Lösungsvorschläge für die Übungsaufgaben der Vorlesung “Statistik“

Thomas Rupp
thomas@7t7.de

6. Dezember 2001

Aufgabe 17 Zu der geometrischen Motivation der Aufgabe schreibe ich hier erstmal nichts. Das soll dem Tutorium bzw. der Vorlesung vorbehalten sein.
Im weiteren bezeichne ein Vektor der Form $(a,\dots,b,\dots)^\top$ immer einen Vektor mit den ersten m Einträgen a und den restlichen n Einträgen b.
Sei $M \subset \mathbb{R}^{m+n}$ der Unterraum, der die Mittelwerte der X_i bzw. der Y_i (bzw. auch deren Erwartungswerte μ_1, μ_2) enthält:

$$M := \{(a,\dots,b,\dots)|a,b \in \mathbb{R}\} =< d > \oplus < e > .$$

Dabei ist d die Diagonale und e die zu d in M orthogonale Komponente. Um uns später die Rechnung zu erleichtern, wählen wir diese geschickt, nämlich die beiden orthogonalen, auf Länge Eins normierten Vektoren

$$d := \left(\frac{1}{\sqrt{m+n}},\dots,\frac{1}{\sqrt{m+n}},\dots\right)^\top$$

und

$$e := \left(\sqrt{\frac{n}{m}}\sqrt{\frac{1}{n+m}},\dots,-\sqrt{\frac{m}{n}}\sqrt{\frac{1}{n+m}}\right)^\top .$$

Das diese normiert und orthogonal sind, folgt aus

$$\begin{aligned}
\|d\|^2 &= \langle d,d\rangle = (m+n)\frac{1}{m+n} = 1\\
\|e\|^2 &= \langle e,e\rangle = m\frac{n}{m(m+n)} + n\frac{m}{n(m+n)} = (m+n)\frac{1}{m+n} = 1\\
\langle d,e\rangle &= m\sqrt{\frac{n}{(m+n)^2 m}} - n\sqrt{\frac{m}{(m+n)^2 n}} = \sqrt{\frac{mn}{(m+n)^2}} - \sqrt{\frac{mn}{(m+n)^2}} = 0
\end{aligned}$$

Das diese wirklich eine Orthogonalbasis für den Raum $M =< (\bar{X},\dots,0,\dots)^\top > \oplus < (0,\dots,\bar{Y},\dots)^\top >$ ist, kann man nachrechnen Es ist

$$(\bar{X},\dots,0,\dots)^\top = \bar{X}(ad+be) \text{ und } (0,\dots,\bar{Y},\dots)^\top = \bar{Y}(a'd+b'e).$$

Wobei $a := \left(\sqrt{m/n/(n+m)}\sqrt{n+m}\right) / \left(\sqrt{n/m/(m+n)} + \sqrt{m/n/(m+n)}\right)$ und $b := 1/\left(\sqrt{n/m/(m+n)} + \sqrt{m/n/(m+n)}\right)$, sowie $a' := \frac{m}{n}a, b' := -b$.

Die von uns gesuchten Räume werden $E :=< e >$ und $F := M^\perp$ sein. Diese haben Dimension 1 und $m+n-2$.

Da $\|\mathcal{P}_E \mathfrak{Z}\|^2 = \frac{|\langle e,\mathfrak{Z}\rangle|^2}{\|e\|^2} \Rightarrow \|\mathcal{P}_E\mathfrak{Z}\| = |\langle e,\mathfrak{Z}\rangle|$, da $\|e\|^2 = 1$. Nun gilt es noch folgendes durchzurechnen:

$$\begin{aligned}
|\langle e,\mathfrak{Z}\rangle| &= \left|\sum_{i=1}^m \left(Z_i\sqrt{\frac{n}{m(m+n)}}\right) - \sum_{i=1}^n \left(Z_i\sqrt{\frac{m}{n(m+n)}}\right)\right|\\
&= \frac{1}{\sigma}\left|\sqrt{\frac{n}{m(m+n)}}\frac{m}{m}\sum_{i=1}^m (X_i-\mu_1) - \sqrt{\frac{m}{n(m+n)}}\frac{n}{n}\sum_{i=1}^n (Y_i-\mu_2)\right|\\
&= \frac{1}{\sigma}\left|\sqrt{\frac{nm^2}{m(m+n)}}(\bar{X}-\mu_1) - \sqrt{\frac{mn^2}{n(m+n)}}(\bar{Y}+\mu_2)\right|\\
&= \frac{1}{\sigma}\sqrt{\frac{mn}{m+n}}\left|\bar{X}-\mu_1-\bar{Y}+\mu_2\right|\\
&= \frac{1}{\sigma}\sqrt{\frac{1}{\frac{m+nn}{mn}}}\left|\bar{X}-\mu_1-\bar{Y}+\mu_2\right| = \left|\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{1}{m}+\frac{1}{n}}\sigma}\right|
\end{aligned}$$

Beim zweiten Teil machen wir uns den Pythagoras doppelt zu Nutze, denn es gilt ja

$$\|\mathcal{P}_F\mathfrak{Z}\|^2 = \|\mathfrak{Z}\|^2 - \|\mathcal{P}_M\mathfrak{Z}\|^2 = \|\mathfrak{Z}\|^2 - \left(\|\mathcal{P}_{<d>}\mathfrak{Z}\|^2 + \|\mathcal{P}_{<e>}\mathfrak{Z}\|^2\right).$$

Also können wir, da d und e ja normiert sind, folgendes durchrechnen

$$\begin{aligned}
\|\mathcal{P}_F\mathfrak{Z}\|^2 &= \langle\mathfrak{Z},\mathfrak{Z}\rangle^2 - \langle d,\mathfrak{Z}\rangle^2 - \langle e,\mathfrak{Z}\rangle^2 \\
&= \langle\mathfrak{Z},\mathfrak{Z}\rangle^2 - \left(\sum_{i=1}^{m+n} Z_i\sqrt{\frac{1}{m+n}}\right)^2 \\
&\quad - \left(\left(\sum_{i=1}^{m} Z_i\sqrt{\frac{n}{m(m+n)}}\right) + \left(-\sum_{i=1}^{n} Z_i\sqrt{\frac{m}{n(m+n)}}\right)\right)^2 \\
&= \langle\mathfrak{Z},\mathfrak{Z}\rangle^2 - \frac{1}{\sigma^2}\frac{1}{m+n}\left(\sum_{i=1}^{m}(X_i-\mu_1) + \sum_{i=1}^{n}(Y_i-\mu_2)\right)^2 \\
&\quad - \frac{1}{\sigma^2}\left(\sum_{i=1}^{m}\sqrt{\frac{n}{m(m+n)}}(X_i-\mu_1) - \sum_{i=1}^{n}\sqrt{\frac{m}{n(m+n)}}(Y_i-\mu_2)\right)^2 \\
&= \langle\mathfrak{Z},\mathfrak{Z}\rangle^2 - \frac{1}{\sigma^2}\left[\frac{1}{m+n}\left[\left(\sum_{i=1}^{m}(X_i-\mu_1)\right)^2 + 2\sum_{i=1}^{m}(X_i-\mu_1)\sum_{i=1}^{n}(Y_i-\mu_2)\right.\right. \\
&\quad \left. + \left(\sum_{i=1}^{n}(Y_i-\mu_2)\right)^2\right] + \left[\frac{n}{m(m+n)}\left(\sum_{i=1}^{m}(X_i-\mu_1)\right)^2 - 2\sqrt{\frac{nm}{mn(m+n)^2}}\right. \\
&\quad \left.\left. \cdot\sum_{i=1}^{m}(X_i-\mu_1)\sum_{i=1}^{n}(Y_i-\mu_2) + \frac{m}{n(m+n)}\left(\sum_{i=1}^{n}(Y_i-\mu_2)\right)^2\right]\right] \\
&= \langle\mathfrak{Z},\mathfrak{Z}\rangle^2 - \frac{1}{\sigma^2(m+n)}\left[m^2(\bar{X}-\mu_1)^2 + n^2(\bar{Y}-\mu_2)^2\right. \\
&\quad \left. + \frac{nm^2}{m}(\bar{X}-\mu_1)^2 + \frac{mn^2}{n}(\bar{Y}-\mu_2)^2\right] \\
&= \langle\mathfrak{Z},\mathfrak{Z}\rangle^2 - \frac{1}{\sigma^2(m+n)}\left[m(\bar{X}-\mu_1)^2(m+n) + n(\bar{Y}-\mu_2)^2(n+m)\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{m}(X-\mu_1)^2 + \sum_{i=1}^{n}(Y-\mu_2)^2 - m(\bar{X}-\mu_1)^2 - n(\bar{Y}-\mu_2)^2\right] \\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{m}\left[(X_i-\mu_1)^2 - (\bar{X}-\mu_1)^2\right] + \sum_{i=1}^{n}\left[(Y_i-\mu_2)^2 - (\bar{Y}-\mu_2)^2\right]\right) \\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{m}(X-\bar{X})^2 + \sum_{i=1}^{n}(Y_i-\bar{Y})^2\right) = \left(\frac{\sqrt{n+m-2}}{\sigma}s_{\mathfrak{X},\mathfrak{Y}}\right)^2
\end{aligned}$$

Den Schritt auf die letzte Zeile kann man entweder elementar nachvollziehen

$$\begin{aligned}
&\sum_{i=1}^{m}(X_i-\mu)^2 (=: c^2) - \sum_{i=1}^{m}(\bar{X}-\mu)^2 (=: b^2) = \sum_{i=1}^{m}(X_i-\bar{X})^2 (=: a^2) \\
\Leftrightarrow\quad &\frac{1}{2}\sum_{i=1}^{m}(X_i^2 - 2X_i\mu + \mu^2 - \bar{X}^2 + 2\bar{X}\mu - \mu^2) = \frac{1}{2}\sum_{i=1}^{m}(X_i^2 - 2\bar{X}X_i + \bar{X}^2) \\
\Leftrightarrow\quad &0 = \sum_{i=1}^{m}\bar{X}^2 - \bar{X}X_i + X_i\mu - \bar{X}\mu = m\bar{X}(\bar{X}-\mu) - \sum_{i=1}^{m}X_i(\bar{X}-\mu) = 0
\end{aligned}$$

oder man sieht es mit dem Pythagoras, da $c^2 = \|X-\mu|^2, b^2 = \|\mu - \bar{X}\|^2$ und $a^2 = \|X - \bar{X}\|^2$ ist. Und man muss natürlich wissen, dass

$$s_{\mathfrak{X},\mathfrak{Y}} := \sqrt{\frac{\sum_{i=1}^{m}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(Y_i-\bar{Y})^2}{m+n-2}}.$$

Wir wissen (zumindest aus der Vorlesung), dass $\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}$ standardnormalverteilt und $\frac{\sum(X_i-\bar{X})^2+\sum(Y_i-\bar{Y})^2}{\sigma^2}$ ist $\chi^2(m+n-2)$ verteilt und unabhängig von ersterem. Daraus folgte, dass $\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{1}{m}+\frac{1}{m}}s_{\mathfrak{X},\mathfrak{Y}}}$ Student$(m+n-2)$ verteilt ist.
Also ist in unserem jetzigen Fall

$$\frac{\|\mathcal{P}_E\mathfrak{Z}\|}{\sqrt{\frac{\|\mathcal{P}_F\mathfrak{Z}\|^2}{m+n-2}}} = \frac{\|\mathcal{P}_E\mathfrak{Z}\|\sqrt{m+n-2}}{\|\mathcal{P}_F\mathfrak{Z}\|} \sim t(m+n-2).$$

Aufgabe 18 Die Hypothese ist die, dass der Median von $\mathfrak{L}(X_1)$ genau θ ist. Da die X_i iid sind, ist der Median unter der Hypothese immer θ. Der Hypothesentest ist dann sehr analog zur Auffindung des Konfidenzintervalls: Setze

$$Z := \sum_{i=1}^{n} I_{\{\theta_i^* > \theta\}} \text{ wobei } \theta^* \text{ der Stichprobenmedian von } X_i \text{ ist.}$$

Wir gehen davon aus, dass auf einzelnen Atomen keine Masse liegt und somit, unter der Hypothese, die Wahrscheinlichkeit, dass $\theta_i^* > \theta = 0.5$ ist.
Und somit ist Z binomial zu den Parametern $0.5, n$ verteilt ist. Die Hypothese wird also unter einem Signifikanzniveau α abgelehnt, falls (mit $Z < \frac{n}{2}$) `pbinom(Z,n,0.5)`$< \frac{\alpha}{2}$ (oder $> 1-\frac{\alpha}{2}$, falls $Z > \frac{n}{2}$).

Aufgabe 19 Wir benötigen folgende Zusammenhänge:

$$X_i \text{ iid } \sim \mathfrak{N}(\mu,\sigma^2) \quad \Rightarrow \quad \frac{\sqrt{n}\bar{X}-\mu}{\sigma} \sim \mathfrak{N}(0,1) \tag{1}$$

$$X_i \text{ iid } \sim \mathfrak{N}(\mu,\sigma^2) \quad \Rightarrow \quad \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2_{n-1} \tag{2}$$

$$Z \sim \mathfrak{N}(0,1), Y \sim \chi^2_k \quad \Rightarrow \quad \frac{Z}{\sqrt{\frac{Y}{k}}} \sim T_k \tag{3}$$

Mit $W := \frac{\sqrt{n}(\bar{X}-\mu)}{s_{\mathfrak{X}}}$ folgt

$$\begin{aligned}
P(-q \le W \le q) &= P(-q \le \frac{\sqrt{n}\bar{X}-\sqrt{n}\mu}{s_{\mathfrak{X}}} \le q) \\
&= P(-qs_{\mathfrak{X}} \le \sqrt{n}\bar{X}-\sqrt{n}\mu \le qs_{\mathfrak{X}}) \\
&= P(-qs_{\mathfrak{X}} - \sqrt{n}\bar{X} \le -\sqrt{n}\mu \le qs_{\mathfrak{X}} - \sqrt{n}\bar{X}) \\
&= P(\frac{qs_{\mathfrak{X}}}{\sqrt{n}} + \bar{X} \ge \mu \ge -\frac{qs_{\mathfrak{X}}}{\sqrt{n}} + \bar{X}) \\
&= P(\bar{X} - \frac{qs_{\mathfrak{X}}}{\sqrt{n}} \le \mu \le \bar{X} + \frac{qs_{\mathfrak{X}}}{\sqrt{n}})
\end{aligned}$$

Es gilt nun also W genauer zu untersuchen. So ist

$$s_{\mathfrak{X}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{X}-X_i)^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}((\bar{Y}+Z)-(Y_i+Z))^2} = s_{\mathfrak{Y}}.$$

Somit gilt nun

$$\begin{aligned} W &= \frac{\sqrt{n}(\bar{Y}+Z-\mu)}{s_{\mathfrak{X}}} = \frac{\sigma}{\sigma}\frac{\sqrt{n}(\bar{Y}-\mu)+\sqrt{n}Z}{s_{\mathfrak{Y}}} = \frac{A+B}{s_{\mathfrak{Y}}} \\ \text{wegen (1)} \quad & A \sim \sigma \cdot \mathfrak{N}(0,1) \sim \mathfrak{N}(0,\sigma^2) \text{ und } B \sim \mathfrak{N}(0, nc\sigma^2) \\ \Rightarrow W &= \frac{C}{s_{\mathfrak{Y}}} \text{ mit } C \sim \mathfrak{N}(0, \sigma^2(1+nc)) \\ &= \frac{\sigma\sqrt{1+nc}A'}{\sigma\sqrt{\frac{1}{n-1}\frac{1}{\sigma^2}\sum(\bar{Y}-Y_i)^2}} \text{ mit } A' \sim \mathfrak{N}(0,1) \\ \text{wegen (3) und (2)} \quad &\sim \sqrt{1+nc}T_{n-1} \end{aligned}$$

Die X_i sind ja nicht unabhängig, aber es gilt jetzt mit $\widetilde{W} \sim t_{n-1}$

$$\begin{aligned} & P(\bar{X} - \frac{qs_{\mathfrak{X}}}{\sqrt{n}} \le \mu \le \bar{X} + \frac{qs_{\mathfrak{X}}}{\sqrt{n}}) = P(-q \le W \le q) \\ = \quad & P(\frac{-q}{\sqrt{1+cn}} \le \widetilde{W} \le \frac{q}{\sqrt{1+cn}}) \end{aligned}$$

Mit $n = 100$ und $q = \texttt{qt(0.975,99)} = 1.984217$ ergibt sich mit

$$P(\frac{-q}{\sqrt{1+100c}} \le \widetilde{W} \le \frac{q}{\sqrt{1+100c}}) = \texttt{pt(}\frac{q}{\sqrt{1+100c}}\texttt{,99)-pt(}\frac{-q}{\sqrt{1+100c}}\texttt{,99)}$$

für die drei c's:

	$c=0$	$c=0.01$	$c=1$
P	0.95	0.836	0.156

Das Fazit der Aufgabe ist analog zu der Aufgabe 16 zu sehen: für unabhängige, normalverteilte ZVs gilt ja $\frac{\sqrt{n}(\bar{X}-\mu)}{s_{\mathfrak{X}}} \sim T_{n-1}$. In Aufgabe 16 hat man gesehen, was passiert, wenn diese nicht normalverteilt sind.
Hier sieht man, was passiert, wenn sie nicht unabhängig sind. Je grösser die "nicht eingeplanten" Abweichungen sind (c), desto eher liegt die T-Approximation falsch[1]. Ebenso mit steigenden n, da somit die "nicht eingeplante" Abweichung immer deutlicher eingeht (die Fehler in $\bar{X}$ sollen sich ja mit steigendem n herausmitteln; aber der Fehler geht jedoch unverändert und somit überproportional mit ein). Je stärker die Varianz von Z ist, desto größer sind die Abweichungen von der Null und somit auch der Fehler.

Jetzt noch der Korrelationskoeffizient (im Hinterkopf behalten wir, dass für normalverteilte ZVs Unabhängigkeit und Unkorreliertheit äquivalent sind):

$$\kappa = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}X_i}\sqrt{\mathrm{Var}X_j}}.$$

Wegen der Unabhängigkeit von Z und Y_i gilt:

$$0 = \mathrm{Cov}(Y_i, Y_j) = \mathbb{E}(Y_iY_j) - \mathbb{E}Y_i\mathbb{E}Y_j \Leftrightarrow \mathbb{E}(Y_iY_j) = \mu^2 \quad (4)$$

$$0 = \mathrm{Cov}(Z, Y_i) = \mathbb{E}(ZY_i) - \mathbb{E}Z\mathbb{E}Y_i \Leftrightarrow \mathbb{E}(ZY_i) = 0 \quad (5)$$

Also ist

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathbb{E}(X_iX_j) - \mathbb{E}X_i\mathbb{E}X_j \\ &= \mathbb{E}(Z^2 + ZY_j + ZY_i + Y_iY_j) - \mu^2 \\ &= \mathbb{E}(Z^2) - 0^2 + \mathbb{E}(ZY_i) + \mathbb{E}(ZY_j) + \mathbb{E}(Y_iY_j) - \mu^2 \\ &= \mathrm{Var}Z \qquad \text{wegen (5) und (4)} \end{aligned}$$

[1] für $c = 0$ ist die "Approximation" exakt; s..u.

Weil $\sqrt{\mathrm{Var}X_i} = \sqrt{c\sigma^2}$ ist κ nun

$$\kappa = \frac{\mathrm{Var}Z}{(1+c)\sigma^2} = \frac{c}{1+c}.$$

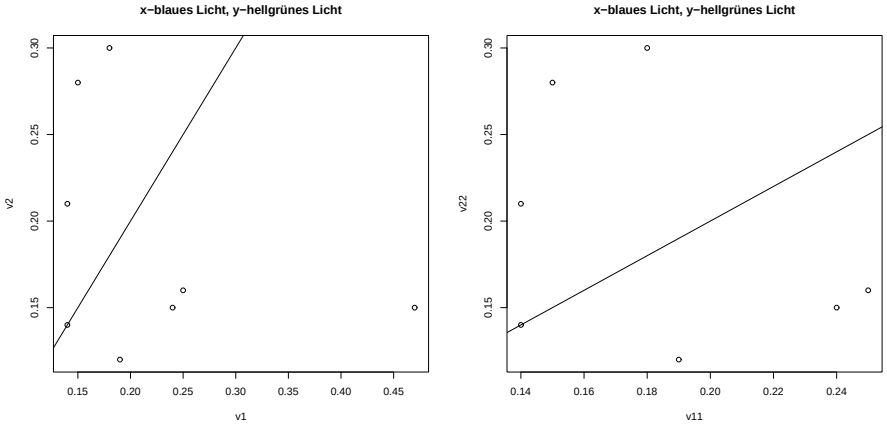
Die X_i sind also nur für $c = 0$ unkorreliert und somit unabhängig. Je grösser c ist, desto grßser ist die Parralelität zwischen den Abweichungen der X_i von μ (klar, wenn Z stärker abweicht wirkt sich das gleichermaßen auf alle X_i aus).
Ganz nebenbei noch bemerkt, dass (wegen $\mathrm{Cov}(X, X) = \mathrm{Var}X$) ganz allgemein gilt:

$$\kappa := \frac{\mathrm{Cov}(X, X)}{\sqrt{\mathrm{Var}X}\sqrt{\mathrm{Var}X}} = \frac{\mathrm{Var}X}{\mathrm{Var}X} = 1.$$

Aufgabe 20 Beim testen ist zu beachten, dass die Paarungen der Daten erhalten bleibt.
Der T-Test mit Paarung prüft die Hypothese, ob beide Datensetze den gleichen Mittelwert haben. Ein `t.test(v1,v2,paired=T)` bringt einen p-Wert von 0.567. Nach "schlimmer" wird es, wenn wir den Vogel 18 ausser acht lassen; dann kommt ein p-Wert von 0.79 heraus. Daraus lässt sich nur ableiten, dass durch diesen Test kein Indiz zur Ablehnung der Hyptohese geliefert wird.

Ein `wilcox.test(jitter(v1),jitter(v2),paired=T`, also ein Wilcoxon-Rang Test unter Beachtung der Paarung, bringt auch lediglich einen p-Wert von 0.74.

Plotten wir einmal die Vögel so, dass jeder Vogel als Punkt im $\mathbb{R}^2$ dargestellt wird. Seine x-Koordinate als Variabilität bei blauen Licht, respektive die y-Koordinate beim hellgrünen Licht (in der zweiten Grafik fehlt der Ausreisser-Vogel Nummer 18).



Wie wir sehen, lässt auch dieses Bild wenig Aufschluss zu: von den einem Ausreisser mal abgesehen, gibt es gleiche Variabilität, als auch eine recht symmetrische Abweichung nach oben und nach unten. Also können wir auch hier keine eindeutige Aussage treffen.

Alles in allem lässt wohl den Schluss nahe liegen, dass zumindest keine deutliche Abweichungen der zugrunde liegenden Verteilung vorliegen. Die Farbe des Lichtes scheint keine besonderen Auswirkungen zu haben. Einen Beweis bzw. eindeutigen Hinweis haben wir dafür jedoch nicht.