Tilastollinen päättely I, kevät 2017
Harjoitus 2 – palautuspäivämäärä 28. 3.

Opastus: Tehtävät liittyvät pääosin mallin rakentamiseen (monisteen jakso 2.1) ja uskottavuusfunktioon sekä suurimman uskottavuuden menetelmään (jaksot 4.1–4.4).

1. Jatkoa harjoituksen 1 tehtäviin 1, 2 ja 5. Yhden havainnon sijasta tehdäänkin ko. jakaumasta kaksi riippumatonta havaintoa, satunnaismuuttujina siis $Y_1 \perp\!\!\!\perp Y_2$, joilla kummallakin on sama pistetodennäköisyysfunktio $f(y;\theta)$.

Oletetaan, että havainnot ovat $y_1 = 4$ ja $y_2 = 3$. Laske vastaavan uskottavuusfunktion $L(\theta; y_1, y_2)$ arvot ja määritä suurimman uskottavuuden estimaatti $\hat{\theta}$.

2. Jatkoa harjoituksen 1 tehtäviin 3 ja 4, joissa metrojunan odotusaikaa kuvattiin $\mathrm{Tas}(0,\theta)$-jakaumalla.

a) Muodosta tilastollisen mallin lauseke (yhteistiheysfunktio) satunnaiskokeelle, jossa tehtiin n riippumatonta odotusajan mittausta $y_1, \ldots, y_n$.

b) Ilmoita aineistoa $\mathbf{y} = (y_1, \ldots, y_n)$ vastaava uskottavuusfunktio $L(\theta; \mathbf{y})$. Totea selvästi, missä joukossa uskottavuusfunktio on määritelty ja missä se on nollasta poikkeava. (Havainnoista $y_1, \ldots, y_n$ suurin on tässä avainasemassa; sille käytetään usein merkintää $y_{(n)}$.) Piirrä kuva.

c) Mikä on parametrin θ suurimman uskottavuuden estimaatin lauseke edellä? Mitkä ovat sen arvot harjoituksen 1 tehtävän 4 kohdissa a–c?

Huom: Törmäät c-kohdassa ongelmaan, mikäli käytit $\mathrm{Tas}(0,\theta)$-jakauman tiheysfunktiosta versiota, joka on $= 1/\theta$ vain avoimella välillä $]0,\theta[$. Kannattaakin valita tiheysfunktiolle versio, joka on $= 1/\theta$ vastaavalla suljetulla välillä. Se määrittelee saman jakauman mutta toimii suurimman uskottavuuden estimoinnissa paremmin.

3. Kolibakteerin pelätään pilanneen erään joen, josta otetaan juomavettä. Asian tutkimiseksi otamme joesta n:stä satunnaisesti valitusta kohdasta koeputkellisen verran vettä ja kustakin näytteestä määritämme kolibakteerien lukumäärät $y_1, \ldots, y_n$.

Klassinen jakauma, jota käytetään kuvaamaan bakteerien lukumäärän jakautumista tilavuusyksikössä vettä, on Poissonin jakauma.[1] Palautetaan todennäköisyyslaskennasta mieleen, että sen ptnf on muotoa $g(x;\mu) = e^{-\mu}\mu^x/x!$, kun $x = 0, 1, 2, \ldots$ ja $\mu > 0$ on jakauman odotusarvo. Mallinnammekin mittaustulokset niin, että niitä vastaavat satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat satunnaisotos eli n riippumatonta havaintoa Poissonin jakaumasta.

Muodosta tilastollisen mallin lauseke (eli yhteispistetodennäköisyysfunktio) ja uskottavuusfunktio. Johda uskottavuusfunktion logaritmia tutkimalla ja huolellisesti perustellen μ:n suurimman uskottavuuden estimaatti. Mikä on parametrin μ käytännön tulkinta?

4. Mallina on satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$, ts. $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2) \perp\!\!\!\perp$. Luennoilla on näytetty, että odotusarvoparametrin μ suurimman uskottavuuden estimaatti on $\hat{\mu} = \bar{y}$, kun havainnot ovat $y_1, \ldots, y_n$ (ks. muistiinpanojen jakso 4.4).

Tarkastellaan vastaavaa satunnaismuuttujaa eli *estimaattoria* $\hat{\mu} = \overline{Y} = (Y_1 + \cdots + Y_n)/n$ (merkiten sitä siis samalla symbolilla kuin estimaattia).

a) Totea, että $\hat{\mu}$ on *harhaton* eli $E(\hat{\mu}) = \mu$. Mitä tämä tulkinnallisesti merkitsee "toistetun aineistonkeruun" ajatuksen näkökulmasta?

[1] Kiinnostuneet voivat katsoa esim. artikkelia http://www.ncbi.nlm.nih.gov/pmc/articles/PMC243648/ ja sen viitteitä.

b) Laske varianssi $\mathrm{var}(\hat{\mu})$ ja totea, että se lähestyy nollaa, kun havaintojen lukumäärä n kasvaa rajatta. Mitä tämä tulos voisi merkitä?

Tämän tehtävän aihepiireistä puhutaan muistiinpanojen luvussa 3 (ks. erityisesti jaksot 3.4 ja 3.5). Emme kuitenkaan syvenny niihin tällä kurssilla laajemmin.

5. Olkoot $y_1, \ldots, y_n$ reaalilukuja ja $\bar{y} = \sum_{i=1}^n y_i/n$ niiden aritmeettinen keskiarvo. Näytä, että jos μ on reaaliluku, niin

$$\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2.$$

Luennoilla tätä hajotelmaa käytetään normaalijakauman su-estimoinnin yhteydessä (ks. muistiinpanojen sivu 29). Vihje: $y_i - \mu = (y_i - \bar{y}) + (\bar{y} - \mu)$.

6. Olkoot $Y_1, \ldots, Y_n$ riippumattomia satunnaismuuttujia, joille $E(Y_i) = \mu$ ja $\mathrm{var}(Y_i) = \sigma^2$ (esimerkiksi $Y_i \sim N(\mu, \sigma^2)$). Osoita edellisen tehtävän avulla, että jos $n \geq 2$, niin tavanomaiselle otosvarianssille

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \overline{Y})^2$$

pätee $E(S^2) = \sigma^2$. Tämä kertoo, että S^2 on *harhaton* estimaattori σ^2:lle.