**Johdatus tilastolliseen päättelyyn, kesä 2015**
**Harjoitus 2 (28.5.)**

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

**1.** Erään diskreetin jakauman pistetodennäköisyysfunktio $g(x;\theta)$ riippuu parametrista $\theta$, jolla on kaksi mahdollista arvoa: 1 ja 2. Funktion $g(x;\theta)$ arvot käyvät ilmi taulukosta alla.

| $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $g(x;1)$ | 0.0 | 0.3 | 0.4 | 0.2 | 0.1 |
| $g(x;2)$ | 0.1 | 0.4 | 0.3 | 0.1 | 0.1 |

Satunnaismuuttujat $Y_1$, $Y_2$ ja $Y_3$ ovat riippumattomia ja noudattavat kukin kyseistä jakaumaa. Niiden havaitut arvot ovat $y_1 = 3$, $y_2 = 2$ ja $y_3 = 4$. Muodosta tätä aineistoa vastaava uskottavuusfunktio ja määritä suurimman uskottavuuden estimaatti $\hat{\theta}$.

**2.** Jatkoa harjoituksen 1 tehtävään 6.

a) Määritä parametrin $\theta$ suurimman uskottavuuden estimaatti $\hat{\theta}$. Ohje: Huomaat törmääväsi 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.

b) Mitkä ovat $\hat{\theta}$:n arvot harjoituksen 1 tehtävän 4 kohdissa a–c?

**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 on käytetty kuvaamaan bakteerien lukumäärän jakautumista tilavuusyksikössä vettä, on Poissonin jakauma.[1] Palautetaan todennäköisyyslaskennasta mieleen, että sen pistetodennäköisyysfunktio 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. (Mikä on nyt parametrin $\mu$ käytännön tulkinta?)

Muodosta tilastollisen mallin lauseke (eli yhteispistetodennäköisyysfunktio) ja uskottavuusfunktio. Johda uskottavuusfunktion logaritmia tutkimalla $\mu$:n suurimman uskottavuuden estimaatti.

**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 $\mu$ 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.** a) Olkoot $y_1, \ldots, y_n$ reaalilukuja ja $\bar{y} = \sum_{i=1}^n y_i/n$ niiden aritmeettinen keskiarvo (otoskeskiarvo) kuten edellä. Näytä, että jos $\mu$ 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 käytettiin normaalijakauman su-estimoinnin yhteydessä (ks. muistiinpanojen sivu 29). Ehdotus: Kirjoita aluksi $y_i - \mu = (y_i - \bar{y}) + (\bar{y} - \mu)$.

b) 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 a-kohdan avulla, että 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 $\sigma^2$:lle.