Blockpraktikum zur Statistik mit R

08. Oktober 2010

Till Breuer, Sebastian Mentemeier und Matti Schneider

Institut für Mathematische Statistik
Universität Münster
WS 2010/11

Gliederung

1. Ein-Stichproben-Fall
   - Parametrische Testverfahren zu Lagealternativen
   - Verteilungsfreie Testverfahren zu Lagealternativen
   - Nichtparametrische Anpassungstest (Goodness-of-fit-Tests)

2. Zwei-Stichproben-Fall
   - Parametrische Zwei-Stichproben-Tests
   - Nichtparametrische Zwei-Stichproben-Anpassungstests

Gliederung

1. Ein-Stichproben-Fall
   - Parametrische Testverfahren zu Lagealternativen
   - Verteilungsfreie Testverfahren zu Lagealternativen
   - Nichtparametrische Anpassungstest (Goodness-of-fit-Tests)
2. Zwei-Stichproben-Fall
   - Parametrische Zwei-Stichproben-Tests
   - Nichtparametrische Zwei-Stichproben-Anpassungstests

Ein-Stichproben-Fall

Es wird ein einziges Merkmal X auf der Basis einer einfachen Zufallsstichprobe $(X_1, \ldots, X_n)$ bzgl. interessierender Fragestellungen getestet, z. B. auf

- die Lage von Mittelwert oder Median im Vergleich zu vermuteten Werten - hierbei wird unterschieden zwischen
  - parametrischen Verfahren
  - verteilungsfreien Verfahren
- die Klasse der zugrundeliegenden Verteilung.

Student'scher t-Test (Ein-Stichproben-Fall)

Annahmen:	$X_1, \ldots, X_n$ u. i. v. mit $X \sim \mathcal{N}(\mu, \sigma^2)$ bzw. beliebig verteilt mit ex. Varianz und großem n
Hypothesen:	(a) H: $\mu = \mu_0$ vs. K: $\mu \neq \mu_0$, (`two.sided`) (b) H: $\mu \leq \mu_0$ vs. K: $\mu > \mu_0$, (`greater`) (c) H: $\mu \geq \mu_0$ vs. K: $\mu < \mu_0$. (`less`)
Teststatistik:	$T(X) = \sqrt{n}\frac{\overline{X}-\mu_0}{S(X)}$
Verteilung unter μ_0:	$t(n-1)$ (Student'sche t-Verteilung mit $n-1$ Freiheitsgraden)
Ablehnungs-bereich:	(a) $\lvert T(X)\rvert > q_{1-\alpha/2}(t(n-1))$, (b) $T(X) > q_{1-\alpha}(t(n-1))$, (c) $T(X) < -q_{1-\alpha}(t(n-1)) = q_\alpha(t(n-1))$.
R-Befehl:	`t.test(x, mu =` μ_0`, alternative="...")`

Gliederung

1. Ein-Stichproben-Fall
   - Parametrische Testverfahren zu Lagealternativen
   - Verteilungsfreie Testverfahren zu Lagealternativen
   - Nichtparametrische Anpassungstest (Goodness-of-fit-Tests)
2. Zwei-Stichproben-Fall
   - Parametrische Zwei-Stichproben-Tests
   - Nichtparametrische Zwei-Stichproben-Anpassungstests

Vorzeichen-Test

Annahmen:	$X_1, \ldots, X_n$ u. i. v. mit stetiger Verteilungsfunktion
Hypothesen:	(a) H: $x_{\text{med}} = \mu_0$ vs. K: $x_{\text{med}} \neq \mu_0$,
	(b) H: $x_{\text{med}} \geq \mu_0$ vs. K: $x_{\text{med}} < \mu_0$.
	(c) H: $x_{\text{med}} \leq \mu_0$ vs. K: $x_{\text{med}} > \mu_0$,
Teststatistik:	$A = \sum_{i=1}^{n} \mathbf{1}_{\{X_i < \mu_0\}}$
Verteilung unter μ_0:	$\mathfrak{B}(n, 0.5)$
Ablehnungs-bereich:	(a) $\min(A, n-A) \leq q_{\alpha/2}(\mathfrak{B}(n, 0.5))$,
	(b) $n - A \leq q_{\alpha}(\mathfrak{B}(n, 0.5))$,
	(c) $A \leq q_{\alpha}(\mathfrak{B}(n, 0.5))$.
R-Befehle:	(a) `binom.test(min(A,n-A),n, alternative="two.sided")`
	(b) `binom.test(n-A,n,alternative="less")`
	(c) `binom.test(A,n,alternative="less")`

Der Wilcoxon-Vorzeichen-Rangtest

Annahmen:	$X_1, \ldots, X_n$ u. i. v. mit stetiger Verteilungsfunktion, symmetrische Verteilung
Hypothesen:	(a) H: $x_{\text{med}} = \mu_0$ vs. K: $x_{\text{med}} \neq \mu_0$, (`two.sided`) (b) H: $x_{\text{med}} \leq \mu_0$ vs. K: $x_{\text{med}} > \mu_0$, (`greater`) (c) H: $x_{\text{med}} \geq \mu_0$ vs. K: $x_{\text{med}} < \mu_0$. (`less`)
Teststatistik:	$W^+ = \sum_{i=1}^n \text{rg}\lvert D_i \rvert Z_i$ mit $D_i = X_i - \mu_0$, $Z_i = \mathbf{1}_{\{D_i > 0\}}$
Verteilung unter μ_0:	Wilcoxon'sche Rangstatistik, für großes n approximativ $\mathcal{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$
Ablehnungs-bereich:	(a) $W^+ < w^+_{\alpha/2}$ oder $W^+ > w^+_{1-\alpha/2}$ (b) $W^+ < w^+_{\alpha}$ (c) $W^+ > w^+_{1-\alpha/2}$
R-Befehl:	`wilcox.test(x, mu =` μ_0`, alternative="...")`

Gliederung

1. Ein-Stichproben-Fall
   - Parametrische Testverfahren zu Lagealternativen
   - Verteilungsfreie Testverfahren zu Lagealternativen
   - Nichtparametrische Anpassungstest (Goodness-of-fit-Tests)

2. Zwei-Stichproben-Fall
   - Parametrische Zwei-Stichproben-Tests
   - Nichtparametrische Zwei-Stichproben-Anpassungstests

χ^2-Anpassungstest für kategoriale Merkmale

Annahmen:	$X_1, \ldots, X_n$ u. i. v. mit Werten in $\{1, \ldots, k\}$
Hypothese:	H: $P(X_1 = i) = p_i$ für $i = 1, \ldots, k$ K: $P(X_1 = i) \neq p_i$ für ein i
Teststatistik:	$\chi^2 := \sum_{i=1}^{k} \frac{(h_i - np_i)^2}{np_i}$ mit $h_i = \lvert\{j : X_j = i\}\rvert$
Verteilung unter H:	approximativ χ^2_{k-1}; Approximation anwendbar, wenn $np_i \geq 1$ für alle i, $np_i \geq 5$ für min. 80% der i
Ablehnungs-bereich:	$\chi^2 > q_{1-\alpha}(\chi^2_{k-1})$
R-Befehl:	`chisq.test(x, p)`, mit p=$(p_1, \ldots, p_k)$

Kolmogoroff-Smirnoff-Test

Annahmen:	$X_1, \ldots, X_n$ u. i. v. mit stetiger Verteilungsfunktion F
Hypothesen:	(a) H: $F = F_0$ vs. K: $F \neq F_0$ (`two.sided`) (b) H: $F \leq F_0$ vs. K: $F > F_0$ (`greater`) (c) H: $F \geq F_0$ vs. K: $F < F_0$ (`less`)
Teststatistik:	(a) $\sup_{t\in\mathbb{R}} \lvert F_{n,x}(t) - F_0(t)\rvert$, (b) $\sup_{t\in\mathbb{R}} F_{n,x}(t) - F_0(t)$ bzw. (c) $\inf_{t\in\mathbb{R}} F_{n,x}(t) - F_0(t)$
Verteilung:	tabelliert
Ablehnungs-bereich:	falls die Statistik zu groß wird
R-Befehl:	`ks.test(x, "F", `θ`, alternative="...")` wobei `F` eine Verteilungsfunktion sein muss, etwa `pnorm`, und θ die zugehörigen Parameter, z.B. μ, σ^2

Shapiro-Wilk-Test

Annahmen:	$X_1, \ldots, X_n$ u. i. v. mit stetiger Verteilungsfunktion F
Hypothese:	H: F ist eine Normalverteilung K: F ist keine Normalverteilung
Teststatistik:	$W = \frac{(\sum_{i=1}^n a_i X_{(i)})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$, dabei ist $\mathbf{a}^{\mathrm{t}} = (a_1, \ldots, a_n)$ durch $\mathbf{a}^{\mathrm{t}} = \frac{\mathbf{m}^{\mathrm{t}} \mathbf{V}^{-1}}{(\mathbf{m}^{\mathrm{t}} \mathbf{V}^{-2} \mathbf{m})^{1/2}}$, wobei $\mathbf{m}$ und $\mathbf{V}$ Erwartungswertvektor bzw. Kovarianzmatrix eines geordneten Vektors von n u. i. v. $\mathcal{N}(0,1)$-Variablen sei
Verteilung:	Shapiro-Wilk-Verteilung
Ablehnungsbereich:	kleines W
R-Befehl:	`shapiro.test(x)`

Gliederung

1. Ein-Stichproben-Fall
   - Parametrische Testverfahren zu Lagealternativen
   - Verteilungsfreie Testverfahren zu Lagealternativen
   - Nichtparametrische Anpassungstest (Goodness-of-fit-Tests)

2. Zwei-Stichproben-Fall
   - Parametrische Zwei-Stichproben-Tests
   - Nichtparametrische Zwei-Stichproben-Anpassungstests

Zwei-Stichproben-Fall

In diesem Fall wird ein Merkmal unter zwei Bedingungen untersucht oder man betrachtet zwei Merkmale, die am selben Merkmalsträger erhoben werden:

1. Zwei unabhängige Zufallsstichproben $(X_{1,1}, \ldots, X_{1,n_1}), (X_{2,1}, \ldots, X_{2,n_2})$, $n_1, n_2 \in \mathbb{N}$, wobei sich die Randbedingungen bei der Entnahme der Stichproben in genau einer Randbedingung unterscheiden.
2. Ein Merkmal unter zwei verschiedenen Bedingungen am selben Merkmalsträger: $(X_{1,1}, X_{1,2}), \ldots, (X_{n,1}, X_{n,2})$ (verbundene Stichproben, *matched pairs*).
3. Zwei Merkmale X und Y am selben Merkmalsträger (unter jeweils gleichen Bedingungen): $(X_1, Y_1), \ldots, (X_n, Y_n)$ (verbundene Stichproben).

Reduktion auf das Ein-Stichproben-Problem

- Das Problem (2) der Messung eines Merkmals unter verschiedenen Bedingungen am selben Merkmalsträger wird im Falle intervallskalierter Merkmale häufig durch Differenzbildung auf das Ein-Stichprobenproblem zurückgeführt.
- Dies wird in R in den Befehlen `t.test` und `wilcox.test` über den Parameter `paired` (=`TRUE` / `FALSE`) gesteuert.
- Wir konzentrieren uns im Folgenden auf den Fall (1).

Exakter Test von Fisher

Annahmen:	unabhängige Zufallsstichproben $X_1, \ldots, X_n$ u. i. v. $\sim \mathfrak{B}(1, \theta_X)$, $Y_1, \ldots, Y_m$ u. i. v. $\sim \mathfrak{B}(1, \theta_Y)$
Hypothesen:	(a) H: $\theta_X = \theta_Y$ vs. K: $\theta_X \neq \theta_Y$, (`two.sided`) (b) H: $\theta_X \leq \theta_Y$ vs. K: $\theta_X > \theta_Y$, (`greater`) (c) H: $\theta_X \geq \theta_Y$ vs. K: $\theta_X < \theta_Y$. (`less`)
Teststatistik:	$(U, V) = \left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right)$
Verteilung:	U hat unter $\theta_X = \theta_Y$ eine $\mathfrak{H}(n, m, U+V)$-Verteilung
Ablehnungs-bereich:	(a) $U > c_\alpha(\mathfrak{H}(n, m, U+V))$ (α-Fraktil) (b) $U < q_\alpha(\mathfrak{H}(n, m, U+V))$ (α-Quantil) (c) U zu groß oder zu klein
R-Befehl:	`fisher.test(T, alternative="...")` wobei `T` die Kontingenztafel von X und Y ist d. h. `T<-matrix(c(U,V,n-U,m-V),2)`

Student'scher t-Test im Zwei-Stichproben-Fall

Annahmen:	unabhängige Zufallsstichproben, gl. Varianz $X_1, \ldots, X_n$ u. i. v. $\sim \mathcal{N}(\mu_X, \sigma^2)$ $Y_1, \ldots, Y_m$ u. i. v. $\sim \mathcal{N}(\mu_Y, \sigma^2)$
Hypothesen:	(a) H: $\mu_X = \mu_Y$ vs. K: $\mu_X \neq \mu_Y$, (`two.sided`) (b) H: $\mu_X \leq \mu_Y$ vs. K: $\mu_X > \mu_Y$, (`greater`) (c) H: $\mu_X \geq \mu_Y$ vs. K: $\mu_X < \mu_Y$. (`less`)
Teststatistik:	$T(X) = \left(\frac{nm}{n+m}\right)^{1/2} \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{n-1}{m+n-2} S(X)^2 + \frac{m-1}{m+n-2} S(Y)^2}}$
Verteilung unter $\mu_X = \mu_Y$:	$t(n+m-2)$ (Student'sche t-Verteilung mit $n+m-2$ Freiheitsgraden)
Ablehnungs-bereich:	(a) $\lvert T(X) \rvert > q_{1-\alpha/2}(t(n+m-2))$, (b) $T(X) > q_{1-\alpha}(t(n+m-2))$, (c) $T(X) < q_{\alpha}(t(n+m-2))$.
R-Befehl:	`t.test(x, y, alternative="...", var.equal=TRUE)`

Gliederung

1. Ein-Stichproben-Fall
   - Parametrische Testverfahren zu Lagealternativen
   - Verteilungsfreie Testverfahren zu Lagealternativen
   - Nichtparametrische Anpassungstest (Goodness-of-fit-Tests)

2. Zwei-Stichproben-Fall
   - Parametrische Zwei-Stichproben-Tests
   - Nichtparametrische Zwei-Stichproben-Anpassungstests

Kolmogoroff-Smirnoff-Test

Annahmen:	$X_1, \ldots, X_n$ u. i. v. mit stetiger Verteilungsfunktion F $Y_1, \ldots, Y_m$ u. i. v. mit stetiger Verteilungsfunktion G
Hypothesen:	(a) H: $F = G$ vs. K: $F \neq G$ (`two.sided`) (b) H: $F \leq G$ vs. K: $F > G$ (`greater`) (c) H: $F \geq G$ vs. K: $F < G$ (`less`)
Teststatistik:	(a) $\sup_{t \in \mathbb{R}} \lvert F_{n,x}(t) - G_{m,y}(t) \rvert$, (b) $\sup_{t \in \mathbb{R}} F_{n,x}(t) - G_{m,y}(t)$ bzw. (c) $\inf_{t \in \mathbb{R}} F_{n,x}(t) - G_{m,y}(t)$
Verteilung:	tabelliert
Ablehnungs-bereich:	falls die Statistik zu groß wird
R-Befehl:	`ks.test(x, y, alternative="...")`