MTTTA1 Tilastomenetelmien perusteet
Luento 24.1.2019

Kertausta ja täydennystä χ^2-yhteensopivuustestistä

H_0: otos peräisin tietystä jakaumasta

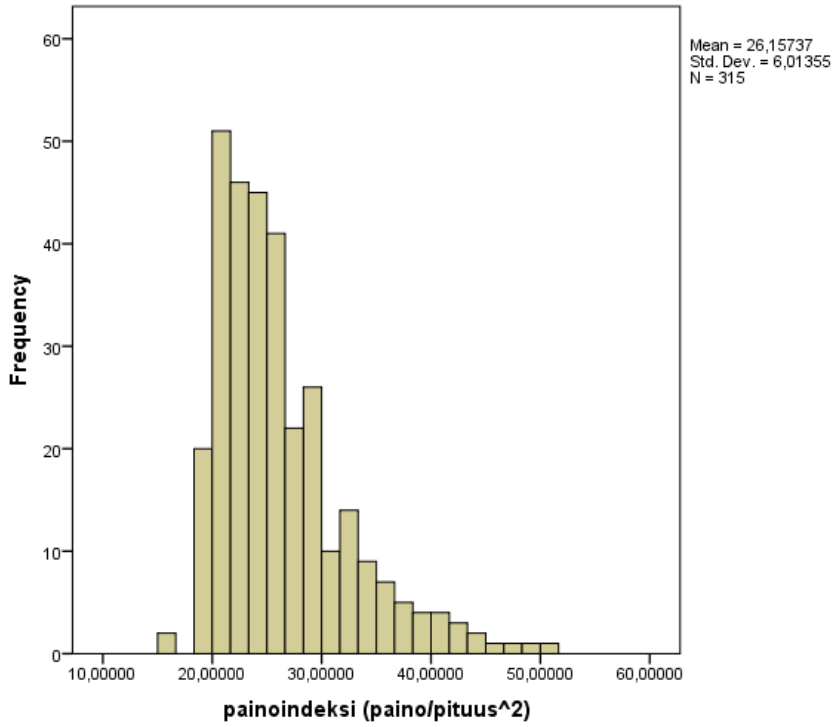
H_1: otos ei peräisin tästä jakaumasta

Jos H_0: otos peräisin tietystä jakaumasta on tosi, niin

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \sim \chi^2_{k-1}.$$

Esim. Plasma-aineisto, sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

y = painoindeksi (paino/pituus^2)



H_0: Otos peräisin $N(26.16, 6.01^2)$:sta

Vaihtoehtoisia testejä normaalisuuden testaamiseksi:

SPSS -> Analyze -> Descriptive Statistics -> Explore ...Plots -> Normality plots with tests

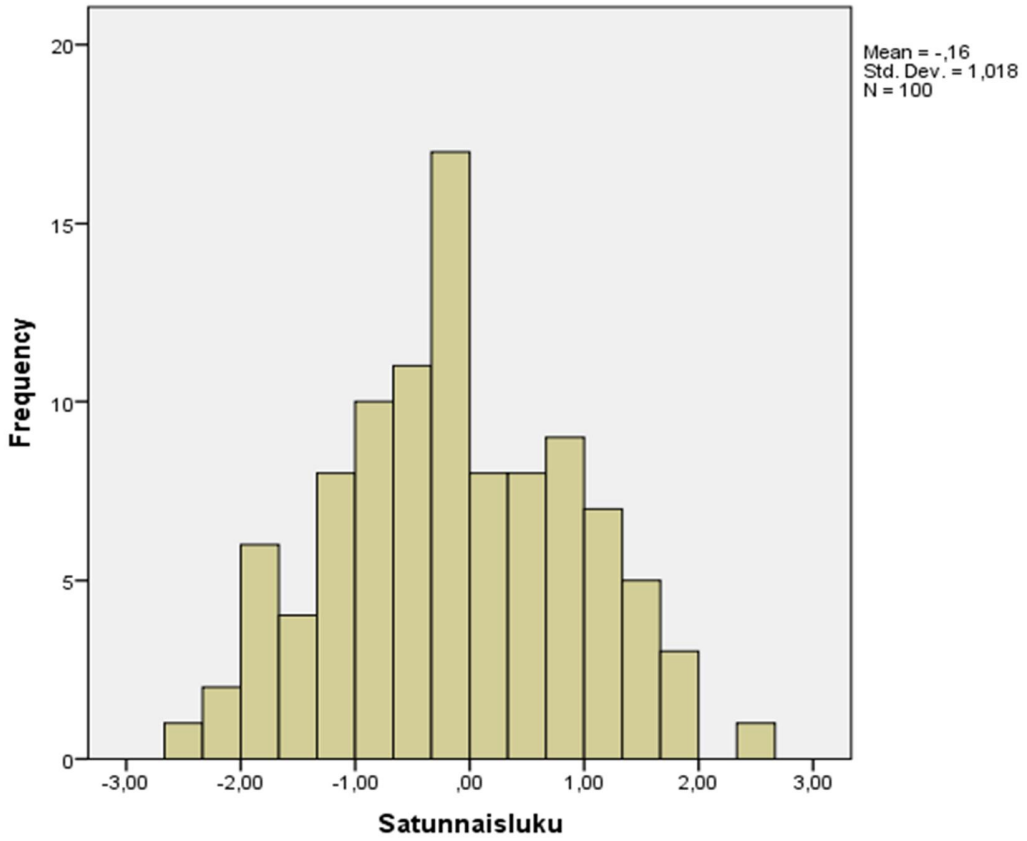
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
painoindeksi (paino/pituus^2)	,127	315	,000	,889	315	,000

a. Lilliefors Significance Correction

H_0 hylätään molemmilla testeillä, koska p-arvot < 0,001. Otos ei peräisin normaalijakaumasta.

Esim. Generoitu 100 lukua N(0, 1):stä, SPSS-funtio RV.NORMAL(0,1)



H_0: Otos peräisin $N(-0{,}16,\ 1{,}018^2)$:sta

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Satunnaisluku	,049	100	,200*	,990	100	,694

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

H_0 hyväksytään molemmilla testeillä, koska p-arvot > 0,05. Satunnaislukugeneraattori OK.

3.2 χ^2-riippumattomuustesti

Ristiintaulukon perusteella riippumattomuuden testaaminen

H_0: X ja Y ovat riippumattomia

H_1: X ja Y ovat riippuvia

Esim. Tampereella myydyt pienet asunnot, aineisto http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Tre_myydyt_asunnot_2009.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Kunto * Sijainti Crosstabulation

% within Sijainti

		Sijainti			
		Keskustassa	Alle 5 km keskustasta	Yli 5 km keskustasta	Total
Kunto	Hyvä	42,9%	29,8%	30,0%	32,7%
	Tyydyttävä	38,1%	42,6%	53,3%	44,9%
	Huono	19,0%	27,7%	16,7%	22,4%
Total		100,0%	100,0%	100,0%	100,0%

H_0: Kunto ja sijainti ovat riippumattomia

H_1: Kunto ja sijainti ovat riippuvia

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,725[a]	4	,605
Likelihood Ratio	2,667	4	,615
Linear-by-Linear Association	,134	1	,714
N of Valid Cases	98		

a. 1 cells (11,1%) have expected count less than 5. The minimum expected count is 4,71.

H_0 hyväksytään, koska p-arvo on 0,605 > 0,05.

Tarkastellaan yleisesti ristiintaulukkoa

		x				
		1	2	$\dots$	J	
y	1	f_{11}	f_{12}	$\dots$	f_{1J}	$f_{1\cdot}$
	2	f_{21}	f_{22}	$\dots$	f_{2J}	$f_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	I	f_{I1}	f_{I2}	$\dots$	f_{IJ}	$f_{I\cdot}$
		$f_{\cdot 1}$	$f_{\cdot 2}$	$\dots$	$f_{\cdot J}$	n

Määritetään ristiintaulukkoon teoreettiset frekvenssit e_{ij} siten, että oletetaan H_0: X ja Y riippumattomia on tosi. Tällöin oltava

$$\frac{e_{ij}}{f_{\cdot j}} = \frac{f_{i\cdot}}{n} \qquad \text{eli} \qquad e_{ij} = \frac{f_{i\cdot} f_{\cdot j}}{n}$$

Jos H_0 on tosi, niin

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2 (I-1)(J-1)$$

Nyt H_0 hylätään riskitasolla α, jos

$$\chi^2_{havaittu} > \chi^2_{\alpha,(I-1)(J-1)}$$

Jos I =2 ja J = 2 (nelikenttä), niin testisuure voidaan laskea myös kaavalla

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{f_{\cdot 1}f_{\cdot 2}f_{1\cdot}f_{2\cdot}}$$

Testiä voidaan käyttää:

a) (I-1)(J-1) = 1
   - n >40
   - 20 ≤ n ≤ 40

   kaikkien teoreettisten frekvenssien oltava ≥ 5.

b) (I-1)(J-1) > 1
   - kaikkien teoreettisten frekvenssien oltava > 1 ja enintään 20 % saa olla alle 5.

Esim. Edellisestä ristiintaulukosta testisuureen laskeminen.

Kunto * Sijainti Crosstabulation

			Sijainti			
			Keskustassa	Alle 5 km keskustasta	Yli 5 km keskustasta	Total
Kunto	Hyvä	Count	9	14	9	32
		Expected Count	6,9 $= e_{11}$	15,3 $= e_{12}$	9,8	32,0
		% within Sijainti	42,9%	29,8%	30,0%	32,7%
	Tyydyttävä	Count	8	20	16	44
		Expected Count	9,4	21,1	13,5	44,0
		% within Sijainti	38,1%	42,6%	53,3%	44,9%
	Huono	Count	4	13	5	22
		Expected Count	4,7	10,6	6,7 $= e_{33}$	22,0
		% within Sijainti	19,0%	27,7%	16,7%	22,4%
Total		Count	21	47	30	98
		Expected Count	21,0	47,0	30,0	98,0
		% within Sijainti	100,0%	100,0%	100,0%	100,0%

$$e_{11} = \frac{32 \cdot 21}{98}, \; e_{12} = \frac{32 \cdot 47}{98}, \ldots, \; e_{33} = \frac{22 \cdot 30}{98}$$

Chi-Square Tests

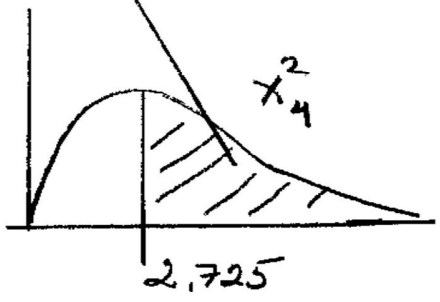
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,725[a]	4	,605
Likelihood Ratio	2,667	4	,615
Linear-by-Linear Association	,134	1	,714
N of Valid Cases	98		

a. 1 cells (11,1%) have expected count less than 5. The minimum expected count is 4,71.

H_0: ei riippuvuutta

$$X^2 = \frac{(9-6,9)^2}{6,9} + \dots + \frac{(5-6,7)^2}{6,7} = 2,725$$

(3-1)(3-1)



H_0 hyväksytään, koska p-arvo $> 0,05$.
Kunnon ja sijainnin välillä ei riippuvuutta.

Päätely taulukon avulla

$$\chi^2_{0,05,4} = 9,49 > 2,725 = \chi^2_{havaittu}$$, jotem H_0 hyväksytään 5%:n riskitasolla

Esim. Monisteesta Leppälä, R., Ohjeita tilastollisen tutkimuksen toteuttamiseksi IBM SPSS Statistics -ohjelmiston avulla, http://urn.fi/URN:ISBN:978-952-03-0501-7, esimerkki 13

Kyselylomake
http://www.sis.uta.fi/tilasto/tiltp3/kevat2003/Aineistoja/arviointi_lomake.pdf

Y = Opintojakson työläys
X = Opintosuunta

H_0: X ja Y ovat riippumattomia
H_1: X ja Y ovat riippuvia

Opintojakson työläys * opsuunta Crosstabulation

			opsuunta		
			hallinto	taloust	Total
Opintojakson työläys	työläs	Count	13	16	29
		Expected Count	8,5	20,5	29,0
		% within opsuunta	68,4%	34,8%	44,6%
	sopiva	Count	5	15	20
		Expected Count	5,8	14,2	20,0
		% within opsuunta	26,3%	32,6%	30,8%
	vähätöinen	Count	1	15	16
		Expected Count	4,7	11,3	16,0
		% within opsuunta	5,3%	32,6%	24,6%
Total		Count	19	46	65
		Expected Count	19,0	46,0	65,0
		% within opsuunta	100,0%	100,0%	100,0%

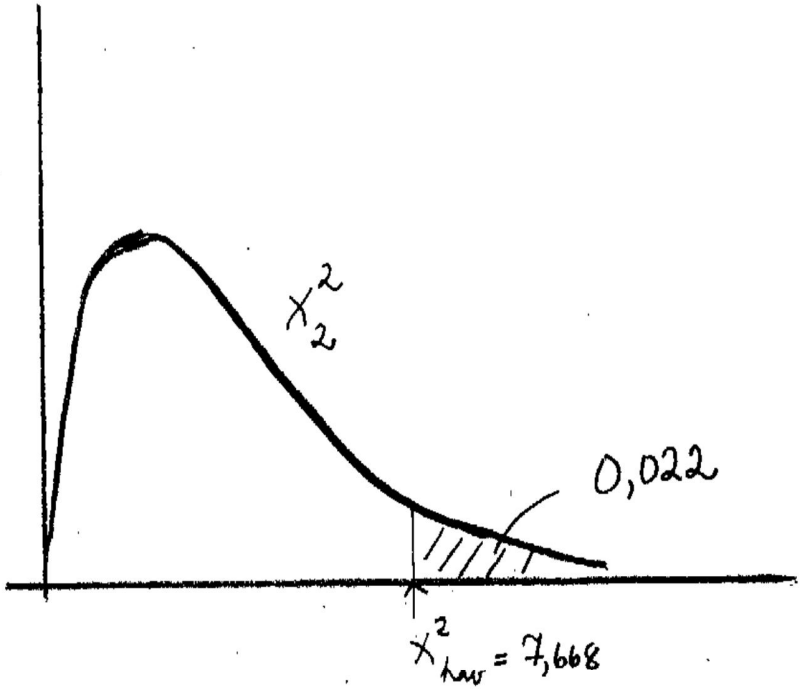
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,668[a]	2	,022
Likelihood Ratio	8,680	2	,013
Linear-by-Linear Association	7,548	1	,006
N of Valid Cases	65		

a. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 4,68.

Testin käyttöön liittyvät oletukset tällä luokituksella kunnossa, vain 16,7 % (1/6) odotetusta frekvensseistä alle 5 ja kaikki > 1.

Pienin riskitaso, jolla H_0 voidaan hylätä, on 0,022. Tätä suuremmilla riskeillä H_0 hylätään, pienemmillä hyväksytään.



$\chi^2_{0,01,2} = 9,21 > 7,668$, H_0 hyväksytään

$\chi^2_{0,025,2} = 7,38 < 7,668$, H_0 hylätään