

Univariat, deskriptiv
dichotom - binär

nominal - keine Ordnung

ordinal - Ordnung

diskret - ~~ab~~ endlich/abzählbar

stetig - überabzählbar

absolute Häufigkeit

$$h_j$$

relative Häufigkeit

$$f_j$$

Histogramm

Intervalle: $[c_0, c_1], [c_1, c_2], \dots$

Breite: $d_j = c_j - c_{j-1}$

Höhe: $\frac{h_j}{d_j}$ oder $f_j d_j$

Unimodal: ein Gipfel

Multimodal: mehrere Peaks

n : Anzahl Merkmalsträger

x_i : Wert i -ter Träger

(x_1, \dots, x_n) Urliste

Empirische Verteilungsfkt.

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq x\}}, x \in \mathbb{R}$$

(das Stufen-Dinkis)

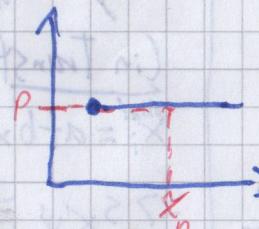
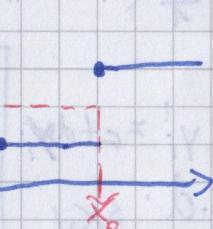
Empirisches Quantil

pe($0,1$), p-Quantil

$$\tilde{x}_p = \begin{cases} x_0 & , n \neq h < np + 1, np \in \mathbb{N} \\ \frac{1}{2}(x_h + x_{h+1}) & , h = np, np \in \mathbb{N} \end{cases}$$

Weiteres

Wie groß muss ich sein, damit der Anteil der Größe p kleiner ist als ich



Lagemasse

$$\text{Mittel: } \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{Median: } \tilde{x}_p = \begin{cases} x_{\lceil \frac{n+1}{2} \rceil} & \text{in gerade} \\ \left(\frac{1}{2} x_{\lfloor \frac{n}{2} \rfloor} + x_{\lceil \frac{n}{2} \rceil} \right) & \text{in ungerade} \end{cases}$$

Eigenschaften Mittel

Schwerpunkt:

$$\sum_{i=1}^n (x_i - \bar{x}_n) = 0$$

Minimum von:

$$f(t) = \sum_{i=1}^n (x_i - t)^2, t \in \mathbb{R}$$

Gepoolt: \bar{x}_{n_1} Mittel x_{11}, \dots, x_{n_1} ; \bar{y}_{n_2} Mittel y_{11}, \dots, y_{n_2}

$$\bar{x}_{n_1+n_2} = \frac{n_1}{n_1+n_2} \bar{x}_{n_1} + \frac{n_2}{n_1+n_2} \bar{y}_{n_2}$$

Eigenschaft Minimum Median:

$$f(t) = \sum_{i=1}^n |x_i - t|, t \in \mathbb{R}$$

Streuungsmaße $\sum_{i=1}^n (x_i - \bar{x}_n)^2 \cdot f(x_i)$

$$\text{Varianz: } s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

$$\text{Standartabw.: } s_n = \sqrt{s_n^2}$$

Spezielle Quantile

$$\tilde{x}_{0,25}: 1\text{-Quantil}; \tilde{x}_{0,75}: 3\text{-Quant.}$$

①

Interquantil-abstand (IQR)

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

Variationskoeff.

$$V = \frac{s_n}{\bar{x}_n}, (\bar{x}_n > 0)$$

Lin Transform

$a, b \in \mathbb{R}$, y_i : Lin Trans. von x_i :

$$y_i = a + b x_i$$

$$\Rightarrow (i) \tilde{y}_{0,5} = a + b \tilde{x}_{0,5}$$

$$(ii) \bar{y}_n = a + b \bar{x}_n$$

$$(iii) s_{n,y}^2 = b^2 s_{n,x}^2$$

Skalentrans.

$b \in \mathbb{R}$, $y_i = b x_i$

$$\Rightarrow V_y = V_x$$

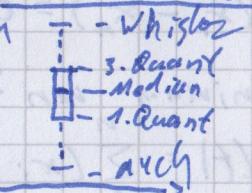
Kontingenzvariablen

h_{ij} : Abz. Häufigkeit (x_i, y_j)

$$h_{i \bullet} = \sum_{j=1}^k h_{ij} = h(x_i)$$

$$\textcircled{2} \quad h_{\bullet j} = h(y_j) = \sum_{i=1}^k h_{ij}$$

Box- & Whisker-Plot



Whisker: 0,05 & 0,95 Quant

in R: min, max Wert mit Abst. von max 1,5 x IQR von 1,3 - &

Bivariate deskriptive Statistik

Kontingenztabelle

	y_1	\dots	y_j	\dots	y_n	
x_1	h_{11}	\dots	h_{1j}	\dots	h_{1n}	$h_{1 \bullet}$
x_i	h_{i1}	\dots	h_{ij}	\dots	h_{in}	$h_{i \bullet}$
x_k	h_{k1}	\dots	h_{kj}	\dots	h_{kn}	$h_{k \bullet}$
	$h_{\bullet 1}$	\dots	$h_{\bullet j}$	\dots	$h_{\bullet n}$	n

Abs: Relat. Häufigkeit (x_i, y_j)

$$f_{ij} = \frac{h_{ij}}{n}$$

$$f_{i \bullet} = \frac{h_{i \bullet}}{n}$$

$$f_{\bullet j} = \frac{h_{\bullet j}}{n}$$

Bedingte Häufigkeiten

$$f(x_i | y_j) = \frac{h_{ij}}{h_{i \bullet}} = \frac{f_{ij}}{f_{i \bullet}}$$

"Häufigkeit von x_i unter y_j "

$$\sum_{j=1}^k f(x_i | y_j) = 1 = \sum_{i=1}^n f(x_i | y_j)$$

empirische Korrelation

(Bravais & Pearson)

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x s_y}} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y}_n)^2}}$$

$$(1) -1 \leq r_{xy} \leq 1$$

$$(2) |r_{xy}| = 1 \Leftrightarrow \exists a, b \in \mathbb{R} \text{ mit } y_i = a + b x_i$$

\Rightarrow linearer Zusammenhang

Rangkorrelation (Spearman)

$R(x_i)$: Rang von x_i

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

(emp) Korrelation

(x_i, y_j) biv. Merkmal

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

(in Transf.

$$x'_i = a + b x_i; y'_i = c + d y_i$$

$$\Rightarrow s_{x'y} = b \cdot d \cdot s_{xy}$$

$$R_{SP} = \frac{\sum_{i=1}^n (R(x_i) - \bar{R}(x))(R(y_i) - \bar{R}(y))}{\sqrt{\sum_{i=1}^n (R(x_i) - \bar{R}(x))^2} \cdot \sqrt{\sum_{i=1}^n (R(y_i) - \bar{R}(y))^2}}$$

$R_{SP} = 1$, wenn monoton

m

empirische Kor.: lineare Zusammenhänge

Rangkor.: monotone Zusammenhänge

Eigenschaften:

$$(i) f_{i \bullet} = \sum_{j=1}^k f_{ij}$$

$$(ii) \sum_{i=1}^n f_{i \bullet} = \sum_{j=1}^k f_{\bullet j} = 1$$

Wahrscheinlichkeiten

Ergebnisraum

Ω : Menge aller möglichen Ergebnisse eines Experiments

Ereignis: Teilmenge von Ω

Elementarere.: $\{\omega\} \subset \Omega$

Wahrscheinlichkeit (Laplace)

A : Menge günstiger Ereignisse

$$P(A) = \frac{|A|}{|\Omega|}$$

Voraussetzung: Jedes Elementarereig. gleich wahrscheinlich

Wahrscheinlichkeit (empirisch)

n : Anzahl Wiederholungen

$h(A)$: Absolute Anzahl Auftreten A

$$P(A) = \lim_{n \rightarrow \infty} \frac{h(A)}{n}$$

Axiome v. Kolmogorow

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

$$2. P(\Omega) = 1$$

$$3. P(A \cup B) = P(A) + P(B), \text{ falls } A \cap B = \emptyset$$

Rechenregeln:

(1) Komplement

$$P(\bar{A}) = 1 - P(A)$$

(2) Differenz:

$$P(A \setminus B) = P(A) - P(A \cap B)$$

(3) Additionsatz:

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

(4) Satz v. d. totalen Wahrsch.:

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

Bedingte Wahrscheinlichkeit

Wahrscheinlichkeit für A unter Bedingung B:

$$P(A|B)$$

Laplace-Raum:

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

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

$$P(\emptyset|B) : \{A : A \subset \Omega \Rightarrow \emptyset, 1\}$$

Rechenregeln bedingte Wahrscheinlichkeiten

(1) Multiplikationssatz:

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

(2) Satz v. d. totalen Wahrscheinlichkeit

$$P(A) = P(A|B) \cdot P(B) + P(A|\bar{B}) \cdot P(\bar{B})$$

(3) Satz v. Bayes:

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

Unabhängige Ereignisse

A, B unabhängig, falls:

$$P(A|B) = P(A|\bar{B}) = P(A)$$

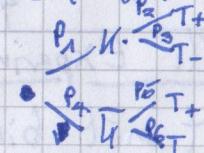
Gütekriterien:

Sensitivität: $P(T_+|K)$

T_+ : Test Positiv
 T_- : Test Negativ

Spezifität: $P(T_-|\bar{K})$

K: Krank



$$\begin{aligned} P_1 &= P(K) \\ P_2 &= P(\bar{K}) \\ P_3 &= P(T_+|K) \\ P_4 &= P(T_-|K) \\ P_5 &= P(T_+|\bar{K}) \\ P_6 &= P(T_-|\bar{K}) \end{aligned}$$

(3)

Diskrete Verteilungen

Zufallsvariable:

Abb. $X: \Omega \rightarrow \mathbb{R}$, d.h. jedem Elementarereignis ω wird eine reelle Zahl zugeordnet

diskret: Wertebereich abzählbar

stetig: Wertebereich nicht überabzählbar

Zähldichte (Wahrscheinlichkeitsfkt.)

$f: \Omega \rightarrow [0, 1]$
Bildet Zufallsvar. auf Wahrscheinlichkeit ab

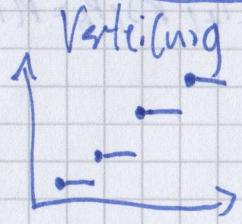
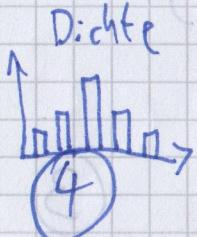
$$\text{z.B. } f(x_i) = \frac{|x_i|}{|\Omega|} \quad f(x) = P(X=x)$$

Verteilungsfunktion:

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

speziell: diskrete Variab.

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



Eigenschaften Verteilungsfkt.

- (i) $F(x)$ definiert $\forall x \in \mathbb{R}$
- (ii) $0 \leq F(x) \leq 1$
- (iii) $F(x)$ monoton wachsend
- (iv) $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$

Unabhängigkeit v. Zufallsvariablen

X & Y unabhängig g.d.w.

$$P(X=x_1, Y=y) = P(X=x_1) \cdot P(Y=y) \quad \forall x_1, y \in \Omega$$

$$f(x, y) = f(x) \cdot f(y), \text{ mit } f$$

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

"gemeingemeine Dichte"

Faltung

X, Y unabhängig mit Zähldichten $f, g \Rightarrow$ Zähldichte h d. Summe

$$h(u) = P(X+Y=u) = \sum_j f(j) \cdot g(u-j)$$

$$= \sum_i f(i) \cdot g(u-i)$$

Bernoulli Experiment

Experiment mit binären Ausgang

$$X = \begin{cases} 1, & \text{A tritt ein} \\ 0, & \text{A tritt nicht ein} \end{cases}$$

Erfolgswahrscheinlichkeit $p = P(X=1)$

$$\Rightarrow X \sim \text{Bernoulli}(p)$$

Zähldichte Bernoulli

$$f(x) = p^x \cdot (1-p)^{1-x}$$

Binomialverteilung

$X = X_1 + X_2 + \dots + X_n$ binomialverteilt, gdw.

(1) X_i binomialverteilt

(2) $p = \text{const. } \forall X_i$

(3) X_i, X_j unabh. $\forall i \neq j$

$$\Rightarrow X \sim \text{Bin}(n, p)$$

Dichte Binomialverteilung

$$f(k) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \forall k \in \{0, \dots, n\}$$

Erwartungswert einer diskreten Zufallsvar.

$T = \{x_1, x_2, \dots\}$ Wertebereich von X

$$\Rightarrow E(X) = \sum_i x_i \cdot f(x_i)$$

Varianz

$$\text{Var}(X) = \sum_i (x_i - E(X))^2 \cdot f(x_i)$$

Spezial: Bernoulli

$$E(X) = 1 \cdot p + 0 \cdot (1-p) = p$$

$$\text{Var}(X) = (1-p)p$$

Spezial: Binom

$$E(X) = n \cdot p$$

$$\text{Var}(X) = n \cdot p \cdot (1-p)$$

Rechenregeln Erwartungswert & Varianz

Erwartungswert:

$$(1) E(aX) = \sum_i ax_i \cdot f(x_i) = a \cdot E(X)$$

$$(2) E(aX+b) = a \cdot E(X) + b$$

$$(3) E(X+Y) = E(X) + E(Y)$$

Varianz:

$$(1) \text{Var}(aX) = a^2 \cdot \text{Var}(X)$$

$$(2) \text{Var}(aX+b) = a^2 \cdot \text{Var}(X)$$

$$(3) \text{Var}(X) = E((X-E(X))^2) = E(X^2) - (E(X))^2$$

Poisson-Verteilung

ZfV. X mit ~~Wahrs.~~ $k \in \mathbb{N}$

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

$$\Rightarrow X \sim Po(\lambda)$$

$$\begin{aligned} \Rightarrow E(X) &= \lambda \\ \text{Var}(X) &= \lambda \end{aligned}$$

Für $X \sim \text{Bin}(n, p)$, großes n , kleines p

$$\Rightarrow X \sim Po(\lambda), \lambda = n \cdot p$$

(5)

Noch mehr Rechenregeln Erwartungswert

$$(4) E(g(X)) = \sum_i g(x_i) f_X(x_i), g: \mathbb{R} \rightarrow \mathbb{R}$$

$$(5) E(aX+b) = aE(X) + b$$

$$(6) E(X) = c, \text{ falls Zähldichte symmetrisch um } c, \text{ d.h. } f(c-x) = f(c+x) \forall x$$

Poisson-Prozess

X_t : Anzahl der Ereignisse im Zeitraum $[0, 0+t]$, $t > 0$.

X_t : Poisson-Prozess mit Intensität λ , falls $\exists: t \quad X_t \sim Po(\lambda t)$

$$\Rightarrow P(X_t = k) = \frac{(t\lambda)^k}{k!} e^{-t\lambda}$$

Wahrs. Eintreten prop. zu t

Wahrs. Eintreten in verschiedenen Zeitintervallen unabhängig: a .

Geometrische Verteilung

X nimmt $k \in \mathbb{N}$ mit Wahrscheinlichkeit

$$P(X=k) = p(1-p)^{k-1}, p \in [0, 1]$$

$$\Rightarrow X \sim G(p)$$

$$\Rightarrow E(X) = \frac{1}{p}, \text{Var}(X) = \frac{1-p}{p^2}$$

$X = \text{"Anzahl Versuche bis zum Eintreten"}$

Varianz

(6) ~~Var(aX+bY)~~

$$(4) \text{Var}(aX+bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$$

falls X, Y unabh.

Stetige Zufall Vertr. Verteilungen

Dichte

stetig integrierbare Fkt $f: \mathbb{R} \rightarrow [0, \infty)$, mit

$$\int_a^b f(x) dx = P(a \leq X \leq b)$$

heißt Dichtefkt.

$$\Rightarrow \int_{-\infty}^{\infty} f(x) dx = 1$$

Verteilungs fkt:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx$$

Eigenschaften:

$$(i) P(a \leq X \leq b) = F(b) - F(a)$$

$$(ii) P(X \leq a) = 1 - F(a)$$

$$(iii) \lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$$

Quantil:

q-Quantil:

$$q \in [0, 1]$$

$\Rightarrow F(x_q) = q$
ist das q-Quantil

Zufallsvektor

$X: \Omega \rightarrow \mathbb{R}^n$ mit $w \rightarrow X(w) = (X_1(w), \dots, X_n(w))$
heißt n-dim Zufallsvektor $X = (X_1, \dots, X_n)$

Multivariate Verteilungsfkt.

$$F_X((x)) = F_X(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$$

Quantilfkt.

Wenn $F(x)$ stetig streng monoton
 \Rightarrow Umkehrfunktion $F^{-1}(q)$ heißt
Quantilsfkt. Für festes q
heißt $F^{-1}(q)$ q-Quantil

Erwartungswert

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

Varianz

$$\text{Var}(x) = \sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x) dx$$

Multivariate Verteilungen

Unabhängigkeit

X, Y unabhängig gdw.

$$\begin{aligned} F_{(X,Y)}(x, y) &= P(X \leq x, Y \leq y) = \\ &= P(X \leq x), P(Y \leq y) \\ &= F_X(x) \cdot F_Y(y) \\ \forall x, y \in \mathbb{R} \Leftrightarrow f(x, y) &= f_X(x) f_Y(y) \end{aligned}$$

Kovarianz

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Rechenregeln:

- (1) $\text{Var}(x) = \text{Cov}(x, x)$
- (2) $\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$
- (3) X, Y unabh. $\Rightarrow \text{Cov}(X, Y) = 0$
- (4) $\text{Cov}(ax + b, cy + d) = ac \text{Cov}(X, Y)$
- (5) $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$

(Zähldichte eines Zufallsvektors)

$$X = (X_1, \dots, X_n) \text{ n-dim ZV}$$

Dichtefkt:

a) X_1, \dots, X_n stetige ZV, d.h. $f: \mathbb{R}^n \rightarrow [0, \infty)$

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_1 \dots dx_n = P(X_1 \leq x_1, \dots, X_n \leq x_n)$$

b) ~~stetige~~ $f: \mathbb{R}^n \rightarrow [0, \infty)$:

$$f(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n)$$

Eigenschaften:

$$a) \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_1 \dots dx_n = 1$$

$$b) \sum \dots \sum f(x_1, \dots, x_n) = 1$$

Zd: Kovarianz

$$\begin{aligned} \text{Cor}(X, Y) &= \int \int (x - E(x))(y - E(y)) f(x, y) dx dy \\ &= \sum \sum (x_i - E(x))(y_j - E(y)) f(x_i, y_j) \end{aligned}$$

Gleichverteilung

Gleichverteilung auf $[0, 1]$

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

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

Gleichverteilungallgemein: Intervall $[a, b]$

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

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

Normalvert.

Normalvert. definiert durch

 μ : Lage Max σ : Breit Kurve, halber Abst. Wende pkt.

$$\Rightarrow N(\mu, \sigma^2)$$

Dichtefkt

$$f_{\mu, \sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-(x-\mu)^2/(2\sigma^2)}$$

Verteilungsfkt

$$F_{\mu, \sigma}(x) = \int_{-\infty}^x f_{\mu, \sigma}(t) dt$$

Bem: nur mittels num. Int.

Erwartungswert/Varianz

$$E(X) = \mu$$

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

Standardisierung

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

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

Berech. P mit Standard.

$$F_{N(\mu, \sigma^2)}(x) = P(X \leq x) = P\left(\frac{x-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = F_{N(0, 1)}\left(\frac{x-\mu}{\sigma}\right)$$

Wichtige Wahrsch.

$$x \sim N(0, 1)$$

$$P(-1 < X < 1) = 0,84 - 0,16 = 0,68$$

$$P(-2 < X < 2) \approx 0,95$$

$$P(-3 < X < 3) \approx 0,997$$

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

$$P(\mu - \sigma < Y < \mu + \sigma) = 0,68$$

$$P(\mu - 2\sigma < Y < \mu + 2\sigma) = 0,95$$

analog

unabh. normat. ZV.Eigenschaften:

$$X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2), a \text{ konst.}$$

$$- aX \sim N(a\mu_X, a^2\sigma_X^2)$$

$$- X+Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Exponentialvert.

X stetig, positiv (0,0)

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

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

$$X \sim \text{Exp}(\lambda)$$

Erwartungswert & Varianz

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

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

Gedächtnislosigkeit: $P(X > x+t | X > x) = P(X > t)$

Ausfallrate: „Rate Ausfall zu einem festen Zeitpunkt“

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

Grenzwertsätze

(nächste Seite)

(7)

große Stichproben ($n \rightarrow \infty$)

Mittel:

$$\bar{X}_n = \frac{1}{n} \sum_i^n X_i \rightarrow E(X_i)$$

emp. Verteilungsfkt:

$$F_n(x) = \frac{1}{n} \sum_i^n \mathbb{1}_{\{X_i \leq x\}} \rightarrow F(x)$$

relative Häufigkeit:

$$h_n(A) = \frac{1}{n} \sum_i^n \mathbb{1}_{\{X_i \in A\}} \rightarrow P(X_i \in A)$$

(falls alle X_i unabh. & ident. verteilt.)

Konvergenz

fast sichere Konvergenz

Folge $(X_n)_{n \geq 1}$ vom ZV konvrgt
fast sichergen X , falls gilt:

$$P(\lim_{n \rightarrow \infty} X_n = x) = P(\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = x(\omega)\}) = 1$$

$$\Rightarrow X_n \xrightarrow{\text{f.s.}} x$$

äquiv:

$$\lim_{m \rightarrow \infty} P(\sup_{n \geq m} |X_n - x| > \epsilon) = 0 \quad \forall \epsilon > 0$$

⑧

Stochastische Konvergenz

$(X_n)_{n \geq 1}$ konv. stoch. gegen X , wenn:

$$\lim_{n \rightarrow \infty} P(|X_n - x| > \epsilon) = 0 \quad \forall \epsilon > 0$$

$$\Rightarrow X_n \xrightarrow{P} x$$

Es gilt:

$$X_n \xrightarrow{\text{f.s.}} x \Rightarrow X_n \xrightarrow{P} x$$

(nur die eine Richtung)

Konv. in Verteil. (schwache Konv.)

$$\lim_{n \rightarrow \infty} P(X_n \leq x) = P(X \leq x) \quad \forall x \in \mathbb{R} \text{ mit } P(X \leq x) \text{ stetig}$$

$$\Rightarrow X_n \xrightarrow{D} x$$

Es gilt:

$$X_n \xrightarrow{P} x \Rightarrow X_n \xrightarrow{D} x$$

Erwartungswert & Varianz

X_1, \dots, X_n unabh. & id. vrt ZV
mit $E(X_i) = \mu$ & $\text{Var}(X_i) = \sigma^2$
dann gilt:

$$(1) E(\bar{X}_n) = \mu$$

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

Schwaches Gesetz d. großen Zahlen

$(X_i)_{i \geq 1}$ Folge von ZV mit
 $E(X_i) = \mu$ & $\text{Var}(X_i) = \sigma^2$

(A) $X_i \wedge X_j$ paarw. konv.:

$$\text{cov}(X_i, X_j) = 0 \quad \forall i \neq j$$

$$\text{Var}(X_i) \leq M < \infty$$

V(B), X_i iid.

$$\Rightarrow \bar{X}_n \xrightarrow{P} \mu$$

Insbesondere (für A):

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{M}{n\epsilon^2} \quad \forall \epsilon > 0$$

starkes Gesetz d. g. Zahlen

$(X_i)_{i \geq 1}$ Folge iid. mit $E(X_i) = \mu$
 $\text{Var}(X_i) = \sigma^2 < \infty \quad \forall i$

$$\Rightarrow \bar{X}_n \xrightarrow{D} \mu$$

Relative Häufigkeiten

$(X_i)_{i \geq 1}$ Folge iid.

$$h_n(A) = \frac{1}{n} \sum_i^n \mathbb{1}_{\{X_i \in A\}} \xrightarrow{\text{f.s.}} P(X \in A)$$

Erwartungswert & Varianz

X_1, \dots, X_n iid ZV mit $E(X_i) = \mu$ & $\text{Var}(X_i) = \sigma^2$

$$(1) E(\sum_i^n X_i) = n\mu$$

$$(2) \text{Var}(\sum_i^n X_i) = n\sigma^2$$

Zentraler Grenzwertsatz $(X_i)_{i \geq 1}$ Folge iid. mit

$$E(X_i) = \mu < \infty \wedge \text{Var}(X_i) = \sigma^2 < \infty$$

$$\Rightarrow Z_n = \frac{1}{\sigma \sqrt{n}} \sum_{i=1}^n X_i - n\mu \xrightarrow{D} N(0, 1)$$

Konvergenz arithm. Mittel geg. $N(0, 1)$
 $(X_i)_{i \geq 1}$ Folge iid mit $E(X_i) = \mu$ und $\text{Var}(X_i) = \sigma^2$

$$\tilde{Z}_n = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{D} N(0, 1)$$

Hauptsatz der Statistik

(Satz v. Glivenko-Cantelli)

 $(X_i)_{i \geq 1}$ Folge iid. mit Verteilung $F(x)$:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

$$\text{mit } F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}$$

emp. Verteil.

Signifikanzniveau

α : Wahrscheinlichkeit, dass die Nullhypothese fälschlicherweise verworfen wird

- Maß für stat. Sicherheit

- l.d. Med: $\alpha = 5\%$ - ggf. auch $1\%, 10\%$

Zusammenfassung:

Mittel:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{a.s.} E(X_i) = \mu$$

und $\bar{X}_n \xrightarrow{D} N(\mu, \frac{\sigma^2}{n})$

rel Häufigkeit:

$$h_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \in A\}} \xrightarrow{a.s.} P(X \in A)$$

emp Verteil.:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

Schließende StatistikAlternativhypothese: H_1

Hypothesen, die nachgewiesen werden soll.

Nullhypothese: H_0
 Gegenstück der Nullhypothese

Prüfgröße T (Teststatistik)

Eine ZV $T = T(X_1, \dots, X_n)$, ~~die auf~~ ~~der~~ ~~aus~~ ~~der~~ ~~Verteilung~~ unter H_0 bekannt ist.

Ablehnungs-^(R) & Annahmebereich ^(A)T in R: H_0 verwirft, H_1 annahm.T in A: H_0 beibeh., H_1 kann n.spezial: H_0 „Gleichheit“ b. ar. (Symmetrie)

$$R = \{T < c_u\} \cup \{T > c_o\}$$

$$A = \{c_u \leq T \leq c_o\}$$

$$\Rightarrow P_{H_0}(T < c_u) \leq \alpha/2 \quad \text{und} \quad P_{H_0}(T > c_o) \leq \alpha/2$$

$$\Rightarrow c_u = \inf(P_{H_0}(T \leq t) \geq \alpha/2) \quad \text{und} \quad c_o = \inf(P_{H_0}(T \geq t) \geq 1 - \alpha/2)$$

$$\Rightarrow c_u = q_{\alpha/2} \quad \text{und} \quad c_o = q_{1 - \alpha/2}$$

„Entdeckungswahrscheinlichkeit“ (POWERT!!)Power: Wahrscheinlichkeit unter H_1 , dass der Test H_0 verwirft

$$\Leftrightarrow P_{H_1}(T \in R)$$

Fehler

	Entscheidung	
	H_0	H_1
H_0	V	1. Art
H_1		2. Art

(9)

zu Fehler:

Fehler 1. Art: fälschliches Verwerfen der Nullhypothese

Fehler 2. Art: fälschliches Beibehalten

Fehlerwahrscheinlichkeit

1. Art: $P_{H_0}(H_0 \text{ verworfen}) \leq \alpha$

kontrolliert durch Sig.-niv.

2. Art: $\beta = P_{H_1}(H_0 \text{ beib.}) = 1 - P_{H_1}(H_0 \text{ verw.})$

β beliebig groß

$(1-\beta) = P_{H_1}(H_0 \text{ verw.}) \rightarrow \text{POW/AH}!!!$

-P-Wert

Wahrscheinlichkeit unter H_0 einen

Wert so groß oder „Extrem“ als gemessen zu erhalten

Entscheidung mittels p-Wert

$p \leq \alpha$ H_0 verworfen

$p > \alpha$ H_0 beibehalten

(10)

wichtige spezielle Testverfahren

χ^2 -Verteilung

X stetig ZV

$$f(x) = \begin{cases} \frac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{x}{2}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

χ^2 -Vert mit n Freiheitsgraden
 $X \sim \chi_n^2$. $\Gamma(K)$ Gammafkt.

Zusammenhang mit norm. Vrt:

X_1, \dots, X_n unabh., standardnorm. vrt.
 $\Rightarrow X_1^2 + \dots + X_n^2 \sim \chi_n^2$

Approximativer Test

Hypothese (zweiseitig):

$H_0: \pi = \pi_0$ v. $H_1: \pi \neq \pi_0$

Prüfgröße:

$T = \#\text{Erfolge} = \sum_i X_i \sim \text{Bin}(n, \pi)$

$\stackrel{\text{approx}}{\sim} N(n\pi, n\pi(1-\pi))$

standardisierte Prüfgröße:

$$\tilde{T} = \frac{T - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} \stackrel{\pi_0}{\sim} N(0,1)$$

Vert. von \tilde{T}

für bel. π : $\tilde{T} \sim N(\mu, \sigma^2)$

$$\mu = \frac{n(\pi - \pi_0)}{\sqrt{n\pi_0(1-\pi_0)}}$$

$$\sigma^2 = \frac{\pi_0(1-\pi)}{\pi_0(1-\pi_0)}$$

Power

POW/AH-Berechnung

POW/AH!!!: $P_{H_1}(H_0 \text{ verw.})$

$$= P_{\pi \neq \pi_0}(\tilde{T} < c_\alpha) + P_{\pi \neq \pi_0}(\tilde{T} > c_\beta)$$

Fallzahl (Berechnung)

Annahme: wahre Rate $\pi_1 \neq \pi_0$
(z.B. $\pi_1 = 0,6$)

Festlegung: zu erreichte POW/AH!!!
 $1 - \beta$, idR $\beta = 0,2$

Ziel: bestimme n so, dass

$$\text{POW/AH}_{\pi_1} > 1 - \beta$$

Berechnung approx Fallzahl:

cloglog.sample.size(0.6, n=NULL, p=0.5, power=0.8)
 $\Rightarrow n = 187$

Berechnung POW/AH!!!:

binom.power(0.6, 187, 0.5, method
= c("asympt")
= c("exact"))

Konfidenzintervall für Parameter θ

Das Intervall $[\Theta_u(x), \Theta_o(x)]$ mit
 $P_\theta(\Theta \in [\Theta_u(x), \Theta_o(x)]) \geq 1-\alpha \quad \forall \theta$
mit x Datensetzer mit
Konfidenzintervall zum Niveau
von $1-\alpha$

Approx Konfidenzintervall für π

Punktschätzer $\hat{\pi} = \frac{1}{n} \sum_i x_i$

untere & obere Grenze konf. iden. int.

$$\pi_{\text{u}}(x) = \hat{\pi} - c \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

$$\pi_{\text{o}}(x) = \hat{\pi} + c \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

mit $c = \phi^{-1}(1 - \frac{\alpha}{2})$ (auch $\varepsilon_{\text{plus}}$ genannt)

($1 - \frac{\alpha}{2}$ -Quantil der Standardnorm. vert.)

$$\text{bzw: } 1-\alpha = 0,95 \Rightarrow c = \phi^{-1}(0,975) \approx 1,96$$

Aquivalenz Test & konf. int.

Test z. Signiniveau 5% \Leftrightarrow 95% Konf. int.

H_0 vs. $H_1 \Leftrightarrow p < 0,05 \Leftrightarrow H_0$ nicht im 95%-KI

H_0 beib. $\Leftrightarrow p \geq 0,05 \Leftrightarrow H_0$ im 95%-KI

$1-\alpha$ -Konf. int. für Param ϑ

$1-\alpha$: α -Irrtumswahrsch.

$\Rightarrow [G_u(x), G_o(x)]$, mit

$P_\theta(\vartheta \in [G_u(x), G_o(x)]) \geq 1-\alpha \quad \forall \vartheta$

ist konf. int. zum Niveau $1-\alpha$

Opt.eigenschaft Konf.int

unpräzise: Länge konf. int. minimal

präzise: Sei: ϑ^* wahr. Parameter:

$$P_\theta(\vartheta^* \in [G_u(x), G_o(x)]) \stackrel{!}{=} \min_{\vartheta \neq \vartheta^*} P_\theta(\vartheta \in [G_u(x), G_o(x)])$$

Punktschätzung

X_1, \dots, X_n iid P_θ , ϑ unbekannt:

$$\Rightarrow T = g(X_1, \dots, X_n)$$

heißt Punktschätzer für ϑ

Gute Schätzung

- schwcht um wahren Wert
- wenig Streuung

Erwartungstreue

$T = g(X_1, \dots, X_n)$ Punktsch.
 T erwartungstreu/unverzerrt, wenn

$$E_\theta(T) = \vartheta$$

Bias

$$\text{Bias}_\vartheta(T) = E_\theta(T) - \vartheta$$

$$E(X_i) = \mu, \text{Var}(X_i) = \sigma^2$$

die Schätze ist Normaldist.

Punktschätzer

Sichrige Schätzer

arit. Mittel: $\bar{X}_n = \frac{1}{n} \sum_i^n X_i$

Stichp.var.: $s_n^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2$

Spezial: X_i iid Bernoulli(π):

schätzt Rate: $\hat{\pi} = \frac{1}{n} \sum_i^n X_i$

$X_i \sim N(\mu, \sigma^2)$
 schätzer für $E = \mu$:

$$\frac{1}{n} \sum_i^n X_i$$

$E(T) = \text{immer für } X \mu \text{ einsetzen}$

(Brüche kann man raus ziehen)

Standardfehler (SE)

Standartabw.

$$SE_{\text{re}}(T) = \sqrt{\text{Var}_{\text{re}}(T)} = \sqrt{E_{\text{re}}[(T - E_{\text{re}}(T))^2]}$$

heißt auch Standardfehler

Mean Squared Error (MSE)

$$MSE_{\text{re}}(T) = E_{\text{re}}[(T - \vartheta)^2]$$

- falls Erwartungstreu:

$$MSE_{\text{re}}(T) = SE_{\text{re}}(T)^2$$

Allgemein gilt:

$$MSE_{\text{re}}(T) = SE_{\text{re}}(T)^2 + \text{Bias}_{\text{re}}(T)^2$$

Konsistenz

- stark konsistent:

$$T_n \xrightarrow{\text{f.s.}} \vartheta$$

- schwach konsistent:

$$T_n \xrightarrow{P} \vartheta$$

Likelihoodfunktion

X_1, \dots, X_n reell. ZV w.

gem. Dichtf. bzw. Wahrhf.

f_w(x_1, \dots, x_n), $\vartheta \in \Theta$ f. x_1, \dots, x_n

dazugeh. Realisierung.

\Rightarrow L. he.flt. $L(\cdot | x_1, \dots, x_n) : \Theta \rightarrow \mathbb{R}$

def durch

$$L(\vartheta | x_1, \dots, x_n) = f_w(x_1, \dots, x_n)$$

Maximum-Likelihood

Lösung $\hat{\vartheta} = \hat{\vartheta}(x_1, \dots, x_n)$

des Maximierungsprobs.

$$\hat{\vartheta} \rightarrow \max_{\vartheta \in \Theta}, \text{ d.h.}$$

$$L(\hat{\vartheta}) \geq L(\vartheta) \forall \vartheta \in \Theta$$

\Rightarrow Maximum-Likelihood-Schätzer:
 $\hat{\vartheta} = \hat{\vartheta}(x_1, \dots, x_n)$

Bemerkungen:

- Logarithmus Like.flt. heißt Log-Likeli.flt.
ans. und mit $L = \ln L$ bez.

- Aufgrund der Monotonie des Log gilt:

\bullet Max von $L(\vartheta) \Leftrightarrow$ Max von $L(\vartheta)$

kleinstes-Quadrat-Methode

y_1, \dots, y_n iid. mit $g(\vartheta) = E_{\text{re}}(Y_i)$

Erwartungswert y_1, \dots, y_n die Realisierungen

$$\Rightarrow \sum_i (y_i - g(\vartheta))^2 \xrightarrow{n \rightarrow \infty} \min$$

heißt kleinstes-Quadrat-Schätzer.

Lineare Regression

Modell

- $x_1, \dots, x_n \in \mathbb{R}$ geg. derm. Verte oder Realisierungen von X
oder nicht beob

- $\varepsilon_1, \dots, \varepsilon_n$ iid ZV mit $E(\varepsilon_i) = 0, \text{Var}(\varepsilon_i) = \sigma^2$

• y_1, \dots, y_n beobachtbare ZV mit
 $y_i = a + b x_i + \varepsilon_i$

~~•~~ y_i bilden Modell der einf.
(in. Reg.)

(ggf. nach ε_i anstellen...)