TODENNÄKÖISYYSLASKENNAN JA TILASTOTIETEEN PERUSKURSSI (MS-A0501/2/3/4). 26.2.–16.4.2024. Aalto-yliopisto. Pekka Pere, Natalia Vesselinova ja Jukka Kohonen.

Kokeen 16.4.2024 kysymysten ratkaisuehdotukset

1.

a)

- $\mathsf{P}(A \text{ tai } B \text{ tai molemmat }) = \mathsf{P}(A \text{ tai } B) - \mathsf{P}(A \text{ ja } B) = \mathsf{P}(A \cup B) - \mathsf{P}(A \cap B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B) = 0.5 + 0.3 - 0.2 = 0.6.$
- $\mathsf{P}(\text{vain } A) = \mathsf{P}(A \text{ muttei } B) = \mathsf{P}(A \cap B^C) = 0.5 - 0.2 = 0.3.$
- $\mathsf{P}(\text{ainakin } A) = \mathsf{P}(A) = 0.5.$
- $\mathsf{P}(\text{korkeintaan } A) = \mathsf{P}(B^C) = 1 - \mathsf{P}(B) = 1 - 0.3 = 0.7.$

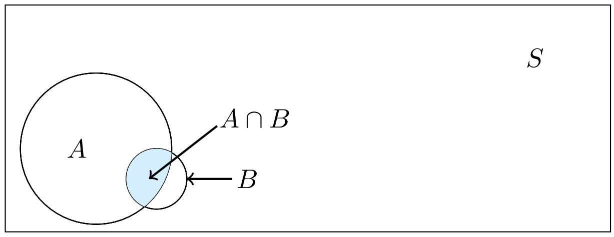
b) Ensimmäinen vaihtoehto “Aku on verovirkailija” on todennäköisempi vaihtoehto, koska rajaaminen Piraattipuolueessa toimiviin verovirkailijoihin voi vain pienentää todennäköisyyttä. Tämä on riittävä vastaus.

Kuvan 1 Venn-diagrammi havainnollistaa (A = “verovirkailijat”, B = “Piraattipuolueen jäsenet” ja S = “suomalaiset”). Periaatteessa olisi mahdollista myös, että vaihtoehdot olisivat yhtätodennäköisiä (kuva 4 havainnollistaa erikoistapaus 3:a).

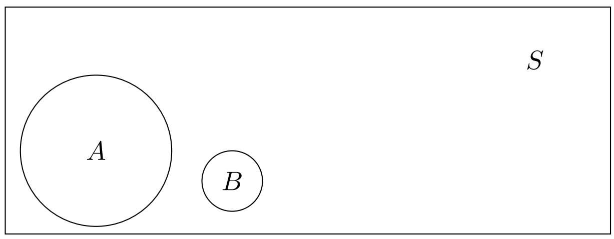
Kolme erikoistapausta:

1. Kukaan Piraattipuolueen jäsen ei ole verovirkailija ($A \cap B = \emptyset$; kuva 2). Akun todennäköisyys olla sekä verovirkailija että Piraattipuoleen jäsen on tällöin erityisen pieni eli nolla.
2. Kaikki piraattipuoluelaiset ovat verovirkailijoita ($B \subset A$; kuva 3). Edelleen Aku on todennäköisemmin verovirkailija kuin sekä verovirkailija että Piraattipuolueen jäsen.
3. Kaikki verovirkailijat kuuluvat Piraattipuolueeseen ($A \subset B$; kuva 3 vaihtaen joukkojen A ja B merkinnät ja merkitys) tai verovirkailijat ja piraattipuoluelaiset ovat täsmälleen samat ihmiset ($A = B$, kuva 4). Tällöin tapahtumien “Aku on verovirkailija” ja “Aku on verovirkailija ja Piraattipuolueen jäsen” todennäköisyydet ovat samat.

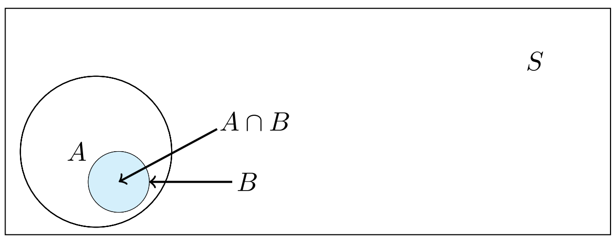
Tehtävä on psykologi Daniel Kahnemanin kirjasta (2011, luku 15) *Thinking, Fast and Slow.* Penguin Books. Lontoo. Kahnemanin raportoimissa kokeissa yliopisto-opiskelijoista vähintään 85 prosenttia arvioi vaihtoehdon (Kahnemanin tehtävään liittyvän vastaavan vaihtoehdon) “Aku on verovirkailija ja Piraattipuolueen jäsen” todennäköisemmäksi, vaikka se ei loogisesti voi olla mahdollista. Kahneman ja kolleganså Amos Tversky sanovat ihmisten tekevän konjunktiovirhepäätelmän (*conjunction fallacy*), kun ihmiset päättelevät kahden tapahtuman leikkauksen (molemmille tapahtumille yhteisten alkeistapahtumien) olevan todennäköisempi kuin toisen tapahtumista. Kahneman palkittiin taloustieteen Nobelilla vuonna 2002 tutkimuksistaan ihmisten päätöksenteosta epävarmuuden vallitessa.



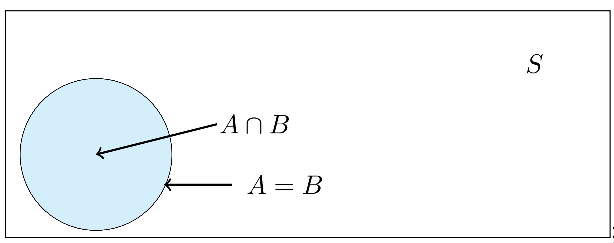
Kuva 1: Piraattipuolueen jäseniä on verovirkailijoissa, mutta kaikki Piraattipuolueen jäsenet eivät ole verovirkailijoita.



Kuva 2: Piraattipuolueen jäsenet ja verovirkailijat ovat erillisiä joukkoja.



Kuva 3: Piraattipuolueen jäsenet ovat (aito) osajoukko verovirkailijoita.



;

Kuva 4: Piraattipuolueen jäsenet ja verovirkailijat ovat sama joukko.

2.

a) Symmetrisen jakauman tilanteessa keskipaino on 0.025. ja 0.975. kvantiilien keskiarvo: $(3700 + 6900)/2 = 5300$. Jakauma ei ole symmetrinen, sillä ko. kvantiilien keskiarvo ei ole keskimääräinen paino 5200. Tiedoista ei voi päätellä, että jakauma on symmetrinen.

Tutkimuksen kuvion 2 osion B histogrammista ilmenee, että ekologisen painon jakauma on hieman vino oikealle.

b) Keskipaino on 0.025. ja 0.975. kvantiilien keskiarvo: $(5100 + 9100)/2 = 7100$. Kyseiset kvantiilit ovat yhtä kaukana odotusarvosta, joten jakauma saattaa olla symmetrinen. Se ei ole välttämättä symmetrinen. Sekä jakauman 0.025. ja 0.975. kvantiileja ulommat että sisemmät kvantiilit voivat sijoittua epäsymmetrisesti odotusarvoon nähden. Tiedoista ei voi päätellä, että painon jakauma on symmetrinen.

c) Merkitään täysikasvuisen *Tyrannosaurus rexin* painoa X:llä. Tehtävän mukaan täysikasvuisen *Tyrannosaurus rexin* paino on normaalijakautunut ja

$$\mathsf{P}(X \leq 5100) = \mathsf{P}(X \geq 9100) = 0.025.$$

Standardinormaalijakauman 0.025. ja 0.975. kvantiili ovat -1.959964 ja 1.959964 (R-komennoilla `qnorm(0.025)` ja `qnorm(0.975)`). Näin ollen pitää päteä, että

$$\mathsf{P}\left(\frac{5100-7100}{\sigma} \leq -1.959964\right) = \mathsf{P}\left(\frac{9100-7100}{\sigma} \geq 1.959964\right) = 0.025.$$

Keskihajonta σ voidaan ratkaista esimerkiksi näin:

$$\begin{aligned} \frac{5100-7100}{\sigma} &= -1.959964 && \Leftrightarrow \\ -2000 &= -1.959964\sigma && \Leftrightarrow \\ \sigma &= \frac{-2000}{-1.959964} = 1020.427. \end{aligned}$$

Täysikasvuisen *tyrannosaurus rexin* painon keskihajonta ja varianssi ovat 1020 kg ja $1020.427^2 = 104\,1271$ kg^2.

3. Waldin 95 %:n luottamusvälin ala- ja yläraja ovat

$$\frac{y}{n} \pm 1.960\sqrt{\frac{\frac{y}{n}\left(1-\frac{y}{n}\right)}{n}}$$

Yllä 1.960 on standardinormaalijakauman 0.975. kvantiili (`qnorm(0.975)`). Tehtävän tilanteessa rajat ovat

$$\frac{1311}{1700} - 1.960\sqrt{\frac{\frac{1311}{1700}\left(1-\frac{1311}{1700}\right)}{1700}} = 0.7512074.$$

ja

$$\frac{1311}{1700} + 1.960\sqrt{\frac{\frac{1311}{1700}\left(1-\frac{1311}{1700}\right)}{1700}} = 0.7911456.$$

Rajojen lasku R:llä:

```
n <- 1700
p <- 1311/n
q <- 1-p
p-1.960*sqrt(p*q/n)
p+1.960*sqrt(p*q/n)

# Vaihtoehtoisesti asenna mosaic-paketti ja komenna binom.test():
install.packages("mosaic")
library(mosaic)
binom.test(x=1311,n=1700,ci.method="Wald")
 - -
## number of successes = 1311, number of trials = 1700, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
## 0.7512077 0.7911452
 - -
```

Luottamusväli on $[0.751, 0.791]$. Se ei kata todellista äänestysosuutta 0.705.

Yksi selitys on, että haastatellut ovat arastelleet myöntää haastattelijalle etteivät äänestäneet. He ovat kertoneet haastattelijalle äänestäneensä vaikkeivät äänestäneet.

4.

a) Tehtävän tilanteessa uskottavuusfunktio on

$$\begin{aligned}
\mathsf{L}(\lambda) &= \prod_{i=1}^{3} p_X(x_i) \\
&= \prod_{i=1}^{3} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\
&= e^{-3\lambda} \frac{\lambda^{x_1+x_2+x_3}}{x_1!x_2!x_3!} \\
&= e^{-3\lambda} \frac{\lambda^{22}}{10!5!7!}.
\end{aligned}$$

Log-uskottavuusfunktio on

$$\begin{aligned}
l(\lambda) &= \ln(\mathsf{L}(\lambda)) \\
&= \ln\left(e^{-3\lambda} \frac{\lambda^{22}}{10!5!7!}\right) \\
&= \ln e^{-3\lambda} + \ln\left(\frac{\lambda^{22}}{10!5!7!}\right) \\
&= -3\lambda + 22 \ln \lambda - \ln(10!5!7!)
\end{aligned}$$

Derivoidaan log-uskottavuusfunktio sen maksimin löytämiseksi.

$$\begin{aligned}
l'(\lambda) &= \frac{\partial l(\lambda)}{\partial \lambda} \\
&= -3 + 22\frac{1}{\lambda}
\end{aligned}$$

Asetetaan derivaatta nollaksi:

$$l'(\lambda) = \frac{\partial l(\lambda)}{\partial \lambda} = 0$$

$$-3 + 22\frac{1}{\lambda} = 0$$

$$\Rightarrow \lambda = \frac{22}{3}.$$

Tarkistetaan, että derivaatan nollakohdassa toinen derivaatta on negatiivinen eli että log-uskottavuusfunktiolla on paikallinen maksimi kyseisessä pisteessä:

$$\begin{aligned}
l''(\lambda) &= \frac{\partial^2 l(\lambda)}{\partial \lambda^2} \\
&= -22\frac{1}{\lambda^2} < 0.
\end{aligned}$$

Tarkistetaan, että maksimi ei ole log-uskottavuusfunktion määrittelyvälin reunoilla: $\lambda \to 0^+$ ja $\lambda \to \infty$. Molemmilla reunoilla $\ln(\mathsf{L}(\lambda)) \to -\infty$. Näin ollen log-uskottavuusfunktio maksimoituu derivaatan nollakohdassa. Vastauksesta saa täydet pisteet, vaikka tätä tarkistusta ei olisi tehnyt.

Suurimman uskottavuuden (SU) estimaattori on $\hat{\lambda} = 22/3 \approx 7.33$. Huomaa, että SU-estimaattori on havaintojen keskiarvo.

b) Normeeraamaton posteriorijakauma (jälkijakauma) on priorijakauman (esijakauman) ja uskottavuusfunktion tulo:

$$p(\lambda) \cdot p(x|\lambda) = p(\lambda) \cdot \mathsf{L}(\lambda).$$

Posteriorijakauma on siis

$$\begin{aligned} &\propto e^{\lambda} \cdot e^{-3\lambda} \lambda^{22} \\ &\propto e^{-4\lambda} \lambda^{22}. \end{aligned}$$

(Gamma-jakauma vailla normeeraavia tekijöitä.)

Kohdan a) tapaan saadaan MAP-estimaatiksi $\hat{\lambda} = 22/4 = 11/2 = 5.5$.

c) Uusi priorijakauma on kohdassa b) johdettu posteriorijakauma. Uuden havainnon huomioiva uskottavuusfunktio on

$$\begin{aligned} \mathsf{L}(\lambda) &= p_X(x_4) \\ &= e^{-\lambda} \frac{\lambda^{20}}{20!}. \end{aligned}$$

Päivitetty posteriorijakauma on

$$e^{-4\lambda} \lambda^{22} \cdot e^{-\lambda} \lambda^{20} = e^{-5\lambda} \lambda^{44},$$

ja uusi MAP-estimaatti on $\hat{\lambda} = 44/5 = 8.8$.