Aalto-yliopisto

MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi Gripenberg, Kiiski
Harjoitus 4
17–21.11.2014, viikko 47

P1. Alla on lueteltu joukko muuttujia:

(1) Mansikoiden C-vitamiinipitoisuus: mg/100g.
(2) Alvarin aukiolta löydetyn kasvin laji.
(3) Paine, joka vaaditaan teräksisen säiliön murtumiseen: N/m^2.
(4) Henkilöiden mielipide metropolihallinnosta mitattuna asteikolla ”kannatan”, ”ei mielipidettä”, ”vastustan”.
(5) Espoo Bluesin sijoitus jääkiekkoliigassa.
(6) Opiskelijan koulutusohjelma.
(7) Opiskelijan kurssin MS-A0502 suorittamisvuosi.

Vastaa seuraaviin kysymyksiin

(a) Millä muuttujilla on nominaali- eli laatueroasteikko?
(b) Millä on ordinaali- eli järjestysasteikko?
(c) Millä muutujilla on intervalli- eli välimatka-asteikko?
(d) Millä muuttujilla on ratio- eli suhdeasteikko?

Huom! *Kaikissa tapauksissa ei ole välttämättä olemassa yksi ainoa oikea vastaus!*

Ratkaisu: (a) Muuttujat (2), (4), (6).
(b) Muuttujat (5) ja ehkä (4).
(c) Muuttuja (7).
(d) Muuttujat (1) ja (3)

P2. Erään talon joillakin asukkailla on seuraavat kuukausitulot (e/kk):

3300	3200	1700	4400	4000	4200	1400	4100	1700
4300	1900	2200	2200	2400	2600	3100	2900	3200
800	1700	2100	2300	2500	2700	1900	1800	1500
1200	700	6400	7500	1600	1800	2000	2200	2500.

Muodosta aineistosta luokiteltu frekvenssijakauma, jonka luokat ovat: $0-1000$, $1001-2000$, $2001-3000$, $3001-4000$, $4001-6000$, ja $6001-8000$. Määritä myös tätä frekvenssijakaumaa vastaavan histogrammikuvion suorakaiteiden korkeudet ottaen huomioon, että pinta-alojen pitää suhtautua toisiinsa kuten vastaavat luokkafrekvenssit. Hahmottele myös ko. histogrammikuvio paperille.

Määritä aineistosta myös seuraavat tunnusluvut:

(a) Minimi ja maksimi.
(b) Vaihteluväli (eli pienin väli $[a, b]$ niin, että kaikki arvot ovat tällä välillä) ja vaihteluvälin pituus.
(c) Mediaani.

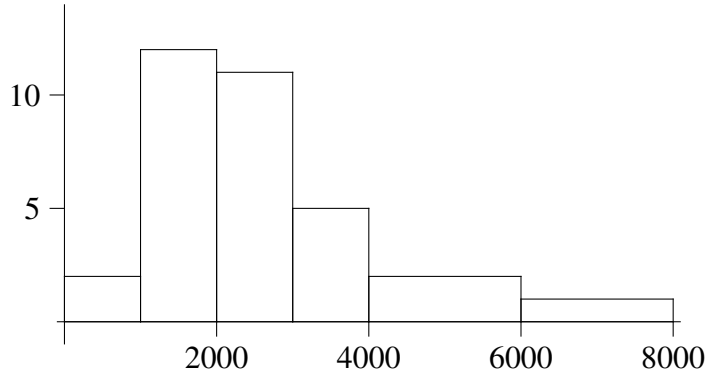
Ratkaisu:

Histogrammi muodostuu suorakaiteista, joiden pinta-alat suhtautuvat toisiinsa kuten vastaavat luokkafrekvenssit (tai suhteelliset luokkafrekvenssit). Luokkafrekvenssit ja vastaavien

suorakaiteiden korkeudet ovat

Luokka	Luokkafrekvenssi	Suorakaiteen korkeus
0-1000	2	2
1001-2000	12	12
2001-3000	11	11
3001-4000	5	5
4001-6000	4	2
6001-8000	2	1

Lopullinen kuvaaja näyttää seuraavanlaiselta (y-akselin yksikkö on asukasta/1000 euroa):



(a) Minimi $= 700$ ja maksimi $= 7500$.

(b) Vaihteluväli on (700, 7500) ja vaihteluvälin pituus on $7500 - 700 = 6800$.

(c) Luokkafrekvenssitaulukosta nähdään, että talossa asuu 14 henkilöä joiden tulot ovat korkeintaan 2000 euroa ja 11 henkilöä joiden tulot ovat vähintään 3001 euroa. Muiden tulot ovat järjestyksessä 2100, 2200, 2200, 2200, 2300, 2400, 2500, 2500, 2600, 2700 ja 2900. Jätetään näistä 3 suurinta lukua pois ($3 = 14 - 11$) ja jäljellä olevien (2100, 2200, 2200, 2200, 2300, 2400, 2500, 2500) mediaani on $(2200 + 2300)/2 = 2250$, joka siis on koko aineiston mediaani, mutta periaatteessa mikä tahansa luku väliltä $[2200, 2300]$ täyttää mediaanille astetetut ehdot.

P3. Sinulla on otos satunnaismuuttujasta, joka oletetaan olevan normaalijakautunut. Otoksen koko on 30 ja sen keskiarvo on 2.5. Miten iso otoksesta laskettu otosvarianssi voi korkeintaan olla jos saat odotusarvolle 98%:n (symmetrisen) luottamusvälin, jonka pituus on korkeintaan 1.8?

Vastaus: 4.01

Ratkaisu: Jos satunnaismuutuja X_j, $j = 1, \ldots, n$ ovat riippumattomia ja $\sim$ $\mathrm{N}(\mu, \sigma^2)$-jakautuneita niin

$$\frac{\overline{X} - \mu}{\sqrt{\frac{1}{n}S^2}} \sim t(n-1),$$

missä $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ on otoksen keskiarvo ja $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ on otosvarianssi

Satunnaismuuttuja joka on $t(29)$-jakautunut saa todennäköisyydellä 0.98 arvon väliltä $[-2.4620, 2.4620]$ koska $F^{-1}_{\mathrm{t}(29)}(0.01) = -2.4620$, $F^{-1}_{\mathrm{t}(29)}(0.99) = 2.4620$ ja $1 - 0.01 - (1 - 0.99) = 0.98$. Tästä seuraa, että odotusarvon (symmetrinen) 98%:n luottamusväli on

$$\left[\overline{x} - 2.4620 \cdot \sqrt{\frac{1}{30}s^2}, \overline{x} + 2.4620\sqrt{\frac{1}{30}s^2}\right]$$

Jotta tämän välin pituus olisi korkeintaan 1.8 pitää olla

$$2 \cdot 2.462 \cdot \sqrt{\frac{1}{30}s^2} \leq 1.8 \quad \Rightarrow \quad s^2 \leq 30\left(\frac{1.8}{2 \cdot 2.462}\right)^2 \approx 4.01.$$

P4. Olkoot $X_i, i = 1, 2, \ldots, n$ riippumattomia satunnaismuuttujia siten, että $\mathrm{E}(X_i) = 0$ ja $\mathrm{Var}(X_i) = \sigma^2$ kun $i = 1, 2, \ldots, n$.
Osoita, että $\mathrm{E}(S^2) = \sigma^2$ kun

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

missä $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ eli otosvarianssi on varianssin **harhaton** estimaattori.

Vihje: Laske $\mathrm{E}\big((n-1)S^2\big)$ *ja huomaa, että koska satunnaismuttujien* X_i *odotusarvot ovat* 0 *niin* $\mathrm{E}(X_i^2) = \mathrm{Var}(X_i)$ *ja* $\mathrm{E}(\overline{X}) = 0$ *joten* $\mathrm{E}(\overline{X}^2) = \mathrm{Var}(\overline{X})$ *ja muista mikä* $\mathrm{Var}(\overline{X})$ *on kun* X_i*:t ovat riippumattomia.*

Huom! *Jos* $\mathrm{E}(X_i) = \mu \neq 0$ *niin voisimme ottaa satunnaismuuttujiksi* $Y_i = X_i - \mu$ *eikä* S^2 *muuttuisi vaikka jokainen* X_i *korvattaisiin* Y_i*:llä joten oletus että odotusarvo on* 0 *ei ole rajoitus, pelkästään yksinkertaistus.*

Ratkaisu:

$$\begin{aligned}
\mathrm{E}\big((n-1)S^2\big) &= \mathrm{E}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \mathrm{E}\left(\sum_{i=1}^{n}\Big(X_i^2 - 2X_i\overline{X} + (\overline{X})^2\Big)\right) \\
&= \mathrm{E}\left(\sum_{i=1}^{n} X_i^2\right) + \mathrm{E}\left((-2)\sum_{i=1}^{n} X_i\overline{X}\right) + \mathrm{E}\left(\sum_{i=1}^{n}\overline{X}^2\right) \\
&= \sum_{i=1}^{n}\mathrm{E}\big(X_i^2\big) + \mathrm{E}\left((-2n)\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\overline{X}\right) + \mathrm{E}\left(n\overline{X}^2\right) \\
&= \left(\sum_{i=1}^{n}\sigma^2\right) + (-2n+n)\mathrm{E}(\overline{X}^2) = n\sigma^2 - n\frac{1}{n}\sigma^2 = (n-1)\sigma^2,
\end{aligned}$$

ja tästä väite seuraa.

P5. Satunnaismuuttujan X tiheysfunktio on

$$f(x,\theta) = \begin{cases} (\theta-1)x^{-\theta}, & x \geq 1, \\ 0, & x < 1, \end{cases}$$

missä parametri $\theta > 1$. Satunnaismuuttujasta on saatu havainnot 2, 5 ja 14.

(a) Estimoi θ momenttimenetelmällä.
(b) Estimoi θ suurimman uskottavuuden menetelmällä.

Vihje: Muista, että $\int_1^\infty x(\theta-1)x^{-\theta}\,\mathrm{d}x = \frac{\theta-1}{\theta-2}$ *kun* $\theta > 2$ *ja että* $\frac{\mathrm{d}}{\mathrm{d}\theta}a^\theta = a^\theta \ln(a)$. *Kun lasket suurimman uskottavuuden estimaattia, voit menetellä kuten tämän viikon stack-tehtävässä 7, sillä erolla että tässä voit ratkaista ääriarvokohdan analyyttisesti.*

Vastaus: ≈ 2.17 ja ≈ 1.61

Ratkaisu: Koska $f(x)$ on tiheysfunktio, sen on toteutettava ehdot $f(x) \geq 0$ ja $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$ ja jos $\theta \leq 1$ kumpikaan näistä ei toteudu.
(a) Estimointi momenttimenetelmällä:

$$\mathrm{E}(X) = \int_1^\infty x(\theta-1)x^{-\theta}\,\mathrm{d}x = \frac{\theta-1}{\theta-2},$$

olettaen, että $\theta > 2$ koska jos $\theta \leq 2$ jakaumalla ei ole (äärellistä) momenttia.

Momenttiestimaattori saadaan yhtälöstä $\mathrm{E}(X) = \overline{x}$ eli

$$\frac{\theta-1}{\theta-2} = \frac{1}{3}(2+5+14) = 7,$$

josta seuraa, että

$$\hat{\theta}_{MM} = \frac{13}{6} \approx 2.17.$$

(b) Estimointi suurimman uskottavuuden menetelmällä: Riippumattoman otoksen $X_i, i = 1,2,\ldots,n$ uskottavuusfunktio on

$$L(\theta; x_1,\ldots,x_n) = f(x_1;\theta)\cdot f(x_2;\theta)\cdot\ldots\cdot f(x_n;\theta) = (\theta-1)^n(x_1x_2\ldots x_n)^{-\theta}.$$

Uskottavuusfunktion maksimi saadaan derivaatan nollakohtana:

$$\begin{aligned} 0 = L'(\theta; x_1,\ldots,x_n) &= n(\theta-1)^{n-1}(x_1x_2\ldots x_n)^{-\theta} - (\theta-1)^n(x_1x_2\ldots x_n)^{-\theta}\ln(x_1x_2\ldots x_n) \\ &= L(\theta; x_1,\ldots,x_n)\left(\frac{n}{\theta-1} - \ln(x_1x_2\ldots x_n)\right), \end{aligned}$$

josta saadaan

$$\hat{\theta}_{SU} = 1 + \frac{n}{\ln(x_1x_2\ldots x_n)}.$$

(Koska funktio $\frac{n}{\theta-1}$ on vähenevä, niin nähdään että kyseessä todella on maksimi.) Tässä tapauksessa $\hat{\theta}_{SU} = 1 + \frac{3}{\ln(2\cdot5\cdot14)} = 1 + \frac{3}{\ln(140)} \approx 1.61$