



Statistika ja modelleerimise baaskoolitus õppejõududele ja juhendajatele

Loeng 3

Eesti Maaülikool
27.-30. august 2012
Tanel Kaart



**Seosed diskreetsete/mittearvuliste
tunnuste vahel – kahemõõtmeline
sagedustabel, χ^2 - ja Fisheri täpne
test, šansside suhe**

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 $x_1, x_2, \dots, x_m$ ja tunnus Y , millel on k erinevat väärtust $y_1, y_2, \dots, y_k$.

Ja olgu valimi maht n , kusjuures igal valimi objektil on mõlemad tunnused mõõdetud.

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

$X \backslash Y$	y_1	y_2	$\dots$	y_k	$n_{i.}$
x_1	n_{11}	n_{12}	$\dots$	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	$\dots$	n_{2k}	$n_{2.}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_m	n_{m1}	n_{m2}	$\dots$	n_{mk}	$n_{m.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.k}$	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 vasta-va 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:

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	10	2	12
Emane	4	9	13
Kokku	14	11	25

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, sõltumatu 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.

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

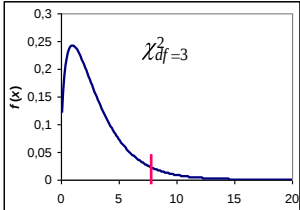
kus $df = (m-1)(k-1)$

$X \backslash Y$	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

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 $p \leq \alpha$, siis on tõestatud H_1 (tunnused on sõltuvad), vastupidisel juhul jäädaakse tunnuste sõltumatuse hüpoteesi juurde.

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



*MS Excelis saab
 χ^2 -jaotuse
 $1-\alpha$ -kvantiili leidmiseks
kasutada funktsiooni
 $CHIINV(\alpha; df)$*

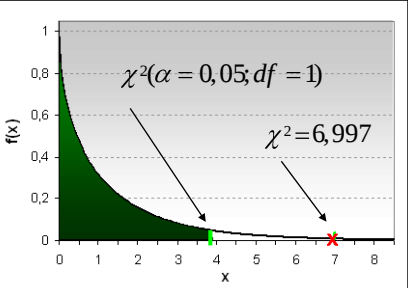
df	$\alpha = 0,05$	$\alpha = 0,01$
1	3,841	6,635
2	5,991	9,210
3	7,815	11,345
4	9,488	13,277
5	11,070	15,068
6	12,592	16,812
7	14,067	18,475
8	15,507	20,090
9	16,919	21,666
10	18,307	23,209
12	21,026	26,217
14	23,685	29,141
16	26,296	32,000
18	28,869	34,805
20	31,410	37,566
25	37,652	45,624
30	43,773	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: $\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - n_i n_j / n)^2}{n_i n_j / n}$



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

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
Katsetulemus	1. võimalus	N_1
	2. võimalus	N_2
	3. võimalus	N_3
	...	...
	k . võimalus	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 veregrupi-lookuse, 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:

$$p = P('01') = (2 \cdot 13 + 23) / (2 \cdot 40) = 0,6125.$$

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

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

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:

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

$X \backslash Y$	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
2	11

1,75E-05

11	1
3	10

0,00077

9	3
5	8

0,06352

8	4
6	7

0,19056

4	8
10	3

0,03176

7	5
7	6

0,30490

3	9
11	2

0,00385

6	6
8	5

0,26679

2	10
12	1

0,000192

5	7
9	4

0,12704

1	11
13	0

2,69E-06

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

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 (25_SLC6A4_2020940)

Frequency Percent Row Pct Col Pct	CC	CG	GG	Total
0	22 11.06 56.67 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	15.3949	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 x 2 sagedustabeli tarvis (suuremate tabelite korral arvutatakse kordajad analoogselt)

	Juhud – 1 (haiged/responders/...)	Kontrollid – 0 (terved/nonresponders/...)	Kokku
Eksponeeritud	<i>a</i>	<i>b</i>	<i>a+b</i>
Mitteeksponeeritud	<i>c</i>	<i>d</i>	<i>c+d</i>
Kokku	<i>a+c</i>	<i>b+d</i>	<i>a+b+c+d=N</i>

- **Sagedused** (*observed frequencies*) *a*, *b*, *c*, *d*.
- **Oodatavad sagedused** assotsiatsiooni puudumise korral (uuritavate tunnuste sõltumatuse eeldusel, st nullhüpoteesi kehtides; *expected frequencies*);
(*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

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

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

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

$$95\% CI_{RR} \approx e^{\ln(RR) \pm 1.96 \times se[\ln(RR)]} = \frac{RR}{e^{1.96 \times se[\ln(RR)]}}; RR \times e^{1.96 \times se[\ln(RR)]}$$

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

Näide.

	Tervete arv	Haigete arv	Oodatav haigete arv	<i>IR</i> _{Haige}	<i>RR</i>	95% <i>CI</i> _{<i>RR</i>}	<i>RR</i>	95% <i>CI</i> _{<i>RR</i>}
							või	
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:

$$OR = \frac{\text{juhu šanss eksponeeritudel}}{\text{juhu šanss mitte eksponeeritudel}}$$

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

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

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

Šansid, šansside suhe

Šansid isastel vs emastel:

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

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

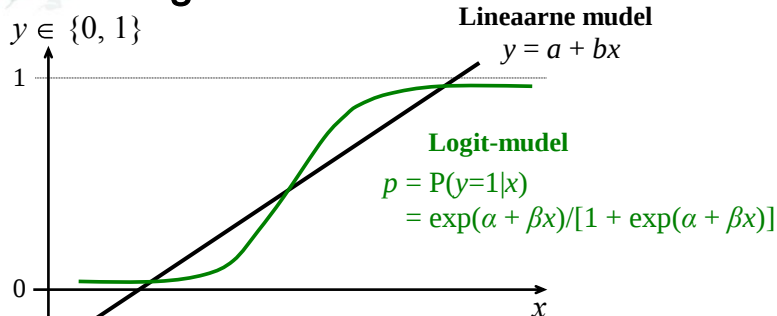
Sugu	Ravi tulemus		Kokku
	Terve	Haige	
Isane	10	2	12
Emane	4	9	13
Kokku	14	11	25

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

	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)

Logistiline regressioon



- Logistilise regressiooni abil leitud tõenäosuste hinnangud jäävad alati 0 ja 1 vahele.
- Logistilise regressiooni mudeli (*logit*-mudeli) kujusid:
$$p = P(y=1|x) = \exp(\alpha + \beta x) / [1 + \exp(\alpha + \beta x)] = 1 / [1 + \exp(-\alpha - \beta x)],$$
$$\ln[p/(1-p)] = \alpha + \beta x,$$
$$\text{logit}(p) = \alpha + \beta x,$$
- p on huvi pakkuva sündmuse Y toimumise tõenäosus: $p = P(y=1)$,
- $p/(1-p)$ on šansside suhe [*odds ratio*],
- $\ln[p/(1-p)]$ on logaritmiline šansside suhe [*log odds ratio*].

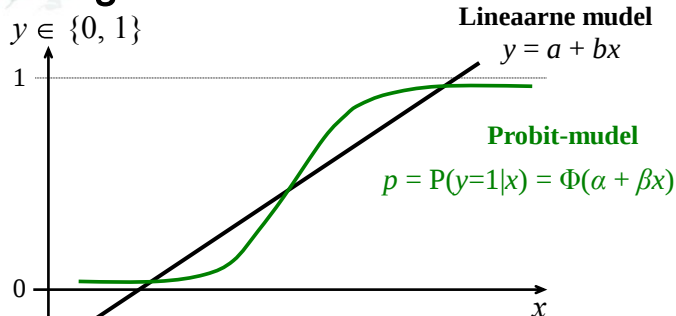
Logistiline regressioon

- Tõenäosuse hinnang avaldub kujul:

$$p = \exp(\alpha + \beta x) / [1 + \exp(\alpha + \beta x)] = 1 / [1 + \exp(-\alpha - \beta x)];$$

- kui $\alpha + \beta x = 0$, siis $p = 0,5$
 - $\alpha + \beta x$ suurenemisel $p \rightarrow 1$,
 - $\alpha + \beta x$ vähenemisel $p \rightarrow 0$.
- Logistilise regressioonivõrrandi kordaja β eksponent, e^β , näitab, kui mitu korda muutub sündmuse toimumise šanss (kui mitu korda muutub šansside suhe) argumenti muutumisel ühe ühiku võrra.
 - Näiteks kui $e^\beta = 2$, siis kaasneb argumenttunnuse väärtuse suurenemisega 1 võrra sündmuse toimumise šansi kahekordne suurenemine.
 - Negatiivse regressioonikordaja β korral on šansside suhe ühest väiksem, $e^\beta < 1$, seega kaasneb argumenttunnuse suurenemisega uuritava sündmuse toimumise šansi vähenemine.

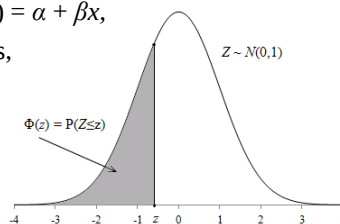
Probit-regressioon



- Probit-regressiooni abil leitud tõenäosuste hinnangud jäävad alati 0 ja 1 vahele.
- Probit-regressiooni mudeli (*probit*-mudeli) kujusid:

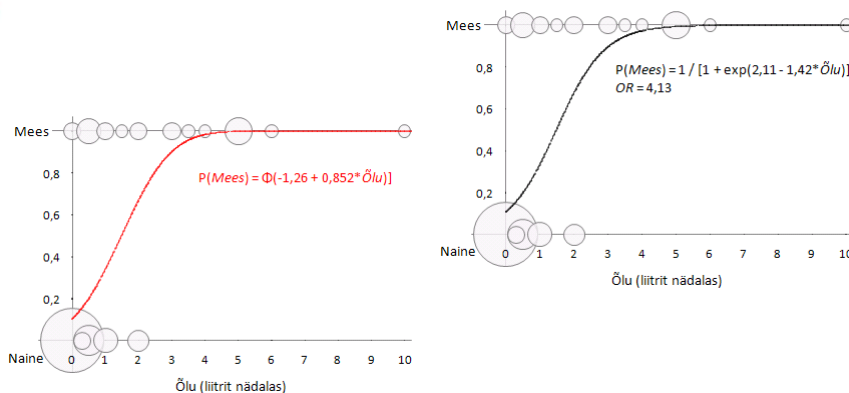
$$p = P(y=1|x) = \Phi(\alpha + \beta x) \leftrightarrow \text{probit}(p) = \Phi^{-1}(p) = \alpha + \beta x,$$
- p on huvipakkuva sündmuse Y toimumise tõenäosus,
- $\Phi(z)$ on standardse normaaljaotuse jaotusfunktsiooni väärtus kohal z :

$$\Phi(z) = P(Z \leq z), \quad Z \sim N(0,1).$$



Logistiline *versus* probit-regressioon

- Tõenäosuste hinnangutel vahe peaaegu puudub, kumba meetodit kasutada, sõltub enamasti mingis valdkonnas välja kujunenud traditsioonidest.
- Logistilise regressiooni täiendav tulemus on šansside suhe.



LD90, LTemp50, ...

- LD90, LTemp50, ... on logit- või probit-mudelil baseeruvad argumenttunnuse hinnangulised väärtused, mille korral uuritav sündmus leiab aset ette antud tõenäosusega.
Näiteks 90%-liselt surmav doos (90% *lethal dose*, LD90)
või 50%-liselt surmav temperatuur (50% *lethal temperature*, LTemp50).

- Logistilise regressiooni mudeli esitusest $\ln[p/(1-p)] = \alpha + \beta x$ järeldeb, et

$$x = \{\ln[p/(1-p)] - \alpha\} / \beta.$$

- Probit-regressiooni mudeli esitusest $\Phi^{-1}(p) = \alpha + \beta x$ järeldeb, et

$$x = [\Phi^{-1}(p) - \alpha] / \beta.$$

Supercooling ability and cold hardness of the pollen beetle *Meligethes aeneus*

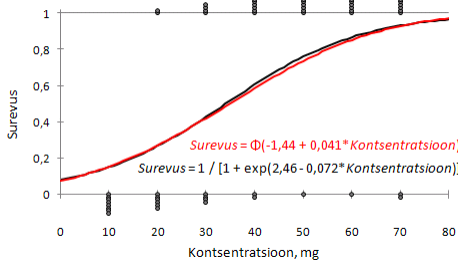
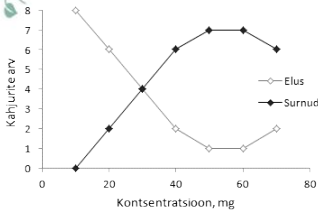
Külli Hiiesaar*, Ingrid H Williams, Marika Mänd, Anne Luik, Katrin Jõgar, Luule Metspalu, Eha Sivilponis, Angela Ploomi & Irja Kivimägi
Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Kreutzwaldi St. 1, 51014 Tartu, Estonia

Table 1 Mean (± SE) lethal time [Ltime₅₀ (h); time required for 50% mortality at constant -7 °C] and lethal temperature [Ltemp₅₀ (°C); temperature required for 50% mortality after 24 h exposure] of *Meligethes aeneus* beetles

Ranking time	n	Ltime ₅₀	95% fiducial limits	n	Ltemp ₅₀	95% fiducial limits
June	333	56.2 ± 0.4a	44.3–65.6	231	-8.6 ± 0.2a	-8.7 to -8.0
August	575	80.0 ± 0.2b	73.0–86.4	294	-9.1 ± 0.9b	-9.3 to -8.8
September	406	182.8 ± 0.1c	120.9–347.8	587	-9.8 ± 1.1b	-10.5 to -8.9
January/February	376	418.1 ± 0.2c ¹	336.0–597.5	374	-9.5 ± 0.7b ²	-10.3 to -8.8

Different letters within a column indicate significant differences between seasonal groups based on the lack of overlap of their 95% fiducial limits.

Logistiline ja probit-regressioon – näide



- Logistilise mudeli kohaselt

$$\text{Surevus} = 1 / [1 + \exp(2,46 - 0,072 * \text{Kontsentratsioon})],$$

mistap 90%-liselt surmav kontsentratsioon

$$\text{LC90} = \{\text{LN}[0,9/(1-0,9)] - 2,46\} / 0,072 = 64,60 \text{ mg.}$$

- Probit-mudeli kohaselt

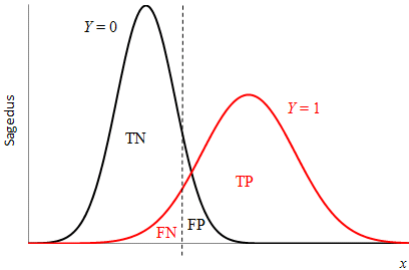
$$\text{Surevus} = \Phi(-1,44 + 0,041 * \text{Kontsentratsioon}),$$

mistap 90%-liselt surmav kontsentratsioon

$$\text{LC90} = [\Phi^{-1}(0,9) + 1,44] / 0,041 = (1,28 + 1,44) / 0,041 = 65,82 \text{ mg.}$$

Testi/mudeli tundlikkus ja spetsiifilisus

- Tõeselt positiivne (*true positive*, TP).
- Valepositiivne (*false positive*, FP).
- Tõeselt negatiivne (*true negative*, TN).
- Valenegatiivne (*false negative*, FN).



Prognosis	Tegelik olek	
	Y = 0	Y = 1
Y = 0	TN	FN
Y = 1	FP	TP

- Tundlikkus (*sensitivity*, *sensitivity*) näitab, kui suure osa (kui mitu protsenti) uuritava sündmuse toimumistest ennustab kasutatud mudel õigesti:

$$\text{Tundlikkus} = TP / (TP + FN).$$

- Spetsiifilisus (*specificity*) näitab, kui suure osa (kui mitu protsenti) uuritava sündmuse mittetoimumistest ennustab kasutatud mudel õigesti:

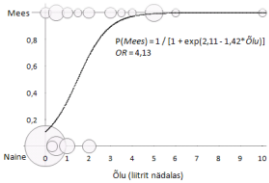
$$\text{Spetsiifilisus} = TN / (TN + FP).$$

Testi/mudeli tundlikkus ja spetsiifilisus

Näide: tudengi sugu *versus* nädalane õllekogus

Logistiline mudel:

$$P(\text{Mees}) = 1 / [1 + \exp(2,11 - 1,42 \cdot \text{Õlu})]$$



Õlu (l)	p	Mees Naine		TP	TN	FP	FN	Tundlikkus	Spetsiifilisus
0	0,108	2	27	1	0	1	0	1	0
≤0,3	0,157	0	2	0,909	0,643	0,357	0,091	0,909	0,643
≤0,5	0,198	4	6	0,909	0,690	0,310	0,091	0,909	0,690
≤1	0,334	2	4	0,727	0,833	0,167	0,273	0,727	0,833
≤1,5	0,505	1	0	0,636	0,929	0,071	0,364	0,636	0,929
≤2	0,675	2	3	0,591	0,929	0,071	0,409	0,591	0,929
≤3	0,896	2	0	0,5	1	0	0,5	0,5	1
≤3,5	0,946	1	0	0,409	1	0	0,591	0,409	1
≤4	0,973	1	0	0,364	1	0	0,636	0,364	1
≤5	0,993	5	0	0,318	1	0	0,682	0,318	1
≤6	0,998	1	0	0,091	1	0	0,909	0,091	1
≤10	1,000	1	0	0,045	1	0	0,955	0,045	1

22 42

ROC-kõver

- ROC-kõver (*receiver operating characteristic curve*) või üldisemalt ROC-analüüs tähendab testi või mudeli headuse hindamist läbi tundlikkuse ja spetsiifilisuse.

Näide: tudengi sugu *versus* nädalane õllekogus.

Logistiline mudel:

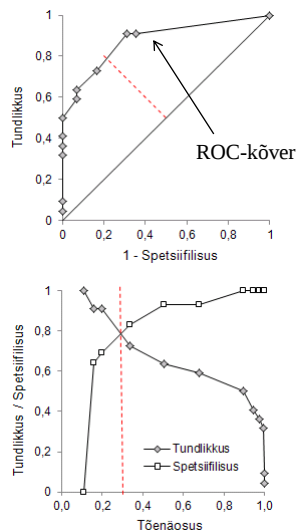
$$P(\text{Mees}) = 1 / [1 + \exp(2,11 - 1,42 \cdot \tilde{O}lu)].$$

Ligikaudne optimaalne logistilisest mudelist hinnatud tõenäosus, eristamaks õlle joomise alusel naisi ja mehi, on 0,3.

Nii sellele väärtusele vastavad tundlikkus kui ka spetsiifilisus on 0,8 ja

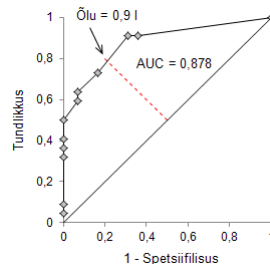
vastav õllekogus on

$$\tilde{O}lu_{30} = \{\text{LN}[0,3/(1-0,3)] - 2,11\} / 1,42 = 0,9 \text{ l.}$$



ROC-kõvera alune pindala

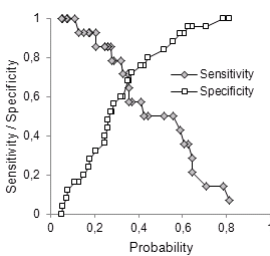
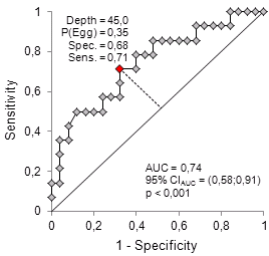
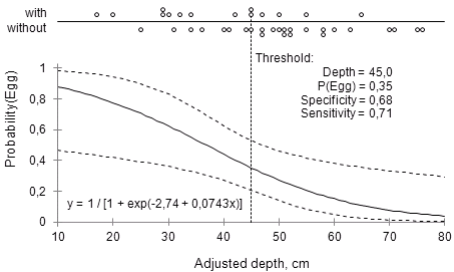
- ROC-kõvera alune pindala (*area under the curve*, AUC) on üks testi /mudeli headuse mõõtet.
- Kui uuritava sündmuse toimumist ennustada näiteks kulli ja kirja viskamise teel, on ennustuse täpsus 0,5 (50%) – seega vastab AUC = 0,5 juhule, kus mudeli argument mingit rolli ei mängi, ehk mingit seost ei ole. Mida enam erineb AUC 0,5-st, seda täpsemini antud mudel ennustab ehk prognoosib (seda parem on mudel).
- Kokkuleppelised piirid, hindamaks testi/mudeli headust:
 - kui $AUC \geq 0,9$, siis on testi/mudeli täpsus suurepärane (*excellent*),
 - $AUC \geq 0,8$ puhul hea (*good*),
 - $AUC \geq 0,7$ puhul rahuldav (*fair*),
 - $AUC \geq 0,6$ puhul kasin (*poor*) ja alla selle ei ole erilist mõtet ennustuse/prognoosi täpsusest rääkida.



Logistiline regressioon ja ROC-kõver

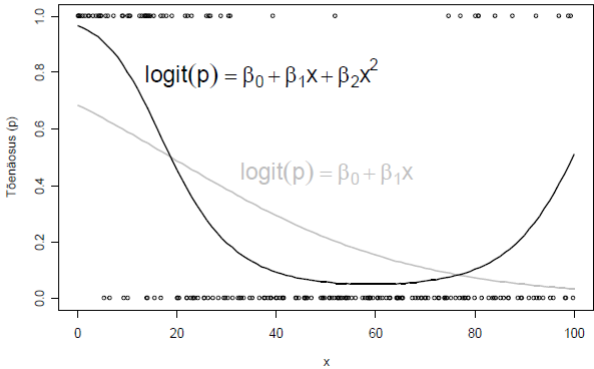
Näide. Kui sügavad veekogud sobivad mudakonnale (*bufo calamita*) kudemiseks?

- Logistilise regressiooni ja ROC-kõvera analüüsi alusel on sobivaimaks piiriks eristamiseks mudakonnale kudemiseks sobivaid ja mitesobivaid veekogusid veekogu maksimaalne sügavus 45 cm,
- seejuures on 45 cm-st madalamad 71% kudeveekogusid ja 45 cm-st sügavamad 68% kudemiseks mittevõlitud veekogusid ning
- kudemiseks sobivaks võib veekogu lugeda juba siis, kui logistilisest mudelist hinnatud kudemiseks sobivuse tõenäosus on üle 0,35;
- $OR = e^{-0,0743} = 0,928$, seega väheneb veekogu sügavuse suurenemisega 1 cm võrra šanss sobida mudakonnale kudemiseks 0,928 korda.



Logistiline regressioon

Mis teha, kui pakutud seos ei sobi?



Joonis Märt Mölsi loengukonspektist

Logistiline mudel kui üldistatud lineaarne mudel (*generalized linear model*)

Mudel hindamaks soo mõju ravi tulemusele

$$\text{logit}(p_{ij}) = \mu + S_i + \varepsilon_{ij},$$

millest

$$\hat{p}_{ij} = \exp(\mu + S_i) / [1 + \exp(\mu + S_i)].$$

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

Effect	sugu	Estimate	Standard Error
Intercept		-1.6094	0.7716
sugu	E	2.4204	0.9804
sugu	I	0	.

Seega

$$P(\text{haige}|\text{sugu}=I) = \exp(-1,61 + 2,42) / [1 + \exp(-1,61 + 2,42)],$$

$$P(\text{haige}|\text{sugu}=E) = \exp(-1,61 + 0) / [1 + \exp(-1,61 + 0)]$$

ja

$$OR = \exp(2,42) = 11,25,$$

st et šanss mitte terveks saada on emaste hulgas 11,25 korda kõrgem.