ÕPIOBJEKT

„Binaarsete tunnuste analüüsimeetodid“

Tanel Kaart

http://ph.emu.ee/~ktanel/ bin_tunnuste_analyys/

www.emu.ee
Eesti Maaülikool
Estonian University of Life Sciences

Euroopa Liit
Euroopa Sotsiaalfond

Eesti tuleviku heaks

Õpiobjektid -> Binaarsete tunnuste analüüsimeetodid

BINAARSETE TUNNUSTE ANALÜÜSIMEETODID

Õpiobjekti kirjeldus
Õpijuhis

1. Sissejuhatus
   - Binaarsete tunnuste olemus ja kodeerimine
   - Binaarsete tunnuste esitus arvutis
   - Binaarne tunnus kui fiktiivne muutuja
2. Binaarse tunnuse seos mittearvulise tunnusega või diskreetse arvtunnusega
   - Kahemõõtmeline sagedustabel
   - Suhtelised sagedused
   - Sõltumatuse juhule vastavad sagedused
   - Juhtude esinemissagedus ja riskisuhe
   - Šansside suhe
   - Hii-ruut test
   - Fisheri täpne test
3. Binaarse tunnuse seos pideva arvtunnusega
   - Logistiline regressioon
   - Probit-regressioon
   - Logit- *vs* probit-regressioon ning tulemuste illustreerimine
   - 50% ja 90% vastuse määr (LD50, LTemp90 jmt)
   - Tundlikkus ja spetsiifilisus
   - ROC-kõver
   - Optimaalne piirväärtus
   - ROC-kõvera alune pindala
   - Diskreetse argumendiga logistiline mudel
4. Enesekontroll
   - Küsimused ja ülesanded
   - Vastused ja lahendused

Lisa
   - Kogu materjal ühe pdf-failina: bin_tunnuste_analyys.pdf

Õpiobjekti kirjeldus

Õppekava: Keskkonnateadus ja rakendusbioloogia (80130), Metsandus (80131), Põllumajandus (80132), Tehnikateadus (80133), Veterinaarmeditsiin ja toiduteadus (80134)

Õppeaine: DK.0007 Matemaatiline statistika ja modelleerimine; VL.0435 Katsetöö metoodika ja statistiline andmetöötlus

Maht: 10 tundi

Sihtrühm: EMÜ doktorikooli tudengid, VLI looma- ja kalakasvatuse magistrandid ja teised asjahuvilised.

Õpiobjekti vajalikkuse põhjused:

Mitmed binaarsete tunnuste analüüsimisel kasutatavad meetodid on aegade jooksul kujunenud traditsioonide läbi muutunud uurimisvaldkonna-spetsiifilisteks. Antud õpiobjekt tutvustab erinevaid binaarsete tunnuste analüüsimeetodeid laiendamaks nende rakendusvaldkondi, kirjeldades selleks meetodite vahelisi seoseid ja erinevusi ning nende kasutatavust erinevate hüpoteeside kontrollimiseks ja parameetrite hindamiseks. Käsitletakse teemasid nagu riski- ja šansside suhe, logistiline ja probit-regressioon, tundlikkus, spetsiifilisus ja ROC-kõverad, 50% ja 90% vastuse määr.

Eesmärk: Õpiobjekti eesmärk on toetada õppeaine omandamist ja olla toeks edasisel teadustööl.

Õpiobjekti läbinu:

- mõistab binaarsete tunnuste analüüsimeetodite olemust,
- suudab valida oma küsimustele ja andmete struktuurile vastava analüüsimeetodi,
- suudab võtta vastu otsuseid ja sõnastada järeldusi tuginedes analüüsitulemustele,
- omab võimalust enesekontrolliks.

Sisu ja tehniline teostus: Tanel Kaart

Eesti Maaülikool
sügissemester 2012

Järgmine >

Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

Sisukord

1. Sissejuhatus

1.1. Binaarsete tunnuste olemus ja kodeerimine

Binaarne tunnus on tunnus, millel on vaid kaks võimalikku väärtust (näiteks sugu väärtustega ’isane’ ja ’emane’).

Katse/uuringu tulemusena tekivad binaarsed tunnused siis, kui uuritakse mingi sündmuse toimumist – a’la lehm tiinestus/ei tiinestunud, koera ravi oli tulemuslik/ei olnud tulemuslik, kahjur suri ära/ei surnud ära jne.

Sageli moodustatakse binaarne tunnus algselt pideval skaalal mõõdetud tunnuse väärtuste alusel – näiteks fikseerides ära piirid normaalse vererõhu tarvis, saab reaalselt mõõdetud vererõhu väärtuste alusel moodustada binaarse tunnuse väärtustega vererõhk on normis/ei ole normis.

Analüüsimiseks esitatakse binaarne tunnus enamasti nö 0-1-tunnusena, st et huvipakkuva sündmuse toimumist tähistatakse arvuga 1 ja sündmuse mittetoimumist arvuga 0. Sellise tähistuse puhul näitab tavaline aritmeetiline keskmine uuritava sündmuse toimumise osakaalu. Näiteks oletades, et kuuest ravitud koerast neli said terveks, ning tähistades terveks saamist ühega ja terveks mittesaamist nulliga – andmestik, kus kaks esimest koera ei saanud terveks ja neli järgnevat said, on siis kujul {0,0,1,1,1,1} –, annab aritmeetiline keskmine (0+0+1+1+1+1)/6=4/6=2/3 terveks saanud koerte osakaalu (2/3 ehk 66,7% ravitud koertest said terveks).

Et suur osa binaarsete tunnuste analüüsimetoodikast on välja kasvanud epidemioloogiast, kasutatakse teooria esitamisel enamasti ka epidemioloogiast pärit mõisteid.

Näiteks indiviide/objekte, kelle/mille puhul leidis aset uuritav sündmus (a’la lehm jäi tiineks, inimene põdes uuritavat haigust, taimekahjur suri ära, talunik läks pankrotti jne), nimetatakse **juhtudeks** (ingl. ***cases***, ***responders***) ja neid, kelle/mille puhul uuritavat sündmust ei toimunud, nimetatakse **kontrollideks** (ingl. ***controls***, ***nonresponders***), mõnikord ka **baas-** ehk **referentsgrupiks** (ingl. ***reference***).

Juhul, kui uuritava sündmuse toimumist potentsiaalselt mõjutav või sellega assotsieeruv faktor on kaheväärtuseline (a’la sai ravimit/ei saanud ravimit, on naine/ei ole naine jne), nimetatakse indiviide/objekte, kelle/mille puhul realiseerus faktori huvipakkuv variant (a’la sai ravimit, on naine jne), **eksponeerituteks** (ingl. ***exposed***), ja ülejäänuid (ei saanud ravimit, ei ole naine jne) **mitteeksponeerituteks** (ingl. ***non-exposed***). Mitteeksponeerituid ja uuritava sündmuse toimumise sagedust neil käsitletakse enamasti nö baasina, mille suhtes hinnatakse uuritava sündmuse toimumist eksponeerituil.

1.2. Binaarsete tunnuste esitus arvutis

Lihtsaim ja universaalseim binaarsete tunnuste esitus eeldab, et andmetabelis vastab igale uurimisobjektile üks rida. Sellisel juhul saab ka iga objekti tarvis kirja panna, kas huvipakkuv sündmus selle objekti puhul toimus või mitte (tunnuse väärtuseks vastavalt 1 või 0).

Näiteks uurides tudengite rahulolu õppetöö korraldusega sõltuvalt nende eksamihindest ja soost, võib andmetabeli üles ehitada järgneval kujul:

Tudengi jrk nr	Eksamihinne	Sugu	Rahul õppetööga (jah/ei)
1	A	N	1
2	C	M	0
3	A	N	0
4	B	N	1
5	B	M	1
6	B	N	0
7	B	N	1
8	A	M	1

Alternatiivne variant sama andmetabeli esitamiseks, mida mitmed arvutiprogrammid aktsepteerivad või suisa nõuavad, on esitus sõltuvalt uurimishüpoteesist. Sellisel juhul koostatakse andmetabel, milles on üks rida iga analüüsitava(te) faktori(te) taseme(te kombinatsiooni) tarvis. Selles reas on kirjas vastava taseme (või tasemete kombinatsiooni) esinemiste arv ning uuritava sündmuse toimumiste arv antud taseme (või tasemete kombinatsiooni) korral.

Näiteks soovides uurida seost eksamihinde ja õppetöö korraldusega rahulolu vahel, jättes kõrvale võimaliku soo mõju, võib andmetabel olla kujul

Eksamihinne	Tudengite arv	Õppetööga rahul olnud tudengite arv
A	3	2
B	4	3
C	1	0

Soovides teostada aga keerulisemaid analüüse, uurides õppetöö korraldusega rahulolemise seotust korraga nii eksamihinde kui ka tudengi sooga, peaks andmetabelis olema üks rida iga andmestikus esineva eksamihinde ja soo väärtuste kombinatsiooni tarvis:

Eksamihinne	Sugu	Tudengite arv	Õppetööga rahul olnud tudengite arv
A	M	1	1
A	N	2	1
B	M	1	1
B	N	3	2
C	M	1	0

Et eksamihindega ’C’ naisterahvaid andmestikus polnudki, puudub eelnevast tabelist ka eksamihinde ’C’ ja soo ’N’ kombinatsioonile vastav rida.

NB! Kui uurimisobjektidel on mõõdetud ka mõne pideva tunnuse väärtus, on ainus variant esitada andmetabel objektiviisi – igale uurimisobjektile vastab üks rida, kus on kirjas nii mõõdetud tunnus(t)e väärtus(ed) kui ka info uuritava sündmuse toimumise/mittetoimumise kohta (kujul 1/0).

Tudengi jrk nr	Eksamihinne	Sugu	Vanus	Rahul õppetööga (jah/ei)
1	A	N	22	1
2	C	M	20	0
3	A	N	21	0
4	B	N	22	1
5	B	M	21	1
6	B	N	23	0
7	B	N	28	1
8	A	M	21	1

1.3. Binaarne tunnus kui fiktiivne muutuja

Ankeetküsitluse puhul, kus ühele küsimusele võib anda mitu vastust, tuleks andmete analüüsimiseks tekitada ühest küsimusest nii mitu vaid kahe üksteist välistava vastusevariandiga alamküsimust, kui palju on algsel küsimusel vastusevariante. Näiteks kui küsimuses talu tegevusala kohta võib talu tegevusalaks märkida ’turismi’, ’loomakasvatuse’ või ’taimekasvatuse’ või ükskõik millise kombinatsiooni antud tegevusaladest, a’la ’turism, loomakasvatus’), tuleks andmetabelisse tekitada kolm nulle ja ühtesid väärtustena kasutavat **fiktiivset binaarset tunnust** (inglise keeles ***dummy variable***), millest esimene märgib seda, kas talu tegevusalaks on turism või mitte, teine seda, kas talu tegevusalaks on loomakasvatus või mitte, ja kolmas seda, kas talu tegevusalaks on taimekasvatus või mitte (vt järgnevat tabelit).

Tegevusala		Turism	Loomakasvatus	Taimekasvatus
Turism		1	0	0
Turism, loomakasvatus	->	1	1	0
Loomakasvatus		0	1	0
Loomakasvatus, taimekasvatus		0	1	1

Mõnikord esitatakse üks mitmeväärtuselise tunnus seeria binaarsete tunnustena ka siis, kui igas andmetabeli reas saab olla vaid üks väärtus. Näiteks uurides teatud veeselgrootsete esinemist sõltuvalt setetest veekogus (liiv, savi, turvas), võib veeru ’Setted’ alusel moodustada kolm nulle ja ühtesid sisaldavat veergu ’Liiv, ’Savi’, ’Turvas’ (vt järgnevat tabelit). Taoline esitus annab võimaluse teostada lihtsalt võrdlusi ’liivased vs mitteliivased’ vmt, samuti on taoline esitus mõttekas mõnede mitmemõõtmeliste statistikameetodite puhul.

Setted		Liiv	Savi	Turvas
Liiv		1	0	0
Savi	->	0	1	0
Liiv		1	0	0
Turvas		0	0	1

2. Binaarse tunnuse seos kategoorilise tunnusega (diskreetse faktoriga)

2.1. Kahemõõtmeline sagedustabel

Kahemõõtmeline sagedustabel näitab, kui mitu korda huvipakkuv sündmus toimus (nö juhud) ja kui mitu korda ei toimunud (kontrollid) diskreetse faktori eri tasemetel. Kaheväärtuselise faktori puhul on tegu 2x2-tabeliga:

	Juht	Kontroll
Eksponeeritud	*a*	*b*
Mitteeksponeeritud	*c*	*d*

Näiteks uurides, kas teatud haigust põdenud koerte tervenemine sõltub koera soost, võib uuritud 25 koera jagada kahemõõtmelisse sagedustabelisse kujul

	Ei saanud terveks	Sai terveks
Emane	9	4
Isane	2	10

Selle tabeli alusel kokku 12-st isasest koerast 10 said ja kaks ei saanud terveks, 13-st emasest koerast aga neli said ja üheksa ei saanud terveks.

Enamasti lisatakse kahemõõtmelisse sagedustabelisse ka summaveerg ja -rida:

	Juht	Kontroll	Kokku
Eksponeeritud	*a*	*b*	*a+b*
Mitteeksponeeritud	*c*	*d*	*c+d*
Kokku	*a+c*	*b+d*	$n = a+b+c+d$

Sama tabel koerte näite tarvis:

	Ei saanud terveks	Sai terveks	Kokku
Emane	9	4	13
Isane	2	10	12
Kokku	11	14	25

Muidugi ei ole kahemõõtmelise sagedustabeli rakendamine piiratud vaid kaheväärtuseliste tunnustega. Üldjuhul võib tunnustel *X* ja *Y*, mille vahelist seost uuritakse, olla vastavalt *m* ja *k* erinevat väärtust $x_1, x_2, \ldots, x_m$ ja $y_1, y_2, \ldots, y_k$. Nende tunnuste võimalikku omavahelist seotust kirjeldav kahemõõtmeline sagedustabel omab siis *m* veergu (iga veerg vastab erinevale tunnuse *X* väärtusele) ja *k* rida (iga rida vastab erinevale tunnuse *Y* väärtusele). Nii tekib *m* x *k* lahtrit, millest igaüks vastab tunnuste väärtuste erinevale kombinatsioonile. Igasse lahtrisse kirjutatakse vastava väärtuspaari esinemise sagedus n_{ij}. Sagedustabeli viimane rida ja veerg saadakse reasagedusi ja veerusagedusi kokku liites. Sellele viitavad ka tähistused $n_{i.}$ ja $n_{.j}$,

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

Punktiga tähistatakse need indeksid, üle mille käis vastava suuruse arvutamisel summeerimine.

Sagedustabeli alumises parempoolses lahtris esitatakse tavaliselt valimi maht *n*. Kuna see arvutatakse ääresageduste summeerimise teel,

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

siis võib vastatava suuruse tähistamiseks kasutada ka sümbolit $n_{..}$.

Kahemõõtmelise sagedustabeli üldkuju on järgmine:

	y_1	y_2	…	y_k	Kokku
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}	$n_{m.}$
Kokku	$n_{.1}$	$n_{.2}$	…	$n_{.k}$	n

2.2. Suhtelised sagedused

Sageli esitatakse kahemõõtmelises sagedustabelis absoluutsete sageduste kõrval ka suhtelised sagedused. Viimaseid on võimalik arvutada kolmel viisil:

- jagades lahtrite sagedused n_{ij} läbi vastavate rea ääresagedusega $n_{i.}$, saame **rea suhtelised sagedused**, mis määravad tunnuse *Y* tingliku jaotuse antud reale vastava *X* väärtuse korral;
- jagades lahtrite sagedused n_{ij} läbi vastava veeru ääresagedusega $n_{.j}$, saame **veeru suhtelised sagedused**, mis määravad tunnuse *X* tingliku jaotuse antud veerule vastava *Y* väärtuse korral;
- jagades lahtrite sagedused n_{ij} läbi valimi mahuga *n*, saame **tabeli suhtelised sagedused**.

Koerte näite puhul avalduvad rea suhtelised sagedused kujul

	Ei saanud terveks	Sai terveks	Kokku
Emane	0,69	0,31	1
Isane	0,17	0,83	1
Kokku	0,44	0,56	1

Tabelis olevad sagedused näitavad, kui palju oli terveks saanud ja terveks mittesaanud koeri erinevate sugude korral. Näiteks suurus 0,83, mis on arvutatud kui terveks saanud isaste koerte arv 10 jagatud kõigi isaste koerte arvuga 12, näitab, et isastest koertest said terveks 83%. Samas oli terveks saanud emaseid koeri vaid 31% kõigist ravitud emastest koertest. Kõigist koertest kokku 56% said terveks ja 44% ei saanud terveks.

Veeru suhtelised sagedused, mis näitavad sugude jaotust eraldi terveks saanud ja terveks mittesaanud koerte hulgas, on järgmised:

	Ei saanud terveks	Sai terveks	Kokku
Emane	0,82	0,29	0,52
Isane	0,18	0,71	0,48
Kokku	1	1	1

Terveks saanud koertest 71% olid isased ja 29% emased, samas kui terveks mittesaanud koertest olid isased vaid 18% ja emased tervelt 82%. Kõigist ravitud koertest 48% olid isased ja 52% emased.

Kogu tabeli suhtelised sagedused näitavad iga väärtuste kombinatsiooni esinemissagedust:

	Ei saanud terveks	Sai terveks	Kokku
Emane	0,36	0,16	0,52
Isane	0,08	0,40	0,48
Kokku	0,44	0,56	1

Kõigist ravitud koertest 40% olid terveks saanud isased, 16% terveks saanud emased, 8% terveks mittesaanud isased ja 36% terveks mittesaanud emased.

2.3. Sõltumatuse juhule vastavad sagedused

Juhul, kui tunnused X ja Y on sõltumatud, peaks mistahes väärtuste kombinatsiooni $x_i y_j$ esinemissagedus n_{ij} võrduma avaldise $n_{i.}n_{.j}/n$ väärtusega. Sõltumatuse juhule vastavate oodatavate sageduste võrdlus tegelike sagedustega võimaldab välja selgitada sõltumatuse juhust enim erinevad väärtuste kombinatsioonid, samuti baseerub taolisel võrdlusel tunnuste X ja Y vahelise seose statistilise olulisuse testimisel kasutatav χ^2-test (punkt 2.5).

Näiteks koerte andmestiku puhul on isase koera valikuks 12 võimalust 25-st, terveks saanud koera valikuks 14 võimalust 25-st. Eeldusel, et terveks saamine ja sugu ei ole seotud, peaks tõenäosus, et juhuslikult valitud koer on terveks saanud isane, avalduma korrutisena (12/25)x(14/25). Kuna kokku oli uuringus 25 koera, peaks neist terveks saanud isaseid olema 25x(12/25)x(14/25) = (12x14)/25 = 6,72.

Analoogselt on leitav, et terveks saanud emaseid peaks soo ja ravi tulemuse sõltumatuse korral olema (13x14)/25 = 7,28, terveks mittesaanud isaseid (12x11)/25 = 5,28 ja terveks mittesaanud emaseid (13x11)/25 = 5,72:

Oodatavad sagedused	Ei saanud terveks	Sai terveks	Kokku
Emane	5,72	7,28	13
Isane	5,28	6,72	12
Kokku	11	14	25

Kõrvutades viimati leitud oodatavaid sagedusi tegelike sagedustega

	Ei saanud terveks	Sai terveks
Emane	9	4
Isane	2	10

on näha, et terveks saanud isaseid koeri ja terveks mittesaanud emaseid koeri on tegelikkuses märksa enam, kui võinuks eeldada soo ja ravi tulemuse sõltumatuse korral, samavõrd vähem võrreldes sõltumatuse juhuga on aga terveks mittesaanud isaseid ja terveks saanud emaseid.

2.4. Juhtude esinemissagedus ja riskisuhe

Juhtude esinemissagedus ehk **risk** (epidemioloogias **haigestumuskordaja**; ingl. ***incidence rate***, ***rate***) leitakse iga riskifaktori taseme tarvis kui juhtude esinemise arv jagatuna kõigi antud tasemele vastavate vaatluste arvuga.

Kui tegu on 2x2-tabeliga kujul

	Juht	Kontroll	Kokku
Eksponeeritud	*a*	*b*	*a*+*b*
Mitteeksponeeritud	*c*	*d*	*c*+*d*
Kokku	*a*+*c*	*b*+*d*	*n* = *a*+*b*+*c*+*d*

avalduvad juhtude esinemissagedused ehk riskid suhetena $a/(a+b)$ ja $c/(c+d)$.

Riskisuhe (ingl. ***risk ratio***, ***relative risk***; *RR*) avaldub suhtena

RR = (juhu risk eksponeeritutel)/(juhu risk mitteeksponeeritutel),

ehk, kasutades tabelis toodud tähistusi,

$$RR = [a/(a+b)] / [c/(c+d)].$$

Et mitteeksponeerituid käsitletakse nö baasgrupina (vt punkt 1.1), on nende puhul riskisuhte väärtuseks üks.

Juhul, kui võrreldavaid gruppe on enam kui kaks, leitakse riskisuhe tavaliselt kas vähima juhtude esinemissagedusega rea (faktori taseme) või siis sisulistel kaalutlustel nö baasina ehk referentsgrupina käsitletava faktori taseme suhtes.

Riskisuhte alusel järelduste sõnastamine lähtub sellest, et

- kui sündmuse toimumise esinemissagedus võrreldavais gruppides on ühesugune, on riskisuhte väärtuseks üks ($RR = 1$),
- kui riskisuhte väärtus on ühest suurem ($RR > 1$), on uuritava sündmuse toimumise sagedus eksponeerituil (katsegrupis) suurem, kui mitteeksponeerituil (kontrollgrupis),
- kui riskisuhte väärtus on aga ühest väiksem ($RR < 1$), on uuritava sündmuse toimumise sagedus eksponeerituil (katsegrupis) väiksem, kui mitteeksponeerituil (kontrollgrupis).

Otsustamaks võrreldavate gruppide vahelise erinevuse statistilise olulisuse üle, võib kasutada riskisuhte 95%-list **usaldusintervalli** (ingl. ***confidence interval***; *CI*):

- kui erinevuse puudumisele (nullhüpoteesile) vastav arv 1 jääb usaldusintervalli sisse, ei ole katse- ja kontrollgrupi vaheline erinevus statistiliselt oluline ($1 \in 95\% CI_{RR}$ -> $p>0{,}05$),
- kui aga usaldusintervall arvu 1 ei sisalda, on katse- ja kontrollgrupi vaheline erinevus statistiliselt oluline ($1 \notin 95\% CI_{RR}$ -> $p<0{,}05$).

95%-line usaldusintervall riskisuhtele on ligikaudu hinnatav valemist

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

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

Koerte näites on terveks mittesaamise risk emastel koertel 9/(9+4) = 0,69 ja isastel koertel 2/(2+10) = 0,17.

Riskisuhe avaldub aga kujul

$$RR = [9/(9+4)] / [2/(2+10)] = 4{,}15.$$

Seega on emastel koertel 4,15 korda suurem risk terveks mitte saada.

95%-lised usalduspiirid riskisuhtele tulevad (0,90; 19,23). Et arv üks sisaldub leitud 95%-lises usaldusintervallis, ei ole olulisuse nivool 0,05 alust väita, nagu erineks terveks mittesaamise risk emastel ja isastel statistiliselt oluliselt.

Valides faktori sugu baastasemeks hoopis isased, tuleb riskisuhteks

$$RR = [2/(2+10)] / [9/(9+4)] = 0{,}24.$$

See tähendab, et isastel koertel on risk terveks mitte saada 0,24 korda väiksem, kui emastel koertel. 95%-line usaldusintrevall tuleb sellisel juhul (0,05; 1,11), mis ei anna samuti alust lugeda soo ja ravi tulemuse seotust tõestatuks.

2.5. Šansside suhe

Sündmuse toimumise **šansid** (ingl. ***odds***) näitavad, mitmel juhul sündmus toimub võrreldes sellega, mitmel juhul ta ei toimu.

Näiteks kui sündmus toimub tõenäosusega 0,2 (20%) ehk ühel juhul viiest, siis selle sündmuse toimumise šansid on üks nelja vastu ehk 1:4.

Koerte näites on emaste koerte šanss mitte terveks saada 9/4 = 2,25 ja isaste koerte šanss mitte terveks saada 2/10 = 0,2. Terveks saamise šansid, millest on justkui loomulikum rääkida, on emaste koerte puhul 4/9 = 0,44 ja isaste koerte puhul 10/2 = 5.

Šansside suhe (ingl. ***odds ratio***; *OR*) näitab, kui mitu korda erineb uuritava sündmuse toimumise šanss eksponeerituil võrreldes mitteeksponeeritutega:

$$OR = (\text{juhu šanss eksponeeritutel})/(\text{juhu šanss mitteeksponeeritutel}),$$

ehk, kasutades 2x2-tabelis toodud tähistusi,

$$OR = (a/b) / (c/d).$$

Et mitteeksponeerituid käsitletakse nö baas- ehk referentsgrupina (vt punkt 1.1), on nende puhul šansside suhte väärtuseks üks.

Juhul, kui võrreldavaid gruppe on enam kui kaks, leitakse šansside suhe tavaliselt kas vähima juhtude esinemissagedusega rea (faktori taseme) või siis sisulistel kaalutlustel nö baasina ehk referentsgrupina käsitletava faktori taseme suhtes.

Šansside suhte alusel järelduste sõnastamine lähtub sellest, et

- kui sündmuse toimumise šanss võrreldavais gruppides on ühesugune, on šansside suhte väärtuseks üks ($OR = 1$),
- kui šansside suhte väärtus on ühest suurem ($OR > 1$), on uuritava sündmuse toimumise šanss (ja sestap ka tõenäosus) eksponeerituil (katsegrupis) suurem, kui mitteeksponeerituil (kontrollgrupis),

- kui šansside suhte väärtus on ühest väiksem ($OR < 1$), on uuritava sündmuse toimumise šanss (ja sestap ka tõenäosus) eksponeerituil (katsegrupis) väiksem, kui mitteeksponeerituil (kontrollgrupis).

Otsustamaks võrreldavate gruppide vahelise erinevuse statistilise olulisuse üle, võib kasutada šansside suhte 95%-list usaldusintervalli:

- kui erinevuse puudumisele (nullhüpoteesile) vastav arv 1 jääb usaldusintervalli sisse, ei ole katse- ja kontrollgrupi vaheline erinevus statistiliselt oluline ($1 \in 95\%CI_{OR}$ -> $p>0,05$),
- kui aga usaldusintervall arvu 1 ei sisalda, on katse- ja kontrollgrupi vaheline erinevus statistiliselt oluline ($1 \notin 95\%CI_{OR}$ -> $p<0,05$).

95%-line usaldusintervall šansside suhtele on ligikaudu hinnatav valemist

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

kus $se[\ln(OR)] = \sqrt{\frac{1}{\text{juhtude arv eksponeeritutel}} + \frac{1}{\text{juhtude arv mitteeksponeeritutel}} + \frac{1}{\text{kontrollide arv eksponeeritutel}} + \frac{1}{\text{kontrollide arv mitteeksponeeritutel}}} = \sqrt{\frac{1}{a} + \frac{1}{c} + \frac{1}{b} + \frac{1}{d}}$.

Uurides koerte näites terveks mitte saamist emastel koertel võrreldes isaste koertega, tuleb vastava šansside suhte väärtuseks (9/4) / (2/10) = 11,25.

Sageli esitatakse šansside suhe koos usalduspiiridega tabeli kujul, kus ära on toodud ka baas- ehk referentsgrupp:

	OR	$95\%CI_{OR}$
Emane	11,25	(1,64; 76,85)
Isane	1	

Emastel koertel on šanss mitte terveks saada 11,25 korda kõrgem, kui isastel koertel, ning kuna šansside suhte 95%-line usaldusintervall ei sisalda võrdseid šansse tähendavat arvu üks, on see erinevus ka statistiliselt oluline ($p<0,05$).

Valik *online*-kalkulaatoreid, mis pakuvad šansside suhte arvutamise võimalust:

- http://vassarstats.net/odds2x2.html
- http://www.quantitativeskills.com/sisa/statistics/two2hlp.htm
- http://department.obg.cuhk.edu.hk/ResearchSupport/Independent_2x2_table.ASP
- http://www.healthstrategy.com/epiperl/epiperl.htm (väga palju erinevaid karakteristikuid)
- http://statpages.org/ctab2x2.html (lisaks väga palju erinevaid karakteristikuid)

2.6. Hii-ruut test

Klassikaliseim kahemõõtmelise sagedustabeli alusel teostav test, otsustamaks tabelis esitatud tunnuste seotuse statistilise olulisuse üle, on **χ^2-test** (ingl. ***chi-square test*** või ***Pearson's goodness-of-fit test***).

χ^2-test võrdleb andmete alusel konstrueeritud sagedustabelit nö ideaalse, sõltumatuse juhule vastava, sagedustabeliga. Viimases avalduvad sagedused kujul $n_{i.}n_{.j}/n$ (vt punkt 2.3).

Testitav hüpoteeside paar on kujul:

> H_0: tunnused on sõltumatud ehk potentsiaalne riskifaktor ei mõjuta uuritava sündmuse toimumist,
> H_1: tunnused on sõltuvad ehk potentsiaalne riskifaktor mõjutab uuritava sündmuse toimumist;

ehk matemaatilisel kujul:

> H_0: $n_{ij} = n_{i.}n_{.j}/n$,
> H_1: $n_{ij} \neq n_{i.}n_{.j}/n$.

Teststatistikut, mis mõõdab erinevust nullhüpoteesile vastava ja tegeliku sagedustabeli vahel ning mis nullhüpoteesi kehtides on ligikaudu χ^2-jaotusega vabadusastmete arvuga $(m-1)(k-1)$, nimetatakse **χ^2-statistikuks** ja see avaldub kujul

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n},$$

kus m ja k on uuritavate tunnuste erinevate väärtuste arvud. Binaarse uuritava tunnuse ja kaheväärtuselise faktortunnuse tarvis konstrueeritava 2x2-sagedustabeli puhul on teststatistik χ^2-jaotusega vabadusastmete arvuga üks.

Seega summeeritakse χ^2-statistiku väärtuse leidmiseks kõigi tegelike ja sõltumatuse juhule vastavate sageduste ruuterinevus, mis on täiendavalt läbi jagatud sõltumatuse juhule vastavate sagedustega.

Otsuse vastu võtmine, kumb hüpoteesidest kehtib, käib tänapäeval enamasti olulisuse tõenäoses p alusel, mis leitakse kui tõenäosus, et vastava χ^2-jaotusega suurus võib omandada teststatistiku väärtusega võrdse või suurema väärtuse. Kui $p \leq 0{,}05$, loetakse traditsiooniliselt tõestatuks alternatiivne hüpotees H_1, kui aga $p > 0{,}05$, jäädakse nullhüpoteesi H_0 juurde.

Kuna χ^2-testil leitav teststatistik on nullhüpoteesi kehtides χ^2-jaotusega vaid ligikaudu, ei sobi χ^2-test väga väikeste valimite analüüsimiseks. Enamasti pannakse χ^2-testi eeldusena kirja, et kõik oodatavad sagedused $n_{i.}n_{.j}/n$ peaksid olema suuremad-võrdsed viiest.

Koerte näite puhul võib testitava hüpoteeside paari sõnastada kujul:

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

χ^2-statistiku väärtuse arvutamiseks vajalikud sagedustabelid tegelike ja nullhüpoteesile vastavate sagedustega on kujul (vt punkt 2.3)

	Ei saanud terveks	Sai terveks
Emane	9	4
Isane	2	10

ja

	Ei saanud terveks	Sai terveks
Emane	5,72	7,28
Isane	5,28	6,72

Suurused $(n_{ij} - n_{i.}n_{.j}/n)^2 / (n_{i.}n_{.j}/n)$ iga sagedustabeli lahtri tarvis on kirjas järgmises tabelis

	Sai terveks	Ei saanud terveks
Isane	1,60	2,04
Emane	1,48	1,88

millest $\chi^2 = 1{,}60+1{,}48+2{,}04+1{,}88 = 6{,}997$.

Ühe vabadusastmega χ^2-jaotuse puhul on sellele teststatistiku väärtusele vastav olulisuse tõenäosus $p = 0{,}008$, mistap võib lugeda tõestatuks alternatiivse hüpoteesi: ravi tulemus on soospetsiifiline (ehk seos ravi tulemuse ja koera soo vahel on statistiliselt oluline).

Valik *online*-kalkulaatoreid, mis pakuvad χ^2-testi teostamise võimalust:

- http://vassarstats.net/ -> Frequency data
- http://www.quantitativeskills.com/sisa/statistics/two2hlp.htm
- http://www.physics.csbsju.edu/stats/contingency.html
- http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Catego.htm
- http://www.healthstrategy.com/epiperl/epiperl.htm (lisaks palju erinevaid karakteristikuid)
- http://statpages.org/ctab2x2.html (lisaks väga palju erinevaid karakteristikuid)

2.7. Fisheri täpne test

Fisheri täpne testi (ingl. ***Fisher exact test***) abil kontrollitav hüpoteeside paar on analoogne χ^2-testi hüpoteeside paariga:

H_0: tunnused on sõltumatud,
H_1: tunnused on sõltuvad;

aga tulemuseni jõudmise metoodika on erinev. Nimelt ei arvutata Fisheri täpse testi puhul teststatistiku väärtust, selle asemel leitakse kõik antud summaarsete rea- ja veerusageduste puhul võimalikud kahemõõtmelised sagedustabelid ning arvutatakse nende esinemise tõenäosused nullhüpoteesi eeldusel (tunnuste sõltumatue eeldusel). Et läbi mängitakse kõik võimalikud variandid, nimetatakse Fisheri täpset testi (ja tegelikult ka teisi kõigi võimalike variantide läbimängimisel baseeruvaid teste) **permutatsioonitestiks**. Tunnuste sõltumatue eeldusel on tõenäosus, et uuritavad indiviidid/objektid on tabeli latrite vahel jaotunud just mingil konkreetsel viisil, leitav hüpergeomeetrilisest jaotusest.

Üldjuhul avaldub mingi konkreetse *m*x*k*-sagedustabeli esinemise tõenäosus hüpergeomeetrilise jaotuse tõenäosusfunktsioonist kujul

$$p_{\text{tabel}} = \frac{\prod_{i=1}^{k} n_{i.}! \;\; \prod_{j=1}^{m} n_{.j}!}{n! \prod_{i,j} n_{i,j}!},$$

kus $n! = n\text{x}(n-1)\text{x}\ldots\text{x}2\text{x}1$ on n-i faktoriaal (definitsiooni kohselt on ka $0! = 1$).

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

$$p_{\text{tabel}} = [(a+b)!(c+d)!(a+c)!(b+d)!] / [n!a!b!c!d!].$$

Otsuse vastuvõtmiseks vajaliku olulisuse tõenäosuse arvutamiseks on Fisheri täpse testi puhul kaks võimalust:

- summeeritakse andmetele vastava sagedustabeli ja kõigi sellest väiksema esinemis-tõenäosusega tabelite tõenäosused;
- summeeritakse andmetele vastava sagedustabeli ja sellest nö samas suunas ekstreemsemate tabelite esinemistõenäosused ja korrutatakse tulemus kahega.

Koerte näite puhul on algne andmetele vastav sagedustabel kujul

	Ei saanud terveks	Sai terveks	Kokku
Emane	**9**	**4**	13
Isane	**2**	**10**	12
Kokku	11	14	25

ja sellise tabeli saamise tõenäosus soo ja ravi tulemuses sõltumatuse korral on [(10+2)!(4+9)!(10+4)!(2+9)!] / [25!10!2!4!9!] = 0,01059.

Alternatiivsed samade rea- ja veerusummadega sagedustabelid ning nende esinemise tõenäosused nullhüpoteesi eeldusel on järgmised (andmetele vastavast tabelist veel ebatõenäolisemad tabelid ja nende tõenäosused on esitatud paksus kirjas, andmetele vastavast tabelis nö samas suunas ekstreemsemate tabelite tõenäosused on täiendavalt allajoonitud):

11	**2**
0	**12**

p_{tabel} = **<u>0,0000175</u>**

10	**3**
1	**11**

p_{tabel} = **<u>0,00077</u>**

8	5
3	9

p_{tabel} = 0,06352

7	6
4	8

p_{tabel} = 0,19056

6	7
5	7

p_{tabel} = 0,26679

5	8
6	6

p_{tabel} = 0,30490

4	9
7	5

p_{tabel} = 0,12704

3	10
8	4

p_{tabel} = 0,03176

2	**11**
9	**3**

p_{tabel} = **0,00385**

1	**12**
10	**2**

p_{tabel} = **0,000192**

0	**13**
11	**1**

p_{tabel} = **0,00000269**

Olulisuse tõenäosuse väärtuseks tuleb sõltuvalt arvutamismeetodist

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.

Mõlemal viisil arvutatud p-väärtuse alusel võib lugeda tõestatuks alternatiivse hüpoteesi: ravi tulemus on soospetsiifiline (ehk seos ravi tulemuse ja koera soo vahel on statistiliselt oluline).

Valik *online*-kalkulaatoreid, mis pakuvad Fisheri täpse testi teostamise võimalust (enamasti vaid 2x2-tabeli tarvis):

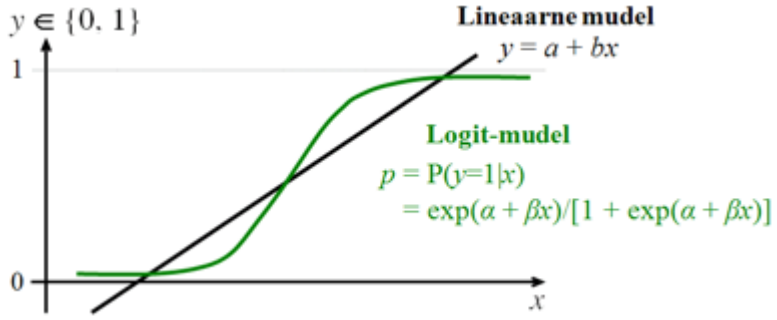
- http://www.physics.csbsju.edu/stats/fisher.form.html
- http://vassarstats.net/ -> Frequency data
- http://www.quantitativeskills.com/sisa/statistics/fishrhlp.htm, http://www.quantitativeskills.com/sisa/statistics/fiveby2.htm
- http://www.langsrud.com/fisher.htm
- http://statpages.org/ctab2x2.html (lisaks väga palju erinevaid karakteristikuid)

3. Binaarse tunnuse seos pideva arvtunnusega

3.1. Logistiline regressioon

Logistiline regressioon (ingl. ***logistic regression***) või üldisemalt **logistiline mudel** ehk **logit-mudel** prognoosib uuritava sündmuse toimumise tõenäosust ja selle muutumist sõltuvalt pideva argumenttunnuse väärtuse muutumisest.

Kuigi binaarse, väärtustega 0 ja 1, tunnuse modelleerimiseks võib kasutada ka lineaarset regressioonivõrrandit kujul $y = a + bx$ (regressioonikordajatele vähimruutude printsiipi rahuldavate hinnangute saamiseks ei ole muid piiranguid, kui et nii uuritav ehk sõltuv tunnus y kui ka argument- ehk sõltumatu tunnus x peavad olema arvulised), ei garanteeri lineaarne regressioonanalüüs saadavate prognooside jäämist mõistlikku vahemikku 0-st 1-ni. Seevastu logistilise regressiooni abil leitud tõenäosuste hinnangud jäävad alati 0 ja 1 vahele (vt järgmine joonis).



Logistilise regressiooni mudeli (logit-mudeli), mis binaarse tunnuse y suhtes on tegelikult mittelineaarne mudel (nagu näha ka eelnevalt jooniselt), võib esitada mitmel erineval viisil.

Üks variant on panna mudel kirja uuritava sündmuse toimumise tõenäosuse $p = P(y=1)$ tarvis kujul

$$p = P(y=1|x) = \exp(\alpha + \beta x)/[1 + \exp(\alpha + \beta x)] = 1 / [1 + \exp(-\alpha - \beta x)].$$

Alternatiivne esitus on logit-funktsioonina kujul

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

Logistilise regressiooni võrrandi parameetrite tõlgendamine lähtub tõsiasjast, et suhe $p/(1-p)$ kujutab enesest huvipakkuva sündmuse toimumise **šanssi** – näitab, kui mitu korda tõenäolisem on uuritava sündmuse toimumine võrreldes sündmuse mittetoimumisega.

- Suurus $\ln[p/(1-p)]$ on siis **logaritmiline šanss** (ingl. *log odds*).
- Juhul, kui uuritava sündmuse toimumine on samaväärne sündmuse mittetoimumisega, st et $p = 1 - p = 0{,}5$, siis võrdub šanss ühega: $p/(1-p) = 1$, ja logaritmiline šanss nulliga: $\ln[p/(1-p)] = 0$ (sest $\ln(1) = 0$). Logistilise regressiooni kontekstis vastab šansi ühega võrdumine olukorrale, kus $\alpha + \beta x = 0$.
- Logistilise regressioonivõrrandi kordaja β eksponent e^{β} näitab, kui mitu korda muutub sündmuse toimumise šanss argumendi muutumisel ühe ühiku võrra. Tuleneb see logistilise regressiooni võrrandist, mille kohaselt $p/(1-p) = e^{\alpha + \beta x}$ ja millest omakorda järeldub, et

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

  (x-i suurenemine ühe võrra muudab šanssi e^{β} korda).

- Seega kujutab kordaja β eksponent e^{β} enesest **šansside suhet**: $e^{\beta} = OR$.

  Näiteks kui $e^{\beta} = OR = 2$, siis kaasneb argumenttunnuse väärtuse suurenemisega ühe võrra sündmuse toimumise šansi kahekordne suurenemine (sündmuse toimumine muutub sündmuse mittetoimumisega võrreldes kaks korda tõenäolisemaks).

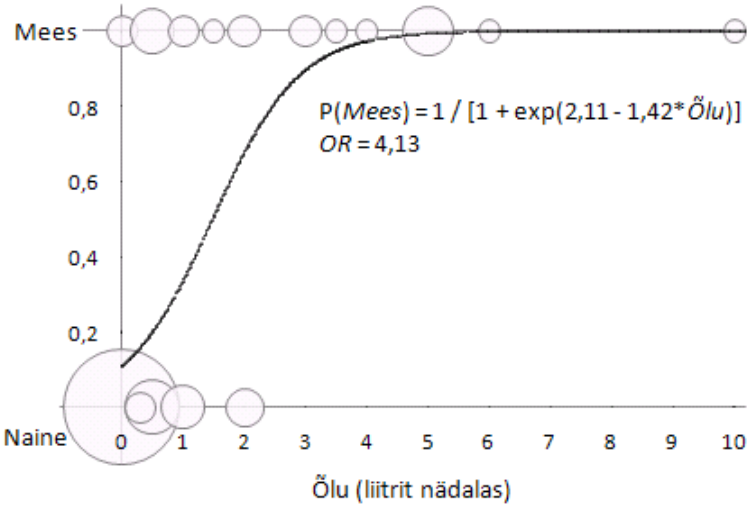
  Negatiivse regressioonikordaja β korral šansside suhe väheneb, sest $e^{\beta} = OR < 1$. St, et mida suurem on argumenttunnuse x väärtus, seda ebatõenäolisem on huvipakkuva sündmuse toimumine võrreldes sündmuse mittetoimumisega.

- Eelnevast tulenevalt on ka loomulik, et kui kordaja β on positiivne, siis argumentunnuse x väärtuse suurenedes suureneb ka uuritava sündmuse tõenäosus (tegu on positiivse seosega), kui aga kordaja β on negatiivne, siis argumentunnuse x väärtuse suurenedes uuritava sündmuse tõenäosus väheneb (tegu on negatiivse seosega).

Vaatame näitena andmestikku 66 tudengi vastustest nende soo ja nädalas keskmiselt joodava õllekoguse kohta (andmed *Exceli* tabelina võib alla laadida aadressilt http://www.emu.ee/~ktanel/bin_tunnuste_analyys/tudeng_ja_6lu.xlsx).

Rakendame logistilist regressioonanalüüsi prognoosimaks meheks olemise tõenäosust nädalas keskmiselt tarbitava õllekoguse alusel.

Andmeid ja analüüsi tulemusi on illustreeritud järgneval joonisel (ringid vastavad erinevatele õllekogustele ja ringi suurus tudengite arvule, pidev must joon on logistilise regressioonivõrrandi graafik ning y-telg vastab meheks olemise tõenäosusele).



Nagu jooniselt näha, on naistudengite hulgas enim õlut mittejoovaid tudengeid, meestudengite tarbitavad õllekogused on suuremad, mistap on loomulik ka logistilise regressioonivõrrandi graafiku suund – mida suurem on nädalas tarbitav õllekogus, seda suurema tõenäosusega on tegu meestudengiga.

Logistilise regressiooni võrrandi parameetrite hinnanguiks on: $\alpha = -2{,}11$ ja $\beta = 1{,}42$. Seega on logistiline regressioonivõrrand esitatav kas joonisel toodud kujul (prognoosimaks meheks olemise tõenäosust)

$$p = \mathrm{P}(Mees) = 1 / (1 + e^{2{,}11 - 1{,}42*Õlu})$$

või siis lineaarse võrrandina logaritmilise šansi tarvis kujul

$$\ln[p/(1-p)] = -2{,}11 + 1{,}42*\text{Õlu}.$$

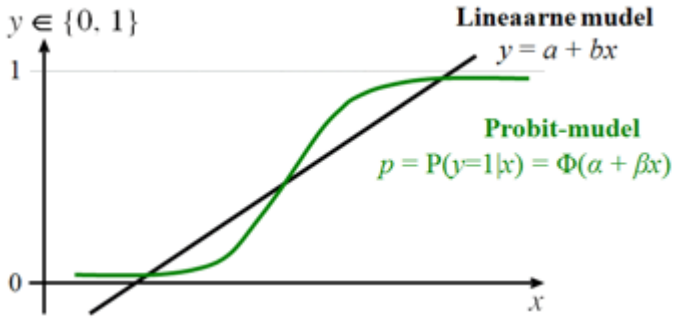
Šansside suhe avaldub kordaja $\beta = 1{,}42$ eksponentfunktsioonina: $OR = e^{1{,}42} = 4{,}13$. Seega suurendab tudengite puhul ühe lisaliitri õlle joomine nädalas meheks olemise šanssi 4,13 korda võrreldes naiseks olemise šansiga.

Logistiline regressioon *online*-kalkulaatori abil:

- http://statpages.org/logistic.html

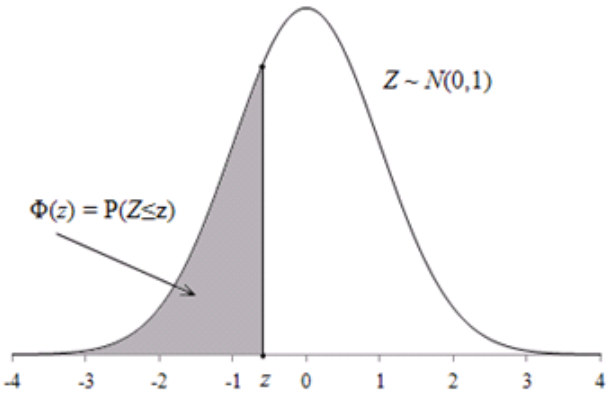
3.2. Probit-regressioon

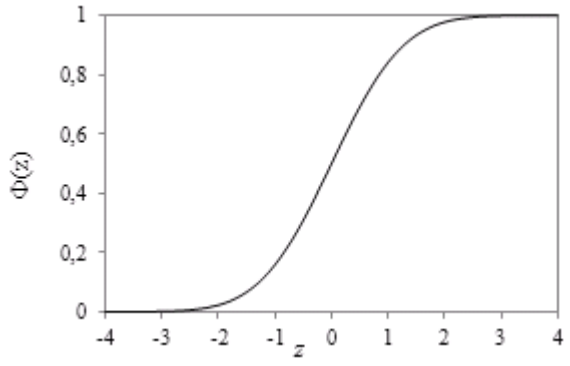
Analoogselt logistilisele regressioonile prognoosib ka **probit-regressioon** (ingl. ***probit-regression***) uuritava sündmuse toimumise tõenäosust ja selle muutumist sõltuvalt pideva argumenttunnuse väärtuse muutumisest ning saadavad prognoosid jäävad alati 0 ja 1 vahele (vt järgmine joonis).



Funktsioonina, mis projitseerib mistahes reaalarvulise väärtuse vahemikku (0,1), kasutab probit-regressioon standardse normaaljaotuse jaotusfunktsiooni, mida traditsiooniliselt tähistatakse tähega Φ.

Jaotusfunktsiooni Φ väärtus kohal z kujutab enesest tõenäosust, et standardse normaaljaotusega juhuslik suurus Z omandab väärtuse, mis on väiksem või võrdne z-st (vt alljärgnev vasakpoolne joonis): $\Phi(z) = P(Z \leq z)$, $Z \sim N(0,1)$. Alljärgneval parempoolsel joonisel on aga esitatud standardse normaaljaotuse jaotusfunktsiooni graafik.





Probit-regressiooni mudel (probit-mudel) kirja panduna sündmuse toimumise tõenäosuse $p = P(y=1)$ tarvis on järgmine:

$$p = P(y=1|x) = \Phi(\alpha + \beta x).$$

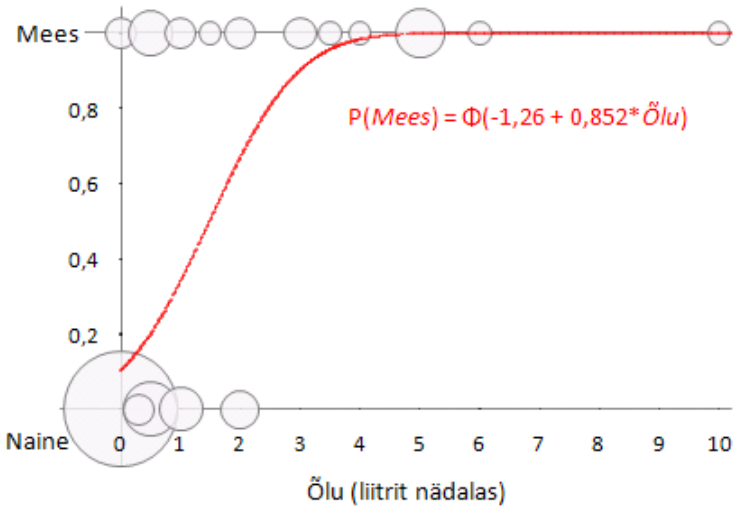
Alternatiivne esitus on lineaarse võrrandina standardse normaaljaotuse jaotusfunktsioon pöördfunktsiooni ehk probit-funktsiooni suhtes (siit ka probit-regressiooni nimetus):

$$\text{probit}(p) = \Phi^{-1}(p) = \alpha + \beta x.$$

Püüdes prognoosida tudengi meheksolemise tõenäosust nädalas keskmiselt tarbitava õllekoguse alusel probit-regressiooniga, on tulemuseks regressioonivõrrand

$$P(Mees) = \Phi(-1{,}26 + 0{,}854*\tilde{O}lu).$$

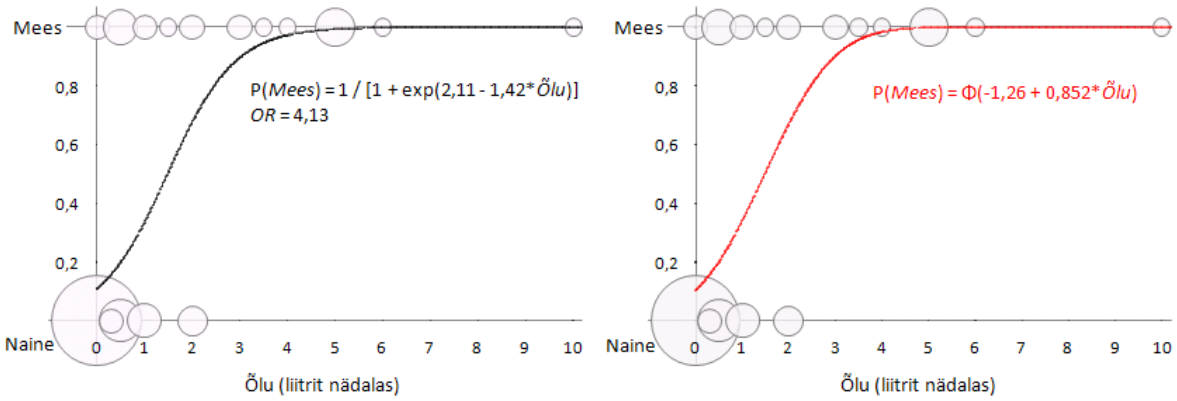
See, et tarbitava õllekoguse kordaja mudelis, 0,854, on positiivne arv, näitab, et mida enam tudeng õlut joob, seda suurema tõenäosusega ta mees on. Hinnatud probit-regressiooni võrrand koos algandmetega on esitatud järgneval joonisel.



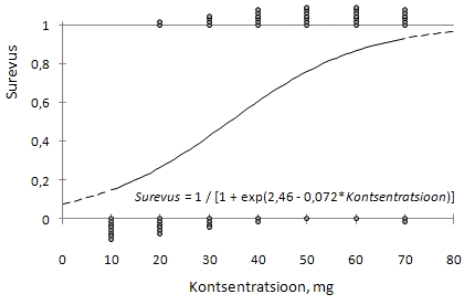
3.3. Logit- *vs* probit-regressioon ning tulemuste illustreerimine

Kas kasutada binaarse tunnuse väärtuste prognoosimiseks logistilist või probit-regressiooni sõltub suuresti uurimisvaldkonnast ja seal valitsevaist traditsioonidest, vahest ka kasutatavast tarkvarast. Tõenäosuste hinnangutel vahe peaaegu puudub (vrdl. ka tudengi meheksolemise tõenäosuse hinnanguid järgmistel joonistel).

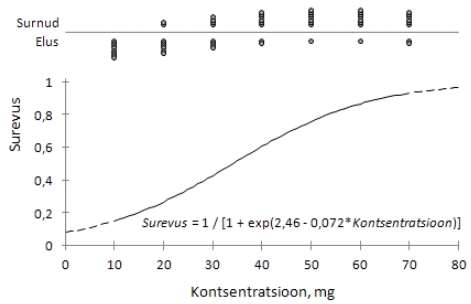
Logistilise regressiooni täiendav tulemus on šansside suhe.



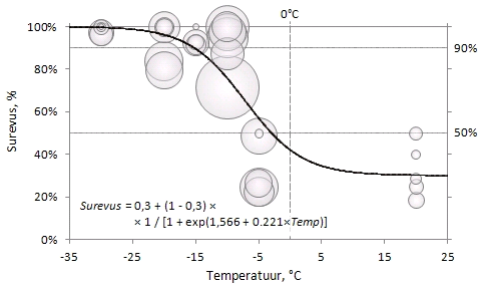
Nii logistilise kui ka probitregressiooni tulemuste illustreerimiseks sobib sarnaselt lineaarsele regressioonanalüüsile kasutada jooniseid, kus on näidatud nii algandmed kui prognoosivõrrandi graafik. Ainult erinevalt lineaarsest regressioonanalüüsist, kus uuritaval tunnusel on enamasti palju erinevaid väärtusi, mis joonisel ei kattu, tuleb binaarse uuritava tunnuse puhul näha pisut lisavaeva esitamaks algandmeid eristatavalt. Lahenduseks on kas kasutada tavalise punktdiagrammi asemel mulldiagrammi, kus mulli suurus näitab mingite ühesuguste väärtuste hulka, või esitada paralleelselt prognoosivõrrandi graafikuga teine mittekattuvate algandmete graafik. Kuidas seda Excelis teha, on õpetatud järgmistel internetilehtedel:



http://www.eau.ee/~ktanel/joonised_excelis/joonis7.php



http://www.eau.ee/~ktanel/joonised_excelis/joonis8.php



http://www.eau.ee/~ktanel/joonised_excelis/joonis9.php

3.4. 50% ja 90% vastuse määr (LD50, Ltemp90 jmt)

LD90, LTemp50 jt mitmel erialal leitavad suurused on logit- või probit-mudelil baseeruvad argumenttunnuse hinnangulised väärtused, mille korral uuritav sündmus leiab aset ette antud tõenäosusega.

Näiteks 90%-liselt surmav doos (ingl. ***90% lethal dose***, LD90) või 50%-liselt surmav temperatuur (ingl. ***50% lethal temperature***, LTemp50) on vastavalt sellised doosid või temperatuurid, mille korral hinnanguliselt 90% või 50% uuritud indiviide (kahjureid, istikuid vmt) sureb.

Nende suuruste leidmisel fikseeritakse huvipakkuv tõenäosus p ning avaldatakse sellele vastav argumenttunnuse väärtus x kas logistilise regressiooni või probit-regressiooni mudeli pöördfunktsioonina.

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

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

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

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

kus $\Phi^{-1}(p)$ on standardse normaaljaotuse pöördfunktsiooni väärtus kohal p. Excelis on $\Phi^{-1}(p)$ leitav funktsiooniga NORM.S.INV(), näiteks funktsioon =NORM.S.INV(0,9) annab tulemuseks 1,28.

> Püüdes tudengite soo prognoosimise näites leida seda nädalast õllekogust, mille puhul võib juba 90%-lise tõenäosusega väita, et tegu on meestudengiga, st hinnata suurust 'Mees90',
>
> tuleneb logistilisest mudelist P(*Mees*) = 1 / [1 + exp(2,11 – 1,42**Õlu*)], et
>
> $$\text{Mees90} = \{\text{LN}[0{,}9/(1-0{,}9)] + 2{,}11\}/1{,}42 = 3{,}03 \text{ l};$$
>
> probit-mudelist P(*Mees*) = Φ(–1,26 + 0,854**Õlu*) tuleneb aga, et
>
> $$\text{Mees90} = [\Phi^{-1}(0{,}9) + 1{,}26]/0{,}854 = 2{,}98 \text{ l}.$$
>
> Seega väitmaks 90%-lise kindlusega, et tudengi näol on tegu meesterahvaga, peab ta jooma kolm liitrit õlut nädalas.

3.5. Tundlikkus ja spetsiifilisus

Üks prognoosimiseks kasutatavate testide, mudelite, algoritmide või tehnoloogiate rakendatavuse peamisi kriteeriume on saadavate prognooside täpsus. Juhul, kui prognoositavaks on binaarse tunnuse väärtus (mingi sündmuse toimumine), on prognoosi korrektsuse hindamiseks vajalikud suurused koondatavad järgmisesse 2x2-tabelisse.

Prognoos	Tegelik olek		
	$Y = 0$ (negatiivne)	$Y = 1$ (positiivne)	Kokku
$Y = 0$ (negatiivne)	*TN*	*FN*	*TN*+*FN*
$Y = 1$ (positiivne)	*FP*	*TP*	*FP*+TP
Kokku	*TN*+*FP*	*FN*+*TP*	*TN*+*FN*+*FP*+*TP*

Selles tabelis

- *TN* märgib nende juhtude arvu, millal prognoosi kohaselt uuritavat sündmust ei oleks tohtinud toimuda ja tegelikult ka ei toimunud – so **tõeselt negatiivsete** juhtude arv (ingl. ***true negative***, *TN*);
- *FN* on ekslikult negatiivseks prognoositud juhtude arv – nö **valenegatiivsete** juhtude arv (ingl. ***false negative***, *FN*);
- *TP* on **tõeselt positiivsete** juhtude arv (ingl. ***true positive***, *TP*);
- *FP* on ekslikult ennustatud sündmuse toimumiste arv – nö **valepositiivsete** arv (ingl. ***false positive***, *FP*).

Tõeselt ja vääralt positiivsete ja negatiivsete juhtude arvu alusel leitakse erinevatel erialadel suur hulk erinevaid prognoosi korrektsuse (ehk testi/mudeli/algoritmi/tehnoloogia toimimise) hindamiseks kasutatavaid karakteristikuid (ingl. ***operating characteristics***), milledest enim kasutatud on tundlikkus ja spetsiifilisus
(vt ka http://en.wikipedia.org/wiki/Sensitivity_and_specificity).

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

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

Mõnes valdkonnas defineeritakse tundlikkusega sama valemi abil **tõeselt positiivsete määr** (ingl. ***true positive rate***, $TPR = TP / (TP+FN)$).

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

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

Karakteristikut üks miinus spetsiifilisus nimetatakse **valepositiivsete määraks** (ingl. ***false positive rate***, $FPR = 1 - [TN / (TN+FP)] = FP / (TN+FP)$).

> Kui rakendada tudengite soo ja õlle tarbimise näites lihtsaimat võimalikku tudengi soo määramise algoritmi – loeme tudengi meheks, kui ta joob õlut, ja naiseks, kui ta õlut ei joo –, saame prognoosi täpsuse hindamiseks järgmise tabeli:

Prognoos	Tudengi tegelik sugu		
	Naine	Mees	Kokku
Ei joo õlut -> naine	27	2	29
Joob õlut -> mees	15	20	35
Kokku	42	22	64

Võttes sündmuse toimumiseks (positiivseks sündmuseks) tudengi meheks osutumise ja sündmuse mittetoimumiseks (negatiivseks sündmuseks) naiseks osutumise, on tõeselt negatiivsete otsustuste arv $TN = 27$ (27 naise kohta otsustati õigesti, et nad ei ole mehed), valenegatiivsete otsuste arv $FN = 2$ (kahe õlut mittejoova meestudengi sugu prognoositi valesti), valepositiivsete otsuste arv $FP = 15$ (15 õlut joona naistudengi sugu prognoositi valesti) ja tõeselt positiivsete otsuste arv $TP = 20$ (20 meestudengit prognoositi õigesti meesteks).

Testi tundlikkus avaldub suhtena

$$\text{Tundlikkus} = 20 / (20+2) = 0{,}909$$

ja spetsiifilisus suhtena

$$\text{Spetsiifilisus} = 27 / (27+15) = 0{,}643.$$

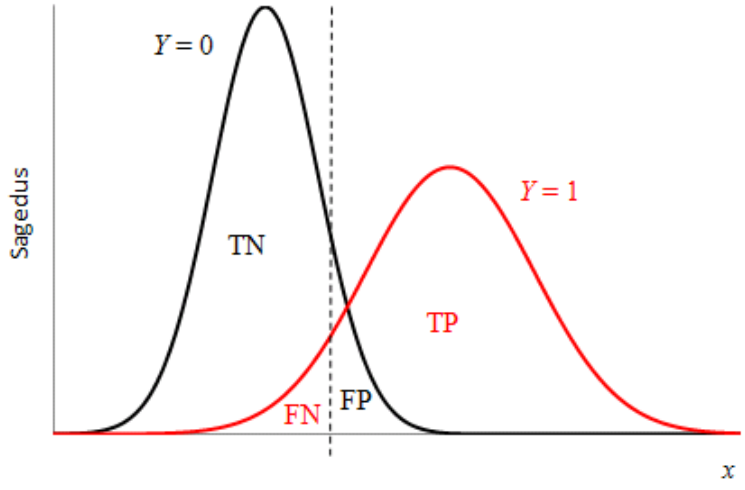
Seega õnnestub vaid õlle joomist ja mittejoomist kasutades õigesti ennustada 90,9% meestudengi ja 64,3% naistudengi sugu.

Tundlikkus ja spetsiifilisus (ja suur hulk muid karakteristikud) *online*-kalkulaatori abil:

- http://statpages.org/ctab2x2.html
- http://vassarstats.net/clin1.html

3.6. ROC-kõver

Juhul, kui uuritava sündmuse toimumise prognoosimine toimub mingi pideva argumendi/protsessi alusel, võib tundlikkuse ja spetsiifilisuse leida iga pideva argumendi väärtuse korral – võib ju klassifitseerimise aluseks olevaks **piirväärtuse**ks (ingl. ***threshold***, puntiirjoon järgneval joonsel) valida mistahes argumendi väärtuse ja lugeda iga kord kokku õigesti ja valesti prognoositud juhtude arvu ning arvutada ka tundlikkuse ja spetsiifilisuse.



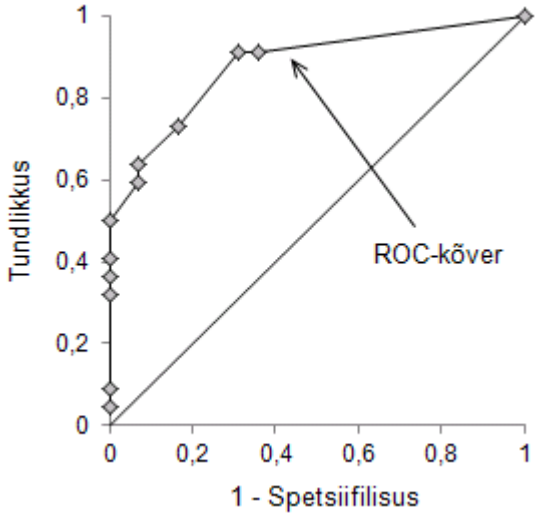
Näiteks prognooside tudengi sugu nädalas keskmiselt tarbitava õllekoguse alusel, on võimalikeks õllekoguse piirväärtusteks, mille kohalt jagada tudengid meesteks ja naisteks, kõik andmestikus esinevad õllekogused. Õigesti ja valesti klassifitseeritud mees- ja naistudengite arvud ning tundlikkused ja spetsiifilisused sõltuvalt piirväärtuseks valitud õllekogusest on esitatud järgnevas tabelis.

Õlu (l)*	TP	TN	FP	FN	Tundlikkus	Spetsiifilisus
0	22	0	42	0	1	0
0,3	20	27	15	2	0,909	0,643
0,5	20	29	13	2	0,909	0,690
1	16	35	7	6	0,727	0,833
1,5	14	39	3	8	0,636	0,929
2	13	39	3	9	0,591	0,929
3	11	42	0	11	0,5	1
3,5	9	42	0	13	0,409	1
4	8	42	0	14	0,364	1
5	7	42	0	15	0,318	1
6	2	42	0	20	0,091	1
10	1	42	0	21	0,045	1

* Nädalas keskmiselt tarbitav õllekogus, millest alates klassifitseeritakse tudeng meheks.

ROC-kõver (ingl. ***receiver operating characteristic curve***) või üldisemalt ROC-analüüs on erinevatele argumendi väärtustele vastavate tundlikkuse ja spetsiifilisuse paaride graafiline esitus hindamaks optimaalseimat piirväärtust ja prognoosi täpsust. Enamasti on sellel joonisel tundlikkuse väärtused y-teljel ning üks miinus spetsiifilisuse e valepositiivsete määra väärtused x-teljel. Joonise diagonaal vastab olukorrale, kus sõltumata argumendi väärtusest on tundlikkus ja spetsiifilisus võrdsed 0,5-ga, ehk mil uuritava sündmuse toimumine on juhuslik sõltumata argumendi väärtusest.

Tudengite näitele vastav ülaltoodud tabeli alusel joonistatud ROC-kõver on järgmine:



Kui iga tudengi puhul ennustada tema sugu näiteks kulli ja kirja viskamise teel, on ennustuse täpsus 0,5 (50%) ja sõltumata piirväärtuseks valitavast õllekogusest peaks ROC-kõvera punktid paiknema joonise diagonaalil.

Nimetus ROC-kõver (ingl. *receiver operating characteristic curve*) on pärit II Maailmasõja päevilt, mil Suurbritannia insenerid võtsid selle meetodi kasutusele hindamaks oma ja vaenlase lennukite jm sõjatehnika eristamise täpsust vastuvõetud radarisignaalide alusel.

Mõnikord tõlgendatakse lühendit ROC-kõver ka (testi/mudeli/algoritmi/tehnoloogia) **suhtelise toimimise karakteristikute kõverana** (ingl. ***relative operating characteristic curve***), sest tegu on kahe nn toimimise karakteristiku (tõeselt positiivsete ja valepositiivsete määra) võrdluskõveraga.

3.7. Optimaalne piirväärtus

Uuritava sündmuse toimumist ja mittetoimumist eristava optimaalseima piirväärtuse leidmiseks on mitmeid võimalusi:

- üks variant on võtta optimaalseimaks väärtuseks maksimaalsele tundlikkuse ja spetsiifilisuse summale vastav väärtus,
- teine variant on leida ROC-kõvera diagonaalist kaugeimal paiknev punkt,
- kolmas variant on kanda nii tundlikkuse kui ka spetsiifilisuse väärtused joonisele prognoositud tõenäosuse suhtes ning valida optimaalseimaks piirväärtuseks tundlikkuse ja spetsiifilisuse kõverate ristumispunktile vastav väärtus.

Pideval skaalal esitatud argumendi ja pideva ROC-kõvera puhul annavad kõik kolm optimaalseima piirväärtuse leidmise varianti sama tulemuse, diskreetse argumendi ja sellelest lähtuvalt murdjoonena esituva ROC-kõvera puhul võivad optimaalseimad piirväärtused pisut erineda.

Tudengite näites vastab maksimaalne tundlikkuse ja spetsiifilisuse summa 0,909+0,690=1,600 nädalas tarbitavale õllekogusele 0,5 liitrit (vt järgnev tabel) – seega on tudengite meheks ja naiseks jagamine täpseim, kui tõmmata nö piir 0,5 liitri juurde, kes tarbivad 0,5 liitrit või enam, on mehed, kes vähem, naised.

Prognoosides aga tudengi meheks osutumise tõenäosust logistilisest mudelist

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

(saadud hinnangud on esitatud järgneva tabeli teises veerus) ning kandes nii tundlikkuse kui ka spetsiifilisuse väärtused joonisele prognoositud tõenäosuste suhtes (vt tabeli järgne vasakpoolne joonis), osutub, et ligikaudne optimaalne logistilisest mudelist hinnatud tõenäosus, eristamaks õlle joomise alusel naisi ja mehi, on 0,3. Tõenäosusele 0,3 vastavad tundlikkus ja spetsiifilisus on mõlemad ligikaudu 0,8 ning vastav mees- ja naistudengite eristamiseks kasutatav õllekogus on arvutatav logistilise regressioonivõrrandi pöördfunktsioonina (vt punkt 3.4):

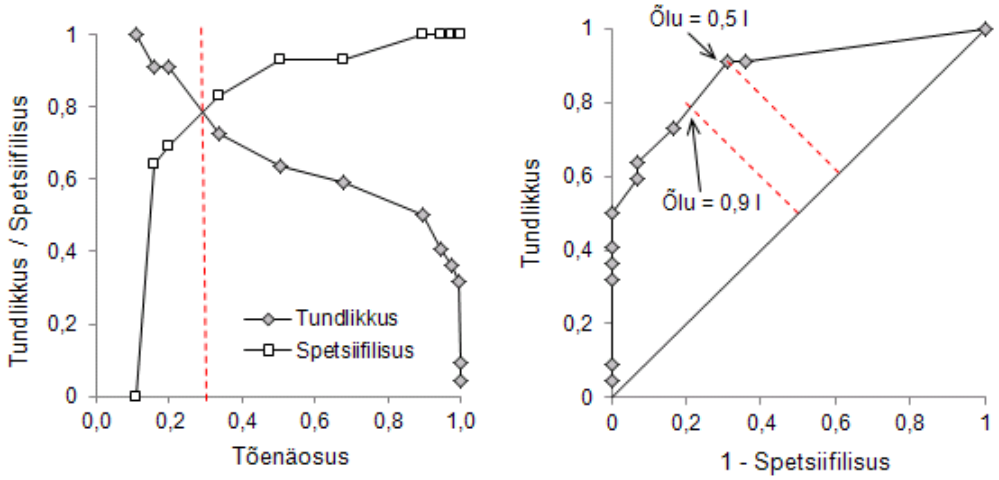
$$\tilde{O}lu_{30} = \{LN[0{,}3/(1–0{,}3)] + 2{,}11\}/1{,}42 = 0{,}9 \text{ l}.$$

Seega võib mees- ja naistudengeid täpseimalt eristavaks õllekoguseks võtta ka 0,9 liitrit.

Nädalas keskmiselt tarbitava õllekoguse ja tudengi soo sõltumatusele vastavast diagonaalist ROC-kõvera joonisel paiknevad mõlemad optimaalsetele piirväärtustele vastavad ROC-kõvera punktid võrdsel kaugusel.

Õlu (l)*	P(*Mees*)	TP	TN	FP	FN	Tundlikkus	Spetsiifilisus	Tundl.+spets.
0	0,108	22	0	42	0	1,000	0,000	1,000
0,3	0,157	20	27	15	2	0,909	0,643	1,552
0,5	0,198	20	29	13	2	0,909	0,690	1,600
1	0,334	16	35	7	6	0,727	0,833	1,561
1,5	0,505	14	39	3	8	0,636	0,929	1,565
2	0,675	13	39	3	9	0,591	0,929	1,519
3	0,896	11	42	0	11	0,500	1,000	1,500
3,5	0,946	9	42	0	13	0,409	1,000	1,409
4	0,973	8	42	0	14	0,364	1,000	1,364
5	0,993	7	42	0	15	0,318	1,000	1,318
6	0,998	2	42	0	20	0,091	1,000	1,091
10	1,000	1	42	0	21	0,045	1,000	1,045

* Nädalas keskmiselt tarbitav õllekogus, millest alates klassifitseeritakse tudeng meheks.



3.8. ROC-kõvera alune pindala

ROC-kõvera alune pindala (ingl. ***area under the curve***, *AUC*) on üks testi /mudeli headuse mõõte. 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 on järgmised:

- kui $AUC \geq 0{,}9$, siis on testi/mudeli täpsus suurepärane (ingl. *excellent*),
- $AUC \geq 0{,}8$ puhul hea (ingl. *good*),
- $AUC \geq 0{,}7$ puhul rahuldav (ingl. *fair*),
- $AUC \geq 0{,}6$ puhul kasin (ingl. *poor*) ja alla selle ei ole erilist mõtet ennustuse/prognoosi täpsusest rääkida.

Kui statistiliste analüüside teostamisel kasutatav tarkvara väljastab lisaks ROC-kõvera aluse pinna suurusele ka selle 95%-lise usaldusaintervalli 95%CI_{AUC}, on viimane kasutatav otsustamaks mudeli statistilise olulisuse üle:

- kui argumenttunnuse ja prognoositava sündmuse sõltumatuse juhule (nullhüpoteesile) vastav arv 0,5 jääb usaldusintervalli sisse, ei ole mudel statistiliselt oluline ($0{,}5 \in 95\%CI_{AUC}$ -> $p>0{,}05$),
- kui aga usaldusintervall arvu 0,5 ei sisalda, on prognoosimudel statistiliselt oluline ($0{,}5 \notin 95\%CI_{AUC}$ -> $p<0{,}05$).

Tudengite näites tuleb ROC-kõvera aluse pinna suuruseks AUC = 0,878, mistap võib tudengite soo prognoosimise täpsust nädalase keskmise õlletarbimise alusel lugeda heaks. Et ROC-kõvera aluse pindala 95%-line usaldusintervall $95\%CI_{AUC}$ = (0,785; 0,970) ei sisalda arvu 0,5, võib seose tudengite õlletarbimise ja soo vahel lugeda ka statistiliselt oluliseks ($p<0{,}05$).

3.9. Diskreetse argumendiga logistiline mudel

Juhul, kui binaarse tunnuse väärtusi soovitakse prognoosida mitte pideva vaid hoopis diskreetse argumendi väärtuste alusel, tuleb ka mudel vastavalt esitada. Seejuures on argumentide suhtes lineaarses mudelis prognoositavaks ikkagi tõenäosuse logit- või probit-funktsiooni väärtus, mudeli paremal poolel sisaldub aga mudeli vabaliige pluss diskreetsete faktorite mõjud, mis on vaja andmetest hinnata.

Peatükkides 2.1-2.7 näitena vaadatud koerte andmestiku puhul on mudel, hindamaks soo mõju ravi tulemusele, logistilise mudelina esitatav kujul

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

kus p_{ij} on i. sugu j. koera terveks mittesaamise tõenäosus (i=1,2; j=1,…, 25), μ on mudeli vabaliige, S_i on i. soo mõju ja e_{ij} on mudeli viga.

Kui mudeli vabaliige ja soo mõjud on andmeist hinnatud, avaldub terveks mittesaamise tõenäosus kujul

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

Võttes isaseks olemise efekti parameetrite ühese hindamise tarvis võrdseks nulliga (isane olemine on sellisel juhul nö baastase ja mudeliga hinnatakse emaseks olemise efekti selle suhtes), tuleb mudeli vabaliikme hinnanguks –1,61 ja emaseks olemise mõju hinnanguks 2,42.

Terveks mittesaamise tõenäosused sõltuvalt koera soost avalduvad kujul

$$\text{P(haige|sugu=E)} = \exp(-1{,}61 + 2{,}42)/[1 + \exp(-1{,}61 + 2{,}42)] = 0{,}69$$

ja

$$\text{P(haige|sugu=I)} = \exp(-1{,}61 + 0)/[1 + \exp(-1{,}61 + 0)] = 0{,}17$$

ning võrduvad arvuliselt terveks mittesaanud koerte suhteliste sagedustega eraldi emaste ja isaste koerte hulgas (vt punkti 2.2 näite esimene tabel).

Šansside suhe, mis avaldub sarnaselt logistilisele regressioonile mudeli parameetri eksponent-funktsioonina:

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

näitab, et šanss mitte terveks saada on emastel koertel 11,25 korda kõrgem, kui isastel koertel. Ka selle, antud juhul logistilisest mudelist leitud parameetri väärtus, on identne peatükis 2.5 kahemõõtmelisest sagedustabelist leitud šansside suhte väärtusega.

4. Enesekontroll

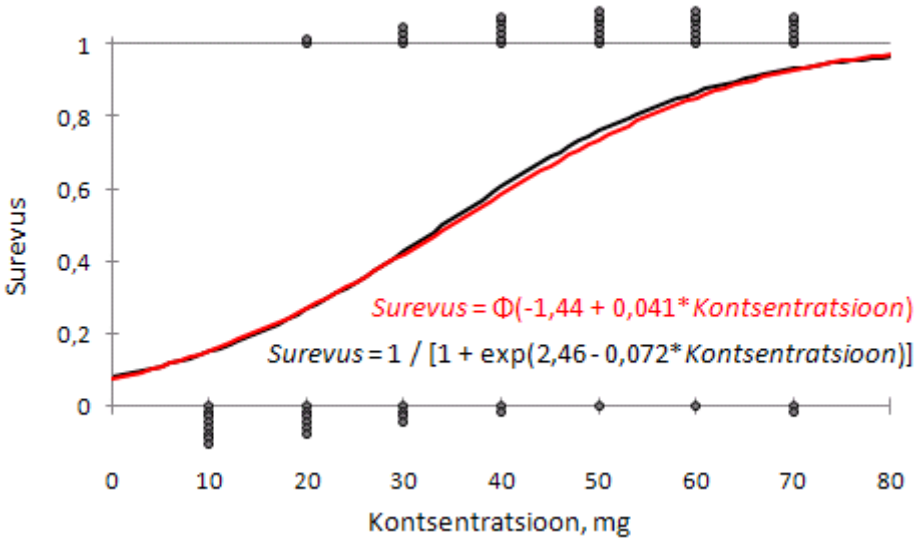
4.1. Küsimused ja ülesanded

1. 16-l mullikal on fikseeritud nende tiinestuvus esimesest seemendusest ning lisaks on nad genotüpiseeritud teatud lookuste osas. Uurimaks seost mullikate tiinestumise ja ühe konkreetse genotüübi vahel, konstrueeriti järgmine sagedustabel.

Genotüüp	Ei tiinestunud	Tiinestus	Kokku
AA	2	5	7
AG	8	1	9
Kokku	10	6	16

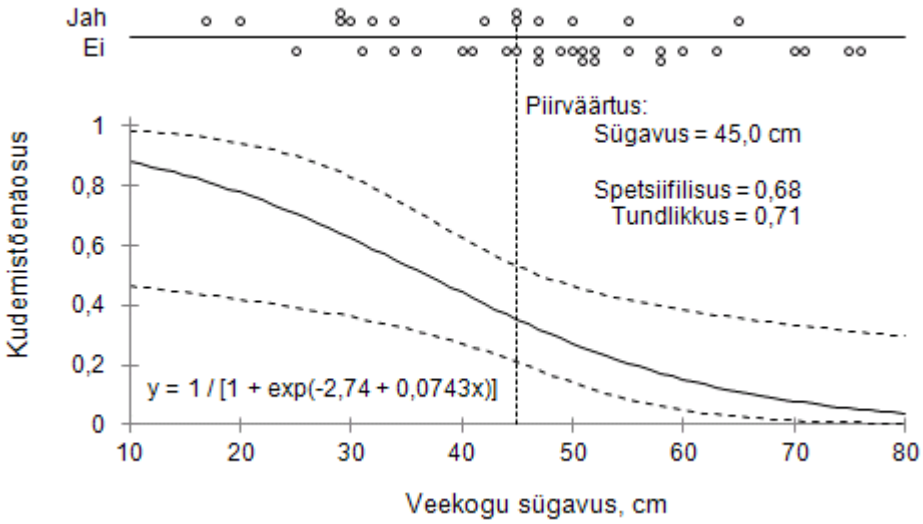
   a) Arvutage toodud tabeli alusel tiinestuvuskordajad („risk tiinestuda“), riskisuhted ja šansside suhted ning sõnastage nende alusel mõned laused.
   b) Arvutage šansside suhte 95%-usaldusintervall ja sõnastage järeldus seose statistilise olulisuse kohta.
   c) Teostage χ^2-test ja sõnastage järeldus seose statistilise olulisuse kohta.
   d) Teostage Fisheri täpne test (näiteks mõne *online*-kalkulaatori abil) ja sõnastage järeldus seose statistilise olulisuse kohta.

2. Uuriti kahjurite surevust sõltuvalt kahjuritõrjevahendi kontsentratsioonist. Uuringu andmed ning logistilise ja probit-regressiooni tulemused on esitatud järgmisel joonisel.



   a) Leidke 90%-lise tõenäosusega surmav kontsentratsioon (LC90, *90% lethal concentration*) nii logistilise kui ka probit-regressiooni alusel.
   b) Kui suur on šansside suhe ja mida see näitab?

3. Uuriti, kui sügavates veekogudes armastab kudeda mudakonn. Uuringu andmed ning logistilise regressioonanalüüsi tulemused on esitatud järgmisel joonisel.



a) Millisele hinnangulisele kudemistõenäosusele vastab kudemiseks sobivaid ja mittesobivaid veekogusid optimaalseimalt eristav sügavuse piirväärtus 45 cm?
b) Mitu korda väheneb kudemise šanss veekogu sügavuse suurenemisel 1 cm võrra?
c) Mida tähendab, et antud mudeli tundlikkus on 0,71 ja spetsiifilisus 0,68?
d) Kui teada on, et ROC-kõvera alune pindala on AUC = 0,74 ja 95%CI_{AUC} = (0,58;0,91), siis kui hästi võimaldab veekogu sügavus prognoosida selle sobivust mudakonnale kudemiseks ja kas vastav seos on statistiliselt oluline?

4.2. Vastused ja lahendused

1. 16-l mullikal on fikseeritud nende tiinestuvus esimesest seemendusest ning lisaks on nad genotüpiseeritud teatud lookuste osas. Uurimaks seost mullikate tiinestumise ja ühe konkreetse genotüübi vahel, konstrueeriti järgmine sagedustabel.

Genotüüp	Ei tiinestunud	Tiinestus	Kokku
AA	2	5	7
AG	8	1	9
Kokku	10	6	16

a) Arvutage toodud tabeli alusel tiinestuvuskordajad („risk tiinestuda“), riskisuhted ja šansside suhted ning sõnastage nende alusel mõned laused.

Vastus: genotüübiga AA lehmade tiinestuvuskordaja on 0,714 ja genotüübiga AG lehmade tiinestuvuskordaja on 0,111, st et genotüübiga AA lehmadest tiinestus 71,4% ja genotüübiga AG lehmadest 11,1%. Riskisuhe $RR = 6{,}43$ tähendab, et genotüübiga AA lehmadel on „risk“ tiinestuda 4,43 korda suurem, võrreldes genotüübiga AG lehmadega. Šansside suhte $OR = 20{,}0$ alusel võib väita, et genotüübiga AA lehmadel on šanss tiinestuda 20 korda suurem, võrreldes genotüübiga AG lehmadega.

b) Arvutage šansside suhte 95%-usaldusintervall ja sõnastage järeldus seose statistilise olulisuse kohta.

Vastus: Šansside suhte 95%-usaldusintervall $95\%CI_{OR} = (1{,}42;\ 282{,}46)$ näitab, et genotüübiga AA lehmade tiinestumine erineb statistiliselt oluliselt genotüübiga AG lehmade tiinestumisest (sest 95%-usaldusintervall ei sisalda arvu 1).

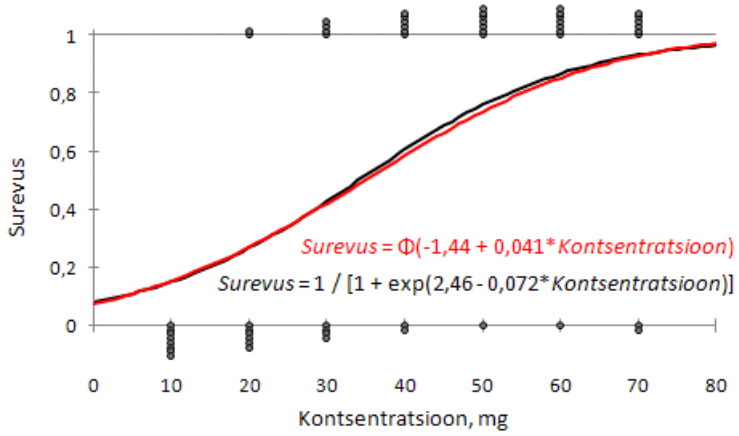
c) Teostage χ^2-test ja sõnastage järeldus seose statistilise olulisuse kohta.

Vastus: $p = 0{,}0134 < 0{,}05$, seega on seos uuritava genotüübi ja lehmade tiinestumise vahel statistiliselt oluline.

d) Teostage Fisheri täpne test (näiteks mõne *online*-kalkulaatori abil) ja sõnastage järeldus seose statistilise olulisuse kohta.

Vastus: ühepoolsele testile vastav olulisuse tõenäosus $p = 0{,}024$, kahepoolsele testile vastav olulisuse tõenäosus sõltuvalt arvtamise metoodikast $p = 0{,}035$ või $p = 0{,}049$. Seega on seos uuritava genotüübi ja lehmade tiinestumise vahel statistiliselt oluline.

2. Uuriti kahjurite surevust sõltuvalt kahjuritõrjevahendi kontsentratsioonist. Uuringu andmed ning logistilise ja probit-regressiooni tulemused on esitatud järgmisel joonisel.



a) Leidke 90%-lise tõenäosusega surmav kontsentratsioon (LC90, *90% lethal concentration*) nii logistilise kui ka probit-regressiooni alusel.

Vastus:
logistilise mudeli kohaselt *Surevus* = 1 / [1 + exp(2,46 – 0,072**Kontsentratsioon*)], mistap 90%-liselt surmav kontsentratsioon

$$LC90 = \{LN[0,9/(1–0,9)] + 2,46\}/0,072 = 64,68 \text{ mg};$$

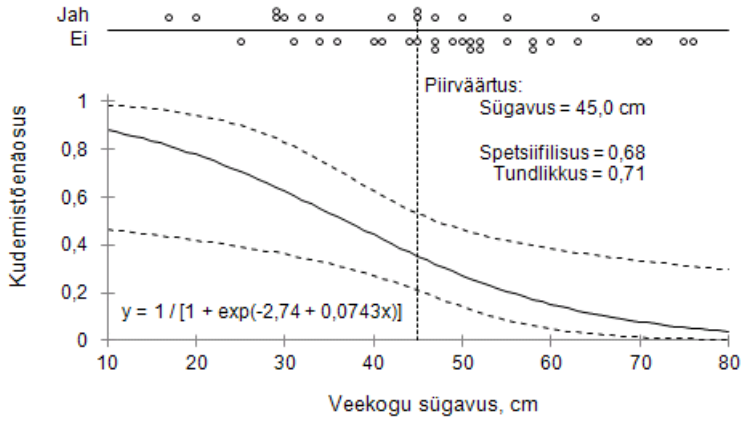
probit-mudeli kohaselt *Surevus* = Φ(–1,44 + 0,041**Kontsentratsioon*), mistap 90%-liselt surmav kontsentratsioon

$$LC90 = [\Phi–1(0,9) + 1,44]/0,041 = (1,28+1,44)/0,041 = 66,38 \text{ mg}.$$

b) Kui suur on šansside suhe ja mida see näitab?

Vastus: $OR = e^{0,072} = 1,075$, seega suureneb kahjuritõrjevahendi kontsentratsiooni suurenemisel 1 mg võrra šanss, et kahjur sureb, 1,075 korda.

3. Uuriti, kui sügavates veekogudes armastab kudeda mudakonn. Uuringu andmed ning logistilise regressioonanalüüsi tulemused on esitatud järgmisel joonisel.



a) Millisele hinnangulisele kudemistõenäosusele vastab kudemiseks sobivaid ja mittesobivaid veekogusid optimaalseimalt eristav sügavuse piirväärtus 45 cm?

Vastus: kudemiseks sobivaks võib veekogu lugeda juba siis, kui logistilisest mudelist hinnatud kudemiseks sobivuse tõenäosus on üle 0,35.

b) Mitu korda väheneb kudemise šanss veekogu sügavuse suurenemisel 1 cm võrra?

Vastus: $OR = e^{-0,0743} = 0,928$, seega väheneb veekogu sügavuse suurenemisega 1 cm võrra šanss sobida mudakonnale kudemiseks 0,928 korda.

c) Mida tähendab, et antud mudeli tundlikkus on 0,71 ja spetsiifilisus 0,68?

Vastus: 45 cm-st madalamad on 71% kudeveekogusid ja 45 cm-st sügavamad on 68% kudemiseks mittevalitud veekogusid.

d) Kui teada on, et ROC-kõvera alune pindala on $AUC = 0,74$ ja $95\%CI_{AUC} = (0,58;0,91)$, siis kui hästi võimaldab veekogu sügavus prognoosida selle sobivust mudakonnale kudemiseks ja kas vastav seos on statistiliselt oluline?

Vastus: otsustades veekogu sobivuse üle mudakonnale kudemiseks üksnes veekogu sügavuse järgi, on prognoosi täpsus rahuldav ($AUC = 0,74$), samas on seos veekogu sügavuse ja kudemistõenäosuse vahel on statistiliselt oluline ($p<0,05$, usaldusintervall ei sisalda arvu 0,5).