Mat-1.2620 Sovellettu todennäköisyyslaskenta B

2. välikoe 14.05.2011 / Kibble

Kirjoita selvästi *jokaiseen koepaperiin* seuraavat tiedot:

– Mat-1.2620 SovTnB 2. vk 14.05.2011
– opiskelijanumero + kirjain
– TEKSTATEN sukunimi ja kaikki etunimet
– koulutusohjelma ja vuosikurssi
– mahdolliset entiset nimet ja koulutusohjelmat
– nimikirjoitus

Sallitut apuvälineet: ***Laskin*** **ja Mellinin** ***kaava-*** **ja** ***taulukkokokoelmat*****.**

Vastausohje: **Vastaa** ***lyhyesti*** **ja** ***ytimekkäästi*****, mutta** ***perustele ratkaisusi*****. Pelkkä lukuarvo vastauksena** ***ei anna*** **pisteitä.**

1. Tehdas valmistaa kuulalaakerin kuulia. Kuulien paino vaihtelee satunnaisesti noudattaen normaalijakaumaa. Kuulien joukosta poimittiin yksinkertainen satunnaisotos. Otoskeskiarvoksi saatiin tällöin 50 g. Tehdään (epärealistinen) oletus, että normaalijakauman varianssi 0.0025 g^2 on *tunnettu*.

Määrää 95 %:n *luottamusvälit* kuulien painon odotusarvolle, jos otoskokona oli

(a) 10

(b) 40

Vertaa saatujen luottamusvälien pituuksia toisiinsa. Miten luottamusvälin pituus käyttäytyy otoskoon funktiona?

Ratkaisu kysymykseen 1:

Määritellään satunnaismuuttujat

X_i = Kuulalaakerin i paino otoksessa, $i = 1, 2, \ldots, n$

Oletuksien mukaan

$$X_1, X_2, \ldots, X_n \perp$$
$$X_i \sim \text{N}(\mu, \sigma^2),\ i = 1, 2, \ldots, n$$

jossa varianssi

$$\sigma^2 = 0.0025\ \text{g}^2$$

on tunnettu.

Otokseen poimittujen ruuvien painojen *aritmeettinen keskiarvo* oli

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 50\ \text{g}$$

Konstruoidaan otoksesta saatujen tietojen perusteella $(1-\alpha)$ %:n *luottamusväli* odotusarvo-parametrille μ. Koska varianssi σ^2 oletettiin *tunnetuksi*, luottamusväli on muotoa

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

jossa

$\bar{X}$ = havaintojen *aritmeettinen keskiarvo* otoksessa

σ^2 = jakauman *varianssi*

n = havaintojen *lukumäärä*

$-z_{\alpha/2}$ ja $+z_{\alpha/2}$ = luottamustasoon $(1-\alpha)$ liittyvät *luottamuskertoimet normaalijakaumasta* N(0,1)

Valitaan *luottamustasoksi*

$$1-\alpha = 0.95$$

Koska siten

$$\alpha = 0.05$$

luottamuskertoimet ovat $-z_{\alpha/2} = -z_{0.025}$ ja $+z_{\alpha/2} = +z_{0.025}$ ja ne toteuttavat yhtälöt

$$\Pr(z \le -z_{\alpha/2}) = \Pr(z \le -z_{0.025}) = \frac{\alpha}{2} = 0.025$$

$$\Pr(z \ge +z_{\alpha/2}) = \Pr(z \ge +z_{0.025}) = \frac{\alpha}{2} = 0.025$$

jossa satunnaismuuttuja z noudattaa *standardoitua normaalijakaumaa*:

$$z \sim \mathrm{N}(0,1)$$

Siten

$$\Pr(-z_{\alpha/2} \le z \le +z_{\alpha/2}) = \Pr(-z_{0.025} \le z \le +z_{0.025}) = 1-\alpha = 0.95$$

Standardoidun normaalijakauman N(0,1) taulukoiden mukaan

$$-z_{0.025} = -1.96$$
$$+z_{0.025} = +1.96$$

Siten 95%:n *luottamusväli* normaalijakauman odotusarvoparametrille μ on muotoa

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 50 \pm 1.96 \times \frac{0.05}{\sqrt{n}}$$

(a) $n = 10$:

Luottamusväliksi saadaan

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 50 \pm 1.96 \times \frac{0.05}{\sqrt{10}} = 50 \pm 0.031 = (49.969, 50.031)$$

(b) $n = 40$:

Luottamusväliksi saadaan

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 50 \pm 1.96 \times \frac{0.05}{\sqrt{40}} = 50 \pm 0.015 = (49.985, 50.015)$$

Jos otantaa toistetaan, niin *luottamustason frekvenssitulkinnan* mukaan otoksista konstruoidut luottamusvälit peittävät (keskimäärin)

95 %:ssa

otoksia parametrin μ tuntemattoman arvon ja (keskimäärin)

5 %:ssa

otoksia ei sitä tee.

(i) Odotusarvon luottamusväli *lyhenee*, jos otoskokoa *kasvatetaan*.

(ii) Jos luottamusvälin pituus halutaan *puolittaa*, pitää havaintojen lukumäärä *nelinkertaistaa*.

2. Maissa A ja B ovat paljon jalava puita, joissa monissa on jalavatautisieni. Tutkimuksessa havaitaan että maassa A 500:n jalavan satunnaisotoksessa 325:lla puulla on jalavatautisieni ja että maassa B 300:n jalavan satunnaisotoksessa 201:lla puulla on jalavatautisieni.

Puu asiantuntijoilla on teoria että sairaiden puiden suhteellinen osuus maassa B on suurempi kuin maassa A koska tauti saapui maahan B ensin. Testaa 1%:n merkitsevyystasolla asiantuntijoiden teoriaa.

<u>Ratkaisu kysymykseen 2:</u>

Olkoon
A = ”Jalava puulla on jalavatautisieni”
ja Pr(A) = p_1 , Jalava puulla maassa A on jalavatautisieni
Pr(A) = p_2 , Jalava puulla maassa B on jalavatautisieni
Määritellään riippumattomat satunnaismuuttujat

Asetetaan *nollahypoteesiksi*
$H_0 : p_1 = p_2 = p$
Ja vaihtoehtoiseksi hypoteesiksi
$H_1 = p_1 < p_2$

Jos nolla hypoteesi pätee, voidaan otokset yhdistää ja p:n harhaton estimaattori on

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{325 + 201}{500 + 300} = 0.6575$$

Määritellään *testisuure*

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

Jos *nollahypoteesi* H_0 *pätee, testisuure z noudattaa suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa*:
$z \sim_a N(0,1)$

Tehtävässä

$$z = \frac{0.65 - 0.67}{\sqrt{0.6575(0.3425)\left(\frac{1}{500}+\frac{1}{300}\right)}} = -0.577$$

Merkitsevyystasoa 0.01 *vastaava kriittinen arvo* on
$z_{0.01} = -2.33$
Koska
z = -0.577 >- 2.33
testisuureen z arvo ei osunut hylkäysalueelle ja nollahypoteesi H_0 *voidaan jättää voimaan* 1 %:n *merkitsevyystasolla*. Ei ole tarpeeksi näyttöä 1%:n merkitsevyystasolla siitä että maassa B suhteellinen osuus jalava puista joilla on jalavatautisieni olisi suurempi kuin maassa A.

3. Eräässä kokeessa verrattiin kahta sademäärän mittaukseen käytettävää laitetta. Kummallakin laitteella mitattiin sademäärät *samalla paikalla* 6 sadepäivän aikana. Mittaustulokset (sademäärät mm:nä) on annettu alla olevassa taulukossa.

Testaa hypoteesia, että mittarit tuottavat keskimäärin samoja mittaustuloksia, kun vaihtoehtoisena hypoteesina on, että mittarit tuottavat keskimäärin eri mittaustuloksia. Käytä testissä 1 %:n merkitsevyystasoa.

Laite	**1**	**2**	**3**	**4**	**5**	**6**
A	1.3	9.6	0.3	1.4	5.9	0.5
B	1.4	10.3	0.3	1.5	6.1	0.6

Ratkaisu kysymykseen 3:

Laite	1	2	3	4	5	6
A	1,3	9,6	0,3	1,4	5,9	0,5
B	1,4	10,3	0,3	1,5	6,1	0,6
D_i	-0,1	-0,7	0	-0,1	-0,2	-0,1
(D_i-Dbar)^2	0,01	0,25	0,04	0,01	0	0,01

Dbar	-0,2		
Sum(D_i-Dbar)^2	0,32	t	-1,93649
s_D^2	0,064		

Riippumattomien otoksien *t*-testiä *ei saa käyttää*, koska *mittaustulokset samasta sateesta eivät ole riippumattomia*: Jos mittarit toimivat edes jossakin määrin luotettavasti, molempien mittareiden pitää antaa *samalle sateelle* toisiaan lähellä olevat mittaustulokset, ts. *A*-mittauksilla ja *B*-mittauksilla on oltava (voimakas) positiivinen korrelaatio.

Koska mittaustulokset riippuvat pareittain toisistaan, tällaisessa *parivertailuasetelmassa* toimitaan seuraavasti: Määrätään havaintoarvojen parikohtaiset *erotukset* ja testataan nollahypoteesia, jonka mukaan *erotukset ovat keskimäärin nollia*.

Olkoot

X_{Ai} = sateen i sademäärä mittarilla A, $i = 1, 2, \ldots, 6$

X_{Bi} = sateen i sademäärä mittarilla B, $i = 1, 2, \ldots, 6$

$D_i = X_{Ai} - X_{Bi}$, $i = 1, 2, \ldots, 6$

Yleinen hypoteesi H on muotoa:

$D_i \sim \mathrm{N}(\mu_D, \sigma_D^2)$, $i = 1, 2, \ldots, 6$

Erotukset $D_1, D_2, \ldots, D_6$ ovat riippumattomia

Nollahypoteesi H_0 on muotoa:

$\mathrm{E}(D_i) = 0$, $i = 1, 2, \ldots, 6$

Sovelletaan *yhden otoksen t-testiä* mittaustulosten erotuksille.

Testisuureena on

$$t = \frac{\bar{D}}{s_D / \sqrt{n}}$$

jossa

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n} (D_i - \bar{D})^2$$

Jos *nollahypoteesi* H_0 *pätee, testisuure t noudattaa Studentin t-jakaumaa* vapausastein ($n - 1$):

$$t \sim t(n-1) = t(5)$$

Itseisarvoltaan suuret testisuureen t arvot johtavat nollahypoteesin hylkäämiseen.

Tehtävän tapauksessa

$$n = 6\,,\ \bar{D} = -0.2\,,\ s_D^2 = 0.064$$

Siten

$$t = \frac{\bar{D}}{s_D / \sqrt{n}} = \frac{-0.2}{\sqrt{0.064} / \sqrt{6}} = -1.936$$

Koska vaihtoehtoisena hypoteesina on 2-suuntainen vaihtoehto $\mathrm{H}_1 : \mu_D \neq 0$, *merkitsevyystasoa* 0.01 *vastaavat kriittiset arvot* ovat

$$-t_{0.005} = -4.032$$

$$+t_{0.005} = +4.032$$

sillä *t*-jakauman taulukoiden mukaan

$$\Pr(t \geq 4.032) = 0.005$$

kun $t \sim t(5)$. Koska

$$-4.032 < t = -1.936 < +4.032$$

testisuureen *t* arvo –1.936 on osunut hyväksymisalueelle ja *nollahypoteesi* H_0 *jää voimaan* 1 %:n *merkitsevyystasolla*.

Johtopäätös:

Mittarit A ja B näyttävät keskimäärin samoja arvoja.

4. Kyselytutkimuksessa haluttiin verrata Vasemmistoliiton (Vas), Sosiaalidemokraattisen puolueen (SDP) ja Vihreän liiton (Vih) kannatuksen jakautumista äänioikeutettujen joukossa kolmessa kunnassa A, B ja C.

Vertailua varten kuntien A, B ja C äänioikeutettujen joukosta poimittiin toisistaan riippumattomat yksinkertaiset satunnaisotokset, joiden koot olivat 300 (kunta A), 180 (kunta B) ja 340 (kunta C) ja otokseen poimituilta kysyttiin mitä puoluetta he äänestäisivät seuraavissa vaaleissa. Tulokset kyselyistä on annettu alla olevassa taulukossa.

Testaa nollahypoteesia, että puolueiden kannatus jakautuu samalla tavalla kunnissa A, B ja C, kun vaihtoehtoisena hypoteesina on, että kannatus ei jakaudu samalla tavalla. Käytä testissä 5 %:n merkitsevyystasoa.

	Puolue	Vas	Sdp	Vih	Otoskoko
Kunta	A	60	160	80	300
	B	40	80	60	180
	C	60	160	120	340

Ratkaisu kysymykseen 4:

Sovelletaan χ^2-*homogeenisuustestiä*.

Sovelletaan testisuuretta

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

jossa odotetut frekvenssit E_{ij} määrätään käyttäen nollahypoteesia

H_0 : Puolueiden kannatus jakaantuu kunnissa A, B ja C samalla tavalla

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* vapausastein $(r - 1)(c - 1)$, jossa

r = frekvenssitaulun rivien lukumäärä

c = frekvenssitaulun sarakkeiden lukumäärä

Vaihtoehtoisena hypoteesina H_1 on se, että puolueiden kannatus ei jakaannu kunnissa A, B ja C samalla tavalla.

χ^2*-testisuureen arvoksi* saadaan

$$\chi^2 = 7.429$$

Laskutoimitukset:

O_{ij}	Vas	SDP	Vih	Otoskoko
A	60	160	80	300
B	40	80	60	180
C	60	160	120	340
Summa	160	400	260	820

E_{ij}	Vas	SDP	Vih	Otoskoko	Tarkistus
A	58,53659	146,3415	95,12195	300	300
B	35,12195	87,80488	57,07317	180	180
C	66,34146	165,8537	107,8049	340	340
Summa	160	400	260	820	820
Tarkistus	160	400	260	820	

χ^2	Vas	SDP	Vih	Otoskoko
A	0,036585	1,274797	2,404003	3,715385
B	0,677507	0,693767	0,150094	1,521368
C	0,606169	0,2066	1,379539	2,192308
Summa	1,320261	2,175163	3,933635	7,42906

Vapausteiden lukumääräksi saadaan

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

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

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

Koska testisuureen arvo

$$\chi^2 = 7.429 < 9.488 = \chi^2_{0.05}$$

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

5 Alla olevassa taulukossa on annettu muuttujien *y* ja *x* havaitut arvot. Havainnoista estimoidaan PNS-menetelmällä tavallinen lineaarinen regressiomalli:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , } \varepsilon_i \sim \mathrm{N}(0, \sigma^2) \text{ , } i = 1, 2, \ldots, n$$

jossa selitettävänä muuttujana on *y* ja selittävänä muuttujana on *x*.

(a) Määrää sen estimoidun mallin selitysaste R^2.

(b) Määrää estimoidun mallin residuaali pisteessä $x = 0$.

i	1	2	3	4	5
y	1	1	0	0	–1
x	–1	0	1	2	3

Ratkaisu kysymykseen 5:

Yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

regressiokertoimien β_0 ja β_1 *PNS-estimaatit* ja otoskorrelaatio r_{xy} saadaan lasketuksi seuraavassa esitettävällä tavalla.

Muuttujien *x* ja *y* havaittujen arvojen *aritmeettiset keskiarvot* $\bar{x}$ ja $\bar{y}$, *otosvarianssit* s_x^2 ja s_y^2, *otoskeskihajonnat* s_x ja s_y, *otoskovarianssi* s_{xy} ja *otoskorrelaatio* r_{xy} saadaan muuttujien *x* ja *y* havaittujen arvojen *summista*, *neliösummista* ja *tulosummasta*:

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

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

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right)$$

$$s_x = \sqrt{s_x^2}$$

$$s_y = \sqrt{s_y^2}$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right)$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Alla on taulukko jos näkyy muuttujien x ja y havaittujen arvojen *summat*, *neliösummat* ja *tulosumma*:

i	*x*	*y*	x^2	y^2	*xy*	*yhat*	*res*	res^2
1	-1	1	1	1	-1	1,2	-0,2	0,04
2	0	1	0	1	0	0,7	0,3	0,09
3	1	0	1	0	0	0,2	-0,2	0,04
4	2	0	4	0	0	-0,3	0,3	0,09
5	3	-1	9	1	-3	-0,8	-0,2	0,04
Summa	5	1	15	3	-4	1	2,22E-16	0,3

M(*x*) = 1
M(*y*) = 0,2
s_x^2 = 2,5 s_x = 1,581139
s_y^2 = 0,7 s_y = 0,83666
s_{xy} = -1,25
r_{xy} = -0,94491

b_1 = -0,5
b_0 = 0,7

s^2 = 0,1

SST = 2,8
SSE = 0,3
SSM = 2,5

R^2 = 0,892857
R = 0,944911

Estimoidun PNS-suoran yhtälö on muotoa

$$y = b_0 + b_1 x$$

jossa b_0 ja b_1 ovat mallin regressiokertoimien β_0 ja β_1 *PNS-estimaattorit*.

Estimaattoreiden b_0 ja b_1 arvot saadaan yllä määrätyistä otostunnusluvuista:

$$b_1 = r_{xy}\frac{s_y}{s_x} = -0.5$$

$$b_0 = \bar{y} - b\bar{x} = 0.7$$

Estimoidun *PNS-suoran yhtälöksi* saadaan siten

$$y = 0.7 - 0.5x$$

(a) Estimoidun mallin *selitysaste* R^2 voidaan laskea usealla eri tavalla. *Yhden selittäjän lineaarisen regressiomallin tapauksessa* (koska mallissa oli mukana vakio) pätee

$$R^2 = r_{xy}^2 = (-0.945)^2 = 0.893$$

(b) Residuaali pisteessä x=0 on

$$e_2 = y_2 - \hat{y} = 1 - 0.7 = 0.3$$