Übersicht zur Statistik (Version 8. Mai 2023)

Roland Schäfer, FSU Jena, Germanistische Sprachwissenschaft

Hinweis: Hier wird eine verkürzte Indexschreibweise für Summen verwendet. $\sum_i$ ist zu lesen als $\sum_{i=1}^{n}$, wobei n jeweils das Maximum für i ist. Mehrfachindexierung wie in $\sum_{i,j}$ ist aufzulösen als $\sum_{i=1}^{n}\sum_{j=1}^{m}$ (also beide Laufvariablen durchzählen von 1 bis Maximum).

1 Deskriptive Statistik

Mittel (arithmetisch)	$\bar{x} = \frac{\sum_i x_i}{n}$
	`mean(x)`
Summen der Quadrate	$SQ(x) = \sum_i (x_i - \bar{x})^2$
	`sum((x-mean(x))^2)`
Varianz	$s^2(x) = \frac{SQ(x)}{n-1}$
	`var(x)`
Standardabweichung	$s(x) = \sqrt{s^2(x)}$
	`sd(x)`
Standardfehler (für Mittelwert $\bar{x}$)	$SF(x) = \frac{s(x)}{\sqrt{n}}$
	`sd(x)/sqrt(length(x))`
Standardfehler (für Anteilswert p)	$SF(p) = \sqrt{\frac{p \cdot (1-p)}{n}}$
	`sqrt(p*(1-p)/n)`
z-Wert für Messwert	$z(x_i) = \frac{x_i - \bar{x}}{s(x)}$
	`(x[i]-mean(x))/sd(x)`
Summe der Produkte	$SP(x,y) = \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})$
Kovarianz	$cov(x,y) = \frac{SP(x,y)}{n-1}$
	`cov(x,y)`
Pearson-Korrelation	$r_P(x,y) = \frac{cov(x,y)}{s(x) \cdot s(y)}$
(und Punkt-Biserielle Korrelation)	`cor(x,y)`
Signifikanztest für Pearson-Korrelation	$t = r\sqrt{\frac{n-2}{1-r^2}}$ mit $df = n - 2$

Spearman-Rang-Korrelation	$r_S = 1 - \frac{6 \sum_{i=1}^{n} (Rang(x_i) - Rang(y_i))^2}{n(n^2-1)}$
	`cor(x,y, method="spearman")`
Konfidenzintervall für Wert v (Anteil p oder Mittel $\bar{x}$)	$KI(v) = v \pm z \cdot SF(v)$

	`sf <-` … (s. o. für Anteil oder Mittel)
	`z <- qnorm(c(0.025, 0.975))`
	`p+sf*z`

2 Statistiken für Zähldaten

2.1 N-Felder-Tests

Die hier besprochenen Tests setzen wir typischerweise für **Zähldaten in Kreuztabellen** ein. Durchnumerierung von Zellen in Zeilen (i) und Spalten (j):

x_{11}	x_{12}
x_{21}	x_{22}

erwartete Häufigkeit (Zelle x_{ij})	$EH(x_{ij}) = \frac{\sum_k x_{ik} \cdot \sum_k x_{kj}}{n}$ für die Summe von Spalte i als $\sum_k x_{ik}$ und die Summe von Zeile j als $\sum_k x_{kj}$
χ^2	$\chi^2 = \sum_{ij} \frac{(x_{ij} - EH(x_{ij}))^2}{EH(x_{ij})}$
	`x <- matrix(c(x11, x21, x12, x22), 2)` `xq <- chisq.test(x)`
Cramérs ϕ (Effektstärke für 2×2-Tabellen)	$\phi = \sqrt{\frac{\chi^2}{n}}$
	`sqrt(xq$statistic/sum(x))`
Cramérs v (Effektstärke für s×z-Tabellen)	$v = \sqrt{\frac{\frac{\chi^2}{n}}{min(s-1, z-1)}}$
	`sz <- min(nrow(x)-1, ncol(x)-1)` `sqrt(xq$statistic/sum(x)/sz)`

ϕ bzw. v werden nach Cramér überlicherweise wiefolgt interpretiert (s. Gravetter & Wallnau S. 603). Wie sehr man diesem Schema folgen will, bleibt jedem selber überlassen.

Freiheitsgrade	Intervall Cramér's v	Einschätzung
V mit $df = 1$ bzw. ϕ	$0.1 < v < 0.3$	schwacher Effekt
	$0.3 < v < 0.5$	mittelstarker Effekt
	$v > 0.5$	starker Effekt
$df = 2$	$0.07 < v < 0.21$	schwacher Effekt
	$0.21 < v < 0.35$	mittelstarker Effekt
	$v > 0.35$	starker Effekt
$df = 3$	$0.06 < v < 0.17$	schwacher Effekt
	$0.17 < v < 0.29$	mittelstarker Effekt
	$v > 0.29$	starker Effekt

Der **exakte Fisher-Test** berechnet die Wahrscheinlichkeit direkt (wie alle exakten Tests) und ist schwierig von Hand zu berechnen. Er kann in R wie der χ^2-Test gerechnet werden, der Funktionsname lautet dabei `fisher.test()` statt `chisq.test()`. Der Zugriff auf `$statistic` funktioniert nicht, weil es keinen Testwert gibt. Als Effektstärke nimmt man gerne das Chancenverhältnis (s. u.), das R sowieso gleich mit ausgibt.

2.2 Chance

Chance	$o(E) = \frac{p(E)}{1-p(E)}$
Chancenverhältnis	$or(E\|A, E\|B) = \frac{o(E\|A)}{o(E\|B)}$

2.3 Fehlerreduktion

Fehlerreduktion	$\lambda = \frac{(\sum_i M_i) - max(Z)}{(\sum_i Z_i) - max(Z)}$ für die modalen Kategorien M und die Zeilensummen Z

2.4 Binomialtest

Binomialtest als z-Test	$z = \frac{X-\mu}{s}$
für die gemessene Anzahl X	mit $\mu = p \cdot n$
und den H0-Anteilswert p	und $s = \sqrt{n \cdot p \cdot (1-p)}$
(kritische Werte: Normalverteilung)	`binom.test(X, n, p)`

3 t-Test für Mittelwerte

3.1 t-Test mit einer Stichprobe

Ganz allgemein ist ein „t-Test" jeder Test, der auf einem Testwert beruht, der t-verteilt ist (Paralleles gilt für χ^2-Tests, F-Tests usw.). Typischerweise meint man mit „t-Test" aber speziell den t-Test für Mittelwertunterschiede, um den es hier jetzt geht. Der **t-Test für Mittelwerte mit einer Stichprobe** ist ein direkter Vergleichstest für Mittelwerte. Er testet ein **Stichprobenmittel** $\bar{x}$ gegen ein (bekanntes oder theoretisch vermutetes) **Grundgesamtheitsmittel** μ auf signifikante Abweichung. Wenn die **Varianz der Grundgesamtheit** und damit der **nicht-geschätzte Standardfehler**

für Stichproben der Größe n aus der Grundgesamtheit bekannt ist, wird der t-Test zum z-Test. Dabei ist er einzige wesentliche Unterschied, dass statt der t-Verteilung die Normalverteilung zur Berechnung der kritischen Werte verwendet wird. In der Praxis ist fast immer der t-Test zu verwenden.

t-Test mit einer Stichprobe	$t(\bar{x}, \mu) = \frac{\bar{x}-\mu}{SF(x)}$ mit $df = n - 1$
	`tt <- t.test(x, mu=Mu)`
	Hier ist `Mu` der angenommene μ, wird händisch eingegeben.
Cohens d	$d = \frac{\bar{x}-\mu}{s(x)}$
(Effektstärke für t-Test)	`(mean(x)-Mu)/sd(x)`
Cohens r^2	$r^2 = \frac{t^2}{t^2+df}$
	`t <- tt$statistic`
	`t^2/(t^2+length(x)-1)`

3.2 t-Test mit zwei Stichproben

Der **t-Test mit zwei Stichproben** vergleicht die **Differenz zwischen zwei Stichprobenmitteln $\bar{x_1}$ und $\bar{x_2}$** mit der **Differenz zwischen den Mitteln in den zwei Grundgesamtheiten μ_1 und μ_2**, aus denen die Stichproben stammen. Unter Annahme der H0 sind die Mittel der Grundgesamtheiten gleich, so dass $\mu_1 - \mu_2 = 0$. Bei der Berechnung des Standardfehlers für die Differenz der Mittelwerte zweier Stichproben $SF(\bar{x}_1 - \bar{x}_2)$ muss die zusammengelegte Varianz (*pooled variance*) s_p^2 zugrundegelegt werden, um unterschiedlichen Stichprobengrößen gerecht zu werden.

t-Test mit zwei Stichproben	$t(\bar{x}_1, \bar{x}_2) = \frac{(\bar{x_1}-\bar{x_2})-(\mu_1-\mu_2)}{SF(x_1-x_2)} = \frac{(\bar{x_1}-\bar{x_2})-0}{SF(x_1-x_2)}$
	mit $df = (n_1 - 1) + (n_2 - 1)$
	`tt <- t.test(x1, x2)`
zusammengelegte Varianz	$s_p^2(x1, x2) = \frac{SQ(x_1)+SQ(x_2)}{(n_1-1)+(n_2-1)}$
Standardfehler für Mitteldifferenzen	$SF(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_p^2(x_1,x_2)}{n_1} + \frac{s_p^2(x_1,x_2)}{n_2}}$
Cohens d	$d = \frac{\bar{x_1}-\bar{x_2}}{\sqrt{s_p^2(x1,x2)}}$
Cohens r^2	$r^2 = \frac{t^2}{t^2+df}$

Cohen's d wird nach Cohen wiefolgt interpretiert (s. Gravetter & Wallnau S. 258).

Intervall Cohen's d	Einschätzung
$0 < d < 0.2$	schwacher Effekt
$0.2 < d < 0.8$	mittelstarker Effekt
$d > 0.8$	starker Effekt

Gemäß Cohen ist r^2 für den t-Test wiefolgt zu interpretieren (s. Gravetter & Wallnau S. 288). Der Nachteil ist ggü. anderen r^2 und R^2, dass die Werte kleiner sind, als man erwartet.

Intervall r^2	Einschätzung
$0.01 < r^2 < 0.09$	schwacher Effekt
$0.09 < r^2 < 0.25$	mittelstarker Effekt
$r^2 > 0.25$	starker Effekt

4 ANOVA

4.1 Einfaktorielle ANOVA

Die ANOVA (*ANalysis Of VAriance*) testet auf Mittelwertunterschiede zwischen beliebig vielen Stichproben, die im ANOVA-Kontext **Gruppen** heißen. Der Unterschied zum t-Test ist in der Anwendung, dass mehr als zwei Gruppen vorliegen können. Mathematisch liegt der Unterschied zum t-Test darin, dass die ANOVA nicht direkt die Mittelwerte vergleicht, sondern die Varianz zwischen den Gruppen (= konzeptuell die Varianz, die auf Kosten des vermuteten Effekts geht) zur Varianz innerhalb der Gruppen (= konzeptuell die Zufallsvarianz) ins Verhältnis setzt. Bei der einfaktoriellen ANOVA entsprechen die k Gruppen (x_i mit $i = 1..k$ und den Gruppengrößen n_i) den **Ausprägungen einer nominalen unabhängigen Variable**, die im ANOVA-Kontext (wie in R) **Faktor** heißt. Die gesamte Stichprobe sei X, ihre Größe N. Die Summen der Gruppen seien T_i und die Gesamtsumme G.

F für einfaktorielle ANOVA	$F = \frac{s^2_{zwischen}}{s^2_{in}}$
Varianz in den Gruppen	$s^2_{in} = \frac{SQ_{in}}{df_{in}}$
Summe der Quadrate in den Gruppen	$SQ_{in} = \sum_i SQ(x_i)$
Freiheitsgrade in den Gruppen	$df_{in} = (N-1) - (k-1)$
Varianz zwischen den Gruppen	$s^2_{zwischen} = \frac{SQ_{zwischen}}{df_{zwischen}}$
Summe der Quadrate zwischen den Gruppen	$SQ_{zwischen} = \sum_i (\frac{T_i^2}{n_i}) - \frac{G^2}{N}$
Freiheitsgrade zwischen den Gruppen	$df_{zwischen} = k - 1$
r^2 für ANOVAs	$\eta^2 = \frac{SQ_{zwischen}}{SQ_{gesamt}}$
Summe der Quadrate gesamt	$SQ_{gesamt} = \sum_i (X_i - \bar{X})$

4.2 Mehrfaktorielle ANOVA

Die **mehrfaktorielle ANOVA** erlaubt die Berechnung des Einflusses zweier oder mehrerer **Hauptfaktoren** auf die Abhängige sowie eventueller **Interaktionen zwischen den Faktoren**. Die Faktoren werden mit A, B usw. bezeichnet, k_A, k_B usw. sind die Zahlen ihrer Ausprägungen. Im Vergleich zur einfaktoriellen ANOVA kommt hinzu, dass die Varianz der Faktoren einzeln berechnet werden muss, um dann durch Subtraktion dieser Varianzen von der gesamten Varianz zwischen den Gruppen die Interaktionsvarianz zu ermitteln. Bei zwei Faktoren mit je zwei Ausprägungen ergibt sich z. B. ein Vier-Gruppen-Design wie in folgender Tabelle, mit x_{ij} als Gruppen und A_i und B_j als Zeilen- und Spaltensummen (konzeptuell die Gruppen der Hauptfaktoren ohne Beachtung des jeweils anderen Faktors).

x_{11}	x_{21}	A_1
x_{21}	x_{22}	A_2
B_1	B_2	X

Die Varianz in den Gruppen und zwischen den Gruppen wird wie bei der einfaktoriellen ANOVA gerechnet. Dann werden F-Werte für jeden Faktor und die Interaktionen gerechnet.

F für ANOVA eines Hauptfaktors (hier A)	$F_A = \frac{s_A^2}{s_{in}^2}$
Hauptfaktorenvarianz (hier A)	$s_A^2 = \frac{SQ_A}{df_A}$
Summe der Quadrate für Hauptfaktoren	$SQ_A = \sum_i (\frac{{T_{Ai}}^2}{n_{Ai}}) - \frac{G^2}{N}$
Freiheitsgrade für Hauptfaktoren	$df_A = k_A - 1$
Effektstärke für Hauptfaktor	$\eta_A^2 = \frac{SQ_A}{SQ_{gesamt} - SQ_B - SQ_{A\times B}}$
F für ANOVA der Interaktion	$F_{A\times B} = \frac{s_{A\times B}^2}{s_{in}^2}$
Interaktionsvarianz	$s_{A\times B}^2 = \frac{SQ_{A\times B}}{df_{A\times B}}$
Summe der Quadrate für Interaktion	$SQ_{A\times B} = SQ_{zwischen} - SQ_A - SQ_B$
Freiheitsgrade für Interaktion	$df_{A\times B} = df_{zwischen} - df_A - df_B$
Effektstärke für Interaktion	$\eta_{A\times B}^2 = \frac{SQ_{A\times B}}{SQ_{gesamt} - SQ_A - SQ_B}$

5 Nichtparametrische Alternativen für t und ANOVA

Es wird eine Doppelindexierung nötig: x_i sind die Werte von Gruppe i, x_{i_j} ist Wert j aus Gruppe i.

Rang von x_{i_j} in der vereinten Stichprobe	$R(x_{i_j})$
Mann-Whitney-U-Test	$U(x_i) = n_1 \cdot n_2 + \frac{n_i(n_i+1)}{2} - \sum_j R(x_{i_j})$
U ist normalverteilt (z-Test).	$U = min(U(x_1), U(x_2), \ldots, U(x_n))$
Ersetzt t-Test mit zwei Stichproben.	`wilcox.test(x, y)`
Kruskal-Wallis-H-Test	$H = \frac{12}{N(N+1)} \cdot \sum_i \frac{(\sum_j R(x_{i_j}))^2}{n_i} - 3(N+1)$
H ist χ^2-verteilt.	$df = k - 1$
Ersetzt ANOVA.	`kruskal.test(x~y)`

6 Lineare Modelle

Bei der Regression spricht man auch von „Regressoren“ statt „Unabhängigen“ und von der „Response-Variable“ statt der „Abhängigen“.

Modellgleichung (allgemeines LM)	$\hat{y} = b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + a$
Koeffizient bei einem Regressor Intercept bei einem Regressor t für Koeffizienten	$b = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SP(x,y)}{SQ(x)}$ $a = \bar{y} - b \cdot \bar{x}$ $t = \frac{b}{SF(b)}$ mit $SF(b) = \frac{\sqrt{\frac{\sum e^2}{n-1}}}{\sqrt{SQ(x)}}$
Residual-Standardfehler für Modell	$SF_{residual} = \sqrt{\frac{\sum e^2}{n-2}}$
Residual-Varianz Regressions-Varianz F-Test für Modell mit Freiheitsgraden	$s^2_{residual} = \frac{(1-r^2) \cdot SQ(y)}{1}$ $s^2_{regression} = \frac{r^2 \cdot SQ(y)}{n-2}$ $F = \frac{s^2_{regression}}{s^2_{residual}}$ $df_1 = 1$ und $df_2 = n - 1$
(alles zusammen)	`summary(lm(y~x))` bzw. `summary(lm(y~x1+x2))`

7 Logistische Regression/Generalisierte Lineare Modelle

Logits (Linearkombination)	$z = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_0$
Modellgleichung (Logit-Link)	$\hat{p}(y = 1) = \frac{1}{1+e^{-z}}$
Chancen des Modells	$\hat{o}(y = 1) = \frac{p(y=1)}{1-p(y=1)} = e^z$
Chancenverhältnis für Regressor x_i	$or(y = 1 \mid x_i) = e^{\beta_i}$
Likelihood des Modells für Daten k	$p_k = (\frac{1}{1+e^{-z_k}})^{y_k} \cdot (1 - \frac{1}{1+e^{-z_k}})^{1-y_k}$
Likelihood des Modells für alle Daten	$L = \prod_k p_k$
Likelihood-Ratio-Test (χ^2-Test)	$LR = (-2 \cdot ln(L_r)) - (-2 \cdot ln(L_f))$ mit L_r/L_f als L des reduzierten/vollen Modells
Cox & Snell und Nagelkerke R^2	$R^2_C = 1 - (\frac{L_0}{L_f})^{\frac{2}{n}}$ $R^2_N = \frac{R^2_C}{R^2_{max}}$ mit $R^2_{max} = 1 - (L_0)^{\frac{2}{n}}$
Schätzung der Dispersion	$\hat{\phi} = \sum(\frac{R_P}{df_R})^2$ mit R_P als Pearson-Residuen und $df_R = n - p$ mit p als Anzahl der geschätzten Parameter

Modell-Anpassung	`m <- glm(y~x1+x2*y3, family="bin")` `summary(m)`
Chancenverh. (*or*) für Koeffizienten	`exp(coef(m))`
Konfidenzint. für *or*	`exp(confint(m))`
Log-Likelihood extrahieren	`logLik(m)`
Nagelkerke R^2	`library(fmsb)` `NagelkerkeR2(m)`

LR-Test	`m0 <- glm(y~1, family="bin")`
	`lr <- (-2*logLik(m0))-(-2*logLik(m))`
	`pchisq(lr, m$rank-m0$rank)`

Modellselektion	`drop1(m)`
Varianzinflationsfaktoren	`library(car)` `vif(m)`

Vorhersagegüte	`pred <- ifelse(predict(m) <= 0.5, 0, 1)`
	`tab <- table(pred, mydata$response)`
	`sum(diag(tab))/sum(tab)`

Fehlerrate in Kreuzzvalidierung (hier $k = 10$)	`library(boot)` `cv.glm(mydata, m, K=10)$delta`

8 Quantilfunktionen in R

Mit den Quantilfunktionen können kritische Werte für bekannte Verteilungen ermittelt werden. Die Syntax in R ist von Verteilung zu Verteilung gesondert zu betrachten. Wir geben hier die Aufrufe für $\alpha = 0.05$ an. Die Freiheitsgrade (`df`) werden hier in den Beispielen typisch gewählt müssen aber natürlich an die gegebene Testsituation angepasst werden.

1. Normalverteilung, beidseitig:

   ```
   qnorm(c(0.025, 0.975))
   ```

   Ausgabe: Die beiden Werte, **zwischen denen** 95% der Werte liegen.

2. t-Verteilung, beidseitig:

   ```
   qt(c(0.025, 0.975), df=9)
   ```

   Ausgabe: Die beiden Werte, **zwischen denen** 95% der Werte liegen.

3. χ^2-Verteilung, einseitig:

   ```
   qchisq(0.95, df=1)
   ```

   Ausgabe: Der Wert, **links dessen** 95% der Werte liegen.

4. F-Verteilung, einseitig:

   ```
   qf(0.95, df1=2, df2=8)
   ```

   Ausgabe: Der Wert, **links dessen** 95% der Werte liegen.
   `df1` ist $df_{zwischen}$ und `df2` ist df_{in}.

9 Dichten einiger wichtiger Verteilungen