Tilastomatematiikka

Toinen välikoe 16.03.2023

1. Laatuinsinööri tutki jalostuksen vaikutusta erääseen suureeseen, teki 2 satunnaisotosta ja sai Excelillä seuraavan tulostuksen.

t-Test: Two-Sample Assuming Equal Variances			
t-testi: 2 otosta olettaen varianssit yhtäsuuriksi			
	Jalostus	*Ilman*	
Mean	3108,125	2902,8	keskiarvo
Variance	42382,41	76875,96	varianssi
Observations	8	10	havainnot
Pooled Variance	61785,03		
Hypothesized Mean Difference	0		
df	16		
t Stat	1,741442		
P(T<=t) one-tail	0,050398		
t Critical one-tail	1,745884		
P(T<=t) two-tail	0,100795		
t Critical two-tail	2,119905		

Kuva 1: *Excel*:n antama tulostus tehtävän 1 datalle

a) Laske otoskeskihajonnat molemmille otoksille. (1p)

b) Määrää tilanteeseen sopiva testimuuttuja. (1p)

c) Laske suureiden odotusarvojen erotuksen 95 % symmetrinen 2-suuntainen luottamusväli. Mitä mieltä olet sen perusteella jalostuksen vaikutuksesta? (4p)

2. Mooren lain mukaan transistorien lukumäärä mikropiirillä kaksinkertaistuu joka toinen vuosi. Taulukossa on transistorien lukumäärä mikropiirillä yksikkönä miljoona transistoria kahdeksalta vuodelta vuoden 2000 alusta alkaen.

t [a]	0	1	2	3	4	5	6	7
N $[10^6]$	37.2	42.6	55.7	151.2	273.8	305.1	582.9	805.8

a) Määrää regressiosuora muuttujien t ja $\ln N$ välille ja laske regressiosuoran selitysaste. Mitä mieltä olet mallista? (2p)

b) Laske a)-kohdan regressiosuoran kulmakertoimen 95 % luottamusväli. Minkä satunnaismuuttujan avulla saat laskettua luottamusvälin? Laske kriittisen alueen raja r_0. (3p)

c) Missä ajassa transistorien lukumäärä kaksinkertaistuu a)-kohdan mukaisen mallin mukaan? (1p)

3. Eräiden satunnaismuuttujien X ja Y yhteisjakauma on

$X\backslash Y$	1	3	5
0	$\frac{1}{3}$	0	0
1	0	$\frac{1}{3}$	0
2	0	0	$\frac{1}{3}$

Taulukko 1: Satunnaisvektorin (X, Y) pistetodennäköisyydet

a) Mikä on satunnaisvektorin (X, Y) arvojoukko? (1p)

b) Laske muuttujien X ja Y korrelaatiokerroin. (4p)

c) Määrää muuttujien X ja Y välinen riippuvuus niin tarkasti kuin se annettujen tietojen ja laskelmien perusteella on mahdollista. (2p)

Tehtävien ratkaisuperiaatteet

1. a) Excel-tulostuksesta näkee molempien otoksien 1 ja 2 otosvarianssit. Otoshajonnat ovat näiden neliöjuuret, eli $s_1 = \sqrt{42382.41} \approx 205.87$ ja $s_2 = \sqrt{76875.96} \approx 277.27$.

   b) Tulosteen mukaan populaatiovarianssit σ_1^2 ja σ_2^2 on oletettu yhtäsuuriksi, jolloin tilanteeseen sopii satunnaismuuttuja (testimuuttuja)

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}} \sim t_{n+m-2},$$

   joka löytyy esimerkiksi kaavakokoelmasta.

   c) Luottamusväli löydetään b)-kohdan testimuuttujan T avulla ehdosta

$$\mathbb{P}(-r_0 \leq T \leq r_0) = 95\,\%.$$

   Koska $n+m-2 = 16$, t-jakauman taulukosta löydetään $r_0 = 2.120$. Tämä löytyy myös Excel-tulosteesta. Puretaan epäyhtälö $-r_0 \leq T \leq r_0$ auki odotusarvojen erotuksen suhteen, jolloin saadaan luottamusväli

$$\overline{X} - \overline{Y} - r_0\sqrt{\tfrac{1}{n} + \tfrac{1}{m}}\sqrt{\tfrac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}} \leq \mu_1 - \mu 2 \leq \overline{X} - \overline{Y} + r_0\sqrt{\tfrac{1}{n} + \tfrac{1}{m}}\sqrt{\tfrac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}$$

   Excel-tulosteesta löydetään otoskeskiarvot, otosvarianssit ja otoskoot. Kun lukuarvot sijoitetaan paikoilleen, saadaan luottamusvälin realisaation alarajaksi otoksessa

$$3108.125 - 2902.8 - 2.120\sqrt{\frac{1}{8} + \frac{1}{10}}\sqrt{\frac{(8-1)42382.41 + (10-1)76875.96}{8+10-2}} \approx -44.6.$$

   Hieman vähemmällä naputtelulla olisi selvitty, jos olisi käytetty Excel-tulosteen yhdistettyä otosvarianssia

$$S_P^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2} \approx 61785.03,$$

   mutta menihän tuo kohtuudella tuota kaavaa ilmankin.
   Vastaavalla tavalla löydetään yläraja, joten otoksesta laskettu luottamusväli on

$$I_{\mu_1 - \mu_2} = [-44.6, 455.3].$$

   Tämän perusteella jalostuksella ei ole tilastollisesti merkittävää vaikutusta. Saman lopputuloksen olisi toki voinut nähdä suoraan myös Excel-tulosteen hypoteesien testauksesta.

2. a) Lasketaan transistorien lukumäärän N logaritmi $\ln N$ ja syötetään näin saatu havaintoaineisto Exceliin, josta *Data Analysis Toolpakin Regression*-toiminnolla saadaan regressiosuoraksi $\ln N = 3.42 + 0.48t$. Selitysaste on $r^2 \approx 0.97$, jonka perusteella malli selittää 97 % $\ln N$:n satunnaisvaihtelusta. Tämän perusteella malli sopii todella hyvin dataan.

c) Sopiva satunnaismuuttuja (testimuuttuja)

$$T = \sqrt{n-1}s_x \frac{\frac{S_{xY}}{s_{xx}} - \beta}{S_r} \sim t_{n-2}$$

löytyy esimerkiksi kaavakokoelmasta. Lasketaan muuttujan avulla kriittisen alueen raja r_0 ehdosta $\mathbb{P}(-r_0 \leq T \leq r_0) = 0.95$. Numeerisesti laskemalla tai t-jakauman taulukosta saadaan $r_0 \approx 2.447$, sillä vapausasteet ovat $n - 2 = 6$. Puretaan epäyhtälö $-r_0 \leq T \leq r_0$ auki β:n suhteen, jolloin saadaan luottamusväli

$$\frac{S_{xY}}{s_{xx}} - r_0 \frac{S_r}{\sqrt{n-1}s_x} \leq \beta \leq \frac{S_{xY}}{s_{xx}} + r_0 \frac{S_r}{\sqrt{n-1}s_x},$$

jonka realisaatio otoksessa on

$$b - r_0 \frac{s_r}{\sqrt{n-1}s_x} \leq \widehat{\beta} \leq b + r_0 \frac{s_r}{\sqrt{n-1}s_x},$$

missä b on a)-kohdassa lasketun regressiosuoran kulmakerroin, s_r on mallin jäännöshajonta ja s_x on ajan otoshajonta. Excel itse asiassa laskee valmiiksi kulmakertoimen keskivirheen $s_r/\sqrt{n-1}s_x$, joten hajonnan s_x laskeminen ei edes ole välttämätöntä. Kulmakertoimen keskivirhe on Excelin mukaan $SE(b) \approx 0.03517$. Niinpä otokselle lasketun luottamusvälin alarajaksi saadaan

$$b - r_0 SE(b) = 0.4798 - 2.571 \cdot 0.03517 \approx 0.394,$$

joka löytyy suoraan myös Excel-tulosteesta. Vastaavalla tavalla löydetään myös yläraja, joten luottamusväli on

$$I_\beta = [0.394, 0.566].$$

c) Olkoon a)-kohdan mallin mukaan $\ln N_1 = a + bt_1$ ja $\ln N_2 = a + bt_2$. Kun $N_2 = 2N_1$, saadaan

$$\ln N_2 - \ln N_1 = \ln(2N_1) - \ln(N_1) = \ln\left(\frac{2N_1}{N_1}\right) = \ln 2 = b(t_2 - t_1),$$

josta saadaan kaksinkertaistumisaika

$$t_2 - t_1 = \frac{\ln 2}{b} \overset{\text{sij.}}{\approx} 1.44.$$

3. a) Taulukosta nähdään, että positiivinen todennäköisyys esiintyy ainoastaan pistepareilla $(0,1), (1,3)$ ja $(2,5)$, joten satunnaisvektorin arvojoukko on $S_{X,Y} = \{(0,1), (1,3), (2,5)\}$.

b) Korrelaatiokertoimen laskukaava

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

löytyy kaavakokoelmasta.
Täytyy siis laskea X:n ja Y:n odotusarvo ja varianssi sekä tulon odotusarvo $\mathbb{E}(XY)$. Koska kullakin rivillä ja sarakkeella on vain yksi nollasta poikkeava pistetodennäköisyys, on X:n ja Y:n reunajakaumat helppo laskea. Muuttujan X jakaumaksi saadaan

x	0	1	2
$\mathbb{P}(X=x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Odotusarvoksi saadaan

$$\mathbb{E}(X) = \sum_{k=0}^{2} k\mathbb{P}(X=k) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} = 1.$$

Varianssille voidaan käyttää laskukaavaa

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \sum_{k=0}^{2} k^2 \cdot \mathbb{P}(X=k) - 1 = \frac{5}{3} - 1 = \frac{2}{3}.$$

Vastaavalla tavalla saadaan Y:n reunajakauma

y	1	3	5
$\mathbb{P}(Y=y)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

odotusarvo $\mathbb{E}(Y) = 3$ ja varianssi $\mathrm{Var}(Y) = \frac{8}{3}$. Tunnusluvut olisi voitu vaihtoehtoisesti laskea suoraan yhteisjakaumasta ja odotusarvot olisi voinut päätellä ”ohtaluulla” symmetrian nojalla.
Lasketaan lopuksi vielä tulon odotusarvo

$$\mathbb{E}(XY) = \sum_i \sum_j ij \cdot \mathbb{P}(\{X=i\} \cap \{Y=j\}) = \frac{1}{3} \cdot 1 \cdot 3 + \frac{1}{3} \cdot 2 \cdot 5 = \frac{13}{3}.$$

Nyt kaikki on laskettu korrelaatiokerrointa varten. Korrelaatio(kerroin) on

$$\rho(X,Y) = \frac{\frac{13}{3} - 3}{\sqrt{\frac{2}{3}}\sqrt{\frac{8}{3}}} = 1.$$

c) Koska korrelaatio on 1, niin X:n ja Y:n välillä vallitsee *lineaarinen riippuvuus*, joten $aX + bY = c$ todennäköisyydellä yksi joillekin vakioille a, b, c. Sijoittamalla yhtälöön pisteet $(0,1)$ ja $(1,3)$ saadaan yhtälöt $b = c$ ja $a+3b = c$, joten $a = -2c$. Valitsemalla $c = 1$ nähdään, että $Y = 2X + 1$ todennäköisyydellä yksi.

Kaavoja

Todennäköisyyden ominaisuuksia

$$\begin{aligned}
&\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),\\
&\mathbb{P}(A \backslash B) = \mathbb{P}(A \cap \overline{B}) = \mathbb{P}(A) - \mathbb{P}(A \cap B),\\
&\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A),\\
&\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B),\\
&\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}\\
&\mathbb{P}(B) = \sum_{k=1}^{n} \mathbb{P}(B|A_k)\mathbb{P}(A_k)
\end{aligned}$$

Odotusarvoja ja variansseja

Ptnf. tai tf.	$\mu_X := \mathbb{E}(X)$	$\sigma_X^2 := \mathrm{Var}(X)$
$\mathbb{P}(X = x)$	$\sum_x x\mathbb{P}(X = x)$	$\sum_x (x - \mu_X)^2 \mathbb{P}(X = x)$
$f_X(x)$	$\int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x$	$\int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\mathrm{d}x$
$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$
$p(1-p)^{x-1}$	$1/p$	$(1-p)/p^2$
$\frac{a^x}{x!}\mathrm{e}^{-a}$	a	a
$1/(b-a)$	$(a+b)/2$	$(b-a)^2/12$
$\theta \mathrm{e}^{-\theta x}$	$1/\theta$	$1/\theta^2$
$\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y), \quad \mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Eräitä testimuuttujia

$$\begin{aligned}
&\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ (likimain, kun ”n on suuri”)},\\
&\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},\\
&\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}} \sim t_{n+m-2},\\
&\sqrt{n-1}s_x \frac{\frac{S_{xY}}{s_{xx}} - \beta}{S_r} \sim t_{n-2}
\end{aligned}$$

Regressio, korrelaatio ja kovarianssi

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}};\ s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y});\ s_{xx} = s_x^2;$$

$$y = a + bx; \quad b = \frac{s_{xy}}{s_{xx}}; \quad a = \overline{y} - b\overline{x};$$

$$s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = \frac{n-1}{n-2}(1 - r^2)s_{yy};$$

$$\sigma_{XY} = \mathrm{Cov}(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))), \quad \sigma_{XX} = \sigma_X^2;$$

$$= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

$$\rho(X,Y) = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$