Tilastollinen päättely II, syksy 2015 – kevät 2016
Esimerkkiratkaisut 2 (10. ja 12. 11. 2015)

1. (Monisteen teht. 2.2.) Olkoot $y_1, \ldots, y_n$ ja a mielivaltaisia reaalilukuja. Varmista, että

$$\sum_{i=1}^{n}(y_i - a)^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{y} - a)^2,$$

kun $\overline{y} = \sum_{i=1}^{n} y_i/n$. [Tulosta käytettiin normaalijakaumamallin uskottavuusfunktion muodostamisessa, ks. monisteen kohta 2.1.4.]

Ratkaisu:

Huomataan ensin, että $y_i - a$ voidaan yhtäpitävästi kirjoittaa muodossa $y_i - a + \overline{y} - \overline{y} = (y_i - \overline{y}) + (\overline{y} - a)$, ja että $\sum_{i=1}^{n} y_i = n\overline{y}$. Sijoittamalla yhtälön vasemmalle puolelle saadaan

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - a)^2 &= \sum_{i=1}^{n}[(y_i - \overline{y}) + (\overline{y} - a)]^2 \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 + 2\sum_{i=1}^{n}(y_i - \overline{y})(\overline{y} - a) + \sum_{i=1}^{n}(\overline{y} - a)^2 \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 + 2(\overline{y} - a)\overbrace{\left(\sum_{i=1}^{n} y_i - n\overline{y}\right)}^{=0} + n(\overline{y} - a)^2 \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{y} - a)^2.
\end{aligned}$$

2. (Monisteen teht. 2.3.) Olkoot $Y_1, \ldots, Y_n \sim \text{Exp}(\lambda)$ $\perp\!\!\!\perp$. Kirjoita vastaava tilastollisen mallin lauseke (yhteistiheysfunktio). Muodosta sitten aineistoa $\mathbf{y} = (y_1, \ldots, y_n)$ vastaavat uskottavuus- ja log-uskottavuusfunktio sekä määritä huolellisesti perustellen parametrin λ suurimman uskottavuuden estimaatti. Hahmottele log-uskottavuusfunktion kuvaajaa.

Ratkaisu:

Määritellään parametriavaruus $\Omega := (0, \infty)$. Riippumattomuuden nojalla satunnaismuuttujien $Y_1, ..., Y_n$ yhteistiheysfunktio on

$$f_{\mathbf{Y}}(\mathbf{y}; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda n \overline{y}}.$$

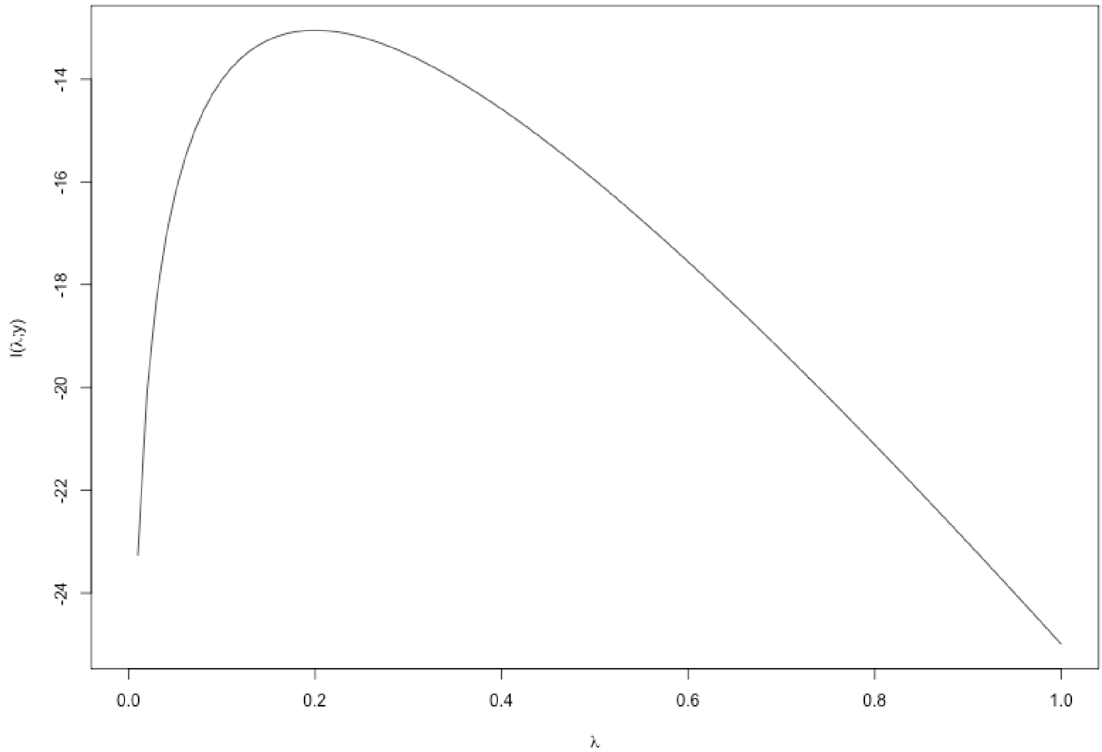
Uskottavuusfunktio on $L(\lambda; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \lambda) = \lambda^n e^{-\lambda n\overline{y}}$. Ottamalla tästä luonnollinen logaritmi saadaan log-uskottavuusfunktio $\ell(\lambda; \mathbf{y}) = n \log \lambda - \lambda n \overline{y}$.

Log-uskottavuusfunktion lokaalit maksimit (ja minimit) löydetään sen derivaatan nollakohdista:

$$\ell'(\lambda; \mathbf{y}) = \frac{n}{\lambda} - n\overline{y} = 0 \Leftrightarrow \lambda = \frac{1}{\overline{y}}.$$

Log-uskottavuusfunktio on tässä tapauksessa aidosti konkaavi, sillä kaikilla $\lambda \in \Omega$ pätee $\ell''(\lambda; \mathbf{y}) = -n/\lambda^2 < 0$. Tästä seuraa, että parametrin λ SU-estimaatti on $\hat{\lambda} = 1/\overline{y}$. Log-uskottavuusfunktion arvoja voi tutkia esimerkiksi piirtämällä R-ohjelmistolla kuvan. Alla on annettu kuvaaja, sekä sen tuottanut koodi.

Koodi:



Kuva 1: Log-uskottavuusfunktio parametrin λ funktiona. Kuvassa on käytetty arvoja $n = \overline{y} = 5$ ja se on piirretty R-ohjelmistolla.

```
y <- 5;
n <- 5;
curve(n*log(x) - x*n*y, from=0, to=1,
    xlab=expression(lambda),
    ylab=expression(paste("l(", lambda, ";y)")));
```

3. (Monisteen teht. 2.6.) Olkoon mallina $Y_1, \ldots, Y_n \sim \text{Tas}(\theta, \theta + 1) \perp\!\!\!\perp$. Johda aineistoa $\mathbf{y} = (y_1, \ldots, y_n)$ vastaava uskottavuusfunktio ja totea, että se saa suurimman arvonsa jokaisessa välin $\big(y_{(n)} - 1, y_{(1)}\big)$ pisteessä, kun merkitään $y_{(1)} = \min(y_1, \ldots, y_n)$ ja $y_{(n)} = \max(y_1, \ldots, y_n)$. Siten su-estimaatti $\hat{\theta}(\mathbf{y})$ ei ole yksikäsitteinen (todennäköisyydellä yksi). Mitä mahtaa tapahtua välin $\big(y_{(n)} - 1, y_{(1)}\big)$ pituudelle, kun havaintojen lukumäärä $n \to \infty$?

Ratkaisu:

Riippumattomuuden nojalla tilastollinen malli (ytf) on muotoa

$$
\begin{aligned}
f_{\mathbf{Y}}(\mathbf{y}; \theta) &= \prod_{i=1}^{n} \frac{1}{(\theta + 1) - \theta} 1_{(\theta, \theta+1)}(y_i) \\
&= \prod_{i=1}^{n} 1_{(\theta, \theta+1)}(y_i) = \begin{cases} 1, \text{ kun } y_{(1)} > \theta \text{ ja } y_{(n)} < \theta + 1 \\ 0, \text{ muuten,} \end{cases}
\end{aligned}
$$

missä $1_{(\theta, \theta+1)}(\cdot)$ on välin $(\theta, \theta + 1)$ indikaattorifunktio.

Huomataan, että ehdot $y_{(1)} > \theta$ ja $y_{(n)} < \theta + 1$ ovat yhtäpitäviä sen kanssa, että $y_{(n)} - 1 <$

$\theta < y_{(1)}$. Siten uskottavuusfunktio voidaan esittää muodossa

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

Tästä seuraa, että uskottavuusfunktio saa maksimiarvonsa $L(\theta) = 1$ jokaisella $\theta \in (y_{(n)} - 1, y_{(1)})$.

Havaintojen lukumäärän kasvaessa rajatta satunnaismuuttujat $Y_{(1)}$ ja $Y_{(n)}$ saavat arvoja mielivaltaisen läheltä välin $(\theta, \theta + 1)$ reunoja ja välin $\big(y_{(n)} - 1, y_{(1)}\big)$ pituus lähenee nollaa.

4. (Vrt. monisteen teht. 2.10.) Ladislaus Bortkiewicz julkaisi vuonna 1898 kirjan *The Law of Small Numbers*, joka käsitteli Poissonin jakaumaa. Kuuluisin kirjassa esitetty aineisto oli hevosenpotkuun kuolleiden miesten vuosittaiset lukumäärät neljässätoista Preussin armeijan yksikössä kahdenkymmenen vuoden (1875–1894) ajalta, yhteensä siis 280 havaintoa. Yhteenveto aineistosta on alla.

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

Bortkiewicz havaitsi, että Poissonin jakauma sopii kuvaamaan hevosenpotkuun kuolleiden miesten määrää kussakin yksikössä vuoden aikana. Olkoon μ kyseisen Poissonin jakauman odotusarvo, ja oletetaan, että kuolleiden lukumäärät eri vuosina ja eri yksiköissä ovat toisistaan riippumattomat. Tällöin aineistoa voidaan mallintaa satunnaismuuttujilla $Y_1, \ldots, Y_{280} \sim P(\mu) \perp\!\!\!\perp$.

Muodosta aineistoa vastaava log-uskottavuusfunktio ja määritä μ:n suurimman uskottavuuden estimaatti.

[Tästä aineistosta on netissä paljon tarinaa (hakusanat esim. "horse kick data"). Koko aineisto löytyy esim. osoitteesta http://www.math.uah.edu/stat/data/HorseKicks.html.]

Ratkaisu:

Olkoon Y_i kuten tehtävänannossa on ehdotettu. Merkitään $Y = (Y_1, \ldots, Y_{280})$ ja $\overline{Y} = 1/280 \sum_{i=1}^{280} Y_i$. Tilastollinen malli voidaan nyt kirjoittaa muotoon

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \mu) = \prod_{i=1}^{280} e^{-\mu} \frac{\mu^{y_i}}{y_i!} = e^{-n\mu} \frac{\mu^{n\overline{y}}}{\prod_{i=1}^{280} y_i}.$$

Kertomalla parametreista riippumaton tekijä $\prod_{i=1}^{280} y_1$ pois saadaan uskottavuusfunktio

$$L(\mu; \boldsymbol{y}) = e^{-n\mu} \mu^{n\overline{y}},$$

josta logaritmoimalla saadaan log-uskottavuusfunktio

$$\ell(\mu; \boldsymbol{y}) = -n\mu + n\overline{y} \log \mu.$$

Derivoimalla parametrin μ suhteen saadaan

$$\ell'(\mu; \boldsymbol{y}) = -n + \frac{n\overline{y}}{\mu} = 0 \Leftrightarrow \mu = \overline{y}.$$

Kyseessä on globaali maksimi, sillä kaikilla $\mu > 0$ pätee

$$\ell''(\mu; \boldsymbol{y}) = -\frac{n\overline{y}}{\mu^2} < 0.$$

Suurimman uskottavuuden estimaatti on parametrille μ on siis $\hat{\mu} = \overline{y} = (91 + 32 * 2 + 3 * 11 + 4 * 2)/280 = 192/280 = 0{,}7$.

5. (Monisteen teht. 2.8.) Tarkastellaan riippumatonta otosta gammajakaumasta: $Y_1, \ldots, Y_n \sim G(\alpha, 1/\beta) \perp\!\!\!\perp$, jossa $\alpha, \beta > 0$.

a) Parametrina on (α, β). Totea, että uskottavuusyhtälöt voidaan saattaa muotoon

$$\begin{cases} \log \alpha - \psi(\alpha) = \log \overline{y} - \frac{1}{n}\sum_{i=1}^{n} \log y_i, \\ \qquad\qquad \beta = \overline{y}/\alpha, \end{cases}$$

jossa $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ on ns. *digammafunktio.* Yhtälöitä ei voi ratkaista suljetussa muodossa.

b) Parametrina on vain β, ja α on tunnettu luku. Mikä on β:n su-estimaatti?

Ratkaisu:

a) Gammajakauman parametrit ovat positiivisia reaalilukuja, joten määritellään parametriavaruus $\Omega := (0, \infty) \times (0, \infty)$. Riippumattomuuden nojalla uskottavuusfunktio on

$$\begin{aligned} L(\alpha, \beta; \boldsymbol{y}) = f_{\mathbf{Y}}(\mathbf{y}; \alpha, \beta) &= \prod_{i=1}^{n} \frac{y_i^{\alpha-1} e^{\frac{-y_i}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} \\ &= \prod_{i=1}^{n} \beta^{-\alpha}\Gamma(\alpha)^{-1} e^{\frac{-y_i}{\beta}} y_i^{\alpha-1} \\ &= \beta^{-n\alpha}\Gamma(\alpha)^{-n} e^{\frac{-n\overline{y}}{\beta}} \left(\prod_{i=1}^{n} y_i\right)^{\alpha-1}. \end{aligned}$$

Tästä saadaan logaritminen uskottavuusfunktio

$$\ell(\alpha, \beta; \boldsymbol{y}) = -n\alpha \log \beta - n \log\left(\Gamma(\alpha)\right) - \frac{n\overline{y}}{\beta} + (\alpha - 1)\sum_{i=1}^{n} \log(y_i).$$

Uskottavuusyhtälöt saadaan johdettua ratkaisemalla logaritmisen uskottavuusfunktion gradientin nollakohdat

$$\begin{aligned} \nabla\ell(\alpha, \beta; \boldsymbol{y}) &= \left(\frac{\partial}{\partial\alpha}\ell(\alpha, \beta), \frac{\partial}{\partial\beta}\ell(\alpha, \beta)\right) \\ &= \left(-n\log\beta - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n}\log(y_i), \frac{-n\alpha}{\beta} + \frac{n\overline{y}}{\beta^2}\right) = \overline{0} \end{aligned}$$

Tästä saadaan yhtälöryhmä

$$\begin{cases} -n\log\beta - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n}\log(y_i) = 0 \\ \frac{-n\alpha}{\beta} + \frac{n\overline{y}}{\beta^2} = 0 \end{cases}$$

$\Leftrightarrow$

$$\begin{cases} -\log\beta - \psi(\alpha) = -\frac{1}{n}\sum_{i=1}^{n}\log(y_i) \\ \beta = \frac{\overline{y}}{\alpha} \end{cases}$$

Sijoittamalla toisen yhtälön tulos ensimmäiseen saadaan väite osoitettua, sillä $-\log \beta = -\log \overline{y} + \log \alpha$.

b) Oletetaan, että $\alpha > 0$ on tunnettu luku ja siten uskottavuusfunktio on vain β:n funktio. Nyt parametriavaruus on $\Omega = (0, \infty)$ ja edellisen kohdan perusteella $\ell'(\beta; \boldsymbol{y}) = 0 \Leftrightarrow \beta = \overline{y}/\alpha$. Derivaattafunktion ainoa nollakohta on globaali maksimi, sillä

$$\ell''(\beta; \boldsymbol{y}) = \frac{n\alpha}{\beta^2} - \frac{2n\overline{y}}{\beta^3} = \frac{n\alpha^3 - 2n\alpha^3}{\overline{y}} = \frac{-n\alpha^3}{\overline{y}} < 0.$$

Siis $\hat{\beta} = \overline{y}/\alpha$.