Tilastomatematiikka

Toinen välikoe 8.3.2021

1. Tarkastellaan Remdesivir-lääkehoidon vaikutusta vakavasti sairaisiin koronapotilaisiin. Hengityskonehoitoa saavilla potilailla Remdesivir-lääkettä sai 5 potilasta ja placeboa sai 4 potilasta. Mitattiin hengityskonehoitoa vaatinut aika kullekin potilaalle ja saatiin seuraavat havainnot

Hoitomuoto	Hoitoaika (vrk)				
Remdesivir	3	6	10.5	12	13.5
Placebo	8	16	19	21	

Oletetaan, että hoitoajat ovat normaalijakautuneita ja että populaatiovarianssit ovat samat. Testaa riskitasoa 5 %, onko lääkityksellä vaikutusta hengityskonehoidon aikaan.

a) Muotoile tilanteeseen sopivat hypoteesit. (1p)

b) Mikä testimuuttuja sopii tähän tilanteeseen? Määrää testimuuttujan saama arvo otoksessa ja kriittisen alueen raja (kynnysarvo) r_0. (3p)

c) Mikä on johtopäätös? Mitä mieltä olet tutkimuksesta? (2p)

2. Tarkastellaan viikoittaisia koronavirustartuntoja Suomessa vuoden 2021 alusta alkaen. Viikolla x havaittujen tartuntojen lukumäärä y yhdeksällä ensimmäisellä viikolla oli

x	1	2	3	4	5	6	7	8	9
y	1665	1821	1972	2474	2454	2672	3044	3478	4513

a) Piirrä muuttujia x ja y vastaava sirontakuvio. Mitä voit sanoa muuttujien välisestä korrelaatiosta ja riippuvuudesta sen perusteella?

b) Valitaan selitettäväksi muuttujaksi $\ln y$ (muuttujan y luonnollinen logaritmi) ja selittäväksi muuttujaksi x. Määrää havaintoja x ja $\ln y$ vastaava regressiosuora.

c) Määrää b)-kohdassa saamasi mallin selitysaste. Mitä voit sanoa mallin sopivuudesta sen perusteella? Laske mallin antama ennuste koronavirustartuntojen määrälle viikolla 10.

3. Eräiden satunnaismuuttujien X ja Y yhteisjakauma on

$X\backslash Y$	0	1	2
0	1/4	1/4	1/4
1	0	1/4	0

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

b) Laske muuttujien X ja Y odotusarvo. (2p)

c) Laske muuttujien X ja Y kovarianssi. Ovatko X ja Y riippumattomia? (3p)

4*. Anna arvio tehtävistä 1-3 saamastasi pistemäärästä. Jos arviosi on pisteen sisällä todellisista pisteistä, saat yhden lisäpisteen.

Kaavoja

Todennäköisyyden ominaisuuksia

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

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

$$\begin{aligned}
&r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}};\ s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y});\ s_{xx} = s_x^2;\\
&y = a + bx; \quad b = \frac{s_{xy}}{s_{xx}}; \quad a = \overline{y} - b\overline{x};\\
&s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = \frac{n-1}{n-2}(1 - r^2)s_{yy};\\
&\sigma_{XY} = \mathrm{Cov}(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))), \quad \sigma_{XX} = \sigma_X^2;\\
&\qquad = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\\
&\rho(X,Y) = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}
\end{aligned}$$

Tehtävien ratkaisuperiaatteet

1. a) Tässä tarkastellaan kahta riippumatonta otosta, joten on järkevää tarkastella odotusarvojen erotusta koskeavia hypoteeseja. Olkoon μ_X Remdesivir-hoitoa vastaavan hoitoajan odotusarvo ja vastaavasti μ_Y placebo-hoitoa vastaavan hoitoajan odotusarvo. Sopivat hypoteesit ovat

$$H_0 : \mu_X - \mu_Y = 0,$$
$$H_1 : \mu_X - \mu_Y < 0,$$

sillä varmaankin oletetaan, että lääkehoidolla saavutetaan jotain hyötyä.

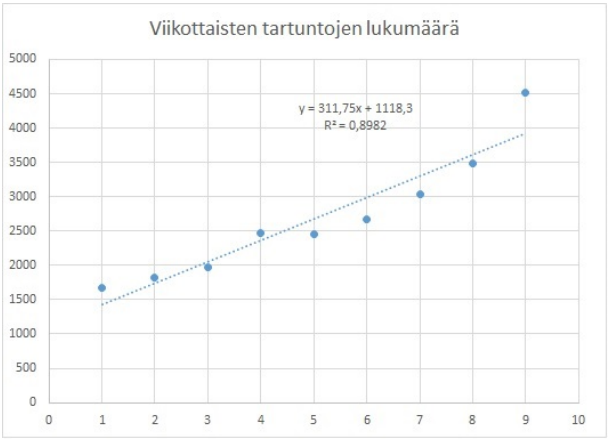
b) Sopiva testimuuttuja löytyy kaavakokoelmasta ja on

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

Testimuuttujan arvo $t = -2.09$ otoksessa saadaan Excelistä. Koska kyseessä on 1-suuntainen "vasemmalle kallellaan oleva" vastahypoteesi, saadaan kriittisen alueen rajaksi $r_0 = -1.895$ joko taulukosta tai Excelistä, sillä vapausasteet ovat $4 + 5 - 2 = 7$.

c) Johtopäätös riskitasolla 5 % on, että meillä on näyttöä nollahypoteesia vastaab. Tämän otoksen perusteella näyttäisi siis siltä, että Remdesivir-hoito vähentää hengityskoneessaoloaikaa. Johtopäätökseen kannattaa suhtautua varauksella, sillä otoskoko on aika pieni. Remdesivir-hoito on vain yksi tekijä muiden joukossa ja populaation homogeenisuuden voi kyseenalaistaa. Myös normaalijakaumaoletukseen voi suhtautua kriittisesti.

2. a) Kuvaajasta nähdään, että muuttujien välillä on voimakas positiivinen korrelaatio ja että riippuvuus näyttäisi tästä huolimatta olevan käyräviivaista.



Kuva 1: Excelin Scatter Chart-toiminnolla piirretty sirontakuvio, jossa on mukana regressiosuoran yhtälö ja sen selitysaste

b) Kuvassa 2 on Excelin antama Regression-tulostus datalle, jossa selittävänä muuttujana on viikon järjestysnumero ja selitettävänä muuttujana viikottaisten tartuntojen logaritmi. Tulostuksesta nähdään regressiosuoran yhtälö $\ln y = 7.27 + 0.11x$.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0,980584					
R Square	0,961546					
Adjusted R Square	0,956052					
Standard Error	0,067086					
Observations	9					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	0,787761	0,787761	175,0352	3,29483E-06	
Residual	7	0,031504	0,004501			
Total	8	0,819265				
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	7,272753	0,048737	149,224	1,6E-13	7,157508293	7,3879984
Viikko	0,114583	0,008661	13,23009	3,29E-06	0,094103791	0,1350629

Kuva 2: Excelin Regression-tulostus tehtävän 2 datalle

c) Regression-tulostuksesta saadaan selitysasteeksi $r^2 = 0.96$, joka on erittäin hyvä. Tämän mukaan malli sopii erinomaisesti dataan. Tämä ei ole sattumaa, sillä matemaattisen mallin perusteella tartunnat noudattavat eksponentiaalista kasvua, lähes koko populaatio on altis tartunnoille.
Mallin mukaan tartuntojen lukumäärälle pätee

$$\hat{y} = \hat{y}(x) = \mathrm{e}^{7.27+0.11x} = \underbrace{\mathrm{e}^{7.27}}_{=C} \cdot \mathrm{e}^{0.11x},$$

josta tartuntojen lukumääräksi viikolla 10 saadaan $\hat{y}(10) \approx 4530$.

3. a) Satunnaisvektorin arvojoukko on $S_{XY} = \{(0,0), (0,1), (0,2), (1,1)\}$.

b) Koska mahdollisia arvoja on vain kourallinen, saadaan reunajakaumat näppärästi täydentämällä taulukkoa summaamalla pistetodennäköisyydet riveittäin ja sarakkeittain yhteen, jolloin saadaan

$X \backslash Y$	0	1	2	$p_i = \sum_i p_{ij}$
0	1/4	1/4	1/4	3/4
1	0	1/4	0	1/4
$q_j = \sum_i p_{ij}$	1/4	1/2	1/4	$\sum_{i,j} p_{ij} = 1$

Taulukko 1: Satunnaisvektorin (X, Y) pistetodennäköisyydet ja reunajakaumien pistetodennäköisyydet

Reunajakaumien odotusarvot voidaan laskea joko suoraan yhteisjakaumasta tai jo lasketuista reunajakaumista. Esitetään tässä oleellisesti molemmat tavat:

$$\mathbb{E}(X) = \sum_{i,j} i \cdot \underbrace{\mathbb{P}(\{X = i\} \cap \{Y = j\})}_{=p_{ij}} = \sum_i i \cdot \underbrace{\sum_j p_{ij}}_{=p_i = \mathbb{P}(X=i)} = \frac{1}{4}.$$

Jos viimeistä edellinen yhtäsuuruus jätetään pois, on odotusarvo laskettu suoraan yhteisjakaumasta. Jos taas kyseinen yhtäsuuruus on mukana, on laskussa hyödynnetty reunajakaumaa.

Vastaavasti saadaan Y:n odotusarvoksi

$$\mathbb{E}(Y) = \sum_{i,j} j \cdot p_{ij} = \sum_j j q_j = 0 \cdot 1/4 + 1 \cdot 1/2 + 2 \cdot 1/4 = 1.$$

c) Määritelmän ja kaavakokoelman mukaan kovarianssi on

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Koska muuttujien X ja Y odotusarvot on jo laskettu, riittää laskea tulon XY odotusarvo. Näppärimmin sen saa taulukosta 1. Nähdään, ettei "eloon jää" kuin yksi yhteenlaskettava, eli termi $ij \cdot p_{ij} = 1 \cdot 1 \cdot p_{11} = 1/4$. Niinpä kovarianssi on

$$\mathrm{Cov}(X, Y) = 1/4 - 1/4 = 0.$$

Muuttujat eivät kuitenkaan ole riippumattomia, sillä esimerkiksi

$$p_{10} = \mathbb{P}(\{X = 1\} \cap \{Y = 0\}) = 0 \neq p_1 \cdot q_0 = \mathbb{P}(X = 1) \cdot \mathbb{P}(Y = 0) = 1/4 \cdot 3/4.$$