Deskriptive Statistik

Stichproben

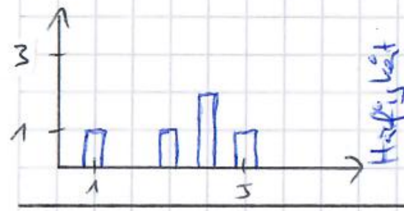
Merkmalstyp

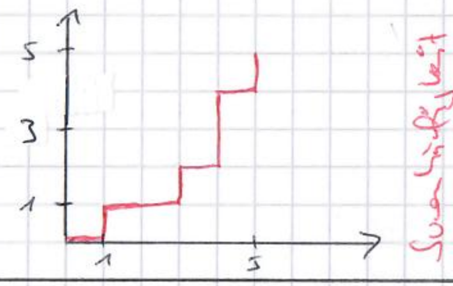
- Kategoriell
  - Ausprägung
  - endlich
  - nominal
    - Kategorie
    - ohne Reihenfolge
    - Bsp. Tiere
  - ordinal
    - Reihenfolge
    - Ränge
    - Bsp. Feedback
- metrisch
  - Zahlen
  - endlich / unendlich
  - diskret
    - endlich
    - abzählbar
    - Bsp. Einwohner
  - stetig
    - in Intervall
    - Bsp. Mietdauer

Absolute Häufigkeit

n = Total Stichprobe

$h(x)$ = Anzahl

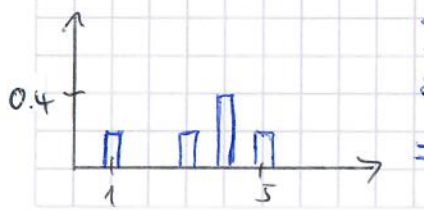




Relative Häufigkeit

in % ⇒ 0–100%
berechnet ⇒ 0–1

$$f(x) = \frac{h(x)}{n}$$

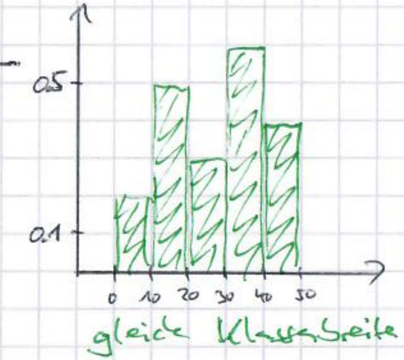


Klassierte Daten & Häufigkeit

i	x	b	n	$h = \frac{n}{b}$
1	[0, 10[	10	2	0.2
2	[10, 20[	10	5	0.5
3	[20, 30[	10	3	0.3
4	[30, 40[	10	6	0.6
5	[40, 50[	10	4	0.4

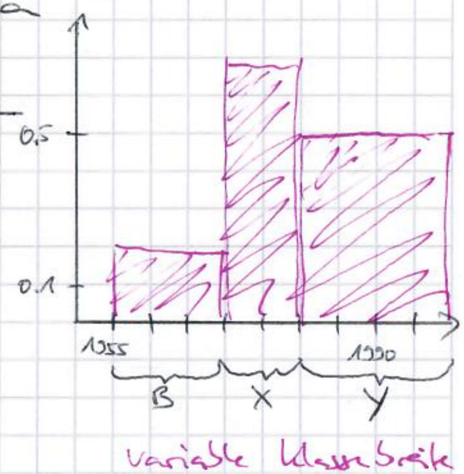
Bsp. zwischen 30 & 40
$= 10 \cdot 0.6 = 6$ Personen

Histogramm



i	x	b	n	$h = \frac{n}{b}$
B	[55, 70[	15	3	0.2
X	[70, 80[	10	7	0.7
Y	[80, 00[	20	10	0.5

Bsp. Baby Boomer
$= 15 \cdot 0.2 = 3$ Personen



Mittelwert = Durchschnitt

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

Bsp. [1, 7, 5, 4, 3]

$$\bar{x} = \frac{1}{5} \cdot (1+7+5+4+3)$$

$$= \frac{1}{5} \cdot 20 = \underline{4}$$

Median

Lagekennwert

$$x_{med} = \begin{cases} x_{[\frac{n+1}{2}]} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]}\right) & \text{falls } n \text{ gerade} \end{cases}$$

Quantile

nächst grössere ganze Zahl

$$R_q = \lceil n \cdot q \rceil$$

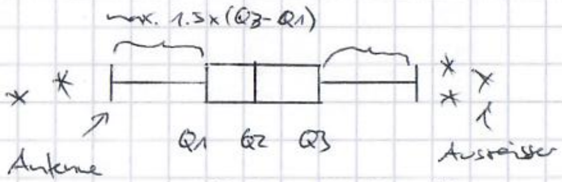
1. Quartil = 0.25 Quantil
2. Quartil = Median
3. Quartil = 0.75 Quantil

Bsp. [0, 1, 3, 3, 4, 4, 5]

3. Quartil = 0.75 Quantil
$n \cdot q = 7 \cdot 0.75 = 5.25 \Rightarrow 6$
$x_{[6]} = 4$

Interquartilsabstand = Q3 − Q1

Boxplot



Bsp. [0, 1, 3, 3, 4, 4, 5]

$Q1 = n \cdot q = 7 \cdot 0.25 = 2.75 \Rightarrow 3.$ Wert $\Rightarrow 3$
Median = 3
$Q4 = n \cdot q = 7 \cdot 0.75 = 5.25 \Rightarrow 6.$ Wert $\Rightarrow 4$
Interquartilsabstand = Q3 − Q1 = 1
5 ⇒ im oberen Whisker / Antenne
0 & 1 ⇒ untere Aussreisser

Modus

= Modalwert

Lagekennwert

Wert der am häufigsten Vorkommt

Streuungskennwerte

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

$$\text{Standardabweichung} = s = \sqrt{s^2}$$

$$\text{korrigierte Varianz} = s^2_{korr} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{korrigierte Standardabweichung} = s_{korr} = \sqrt{s^2_{korr}}$$

Pearson-Korrelationskoeffizient

(auch korrigiert möglich)

$$\left.\begin{array}{l} s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \end{array}\right\} \text{Standardabweichungen}$$

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

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} \qquad \text{Korrelationskoeffizient}$$

Bsp. $x = [1, 2, 3]$ & $y = [-1, 2, 4]$

$$\bar{x} = \frac{1}{3}(1+2+3) = 2 \qquad \bar{y} = \frac{1}{3}(-1+2+4) = \frac{5}{3}$$

$$s_x^2 = \frac{1}{3}\left[(1-2)^2 + (2-2)^2 + (3-2)^2\right] = \frac{2}{3} \qquad s_x = \sqrt{\frac{2}{3}}$$

$$s_y^2 = \frac{1}{3}\left[\left(-1-\frac{5}{3}\right)^2 + \left(2-\frac{5}{3}\right)^2 + \left(4-\frac{5}{3}\right)^2\right] = \frac{38}{9} \qquad s_y = \sqrt{\frac{38}{9}}$$

$$s_{xy} = \frac{1}{3}\left[(1-2)\left(-1-\frac{5}{3}\right) + (2-2)\left(2-\frac{5}{3}\right) + (3-2)\left(4-\frac{5}{3}\right)\right] = \frac{4}{3}$$

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{2 \cdot \sqrt{57}}{19}$$

Spearman-Rangkorrelationskoeffizient

zwischen Ränge & nicht Zahlen

Bsp.

	1	2	3	4	5	6	
x_i	59	35	43	23	42	27	
y_i	14.6	11.8	14.3	13.0	14.2	11.0	
$rg(x_i)$	6	3	5	1	4	2	$\boxed{\overline{rg(x)} = 3.5}$
$rg(x_i) - \overline{rg(x)}$	2.5	-0.5	1.5	-2.5	0.5	-1.5	
$rg(y_i)$	6	2	5	3	4	1	$\boxed{\overline{rg(y)} = 3.5}$
$rg(y_i) - \overline{rg(y)}$	2.5	-1.5	1.5	-0.5	0.5	-2.5	

$$r_{sp} = \frac{\sum (rg(x_i) - \overline{rg(x)})(rg(y_i) - \overline{rg(y)})}{\sqrt{\sum (rg(x_i) - \overline{rg(x)})^2} \cdot \sqrt{\sum (rg(y_i) - \overline{rg(y)})^2}}$$

$$= \frac{14.5}{\sqrt{17.5} \cdot \sqrt{17.5}} = 0.83$$

$\Rightarrow$ Fall gleiche Zahlen, dann Durchschnitt der Ränge nehmen

Kombinatorik

$n! = n\,(n-1)!$ Fakultät

Bsp. $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$

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

Variation (mit Reihenfolge)		Kombination (ohne Reihenfolge)	
mit W.	ohne W.	mit W.	ohne W.
n^k	$\frac{n!}{(n-k)!}$	$\binom{n+k-1}{k}$	$\binom{n}{k}$
Bsp. 1	Bsp. 2	Bsp. 3	Bsp. 4

} Wiederholung

Bsp. 1 $\Rightarrow$ Zahlenschlossproblem

6 Kränze & 10 Ziffern $\Rightarrow$ $\underline{10^6}$ Möglichkeiten

Bsp. 2 $\Rightarrow$ Schwimmwettkampf

3 Plätze & 10 Schwimmer $\frac{10!}{(10-3)!} = \underline{720}$ Möglichkeiten

Bsp. 3 $\Rightarrow$ Tell schiessen

3 Versuche & 11 Bereiche $\Rightarrow$ $\binom{13}{3} = \underline{286}$ Möglichkeiten

Bsp. 4 $\Rightarrow$ Lotto

6 aus 49 ohne Zurücklegen $\Rightarrow$ $\frac{49!}{43! \cdot 6!} = \underline{13'983'816}$ Möglichkeiten

diskrete Wahrscheinlichkeitsräume

Ergebnisraum $\Omega = \{\ldots\}$ (Omega)

Zähldichte $\rho = \ldots$ (Rho)

Bsp. Einmal würfeln mit fairem Würfel

$\Omega = \{1, 2, 3, 4, 5, 6\}$, Wahrscheinlichkeit für die Zahl 5 $\Rightarrow \rho = 1/6$

Laplace-Raum

$P(M) = \frac{|M|}{|\Omega|}$ $|M|$ = Anzahl Element der Menge M

Eigenschaften:

$P(\{\}) = 0$ unmögliches Ergebnis

$P(\Omega) = 1$ sicheres Ereignis

$P(\Omega \setminus A) = 1 - P(A)$ komplementäres Ereignis

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

$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + \ldots$

Sigma-Additivität

Zufallsvariablen

Am Beispiel vom zweimaligen Würfeln mit einem fairen Würfel.

Ergebnisraum $\Omega = \{(1,1), (1,2) \ldots (6,6)\}$

$\Rightarrow 6 \cdot 6$ Elemente $= 36$

Wahrscheinlichkeit für $\rho(1,1) = \rho(5,6) = 1/36$

$P(X=2) \Rightarrow$ Summe der beiden Würfel

$f(2) = P(X=2) = 1/36$

PDF

$f(2.3) = P(X=2.3) = 0$

$f(11) = P(X=11) = 2/36 = 1/18$

$f: \mathbb{R} \to [0,1], f(x) = P(X=x)$

$F(4) = P(X \le 4) = P(X=4) + P(X=3) + P(X=2) = 6/36 = 1/6$

CDF

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

$\vdots$

$P(7 \le X \le 10) = F(10) - F(6) = 33/36 - 15/36 = 1/2$

$P(X > 10) = 1 - F(10) = 1 - 33/36 = 1/12$

$P(X=3$ oder $X=7) = f(3) + f(7) = 2/36 + 6/36 = 2/9$

$P(X \neq 4) = 1 - f(4) = 1 - 3/36 = 11/12$

Kenngrössen

$\mu = E(X)$ = Erwartungswert

$\sigma^2 = V(X)$ = Varianz

$\sigma = S(X)$ = Standardabweichung

An einem Beispiel mit einer Zufallsvariabel

x	-1	0	1	2
$f(x)$	0.2	0.1	0.4	0.3

$$\mu = 0.2\cdot(-1) + 0.1\cdot 0 + 0.4\cdot 1 + 0.3\cdot 2 = \underline{0.8}$$

$$\sigma^2 = 0.2\,(-1-0.8)^2 + 0.1\,(0-0.8)^2 + 0.4\,(1-0.8)^2 + 0.3\,(2-0.8)^2 = \underline{1.16}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.16} = \underline{1.077}$$

Bedingte Wahrscheinlichkeiten

Wahrscheinlichkeit für B, wenn A eintritt

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

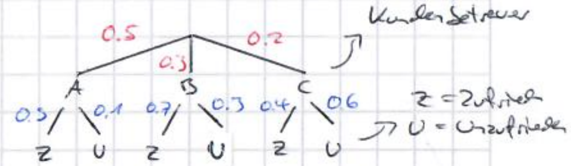
Multiplikationssatz

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

Satz von Bayes

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

Beispiel einer bedingten Wahrscheinlichkeit



Kundenbetreuungsschlüssel

Zufriedenheitsindex

W für A & Z

$$\Rightarrow P(A\cap Z) = P(A)\cdot P(Z|A) = 0.5\cdot 0.9 = \underline{0.45}$$

W für U

$$\Rightarrow P(U) = 0.5\cdot 0.1 + 0.3\cdot 0.3 + 0.2\cdot 0.6 = \underline{0.26}$$

W für Z & C

$$\Rightarrow P(C|Z) = \frac{P(C\cap Z)}{P(Z)} = \frac{0.4\cdot 0.2}{0.5\cdot 0.9 + 0.3\cdot 0.7 + 0.2\cdot 0.4} = \underline{0.108}$$

W für U & B

$$\Rightarrow P(B|U) = \frac{P(B\cap U)}{P(U)} = \frac{0.3\cdot 0.3}{0.26} = \underline{0.35}$$

Stochastische Unabhängigkeit

$\boxed{\text{abhängig}}$

stochastisch $\boxed{\text{unabhängig}}$ wenn $P(A\cap B) = P(A)\cdot P(B)$

Beispiel mit Würfel, einmaliges Würfeln

$$P(\text{grade} \cap \geq 3) = \frac{|\{4,6\}|}{|\Omega|} = \frac{1}{3}$$

$$P(\text{grade})\cdot P(\geq 3) = \frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$$

$$P(\geq 3 \cap 3\text{teilt}) = \frac{|\{3,6\}|}{|\Omega|} = \frac{1}{3}$$

$$P(\geq 3)\cdot P(3\text{teilt}) = \frac{2}{3}\cdot\frac{1}{3} = \frac{2}{9}$$

grade : $\{2,4,6\}$

≥ 3 : $\{3,4,5,6\}$

3teilt : $\{3,6\}$

wenn X & Y unabhängig

$$\Rightarrow E(X\cdot Y) = E(X)\cdot E(Y)$$

$$\Rightarrow V(X+Y) = V(X) + V(Y)$$

Diskrete Verteilungen

diskret => nur ganzzahlige Werte

Hypergeometrische Verteilung:

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

N = total Objekte
M = bestimmte Objekte
n = Stichprobengrösse
x = Anzahl bestimmter Objekte in Stichprobe

$\mu = E(X) = n \cdot \frac{M}{N}$ $\quad \sigma = S(X) = \sqrt{\sigma^2}$

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

Bernoulliverteilung: 1 oder 0

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

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

$\mu = E(X) = n\cdot p$ / $\sigma^2 = V(X) = n \cdot p \cdot q$

Poissonverteilung: Wahrscheinlichkeit für das Eintreten von einer bestimmten Anzahl gleichartiger Ereignisse

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

$\mu = E(X) = \lambda$ / $\sigma^2 = V(X) = \lambda$

Beispiel => 120 Anrufe pro Stunde, innerhalb einer Minute?

kein Anruf => $P(X=0) = \frac{2^0}{0!}e^{-2} = e^{-2} = \underline{0.135}$

1 Anruf => $P(X=1) = \frac{2^1}{1}e^{-2} = 2e^{-2} = \underline{0.271}$

stetige Verteilungen

stetig => im Intervall, auch nicht ganzzahlige Werte

Gaussverteilung / Normalverteilung: $\sigma > 0$

$$\varphi_{\mu\sigma}(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

wenn $\mu = 0$ & $\sigma = 1$

$$\Rightarrow \varphi(x) = \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}x^2} \quad \text{(Standardnormalverteilung)}$$

$E(X) = \mu$ / $V(X) = \sigma^2$ $\quad$ => Tabelle 1

Bsp. standardnormalverteilt $P(X \le 0.45) = \underline{0.6736}$
normalverteilt => $U = \frac{x-\mu}{\sigma}$

Zentraler Grenzwertsatz: $n \to \infty$

$S_n = \sum_{i=1}^{n} X_i$

$$U_n = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \qquad \lim_{n\to\infty} F_n(u) = \Phi(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-\frac{1}{2}t^2}\,dt$$

Beispiel => T_i 1 für $1<t<2$, sonst 0

$\mu = \underline{1.5}$ / $\sigma^2 = \int_1^2 \left(t-\frac{3}{2}\right)^2 \cdot 1\,dt = \left[\frac{t^3}{3} - \frac{3}{2}t^2 + \frac{9}{4}t\right]_1^2 = \underline{\frac{1}{12}}$

$T = 6\cdot 1.5 = \underline{9}$ / $\sigma^2 = 6 \cdot \frac{1}{12} = \underline{\frac{1}{2}}$

$P(T>10) = 1 - \Phi\left(\frac{10-9}{\sqrt{\frac{1}{2}}}\right) = 1 - \Phi(1.414) = \underline{\underline{0.0787}}$

Lineare Regression

Regressionsgerade $g(x) = m x + d$

Steigung $m = \frac{s_{xy}}{s_x^2}$

y-Achsenabschnitt $d = \bar{y} - m\bar{x}$

Residuenvarianz $s_\epsilon^2 = s_y^2 - \frac{s^2_{xy}}{s_x^2}$

Varianz x_i-Werte $s_x^2 = \left(\frac{1}{n}\sum x_i^2\right) - \bar{x}^2$

Varianz y_i-Werte $s_y^2 = \left(\frac{1}{n}\sum y_i^2\right) - \bar{y}^2$

Kovarianz $s_{xy} = \left(\frac{1}{n}\sum x_i y_i\right) - \bar{x}\cdot\bar{y}$

prognostizierte Varianz $s_{\hat{y}}^2 = s_y^2 - s_\epsilon^2$

Bestimmtheitsmass $R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$

= Korrelationskoeffizient^2 $R^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2}$ r_{xy}^2

Nicht lineares Verhalten => Linearisieren

$y = q \cdot x^m$	$\log(y) = \log(q) + m \cdot \log(x)$
$y = q \cdot m^x$	$\log(y) = \log(q) + \log(m) \cdot x$
$y = q \cdot e^{mx}$	$\ln(y) = \ln(q) + m \cdot x$
$y = \frac{1}{q + m \cdot x}$	$V = q + m \cdot x \,/\, V = \frac{1}{y}$
$y = q + m \cdot \ln(x)$	$y = q + m \cdot U \,/\, U = \ln(x)$

Schliessende Statistik

=> Blatt 7.3.3

Beispiel 95% - Vertrauensintervall:

$\sigma^2 = 118 \,/\, \mu$ für 95% = ?

i	1	2	3	4	5	6	7	8
x_i	140	162	128	136	148	147	151	128

$\bar{x} = \frac{1}{8}(140 + 162 + 128 + 136 + 148 + 147 + 151 + 128) = \underline{138}$

$e = c \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{10.86}{\sqrt{8}} = \underline{7.53}$

$\phi(c) = \frac{1+\gamma}{2} = \frac{1+0.95}{2} = 0.975$

Tabelle 2

$p = \phi(c) = 0.975 \Rightarrow \underline{c = 1.960}$

$\mu : [\bar{x} - e ; \bar{x} + e] = \underline{\underline{[130.47 ; 145.53]}}$

Beispiel Irrtumswahrscheinlichkeit $\alpha = 1\%$:

$n = 500 \,/\, x = 25 \,/\, \mu = ? \,/\, \gamma = 0.99$

Problem 4 $\Rightarrow n\hat{p}(1-\hat{p}) = 500 \cdot 0.05 \cdot 0.95 = 23.75 > 9$ ✓

$\hat{p} = \bar{x} = \frac{25}{500} = 0.05$

$p = \phi(c) = \frac{1+\gamma}{2} = 0.995 \Rightarrow \underline{c = 2.576}$

$e = c \cdot \sqrt{\frac{\bar{x} \cdot (1-\bar{x})}{n}} = 2.576 \cdot \sqrt{\frac{0.05 \cdot 0.95}{500}} = 0.0251$

$\mu : [\bar{x} - e ; x + e] = \underline{\underline{[0.0249 ; 0.0751]}}$