Einführung in die angewandte Statistik mit R

10. Übungsblatt

Zusammenfassung der zugrundeliegenden Theorie

Einstichproben-Tests für den Erwartungswert einer zugrundeliegenden Normalverteilung

Gegeben seien Realisationen $x_1, \dots, x_n \in \mathbb{R}$ stochastisch unabhängiger, jeweils mit Parametern μ und σ^2 normalverteilter Zufallsvariablen $X_1, \dots, X_n$ (mit $n \geq 2$). Hierbei seien $\mu \in \mathbb{R}$ und $\sigma \in (0, \infty)$. Weiter sei $\alpha \in (0,1)$, und es bezeichnen

$$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{und} \quad \hat{\sigma} := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

das arithmetische Mittel bzw. die Stichproben-Standardabweichung zu $X_1, \dots, X_n$.

(1) Einstichproben-Gauß-Test für den Erwartungswert μ zum Niveau α (bei bekannter Varianz σ^2)

Teststatistik

$$T = \sqrt{n}\, \frac{\overline{X} - \mu_0}{\sigma} \quad (\mu_0 \in \mathbb{R} \text{ vorgegebener Sollwert})$$

Es bezeichne T_{real} die zugehörige realisierte Teststatistik, die sich aus der Ersetzung von $X_1, \dots, X_n$ durch $x_1, \dots, x_n$ ergibt.

Testentscheidungen

Hypothesen		Testentscheidung	Zugehöriger p-Wert
H_0	H_1	Ablehnung von H_0, falls	
$\mu \leq \mu_0$	$\mu > \mu_0$	$T_{\text{real}} > u_{1-\alpha}$	$P_{\mu_0}(T > T_{\text{real}})$
$\mu \geq \mu_0$	$\mu < \mu_0$	$T_{\text{real}} < -u_{1-\alpha} = u_\alpha$	$P_{\mu_0}(T < T_{\text{real}})$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert T_{\text{real}} \rvert > u_{1-\alpha/2}$	$P_{\mu_0}(\lvert T \rvert > \lvert T_{\text{real}} \rvert)$

Hierbei bezeichnen u_β das β-Quantil der Standard-Normalverteilung für $\beta \in (0,1)$ und P_{μ_0} das zugrundeliegende Wahrscheinlichkeitsmaß bei vorliegendem Parameter μ_0.

(2) Einstichproben-t-Test für den Erwartungswert μ zum Niveau α (bei unbekannter Varianz σ^2)

Teststatistik

$$T = \sqrt{n}\,\frac{\overline{X} - \mu_0}{\hat{\sigma}} \qquad (\mu_0 \in \mathbb{R} \text{ vorgegebener Sollwert})$$

Wiederum bezeichne T_{real} die zugehörige realisierte Teststatistik, die sich aus der Ersetzung von $X_1, \ldots, X_n$ durch $x_1, \ldots, x_n$ ergibt.

Testentscheidungen

Hypothesen		Testentscheidung	Zugehöriger p-Wert
H_0	H_1	Ablehnung von H_0, falls	
$\mu \leq \mu_0$	$\mu > \mu_0$	$T_{\text{real}} > t_{1-\alpha}(n-1)$	$P_{\mu_0}(T > T_{\text{real}})$
$\mu \geq \mu_0$	$\mu < \mu_0$	$T_{\text{real}} < -t_{1-\alpha}(n-1) = t_\alpha(n-1)$	$P_{\mu_0}(T < T_{\text{real}})$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert T_{\text{real}}\rvert > t_{1-\alpha/2}(n-1)$	$P_{\mu_0}(\lvert T\rvert > \lvert T_{\text{real}}\rvert)$

Hierbei bezeichnen $t_\beta(k)$ das β-Quantil der t-Verteilung mit Freiheitsgrad $k \in \mathbb{N}$ für $\beta \in (0,1)$ und wiederum P_{μ_0} das zugrundeliegende Wahrscheinlichkeitsmaß bei vorliegendem Parameter μ_0.

Zweistichproben-Tests für den Vergleich der Erwartungswerte zweier verbundener (gepaarter) Stichproben bei zugrundeliegender Normalverteilung

Gegeben seien Datenpaare $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ mit $n \geq 2$, die als Realisationen stochastisch unabhängiger, identisch verteilter Zufallsvektoren $(X_1, Y_1), \ldots, (X_n, Y_n)$ aufgefasst werden können. (Man beachte, dass hiermit *nicht* die stochastische Unabhängigkeit von X_i und Y_i für $i \in \{1, \ldots, n\}$ vorausgesetzt wird.)

Weiter sei $\Delta_i := X_i - Y_i$ für $i \in \{1, \ldots, n\}$ normalverteilt mit Parametern

$$\mu_\Delta := \mu_1 - \mu_2 := \mathrm{E}(X_1) - \mathrm{E}(Y_1) \in \mathbb{R} \quad \text{und} \quad \sigma_\Delta^2 := \mathrm{Var}(X_1 - Y_1) > 0\,.$$

Zu vorgegebenem Sollwert $\delta_0 \in \mathbb{R}$ werden die folgenden Hypothesen betrachtet:

$$\begin{aligned}
H_0 &: \mu_\Delta = \mu_1 - \mu_2 \leq \delta_0 \quad \text{gegen} \quad H_1 : \mu_\Delta = \mu_1 - \mu_2 > \delta_0 \\
H_0 &: \mu_\Delta = \mu_1 - \mu_2 \geq \delta_0 \quad \text{gegen} \quad H_1 : \mu_\Delta = \mu_1 - \mu_2 < \delta_0 \\
H_0 &: \mu_\Delta = \mu_1 - \mu_2 = \delta_0 \quad \text{gegen} \quad H_1 : \mu_\Delta = \mu_1 - \mu_2 \neq \delta_0
\end{aligned}$$

Von besonderem Interesse ist hierbei der Fall $\delta_0 = 0$, der dem direkten Vergleich der beiden Erwartungswerte μ_1 und μ_2 entspricht.

Da die Differenzen $\Delta_1, \ldots, \Delta_n$ die Voraussetzungen der zuvor angegebenen Einstichproben-Tests für den Erwartungswert einer zugrundeliegenden Normalverteilung erfüllen, wird die Situation verbundener (gepaarter) Stichproben hierauf zurückkgeführt.

Für die zugehörigen realisierten Differenzen $\delta_1 := x_1 - y_1, \ldots, \delta_n := x_n - y_n$ werden der Einstichproben-Gauß-Test bzw. der Einstichproben-t-Test durchgeführt je nachdem, ob die Varianz σ_Δ^2 *bekannt* oder *unbekannt* ist.

Zweistichproben-Tests für den Vergleich zweier unabhängiger Stichproben bei zugrundeliegender Normalverteilung

Gegeben seien Realisationen $x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2} \in \mathbb{R}$ (gemeinsam) stochastisch unabhängiger Zufallsvariablen $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$ mit $n_1, n_2 \geq 2$, $X_i \sim N(\mu_1, \sigma_1^2)$ und $Y_j \sim N(\mu_2, \sigma_2^2)$ für $i \in \{1, \dots, n_1\}$ bzw. $j \in \{1, \dots, n_2\}$. Hierbei seien $\mu_1, \mu_2 \in \mathbb{R}$ sowie $\sigma_1, \sigma_2 > 0$ jeweils unbekannt. Weiter sei $\alpha \in (0,1)$, und es bezeichnen

$$\overline{X} := \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \quad , \quad \overline{Y} := \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j \quad \text{und}$$

$$\hat{\sigma}_1 := \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \overline{X})^2} \quad , \quad \hat{\sigma}_2 := \sqrt{\frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \overline{Y})^2}$$

die arithmetischen Mittel und die Stichproben-Standardabweichungen zu $X_1, \dots, X_{n_1}$ bzw. zu $Y_1, \dots, Y_{n_2}$.

(1) F-Test zum Vergleich der Varianzen zweier unabhängiger Stichproben zum Niveau α

Teststatistik

$$F := \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

Es bezeichne F_{real} die zugehörige realisierte Teststatistik, die sich aus der Ersetzung von $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$ durch $x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}$ ergibt.

Testentscheidungen

Hypothesen		Testentscheidung	Zugehöriger p-Wert
H_0	H_1	Ablehnung von H_0, falls	
$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$F_{\text{real}} > F_{1-\alpha}(n_1 - 1, n_2 - 1)$	$P_0(F > F_{\text{real}})$
$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F_{\text{real}} < F_{\alpha}(n_1 - 1, n_2 - 1)$	$P_0(F < F_{\text{real}})$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F_{\text{real}} < F_{\alpha/2}(n_1 - 1, n_2 - 1)$ oder $F_{\text{real}} > F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$	$2\,P_0(F > F_{\text{real}})$, falls $F_{\text{real}} \geq 1$ $2\,P_0(F < F_{\text{real}})$, falls $F_{\text{real}} < 1$

Hierbei bezeichnen $F_\beta(k_1, k_2)$ das β-Quantil der F-Verteilung mit erstem Freiheitsgrad $k_1 \in \mathbb{N}$ und zweitem Freiheitsgrad $k_2 \in \mathbb{N}$ für $\beta \in (0,1)$ sowie P_0 das zugrundeliegende Wahrscheinlichkeitsmaß bei vorliegender Gleichheit $\sigma_1^2 = \sigma_2^2$.

(Man beachte hierbei, dass die Verteilung der Zufallsvariablen F für $\sigma_1^2 = \sigma_2^2$ gemäß C 9.9 und C 9.13 unabhängig von der speziellen Wahl von $\mu_1, \mu_2 \in \mathbb{R}$ und $\sigma_1 = \sigma_2 \in (0, \infty)$ ist.)

(2) Zweistichproben-t-Test zum Vergleich der Erwartungswerte zweier unabhängiger Stichproben zum Niveau α für $\sigma_1^2 = \sigma_2^2$

Teststatistik

$$\hat{D} := \frac{\overline{X} - \overline{Y}}{\hat{\sigma}_{\text{pool}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{mit} \quad \hat{\sigma}_{\text{pool}} := \sqrt{\frac{(n_1 - 1)\, \hat{\sigma}_1^2 + (n_2 - 1)\, \hat{\sigma}_2^2}{n_1 + n_2 - 2}}$$

Es bezeichne $\hat{D}_{\text{real}}$ die zugehörige realisierte Teststatistik, die sich aus der Ersetzung von $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ durch $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ ergibt.

Testentscheidungen

Hypothesen		Testentscheidung	Zugehöriger p-Wert
H_0	H_1	Ablehnung von H_0, falls	
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$\hat{D}_{\text{real}} > t_{1-\alpha}(n_1 + n_2 - 2)$	$P_0(\hat{D} > \hat{D}_{\text{real}})$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$\hat{D}_{\text{real}} < -t_{1-\alpha}(n_1 + n_2 - 2)$	$P_0(\hat{D} < \hat{D}_{\text{real}})$
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$\lvert\hat{D}_{\text{real}}\rvert > t_{1-\alpha/2}(n_1 + n_2 - 2)$	$P_0(\lvert\hat{D}\rvert > \lvert\hat{D}_{\text{real}}\rvert)$

Hierbei bezeichnen $t_\beta(k)$ das β-Quantil der t-Verteilung mit Freiheitsgrad $k \in \mathbb{N}$ für $\beta \in (0, 1)$ und P_0 das zugrundeliegende Wahrscheinlichkeitsmaß bei vorliegenden Gleichheiten $\mu_1 = \mu_2$ und $\sigma_1^2 = \sigma_2^2$ der Erwartungswerte und Varianzen.

(3) Zweistichproben-Test von Welsh zum Vergleich der Erwartungswerte zweier unabhängiger Stichproben zum Niveau α für σ_1^2, σ_2^2 beliebig

Teststatistik

$$T_W := \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{n_1}\, \hat{\sigma}_1^2 + \frac{1}{n_2}\, \hat{\sigma}_2^2}}$$

Bei Gleichheit der Erwartungswerte μ_1 und μ_2 gilt approximativ

$$T_W \sim t(k) \quad \text{mit} \quad k := \left\lfloor \frac{\left(\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2\right)^2}{\frac{1}{n_1-1}\left(\hat{\sigma}_1^2/n_1\right)^2 + \frac{1}{n_2-1}\left(\hat{\sigma}_2^2/n_2\right)^2} \right\rfloor ,$$

wobei diesmal $\hat{\sigma}_1^2$ und $\hat{\sigma}_2^2$ die realisierten Stichproben-Varianzen bezeichnen. Die (untere) Gaußklammer wird hierbei benötigt, damit für die Quantil-Berechnung ein ganzzahliger Freiheitsgrad k der relevanten t-Verteilung $t(k)$ resultiert.

Hiermit ergeben sich – analog zum Zweistichproben-t-Test – mit P_0 als zugrundeliegendem Wahrscheinlichkeitsmaß bei vorliegender Gleichheit $\mu_1 = \mu_2$ die folgenden Entscheidungs-Kriterien, wobei $T_{W,\text{real}}$ die realisierte Teststatistik bezeichne:

Hypothesen		Testentscheidung	Zugehöriger p-Wert
H_0	H_1	Ablehnung von H_0, falls	
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$T_{W,\text{real}} > t_{1-\alpha}(k)$	$P_0(T_W > T_{W,\text{real}})$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$T_{W,\text{real}} < -t_{1-\alpha}(k)$	$P_0(T_W < T_{W,\text{real}})$
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$\lvert T_{W,\text{real}}\rvert > t_{1-\alpha/2}(k)$	$P_0(\lvert T_W\rvert > \lvert T_{W,\text{real}}\rvert)$

Anpassungs-Test von Shapiro und Wilk
zur Überprüfung der Normalverteilungs-Annahme

Es seien $X_1, \ldots, X_n$ stochastisch unabhängige, identisch verteilte Zufallsvariablen mit $n \geq 2$, unbekannter Verteilungsfunktion F und existierender Varianz $\sigma^2 \in (0, \infty)$.

Der auf *S.S. Shapiro* und *M.B. Wilk* zurückgehende Anpassungstest zur Überprüfung einer zugrundeliegenden Normalverteilung basiert auf einem Regressionsmodell für die aufsteigend angeordneten Zufallsvariablen $X_{(1)}, \ldots, X_{(n)}$ und dem Vergleich des zugehörigen Regressionsschätzers für die unbekannte Varianz σ^2 mit der Stichprobenvarianz (als erwartungstreuem Schätzer für σ^2).

Falls $X_1, \ldots, X_n$ jeweils normalverteilt mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in (0, \infty)$ sind, liegt das folgende Regressionsmodell vor:

$$X_{(i)} = \mu + \sigma\, m_i + \varepsilon_i \ , \quad i \in \{1, \ldots, n\} \ ,$$

mit $m_i := \frac{1}{\sigma}\big(E(X_{(i)}) - \mu\big)$ und $\varepsilon_i := X_{(i)} - E\big(X_{(i)}\big)$ für $i \in \{1, \ldots, n\}$. Man beachte hierbei, dass die Zufallsvariablen $\varepsilon_1, \ldots, \varepsilon_n$ im Allgemeinen korreliert sind. Somit liegt ein *allgemeines lineares Regressionsmodell* vor (vgl. z.B. L. Fahrmeir, A. Hamerle, G. Tutz, *Multivariate statistische Verfahren*, 2. Aufl., de Gruyter, Berlin, 1996, Kap. 4).

In diesem Modell ergibt sich gemäß der *verallgemeinerten Methode der kleinsten Quadrate* der folgende Schätzer für σ:

$$\tilde{\sigma} := \frac{m'\, V^{-1}\, X}{m'\, V^{-1}\, m}$$

mit $m := (m_1, \ldots, m_n)'$, $X := (X_{(1)}, \ldots, X_{(n)})'$ und $V := \frac{1}{\sigma^2}\big(\mathrm{Kov}(\varepsilon_i, \varepsilon_j)\big)_{1 \leq i,j \leq n}$.

Entsprechend der eingangs erläuterten Motivation zum sogenannten Shapiro-Wilk-Test ist die zugehörige Teststatistik damit gegeben durch:

$$W := \frac{\tilde{\sigma}^2}{(n-1)\,\hat{\sigma}^2}\, \frac{(m'\, V^{-1}\, m)^2}{m'\, V^{-1}\, V^{-1}\, m} = \frac{\left(\sum_{i=1}^{n} a_i\, X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

mit $\hat{\sigma}^2$ als Stichproben-Varianz zu $X_1, \ldots, X_n$ und

$$(a_1, \ldots, a_n) := \frac{m'\, V^{-1}}{\sqrt{m'\, V^{-1}\, V^{-1}\, m}} \ .$$

Für $n \leq 50$ sind die – von n abhängigen – Koeffizienten $a_1, \ldots, a_n$ und Quantile der Verteilung von W vertafelt in S.S. Shapiro, M.B. Wilk, *An analysis of variance test for normality (complete samples)*, Biometrika 52, 1965, p. 591 - 611. Der in R verwendete Algorithmus zur Bestimmung der p-Werte des Shapiro-Wilk-Tests basiert auf P. Royston, *Remark AS R94: A remark on Algorithm AS 181: The W test for normality*, Applied Statistics 44, 1995, p. 547 - 551.

Der p-Wert (p-value)

Führt man einen statistischen Test mit einem Statistik-Programm durch, so gibt man *kein* Signifikanzniveau α ein, und man erhält als Ergebnis auch *keine Entscheidung* für eine der beiden Hypothesen. Stattdessen liefert das Programm (u.a.) den sogenannten p-Wert, dessen Verwendung im weiteren Verlauf näher erläutert wird.

Allgemein erhält man die Testentscheidung bei gegebenem Signifikanzniveau $\alpha \in (0, 1)$ mit Hilfe des p-Wertes wie folgt:

$$\boxed{\begin{array}{lcl} p\text{-Wert} < \alpha & \Longrightarrow & \text{Ablehnung von } H_0 \\ p\text{-Wert} \geq \alpha & \Longrightarrow & \text{Annahme von } H_0 \end{array}}$$

Der p-Wert ist somit das größte Signifikanzniveau, für welches die Nullhypothese H_0 noch angenommen wird.

Die Testentscheidung mit Hilfe des p-Wertes stellt *kein* neues Entscheidungs-Kriterium dar, sondern ist äquivalent zu der Bedingung, die sich aus dem Vergleich der betreffenden Teststatistik mit der zugehörigen Testschranke gemäß der *Neyman/Pearson*-Theorie ergibt. Die spezielle Darstellung des p-Wertes ist dabei abhängig von der jeweils betrachteten Test-Situation.

Die korrekte Vorgehensweise bei der Durchführung von Signifikanztests mit Hilfe des p-Wertes im Sinne der Test-Theorie von *Neyman/Pearson* sollte gemäß des folgenden Ablaufs erfolgen:

(i) Vorgabe eines Signifikanzniveaus $\alpha \in (0, 1)$
(vom Anwender **vorweg** festzulegen!).

(ii) Vergleich des berechneten p-Wertes mit dem Niveau α und Testentscheidung gemäß der oben angegebenen Vorschrift.

Relevante R-Funktionen im Überblick

Einlesen von Datensätzen

R-Funktion	Beschreibung
`read.table()`	Einlesen eines Datensatzes aus einer externen ASCII-Datei; das Ergebnis ist ein Objekt der Klasse `data.frame`.

Lage- und Streuungsmaße

R-Funktion	Beschreibung
`mean()`	Berechnung des arithmetischen Mittels.
`sd()`	Berechnung der Stichproben-Standardabweichung.
`var()`	Berechnung der Stichproben-Varianz bei Eingabe eines Vektors bzw. der Stichproben-Kovarianzmatrix bei Eingabe einer Matrix.

Statistische Tests/Konfidenzintervalle

R-Funktion	Beschreibung
`t.test()`	Berechnung von Konfidengrenzen bzw. Durchführung statistischer Tests für den Erwartungswert unter Normalverteilungsannahme bei unbekannter Varianz.
`var.test()`	Durchführung des F-Tests zum Vergleich der Varianzen zweier Stichproben unter Normalverteilungsannahme.
`shapiro.test()`	Durchführung des Anpassungs-Tests von Shapiro und Wilk zur Überprüfung der Normalverteilungs-Annahme.

Quantile (zu einer Wahrscheinlichkeitsverteilung)

R-Funktion	Beschreibung
`qvert()`	Berechnung von Quantilen zu einer bestimmten Wahrscheinlichkeitsverteilung; hierbei steht `vert` stellvertretend für die in R verwendete Bezeichnung der betreffenden Verteilungsklasse.

Übungs-Aufgaben

Aufgabe P 23

Ein Vertrag über die Lieferung von Margarine in 250g–Packungen sieht vor, dass der Abnehmer einen Preisnachlass fordern kann, wenn er zum Signifikanzniveau $\alpha = 10\%$ nachweist, dass der zu erwartende Packungsinhalt μ weniger als 250 g beträgt.

Hierzu entnimmt er der Lieferung (zufällig und unabhängig voneinander) 10 Packungen und bestimmt den jeweiligen Inhalt. Die Überprüfung ergibt die folgenden Mengen (in g):

$$249\ ,\ 251\ ,\ 247\ ,\ 250\ ,\ 248\ ,\ 252\ ,\ 250\ ,\ 246\ ,\ 250\ ,\ 248\ .$$

(a) Geben Sie die angegebenen Mengen in R als Komponenten des Vektors `margarine` ein, und vergleichen Sie (zur Kontrolle) die Einträge dieses Vektors mit den Werten aus der Aufgabenstellung.

Überprüfen Sie zunächst, ob diese Mengenangaben als Realisationen (stochastisch unabhängiger) jeweils $N(\mu, \sigma^2)$-verteilter Zufallsvariablen mit (unbekannten) Parametern $\mu \in \mathbb{R}$ und $\sigma > 0$ angesehen werden können. Führen Sie hierzu den Shapiro-Wilk-Test durch, indem Sie die folgende Programmzeile eingeben:

```
> shapiro.test(margarine)
```

Entscheiden Sie anhand des Testergebnisses zum Signifikanzniveau $\alpha = 10\%$, ob die Normalverteilungs-Annahme verworfen werden muss.

(b) Überprüfen Sie anhand der gegebenen Daten, ob der unbekannte Erwartungswert μ mit dem Sollwert $\mu_0 = 250$ übereinstimmt. Führen Sie hierzu mittels R den (zweiseitigen) Einstichproben-t-Test zu den Hypothesen

$$H_0 : \mu = \mu_0 \quad \text{gegen} \quad H_1 : \mu \neq \mu_0$$

für die oben angegebenen Mengen durch, indem Sie die folgende Programmzeile eingeben:

```
> t.test(margarine, mu=250)
```

Wie lautet die Testentscheidung bei gegebenem Signifikanzniveau $\alpha = 10\%$?

(c) Überprüfen Sie, ob der Abnehmer der Margarine–Lieferung den Preisnachlass fordern kann. Führen Sie hierzu mittels R den (einseitigen) Einstichproben-t-Test zu den Hypothesen

$$H_0 : \mu \geq \mu_0 \quad \text{gegen} \quad H_1 : \mu < \mu_0$$

für die oben angegebenen Mengen durch, indem Sie die Eingabe aus Teil (b) um die Option `alternative="less"` ergänzen.

(d) Berechnen Sie (zum Vergleich mit den Ergebnissen aus (b) und (c)) die (realisierte) Teststatistik des Einstichproben-t-Tests zu den gegebenen Daten mit Hilfe geeigneter, in R implementierter, statistischer Funktionen.

Aufgabe P 24

1982 wurde eine Untersuchung darüber durchgeführt, ob Kinder, deren Eltern in einer Fabrik arbeiteten, die Blei zur Herstellung von Batterien verwendete, einer höheren Bleibelastung ausgesetzt waren. Hierzu wurden von 33 Kindern aus solchen Familien die Bleigehalte im Blut bestimmt und mit denen von 33 gleichaltrigen Kontrollkindern aus der Nachbarschaft verglichen. Die Eltern dieser Kontrollkinder waren in Industriezweigen beschäftigt, die kein Blei verarbeiteten. Die zugehörigen Messwerte (in μg/dl) sind in der Datei `blei.dat` (unter RWTHmoodle, R-Praktikum, Datensätze zum Praktikum) enthalten.

(Quelle: Morton, D., et al. (1982), *Lead absorption in children of employees in a lead–related industry*, American Journal of Epidemiology 155, 549 – 555.)

Anhand der gegebenen Messwerte soll überprüft werden, ob der Arbeitsplatz der Eltern einen Einfluss auf den Bleigehalt im Blut der Kinder hatte. Da dieser Datensatz verbundene Stichproben enthält (Kinder, deren Eltern in einer Blei verarbeitenden Fabrik arbeiteten, im Vergleich zu Kontrollkindern, deren Eltern keiner besonderen Bleibelastung ausgesetzt waren), wird der Einstichproben-t-Test auf die Differenzwerte der Bleigehalte eines jeden Beobachtungspaares angewendet (unter der Annahme, dass diesen Differenzwerten eine Normalverteilung zugrunde liegt; vgl. S. 2).

(a) Erzeugen Sie aus den gemessenen Bleigehalten der Datei `blei.dat` die Datentabelle `blei`, und geben Sie die Messwerte in der Konsole aus. Verwenden Sie beim Einlesen die Option `header=TRUE`, um die Merkmalbezeichnungen aus der Datei `blei.dat` zu übernehmen.

(b) Erzeugen Sie den Vektor `diff`, der als Komponenten die Differenzwerte aus erster und zweiter Spalte der Datentabelle `blei` enthält, und führen Sie für diese Differenzwerte den Shapiro-Wilk-Test zur Überprüfung der Normalverteilungs-Annahme durch. Welche Testentscheidung ergibt sich zum Niveau $\alpha = 10\%$?

Führen Sie anschließend entsprechend der vorhergehenden Erläuterungen mittels R zu diesen Differenzwerten den Einstichproben-t-Test zum Sollwert $\mu_0 = 0$ durch. Entscheiden Sie anhand des Testergebnisses zum Niveau $\alpha = 5\%$, ob sich für die Kinder, deren Eltern in der Blei verarbeitenden Fabrik arbeiteten, im Vergleich zu den Kontrollkindern ein signifikant höherer erwarteter Bleigehalt ergibt.

Aufgabe P 25

In der Produktion von Eisenbahn-Waggons werden zur Durchführung von Bohrungen an Stahlträgern zwei unterschiedliche Bohrstationen B_1 und B_2 verwendet. Anhand eines geeigneten statistischen Tests soll überprüft werden, ob sich die für die Bohrvorgänge benötigten Zeiten an beiden Stationen signifikant voneinander unterscheiden.

Hierzu wurden 13 Bohrungen mit der Bohrstation B_1 und 11 Bohrungen mit der Bohrstation B_2 durchgeführt. Es ergaben sich folgende Zeiten für die Bohrvorgänge (in s):

Bohrstation B_1	67.5	59.0	51.2	61.1	51.7	55.9	55.2	55.6	54.4	60.2	61.8	60.7	68.1
Bohrstation B_2	61.8	77.7	66.7	59.6	70.8	69.5	66.4	61.1	62.9	68.5	75.2		

Die oben angegebenen Zeiten sind in der Datei `bohrzeiten.dat` (unter RWTHmoodle, R-Praktikum, Datensätze zum Praktikum) enthalten.

(a) Lesen Sie die Zeiten aus der Datei bohrzeiten.dat in R ein, erzeugen Sie hieraus eine Datentabelle mit der Bezeichnung `bohrzeiten`, und geben Sie diese Datentabelle in der Konsole aus.

Verwenden Sie beim Einlesen die Option `header=TRUE`, um die Merkmalbezeichnungen aus der Datei bohrzeiten.dat zu übernehmen, und die Option `fill=TRUE`, damit R die „fehlenden" Zeiten der Bohrstation B_2 korrekt (jeweils mit Wert `NA` für *not available*) in die Datentabelle übernimmt.

(b) Die Durchführung der statistischen Tests in den Aufgabenteilen (c) und (d) erfordert, dass die Bohrzeiten als Realisationen (gemeinsam) stochastisch unabhängiger Zufallsvariablen $X_1, \ldots, X_{13}, Y_1, \ldots, Y_{11}$ aufgefasst werden können mit $X_i \sim N(\mu_1, \sigma_1^2)$ bzw. $Y_j \sim N(\mu_2, \sigma_2^2)$ für $i \in \{1, \ldots, 13\}$, $j \in \{1, \ldots, 11\}$, wobei $\mu_1, \mu_2 \in \mathbb{R}$ und $\sigma_1, \sigma_2 > 0$ jeweils unbekannt sind.

Überprüfen Sie zunächst **für beide Stichproben** die Normalverteilungsannahmen mit Hilfe des Shapiro-Wilk-Tests.

(c) Überprüfen Sie weiterhin anhand des F-Tests, ob sich die beiden Varianzen σ_1^2 und σ_2^2 signifikant unterscheiden. Geben Sie hierzu in R die folgende Programmzeile ein:

```
> var.test(bohrzeiten[,1],bohrzeiten[,2])
```

(d) Überprüfen Sie anhand des Zweistichproben-t-Tests, ob sich für die Station B_1 signifikant kürzere erwartete Bohrzeiten als für die Station B_2 ergeben.

Führen Sie den Zweistichproben-t-Test in R durch, indem Sie in die Funktion `t.test()` als erste beiden Argumente die Vektoren (Spalten) mit den Zeiten der beiden Bohrstationen eingegeben und `var.equal=TRUE` als zusätzliche Option ergänzen, um R die Situation gleicher Varianzen für beide Stichproben anzuzeigen.

(e) In der vorliegenden Situation ist die (realisierte) Teststatistik des Zweistichproben-t-Tests gegeben durch

$$\hat{D} := \frac{\overline{x} - \overline{y}}{\hat{\sigma}_{\text{pool}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{mit} \quad \hat{\sigma}_{\text{pool}} := \sqrt{\frac{(n_1 - 1)\,\hat{\sigma}_1^2 + (n_2 - 1)\,\hat{\sigma}_2^2}{n_1 + n_2 - 2}}$$

sowie $n_1 = 13$ und $n_2 = 11$. Hierbei bezeichnen $\overline{x}$, $\overline{y}$ die arithmetischen Mittelwerte und $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ die Stichprobenvarianzen zu den Zeiten der Bohrstationen B_1 bzw. B_2.

Berechnen Sie (zum Vergleich mit dem Ergebnis aus (d)) die (realisierte) Teststatistik $\hat{D}$ mit Hilfe geeigneter, in R implementierter, statistischer Funktionen.

Entscheiden Sie in (b) – (d) jeweils, ob die betreffende Nullhypothese für übliche Sinifikanzniveaus $\alpha \in (0, 1)$ verworfen wird.