

Statistik I

Ajay Bharadwaj
Frühlingssemester 2024

Vorwort

Dieses Skript wurde für die Vorlesung «Statistik I» im Frühlingsemester 2023 geschrieben und basiert auf die Vorlesungsfolien und Aufgaben von Dr. Markus Kalisch.

Das Skript erhebt keinen Anspruch auf Vollständigkeit sowie Richtigkeit und ersetzt auf keinen Fall den Unterrichtsstoff. Im Zweifelsfall gelten die Informationen aus der Vorlesung.

Dieses Skript dient ausschliesslich Studierende der ETH Zürich, welche die «Statistik I» Vorlesung von Dr. Markus Kalisch besuchen und darf nicht weiterverbreitet werden.

Aktuelle Version: 13. Mai 2024

Inhaltsverzeichnis

1	Wahrscheinlichkeitsmodelle.....	5
1.1	Mengenlehre.....	5
1.2	Laplace-Modell.....	5
2	Bedingte Wahrscheinlichkeit und Odds.....	6
2.1	Unabhängigkeit.....	6
2.2	Abhängigkeit.....	6
2.3	Odds & LogOdds.....	6
3	Zufallsvariablen & Verteilungen.....	7
3.1	Kennzahlen einer Verteilung.....	7
3.2	Binomialverteilung.....	7
4	Diskrete Verteilungen.....	9
4.1	Poisson Verteilung $\rightarrow X \sim \text{Pois}(\lambda)$	9
4.2	Uniforme Verteilung $\rightarrow X \sim \text{Unif}(n)$	9
4.3	Hypergeometrische Verteilung $\rightarrow X \sim \text{Hyper}(N, n, m)$	9
4.4	Parameterschätzung.....	9
5	Einseitiger Binomialtest.....	10
5.1	Binomialtest.....	10
5.2	Verwerfungsbereich.....	10
5.3	Fehlerarten.....	10
6	Zweiseitiger Binomialtest.....	11
6.1	P-Wert.....	11
6.2	Vertrauensintervall.....	11
6.3	Verwerfungsbereich.....	11
7	Deskriptive Statistik.....	12
7.1	Kennzahlen.....	12
7.2	Quantile q_α	12
7.3	Empirische Korrelation.....	12
7.4	Graphische Darstellungsweisen.....	13

8	Kontinuierliche Verteilungen & ZGS	14
8.1	Kontinuierliche Verteilungen.....	14
8.2	Normalverteilung.....	14
8.3	Lineare Transformation	15
8.4	Zentraler Grenzwertsatz	15
9	Ein-Stichprobe t-Test	16
9.1	Gesetz der grossen Zahlen.....	16
9.2	z-Test.....	16
9.3	t-Test	17
10	Zwei-Stichproben t-Test.....	18
10.1	Gepaarte / ungepaarte Stichproben	18
10.2	Zwei Stichproben.....	18
11	Lineare Regression.....	19
11.1	Parameterschätzungen	19
11.2	Plots.....	20

1 Wahrscheinlichkeitsmodelle

1.1 Mengenlehre

1. Grundraum $\Omega = \{\text{alle möglichen Elementarereignisse } \omega_i\}$
2. Ereignisse sind Teilmengen des Grundraums
3. Wahrscheinlichkeit für Ereignis A ist $P[A]$

Bsp. Würfel:

$$\Omega = \{1,2,3,4,5,6\}$$

$$\omega_1 = 1 ; \omega_2 = 2 ; \dots$$

Ereignis A: *gerade Zahlen* $\rightarrow A = \{2,4,6\}$

$P[A] \geq 0$ Wahrscheinlichkeiten können nicht negativ sein

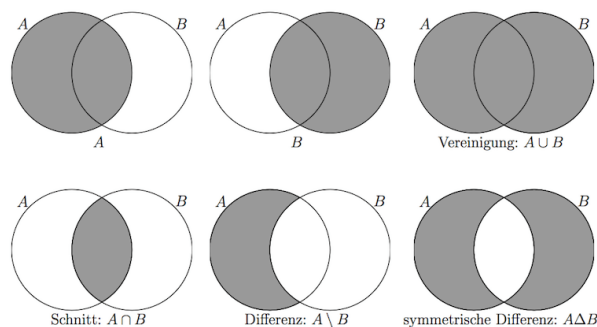
$P[\Omega] = 1$ Die gesamte Wahrscheinlichkeit ist immer 100%

$P(A \cap B) \rightarrow A \text{ und } B$

$P(A \cup B) \rightarrow A \text{ oder } B \rightarrow P(A) + P(B) - P(A \cap B)$

$P(A \setminus B) \rightarrow A \text{ ohne } B \rightarrow P(A) - P(A \cap B)$

$P(A^C) \rightarrow \text{nicht } A \text{ (Gegenereignis)} \rightarrow 1 - P(A)$



Wenn A und B keine Überschneidung haben (Schnittmenge = 0), dann sind A und B *disjunkt* und es gelten folgende Beziehungen:

$$P(A \cap B) = 0 = \{ \}$$

$$P(A \cup B) = P(A) + P(B)$$

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

1.2 Laplace-Modell

Wenn alle Elementarereignisse die gleiche Wahrscheinlichkeit haben (*Bsp. Fairer Würfel*), kann folgende Formel benutzt werden für die Berechnung von $P[A]$:

$$P[A] = \frac{\text{Anzahl günstige Elementarereignisse}}{\text{Gesamte Anzahl Elementarereignisse}}$$

2 Bedingte Wahrscheinlichkeit & Odds

2.1 Unabhängigkeit

Falls das Auftreten von Ereignis A die Wahrscheinlichkeit von Ereignis B nicht beeinflusst, sind sie unabhängig. Das gilt zum Beispiel, wenn zweimal hintereinander gewürfelt wird. Das Ergebnis vom zweiten Versuch hängt nicht vom Ergebnis des ersten Versuchs ab.

$$P[A \cap B] = P[A] * P[B]$$

$$P[A] = P[A|B] = P[A|B^c]$$

2.2 Abhängigkeit

Falls die Wahrscheinlichkeit für Ereignis B von Ereignis A beeinflusst wird, dann sind sie abhängig. Ereignis A könnte zum Beispiel «früh schlafen gehen» sein und Ereignis B wäre «Morgenvorlesung besuchen». Diese zwei Ereignisse sind abhängig, weil wenn jemand zu spät schlafen geht, sinkt die Wahrscheinlichkeit für einen Vorlesungsbesuch am Morgen.

$$P[A|B] \neq P[B|A]$$

$$\text{Satz von Bayes: } P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[B|A] * P[A]}{P[B]}$$

$$\text{Gegenereignis: } P[A|B] = 1 - P[A^c|B]$$

$$\text{Satz der totalen Wahrscheinlichkeit: } P[B] = \sum P[B|A_i] * P[A_i]$$

$$\text{Falls nur } A \text{ \& } A^c \text{ existieren: } P[B] = P[B|A] * P[A] + P[B|A^c] * P[A^c]$$

2.3 Odds & LogOdds

$$\text{odds} = \frac{p}{1-p} = \frac{P[A]}{P[A^c]}$$

$$\text{logOdds} = \ln(\text{odds})$$

$$\text{odds} = e^{\text{logOdds}}$$

3 Zufallsvariablen & Verteilungen

3.1 Kennzahlen einer Verteilung

X ist eine Funktion, welche eine Zufallsvariable x liefert.

$$\sum_{\text{alle } x_i} P[X = x_i] = 1$$

$$\text{Erwartungswert } E(X) = \sum x_i * P(X = x_i)$$

$$\text{Varianz } \text{Var}(X) = \sum (x_i - E(X))^2 * P(X = x_i)$$

$$\text{Standardabweichung } \sigma(X) = \sqrt{\text{Var}(X)}$$

$$P[X \leq x_j] = \sum_{i \leq j} P[X = x_i]$$

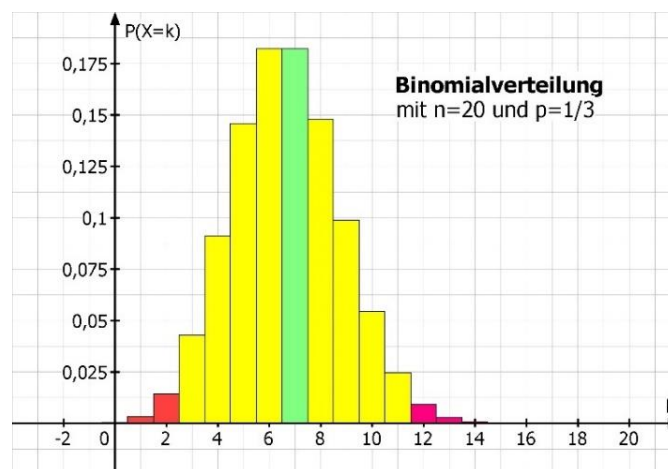
$$\text{Bsp. } P[X \leq 3] = P[X = 1] + P[X = 2] + P[X = 3]$$

3.2 Binomialverteilung

Die Variable x zählt die Anzahl der Erfolge in einer Serie von gleichartigen, unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben. Die Variable n sagt dabei aus, wie viele Versuche insgesamt durchgeführt werden. Die Erfolgswahrscheinlichkeit wird mit p oder π angegeben.

Falls sich die Erfolgswahrscheinlichkeit nach einem Versuch ändert, kann die Binomialverteilung nicht eingesetzt werden.

Beispiel für $X \sim \text{Bin}(20, 0.33)$



Formeln für die Binomialverteilung

$$P[X = x]$$

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

Taschenrechner: $\text{binomPdf}(n, p, x)$

$$P[X \leq x]$$

$$P[X \leq x] = \sum_{i=0}^x \binom{n}{i} \pi^i (1 - \pi)^{n-i}$$

Taschenrechner: $\text{binomCdf}(n, p, x)$

$$P[X > x]$$

Taschenrechner: $P[X > x] = 1 - P[X \leq x]$ (Gegenereignis)

Kennzahlen Binomialverteilung

$$E[x] = n * \pi$$

$$\text{Var}[x] = n * \pi * (1 - \pi)$$

$$\sigma[x] = \sqrt{\text{Var}[x]}$$

$$\text{Binomialkoeffizient: } \binom{n}{k} = \frac{n!}{k! * (n - k)!}$$

4 Diskrete Verteilungen

4.1 Poisson Verteilung $\rightarrow X \sim \text{Pois}(\lambda)$

Seltene Ereignisse in einem fixen Zeitraum. *Bsp. Unfallraten, radioaktive Zerfälle*

Für $P(X = x)$ & $P(X \leq x) \rightarrow \text{TR PoissonPdf} \& \text{PoissonCdf}$

$$E(X) = \lambda$$

$$\text{Var}(X) = \lambda = \sigma^2$$

grosse n und kleine $\pi \rightarrow \text{Bin}(n, \pi) \approx \text{Pois}(\lambda)$

*unabhängige X & $Y \rightarrow E(X + Y) = E(X) + E(Y)$ und $E(a * X) = a * E(X)$*

4.2 Uniforme Verteilung $\rightarrow X \sim \text{Unif}(n)$

Alle Ereignisse müssen gleich wahrscheinlich sein. *Bsp. Fairer Würfel*

$$P(X = x) = \frac{1}{n} \quad (n \rightarrow \text{Anzahl Ereignisse})$$

$$E(X) = \frac{n + 1}{2}$$

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

4.3 Hypergeometrische Verteilung $\rightarrow X \sim \text{Hyper}(N, n, m)$

Gewinnwahrscheinlichkeiten sind nicht konstant und sind vom vorherigen Resultat abhängig.

Bsp. Markierte Kugeln aus einer Urne ziehen, ohne diese zurückzulegen.

$m \rightarrow \text{markierte}$

$n \rightarrow \text{gezogen}$

$x \rightarrow \text{markierte von den gezogenen}$

$N \rightarrow \text{Totale Anzahl}$

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

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

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

4.4 Parameterschätzung

$$\hat{\pi} = \frac{E(X)}{n} \rightarrow \text{hier ist } E(X) \text{ der beobachtete Wert für } x$$

5 Einseitiger Binomialtest

5.1 Binomialtest

Ein Modell wird mit der Binomialverteilung aufgestellt und danach werden mehrere Versuche durchgeführt. Ziel ist es, zu prüfen, ob der beobachtete Testwert «realistisch» ist.

Bsp. Ist ein spezifischer Würfel fair, oder werden übermässig viele 6er gewürfelt?

Signifikanzniveau $\rightarrow$ meistens $\alpha = 0.05$

Wenn das beobachtete Ereignis eine kleinere Wahrscheinlichkeit als das Signifikanzniveau aufweist, wird es als «unrealistisch» angesehen und die Nullhypothese wird verworfen.

5.2 Verwerfungsbereich

$$H_A: \pi \neq \pi_o \text{ (beidseitig)} \rightarrow K = [0, c_u] \cup [c_o, n]$$

$$H_A: \pi > \pi_o \text{ (rechtsseitig)} \rightarrow K = [c_r, n]$$

$$H_A: \pi < \pi_o \text{ (linksseitig)} \rightarrow K = [0, c_l]$$

Die Grenzen können mit einsetzen und ausprobieren im Taschenrechner gefunden werden.

$$P[T \leq c_u] \leq \frac{\alpha}{2} \quad \text{und} \quad P[T \geq c_o] \leq \frac{\alpha}{2}$$

$$P[T \geq c_r] \leq \alpha$$

$$P[T \leq c_l] \leq \alpha$$

Falls die gemessene Teststatistik sich im Verwerfungsbereich befindet, kann die Nullhypothese H_0 auf dem Signifikanzniveau α verworfen werden.

Achtung: Nullhypothese kann nicht verworfen werden $\neq$ Beweis, dass die Nullhypothese stimmt

5.3 Fehlerarten

$\alpha = 0.05$	H_0 wahr	H_0 falsch
H_0 keep	Spezifität $\rightarrow 1 - \alpha$	Fehler 2. Art $\rightarrow \beta$
H_0 discard	Fehler 1. Art $\rightarrow \leq \alpha$	Sensitivität / Macht $\rightarrow 1 - \beta$

Fehler 1. Art: Nullhypothese wird verworfen, obwohl sie stimmt

Fehler 2. Art: Nullhypothese wird nicht verworfen, obwohl sie nicht stimmt

Wenn der eine Fehler kleiner wird, wird der andere Fehler grösser $\rightarrow$ Kompromiss finden

Macht ist die Wahrscheinlichkeit, dass die Nullhypothese verworfen wird, wenn sie in Realität nicht stimmt, sondern eine Alternativhypothese richtig ist. Wenn alle anderen Parameter gleichbleiben, kann die Macht mit einer Erhöhung der Stichprobengrösse verbessert werden.

Ziel: Macht von einem Test sollte möglichst gross sein

6 Zweiseitiger Binomialtest

6.1 P-Wert

Der P-Wert gibt an, wie wahrscheinlich die Ergebnisse der Stichprobe (und extremere Ergebnisse) sind, unter der Annahme, dass die Nullhypothese stimmt.

$$P\text{-Wert} \leq \alpha \rightarrow H_0 \text{ wird verworfen}$$

Ist auch eine Angabe der Signifikanz für die Alternativhypothese. Desto kleiner der P-Wert, desto signifikanter ist der Unterschied.

6.2 Vertrauensintervall

Dieses Intervall enthält den wahren π -Wert mit 95% Wahrscheinlichkeit. Falls unser geschätzter π_0 -Wert sich nicht im Intervall befindet, kann die Nullhypothese verworfen werden.

$$95\% \text{ Vertrauensintervall: } \frac{x}{n} \pm 1.96 * \sqrt{\frac{x}{n} * \left(1 - \frac{x}{n}\right) * \frac{1}{n}}$$

$$95\% \text{ VI} \rightarrow \alpha = 0.05$$

$\sqrt{n}$ – Gesetz: für n – mal so viele Beobachtungen, wird das VI um $\sqrt{n}$ kleiner

Bsp. Wenn n viermal so gross wird, wird das VI halb so gross und weist eine höhere Genauigkeit aus.

6.3 Verwerfungsbereich

$$H_A: \pi \neq \pi_0 \text{ (beidseitig)} \rightarrow K = [0, c_u] \cup [c_o, n]$$

$$H_A: \pi > \pi_0 \text{ (rechtsseitig)} \rightarrow K = [c_r, n]$$

$$H_A: \pi < \pi_0 \text{ (linksseitig)} \rightarrow K = [0, c_l]$$

$$P[T \leq c_u] \leq \frac{\alpha}{2} \quad \text{und} \quad P[T \geq c_o] \leq \frac{\alpha}{2}$$

Normalapproximation

gute Approximation: n gross & $\pi \approx 0.5$

$$\text{beidseitig: } c_{u,o} = n\pi_0 \pm 1.96\sqrt{n\pi_0(1-\pi_0)} \quad (u: \text{abrunden}, o: \text{aufrunden})$$

$$\text{rechtsseitig: } c_r = n\pi_0 + 1.64\sqrt{n\pi_0(1-\pi_0)} \quad (\text{aufrunden})$$

$$\text{linksseitig: } c_l = n\pi_0 - 1.64\sqrt{n\pi_0(1-\pi_0)} \quad (\text{abrunden})$$

Verwerfungsbereich $\neq$ Vertrauensintervall

7 Deskriptive Statistik

7.1 Kennzahlen

$$\text{Arithmetisches Mittel: } \bar{x} = \frac{1}{n} * \sum x_i$$

$$\text{Empirische Standardabweichung: } s_X = \sqrt{\text{Var}} = \sqrt{\frac{1}{n-1} * \sum (x_i - \bar{x})^2}$$

7.2 Quantile q_a

$a * 100\%$ der Datenpunkte sind kleiner als q_a

- Daten aufsteigend sortieren
- Fall 1: $a * n \rightarrow$ ganze Zahl $\rightarrow \frac{1}{2}(x_{a*n} + x_{a*n+1})$
- Fall 2: $a * n \rightarrow$ keine ganze Zahl $\rightarrow x_{a*n+\frac{1}{2}}$ (auf ganze Zahl runden)

$$\text{Empirischer Median} = q_{0.5}$$

$$q_{0.25} \rightarrow 1. \text{ Quartil} \quad q_{0.75} \rightarrow 3. \text{ Quartil}$$

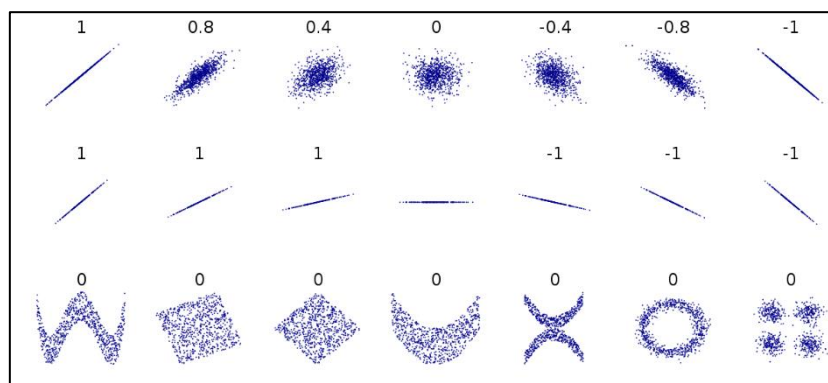
$$\text{Interquartilrange } IQR = q_{0.75} - q_{0.25}$$

7.3 Empirische Korrelation

Die Korrelation ist eine dimensionslose normierte Zahl zwischen -1 und 1. Eine Korrelation von null heisst, dass die zwei Datensätze unabhängig voneinander sind. Das Vorzeichen sagt aus, ob eine positive oder negative Korrelation besteht.

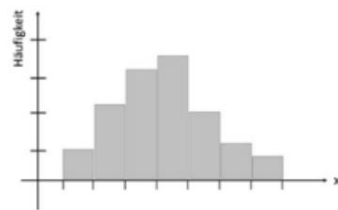
$$r = \hat{\rho} = \frac{s_{XY}}{s_X * s_Y}$$

$$s_{XY} = \frac{1}{n-1} \sum (x_i - \bar{x}) * (y_i - \bar{y})$$

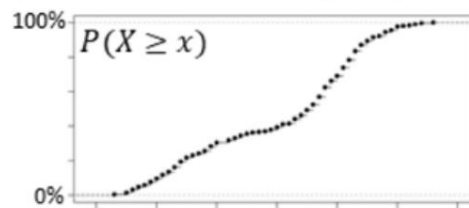


7.4 Graphische Darstellungsweisen

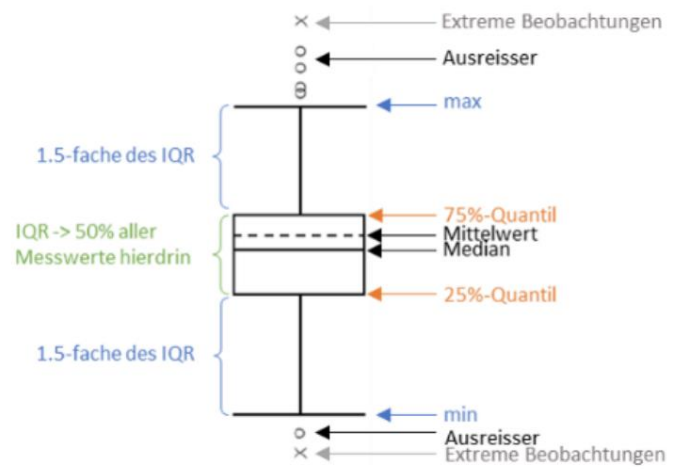
[1] Histogramm



[2] Kumulative Verteilungsfkt. (ECDF)



[3] Boxplot



8 Kontinuierliche Verteilungen & ZGS

8.1 Kontinuierliche Verteilungen

Allgemein

Die Zufallsvariable X ist stetig und kann kontinuierliche Werte annehmen. Weil es dann «unendliche» einzelne Ereignisse gibt, ist die Punktwahrscheinlichkeit = 0.

$$P[X = x] = 0$$

Uniforme Verteilung

$$P[X \leq x] = \frac{x - a}{b - a} \quad (a \leq x \leq b)$$

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

Exponentialverteilung

$$P[X \leq x] = 1 - e^{-\lambda x} \quad (\lambda \text{ ist Zeit zwischen Ereignisse})$$

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

8.2 Normalverteilung

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

$$\mu \rightarrow \text{Erwartungswert} = E(X)$$

$$\sigma^2 \rightarrow \text{Varianz} = \text{Var}(X)$$

$$N(0,1) \rightarrow \text{Standard - Normalverteilung} \rightarrow Z \sim N(0,1)$$

Die P-Werte einer Standard-Normalverteilung sind tabelliert.

Wenn eine andere Normalverteilung gegeben ist, kann diese standardisiert werden, um die P-Werte aus der Tabelle ablesen zu können.

$$z = \frac{x - \mu}{\sigma} \sim N(0,1)$$

Die Normalverteilung ist symmetrisch $\rightarrow P(X \leq -x) = 1 - (X \leq x)$

8.3 Lineare Transformation

Wenn Y eine neue transformierte Zufallsvariable von X ist.

$$Y = a + b * X$$

$$E(Y) = a + b * E(X)$$

$$Var(Y) = b^2 * Var(X)$$

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

$$Quantil\ q_Y^a = a + b * q_X^a$$

Wenn mehrere X zur Transformation beitragen:

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

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

8.4 Zentraler Grenzwertsatz

Mittelwert einer beliebigen Verteilung nähert sich mit zunehmendem Stichprobenumfang der Normalverteilung an. Je grösser die Stichprobe ist, desto näher an einer Normalverteilung.

$$S_n \sim N(n * E(X_i), n * Var(X_i))$$

$$\mu = n * E(X_i)$$

$$\sigma^2 = n * \sigma_{x_i}^2$$

Meistens muss dann diese neue Normalverteilung standardisiert werden, damit P-Werte aus der Tabelle abgelesen werden können.

9 Ein-Stichprobe t-Test

9.1 Gesetz der grossen Zahlen

Mit steigender Stichprobenumfang nähert sich die relativ beobachtete Häufigkeit seinem Erwartungswert. Die Standardabweichung des arithmetischen Mittels ist abhängig von der Stichprobengrösse und der Standardabweichung einer einzelnen Beobachtung.

$$\sigma_{\bar{X}_n} = \frac{\sigma_x}{\sqrt{n}}$$

9.2 z-Test

Allgemein

σ_x ist bekannt $\rightarrow$ basiert auf der Normalverteilung

Werte für den Verwerfungsbereich / Vertrauensintervall müssen «umgekehrt» aus der Normalverteilungstabelle abgelesen werden.

Teststatistik Z

Annahme $\rightarrow \mu_0 = 0$ unter der Nullhypothese

μ ist das arithmetische Mittel der Differenz zum Vergleich

$\rightarrow$ wenn kein signifikanter Unterschied besteht, muss $\mu = 0$ sein

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma_{\bar{X}_n}} = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma_x} \rightarrow \sigma_{\bar{X}_n} = \frac{\sigma_x}{\sqrt{n}}$$

Verwerfungsbereich

$$H_A: \mu_A \neq \mu_0 \rightarrow \left(-\infty, -z_{1-\frac{\alpha}{2}}\right) \cup \left(z_{1-\frac{\alpha}{2}}, \infty\right)$$

$$H_A: \mu_A < \mu_0 \rightarrow (-\infty, -z_{1-\alpha})$$

$$H_A: \mu_A > \mu_0 \rightarrow (z_{1-\alpha}, \infty)$$

Vertrauensintervall

$$H_A: \mu_A \neq \mu_0 \rightarrow VI = \bar{X}_n \pm z_{1-\frac{\alpha}{2}} * \frac{\sigma_x}{\sqrt{n}}$$

9.3 t-Test

Allgemein

σ_x ist unbekannt $\rightarrow$ muss geschätzt werden

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

Werte für den Verwerfungsbereich / Vertrauensintervall müssen aus der t-Verteilungstabelle abgelesen werden.

Teststatistik T

Annahme $\rightarrow \mu_0 = 0$ unter der Nullhypothese

μ ist das arithmetische Mittel der Differenz zum Vergleich

$\rightarrow$ wenn kein signifikanter Unterschied besteht, muss $\mu = 0$ sein

$$T = \frac{\bar{X}_n - \mu_0}{\hat{\sigma}_{\bar{X}_n}} = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\hat{\sigma}_x} \rightarrow \hat{\sigma}_{\bar{X}_n} = \frac{\hat{\sigma}_x}{\sqrt{n}}$$

Verwerfungsbereich

degrees of freedom = $n - 1$

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

$$H_A: \mu_A < \mu_0 \rightarrow \left(-\infty, -t_{n-1; 1-\alpha}\right)$$

$$H_A: \mu_A > \mu_0 \rightarrow \left(t_{n-1; 1-\alpha}, \infty\right)$$

Vertrauensintervall

$$H_A: \mu_A \neq \mu_0 \rightarrow VI = \bar{X}_n \pm t_{n-1; 1-\frac{\alpha}{2}} * \frac{\hat{\sigma}_x}{\sqrt{n}}$$

10 Zwei-Stichproben t-Test

10.1 Gepaarte / ungepaarte Stichproben

gepaart → gleiche Stichprobengrösse ($n = m$)

UND

gepaart → klare Zuordnung möglich

Bsp. Bei mehreren Patienten den Puls messen vor/nach dem Medikament. (gepaart)

Bsp. Mehrere Temperaturmessungen auf zwei unterschiedlichen Bergen. (ungepaart)

Gepaart: Differenz der zwei Messungen nehmen und ein-Stichproben t-Test aus dem vorherigen Kapitel durchführen.

Ungepaart: Zwei Messungen können nicht zueinander geordnet werden, somit ist keine Differenz vorhanden und es muss der zwei-Stichproben t-Test durchgeführt werden.

10.2 Zwei Stichproben

Allgemein

Annahme: $\sigma_x^2 = \sigma_y^2$ (gleiche Varianz)

n = Stichprobengrösse der Variable X

m = Stichprobengrösse der Variable Y

Teststatistik T

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

$$S_{pool}^2 = \frac{1}{n + m - 2} ((n - 1) * \hat{\sigma}_x^2 + (m - 1) * \hat{\sigma}_y^2)$$

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

$$\hat{\sigma}_y^2 = \frac{1}{m - 1} \sum_{i=1}^m (Y_i - \bar{Y}_m)^2$$

Verwerfungsbereich

degrees of freedom = $n + m - 2$

$$H_A: \mu_A \neq \mu_0 \rightarrow \left(-\infty, -t_{n+m-2; 1-\frac{\alpha}{2}}\right) \cup \left(t_{n+m-2; 1-\frac{\alpha}{2}}, \infty\right)$$

$$H_A: \mu_A < \mu_0 \rightarrow \left(-\infty, -t_{n+m-2; 1-\alpha}\right)$$

$$H_A: \mu_A > \mu_0 \rightarrow \left(t_{n+m-2; 1-\alpha}, \infty\right)$$

11 Lineare Regression

11.1 Parameterschätzungen

Allgemein

$$y = \beta_0 + \beta_1 * x$$

Diese Funktion kann gebraucht werden, um Prognosen zu machen.

t-Test

Ziel: Ist β_0 bzw. β_1 signifikant unterschiedlich von 0?

$$T = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$$

$$\text{Verwerfungsbereich} \rightarrow \left(-\infty, -t_{n-2; 1-\frac{\alpha}{2}}\right) \cup \left(t_{n-2; 1-\frac{\alpha}{2}}, \infty\right)$$

$$\text{Vertrauensintervall} \rightarrow \hat{\beta}_i \pm \hat{\sigma}_{\hat{\beta}_i} * t_{n-2; 1-\frac{\alpha}{2}}$$

R-Output

```
Call:
lm(formula = happiness ~ income, data = income.data)

Residuals:
    Min       1Q   Median       3Q      Max
-2.02479 -0.48526  0.04078  0.45898  2.37805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20427    0.08884   2.299   0.0219 *
income       0.71383    0.01854  38.505  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7181 on 496 degrees of freedom
Multiple R-squared:  0.7493, Adjusted R-squared:  0.7488
F-statistic: 1483 on 1 and 496 DF, p-value: < 2.2e-16
```

$$y = \beta_0 + \beta_1 * x$$

$$T = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$$

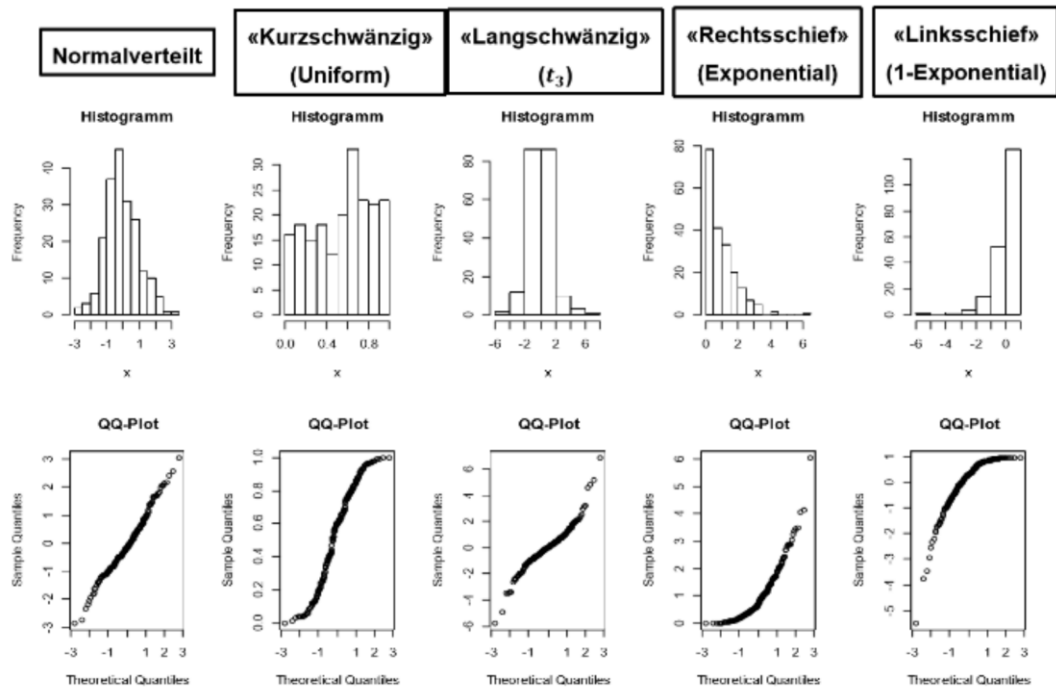
$$\text{degrees of freedom} = n - 2$$

$$\text{Approximatives 95\% - VI} \rightarrow \hat{\beta}_i \pm 2 * \hat{\sigma}_{\hat{\beta}_i}$$

Exaktes 95% - VI $\rightarrow$ statt 2 den Wert von $t_{df; 0.975}$ einsetzen

11.2 Plots

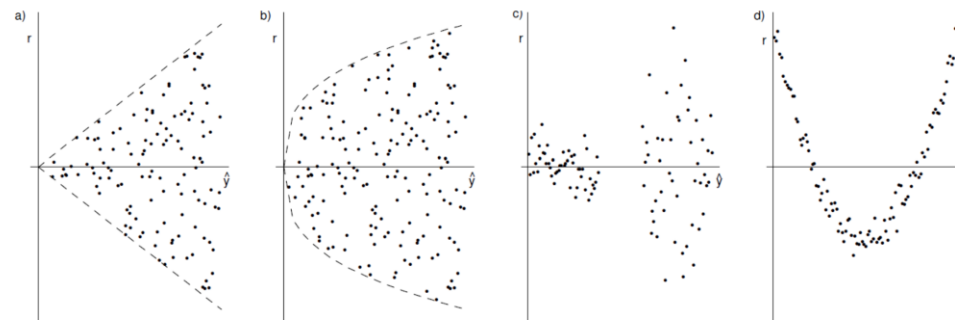
QQ-Plot



14

Tukey-Anscombe Plot

schlecht



gut

