Matemaatiline statistika ja modelleerimine

Sagedustabelite analüüs.
χ^2-test. Fisher'i täpne test

EMÜ doktorikool
DK.0007

Tanel Kaart

Kahemõõtmelise sagedustabeli üldkuju

Kahemõõtmeline sagedustabel võimaldab uurida kahe nominaaltunnuse või diskreetse arvtunnuse vahelist seost.

Näide. Kurja tõppe haigestunud 25 koera raviti samal meetodil. Tulemusena sai 14 koera terveks ja 11-l ei õnnestunud haigust välja ravida. Võttes arvesse ka ravitud koerte soo, kas võib väita, et antud haiguse ja ravimeetodi korral ravi tulemus sõltub koera soost?

Kahemõõtmeline sagedustabel:

Sugu	Ravi tulemus		Kokku
	Terve	Haige	
Isane	10	2	12
Emane	4	9	13
Kokku	14	11	25

Kahemõõtmelise sagedustabeli üldkuju

Olgu vaatluse all tunnus X, millel on m erinevat väärtust $x_1, x_2, \ldots, x_m$ ja tunnus Y, millel on k erinevat väärtust $y_1, y_2, \ldots, y_k$.
Ja olgu valimi maht n, kusjuures igal valimi objektil on mõlemad tunnused mõõdetud.

$$n_{i.} = \sum_{j=1}^{k} n_{ij},\ n_{.j} = \sum_{i=1}^{m} n_{ij},\ n = \sum_{j=1}^{k} n_{.j} = \sum_{i=1}^{m} n_{i.}$$

X \ Y	y_1	y_2	…	y_k	$n_{i.}$
x_1	n_{11}	n_{12}	…	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2k}	$n_{2.}$
…	…	…	…	…	…
x_m	n_{m1}	n_{m2}	…	n_{mk}	$n_{m.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	…	$n_{.k}$	n

Rea suhtelised sagedused saadakse, jagades lahtrite sagedused läbi vastava rea ääresagedusega: $u_{ij} = n_{ij} / n_{i.}$.

Veeru suhtelised sagedused saadakse, jagades lahtrite sagedused läbi vastava veeru ääresagedusega: $s_{ij} = n_{ij} / n_{.j}$.

Tabeli suhtelised sagedused saadakse, jagades lahtrite sagedused läbi valimi mahuga: $t_{ij} = n_{ij} / n$.

Suhtelised sagedused

Kahemõõtmeline sagedustabel:

Sugu	Ravi tulemus		Kokku
	Terve	Haige	
Isane	10	2	12
Emane	4	9	13
Kokku	14	11	25

Suhtelised sagedused:

Sugu	Ravi tulemus		Kokku
	Terve	Haige	
Isane	0,83	0,17	1,00
Emane	0,31	0,69	1,00
Kokku	0,56	0,44	1,00

Sugu	Ravi tulemus		Kokku
	Terve	Haige	
Isane	0,71	0,18	0,48
Emane	0,29	0,82	0,52
Kokku	1,00	1,00	1,00

Sugu	Ravi tulemus		Kokku
	Terve	Haige	
Isane	0,40	0,08	0,48
Emane	0,16	0,36	0,52
Kokku	0,56	0,44	1,00

Hii-ruut test (χ^2-test)

[*chi-square test* või (*Pearson's*) *goodness-of-fit test*]

Võrreldakse andmete alusel konstrueeritud sagedustabelit nn ideaalse, sõltumatuse juhule vastava, sagedustabeliga. Viimases peaksid ridade suhtelised sagedused võrduma summaarse suhteliste sageduste reaga ja veergude suhtelised sagedused summaarse suhteliste sageduste veeruga, ehk $n_{ij} = n_{i.}n_{.j}/n$.

H_0 – tunnused on sõltumatud, st $n_{ij} = n_{i.}n_{.j}/n$,
H_1 – tunnused on sõltuvad.

X \ Y	y_1	…	y_k	$n_{i.}$
x_1	n_{11}	…	n_{1k}	$n_{1.}$
…	…	…	…	…
x_m	n_{m1}	…	n_{mk}	$n_{m.}$
$n_{.j}$	$n_{.1}$	…	$n_{.k}$	n

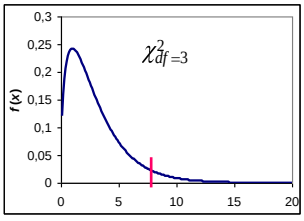
Teststatistik: $\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} \underset{H_0}{\sim} \chi^2_{df}$,

kus $df = (m-1)(k-1)$

Eeldused: kõik nullhüpoteesile vastavad sagedused ≥ 5 $n_{i.}n_{.j}/n \geq 5$, iga i, j ja iga uuritav objekt saab omada vaid üht väärtuste kombinatsiooni

Otsuse vastuvõtmine: kui teststatistiku väärtus on suurem kui χ^2-jaotuse vastav kriitiline väärtus $(\chi^2 \geq h_{1-\alpha,(m-1)(k-1)})$, või kui $\boldsymbol{p \leq \alpha}$,
siis on tõestatud H_1 (tunnused on sõltuvad), vastupidisel juhul jäädakse tunnuste sõltumatuse hüpoteesi juurde.

χ^2-jaotuse 1–α-kvantiilide $(h_{1-\alpha,df})$ väärtused



*MS Excel*is saab χ^2-jaotuse 1– α-kvantiili leidmiseks kasutada funktsiooni *CHIINV*(α;*df*)

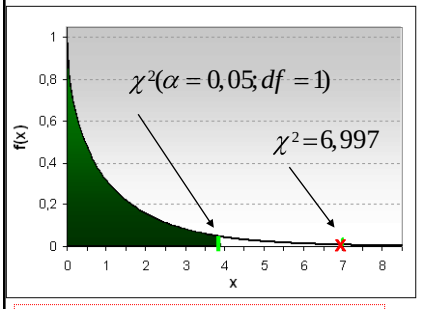
df	α = 0,05	α = 0,01
1	3,841	6,635
2	5,991	9,210
3	7,815	11,345
4	9,488	13,277
5	11,070	15,068
6	12,592	16,812
7	14,067	18,475
8	15,507	20,090
9	16,919	21,666
10	18,307	23,209
12	21,026	26,217
14	23,685	29,141
16	26,296	32,000
18	28,869	34,805
20	31,410	37,566
25	37,652	45,624
30	43,773	50,892
35	49,802	57,342
40	55,758	63,691
45	61,656	69,957
50	67,505	76,154
60	79,082	88,379
70	90,531	100,425
100	124,32	135,807

Hii-ruut test (χ^2-test) kahemõõtmelise sagedustabeli korral

Näide. Sugu *versus* ravi tulemus?

H_0 – ravi tulemus ei sõltu koera soost,
H_1 – ravi tulemus on soospetsiifiline.

Teststatistik: $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - n_i n_j / n)^2}{n_i n_j / n}$



MS Excel leiab vastava *p*-väärtuse valemiga CHIDIST(χ^2;*df*)

n_{ij}

Sugu	Ravi tulemus		Kokku
	Terve	Haige	
Isane	10	2	12
Emane	4	9	13
Kokku	14	11	25

$\frac{n_i n_j}{n}$

	Terve	Haige
Isane	6,72	5,28
Emane	7,28	5,72

$\frac{(n_{ij} - n_i n_j / n)^2}{n_i n_j / n}$

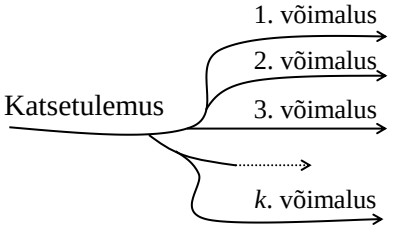
	Terve	Haige
Isane	1,60	2,04
Emane	1,48	1,88

$\chi^2 = 6,997$

=> H_1 – ravi tulemus on soospetsiifiline (p=0,0082).

Hii-ruut test (χ^2-test) üldiselt

H_0 – vaatlustulemused on kooskõlas teooria poolt ennustatuga,
H_1 – vaatlustulemused ei kinnita teooriat.



	Vaadeldud juhtude arv	Teooria poolt ennustatud juhtude arv
1. võimalus	n_1	N_1
2. võimalus	n_2	N_2
3. võimalus	n_3	N_3
…	…	…
k. võimalus	n_k	N_k

Teststatistik: $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - N_i)^2}{N_i} \underset{H_0}{\sim} \chi^2_{df}$

Vabadusastmete arv (*df*, *degrees of freedom*) = erinevate võimalike tulemuste (väärtuste) arv – valimi põhjal hinnatud teoreetiliste parameetrite arv

Eeldused: mida rohkem vaatlusi, seda täpsem; soovituslikult vähemalt 80% nullhüpoteesile vastavaist sagedustest ≥ 5 ja mitte ükski < 1.

χ^2-test – näiteid geneetikast (HW seadus)

Hardy-Weinbergi seadus. See populatsioonigeneetika põhiseadus väidab, et kui populatsioon on piisavalt suur, paarumine on juhuslik ning puuduvad looduslik ja kunstlik valik, migratsioon jmt, siis püsivad geeni- ja genotüübisagedused põlvkonniti konstantsed.

Lihtsaim viis seda populatsiooni geneetilise tasakaalu seadust matemaatiliselt formuleerida, on võtta vaatluse alla üks kahe esinemisvormiga (kahealleelne) geen (alleelide tähisteks traditsiooniliselt *a* ja *A*) ning eeldada, et alleeli *A* esinemissagedus populatsioonis on *p* (tõenäosus, et populatsioonist juhuslikult valitud geen on *A*, on *p*). Siis juhul, kui populatsioon on Hardy-Weinbergi tasakaalus, peaks genotüüpide jaotus olema järgmine:

genotüüp	esinemistõenäosus
AA	p^2
Aa	$2p(1-p)$
aa	$(1-p)^2$

Näide. Selgitamaks, kas paljude populatsioonigeneetikas (ja loomade aretuses) rakendatavate meetodite eelduseks olev Hardy-Weinbergi seadus kehtib ka tänapäevastes aretuspopulatsioonides, viidi läbi veiste veregruppide uuring.

Järgnevas tabelis on näitena toodud 40 eesti punast tõugu lehma dialleelse veregrupilookuse, tähisega EAF ning alleelidega vastavalt ’01’ ja ’02’, genotüübisagedused.

χ^2-test – näiteid geneetikast (HW seadus)

01/01-tüüpi isendeid: 13
01/02-tüüpi isendeid: 23
02/02-tüüpi isendeid: 4

Kontrollimaks hüpoteesi leitud genotüübisageduste vastavusest Hardy-Weinbergi seadusele (H_0), tuleb leida alleeli ’01’ esinemistõenäosus:

$$p = \mathrm{P}(\text{’01’}) = (2*13+23)/(2*40) = 0{,}6125.$$

Eelmise slaidi valemite alusel saab leida, kui palju ühe või teise genotüübiga isendeid pidanuks valimisse sattuma Hardy-Weinbergi seaduse kehtides, ning viia läbi χ^2-test:

	tegelik	oodatav prop.	oodatav arv	erinevus
01/01-tüüpi isendeid:	13	0,375	15	-2
01/02-tüüpi isendeid:	23	0,475	19	4
02/02-tüüpi isendeid:	4	0,150	6	-2

Teststatistik: $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - N_i)^2}{N_i} = \frac{(-2)^2}{15} + \frac{4^2}{19} + \frac{(-2)^2}{6} = 1{,}786.$

Vabadusastmete arv $df = 3-2 = 1$, sest andmetest (3 genotüübisagedust) peame hindama 2 parameetrit: valimi suuruse *n* ja alleeli ’01’ esinemissageduse *p*.

Et χ^2-jaotuse kriitiline väärtus $df = 1$ ja $\alpha = 0{,}05$ korral on 3,841 > 1,786, siis järeldame, et erinevus Hardy-Weinbergi tasakaalus oleva populatsiooni ja eesti punast tõugu veiste populatsiooni vahel on väike ning jääme nullhüpoteesi juurde.

Fisher'i täpne test [*Fisher's exact test*]

Konstrueeritakse kõikvõimalikud antud rea- ja veerusagedustega tabelid ja leitakse neist igaühe esinemistõenäosused mitmemõõtmelise hüpergeomeetrilise jaotuse tõenäosusfunktsioonist lähtuvalt:

$$p = \frac{\prod_{i=1}^{k} n_{i.}! \quad \prod_{j=1}^{m} n_{.j}!}{n! \prod_{i,j} n_{i,j}!}$$

X \ Y	y_1	y_2	…	y_k	$n_{i.}$
x_1	n_{11}	n_{12}	…	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2k}	$n_{2.}$
…	…	…	…	…	…
x_m	n_{m1}	n_{m2}	…	n_{mk}	$n_{m.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	…	$n_{.k}$	n

Andmete alusel konstrueeritud sagedustabeli ja sellest veel ekstreemsemate tabelite esinemistõenäosuste summa kujutabki enesest olulisuse tõenäosust (tõenäosust saada ilmnenud struktuuriga andmed juhuslikult).

2x2-tabelite puhul avaldub konkreetse, fikseeritud rea- ja veerusummadega tabeli tõenäosus kujul

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

$$p = [(a+b)!(c+d)!(a+c)!(b+d)!] / [n!a!b!c!d!].$$

Fisher'i täpne test [*Fisher's exact test*]

Näide.

Sugu	Ravi tulemus		
	Terve	Haige	Kokku
Isane	10	2	12
Emane	4	9	13
Kokku	14	11	25

0,01059

12	0	**1,75E-05**
2	11	

11	1	**0,00077**
3	10	

9	3	0,06352
5	8	

8	4	0,19056
6	7	

7	5	0,30490
7	6	

6	6	0,26679
8	5	

5	7	0,12704
9	4	

4	8	0,03176
10	3	

3	9	0,00385
11	2	

2	10	0,000192
12	1	

1	11	2,69E-06
13	0	

$p = 2\ (0{,}01059+0{,}00077+0{,}0000175) = 0{,}02275$

või

$p = (0{,}01059+0{,}00077+0{,}0000175)$
$+(0{,}00385+0{,}000192+0{,}00000269) = 0{,}01542$

Fisher'i täpne test *versus* χ^2-test

Näide. Uue depressiooniravimi uuringud, 123 patsienti grupeerituna ravimi mõju järgi (mõjus/ei mõjunud), 252 valdavalt dialleelset geneetilist markerit (3 genotüüpi).

- Analüüsid teostati statistikapaketiga SAS.
- Mõlemad testid andsid 10 statistiliselt olulist (p<0,05) seost, kusjuures 9 selliselt tuvastatud markerit langesid kokku.
- Fisher'i täpne test andis 167 juhul suurema ja 85 juhul väiksema p-väärtuse kui χ^2-test.

The SAS System

The FREQ Procedure

Table of RESPON by _25_SLC6A4_2020940

RESPON(RESPON) _25_SLC6A4_2020940(225_SLC6A4_2020940)

Frequency / Percent / Row Pct / Col Pct	CC	CG	GG	Total
0	22 11.06 66.67 15.49	8 4.02 24.24 14.81	3 1.51 9.09 100.00	33 16.58
1	120 60.30 72.29 84.51	46 23.12 27.71 85.19	0 0.00 0.00 0.00	166 83.42
Total	142 71.36	54 27.14	3 1.51	199 100.00

Frequency Missing = 4

The SAS System

The FREQ Procedure

Statistics for Table of RESPON by _25_SLC6A4_2020940

Statistic	DF	Value	Prob
Chi-Square	2	15.3349	0.0005
Likelihood Ratio Chi-Square	2	11.0307	0.0040
Mantel-Haenszel Chi-Square	1	2.4629	0.1166
Phi Coefficient		0.2776	
Contingency Coefficient		0.2675	
Cramer's V		0.2776	

WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

Table Probability (P)	7.395E-04
Pr <= P	0.0080

Effective Sample Size = 199
Frequency Missing = 4

Sagedustabelist leitavad suurused

Enamasti defineeritakse 2 2 sagedustabeli tarvis
(suuremate tabelite korral arvutatakse kordajad analoogselt)

	Juhud – 1 (haiged/*responders*/…)	Kontrollid – 0 (terved/*nonresponders*/…)	Kokku
Eksponeeritud	a	b	$a+b$
Mitteeksponeeritud	c	d	$c+d$
Kokku	$a+c$	$b+d$	$a+b+c+d=N$

- **Sagedused** (*observed frequences*) a, b, c, d.
- **Oodatavad sagedused** assotsiatsiooni puudumise korral (uuritavate tunnuste sõltumatuse eeldusel, st nullhüpoteesi kehtides; *expected frequences*);
($a+b$) ($a+c$)/N, … – võimaldavad välja selgitada sõltumatuse juhust enim erinevad väärtuste kombinatsioonid;
- **Juhtude esinemissagedus** (haigestumuskordaja, *incidence rate*, *rate*) iga riskifaktori taseme tarvis – $a/(a+b)$, $c/(c+d)$.

Sagedustabelist leitavad suurused epidemioloogias

	Juhud	Kontr.	Kokku
Eksp.	a	b	$a+b$
Mitteeksp.	c	d	$c+d$
Kokku	$a+c$	$b+d$	N

- **Riskisuhe** (*RR, risk ratio, relative risk*):

$$RR = \frac{\text{juhu risk eksponeeritutel}}{\text{juhu risk mitte eksponeeritutel}};$$

kui võrreldavaid gruppe on enam kui 2, siis leitakse *RR* tavaliselt vähima juhtude esinemissagedusega rea (grupi) suhtes;
näiteks eeldades, et $a/(a+b)>c/(c+d)$: $RR_1=[a/(a+b)]/[c/(c+d)]$, $RR_2=1$.

$$95\%\ CI_{RR} \approx e^{\ln(RR)\pm 1{,}96\times se[\ln(RR)]} = \frac{RR}{e^{1{,}96\times se[\ln(RR)]}}; RR\times e^{1{,}96\times se[\ln(RR)]}$$

$$se[\ln(RR)] = \sqrt{\frac{1}{\text{juhtude arv eksponeeritutel}} + \frac{1}{\text{juhtude arv mitteeksponeeritutel}}}$$

Näide.

	Tervete arv	Haigete arv	Oodatav haigete arv	IR_{Haige}	RR	95% CI_{RR}	RR *või*	95% CI_{RR}
Isane	10	2	5,28	0,182	0,22	(0,05; 1,03)	1	
Emane	4	9	5,72	0,818	1		4,50	(0,97; 20,83)

Šansid, šansside suhe [*Odds, odds ratio*]

Sündmuse toimumise **šansid** [*odds*] näitavad, mitmel juhul sündmus toimub võrreldes sellega, mitmel juhul ta ei toimu.
Näiteks kui sündmus toimub tõenäosusega 0,2 (20%) e ühel juhul viiest, siis selle sündmuse toimumise šansid on üks nelja vastu e 1:4.

Šansside suhe (*OR, odds ratio*) näitab, kui mitu korda erineb uuritava sündmuse toimumise šanss ühes grupis võrreldes teis(te)ga:

$$OR = \frac{\text{juhu shanss eksponeeritutel}}{\text{juhu shanss mitte eksponeeritutel}}.$$

Näiteks 2x2-tabeli korral võib leida $OR_1=(a/b)/(c/d)$, $OR_2=1$.

$$95\%\ CI_{OR} \approx e^{\ln(OR)\pm 1{,}96\times se[\ln(OR)]} = \frac{OR}{e^{1{,}96\times se[\ln(OR)]}}; OR\times e^{1{,}96\times se[\ln(OR)]}$$

$$se[\ln(OR)] = \sqrt{\frac{1}{\text{juhtude arv eksponeeritutel}} + \frac{1}{\text{juhtude arv mitteeksponeeritutel}} + \frac{1}{\text{kontrollide arv eksponeeritutel}} + \frac{1}{\text{kontrollide arv mitteeksponeeritutel}}}$$

Šansid, šansside suhe

Šansid isastel *vs* emastel:

$OR_{\text{Haige}} = (2/10) / (9/4) = 8/90 = 0{,}089$

$OR_{\text{Terve}} = (10/2) / (4/9) = 90/8 = 11{,}25$

Sugu	Ravi tulemus		
	Terve	Haige	Kokku
Isane	10	2	12
Emane	4	9	13
Kokku	14	11	25

$$95\%\ CI_{OR\ \text{Haige}} \approx \frac{0{,}089}{e^{1{,}96\times\sqrt{1/2+1/9+1/10+1/4}}}; 0{,}089 \times e^{1{,}96\times\sqrt{1/2+1/9+1/10+1/4}} = 0{,}013; 0{,}607$$

	Terve olemise šanss	Haige olemise šanss	Šansside suhe (OR_{Haige})	95% CI_{OR}
Isane	5,00	0,20	0,089	(0,013; 0,607)
Emane	0,44	2,25	1	

	Šansside suhe (OR_{Haige})	95% CI_{OR}
Isane	1	
Emane	11,25	(1,64; 76,85)

Šansside suhe või riskisuhe?

Lihtsate nn katsepõhiste uuringute korral riskisuhe (või šansside suhe), keerulisemate mudelipõhiste uuringute korral šansside suhe.

Relationships Among Milk Yield, Body Condition, Cow Weight, and Reproduction in Spring-Calved Holstein-Friesians

F. Buckley*, K. O'Sullivan†, J. F. Mee*, R. D. Evans*, and P. Dillon*
*Moorepark Dairy Production Research Centre, Fermoy, Co. Cork, Ireland
†Statistical Laboratory, Dept. of Statistics, University College Cork, Ireland

J. Dairy Sci. 86:2308–2319

Table 5. Association between milk production, BCS, and BW variables, and SR21.[1]

Model[2]	OR	95% CI	*P*-value
Model relating milk production variables to likelihood of SR21	n = 2753		$R^2 = 0.164$
Estimated 200-d milk protein content (g/kg)			
<31.8	1		
31.8 to 33.0	1.20	0.90–1.62	NS
33.1 to 34.4	1.52	1.10–2.09	0.011
>34.4	1.54	1.10–2.14	0.012
Protein-to-fat ratio at herd SBD			
<0.81	1		
0.81 to 0.90	1.34	0.98–1.84	0.066
0.91 to 1.00	1.11	0.80–1.52	NS
>1.00	1.45	1.03–2.05	0.036
Model relating BCS variables to likelihood of SR21	n = 2204		$R^2 = 0.131$
Average BCS between 60 and 100 d of lactation (BCS units)			
≤2.50	0.59	0.44–0.78	<0.001
2.75 to 3.0	1		
≥3.25	0.90	0.63–1.31	NS
Model relating BW variables to likelihood of SR21	n = 1483		$R^2 = 0.194$
BW at herd SBD (kg)			
<483	1		
483 to 529	1.33	0.82–2.17	NS
530 to 576	1.20	0.70–2.07	NS
>576	1.90	1.00–3.60	0.048
BW loss from precalving to nadir (kg)			
>131	1		
110 to 131	1.81	1.15–2.86	0.011
88 to 109	1.01	0.65–1.55	NS
<88	1.17	0.72–1.90	NS
BW gain from herd SBD to 90 d thereafter (kg)			
<17	1		
17 to 34	1.08	0.69–1.70	NS
35 to 52	1.64	1.00–2.69	0.052
>32	1.60	0.91–2.82	0.100

[1]SR21 = Submission in the first 3 wk of the breeding season, OR = odds ratio, CI = confidence interval, SBD = herd start of breeding date, n = number of cows included in analysis, NS = $P > 0.10$.
[2]All models were adjusted for herd, calving period, lactation number, proportion of Holstein-Friesian genes, breeding value for milk yield, and degree of calving assistance.