Aalto-yliopisto

Gripenberg, Kiiski

MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi
2. välikoe 10.12.2014

Kirjoita jokaiseen koepaperiin nimesi, opiskelijanumerosi ym. tiedot!
Laskin, I. Mellinin tilastolliset taulukot, Matlab/Octave-funktiolista ja muistiinpanolappu ovat sallittuja apuvälineitä!
Kirjoita välivaiheet näkyviin!
Käytä joko taulukoita tai kirjoita millä Matlab/Octave-komennoilla voisit laskea tarvittavat kertymäfunktioiden tai niiden käänteisfunktioiden arvot ja minkälaisia päätelmiä, esimerkiksi merkitsevyystasojen suhteen, näiden arvojen perusteella voisit tehdä.

1. Porkkanapussin painon odotusarvon selvittämiseksi otettiin 20 pussin otos, pussit punnittiin jolloin niiden painojen keskiarvoksi tuli 501 g ja otosvarianssiksi 45 g^2. Kun sitten tavanomaiseen tapaan laskettiin porkkanapussin painon odotusarvon (symmetrinen) luottamusväli, olettaen että pussin paino on $\mathrm{N}(\mu, \sigma^2)$-jakautunut, saatiin tulokseksi $[498.41, 503.59]$.

(a) Määritä tämän luottamusvälin luottamustaso.
(b) Mitä tämä luottamusväli ja luottamustaso merkitsevät?

Ratkaisu: (a) Jos satunnaismuuttujasta $X \sim \mathrm{N}(\mu, \sigma^2)$ on otettu otos X_j, $j = 1, 2, 3 \ldots, n$ niin

$$\left[\overline{X} - F^{-1}_{\mathrm{t}(n-1)}(1 - \tfrac{1}{2}\alpha)\sqrt{\frac{S^2}{n}}, \overline{X} + F^{-1}_{\mathrm{t}(n-1)}(1 - \tfrac{1}{2}\alpha)\sqrt{\frac{S^2}{n}}\right]$$

on odotusarvon μ symmetrinen luottamusväli luottamustasolla $1 - \alpha$. Tässä $\overline{X}$ on keskiarvo ja S^2 on otosvarianssi. Luottamusvälin pituus on siis

$$2F^{-1}_{\mathrm{t}(n-1)}(1 - \tfrac{1}{2}\alpha)\sqrt{\frac{S^2}{n}},$$

ja koska tässä tapauksessa $n = 20$, $s^2 = 45$ ja luottamsuvälin pituus on $503.59 - 498.41 = 5.18$ niin

$$F^{-1}_{\mathrm{t}(20-1)}(1 - \tfrac{1}{2}\alpha) = \frac{5.18}{2 \cdot \sqrt{\frac{45}{20}}} = 1.7267.$$

Näin ollen luottamustaso on

$$1 - \tfrac{1}{2}\alpha = F_{\mathrm{t}(19)}(1.7267) = 0.94978$$

josta saadaan luottamsutasoksi

$$1 - \alpha = 2 \cdot 0.94978 - 1 \approx 0.9.$$

(b) Jos porkkanapusseista otetaan otos ja lasketaan painon odotusarvon luottamusväli, niin todennäköisyys, että saadaan väli johon odotusarvo kuuluu on luottamustaso. Sitten kun väli on laskettu ja tulokseksi on saatu esimerkiksi $[498.41, 503.59]$ kuten tässä, niin odotusarvo, joka on luku, eikä satunnaismuuttuja, joko kuuluu tähän väliin tai sitten ei, mutta jos tämä toistetaan esimerkiksi 1000 kertaa ja luottamustaso on 0.9 kuten tässä niin noin 900 kertaa käy niin, että odotusarvo kuuluu luottamusväliin.

2. Paikkakunnat A ja B ovat erään hankkeen vaikutuspiirissä. Hankkeen kannattajat ovat ahkerasti tehneet työtä hankeen puolesta paikkakunnalla A joten oletetaan yleisesti, että sen kannatus on suurempi paikkakunnalla A kuin paikkakunnalla B. Kampanjoinnin vaikutuksen selvittämiseksi kysyttiin satunnaisesti valituilta henkilöiltä mitä mieltä he ovat tästä hankkeesta. Paikkakunnalla A haastateltiin 130 henkilöä ja heistä 70 sanoivat kannattavansa hanketta ja paikkakunnalla B haastateltiin 120 henkilöä ja heistä 65 sanoivat vastustavansa tätä hanketta.

(a) Testaa merkitsevyystasolla 0.05 nollahypoteesia, että hankkeen kannatus paikkakunnalla B on vähintään yhä suuri kuin paikkakunnalla A.

(b) Onko (a)-kohdassa valittu nollahypoteesi järkevä? Perustele!

Ratkaisu: (a) Olkoon p_A todennäköisyys, että henkilö paikkakunnalla A kannattaa hanketta ja p_B todennäköisyys, henkilö paikkakunnalla B kannattaa hanketta. Nollahypoteesi on siis $p_A \leq p_B$ ja laskuissa käytämme ääritapausta $p_A = p_B$. Nollahypoteesi ei kerro mikä tämä todennäköisyys on mutta voimme käyttää sen estimaattina $\hat{p} = (70 + 55)/(130 + 120) = 0.5$.

Olkoon X satunnaismuuttuja joka saa arvon 1 jos henkilö paikkakunnalla A kannattaa hanketta ja muuten 0 ja Y on vastaavasti satunnaismuuttuja joka saa arvon 1 jos henkilö paikkakunnalla B kannattaa hanketta ja muuten 0. Meillä on nyt otokset X_j, $j = 1, \ldots, 130$ ja Y_j, $j = 1, \ldots, 120$ ja testimuuttujaksi otamme

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_y} + \frac{1}{n_y})}} \sim_a \mathrm{N}(0,1).$$

Testimuuttujan arvoksi tulee tässä tapauksessa

$$\frac{\frac{70}{130} - \frac{55}{120}}{\sqrt{0.5 \cdot 0.5 \cdot (1/130 + 1/120)}} = 1.2659,$$

ja p-arvoksi tulee, koska nollahypoteesi on $p_A \leq p_B$,

$$p \approx 1 - F_{\mathrm{N}(0,1)}(1.2659) = 1 - \texttt{normcdf}(1.2659) = 0.103,$$

Koska $p > 0.05$ nollahypoteesia ei hylätä.

(b) Nollahypoteesi $p_A \leq p_B$ on järkevä vastaväitteenä oletukselle että kannatus paikkakunnalla A on suurempi kuin paikkakunnalla B. Jos osoittautuisi, ettei kampanjoinnilla olisi toivottua vaikutusta hankkeen kannattajilla ei olisi syytä toteuttaa samanlaista kampanjaa paikkakunnalla B. Sen sijaan perusteluksi ei käy että kannatus paikkakunnalla A osoittautui suuremmaksi kuin paikkakunnalla B.

---

3. Veren kolesterolin alentamiseksi kehitettiin lääke ja sen vaikutuksen selvittämiseksi mitattiin koehenkilöiden veren kokonaiskolesterolitaso ennen lääkkeen ottamista ja kun lääkettä oli otettu säännöllisesti kolme kuukautta. Tulokset olivat (yksikkönä mmol/l):

Ennen	5.5	7.6	5.7	8.9	5.9	9.8	7.5	6.3
Jälkeen	5.0	6.8	5.1	7.3	5.7	9.1	6.9	6.5

(a) Jos X on koehenkilön kolesterolitaso ennen lääkkeen ottamista ja Y sen jälkeen niin minkälaisella testillä voisit hylätä nollahypoteesin, että X ja Y ovat riippumattomia ja normaalijakautuneita. Sinun ei tarvitse laskea testimuuttujan numeerista arvoa tai testin p-arvoa.

(b) Formuloi järkevä nollahypoteesi koskien kolesterolitason muutoksia ja testaa tätä nollahypoteesia merkitsevyystasolla 0.01.

Ratkaisu: (a) Jos X ja Y ovat riippumattomia ja normaalijakutuneita niin niiden välinen korrelaatio $\rho_{XY} = 0$ joten tämä on nollahypoteesi jota testataan. Silloin testimuuttujaksi otetaan $\frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}}$ joka on t$(n-2)$ jakautunut joten on laskettava korrelaatiokerroin $r_{xy} = \frac{\sum_{j=1}^n (x_j-\overline{x})(y_j-\overline{y})}{\sqrt{\sum_{j=1}^n (x_j-\overline{x})^2 \sum_{j=1}^n (y_j-\overline{y})^2}}$.

(b) Nollahypoteesina voisi olla että kolesterolitason muutos $U = Y - X$ on satunnaismuuttuja joka on N(μ, σ^2) jakautunut ja eri henkilöillä muutokset ovat toisistaan riippumattomia. Vastaväitteenä tavoiteelle alentaa kolesterolitasoa otetaan nollahypoteesiksi $\mu \geq 0$.

Satunnaismuuttujan arvoiksi saadaan

u_j	-0.5	-0.8	-0.6	-1.6	-0.2	-0.7	-0.6	0.2

Keskiarvoksi tulee $\overline{u} = \frac{1}{8}(-0.5 - 0.8 - 0.6 - 1.6 - 0.2 - 0.7 - 0.6 + 0.2) = -0.6$ jolloin otosvarianssiksi tulee

$$s_u^2 = \frac{1}{7}(-0.5+0.6)^2 + (-0.8+0.6)^2 + (-0.6+0.6)^2 + (-1.6+0.6)^2 + (-0.2+0.6)^2 + (-0.7+0.6)^2 + (-0.6+0.6)^2 + (0.2+0.6)^2) = \frac{1.86}{7} = 0.26571.$$

Testimuuttujana meillä on

$$\frac{U-0}{\sqrt{\frac{S^2}{n}}} \sim \text{t}(7),$$

ja testimuuttujan arvoksi tulee

$$\frac{-0.6}{\sqrt{\frac{0.26571}{8}}} = -3.2922.$$

p-arvoksi tulee siten kun nollahypoteesi on $\mu \geq 0$,

$$p = F_{\text{t}(7)}(-3.922) = \texttt{tcdf}(-3.922, 7) = 0.0066.$$

Koska $p < 0.01$ niin nollahypoteesi hylätään merkitsevyystasolla 0.01.

---

4. Haluttiin selvittää mozzarella-juuston kulutuksen ja insinööritieteiden tohtoritutkintojen välistä yhteyttä eräässä maassa ja tätä varten kerättiin eri vuosilta seuraavat havainnot (x_j, y_j), missä x on mozzarella-juuston kulutus (kg/henkilö) ja y on tohtoritutkintojen lukumäärä:

x	4.2	4.4	4.4	4.4	4.5	4.6	4.8	5.0	4.8	4.8
y	480	501	540	552	547	622	655	701	712	708

Oletetaan että, $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ missä satunnaismuuttujat ε_j ovat riippumattomia ja N$(0, \sigma^2)$-jakautuneita. Määritä parametrien β_0 ja β_1 pienimmän neliösumman estimaatit ja testaa nollahypoteesia $\beta_1 = 200$ merkitsevyystasolla 0.01.

Seuraavassa taulukossa on annettu joitakin tunnuslukuja.

$\overline{x}$	$\overline{y}$	$\sum_{j=1}^{10}(x_j-\overline{x})^2$	$\sum_{j=1}^{10}(y_j-\overline{y})^2$	$\sum_{j=1}^{10}(x_j-\overline{x})(y_j-\overline{y})$
4.59	601.8	0.569	70800	190.28

Ratkaisu: Koska

$$s_y^2 = \frac{1}{9}\sum_{j=1}^{10}(y_j - \overline{y})^2 = 7866.7$$

$$r_{xy} = \frac{\sum_{j=1}^{10}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{10}(x_j - \overline{x})^2 \sum_{j=1}^{10}(y_j - \overline{y})^2}} = 0.94803,$$

$$s^2 = \frac{9}{8}s_y^2(1 - r_{xy}^2) = 895.97,$$

niin parametrien β_1 ja β_0 pienimmän neliösumman estimaateiksi tulevat

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{j=1}^{10}(x_j - \overline{x})(y_j - \overline{y})}{\sum_{j=1}^{10}(x_j - \overline{x})^2} = 334.41,$$

$$b_0 = \overline{y} - b_1\overline{x} = -933.14.$$

Kun testataan nollahypoteesia β_1 niin testimuuttujana käytetään

$$W_1 = \frac{B_1 - 200}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \sim \mathrm{t}(n-2).$$

Tämän testimuuttujan arvoksi tulee tässä tapauksessa

$$w_1 = \frac{134.41}{\sqrt{\frac{895.97}{0.569}}} = 3.3872.$$

Koska nollahypoteesi on $\beta_1 = 200$ eli vaihtoehto on kakssisuuntainen niin p-arvoksi tulee

$$p = 2(1 - F_{\mathrm{t}(8)}(3.3872)) = \texttt{2} * (\texttt{1} - \texttt{tcdf}(\texttt{3.3872}, \texttt{8})) = 0.00954.$$

Koska $p < 0.01$ niin nollahypoteesi hylätään.
