MTTTA1 Tilastomenetelmien perusteet
Luento 22.1.2019

Luku 3

χ^2-yhteensopivuus- ja riippumattomuustestit

3.1 χ^2-yhteensopivuustesti

H_0: otos peräisin tietystä jakaumasta

H_1: otos ei peräisin tästä jakaumasta

Esim. H_0: otos peräisin normaalijakaumasta
H_0: otos peräisin tasajakaumasta

Esim. Eräällä kurssilla opiskelijat generoivat satunnaislukuja vastaamalla kysymyksiin:

1. Ravistele päätäsi ja arvo yksi kokonaisluku

	1	2	3	4	5	6	7	8	9	10	
heittotulos:	*2*	*3*	*6*	*3*	*4*	*7*	*6*	*5*	*3*	*1*	*n=40*

2. Ravistele päätäsi uudelleen ja arvo yksi kokonaisluku

	1	2	3	4	5	6	7	8	9	10	
heittotulos:	*1*	*2*	*9*	*7*	*5*	*4*	*2*	*5*	*4*	*1*	*n=40*

3. Ravistele päätäsi ja heitä rahaa

	klaava	kruuna	
heittotulos:	*21 (52,5 %)*	*19*	*n=40*

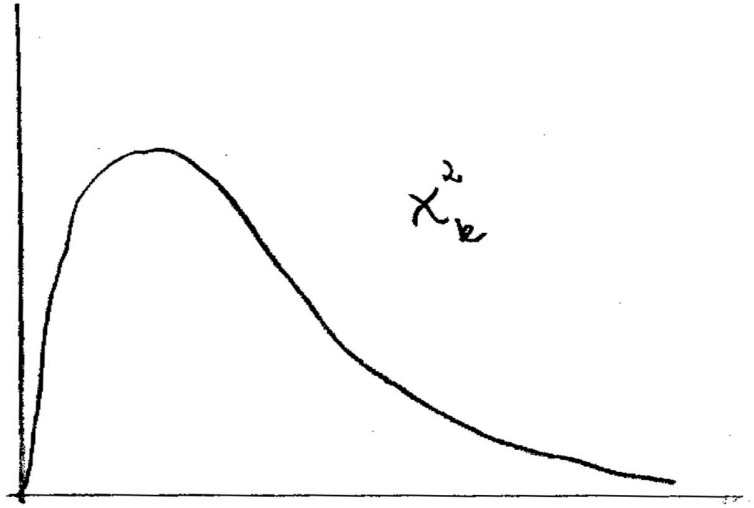
4. Ravistele päätäsi uudelleen ja heitä rahaa

	klaava	kruuna	
heittotulos:	*13 (32,5 %)*	*27*	*n=40*

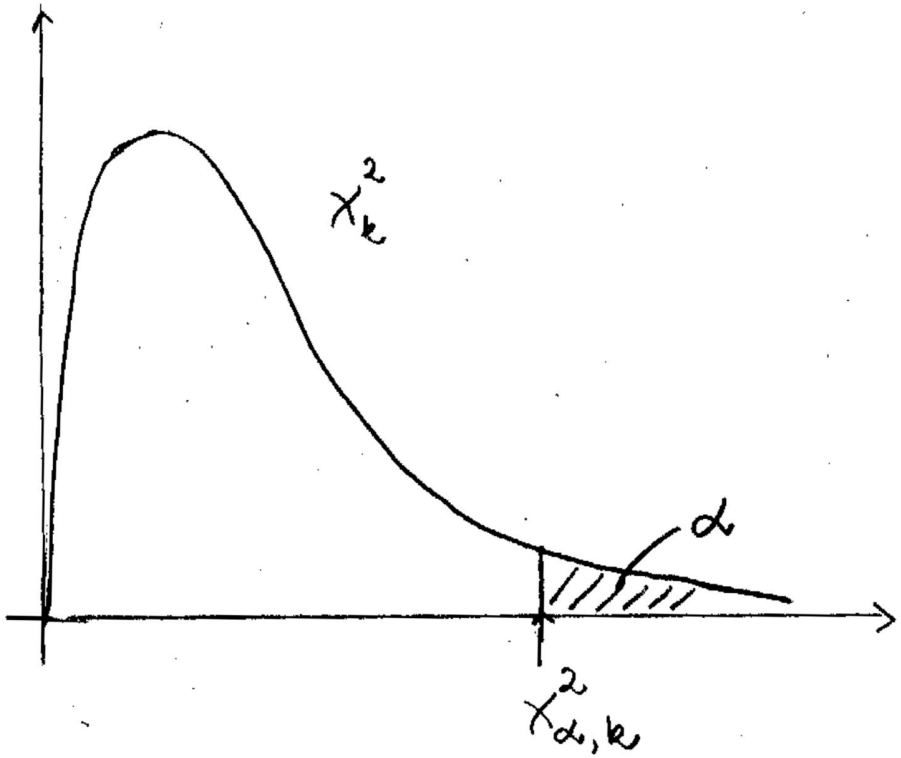
Voidaanko ajatella, että ensimmäinen kokonaisluvun valinta on otos diskreetistä tasajakaumasta? Jos olisi, niin jokainen numero olisi esiintynyt 4 kertaa. Voidaanko ajatella, että rahanheiton tulos on otos jakaumasta, jossa klaavoja 50 %? Jos olisi, niin klaavoja pitäisi olla 20 ja kruunia 20.

Olkoot riippumattomat Z_i ~N(0, 1), i = 1, …, k. Tällöin $Z_1^2 + Z_2^2 + \ldots + Z_k^2$ noudattaa nk. χ^2 – jakaumaa vapausastein k, merkitään χ_k^2. Tällöin $E(\chi_k^2) = k$, $Var(\chi_k^2) = 2k$.

χ^2 – jakauman tiheysfunktion kuvaaja, muoto riippuu vapausasteista



Määritellään $\chi^2_{\alpha,k}$ siten, että $P\left(\chi^2_k \geq \chi^2_{\alpha,k}\right) = \alpha$.



Näitä arvoja on taulukoitu,
ks. http://www.sis.uta.fi/tilasto/mttta1/kevat2019/chi.pdf

Tarkastellaan muuttujan frekvenssijakaumaa. Oletetaan, että jakaumassa on k kappaletta luokkia ja näiden luokkien frekvenssit f_1, f_2, ..., f_k.

Testataan sitä, ovatko havaitut frekvenssit sopusoinnussa H_0:n mukaisten nk. teoreettisten eli odotettujen frekvenssien e_1, e_2, ..., e_k kanssa.

Jos

H_0: otos peräisin tietystä jakaumasta

on tosi, niin

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

H_0 hylätään riskitasolla α, jos $\chi^2_{havaittu} > \chi^2_{\alpha,k-1}$.

Testiä voidaan käyttää, jos kaikki teoreettiset frekvenssit ovat > 1 ja enintään 20 % < 5.

Esim. Rahanheitto

H_0: Otos peräisin jakaumasta, jossa klaavoja ja kruunia yhtä paljon

1. rahanheitto

	f_i	e_i
klaavoja	21	20
kruunia	19	20

$$\chi^2_{havaittu} = \frac{(21-20)^2}{20} + \frac{(19-20)^2}{20} = 0{,}1$$

$\chi^2_{0.05,2-1}$ = 3,84 > $\chi^2_{havaittu}$ = 0,1 , H_0 hyväksytään 5%:n riskitasolla. Voidaan siis ajatella, että rahanheitto tehty satunnaisesti.

2. rahanheitto

	f_i	e_i
klaavoja	13	20
kruunia	27	20

$\chi^2 = \frac{(13-20)^2}{20} + \frac{(27-20)^2}{20} = 4,9$

Koska

$\chi^2_{0.05,2-1}$ = 3,84 < $\chi^2_{havaittu}$ = 4,9 < $\chi^2_{0.025,2-1}$ = 5,02 ,

niin 0,025 < p-arvo < 0,05.

Esim. Ystäväsi väittää, että suomalaisista 10 % on vasenkätisiä. Tutkit asiaa ja valitset satunnaisesti 400 suomalaista, joista 56 on vasenkätisiä. Uskotko ystäväsi väitteen?

H_0: 10 % suomalaisista on vasenkätisiä

	f_i	e_i
vasenkätisiä	56	0,1·400 = 40
ei-vasenkätisiä	344	0,9·400 = 360

$$\chi^2_{havaittu} = \frac{(56-40)^2}{40} + \frac{(344-360)^2}{360} = 7{,}11$$

$\chi^2_{0,01,1} = 6{,}63$
$\chi^2_{0,005,1} = 7{,}88$

H_0 hylätään 1 %:n riskitasolla, mutta ei 0,5 %:n riskitasolla, siis 0,005 < p-arvo < 0,01.

Laskuri http://vassarstats.net/csfit.html ja p-arvon arviointi http://vassarstats.net/csqsamp.html, p ≈ 0,008151

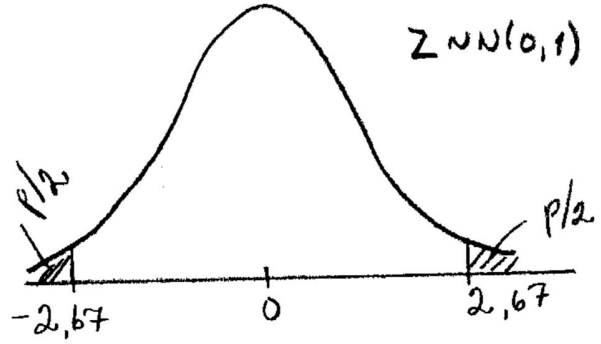
Toisin

$H_0: \pi = 10$

$H_1: \pi \neq 10$

$$z = \frac{14 - 10}{\sqrt{10 \cdot 90/400}} = 2{,}67$$

p-arvo = $2(1-\Phi(2{,}67)) = 2(1-0{,}9962) = 0{,}0076$



Jos χ^2-yhteensopivuustestissä luokkien lukumäärä on kaksi, niin $\chi^2 = Z^2$. Edellisessä esimerkissä $7{,}11 \approx 2{,}67^2$.

Esim. 3.1.4 Nopanheitto,
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=26

H_0: Otos peräisin Tasd(1, 6):sta

silmäluku	f_i	e_i
1	8	122/6 = 20,3
2	5	122/6
3	17	122/6
4	27	122/6
5	26	122/6
6	39	122/6

$$\chi^2_{havaittu} = \frac{(8 - 20{,}3)^2}{20{,}3} + \cdots + \frac{(39 - 20{,}3)^2}{20{,}3} = 40{,}6$$

$$> \chi^2_{0.005,6-1} = 16{,}75$$

H_0 hylätään, nopanheitto ei ole tapahtunut satunnaisesti.

Esim. 3.1.2 Asiakkaiden laskujen maksutavat, http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=25

H_0: ei tapahtunut muutosta

H_1: on tapahtunut muutos

	f_i	e_i	
ajoissa	287	0,8x400	= 320
1 kk myöhässä	49	0,1x400	= 40
2 kk myöhässä	30	0,06x400	= 24
yli 2 kk myöhässä	34	0,04x400	= 16

$$\chi^2_{hav.} = \frac{(287 - 320)^2}{320} + \cdots + \frac{(34 - 16)^2}{16} = 27{,}58 > \chi^2_{0.005,4-1} = 12{,}84$$

Päätellään muutosta tapahtuneen.

Laskuri http://vassarstats.net/csfit.html

Pelkän p-arvon määrittäminen
http://onlinestatbook.com/2/calculators/chi_square_prob.html

Esim. 3.1.5 Onko painoindeksi normaalisti jakautunut?
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=26

H_0: Otos peräisin $N(25.58, 4.66^2)$:sta

Painoindeksi	frekv.	odotettu frekv.
alle 20,1	9	11,5 = e_1
20,1-21,4	15	6,3
21,4-25,5	26	30,3
25,5-28,5	23	23,6
28,5-32,2	15	18,1
yli 32,2	9	7,5
	97	97

$e_1 = 97 \cdot P(X \leq 20{,}1) = 97 \cdot \Phi((20{,}1-25{,}58)/4{,}66)$
$= 97 \cdot \Phi(-1{,}18) = 97 \cdot (1- \Phi(1{,}18)) = 97 \cdot 0{,}119 =$
$11{,}5$

Vastaavalla tavalla lasketaan muidenkin luokkien odotetut frekvenssit.

Saadaan

$$\chi^2_{havaittu} = \frac{(9 - 11{,}5)^2}{11{,}5} + \cdots + \frac{(9 - 7{,}5)^2}{7{,}5} = 13{,}94$$

$> \chi^2_{0.005,6-2-1} = 12{,}84$

Päätellään, että otos ei peräisin normaalijakaumasta.

Huom! Vapausasteet pienenevät estimoitujen parametrien verran.

Laskurin http://vassarstats.net/csfit.html antama tulos, vapausasteissa ei huomioitu estimointia.

Chi-Square "Goodness of Fit" Test

The logic and computational details of chi-square tests are described in Chapter 8 of Concepts and Applications.

This unit will calculate the value of chi-square for a one-dimensional "goodness of fit" test, for up to 8 mutually exclusive categories labeled A through H. To enter an observed cell frequency, click the cursor into the appropriate cell, then type in the value. Expected values can be entered as either frequencies or proportions. If you enter the expected values as proportions, the entries can take the form of either decimal fractions such as .25, or common fractions such as 1/4. Whenever possible, it is better to enter common fractions rather than rounded decimal fractions: 1/3 rather than .3333; 1/6 rather than .1667; and so forth.

When all observed and expected values have been entered, click the «Calculate» button. To perform a new analysis with a new set of data, click the «Reset» button.

Cate-gory	Observed Frequency	Expected Frequency	Expected Proportion	Percentage Deviation	Standardized Residuals
A	9	11.5	0.1185567	-21.74%	-0.74
B	15	6.3	0.0649484	+138.1%	+3.47
C	26	30	0.3092783	-13.33%	-0.73
D	23	23.6	0.2432989	-2.54%	-0.12
E	15	18.1	0.1865979	-17.13%	-0.73
F	9	7.5	0.0773195	+20%	+0.55
G				----	----
H				----	----

Sums:

Observed Frequencies: 97

Expected Frequencies: 97

Expected Proportions: 1.0

Reset Calculate

[Note that for df=1, the calculated value of chi-square is corrected for continuity.]

[For df=1, this is the uncorrected value of chi-square.]

chi-square = 13.94

df = 5

P = 0.016

[P is non-directional]