad \sim \mathcal{N}(\mu_Y, \sigma^2_Y) \end{aligned}$$

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

$$X_1,\ldots,X_n \; iid \quad \sim F_X$$
$$Y_1,\ldots,Y_m \; iid \quad \sim F_Y$$

Wobei F_X eine beliebige Verteilungsfunktion und $F_Y(x) = F_X(x-\delta)$ (d.h. Verteilungen sind identisch aber um δ verschoben). Die Berechnung des P-Werts sollte mit dem Computer erfolgen.

4.7 Multiples Testen: Bonferroni Korrektur

Gesucht ist eine Liste mit der Eigenschaft $P(\text{mindestens ein Fehler 1. Art}) \leq \alpha$. Die Bonferroni Korrektur setzt das Signifikanzniveau auf $\frac{\alpha}{m}$, wobei m die Anzahl Tests ist. Der Nachteil besteht darin, dass die Liste zu "konservativ" sein kann und keine beobachteten Werte mehr enthält.

$$P\left(\bigcup_{i=1}^{m} F_i\right) \leq \sum_{i=1}^{m} P(F_i) = \sum_{i=1}^{m} \frac{\alpha}{m} = \alpha$$

5 Regression

5.1 Einfache Lineare Regression

Aus dem Datensatz soll ein linearer Zusammenhang gefunden werden. Dabei sind die Fehler um die Gerade herum normal verteilt. Das Modell kann die folgende Form haben:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i\,, \quad \epsilon_i \sim \mathcal{N}(0,\sigma^2) \quad iid$$

Sind die β's nicht wie oben linear (z.B. keine $exp(\beta x_i)$ oder $log(\beta_0 + \beta_1 x_i + \epsilon_i)$), so ist das Modell ebenfalls nicht linear.
 # Datenpunkte = degrees of freedom (dof) + # β's
Koeffizienten: $\hat{\beta} = \sigma(\hat{\beta}) * t(\hat{\beta})$
95%-VI
genau: $VI(\beta) = \beta \pm t_{df,0.975} * \sigma(\beta)$
approximativ: $VI(\beta) = \beta \pm 2 * \sigma(\beta)$
Verwerfungsbereich:

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

p-Wert: Bsp: $t(\beta_0) = \beta_0/\sigma(\beta_0) = -0.419/0.246 = -1.7$ $-t_{47,1-(\text{p-Wert}/2)} = -1.7 \rightarrow t_{47,1-(\text{p-Wert}/2)} = 1.7 \rightarrow$ Tabelle $t_{47,0.95} = 1.7 \rightarrow \text{p-Wert}/2 = 0.05 \rightarrow \text{p-Wert} = 0.1$
Erwartetes y_i: x, β_0 und β_1 in $y_i = \beta_0 + \beta_1 x_i$ einsetzen

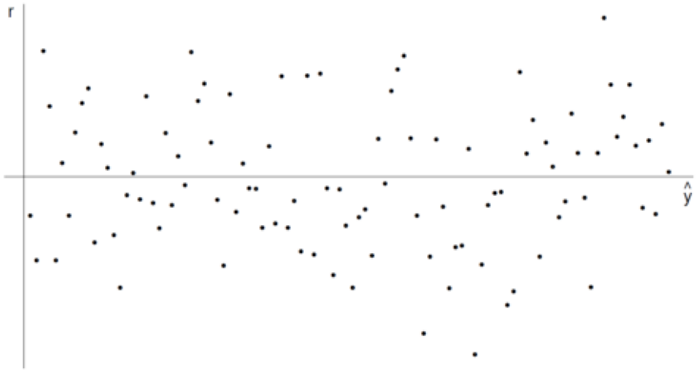
5.2 R-Output bei Linearer Regression

`Estimate:` $\hat{\beta}_0$
`Std. Error:` Standardfehler $\sigma(\hat{\beta})$
`t value:` t-Wert
`Pr(>|t|):` p-Wert
`(Intercept)` β_0
Zeile darunter: β_1

5.3 Tukey-Anscombe-Plot

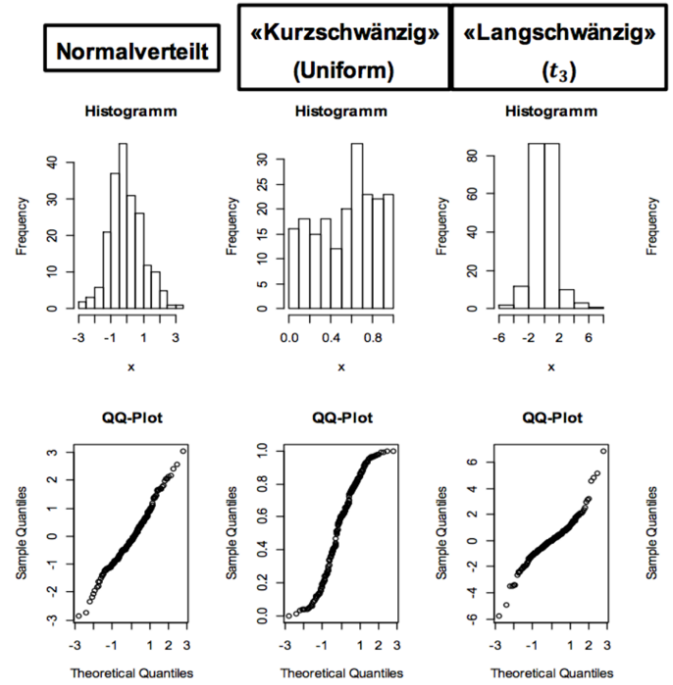
Der Tukey-Anscombe-Plot zeigt die Fehlervarianz über die ganze Breite der Daten an.

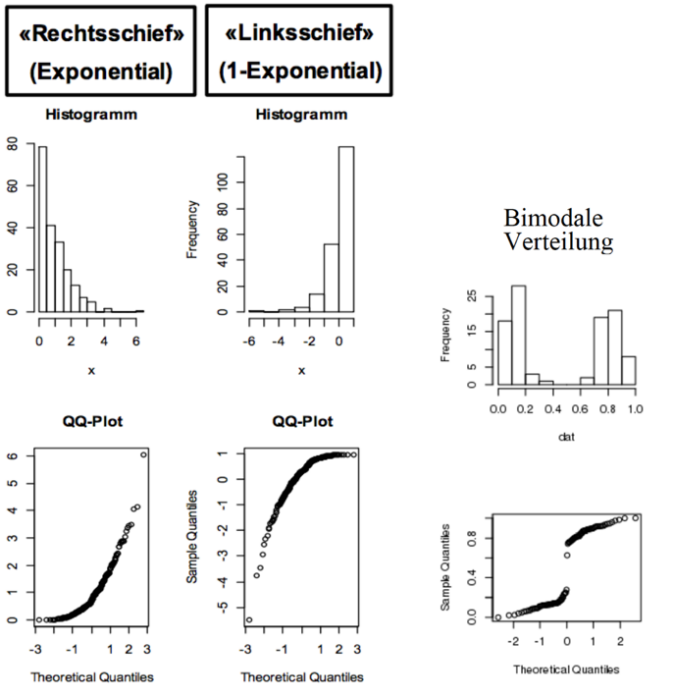
5.3.1 Konstante Fehlervarianz



5.3.2 Nicht Konstante Fehlervarianz

5.4 QQ-Plot und Histogramm: Verteilungen





Rechtsschief: Median < Erwartungswert, rechts flacher
Linksschief: Median > Erwartungswert, links flacher

6 Tabelle der Kumulativen Normalverteilung

6.1 $\Phi(z)$

Die kumulative Normalverteilung Φ hängt von z ab. Wird also ein bestimmter Wert (z.B. $z = 1.96$) gesucht, sucht man in der ersten Spalte den Wert bis und mit der ersten Nachkommastelle (hier 1.9). Denn gesuchte Wert von Φ findet man in der Spalte der zweiten Nachkommastelle (hier .06). In unserem Beispeil beträgt dieser 0.975.
Es gilt $\Phi(-z) = 1 - \Phi(z)$. Ausserdem sind bei der Normalverteilung $<$ und $\leq$ als auch $>$ und $\geq$ austauschbar, da die zusätzliche $=$ Bedingung den Wert nicht verändert.
Negative z: Im Folgenden steht $-z$ für $z < 0$. Ist $P(Z > -z)$ gesucht, gilt $P(Z > -z) = P(Z < z)$, da
$P(Z > z) = 1 - P(Z < -z) = 1 - (1 - P(Z < z))$

6.2 $\Phi^{-1}(z)$

Sucht man einen Wert der Umkehrfunktion, also z, so Sucht man den bekannten Wert von Φ in der Tabelle (um das Bsp von eben wieder aufzugreifen: $\Phi(z) = 0.975$). Von dort findet man den gesuchten z-Wert durch kombinieren der Zeilen- und Spalten-Indizes
(hier: $z = 1.9 + .06 = 1.96$).

7 Beispielaufgaben

7.1 Vergleiche von P(X) und P(Z)

Angenommen $X \sim t_5$, $Z \sim \mathcal{N}(0,1)$. Dann ist $P(X \leq 2)$ grösser als $P(Z \leq 2)$.

t-Tabelle $\rightarrow t_{5;0.95} \approx 2 \rightarrow P(X \leq 2) \approx 0.95$
z-Tabelle $\rightarrow P(Z \leq 2) \approx 0.9772 \implies P(Z \leq 2) > P(X \leq 2)$.
Damit ist die Aussage **falsch**.

7.2 p-Wert bei t-Test berechnen

7.2.1 Bsp 1

Bei einem zweiseitigen t-Test mit $n = 20$ Beobachtungen ist der Wert der Teststatistik 1.729. Der P-Wert ist dann etwa 0.1.

Verteilung der Teststatstik: $T \sim t_{n-1} = t_{19}$
Wert von T: $t = 1.729 \rightarrow$ t-Tabelle $\rightarrow P(T \leq 1.729) \approx 0.95$
p-Wert: $p = P(T \geq t) + P(T \leq -t) = 2 * P(T \geq t)$
Es gilt: $P(T \geq t) = 1 - P(T \leq t)$
$\Rightarrow p = 2 * P(T \geq t) = 2 * (1 - P(T \leq t)) = 2 * (1 - 0.95) \approx 0.1$
Somit ist die Aussage **richtig**.

7.2.2 Bsp 2

Angenommen bei einem zweiseitigen Zwei-Stichproben t-Test mit $n_1 = 10$ und $n_2 = 6$ Beobachtungen ist der beobachtete Wert der Teststatistik $t = 2.624$. Der P-Wert ist dann 0.1.

Freiheitsgrade $df = n_1 + n_2 - 2 = 14$
Verteilung: $T \sim t_{14} \rightarrow$ t-Tabelle $\rightarrow t_{14;0.99} = 2.624$
p-Wert bei **zweiseitigem Test**: $p = 2 * 0.01 = 0.02$
Die Aussage ist also **falsch**.

7.3 Textaufgaben zur Normalverteilung

7.3.1 Bsp 1

Eine Gaskartusche hält im Schnitt für 1h mit Standardabweichung 0.1h. Die Brennzeit der Kartuschen ist unabhängig voneinander. Die Wahrscheinlichkeit, dass 21 Kartuschen genug sind für 20h Brennzeit, ist grösser als 95%.

Gesamtbrennzeit $S = \sum_{i=1}^{21} S_i$
Erwartungswert $E(S_i) = 1h$ und Varianz $Var(S_i) = 0.1^2$
ZGS: $S \sim \mathcal{N}(n * 1, n * 0.1^2) = \mathcal{N}(21, 0.21)$
Wahrscheinlichkeit:

$$\begin{aligned} P(S > 20h) &= P\left(\frac{S-21}{\sqrt{0.21}} > \frac{20-21}{\sqrt{0.21}}\right) \\ &= P(Z > -2.1822) = P(Z \leq 2.1822) \\ &\approx 0.985 > 0.95 \end{aligned}$$

Die Aussage ist **richtig**.

7.3.2 Bsp 2

Ein durchschnittlicher Lachs ist 10 kg schwer mit einer Standardabweichung von 2 kg. Das Gewicht der Fische ist unabhängig voneinander. Ein Fischerboot fängt 30 Lachse an einem Tag. Die Wahrscheinlichkeit, dass der Fang mehr als 330 kg wiegt, ist kleiner als 1%.

Gesamtgewicht $Z = \sum_{i=1}^{30} X_i$
Erwartungswert $E(X_i) = 10$ kg und Varianz $Var(X_i) = 2^2$
ZGS: $X \sim \mathcal{N}(n * 10, n * 2^2) = \mathcal{N}(300, 120)$
Wahrscheinlichkeit:

$$\begin{aligned} P(X > 330) &= P\left(\frac{Z-300}{\sqrt{120}} > \frac{330-300}{\sqrt{120}}\right) \\ &= P(\frac{Z-300}{\sqrt{120}} > 2.74) = P(Z^* \leq 2.74) \\ &= 1 - 0.9969 = 0.0031 < 0.01 \end{aligned}$$

$Z^* \sim \mathcal{N}(0,1)$ ist das standardisierte Gesamtgewicht
Die Aussage ist **richtig**.

8 TI-83/84

8.1 binompdf und binomcdf

DISTR: 2nd → VARS

- binompdf:
  Wahrscheinlichkeitsfkt. der Binomialverteilung
- binomcdf:
  Verteilungsfkt. der Binomialverteilung

$X \sim \text{Binom}(n, \pi)$

- $P[X = x] \Rightarrow \text{binompdf}(n, \pi, x)$
- $P[X \leq x] \Rightarrow \text{binomcdf}(n, \pi, x)$
- $P[X < x] \Rightarrow \text{binomcdf}(n, \pi, x-1)$
- $P[X \geq x] \Rightarrow 1-\text{binomcdf}(n, \pi, x-1)$
- $P[X > x] \Rightarrow 1-\text{binomcdf}(n, \pi, x)$

8.2 Binomialkoeffizient und Fakultät

MATH → PRB
$\binom{n}{x} \Rightarrow n$ nCr x **oder** $\frac{n!}{x!(n-x)!}$

8.3 Tests

Daten müssen in Listen gespeichert werden.
z.B.: $\{15, 8, -1, 2\} \rightarrow L_1$
{: 2nd → (}: 2nd →)
→: STO> L_x: 2nd → STAT (LIST)
$\text{Freq}_{(1/2)}$ immer = 1
STAT → TESTS

- z-Test...
- t-Test...
- 2-SampTTest...: ungepaarter t-Test
  (immer mit Pooled: Yes)