Tilastomatematiikka

Toinen välikoe 17.03.2022

1. Tarkasteltiin kahden eri rengasmerkin kulumista ajamalla renkaat loppuun ja mittaamalla renkaiden eliniät yksikkönä tuhat kilometriä. Saatiin seuraavat havainnot.

Merkki 1	32.1	33.5	36.2	36.9	37.2	38.2	45.3	48.4
Merkki 2	32.0	33.2	34.3	35.5	37.8	38.0	42.3	47.8

Muotoile sopivat hypoteesit ja testaa ne 99 % luottamustasolla, kun olemme kiinnostuneita onko renkaiden keskimääräisessä kulumisessa eroa merkkien välillä. Kumpaa rengasmerkkiä suosittelisit havaintojen perusteella? Oletetaan, että rengasmerkkien $i = 1, 2$ eliniät noudattavat likimain normaalijakaumaa $N(\mu_i, \sigma^2)$.

2. Tutkittiin kaasun paineen p ja tilavuuden V välistä yhteyttä ja saatiin seuraava havaintoaineisto

V [cm^3]	50	60	70	90	100
p [kg/cm^3]	64.6	51.2	40.4	25.8	7.9

a) Piirrä havaintoja vastaava sirontakuvio. Mitä voit tämän perusteella sanoa muuttujien välisestä riippuvuudesta? (1p)

b) Oletetaan, että paine p ja tilavuus V noudattavat tilanyhtälöä

$$pV^n = C, \tag{1}$$

missä n ja C ovat tuntemattomia parametreja. Estimoi parametreja määräämällä regressiosuora muuttujien $y = \ln p$ ja $x = \ln V$ välille. Mikä on regressiosuoran selitysaste ja sen tulkinta? (3p)

c) Laske (b)-kohdasta saadun mallin (1), missä n ja C ovat regressiosuoran avulla saadut estimaatit, mukainen ennuste paineelle p, kun tilavuus on $V = 40$. (2p)

3. Eräiden satunnaismuuttujien X ja Y yhteisjakauma on

$X\backslash Y$	-1	0	1
-1	1/25	5/25	4/25
0	2/25	4/25	4/25
1	2/25	1/25	2/25

Taulukko 1: Satunnaisvektorin (X, Y) pistetodennäköisyydet

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

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

c) Ovatko muuttujat X ja Y riippumattomat? (1p)

Tehtävien ratkaisuperiaatteet

1. Oletetaan, että otokset ovat toisistaan riippumattomat. Tämän oletuksen paikkansa pitävyys riippuu itse testauksesta. Saattaisi olla perusteltua verrata havaintoja pareittain toisiinsa, jos renkaat on pareittain ajettu loppuun samalla autolla samoissa olosuhteissa. Koska emme tätä tiedä, oletetaan, että koejärjestely on tehty niin, että otokset ovat riippumattomat.

   Koska haluamme tietää, onko renkaiden keskimääräisessä kulumisessa eroa, tehdään odotusarvojen erotuksen testi. Testataan hypoteesit

   $$H_0 : \mu_1 - \mu_2 = 0,$$
   $$H_1 : \mu_1 - \mu_2 \neq 0.$$

   Tehtävänannon mukaan populaatiovarianssi σ^2 on molemmille otoksille sama, joten tilanteeseen sopiva testimuuttuja

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

   löytyy kaavakokelmasta.

   Koska vastahypoteesi on kaksisuuntainen ja luottamustaso on 99 %, määrätään joko taulukosta tai numeerisesti kynnysarvo r_0 ehdosta $\mathbb{P}(|T| \leq r_0) = 0.99$. Koska $n = m = 8$, ovat vapausasteet 14, joten t-jakauman taulukon tai Excelin mukaan $r_0 = 2.977$.

   Nollahypoteesin vallitessa testimuuttujan arvoksi otoksessa saadaan $\widehat{T} = 0.318$.

   Johtopäätös: Koska Excelin tulostama p-arvo $\mathbb{P}(|T| > \widehat{T}) = 0.76 > 0.01$ tai yhtälailla koska $|\widehat{T}| = 0.318 < 2.977 = r_0$, ei nollahypoteesia ole syytä hylätä. Tämän otoksen perusteella kumpikaan rengasmerkki ei ole toista parempi, joten käyttäjän kannalta on yhdentekevää kumman rengasmerkin valitsee, jos kriteerinä käyttää kulumista. Kannattaa huomata, että pelkkä keskiarvojen vertailu ei riitä, koska kyseessä on merkkiä kohden vain yksi otos, jonka otoskoko on suhteellisen pieni. Jossakin toisessa otoksessa keskiarvot voivat olla aivan toisenlaiset.

   Alla on *Excelin Data Analysis*-työkalusta löytyvä tulostus odotusarvojen erotuksen hypoteesien testaukselle, mistä löytyy samat asiat kuin yllä on kuvattu.

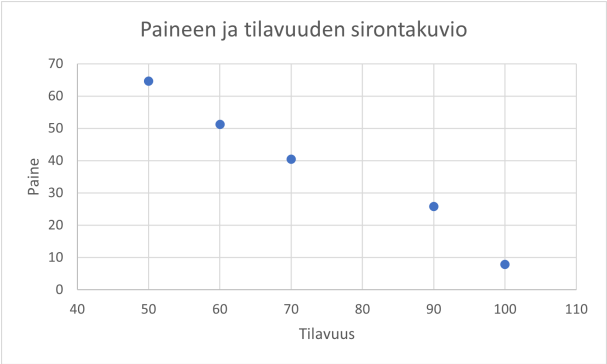
2. a) Kuvassa 1 on Excelin piirtämä sirontakuvio tehtävän datalle. Excelin automaattisesti piirtämän kuvion akseleita kannattaa usein skaalata uudelleen, jotta saa havainnollisemman kuvan. Sama pätee millä tahansa muullakin softalla tai käsin piirrettyyn kuvaan. Sirontakuvion perusteella muuttujien välillä näyttäisi olevan *lineaarinen riippuvuus*. Toki piirtotarkkuuden rajoissa myös muut riippuvuudet ovat mahdollisia. Joka tapauksessa muuttujilla näyttää olevan vahva riippuvuus toisistaan.

   b) Siirrytään tehtävänannon siivittämänä tarkastelemaan uusia muuttujia $y = \ln p$ ja $x = \ln V$. *Excelin Data Analysis*-työlalun *Regression*-toiminnolla saadaan regressiosuora

   $$\ln p = 14.70 - 2.64 \ln V.$$

t-Test: Two-Sample Assuming Equal Variances		
	Merkki 1	*Merkki 2*
Mean	38,475	37,6125
Variance	31,405	27,47839
Observations	8	8
Pooled Variance	29,4416964	
Hypothesized Mean Difference	0	
df	14	
t Stat	0,31791256	
P(T<=t) one-tail	0,37762185	
t Critical one-tail	2,62449407	
P(T<=t) two-tail	0,7552437	
t Critical two-tail	2,97684273	

Taulukko 2: Excelin antama tulostus tehtävän 1 datalle



Kuva 1: Excelin piirtämä tehtävän 2 sirontakuvio

Mallin selitysaste on Excelin mukaan 82 %, joka tarkoittaa, että $x = \ln V$ selittää 82 % $y = \ln p$:n satunnaisvaihtelusta. Selitysaste on hyvä, eli sen mukaan malli sopii hyvin dataan. Kannattaa kuitenkin huomata, että pelkkä selitysaste on huono mittari kuvaamaan mallien välistä paremmuutta.

c) Edellisen kohdan mukaan $\ln p = 14.70 - 2.64 \ln V$, josta eksponenttifunktion ja logaritmifunktion laskusääntöjen perusteella saadaan

$$p = \mathrm{e}^{\ln p} = \mathrm{e}^{14.70 - 2.64 \ln V} = \mathrm{e}^{14.70 + \ln V^{-2.64}} = \mathrm{e}^{14.70} \cdot V^{-2.64}.$$

Kun $V = 40$, saadaan mallin antamaksi ennusteeksi

$$\widehat{p} = \mathrm{e}^{14.70} \cdot 40^{-2.64} \approx 143.$$

Ennuste voi olla järkeväkin, mutta yleisesti ekstrapolointi havaintovälin ulkopuolelle voi ”mennä pahasti metsään”.

3. a) Taulukon mukaan arvojoukko on

$$S_{X,Y} = \{(-1,-1), (-1,0), (-1,1), \ldots, (1,1)\},$$

jossa on 9 alkiota.

b) Täydennetään tehtävänannon yhteisjakaumaa reunajakaumilla laskemalla pistetodennäköisyydet riveittäin ja sarakkeittain yhteen (eli eliminoimalla toinen muuttuja pois)

$X\backslash Y$	-1	0	1	$p_i = \mathbb{P}(X = i)$
-1	$1/25$	$5/25$	$4/25$	$2/5$
0	$2/25$	$4/25$	$4/25$	$2/5$
1	$2/25$	$1/25$	$2/25$	$1/5$
$q_j = \mathbb{P}(Y = j)$	$1/5$	$2/5$	$2/5$	$\Sigma = 1$

Käytetään kovarianssin laskemiseen kaavakokoelmasta löytyvää kaavaa

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

Koordinaattimuuttujien X ja Y odotusarvot on nyt helppo laskea täydennetystä taulukosta. Odotusarvoiksi saadaan

$$\mathbb{E}(X) = \sum_i i \cdot p_i = (-1) \cdot \frac{2}{5} + 0 \cdot \frac{2}{5} + 1 \cdot \frac{1}{5} = -\frac{1}{5}$$

ja

$$\mathbb{E}(Y) = \sum_j j \cdot q_j = (-1) \cdot \frac{1}{5} + 0 \cdot \frac{2}{5} + 1 \cdot \frac{2}{5} = \frac{1}{5}.$$

Tulon odotusarvon laskennassa huomataan kuten edellä, että tulo XY on nolla, jos jompi kumpi tulon tekijöistä on nolla. Siten suurin osa termeistä ”kuolee pois”. Jäljelle jää 4 yhteenlaskettavaa ja saadaan

$$\mathbb{E}(XY) = \sum_{i,j} ij \cdot p_{ij} = (-1)\cdot(-1)\cdot\frac{1}{25} + (-1)\cdot 1\cdot\frac{4}{25} + 1\cdot(-1)\cdot\frac{2}{25} + 1\cdot 1\cdot\frac{2}{25} = -\frac{3}{25}.$$

Edellisten laskujen perusteella kovarianssi on

$$\mathrm{Cov}(X, Y) = -\frac{3}{25} - \left(-\frac{1}{5}\right) \cdot \frac{1}{5} = -\frac{2}{25}.$$

c) Edellisen kohdan perusteella kovarianssi ei häviä, joten *muuttujat eivät ole riippumattomat*. Sama asia voidaan todeta myös taulukosta. Esimerkiksi

$$p_{11} = \mathbb{P}(\{X = 1\} \cap \{Y = 1\}) = \frac{1}{25} \neq \frac{2}{5} \cdot \frac{1}{5} = \mathbb{P}(X = 1) \cdot \mathbb{P}(Y = 1).$$

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

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