Sovellettu todennäköisyyslaskenta B

Antti Rasila

16. marraskuuta 2007

χ^2-yhteensopivuustesti 1/5

- Tässä esitettävä χ^2-yhteensopivuustesti on sikäli historiallinen, että se on ensimmäinen todennäköisyyksiin perustuva tilastollinen päättelytesti.
- Testin esitti vuonna 1900 Karl Pearson (1857-1936). Testi merkitsi alkua modernille tilastotieteelle.
- Yhteensopivuustestissä tarkastellaan, onko satunnaismuuttujasta X tehdyt havainnot sopusoinnussa X:n jakaumasta tehtyjen ennakko-oletusten kanssa.
- Menetelmä eroaa aikaisemmin tällä kurssilla esitetyistä testeistä siinä, että nollahypoteesi ei koske jakauman parametreja. Yhteensopivuustesti kuuluu *epäparametrisiin testeihin*.
- Yleisemmin χ^2-testillä voidaan tutkia, onko havaintoaineisto syntynyt hypoteesin edellyttämällä tavalla, esim. onko havaintoaineisto peräisin tietystä jakaumasta, ovatko havaintoaineistot peräisin samasta jakaumasta, ovatko kaksi mitattua suuretta riippumattomia, jne.

χ^2-yhteensopivuustesti 2/5

- Oletetaan, että havainnot $X_1, \ldots, X_n$ muodostavat satunnaisotoksen perusjoukosta S.
- Nollahypoteesi H_0: Havainnot $X_1, \ldots, X_n$ noudattavat todennäköisyysjakaumaa $f(x;\theta)$, jonka parametrit eivät välttämättä ole tunnettuja.
- Vaihtoehtoinen hypoteesti H_1: Havainnot $X_1, \ldots, X_n$ eivät noudata nollahypoteesin määräämää todennäköisyysjakaumaa.
- Luokitellaan havainnot $X_1, \ldots, X_n$ toisensa poissulkeviin luokkiin, joiden lukumäärä on m. Olkoot
$$O_k, \qquad k = 1, 2, \ldots, m$$
luokkia k vastaavat havaintojen frekvenssit.
- Huomaa, että
$$\sum_{i=1}^{k} O_k = n.$$

χ^2-yhteensopivuustesti 3/5

- Oletetaan, että nollahypoteesti H_0 määrää satunnaismuuttujan X jakauman tyypin, mutta jakauman parametrit ovat tuntemattomia. Oletetaan, että P_k on todennäköisyys sille, että X saa arvon luokasta k, kun nollahypoteesi H_0 pätee.
- Luokkaan k kuuluvien havaintojen odotettu frekvenssi E_k on
$$E_k = nP_k, \qquad k = 1, 2, \ldots, ,.$$
- Jakaumassa esiintyvät tuntemattomat parametrit täytyy estimoida havainnoista.
- *Odotetut solufrekvenssit* E_k toteuttavat yhtälön
$$\sum_{k=1}^{m} E_k = n.$$

χ^2-yhteensopivuustesti 4/5

- Määritellään χ^2-testisuure
$$\chi^2 = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k},$$
missä O_k on havaittu frekvenssi, E_k odotettu frekvenssi ja m luokkien lukumäärä.
- Testisuure voidaan kirjoittaa myös muotoon
$$\chi^2 = \sum_{k=1}^{m} \frac{(\hat{p}_k - P_k)^2}{P_k},$$
missä $\hat{p}_k$ on suhteellinen frekvenssi luokassa k, P_k tn. sille, että havainto kuuluu luokkaan k (jos nollahypoteesin pätee) ja m luokkien lukumäärä.
- Jos H_0 pätee, testisuure noudattaa suurissa otoksissa approksimatiivisesti χ^2-jakaumaa vapausastein $f = m - 1 - p$, missä p frekvenssien E_k määräämiseksi estimoitujen parametrien lukumäärä.

χ^2-yhteensopivuustesti 5/5

- Approksimaatio on tavallisesti riittävän hyvä, jos odotetut frekvenssit E_k toteuttavat ehdot
$$E_k > 5, \qquad k = 1, 2, \ldots, m.$$
- Testisuuren χ^2 normaaliarvo, eli odotusarvo nollahypoteesin pätiessä on
$$\mathrm{E}(\chi^2) = f, \qquad f = m - 1 - p.$$
- Normaaliarvoa merkitsevästi suuremmat testisuureen arvot viittaavat siihen, että H_0 ei päde.
- Normaaliarvoa merkitsevästi pienemmät testisuureen arvot viittaavat siihen, että havaintojen ja nollahypoteesin määräämän jakauman yhteensopivuus on *liian hyvää*: havainnot on mahdollisesti väärennetty.

Esimerkki (Lainen) 1/2

- Tutkittaessa nopan virheettömyyttä suoritettiin $n = 120$ heittoa. Saadut silmälukujen määrät olivat

i	1	2	3	4	5	6
n_i	23	17	14	24	16	26

- Virheettömällä nopalla $Pr(X = i) = 1/6$, mitään parametreja ei tarvitse estimoida.
- Kunkin silmäluvun odotetaan siis esiintyvän 120 heitossa $120/6 = 20$ kertaa, eli $E_i = 20$, kun $i = 1, \ldots, 6$.
- χ^2-testisuure saa arvon
$$\chi^2 = \frac{(23-20)^2}{20} + \frac{(17-20)^2}{20} + \ldots + \frac{(26-20)^2}{20} = 6.10.$$

Esimerkki (Laininen) 2/2

- Luokkia on $m = 6$ kappaletta, eikä parametreja tarvinnut estimoida, joten $p = 0$. Testisuure on siis asymptoottisesti χ^2-jakautunut vapausasteilla $\nu = m - p - 1 = 5$.
- Testin p-arvoksi saadaan (tietokoneella)

$$Pr(\chi^2 \geq 6.10) = 0.2966.$$

- Taulukosta voidaan katsoa, että $\chi^2_{0.05} = 11.1$ vapausasteilla 5. Koska testisuudeen arvo 6.10 on tätä pienempi, $Pr(\chi^2 \geq 6.10) > 0.05$.
- Noppa hyväksytään virheettömäksi.

Homogeenisuuden testaaminen 1/4

- Oletetaan, että perusjoukko S on jaettu r:ään ryhmään, joista on poimittu toisistaan riippumattomat satunnaisotokset.
- Nollahypoteesi H_0: Havainnot noudattavat jokaisessa ryhmässä $i = 1, \ldots, r$ samaa jakaumaa.
- Vaihtoehtoinen hypoteesi H_1: Havainnot eivät ryhmissä $i = 1, \ldots, r$ eivät noudata samaa jakaumaa.
- Poimitaan ryhmistä toisistaan riippumattomat satunnaisotokset, joiden koot ovat

$$n_i, \qquad i = 1, 2, \ldots, r.$$

- Luokitellaan havainnot jokaisesta otoksesta samaa luokitusta käyttäen toisensa poissulkeviin luokkiin, joiden lukumäärä on c. Määrätään otoksen i luokkaan j kuuluvien havaintojen havaittu frekvenssi

$$O_{ij}, \qquad i = 1, 2, \ldots, r, \ j = 1, 2, \ldots, c.$$

Homogeenisuuden testaaminen 2/4

Nämä voidaan esittää taulukkona:

	1	2	...	c	Σ
1	O_{11}	O_{12}	...	O_{1c}	n_1
2	O_{21}	O_{22}	...	O_{2c}	n_2
...	...	...	...	...	...
r	O_{r1}	O_{r2}	...	O_{rc}	n_r
Σ	C_1	C_2	...	C_c	n

Tässä r on ryhmien lukumäärä, c on luokkien lukumäärä O_{ij} havaittu frekvenssi ryhmän i luokassa j, n_i otoskoko ryhmässä i, C_j havaittu frekvenssi yhdistetyn havaintoaineiston luokassa j ja n havaintojen kokonaislukumäärä.

Homogeenisuuden testaaminen 3/4

- Erityisesti pätee, että
  (i) $\sum_{j=1}^{c} E_{ij} = n_i$,
  (ii) $\sum_{i=1}^{r} O_{ij} = C_j$, ja
  (iii)

$$\sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} = \sum_{i=1}^{r} n_i = \sum_{j=1}^{c} C_j = n.$$

- Määrätään nollahypoteesin pätiessä odotetut solufrekvenssit E_{ij} yhtälöillä

$$E_{ij} = nP_{ij} = \frac{n_i C_j}{n}, \qquad i = 1, \ldots, r, \ j = 1, \ldots, c.$$

- Kuten edellä, näistä voidaan muodostaa taulukko.

Homogeenisuuden testaaminen 4/4

- Määritellään χ^2-testisuure

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- Jos nollahypoteesi pätee, testisuure noudattaa suurissa otoksissa approksimatiivisesti χ^2-jakaumaa vapausastein $f = (r-1)(c-1)$, missä r on ryhmien ja c luokkien määrä.
- Approksimaatio on tavallisesti riittävän hyvä, jos $E_{ij} > 1$ ja keskimääräiset frekvenssit $C_j/r > 5$.
- Testisuureen normaaliarvo eli odotusarvo nollahypoteesin pätiessä on $\mathrm{E}(\chi^2) = f$.
- Normaaliarvoa merkitsevästi suuremmat χ^2-testisuureen arvot viittaavat siihen, että nollahypoteesi ei päde.

Esimerkki (Laininen) 1/2

- Tehtaassa valmistetaan tuotetta kolmella tuotantolijalla. Valmistetut tuotteet luokitellaan luokkiin 0 (virheetön), 1 (yksi virhe) ja 2 (vähintään kaksi virhettä). Tutkitaan, onko virheiden lukumäärien jakaumissa eroja eri tuotantolinjojen välillä.
- Otetaan jokaiselta tuotantolinjalta 200 yksikköä tuotetta ja lasketaan virheellisyysluokkien frekvenssit:

Vikoja	0	1	2	Σ
Linja 1	169	18	13	200
Linja 2	144	23	33	200
Linja 3	180	9	11	200
Σ	493	50	57	600

- Silmämääräisesti näyttäisi siltä, että Linja 2 tuottaa muita ennemmän virheitä. Onko havainto tilastollisesti luotettava?

Esimerkki (Laininen) 2/2

- Jos linjojen välillä ei ole eroja, on jokaisella linjalla virheettömien tuotteiden lukumäärän estimaatti $200 \cdot 493/600 = 164.33$, yhden virheen sisältävien estimaatti $200 \cdot 50/600 = 16.67$ ja vähintään kaksi virhettä sisältävien $200 \cdot 57/600 = 19.00$.
- Testisuureen arvo on tällöin

$$\chi^2 = \frac{(169 - 164.33)^2}{164.33} + \frac{(18 - 16.67)^2}{16.67} + \frac{(11 - 19.00)^2}{19.00} = 25.76.$$

- Vapausasteparametri $\nu = (3-1)(3-1) = 4$.
- Testin p-arvoksi saadaan

$$Pr(\chi^2 \geq 25.76) = 0.0000,$$

joten linjojen välillä on tilastollisesti merkitseviä eroja.