MTTTP1

SELITYKSIÄ JA ESIMERKKEJÄ KAAVAKOKOELMAN KAAVOIHIN LIITTYEN

Aineisto kaavojen (1) – (3), (9) ja (11) esimerkkeihin. Lepakot paikallistavat hyönteisiä lähettämällä korkeataajuista ääntä. Ne pystyvät paikallistamaan hyönteiset kaiun kuulemiseen kuluvan ajan perusteella. Tutkijat arvelevat, että keskimääräinen tunnistusmatka olisi yli 35 cm. He keräsivät aineiston mitaten etäisyydet (cm), joista lepakot löysivät hyönteisiä. Mitatut etäisyydet olivat 62, 52, 68, 23, 40.

(1) Muuttujan *x* keskiarvo $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Esim.

Etäisyyksien keskiarvo $\bar{x}$ = (62+52+68+23+40)/5 = 49.

(2) Muuttujan *x* varianssi $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{SS_x}{n-1}$.

Voidaan merkitä myös s^2.

Esim.

Etäisyyksien varianssi s^2 ={ $(62-49)^2$+$(52-49)^2$+$(68-49)^2$+$(23-49)^2$+ $(40-49)^2$}/4 = 1296/4 = 324. Nyt siis *SS* (= *SSx*) = 1296, *n* = 5.

Toisin

$\sum x_i^2 = 62^2 + 52^2 + 68^2 + 23^2 + 40^2 = 13301$

$n\bar{x}^2 = 5 \cdot 49^2 = 12005$

s^2 =(13301-12005)/(5-1) = 1296/4 = 324

(3) Muuttujan *x* keskihajonta $s_x = \sqrt{s_x^2}$

Esim.

Etäisyyksien keskihajonta *s* = $\sqrt{324}$ = 18.

(4) Korrelaatiokerroin

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

$$= \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

Mittaa kahden muuttujan *x* ja *y* välillä lineaarisen riippuvuuden voimakkuutta, käsin laskeminen ei olennaista, SPSS laskee. Voidaan merkitä myös *rxy*.
Asiasta kiinnostuneille lisäesimerkki korrelaatiokertoimen laskemisen yhteydessä tarvittavien summien ja neliösummien laskusta
http://www.sis.uta.fi/tilasto/tiltp1/syksy2003/moniste_5.pdf.
Korrelaatiokertoimeksi saadaan r = $\frac{-10}{\sqrt{16 \cdot 2400}}$ = – 0,051.

(5) Normaalijakauma

$X \sim N(\mu, \sigma^2), E(X) = \mu, Var(X) = \sigma^2, Z \sim N(0,1), P(Z \le z) = \Phi(z)$

Satunnaismuuttuja X noudattaa normaalijakaumaa odotusarvona μ ja varianssina σ^2. Satunnaismuuttuja Z noudattaa normaalijakaumaa, jonka odotusarvo 0 ja varianssi 1, nk. standardoitu normaalijakauma, jonka kertymäfunktion $\Phi(z)$ arvoja on taulukoita.

Standardoidun normaalijakauman taulukkoarvoja.

Z ~ N(0, 1)

z	1,6449	1,9600	2,3264	2,5758	3,0902	3,2905
Φ(z) = P(Z ≤ z)	0,9500	0,9750	0,9900	0,9950	0,9990	0,9995
P(Z≥z) = 1 - P(Z≤z)=P(Z≤-z)	0,0500	0,0250	0,0100	0,0050	0,0010	0,0005

Esimerkiksi Φ(1,96) = P(Z ≤ 1,96) = 0,975, P(Z ≥ 1,96) = 0,025 eli $z_{0,025}$ = 1,96, P(Z ≤ -1,96) = 0,025.

(6) Otoskeskiarvon odotusarvo ja varianssi

$E(\bar{X}) = \mu, Var(\bar{X}) = \sigma^2 / n$

Teoreettinen tulos, jolla pystytään arvioimaan otoskeskiarvon vaihtelua. Käytetään hyväksi mm. odotusarvon testauksessa ja odotusarvon luottamusvälin määrityksessä. Varianssi ja näin myös keskihajonta (=otoskeskiarvon keskivirhe = $\sigma/\sqrt{n}$) joudutaan käytännössä estimoimaan. Estimoitu keskivirhe on $s/\sqrt{n}$.

(7) Studentin t-jakaumaa noudattava satunnaismuuttuja $t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$

Käytetään mm. odotusarvon luottamusvälin määrittämisessä sekä odotusarvojen testauksessa, ks. kaavat (9), (11), (13). Jakauman taulukkoarvoja on käytettävissä (ks. s. 5).

(8) Arvioidaan tietynlaisten alkioiden prosenttiosuutta populaatiossa.
100(1–α) %:n luottamusväli prosenttiosuudelle $p \pm z_{\alpha/2}\sqrt{p(100-p)/n}$
Kyseessä väli, jolla arvellaan kyseisen prosenttiluvun olevan. Otoksesta laskettu prosenttiosuus on *p*, otoskoko *n*, $z_{\alpha/2}$ normaalijakauman taulukkoarvo.

Esim.

Erään puolueen kannatuksen arviointi. Kyselyyn vastasi 200 henkilöä, joista 40 puolueen kannattajia. Nyt $n = 200$, $p = 100 \cdot 40/200 = 20$. Määritettäessä 95%:n luottamusväliä $\alpha = 0{,}05$, $\alpha/2 = 0{,}05/2 = 0{,}025$, $z_{0{,}025} = 1{,}96$ (normaalijakauman taulukosta s. 2), luottamusvälin

alaraja $20 - 1{,}96\sqrt{20(100-20)/200} = 14{,}5$,
yläraja $20 + 1{,}96\sqrt{20(100-20)/200} = 25{,}5$.

Arvellaan siis todellisen kannatusprosentin olevan tällä välillä.

(9) Arvioidaan populaation keskiarvoa eli odotusarvoa.
100(1–α) %:n luottamusväli odotusarvolle (varianssi tuntematon) $\overline{X} \pm t_{\alpha/2;n-1}s/\sqrt{n}$
Kyseessä väli, jolla arvellaan odotusarvon olevan.

Esim.

Lepakoiden tunnistusmatka. Määritettäessä 95%:n luottamusväliä $\alpha = 0{,}05$, $\alpha/2 = 0{,}05/2 = 0{,}025$, $t_{0{,}025,\,5\text{-}1} = 2{,}776$ (Studentin t-jakauman taulukosta, s. 5), luottamusvälin

alaraja $49 - 2{,}776 \cdot 18/\sqrt{5} = 26{,}6$,
yläraja $49 + 2{,}776 \cdot 18/\sqrt{5} = 71{,}3$.

Arvellaan siis lepakoiden tunnistusmatkan olevan keskimäärin 27 cm – 71 cm.

(10) Tutkitaan, voisiko populaatiossa olla tietynlaisia alkioita väitetty prosenttiosuus.

$H_0 : \pi = \pi_0$, $Z = \dfrac{p - \pi_0}{\sqrt{\pi_0(100-\pi_0)/n}} \overset{likimain}{\sim} N(0,1)$, kun H_0.

Esim.

Eräs puolue väittää kannatuksensa olevan 22 %. Nyt $H_0 : \pi = 22\%$.
Tutkimuksessa kyselyyn vastasi 200 henkilöä, joista 40 puolueen kannattajia.
Nyt $n = 200$, $p = 100 \cdot 40/200 = 20$, joten $z_{havaittu} = \dfrac{20-22}{\sqrt{22(100-22)/200}} = -0{,}68$.

Tämä ihan tavanomainen arvo normaalijakaumasta, joten voidaan uskoa väite. Harvinaisten arvojen raja esim. 5 %:n riskitasolla yksisuuntaisessa testissä -1,6449 tai kaksisuuntaisessa -1,96 ($z_{0{,}05} = 1$, 6449, $z_{0{,}025} = 1{,}96$), laskettu arvo ei kuulu harvinaisten arvojen joukkoon.

(11) Tutkitaan, voisiko populaation odotusarvo olla väitetty luku.

$H_0 : \mu = \mu_0, \; t = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$, kun H_0 tosi.

Esim.

Lepakoiden tunnistusmatka. Tutkitaan voisiko keskimääräinen matka olla 35 cm vain olisiko se pidempi.
$H_0 : \mu = 35, \; H_1 : \mu > 35$

Nyt $t_{havaittu} = \dfrac{49-35}{18/\sqrt{5}} = 1{,}74 <$ $t_{0{,}05,\,5\text{-}1}$ = 2,132, joten 5 %:n riskitasolla tarkasteluna ei harvinaisten arvojen joukkoon kuuluva. Uskotaan väittämä, että keskimääräinen tunnistusmatka on 35 cm. Otos ei siis tue tutkijoiden arvelua.

(12) Tutkitaan kahden muuttujan välistä riippumattomuutta ristiintaulukon avulla. Ristiintaulukosta riippumattomuuden testaus: $\chi^2 \sim \chi^2_{(I-1)(J-1)}$, kun ei riippuvuutta.

SPSS-laskee testisuureen ja *p*-arvon, jonka avulla tehdään päättely. Nollahypoteesi on: ei riippuvuutta. Pieni *p*-arvo (esim. pienempi kuin 0,05) johtaa nollahypoteesin hylkäämiseen. Tällöin päätellään riippuvuutta olevan.

(13) Tutkitaan kahden populaation odotusarvojen yhtäsuuruutta.
$H_0 : \mu_1 = \mu_2, \; t \sim t_{n+m-2}$, kun tosi H_0 (oletetaan riippumattomat otokset ja populaatioiden varianssit yhtä suuriksi, mutta tuntemattomiksi).

SPSS-laskee testisuureen ja *p*-arvon, jonka avulla tehdään päättely. Pieni *p*-arvo (esim. pienempi kuin 0,05) johtaa nollahypoteesin hylkäämiseen. Tällöin päätellään, että odotusarvot eivät samoja. Tarkastellaan siis muuttujan keskiarvoja kahdessa ryhmässä.

(14) Tutkitaan, onko kahden muuttujan välillä lineaarista riippuvuutta.
H_0 : populaatiossa kahden muuttujan korrelaatiokerroin (ϱ) on nolla,

$t = \dfrac{r_{xy}}{\sqrt{(1-r_{xy}^2)/(n-2)}} \sim t_{n-2}$, kun H_0 tosi.

SPSS antaa korrelaatioiden laskun yhteydessä *p*-arvon, jonka avulla tehdään päättely. Pieni *p*-arvo (esim. pienempi kuin 0,05) johtaa nollahypoteesin hylkäämiseen. Tällöin päätellään lineaarista riippuvuutta olevan.

Studentin t–jakauman taulukkoarvoja $t_{\alpha;df}$, joille $P(t_{df} \geq t_{\alpha;df}) = \alpha$.

df	α = 0,05	α = 0,025	α = 0,01	α = 0,005
1	6,314	12,706	31,821	63,656
2	2,920	4,303	6,965	9,925
3	2,353	3,182	4,541	5,841
4	2,132	2,776	3,747	4,604
5	2,015	2,571	3,365	4,032
6	1,943	2,447	3,143	3,707
7	1,895	2,365	2,998	3,499
8	1,860	2,306	2,896	3,355
9	1,833	2,262	2,821	3,250
10	1,812	2,228	2,764	3,169
11	1,796	2,201	2,718	3,106
12	1,782	2,179	2,681	3,055
13	1,771	2,160	2,650	3,012
14	1,761	2,145	2,624	2,977
15	1,753	2,131	2,602	2,947
16	1,746	2,120	2,583	2,921
17	1,740	2,110	2,567	2,898
18	1,734	2,101	2,552	2,878
19	1,729	2,093	2,539	2,861
20	1,725	2,086	2,528	2,845
21	1,721	2,080	2,518	2,831
22	1,717	2,074	2,508	2,819
23	1,714	2,069	2,500	2,807
24	1,711	2,064	2,492	2,797
25	1,708	2,060	2,485	2,787
26	1,706	2,056	2,479	2,779
27	1,703	2,052	2,473	2,771
28	1,701	2,048	2,467	2,763
29	1,699	2,045	2,462	2,756
30	1,697	2,042	2,457	2,750
40	1,684	2,021	2,423	2,704
60	1,671	2,000	2,390	2,660
120	1,658	1,980	2,358	2,617
∞	1,645	1,960	2,326	2,576

Esimerkiksi $t_{0,05;10} = 1,812$, siis $P(t_{10} \geq 1,812) = 0,05$. $P(t_{10} \leq -1,812) = 0,05$.