Lösungsvorschläge für die Übungsaufgaben der Vorlesung “Statistik“

Thomas Rupp
thomas@7t7.de

20. Dezember 2001

Aufgabe 29 Sehen wir uns die Differenz der quadratischen Fehler der Schätzer s^2 und $\hat{\sigma}^2$ einmal an:

$$\begin{aligned}
& \mathbb{E}\left[(s^2-\sigma^2)^2\right] - \mathbb{E}\left[(\hat{\sigma}^2-\sigma^2)^2\right] \\
= & \overbrace{\left(\mathbb{E}(s^2)-\sigma^2\right)^2 + \mathrm{Var}(S^2)}^{28a)} - \overbrace{\left[\frac{\sigma^4}{n^2} + \left(\frac{n-1}{n}\right)^2 \frac{1}{n}\left(m_4 - \frac{n-3}{n-1}\sigma^4\right)\right]}^{28c)} \\
= & 0^2 + \underbrace{\frac{1}{n}\left(m_4 - \frac{n-3}{n-1}\sigma^4\right)}_{24)} - \frac{\sigma^4}{n^2} - \left(\frac{n-1}{n}\right)^2 \frac{1}{n}\left(m_4 - \frac{n-3}{n-1}\sigma^4\right)
\end{aligned}$$

Ein `simplify(..)` dieses Ausdruckes ergibt

$$= \frac{2n^2 m_4 - 3n^2\sigma^4 + 8n\sigma^4 - 3nm_4 + m - 3\sigma^4}{n^3(n-1)}.$$

$\mathbb{E}\left[(s^2-\sigma^2)^2\right]$ ist genau dann größer, wenn der obige Ausdruck positiv ist:

$$\cdots > 0 \Leftrightarrow m_4(2n^2-3n+1) + \sigma^4(8n-3n^2-3) > 0 \Leftrightarrow \frac{m_4}{\sigma^4} > \frac{3n^2-8n+3}{(2n-1)(n-1)}.$$

Welcher Varianz-Schätzer besser ist, hängt allgemein also zum einen von der Anzahl der beobachteten Daten (n) ab, zum anderen aber auch von σ und m_4; also auch von der Verteilung.
Nimmt man sich aber bestimmte Verteilungen heraus, so kann man folgendes sehen:

$\mathfrak{N}(\mu, \sigma^2)$:
Für normalverteilte X_i gilt (Berechnung wie in Aufgabe 24; hier nur die Ergebnisse, die MAPLE liefert): $\sigma^2 = \sigma^2$ (welch Wunder) und $m_4 = 3\sigma^4$. Also gilt

$$\frac{m_4}{\sigma^4} = 3 > \frac{3}{2} \geq \frac{3n^2-8n+3}{(2n-1)(n-1)} \forall n > 1.$$

$U(a,b)$:
Für die uniforme Verteiltung gilt $\sigma^4 = \frac{(b-a)^4}{144}$ und $m_4 = \frac{(a-b)^4}{80}$. Somit gilt für den Quotienten

$$\frac{m_4}{\sigma^4} = \frac{9}{5} > \frac{3}{2} \geq \frac{3n^2-8n+3}{(2n-1)(n-1)} \forall n > 1.$$

Also hat für die normale und die uniforme Verteilung immer der ML-Schätzer für die Varianz den kleineren, erwarteten quadratischen Fehler.

$\frac{1}{2}(\delta_a + \delta_b)$:

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \frac{a+b}{2})^2\right) = \frac{1}{2}\left[\left(\frac{a-b}{2}\right)^2 + \left(\frac{b-a}{2}\right)^2\right] \Rightarrow \sigma^4 = \left(\frac{a-b}{2}\right)^4.$$

$$m_4 = \frac{1}{2}\left[\left(\frac{a-b}{2}\right)^4 + \left(\frac{b-a}{2}\right)^4\right] = \left(\frac{a-b}{2}\right)^4.$$

Hier gilt also, im Gegensatz zu vorher:

$$\frac{m_4}{\sigma^4} = 1 \geq \frac{3n^2-8n+3}{(2n-1)(n-1)} \forall n > 4.$$

Der ML-Schätzer hat also erst ab einer Stichprobengröße von $n = 4$ den kleineren, erwarteten Fehler.

Aufgabe 30 Wie sieht die benötigte Teststatistik aus? Dazu brauchen wir erstmal die Erwartungswerte!
Die Anzahl der Versuche n ist 26306. Wie ist die Wahrscheinlichkeit p_j, dass von 12 fairen Würfeln genau j Stück 5 oder 6 Augen zeigen? Die Wahrscheinlichkeit ist natürlich binomial verteilt:

$$p_j = \binom{12}{j} \left(\frac{2}{6}\right)^j \left(\frac{4}{6}\right)^{12-j}.$$

Die χ^2-verteilte Teststatistik ($13-1$ Freiheitsgrade, da der Raum der p_j Dimension Null hat, ist dann

$$T = \sum_{j=0}^{12} \frac{(k_j - np_j)^2}{np_j} = 41.312.$$

Da `pchisq(41.312,12)`$= 0.9999565$ kann die Hypothese, dass es sich um faire Würfel handelt sogar auf einem 99.99% Niveau verworfen werden.

Aufgabe 31
Alles spielt sich nun im $\mathbb{R}^n$ ab.

a)

$\mathfrak{Y}$ ist $\mathfrak{N}(0, C)$-verteilt. Weil C regulär ist, existiert ein A, so dass $AA^\top = C$ ist. Multipliziert man eine standardnormalverteilte ZV $\mathfrak{Z}$ mit der Wurzel der Varianz einer anderen normalverteilten ZV, dann ist das Produkt verteilt wie eben diese: $\mathfrak{Z} \sim \mathfrak{N}(0, I) \Rightarrow A\mathfrak{Z} = \mathfrak{Y}$. Daher folgt

$$\begin{aligned}
\mathfrak{Y}^\top C^{-1} \mathfrak{Y} &= (A\mathfrak{Z})^\top C^{-1} (A\mathfrak{Z}) \\
&= \mathfrak{Z}^\top A^\top C^{-1} A\mathfrak{Z} \\
&= \mathfrak{Z}^\top A^\top (AA^\top)^{-1} A\mathfrak{Z} \\
&= \mathfrak{Z}^\top A^\top (A^\top)^{-1} A^{-1} A\mathfrak{Z} \\
&= \mathfrak{Z}^\top \mathfrak{Z} = \|\mathfrak{Z}\|^2 \\
&= Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n)
\end{aligned}$$

Wie man leicht einsieht spielt es keine Rolle, ob $\mathfrak{Y}$ zentriert war.

b)

Sei $\mathfrak{X} = \vec{\mu} + \mathfrak{E}, \mathfrak{E} \sim \mathfrak{N}(0, C), C$ regulär. Nach Aufgabenteil b) ist $\mathfrak{E}^\top C^{-1} \mathfrak{E}$ dann $\chi^2(n)$ verteilt.
Wir schätzen das $\vec{\mu}$ jetzt dadurch, indem wir annehmen, dass $\mathfrak{X} - \vec{y}$ zentriert ist und testen, ob wie diese Hypothese annehmen können; nun spielt also, unter der Hypothese, $(\mathfrak{X} - \vec{y}) = \mathfrak{E}$ die Rolle von $\mathfrak{Y}$ aus Aufgabenteil a):
Das $(1-\alpha)$ Konfidenzintervall ist

$$K(\mathfrak{X}) := \{y \in \mathbb{R}^n | (\mathfrak{X} - \vec{y})^\top C^{-1} (\mathfrak{X} - \vec{y}) \leq \text{qchisq(1-}\alpha\text{,n)} =: q\}.$$

Dies ist in der Tat das gesuchte Intervall, da für den wahren $\vec{\mu}$ gilt

$$P(\vec{\mu} \in K) = P((\mathfrak{X} - \vec{\mu})^\top C^{-1} (\mathfrak{X} - \vec{\mu}) =\leq q) = P(\underbrace{\mathfrak{E}^\top C^{-1} \mathfrak{E}}_{\|\mathfrak{Z}\|^2} \leq q) = 1 - \alpha.$$

Wenn C die Einheitsmatrix ist, ist die Menge K die Normkugel im $\mathbb{R}^n$; ansonsten wird die Menge (zu einem Ellipsoid) verzerrt.

Aufgabe 32
Hier geht es natürlich um den χ^2-Test. Mir sind hierzu zwei mögliche Ansätze eingefallen. Ersteinmal der klassische:

a)

Die G_i's sind alle iid und gleich 0.5. Die Wahrscheinlichkeit, das ein Kind männlich ist, ist $p_j = p = 0.5 \quad \forall j$. Der Raum F der möglichen p_j ist also nulldimensional. Die erwartete Anzahl ist für jeden Ausgang gleich; nämlich $\left(\frac{1}{2}\right)^4 \cdot n = 2283.5$. Dabei ist natürlich $n := 36536$.
Zur besseren Übersicht können wir mit $m = 0, w = 1$ die Ausgänge als Binärzahlen auffassen. Somit ist die Bezeichnung k_j für j-ten Datensatz eindeutig.
Die Teststatistik T ist hier also $\chi^2(15)$ verteilt $(16 - 1 - \dim F)$.
Sei $T(y, x, c, v) : \mathbb{R}^4 - \mathbb{R}$ die Funktion, welche die Teststatistik auswertet. Die Funktion hat also die Form

$$\begin{aligned} T(y,x,c,v) &= \frac{(2574 - nyxcv)^2}{nyxcv} + \frac{(2469 - nyxc(1-v))^2}{nyxc(1-v)} + \dots \\ &\quad \dots + \frac{(2035 - n(1-y)(1-x)(1-c)(1-v))^2}{n(1-y)(1-x)(1-c)(1-v)}. \end{aligned}$$

Dann haben wir:

$$T(p,p,p,p) = \frac{685796}{4567} \approx 150.16.$$

Wegen $\texttt{pchisq}(150.16, 15) \approx 1$ können wir diese Hypothese ablehnen.

b)

Nun sind die $p_j = P$ zwar immernoch gleich und iid, jedoch nicht fest. Somit kommen sie aus einem eindimensionalen Raum und beim Testen haben wir nurnoch 14 Freiheitsgrade.
Die ML-Schätzer $\hat{p}_j$ ergeben sich aus den Mittelwerten der Wahrscheinlichkeiten, dass das j-te Kind männlich ist. Da hier alle identisch sein sollen, mitteln wir sie anschliessend:

$$\hat{p}_i = \sum_{i=0}^{15} k_j \cdot I_{\{G_j^i = 0\}} \cdot \frac{1}{n}.$$

Dabei ist G_j^i das Geschlecht des i-ten Kindes aus dem j-ten Datensatz. Somit ist $\hat{p}_1 = \frac{9487}{18268}, \hat{p}_2 = \frac{2355}{4567}, \hat{p}_3, = \frac{18745}{36536}$ und $\hat{p}_4 = \frac{9379}{18268}$. Und daraus bekommen wir unseren ML-Schätzer $\hat{p} = \frac{\sum_{j=1}^4 \hat{p}_j}{4} = \frac{75317}{146144} \approx 0.51536$.
Damit ergibt sich dann die Teststatistik

$$T(\hat{p}, \hat{p}, \hat{p}, \hat{p}) \approx 11.995.$$

Wegen $\texttt{pchisq}(11.995, 14) \approx 0.39$ können wir diese Hypothese nicht ohne weiteres ablehnen.

c)

Wir haben die ML-Schätzer schon in der b) berechnet und können einfach nachrechnen, dass

$$T = \sum_{j=0}^{15} \frac{(k_j - n\hat{p}_j)^2}{n\hat{p}_j} \approx 21.197.$$

Der Raum der möglichen Wahrscheinlichkeiten ist nun vierdimensional, und daher ist $T \sim \chi^2(16 - 1 - 4)$ verteilt:

$$T(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4) \approx 21.1969.$$

Wegen `pchisq`$(21.1969, 11) \approx 0.9686$ können wir diese Hypothese auf einem Niveau von 5% noch ablehnen.

Eine andere Möglichkeit wie man vorgehen kann, ist diejenige, das man sich einfach ansieht welche Wahrscheinlichkeiten $p = (p_1, p_2, p_3, p_4)$ eine Hyppthese als unwahrscheinlich klassifizieren.
Da wir für konkrete p's die χ^2-Verteilung mit 15 Freiheitsgraden betrachten müssen, lehnen wir alle diejenigen p's ab, für die

$$T(p_1, p_2, p_3, p_4) > \texttt{qchisq}(0.95, 15) \approx 25 \text{ gilt.}$$

Wenn $p_1 = p_2 = p_3 = p_4$ lehnen wir auf 5%-Niveau ab, wenn $T(p_1, p_1, p_1, p_1) \geq 25 \Leftrightarrow p < 0.5106458$ oder $p > 0.520071$ bzw. liefert uns der Test keinen Grund abzulehnen, wenn $0.5106458 < p < 0.520071$.

Wenn nun die p's verschieden sein können, wird eine einfache Darstellung von

$$\{p = (p_1, p_2, p_3, p_4) | T(p_1, p_2, p_3, p_4) > 25\}.$$

wesentlich schwieriger.
Auf einen 5%-Niveau sind alle Ausgänge abzulehnen, wenn nicht die folgenden Bedingungen erfüllt sind:

$$0.509 < p_1 < 0.529, \quad 0.505 < p_2 < 0.526,$$
$$0.503 < p_3 < 0.523, \quad 0.503 < p_4 < 0.524.$$

Diese sind erstmal suspekt und müssen noch näher geprüft werden. Welche noch abgelehnt werden, lässt sich in akzeptabler Länge hier jedoch nicht wiedergeben. Meine 32MB Datei mit den Koordinaten meines Konfidenzbereichs hier wiederzugeben ist wenig sinnvoll.

Hier lässt sich nun auch etwas zu der Behauptung der d) sagen. Die Annahmebereiche zeigen deutlich, dass unter den jeweiligen Hypothesen die plausiblen Wahrscheinlichkeiten über 0.5 liegen. Daher müssen mehr Söhne als Töchter produziert werden.
Die theoretische Wahrscheinlichkeit von 0.5 für alle trifft also wohl nicht zu. Ob das nun aber heisst, dass die Wahrscheinlichkeit für alle Famillien höher liegt (wahrscheinlich Abhängig von der "Position" des Kindes), oder nur deswegen, weil er bei einigen deutlich höher und anderen niedriger ist, lässt sich nicht klären.