TODENNÄKÖISYYSLASKENNAN JA TILASTOTIETEEN PERUSKURSSI. 26.2.–16.4.2024. Aalto-yliopisto. Yliopistonlehtori Pekka Pere, Tutkijatohtori Natalia Vesselinova.

Uusintakokeen 7.6.2024 kysymysten ratkaisuehdotukset

1.

a) Merkitään A:lla, B:llä ja D:llä tapahtumia "potilas paranee psykoanalyysilla", "potilas paranee lääkityksellä" ja "potilas paranee psykoanalyysilla tai lääkityksellä tai molemmilla". Psykoanalyysi ja lääkitys toimivat tai eivät toisistaan riippumatta. Potilas paranee yhtäaikaisella psykoanalyysi- ja lääkityshoidolla todennäköisyydellä 0.72:

$$\begin{aligned}
\mathsf{P}(D) &= \mathsf{P}(A \cup B) \\
&= \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B) \\
&\overset{\text{riippum.}}{=} \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A)\mathsf{P}(B) \\
&= 0.6 + 0.3 - 0.6 \times 0.3 \\
&= 0.72.
\end{aligned}$$

Tehtävä voidaan ratkaista myös vastatapahtumien avulla. Merkitään $A^C =$ "potilas ei parane psykoanalyysilla", $B^C =$ potilas ei parane lääkityksellä ja $D^C =$ "potilas ei parane psykoanalyysilla eikä lääkityksellä". Näillä merkinnöillä $D^C = A^C \cap B^C$ (piirrä Venn-diagrammi!). Jos tapahtumat A ja B ovat riippumattomia, niin niiden vastatapahtumat A^C ja B^C ovat riippumattomia (tämä on oma harjoitustehtävänsä). Riippumattomuuden perusteella

$$\begin{aligned}
\mathsf{P}(D) &= 1 - \mathsf{P}(D^C) \\
&= 1 - \mathsf{P}(A^C \cap B^C) \\
&\overset{\text{riippum.}}{=} 1 - \mathsf{P}(A^C)\mathsf{P}(B^C) \\
&= 1 - (1 - 0.6) \times (1 - 0.3) \\
&= 0.72.
\end{aligned}$$

b) Merkitään C:llä tapahtumaa "potilas paranee yhdistetyllä psykoanalyysilla ja lääkityksellä". Kokonaistodennäköisyyden kaavan mukaan

$$\begin{aligned}
\mathsf{P}(D) &= \mathsf{P}(D \mid A)\mathsf{P}(A) + \mathsf{P}(D \mid B)\mathsf{P}(B) + \mathsf{P}(D \mid C)\mathsf{P}(C) \\
&= 0.6 \times 0.05 + 0.3 \times 0.9 + 0.8 \times 0.05 \\
&= 0.34
\end{aligned}$$

Potilas paranee todennäköisyydellä 0.34.

c) Merkitään $E =$ "potilas paranee ilman hoitoa" ja $F = \mathsf{P}(D \cup E)$ eli että "potilas paranee hoidolla tai ilman hoitoa". Potilas paranee nyt todennäköisyydellä 0.776:

$$\begin{aligned}
\mathsf{P}(F) &= \mathsf{P}(D \cup E) \\
&= \mathsf{P}(D) + \mathsf{P}(E) - \mathsf{P}(D \cap E) \\
&\overset{\text{riippum.}}{=} \mathsf{P}(D) + \mathsf{P}(E) - \mathsf{P}(D)\mathsf{P}(E) \\
&= 0.72 + 0.2 - 0.72 \times 0.2 \\
&= 0.776.
\end{aligned}$$

Sama todennäköisyys saadaan myös yleisestä yhteenlaskusäännöstä:

$$\begin{aligned}
\mathsf{P}(F) &= \mathsf{P}(A \cup B \cup E) \\
&= \mathsf{P}(A) + \mathsf{P}(B) + \mathsf{P}(E) - \mathsf{P}(A \cap B) - \mathsf{P}(A \cap E) - \mathsf{P}(B \cap E) \\
&\quad + \mathsf{P}(A \cap B \cap E) \\
&\overset{\text{riippum.}}{=} \mathsf{P}(A) + \mathsf{P}(B) + \mathsf{P}(E) - \mathsf{P}(A)\mathsf{P}(B) - \mathsf{P}(A)\mathsf{P}(E) - \mathsf{P}(B)\mathsf{P}(E) \\
&\quad + \mathsf{P}(A)\mathsf{P}(B)\mathsf{P}(E) \\
&= 0.6 + 0.3 + 0.2 - 0.6 \times 0.3 - 0.6 \times 0.2 - 0.3 \times 0.2 \\
&\quad + 0.6 \times 0.3 \times 0.2 \\
&= 0.776.
\end{aligned}$$

2.

a) Yhden arvan voiton X_1 odotusarvo on

$$\mu_1 = \mathsf{E}(X_1) = \frac{1}{3200} \times 900 + \frac{400}{3200} \times 8.31 + \frac{2799}{3200} \times 0 = 1.32$$

taaleria.

b) Yhden arvan voiton X_1 varianssi on

$$\begin{aligned}
\mathsf{V}(X_1) = \mathsf{E}[(X_1 - \mu_1)^2] &= \frac{1}{3200} \times (900 - 1.32)^2 + \frac{400}{3200} \times (8.31 - 1.32)^2 \\
&\quad + \frac{2799}{3200} \times (0 - 1.32)^2 \\
&= 260.0146 \approx 260.0
\end{aligned}$$

taaleria toiseen. Voiton keskihajonta on $\mathsf{SD}(X_1) = (260.0146)^{1/2} = 16.12497 \approx 16.1$ taaleria.

c) Yhden arvan tuoton X_2 odotusarvo on

$$\mu_2 = \mathsf{E}(X_2) = \mathsf{E}(X_1 - 1.5) = \mu_1 - 1.5 = 1.32 - 1.5 = -0.18$$

taaleria. Työläämmin tuoton odotusarvon voi laskea näin:

$$\mu_2 = \mathsf{E}(X_2) = \frac{1}{3200} \times (900 - 1.5) + \frac{400}{3200} \times (8.31 - 1.5) + \frac{2799}{3200} \times (0 - 1.5) = -0.18.$$

d) Vakion vähentäminen satunnaismuuttujasta ei muuta sen varianssia. Tuoton varianssi ja keskihajonta ovat siten 260.0 taaleria toiseen ja 16.1 taaleria. Tämä on riittävä vastaus. Laskemalla saadaan samat tulokset:

$$\begin{aligned}
\mathsf{V}(X_2) &= \mathsf{E}[(X_2 - \mu_2)^2] \\
&= \frac{1}{3200} \times [900 - 1.5 - (-0.18)]^2 + \frac{400}{3200} \times [8.31 - 1.5 - (-0.18)]^2 \\
&\quad + \frac{2799}{3200} \times [0 - 1.5 - (-0.18)]^2 \\
&= \frac{1}{3200} \times (900 - 1.32)^2 + \frac{400}{3200} \times (8.31 - 1.32)^2 \\
&\quad + \frac{2799}{3200} \times (0 - 1.32)^2
\end{aligned}$$

$$= 260.0146 \approx 260.0,$$

ja $\mathsf{SD}(X_2) = (260.0146)^{1/2} = 16.12497 \approx 16.1$.

e) Kaikki arvat myytiin. Myyntitulot ovat $3\,200 \times 1.5 = 4\,800$ taaleria. Sairaala saa myyntituloista $0.12 \times 4\,800 = 576$ taaleria. Myynnissä ei ole satunnaisvaihtelua, joten sairaalan saaman avustuksen X_3 odotusarvo on samainen 576 taaleria. Lasku:

$$\mu_3 = \mathsf{E}(X_3) = 576 \times 1 = 576.$$

f) Koska avustuksessa ei ole satunnaisvaihtelua, sairaalan saaman avustuksen varianssi ja keskihajonta ovat 0 taaleria toiseen ja 0 taaleria. Laskut:

$$\begin{aligned} \mathsf{E}[(X_3-\mu_3)^2] &= (576-576)^2 \times 1 = 0 \qquad \text{ja} \\ \mathsf{SD}(X_3) &= \sqrt{0} = 0. \end{aligned}$$

3.

a) Kunkin haastattelun tulos X_i on Bernoulli-koe (1 = äänesti; 0 = ei äänestänyt). Äänestysosuus $\hat{\pi} = \sum_{i=1}^{n} x_i$ on keskiarvo riippumattomista Bernoulli-jakautuneista satunnaismuuttujista. Bernoulli-satunnaismuuttujan odotusarvo ja varianssi ovat π ja $\pi(1-\pi)$.

Keskeisen raja-arvolauseen mukaan standardoitu satunnaismuuttuja

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

noudattaa suurilla havaintomäärillä likimain standardinormaalijakaumaa, jos X_i:t $(i = 1, \ldots, n)$ ovat riippumattomia satunnaismuuttujia, joilla on odotusarvo $\mathsf{E}(X_i) = \mu$ ja varianssi $\mathsf{V}(X_i) = \sigma^2 > 0$ ja $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$.

Tehtävän tilanteessa $\mu = \pi$ ja $\sigma^2 = \pi(1-\pi)$. Haastateltuja on suuri määrä, joten keskeisen raja-arvolauseen perusteella

$$\frac{\hat{\pi}-\pi}{\pi(1-\pi)/\sqrt{n}}$$

on likimain standardinormaalijakautunut.

b) Testataan, onko osuus nollahypoteesin mukainen. Nolla- ja vastahypoteesi ovat H_0: $\pi = 0.705$ ja H_1: $\pi \neq 0.705$. Testisuure on

$$Z = \frac{\hat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

Tehtävän tilanteessa

$$Z = \frac{0.7711765 - 0.705}{\sqrt{0.705(1-0.705)/1700)}} \approx 5.983.$$

R:llä numeerinen arvo saadaan näin:

```
(0.7711765-0.705)/(sqrt(0.705*(1-0.705)/1700))
## [1] 5.983051
```

c) Kriittiset arvot ovat standardinormaalijakauman 0.005. ja 0.995. kvantiilit -2.576 ja 2.576 (R-komento `qnorm(0.995)`). Testisuureen arvo 5.983 on suurempi kuin kriittinen arvo 2.576, joten nollahypoteesi hylätään. Päätellään, että vastaajien ilmoittaman mukainen äänestysosuus poikkeaa tilastollisesti merkitsevästi 70.5 %:sta.

Ei päätellä, että äänioikeutettujen äänestysosuus ei olisi 70.5 %, sillä se tiedetään todeksi äänestysosuudeksi. Jos sattuman eli hyvin poikkeuksellisen kyselytuloksen mahdollisuus sivuutetaan, aineiston poikkeuksellisuuden täytyy johtua muusta seikasta.

Yksi selitys voi olla, että ESS-kyselyn otos on valikoitunut. Ehkä siihen kieltäytyvät osallistumasta kansalaiset, jotka ovat muutenkin tai ainakin äänestysaktiivisuudeltaan muita passiivisempia. Kyselyyn vastaisivat tällöin erityisesti äänestävät kansalaiset. Yhteiskunnasta syrjäytyneet eivät ehkä äänestä vaaleissa eivätkä suostu haasteltavaksi tai kysely ei tavoita heitä. ESS-kyselyissä kerrottu äänestysaktiivisuus muodostuisi näin suuremmaksi kuin mitä se todellisuudessa on.

Toinen mahdollinen selitys on, että ESS on tehty henkilökohtaisena haastatteluna. Äänestämiskysymys on saatettu esittää sanallisena kysymyksenä, johon haastateltava on vastannut. Ihmiset saattavat nolostella äänestämättömyyttään ja väittää haastattelijalle paikkansapitämättömästi, että ovat äänestäneet. Äänestäminen on mahdollisesti arkaluonteinen aihe, jota kannattaisi selvittää hienovaraisemmin. Olavi Borg (2010, 80) kertoo verranneensa eduskuntavaaleja 1962 koskevan kyselytutkimuksensa tietoja vaalitilastoihin. Moni joka oli kyselyssä ilmoittanut äänestäneensä ei ollut silti äänestänyt.

Päätelmä on, että otannassa on vikaa. Nollahypoteesin mukaisessa osuuden suuruudessa ei ole, sillä se tiedetään oikeaksi.

4.

a) Normeeramaton posteriorijakauma (jälkijakauma) on priorijakauman (esijakauman) ja uskottavuusfunktion tulo

$$\text{posteriorijakauma} \propto \text{priorijakauma} \cdot \text{uskottavuus}.$$

Merkitään estimoitavaa parametria λ:lla, priorijakaumaa $f(\lambda)$:lla ja uskottavuusfunktiota $\mathsf{L}(\lambda)$:lla. Näillä merkinnöillä

$$\text{priori} \cdot \text{uskottavuus} = f(\lambda) \cdot \mathsf{L}(\lambda).$$

Havaintojen riippumattomuuden ja niiden Poisson-jakautuneisuuden perusteella uskottavuusfunktio on

$$\begin{aligned}
\mathsf{L}(\lambda) &= \prod_{i=1}^{5} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\
&= e^{-5\lambda} \frac{\lambda^{x_1+x_2+\cdots+x_5}}{x_1!x_2!x_3!x_4!x_5!} \\
&= e^{-5\lambda} \frac{\lambda^{500}}{100!110!90!50!150!} \\
&\propto e^{-5\lambda}\lambda^{500}.
\end{aligned}$$

Kolmannella rivillä on sijoitettu havaitut potilasmäärät, ja viimeisellä rivillä on huomioitu vain uskottavuusfunktion kannalta relevantit termit.

Normeeramaton posteriorijakauma on

$$\begin{aligned}
f(\lambda) \cdot \mathsf{L}(\lambda) &= e^{-\lambda} e^{-5\lambda} \lambda^{500} \\
&= e^{-6\lambda} \lambda^{500}.
\end{aligned}$$

(Huomautus: Tämä on normeerausta vaille Gamma-jakauma.)

b) Uusi priorijakauma on kohdassa a) johdettu posteriorijakauma. Uusiin havaintoihin liittyvä uskottavuusfunktio on

$$\begin{aligned} \mathsf{L}(\lambda) &= e^{-\lambda}\frac{\lambda^{250}}{250!} \cdot e^{-\lambda}\frac{\lambda^{200}}{200!} \\ &= e^{-2\lambda}\frac{\lambda^{450}}{250!200!} \\ &\propto e^{-2\lambda}\lambda^{450}. \end{aligned}$$

Päivitetty normeeramaton posteriorijakauma on

$$e^{-6\lambda}\lambda^{500} \cdot e^{-2\lambda}\lambda^{450} = e^{-8\lambda}\lambda^{950}.$$