Tanel Kaart
sügis, 2009

Statistiline andmetöötlus

VL.0435

Loeng 5

- ✓ Kahemõõtmeline sagedustabel
- ✓ χ^2- ja Fisheri täpne test
- ✓ 0-1 tüüpi tunnuste analüüs – šansside suhe ja logistiline regressioon

http://www.eau.ee/~ktanel/VL_0435/

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

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

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.

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

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

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: $t_{ij} = n_{ij} / n$.

Kahemõõtmelise sagedustabeli üldkuju

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) kahemõõtmelise sagedustabeli korral

[*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

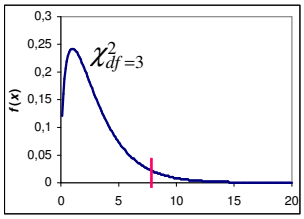
Teststatistik: $\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(n_{ij} - n_i n_{.j}/n)^2}{n_i n_{.j}/n} \underset{H_0}{\sim} \chi^2_{df}$,

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

Eeldused: kõik nullhüpoteesile vastavad sagedused ≥ 5 $\left(n_{i.} n_{.j} / n \geq 5, \text{iga } i, j\right)$ 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	$\alpha = 0{,}05$	$\alpha = 0{,}01$
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: $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - n_i n_j/n)^2}{n_i n_j/n}$

n_{ij}

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

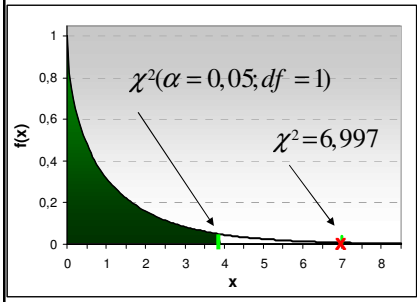
$\frac{n_i n_j}{n}$

	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

$\chi^2 = 6,997$

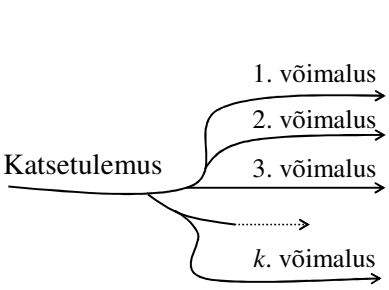


MS Excel leiab vastava *p*-väärtuse valemiga CHIDIST(χ^2;*df*)

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

Hii-ruut testi olemus

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.

Hii-ruut test – näiteid geneetikast

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:

$$p = \frac{\left(\prod_{i=1}^{k} n_{i.}!\right)\left(\prod_{j=1}^{m} n_{.j}!\right)}{n!\prod_{i,j} n_{i,j}}$$

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

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

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

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	

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

või

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

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

Sündmuse toimumise šansid [*odds*] näitavad, mitmel juhul sündmus toimub võrreldes sellega, mitmel juhul ta ei toimu.

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

a	*b*	*a+b*
c	*d*	*c+d*
a+c	*b+d*	*n*

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

$OR_{Terve} = (a/b)/(c/d)$
$= (10/2) / (4/9) = 90/8$
$= 11,25$

Šansid jm epidemioloogias kasutatavad karakteristikud

Näide. Kas luts on sagedamini *Diphyllobothrium latum* (hariliku laiussi) plerotserkoididega nakatunud kui haug?

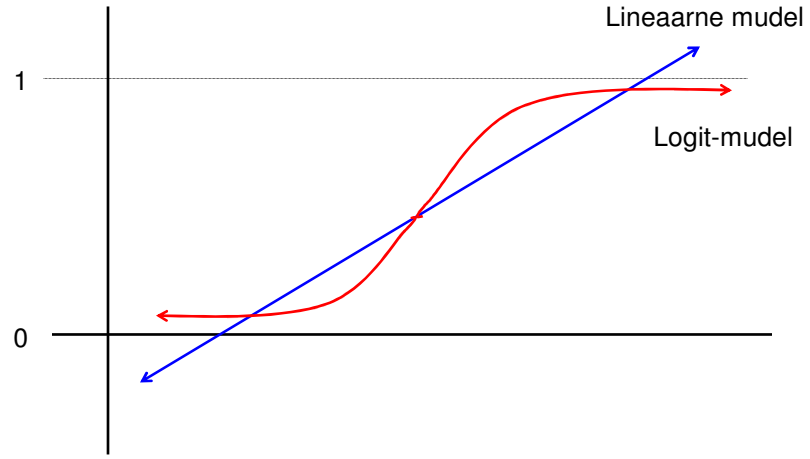
	Nakatunud	Terve	Kokku
Luts	59	78	137
Haug	28	186	214
Kokku	87	264	351

	Oodatav nakatunute arv	Haigestumus-kordaja	Riskisuhe	Nakatumise šanss	Šansside suhe
Luts	34	0,431	3,29	0,756	5,03
Haug	53	0,131	1	0,151	1

Järeldused (näiteks).

- Lutsudest oli nakatunud 43,1%, haugidest 13,1%.
- Lutsude šansid nakatuda on 5,03 korda suuremad võrreldes haugidega.

Logistiline regressioon



Logistilise regressiooni mudel

- Logistilise regressiooni mudel e logit-mudel:

$$\ln[p/(1-p)] = \alpha + \beta X + e$$

  - p on sündmuse Y toimumise (esinemise) tõenäosus, $P(Y = 1)$
  - $p/(1-p)$ on šansside suhe [*OR*]
  - $\ln[p/(1-p)]$ on logaritmiline šansside suhe e "logit"

- Logistiline regressioon garanteerib hinnatud tõenäosuste jäämise vahemikku 0-st 1-ni.
- Tõenäosuse hinnang avaldub kujul:

$$p = \exp(\alpha + \beta X) / [1 + \exp(\alpha + \beta X)]$$

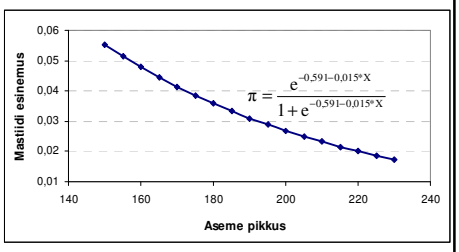
  - kui $\alpha + \beta X = 0$, siis $p = 0{,}5$
  - $\alpha + \beta X$ suurenemisel p läheneb 1-le
  - $\alpha + \beta X$ vähenemisel p läheneb 0-le

Logistilise regressiooni mudel – parameetrite tõlgendus

- $\ln[p/(1-p)] = \alpha + \beta X$
- Mida suurem on β, seda järsem on tõus graafiku keskosas.
  - Kui β on negatiivne, siis graafik langeb (negatiivne seos).
  - Parameeter α nihutab graafikut vasakule-paremale.
- $OR = e^{\beta}$
  - seega näitab šansside suhe, kuidas muutuvad sündmuse šansid (kui palju suurenevad šansid saada tulemust $Y = 1$), kui argument muutub ühiku võrra.
- Üldiselt: kui argumendi muutus on c ühikut, siis šansside suhe muutub $e^{c\beta}$ korda.

Logistiline regressioon. Näide

- Mastiidi esinemise tõenäosus (π) prognoosituna aseme pikkuse (X) järgi:



- SAS-i protseduur LOGIT:

```
proc logistic data=analyys.andmed1
 descending;
 model mastiit = aseme_pi;
run;
```

- Programmi väljund

```
                    The SAS System
                The LOGISTIC Procedure
                        ......
 Analysis of Maximum Likelihood Estimates

                  Parameter  Standard      Wald        Pr >     Standardized  Odds
Variable  DF      Estimate     Error  Chi-Square  Chi-Square      Estimate     Ratio
INTERCPT  1       -0.5809    0.7074      0.6743      0.4116             .         .
ASEME_PI  1       -0.0148   0.00409     13.0188      0.0003      -0.156026     0.985
```

$OR = e^{-0,0148} = 0,985$