

Formelsammlung "Biometrie und Methodik"

1 Beschreibende Statistik

Univariate Datenanalyse	(Beobachtungswerte x_1, \dots, x_n)
Absolute Häufigkeit	$n_k =$ Anzahl der ω_i mit $x_i = a_k$
Relative H'keit/ Empirische Verteilung	$h_k = n_k/n$
Empirische Verteilungsfunktion	$F_n(x) = \sum_{a_k \leq x} h_k = \frac{1}{n} \cdot$ Anzahl der x_i mit $x_i \leq x$
Arithmetisches Mittel	$\bar{x} = \frac{1}{n} \sum_i x_i = \sum_k h_k a_k$
Empirische Varianz	$\sigma^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2 = \sum_k h_k (a_k - \bar{x})^2$
Stichprobenvarianz	$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 = \frac{n}{n-1} \sum_k h_k (a_k - \bar{x})^2$
Variationskoeffizient	$v = \sigma/\bar{x}$
Multivariate Datenanalyse	(Beobachtungswerte $(x_1, y_1), \dots, (x_n, y_n)$)
Absolute Häufigkeit	$n_{kl} =$ Anzahl der ω_i mit $x_i = a_k$ und $y_i = b_l$
Häufigkeiten der einzelnen Merkmale	$n_k^X = \sum_l n_{kl}$, $n_l^Y = \sum_k n_{kl}$
Relative H'keit/ Empirische Verteilung	$h_{kl} = n_{kl}/n$
Randverteilungen	$h_k^X = \sum_l h_{kl}$, $h_l^Y = \sum_k h_{kl}$
Bedingte relative Häufigkeiten	$h_{k l} = h_{kl}/h_l^Y$, $\tilde{h}_{l k} = h_{kl}/h_k^X$
Quadratische Kontingenz	$\chi^2 = \sum_{k,l} \frac{(n_{kl} - \hat{n}_{kl})^2}{\hat{n}_{kl}}$ mit $\hat{n}_{kl} = \frac{1}{n} n_k^X n_l^Y$
Mittlere quadratische Kontingenz	$\phi^2 = \frac{\chi^2}{n} = \sum_{k,l} \frac{(h_{kl} - h_k^X h_l^Y)^2}{h_k^X h_l^Y}$
Cramérsches Kontingenzmaß	$C = \sqrt{\frac{\phi^2}{\min(r-1, s-1)}} = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, s-1)}}$ (wobei r, s Zeilen-/Spaltenzahl in Kontingenztabelle)
Empirische Kovarianz	$c_{x,y} = \frac{1}{n} \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \overline{xy} - \bar{x} \cdot \bar{y}$ $c_{x,y} = \sum_k \sum_l h_{kl} \cdot (a_k - \bar{x}) \cdot (b_l - \bar{y})$
Korrelationskoeffizient	$\rho_{x,y} = \frac{c_{x,y}}{\sigma_x \sigma_y}$
Regressionsgerade	$f(x) = \hat{\alpha} + \hat{\beta}x$, $\hat{\beta} = \frac{c_{x,y}}{\sigma_x^2}$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Für **klassierte Daten** gelten die Formeln zur Berechnung von Mittelwert, empirischer Varianz und Kovarianz **näherungsweise**, wenn man für a_k den Klassenmittelpunkt, und für h_k die relative Häufigkeit der Klasse einsetzt.

2 Wahrscheinlichkeiten

Gegenereignis	$P[A \text{ tritt nicht ein}] = 1 - P[A \text{ tritt ein}]$
Additionssatz	$P[A \text{ tritt ein oder } B \text{ tritt ein}]$ $= P[A \text{ tritt ein}] + P[B \text{ tritt ein}] - P[A \text{ tritt ein und } B \text{ tritt ein}]$
Bedingte W'keit	$P[A B] = \frac{P[A \text{ und } B]}{P[B]} = \frac{P[A \text{ und } B \text{ treten beide ein}]}{P[B \text{ tritt ein}]}$
Fallunterscheidung	$P[A] = \sum_{i=1}^n P[A H_i] \cdot P[H_i]$ wobei H_i disjunkte Fälle, von denen genau einer eintritt
Satz von Bayes	$P[H_i A] = \frac{P[A H_i] \cdot P[H_i]}{\sum_{j=1}^n P[A H_j] \cdot P[H_j]} = \text{const.} \cdot P[A H_i] \cdot P[H_i]$ wobei H_i disjunkte Fälle, von denen genau einer eintritt
Multiplikativität	$P[A \text{ und } B \text{ treten ein}] = P[A] \cdot P[B A] = P[B] \cdot P[A B]$
Unabhängigkeit	$P[A \text{ und } B \text{ treten ein}] = P[A] \cdot P[B]$

3 Zufallsvariablen und ihre Verteilung

3.1 Diskrete Zufallsvariablen

Verteilung/Massenfkt.	$p_X(a_k) = P[X = a_k]$
Verteilungsfunktion	$F_X(y) = P[X \leq y] = \sum_{a_k \leq y} p_X(a_k)$
p-Quantil	$P[X < x_{(p)}] \leq p \leq P[X \leq x_{(p)}]$
Erwartungswert	$E[X] = \sum_k a_k \cdot p_X(a_k)$ $E[g(X)] = \sum_k g(a_k) \cdot p_X(a_k)$
Varianz	$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$
Standardabweichung	$\sigma(X) = \sqrt{Var(X)}$

3.2 Stetige Zufallsvariablen

Dichtefunktion	$P[a \leq X \leq b] = P[a < X < b] = \int_a^b f_X(x) dx$
Verteilungsfunktion	$F_X(y) = P[X \leq y] = \int_{-\infty}^y f_X(x) dx$
Zusammenhang	$f_X = F_X'$
p-Quantil	$F_X(x_{(p)}) = P[X \leq x_{(p)}] = p$
Erwartungswert	$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$ $E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) dx$
Varianz	$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$ $= \int_{-\infty}^{\infty} (x - E[X])^2 \cdot f_X(x) dx$
Standardabweichung	$\sigma(X) = \sqrt{Var(X)}$

3.3 Rechenregeln

Erwartungswert	$E[a \cdot X + b \cdot Y] = a \cdot E[X] + b \cdot E[Y]$
Varianz	$Var(a \cdot X) = a^2 \cdot Var(X)$ $\sigma(a \cdot X) = a \cdot \sigma(X)$ $Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$ X, Y unkorreliert $\Rightarrow Var(X + Y) = Var(X) + Var(Y)$
Standardisierung	$Y = \frac{X - E[X]}{\sigma(X)}$ ist ZV mit $E[Y] = 0$ und $\sigma(Y) = 1$

4 Spezielle Verteilungen

Gleichverteilung (auf $\{a_1, \dots, a_m\}$)	$p(a_i) = \frac{1}{m} = \frac{1}{\text{Anzahl der möglichen Werte}}$ $\text{Erw.wert} = \frac{m+1}{2}, \quad \text{Varianz} = \frac{m^2-1}{12}$
Empirische Verteilung (der Daten x_1, \dots, x_n)	$p(a_k) = \text{rel. Häufigkeit von } a_k \text{ unter } x_1, \dots, x_n$ $\text{Erw.wert} = \bar{x}_n = \frac{1}{n} \sum_i x_i$ $\text{Varianz} = \sigma_n^2 = \frac{1}{n} \sum_i (x_i - \bar{x}_n)^2$
Bernoulli (p)	$p(1) = p, \quad p(0) = 1 - p$ $\text{Erw.wert} = p, \quad \text{Varianz} = p \cdot (1 - p)$
Bin (n, p)	$p(k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} \quad (k = 0, 1, \dots, n)$ $\binom{n}{k} = \frac{n \cdot (n-1) \cdot \dots \cdot (n-k+1)}{k \cdot (k-1) \cdot \dots \cdot 1} = \frac{n!}{k!(n-k)!}$ $\text{Erw.wert} = n \cdot p, \quad \text{Varianz} = n \cdot p \cdot (1 - p)$
Hypergeom (\hat{n}, \hat{k}, n)	$p(k) = \frac{\binom{\hat{k}}{k} \cdot \binom{\hat{n}-\hat{k}}{n-k}}{\binom{\hat{n}}{n}} \quad (k = 0, 1, \dots, n)$ $\text{Erw.wert} = n \cdot \frac{\hat{k}}{\hat{n}}, \quad \text{Varianz} = n \cdot \frac{\hat{k}}{\hat{n}} \cdot \frac{\hat{n}-\hat{k}}{\hat{n}} \cdot \frac{\hat{n}-n}{\hat{n}-1}$
Poisson (λ)	$p(k) = \frac{1}{k!} \lambda^k \cdot e^{-\lambda} \quad (k = 0, 1, 2, \dots)$ $\text{Erw.wert} = \lambda, \quad \text{Varianz} = \lambda$
Exp (λ)	$f(t) = \lambda \cdot e^{-\lambda \cdot t} \quad \text{für } t > 0, \quad f(t) = 0 \quad \text{für } t \leq 0$ $\text{Erw.wert} = \frac{1}{\lambda}, \quad \text{Standardabw.} = \frac{1}{\lambda}$
N ($0, 1$)	$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ $\text{Erw.wert} = 0, \quad \text{Varianz} = 1$
N (m, σ^2)	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/2\sigma^2}$ $\text{Erw.wert} = m, \quad \text{Varianz} = \sigma^2$
Lineare Transformationen	$X \sim N(m, \sigma^2) \quad \implies \quad X + a \sim N(m + a, \sigma^2)$ $b \cdot X \sim N(b \cdot m, b^2 \cdot \sigma^2)$

5 Approximation von Verteilungen

Poissonapproximation der Binomialverteilung	Für große n und für p mit $n \cdot p = \lambda$ (also $p = \lambda/n$) gilt $\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \approx \frac{1}{k!} \lambda^k \cdot e^{-\lambda}$
Gesetz der großen Zahlen ($n \rightarrow \infty$)	Für X_1, X_2, \dots unabh., ident. vert., $E[X_i] = m$, $Var(X_i) = \sigma^2$ gilt: $\bar{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \dots + X_n) \rightarrow m$
Anwendung auf rel. H'keiten	$h_n(a) \rightarrow p(a)$
Zentraler Grenzwertsatz	Für X_1, X_2, \dots unabh., ident. vert., $E[X_i] = m$, $Var(X_i) = \sigma^2$ gilt: $\bar{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \dots + X_n) \approx N\left(m, \frac{\sigma^2}{n}\right)$ für große n
Normalapproximation der Binomialverteilung	$X \sim Bin(n, p)$, $Y = (X - np) / \sqrt{np(1-p)}$ $P[a \leq Y \leq b] \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \Phi(b) - \Phi(a)$
Anwendbarkeit (Faustregel)	$n \cdot p \geq 10$ und $n \cdot (1-p) \geq 10$
Normalapproximation für W'keiten der σ -Umgebungen	$P[X - E[X] \leq \sigma(X)] \approx 68.2\%$ (exakt für $X \sim N(0, 1)$) $P[X - E[X] \leq 2\sigma(X)] \approx 95.4\%$ (exakt für $X \sim N(0, 1)$) $P[X - E[X] \leq 3\sigma(X)] \approx 99.7\%$ (exakt für $X \sim N(0, 1)$)

6 Punktschätzer und Teststatistiken

Einstichprobenproblem	Statistiken	Verteilung unter H_0
Schätzung des Mittelwerts	$\bar{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \dots + X_n)$	
Schätzung der Varianz	$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$	
Gauss-Statistik	$Z_n = \frac{\bar{X}_n - m_0}{\sigma/\sqrt{n}}$	$\sim N(0, 1)$ im Gaussmodell
t-Statistik	$T = \frac{\bar{X}_n - m_0}{S_n/\sqrt{n}}$	$\sim t(n-1)$ im Gaussmodell
V-Statistik (Vorzeichentest)	$V =$ Anzahl der pos. Vorzeichen von $X_i - \mu_0$	$\sim Bin(n, 0.5)$ bei stet. Verteilung
W-Statistik (Wilcoxon-test)	$W =$ Summe der Ränge der pos. Differenzen $X_i - m_0$	s. Lit., Vorauss. stet. symm. Vert.
Verbundenes Zweistichprobenproblem	Statistiken ($U_i = X_i - Y_i$)	Verteilung unter H_0
Schätzung der Differenz der Mittelwerte	$\bar{U}_n = \bar{X}_n - \bar{Y}_n$	
Schätzung der Varianz	$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (U_i - \bar{U}_n)^2$	
Gauss-Statistik	$Z_n = \frac{\bar{U}_n}{\sigma/\sqrt{n}}$	$\sim N(0, 1)$ im Gaussmodell
t-Statistik	$T = \frac{\bar{U}_n}{S_n/\sqrt{n}}$	$\sim t(n-1)$ im Gaussmodell
V-Statistik (Vorzeichentest)	$V =$ Anzahl der pos. Vorzeichen von U_i	$\sim Bin(n, 0.5)$ bei stet. Verteilung
W-Statistik (Wilcoxon-test)	$W =$ Summe der Ränge der pos. Differenzen U_i	s. Lit., Vorauss. stet. Verteilungen unterscheiden sich nur in Lage
Unverbundenes Zweistichprobenprobl.	Statistiken	Verteilung unter H_0
Schätzung der Mittelwerte	$\bar{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \dots + X_n)$ $\bar{Y}_m = \frac{1}{m} \cdot (Y_1 + Y_2 + \dots + Y_m)$	
Gepoolte Schätzung der Varianz	$S_{pool}^2 = \frac{1}{n+m-2} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{i=1}^m (Y_i - \bar{Y}_m)^2 \right)$	
Gauss-Statistik	$Z_n = \frac{\bar{X}_n - \bar{Y}_m}{\sigma/\sqrt{n}}$	$\sim N(0, 1)$ im Gaussmodell
t-Statistik	$T = \frac{\bar{X}_n - \bar{Y}_m}{S_{pool} \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$	$\sim t(n+m-2)$ im Gaussmodell
U-Statistik (Mann-Whitney-U-Test)	$U =$ siehe Literatur (Rangstatistik)	s. Lit., Vorauss. stet. Verteilungen unterscheiden sich nur in Lage

Einfaktorielle Varianzanalyse

Schätzer für i-tes Gruppenmittel

Schätzer für Gesamtmittelwert

Quadratsumme zwischen Gruppen

Quadratsumme innerhalb der Gruppen

Mittlere Streuung zwischen Gruppen

Mittlere Streuung innerhalb Gruppen

F-Statistik

H-Statistik (Kruskal-Wallis-Test)

Statistiken

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{k} \sum_{i=1}^k n_i \cdot \bar{Y}_i$$

$$QS_b = \sum_{i=1}^k n_i \cdot (\bar{Y}_i - \bar{Y})^2$$

$$QS_w = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

$$MQS_b = \frac{1}{k-1} \cdot QS_b$$

$$MQS_w = \frac{1}{n-k} \cdot QS_w$$

$$F = \frac{MQS_b}{MQS_w}$$

 H = siehe Literatur
(Rangstatistik)**Verteilung unter H_0** $(n_i$ Daten y_{i1}, \dots, y_{in_i} in Gruppe i) $(k$ Gruppen) $\sim \chi^2(k-1)$ im Gaussmodell $\sim \chi^2(n-k)$ im Gaussmodell $\sim F(k-1, n-k)$ im Gaussmodells. Lit., Vorauss. stet. Verteilungen
unterscheiden sich nur in Lage**ANOVA-Tabelle:**

Quelle der Variation	QS	Freiheitsgrade	MQS	f
zwischen den Gruppen	QS_b	$k-1$	MQS_b	MQS_b/MQS_w
innerhalb der Gruppen	QS_w	$n-k$	MQS_w	
Gesamt	QS_t	$n-1$		

Anpassungs- und Unabhängigkeitstest Statistiken

Chi Quadrat-Statistik (Anpassungstest)

$$\chi^2 = \sum_{l=1}^r \frac{(n_l - \hat{n}_l)^2}{\hat{n}_l}$$

n_l = Häufigkeit der Klasse K_l
 \hat{n}_l = theoretische H'keit unter H_0

Chi Quadrat-Statistik (Unabhängigkeitstest)

$$\chi^2 = \sum_l \frac{(n_{kl} - \hat{n}_{kl})^2}{\hat{n}_{kl}}$$

n_{kl} = H'keit der Komb. (a_k, b_l)
 $\hat{n}_{kl} = \frac{1}{n} n_k^X n_l^Y$ theoret. H'keit

Verteilung unter H_0 $\approx \chi^2(r-1)$ für große n $\approx \chi^2((r-1) \cdot (s-1))$ für große n

7 Konfidenzintervalle (im Gaussmodell)

Einstichprobenproblem	Konfidenzintervall
Mittelwert bei bekannter Varianz	$\bar{X}_n \pm z_{1-\alpha/2} \cdot \sigma / \sqrt{n}$
Mittelwert bei unbekannter Varianz	$\bar{X}_n \pm t_{n-1, 1-\alpha/2} \cdot S_n / \sqrt{n}$
Verbundenes Zweistichprobenproblem	Konfidenzintervall ($U_i = X_i - Y_i$)
Differenz der Mittelwerte bei bekannter Varianz	$\bar{U}_n \pm z_{1-\alpha/2} \cdot \sigma / \sqrt{n}$
Differenz der Mittelwerte bei unbekannter Varianz	$\bar{U}_n \pm t_{n-1, 1-\alpha/2} \cdot S_n / \sqrt{n}$
Unverbundenes Zweistichprobenproblem	Konfidenzintervall
Differenz der Mittelwerte bei bekannter Varianz	$\bar{X}_n - \bar{Y}_m \pm z_{1-\alpha/2} \cdot \sigma \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}$
Differenz der Mittelwerte bei unbekannter Varianz	$\bar{X}_n - \bar{Y}_m \pm t_{n+m-2, 1-\alpha/2} \cdot S_{pool} \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}$

8 Hypothesentests

- U = Teststatistik (z.B. $U = Z, T, W, \chi^2$, etc.)
- u = beobachteter Wert der Teststatistik
- α = Signifikanzniveau
- u_α = α -Quantil der Teststatistik

Testentscheidung über Quantile	H_0 wird verworfen, falls
Rechtsseitige Alternative	$u \geq u_{1-\alpha}$
Linksseitige Alternative	$u \leq u_\alpha$
Beidseitige Alternative (allgemein)	$u \geq u_{1-\alpha/2}$ oder $u \leq u_{\alpha/2}$
Beidseitige Alternative (bei symmetrischer Verteilung von U mit Mittelwert 0)	$ u \geq u_{1-\alpha/2}$
Testentscheidung über p-Wert	H_0 wird verworfen, falls
Beliebige Alternative	$\alpha \geq p$
Berechnung des p-Werts	p-Wert
Rechtsseitige Alternative	$p = P_{H_0} [U \geq u]$
Linksseitige Alternative	$p = P_{H_0} [U \leq u]$
Beidseitige Alternative (bei symmetrischer Verteilung von U mit Mittelwert c)	$p = P_{H_0} [U - c \geq u - c]$

γ -Quantile der Student- t_n -Verteilung							
n	γ						
	0,90	0,95	0,975	0,99	0,995	0,999	0,9995
1	3,078	6,314	12,706	31,821	63,656	318,289	636,578
2	1,886	2,920	4,303	6,965	9,925	22,328	31,600
3	1,638	2,353	3,182	4,541	5,841	10,214	12,924
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,894	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,689
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,660
∞	1,282	1,645	1,960	2,327	2,577	3,092	3,293

Die letzte Zeile " ∞ " enthält die Quantile der Standard-Normalverteilung und gilt in guter Näherung für die t_n -Verteilung mit $n \geq 30$