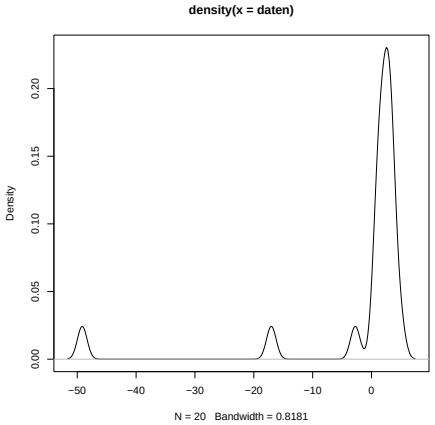
Lösungsvorschläge für die Übungsaufgaben der Vorlesung “Statistik“

Thomas Rupp
thomas@7t7.de

7. Dezember 2001

Aufgabe 21



Wie wissen von der Verteilung nur, dass sie symmetrisch um δ ist. Die beiden deutlichen Ausreisser raten mir auch davon ab, eine Normalverteilung anzunehmen.
Ein `plot(density(daten))` hilft hier auch nicht wirklich (siehe rechte Abbildung).
Bemühen wir den Wilcoxon'schen Test, der uns auch das gesuchte Konfidenzintervall angibt. Hier kommt der `median(daten)`$=$ 2.165 als Schätzwert von δ in Frage.

Ein `wilcox.test(jitter(daten),mu=median(daten),conf.int=T)` liefert als p-Wert 0.78, was eigentlich auch zu erwarten war. Als Schätzung für δ liefert der Test den Pseudomedian 1.97 und das dazugehörige Konfidenzintervall $[0.22, 2.75]$.

Will man das nicht mit R, sondern per Hand lösen, geht das natürlich auch. Berechnen eines Konfidenzintervalls für den Median:

Versehe alle Werte mit ihrem Rang. Der Median der Daten ist 2.165. Nun hat jede gezogene (mit zurücklegen) Zahl eine Wahrscheinlichkeit 0.5 kleiner als der Median zu sein; die Anzahl dieser Zahlen ist also binomial verteilt.
Gesucht ist nun die Zahl a, so dass $P(X < a) \leq \frac{\alpha}{2} = 0.025$. Dies ist $a = 5$. Also ist $[r(5), r(20 - 5 + 1)]$ aus Symmetriegründen ein 0.95er Konfidenzintervall. Dabei ist $r(\cdot)$ die Abbildung vom Rang auf ihr Urbild. Zu Rang 5 gehört die Zahl 0.96 zu 16 gehört die 3.2.
Das per Hand ermittelte Konfidenzintervall ist also: $[0.96, 3.2]$.

Die Diskrepanz kommt dadurch, dass R beim Wilcoxon-Test glättet, was wir hier nicht getan haben.

Aufgabe 22
Wir arbeiten in dieser Aufgabe mit den drei Parametern n, γ, δ mit $n > 0, 5/6 < \gamma, \delta < 1$.
Dabei ist n die Anzahl der Versuche, die wir beim Test gestatten. Den dadurch entstehende Raum der möglichen erfolgreichen Vorhersagen $A := \{0, 1, \ldots, n\}$ partitionieren wir in

$$A_0 := \{\lfloor 0.3 \cdot n \cdot \gamma \rfloor, \ldots, n\} \text{ und } A_1 := \{0, 1, \ldots, \lfloor 0.3n\gamma \rfloor - 1\}.$$

Aus dem Raum der Parameter $\Theta := \{\pi | 0 \leq \pi \leq 1\}$ nehmen wir die folgenden beiden Mengen heraus:

$$\vartheta_0 := \{\pi | \pi \geq 0.3\} \text{ und } \vartheta_1 := \{\pi | \pi < 0.3 \cdot \delta\}.$$

Schliesslich ist noch X die Anzahl der richtigen Vorhersagen.

Jetzt ist auch ersichtlich, wozu wir γ und δ benötigen. Wenn $\gamma = 1$ wäre, dann geht die Wahrscheinlichkeit, dass der Hellseher "falsch" liegt, gegen 0.5; wir wollen aber < 0.05 erreichen. Dasselbe gilt bei der Ratio für $\delta = 1$. Auch sollte $5/6 < \gamma, \delta$

einleuchtend sein[1]. Ob wir aber gegen die Hypothese testen, ob seine Ratewahrscheinlichkeit kleiner als 0.3 ist oder genau 0.25 oder ≤ 0.25 ist nicht festgelegt.

Nun sollen sowohl der Fehler erster Art, als auch zweiter Art kleiner als 0.05 sein.

Angenommen ϑ_0 stimmt, dann wird die Behauptung des Hellsehers fälschlicherweise abgelehnt, falls $\{X < \lfloor 0.3n\gamma \rfloor\}$ eintritt. Dies geschieht im schlimmsten Fall mit

$$P_{\vartheta_0}(X \notin A_0) = \texttt{pbinom}(\lfloor 0.3n\gamma \rfloor - 1, n, 0.3) \overset{!}{<} 0.05.$$

Nun angenommen das ϑ_1 zutrifft. Die Ratio wird verworfen, wenn $\{X \notin A_1\}$ eintritt:

$$P_{\vartheta_1}(X \notin A_1) = 1 - \texttt{pbinom}(\lfloor 0.3n\gamma \rfloor - 1, n, 0.3\delta) \overset{!}{<} 0.05.$$

Nun müssen γ, δ nurnoch so bestimmt werden, dass n möglichst klein wird.
Die auf 3 Stellen gerundete kleinste Lösung für n ist $\gamma = 0.917, \delta = 0.833$ und $n = 869$.
Die Aussage des Hellsehers wird also angenommen, falls er bei 869 Versuchen mindestens 239 mal richtig liegt, was einer noch zulässigen Quote von 0.2751 entspricht.
Der Fehler erster Art beträgt dann 0.04924, der Fehler zweiter Art 0.04929.
Wir testeten also seine Hypthese "Er rate mit mind. 30%.", gegen unsere Hypothese, er rate mit den üblichen 25%).

Wollen wir gegen die Hypothese testen, dass er mit weniger als 0.3 rät, so werden mehr Tests nötig (bei ≥ 0.3 gegen < 0.3 sind es unendlich viele). Bei $\delta = 0.899$ und $\gamma = 0.95$ sind schon 2426 Versuche nötig.

Aufgabe 23
Wir folgen dem Geist vom Lemma von Neyman und Pearson und sehen uns den Quotienten $\frac{f_0(10)}{f_1(10)} = \frac{5}{10} = \frac{1}{2} =: c$ an.
Nun gilt es alle $b \in B$ zu finden für die ebenfalls $\frac{f_0(b)}{f_1(b)} > \frac{1}{2}$ gilt (">" deswegen, damit die 10 gerade so nicht in A_0 liegt; da wir ja wissen wollen, wie wahrscheinlich es ist, ein solches oder unwahrscheinlicheres Ergebnis zu erhalten).
Dies sind

$$B' := \left\{ b | f_0(b) > \frac{1}{2} f_1(b) \right\} = \{1, 2, 5, 6, 7, 8, 9, 11, 12, 13, 15\}.$$

Der p-Wert α ist dann definiert als

$$\alpha := P_{f_0}(X \notin B') = P_{f_0}(X \in B \backslash B') = f_0(3) + f_0(4) + f_0(10) + f_0(14) = 0.25.$$

Zu beachten ist hier, das P_{f_0} die Wahrscheinlichkeit unter der Verteilung f_0 ist und keineswegs die Gleichverteilung auf B!
Wir aussagekräftig ein p-Wert von 0.2 ist, steht auf einem anderen Blatt.

Aufgabe 24
a)

[1] Ein Test, bei dem weniger als $0.25n$ richtige Tipps zur Annahme führen ist nicht intuitiv, ebensowenig wie ein Test gegen die Hypothese, dass die Ratewahrscheinlichkeit kleiner als 0.25 ist.

Sehen wir uns die beiden Seiten erstmal einzeln an:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}x_i^2-2x_i\bar{x}+\bar{x}^2=n\bar{x}^2-2\bar{x}n\frac{1}{n}\sum_{i=1}x_i+\sum_{i=1}x_i^2\\
&= \left(\sum_{i=1}^{n}x_i^2\right)-n\bar{x}^2\\
\frac{1}{n}\sum_{i<j}(x_i-x_j)^2 &= \frac{1}{n}\sum_{i<j}(x_i^2-2x_ix_j+x_j^2)\\
&= \frac{1}{n}\left[(x_1^2-2x_1x_2+x_2^2)+(x_1-2x_1x_3+x_3^2)+(x_2^2-2x_2x_3+x_3^2)\right.\\
&\quad \left.+\cdots+(x_1^2-2x_1x_n+x_n^2)+\cdots+(x_{n-1}^2-2x_{n-1}x_n+x_n^2)\right]\\
&= \frac{1}{n}\left[\sum_{i=1}^{n-1}x_1^2+\sum_{i=1}^{n-2}x_2^2+\cdots+\sum_{i=1}^{n-(n-1)}x_{n-1}^2+\sum_{i=2}^{n}(i-1)x_i^2-2\sum_{i<j}x_ix_j\right]\\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(n-i)x_i^2+\sum_{i=1}^{n}(i-1)x_i^2-2\sum_{i<j}x_ix_j\right]\\
&= \frac{1}{n}\left[\sum_{i=1}^{n}nx_i^2-\sum_{i=1}^{n}x_i^2-2\sum_{i<j}x_ix_j\right]\\
&= \sum_{i=1}^{n}x_i^2-\frac{1}{n}\left[\sum_{i=1}^{n}x_i^2+2\sum_{i<j}x_ix_j\right]
\end{aligned}$$

Also bleibt noch folgendes zu zeigen

$$\begin{aligned}
& n^2\bar{x}^2=\sum_{i=1}^{n}x_i^2+2\sum_{i<j}x_ix_j\\
\Leftrightarrow\quad & \left(\sum_{i=1}^{n}x_i\right)^2-\sum_{i=1}^{n}x_i^2=2\sum_{i<j}x_ix_j\\
\Leftrightarrow\quad & (x_1+x_2+\ldots x_n)^2-x_1^2-\sum_{i=1}^{n}x_i^2=2\sum_{i<j}x_ix_j
\end{aligned}$$

Hier kann man aufhören und sehen, dass sich die Quadrate nach dem quadrieren gerade durch die Summe aufheben und nurnoch die Mischterme übrig bleiben.

b)

Da $Z_i-Z_j=(X_i-\mu)-(X_j-\mu)=X_i-X_j$ können wir bei i),ii) und iii) die X durch deren Zentrierung ersetzung: $\mathbb{E}(Z)=0,\mathbb{E}(Z^2)=\sigma^2,\mathbb{E}(Z^4)=m_4$.

i)

$$\begin{aligned}
\mathrm{Var}\left((Z_1-Z_2)^2\right) &= \mathbb{E}\left((Z_1-Z_2)^4\right)-\left(\mathbb{E}\left[(Z_1-Z_2)^2\right]\right)^2\\
&= \mathbb{E}(Z_1^4)-4\mathbb{E}(Z_1^3Z_2)+6\mathbb{E}(Z_1^2Z_2^2)-4\mathbb{E}(Z_1Z_2^3)+\mathbb{E}(Z_2^4)\\
&\quad -\left(\mathbb{E}(Z_1^2)-2\mathbb{E}(Z_1Z_2)+\mathbb{E}(Z_2^2)\right)^2\\
&= m_4-0+6\sigma^4-0+m_4-\underbrace{(\sigma^2-0+\sigma^2)^2}_{4\sigma^2}\\
&= 2(m_4+\sigma^4)
\end{aligned}$$

ii)

$$\begin{aligned}
\mathrm{Cov}\left[(Z_1-Z_2)^2,(Z_1-Z_3)^2\right] &= \mathbb{E}\left[(Z_1-Z_2)^2(Z_1-Z_3)^2\right] - \overbrace{\mathbb{E}(Z_1-Z_2)^2\mathbb{E}\left[(Z_1-Z_3)^2\right]}^{4\sigma^2} \\
&= \mathbb{E}(Z_1^4 - 2Z_1^3Z_3 + Z_1^2Z_3^2 - 2Z_1Z_3 + 4Z_1^2Z_2Z_3 - 2Z_1Z_2Z_3^2 \\
&\quad + Z_2^2Z_3^2 - 2Z_2^2Z_1Z_3 + Z_2^2Z_3^2) - 4\sigma^2 \\
&= m_4 - 0 + \sigma^4 - 0 + 0 - 0 + \sigma^4 - 0 + \sigma^4 - 4\sigma^4 \\
&= m_4 - \sigma^4
\end{aligned}$$

iii)

$$\mathrm{Var}\left[\sum_{i<j}(Z_i-Z_j)^2\right] = \sum_{\substack{i<j\\k<l}}\mathrm{Cov}\left[(Z_i-Z_j)^2,(Z_k-Z_l)^2\right]$$

Nun gilt aber (unter $i<j, k<l$) für die einezlnen Kovarianzen folgendes:

$$\mathrm{Cov}\left[(Z_i-Z_j)^2,(Z_k-Z_l)^2\right] = \begin{cases} 2(m_4+\sigma^4) & i=l \text{ und } j=k \\ m_4-\sigma^4 & i=k, j\neq l \text{ oder } j=l, i\neq k \\ & \text{oder } i=l, j\neq k \text{ oder } j=k, i\neq l \\ 0 & \text{sonst.} \end{cases}$$

Wieviel gültige Kombinationsmöglichkeiten gibt es für den ersten Fall?
Genau $\sum_{i=1}^{n-1}(n-i) = \frac{1}{2}n(n-1)$ viele (wenn $i=j=a$, dann gibt es für $j=k$ noch $n-a$ viele Möglichkeiten; wegen $i<j\leq n$ geht die Summe nur bis $n-1$).

Wie sieht es mit dem zweiten Fall aus? Hier sind 4 Summen zu bilden:

$i=k$: $\sum_{i=1}^{n-1}(n-i)(n-i-1)$

Für jedes feste $i=k$ kann man noch $n-i$ Ziffern für j und dann noch (da $j\neq l$) $n-i-1$ viele für l auswählen.

$j=l$: $\sum_{i=1}^{n}(i-1)(i-2)$

Wenn $j=l$ fest ist, bleiben $i-1$ Möglichkeiten für i und dann noch $i-2$ für k übrig.

$i=l$: $\sum_{i=2}^{n-1}(n-i)(i-1)$

Für jedes feste $i=l$ bleiben $n-i$ Möglichkeiten für j und $i-1$ für k übrig.

$j=k$: $\sum_{i=2}^{n-1}(i-1)(n-i)$

Dasselbe wie eben, nur mit vertauschten Rollen.

Addieren dieser Summen fördert $n^3-3n^2+2n = n(n-1)(n-2)$ zutage.

Setzen wir das Ergebnis zusammen, erhalten wir das gewünschte

$$n(n-1)(m_4+\sigma^4) + n(n-1)(n-2)(m_4-\sigma^4).$$

Und letzendlich noch

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] &= \mathrm{Var}\left[\frac{1}{n(n-1)}\sum_{i<j}^{n}(X_i-X_j)^2\right] \\
&= \left(\frac{1}{n(n-1)}\right)^2 \mathrm{Var}\left[\sum_{i<j}^{n}(X_i-X_j)^2\right] \\
&= \frac{1}{n}\frac{1}{n-1}(m_4+\sigma^4+(n-2)(m_4-\sigma^4) \\
&= \frac{m_4(n-1)-\sigma^4(n-3)}{n(n-1)} \\
&= \frac{1}{n}\left[m_4-\sigma^4\frac{n-3}{n-1}\right]
\end{aligned}$$