Tilastollinen päättely, syksy 2013 - kevät 2014

Harjoitus 2

Esimerkkiratkaisut

1. Olkoon $Y_1, \ldots, Y_n \sim Exp(\lambda)$ $\perp\!\!\!\perp$. Kirjoita vastaavan tilastollisen mallin lauseke (ytf). Muodosta sitten aineistoa $\boldsymbol{y} = (y_1, \ldots, y_n)$ vastaava uskottavuus- ja log-uskottavuusfunktio sekä määritä huolellisesti perustellen parametrin λ suurimman uskottavuuden estimaatti. Hahmottele log-uskottavuusfunktion kuvaajaa.

   Ratk. Riippumattomuusoletuksen perusteella yhteistiheysfunktioksi saadaan

   $$f_{\boldsymbol{Y}}(\boldsymbol{y}; \lambda) = \prod_{i=1}^{n} f_{Y_i}(y_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i} = \lambda^n e^{-\lambda n \bar{y}},$$

   jossa $\bar{y} = (y_1 + \cdots + y_n)/n$. Uskottavuusfunktioksi saadaan

   $$L(\lambda; \boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \lambda) = \lambda^n e^{-\lambda n \bar{y}},$$

   jolloin log-uskottavuusfunktioksi tulee

   $$l(\lambda; \boldsymbol{y}) = \log L(\lambda; \boldsymbol{y}) = n \log(\lambda) - \lambda n \bar{y}.$$

   Etsitään suurimman uskottavuuden estimaattia log-uskottavuusfunktion derivaatan nollakohdista:

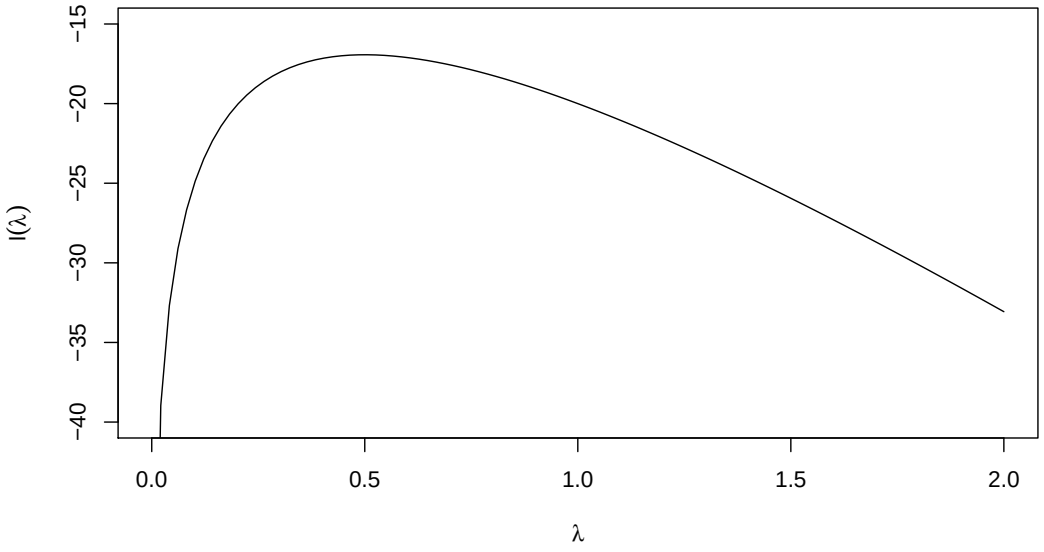
   $$l'(\lambda; \boldsymbol{y}) = \frac{d}{d\lambda} l(\lambda; \boldsymbol{y}) = \frac{n}{\lambda} - n\bar{y} \doteq 0 \Leftrightarrow \lambda = \frac{1}{\bar{y}}.$$

   Koska parametriavaruus on yhtenäinen ja avoin väli ($\Omega = (0, \infty)$), uskottavuusyhtälöllä on vain yksi ratkaisu ($1/\bar{y}$) ja log-uskottavuusfunktion toinen derivaatta

   $$l''(\lambda; \boldsymbol{y}) = \frac{d^2}{d\lambda^2} l(\lambda; \boldsymbol{y}) = -\frac{n}{\lambda^2}$$

   on negatiivinen kaikilla $\lambda > 0$ (joten se on negatiivinen myös kohdassa $1/\bar{y}$), niin derivaatan nollakohta on uskottavuusfunktion globaali maksimikohta ja suurimman uskottavuuden estimaatti on siis

   $$\hat{\lambda} = \frac{1}{\bar{y}}.$$



Kuva 1: Tehtävän 2 log-uskottavuusfunkion kuvaaja, kun $n = 10$ ja $\bar{y} = 2$.

2. (a) Olkoon $f(y;\theta) = \theta/y^{\theta+1}$, kun $y > 1$ (ja $= 0$ muulloin). Varmista, että f on erään jatkuvan jakauman tiheysfunktio, kun θ on positiivinen parametri.

   (b) Oletetaan, että satunnaismuttujat $Y_1, \ldots, Y_n$ ovat riippumattomia ja noudattavat em. jakaumaa. Muodosta syntyvän tilastollisen mallin ytf, ilmoita sen log-uskottavuusfunktio ja etsi parametrin suurimman uskottavuuden estimaatti, kun aineisto on $\boldsymbol{y} = (y_1, \ldots, y_n)$.

Ratk.

(a) $f(y;\theta)$ on tiheysfunktio, jos (i) $f(y;\theta) \geq 0$ kaikilla $y \in \mathbb{R}$ ja $\theta > 0$ ja (ii) $\int_{-\infty}^{+\infty} f(y;\theta)dy = 1$ kaikilla $\theta > 0$.

Koska

$$f(y;\theta) = \frac{\theta}{y^{\theta+1}} 1_{(1,\infty)}(y) \geq 0 \quad \text{kaikilla } y \in \mathbb{R}, \text{ kun } \theta > 0$$

ja

$$\int_{-\infty}^{+\infty} f(y;\theta)dy = \int_1^{\infty} \frac{\theta}{y^{\theta+1}} dy = \left|_1^{\infty} -\frac{1}{y^{\theta}} = -(0-1) = 1, \quad \text{kun } \theta > 0,\right.$$

niin kyseessä jatkuvan jakauman tiheysfunktio.

(b) Koska havainnot ovat riippumattomia, niin tilastollisen mallin ytf on

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\theta) \;=\; \prod_{i=1}^{n} \frac{\theta}{y_i^{\theta+1}} = \frac{\theta^n}{\left(\prod_{i=1}^n y_i\right)^{\theta+1}} = \frac{1}{\prod_{i=1}^n y_i} \cdot \frac{\theta^n}{\left(\prod_{i=1}^n y_i\right)^{\theta}}$$

Valitaan $c(\boldsymbol{y}) = \prod_{i=1}^{n} y_i$. Uskottavuusfunktioksi saadaan nyt

$$L(\theta; \boldsymbol{y}) \quad = \quad c(\boldsymbol{y}) f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta) = \frac{\theta^n}{\left(\prod_{i=1}^{n} y_i\right)^{\theta}}$$

ja log-uskottavuusfunktioksi

$$l(\theta; \boldsymbol{y}) \quad = \quad n \log(\theta) - \theta \log\left(\prod_{i=1}^{n} y_i\right) = n \log(\theta) - \theta \sum_{i=1}^{n} \log(y_i)$$

Etsitään suurimman uskottavuuden estimaatti $l(\theta; \boldsymbol{y})$:n derivaatan avulla.

$$l'(\theta; \boldsymbol{y}) \quad = \quad \frac{n}{\theta} - \sum_{i=1}^{n} \log(y_i) \doteq 0 \quad \Rightarrow \quad \hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log(y_i)}.$$

Koska parametriavaruus on yhtenäinen ja avoin väli ($\Omega = (0, \infty)$), uskottavuusyhtälöllä on vain yksi ratkaisu ($\hat{\theta}$) ja log-uskottavuusfunktion toinen derivaatta

$$l''(\theta; \boldsymbol{y}) \quad = \quad -\frac{n}{\theta^2}$$

on negatiivinen kaikilla $\theta > 0$ (joten se on negatiivinen myös kohdassa $\hat{\theta}$), niin derivaatan nollakohta on uskottavuusfunktion globaali maksimikohta ja suurimman uskottavuuden estimaatti on siis

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log(y_i)}.$$

3. Olkoon mallina $Y_1, \ldots, Y_n \sim Tas(\theta, \theta + 1)$ ⫫. Johda aineistoa $\boldsymbol{y} = (y_1, \ldots, y_n)$ vastaava uskottavuusfunktio ja totea, että se saa suurimman arvonsa jokaisessa välin $\left(y_{(n)} - 1, y_{(1)}\right)$ pisteessä, kun merkitään $y_{(1)} = \min(y_1, \ldots, y_n)$ ja $y_{(n)} = \max(y_1, \ldots, y_n)$. Siten suurimman uskottavuuden estimaatti $\hat{\theta}(\boldsymbol{y})$ ei ole yksikäsitteinen (todennäköisyydellä yksi).

Ratk. Mallia vastaava yhteistiheysfunktio on

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta) = \prod_{i=1}^{n} \frac{1}{(\theta + 1) - \theta} 1_{(\theta, \theta+1)}(y_i) = \prod_{i=1}^{n} 1_{(\theta, \theta+1)}(y_i),$$

jossa

$$1_A(y) = \begin{cases} 1 & \text{kun } y \in A, \\ 0 & \text{kun } y \notin A \end{cases}$$

on indikaattorifunktio. Nähdään, että yhteistiheysfunktion arvo on nollasta eroava (ts. ykkönen) silloin ja vain silloin kun $y_1 \in (\theta, \theta + 1)$, $y_2 \in (\theta, \theta + 1)$, $\cdots$, $y_{n-1} \in (\theta, \theta + 1)$ ja $y_n \in (\theta, \theta + 1)$. Tämä on voimassa täsmälleen silloin

kun $y_{(1)} > \theta$ ja $y_{(n)} < \theta + 1$ eli kun $\theta \in (y_{(n)} - 1, y_{(1)})$. Uskottavuusfunktioksi saadaan nyt

$$L(\theta; \boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta) = \begin{cases} 1 & \text{kun } \theta \in (y_{(n)} - 1, y_{(1)}), \\ 0 & \text{kun } \theta \notin (y_{(n)} - 1, y_{(1)}) \end{cases}$$

Uskottavuusfunktio saavuttaa siis suurimman arvonsa 1 jokaisessa välin $(y_{(n)} - 1, y_{(1)})$ pisteessä, joten suurimman uskottavuuden estimaatti ei ole yksikäsitteinen.

4. Vuonna 1898 ilmestyneessä kuuluisassa tilastossa oli raportoitu hevosenpotkuun kuolleiden miesten vuosittaiset lukumäärät neljässätoista Preussin armeijan yksikössä kahdenkymmenen vuoden ajalta, yhteensä siis 280 havaintoa. Yhteenveto tuloksista on alla.

Kuolleita	0	1	2	3	4	≥ 5
Havaintoja	144	91	32	11	2	0

Oletetaan, että kuolleiden lukumäärä yhtenä vuonna yhdessä yksikössä noudattaa Poisson-jakaumaa ja on riippumaton sekä muiden yksiköiden että muiden vuosien lukumääristä. Olkoon μ kyseisen Poisson-jakauman odotusarvo. Muodosta aineistoa vastaavan log-uskottavuusfunktion lauseke ja etsi μ:n suurimman uskottavuuden estimaatti. Mikä on suurimman uskottavuuden estimaatti todennäköisyydelle, että yhtään miestä ei kuole tietyssä yksikössä vuoden aikana?

Ratk. Olkoon

$X_{ij} =$ "Yksikössä i vuonna j kuolleiden lukumäärä", $i = 1, \ldots, 14, j = 1, \ldots, 20,$

jossa $X_{ij} \sim P(\mu)$ ovat riippumattomia. Tilastollinen malli on nyt

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \mu) = \prod_{i=1}^{14} \prod_{j=1}^{20} e^{-\mu} \frac{\mu^{x_{ij}}}{x_{ij}!} = e^{-n\mu} \frac{\mu^{\sum_{i,j} x_{ij}}}{\prod_{i,j} x_{ij}!},$$

jossa $n = 14 \cdot 20 = 280$. Uskottavuusfunktioksi voidaan valita tällöin

$$L(\mu; \boldsymbol{x}) = \left(\prod_{i,j} x_{ij}! \right) f_{\boldsymbol{X}}(\boldsymbol{x}; \mu) = e^{-n\mu} \mu^{\sum_{i,j} x_{ij}},$$

joten logaritmiseksi uskottavuusfunktioksi saadaan

$$l(\mu; \boldsymbol{x}) = -n\mu + \left(\sum_{i,j} x_{ij} \right) \log(\mu).$$

Johdetaan suurimman uskottavuuden estimaatti etsimällä funktion $l(\mu; \boldsymbol{x})$ derivaatan nollakohta.

$$l'(\mu; \boldsymbol{x}) = -n + \frac{\sum_{i,j} x_{ij}}{\mu} \doteq 0 \Rightarrow \mu = \frac{1}{n} \sum_{i,j} x_{ij} = \bar{x}.$$

Todetaan vielä, että kyseessä on maksimikohta. Kun $\sum_{i,j} x_{ij} > 0$, niin

$$l''(\mu; \boldsymbol{x}) = -\frac{\sum_{i,j} x_{ij}}{\mu^2} < 0 \quad \text{kaikilla } \mu > 0,$$

joten derivaatan nollakohdassa on maksimikohta ($\frac{1}{n} \sum_{i,j} x_{ij} > 0$). Kun $\sum_{i,j} x_{ij} = 0$, niin logaritminen uskottavuusfunktio on

$$l(\mu; \boldsymbol{x}) = -n\mu,$$

joka saavuttaa maksimiarvon kohdassa $\mu = 0 = n^{-1} \sum_{i,j} x_{ij}$. Parametrin μ suurimman uskottavuuden estimaatti on siis

$$\hat{\mu} = \frac{1}{n} \sum_{i,j} x_{ij} = \bar{x}.$$

Kun $y_0 = 144$, $y_1 = 91$, $y_2 = 32$, $y_3 = 11$, $y_4 = 2$ ja $y_5 = 0$, niin suurimman uskottavuuden estimaatiksi saadaan

$$\hat{\mu} = \frac{144 \cdot 0 + 91 \cdot 1 + 32 \cdot 2 + 11 \cdot 3 + 2 \cdot 4}{144 + 91 + 32 + 11 + 2} = \frac{196}{280} = 0.7.$$

Lasketaan seuraavaksi suurimman uskottavuuden estimaatti todennäköisyydelle, että yhtään miestä ei kuole tietyssä yksikössä vuoden aikana. Estimoitava parametri on nyt

$$p_0 = P(X_{ij} = 0) = e^{-\mu} \frac{\mu^0}{0!} = e^{-\mu}.$$

Funktio $p_0(\mu) = e^{-\mu}$ on bijektio, koska

$$p_0 = e^{-\mu} \quad \Leftrightarrow \quad \mu = -\log(p_0).$$

Käyttämällä suurimman uskottavuuden estimaatin invarianssiominaisuutta, parametrin p_0 suurimman uskottavuuden estimaatiksi saadaan

$$\hat{p_0} = e^{-\hat{\mu}} = e^{-0.7} \approx 0.4966.$$