8 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.

8.1 Pearsonin testisuure

- Tilastollinen malli koostuu 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.
- Yhteisjakauma tiedetään, mikäli tunnetaan eri vaihtoehtojen $1, \ldots, k$ eli eri luokkien todennäköisyydet

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

- Luokkien todennäköisyyksien summa on yksi, joten mallissa on $k - 1$ 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

$$p_k = 1 - p_1 - p_2 - \cdots - p_{k-1}. \tag{2}$$

Binomikoe on multinomikokeen erikoistapaus

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

Multinomikokeessa tarkastellaan frekvenssejä

- Olkoon n_i luokan i havaittu frekvenssi, eli n_i on niiden indeksien h lukumäärä, joille $y_h = i$.
- Merkitsemme vastaavia satunnaismuuttujia symboleilla N_i.
- Binomijakauman määritelmän perusteella

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

joten $EN_i = np_i$.

Frekvenssien analysoinnissa pitää huomioida niiden riippuvuus toisistaan

- Yksittäisten kokeiden lopputulokset Y_h ovat riippumattomia.
- Jos i ja j ovat eri luokkia, niin frekvenssit N_i ja N_j ovat riippuvia satunnaismuuttujia, mikä pitää analyysissä ottaa huomioon.
- Kutsutaan nyt binomikokeessa onnistumista luokaksi yksi ja epäonnistumista luokaksi kaksi. Vastaavien frekvenssien jakaumat ovat

$$N_1 \sim \mathrm{Bin}(n, p), \qquad N_2 \sim \mathrm{Bin}(n, 1-p),$$

  mutta koska $N_2 = n - N_1$, niin satunnaismuuttujan N_2 arvo tiedetään, jos satunnaismuuttujan N_1 arvo tiedetään.
- Jos luokkia on enemmän kuin kaksi, niin frekvenssien keskinäinen riippuvuus ei ole yhtä äärimmäistä, mutta riippuvuus säilyy kuitenkin myös tässä tapauksessa.

Pearsonin testisuure

- Pearsonin testisuure vertaa luokkien havaittuja frekvenssejä n_i niiden odotettuihin frekvensseihin np_i seuraavalla tavalla:

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\,p_i)^2}{n\,p_i}. \tag{3}$$

- Tavanomaisempi tapa mitata vektorin $(n_1, \ldots, n_k)$ ja vektorin $(np_1, \ldots, np_k)$ välistä etäisyyttä olisi esim. laskea vektoreiden euklidinen etäisyys, jolloin ensin summataan komponenttien neliöidyt erotukset $(n_i - np_i)^2$, ja lopuksi tuloksesta lasketaan neliöjuuri.
- Tällainen vertailu ei tässä tapauksessa ole hyvä ajatus, sillä todennäköisemmissä luokissa on odotettavissa enemmän satunnaisvaihtelua kuin harvinaisemmissa luokissa. Tätä varten Personin testisuureessa erotuksen neliö jaetaan luokan odotetulla frekvenssillä.

Oi ei -kaava Pearsonin testisuureelle

- Pearsonin testisuure voidaan ilmaista myös kaavalla

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{4}$$

  jossa $O_i = n_i$ on havaittu frekvenssi (engl. *observed frequency*), ja $E_i = np_i$ on odotettu frekvenssi (engl. *expected frequency*).
- Karl Pearson esitti v. 1900 perustelun sille, miksi hänen nimeään kantavalla testisuureella on suuressa otoksessa likimain χ^2-jakauma.
- Pearsonin esittämä testi on ensimmäinen tunnettu tilastollinen testi, ja se avasi uuden aikakauden tilastollisen päättelyn historiassa.

Yhteensopivuustesti yksinkertaiselle hypoteesille

- Yksinkertaisimmassa yhteensopivuustestissä (engl. *goodness of fit test*) testataan yksinkertaista nollahypoteesia

$$H_0 : p_1 = \pi_1, \ldots, p_{k-1} = \pi_{k-1}, p_k = \pi_k, \tag{5}$$

  jossa luvut (π_i) ovat jonkin teorian mukaisia luokkien todennäköisyyksiä, jotka oletetaan tunnetuiksi.
- Aineistosta lasketaan havaitut frekvenssit $O_i = n_i$.
- Nollahypoteesin vallitessa odotetut frekvenssit ovat

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

Yhteensopivuustesti yksinkertaiselle nollahypoteesille (jatkoa)

- Suuret Pearsonin testisuureen X^2 arvot ovat nollahypoteesille kriittisiä.
- Nollahypoteesi hylätään merkitsevyystason $0 < \alpha < 1$ testissä, mikäli lasketun testisuureen arvo on suurempi kuin χ^2_{k-1} jakauman α-yläkvantiili.
- Huomaa, että χ^2_ν-jakauman vapausasteluku ν on tässä tapauksessa

$$\nu = k - 1, \tag{6}$$

  eli se on yhtä kuin mallin vapaiden parametrien lukumäärä $k - 1$.
- Testin p-arvo on $1 - F(t)$, missä F on χ^2_{k-1}-jakauman kertymäfunktio, ja t on testisuureen laskettu arvo.

Esimerkki: nopan harhattomuuden testaaminen

Simuloidaan $n = 2000$ nopanheittoa harhaisesta nopasta, jolle silmälukujen todennäköisyydet ovat

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

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

Nopan harhattomuuden testaaminen (jatkoa)

Testaan nyt eräässä tällaisessa simuloinnissa saatuja frekvenssejä, kun (epätoden) nollahypoteesin mukaan kaikilla silmäluvuilla on sama todennäköisyys 1/6.
Havaitut frekvenssit, odotetut frekvenssit sekä suureet $(O_i - E_i)^2/E_i$ ovat

silmäluku	1	2	3	4	5	6	summa
havaittu O_i	311	318	306	342	316	407	2000
odotettu E_i	333.3	333.3	333.3	333.3	333.3	333.3	2000
$(O_i - E_i)^2/E_i$	1.496	0.705	2.241	0.225	0.901	16.280	21.85

Testaus oman koodin avulla

Allaolevassa koodissa lasketaan Pearsonin X^2-testisuureen arvo.

```
obs <- c(`1` = 311, `2` = 318, `3` = 306, `4` = 342, `5` = 316,
    `6` = 407)
n <- sum(obs)
alpha <- 0.05
p0 <- c(1, 1, 1, 1, 1, 1)/6
expected <- n * p0
print(x2 <- sum((obs - expected)^2/expected))

## [1] 21.85
```

Kriittinen arvo ja *p*-arvo omalla koodilla

Lasketaan χ^2_{k-1}-jakauman kriittinen arvo sekä testisuuretta vastaava *p*-arvo.

```
nu <- length(p0) - 1
print(crit <- qchisq(alpha, df = nu, lower = FALSE))

## [1] 11.07

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

## [1] 0.0005591
```

Nollahypoteesi hylätään merkitsevyystasolla $\alpha = 0.05$.

Testin tekeminen valmiilla funktiolla

Sama testi saadaan suoritettua myös soveltamalla R:n funktiota `chisq.test`.

```
chisq.test(obs)

##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 21.85, df = 5, p-value = 0.0005591
```

Esimerkki: Pearsonin testisuure binomikokeessa

- Oletetaan, että onnistumisten lukumäärä K noudattaa binomijakaumaa $\mathrm{Bin}(n, p)$.
- (Kuvitteellisista) tuloksista saadaan taulukko

luokka	onnistuminen	epäonnistuminen	yhteensä
havaittu	K	$n - K$	n
odotettu	np	$n(1-p)$	n

- Tällöin

$$X^2 = \frac{(K-np)^2}{np} + \frac{[(n-K)-n(1-p)]^2}{n(1-p)} = \frac{(K-np)^2}{np(1-p)}.$$

- Teorian mukaan tämän satunnaismuuttujan pitäisi likimain noudattaa jakaumaa χ^2_1.

Miksi χ^2_1 on hyvä approksimaatio?

- Pearsonin testisuureen kaava oli $\dfrac{(K-np)^2}{np(1-p)}$.
- Binomijakauman normaalijakauma-approksimaation mukaan satunnaismuuttuja

$$\frac{K-np}{\sqrt{np(1-p)}}$$

noudattaa suurella otoskoolla n osapuilleen standardinormaalijakaumaa $N(0,1)$. Pearsonin testisuure on tämän satunnaismuuttujan neliö.

- Todennäköisyyslaskennasta tiedetään, että mikäli $Z \sim N(0,1)$, niin

$$Z^2 \sim \chi^2_1.$$

Milloin χ^2-approksimaatio on riittävän hyvä?

- Pearsonin X^2-testisuureeseen perustuva testi on likimääräinen, sillä se perustuu likimääräiseen suuren otoskoon jakaumatulokseen.
- Milloin tämä 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.

Yhteensopivuustesti yhdistetylle hypoteesille

- Usein yhteensopivuustestissä nollahypoteesi on yhdistetty, ja se voidaan esittää kaavalla

$$H_0 : p_1 = p_1(\theta), \ldots, p_k = p_k(\theta), \qquad \text{jollekin } \theta \in \Theta_0. \quad (7)$$

- Tällöin odotettuja frekvenssejä ei saada laskettua ennen kuin parametrin θ arvo on estimoitu.
- Oletetaan, että θ estimoidaan havaituista frekvensseistä suurimman uskottavuuden menetelmällä, ja että SU-estimaatti on $\hat{\theta}$.

Yhteensopivuustesti yhdistetylle hypoteesille

- Testisuureena voidaan käyttää Pearsonin testisuuretta,

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

- Odotetut frekvenssit lasketaan käyttämällä parametrin tilalla sen SU-estimaattia

$$E_i = n\, p_i(\hat{\theta}), \qquad i = 1, \ldots, k. \tag{8}$$

- Jos parametrilla θ on d vapaata komponenttia (ts. komponenttia, joita mikään sidosehto ei kytke toisiinsa) niin tällöin testisuureella (satunnaismuuttujaksi ymmärrettynä) on suurella otoskoolla osapuilleen χ^2-jakauma vapausasteluvulla

$$\nu = k - 1 - d. \tag{9}$$

Uskottavuusfunktio

- Yhden havaintosatunnaismuuttujan Y_h ptnf on

$$P(Y_h = y_h) = \begin{cases} p_1 & \text{jos } y_h = 1, \\ p_2 & \text{jos } y_h = 2, \\ \vdots & \\ p_k & \text{jos } y_h = k. \end{cases}$$
$$= p_1^{1(y_h=1)} \, p_2^{1(y_h=2)} \, \cdots p_k^{1(y_h=k)}$$

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

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

Uskottavuusfunktio (jatkoa)

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

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

- Kun luokkien todennäköisyyksien p_i potenssit yhdistetään, uskottavuusfunktiolle saadaan lauseke

$$L(p_1, \ldots, p_{k-1}) = p_1^{n_1} \, p_2^{n_2} \, \ldots \, p_k^{n_k}, \tag{10}$$

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

- Jos luokkien todennäköisyydet p_i riippuvat parametrin θ arvosta, niin uskottavuusfunktioksi saadaan

$$L(\theta) = p_1(\theta)^{n_1} \, p_2(\theta)^{n_2} \, \ldots \, p_k(\theta)^{n_k}, \quad \theta \in \Theta.$$

Alfahiukkasten lukumäärät

- On toistettu 2608 kertaa tietyn pituinen mittaus, jossa on laskettu hiukkaslaskuriin osuvien alfahiukkasten lukumäärä.
- Sitten on taulukoitu niiden mittausten lukumäärä, joissa laskuri on havainnut i kappaletta alfahiukkasia, kun $i = 0, 1, \ldots, 9$.
- Ne tapaukset joissa on havaittu $i \geq 10$ alfahiukkasta on yhdistetty yhdeksi luokaksi.
- Havaitut frekvenssit ovat

luokka	0	1	2	3	4	5	6	7	8	9	≥ 10	yht.
havaittu	57	203	383	525	532	408	273	139	45	27	16	2608

Yhteensopivuus Poissonin jakauman kanssa

- Tutkitaan, noudattaako alfahiukkasten lukumäärä Poissonin jakaumaa.
- Poissonin jakauma on diskreetti todennäköisyysjakauma, jonka pistetodennäköisyysfunktio on

$$g(x;\theta) = \mathrm{e}^{-\theta}\,\frac{\theta^x}{x!}, \qquad x = 0, 1, 2, \ldots,$$

jossa parametri $\theta > 0$ on jakauman odotusarvo.

Luokkien todennäköisyydet, kun mallina on Poissonin jakauma

- $x \mapsto g(x;\theta)$ on Poissonin jakauman ptnf.
- Luokat $0,\ldots,9$ vastaavat havaittujen alfahiukkasten lukumääriä, joten niiden todennäköisyydet ovat

$$p_i(\theta) = g(i;\theta), \qquad i = 0,1,\ldots,9.$$

- Viimeinen luokka vastaa sitä tapahtumaa, että laskuri havaitsee vähintään 10 alfahiukkasta, ja sen todennäköisyys on

$$p_{10}(\theta) = 1 - \sum_{i=0}^{9} g(i;\theta).$$

- Poissonin jakauman parametrin SU-estimaatiksi saadaan (tietokoneen avulla) $\hat{\theta} = 3.8703$.
- Kun tämä sijoitetaan Poissonin jakauman pistetodennäköisyysfunktioon, saadaan Pearsonin testisuure X^2 laskettua.

Testisuureen arvo

luokka	havaittu	odotettu	$(O_i - E_i)^2/E_i$
0	57	54.38	0.13
1	203	210.47	0.27
2	383	407.30	1.45
3	525	525.46	0.00
4	532	508.42	1.09
5	408	393.55	0.53
6	273	253.86	1.44
7	139	140.36	0.01
8	45	67.90	7.73
9	27	29.20	0.17
≥ 10	16	17.08	0.07
summa	2608	2608	12.88

Lopputulos

- Pearsonin testisuureen arvo on $X^2 = 12.88$.
- Sitä verrataan χ^2_ν-jakaumaan, jonka vapausasteluku on

$$\nu = k - 1 - d = 11 - 1 - 1 = 9.$$

- Testin p-arvoksi saadaan $p = 0.17$.
- Tämä ei ole erityisen pieni: jos Poisson-malli pitää paikkaansa, niin 17 %:n todennäköisyydellä saadaan testisuureelle arvoja joiden mukaan havaitut ja odotetut frekvenssit poikkeavat ainakin näin paljon toisistaan. Nollahypoteesi jää voimaan, jos käytetään esim. 5 %:n merkitsevyystasoa.

8.2 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.

Kontingenssitaulukko

- Havaitut frekvenssit

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

voidaan esittää taulukkona, jossa i indeksoi vaakarivejä ja j sarakkeita.

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

- Tälläistä taulukkoa kutsutaan *kontingenssitaulukoksi* (engl. *contingency table*).
- Taulukon rivin i summaa merkitään $n_{i\bullet}$ ja sarakkeen j summaa $n_{\bullet j}$, ts. alaindeksi piste tarkoittaa summaamista kyseisen indeksin yli.

Riippumattomuuden testaaminen

- Testaaan nollahypoteesia, jonka mukaan satunnaismuuttujat X_h ja Y_h ovat riippumattomat.
- Tämä tarkoittaa sitä, että yhteisjakauman pistetodennäköisyysfunktio saadaan kertomalla keskenään vastaavat reunajakaumien pistetodennäköisyydet. Ts. kaikilla i ja j

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

- 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.

SU-estimaatit

- Luvun lopussa osoitetaan, että nollahypoteesin vallitessa suurimman uskottavuuden estimaatit ovat

$$\hat{\gamma}_i = \frac{n_{i\bullet}}{n}, \qquad i = 1, \ldots, r \tag{11}$$

$$\hat{\delta}_j = \frac{n_{\bullet j}}{n}, \qquad j = 1, \ldots, c. \tag{12}$$

- Myös tässä tilanteessa todennäköisyysparametrien suurimman uskottavuuden estimaatit ovat vastaavat suhteelliset frekvenssit.

Vapausasteluku

- Kun nollahypoteesi pitää paikkansa, niin tuntemattomia vapaita parametreja on

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

kappaletta, sillä molemmat reunatodennäköisyysfunktiot summautuvat ykköseksi.

- Khiin neliön testissä vapausasteiden lukumääräksi saadaan kaavan (9) mukaan

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

Riippumattomuden testaus käytännössä

- Havaitut frekvenssit ovat $O_{ij} = n_{ij}$.
- Jos rivi- ja sarakeluokittelut ovat riippumattomat (nollahypoteesi), niin odotetut frekvenssit ovat

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

- Pearsonin χ^2-testisuure on

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

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

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

Esimerkki: ABO-veriryhmien geneettinen perusta

- Ihmiset luokitellaan eri veriryhmiin veren ominaisuuksien perusteella.
- Tunnetuin veriryhmäjärjestelmä on ns. ABO-veriryhmäjärjestelmä, jossa 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

Miten ABO-veriryhmä määräytyy geeneistä?

- 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.

Aineisto ja testi

- 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 $X^2 = 15.73$.
- Tätä 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.
- ABO-veriryhmä määräytyy oikeasti 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.

Veriryhmäaineiston analysointi R:llä itsetehdyllä koodilla

```
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] 0.5616 0.4384
```

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

```
## [1] 0.90284 0.09716
```

Testisuureen laskeminen omalla R-koodilla

```
observed <- bloodgroups
print(expected <- n * (gamma.hat %o% delta.hat))
```

```
##      [,1]  [,2]
## [1,]  214 23.03
## [2,]  167 17.97
```

```
print(x2 <- sum((observed - expected)^2/expected))
```

```
## [1] 15.73
```

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

```
## [1] 3.841
```

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

```
## [1] 7.298e-05
```

Riippumattomuuden testaaminen valmiilla funktiolla

Testin voi suorittaa funktiolla `chisq.test`, joka käyttää oletusarvoisesti ns. jatkuvuuskorjausta. Alla olevassa koodissa pyydetään erikseen olemaan käyttämättä jatkuvuuskorjausta, jotta tuloksia voitaisiin suoraan verrata itse tehtyjen laskujen kanssa.

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

##
##  Pearson's Chi-squared test
##
## data:  bloodgroups
## X-squared = 15.73, df = 1, p-value = 7.298e-05
```

8.3 Homogeenisuuden testaaminen

- Nytkin oletetaan, että otosyksiköillä on kaksi diskreettiä ominaisuutta x ja y.
- Ominaisuuden x:n mahdolliset arvot ovat $1, \ldots, r$ ja y:n mahdolliset arvot ovat $1, \ldots, c$.
- 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 homogeenisuuden testaamisessa

- Nollahypoteesi on

$$H_0 : p_{ij} = \pi_j, \qquad \text{kaikilla } i = 1, \ldots, r \text{ ja } j = 1, \ldots, c, \quad (13)$$

missä

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

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

- Todennäköisyyksien (π_i) suurimman uskottavuuden estimaateiksi saadaan

$$\hat{\pi}_j = \frac{n_{\bullet j}}{n}, \qquad j = 1, \ldots, c, \quad (14)$$

jotka ovat y-arvojen suhteelliset frekvenssit laskettuna kaikista ositteista.

Homegeenisuuden testaaminen käytännössä

- Testissä muodostetaan taulukko, jossa vaakariville i tulee ositteesta i lasketut frekvenssit, ja vaakarivin i frekvenssien summa $n_{i\bullet}$ on ositteesta i tehdyn otoksen koko n_i.
- Osoittautuu, että tämän jälkeen testaus voidaan tehdä aivan samalla tavalla kuin edellisessä jaksossa (riippumattomuuden testaaminen).
- Pearsonin testisuuretta verrataan χ^2_ν-jakaumaan, jossa vapausasteluku on

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

8.4 Uskottavuusosamäärän testisuure

- Tämän luvun tilanteissa käytetään usein myös muita testisuureita kuin Pearsonin testisuuretta.
- Olkoon uskottavuusfunktio $L(\boldsymbol{\theta}; \mathbf{Y})$, logaritminen uskottavuusfunktio $\ell(\boldsymbol{\theta}; \mathbf{Y})$ ja suurimman uskottavuuden estimaattori $\hat{\boldsymbol{\theta}}(\mathbf{Y})$.
- Asymptoottisen teorian perusteella tiedetään, että jos havaintosatunnaisvektorilla $\mathbf{Y}$ on parametria $\boldsymbol{\theta}$ vastaava jakauma, niin tällöin *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 osapuilleen χ^2-jakauma, jossa vapausasteluku on yhtä kuin mallin vapaiden parametrien lukumäärä.

Uskottavuusosamäärän testisuure multinomikokeessa

- Oletamme nyt, että luokkien todennäköisyydet $p_1, \ldots, p_{k-1}$ ovat vapaita parametreja.
- Uskottavuusosamäärän testisuureessa tarvittavat log-uskottavuusarvot ovat kaavan (10) mukaan

$$\ell(\hat{p}_1, \ldots, \hat{p}_{k-1}; \mathbf{y}) = \sum_{i=1}^{k} n_i \log \hat{p}_i,$$

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

- SU-estimaatit ovat vastaavat suhteelliset frekvenssit

$$\hat{p}_1 = \frac{n_1}{n}, \ldots, \hat{p}_{k-1} = \frac{n_{k-1}}{n}, \hat{p}_k = \frac{n_k}{n},$$

Uskottavuusosamäärän testisuure

- Uskottavuusosamäärän testisuure on

$$W = 2 \sum_{i=1}^{k} n_i \log \frac{\hat{p}_i}{p_i} = 2 \sum_{i=1}^{k} n_i \log \frac{n_i}{n\, p_i} \tag{15}$$

$$= 2 \sum_{i=1}^{k} O_i \log \frac{O_i}{E_i}, \tag{16}$$

  Jos jokin n_i tai O_i on nolla, niin tässä kaavassa pitää tulkita $0 \log 0 = 0$.

- Suurella otoskoolla testisuuretta vastaavalla satunnaismuuttujalla on osapuilleen χ^2_{k-1}-jakauma.

Uskottavuusosamäärän testisuure muissa tilanteissa

- Myös muissa tämän kappaleen tilanteissa voidaan myös käyttää uskottavuusosamäärän testisuuretta, mutta tällöin tarvitaan sisäkkäisten mallien vertailuun tarkoitettua uskottavuusosamäärän testisuuretta.
- Se on kaikissa tämän kappaleen tilanteissa muotoa

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

tai kaksiulotteisille taulukoille muotoa

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

jossa odotetut frekvenssit lasketaan samoin kuin Pearsonin testisuureelle.

- Uskottavuusosamäärän testisuuretta verrataan täsmälleen samaan χ^2-jakaumaan kuin Pearsonin testisuuretta.

8.5 Suurimman uskottavuuden estimaatit

Nyt johdamme tässä kappaleessa ilmoitetut suurimman uskottavuuden estimaattien kaavat. Ne olisi mahdollista johtaa monella tavalla.

- Voisimme eliminoida yhden todennäköisyysparametereista käyttämällä sitä tietoa, että ne summautuvat ykköseksi.
- Toinen mahdollisuus olisi käyttää Lagrangen keinoa rajoitteellisten optimointitehtävien ratkaisemiseksi.
- Nyt tulokset kuitenkin johdetaan käyttämällä juuri tähän tilanteeseen sopivaa epäyhtälöä.

Tällä konstilla johdoista tulee paljon yksinkertaisempia kuin yleisemmillä keinoilla.

Epäyhtälö logaritmifunktiolle

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

$$\log(x) \leq x - 1, \qquad \text{kaikilla } x > 0. \tag{17}$$

Yhtäsuuruus saavutetaan ainoastaan pisteessä $x = 1$. Tämän väitteen voi tarkistaa helposti analyysin keinoilla.

Apulause

Jos k ei-negatiivista lukua $n_i \geq 0$ summautuu 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 $0 \log 0 = 0$.

Apulauseen 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 $log(x) \leq x - 1$.

Jakson 8.1 SU-estimaatit apulauseen avulla

- Jakson 8.1 alun tilanteessa luokkien todennäköisyydet ovat tuntemattomia parametreja.
- Suurimman uskottavuuden estimaatit saadaan suoraan apulauseesta, sillä

$$\ell(p_1, \ldots, p_{k-1}) = \sum_{i=1}^{k} n_i \log(p_i)$$
$$\leq \sum_{i=1}^{k} n_i \log \frac{n_i}{n} = \sum_{i=1}^{k} n_i \log \hat{p}_i,$$

  jossa SU-estimaatit ovat

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

- Näitä kaavoja käytettiin jaksossa 8.4, kun johdettiin uskottavuusosamäärän testisuure.

Jakson 8.2 SU-estimaatit

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

$$\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \, \log p_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \, \log(\gamma_i \, \delta_j) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \, [\log \gamma_i + \log \delta_j]$$

$$= \sum_{i=1}^{r} \log \gamma_i \sum_{j=1}^{r} n_{ij} + \sum_{j=1}^{c} \log \delta_j \sum_{i=1}^{r} n_{ij}$$

$$= \sum_{i=1}^{r} n_{i\bullet} \log \gamma_i + \sum_{j=1}^{c} n_{\bullet j} \log \delta_j$$

$$\leq \sum_{i=1}^{r} n_{i\bullet} \log \frac{n_{i\bullet}}{n} + \sum_{j=1}^{c} n_{\bullet j} \log \frac{n_{\bullet j}}{n},$$

mistä nähdään, että suurimman uskottavuuden estimaateilla on kaavat (11) ja (12).

Jakson 8.3 SU-estimaatit

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

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

joten suurimman uskottavuuden estimaattien kaavat (14) saadaan suoraan apulauseesta, kun ensin siirrytään tarkastelemaan uskottavuusfunktion logaritmia.