Dr. Christof Luchsinger

Übungsblatt 5 zur Vorlesung

"Statistische Methoden"

Testtheorie: $\theta \leq \theta_0$ vs $\theta > \theta_0$, MLQ, UMP, exponentielle Familie

Herausgabe des Übungsblattes: Woche 13, Abgabe der Lösungen: Woche 14 (bis Freitag, 1615 Uhr), Besprechung: Woche 15

Standard

Aufgabe 20 [exponentielle Familie, MLQ und minimal-suffiziente Statistik] [1 Punkt]

Zeigen Sie, dass die Gamma-Verteilung (1.4.2.3) und die Binomialverteilung (1.4.1.2) zur exponentiellen Familie gehören und berechnen Sie natürliche Parameter und minimal suffiziente Statistiken dazu. Tipp: n ist jeweils *nicht* selber Parameter und kann als bekannte Zahl aufgefasst werden.

Aufgabe 21 [Satz 4.1 bzw Satz 4.9] [4+1 Punkte]

Ein Hersteller von Glühbirnen behauptet, die von Ihnen produzierten Glühbirnen hätten eine durchschnittliche Lebensdauer von 1000 Stunden. Eine Konsumentenschutzorganisation bezweifelt dies. Bevor sie an die Öffentlichkeit geht, will sie aber mit dem Hersteller zusammen eine zufällige Stichprobe vom Umfang 2000 ausgiebig testen (brennen lassen bis kaputt). Man einigt sich darauf, davon auszugehen, dass die Glühbirnen unabhängig voneinander brennen und die Lebensdauer exponentialverteilt modelliert werden kann.

a) Entwickeln Sie mit Hilfe des Lemmas von Neyman-Pearson einen Test, indem Sie vorerst davon ausgehen, dass $\lambda_0 = 1/1000$ und $\lambda_1 = 1/950$ (Sie werden sehen, dass Sie λ_1 nie wirklich brauchen). Nehmen Sie $\alpha = 0.05$ und berechnen Sie das K, genauer das K'.

Tipps: Beispiel 1 aus 4.1.2, 1.4.2.3 und qgamma(0.05,2000,0.001).

b) In der darauffolgenden Untersuchung erhielt man eine durchschnittliche Brenndauer von 967.5 Stunden. Was raten Sie als statistischer Consultant der Konsumentenschutzorganisation?

Aufgabe 22 [NP-Lemma im diskreten Fall (Satz 4.2 bzw Satz 4.9)] [4 + 1 Punkte]

Herr Meier besucht einen Banker in der Bahnhofstrasse in Zürich. Herr Meier sagt, dass er in 60 % der Fälle voraussagen kann, ob der CHF / $-Kurs morgen höher oder tiefer liegt als heute (gleichen Kurs schliessen wir mal aus). Der Banker will Herrn Meier während 10 Handelstagen testen, bevor er ihm die Verantwortung für das Devisengeschäft überträgt. Für den Banker kann man gerade so gut eine Münze werfen, um zu prognostizieren, ob der Kurs morgen höher oder tiefer liegt. Der Banker versteht was von Statistik und wird auf dem 5 % - Niveau einen Test durchführen.

a) Wie wird dieser Test voraussichtlich aussehen? Sie werden die Befehle pbinom(7,10,0.5) und pbinom(8,10,0.5) brauchen. Tipp: Gleichung (4.2).

b) Herr Meier hat noch einen Bruder. Der sagt in genau 20 % der Fälle korrekt voraus, ob der Kurs sinkt oder steigt. Angenommen er kann das wirklich. Wie kann der Banker den Bruder geschickt einsetzen? Die Lösung dieses Problems ist nicht nur eine mathematische Spielerei, sondern ein praktisches statistisches Prinzip.

Honours

Aufgabe 23 [Nehmen Krankheitsfälle signifikant zu?] [1+1+2 Punkte]

Zur Modellierung von Krankheitsfällen (z.B. Creutzfeldt-Jakob CJD) pro Jahr in einem Land kann man zum Beispiel eine Poisson-Zufallsgrösse (vgl 1.4.1.5) einsetzen. In der Vorlesung Angewandte Stochastik werden wir sehen, dass dies nicht nur gut zu realen Daten passt, sondern auch aus theoretischen Gründen sinnvoll ist. Solche Übereinstimmung (praktisch passend und theoretisch fundiert) ist immer sehr wertvoll. Ansonsten hat man eine ad hoc Anpassung eines Modells an einen konkreten Datensatz - wenn wir einen neuen Datensatz erhalten, stimmt das Modell eventuell überhaupt nicht mehr, man spricht deshalb von "ad hoc"-erie.

Wir werden jetzt die Anzahl N_j von (gemeldeten) Krankheitsfällen in Jahr $j, 1 \leq j \leq n$, mit unabhängigen poissonverteilten Zufallsgrössen modellieren. Dabei sei der Parameter in Jahr j gleich θ^j (Potenz, "hoch j", nicht Index).

a) Was ist hier die minimal suffiziente Statistik der gemeinsamen Wahrscheinlichkeitsfunktion für θ über alle n Jahre?

b) Ist MLQ erfüllt ($\theta_0 = 1$ vs $\theta_1 > 1$ beliebig)?

c) Geben Sie einen UMP-Test der Hypothesen $\theta = 1$ vs $\theta > 1$ an (konkrete Zahlen nicht ausrechnen). Was raten Sie als statistischer Consultant der Gesundheitsbehörde, wenn Sie die Alternativ-Hypothese annehmen müssen (was ist dann konkret los)?

Technisches Detail: die gleiche Verteilung der Zufallsgrössen der verschiedenen Jahre wird nicht gefordert. Die bisherige Theorie kann trotzdem eingesetzt werden (freiwillige HA: gehen Sie dazu die bisherige Theorie durch).

Bemerkung zur Modellierung: Wir haben Unabhängigkeit der Anzahl Fälle pro Jahr gefordert. Das heisst unter anderem, dass wenn wir in Jahr i massiv mehr als die erwarteten θ^i Fälle haben, so heisst dies keineswegs, dass wir in Jahr $i+1$ ebenfalls massiv mehr als die erwarteten θ^{i+1} Fälle haben sollten. Damit eignet sich dieses Modell eindeutig nicht für ansteckende Krankheiten wie SARS; wenn wir dort in Woche i mehr als die ursprünglich erwarteten Fälle haben, so werden wir wohl auch in Woche $i+1$ mehr als die erwarteten Fälle haben, weil der "Überschuss" von Woche i auch fleissig "Nachkommen" produziert. CJD ist keine ansteckende Krankheit. Dass wir geometrisches Wachstum der erwarteten Fälle haben, kann hinterfragt werden - geometrisches Wachstum wäre bei ansteckenden Krankheiten ohne (oder mit ungenügenden) Gegenmassnahmen in der Anfangsphase eher angebracht.

Übungsblatt 5 zur Vorlesung "Statistische Methoden"

Olivier Warin

1. April 2011

Aufgabe 20 [exponentielle Familie, MLQ und minimal-suffiziente Statistik]
Sei n eine feste natürliche Zahl.

- **Behauptung:** Die $\Gamma(n,\lambda)$-Verteilung ($\lambda > 0$) gehört zur exponentiellen Familie. Ausserdem ist λ ein natürlicher Parameter und eine minimal suffiziente Statistik für λ von einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_m)$ lautet

$$T(\boldsymbol{x}) \;=\; -\sum_{i=1}^{m} x_i.$$

Beweis: Die Dichtefunktion f einer $\Gamma(n,\lambda)$-verteilten Zufallsgrösse lautet wie folgt:

$$f(x,\lambda) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{\Gamma(n)} \;=\; \underbrace{\frac{\lambda^n}{\Gamma(n)}}_{=:c(\lambda)} \underbrace{x^{n-1}}_{=:h(x)} e^{-\lambda x} \;=\; c(\lambda)h(x)\exp(\lambda t_1(x)),$$

wobei $t_1(x) = -x$. Nach Definition 4.10 gehört die $\Gamma(n,\lambda)$-Verteilung also zur exponentiellen Familie. Ausserdem ist (ebenfalls nach Definition 4.10) λ ein natürlicher Parameter.

Mit dem Abschnitt 4.2.3 folgt sofort, dass die Statistik

$$T(\boldsymbol{x}) = \sum_{i=1}^{m} t_1(x_i) \;=\; -\sum_{i=1}^{m} x_i$$

von einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_m)$ eine minimal suffiziente Statistik für den natürlichen Parameter λ ist.

■

- **Behauptung:** Die $\mathrm{Bin}(n,\theta)$-Verteilung ($\theta \in (0,1)$) gehört zur exponentiellen Familie. Ausserdem ist $\log \frac{\theta}{1-\theta}$ ein natürlicher Parameter und eine minimal suffiziente Statistik für diesen natürlichen Parameter von einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_m)$ lautet

$$T(\boldsymbol{x}) \;=\; \sum_{i=1}^{m} x_i.$$

Beweis: Die Wahrscheinlichkeitsfunktion p eine $\mathrm{Bin}(n,\theta)$-verteilten Zufallsgrösse lautet wie folgt:

$$p(x,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \;=\; \underbrace{(1-\theta)^n}_{=:c(\theta)} \underbrace{\binom{n}{x}}_{=:h(x)} \exp\left(x \log \frac{\theta}{1-\theta}\right) \;=\; c(\theta)h(x)\exp(\theta_1 t_1(x)),$$

wobei $\theta_1 = \log \frac{\theta}{1-\theta}$ und $t_1(x) = x$.

Mit Definition 4.10 und dem Abschnitt 4.2.3 folgt damit, analog wie im ersten Beweis von dieser Aufgabe, die Behauptung.

■

Aufgabe 21 [Satz 4.1 bzw. Satz 4.9]
Ein Hersteller von Glühbirnen behauptet, die von ihnen produzierten Glühbirnen hätten eine durchschnittliche Lebensdauer von 1000 Stunden. Eine Konsumentenschutzorganisation bezweifelt dies. Bevor sie an die Öffentlichkeit geht, will sie aber mit dem Hersteller zusammen eine zufällige Stichprobe vom Umfang 2000 ausgiebig testen (brennen lassen bis kaputt). Man einigt sich darauf, davon auszugehen, dass die Glühbirnen unabhängig voneinander brennen und die Lebensdauer exponentialverteilt modelliert werden kann.

Seien $n := 2000$, $\alpha := 0.05$ und für $i = 1, \ldots, n$ sei X_i die Lebensdauer der Glühbirne i (in Stunden). Nach Obigem, gehen wir davon aus, dass die X_i's iid und exponentialverteilt sind. Sei $\lambda > 0$, so dass $X_1 \sim \mathrm{Exp}(\lambda)$.

a) Wir testen nun (auf dem Niveau α) die Hypothesen $\mathcal{H}_0 : E[X_1] = 1000$ bzw. $\mathcal{H}_1 : E[X_1] = 950$ gegeneinander. Seien $\lambda_0 := {}^1/_{1000}$ und $\lambda_1 := {}^1/_{950}$.

Da $E[X_1] = {}^1/_\lambda$ können wir die Hypothesen $\mathcal{H}_0$ und $\mathcal{H}_1$ auch wie folgt formulieren:

$$\mathcal{H}_0 : \lambda = \lambda_0, \quad \mathcal{H}_1 : \lambda = \lambda_1.$$

Seien $f_0, f_1 : \mathbb{R}^n \to \mathbb{R}$ die gemeinsamen Dichten von $X_1, \ldots, X_n$ unter $\mathcal{H}_0$ bzw. $\mathcal{H}_1$. Es gilt also für $i = 0, 1$:

$$f_i(x_1, \ldots, x_n) \overset{\text{II}}{=} \prod_{k=1}^{n} \lambda_i e^{-\lambda_i x_k} = \lambda_i^n \exp\left(-\lambda_i \sum_{k=1}^{n} x_k\right) \quad (x_1, \ldots, x_n \geqslant 0).$$

Wir schliessen: Für $K > 0$ und $x_1, \ldots, x_n \geqslant 0$ gilt:

$$\frac{f_1(x_1, \ldots, x_n)}{f_0(x_1, \ldots, x_n)} > K \Leftrightarrow \frac{\lambda_1^n}{\lambda_0^n} \exp\left((\lambda_0 - \lambda_1) \sum_{k=1}^{n} x_k\right) > K \Leftrightarrow \sum_{k=1}^{n} x_k < \frac{\log(K \lambda_0^n \lambda_1^{-n})}{\lambda_0 - \lambda_1}.$$

Wir setzen also $K' := \frac{\log(K \lambda_0^n \lambda_1^{-n})}{\lambda_0 - \lambda_1}$. Nach dem Lemma von Neyman-Pearson (Satz 4.1) müssen wir also $K' \geqslant 0$ so bestimmen, dass gilt:

$$P_0\left[\sum_{k=1}^{n} X_k < K'\right] = \alpha.$$

Damit hätten wir auch direkt beginnen können, da wir hier bekanntlich MLQ haben und da $\sum_{k=1}^n x_k$ eine minimal suffiziente Statistik für λ ist.

Nach 1.4.2.3 hat unter der $\mathcal{H}_0$-Hypothese $\sum_{k=1}^n X_k$ eine $\Gamma(n, \lambda_0)$-Verteilung. Wir schliessen:

$$K' \underset{\text{R}}{\doteq} \texttt{qgamma(0.05,2000,1/1000)} \underset{\text{R}}{\doteq} 1927013.$$

Unser Test (nach dem Lemma von Neyman-Pearson) lautet also wie folgt ($(x_1, \ldots, x_n)$ sei dabei eine entsprechende Stichprobe):

- Falls $\sum_{k=1}^n x_k < K' \doteq 1927013$ lehne $\mathcal{H}_0$ ab.
- Falls $\sum_{k=1}^n x_k \geqslant K' \doteq 1927013$ lehne $\mathcal{H}_0$ nicht ab.

b) In der darauffolgenden Untersuchung erhielt man eine durchschnittliche Brenndauer von 967.5 Stunden. Wir bezeichnen die durch diese Untersuchung erhaltene Stichprobe mit $(x_1, \ldots, x_n)$. Es gilt also $\frac{1}{n} \sum_{k=1}^{n} x_k = 967.5$. Wir schliessen:

$$\sum_{k=1}^{n} x_k = n \cdot 967.5 = 1935000 > K' \doteq 1927013$$

Also nehmen wir, nach dem Test aus a), die Hypothese $\mathcal{H}_0$ an. Also raten wir als statistischer Consultant der Konsumentenschutzorganisation, dass sie mit ihrer Vermutung nicht an die Öffentlichkeit gehen soll bzw. sie soll gegen den Hersteller *keine* Klage erheben.

Aufgabe 22 [NP-Lemma im diskreten Fall (Satz 4.2 bzw. Satz 4.9)]
Herr Meier besucht einen Banker an der Bahnhofstrasse in Zürich. Herr Meier sagt, dass er eine Software entwickelt hat, mit der er in 60% der Fälle korrekt voraussagen kann, ob der CHF / $-Kurs morgen höher oder tiefer liegt als heute (gleichen Kurs schliessen wir aus). Der Banker will Herrn Meier während 10 Handelstagen testen. Für den Banker kann man gerade so gut eine Münze werfen, um zu prognostizieren, ob der Kurs morgen höher oder tiefer liegt.

a) Sei θ die Wahrscheinlichkeit, dass die Software von Herr Meier richtig liegt. Setze $n := 10$, $\alpha := 0.05$, $\theta_0 := 0.5$ und $\theta_1 := 0.6$. Seien weiter für $i = 1, \ldots, n$

$$X_i := \begin{cases} 1, & \text{falls Herr Meiers Software am Tag } i \text{ richtig liegt} \\ 0, & \text{sonst.} \end{cases}$$

Wir nehmen an, dass die X_i's unabhängig sind.

Wir wollen nun mit Hilfe des Lemmas von Neyman-Pearson (Satz 4.2) die Hypothesen $\mathcal{H}_0 : \theta = \theta_0$ vs. $\mathcal{H}_1 : \theta = \theta_1$ auf dem α-Niveau testen. Nach Vorlesung (genauer nach Gleichung (4.2)) müssen wir dazu $K' \in \mathbb{N}$ und $\gamma \in [0, 1)$ finden, mit

$$P_0\left[\sum_{i=1}^n X_i > K'\right] + \gamma P_0\left[\sum_{i=1}^n X_i = K'\right] = \alpha.$$

Damit hätten wir auch direkt beginnen können, da wir hier bekanntlich MLQ haben und da $\sum_{k=1}^n x_k$ eine minimal suffiziente Statistik für θ ist.

Unter der $\mathcal{H}_0$-Hypothese hat $\sum_{i=1}^n X_i$ eine $\mathrm{Bin}(n, \theta_0)$-Verteilung. Es gilt also

$$P_0\left[\sum_{i=1}^n X_i > 7\right] \underset{\text{R}}{\doteq} \texttt{pbinom(7,10,0.5,lower.tail=FALSE)} \underset{\text{R}}{\doteq} 0.05468750 > \alpha$$
$$P_0\left[\sum_{i=1}^n X_i > 8\right] \underset{\text{R}}{\doteq} \texttt{pbinom(8,10,0.5,lower.tail=FALSE)} \underset{\text{R}}{\doteq} 0.01074219 < \alpha.$$

Wir schliessen: $K' = 8$ und

$$\gamma = \frac{\alpha - P_0\left[\sum_{i=1}^n X_i > K'\right]}{P_0\left[\sum_{i=1}^n X_i = K'\right]} \underset{\text{R}}{\doteq} \texttt{(0.05 - pbinom(8,10,0.5,FALSE))/dbinom(8,10,0.5)} \underset{\text{R}}{\doteq} 0.8933333.$$

Nach dem Lemma von Neyman-Pearson (Satz 4.2) lautet unser Test also wie folgt ($(x_1, \ldots, x_n)$ sei dabei eine entsprechende Stichprobe)):

- Falls $\sum_{i=1}^n x_i > K' = 8$, lehne $\mathcal{H}_0$ ab.
- Falls $\sum_{i=1}^n x_i = K' = 8$, lehne $\mathcal{H}_0$ mit Wahrscheinlichkeit $\gamma \doteq 0.8933333$ ab.
- Falls $\sum_{i=1}^n x_i < K' = 8$, lehne $\mathcal{H}_0$ nicht ab.

b) Herr Meier hat noch einen Bruder. Der sagt in genau 20% der Fälle korrekt voraus, ob der Kurs sinkt oder steigt. Nehmen wir an, dass er das wirklich kann. Der Banker kann dies nun geschickt einsetzen indem er immer das Gegenteil annimmt, was der Bruder von Herrn Meier sagt. Somit weiss der Banker in 80% der Fälle ob der Kurs sinkt oder steigt.

Aufgabe 23 [Nehmen Krankheitsfälle signifikant zu?]
Zur Modellierung von Krankheitsfällen (z.B. Creutzfeldt-Jakob CJD) pro Jahr in einem Land kann man zum Beispiel eine Poisson-Zufallsgrösse (vgl. 1.4.1.5) einsetzen.

Wir werden jetzt die Anzahl N_j von (gemeldeten) Krankheitsfällen in Jahr j, $1 \leqslant j \leqslant n$, mit unabhängigen poissonverteilten Zufallsgrössen modellieren. Dabei sei der Parameter in Jahr j gleich θ^j (Potenz, "hoch j", nicht Index).

a) Die gemeinsame Wahrscheinlichkeitsfunktion $p(\,\cdot\,;\theta)$ lautet wie folgt ($\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathbb{N}_0^n$):

$$p(\boldsymbol{x};\theta) \overset{\text{II}}{=} \prod_{j=1}^{n} e^{-\theta^j} \frac{\theta^{jx_j}}{x_j!} = e^{-\sum_{j=1}^n \theta^j} \theta^{\sum_{j=1}^n jx_j} \prod_{j=1}^{n} \frac{1}{x_j!}.$$

Somit können wir den entsprechenden Likelihood-Quotienten berechnen ($\boldsymbol{y} = (y_1, \cdots, y_n) \in \mathbb{N}_0^n$):

$$\frac{p(\boldsymbol{y};\theta)}{p(\boldsymbol{x};\theta)} = \theta^{\sum_{j=1}^n jy_j - \sum_{j=1}^n jx_j} \prod_{j=1}^{n} \frac{x_j!}{y_j!}.$$

Dieser Ausdruck ist klar genau dann unabhängig von θ, wenn $\sum_{j=1}^n jx_j = \sum_{j=1}^n jy_j$. Also ist nach Satz 3.5 eine minimal suffiziente Statistik für θ gegeben durch

$$t(\boldsymbol{x}) = \sum_{j=1}^{n} jx_j.$$

b) Sei $\theta_0 = 1$ und $\theta_1 > 1$ beliebig. In den Notationen von Definition 4.4 gilt nun

$$g_{\theta_0\theta_1}(t(\boldsymbol{x})) := \frac{p(\boldsymbol{x},\theta_1)}{p(\boldsymbol{x},\theta_0)} \overset{\text{a)}}{=} \frac{e^{-\sum_{j=1}^n \theta_1^j} \theta_1^{\sum_{j=1}^n jx_j} \prod_{j=1}^n \frac{1}{x_j!}}{e^{-\sum_{j=1}^n 1^j} 1^{\sum_{j=1}^n jx_j} \prod_{j=1}^n \frac{1}{x_j!}} = \underbrace{\exp\left(n - \sum_{j=1}^{n} \theta_1^j\right)}_{>0} \theta_1^{t(\boldsymbol{x})}.$$

Da $\theta_1 > 1$ ist also $g_{\theta_0\theta_1}(t)$ streng monoton wachsend in t, das heisst MLQ ist erfüllt (Definition 4.4).

c) Wir testen nun die Hypothesen $\mathcal{H}_0 : \theta = \theta_0$ gegen $\mathcal{H}_1 : \theta > 1$.

Wir definieren noch $\boldsymbol{N} = (N_1, \ldots, N_n)$. Da MLQ erfüllt ist brauchen wir (entsprechend Satz 4.2) nur ein $K' \in \mathbb{N}_0$ und ein $\gamma \in [0, 1)$ zu finden, so dass gilt:

$$P_0[t(\boldsymbol{N}) > K'] + \gamma P_0[t(\boldsymbol{N}) = K'] = \alpha.$$

Dies kann man mit konkreten Zahlen leicht z.B. mit Hilfe von R machen.

Der UMP-Test sieht dann (nach Vorlesung) wie folgt aus: ($\boldsymbol{x} = (x_1, \cdots, x_n)$ sei dabei eine entsprechende Stichprobe):

- Falls $t(\boldsymbol{x}) > K'$ lehne $\mathcal{H}_0$ ab.
- Falls $t(\boldsymbol{x}) = K'$ lehne $\mathcal{H}_0$ mit Wahrscheinlichkeit γ ab.
- Falls $t(\boldsymbol{x}) < K'$ lehne $\mathcal{H}_0$ nicht ab.

Falls wir die Hypothese $\mathcal{H}_0$ ablehnen und daher die Hypothese $\mathcal{H}_1$ annehmen, bedeutet dies, dass wir uns für $\theta > 1$ entschieden haben. In diesem Fall würden die Rate der Erkrankungen zunehmen. Folglich wäre es günstig der Gesundheitsbehörde zu raten Massnahmen zur Eindämmung der Krankheit zu ergreifen.