MS-A0504

Todennäköisyyslaskennan ja tilastotieteen peruskurssi

6B Osituskaavan ehdollistaminen ja Simpsonin paradoksi

Pekka Pere
matematiikan ja systeemianalyysin laitos
perustieteiden korkeakoulu
Aalto-yliopisto

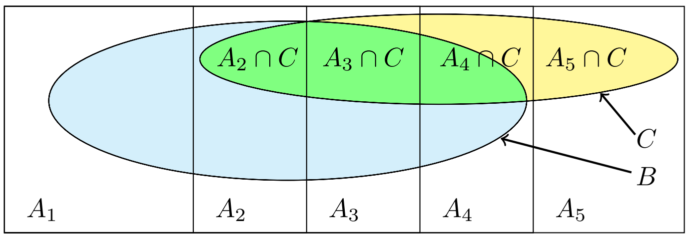
lukuvuosi 2023–2024
periodi I

1 Osituskaavan eli kokonaistodennäköisyyden kaavan ehdollistaminen

Otosavaruus S on ositettu erillisiin tapahtumiin A_i, $i = 1, \ldots, n$. Tapahtuman B todennäköisyys ehdolla tapahtuma C on

$$\mathsf{P}(B \mid C) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \mid C). \tag{1}$$

Todennäköisyys lasketaan nyt tapahtuman C rajaamassa otosavaruudessa (vrt. osituskaava eli kokonaistodennäköisyyden kaava). Oletus on, että $\mathsf{P}(C) > 0$ ja $\mathsf{P}(A_i \cap C) > 0$, $i = 1, \ldots, n$ (muuten kaikkia kaavan ehdollisia todennäköisyyksiä ei ole määritelty). Perustelu: $\mathsf{P}(B \mid C) = \mathsf{P}(B \cap C)/\mathsf{P}(C) = \sum_{i=1}^{n} \mathsf{P}(B \cap A_i \cap C)/\mathsf{P}(C) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \cap C)/\mathsf{P}(C) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \mid C)\mathsf{P}(C)/\mathsf{P}(C) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \mid C)$. Kuvan 1 Venn-diagrammin tilanteessa sininen ellipsi on tapahtuma B, keltainen ellipsi on tapahtuma C. Todennäköisyys $\mathsf{P}(B \mid C)$ on ellipsien vihreän leikkauksen alan suhde keltaisen ellipsin alaan. Kuvassa $\mathsf{P}(A_1 \cap C) = 0$, joten kaavassa (1) tulee asettaa $i = 2, \ldots, 5$. Lisäksi kuvan tilanteessa pätee $\mathsf{P}(B \mid A_2 \cap C) = 1$ ja $\mathsf{P}(B \mid A_5 \cap C) = 0$.



Kuva 1: Ehdollisen kokonaistodennäköisyyden lasku.

2 Simpsonin paradoksi

Kuvan 2 piirtäjä on sivistynyt ja tuntee *Simpsonin paradoksin*. Sen kuvasivat Udny Yule 1903 ja Edward Simpson 1951. Simpsonin esimerkeissä osa-aineistoissa on eroja mutta yhdistetyssä aineistossa ei. Ensimmäinen empiirinen esimerkki paradoksista on julkaistu 1934. Paradoksi on jo ennen nimeämistään tuottanut 1800-luvulla jupinan “valhe, emävalhe, tilasto” (Wainer ja Brown 2004). Paradoksin juju selviää esimerkistä.[2]

Esimerkki.[3] Kaksi maisteria. B ja B^C ovat juuri valmistuneet maistereiksi. He ovat molemmat opiskelleet aineita C ja C^C. B:n pääaine oli C; B^C:n C^C. Merkitään hyvää arvosanaa A:lla ja huonoa A^C:lla. B:n ja B^C:n suorittamat kurssit arvosanojen mukaan ryhmiteltynä ovat alla:

[1] Kuva on artikkelista A. Edwards (2007): A Cautionary Tale. *Significance*, maaliskuu 2007, 47–48. Kiitän *Significance*-lehteä (`www.significancemagazine.com`) luvasta painaa kuva.

[2] Jakso perustuu Blitzsteinin ja Hwangin (2015) kirjaan ja artikkeleihin Baker ja Kramer (2001), Norton ja Divine (2015) sekä Wainer ja Brown (2004).

[3] Blitzstein ja Hwang (2015, 67–69).

Kuva 2: Simpsonin paradoksi.[1]

	C			C^C			yhteensä		
	A	A^C	Σ	A	A^C	Σ	A	A^C	Σ
B	70	20	90	10	0	10	80	20	100
B^C	2	8	10	81	9	90	83	17	100

B on opiskellut enimmäkseen C:tä ja B^C enimmäkseen C^C:ia. Vinkeää todistuksissa on, että

- C:ssä B:n hyvien arvosanojen osuus $70/90 = 7/9$ on suurempi kuin B^C:n $2/10 = 1/5$.
- C^C:ssa B:n hyvien arvosanojen osuus $10/10 = 1$ on suurempi kuin B^C:n $81/90 = 9/10$.
- kokonaisuutena katsoen hyvien arvosanojen osuus $80/100$ B:llä on silti pienempi kuin $83/100$ B^C:lla! $\square$

Jos osajoukoissa tapahtumien todennäköisyyksien suuruusjärjestys on päinvastainen kuin koko joukossa, on kohdattu Simpsonin paradoksi. Kolmen tapahtuman A, B ja C tilanteessa pätevät tällöin joko epäyhtälöt

$$\begin{aligned} \mathsf{P}(A \mid C \cap B) &> \mathsf{P}(A \mid C \cap B^C) \quad \text{ja} \\ \mathsf{P}(A \mid C^C \cap B) &> \mathsf{P}(A \mid C^C \cap B^C) \quad \text{mutta} \\ \mathsf{P}(A \mid B) &< \mathsf{P}(A \mid B^C) \end{aligned}$$

tai päinvastaiset epäyhtälöt. Simpsonin paradoksin voi ajatella ilmenevän myös, jos osajoukoissa pätee yhtäsuuruus muttei koko joukossa tai päinvastoin.

Selitys epäyhtälöille edellä on ehdollistettu kokonaistodennäköisyyden laki (1). Sen mukaan

$$\mathsf{P}(A \mid B) = \mathsf{P}(A \mid C \cap B)\mathsf{P}(C \mid B) + \mathsf{P}(A \mid C^C \cap B)\mathsf{P}(C^C \mid B)$$

ja

$$\mathsf{P}(A \mid B^C) = \mathsf{P}(A \mid C \cap B^C)\mathsf{P}(C \mid B^C) + \mathsf{P}(A \mid C^C \cap B^C)\mathsf{P}(C^C \mid B^C).$$

Vaikka pätisi

$$\mathsf{P}(A \mid C \cap B) > \mathsf{P}(A \mid C \cap B^C) \text{ ja } \mathsf{P}(A \mid C^C \cap B) > \mathsf{P}(A \mid C^C \cap B^C),$$

niin $\mathsf{P}(C \mid B)$ tai $\mathsf{P}(C^C \mid B)$ voi olla niin pieni, tai $\mathsf{P}(C \mid B^C)$ tai $\mathsf{P}(C^C \mid B^C)$ niin suuri, että

$$\mathsf{P}(A \mid B) < \mathsf{P}(A \mid B^C).$$

Esimerkki. Kaksi maisteria (jatkoa). Sijoitetaan tiedot osuuksista ehdollisen kokonaistodennäköisyyden kaavaan:

$$\begin{aligned}
\mathsf{P}(A \mid B) &= \mathsf{P}(A \mid C \cap B)\mathsf{P}(C \mid B) + \mathsf{P}(A \mid C^C \cap B)\mathsf{P}(C^C \mid B) \\
&= \frac{70}{90} \times \frac{90}{100} + \frac{10}{10} \times \frac{10}{100} \\
&= \frac{80}{100} \\
&< \\
&= \frac{83}{100} \\
&= \frac{2}{10} \times \frac{10}{100} + \frac{81}{90} \times \frac{90}{100} \\
&= \mathsf{P}(A \mid C \cap B^C)\mathsf{P}(C \mid B^C) + \mathsf{P}(A \mid C^C \cap B^C)\mathsf{P}(C^C \mid B^C) \\
&= \mathsf{P}(A \mid B^C).
\end{aligned}$$

B:n todistuksessa aineen C suuri osuus $90/100 = \mathsf{P}(C \mid B)$ painottaa kohtuullista arvosanasuhdetta $70/90$ ja aineen C^C pieni osuus $10/100 = \mathsf{P}(C^C \mid B)$ mitätöi loistavaa arvosanasuhdetta $10/10$. B^C:n todistuksessa aineen C pieni osuus mykistää huonoa arvosanasuhdetta $2/10$ ja aineen C^C suuri osuus $90/100 = \mathsf{P}(C^C \mid B^C)$ rummuttaa hienoa arvosanasuhdetta $81/90$. Seuraus on epäyhtälö yllä. $\square$

Esimerkki. Hyvä miehille, hyvä naisille, huono ihmisille (Baker ja Kramer 2001). Verrataan syöpähoitoja B ja B^C. Merkitään miehiä C:llä, naisia C^C:lla, syövästä paranemista A:lla ja syöpään kuolemista A^C:lla. Sivun 3 taulukon mukaan B on B^C:a parempi hoito sekä miehille että naisille. Ihmisille B vaikuttaa huonommalta hoidolta. $\square$

Esimerkki. Hyvä miehille, hyvä naisille, huono ihmisille (jatkoa). Syöpäesimerkissä B^C-hoito vaikutti paremmalta ihmisille ylipäänsä. Huolellisempi tarkastelu paljastaa, että B on parempi hoito. Huomataan, että syöpähoidot tehoavat erilailla miehiin ja naisiin ja että heitä on aivan eri osuudet syöpähoidoissa (s:n 3 taulukko). Idea: Standardoidaan hoitoihin osallistuvien sukupuolten osuudet samoiksi ja verrataan hoitojen tehokkuutta standardoiduilla osuuksilla.

Kuva 3 havainnollistaa. Siinä on laskettu onnistuneiden hoitojen osuudet, jos hoito toimii sivun 3 taulukon mukaisilla sukupuolikohtaisilla osuuksilla ja naisten (C^C) osuus tutkimuksessa vaihtelee välillä $[0, 1]$.

Punainen suora

$$(1-w) \times \frac{7}{9} + w \times \frac{1}{1} = \frac{7}{9} + \frac{2}{9}w$$

kuvaa paranemisten osuutta hoidolla B ja sininen suora

$$(1-w) \times \frac{2}{10} + w \times \frac{81}{90} = \frac{1}{5} + \frac{7}{10}w$$

hoidolla B^C. Yllä w on naisten osuus hoidossa. Suorat ovat painotettuja keskiarvoja miesten ja naisten paranemisosuuksista kyseisellä hoitomuodolla. Kuvaan on merkitty pisteillä ja katkoviivoilla sivun 3 taulukon mukaisia osuuksia:

$$(1 - \frac{10}{100}) \times \frac{7}{9} + \frac{10}{100} \times \frac{1}{1} = 0.9 \times \frac{7}{9} + 0.1 = 0.8$$

(C^C:n osuus 0.1) ja

$$(1 - \frac{90}{100}) \times \frac{2}{10} + \frac{90}{100} \times \frac{81}{90} = 0.1 \times 0.2 + 0.9 \times 0.9 = 0.83$$

(C^C:n osuus 0.9). Punainen suora on sinisen suoran yläpuolella kaikilla naisosuuksilla.

Naisten osuus hoidoissa määrää, kumpi hoidoista vaikuttaa useammin parantavan syövän ihmisillä. Molemmat hoidot toimivat naisilla paremmin kuin miehillä ($10/10 > 70/90$ ja $81/90 > 2/10$), ja erityisen hyvin naisilla toimii hoito B (parantuneiden osuus $10/10 = 1$). Miehillä hoito B^C toimii erityisen huonosti (parantuneiden osuus $2/10 = 0.2$). Jos naisten osuus on hyvin pieni (0.1) hoidossa B ja hyvin suuri (0.9) hoidossa B^C, niin kokonaiskuva hoitojen tehokkuudesta vääristyy ja hoito B^C näyttää hoitoa B paremmalta ($83/100 > 80/100$). Jos naisten ja miesten osuudet hoitoihin osallistumisessa poikkeaisivat vähemmän, hoito B olisi yleensä parempi. Kuvaan on piirretty esimerkkinä tilanne, jossa naisten ja miesten osuudet hoidoissa ovat yhtäsuuret. Hoito B on tällöin selvästi parempi ($0.889 > 0.55$): Hoidollaa B parantumisosuus on

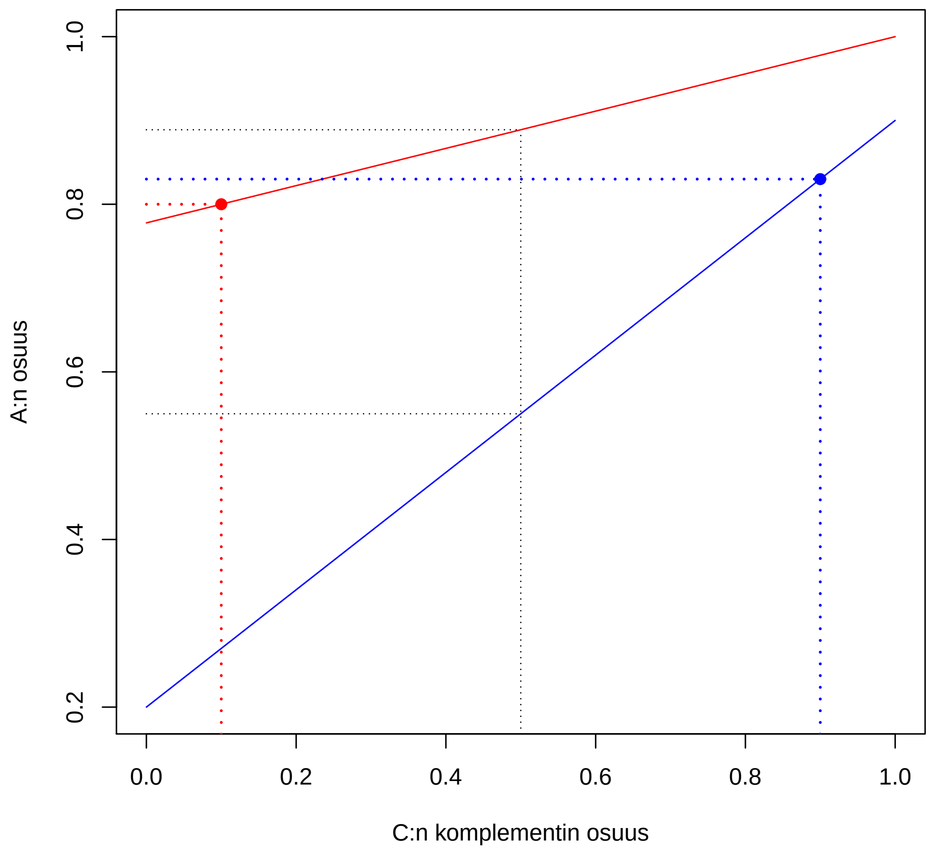
$$0.5 \times \frac{70}{90} + 0.5 \times \frac{10}{10} \approx 0.889$$

ja hoidolla B^C

$$0.5 \times \frac{2}{10} + 0.5 \times \frac{81}{90} = 0.55.$$

Jos miehet ja naiset voidaan ohjata vapaasti kumpaan tahansa hoitoon, heidät kannattaa ohjata hoitoon B. $\square$

Esimerkki. Hyvä miehille, hyvä naisille, huono ihmisille (jatkoa). Merkitään A:lla ja A^C:lla pitkä- ja lyhytikäisyyttä sekä B:llä ja B^C:lla geenin B omaamista ja puuttumista. Miehiä ja naisia osoitetaan edelleen C:llä ja C^C:lla. Geeni on yleinen miehillä mutta harvinainen naisilla, ja geenin omaavat miehet ja naiset ovat pitkäikäisempiä kuin omaamattomat (s:n 3 taulukko). Väestössä geenin omaamattomat ovat silti pitkäikäisempiä kuin geenin omaavat. Geenin omaamisen osuutta sukupuolittain tai sukupuolien osuutta väestössä ei voi muuttaa. Geeni B on siten ylitsepääsemättömästi hyvä naisille, hyvä miehille mutta huono ihmisille. $\square$



Kuva 3: A:n osuuden riippuvuus C^C:n osuudesta.

Simpsonin paradoksin tulkinta on tapauskohtainen. Ylipäänsä ehdollistettu (tarkemmin luokiteltu) tieto on informatiivisempaa, ja siihen kannattaa kiinnittää huomiota enemmän kuin ehdollistamattomaan (vähemmän tarkasti luokiteltuun). Silti luokittelematon tieto kuvaa kokonaisuuden.

Esimerkki.[4] Empiirisiä esimerkkejä Simpsonin paradoksista:

- Urheilussa pelaaja voi pelata toista pelaajaa paremmin kauden ensimmäisellä ja toisella kaudella mutta kokonaisuutena huonommin.

[4] P.J. Bickel, E.A. Hammel ja J.W. O'Connell (1975): Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, 187, 398–404. W.G. Cochran (1968): The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24, 295–313. S. Holm (2013): Tutkielmien arvosanat Helsingin yliopistossa. Pro gradu -tutkielma (tilastotiede). Matemaattis-luonnontieteellinen tiedekunta. Helsingin yliopisto. `https://helda.helsinki.fi/handle/10138/39493` (haettu 28.10.2016).

- Tupakoijien kuolleisuus on kaikissa ikäryhmissä suurempi kuin tupakoimattomien. Tupakoijien kuolleisuus ylipäänsä on kuitenkin pienempi kuin tupakoimattomien, koska tupakoijat ovat keskimäärin tupakoimattomia nuorempia. (Cochran 1968.)
- Yhdysvalloissa Floridan osavaltiossa 1976–1986 valkoihoiset tuomittiin mustaihoisia useammin kuolemaan mutta aineiston alaryhmissä mustaihoiset. (Agresti 2019.)
- Berkeleyn yliopisto haastettiin oikeuteen sukupuolisyrjinnästä vuoden 1973 opiskelijavalinnan takia. Mieshakijoista oli hyväksytty selvästi suurempi osuus kuin naishakijoista. Asiaa tutkittaessa ilmeni, että monilla laitoksilla naisten hyväksymisprosentit olivat suurempia kuin miesten. Selitys miesten suurempaan hyväksymisprosenttiin ylipäänsä oli, että he pyrkivät oppiaineisiin (esim. matemaattisiin tai teknisiin tieteisiin), joille oli helpompi päästä (hyväksyttyjen osuudella mitattuna) kuin naisten suosimiin oppiaineisiin (esim. psykologia). (Bickel ym. 1978.)
- Helsingin yliopiston valtiotieteellisessä tiedekunnassa miehet saavat pro gradu -tutkielmista parempia arvosanoja kuin naiset. Ero häviää, jos verrataan arvosanoja oppiaineittain. Ilmiö selittyy näin: Taloustieteessä myönnetään korkeita arvosanoja ja valtaosa opiskelijoista on miehiä. Sosiaalityössä arvosanat ovat huonompia ja opiskelijoiden enemmistö on naisia. Molemmissa oppiaineissa sukupuolet saavat yhtälailla arvosanoja. (Holm 2013.) □