Luku 7

Yhteensopivuuden ja riippumattomuuden testaaminen

Tässä kappaleessa tarkastelemme eräitä kuuluisia frekventistisen tilastotieteen testejä, joilla voidaan tutkia diskreettien havaintojen sopivuutta erilaisten tilastollisten mallien kanssa. Kaikki tämän kappaleen testit ovat likimääräisiä, ja niiden käyttö vaatii suurta otoskokoa.

7.1 Kaksi likimäärin χ^2-jakautunutta testisuuretta

Tilastollinen malli koostuu nyt diskreeteistä satunnaismuuttujia $Y_1, \ldots, Y_n$, joista kukin voi saada minkä tahansa arvoista $1, 2, \ldots, k$. Oletamme, että satunnaismuuttujat Y_h ovat riippumattomia ja samoin jakautuneita, jolloin niiden yhteisjakauma tiedetään, mikäli tunnetaan eri vaihtoehtojen $1, \ldots, k$ eli eri luokkien todennäköisyydet

$$p_i = P(Y_h = i), \qquad i = 1, \ldots, k \tag{7.1}$$

Näiden todennäköisyyksien summa on yksi, joten mallissa on vain $k-1$ vapaata parametria, joiksi voidaan valita $k-1$ ensimmäisen luokan todennäköisyydet $p_1, \ldots, p_{k-1}$. Tämän jälkeen p_k voidaan laskea muiden p_i funktiona kaavalla

$$p_k = 1 - p_1 - p_2 - \cdots - p_{k-1}, \tag{7.2}$$

jota voidaan jatkossa esitettävissä kaavoissa pitää suureen p_k määritelmänä.

Binomikoe on tämän mallin erikoistapaus, jossa vaihtoehtoja on vain kaksi, ja joista toista pidetään onnistumisena ja toista epäonnistumisena. Tämän jakson mallia voidaan kutsua multinomikokeeksi.

Havaintoja $y_1, \ldots, y_n$ vastaava uskottavuusfunktio on

$$L(p_1, \ldots, p_{k-1}) = \prod_{h=1}^{n} p_1^{1(y_h=1)} p_2^{1(y_h=2)} \cdots p_k^{1(y_h=k)}$$

Tässä $1(y_h = i)$ on osoitinmuuttuja sille, että y_h:n arvo on i, eli

$$1(y_h = i) = \begin{cases} 1 & \text{mikäli } y_h = i, \\ 0 & \text{muuten.} \end{cases}$$

Kun luokkien todennäköisyyksien p_i potenssit yhdistetään, uskottavuusfunktiolle saadaan lauseke

$$L(p_1, \ldots, p_{k-1}) = p_1^{n_1}\, p_2^{n_2}\, \ldots\, p_k^{n_k}, \tag{7.3}$$

jossa n_i on niiden indeksien h lukumäärä, joille $y_h = i$ eli n_i on luokan i havaittu frekvenssi,

$$n_i = \sum_{h=1}^{n} 1(y_h = i), \qquad i = 1, \ldots, k.$$

Merkitsemme vastaavia satunnaismuuttujia symboleilla N_i. Binomijakauman määritelmän perusteella

$$N_i \sim \mathrm{Bin}(n, p_i), \qquad i = 1, \ldots, k.$$

Uskottavuusfunktion kaavan (7.3) perusteella logaritminen uskottavuusfunktio on

$$\ell(p_1, \ldots, p_{k-1}) = \sum_{i=1}^{k} n_i\, \log(p_i), \tag{7.4}$$

ja log tarkoitta luonnollista logaritmia. On mahdollista osoittaa, että parametrien p_i suurimman uskottavuuden estimaatit ovat vastaavat suhteelliset frekvenssit

$$\hat{p}_i = \frac{n_i}{n}, \qquad i = 1, \ldots, k. \tag{7.5}$$

Karl Pearson esitti v. 1900 perustelun sille, miksi tässä tilanteessa suureella

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - n\,p_i)^2}{n\,p_i} \tag{7.6}$$

on asymptoottisesti χ^2_{k-1}-jakauma, jossa vapausasteparametri on yhtä kuin vapaiden parametrien p_i lukumäärä eli $k-1$. Pearsonin χ^2-testisuure ilmaistaan usein kaavalla

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{7.7}$$

jossa $O_i = N_i$ on havaittu frekvenssi (engl. *observed frequency*), ja $E_i = np_i$ on odotettu frekvenssi (engl. *expected frequency*) eli satunnaismuuttujan N_i odotusarvo.

Pearsonin tulokseen perustuva testi on ensimmäinen tunnettu tilastollinen testi, ja se avasi uuden aikakauden (frekventistisen) tilastollisen päättelyn historiassa.

Tässä tilanteessa käytetään usein myös muita testisuureita kuten esim. uskottavuusosamäärän testisuuretta. Merkitään nyt tilastollisen mallin uskottavuusfunktiota $L(\boldsymbol{\theta}; \mathbf{Y})$, sen logaritmista uskottavuusfunktiota $\ell(\boldsymbol{\theta}; \mathbf{Y})$ ja suurimman uskottavuuden estimaattoria symbolilla $\hat{\boldsymbol{\theta}}(\mathbf{Y})$. Suurimman uskottavuu-

den estimaattorien asymptoottisen teorian perusteella tiedetään, että *uskottavuusosamäärän testisuureella* (engl. *likelihood ratio statistic*)

$$W = 2[\ell(\hat{\boldsymbol{\theta}}(\mathbf{Y}); \mathbf{Y}) - \ell(\boldsymbol{\theta}; \mathbf{Y})] = 2 \log \frac{L(\hat{\boldsymbol{\theta}}(\mathbf{Y}); \mathbf{Y})}{L(\boldsymbol{\theta}; \mathbf{Y})}$$

on tiettyjen oletusten vallitessa χ^2-jakauma, jossa vapausasteluku on yhtä kuin mallin vapaiden parametrien lukumäärä.

Tämän jakson tilastolliselle mallille uskottavuusosamäärän testisuureessa tarvittavat log-uskottavuusarvot ovat

$$\ell(\hat{p}_1(\mathbf{Y}), \ldots, \hat{p}_{k-1}(\mathbf{Y}); \mathbf{Y}) = \sum_{i=1}^{k} N_i \, \log \frac{N_i}{n}$$

$$\ell(p_1, \ldots, p_{k-1}; \mathbf{Y}) = \sum_{i=1}^{k} N_i \, \log p_i$$

joten itse testisuure on

$$W = 2 \sum_{i=1}^{k} N_i \, \log \frac{N_i}{n \, p_i} = 2 \sum_{i=1}^{k} O_i \, \log \frac{O_i}{E_i}, \tag{7.8}$$

ja sillä on suurella otoskoolla osapuilleen χ^2_{k-1}-jakauma. Jos jokin N_i tai O_i on nolla, niin tässä kaavassa pitää tulkita $0 \log 0 = 0$.

Yhteensopivuustestissä (engl. *goodness of fit test*) testataan nollahypoteesia

$$H_0 : p_1 = \pi_1, \ldots, p_{k-1} = \pi_{k-1}, p_k = \pi_k,$$

jossa tunnetut luvut (π_i) ovat jonkin teorian mukaisia luokkien todennäköisyyksiä. Aineistosta lasketaan havaitut frekvenssit $O_i = n_i$. Nollahypoteesin vallitessa odotetut frekvenssit ovat

$$E_i = n \, \pi_i, \qquad i = 1, \ldots, k.$$

Testissä lasketaan joko Pearsonin testisuureen X^2 arvo kaavalla (7.7) tai uskottavuusosamäärän testisuureen W arvo kaavalla (7.8). Suuret testisuureen arvot ovat nollahypoteesille kriittisiä. Nollahypoteesi hylätään merkitsevyystason $0 < \alpha < 1$ testissä, mikäli lasketun testisuureen (valinnan mukaan X^2 tai W) arvo on suurempi kuin χ^2_{k-1} jakauman α-yläkvantiili. Testin p-arvo on

$$1 - F(t),$$

jossa F on χ^2_{k-1}-jakauman kertymäfunktio, ja t on testisuureen laskettu arvo.

Sekä Pearsonin χ^2-testisuureeseen perustuva testi että uskottavuusosamäärän testisuureeseen perustuva testi ovat likimääräisiä, sillä ne molemmat perustuvat likimääräiseen suuren otoskoon jakaumatulokseen. Tavallisesti nämä kaksi testisuuretta saavat likimain saman arvon, ja testin päätös ei riipu siitä, kumpaa niistä käytetään.

Milloin jakauma-approksimaatio on tarpeeksi hyvä? Kirjallisuudessa löytyy tähän tilanteeseen erilaisia suosituksia. Testiä sovelletaan huolta vailla esim. silloin, jos kaikille luokille niiden odotetut frekvenssit ovat vähintään viisi. Jos

joidenkin luokkien odotetut frekvenssit ovat liian pieniä, niin sitten kyseisiä luokkia voidaan yhdistää keskenään ennen testin soveltamista.

Yhteensopivuustestien voima kasvaa otoskoon kasvaessa. Jos nollahypoteesi ei pidä tarkasti paikkaansa, niin kyllin suurella otoskoolla se jossakin vaiheessa hylätään, vaikka poikkeama olisi niin pieni, että sillä ei ole käytännön kannalta merkitystä. Poikkeamat nollahypoteesista voivat olla monenlaisia, eikä ole helppoa määritellä skalaariparametria, joka jollakin tavalla mittaisi käytännön kannalta oleellista poikkeamaa nollahypoteesista.

Esimerkki 7.1. (Nopan harhattomuuden testaaminen) Simuloidaan ensin $n = 2000$ nopanheittoa harhaisesta nopasta, jolle silmälukujen todennäköisyydet ovat

$$P(Y = 1) = P(Y = 2) = P(Y = 3) = P(Y = 4) = P(Y = 5) = \frac{4}{25},$$
$$P(Y = 6) = \frac{5}{25}.$$

R:llä simulointi ja havaittujen frekvenssien laskeminen sujuvat seuraavasti.

```
> n <- 2000
> p <- c(4, 4, 4, 4, 4, 5)
> p <- p / sum(p)
> y <- sample(1:6, size = n, replace = TRUE, prob = p)
> y[1:20]

 [1] 5 6 2 5 3 5 2 4 6 6 6 5 4 2 2 2 4 2 5 2

> print(observed <- table(y))

y
  1   2   3   4   5   6
330 318 303 348 316 385
```

Seuraavaksi lasketaan sekä Pearsonin X^2-testisuureen että uskottavuusosamäärän testisuureen W arvo, χ^2_{k-1}-jakauman kriittinen arvo sekä testisuureita vastaavat p-arvot. Testattavan nollahypoteesin mukaan kaikkien silmälukujen todennäköisyys on $1/6$.

```
> alpha <- 0.05
> p.theory <- c(1, 1, 1, 1, 1, 1) / 6
> expected <- n * p.theory
> print(x2 <- sum((observed - expected)^2 / expected))

[1] 13.054

> print(w <- 2 * sum(observed * log(observed / expected)))

[1] 12.77532

> nu <- length(p) - 1
> print(crit <- qchisq(alpha, df = nu, lower = FALSE))

[1] 11.0705
```

```
> print(p.value.x2 <- pchisq(x2, df = nu, lower = FALSE))

[1] 0.02287795

> print(p.value.w <- pchisq(w, df = nu, lower = FALSE))

[1] 0.0255772
```

Nollahypoteesi hylätään merkitsevyystasolla $\alpha = 0.05$ kumpaa tahansa testisuuretta käyttäen. Pearsonin χ^2-testisuuretta käyttävä testi saadaan suoritettua myös R:n funktiolla `chisq.test`.

```
> chisq.test(observed)

        Chi-squared test for given probabilities

data:  observed
X-squared = 13.054, df = 5, p-value = 0.02288
```

△

Usein nollahypoteesi on yhdistetty, ja tällöin odotettuja frekvenssejä ei saada laskettua ennen kuin jonkin haittaparametrin $\boldsymbol{\varphi}$ arvo on estimoitu. Mikäli $\boldsymbol{\varphi}$ estimoidaan suurimman uskottavuuden menetelmällä, ja estimaattori on $\hat{\boldsymbol{\varphi}}$, niin tällöin voidaan osoittaa, että testisuureena voidaan käyttää joko Pearsonin χ^2-testisuuretta tai uskottavuusosamäärän testisuuretta, joilla on ennestään tutut kaavat

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{7.9}$$

$$W = 2 \sum_{i=1}^{k} O_i \log \frac{O_i}{E_i}. \tag{7.10}$$

Uusi asia on se, että odotetut frekvenssit pitää laskea käyttämällä haittaparametrin tilalla sen suurimman uskottavuuden estimaattoria

$$E_i = n\, p_i(\hat{\boldsymbol{\varphi}}), \qquad i = 1, \ldots, k. \tag{7.11}$$

Jos haittaparametrilla $\boldsymbol{\varphi}$ on d rajoittamatonta komponenttia, niin tällöin kummalla tahansa testisuureella on suurella otoskoolla osapuilleen χ^2-jakauma vapausasteluvulla

$$\nu = k - 1 - d. \tag{7.12}$$

Tämän tuloksen perusteli ensimmäisenä Fisher 1920-luvulla. Hän osoitti samalla, että K. Pearson oli tehnyt virheen omissa laskuissaan vapausasteluvun suuruudesta.

Näemme tälle menetelmälle sovelluksen seuraavassa jaksossa.

Taulukko 7.1 Kontingenssitaulukko, jolla on r riviä ja c saraketta.

n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	n

7.2 Riippumattomuuden testaaminen kontingenssitaulukossa

Tarkastellaan n otosyksikköä, joista kustakin mitataan kaksi diskreettiä ominaisuutta (x_h, y_h), jossa $h = 1, \ldots, n$. Ominaisuus x_h saa yhden arvoista $1, \ldots, r$ ja ominaisuus y_h yhden arvoista $1, \ldots, c$.

Tilastollinen malli koostuu n riippumattomasta ja samoin jakautuneesta satunnaismuuttujaparista (X_h, Y_h), joille

$$p_{ij} = P(X_h = i, Y_h = j), \qquad i = 1, \ldots, r, \quad j = 1, \ldots, c.$$

Tilanne on muuten samanlainen kuin edellisessä jaksossa, mutta nyt luokkia indeksoidaan kahdella indeksillä i ja j eikä enää yhdellä indeksillä. Luokkia on rc, joten vapaita parametreja p_{ij} on $rc - 1$, mikäli niitä ei rajoiteta lisäämällä malliin oletuksia. Havaitut frekvenssit ovat

$$n_{ij} = \sum_{h=1}^{n} 1(x_h = i, y_h = j), \qquad i = 1, \ldots, r, \quad j = 1, \ldots, c.$$

ja ne voidaan esittää taulukkona, jossa i indeksoi vaakarivejä ja j sarakkeita. Tälläistä taulukkoa kutsutaan *kontingenssitaulukoksi* (engl. *contingency table*). Taulukon rivin i summaa merkitään $n_{i\cdot}$ ja sarakkeen j summaa $n_{\cdot j}$, ts. alaindeksi piste tarkoittaa summaamista kyseisen indeksin yli, eli

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}, \qquad i = 1, \ldots, r$$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}, \qquad j = 1, \ldots, c.$$

Testattavan nollahypoteesin mukaan satunnaismuuttujat X_h ja Y_h ovat riippumattomat, mikä tarkoittaa sitä, että kaikilla i ja j niiden yhteisjakauman pistetodennäköisyysfunktio saadaan kertomalla keskenään vastaavat reunajakaumien pistetodennäköisyydet, eli

$$p_{ij} = P(X_h = i, Y_h = j) = P(X_h = i)\, P(Y_h = j) = \gamma_i\, \delta_j, \quad \text{missä}$$
$$\gamma_i = P(X_h = i), \qquad \delta_j = P(Y_h = j).$$

Reunajakaumien pistetodennäköisyydet $\gamma_1, \ldots, \gamma_r$ ja $\delta_1, \ldots, \delta_c$ ovat tuntemattomia parametreja. Nollahypoteesi voidaan ilmaista myös sanomalla, että rivi- ja sarakeluokittelut ovat riippumattomia.

Voidaan osoittaa, että nollahypoteesin vallitessa suurimman uskottavuuden estimaatit ovat

$$\hat{\gamma}_i = \frac{n_{i\cdot}}{n}, \qquad i = 1, \ldots, r \tag{7.13}$$

$$\hat{\delta}_j = \frac{n_{\cdot j}}{n}, \qquad j = 1, \ldots, c. \tag{7.14}$$

Jälleen kerran todennäköisyysparametrien suurimman uskottavuuden estimaatit ovat vastaavat suhteelliset frekvenssit.

Kun nollahypoteesi pitää paikkansa, niin tuntemattomia rajoittamattomia parametreja on

$$d = r - 1 + c - 1$$

kappaletta, sillä molemmat reunatodennäköisyysfunktiot summautuvat ykköseksi. Khiin neliön testissä vapausasteiden lukumääräksi saadaan

$$\nu = rc - 1 - (r - 1 + c - 1) = (r - 1)(c - 1).$$

Havaitut frekvenssit ovat $O_{ij} = n_{ij}$, ja nollahypoteesin vallitessa odotetut frekvenssit ovat

$$E_{ij} = n\,\hat{\gamma}_i\,\hat{\delta}_j = \frac{n_{i\cdot}\,n_{\cdot j}}{n}$$

Pearsonin χ^2-testisuure on

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

ja uskottavuusosamäärän testisuure on

$$W = 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

Koon $0 < \alpha < 1$ testissä testisuuretta verrataan χ^2_ν jakauman α-yläkvantiiliin, jossa vapausasteluku

$$\nu = (r - 1)(c - 1).$$

Esimerkki 7.2. (ABO-veriryhmien geneettinen perusta [1, Section 4.5]) Ihmisiä luokitellaan eri veriryhmiin veren ominaisuuksien perusteella. Tunnetuin veriryhmäjärjestelmä on ns. ABO-veriryhmäjärjestelmä, jossa kunkin yksilön veriryhmä on joko A, B, AB tai O. Veriryhmä määritetään tarkistamalla, onko henkilön veressä antigeeniä A tai antigeeniä B. Veriryhmät nimetään tuloksen perusteella seuraavan taulukon mukaisesti

	Ei B	On B
Ei A	O	B
On A	A	AB

Vielä 1920-luvulla oli epäselvää, minkälaisesta geneettisestä mekanismista ABO-veriryhmät määräytyvät. Yksi mahdollinen selitys oli kahden riippumattoman lokuksen malli, jossa lokuksen yksi alleeli määrää, onko veressä antigeeniä A vai ei, ja lokuksen kaksi alleeli määrää, onko veressä antigeeniä B vai ei. Jos kahden riippumattoman lokuksen malli on tosi ja jos otamme populaatiosta satunnaisotoksen, niin tällöin veriryhmistä muodostetussa kontingenssitaulukossa rivi- ja sarakeluokittelut ovat riippumattomia. Englannissa 1930-luvulla tehdystä otoksesta saatiin seuraava veriryhmien kontingenssitaulukko

	Ei B	On B
Ei A	202	35
On A	179	6

Tästä taulukosta laskettuna Pearsonin χ^2 testisuureen arvo on $X^2 = 15.73$ ja uskottavuusosamäärän testisuureen arvo on $W = 17.67$. Näitä verrataan χ^2_1-jakauman α-yläkvantiiliin. Esim. jos merkitsevyystaso on $\alpha = 0.05$, niin tämä kriittinen arvo on 3.84, joten kahden lokuksen malli tulee testissä hylättyä. R:llä testisuureiden arvot voidaan laskea seuraavasti.

```
> print(bloodgroups <- matrix(c(202, 179, 35, 6), 2, 2))

     [,1] [,2]
[1,]  202   35
[2,]  179    6

> print(n <- sum(bloodgroups))

[1] 422

> print(gamma.hat <- rowSums(bloodgroups) / n)

[1] 0.5616114 0.4383886

> print(delta.hat <- colSums(bloodgroups) / n)

[1] 0.9028436 0.0971564

> observed <- bloodgroups
> expected <- n * (gamma.hat %o% delta.hat)
> print(x2 <- sum((observed - expected)^2 / expected))

[1] 15.73193

> print(w <- 2 * sum(observed * log(observed / expected)))

[1] 17.66604

> nu <- (2 - 1) * (2 - 1)
> print(crit <- qchisq(0.05, df = nu, lower = FALSE))

[1] 3.841459
```

Testin voi suorittaa myös funktiolla `chisq.test`, mikäli käytetään Pearsonin χ^2-testisuuretta. Tämä funktio käyttää oletusarvoisesti ns. jatkuvuuskorjausta. Alla olevassa koodissa pyydetään erikseen olemaan käyttämättä jatkuvuuskorjausta, jotta tuloksia voitaisiin suoraan verrata aikaisempien laskujen kanssa.

```
> chisq.test(bloodgroups, correct = FALSE)

        Pearson's Chi-squared test

data:  bloodgroups
X-squared = 15.7319, df = 1, p-value = 7.298e-05
```

Nykyään tiedetään, että ABO-veriryhmä määräytyy yhdestä lokuksesta, jolla voi olla kolme alleelia A, B tai O, joista A ja B ovat dominoivia ja O resessiivinen. Nollahypoteesin hylkääminen on oikea päätös. △

7.3 Homogeenisuuden testaaminen

Kuten edellisessä jaksossa, nytkin oletetaan, että otosyksiköillä on kaksi diskreettiä ominaisuutta x ja y, ja x:n mahdolliset arvot ovat $1, \ldots, r$ ja y:n mahdolliset arvot $1, \ldots, c$. Populaatio jaetaan ominaisuuden x arvojen määräämiin ositteisiin siten, että ositteessa i ominaisuuden x arvo on i. Kustakin osasta tehdään riippumaton kokoa n_i oleva otos

$$Y_{ih}, \qquad h = 1, \ldots, n_i.$$

Tavoitteena on testata, ovatko ositteiden jakaumat samat eli ovatko ositteet homogeenisia. Nollahypoteesi on

$$H_0 : p_{ij} = \pi_j, \qquad \text{kaikilla } i = 1, \ldots, r \text{ ja } j = 1, \ldots, c, \tag{7.15}$$

missä

$$p_{ij} = P(Y_{ih} = j),$$

ja todennäköisyydet $(\pi_1, \ldots, \pi_c)$ ovat tuntemattomia.

Taas muodostetaan kontingenssitaulukko, jossa vaakariville i tulee ositteesta i lasketut frekvenssit, ja vaakarivin i frekvenssien summa $n_{i\cdot}$ on ositteesta i tehdyn otoksen koko n_i. Osoittautuu, että tämän jälkeen testaus voidaan tehdä aivan samoilla kaavoilla kuin edellisessä jaksossa. Todennäköisyyksien (π_i) suurimman uskottavuuden estimaateiksi saadaan

$$\hat{\pi}_j = \frac{n_{\cdot j}}{n}, \qquad j = 1, \ldots, c. \tag{7.16}$$

7.4 Suurimman uskottavuuden estimaatit

Tässä jaksossa johdamme tässä kappaleessa ilmoitetut suurimman uskottavuuden estimaattien kaavat. Kaavat olisi mahdollista johtaa monella menetelmällä. Voisimme eliminoida yhden todennäköisyysparametereista käyttämällä sitä tietoa, että ne summautuvat ykköseksi. Toinen mahdollisuus olisi käyttää Lagrangen keinoa rajoitteellisten optimointitehtävien ratkaisemiseksi. Tässä jaksossa tulokset kuitenkin johdetaan käyttämällä juuri tähän tilanteeseen sopivaa epäyhtälöä. Tällä konstilla johdoista tulee paljon yksinkertaisempia kuin yleisemmillä keinoilla.

Käytämme hyväksi aputulosta, joka sanoo, että luonnollisen logaritmin $\log(x)$ kuvaaja jää pisteeseen $x = 1$ piirretyn tangenttinsa alapuolelle.

$$\log(x) \leq x - 1, \qquad \text{kaikilla } x > 0. \tag{7.17}$$

Yhtäsuuruus saavutetaan ainoastaan pisteessä $x = 1$. Tämän väitteen voi tarkistaa helposti analyysin keinoilla. Estimaatien kaavat voidaan perustella helposti seuraavan lauseen avulla.

Lause 7.1. *Jos k ei-negatiivista lukua $n_i \geq 0$ summautuvat luvuksi $n > 0$, eli*

$$\sum_{i=1}^{k} n_i = n,$$

niin tällöin mille tahansa luvuille $p_i \geq 0$ jotka summautuvat ykköseksi pätee

$$\sum_{i=1}^{k} n_i \log p_i \leq \sum_{i=1}^{k} n_i \log \frac{n_i}{n},$$

missä käytämme tarvittaessa sopimusta $0 \log 0 = 0$.

Todistus. Esitän todistuksen vain siinä tapauksessa, jossa kaikki $n_i > 0$. Yleisen tapauksen saa todistettua helposti samaan tapaan. Väitetyn epäyhtälön vasemman ja oikean puolen erotus on

$$\begin{aligned}\sum \left(n_i \log p_i - n_i \log \frac{n_i}{n} \right) &= \sum n_i \log \frac{p_i}{n_i/n} \leq \sum n_i \left(\frac{p_i}{n_i/n} - 1 \right) \\ &= n \sum p_i - \sum n_i = n - n = 0,\end{aligned}$$

missä sovelsimme epäyhtälöä (7.17). □

Suurimman uskottavuuden estimaattien kaava (7.5) seuraa suoraan tästä lauseesta.

Jos testataan riippumattomuutta kontingenssitaulukossa, niin logaritminen uskottavuusfunktio on nollahypoteesin $p_{ij} = \gamma_i \, \delta_j$ vallitessa

$$\begin{aligned}\sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \, \log p_{ij} &= \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \, \log(\gamma_i \, \delta_j) = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \, [\log \gamma_i + \log \delta_j] \\ &= \sum_{i=1}^{r} \log \gamma_i \sum_{j=1}^{r} n_{ij} + \sum_{j=1}^{c} \log \delta_j \sum_{i=1}^{r} n_{ij} \\ &= \sum_{i=1}^{r} n_{i\cdot} \log \gamma_i + \sum_{j=1}^{c} n_{\cdot j} \log \delta_j \\ &\leq \sum_{i=1}^{r} n_{i\cdot} \log \frac{n_{i\cdot}}{n} + \sum_{j=1}^{c} n_{\cdot j} \log \frac{n_{\cdot j}}{n},\end{aligned}$$

mistä nähdään, että suurimman uskottavuuden estimaateilla on kaavat (7.13) ja (7.14).

Homogeenisuuden testauksessa uskottavuusfunktio on nollahypoteesin $p_{ij} = \pi_j$ vallitessa

$$L(\pi_1, \ldots, \pi_{c-1}) = \prod_{i=1}^{r} \pi_1^{n_{i1}} \, \pi_2^{n_{i2}} \ldots \pi_c^{n_{ic}} = \pi_1^{n_{\cdot 1}} \, \pi_2^{n_{\cdot 2}} \ldots \pi_c^{n_{\cdot c}},$$

joten suurimman uskottavuuden estimaattien kaavat (7.16) saadaan suoraan lauseesta 7.1, kun ensin siirrytään tarkastelemaan uskottavuusfunktion logaritmia.

Kirjallisuutta

[1] A. C. Davison. *Statistical Models.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 2003.