Tilastomatematiikka

Toinen välikoe 12.03.2020

1. Oletetaan, että käytössä oleva lääke auttaa 70 % potilaista. Jos otetaan uusi lääke käyttöön ja se auttaa 148 potilasta 200 potilaasta, joilla sitä testattiin, niin voidaanko sanoa, että uusi lääke on parempi kuin käytössä oleva? Käytä riskitasoa 5 %.

   a) Muotoile tilanteeseen sopivat hypoteesit.

   b) Mikä testimuuttuja sopii tähän tilanteeseen? Määrää testimuuttujan avulla kriittisen alueen raja (kynnysarvo) r_0.

   c) Tee johtopäätös sekä kynnysarvon että p-arvon perusteella.

2. Tutkittiin kvanttitunneloitumista kiinteässä heliumissa tarkastelemalla, kuinka suuri osuus y epäpuhtauksia läpäisee kiinteän heliumin eri lämpötiloissa x [$^\circ C$] ja saatiin seuraava havaintoaineisto

x	-260.5	-255.7	-264.6	-265.0	-270.0	-272.0	-272.5	-272.6	-272.8	-272.9
y	0.425	0.224	0.453	0.475	0.705	0.860	0.935	0.961	0.979	0.990

   a) Määrää havaintoja vastaava regressiosuora ja korrelaatiokerroin.

   b) Mikä on regressiosuoran selitysaste? Mitä voit sanoa mallin sopivuudesta tämän perusteella?

   c) Laske mallin antama ennuste, kun $x = -275$. Mitä voit sanoa mallin antamasta ennusteesta?

Valitse vain toinen seuraavista tehtävistä. Tähtitehtävästä 3* voi saada 8 pistettä.

3. Tutkittiin koronaviruksen COVID-19 itämisaikaa tarkastelemalla 25 virustartunnan saanutta, joiden itämisajan keskiarvoksi saatiin 4.5 päivää ja keskihajonnaksi 2.3 päivää. Oletetaan, että itämisajat ovat toisistaan riippumattomia ja samalla tavalla normaalijakautuneita.

   a) Määrää itämisajalle sopiva piste-estimaatti. (1p)

   b) Laske itämisajan odotusarvon 95 % luottamusväli. (4p)

   c) Mitä voit sanoa käytetyistä mallioletuksista? (1p)

3*. Olkoot $X, Y \sim N(0,1)$ riippumattomia satunnaismuuttujia ja olkoon $Z = X + Y$ satunnaismuuttujien X ja Y summa.

   a) Määrää satunnaisvektorin (X, Z) kovarianssimatriisi. (5p)

   b) Tarkastellaan lineaarista muunnosta $(V, W) = (X, Z) \cdot \mathbf{A}$, missä $\mathbf{A}$ on matriisi

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

   Päättele mitä jakaumaa (V, W) noudattaa laskemalla yhteisjakauman kertymäfunktio ja tiheysfunktio. (3p)

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) Koska olemme kiinnostuneita, onko uusi lääke parempi ja toisaalta lääkkeen auttamien suhteellinen osuus, asetetaan binomijakauman parametrille p hypoteesit

$$H_0 : p = 0.70,$$
$$H_1 : p > 0.70.$$

b) Merkitään X = ”lääkkeen auttavien lkm. otoksessa” ja oletetaan, että lääke auttaa kutakin potilasta muista potilaista riippumatta samalla tavalla, jolloin $X \sim \text{Bin}(n, p)$, missä $n = 200$ ja p on estimoitava parametri. Sopiva estimaattori p:lle on $p^* = \frac{X}{n}$, jolle $\mathbb{E}(p^*) = p$ ja $\text{Var}\left(\frac{p(1-p)}{n}\right)$. Sopiva testimuuttuja löydetään standardoimalla

$$Z = \frac{p^* - \mathbb{E}(p^*)}{\sqrt{\text{Var}(p^*)}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\text{likimain}}{\sim} N(0,1)$$ keskeisen raja-arvolauseen nojalla.

Koska vastahypoteesi $H_1 : p > 0.70$ on oikealle yksisuuntainen, kynnysarvo r_0 saadaan ehdosta

$$\mathbb{P}(Z > r_0) = 0.05 \Leftrightarrow \mathbb{P}(Z \leq r_0) = 0.95 \Rightarrow r_0 = 1.645.$$

c) Testimuuttujan arvo otoksessa on

$$z \overset{H_0 \text{ tosi}}{=} \frac{\frac{148}{200} - 0.7}{\sqrt{\frac{0.7(1-0.7)}{200}}} \approx 1.23.$$

Koska $z = 1.23 < 1.645 = r_0$, johtopäätös on H_0, eli uutta lääkettä ei voida pitää parempana merkitsevyystasolla 5 %.
Samaan johtopäätökseen päädytään myös p-arvon p perusteella, sillä

$$p = \mathbb{P}(Z > 1.23) = 1 - \Phi(1.23) \approx 0.10 \geq 0.05.$$

2. a) Syöttämällä data laskimeen ja käyttämällä laskimen tilastollisia toimintoja saadaan regressiosuoraksi $y = -11.325 - 0.045x$ ja korrelaatiokertoimeksi $r \approx -0.97$.

b) Selitysaste on $r^2 \approx 0.94$, joten malli selittää 94 % osuuden y satunnaisvaihtelusta, mikä on varsin hyvä tulos.

c) Kun $x = -275$, saadaan mallin antamaksi ennusteeksi $y \approx 1.05$, mikä on järjetön tulos. Mallin käytössä on siis syytä olla varovainen havaintoalueen ulkopuolella.

3. a) Odotusarvolle sopiva piste-estimaatti on keskiarvo $\overline{x} = 4.5$.

b) Koska populaatiohajonta σ on tuntematon, niin sopiva testimuuttuja on

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Määrätään 2-suuntainen symmetrinen 95 % luottamusväli laskemalla ensin sellainen r_0, että $\mathbb{P}(-r_0 \leq T \leq r_0) = 0.95$. Koska vapausasteiden lukumäärä on $n - 1 = 24$, saadaan t-jakauman taulukosta $r_0 = 2.064$.
Muokataan epäyhtälöä

$$-r_0 \leq T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq r_0 \Leftrightarrow \overline{X} - r_0 \cdot S/\sqrt{n} \leq \mu \leq \overline{X} + r_0 \cdot S/\sqrt{n}.$$

Sijoittamalla havaintoarvot $\overline{x} = 4.5$ ja $s = 2.3$ oikean puolen epäyhtälöön saadaan luottamusväliksi

$$I_\mu = [4.5 - 2.064 \cdot 2.3/\sqrt{25}, 4.5 + 2.064 \cdot 2.3/\sqrt{25}] = [3.55, 5.45]$$

kahden desimaalin tarkkuudella.

c) Riippumattomuusoletus lienee kunnossa, sillä tartunnan itäminen etenee yksilössä omana prosessinaan, mihin toisen yksilön itämisajalla ei ole vaikutusta. Samalla tavalla jakautuneisuus ei välttämättä ole totta, sillä yksilötasolla itämisaika saattaa olla hyvinkin erilainen. Samasta syystä hajonta voi vaihdella yksilöstä toiseen. Normaalijakaumaoletus voi olla kyseenalainen, sillä itämisaika on positiivinen ja otoksen keskiarvo on aika lähellä nollaa sekä hajonta kohtuullisen suurta. Lisäksi itämisajan määrittämisessä saattaa olla suurta epävarmuutta.

3*. a) Määritelmän mukaan kovarianssimatriisi on

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{XX} & \sigma_{XZ} \\ \sigma_{ZX} & \sigma_{ZZ} \end{pmatrix},$$

missä

$$\sigma_{AB} = \mathrm{Cov}(A, B) = \mathbb{E}(AB) - \mathbb{E}(A)\mathbb{E}(B), \quad A, B \in \{X, Z\},$$

on muuttujien A ja B välinen kovarianssi. Huomaa, että $\sigma_{AA} = \mathrm{Var}(A)$.
Oletuksien $X, Y \sim N(0, 1)$ ja $Z = X + Y$ mukaan

$$\sigma_{XX} = \mathrm{Var}(X) = 1 \quad \text{ja} \quad \sigma_{ZZ} = \mathrm{Var}(X + Y) = \sigma_{XX} + \sigma_{YY} = 2,$$

sillä X ja Y ovat riippumattomia. Edelleen riippumattomuuden ja symmetrian nojalla

$$\sigma_{ZX} = \sigma_{XZ} = \mathbb{E}(XZ) - \mathbb{E}(X)\mathbb{E}(Z) = \mathbb{E}(X(X + Y)) = \mathbb{E}(X^2) - \mathbb{E}(X)\mathbb{E}(Y) = 1,$$

sillä $\mathbb{E}(X) = 0$ ja $1 = \sigma_{XX} = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.
Edellisen perusteella satunnaisvektorin (X, Z) kovarianssimatriisi on

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

b) Tehtävänannon perusteella $V = X$ ja $W = -X + Z = -X + X + Y = Y$. Koska X ja Y ovat riippumattomat, saadaan kertymäfunktioksi

$$F_{V,W}(v, w) = \mathbb{P}(\{X \leq v\} \cap \{Y \leq w\}) = \mathbb{P}(X \leq v)\mathbb{P}(Y \leq w) = F_X(v)F_Y(w).$$

Oletuksen $X, Y \sim N(0, 1)$ nojalla saadaan derivoimalla edellisestä

$$\begin{aligned} f_{V,W}(v, w) &= \frac{\partial^2}{\partial v \partial w} F_{V,W}(v, w) = f_X(v)f_Y(w) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}v^2} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}w^2} \\ &= \frac{1}{2\pi} \mathrm{e}^{-\frac{1}{2}(v^2 + w^2)}, \end{aligned}$$

joten (V, W) noudattaa 2-ulotteista standardinormaalijakaumaa.