Kombinatorik - Wiederholung

abzählbare diskrete Mengen und Tupel (Vektoren)

1. $A \cap B = \emptyset \Rightarrow |A \cup B| = |A| + |B|$
2. $|A \times B| = |A| \cdot |B|$
3. $\exists f : A \mapsto B$ (bijektiv) $\Rightarrow |A| = |B|$

Einstiegsbeispiele:

Binärzahlen entsprechen bijektiv n Mengen $B = \{0,1\}$ mit $|B| = 2$

Dadurch: $\tilde{B} = \underbrace{B \times B \times \ldots \times B}_{n \text{ Mal}} = |B|^n = 2^n$

Jede Potenzmenge $P(A)$ lässt sich auch als Binärzahl ausdrücken mit

$|A| = n$ Stellen

wenn Element enthalten : true / 1

wenn nicht : false / 0

Dadurch: $|P(A)| = 2^{|A|}$

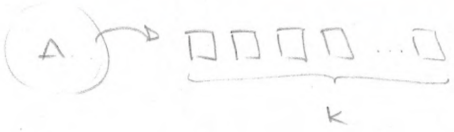
Es gibt 2 Problemarten:

- Auswahlprobleme (Mengen ◯, Reihenfolge unwichtig)
- Anordnungsprobleme (Tupel (□□□□), Reihenfolge wichtig)

Grundmenge

- alle Elemente $k = n$
  - Permutation
    Umordnung
    - ohne WH: $n!$
    - **mit WH**: $\frac{k!}{m_1! \, m_2! \ldots m_n!}$ ↑ Nicht alle Elemente unterscheidbar
- Stichprobe $k \leq n$
  - Variation
    Anordnung
    - ohne WH: $\frac{n!}{(n-k)!} = \binom{n}{k} \cdot k!$
    - mit WH: n^k
  - Kombination
    Auswahl
    - ohne WH: $\binom{n}{k}$
    - mit WH: $\binom{n+k-1}{k}$

Variation mit Wiederholung / mit zurücklegen



$$|A|^k = n^k$$

Beispiel.

$A = \{1, 2, 3\}$

$|A| = n = 3$ $\qquad n^k = 3^{12}$

$k = 12$

Variation ohne Wiederholung / ohne zurücklegen

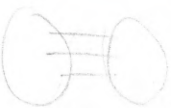
$$n(n-1)(n-2)\ldots(n-k+1) = \frac{n!}{(n-k)!}$$

Beispiel:

3 unterschiedliche Buchstaben aus dem Alphabet

$$26 \cdot 25 \cdot 24 = \frac{26!}{(26-3)!}$$

Permutationen einer Menge



A $\qquad$ A

a_n $\qquad$ $\pi(a_n)$

$$n!$$

Beispiel:

$A = \{B, W, S\}$ $\quad |A| = 3 \quad 3!$

Permutationen einer Multimenge

Weiterhin eine Umordnung, aber Elemente sind nicht alle unterscheidbar

$$\frac{(k_1 + k_2 + k_3 \dots + k_n)!}{k_1!\, k_2!\, k_3 \dots k_n!}$$

← Mit jedem Zug 1 Element weniger

← Weil nicht unterscheidbar, wegen Produktregel Wiederholungen ausschließen

Beispiel:

$A = \{\underbrace{B, B}_{\text{nicht unterscheidbar}}, S, W\}$ → # der nicht unterscheidbaren Anordnungen

$B_1 B_2$

$B_2 B_1$ $\quad 2! = 2$

Beispiel:

$A = \{W, W, S, B, B, B\}$

$$\frac{(2+1+3)!}{2!\, 1!\, 3!}$$

$$\frac{4!}{2!}$$ → wichtig! Die 2 Bier werden nicht wie ein Block behandelt

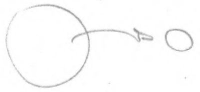
$\boxed{B\,B}\ \boxed{S}\ \boxed{W} = 3!$

denn:

B S W B

S B W B

Kombination ohne Wiederholung



alle möglichen Teilmengen einer Supermenge

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

Analogie : Permutation einer Multimenge

$$A = \{a_1, a_2, a_3, \dots, a_n\}$$

keine Reihenzahl, Anordnung ist egal $\longleftarrow$ $T = \{1, 0, 0, 1, \dots, 0\}$

Die # der Einsen wird vorgegeben.

k Einsen

n-k Nullen

Beispiel:

$|A| = 45$

$|B| = 6$

Es werden 6 Kugeln „auf einmal" gezogen:

$$\frac{45!}{6!\,(45-6)!}$$

Kombination mit Wiederholung

Jedes Element aus der Auswahlmenge darf in der Submenge beliebig oft vorkommen

$A = \{ a_1, a_2, a_3, \dots a_n \}$

$T = \{ 1, 0, 3, \quad 6 \}$ → Vorsicht: $|T|$ = # der untersch. Elementarten

$k = \sum$ der Elemente in T

Analogie:

beliebige Teilmenge $k = 7$

aus 4 verschiedenen Elementen $n = 4$

a_1 a_1 a_1 a_2 a_2 a_4 a_4

Wir setzen alle auf weiß und setzen schwarze Trennkugeln

kein a_3

(Wir haben $k + (n-1)$ Kugeln insgesamt in unserem Tupel (Anordnung wichtig)

Wir haben immer 3 schwarze und 7 weiße nicht unterscheidbare Kugeln

↓

Permutationen einer Multimenge:

$n' = k + (n-1)$

$k' = (n-1)$

$$\frac{n+k-1}{[(n+k-1)-(k-1)]!\,(n-1)!} =$$

$$= \binom{n+k-1}{k}$$

Beispiel:

$k = 7$ → mögliche Kompositionen (Summen)

$3+3+1$ oder $1+1+1+1+3+1$

→ Summe bleibt gleich!

Elementare Wahrscheinlichkeitsrechnung:

Ω ... Ereignisraum

zB bei Würfel: $\Omega = \{1, 2, 3, 4, 5, 6\}$

$A \subseteq \Omega$ ... Ereignis besteht aus Elementarereignissen

Laplace - Experiment

- Zufallsexperiment hat endlich großen Ereignisraum
- jedes Elementar-Ereignis ist gleich wahrscheinlich

$P(A) = \frac{k}{n} = \frac{\text{günstige Fälle}}{\text{mögliche Fälle}}$ $\longrightarrow$ klassische Laplace'sche Wahrscheinlichkeitsdefinition

$P(A) = \lim_{n \to \infty} h_n(A)$ $\longrightarrow$ relative Häufigkeit als statistische Wahrscheinlichkeit

Es muss gelten: Wahrscheinlichkeitsaxiome

$0 \leq P(A) \leq 1$

$P(\Omega) = 1$

$P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \{\}$

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

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

$P(A) \leq P(B)$ wenn $A \subseteq B$

Bedingte Wahrscheinlichkeit

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

Wahrschein des Ereignisses B unter der Bedingung, dass A auch eintritt

$\forall B \in A$

Beispiel

2 faire Würfel werden geworfen

gesucht ist die WSI dass die Augensumme 5 geworfen wird, unter der Bedingung, dass wenigstens 1x die 1 dabei ist:

$|A|$ = "zumindest 1 x die Eins" $= 1 \cdot 6 + 6 \cdot 1 - \underset{(1,1)}{1} = 11$

$|B|$ = "Augensumme 5" $= 4$

$P(A) = \frac{11}{36}$

$P(B) = \frac{4}{36}$ $\longrightarrow$ $P(B|A) = \frac{2}{36} / \frac{11}{36} = \frac{2}{11}$

$P(A \cap B) = \frac{2}{36}$

$$P(A \cap B) = P(A) \cdot P(B|A) = P(B)\, P(A|B)$$

Beispiel

5 alte, 10 neue Batterien (von außen nicht unterscheidbar)

WSI für 2 Batterien hintereinander die neu sind

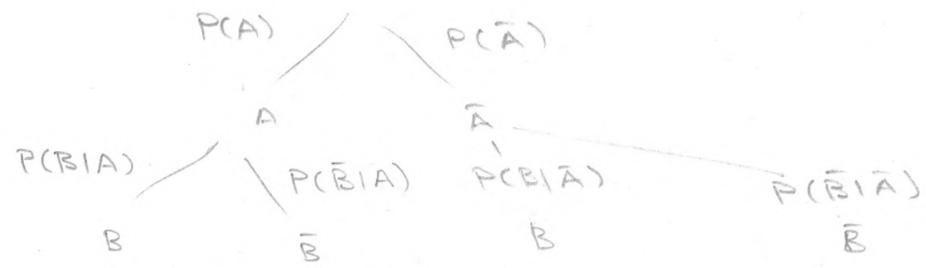
$P(A$ = erste Batterie ist neu$) = \frac{10}{15}$

$P(B$ = zweite Batter ist neu$) = ?$ abhängig davon ob zuerst eine neue genommen wurde

$P(B|A) = \frac{9}{14}$

$P(A \cap B) = P(A) \cdot P(B|A) = \frac{10}{15} \cdot \frac{9}{14} = 0{,}43$

Wahrscheinlichkeitsbaum für abhängige Ereignisse



Axiome für unabhängige Ergebnisse

$P(B|A) = P(B)$ $\qquad P(A) > 0$

$P(A|B) = P(A)$ $\qquad P(B) > 0$

$P(A \cap B) = P(A) \cdot P(B)$

Formel für die totale Wahrscheinlichkeit

beliebige Ereignispartition E_i wobei $\Omega = E_1 \cup \ldots \cup E_n$
und $E_j \cap E_k = \emptyset$ und $A \subseteq \Omega$

$$P(A) = \sum_{k=1}^{n} P(A \cap E_k) =$$

$$= \sum_{k=1}^{n} P(E_k) P(A|E_k)$$

Für ein Ereignis muss man manchmal Ereignispartitionen (wenn abhängig, dann sind das die Äste des Baumes) addieren.

Formel von Bayes

$$P(E_j|A) = \frac{P(E_j) P(A|E_j)}{P(A)} = \frac{P(E_j) P(A|E_j)}{\sum_{k=1}^{n} P(E_k) P(A|E_k)}$$

Vereinfacht: Wenn es zB nur 2 Ereignisse gibt (Binärbaum):

$$P(E|A) = \frac{P(E) P(A|E)}{P(E) P(A|E) + P(\bar{E}) P(A|\bar{E})}$$

Elementare Wahrscheinlichkeitsrechnung

Ω Ergebnis / Ereignis - Menge (Menge aller Elementarereignisse)

$A \subseteq \Omega$ Ereignis

$A^c = \Omega / A$ Gegenereignis

$A \cap B$ A und B

$A \cup B$ A oder B

Laplace Experiment

In einem Laplace Experiment gilt: $P(A) = \frac{k}{n} = \frac{\text{günstige Fälle}}{\text{mögliche Fälle}}$

Wahrscheinlichkeit $P(A)$

$0 \leq P(A) \leq 1$

$P(\Omega) = 1$

$P(A \cup B) = P(A) + P(B)$ wenn $A \cap B = \{\}$ (unvereinbar)

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ wenn $A \cap B \neq \{\}$ (vereinbar)

$P(A) \leq P(B)$ wenn $A \subseteq B$

> Für allen Elementarereignissen gilt:

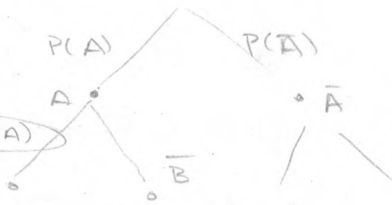
$$\sum_{i=1}^{n} P(\omega_i) = 1$$

$$P(A) = \sum_{\omega_i \in A} P(\omega_i)$$

Bedingte Wahrscheinlichkeit

B unter der Bedingung, dass A eintritt

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \longrightarrow P(B \mid A)$$



Multiplikationssatz

$$P(A \cap B) = P(A)\,P(B \mid A) = P(B) \cdot P(A \mid B)$$

(Umformung des vorherigen Satzes)

Unabhängigkeit, Multiplikationssatz

wenn $P(A) > 0$ und $P(B) > 0$

$$P(B \mid A) = P(B)$$
$$P(A \mid B) = P(A)$$
$$P(A \cap B) = P(A) \cdot P(B)$$

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

→ Beispiel für Abhängigkeit

Formel für totale Wahrscheinlichkeit

Partitionen $E_1 \dots E_n$ von Ω

$$P(A) = \sum_{k=1}^{n} P(A \cap E_k) =$$

$$= \sum_{k=1}^{n} P(E_k) \cdot P(A \mid E_k)$$

→ Alle Paths im Baum zu A

Formel von Bayes

$$P(E_j \mid A) = \frac{P(E_j)\,P(A \mid E_j)}{P(A)} =$$

$$= \frac{P(E_j)\,P(A \mid E_j)}{\sum_{k=1}^{n} P(E_k) \cdot P(A \mid E_k)}$$

→ wenn es nur 2 Ereignisse gibt

$$P(E \mid A) = \frac{P(E)\,P(A \mid E)}{P(E)\,P(A \mid E) + P(\bar{E})\,P(A \mid \bar{E})}$$

Zufallsvariablen

Zufallsvariable $X : \Omega \mapsto \mathbb{R}$

$$\omega \in \Omega \mapsto X(\omega)$$

Ordnet jedem Elementarereignis, eine reelle Zahl zu.

ZB: Augenzahl von 2 Würfeln

$$\Omega = \{ (i,j) \mid i,j = 1, \dots, 6 \}$$

$$P(\{(i,j)\}) = \frac{1}{36}$$

$$X(i,j) = i + j \in \{2, \dots, 12\} = X$$

die zugeordnete Zahl zu einem Ereignis

$$X(1,1) = 2$$
$$X(2,4) = 6$$

Realisation / Realisierung kann disket oder stetig sein

Wahrscheinlichkeitsverteilung

Der Auftritt jeder Realisation hat eine Wahrscheinlichkeit.

$$X = x_i$$
$$\{\omega \mid X(\omega) = x_i\}$$

$$P(X = x_i) = p_i$$

p_i heißt „Wahrscheinlichkeitsverteilung der Zufallszahlen"

Verteilungsfunktion von X

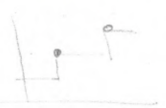
$$F(x) = P(X \leq x) \qquad x \in \mathbb{R}$$

$$F(x) = \sum_{i = x_i \leq x} p_i$$

$$1 - F(x) = P(X > x)$$

Eigenschaften wenn diskret (Treppenfunktion:

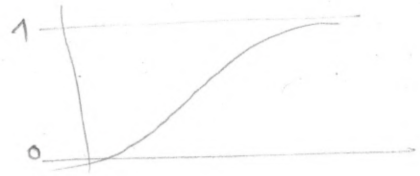
- Wächst monoton von 0 bis 1
- rechtsseitig stetig (an Sprungstelle zählt der obere Wert)



} Sprunghöhe

- Sprunghöhe ist genau die Eintrittswahrsch. an jenem Punkt, dass hinzugefügt wird.

Stetige Verteilungsfunktion von X



$F(x) = P(X \leq x)$

ist stetig wenn Zufallsvariable / Realisierung davon stetig ist.

$f(x) = F'(x)$ wenn diferenzierbar

$f(x)$ ist die Wahrscheinlichkeitsdichtefunktion!

Wahrscheinlichkeits-Dichtefunktion

$$\underbrace{F(x)}_{\text{Verteilungsfunktion}} = \int_{-\infty}^{x} \underbrace{f(t)}_{\text{Wahrscheinlichkeits-Dichtefunktion}} \, dt = P(X \leq x)$$

↓

Quasi das Äquivalent von p_i in

$$P(X \leq x) = \sum_{i : x_i \leq x} p_i$$

Quasi Wahrscheinlichkeit an einem konkreten Punkt.

Eigenschaften:

- nicht negativ
- $\int_{-\infty}^{\infty} f(x) \, dx = 1$

Das Integral an einem best. Punkt ist 0,
Zufallsvariable X ist stetig wenn

$$\underbrace{P(X = x)}_{\text{mit Integral definiert}} = 0 \quad \forall x \in \mathbb{R}$$

Daher:

$$P(X < x) = P(X \leq x)$$
$$P(X > x) = P(X \geq x)$$

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(t) \, dt$$

Beispiel

Straßenbahn fährt alle 10 Minuten

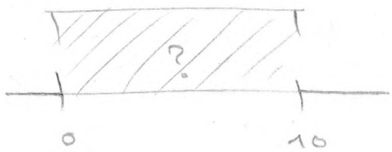
$$f(x) = \begin{cases} k, & \text{für } 0 < x < 10 \\ 0, & \text{sonst} \end{cases}$$

$\longrightarrow$ Wartezeit X ist eine stetige Zufallsvariable zwischen 0 und 10.

Wobei k eine Konstante ist.
(Wahrscheinlichkeitsdichte)

1) Bestimmen Sie k:

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{10} k\,dx = 10k$$



$$1 = 10k$$

$$\frac{1}{10} = 0{,}1 = k$$

$$\rightarrow f(x) = \begin{cases} 0{,}1, & \text{für } 0 < x < 10 \\ 0, & \text{sonst} \end{cases}$$

2) Geben Sie die Verteilungsfunktion F an

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{0}^{x} 0{,}1\,dt = 0{,}1t \Big|_{0}^{x} = 0{,}1x - 0 = 0{,}1x$$

Für $x < 0$ ist $F(x)$ auch 0
Zwischen 0 bis 10 gilt aber

Für $x > 10$ gilt: $F(10) + F(x) = 1 + 0 = 1$

Also:

$$F(x) = \begin{cases} 0 & \text{für } x \leq 0 \\ 0{,}1x & \text{für } 0 \leq x < 10 \\ 1 & \text{für } x \geq 10 \end{cases}$$

3) Wie groß ist die Wahrscheinlichkeit höchstens 3 Min zu warten?

$$F(x) = 0{,}1x$$

$$P(X \leq 3) = F(3) = 0{,}3 = 30\,\%$$

4) Wahrscheinlichkeit mindestens 2 min zu warten?

$$F(x) = 0{,}1\,x$$

$$1 - F(2) = P(X \geq 2)$$

$$1 - 0{,}2 = 0{,}8 = 80\,\%$$

5) Wahrscheinlichkeit zwischen 5 und 9 min zu warten?

$$P(5 < x < 9) = F(9) - F(5) = 0{,}1 \cdot 9 - 0{,}1 \cdot 5 = 0{,}4 = 40\,\%$$

Die Wahrscheinlichkeit, dass eine stetige Zufallsvariable einen Wert in einem kleinen Intervall annimmt:

$$P(a - \Delta a \leq X \leq a + \Delta a) \approx f(a) \cdot 2\Delta a$$

$$F(a + \Delta a) - F(a - \Delta a) = \int_{a-\Delta a}^{a+\Delta a} f(t)\,dt \approx \int_{a-\Delta a}^{a+\Delta a} f(a)\,dt =$$

$$f(a) \int_{a-\Delta a}^{a+\Delta a} 1\,dt = 2\,f(a)\,\Delta a$$

p - Quantil

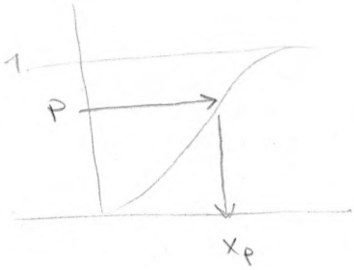
Gegeben sei eine diskrete / stetige Zufallsvariable X

$p \in (0,1)$

$x_p \in \mathbb{R}$

$$F(x_p) = p$$

$$F^{-1}(p) = x_p$$



Unabhängigkeit

Unabhängige Zufallsvariablen X,Y wenn die Ereignisse $X \in A$, $Y \in B$:

$$P((X \in A) \cap (Y \in B)) = P(X \in A) \cdot P(Y \in B)$$

$$P((X \leq x) \cap (Y \leq y)) = P(X \leq x) \cdot P(Y \leq y)$$

Für diskrete gilt zusätzlich noch:

$$P((X = x_i) \cap (Y = y_i)) = P(X = x_i) \cdot P(Y = y_i)$$

Erwartungswert und Varianz

Erwartungswert

$$\mu = E(X) = \begin{cases} \sum_i x_i p_i & \text{falls } X \text{ diskret ist} \\ \int_{-\infty}^{\infty} x f(x)\, dx & \text{falls } X \text{ stetig ist} \end{cases}$$

Angenommen Dichtefunktion ist symmetrisch um Wert c:

$$f(c-x) = f(c+x)$$

dann ist

$$E(X) = c$$

Weiters gilt:

$$E(X+Y) = E(X) + E(Y)$$

$$E(aX) = a\,E(X)$$

$$E(aX+b) = a\,E(X) + b$$

$$E(g(X)) = \begin{cases} \sum_i g(x_i) p_i & \text{falls diskret} \\ \int_{-\infty}^{\infty} g(x) f(x)\, dx & \text{falls stetig} \end{cases}$$

Beispiel

$$E(X \cdot Y) = E(X) \cdot E(Y) \quad \text{falls unabhängig}$$

Beispiel:

X = Anzahl der Störfälle in der Produktion pro Tag

x_i	0	1	2	3	4
p_i	0,3	0,4	0,2	0,08	0,02

Die Behebung der Störung kostet:

$$g(x) = 6 - \frac{5}{1+x} \quad (\text{in } 1000\,€)$$

1) Berechne die Kosten die im Schnitt täglich zu erwarten sind!

$g(x_i)$	1,00	3,50	4,33	4,75	5,00
p_i	0,3	0,4	0,2	0,08	0,02

$$E(g(x)) = 3{,}05 = \sum_i g(x_i) \cdot p_i$$

2) Berechne die Kosten des Erwartungswertes

$$g(E(X)) = 3{,}64 = g(1{,}12)$$

Varianz

$$\sigma^2 = Var(X) = E((X - \overbrace{E(X)}^{\mu})^2) = \begin{cases} \sum_i (x_i - E(x))^2 \, p_i & \text{diskret} \\ \int_{-\infty}^{\infty} (x - E(x))^2 f(x)\,dx & \text{stetig} \end{cases}$$

Quadratische Abweichung vom Erwartungswert

Varianz = 0 wenn X nur Erwartungswert einnehmen kann: „entartete Zufallsvariablen"

Standardabweichung

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

weiters:

$$\sigma^2 = E(X^2) - \underbrace{E(X)^2}_{\mu^2}$$

wichtig:

$$E(X^2) \neq E(X)^2 = \mu^2$$

Das k-te Moment

$m_k(X) = E(X^k)$

$m_0(X) = E(X^0) = E(1) = 1$

$m_1(X) = E(X)$

$m_2(X) = E(X^2) = \sigma^2 + \mu^2$

Varianz Erwart.^{2}

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

$\mu^2 = E(X)^2$

Das k-te zentrale Moment

$m_k(X-\mu) = E((X-\mu)^k)$

$m_0(X-\mu) = E((X-\mu)^0) = 1$

$m_1(X-\mu) = E(X-\mu)^1 = 0$

$m_2(X-\mu) = E((X-\mu)^2) = \sigma^2$

Varianz

Daraus ergibt sich:

wenn wir statt X schreiben „aX + b":

$Var(X) = E((X-\mu)^2)$

$Var(aX+b) = E(((aX+b)-\mu)^2)$

$\Delta = a^2$

Änderung der Varianz um a^2

Satz:

$Y = aX + b$

$Var(Y) = a^2 \, Var(X)$

$\sigma_Y = |a| \sigma_X$

→ Im Erwartungswert aber

$E(aX+b) = a\,E(X) + \underbrace{E(b)}_{=b}$

Beweis:

$Var(aX+b) = E((aX+\cancel{b} - a\mu - \cancel{b})^2) = E((aX-a\mu)^2) =$

$E(a^2(X-\mu)^2) = a^2 E((X-\mu)^2) = a^2 \, Var(X)$

Standardisierte Zufallsvariable

$$Z = \frac{X - \mu}{\sigma} \longrightarrow \begin{array}{l} E(Z) = 0 \\ Var(Z) = 1 \end{array}$$

Zusammenhang zwischen Verteilungsfunktionen:

$$F_X(x) = F_Z\left(\frac{x-\mu}{\sigma}\right) \quad \text{bzw} \quad F_Z(z) = F_X(\sigma z + \mu)$$

Für die p Quantille gilt:

$$x_p = \sigma z_p + \mu$$

Für die Dichtefunktion gilt

$$f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right)$$

Satz

$$Var(X+Y) = Var(X) + 2\underbrace{Cov(X,Y)}_{} + Var(Y)$$

$$cov(X,Y) = E((X-\mu_x)(Y-\mu_y)) = E(XY) - \mu_x\mu_y$$

Wenn sie unabhängig sind gilt denn aber:

$$Cov(X,Y) = 0$$

$$Var(X+Y) = Var(X) + Var(Y)$$

Es gilt weiter

$$Cov(X,Y)^2 \leq Var(X)\, Var(Y)$$

Kovarianz

Es gilt:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \underbrace{2\,\mathrm{Cov}(X,Y)} + \mathrm{Var}(Y)$$

$$\mathrm{Cov}(X,Y) = E\left((X-\mu_X)(Y-\mu_Y)\right) =$$
$$= E(X \cdot Y) - \mu_X \mu_Y$$

Für $\mathrm{Cov}(X,Y)$ gilt:

$$\mathrm{Cov}(X,Y) \leq \mathrm{Var}(X) \cdot \mathrm{Var}(Y)$$

(Gleichheit wenn $Y = aX + b$)

Wenn X und Y unabhängig sind:

$$\mathrm{Cov}(X,Y) = 0$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

> Vorsichtig: aus $\mathrm{Cov}(X,Y) = 0$ folgt nicht, dass X und Y unabhängig sind.

Wenn $\mathrm{Cov}(X,Y) = 0$ ⟶ X, Y sind unkorreliert

$\mathrm{Cov}(X,Y) > 0$ X, Y sind positiv korreliert

$\mathrm{Cov}(X,Y) < 0$ X, Y sind negativ korreliert

Korrelations-Koeffizient

„Kovarianz der standardisierten Zufallsvariablen"

$$\rho_{XY} = \mathrm{Cov}\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Ungleichung von Tschebyscheff

$E(X) = \mu$

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

beliebiges $c > 0$

$$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

$$P(|X-\mu| < c) \geq 1 - \frac{\sigma^2}{c^2}$$

Verteilungen die häufig vorkommen:

Gleichverteilung / Rechteckverteilung

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

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

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

Exponentialverteilung

$$F(x) = \begin{cases} 1 - e^{-kx} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$$

$$f(x) = \begin{cases} k\, e^{-kx} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$$

$$\mu = \frac{1}{k} \qquad \sigma^2 = \frac{1}{k^2}$$

Das Gesetz der großen Zahlen

Das n-fache unabhängige Wiederholen aus dem:

Produktraum Ω^n

Produkt-Wahrscheinlichkeiten $P((X_1 \in A_1) \cap \dots \cap (X_n \in A_n)) =$

$P(X_1 \in A_1) \dots P(X_n \in A_n)$

Arithmetisches Mittel:

$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ (Zufallsvariable)

$\forall X_i : E(X_i) = \mu \quad Var(X_i) = \sigma^2$

Es gilt:

$E(\bar{X}) = \mu$

$Var(\bar{X}) = \frac{\sigma^2}{n}$

Mögliche Wege X_i zu erhalten:

- Ziehen mit Zurücklegen aus endlicher Übermenge
- Ziehen ohne Zurücklegen aus unendlicher Übermenge / sehr großer
- Laplace-Zufalls-Experiment

  wobei μ und σ festgelegt sein müssen

Alle X_i haben Realisationen x_i.

Alle X_i sind unabhängige, identisch verteilte Zufallsvariablen.

Es gilt:

$\bar{X}$ konvergiert stochastisch für wachsenden Stichprobenumfang gegen μ.

$\lim_{n \to \infty} P(|\bar{X} - \mu| < \varepsilon) = 1 \qquad \varepsilon > 0$

Daraus folgt: Theorem von Bernoulli

$P(A) = p \rightarrow$ Experiment n-Mal wiederholt

$f_n \rightarrow$ relative Häufigkeit des Eintretens von A

$\varepsilon > 0$

$\lim_{n \to \infty} P(|f_n - p| \leq \varepsilon) = 1$

(f_n wird quasi zu p)

Hauptsatz der Statistik: $\overline{F}(x)$ konvergiert stochastisch gegen $F(x)$

Empirische Verteilungsfunktion: $\overline{F}(x)$

Verteilungsfunktion $F(x)$

$$\lim_{n \to \infty} P\left(|\overline{F}(x) - F(x)| < \varepsilon \right) = 1$$

Spezielle diskrete Verteilungen

(die oft vorkommen)

Hypergeometrische Verteilung

Ziehung ohne Zurücklegen

Übermenge aus N Elementen

Untermenge aus M Elementen

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

x kann die Werte 0 bis n annehmen

$x > M$

$n - x > N - M$

Kurzschreibweise: $X \sim H(n; M; N)$

Beispiel:

in einem Behälter: 20 Kugeln

4 Blau

16 Rot

ohne Zurücklegen 5 Kugeln nehmen

P(genau 2 Blaue Kugeln) =

$$\frac{\binom{4}{2}\binom{16}{3}}{\binom{20}{5}}$$

$$P(X=x) = \frac{\binom{4}{x}\binom{16}{5-x}}{\binom{20}{5}}$$

Es gilt:

$$\mu = E(X) = n\frac{M}{N}$$

$$\sigma^2 = Var(X) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}$$

Binomialverteilung

2 mögliche Ausgänge

- Erfolg / Treffer $P(A) = p$ Erfolgswahrsch.
- Misserfolg $P(\bar{A}) = p-1$

n Wiederholungen:

Bernoulli Kette der Länge n

X = Anzahl der Versuchsdurchführungen bis A eintritt

$x = 0, 1, 2, \dots, n$

$P(X=x) = \binom{n}{x} p^x q^{n-x}$

$X \sim Bi(n;p)$

$\mu = E(X) = np$

$\sigma^2 = Var(X) = npq = np(p-1)$

Symmetrie:

$X \sim Bi(n;p)$

$Y \sim Bi(n;p-1)$

daraus folgt:

$P(X=x) = P(Y=y)$

$y = n-x$

Addition:

$X \sim Bi(m;p)$

$Y \sim Bi(n;p)$

$X+Y \sim Bi(m+n;p)$

Approximation von der hypergeometrischen Verteilung durch die Binomialverteilung

wenn $n \lesssim \frac{N}{20}$

denn $p = \frac{M}{N}$

Poisson - Verteilung

$$P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad \lambda > 0$$

$X \sim Po(\lambda)$

$\mu = E(X) = \lambda$

$\sigma^2 = Var(X) = \lambda$

„Poisson - Prozess"
in einem bestimmten Zeitraum

falls $X_t \sim Po(t\lambda)$ $\longrightarrow$ $P(X_t = x) = \frac{(t\lambda)^x}{x!} e^{-t\lambda}$

Dann gilt:

$X_t - X_s \sim Po((t-s)\lambda)$

Satz:

Sei $X \sim Bi(n; p) \rightarrow \lambda = np$

$$\lim_{\substack{n \to \infty \\ p \to 0}} P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Falls $n \gtrapprox 50$ und $p \lessapprox 0{,}1$ dann kann eine Binomialverteilung mit $Po(\lambda = np)$ approximiert werden.

Spezielle stetige Verteilungen

Normalverteilung / Gaußverteilung

X ist normalverteilt wenn Dichtefunktion $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Gauss'sche Glockenkurve

$X \sim N(\mu; \sigma^2)$.

- Symmetrisch zu $x = \mu$ (Maximum)
- 2 Wendepunkte an $\mu \pm \sigma$

$F(x) = P(X \leq x) =$

$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$

kann nur numerisch berechnet werden. Deshalb liest man $F(x)$ oft aus Tabellen ab.

Weil man nicht für alle σ und μ Tabellen erstellen kann leitet man alles von der „Standard-Normal-Verteilung" ab.

Für die Momente der Stand. norm. Vert. gilt:

$$E(Z^n) = \begin{cases} 0 & \text{falls } n \text{ ung.} \\ 1 \cdot 3 \cdot 5 \ldots (n-1) & \text{falls } n \text{ ger.} \end{cases}$$

$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ $\quad \mu = 0$, $\sigma^2 = 1$

$\int$

$\Phi(z) = P(Z \leq z) =$

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$

Zusammenhang:

Sei $X \sim N(\mu; \sigma^2)$

$Y = aX + b$ wobei $Y \sim N(?;?)$

dann:

$\mu_Y = E(Y) = a\mu_x + b$

$\sigma^2_Y = Var(Y) = a^2 \sigma^2_x$

Es gilt:

$\Phi(-z) = 1 - \Phi(z)$

Deshalb tabelliert man nur positive z Werte.

Satz:

Sei $X \sim N(\mu; \sigma^2)$

$Z = \frac{X-\mu}{\sigma}$ ⟶ Zugehörige standardisierte Zufallsvariable

- $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ $f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$

- Für p Quantile gilt: $x_p = \sigma z_p + \mu$

Satz:

Sei $0 < p < 1$ z_p = Zugehöriges p Quantil der Standardnormalverteilung, dann gilt:

$z_p = -z_{1-p}$

Satz:

Intervalle um μ:

$[\mu - c; \mu + c]$

$P(\mu - c \leq X \leq \mu + c) = F(\mu + c) - F(\mu - c) = \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) =$

$= 2\,\Phi\left(\frac{c}{\sigma}\right) - 1$

c wird meistens als Vielfaches von σ angegeben: $c = k \cdot \sigma$ $k > 0$

„$k\sigma$"-Intervall

$\mu - \sigma$ bis $\mu + \sigma$: 68,3 %

$\mu - 2\sigma$ bis $\mu + 2\sigma$: 95,5 %

$\mu - 3\sigma$ bis $\mu + 3\sigma$: 99,7 %

⟶ Umgekehrt: Intervallbreite aus Prozenten bestimmen:

$c = \sigma \cdot z_{\frac{1+p}{2}}$

(Verwendet bei Konfidenz-Intervallen)

Normalverteilung als Näherung

$X \sim N(\mu_X ; \sigma_X^2)$

$Y \sim N(\mu_Y ; \sigma_Y^2)$

$X + Y \sim N(\mu_X + \mu_Y ; \sigma_X^2 + \sigma_Y^2)$ (wenn unabhängig)

Zentraler Grenzwert-Satz

Seien $X_1 \dots X_n$ unabhängig, identisch verteilt (nicht normalverteilung notwendig) mit μ und σ^2

Summe $S = X_1 + X_2 + \dots + X_n$

mit $n\mu$ und $n\sigma^2$

$$Z = \frac{S - n\mu}{\sqrt{n}\,\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$\bar{X} = \frac{1}{n} S$

$$\lim_{n \to \infty} P(Z \leq z) = \Phi(z)$$

für hinreichend großes n ist Z praktisch standard-normalverteilt

Alle Zufallsvariablen sind normalverteilt weil sie sich nur um eine lineare Transf. unterscheiden. (für $n \to \infty$)

man sagt:

$$\lim_{n \to \infty} S = X_1 + \dots + X_n \sim N(n\mu ; n\sigma^2)$$

$$\lim_{n \to \infty} \frac{1}{n} S \sim N\left(\mu ; \frac{\sigma^2}{n}\right)$$

$$\lim_{n \to \infty} Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0;1)$$

(eigentlich kann man das so nicht schreiben weil $\lim_{n \to \infty} S = \infty$)

↓

man schreibt

$S \overset{a}{\sim} N(n\mu ; n\sigma^2)$

approximativ

~~Zentraler Grenzwert-Satz~~

$\cancel{E(S) = n\mu}$

$\cancel{Var(S) = n\sigma^2}$

$$\cancel{Z = \frac{S - n\mu}{\sqrt{n}\,\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}$$

↓

$$\cancel{\lim_{n \to \infty} P(Z \leq z) = \Phi(z)}$$

~~für hinreichend großes n ist Z (von S) praktisch normalverteilt.~~

Aus dem zentralen Grenzwert-Satz folgt:

Binomialverteilung lässt sich mit Normalverteilung annähern

$X \sim Bi(n;p)$

falls n groß genug ist gilt:

$\mu = n \cdot p$

$\sigma = n \cdot p(1-p) \gtrapprox 9$ (faustregel)

„Stetigkeits-Korrektur“.

$$F_B(x) \approx F_N(x+0{,}5) = \Phi\left(\frac{x+0{,}5-np}{\sqrt{n\,p(1-p)}}\right)$$

Poisson-Verteilung —"—

$\mu = \lambda$

$\sigma^2 = \lambda$ wenn $\lambda \gtrapprox 9$

$$F_P(x) \approx F_N(x+0{,}5) = \Phi\left(\frac{x+0{,}5-\lambda}{\sqrt{\lambda}}\right)$$

Drei wichtige Prüfverteilungen

Chi-Quadrat-Verteilung
t-Verteilung
F-Verteilung

Eine Prüfgröße ist eine Zufallsvariable. Zum berechnen von Werten aus Stichproben

Definition

$Z_1, \dots, Z_m$ sind unabhängige standardnormalverteilte Zufallsvariablen

$X = Z_1^2 + \dots + Z_m^2$

→ Chi-Quadrat-Verteilung χ^2-Verteilung mit m „Freiheitsgraden“

$X \sim \chi^2(m)$

$E(X) = m$

$Var(X) = 2m$

Ab $m=3$ sind sie unimodal (genau 1 Maximum)

Nach dem Z. Grenzwert-Satz:

mit $\lim_{m \to \infty}$ nähert es sich der gauß'schen Glockenkurve an

Approximation durch Normalverteilung

$$\chi^2_{m;p} \approx m\left(1 - \frac{2}{9m} + z_p\sqrt{\frac{2}{9m}}\right)^3 \quad \text{für } m \gtrapprox 30$$

p Quantil der Standard-Verteilung

Satz:

Sei S^2 die Varianz,

$\bar{x}$ das arithmetische Mittel,

einer Stichprobe $X_1 \dots X_n$ vom Umfang n

$$\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} \longrightarrow$$ Eine χ^2-Verteilung mit $m = n-1$ Freiheitsgraden.

Addition von χ^2

$X_1 \sim \chi^2(m_1)$

$X_2 \sim \chi^2(m_2)$

$(X_1 + X_2) \sim \chi^2(m_1 + m_2)$

T-Verteilung (Student-Verteilung)

$$T = \frac{Z}{\sqrt{X/m}} \qquad T \sim t(m)$$

$E(T) = 0 \quad m > 1$

$Var(T) = \frac{m}{m-2} \quad m > 2$

Wenn m groß genug ist konvergiert die Dichtekurve gegen eine Standard-Normalverteilung

$$t_{m;p} \approx \sqrt{\frac{m}{m+1}} \cdot z_p$$

bessere Approximation:

$$t_{m;p} \approx z_p\left(1 + \frac{1 + z_p^2}{4m}\right) \quad m \gtrapprox 30$$

Satz:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$ $\bar{X}$ ... Arithmetisches Mittel von $X_1 \dots X_n$ aus normalverteilten Grundgesamtheit

T-Verteilung mit $m = n-1$ Freiheitsgraden

F - Verteilung

X_1 und X_2 sind χ^2 - verteilt

$X_1 \sim \chi^2(m_1)$

$X_2 \sim \chi^2(m_2)$

F - Verteilung / Fischer - Verteilung

$$X = \frac{X_1/m_1}{X_2/m_2}$$

$$X \sim F(m_1; m_2)$$

$$E(X) = \frac{m_2}{m_2 - 2} \qquad m_2 > 2$$

$$Var(X) = \frac{2m_2^2(m_1 + m_2 - 2)}{m_1(m_2-4)(m_2-2)^2} \qquad m_2 > 4$$

Es gilt für Quantillen $> 0{,}5$

$$F_{m_1; m_2; 1-p} = \frac{1}{F_{m_2; m_1; p}}$$

$$F(X \leq x) = F\left(\frac{1}{X} \geq \frac{1}{x}\right) = 1 - F\left(\frac{1}{X} \leq \frac{1}{x}\right)$$

Weiters:

Seien S_1^2 und S_2^2 Varianzen von Stichproben mit Umfang n_1 und n_2

$$X = \frac{S_1^2}{S_2^2} \frac{\sigma_2^2}{\sigma_1^2} \qquad m_1 = n_1 - 1$$

$$m_2 = n_2 - 1$$

Random variables and distributions

Probability space $(\Omega, \mathcal{P}(\Omega), P)$

An event A is the sum of „elementary events": $P(A) = \sum_{\omega \in A} P(\omega)$

Random variable

let:

$\Omega = \{(i,j) \mid i,j = 1, \dots, 6\}$

$P(\{i,j\}) = \frac{1}{36} = P(i) \cdot P(j)$

in:

„Payof-Function" (random variable)

$$X(i,j) = \begin{cases} 500 & \text{if } i+j=7 \\ -100 & \text{else} \end{cases}$$

$X: \omega \in \Omega \mapsto \mathbb{R}$

$X(\omega) = x$

„realization of X in image space

$\{x \mid X(\omega) = x, \omega \in \Omega\}$

image space can be discrete or continuous

Probability mass function (pmf) – discrete

„Wahrscheinlichkeitsgewichtungsfunktion"

$p(a) = P(X = a)$ $\quad 0 \leq p(a) \leq 1$ $\quad \sum_{a=1}^{n} p(a) = 1$

probability density function (pdf) – continuous

„Wahrscheinlichkeitsdichtefunktion"

$c \leq d$

$P(c \leq X \leq d) = \int_c^d f(x)\, dx$ $\quad f(x) \geq 0$ $\quad \int_{-\infty}^{+\infty} f(x)\, dx = 1$

we have to integrate it first to get a probability

so f(x) itself doesn't have to be [0;1]

cumulative distribution function

(cdf)

„Verteilungsfunktion"

$$F_X(x) = P(X \leq x) \quad \text{for all } x \in \mathbb{R}$$

- $0 \leq F(x) \leq 1$
- monotonically increasing
  $x \leq y$
  $F(x) \leq F(y)$
- $\lim_{x \to +\infty} F(x) = 1$
- $\lim_{x \to -\infty} F(x) = 0$

$$P(X > x) = 1 - P(X \leq x) = 1 - F_X(x)$$

$$P(a \leq X \leq b) = F_X(b) - F_X(a)$$

Inverse function of cdf

since cdf is not strictly monotonically increasing, one needs the notion of the „generalized inverse" of F

$$F^{-1}(p) := \inf\{x \mid F(x) \geq p\} \quad \text{for } p \in (0;1)$$

Quantile function

$$x_p = F^{-1}(p) \iff F(x_p) = p \quad \text{for } p \in (0;1)$$

p - Quantil of F

$$F_X(x_p) = P(X \leq x_p) = \int_{-\infty}^{x_p} f(t)\,dt = p \iff x_p = F_X^{-1}(p)$$

$x_{0,5}$ ... median

$x_{0,25}$ ... lower quartile

$x_{0,75}$ ... upper quartile

Expected value

discrete

$$E(X) = \sum_i x_i \, p(x_i) = x_1 \cdot p(x_1) + \dots + x_n \cdot p(x_n)$$

continuous

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)$$ → sum of integral should be absolute convergent

"weighted mean" or "central tendency"

"Moment of order one of X"

$E(X^k)$ ..... moment of order k of X

$E((X-E(X))^k)$ ... central moment of order k of X

$E((X-E(X))^2)$ ... $Var(X)$ ... variance of X

$\sigma = \sqrt{Var(X)}$ standard deviation of X

Properties

$E(aX+b) = a\,E(X) + b$

$E(aX+bY) = a\,E(X) + b\,E(Y)$

$E(h(X)) = \sum h(x_i) \cdot p(x_i)$

$E(h(X)) = \int_{-\infty}^{+\infty} h(x) \cdot f(x) \, dx$

Transformations

X is a continuous random variable, its f_X (density) is unknown

What is the distribution of a transformation?

$$Y = g(X) \quad g: \mathbb{R} \mapsto \mathbb{R}$$

1. determine F_X
2. determine F_Y for $Y = g(X)$
3. determine $f_Y(y) = F'_Y(y)$

Common families of distributions

discrete distributions

- Bernoulli
- Binomial
- Geometric
- Poisson

continuous distributions

- Uniform
- Exponential
- Normal (Gaussian)

(- χ^2 - distribution
- t - distribution)

Bernoulli Distribution

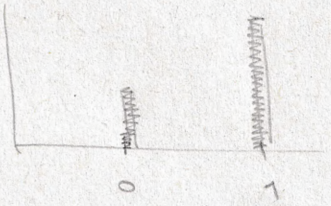
$X \sim ber(p)$

$\downarrow$

success $= p \quad = P(X=1)$

failiure $= q = 1-p = P(X=0)$

X can either be 0 or 1



	X	0	1
„pmf"	p(x)	1-p	p
„cdf"	F(x)	1-p	1

$E(X) = p$

$Var(X) = p(1-p)$

Binomial Distribution

$X \sim B(n,p)$

$\downarrow$

number of successes in n independent bernoulli trials (ber(p))

Therefore $B(1,p) = ber(p)$

X can be $0, ..., n$

$$p(x) = P(X=x) = \binom{n}{x} p^x q^{n-x} \quad \text{for } x \in \{0, \dots, n\}$$

with the binomial formule we get:

$$\sum_{x=0}^{n} p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} =$$

$$(p+1-p)^n = 1$$

$E(X) = np$

$Var(X) = npq$

Geometric distribution

with parameter p

Total # of attempts before a success.

$X \sim$ Geometric distribution (p)

$$p(x) = P(X = x) = (1-p)^x \cdot p$$

$$x \in \{0, 1, 2, \dots\}$$

$$E(X) = \frac{1-p}{p}$$

$$Var(X) = \frac{1-p}{p^2}$$

$$\sum_{x=0}^{\infty} k^x = \frac{1}{1-k} \quad \text{for } |k| < 1$$

$$\sum_{x=0}^{\infty} p(x) = p \sum_{x=0}^{\infty} (1-p)^x =$$

$$= \frac{p}{1-(1-p)} = 1$$

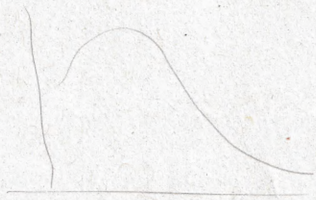
Poisson distribution

$X \sim P(\lambda)$

$X \sim Poi(\lambda)$

λ is the „intensity parameter"

$$P(k) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x \in \{0, 1, 2, \dots\}$$



$$E(X) = \lambda$$

$$Var(X) = \lambda$$

$$\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda} \quad \lambda \in \mathbb{R}$$

therefore

$$\sum_{x=0}^{\infty} p(x) = e^{-\lambda} \cdot \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = 1$$

k is interpreted as the number of times an event occurs in an intervall

Example of poisson distribution

X = number of calls per minute

Consider a device that handles 5 calls every 3 minutes : $\lambda = \frac{5}{3}$

What is the probability, that there will be no calls in the next minute? —"— At least 3 calls?

P(no calls in the next minute) =

$$P(X=0) = \frac{e^{-\frac{5}{3}} \left(\frac{5}{3}\right)^0}{0!} = 0,189$$

P(at least three calls in the next minute) =

$$P(X \geq 3) = 1 - \sum_{x=0}^{2} P(X=k) = 1 - \sum_{x=0}^{2} \frac{e^{-\frac{5}{3}} \left(\frac{5}{3}\right)^x}{x!} = 0,234$$

Binomial - Poisson - Relationship

$P(\lambda)$ as the limit of $B(n,p)$

$$X_n \sim B\left(n, \frac{\lambda}{n}\right)$$

$$\lim_{n\to\infty} P(X_n = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

many identical and independent experiments with a low succes-probability

↓

Number of successes $\overset{a}{\approx}$ poisson

$$X_n \sim B\left(n, \frac{\lambda}{n}\right)$$

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{(np)^x e^{-np}}{x!}$$

(rule of thumb)

for $n \geq 50$

$p \leq \frac{1}{10}$

$np \leq 10$

(continuous distributions)

uniform distribution

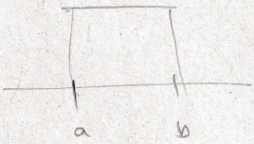
$X \sim \mathcal{U}(a,b)$

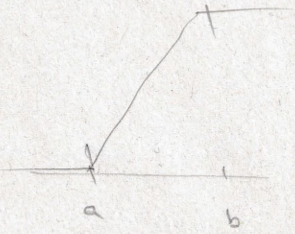
$$f(x) = \begin{cases} \frac{1}{b-a} & x \in (a,b) \\ 0 & \text{sonst} \end{cases}$$
(pdf)

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

$E(X) = \frac{a+b}{2}$

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





exponential distribution

$X \sim exp(\lambda)$ or

$X \sim exp(\tau)$ $\left(\tau = \frac{1}{\lambda}\right)$

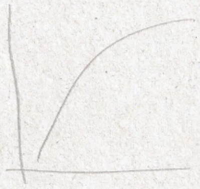
for λ:

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

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

$E(X) = \tau = \frac{1}{\lambda}$

$Var(X) = \tau^2 = \frac{1}{\lambda^2}$



Normal (Gaussian) Distribution

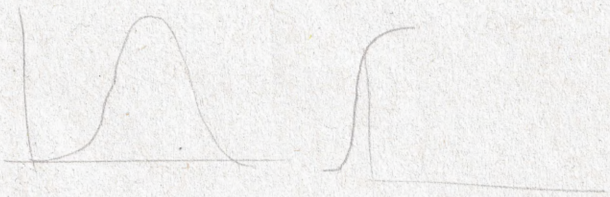
$X \sim \mathcal{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in \mathbb{R}$$

(pdf)

$E(X) = \mu$

$Var(X) = \sigma^2$



$\Phi(z) =$

$$= P(Z \leq z) = \int_{-\infty}^{z} f(t)\, dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}\, dt$$

← Standard normal distrib.

$Z \sim \mathcal{N}(0,1)$

$\Phi(-z) = 1 - \Phi(z)$

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

Standardization:

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

→ make standardized norm. distribution, out of any normal random variable X

The 68-95-99,7-Rule

$P(\mu-\sigma \leq X \leq \mu+\sigma) = P(-1 \leq \frac{X-\mu}{\sigma} \leq 1) = P(|Z| \leq 1) \approx 0{,}68$

$\mu-2\sigma \leq X \leq \mu+2\sigma$ $\quad P(|Z| \leq 2) \approx 0{,}95$

$\mu-3\sigma \leq X \leq \mu+3\sigma$ $\quad P(|Z| \leq 3) \approx 0{,}997$

Properties:

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with cdf F_X

Affine Transformations:

$$Y = a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$$

Quantile p

$$x_p = \mu + \sigma z_p \leftarrow z_p = \Phi^{-1}(p)$$

$$z_p = -z_{1-p}$$

Mean of multiple independent identically distributed (iid) random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad \bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

(normally distributed)

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Statistics and probability theory:

CLT   Central limit theorem   Zentraler Grenzwertsatz

LLN   Law of large numbers   Gesetz der großen Zahlen

Sample mean

$X_1, X_2, \dots, X_n \longrightarrow$ i.i.d: independently identically distributed random variables with sample size n

→ not necessarily normal distrib.

Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$E(\bar{X}) = E(X_1) = \mu$$
$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Especially, when $X_1, \dots X_n$ are normally distributed, the sample mean is also normally distributed

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

CLT: About $\bar{X}_n$'s distribution - For large n, $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ approximately

LLN: About $\bar{X}_n$'s value - As n grows $\bar{X}_n$ tends to be in the neighbourhood of μ

Formally

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$$

Formaly

$$S_n = X_1 + X_2 + X_3 \dots + X_n \quad \text{(sum)}$$
$$\bar{X}_n = \frac{1}{n} S_n \quad \text{(sample mean)}$$

for large n:

$$S_n \approx N(n\mu, n\sigma^2)$$
$$\bar{X}_n \approx N(\mu, \frac{\sigma^2}{n})$$

Histograms

frequency-histogram $\rightarrow$ height of bar $= \sum$ continuous values

density - histogram $\rightarrow$ Area

binning:
dividing continuous function in bins with equal width.

LLN:
Density histogram converges to pdf

Normal approximation of Bernoulli $B(n,p)$

$S_n = X_1 + \dots + X_n \rightarrow$ iid with ber(p)

$E(S_n) = np$

$Var(S_n) = np(1-p)$

„Continuity correction"

$$P(a \leq S_n \leq b) \approx \Phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

$a \leq b$, where $a,b \in \{0,1,\dots,n\}$

$\hookrightarrow$ can be equal because of $\pm\frac{1}{2}$

Approximation

$$P(S_n \leq x) \approx \Phi\left(\frac{x+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

Rule of thumb:
$\min\{np, np(1-p)\} \geq 10$

Normal approximation of Poisson

$\lambda > 15$ $$P(X \leq x) = \Phi\left(\frac{x+\frac{1}{2}-\lambda}{\sqrt{\lambda}}\right)$$

Covarience and correletion

$\underline{\mathrm{Cov}}(X,Y) = E((X-E(X))\cdot(Y-E(Y))) = E(XY) - E(X)E(Y)$

$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$

$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$

if independent:

$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

$\mathrm{Cov}(X,Y) = 0$ (but if $\mathrm{Cov}(X,Y)=0$ that doesn't imply independency)

→ Measures linear dependence!

Correlation coefficient

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

$$= \mathrm{Cov}\left(\frac{X-E(X)}{\sqrt{\mathrm{Var}(X)}}, \frac{Y-E(Y)}{\sqrt{\mathrm{Var}(Y)}}\right)$$

Covariance of standardized version

$-1 \leq \rho(X,Y) \leq 1$

$\rho(X,Y) = 1 \iff Y = aX+b \quad a>0$

$\rho(X,Y) = -1 \iff Y = aX+b \quad a<0$

Joint probability table

Example:

flip coin 3 times

$X :=$ # Heads in the first 2 throws

$Y :=$ # Heads in the last 2 throws

$2^3 = 8$

$|\Omega| = 8$

$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	2	2	1	1	1	1	0	0
Y	2	1	1	0	2	1	1	0

$P(X=0) = 2/8 = 1/4$

$P(X=1) = 4/8 = 1/2$

$P(X=2) = 2/8 = 1/4$

$P(Y=0) = 2/8 = 1/4$

$P(Y=1) = 4/8 = 1/2$

$P(Y=2) = 2/8 = 1/4$

Simultanious occurences:

X\Y	0	1	2	
0	1/8	1/8	0	1/4
1	1/8	1/4	1/8	1/2
2	0	1/8	1/8	1/4
	1/4	1/2	1/4	1

From marginal distributions we compute:

$E(X) = 1 \quad = 0 \cdot 1/4 + 1 \cdot 1/2 + 2 \cdot 1/4 = 1$

$E(Y) = 1$

$E(XY) = 1 \cdot 2/8 + 2 \cdot 1/8 + 2 \cdot 1/8 + 4 \cdot 1/8 = \frac{5}{4}$

$Cov(X,Y) = E(XY) - E(X)E(Y) = 1/4$

$Var(X) = 1/2$

$Var(Y) = 1/2$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = 1/2$$

X \ Y	0	1	2
0	1/8	1/8	0
1	1/8	2/8	1/8
2	0	1/8	1/8

$$\left(\begin{array}{l} 0 \cdot 0 \cdot 1/8 + 0 \cdot 1 \cdot 1/8 + 0 \cdot 2 \cdot 0 + \\ 1 \cdot 0 \cdot 1/8 + 1 \cdot 1 \cdot 2/8 + 1 \cdot 2 \cdot 1/8 + \\ 2 \cdot 0 \cdot 0 + 2 \cdot 1 \cdot 1/8 + 2 \cdot 2 \cdot 1/8 \end{array} \right)$$

Alternative Approach:

Properties of covariance

X_i is flip $i = 1,2,3$

$X_i \sim Ber(1/2)$

$\longrightarrow$ $E(X_i) = \frac{1}{2}$

$Var(X_i) = 1/4$

$Cov(X_1, X_2) = Cov(X_1, X_3) =$

$= Cov(X_2, X_3) = 0$

The 3 tosses are independent

$X = X_1 + X_2$

$Y = X_2 + X_3$

$Cov(X,Y) = Cov(X_1 + X_2, X_2 + X_3)$

$= Cov(X_1, X_2) + Cov(X_1, X_3) + Cov(X_2, X_2) +$
$Cov(X_2, X_3)$

$= Cov(X_2, X_2) = Var(X_2) = 1/4$

Teschl:

Beschreibende Statistik & Zusammenhangsanalysen

Deskriptive Statistik / beschreibende

Beschreibung von Daten

Darstellung von Daten

Kenngrößen

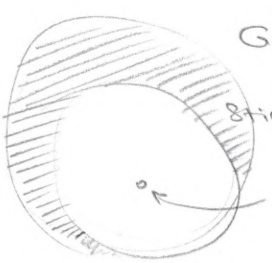
Validierung

Explorative Statistik

Muster finden

Induktive / Schließende / inferentielle / beurteilende Statistik

Mit Wahrscheinlichkeitsrechnung allgemeine ~~Aussagen~~ treffen



Grundgesamtheit

Stichprobe ⟶ Repräsentative Stichprobe = Zufallsstichprobe

Statistische Einheit
(kleinste Entität) und die jeweiligen „Merkmale" / „Ausprägungen" / „Variablen"

↓

Quantitativ (Zahl; diskret / stetig)

Qualitativ (Qualität, Enum)

Skalierungen /
Skalenniveaus:

- nominalskaliert
  Namen ohne natürliche Reihenfolge
- ordinalskaliert
  Namen die sich ordnen lassen
- kardinalskaliert / metrisch skaliert
  Zahlen
  intervallskaliert: keine Verhältnisse, Nullpunkt willkürlich gewählt
  zB Kalenderzeiten
  Verhältnisskaliert: zB Alter, Größe, Einkommen

Häufigkeitsverteilung einer Stichprobe

Univariat - Nur 1 Merkmal beobachten
multivariat - Mehrere Merkmale beobachten

Urliste - Dataset

<u>relative Häufigkeit</u>: $f_i = \frac{h_i}{n}$ → wobei $\sum_i f_i = 1$

<u>empirische Verteilungsfunktion</u>: $F(x) = \sum_{i: a_i \leq x} f_i$

für Histogramme definiert man „bins" (Klassen) folgendermaßen:

i-te Klasse: $f_i = \frac{h_i}{n}$ ← # im Bin

→ „relative Häufigkeit der i-ten Klasse"

Die Wahl der richtigen Klassengröße:

In Histogrammen: Fläche repräsentativ.

Deshalb:

Rechteck-Höhe $= \frac{h_i}{\Delta x_i}$ ← Bin-Breite relative Häufigkeit: $f_i = \frac{h_i}{n}$ ← Häufigkeit von Merkmal i

wobei $\sum_i f_i = 1$

empirische Verteilungsfunktion: $F(x) = \sum_{i: a_i \leq x} f_i$

Kennwerte einer Stichprobe

Lagekennwerte

arithmetisches Mittel $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i=1}^{k} h_i a_i = \sum_{i=1}^{k} f_i a_i$

$x_i = h_i \cdot a_i$ → Werte

Median / Zentralwert

$$\tilde{x} = \begin{cases} x_{m+1} & n = 2m+1 \quad \text{(ungerade Anzahl)} \\ \frac{1}{2}(x_m + x_{m+1}) & n = 2m \quad \text{(gerade Anzahl)} \end{cases}$$

Der Median ist unempfindlich gegenüber Ausreißern.

Modalwert

Wert welcher unter den nominalskalierten Merkmalen (keine Zahl, keine Ordnung) am häufigsten vorkommt.

p-Quantil

$0 < p < 1$

$$\tilde{x}_p = \begin{cases} x_{\lfloor np \rfloor + 1} & np \notin \mathbb{N} \\ \frac{1}{2}(x_{np} + x_{np+1}) & np \in \mathbb{N} \end{cases}$$

Wobei $\tilde{x}_{0,25}, \tilde{x}_{0,75}$ Quartile

$\tilde{x}_{0,1}, \tilde{x}_{0,2}, \tilde{x}_{0,3}, \ldots$ Dezile

$\tilde{x}_{0,01}, \tilde{x}_{0,02} \ldots$ Perzentile

Streuungskennwerte

Stichproben / Empirische Varianz

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{h_i}_{f_i} (a_i - \bar{x})^2$$

Stichproben / Empirische Standardabweichung

$s = \sqrt{s^2}$ → Gleiche Einheit wie einzelnen Werte

Satz:

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right) = \frac{1}{n-1} \left(\sum_{i=1}^{n} (h_i\, a_i)^2 - n \cdot \bar{x}^2 \right)$$

Spannweite

$R = x_{max} - x_{min}$

IQR - Interquartilabstand

$d_Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$

↓

Boxplot mit Whiskers

Lineare Korrelation

Erkennung von Zusammenhängen in einer zweidimensionalen Stichprobe.

Streudiagramm: (x,y) - Wolke

Über das Ausmaß eines „linearen" Zusammenhangs:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$ ← Empirischer Korrelationskoeffizient

$-1 \leq r_{xy} \leq +1$

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Empirische Kovarianz

Wenn

$r_{xy} > 0$: pos korrel.

$r_{xy} < 0$: neg korrel.

$r_{xy} \approx 0$: keine korrel.

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

Empirische Standardabweichungen

Lineare Regression

Zusammenhänge, wenn y abhängig von x

Man drückt y durch f(x) aus

$y = f(x) = kx + d$ → Annäherung durch eine Gerade: lineare Regression durch Regressionsgerade

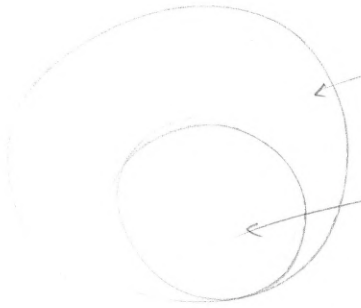
Gauss'sche Methode der kleinsten Quadrate

Wir suchen die minimale: $\sum_{i=1}^{n} (y_i - f(x_i))^2$

$k = r_{xy} \frac{s_y}{s_x}$ $\quad d = \bar{y} - k\bar{x}$

↑ arithmet. Mittel

Schließende Statistik



Grundgesamtheit Umfang = N

Stichprobe von der wir uns Merkmal X anschauen

Stichproben-Umfang = n

Realisationen aus der Stichprobe:

$\{x_1, x_2, x_3, x_4, \dots\}$

Wir schließen von den Realisationen der Stichprobe auf die Grundgesamtheit.

Definition

Eine Zufallsstichprobe vom Umfang n ist eine Folge $X_1 \dots X_n$ von i.i.d Zufallsvariablen

Stichproben-Variablen, (Merkmalsausprägung des i-ten Elements)

mit Realisationen $x_1 \dots x_n$

wenn N hinreichend groß ist,
(Eigentlich wenn N unendlich groß ist)
→ Ziehen ohne Zurücklegen

Wir wollen etwas über X in der Grundgesamtheit wissen: θ

So ein Parameter ist eine feste, uns aber unbekannte Zahl

NICHT vom Zufall abhängig

Von Stichprobe auf die Grundgesamtheit schließen:

- Punktschätzungen
- Intervallschätzungen
- Hypothesentests

Punktschätzungen

Stichprobenfunktion / Schätzfunktion für Θ

$$\underbrace{T(X_1, \dots, X_n)}_{\text{Zufallsvariable}} : \omega \in \Omega \longmapsto T(\omega) \in \mathbb{R}$$

Realisationen sind Schätzwerte

Der Schätzwert hängt von der gezogenen Stichprobe ab.
Deshalb werden wir nicht davon ausgehen, dass es Θ genau trifft.

Wir erwarten aber, dass die Schätzwerte im Mittel richtig sind.
↑ Stichprobe $\Rightarrow$ ↑ Präzision

Definition

Schätzfunktion T heißt erwartungstreu / unverzerrt (unbiased) wenn

$$E(T) = \Theta$$

konsistent wenn sie stochastisch gegen Θ konvergiert:

$$\lim_{n \to \infty} P(|T - \Theta| < \varepsilon) = 1 \quad \text{für beliebiges } \varepsilon > 0$$

konsistent im quadratischen Mittel wenn

$$\lim_{n \to \infty} E((T-\Theta)^2) = 0$$

↳ Quadratische Abweichung

Wegen

$$E((T-\Theta)^2) = E(T^2) - 2\Theta E(T) + \Theta^2 = Var(T) + (E(T) - \Theta)^2$$

Konsistenz im quadratischen Mittel $\Rightarrow$ wenn $\lim_{n \to \infty}$ dann $Var(T) \to 0$

$(E(T) - \Theta) \to 0$

Wenn

$$E(T) = \Theta \quad \text{und} \quad \lim_{n \to \infty} Var(T) = 0$$

dann bedeutet das T ist auch konsistent im quadratischen Mittel

Wenn quadratisch konsistent $\Rightarrow$ konsistent

Satz 27.44

$X_1 \dots X_n$ sind iid mit μ und σ^2

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Zufallsvariable!

$E(\bar{X}) = \mu$

$Var(\bar{X}) = \frac{\sigma^2}{n}$

Das gilt auch für Schätzungsfunktionen

$$\bar{T} = \frac{1}{n} \sum_{i=1}^{n} T_i \longrightarrow E(T_i) = \theta$$

Dadurch gilt:

$\bar{X}$ .... arithmetisches Mittel für μ

$\bar{P}$ ... empirischer Anteil für p $\longrightarrow$ Alle sind unbiased und konsistent im Quadratischen

$\bar{F}$ .... empirische Verteilungsfunktion für F

$$\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i \quad \text{wobei} \quad P_i \begin{cases} \text{Eigenschaft vorhanden} & 1 \\ \text{Eigenschaft nicht vorhanden} & 0 \end{cases}$$

Satz 27,47

Hauptsatz der Statistik

$X_1 \dots X_n$ sind iid mit Verteilungsfunktion $F(X)$ und empirischer Verteilungsfunktion $\bar{F}(x)$

$\varepsilon > 0$ und $x \in \mathbb{R}$:

$$\lim_{n \to \infty} P\left(|\bar{F}(x) - F(x)| < \varepsilon \right) = 1$$

„$\bar{F}(x)$ konvergiert stochastisch gegen $F(x)$"

Schätzfunktion S^2 für σ^2 (Varianz)

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 \right) - \frac{n}{n-1} \bar{x}^2$$

$\downarrow$

Nur Erwartungstreu wenn man durch $n-1$ dividiert & nicht durch n

$\downarrow$

$$\underbrace{E\left(s^2 \cdot \frac{n-1}{n}\right)}_{} = \sigma^2 - \frac{\sigma^2}{n}$$

Asymptotisch erwartungstreu

S^2 ist erwartungstreu & konsistent

Standardabweichung S ist auch asymptotisch erwartungstreu und konsistent.

Intervallschätzungen (für μ bei bekanntem σ) (Normalverteiltes Merkmal)

Punktschätzung sagt ob θ genau getroffen wird oder nicht.
Es wird nichts über die Abweichung gesagt.

Bei Intervallschätzung:

Intervall bestimmen, dass θ mit hoher Wahrscheinlichkeit überdeckt.

$[g_u(X_1, \dots, X_n) ; g_o(X_1 \dots X_n)]$

Deckt mit Wahrscheinlichk. $1-\alpha$ ab.

↓

$P(\theta \in [g_u(X_1 .. X_n) ; g_o(X_1 .. X_n)])$
$= 1 - \alpha$

↓

Konfidenzintervall zum Niveau $1-\alpha$ (Konfidenzniveau)

- Man nennt α „Irrtumswahrsch."
- meistens 0,95, 0,9, 0,99 als Konfidenzniveau

Wichtig:

Streng genommen ist θ fix und die Grenzen werden vom Zufall bestimmt,

Eigentlich ist Konfidenzniveau die Wahrscheinlichkeit, dass θ getroffen wird.

Wir betrachten Konfidenzintervalle für:

- Erwartungswert bei Normalverteilung mit bekannter / unbekannter Standard-Abweichung
- —"— bei beliebiger Verteilung
- —"— bei großem Stichprobenumfang

:

→ Immer ähnliche Vorgehensweise!

Bei unseren Beispielen aber Normalverteilung mit bekannter σ und μ

Normalverteilung

Beispiel

$\sigma = 2$ soll aber 100 sein.

(Wir wollen weder, dass $\mu < 100$ noch $\mu > 100$

Stichprobe:

$n = 10$

$\bar{X} = 98,9$ → gesucht ist μ

wir kennen σ

$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ → Ebenso normalverteilt

$\mu_{\bar{X}} = \mu$

$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$

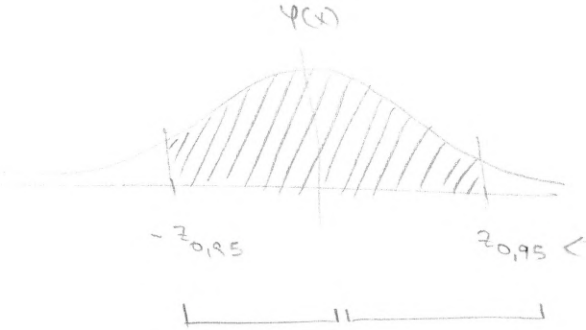
Wählen wir:

$1 - \alpha = 0,9$

Dann brauchen wir einen z Wert, sodass

$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

Zugehörige standardisierte Zufallsvariable



Das $(1-\frac{\alpha}{2})$ Quantil $= z_{0,95}$

$z_{1-\frac{\alpha}{2}} = z_{0,95} = 1,645$

Nun wissen wir, dass Z mit 90%iger Wahrscheinlichkeit, einen Wert zwischen $-1,645$ und $+1,645$ annimmt:

$$0,90 = P(-1,645 \leq Z \leq 1,645) = P\left(-1,645 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq 1,645\right)$$

$$= P\left(-1,645 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1,645 \frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\bar{X} - 1,645 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,645 \frac{\sigma}{\sqrt{n}}\right)$$

→ Mit 90% wahrsch. liegt μ im Intervall $\left[\bar{X} - 1,645 \frac{\sigma}{\sqrt{n}} ; \bar{X} + 1,645 \frac{\sigma}{\sqrt{n}}\right]$

↑

weil Grenzen $\bar{X}$ enthalten sind sie Zufallsvariablen

(Sie hängen von der Stichprobe ab)

Wenn wir $\bar{x} = 98,9$ und $n=10$ und $\sigma = 2$ einsetzen:

$$\left[98,9 - 1,645 \frac{2}{\sqrt{10}} \; ; \; 98,9 + 1,645 \frac{2}{\sqrt{10}}\right] = [97,9 \, , \, 99,9]$$

↑
Enthält mit 90% Wsl μ

Der gesuchte Sollwert 100 ist nicht enthalten.

Zusammengefasst:

Konfidenzintervall für μ eines normalverteilten Merkmals bei bekanntem σ:

1. Konfidenzniveau $(1-\alpha)$ bestimmen
2. Stichprobe vom Umfang n ziehen, $\bar{x}$ berechnen
3. Quantil $z_{1-\frac{\alpha}{2}}$ bestimmen der Standardnormalverteilung

Resultat:

$\left[\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$ deckt mit Wsl $\alpha - 1$ den gesuchten Erwartungswert ab.

Wichtig:

- Konfidenzintervall ist zentriert um $\bar{x}$, nicht μ
  Wir sagen aber trotzdem die richtige Wahrscheinlichk.

- Gesamtbreite des Intervalls:

$$L = 2 z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

→ Daraus folgt

↓ n ⇒ ↑ L, weil dann mehr Information zum einschätzen benötigt.

↑ $(1-\alpha)$ ⇒ ↑ L, weil größere Fläche zwischen Quantilen

(also 0,99 ist breiter als 0,9 z.B.)

- Manchmal ist erwünscht:

$$2 z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq L_{max}$$

$$\rightarrow n \geq \left(\frac{2 z_{1-\frac{\alpha}{2}} \sigma}{L_{max}}\right)^2$$

Bisher wurde das zweiseitige Intervall behandelt.

Man kann es aber auch einseitig lösen:

$$\left(-\infty\,;\ \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right] \quad \text{oder} \quad \left[\bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\,;\ \infty\right)$$

Vorgehen dabei analog:

$$1-\alpha = P\left(Z \le z_{1-\alpha}\right) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha}\right) =$$

$$= P\left(\bar{X}-\mu \le z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = P\left(\mu \ge \bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

$$\downarrow$$

$$\left[\bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\,;\ \infty\right)$$

Intervallschätzungen (für μ bei unbekanntem σ) (Normalverteiltes Merkmal)

$\bar{X} \sim N(\mu; \frac{\sigma^2}{n})$

Da σ^2 nicht bekannt schätzen wir es durch die „erwartungstreue Stichproben-Varianz".

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

> Wir standardisieren damit $\bar{X}$

$Z \neq \frac{\bar{X} - \mu}{S/\sqrt{n}}$ nicht standardnormalverteilt, sondern t-verteilt, mit $n-1$ Freiheitsgraden

Wir bestimmen aus der t-Verteilung:

Quantil $t_{n-1;\,1-\frac{\alpha}{2}}$

Schritte

1. Vertrauensniveau / Konfidenzniveau $1-\alpha$ wählen (zB 0,9, 0,95 ...)
2. Stichprobe vom Umfang n ziehen
   $\bar{X}$ berechnen und S
3. t Quantil bestimmen
   $t_{n-1;\,1-\frac{\alpha}{2}}$

$\left[\bar{x} - t_{n-1;\,1-\alpha/2} \frac{S}{\sqrt{n}} ;\ \bar{x} + t_{n-1;\,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$ Konfidenzintervall

überdeckt μ mit Wahrscheinlichkeit $1-\alpha$

Beispiel

Niveau $= 0{,}95 = \alpha - 1$ $\quad n = 10$

Daraus folgt:

$1 - \frac{\alpha}{2} = 0{,}975$

Quantil der t-Verteilung mit $n-1 = 10-1 = 9$ Freiheitsgraden ist

$t_{9;0{,}975} = 2{,}262$

$\downarrow$

Konfidenzintervall

$\left[98{,}9 \pm 2{,}262 \frac{2{,}83}{\sqrt{10}}\right] = [97{,}3 \,;\, 100{,}5]$

enthält mit 95%iger Wahrscheinlichkeit μ

Konfidenzintervall für μ mit großem Stichprobenumfang (beliebige Verteil.)

Wenn n groß genug ist, spielt die Verteilung keine Rolle

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\rightarrow$ Nach zentralem Grenzwertsatz: $\bar{X}$ für hinreichend große n annähernd normalverteilt.

Falls σ bekannt: $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

Falls σ unbekannt: $Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$

Schritte

1. Wähle $(1-\alpha)$ (zB 0,9, 0,95, ...)
2. Ziehe Stichprobe mit $n \gtrsim 30$ und berechne $\bar{x}$ und gegebenenfalls s falls σ unbek.
3. Bestimme Quantil

   $z_{1-\frac{\alpha}{2}}$

Dann bedeckt Konfidenzintervall μ mit Wahrscheinlichk. $1-\alpha$

$\left[\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma \text{ (oder } s)}{\sqrt{n}}\right]$

Beispiel:

$$1-\alpha = 0{,}95$$
$$1-\frac{\alpha}{2} = 0{,}975$$
$$z_{0{,}975} = 1{,}96$$

$$\left[20 \pm 1{,}96 \frac{1{,}5}{\sqrt{100}}\right] = [19{,}71 ; 20{,}29]$$

Konfidenzintervall für den Vergleich der Erwartungswerte von zwei Normalvert.

Erwartungswerte von Stichproben vergleichen

Lösung mit Hilfe eines Konfidenzintervalls für die Differenz der Erwartungswerte.

Stichprobe 1 $\sim N(\mu_1 ; \sigma_1^2)$
Umfang n_1
$\bar{X}_1 \sim N\left(\mu_1 ; \frac{\sigma_1^2}{n_1}\right)$

Stichprobe 2 $\sim N(\mu_2 ; \sigma_2^2)$
Umfang n_2
$\bar{X}_2 \sim N\left(\mu_2 ; \frac{\sigma_2^2}{n_2}\right)$

$\bar{X}_1 - \bar{X}_2$ ist eine erwartungstreue Schätzfunktion für $\mu_1 - \mu_2$

Nach Additions und Linearitätssatz: (für Normalvert.)

$$\left(\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) \sim N(0;1)$$

(Normalverteilt) Standardisiert

Schritte

1. Konfid. Niveau $1-\alpha$ (zB 0,9, 0,95 ...) wählen
2. Aus jeder Grundgesamtheit Stichprobe 1 und Stichprobe 2 ziehen
   Umfänge: n_1, n_2
   Arithm. Mittel: $\bar{x}_1$, $\bar{x}_2$
3. Quantil der Normalverteilung - Standardisiert bestimmen: $z_{1-\frac{\alpha}{2}}$

$$\left[\bar{x}_1 - \bar{x}_2 \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

überdeckt mit Wsl $\alpha - 1$ den gesuchten Parameter $\mu_1 - \mu_2$

In der Praxis sind Varianzen nicht bekannt (σ_1, σ_2),
Man kann aber zeigen, dass sie gleich sind, also: $\sigma_1 = \sigma_2$

Konfidenzintervall für den Vergleich der Erwartungswerte von 2 Norm. vert. mit unbekannten aber gleichen Standardabweichungen

Wenn $\sigma_1 = \sigma_2$ kann man zeigen:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{n_1+n_2}{n_1 \cdot n_2}\right)\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\right)}} \sim t(n_1+n_2-2)$$

Schritte

1. Vertrauensniveau $1-\alpha$ bestimmen
2. Zwei Stichproben ziehen (wie vorhin nur mit $\underbrace{S_1^2, S_2^2}_{\text{„Stichproben Varianzen"}}$)
3. Quantil $t_{n_1+n_2-2;\, 1-\alpha/2}$ bestimmen, mit t-Verteilung und n_1+n_2-2 Freiheitsgraden

---

$$[g_-, g_+] \quad g_\pm = \bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2;\, 1-\alpha/2} \cdot \sqrt{\left(\frac{n_1+n_2}{n_1 \cdot n_2}\right)\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\right)}$$

überdeckt Parameter $\mu_1 - \mu_2$ mit Wahrscheinlichkeit $1-\alpha$.

➤ Case $n_1 = n_2 = n$ vereinfacht Formel drastisch

$$\left[\bar{x}_1 - \bar{x}_2 \pm t_{2(n-1);\, (1-\frac{\alpha}{2})} \sqrt{\frac{S_1^2 + S_2^2}{n}}\right]$$

Wenn Konfidenzintervall 0 enthält kann es sein dass $\mu_1 - \mu_2 = 0$

„Verbundene Stichproben" (Fortsetzung)

Man definiert eine neue Stichprobe basierend auf andere.

zB.: $D = X_1 - X_2$

$\downarrow$ berechnen den Beziehung für einzelne X_i Werte

$\bar{D} = \bar{X}_1 - \bar{X}_2$

Wenn die Varianzen nicht gleich sind, ist die Berechnung aber kompliziert!

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

ist näherungsweise t-verteilt mit

$$m = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Freiheitsgraden.
(nicht ganzzahlig)

Konfidenzintervall für σ^2 (bei Normalverteilung)

Erwartungswert braucht nicht bekannt sein:

Stichprobe $X_1 \dots X_n$ wobei X_i iid verteilt sind nach $N(\mu; \sigma^2)$

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{n-1}{\sigma^2} s^2$$

$\rightarrow$ χ^2-verteilt mit $n-1$ Freiheitsgraden

Schritte

1. Konfidenzniveau bestimmen $1-\alpha$
2. Stichprobe vom Umfang n ziehen und s^2 berechnen
3. Quantille $\chi^2_{n-1; \frac{\alpha}{2}}$ und $\chi^2_{n-1; 1-\frac{\alpha}{2}}$ bestimmen mit $n-1$ Freiheitsg.

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}} ; \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}\right]$$

$\rightarrow$ überdeckt mit Wahrsch. $1-\alpha$ die Varianz σ^2

Weil χ nicht symmetrisch ist müssen wir Quantillen so bestimmen.

Beispiel

$n = 10$

normalverteilte Grundgesamtheit

$S^2 = 0{,}25$

$\alpha - 1 = 0{,}95$

$\frac{\alpha}{2} = 0{,}025 \quad 1 - \frac{\alpha}{2} = 0{,}975$

χ^2-Verteilung mit $n - 1 = 9$ Freiheitsgraden:

$\chi^2_{9;0,025} = 2{,}7$

$\chi^2_{9;0,975} = 19{,}02$

$$\left[\frac{9 \cdot 0{,}25}{19{,}02} ; \frac{9 \cdot 0{,}25}{2{,}7}\right] =$$

$[0{,}118 , 0{,}833]$

Konfidenzintervall für einen Anteilswert

Wir interessieren uns für $p = P(A)$ falls Grundgesamtheit unendlich.

A: Eigenschaft A — entweder 0 oder 1

bzw für den Anteil p an Elementen in Grundgesamtheit mit dieser Eigenschaft falls Grundgesamtheit endlich ist.

!

$$\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i$$

→ Wie ist $\bar{P}$ verteilt?

Nach $Bi(n;p)$!

$\bar{P} \sim Bi(n;p)$

$\mu = np$

$\sigma^2 = np(1-p)$

→ Für großes n normalverteilt

$\sim N(np ; np(1-p))$

bzw

$$\sim N\left(p ; \frac{p(1-p)}{n}\right)$$

Vorgang:

$$\bar{P} \sim Bi(n, p)$$

mit großen n auch $\sim N\left(p ; \frac{p(1-p)}{n}\right)$

Wir suchen zugehörige standardisierte Normalverteilung:

$$\frac{\bar{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0;1)$$

Wir legen α fest:

$$\frac{(\bar{P} - p)^2}{\frac{p(1-p)}{n}} \leq z^2 \rightarrow p^2 - 2\frac{n}{n+z^2}\left(\bar{P} + \frac{z^2}{2n}\right) p + \frac{n}{n+z^2}\bar{P}^2 \leq 0$$

(Quadratische Gleichung)

$$g \pm = \frac{n}{n+z^2}\left(\bar{P} + \frac{z^2}{2n} \pm z\sqrt{\frac{\bar{P}(1-\bar{P})}{n} + \frac{z^2}{4n^2}}\right)$$

Zusammengefasst:

Approximatives Konfidenzintervall für eine Wahrscheinlichkeit bzw. einen Anteilswert p bei großem Stichprobenumfang ($n \gtrapprox 20$):

1. Bestimmen $1-\alpha$
2. Ziehen Stichprobe Umfang n und berechnen $\bar{P}$ der Elemente
3. Bestimmen Quantil $z_{1-\alpha/2}$

Konfidenzintervall deckt mit Wahrscheinlichkeit $1-\alpha$ den gesuchten Parameter p ab

(Grenzen sind als Näherungswerte zu verstehen)

Wichtig: bei $n\bar{p}(1-\bar{p}) \gtrapprox 9$ kann man einfachere Formel benutzen

Hypothesen - Tests

Wir wollen weiterhin durch die Stichprobe Informationen über die Grundgesamtheit, Diesmal durch „Tests"

Angenommen:

$\bar{x} = 98,9$
μ soll aber 100 sein. (Sollwert von $\mu = 100$)

Wir stellen die Hypothese auf: $\mu = 100$ („Nullhypothese H_0")
Steht aufrecht so lange nicht „schwerwiegende" Beweise dagegen sprechen.

Gegenbehauptung: $\mu \neq 100$ („Alternativhypothese H_1")

Auf Grund der Stichprobe soll bestimmt werden, ob wir H_0 beibehalten oder zu Gunsten von H_1 verwerfen;

H_1 wenn Stichprobe stark vom Sollwert abweicht.
Abweichung sehr stark <=> sehr unwahrscheinlich dass Stichproben-Mittel durch 100 entstanden ist

konstante Abweichung <=> Wahrscheinlichkeit
„starke Abweichung" „sehr unwahrscheinlich"
Definieren wir als c Definieren wir als $\alpha = 5\%$
Signifikanzniveau des Tests

↓

Wahrscheinlichkeit, dass $\bar{X}$ bei Gültigkeit von H_0 um c von μ abweicht ist gleich α %.

$$P(|\bar{X} - \mu| > c) = \alpha$$

Herleiten der „starken Abweichung"

Wir wissen:
Wenn $\sigma = 2$ dann führt diese Bedingung zum kritischen Wert:

$n = 10$
$\sigma = 2$
$\alpha = 0,05$

$$c = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 1,96 \cdot \frac{2}{\sqrt{10}} = 1,24$$

(Intervalle davon bestimmt durch $\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$)

$\alpha = 0{,}05$ Wahrsch.

$c = 1{,}24$ Abweich.

Entscheidungsregel:

$\bar{x} < \mu - c = 100 - 1{,}24 = 98{,}8$

$\bar{x} > \mu + c = 100 + 1{,}24 = 101{,}2$

Dann Abweichung zu groß $\Rightarrow$ zu unwahrscheinlich, dass μ der Sollwert ist $\Rightarrow$ Wir verwerfen H_0 und nehmen H_1 an.

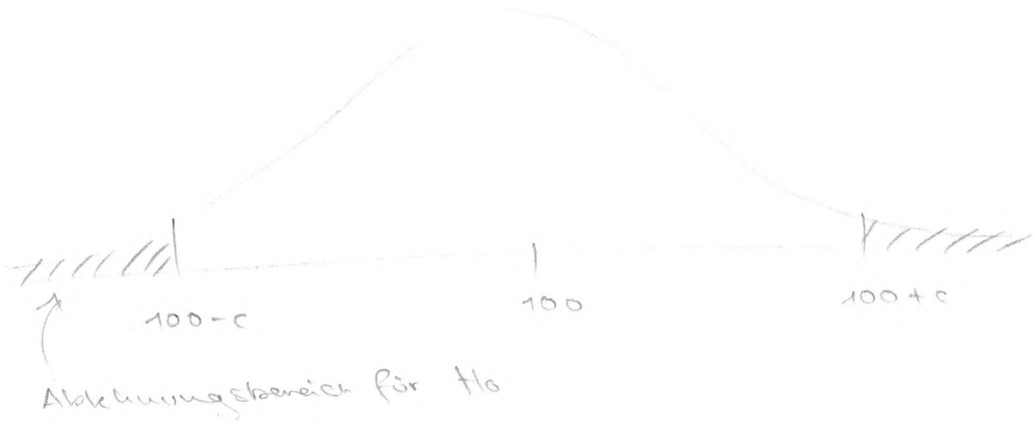
In anderen Worten:

Wir verwerfen H_0 wenn μ nicht im Konfidenzintervall zum Vertrauensniveau $1-\alpha$ liegt.

$\bar{x} = 98{,}9$

$\hookrightarrow$ Liegt nicht im Ablehnungsintervall $(-\infty; 98{,}8) \cup (101{,}2; \infty)$

Wir behalten daher H_0 bei.



Zusammenfassung

Ein statistisches Testproblem besteht aus:

- Nullhypothese H_0
- Alternativhypothese H_1

schließen sich aus

Wenn Hypothesen Aussagen über Parameter θ machen: parametrischer Test.

Hypothese heißt „einfach" wenn:

- Einziger Parameterwert $\theta = \theta_0$

Hypothese heißt „zusammengesetzt" wenn:

- mehrere Parameterwerte $\theta \neq \theta_0$, $\theta \geq \theta_0$ oder $\theta \leq \theta_0$

Test besteht aus:

Zufallsstichprobe und einer

- Prüfgröße / Teststatistik $T(X_1 \dots X_n)$

Ablehnungsbereich / Verwerfungsbereich / kritischer Bereich

umfasst Werte, die mit einer Wahrscheinlichkeit $\leq \alpha$ auftreten

„Signifikanz-Niveau"

0,10
0,05
0,01

Evaluierung des Tests:

Prüfwert im Ablehnungsbereich $\rightarrow H_1$

ansonsten $\rightarrow H_0$

Was bedeutet Signifikanzniveau α eigentlich?

$\alpha = 5\%$ zB bedeutet dass wenn man sehr viele Tests durchführen würde (mit gleichem Umfang n) bei 5% der Tests ein $\bar{x}$ finden, dass im Ablehnungsbereich liegt.

Die Wahrscheinlichkeit liegt bei 5%.

Fehler in der Evaluation von statistischen Tests:

	Entscheidung für: H_0	H_1
H_0 wahr	ok	Fehler 1. Art (α-Fehler)
H_1 wahr	Fehler 2. Art (β-Fehler)	ok

Man sagt auch

Fehler 1. Art = falsch negativ → Wahrscheinlichkeit = α

Fehler 2. Art = falsch positiv → Wahrscheinlichkeit kann nicht vorgegeben werden.

Wenn möglich versucht man aber Fehler der 1. Art wahrscheinlicher zu halten als Fehler der 2. Art

$\underbrace{\text{falsch negativ}}_{\text{(eigentl. pos)}}$ ist gefährlicher als $\underbrace{\text{falsch positiv}}_{\text{(eigentl. neg)}}$

β α

Je kleiner α ist desto wahrscheinlicher ist, dass H_0 beibehalten wird. Dadurch verliert es an Aussagekraft.

Wenn $\alpha = 0$, dann will man mit Sicherheit ein wahres H_0 nicht irrtümlich verwerfen. → Dann wäre $\beta = 1$ (Wahrscheinlichkeit falsches H_0 zu behalten)

→ Man muss abwiegen!

Fortsetzung des Beispiels:

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$ — kann man auch anders formulieren $\rightarrow$ $H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$

H_1 muss nicht die verneinte Aussage zu H_0 sein!
Sie muss H_0 nur ausschließen.

In diesem Fall wäre eine „große Abweichung" nur nach oben möglich"
↓
große Abweichung von der Prüfgröße $\bar{X}$ von μ_0

Der Ablehnbereich ist dann nur der rechte Rand der $N(\mu_0 ; \frac{\sigma^2}{n})$ - Verteilung von $\bar{X}$. Daraus folgt:

$\rightarrow$ $H_0 : \mu \leq \mu_0$

$H_1 : \mu > \mu_0$

Einseitigkeit & Zweiseitigkeit

Ein Testproblem heißt:

- zweiseitig wenn

  $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$

- einseitig wenn

  $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$

  bzw.

  $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$

einseitig!

> In der Praxis sind die Tests standardisiert und der Ablehnbereich auch vorgegeben.

Der Test wird dann Gauß-Test oder z-Test oder u-Test genannt.

Gauß-Test

für μ bei bekanntem σ : (Normalverteiltes Merkmal)

Schritte:

1. Formuliere Hypothesen

   a) $H_0: \mu = \mu_0$ gegen $H_1: \mu \neq \mu_0$

   b) $H_0: \mu \geq \mu_0$ gegen $H_1: \mu < \mu_0$

   c) $H_0: \mu \leq \mu_0$ gegen $H_1: \mu > \mu_0$

2. Wähle Signifikanzniveau α

3. Stichprobe vom Umfang n ziehen, $\bar{x}$ berechnen und den zugehörigen standardisierten Prüfwert,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$ → ist normalverteilt!

4. Bestimme entsprechendes Quantil

   a) $z_{1-\frac{\alpha}{2}}$

   b) $z_{1-\alpha}$

   c) $z_{1-\alpha}$

5. Entscheidungsregel:

   H_0 ist zu verwerfen wenn

   a) $|z| > z_{1-\frac{\alpha}{2}}$

   b) $z < -z_{1-\alpha}$

   c) $z > z_{1-\alpha}$

→ Wichtig:

Gleicher Vorgang wenn nicht normalverteilt, aber $n \gtrsim 30$

Beispiel:

Gauß-Test:

Umfang $n = 10$ $\quad \sigma = 2$ $\quad H_0: \mu \geq 100$

$\bar{x} = 98,9$ $\quad H_1: \mu < 100$

$\alpha = 0,01$

Vorgang:

$\bar{x} = 98,9$

$$z = \frac{98,9 - 100}{2/\sqrt{10}} = -1,739$$

$$z_{1-0,01} = z_{0,99} = 2,326$$

Daraus folgt:

Ablehnungsbereich: $z < -2,326$

Da z nicht im Ablehnungsbereich liegt, wird H_0 behalten.

Frage:

Ab welchem Signifikanzniveau müsste die Nullhypothese bei dieser Stichprobe verworfen werden?

Sobald $z = -1,739 < -z_{1-\alpha}$ liegt

$\hookrightarrow -1,739 = -z_{1-\alpha} \; (= z_\alpha)$

$$\alpha = \Phi(-1,739) = 1 - \Phi(1,739) = 0,041$$

Conclusion:

Würde man $\alpha \geq 0,041$ vorgeben so müsste H_0 verworfen werden in diesem Fall

Dieses berechnete Wert heißt eigentlich p-Wert

P-Wert

Wahrscheinlichkeit vom Signifikanzniveau, ab der die aktuelle Lösung für H_1 verworfen werden müsste.

„Die Entscheidungsregel lautet:
H_0 wird für H_1 verworfen, falls der P-Wert kleiner als α ist"

→ P-Wert $< \alpha$ bedeutet: bei einem kleineren α Wert müsste H_0 verworfen werden.

t-Test

für μ bei unbekanntem σ. (Normalverteiltes Merkmal)

1. Formuliere Hypothesen

   a) $H_0: \mu = \mu_0$ gegen $H_1: \mu \neq \mu_0$

   b) $H_0: \mu \geq \mu_0$ gegen $H_1: \mu < \mu_0$

   c) $H_0: \mu \leq \mu_0$ gegen $H_1: \mu > \mu_0$

2. Wähle Signifikanzniveau α
3. Ziehe Stichpr. mit Umfang n, berechne $\bar{x}$, s und Prüfwert t:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

4. Bestimme Quantil in T-Verteilung

   a) $t_{n-1;\,1-\frac{\alpha}{2}}$ b) c) $t_{n-1;\,1-\alpha}$

5. Entscheidungsregel: H_0 ist zu verwerfen falls

   a) $|t| > t_{n-1;\,1-\frac{\alpha}{2}}$

   b) $t < -t_{n-1;\,1-\alpha}$

   c) $t > t_{n-1;\,1-\alpha}$

wenn $n \geq 30$ ist die Verteilung der Merkmale egal

χ^2-Test

für σ^2 (normalverteiltes Merkmal) → NUR möglich wenn normalverteilt

1. Formuliere Hypothese.

   a) $H_0: \sigma^2 = \sigma_0^2$ $\quad H_1: \sigma^2 \neq \sigma_0^2$

   b) $H_0: \sigma^2 \geq \sigma_0^2$ $\quad H_1: \sigma^2 < \sigma_0^2$

   c) $H_0: \sigma^2 \leq \sigma_0^2$ $\quad H_1: \sigma^2 > \sigma_0^2$

2. Wähle Sign. α

3. Berechne Prüfwert

   $$y = \frac{n-1}{\sigma_0^2} s^2$$

   → Prüfgröße Y ist χ^2-verteilt mit $n-1$ Freiheitsgrade

4. Bestimme Quantillen der χ^2-Verteilung

   a) $\chi^2_{n-1;\frac{\alpha}{2}}$ und $\chi^2_{n-1;1-\frac{\alpha}{2}}$

   b) $\chi^2_{n-1;\alpha}$

   c) $\chi^2_{n-1;1-\alpha}$

5. Entscheidungsregel: H_0 verwerfen falls

   a) $y < \chi^2_{n-1;\frac{\alpha}{2}}$ oder $y > \chi^2_{n-1;1-\frac{\alpha}{2}}$

   b) $y < \chi^2_{n-1;\alpha}$

   c) $y > \chi^2_{n-1;1-\alpha}$

χ^2 - Anpassungstest

Testet ob eine bestimmte Form der Verteilung vorliegt:
(setzt große Stichprobe voraus)

Wir bilden Partitionen und vergleichen Σ der Ist- und Soll-Werte in jeder Partition!

Partitionen A_i mit $1 \leq i \leq k$

zugehörige Wahrsch. $p_i = P(X \in A_i)$

⇒ Stichprobe vom Umfang n

Vergleich von:

- beobachteter Anzahl: n_i
- erwarteter Anzahl: $n \cdot p_i$

$$Y = \sum_{i=1}^{k} \frac{(n_i - n p_i)^2}{n p_i} = \frac{1}{n} \left(\sum_{i=1}^{k} \frac{n_i^2}{p_i} \right) - n$$

⇒ $Y \overset{a}{\sim} \chi^2(k-1)$ für $n \longrightarrow \infty$

1. Formuliere Hypothesen

   $H_0 : P(X \in A_i) = p_i$ für $1 \leq i \leq k$

   $H_1 : P(X \in A_i) \neq p_i$ für mindestens ein i

2. Wähle α
3. Ziehe Stichprobe vom Umfang n, stelle Anzahl von A_i (k - Partitionen) fest.

   Berechne Y ⇒ näherungsweise χ^2-verteilt mit $k-1$ Freiheitsgraden falls $n p_i \gtrsim 5$ für alle i

4. Quantil der χ^2-Verteilung:

   $\chi^2_{k-1;1-\alpha}$

5. Entscheidungsregel:

   H_0 verwerfen, falls $Y > \chi^2_{k-1;1-\alpha}$

t-Distribution

$X_1 \dots X_n$

$X_1 \sim N(\mu, \sigma^2)$

$$\cancel{\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)}$$

$$\boxed{\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1)}$$

SEM

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

density

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}}$$

$x \in \mathbb{R}$

> t-Distributed with $(n-1)$ degrees of freedom

law of large numbers:
better estimation as n grows

R commands:
rt()
dt()
t.test()

Example

$\mu_0 = 30$

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \rightarrow T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

estimation!

$\bar{x} = 16,5$

$\alpha = 0,05$

$n = 16$

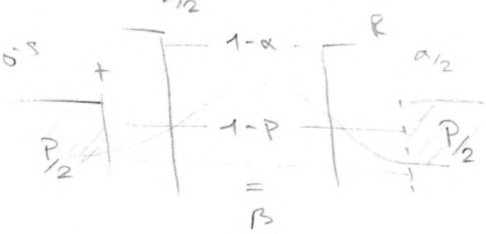
$H_0: \mu = 30 = \mu_0$

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \approx \frac{16,5 - 30}{8,7 / 4} \approx -6,2$$

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \overset{H_0}{\sim} t(15)$$

$$p = \mathbb{P}_{H_0}(|T| \geq |t|) \approx 1,5 \cdot 10^{-5}$$

$\underbrace{\qquad\qquad}_{\text{two sided test}}$



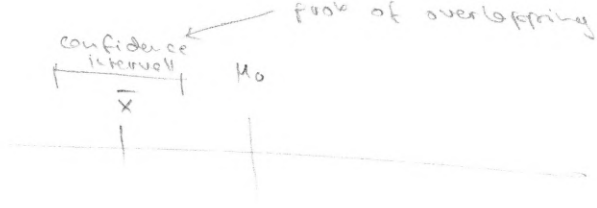
because $p \leq \alpha \Leftrightarrow t \in R$
the observed discrepancy was significant

Confidence intervall

$$I := \left(\bar{x} - q_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}} \; ; \; \bar{x} + q_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}}\right)$$

overlaps μ_0 with probability $1-\alpha$

fail of overlapping μ_0



When Data isn't bell shaped : Asymptotic normality of the mean

$n = 36$

$H_0 : \mu = \mu_0$

$\bar{X} \sim N(\mu, \sigma^2/n)$

central limit theorem

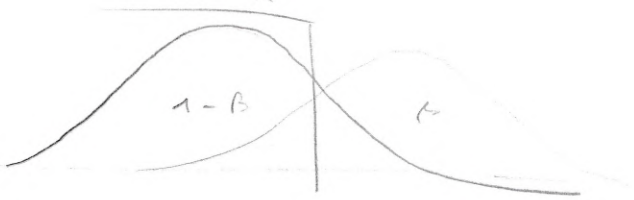
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \longrightarrow N(0,1) \quad \text{for } n \to \infty$$

Confidence intervall

$$I = \left[\bar{X} - q_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} , \bar{X} + q_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \right]$$

Errors and test powers:

Null hypothesis	H_0 rejected	H_0 not rejected
H_0 holds true	α error ($=\alpha$)	($= 1-\alpha$)
H_0 does not hold true	($= 1-\beta$)	β error ($=\beta$)

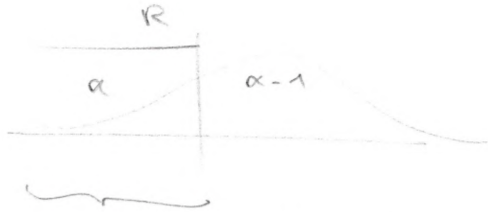


if H_0 does not hold true

β error:

H_0 is not rejected, although H_0 is not true

if H_0 does hold true



probability of rejecting $z = \alpha$ (falsely)

$P_{H_0}(Z \in R) = \alpha$

α error:

H_0 is rejected, although H_0 is true

Was ist der P-Wert?

Beispiel: Stichprobe

$n = 10$ $\quad H_0: \mu \geq 100$ (right-sided test)

$\sigma = 2$ $\quad H_1: \mu < 100$

$\bar{x} = 98{,}9$ $\quad \alpha = 0{,}01$

→ Ab welchem Signifikanzniveau müsste man H_0 verwerfen?

$$z = \frac{98{,}9 - 100}{\frac{2}{\sqrt{10}}} = -1{,}739$$

Das gesuchte Quantil

$z_{1-\boxed{0{,}01}} = z_{0{,}99} = 2{,}326$

Ablehnungsbereich

$R: z < -2{,}326$ bzw $(-\infty; -2{,}326] = R$

Da aber $-1{,}739 > -2{,}326$ behalten wir H_0

Wir suchen jenes α für das gelten würde:

$z = -1{,}739 < -z_{1-\alpha}$

Deshalb bestimmen wir:

$z = -1{,}739 = -z_{1-\alpha} = z_\alpha$ → $\alpha = \Phi(-1{,}739) =$

standard. CDF von der Normalverteil.

$\Phi(z_\alpha) = \alpha$ $\quad = 1 - \Phi(1{,}739) = 0{,}041$

gleicher Vorgang wie in Folien!

$p = 0{,}041 = 1 - \beta$

Bei einem Signifikanzniveau von $\geq 0{,}041$,

also wenn $\alpha > 0{,}041 = p$

muss man H_0 verwerfen

$P(Z \geq -1{,}739) = 0{,}041$

Hint:

$1 - \Phi(z) = \Phi(-z)$

$z_{1-\alpha} = -z_\alpha$

$-z_{1-\alpha} = z_\alpha$

Beispiel:

$\Phi(z_{0{,}2}) = 0{,}2$

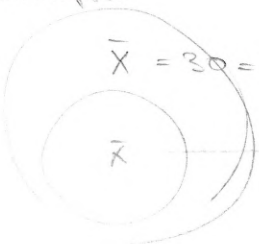
$\Phi(-z_{0{,}8}) = 1 - \Phi(z_{0{,}8}) = 1 - 0{,}8 = 0{,}2$

Basic ideas of hypothesis testing

if we have n iid random variables $X_1 \dots X_n$

$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \rightarrow \bar{X} \sim N(\mu, \sigma^2/\sqrt{n})$

Assumption



$\bar{X} = 30 = \mu_0 \leftarrow$ total population

$\rightarrow$ Determined through our sample

z-Test: unknown expectation but known variance

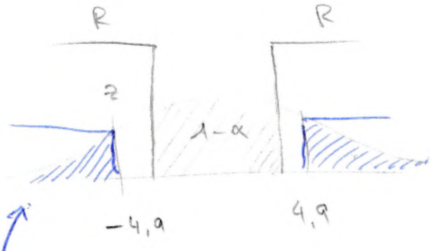
$n = 16$

$X_1 \sim N(\mu, 11^2)$

$H_0: \mu = 30 = \mu_0$ (our claim)

$H_1: \mu \neq 30 = \mu_0$

based on α



based on p

= p

= $1-\alpha$

$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = -4,9$

p-value: quantifies discrepency

probability of finding a value which is as extreme as z if H_0 holds true

$p = \boxed{\mathbb{P}_{H_0}(|Z| \geq |z|)} \approx 9 \cdot 10^{-7}$

If p is very small we say the observed discrepency was "significant" ($p < 10^{-6}$) and reject H_0

if H_0 is true, then something very unlikely has occured

$p = 1 - \beta$

$1 - p = 1 - (1-\beta) = 1 - 1 + \beta = \beta$

reject H_0 if

$p < \alpha \Rightarrow z \in R$

keep H_0 if

$p > \alpha \Rightarrow z \notin R$

Alternative point of view:

We reject H_0 if and only if $z \in R$

$P_{H_0}(Z \in R) = \alpha \qquad \alpha = 5\% \longrightarrow R \approx (-\infty, -1{,}96] \cup [1{,}96, \infty)$

Quantiles

Meaning if H_0 holds true we falsely reject with a probability of α

The point is:
While z is fixed, α decides how „strict" we are

$\alpha = 0{,}001$
$1-\alpha = 0{,}999$

$\alpha = 0{,}05$
$1-\alpha = 0{,}95$

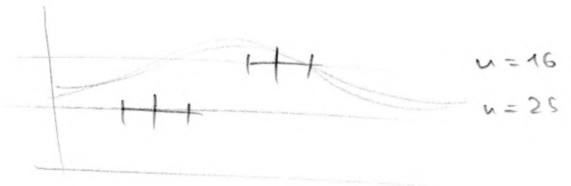
Two sided and one sided testing

two sided: $H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$
$R = (-\infty, q_{\frac{\alpha}{2}}] \cup [q_{1-\frac{\alpha}{2}}; \infty)$
$p = \mathbb{P}_{H_0}(|Z| \geq |z|)$

left sided: $H_0: \mu \geq \mu_0$
$H_1: \mu < \mu_0$
$R = (-\infty, q_\alpha]$
$p = P_{H_0}(Z \leq z)$

right sided: $H_0: \mu \leq \mu_1$
$H_1: \mu > \mu_0$
$R = [q_\alpha, \infty)$
$p = P_{H_0}(Z \geq z)$

Example : Judgement of the discrepancy (two sided test)

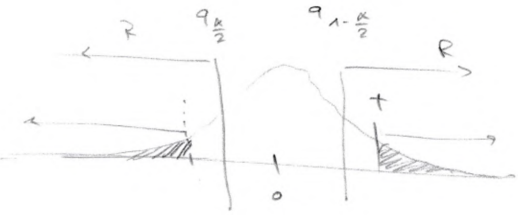


$T \sim t(27,3)$

$t \approx 3,1$ if H_0 holds true $\leftarrow$ $H_0: \mu_x = \mu_y$ $T \sim t(27,3)$

$\alpha = 0,05$ $t \approx 3,1$

$p = \mathbb{P}_{H_0}(|T| \geq |t|) \approx 0,004$



$t \in R \Leftrightarrow p \leq \alpha$

reject H_0 on the 5% - level

Interpretation:

A discrepancy as extreme as observed in the data appears in about only 4 out of 1000 cases; if the groups are taken from the same population

In this sense: the data speak against the H_0 hypothesis.

Generalization:

Let's put things in perspective

$$H_0: \mu_1 = \mu_2 \xrightarrow{\text{approximately}} T = \frac{\bar{Y} - \bar{X}}{\sqrt{SEM_y^2 + SEM_x^2}} \sim t(\nu)$$

lets try $d_0 \neq 0$

$H_0: d := \mu_2 - \mu_1 = 0$ (no difference)

$H_0: d = d_0$ (arbitrary difference)

Under $H_0: d = d_0$

$$T := \frac{\bar{Y} - \bar{X} - d_0}{\sqrt{SEM_y^2 + SEM_x^2}} \sim t(\nu)$$

Or put differently:

$$I := \left((\bar{Y} - \bar{X}) - q_{1-\alpha/2} \cdot \sqrt{SEM_y^2 + SEM_x^2} \,;\, (\bar{y} - \bar{x}) + q_{1-\frac{\alpha}{2}} \cdot \sqrt{SEM_y^2 + SEM_x^2} \right)$$

Interval overlaps d_0 with probability $1-\alpha$

(Das gesamte Ziel des Zweistichproben t-Tests ist es herauszufinden, wie wahrscheinlich es ist, dass unsere Stichproben von den selben Mittelwerten abhängen

↓

abhängig von dem Abstand zwischen $\bar{x}$ und $\bar{y}$)

Welch-Test vs. Student-Test

In the Student test we assume that:

$$\sigma_1^2 = \sigma_2^2$$

$$T = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{1}{n_2} + \frac{1}{n_1}} \cdot S_p} \sim t(n_2 + n_1 - 2)$$

S_p^2 is called pooled empirical variance

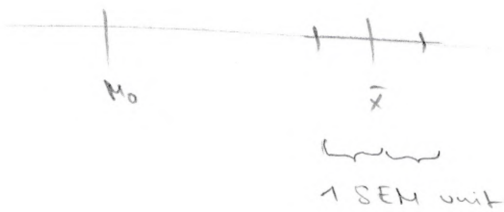
Advantages of Student: precision, degrees of freedom are easy to calculate

Disadvantages of Student: spread must be similar in both groups

Surrounding the two-sample t-test

Reminder: one sample

$n = 25$



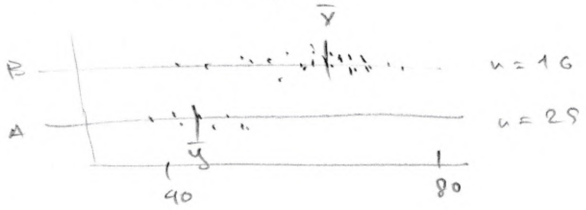
„Standard error of the mean" SEM

$$SEM = \frac{s}{\sqrt{n}}$$

derived from standard deviation of the mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Now with two samples:

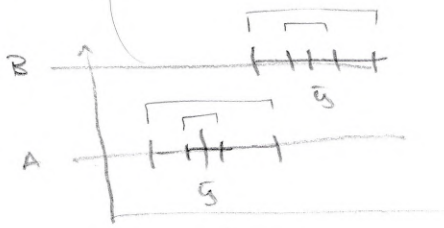


t measures discrepancy betw. values and mean through SEM units

$$T = \frac{\bar{x} - \mu_0}{SEM}$$

there now is a discrepancy between the sample values of the same total population.

How do we measure the discrepancy? → through SEM units
(between the sample means)



Small SEM => huge discrepancy

big SEM => small discrepancy

$$t = \frac{\bar{y} - \bar{x}}{\sqrt{sem_y^2 + sem_x^2}}$$

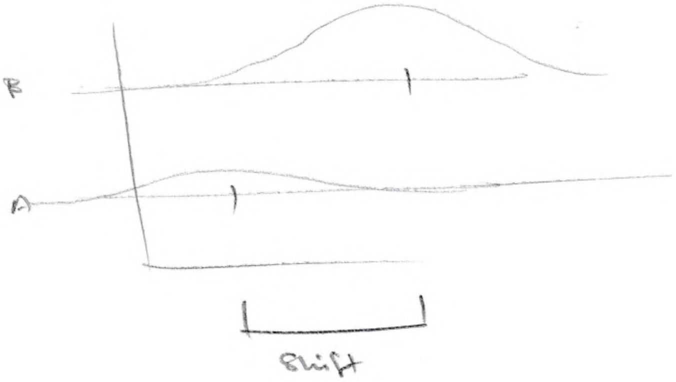
Discrepancy of the means, in relation to their standard errors

We need a Model:

Model: $X_1 \dots X_{n_1}$ $\quad X_i \sim N(\mu_1, \sigma_1^2)$

$y_1 \dots X_{n_2}$ $\quad Y_i \sim N(\mu_2, \sigma_2^2)$

- independent
- identically distributed
- normal distribution



each group has their own unknown parameters for the normal distribution

↓

There is the possibility for a shift.

Question

In our model: How unlikely is the observed discrepancy, (in example resp. $t \approx 3{,}1$) if H_0 holds true (there is no shift)

$H_0: \mu_1 = \mu_2$

$$T := \frac{\bar{Y} - \bar{X}}{\sqrt{SEM_y^2 + SEM_y^2}} \sim t(v)$$

t distribution, with v degrees of freedom

two sided:

$H_1: \mu_1 \neq \mu_2$

one sided - right:

$H_1: \mu_1 < \mu_2$

one sided - left

$H_1: \mu_1 > \mu_2$

(This is plausible, because if the samples are taken from the same population and $\mu_1 = \mu_2$ then

$$\frac{(\bar{Y} - \bar{X}) - (\mu_2 - \mu_1)}{\sqrt{\sigma_2^2/n_2 + \sigma_1^2/n_1}} \sim N(0,1)$$

But $\mu_2 - \mu_1 = 0$

usually approximated with S

One can roughly say:

$|t| \approx 1$ typical val

$|t| \gtrapprox 3$ rare

One sample t-test (Reminder)

„t-Test für μ bei unbekanntem σ eines normalverteilten Merkmals" (siehe Tegehl)

Hypothesis

two sided : $H_0: \mu = \mu_0$ $\quad H_1: \mu \neq \mu_0$

right sided : $H_0: \mu \geq \mu_0$ $\quad H_1: \mu < \mu_0$

left sided : $H_0: \mu \leq \mu_0$ $\quad H_1: \mu > \mu_0$

Question:

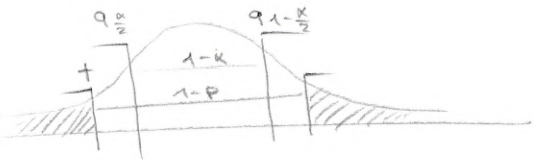
Is our mean too far away from μ ?

$$T = \frac{\bar{X} - \mu_0}{SEM} \qquad t = \frac{\bar{x} - \mu_0}{SEM} \leftarrow \frac{s}{\sqrt{n}} \qquad \underbrace{\sim t(\overbrace{n-1}^{\text{degrees of freedom}})}_{\text{Distribution}}$$

$$p = P_{H_0}(|T| \geq |t|)$$

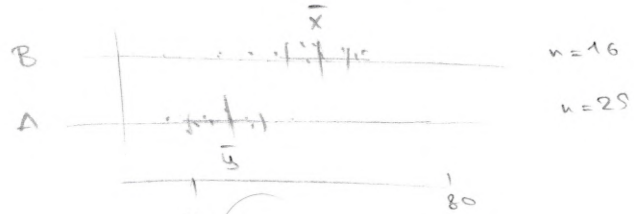
$p \leq \alpha \iff t \in R$

$p > \alpha \iff t \notin R$



$1 - p = \beta$

two sample t-test (Reminder)



Model:

$x_i \leftarrow X_i \sim N(\mu_1, \sigma_1^2)$

$y_i \leftarrow \underbrace{Y_i}_{} \sim N(\mu_2, \sigma_2^2)$

iid normal variables

$$t = \frac{\bar{y} - \bar{x}}{\sqrt{SEM_y^2 + SEM_x^2}} \qquad T = \frac{\bar{Y} - \bar{X}}{\sqrt{SEM_Y^2 + SEM_X^2}} \sim t(\nu)$$

$\uparrow$

(R knows ν)

Degrees of freedom

Question:

How unlikely is the observed discrepancy between $\bar{x}$ and $\bar{y}$ if $\mu_1 = \mu_2$

(Ho holds true)?

no shift

||

(our samples are from the same population)

Hypothesis

two sided : $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 \neq \mu_2$

right sided : $H_0: \mu_1 \geq \mu_2$ $\quad H_1: \mu_1 < \mu_2$

left sided : $H_0: \mu_1 \leq \mu_2$ $\quad H_1: \mu_1 > \mu_2$

Why is it plausible, that its a t-distribution?

$$\frac{(\bar{Y}-\bar{X}) - (\mu_2 - \mu_1)}{\sqrt{\sigma_2^2/n_2 + \sigma_1^2/n_1}} \sim N(0,1)$$

Because $\bar{X}$ and $\bar{Y}$ are independent, they can be standardized!

We dont have these values and therefore get the t-distribution

Since we assume that $\mu_1 = \mu_2$ this expression vanishes

Generalized:

$H_0: \mu_1 = \mu_2 \longrightarrow H_0 \; d = d_0 (:= \mu_2 - \mu_1) = 0$ no difference

$H_0 \; d = d_0 = x \in \mathbb{R}$ arbitrary difference

Alternative: Student test

Until now: Welch-test

In the student test we assume that:

$\sigma_1^2 = \sigma_2^2$

$$T = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{1}{n_2} + \frac{1}{n_1} \cdot S_p}} \sim t(n_2 + n_1 - 2)$$

S_p^2 „empirical variance" must be equal in both samples

Advantage: degrees of freedom is easy to calculate

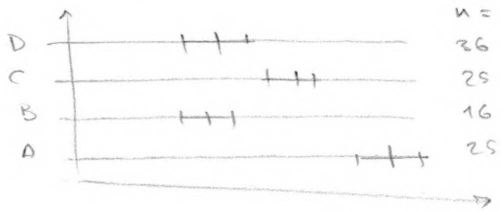
ANOVA:
Analysis of variance

We learned from the t-statistic, with two groups that:

↑ discrepancy between means $\Leftarrow$ ↓ variability within groups
of groups

Now with more groups:

$k = 4$



First approach: pairwise t-tests

number of pairs $m = \binom{k}{2}$ here: $\binom{4}{2} = 6$ t-tests, $d_0 = 0$ (no shift)

groups:	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
p-value:	✗	✓	✓	✓	✗	✗

„The multiple testing problem"

- too many pairs to test $O(n^2)$
- probability of α-error cumulates! $1-(1-\alpha)^m$ → almost 1 after 100 pair-tests

↓

Bonferroni correction

a) correct α: $\alpha^* = \frac{\alpha}{m}$

b) correct p: $p^* = m \cdot p$

Second approach: Analysis of Variances ANOVA

Test all groups „simultaniously"

$k = 4$

$m = 6$ pairwise tests

normal distrib.

Model:

group 1 $X_{1,1} \dots X_{1,n_1}$

$X_{2,1} \dots X_{2,n_2}$

$\vdots$

group k $X_{k,1} \dots X_{k,n_k}$

iid (independent and identically distributed) $X_{i,j} \sim N(\mu_i, \sigma^2)$

with same constant σ^2 for all variables

(remember: students two sample t-test)

Hypothesis:

$H_0 : d = d_0 = 0$ for $\forall \mu_i$ ... $\mu_1 = \mu_2 = \mu_3 = \mu_4 \dots = \mu_k$

H_1 : at least one pair $\mu_i \neq \mu_j$

Testing:

F-statistic: „analysis of variances" → Fisher Distribution

↑ variability between groups ⇐ ↓ variability within groups

$$F = \frac{\text{'variability between groups'}}{\text{'variability within groups'}} = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{X}_{i,\cdot} - \bar{X}\right)^2}{\frac{1}{n-k} \sum_{i,j} \left(X_{i,j} - \bar{X}_{i,\cdot}\right)^2}$$

→ mean of group i − global mean

→ group point − group mean

where:

index of group $i = 1, \dots, k$

mean of ith group:

$$\bar{X}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}$$

global mean:

$$\bar{X} = \frac{1}{n} \sum_{i,j} X_{i,j}$$

$(n = n_1 + n_2 + \dots + n_k)$

F-Tests are always one sided:

f Large → reject H_0

$$F \overset{H_0}{\sim} F(k-1, n-k)$$

numerator, denominator] degrees of freedom