
Mat-1.2620 Sovellettu todennäköisyyslaskenta B**11. harjoitukset / Ratkaisut**

Aiheet: Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen
Tilastollinen riippuvuus ja korrelaatio

Avainsanat:

Havaittu frekvenssi, Heterogeenisuus, Hylkäysalue, Hyväksymisalue, Homogeenisuus, Jakaumaoletus, Keskihajonta, χ^2 -jakauma, χ^2 -homogeenisuustesti, χ^2 -riippumattomuustesti, χ^2 -yhteensopivuustesti, Kaksi-ulotteinen normaalijakauma, Korrelaatio, Korreloimattomuus, Kovarianssi, Kriittinen arvo, Merkitsevyystaso, Momenttimenetelmä, Nollahypoteesi, Normaali-jakauma, Odotettu frekvenssi, Otos, Parametri, p -arvo, Riippumattomuus, Riippuvuus, Sovite, Suurimman uskottavuuden menetelmä, Testi, Vaihtoehtoinen hypoteesi, Vapausasteet, Yhteensopivuus, z -muunnos

Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen **χ^2 -yhteensopivuustesti****Testausasetelma χ^2 -yhteensopivuustestissä**

Tarkastellaan tutkimusasetelmaa, jossa perusjoukon S alkioita kuvataan *faktorilla* eli *tekijällä* A , joka saa olla *laatuero-*, *järjestys-*, *välimatka-* tai *suhdeasteikollinen muuttuja*.

Oletetaan, että perusjoukosta S on poimittu *yksinkertainen satunnaisotos*. Haluamme *testata jakaumaoletusta*, jonka mukaan tekijä A *noudattaa perusjoukossa S jotakin määrättyä todennäköisyysjakaumaa*.

Testattava oletus voidaan muotoilla seuraavilla tavoilla:

- (i) Voidaanko havaintojen jakaumaa *kuvata* oletuksen määrittelemällä todennäköisyysjakaumalla?
- (ii) Voiko otos olla oletuksen määrittelemän todennäköisyysjakauman *generoima* eli *tuottama*?

Yhteensopivuustestissä tutkitaan ovatko otos ja tehty jakaumaoletus *yhteensopivia*.

χ^2 -yhteensopivuustestissä havaintojen ja havaintojen jakaumasta tehdyn oletuksen *yhteensopivuutta mitataan* seuraavalla tavalla:

- (1) Valitaan havainnoille sopiva *luokitus*.
- (2) Määrätään *havaintojen luokkafrekvenssit*.
- (3) Määrätään tehdyn jakaumaoletuksen mukaiset *odotetut luokkafrekvenssit*.
- (4) *Verrataan* havaittuja ja odotettuja luokkafrekvenssejä toisiinsa χ^2 -testisuureella.

Hypoteesit χ^2 -yhteensopivuustestissä

Yleinen hypoteesi H :

Havainnot $X_1, X_2, \dots, X_n$ on saatu poimimalla *yksinkertainen satunnaisotos* perusjoukosta S .

Nollahypoteesi H_0 :

Havainnot $X_1, X_2, \dots, X_n$ noudattavat *todennäköisyysjakaumaa* $f(x; \theta)$, jonka parametrit eivät välttämättä ole tunnettuja.

Vaihtoehtoinen hypoteesi H_1 :

Havainnot $X_1, X_2, \dots, X_n$ *eivät noudata* nollahypoteesin H_0 määrittelemää todennäköisyysjakaumaa.

Havaitut luokkafrekvenssit

Luokitellaan havainnot $X_1, X_2, \dots, X_n$ *toisensa poissulkeviin luokkiin*, joiden lukumäärä olkoon m .

Olkoon

$$O_k, k = 1, 2, \dots, m$$

niiden havaintojen *frekvenssi* eli *lukumäärä* jotka kuuluvat luokkaan k . Frekvenssi O_k on luokkaan k kuuluvien havaintojen **havaittu frekvenssi**.

Oletetaan, että havainnot $X_1, X_2, \dots, X_n$ ovat *diskreetin* satunnaismuuttujan X havaittuja arvoja ja, että satunnaismuuttujan X *mahdolliset arvot* ovat

$$y_1, y_2, \dots, y_m$$

Luokitellaan havainto X_i luokkaan k , jos

$$X_i = y_k, i = 1, 2, \dots, n, k = 1, 2, \dots, m$$

Luokkaan k kuuluvien havaintojen X_i *havaittu frekvenssi* O_k on niiden havaintojen *lukumäärä*, jotka saavat arvon y_k .

Oletetaan, että havainnot $X_1, X_2, \dots, X_n$ ovat *jatkuvan* satunnaismuuttujan X havaittuja arvoja ja, että

$$X \in (a, b)$$

Jaetaan väli (a, b) pisteillä

$$a = a_0 < a_1 < a_2 < \dots < a_{m-1} < a_m = b$$

pistevieraisiin *osaväleihin*

$$(a_{k-1}, a_k], k = 1, 2, \dots, m$$

Luokitellaan havainto X_i luokkaan k , jos

$$X_i \in (a_{k-1}, a_k], i = 1, 2, \dots, n, k = 1, 2, \dots, m$$

Luokkaan k kuuluvien havaintojen X_i *havaittu frekvenssi* O_k on niiden havaintojen *lukumäärä*, jotka kuuluvat väliin k .

Havaitut luokkafrekvenssit O_k voidaan esittää *frekvenssitaulukkona* seuraavassa muodossa:

Luokka	1	2	...	m	Summa
Havaittu frekvenssi	O_1	O_2	...	O_m	n

Frekvenssiä O_k kutsutaan *havaituksi solufrekvenssiksi* *frekvenssitaulukon* solussa k .

Havaitut solufrekvenssit O_k toteuttavat yhtälön

$$\sum_{k=1}^m O_k = n$$

Odotetut luokkafrekvenssit

Oletetaan, että havainnot $X_1, X_2, \dots, X_n$ ovat satunnaismuuttujan X havaittuja arvoja ja, että havainnot on *luokiteltu* toisensa poissulkeviin luokkiin, joiden lukumäärä on m .

Oletetaan, että *nollahypoteesi* H_0 *täysin määrää satunnaismuuttujan* X *jakauman*.

Olkoon P_k *todennäköisyys* sille, että satunnaismuuttuja X saa arvon luokasta k , kun *nollahypoteesi* H_0 *pätee*.

Tällöin luokkaan k kuuluvien havaintojen **odotettu frekvenssi** E_k on

$$E_k = nP_k, k = 1, 2, \dots, m$$

Oletetaan, että *nollahypoteesi* H_0 *määrää satunnaismuuttujan* X *jakauman tyypin, mutta jakauman parametrit ovat tuntemattomia*.

Jos jakauman parametreja ei tunneta, jakaumasta ei voida määrätä todennäköisyyksiä, ellei jakauman parametreja ensin *estimoida* havainnoista.

Olkoon P_k tällöin *estimoitu todennäköisyys* sille, että satunnaismuuttuja X saa arvon luokasta k , kun *nollahypoteesi* H_0 *pätee*.

Tällöin luokkaan k kuuluvien havaintojen **odotettu frekvenssi** E_k on

$$E_k = nP_k, k = 1, 2, \dots, m$$

Oletetaan, että X on *diskreetti* satunnaismuuttuja, jonka mahdolliset arvot ovat

$$y_1, y_2, \dots, y_m$$

Tällöin

$$P_k = \Pr(X = y_k), k = 1, 2, \dots, m$$

jossa todennäköisyys $\Pr(X = y_k)$ määrätään olettaen, että *nollahypoteesi* H_0 *pätee*.

Todennäköisyydet $\Pr(X = y_k)$ voidaan määrätä satunnaismuuttujan X *kertymäfunktion* tai *piste-todennäköisyysfunktion* avulla.

Oletetaan, että X on *jatkuva* satunnaismuuttuja, joka saa arvoja väliltä (a,b) ja, että väli (a,b) on jaettu pisteillä

$$a = a_0 < a_1 < a_2 < \dots < a_{m-1} < a_m = b$$

pistevieraisiin osaväleihin

$$(a_{k-1}, a_k], k = 1, 2, \dots, m$$

Tällöin

$$P_k = \Pr(a_{k-1} < X \leq a_k), k = 1, 2, \dots, m$$

jossa todennäköisyys $\Pr(a_{k-1} < X \leq a_k)$ määrätään olettaen, että *nollahypoteesi* H_0 pätee.

Todennäköisyydet $\Pr(a_{k-1} < X \leq a_k)$ voidaan määrätä satunnaismuuttujan X *kertymäfunktion* tai *tiheysfunktion* avulla.

Odotetut frekvenssit E_k voidaan esittää *frekvenssitaulukkona* seuraavassa muodossa:

Luokka	1	2	...	m	Summa
Odotettu frekvenssi	E_1	E_2	...	E_m	n

Frekvenssiä E_k kutsutaan *odotetuksi solufrekvenssiksi* *frekvenssitaulukon* solussa k .

Odotetut solufrekvenssit E_k toteuttavat yhtälön

$$\sum_{k=1}^m E_k = n$$

Testisuure χ^2 -yhteensopivuustestissä

Testi nollahypoteesille H_0 *perustuu* havaittujen frekvenssien O_k ja odotettujen frekvenssien E_k *vertaamiseen*. Jos havaittujen frekvenssien O_k ja odotettujen frekvenssien E_k *jakaumat muistuttavat toisiaan*, havainnot ovat *sopuoinnussa* nollahypoteesin H_0 kanssa.

Määritellään χ^2 -testisuure

$$\chi^2 = \sum_{k=1}^m \frac{(O_k - E_k)^2}{E_k}$$

jossa

O_k = *havaittu frekvenssi* luokassa k

E_k = *odotettu frekvenssi* luokassa k

m = luokkien lukumäärä

Testisuure χ^2 *mittaa* havaittujen ja odotettujen frekvenssien *jakaumien yhteensopivuutta* tai *etäisyyttä* ja siksi sitä kutsutaan usein χ^2 -*etäisyydeksi*.

χ^2 -testisuure voidaan kirjoittaa myös muotoon

$$\chi^2 = n \sum_{k=1}^m \frac{(\hat{p}_k - P_k)^2}{P_k}$$

jossa

$\hat{p}_k = O_k/n = \text{havaittu suhteellinen frekvenssi luokassa } k$

$P_k = \text{todennäköisyys, että havainto kuuluu luokkaan } k, \\ \text{kun nollahypoteesi } H_0 \text{ pätee}$

$m = \text{luokkien lukumäärä}$

Jos nollahypoteesi H_0 pätee, testisuure χ^2 noudattaa suurissa otoksissa approksimatiivisesti χ^2 -jakaumaa vapausastein $f = (m - 1 - p)$:

$$\chi^2 \sim_a \chi^2(f)$$

jossa

$$f = m - 1 - p$$

$m = \text{luokkien lukumäärä}$

$p = \text{odotettujen frekvenssien } E_k \text{ määrittämiseksi estimoitujen parametrien lukumäärä}$

Approksimaatio on tavallisesti riittävän hyvä, jos odotetut frekvenssit E_k toteuttavat ehdot

$$E_k > 5, k = 1, 2, \dots, m$$

Ehdot saadaan toteutumaan valitsemalla havainnoille sopiva luokitus.

Testisuureen χ^2 normaaliarvo eli odotusarvo nollahypoteesin H_0 pätiessä on

$$E(\chi^2) = f$$

jossa

$$f = m - 1 - p$$

Normaaliarvoaan merkitsevästi suuremmat χ^2 -testisuureen arvot viittaavat siihen, että nollahypoteesi H_0 ei päde. Normaaliarvoaan merkitsevästi pienemmät χ^2 -testisuureen arvot viittaavat siihen, että nollahypoteesi H_0 pätee liian hyvin: Havainnot saattavat olla väärennetyjä.

Homogeenisuuden testaaminen

Testausasetelma χ^2 -homogeenisuustestissä

Tarkastellaan tutkimusasetelmaa, jossa perusjoukon S alkioita kuvataan *yhellä faktorilla* eli tekijällä A , joka saa olla *laatuero-, järjestys-, välimatka- tai suhteasteikollinen muuttuja*.

Jaetaan perusjoukko S jakaa *kahteen tai useampaan ryhmään*, poimitaan ryhmistä *toisistaan riippumattomat yksinkertaiset satunnaisotokset* ja tarkastellaan tekijän A vaihtelua otoksissa.

Tehdään oletus, että *tekijä A noudattaa kaikissa ryhmissä samaa*, tarkemmin määrittelemätöntä *todennäköisyysjakaumaa*.

Haluamme *testata* tehtyä *jakaumaoletusta*:

- (i) Voidaanko eri otoksista (ryhmistä) saatuja havaintoarvojen jakaumia kuvata *samalla todennäköisyysjakaumalla*?
- (ii) Voivatko otokset olla *saman todennäköisyysjakauman generoimia* eli *tuottamia*?

Jos tehty jakaumaoletus *pätee* ja *tekijä A noudattaa siis kaikissa ryhmissä samaa jakaumaa*, perusjoukko on **homogeeninen** ja *perusjoukkoa ei tarvitse jakaa tekijää A koskevissa tarkasteluissa erillisiksi ryhmiksi*.

Jos tehty jakaumaoletus *ei päde* ja *tekijä A noudattaa siis eri ryhmissä eri jakaumia*, perusjoukko on **heterogeeninen** ja *ryhmiä on syytä tarkastella tekijää A koskevissa tarkasteluissa erillisinä*.

Tällaisten jakaumaoletuksen testaamiseen tarkoitettuja testejä kutsutaan **homogeenisuustesteiksi**.

χ^2 -homogeenisuustestin suorittaminen

χ^2 -homogeenisuustestissä havaintojen ja niiden jakaumasta eri ryhmissä tehdyn homogeenisuusoletuksen *yhteensopivuutta* mitataan seuraavalla tavalla:

- (1) Valitaan havainnoille *yhteinen luokitus*.
- (2) Määrätään *havaintojen luokkafrekvenssit* jokaisesta otoksesta.
- (3) Määrätään tehdyn homogeenisuusoletuksen mukaiset *odotetut luokkafrekvenssit*.
- (4) *Verrataan* havaittuja ja odotettuja luokkafrekvenssejä toisiinsa χ^2 -testisuureella.

Hypoteesit χ^2 -homogeenisuustestissä

Yleinen hypoteesi H :

Perusjoukosta voidaan jakaa r ryhmään, joista on poimittu (toisistaan riippumattomat) *yksinkertaiset satunnaisotokset*.

Nollahypoteesi H_0 :

Otokset $i = 1, 2, \dots, r$ on poimittu *samasta todennäköisyysjakaumasta*.

Vaihtoehtoinen hypoteesi H_1 :

Otokset $i = 1, 2, \dots, r$ on poimittu *eri todennäköisyysjakaumista*.

Havaitut frekvenssit

Oletetaan, että tutkimuksen kohteena oleva perusjoukko S on jaettu r ryhmään.

Poimitaan ryhmistä toisistaan riippumattomat yksinkertaiset satunnaisotokset, joiden koot ovat

$$n_i, i = 1, 2, \dots, r.$$

Luokitellaan havainnot jokaisessa otoksessa samaa luokitusta käyttäen toisensa poissulkeviin luokkiin, joiden lukumäärä olkoon c ja määrätään ryhmän i luokkaan j kuuluvien havaintojen **havaittu frekvenssi** eli *lukumäärä*

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

Muodostetaan havaituista frekvensseistä O_{ij} ($r \times c$)-frekvenssitaulukko $[O_{ij}]$:

		Luokat				
Ryhmät		1	2	...	c	Summa
	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
	Summa	C_1	C_2	...	C_c	n

Taulukossa:

r = ryhmien lukumäärä

c = luokkien lukumäärä

O_{ij} = havaittu frekvenssi ryhmän i luokassa j , $i = 1, 2, \dots, r$, $j = 1, 2, \dots, c$

n_i = otoskoko ryhmässä i

C_j = havaittu frekvenssi yhdistetyn havaintoaineiston luokassa j

n = havaintojen kokonaislukumäärä

Frekvenssiä O_{ij} kutsutaan tavallisesti **havaituksi solufrekvenssiksi** frekvenssitaulun solussa (i, j) .

Havaittujen frekvenssien O_{ij} frekvenssitaulukossa pätee:

(i) *Rivisummat* yhtyvät ryhmäkohtaisiin otoskokoihin:

$$\sum_{j=1}^c O_{ij} = n_i, i = 1, 2, \dots, r$$

(ii) *Sarakesummat* yhtyvät yhdistetyn havaintoaineiston luokkafrekvensseihin:

$$\sum_{i=1}^r O_{ij} = C_j, j = 1, 2, \dots, c$$

(iii) *Havaintojen kokonaislukumäärä:*

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

Nollahypoteesin tulkinta χ^2 -homogeenisuustestissä

Jos nollahypoteesi H_0 pätee, havaintojen pitää jakautua (satunnaisvaihtelua lukuun ottamatta) jokaisessa ryhmässä $i = 1, 2, \dots, r$ samalla tavalla luokkiin $j = 1, 2, \dots, c$.

Jos nollahypoteesi H_0 pätee, havaintojen jakautuminen luokkiin $j = 1, 2, \dots, c$ ei saa riippua siitä, mihin ryhmään $i = 1, 2, \dots, r$ ne kuuluvat.

Jos nollahypoteesi H_0 pätee, todennäköisyys, että havainto kuuluu luokkaan $j = 1, 2, \dots, c$ ei saa riippua siitä, mihin ryhmään $i = 1, 2, \dots, r$ se kuuluu.

Odotettujen frekvenssien määrittäminen

Olkoon x on tarkastelun kohteena olevan perusjoukon S alkio.

Määritellään seuraavat todennäköisyydet:

$$\begin{aligned} p_{ij} &= \Pr(x \text{ kuuluu ryhmään } i \text{ ja luokkaan } j) \\ p_{j|i} &= \Pr(x \text{ kuuluu luokkaan } j \mid x \text{ kuuluu ryhmään } i) \\ p_{i\cdot} &= \Pr(x \text{ kuuluu ryhmään } i) \\ p_{\cdot j} &= \Pr(x \text{ kuuluu luokkaan } j) \\ i &= 1, 2, \dots, r, \quad j = 1, 2, \dots, c \end{aligned}$$

Todennäköisyyslaskennan yleisen tulosäännön mukaan

$$p_{ij} = p_{j|i} p_{i\cdot}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

Jos nollahypoteesi H_0 pätee, todennäköisyys, että perusjoukon S alkio x kuuluu luokkaan j ei saa riippua siitä, mihin ryhmään i alkio x kuuluu.

Siten nollahypoteesi H_0 perusjoukon homogeenisuudesta voidaan ilmaista muodossa

$$H_0 : p_{j|i} = p_{\cdot j}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

tai muodossa

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

Todennäköisyydet p_{ij} , $p_{i\cdot}$, $p_{\cdot j}$ voidaan estimoida havaituista frekvensseistä O_{ij} kaavoilla

$$\hat{p}_{ij} = \frac{O_{ij}}{n} \quad \hat{p}_{i\cdot} = \frac{n_i}{n} \quad \hat{p}_{\cdot j} = \frac{C_j}{n}$$

Jos nollahypoteesi H_0 pätee, solutodennäköisyydet p_{ij} voidaan estimoida kaavoilla

$$P_{ij} = \frac{n_i}{n} \times \frac{C_j}{n} = \hat{p}_{i\cdot} \hat{p}_{\cdot j}$$

Määrätään *nollahypoteesin* H_0 *pätiessä odotetut solufrekvenssit* E_{ij} yhtälöillä

$$E_{ij} = nP_{ij} = \frac{C_j}{n} \times n_i = \frac{n_i C_j}{n}, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

Muodostetaan *odotetuista frekvensseistä* E_{ij} $(r \times c)$ -*frekvenssitaulukko* $[E_{ij}]$:

		Luokat				
Ryhvät		1	2	...	c	Summa
	1	E_{11}	E_{12}	...	E_{1c}	n_1
	2	E_{21}	E_{22}	...	E_{2c}	n_2
	...	...	...	...	...	...
	r	E_{r1}	E_{r2}	...	E_{rc}	n_r
	Summa	C_1	C_2	...	C_c	n

Taulukossa

r = ryhmien lukumäärä

c = luokkien lukumäärä

E_{ij} = *odotettu frekvenssi* ryhmän i luokassa j , $i = 1, 2, \dots, r, j = 1, 2, \dots, c$

n_i = *otoskoko* ryhmässä i

C_j = *havaittu frekvenssi* yhdistetyn havaintoaineiston luokassa j

n = *havaintojen kokonaislukumäärä*

Frekvenssiä E_{ij} kutsutaan tavallisesti *odotetuksi solufrekvenssiksi frekvenssitaulun solussa* (i, j) .

Odotettujen frekvenssien E_{ij} *frekvenssitaulukossa* pätee:

(i) *Rivisummat* yhtyvät ryhmäkohtaisiin otoskokoihin:

$$\sum_{j=1}^c E_{ij} = n_i, i = 1, 2, \dots, r$$

(ii) *Sarakesummat* yhtyvät yhdistetyn havaintoaineiston luokkafrekvensseihin:

$$\sum_{i=1}^r E_{ij} = C_j, j = 1, 2, \dots, c$$

(iii) *Havaintojen kokonaislukumäärä*:

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

Testisuure χ^2 -homogeenisuustestissä

Määritellään χ^2 -testisuure

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

jossa

O_{ij} = havaittu frekvenssi solussa (i, j)

E_{ij} = odotettu frekvenssi solussa (i, j)

r = ryhmien lukumäärä

c = lukkien lukumäärä

Testisuure χ^2 mittaa havaittujen ja odotettujen frekvenssien *jakaumien yhteensopivuutta* tai *etäisyyttä* ja siksi sitä kutsutaan usein χ^2 -etäisyydeksi.

Homogeenisuustestin χ^2 -testisuure voidaan kirjoittaa myös muotoon

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^c \frac{(\hat{p}_{ij} - P_{ij})^2}{P_{ij}}$$

jossa

$$\hat{p}_{ij} = \frac{O_{ij}}{n}$$

$$P_{ij} = \frac{E_{ij}}{n} = \frac{n_i}{n} \times \frac{C_j}{n} = \hat{p}_{i\cdot} \hat{p}_{\cdot j}$$

$$\hat{p}_{i\cdot} = \frac{n_i}{n}$$

$$\hat{p}_{\cdot j} = \frac{C_j}{n}$$

$$i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

Jos nollahypoteesi H_0 pätee, testisuure χ^2 noudattaa suurissa otoksissa approksimatiivisesti χ^2 -jakaumaa vapausastein $f = (r - 1)(c - 1)$:

$$\chi^2 \sim_a \chi^2(f)$$

jossa

r = ryhmien lukumäärä

c = luokkien lukumäärä

Approksimaatio on tavallisesti riittävän hyvä, jos odotetut frekvenssit E_{ij} ja keskimääräiset odotetut frekvenssit C_j/r toteuttavat ehdot

$$E_{ij} > 1, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

$$C_j / r > 5, j = 1, 2, \dots, c$$

Ehdot saadaan toteutumaan valitsemalla havainnoille *sopiva luokitus*.

Testisuureen χ^2 *normaaliarvo* eli *odotusarvo nollahypoteesin* H_0 *pätiessä* on

$$E(\chi^2) = f$$

jossa

$$f = (r - 1)(c - 1)$$

Normaaliarvoaan *merkittävästi suuremmat* χ^2 -testisuureen arvot viittaavat siihen, että nollahypoteesi H_0 *ei päde*.

Riippumattomuuden testaaminen

Testausasetelma χ^2 -riippumattomuustestissä

Tarkastellaan tutkimusasetelmaa, jossa perusjoukon S alkioita kuvataan *kahdella faktorilla* eli tekijällä A ja B , jotka saavat olla *laatuero-, järjestys-, välimatka- tai suhdeasteikollisia muuttujia*.

Poimitaan perusjoukosta S *yksinkertainen satunnaisotos* ja tarkastellaan tekijöiden A ja B vaihtelua otoksessa.

Tehdään oletus, että *tekijät A ja B ovat riippumattomia*.

Haluamme *testata* tehtyä *riippumattomuusoletusta*:

Ovatko havainnot *sopusoinnussa* tehdyn riippumattomuusoletuksen kanssa?

Jos tehty oletus *pätee* ja *tekijät A ja B ovat siis riippumattomia*, *tekijöitä A ja B voidaan tarkastella erillisinä*. Jos tehty oletus *ei päde* ja *tekijät A ja B eivät siis ole riippumattomia*, *tekijät A ja B ovat assosioituneita*.

Riippumattomuusoletuksen testaamiseen tarkoitettuja testejä kutsutaan **riippumattomuustesteiksi**.

χ^2 -riippumattomuustestin suorittaminen

χ^2 -riippumattomuustestissä havaintojen ja tehdyn riippumattomuusoletuksen *yhteensopivuutta* mitataan seuraavalla tavalla:

- (1) Valitaan havainnoille sopivat *luokitukset* tekijöiden A ja B suhteen.
- (2) Luokitellaan havainnot tekijöiden A ja B suhteen *ristiin* ja määrätään *havaitut luokkafrekvenssit*.
- (3) Määrätään tehdyn riippumattomuusoletuksen mukaiset *odotetut luokkafrekvenssit*.
- (4) *Verrataan* havaittuja ja odotettuja luokkafrekvenssejä toisiinsa χ^2 -testisuureella.

Hypoteesit χ^2 -riippumattomuustestissä

Yleinen hypoteesi H :

Perusjoukosta on poimittu *yksinkertainen satunnaisotos* ja havaintoyksiköt on voitu *luokitella ristiin kahden tekijän A ja B suhteen*.

Nollahypoteesi H_0 :

Tekijät A ja B ovat riippumattomia.

Vaihtoehtoinen hypoteesi H_1 :

Tekijät A ja B eivät ole riippumattomia.

Havaitut frekvenssit

Poimitaan tutkimuksen kohteena olevasta perusjoukosta S *yksinkertainen satunnaisotos*, jonka koko on n .

Luokitellaan havaintoyksiköt *tekijän A suhteen* toisensa poissulkeviin luokkiin, joiden lukumäärä on r . Määrätään tekijän A luokkaan i kuuluvien havaintojen *havaittu frekvenssi* eli lukumäärä R_i , kun $i = 1, 2, \dots, r$.

*Luokitellaan havaintoyksiköt tekijän B suhteen toisensa poissulkeviin luokkiin, joiden lukumäärä on c . Määrätään tekijän B luokkaan j kuuluvien havaintojen *havaittu frekvenssi* eli lukumäärä C_j , kun $j = 1, 2, \dots, c$.*

*Luokitellaan havaintoyksiköt tekijöiden A ja B suhteen ristiin toisensa poissulkeviin luokkiin, joiden lukumäärä on $r \times c$. Määrätään tekijän A luokkaan i ja tekijän B luokkaan j kuuluvien havaintojen **havaittu frekvenssi** eli *lukumäärä* O_{ij} , kun $i = 1, 2, \dots, r$ ja $j = 1, 2, \dots, c$.*

Muodostetaan *havaituista frekvensseistä* O_{ij} ($r \times c$)-*frekvenssitaulukko* [O_{ij}] :

		<i>B</i> -luokat				
		1	2	...	<i>c</i>	Summa
<i>A</i> -luokat	1	O_{11}	O_{12}	...	O_{1c}	R_1
	2	O_{21}	O_{22}	...	O_{2c}	R_2
	...	...	...	...	...	...
	<i>r</i>	O_{r1}	O_{r2}	...	O_{rc}	R_r
	Summa	C_1	C_2	...	C_c	<i>n</i>

Taulukossa:

r = A -luokkien lukumäärä

c = B -luokkien lukumäärä

O_{ij} = *havaittu frekvenssi* luokassa, jonka määrää A -luokka i ja B -luokka j ,
 $i = 1, 2, \dots, r, j = 1, 2, \dots, c$

R_i = *havaittu frekvenssi* A -luokassa i

C_j = *havaittu frekvenssi* B -luokassa j

n = havaintojen kokonaislukumäärä

Frekvenssiä O_{ij} kutsutaan tavallisesti *havaituksi solufrekvenssiksi frekvenssitaulun solussa* (i, j) .

Havaittujen frekvenssien O_{ij} *frekvenssitaulukossa* pätee:

(i) *Rivisummat* yhtyvät havaittuihin frekvensseihin A -luokituksessa:

$$\sum_{j=1}^c O_{ij} = R_i, i = 1, 2, \dots, r$$

(ii) *Sarakesummat* yhtyvät havaittuihin frekvensseihin B -luokituksessa:

$$\sum_{i=1}^r O_{ij} = C_j, j = 1, 2, \dots, c$$

(iii) *Havaintojen kokonaislukumäärä:*

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

Nollahypoteesin tulkinta χ^2 -riippumattomuustestissä

Jos nollahypoteesi H_0 pätee, havaintojen jakautuminen A -luokkiin ei saa riippua siitä, mihin B -luokkaan havainnot kuuluvat, toisin sanoen, jos nollahypoteesi H_0 pätee, todennäköisyys, että havainto kuuluu A -luokkaan $i = 1, 2, \dots, r$ ei saa riippua siitä, mihin B -luokkaan $j = 1, 2, \dots, c$ se kuuluu.

Jos nollahypoteesi H_0 pätee, havaintojen jakautuminen B -luokkiin ei saa riippua siitä, mihin A -luokkaan havainnot kuuluvat, toisin sanoen, jos nollahypoteesi H_0 pätee, todennäköisyys, että havainto kuuluu B -luokkaan $j = 1, 2, \dots, c$ ei saa riippua siitä, mihin A -luokkaan $i = 1, 2, \dots, r$ se kuuluu.

Odotettujen frekvenssien määrääminen

Olkoon x on tarkastelun kohteena olevan perusjoukon S alkio.

Määritellään seuraavat todennäköisyydet:

$$\begin{aligned} p_{ij} &= \Pr(x \text{ kuuluu } A\text{-luokkaan } i \text{ ja } B\text{-luokkaan } j) \\ p_{i\cdot} &= \Pr(x \text{ kuuluu } A\text{-luokkaan } i) \\ p_{\cdot j} &= \Pr(x \text{ kuuluu } B\text{-luokkaan } j) \\ i &= 1, 2, \dots, r, \quad j = 1, 2, \dots, c \end{aligned}$$

Jos nollahypoteesi H_0 pätee, tapahtumat

$$\begin{aligned} \{x \in S \mid x \text{ kuuluu } A\text{-luokkaa } i\} \\ \{x \in S \mid x \text{ kuuluu } B\text{-luokkaa } j\} \end{aligned}$$

ovat riippumattomia kaikille $i = 1, 2, \dots, r, j = 1, 2, \dots, c$.

Siten nollahypoteesi H_0 tekijöiden A ja B riippumattomuudesta voidaan ilmaista muodossa

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

Todennäköisyydet $p_{ij}, p_{i\cdot}, p_{\cdot j}$ voidaan estimoida havaituista frekvensseistä O_{ij} kaavoilla

$$\hat{p}_{ij} = \frac{O_{ij}}{n} \quad \hat{p}_{i\cdot} = \frac{R_i}{n} \quad \hat{p}_{\cdot j} = \frac{C_j}{n}$$

Jos nollahypoteesi H_0 pätee, solutodennäköisyydet p_{ij} voidaan estimoida kaavoilla

$$P_{ij} = \frac{R_i}{n} \times \frac{C_j}{n} = \hat{p}_{i\cdot} \hat{p}_{\cdot j}$$

Määrätään nollahypoteesin H_0 pätiessä odotetut solufrekvenssit E_{ij} yhtälöillä

$$E_{ij} = n P_{ij} = \frac{R_i C_j}{n}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

Muodostetaan *odotetuista frekvensseistä* E_{ij} ($r \times c$)-frekvenssitaulukko $[E_{ij}]$:

		B-luokat				
		1	2	...	c	Summa
A-luokat	1	E_{11}	E_{12}	...	E_{1c}	R_1
	2	E_{21}	E_{22}	...	E_{2c}	R_2
	...	...	...	...	...	...
	r	E_{r1}	E_{r2}	...	E_{rc}	R_r
	Summa	C_1	C_2	...	C_c	n

Taulukossa

r = A-luokkien lukumäärä

c = B-luokkien lukumäärä

E_{ij} = *odotettu frekvenssi* luokassa, jonka määrää A-luokka i ja B-luokka j ,
 $i = 1, 2, \dots, r, j = 1, 2, \dots, c$

R_i = *havaittu frekvenssi* A-luokassa i

C_j = *havaittu frekvenssi* B-luokassa j

n = havaintojen kokonaislukumäärä

Frekvenssiä E_{ij} kutsutaan tavallisesti *odotetuksi solufrekvenssiksi frekvenssitaulun solussa* (i, j) .

Odotettujen frekvenssien E_{ij} frekvenssitaulukossa pätee:

(i) *Rivisummat* yhtyvät havaittuihin frekvensseihin A-luokituksessa:

$$\sum_{j=1}^c E_{ij} = R_i, i = 1, 2, \dots, r$$

(ii) *Sarakesummat* yhtyvät havaittuihin frekvensseihin B-luokituksessa:

$$\sum_{i=1}^r E_{ij} = C_j, j = 1, 2, \dots, c$$

(iii) *Havaintojen kokonaislukumäärä*:

$$\sum_{i=1}^r \sum_{j=1}^c E_{ij} = \sum_{i=1}^r R_i = \sum_{j=1}^c C_j = n$$

Testisuure χ^2 -riippumattomuustestissä

Määritellään χ^2 -testisuure

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

jossa

O_{ij} = havaittu frekvenssi solussa (i, j)

E_{ij} = odotettu frekvenssi solussa (i, j)

r = A-luokkien lukumäärä

c = B-luokkien lukumäärä

Testisuure χ^2 mittaa havaittujen ja odotettujen frekvenssien *jakaumien yhteensopivuutta* tai *etiäisyyttä* ja siksi sitä kutsutaan usein χ^2 -etiäisyydeksi.

χ^2 -testisuure voidaan kirjoittaa myös muotoon

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^c \frac{(\hat{p}_{ij} - P_{ij})^2}{P_{ij}}$$

jossa

$$\hat{p}_{ij} = \frac{O_{ij}}{n}$$

$$P_{ij} = \frac{E_{ij}}{n} = \frac{R_i}{n} \times \frac{C_j}{n} = \hat{p}_{i\cdot} \hat{p}_{\cdot j}$$

$$\hat{p}_{i\cdot} = \frac{R_i}{n}$$

$$\hat{p}_{\cdot j} = \frac{C_j}{n}$$

$$i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

Jos nollahypoteesi H_0 pätee, testisuure χ^2 noudattaa suurissa otoksissa *approksimatiivisesti* χ^2 -jakaumaa vapausastein $f = (r - 1)(c - 1)$:

$$\chi^2 \sim_a \chi^2(f)$$

jossa

r = A-luokkien lukumäärä

c = B-luokkien lukumäärä

Approksimaatio on tavallisesti *riittävän hyvä*, jos odotetut frekvenssit E_{ij} ja keskimääräiset odotetut frekvenssit R_i/c ja C_j/r toteuttavat ehdot

$$E_{ij} > 1, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

$$R_i / c > 5, i = 1, 2, \dots, r$$

$$C_j / r > 5, j = 1, 2, \dots, c$$

Ehdot saadaan toteutumaan valitsemalla havainnoille *sopiva luokitus*.

Testisuureen χ^2 normaaliarvo eli odotusarvo nollahypoteesin H_0 pätiessä on

$$E(\chi^2) = f$$

jossa

$$f = (r - 1)(c - 1)$$

Normaaliarvoaan *merkittävästi suuremmat* χ^2 -testisuureen arvot viittaavat siihen, että nollahypoteesi H_0 *ei päde*.

χ^2 -homogeenisuustestin ja χ^2 -riippumattomuustestin vertailu

χ^2 -homogeenisuustesti ja χ^2 -riippumattomuustesti tehdään teknisesti *täsmälleen samalla tavalla* ja niinpä frekvenssitaulukosta *ei voi sellaisenaan nähdä* kummasta testausasetelmasta on kyse:

- *Odotetut frekvenssit määrätään samalla kaavalla.*
- *Testisuureet lasketaan samalla kaavalla.*
- *Testisuureet noudattavat nollahypoteesin pätiessä approksimatiivisesti samaa jakaumaa.*

Testien testausasetelmat ovat kuitenkin *täysin erilaiset*.

Homogeenisuustestin testausasetelma:

- (i) Perusjoukko koostuu r ryhmästä ja testissä tarkastellaan perusjoukon alkioiden *jakautumista* luokkiin eri ryhmissä, kun alkiot on luokittelu yhden tekijän A suhteen käyttäen kaikissa ryhmissä *samaa* luokitusta.
- (ii) Havaintoaineisto muodostuu *toisistaan riippumattomista ryhmäkohtaisista satunnais-otoksista*.
- (iii) Sekä ryhmäkohtaiset otoskoot n_i että havaintojen kokonaislukumäärä n ovat *kiinteitä* eli *ei-satunnaisia* (eli valittuja) lukuja, kun taas *sattuma määrää* miten havainnot *jakautuvat* luokkiin ryhmien sisällä.

Riippumattomuustestin testausasetelma:

- (i) Testissä tarkastellaan *kahden tekijän* A ja B *assosiaatiota* eli *riippuvuutta*, kun havainnot on luokiteltu *ristiin* tekijöiden A ja B suhteen.
- (ii) Havaintoaineisto muodostuu *yhdestä satunnaisotoksesta*.
- (iii) Vain havaintojen kokonaislukumäärä n on *kiinteä* eli *ei-satunnainen* (eli valittu) luku, kun taas *sattuma määrää* miten havainnot *jakautuvat* luokkiin *tekijöiden* A ja B *ristiluokituksen suhteen*.

Tilastollinen riippuvuus ja korrelaatio

Otos kaksiulotteisesta normaalijakaumasta

Oletetaan, että satunnaismuuttujien x ja y pari (x, y) noudattaa *kaksiulotteista normaalijakaumaa* eli

$$(x, y) \sim N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$$

jossa

$$\mu_x = E(x)$$

$$\mu_y = E(y)$$

$$\sigma_x^2 = \text{Var}(x) = E[(x - \mu_x)^2]$$

$$\sigma_y^2 = \text{Var}(y) = E[(y - \mu_y)^2]$$

$$\sigma_{xy} = \text{Cov}(x, y) = E[(x - \mu_x)(y - \mu_y)]$$

$$\rho_{xy} = \text{Cor}(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Olko

$$y_1, y_2, \dots, y_n$$

muuttujan y havaitut arvot ja

$$x_1, x_2, \dots, x_n$$

muuttujan x havaitut arvot ja oletetaan, että havaintoarvojen x_i ja y_i parit

$$(x_i, y_i), i = 1, 2, \dots, n$$

muodostavat *yksinkertaisen satunnaisotoksen kaksiulotteista normaalijakaumasta*

$$N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$$

Tällöin

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \perp$$

$$(x_i, y_i) \sim N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}), i = 1, 2, \dots, n$$

Kaksiulotteisen normaalijakauman parametrien estimointi

Kaksiulotteisen normaalijakauman parametrien *momentiestimaattorit* tai *suurimman uskottavuuden estimaattorit* ovat

$$\hat{\mu}_x = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\mu}_y = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\hat{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n} s_x^2$$

$$\hat{\sigma}_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{n-1}{n} s_y^2$$

$$\hat{\sigma}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{n-1}{n} s_{xy}$$

$$\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{s_{xy}}{s_x s_y} = r_{xy}$$

jossa

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Korreloimattomuuden testaaminen

Olkoon *nollahypoteesina*

$$H_0 : \rho_{xy} = 0$$

Määritellään *t-testisuure*

$$t = \sqrt{n-2} \frac{r_{xy}}{\sqrt{1-r_{xy}^2}}$$

Jos nollahypoteesi

$$H_0 : \rho_{xy} = 0$$

pätee, testisuure *t* noudattaa *Studentin t-jakaumaa* vapausastein $(n-2)$:

$$t \sim t(n-2)$$

Testisuureen *t* normaaliarvo = 0, koska *nollahypoteesin pätiessä*

$$E(t) = 0$$

Siten itseisarvoltaan *suuret* testisuureen *t* arvot viittaavat siihen, että *nollahypoteesi* H_0 *ei päde*.

Testi korrelaatiokertoimelle

Olkoon *nollahypoteesina*

$$H_0 : \rho_{xy} = \rho_0$$

Määritellään *Fisherin z-muunnos* kaavalla

$$z = f(u) = \frac{1}{2} \log \left(\frac{1+u}{1-u} \right)$$

Sovelletaan *Fisherin z-muunnosta* $z = f(u)$ otoskorrelaatiokertoimeen r_{xy} :

$$z = f(r_{xy}) = \frac{1}{2} \log \left(\frac{1+r_{xy}}{1-r_{xy}} \right)$$

Voidaan osoittaa, että satunnaismuuttuja z noudattaa suurissa otoksissa *approksimatiivisesti normaalijakaumaa*:

$$z \sim_a N(\mu_z, \sigma_z^2)$$

jossa

$$\mu_z = f(\rho_{xy}) = \frac{1}{2} \log \left(\frac{1 + \rho_{xy}}{1 - \rho_{xy}} \right)$$

$$\sigma_z^2 = \frac{1}{n-3}$$

Approksimaatio on käytännössä *riittävän hyvä*, jos $n > 25$.

Muodostetaan *testisuure*

$$v = \frac{z - \mu_z^0}{\sigma_z}$$

jossa

$$z = f(r_{xy}) = \frac{1}{2} \log \left(\frac{1 + r_{xy}}{1 - r_{xy}} \right)$$

$$\mu_z^0 = f(\rho_0) = \frac{1}{2} \log \left(\frac{1 + \rho_0}{1 - \rho_0} \right)$$

$$\sigma_z^2 = \frac{1}{n-3}$$

Jos *nollahypoteesi* H_0 pätee,

$$v \sim_a N(0,1)$$

Testisuureen v *normaaliarvo* $= 0$, koska *nollahypoteesin* pätiessä

$$E(v) = 0$$

Siten itseisarvoltaan *suuret* testisuureen v arvot viittaavat siihen, että *nollahypoteesi* H_0 *ei päde*.

Tehtävä 11.1.

Kahdelle henkilölle, A ja B, on kummallekin annettu noppa ja kumpaakin on pyydetty heittämään sitä 120 kertaa. A ja B kertovat saaneensa heittojen tuloksena alla esitetyt silmälukujen jakaumat.

- (a) Tutki χ^2 -yhteensopivuustestin avulla voidaanko A:n ja B:n heittämiä noppia pitää virheettöminä eli symmetrisinä? Tämä tapahtuu testaamalla χ^2 -yhteensopivuustestillä nollahypoteesia, että nopanheiton tulos noudattaa diskreettiä tasaista jakaumaa. Tällöin *suuret* χ^2 -testisuureen arvot johtavat nollahypoteesin hylkäämiseen.
- (b) Tutki χ^2 -yhteensopivuustestin avulla, kuinka todennäköistä on se, että A ja B ovat rehellisiä kertoessaan heittäneensä noppaa? Tämä tapahtuu testaamalla χ^2 -yhteensopivuustestillä nollahypoteesia, että nopanheiton tulos noudattaa diskreettiä tasaista jakaumaa. Tällöin *pienet* χ^2 -testisuureen arvot viittaavat siihen, että tulokset ovat ”liian hyviä” ollakseen todellisia.

Käytä (a)-kohdan testeissä sekä 1 %:n että 5 %:n merkitsevyystasoja ja (b)-kohdan testeissä 1 %:n merkitsevyystasoa.

Silmäluku	1	2	3	4	5	6
A:n tulokset	12	16	20	17	22	33
B:n tulokset	19	21	19	21	19	21

Tehtävä 11.1. – Mitä opimme?

Tehtävässä sovelletaan χ^2 -yhteensopivuustestiä.

Tehtävä 11.1. – Ratkaisu:

Kahdelle henkilölle A ja B on kummallekin annettu noppa ja kumpaakin on pyydetty heittämään sitä 120 kertaa. A ja B kertovat saaneensa heittojen tuloksena alla esitetyt silmälukujen jakaumat.

Silmäluku	1	2	3	4	5	6
A:n tulokset	12	16	20	17	22	33
B:n tulokset	19	21	19	21	19	21

Tehtävän *nollahypoteesina* on

$$H_0 : \Pr(\text{Saadaan silmäluku } i) = p = 1/6, i = 1, 2, 3, 4, 5, 6$$

Siten *odotetuiksi frekvensseiksi* saadaan:

$$E_i = np = 20, i = 1, 2, 3, 4, 5, 6$$

A:n käyttämälle nopalle χ^2 -testisuureen arvoksi saadaan

$$\chi^2 = 13.1$$

Laskutoimitukset (Microsoft Excel –ohjelmistolla):

i	O_i	E_i	$(O_i - E_i)^2/E_i$
1	12	20	3.2
2	16	20	0.8
3	20	20	0
4	17	20	0.45
5	22	20	0.2
6	33	20	8.45
Summa	120	120	13.1

B:n käyttämälle nopalle χ^2 -testisuureen arvoksi saadaan

$$\chi^2 = 0.3$$

Laskutoimitukset (Microsoft Excel –ohjelmistolla):

i	O_i	E_i	$(O_i - E_i)^2/E_i$
1	19	20	0.05
2	21	20	0.05
3	19	20	0.05
4	21	20	0.05
5	19	20	0.05
6	21	20	0.05
Summa	120	120	0.3

Koska $k = 6$ ja $m = 0$ (yhtään parametria ei ole estimoitu), χ^2 -testisuureen vapausasteiden lukumäärä on molemmissa tapauksissa

$$f = k - 1 - m = 6 - 1 - 0 = 5$$

χ^2 -jakauman taulukoiden mukaan 1 %:n merkitsevyystasoa vastaavat kriittinen arvo $\chi^2_{0.01}$ saadaan seuraavalla tavalla: Erotetaan todennäköisyysmassa 0.01 jakauman oikealle hännälle.

Koska

$$\Pr(\chi^2 \geq 15.09) = 0.01$$

niin

$$\chi^2_{0.01} = 15.09$$

Etsitään vielä piste $\chi^2_{0.99}$, joka erottaa todennäköisyysmassan 0.01 jakauman vasemmalle hännälle.

Koska

$$\Pr(\chi^2 \geq 0.554) = 0.99$$

niin

$$\Pr(\chi^2 \leq 0.554) = 0.01$$

ja siten

$$\chi_{0.99}^2 = 0.554$$

χ^2 -jakauman taulukoiden mukaan 5 %:n merkitsevyystasoa vastaava *kriittinen arvo* $\chi_{0.05}^2$ saadaan seuraavalla tavalla: Erotetaan todennäköisyysmassa 0.05 jakauman *oikealle* hänälle.

Koska

$$\Pr(\chi^2 \geq 11.07) = 0.05$$

niin

$$\chi_{0.05}^2 = 11.07$$

Etsitään vielä piste $\chi_{0.95}^2$, joka erottaa todennäköisyysmassan 0.05 jakauman *vasemmalle* hänälle.

Koska

$$\Pr(\chi^2 \geq 1.15) = 0.95$$

niin

$$\Pr(\chi^2 \leq 1.15) = 0.05$$

ja siten

$$\chi_{0.95}^2 = 1.15$$

- (a) *Normaaliarvoonsa nähden "liian" suuret χ^2 -testisuureen arvot viittaavat siihen, että havainnot eivät noudata nollahypoteesin kiinnittämää jakaumaa.*

A:n noppa ja 1 %:n merkitsevyystaso:

Koska χ^2 -testisuureen arvo

$$\chi^2 = 13.1 < 15.09 = \chi_{0.01}^2$$

niin *nollahypoteesi jää voimaan* 1 %:n merkitsevyystasolla A:n nopan tapauksessa.

B:n noppa ja 1 %:n merkitsevyystaso:

Koska χ^2 -testisuureen arvo

$$\chi^2 = 0.3 < 15.09 = \chi_{0.01}^2$$

niin *nollahypoteesi jää voimaan* 1 %:n merkitsevyystasolla B:n nopan tapauksessa.

A:n noppa ja 5 %:n merkitsevyystaso:

Koska χ^2 -testisuureen arvo

$$\chi^2 = 13.1 > 11.07 = \chi_{0.05}^2$$

niin *nollahypoteesi hylätään 5 %:n merkitsevyystasolla A:n nopan tapauksessa.*

B:n noppa ja 5 %:n merkitsevyystaso:

Koska χ^2 -testisuureen arvo

$$\chi^2 = 0.3 < 11.07 = \chi_{0.05}^2$$

niin *nollahypoteesi jää voimaan 5 %:n merkitsevyystasolla B:n nopan tapauksessa.*

- (b) *Normaaliarvoonsa nähden "liian" pienet χ^2 -testisuureen arvot viittaavat siihen, että havainnot noudattavat jossakin mielessä "liian hyvin" nollahypoteesin kiinnittämää jakaumaa.*

A:n noppa ja 1 %:n merkitsevyystaso:

Koska χ^2 -testisuureen arvo

$$\chi^2 = 13.1 > 0.554 = \chi_{0.99}^2$$

niin voimme päätellä, että *A on todennäköisesti todella heittänyt noppaa.*

B:n noppa ja 1 %:n merkitsevyystaso:

Koska χ^2 -testisuureen arvo

$$\chi^2 = 0.3 < 0.554 = \chi_{0.99}^2$$

niin, *on syytä epäillä sitä, että B on heittänyt noppaa. B on saattanut keksiä heittotulokset, mutta on aliarvioinut satunnaisvaihtelun merkityksen.*

Tehtävä 11.2.

Geiger-mittari laskee radioaktiivisen aineen emissioiden lukumääriä. Emissioiden lukumäärä on lyhyellä aikavälillä satunnaismuuttuja, jonka voidaan olettaa noudattavan Poissonin jakaumaa.

Erästä ainetta tutkittaessa emissioiden lukumäärät rekisteröitiin 101:llä samanmittaisella lyhyellä aikavälillä. Alla olevassa taulukossa on annettu emissioiden lukumäärien frekvenssit.

Tutki χ^2 -yhteensopivuustestin avulla onko Poisson-jakaumaoletus sopusoinnussa havaintojen kanssa. Käytä testissä 5 %:n merkitsevyystasoa.

Emissioiden lkm	0	1	2	3	4	5
Frekvenssi	40	34	18	5	2	2

Tehtävä 11.2. – Mitä opimme?

Tehtävässä sovelletaan χ^2 -yhteensopivuustestiä.

Tehtävä 11.2. – Ratkaisu:

Erästä radioaktiivista ainetta tutkittaessa Geiger-mittarilla rekisteröitiin emissioiden lukumäärät 101:llä samanmittaisella lyhyellä aikavälillä. Alla olevassa taulukossa on annettu emissioiden lukumäärien frekvenssit.

Emissioiden lkm	0	1	2	3	4	5
Frekvenssi	40	34	18	5	2	2

Tehtävän *nollahypoteesina* on

H_0 : Emissioiden lukumäärä noudattaa *Poisson-jakaumaa*

Poisson-jakauman *pistetodennäköisyysfunktio* on

$$f(x) = \Pr(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

Intensiteettiparametrin λ *suurimman uskottavuuden estimaattori* on havaintojen *aritmeettinen keskiarvo*:

$$\bar{x} = \frac{1}{n} \sum_{i=0}^5 i O_i = \frac{1}{101} \times 103 = 1.0198$$

jossa

n = havaintojen kokonaislukumäärä = 101

O_i = havaittu frekvenssi luokassa i

Odotetuiksi frekvensseiksi saadaan:

$$E_0 = \Pr(X=0) \times n = 36.43$$

$$E_1 = \Pr(X=1) \times n = 37.15$$

$$E_2 = \Pr(X=2) \times n = 18.94$$

$$E_3 = \Pr(X=3) \times n = 6.44$$

$$E_4 = \Pr(X=4) \times n = 1.64$$

$$E_5 = \Pr(X=5) \times n = 0.33$$

Koska

$$E_4 = 1.64 < 5$$

$$E_5 = 0.33 < 5$$

niin luokat 4 ja 5 yhdistetään luokkaan 3.

χ^2 -testisuureen arvoksi saadaan

$$\chi^2 = 0.705$$

Laskutoimitukset (Microsoft Excel –ohjelmistolla):

i	O_i	E_i	$(O_i - E_i)^2/E_i$
0	40	36.43	0.350
1	34	37.15	0.267
2	18	18.94	0.047
3 tai yli	9	8.41	0.041
Summa	101	100.93	0.705

Koska $k = 4$ ja $m = 1$ (yksi parametri on estimoitu), χ^2 -testisuureen vapausasteiden lukumäärä on

$$f = k - 1 - m = 4 - 1 - 1 = 2$$

χ^2 -jakauman taulukoiden mukaan 5 %:n merkitsevyystasoa vastaava kriittinen arvo $\chi_{0.05}^2$

saadaan seuraavalla tavalla: Erotetaan todennäköisyysmassa 0.05 jakauman oikealle hännälle.

Koska

$$\Pr(\chi^2 \geq 5.99) = 0.05$$

niin

$$\chi_{0.05}^2 = 5.99$$

Koska χ^2 -testisuureen arvo

$$\chi^2 = 0.705 < 5.99 = \chi_{0.05}^2$$

niin nollahypoteesi jää voimaan 5 %:n merkitsevyystasolla.

Johtopäätös:

Havainnot saattavat noudattaa Poisson-jakaumaa.

Tehtävä 11.3.

Erään tehtaan johto epäili, että työntekijöiden keskuuteen oli levinnyt tapa venyttää viikonlopun viettoja perjantaihin ja maanantaihin ilmoittautumalla sairaiksi. Asiaa tutkittiin neljän viikon ajan rekisteröimällä poissaolojen lukumäärät jokaisena työpäivänä. Keskiarvotiedot ko. ajanjaksolta on annettu alla olevassa taulukossa.

- (a) Testaa χ^2 -yhteensopivuustestin avulla nollahypoteesia, että poissaolojen lukumäärä jakautuu tasaisesti työpäiville.
- (b) Yhdistä sekä perjantain ja maanantain että tiistain, keskiviikon ja torstain havainnot. Miten (a)-kohdan nollahypoteesia on tällöin modifioitava? Testaa χ^2 -yhteensopivuustestin avulla modifioitua nollahypoteesia.

Käytä testeissä 5 %:n merkitsevyystasoa.

Viikonpäivä	ma	ti	ke	to	pe	Summa
Poissaolojen lukumäärä (ka)	49	35	32	39	45	200

Tehtävä 11.3. – Mitä opimme?

Tehtävässä sovelletaan χ^2 -yhteensopivuustestiä.

Tehtävä 11.3. – Ratkaisu:

- (a) Määritetään odotetut frekvenssit E_j käyttäen nollahypoteesia

H_0 : Poissaolot jakautuvat tasaisesti eri viikonpäiville

Odotetut frekvenssit E_j :

Viikonpäivä	ma	ti	ke	to	pe	Summa
Poissaolojen lukumäärä	40	40	40	40	40	200

χ^2 -testisuureen arvoksi saadaan

$$\chi^2 = 4.90$$

Laskutoimitukset (Microsoft Excel –ohjelmistolla):

Päivä	ma	ti	ke	to	pe	Summa
O_i	49	35	32	39	45	200
E_i	40	40	40	40	40	200
Khi^2	2.025	0.625	1.6	0.025	0.625	4.90

Koska $k = 5$ ja $m = 0$ (yhtään parametria ei ole estimoitu), χ^2 -testisuureen *vapausasteiden lukumäärä* on

$$f = k - 1 - m = 5 - 1 - 0 = 4$$

χ^2 -jakauman taulukoiden mukaan 5 %:n merkitsevyystasoa vastaava *kriittinen arvo* $\chi^2_{0.05}$ saadaan seuraavalla tavalla: Erotetaan todennäköisyysmassa 0.05 jakauman *oikealle* hänälle. Koska

$$\Pr(\chi^2 \geq 9.488) = 0.05$$

niin

$$\chi^2_{0.05} = 9.488$$

Koska χ^2 -testisuureen arvo

$$\chi^2 = 4.90 < 9.488 = \chi^2_{0.05}$$

niin *nollahypoteesi jää voimaan* 5 %:n merkitsevyystasolla.

Johtopäätös:

Poissaolot saattavat hyvinkin jakautua tasaisesti eri viikonpäiville.

- (b) Yhdistämällä sekä perjantain ja maanantain havainnot että tiistain, keskiviikon ja torstain havainnot saadaan seuraava havaittujen frekvenssien taulukko:

Viikonpäivä	pe-ma	ti-ke-to	Summa
Poissaolojen lukumäärä (ka)	94	106	200

Vastaava yhdistäminen (a)-kohdan *odotettujen frekvenssien* taulukossa tuottaa taulukon

Viikonpäivä	pe-ma	ti-ke-to	Summa
Poissaolojen lukumäärä	80	120	200

χ^2 -testisuureen arvoksi saadaan

$$\chi^2 = 4.08$$

Laskutoimitukset (Microsoft Excel –ohjelmistolla):

Päivä	pe-ma	ti-ke-to	Summa
O_i	94	106	200
E_i	80	120	200
Khi^2	2.45	1.63	4.08

Koska $k = 2$ ja $m = 0$ (yhtään parametria ei ole estimoitu), χ^2 -testisuureen *vapausasteiden lukumäärä* on

$$f = k - 1 - m = 2 - 1 - 0 = 1$$

χ^2 -jakauman taulukoiden mukaan 5 %:n merkitsevyystasoa vastaava *kriittinen arvo* $\chi_{0.05}^2$ saadaan seuraavalla tavalla: Erotetaan todennäköisyysmassa 0.05 jakauman *oikealle* hänälle. Koska

$$\Pr(\chi^2 \geq 3.841) = 0.05$$

niin siten

$$\chi_{0.05}^2 = 3.841$$

Koska χ^2 -testisuureen arvo

$$\chi^2 = 4.08 > 3.841 = \chi_{0.05}^2$$

niin *nollahypoteesi hylätään* 5 %:n merkitsevyystasolla.

Johtopäätös:

Poissaolot eivät jakaudu tasaisesti eri viikonpäiville.

Huomautus:

Tehtävästä 11.3. nähdään, että käytetty *luokitus saattaa vaikuttaa χ^2 -yhteensopivuustestin tulokseen*.

Tehtävä 11.4.

Henkilöille A ja B on kummallekin annettu noppa ja kumpaakin on pyydetty heittämään sitä 120 kertaa. A ja B kertovat saaneensa heittojen tuloksena alla esitetyt silmälukujen jakaumat.

Tutki χ^2 -homogeenisuustestin avulla, onko mahdollista, että A ja B ovat käyttäneet samaa noppaa (tai oikeammin: noppia, joiden silmälukujen todennäköisyydet ovat samat).

Käytä testissä 5 %:n merkitsevyystasoa.

Silmäluku	1	2	3	4	5	6
A:n tulokset	16	18	19	22	19	26
B:n tulokset	12	16	20	17	22	33

Tehtävä 11.4. – Mitä opimme?

Tehtävässä sovelletaan χ^2 -homogeenisuustestiä.

Vrt. tässä tarkasteltavaa χ^2 -homogeenisuustestiä tehtävissä 11.5. ja 11.6. sovellettavaan χ^2 -riippumattomuustestiin.

Tehtävä 11.4. – Ratkaisu:

Sovelletaan testisuuretta

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

jossa odotetut frekvenssit E_{ij} määrätään käyttäen nollahypoteesia

H_0 : A ja B ovat käyttäneet samaa noppaa (tai oikeammin: noppia, joiden silmälukujen todennäköisyydet ovat samat)

Siten *odotetut frekvenssit* saadaan kaavalla

$$E_{ij} = n_i C_j / n$$

jossa

n_i = i . ryhmän *otoskoko*

C_j = j . sarakesumma

n = kokonaissumma

Nollahypoteesin H_0 pätiessä testisuure χ^2 noudattaa suurissa otoksissa approksimatiivisesti χ^2 -jakaumaa mukaan vapausastein $(r-1)(c-1)$, jossa

r = frekvenssitaulun rivien lukumäärä

c = frekvenssitaulun sarakkeiden lukumäärä

Olkoon *vaihtoehtoisena hypoteesina* H_1 se, että A ja B eivät ole käyttäneet samaa noppaa (tai oikeammin: A ja B ovat käyttäneet noppia, joiden silmälukujen todennäköisyydet eroavat tosistaan).

χ^2 -testisuureen arvoksi saadaan

$$\chi^2 = 2.406$$

Laskutoimitukset (Microsoft Excel -ohjelmalla):

O_{ij}	1	2	3	4	5	6	Summa
A	16	18	19	22	19	26	120
B	12	16	20	17	22	33	120
Summa	28	34	39	39	41	59	240

E_{ij}	1	2	3	4	5	6	Summa
A	14	17	19.5	19.5	20.5	29.5	120
B	14	17	19.5	19.5	20.5	29.5	120
Summa	28	34	39	39	41	59	240

Khi^2	1	2	3	4	5	6	Summa
A	0.285714	0.058824	0.012821	0.320513	0.109756	0.415254	1.202881
B	0.285714	0.058824	0.012821	0.320513	0.109756	0.415254	1.202881
Summa	0.571429	0.117647	0.025641	0.641026	0.219512	0.830508	2.405763

Vapausteiden lukumääräksi saadaan

$$f = (r - 1)(c - 1) = (2 - 1)(6 - 1) = 5$$

Siten 5 %:n merkitsevyystasoa vastaava kriittiseksi arvoksi $\chi^2_{0.05}$ saadaan χ^2 -jakauman taulukoista

$$\chi^2_{0.05} = 11.07$$

Koska testisuureen arvo

$$\chi^2 = 2.41 < 11.07 = \chi^2_{0.05}$$

niin testisuureen χ^2 arvo on jäänyt hyväksymisalueelle ja *nollahypoteesi* H_0 voidaan jättää voimaan merkitsevyystasolla 0.05.

Johtopäätös:

A:n ja B:n tulokset voivat olla peräisin samasta nopasta (tai oikeammin: nopista, joiden silmälukujen todennäköisyydet ovat samat).

Tehtävä 11.5.

Erään rokotuskokeen tulokset on annettu alla.

- (a) Testaa nollahypoteesia, että rokotettujen ja rokottamattomien sairastavuudessa ei ole eroa suhteellisten osuuksien vertailutestiä käyttäen.
- (b) Testaa nollahypoteesia, että sairastavuus ei riipu rokotuksesta χ^2 -riippumattomuustestiä käyttäen.

Käytä molemmissa testeissä 5 %:n merkitsevyystasoa.

Vertaa (a)- ja (b)-kohtien testien tuloksia toisiinsa. Voivatko testit johtaa eri tuloksiin?

Sairastuminen → Rokotus ↓	Sairastunut: S	Ei sairastunut: ei-S
Rokotettu: R	9	42
Ei rokotettu: ei-R	17	28

Tehtävä 11.5. – Mitä opimme?

Tehtävässä sovelletaan samaan aineistoon *suhteellisten osuuksien vertailutestiä* riippumattomille otoksille ja *χ^2 -riippumattomuustestiä*.

Lisätietoja suhteellisten osuuksien vertailutestistä: ks. 10. harjoitukset.

Vrt. tässä tarkasteltavaa *χ^2 -riippumattomuustestiä* tehtävässä 11.4. sovellettavaan *χ^2 -homogeenisuustestiin*.

Tehtävä 11.5. – Ratkaisu:

- (a) Olkoon

$$A = \text{”Henkilö sairastuu”}$$

ja olkoon

$$\Pr(A) = p_1, \text{ jos henkilö on rokotettu}$$

$$\Pr(A) = p_2, \text{ jos henkilöä ei ole rokotettu}$$

Määritellään riippumattomat satunnaismuuttujat

$$X_{ik} = \begin{cases} 1, & \text{jos } i. \text{ henkilö, jolle on tehty toimepide } k, \text{ sairastuu} \\ 0, & \text{jos } i. \text{ henkilö, jolle on tehty toimepide } k, \text{ ei sairastu} \end{cases}$$

$$i = 1, 2, \dots, n_k, k = 1, 2$$

jossa

$$k = 1 \leftrightarrow \text{henkilö rokotetaan}$$

$$k = 2 \leftrightarrow \text{henkilöä ei rokoteta}$$

Tällöin

$$X_{ik} \sim \text{Ber}(p_k), i = 1, 2, \dots, n_k, k = 1, 2$$

Asetetaan *nollahypoteesi*

$$H_0 : p_1 = p_2$$

Määritellään testisuure (ks. 10. harjoitukset)

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Testisuureen lausekkeessa

n_1 = Rokotettujen lukumäärä

$\hat{p}_1$ = Sairastuneiden *suhteellinen osuus* rokotettujen joukossa

n_2 = Ei-rokotettujen lukumäärä

$\hat{p}_2$ = Sairastuneiden *suhteellinen osuus* ei-rokotettujen joukossa

ja

$\hat{p}$ = Sairastuneiden suhteellinen osuus, kun rokotettuja ja ei-rokotettuja käsitellään yhtenä ryhmänä

Huomaa, että

$$\hat{p}_1 = f_1 / n_1$$

$$\hat{p}_2 = f_2 / n_2$$

jossa

f_1 = Sairastuneiden *lukumäärä* rokotettujen joukossa

f_2 = Sairastuneiden *lukumäärä* ei-rokotettujen joukossa

ja

$$\hat{p} = \frac{f_1 + f_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Jos *nollahypoteesi* H_0 pätee, testisuure z noudattaa suurissa otoksissa *approksimatiivisesti standardoitua normaalijakaumaa*:

$$z \sim_a N(0,1)$$

Tehtävän tapauksessa:

$$n_1 = 9 + 42 = 51$$

$$\hat{p}_1 = \frac{f_1}{n_1} = \frac{9}{51} = 0.1765$$

$$n_2 = 17 + 28 = 45$$

$$\hat{p}_2 = \frac{f_2}{n_2} = \frac{17}{45} = 0.3778$$

ja edelleen

$$\hat{p} = \frac{f_1 + f_2}{n_1 + n_2} = \frac{9 + 17}{51 + 45} = \frac{26}{96} = 0.2708$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.177 - 0.378}{\sqrt{0.271 \times (1 - 0.271) \left(\frac{1}{51} + \frac{1}{45}\right)}} = -2.215$$

Valitaan *vaihtoehtoiseksi hypoteesiksi* se, että sairastuvuus on rokotettujen joukossa *pienempää* kuin ei-rokotettujen joukossa eli olkoon vaihtoehtoisena hypoteesina *yksisuuntainen vaihtoehto*

$$H_1 : p_1 < p_2$$

Tällöin 5 %:n merkitsevyystasoa vastaavaksi *kriittiseksi arvoksi* $-z_{0.05}$ saadaan normaalijakauman taulukosta

$$-z_{0.05} = -1.65$$

Koska

$$z = -2.21 < -1.65 = -z_{0.05}$$

niin testisuureen z arvo on jäänyt hylkäysalueelle ja nollahypoteesi H_0 voidaan hylätä ja *vaihtoehtoinen hypoteesi* H_1 hyväksyä merkitsevyystasolla 0.05.

Johtopäätös:

Sairastuvuus rokotettujen joukossa on *pienempää* kuin rokottamattomien joukossa.

- (b) Sovelletaan χ^2 -testisuuretta

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

jossa odotetut frekvenssit E_{ij} määrätään käyttäen nollahypoteesia

$$H_0 : \text{Sairastumistodennäköisyys ei riipu rokotuksesta}$$

Siten *odotetut frekvenssit* E_{ij} saadaan kaavalla

$$E_{ij} = R_i C_j / n$$

jossa

$$R_i = i. \text{ rivisumma}$$

$$C_j = j. \text{ sarakesumma}$$

$$n = \text{kokonaissumma}$$

Nollahypoteesin H_0 pätiessä testisuure χ^2 noudattaa suurissa otoksissa approksimatiivisesti χ^2 -jakaumaa vapausastein $(r-1)(c-1)$, jossa

$$r = \text{frekvenssitaulun rivien lukumäärä}$$

$$c = \text{frekvenssitaulun sarakkeiden lukumäärä}$$

Olkoon *vaihtoehtoisena hypoteesina* H_1 se, että sairastumistodennäköisyys riippuu rokotuksesta.

χ^2 -testisuureen arvoksi saadaan

$$\chi^2 = 4.906$$

Laskutoimitukset (Microsoft Excel –ohjelmalla):

O_{ij}	S	ei-S	Summa
R	9	42	51
ei-R	17	28	45
Summa	26	70	96

E_{ij}	S	ei-S	Summa
R	13.8125	37.1875	51
ei-R	12.1875	32.8125	45
Summa	26	70	96

Tarkistus

51

45

Tarkistus

26

70

Khi^2	S	ei-S	Summa
R	1.676753	0.622794	2.299548
ei-R	1.900321	0.705833	2.606154
Summa	3.577074	1.328627	4.905701

Vapausasteiden lukumääräksi saadaan

$$f = (r-1)(c-1) = (2-1)(2-1) = 1$$

Siten 5 %:n merkitsevyystasoa vastaavaksi *kriittiseksi arvoksi* $\chi^2_{0.05}$ saadaan χ^2 -jakauman taulukoista

$$\chi^2_{0.05} = 3.84$$

Koska testisuureen arvo

$$\chi^2 = 4.91 > 3.84 = \chi_{0.05}^2$$

niin testisuureen χ^2 arvo on joitunut hylkäysalueelle ja *nollahypoteesi* H_0 voidaan hylätä ja *vaihtoehtoinen hypoteesi* H_1 hyväksyä merkitsevyystasolla 0.05.

Johtopäätös:

Sairastumistodennäköisyys riippuu siitä onko henkilö rokotettu vai ei.

Huomautuksia:

- (i) (a)-kohdan z -testisuureen arvon neliö *yhtyy* (b)-kohdan χ^2 -testisuureen arvoon. Tämä ei ole sattumaa, vaan perustuu χ^2 -testisuureen ominaisuuksiin 2×2 -frekvenssitaulussa.
- (ii) Huomaa kuitenkin, että **z -testin ja χ^2 -testin ekvivalenssia 2×2 -frekvenssitaulun tapauksessa ei voida yleistää $r \times c$ -tauluihin, joissa $r > 2$ ja/tai $c > 2$.**

Tehtävä 11.6.

Alla olevassa taulukossa on annettu tulokset kyselytutkimuksesta, jonka kohteena olivat USA:n kansalaiset. Kyselyn kohteet poimittiin yksinkertaisella satunnaisotannalla ja otokseen poimituilta kysyttiin puoluekanta ja suhtautumista käsiaseiden hallussapidon rajoituksiin.

Onko suhtautuminen aserajoituksiin riippumatonta puoluekannasta?

Käytä asian tutkimisessa χ^2 -riippumattomuustestiä 0.5 %:n merkitsevyystasolla.

Suhtautuminen aserajoituksiin → Puoluekanta ↓	Puoltaa	Ei kantaa	Vastustaa
Demokraatti	110	26	64
Republikaani	90	14	116
Riippumaton	55	10	35

Tehtävä 11.6. – Mitä opimme?

Tehtävässä sovelletaan χ^2 -riippumattomuustestiä.

Vrt. tässä tarkasteltavaa χ^2 -riippumattomuustestiä tehtävässä 11.4. sovellettavaan χ^2 -homogeenisuustestiin.

Tehtävä 11.6. – Ratkaisu:

Sovelletaan testisuuretta

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

jossa odotetut frekvenssit E_{ij} määrätään käyttäen nollahypoteesia

H_0 : Suhtautuminen aserajoituksiin ei riipu puoluekannasta

Siten odotetut frekvenssit saadaan kaavalla

$$E_{ij} = R_i C_j / n$$

jossa

$$R_i = i. \text{ rivisumma}$$

$$C_j = j. \text{ sarakesumma}$$

$$n = \text{kokonaissumma}$$

Nollahypoteesin H_0 pätiessä χ^2 -testisuure noudattaa suurissa otoksissa approksimatiivisesti χ^2 -jakauman mukaan vapausastein $(r-1)(c-1)$, jossa

r = frekvenssitaulun rivien lukumäärä

c = frekvenssitaulun sarakkeiden lukumäärä

Olkoon vaihtoehtoisena hypoteesina H_1 se, että suhtautuminen aserajoituksiin riippuu puoluekannasta.

Havaitut frekvenssit O_{ij} :

O_{ij}	Puoltaa	Ei kantaa	Vastustaa	Yhteensä
Dem	110	26	64	200
Rep	90	14	116	220
Riip	55	10	35	100
Yhteensä	255	50	215	520

Odotetut frekvenssit E_{ij} :

E_{ij}	Puoltaa	Ei kantaa	Vastustaa	Yhteensä
Dem	98.1	19.2	82.7	200
Rep	107.9	21.1	91.0	220
Riip	49.0	9.6	41.4	100
Yhteensä	255	50	215	520

χ^2 -testisuureen arvoksi saadaan

$$\chi^2 = 22.05$$

Laskutoimitukset (Microsoft Excel -ohjelmalla):

O_{ij}	Puoltaa	Ei kantaa	Vastustaa	Yhteensä
Dem	110	26	64	200
Rep	90	14	116	220
Riip	55	10	35	100
Summa	255	50	215	520

E_{ij}	Puoltaa	Ei kantaa	Vastustaa	Yhteensä
Dem	98.1	19.2	82.7	200
Rep	107.9	21.2	91.0	220
Riip	49.0	9.6	41.3	100
Summa	255	50	215	520

Khi^2	Puoltaa	Ei kantaa	Vastustaa	Yhteensä
Dem	1.45	2.38	4.23	8.06
Rep	2.96	2.42	6.89	12.28
Riip	0.72	0.02	0.97	1.71
Summa	5.14	4.82	12.09	22.05

Vapausteiden lukumääräksi saadaan

$$f = (r - 1)(c - 1) = (3 - 1)(3 - 1) = 4$$

Siten 0.5 %:n merkitsevyystasoa vastaava *kriittiseksi arvoksi* $\chi^2_{0.005}$ saadaan χ^2 -jakauman taulukoista

$$\chi^2_{0.005} = 14.86$$

Koska testisuureen arvo

$$\chi^2 = 22.05 > 14.86 = \chi^2_{0.005}$$

niin testisuureen χ^2 arvo *on joutunut hylkäysalueelle* ja *nollahypoteesi* H_0 *voidaan hylätä* ja *vaihtoehtoinen hypoteesi* H_1 *voidaan hyväksyä* merkitsevyystasolla 0.005.

Johtopäätös:

Puoluekanta ja suhtautuminen aserajoituksiin riippuvat toisistaan.

Tehtävä 11.7.

Eräässä 42:n kunnan otoksessa Suomen kuntien joukosta suhteellisen rikollisuuden (muuttuja x ; rikoksia per 1000 asukasta) ja asukastiheyden (muuttuja y ; asukasta per km^2) välisen otoskorrelaatiokertoimen arvoksi saatiin $r = 0.157$.

Oletetaan, että havaintojen parit

$$(x_i, y_i), i = 1, 2, \dots, n$$

jossa indeksi i viittaa kuntaan i , noudattavat kaksiulotteista normaalijakaumaa

$$N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$$

Testaa 5 %:n merkitsevyystasoa käyttäen nollahypoteesia, että muuttujat x ja y ovat korreloimattomia, kun vaihtoehtoiseksi hypoteesiksi valitaan kaksisuuntainen vaihtoehto.

Tehtävä 11.7. – Mitä opimme?

Tehtävässä tarkastellaan *korreloimattomuuden testaamista*.

Lisätietoja kaksiulotteisesta normaalijakaumasta: ks. 12. harjoitukset.

Yleinen testi korrelaatiokertoimelle: ks. tehtävää 11.8.

Tehtävä 11.7. – Ratkaisu:

t -testisuure nollahypoteesille

$$H_0 : \rho_{xy} = 0$$

on muotoa

$$t = \sqrt{n-2} \frac{r_{xy}}{\sqrt{1-r_{xy}^2}}$$

jossa r_{xy} on muuttujien x ja y havaituista arvoista määrätty Pearsonin otoskorrelaatiokerroin.

Jos nollahypoteesi H_0 pätee, testisuure t on jakautunut Studentin t -jakauman mukaan vapausastein $(n-2)$:

$$t \sim t(n-2)$$

Tehtävän tapauksessa testisuureen arvoksi saadaan

$$t = \sqrt{n-2} \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} = \sqrt{42-2} \times \frac{0.157}{\sqrt{1-0.157^2}} = 1.005$$

Vapausasteiden lukumääräksi saadaan

$$df = n - 2 = 40$$

Koska vaihtoehtoiseksi hypoteesiksi on valittu kaksisuuntainen vaihtoehto

$$H_1 : \rho_{xy} \neq 0$$

niin 5 %:n merkitsevyystasoa vastaavaksi *kriittisiksi rajoiksi* $-t_{0.025}$ ja $+t_{0.025}$ saadaan Studentin t -jakauman taulukoista

$$-t_{0.025} = -2.021$$

$$+t_{0.025} = +2.021$$

Koska

$$-t_{0.025} = -2.021 < t = 1.01 < +2.021 = +t_{0.025}$$

testisuureen t arvo *on jäänyt hyväksymisalueelle* ja nollahypoteesi H_0 *voidaan jättää voimaan* 5 %:n merkitsevyystasolla.

Johtopäätös:

Otoksesta saatujen tietojen perusteella kunnan suhteellisen rikollisuuden ja asukasluvun välillä *ei ole korrelaatiota*.

Tehtävä 11.8.

Menestyminen opinnoissa saattaa vaikuttaa vastavalmistuneen alkupalkkaan.

Asiaa tutkittiin eräässä USA:n yliopistossa poimimalla vastavalmistuneiden joukosta yksinkertainen satunnaisotos, jonka koko oli 15. Otokseen poimituilta opiskelijoilta kysyttiin heidän arvosanapisteidensä keskiarvoa (muuttuja x) ja alkupalkkaa (muuttuja y ; 1000 \$).

Otosta kuvaavat perustunnusluvut olivat:

$$\bar{x} = 3.04 \qquad \bar{y} = 18.05$$

$$s_x^2 = 0.063 \qquad s_y^2 = 5.81$$

$$r_{xy} = 0.848$$

Oletetaan, että havaintojen parit

$$(x_i, y_i), i = 1, 2, \dots, n$$

noudattavat kaksiulotteista normaalijakaumaa $N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$.

Testaa 5 %:n merkitsevyystasolla nollahypoteesia

$$H_0 : \rho_{xy} = 0.5$$

kun vaihtoehtoiseksi hypoteesiksi valitaan se, että arvosanapisteiden keskiarvon ja alkupalkan välisen korrelaatiokertoimen arvo on suurempi kuin 0.5.

Tehtävä 11.8. – Mitä opimme?

Tehtävässä tarkastellaan yleistä testiä *Pearsonin tulomomenttikorrelaatiokertoimelle*.

Lisätietoja kaksiulotteisesta normaalijakaumasta: ks. 12. harjoitukset.

Korreloimattomuuden testaaminen: ks. tehtävää 11.7.

Tehtävä 11.8. – Ratkaisu:

Testisuure nollahypoteesille

$$H_0 : \rho_{xy} = \rho_0 = 0.5$$

on muotoa

$$v = \frac{\frac{1}{2} \log \left(\frac{1+r_{xy}}{1-r_{xy}} \right) - \frac{1}{2} \log \left(\frac{1+\rho_0}{1-\rho_0} \right)}{\sqrt{\frac{1}{n-3}}}$$

jossa r_{xy} on muuttujien x ja y havaituista arvoista määrätty Pearsonin otoskorrelaatiokerroin.

Jos nollahypoteesi H_0 pätee,

$$v \sim_a N(0,1)$$

Tehtävän tapauksessa

$$\frac{1}{2} \log \left(\frac{1+r_{xy}}{1-r_{xy}} \right) = \frac{1}{2} \log \left(\frac{1+0.848}{1-0.848} \right) = 1.249$$

$$\frac{1}{2} \log \left(\frac{1+\rho_0}{1-\rho_0} \right) = \frac{1}{2} \log \left(\frac{1+0.5}{1-0.5} \right) = 0.549$$

joten

$$v = \frac{1.249 - 0.549}{\sqrt{\frac{1}{15-3}}} = 2.42$$

Koska vaihtoehtoiseksi hypoteesiksi on valittu yksisuuntainen vaihtoehto

$$H_0 : \rho > 0.5$$

niin 5 %:n merkitsevyystasoa vastaavaksi *kriittiseksi rajoiksi* $+z_{0.05}$ saadaan normaalijakauman taulukoista

$$+z_{0.05} = +1.64$$

Koska

$$+z_{0.05} = +1.64 < +2.42 = z$$

testisuureen z arvo on joutunut hylkäysalueelle ja nollahypoteesi H_0 voidaan hylätä ja vaihtoehtoinen hypoteesi H_1 voidaan hyväksyä 5 %:n merkitsevyystasolla.

Johtopäätös:

Otoksesta saatujen tietojen perusteella arvosanapisteiden keskiarvon ja alkupalkan välisen otoskorrelaatiokertoimen arvo on suurempi kuin 0.5.