Praktikum 6

Salvestage kursuse kodulehelt omale arvutisse andmestik *lehmageen.xls*.

Praktikum püüab kirjeldada mõningaid võimalusi tunnuste vaheliste seoste uurimiseks.

Kommentaarid andmestiku kohta

Konkreetselt antud andmestiku näol on tegu mitmete erinevate statistiliste analüüsimeetodite illustreerimiseks genereeritud kunstliku andmestikuga, mille aluseks on mitmed erinevad aegade jooksul analüüsitud reaalsed andmestikud.

Samas võib enesele ette kujutada ka situatsiooni, et tegu on mingi väikesemahulise (ja väga kalli!) uuringu tulemustega.

Uuritud on 16 lehma, kellel kõigil on teada 1. laktatsiooni piimatoodang (kg), somaatiliste rakkude arv (tuh./ml), tiinestumiseks kulunud seemenduste arv ja genotüübid 5 SNP kohta.

Ülesanne 1

Uurige andmestikus sisalduvate arvutunnuste vahelisi seoseid.

- Arvutage kõigi arvtunnuste vahelised lineaarsed korrelatsioonikordajad ja testige nende statistilist olulisust.
  Vähemalt ühe seose kohta teostage sama analüüs ka internetilehe http://faculty.vassar.edu/lowry/VassarStats.html abil.
- Arvutage kõigi arvtunnuste vahelised astakkorrelatsioonikordajad.
  - Selleks leidke esmalt kõigile tunnustele nende väärtuste astakud, kasutades nii *Exceli* funktsiooni `RANK` kui ka sellel baseeruvat pikemat, aga tulemuseks korrektsed astakud andvat valemit. Püüdke tunnuse *SMARV* (seemenduste arv) astakute põhjal aru saada, mis seal vahet on.
  - Arvutage astakkorrelatsioonikordajad korrektsete astakute alusel nii valemi (*) alusel (vt teooria osa järgm. lehel) kui ka astakute vahelise lineaarse korrelatsioonikordajana. Miks tulevad piimatoodangu ja somaatiliste rakkude arvu astakkorrelatsioonikordajad seemenduste arvuga erinevate arvutusmeetodite korral erinevad?
  - Arvutage seemenduste arvu ja piimatoodangu vaheline astakkorrelatsioonikordaja ka internetilehe http://faculty.vassar.edu/lowry/VassarStats.html abil. Kirjutage tulemustest välja ka astakkorrelatsioonikordajale vastav p-väärtus ja sõnastage lõppjäreldus.
  - Miks on piimatoodangu ja somaatiliste rakkude (või piimatoodangu ja seemenduste arvu) vaheline astakkorrelatsioonikordaja poolt kirjeldatud seos märgatavalt tugevam kui lineaarne seos (arusaamiseks võite uurida nii SRA histogrammi kui ka piimatoodangu ja SRA vahelist hajuvusdiagrammi).
- Arvutage seemenduste arvu ja SRA vahelised lineaarsed korrelatsioonikordajad markeri *GT4* mõlema genotüübi tarvis. Kas seos on erinevate genotüüpide korral erinev? Testige seose (lineaarsete korrelatsioonikordajate) erinevuse statistilist olulisust internetilehe http://faculty.vassar.edu/lowry/VassarStats.html abil.

Teooria ülesande 1 juurde

Kui uuritavate arvtunnuste näol on tegu pidevate ja sümmeetrilise jaotusega tunnustega, on loomulikem seosekordaja **lineaarne** (e **Pearsoni**) **korrelatsioonikordaja**, mis mõõdab kahe arvtunnuse vahelise lineaarse seose tugevust ja suunda.

Kui aga tunnused on ebasümmeetrilise jaotusega (või ilmneb hajuvusdiagrammilt küll kasvav või kahanev, aga mittelineaarne seos), on mõttekas kasutada seose kirjeldamiseks lineaarse korrelatsioonikordaja asemel **astak-** (e **Spearmani**) **korrelatsioonikordajat** (*Spearman rank correlation coefficient*). Viimane on juhul, kui kõik väärtused on erinevad, arvutatav valemist

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(x_{(i)} - y_{(i)})^2}{n(n^2-1)}, \qquad (*)$$

kus kus $x_{(i)}$ on tunnuse X väärtuse x_i astak (järjekorranumber väärtuste kasvavalt järjestatud reas) ja $y_{(i)}$ on tunnuse Y väärtuse y_i astak. Võrdsete väärtuste korral annab toodud valem vale tulemuse, mistõttu on õige leida astakkorrelatsioonikordaja, kui tavaline lineaarne korrelatsioonikordaja uuritavate tunnuste väärtuste astakute vahel.
Astakkorrelatsioonikordaja mõõdab kahe arvtunnuse vahelise monotoonse seose tugevust ja suunda.

Ülesande 1 tööjuhend

1. Arvutage kõigi arvtunnuste vahelised lineaarsed korrelatsioonikordajad ja testige nende statistilist olulisust.

Üksikute lineaarsete korrelatsioonikordajate arvutamiseks on mugav kasutada funktsiooni `CORREL`.

Korrelatsioonikordaja statistilise olulisuse testimiseks (*p*-väärtuse arvutamiseks) *Excelis* eraldi vahendit pole, küll on *p*-väärtus arvutatav nö klassikalisel viisil ja käsitsi,

– leides esmalt teststatistiku väärtuse valemist

$$t = r\sqrt{n-2}\Big/\sqrt{1-r^2}\,,$$

kus r on korrelatsioonikordaja väärtus ja n korrelatsioonikordaja arvutamiseks kasutatud väärtuste paaride arv,

– ja seejärel, teades, et leitud teststatistik peaks nullhüpoteesi (seost ei ole) kehtides olema ligikaudu t-jaotusega parameetriga n–2, arvutades antud t-jaotuse korral teststatistiku väärtusega võrdsete või sellest ekstreemsemate väärtuste esinemise tõenäosuse. Viimane ongi otsitav *p*-väärtus ja *Excelis* on see leitav funktsiooniga `TDIST`, mille esimeseks argumendiks on teststatistiku t väärtus, teiseks argumendiks t-jaotuse parameeter n–2 ja kolmandaks argumendiks arv 2, mis näitab, et soovime testida kahepoolset hüpoteesi (kui ikka veel ei tea, mida see tähendab, siis küsi julgelt!).

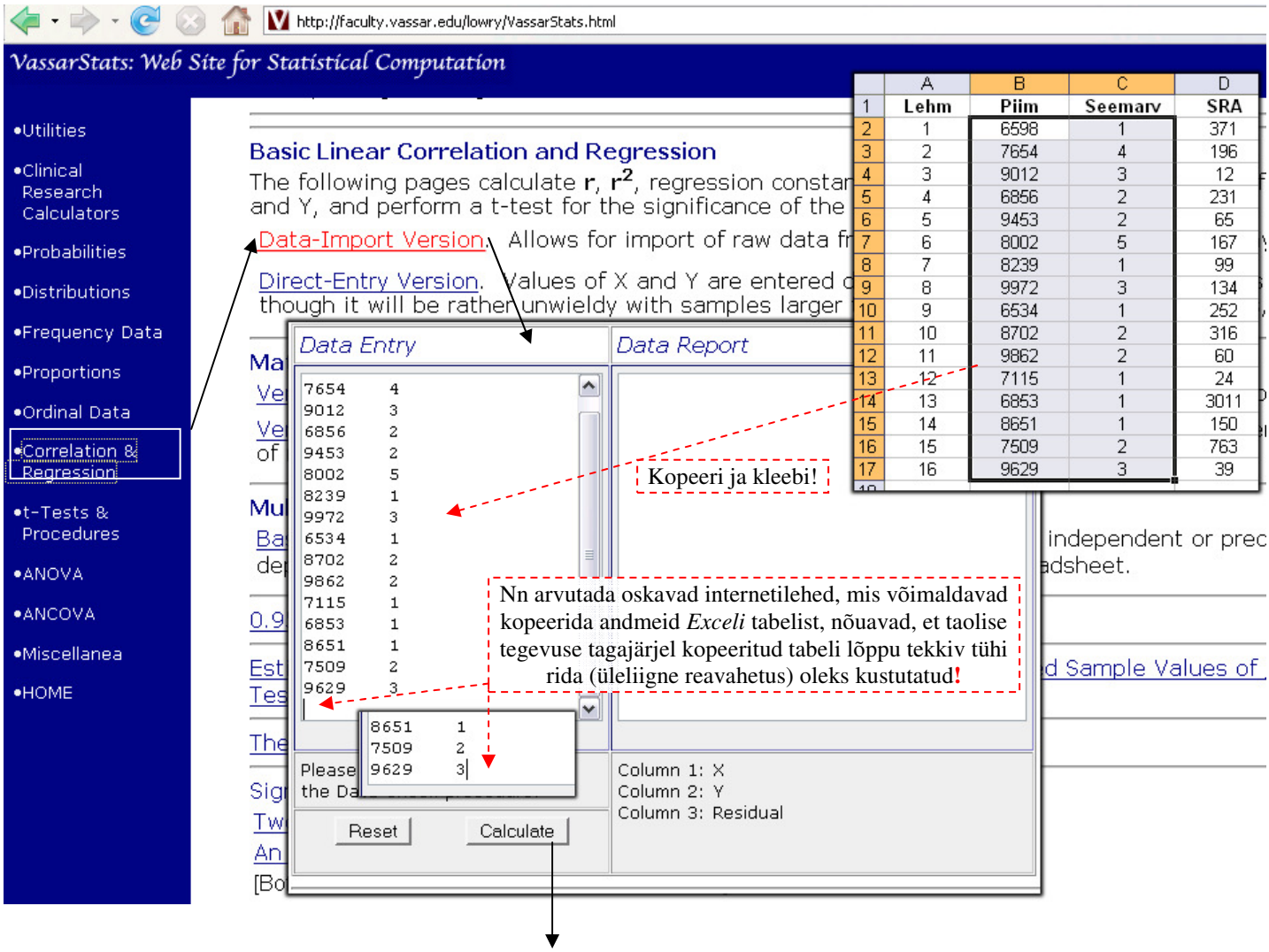
	A	B	C	D	E	F	G
1	**Lehm**	**Piim**	**Seemarv**	**SRA**	**GT1**	**GT2**	**GT3**
2	1	6598	1	371	GC	CT	CC
3	2	7654	4	196	GG	CC	AC
4	3	9012	3	12	CC	CT	CC
5	4	6856	2	231	GG	CC	CC
6	5	9453	2	65	CC	CT	AA
7	6	8002	5	167	GC	CC	CC
8	7	8239	1	99	CC	CC	AC
9	8	9972	3	134	GC	CC	CC
10	9	6534	1	252	CC	CT	CC
11	10	8702	2	316	GG	CT	AC
12	11	9862	2	60	GG	CC	AA
13	12	7115	1	24	CC	CT	CC
14	13	6853	1	3011	CC	CC	CC
15	14	8651	1	150	GC		
16	15	7509	2	763	GC		
17	16	9629	3	39	GG		
18							
19							
20		**Piim<->SA**	**Piim<->SRA**	**SA<->SRA**			
21	***r***	0,3512	-0,3933	=CORREL(C2:C17;D2:D17)			
22	***n***	16	16	16			
23	***t***	1,4036	1,6007	=ABS(D21*SQRT(D22-2)/SQRT(1-D21*D21))			
24	***p***	0,1822	0,1318	=TDIST(D23;D22-2;2)			

Absoluutväärtust arvutavat funktsiooni `ABS` on vaja rakendada garanteerimaks teststatistikule positiivset väärtust (*Excel* ei oska muidu p-väärtust leida).

t n–2 $\left| r\sqrt{n-2}\Big/\sqrt{1-r^2} \right|$

2. Vähemalt ühe seose kohta teostage sama analüüs ka internetilehe http://faculty.vassar.edu/lowry/VassarStats.html abil.

Uurime näiteks piimatoodangu ja seemenduste arvu vahelist seost.



Tulemuseks on hulk karakteristikuid nii lineaarse korrelatsioonikordaja kui ka lineaarse regressioon-analüüsi kohta (vt järgmine lk).

Otsige väljundist üles eelnevalt *Excelis* arvutatud korrelatsioonikordaja, selle baasil leitud test-statistiku väärtus ja seose statistilist olulisust väljendav *p*-väärtus. Kas saate aru ka muudest väljastatud numbritest? Kui ei saa, küsige!

Data Summary

ΣX = 130641 ΣX^2 = 1088257679

ΣY = 34 ΣY^2 = 94

ΣXY = 285219

	X	Y
N	16	
Mean	8165.0625	2.125
Variance	1437716.5958	1.45
Std.Dev.	1199.0482	1.2042
Std.Err.	299.7621	0.301

r	r^2	Slope	Y Intercept	Std. Err. of Estimate
0.3512	0.1234	0.000353	-0.755058	1.167

t	df	P		
1.4	14		one-tailed	0.0916415
			two-tailed	0.183283

0.95 and 0.99 Confidence Intervals for rho

	Lower Limit	Upper Limit
0.95	-0.174	0.721
0.99	-0.334	0.793

0.95 and 0.99 Confidence Intervals for the Slope of the Regression

	Lower Limit	Upper Limit
0.95	-0.0002	0.0009
0.99	-0.0004	0.0011

3. Arvutage kõigi arvtunnuste vahelised astakkorrelatsioonikordajad.

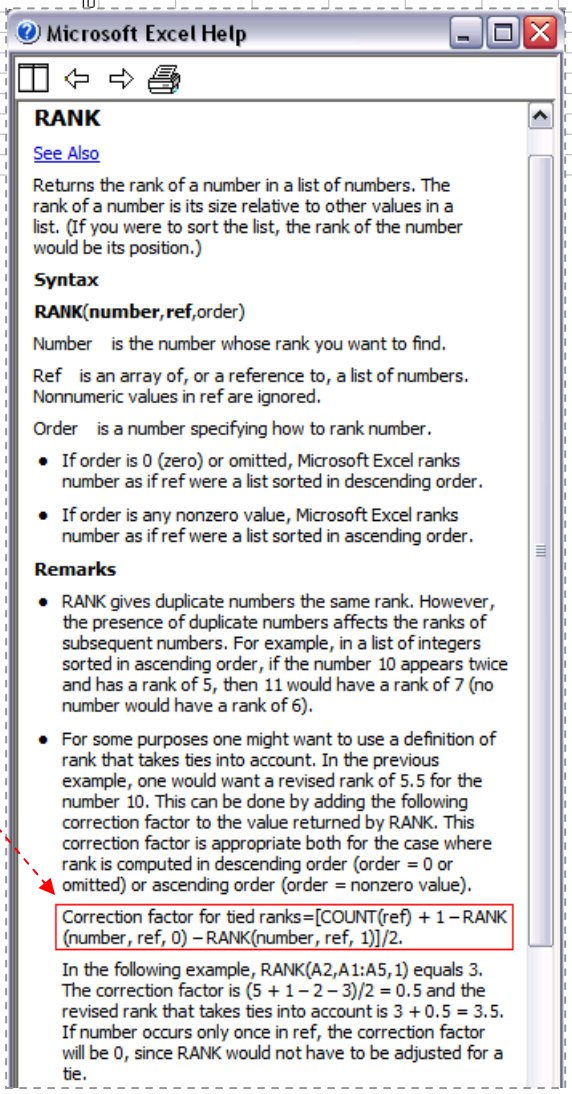
Selleks leidke esmalt kõigile tunnustele nende väärtuste astakud, kasutades nii *Exceli* funktsiooni `RANK` kui ka sellel baseeruvat pikemat, aga tulemuseks korrektsed astakud andvat valemit.

Püüdke tunnuse *SMARV* (seemenduste arv) astakute põhjal aru saada, mis seal vahet on.

	A	B	C	D	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Lehm	Piim	Seemarv	SRA		r(Piim)	r(SA)	r(SRA)	correct_r(Piim)	correct_r(SA)	correct_r(RSA)							
2	1	6598	1	371		15	11	=RANK(D2;D$2:D$17)		13,5	=RANK(D2;D$2:D$17)+(COUNT(D$2:D$17)+1-RANK(D2;D$2:D$17;0)-RANK(D2;D$2:D$17;1))/2							
3	2	7654	4	196		10	2	7	10	2	7							
4	3	9012	3	12		5	3	16	5	4	16							
5	4	6856	2	231		13	6	6	13	8	6							
6	5	9453	2	65														
7	6	8002	5	167														
8	7	8239	1	99		8	11	11	8	13,5	11							
9	8	9972	3	134		1	3	10	1	4	10							
10	9	6534	1	252		16	11	5	16	13,5								
11	10	8702	2	316		6	6	4	6	8								
12	11	9862	2	60		2	6	13	2	8								
13	12	7115	1	24		12	11	15	12	13,5								
14	13	6853	1	3011		14	11	1	14	13,5								
15	14	8651	1	150		7	11	9	7	13,5								
16	15	7509	2	763		11	6	2	11	8								
17	16	9629	3	39		3	3	14	3	4								

=RANK(D2;D$2:D$17)+(COUNT(D$2:D$17)+1-RANK(D2;D$2:D$17;0)-RANK(D2;D$2:D$17;1))/2

Korrektuuri liige, mis liidetuna *Exceli* funktsiooni `RANK` tulemusele võimaldab arvutada korrektse võrdseid väärtusi arvesse võtva astaku; igale väärtusele vastav korrektuuriliige leitakse kujul: [(vaatluste arv) + 1 – (väärtuse jrk. nr. kasvavalt järjestud reas) – (väärtuse jrk. nr. kahanevalt järjestatud reas)] / 2



4. Arvutage astakkorrelatsioonikordajad korrektsete astakute alusel nii valemi (*) alusel kui ka astakute vahelise lineaarse korrelatsioonikordajana. Miks tulevad piimatoodangu ja somaatiliste rakkude arvu astakkorrelatsioonikordajad seemenduste arvuga erinevate arvutusmeetodite korral erinevad?

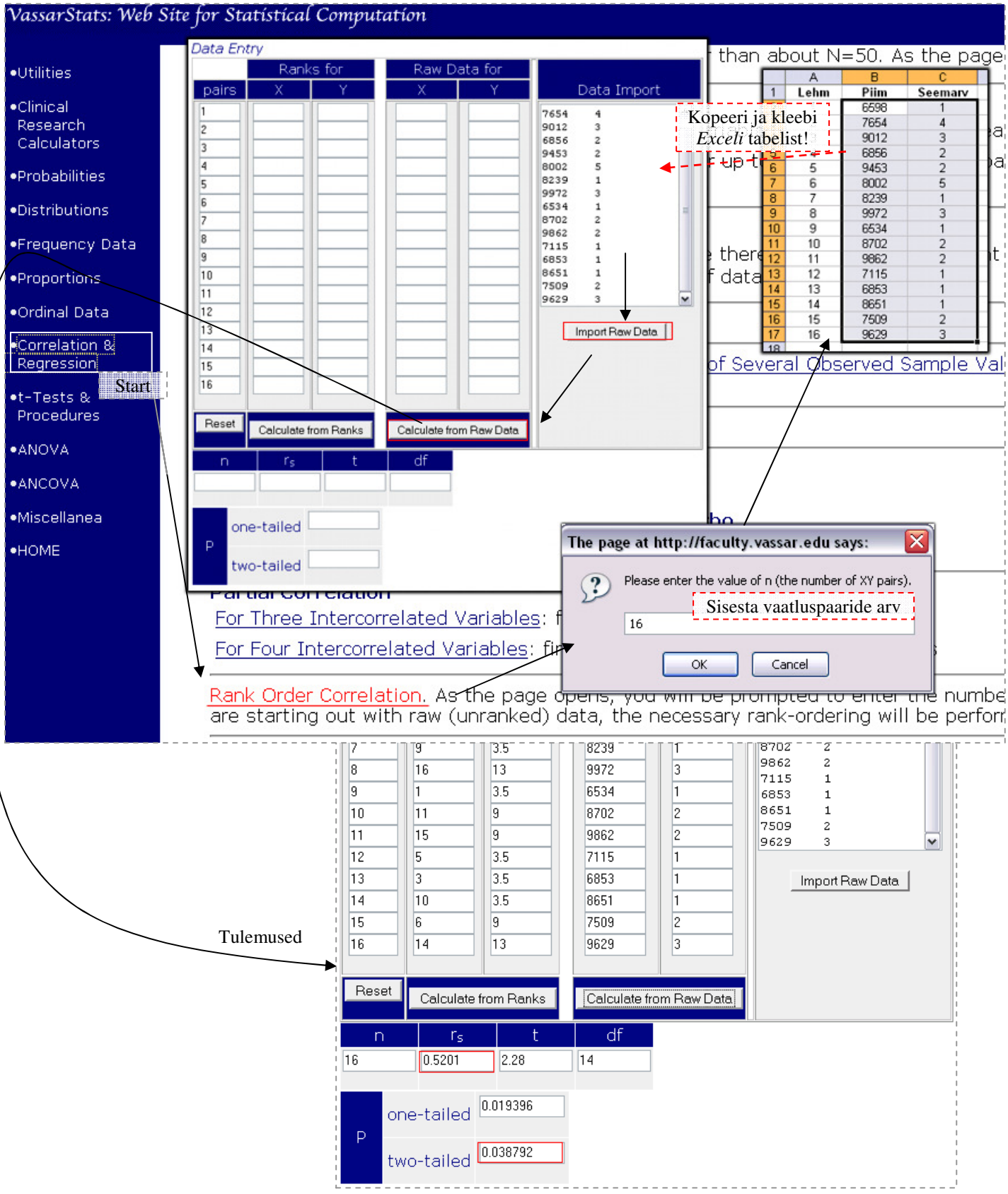
	K	L	M	N	O	P	Q	R	S	T
1	r(Piim)	r(SA)	r(SRA)	correct_r(Piim)	correct_r(SA)	correct_r(SRA)		[r(Piim)-r(SA)]^2	[r(Piim)-r(SRA)]^2	[r(SA)-r(SRA)]^2
2	15	11	3	15	13,5	3		=(N2-O2)^2	144	110,25
3	10	2	7	10	2	7		64	121	25
4	5	3	16	5	4	16		1	121	144
5	13	6	6	13	8	6		25	49	4
6	4	6	12	4	8	12		16	64	16
7	9	1	8	9	1	8		64	1	49
8	8	11	11	8	13,5	11		30,25	9	6,25
9	1	3	10	1	4	10		9	81	36
10	16	11	5	16	13,5	5		6,25	121	72,25
11	6	6	4	6	8	4		4	4	16
12	2	6	13	2	8	13		36	121	25
13	12	11	15	12	13,5	15		2,25	9	2,25
14	14	11	1	14	13,5	1		0,25	169	156,25
15	7	11	9	7	13,5	9		42,25	4	20,25
16	11	6	2	11	8	2		9	81	36
17	3	3	14	3	4	14		1	121	100
18										
19										
20		astakkorrelatsioonikorda		Piim<->SA	Piim<->SRA	SA<->SRA		Piim<->SA	Piim<->SRA	SA<->SRA
21			ρ	=CORREL(N2:N17;O2:O17)		-0,2585		=1-6*SUM(R2:R17)/(COUNT(R2:R17)*(COUNT(R2:R17)^2-1))		

Korrektsete astakute vaheline ruuterinevus $(x_{(i)} - y_{(i)})^2$

Korrektsete (st korduvate väärtuste osas korrigeeritud) astakute vaheline lineaarne korrelatsioonikordaja = astakkorrelatsioonikordaja

Astakkorrelatsioonikordajad arvutatuna mittekorduvaid väärtusi eeldava valemi (*) alusel:
$\rho = 1 - 6 \times \sum_{i=1}^{n} (x_{(i)} - y_{(i)})^2 / n(n^2 - 1)$

5. Arvutage seemenduste arvu ja piimatoodangu vaheline astakkorrelatsioonikordaja ka internetilehe http://faculty.vassar.edu/lowry/VassarStats.html abil. Kirjutage tulemustest välja leitud astakkorrelatsioonikordajale vastav *p*-väärtus ja sõnastage lõppjäreldus.

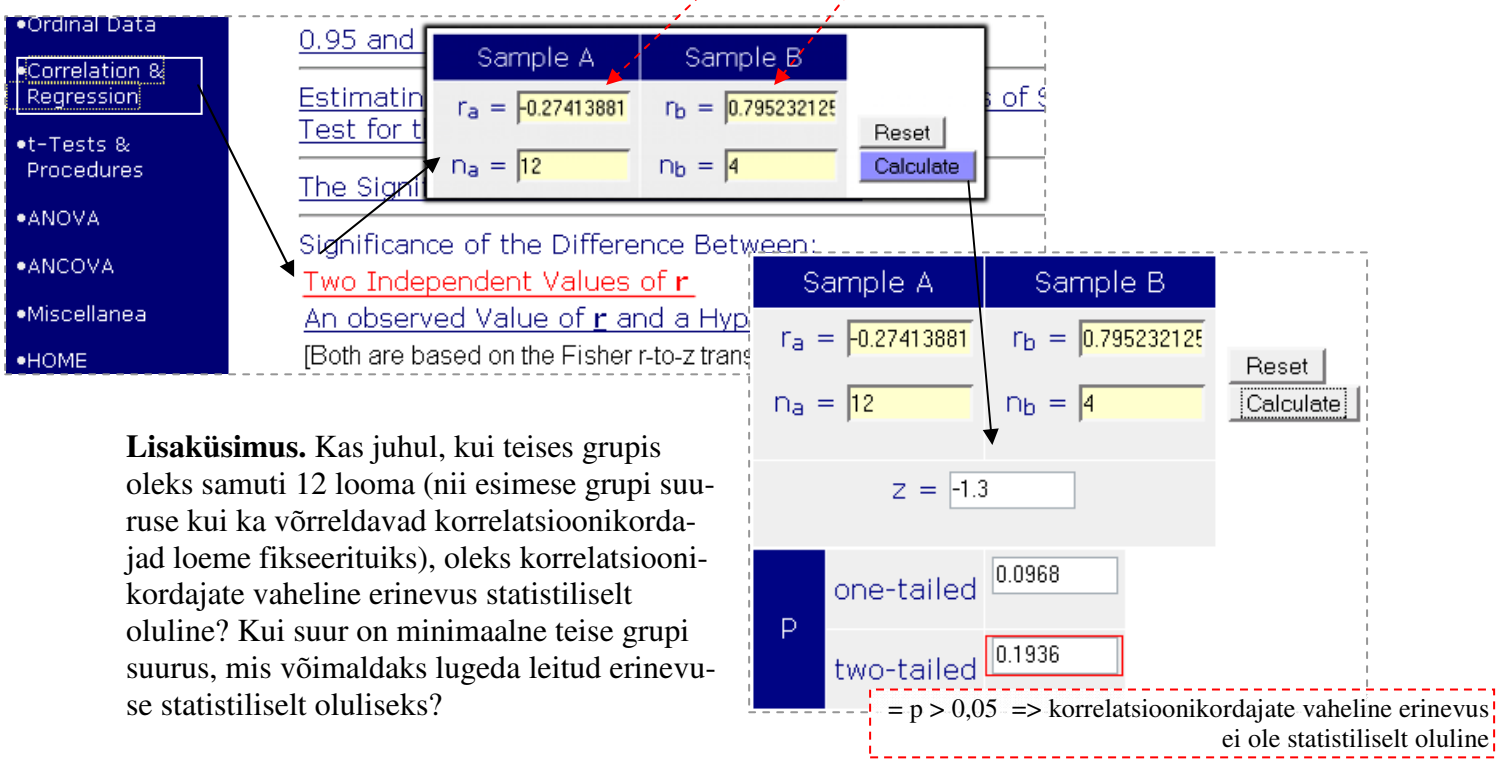


6. Miks on piimatoodangu ja somaatiliste rakkude (või piimatoodangu ja seemenduste arvu) vaheline astakkorrelatsioonikordaja poolt kirjeldatud seos märgatavalt tugevam, kui lineaarne seos (arusaamiseks võite uurida nii SRA histogrammi kui ka piimatoodangu ja SRA vahelist hajuvusdiagrammi)?

7. Arvutage seemenduste arvu ja SRA vahelised lineaarsed korrelatsioonikordajad markeri *GT4* mõlema genotüübi tarvis. Kas seos on erinevate genotüüpide korral erinev? Testige seose (lineaarsete korrelatsioonikordajate) erinevuse statistilist olulisust internetilehe http://faculty.vassar.edu/lowry/VassarStats.html abil.

Esmalt oleks mõttekas teha algandmete tabelist uuele töölehele koopia ja sorteerida kopeeritud andmestik *GT4* järgi.
Järgnevalt on vaja arvutada eraldi korrelatsioonikordajad nii *GT4*-genotüübi 'AA' kui ka genotüübi 'AG' tarvis ning sisestada *VasserStats*-lehel korrelatsioonikordajate erinevuse testimise leheküljel vastavaisse lahtrisse nii võrreldavad kordajad kui ka gruppide suurused.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Lehm**	**Piim**	**Seemarv**	**SRA**	**GT1**	**GT2**	**GT3**	**GT4**	**GT5**				
2	4	6856	2	231	GG	CC	CC	AA	AA		**r(SA, SRA | GT4='AA')**		
3	7	8239	1	99	CC	CC	AC	AA	AA		-0,27414		
4	8	9972	3	134	GC	CC	CC	AA	AA				
5	13	6853	1	3011	CC	CC	CC	AA	TT		**r(SA, SRA | GT4='AG')**		
6	15	7509	2	763	GC	CC	CC	AA	AT		=CORREL(C14:C17;D14:D17)		
7	1	6598	1	371	GC	CT	CC	AA	AT				
8	3	9012	3	12	CC	CT	CC	AA	TT				
9	5	9453	2	65	CC	CT	AA	AA	AT				
10	9	6534	1	252	CC	CT	CC	AA	AA				
11	10	8702	2	316	GG	CT	AC	AA	AA				
12	12	7115	1	24	CC	CT	CC	AA	AT				
13	14	8651	1	150	GC	CT	CC	AA	TT				
14	2	7654	4	196	GG	CC	AC	AG	TT				
15	6	8002	5	167	GC	CC	CC	AG	AT				
16	11	9862	2	60	GG	CC	AA	AG	TT				
17	16	9629	3	39	GG	CC	CC	AG	AA				



Lisaküsimus. Kas juhul, kui teises grupis oleks samuti 12 looma (nii esimese grupi suuruse kui ka võrreldavad korrelatsioonikordajad loeme fikseerituiks), oleks korrelatsioonikordajate vaheline erinevus statistiliselt oluline? Kui suur on minimaalne teise grupi suurus, mis võimaldaks lugeda leitud erinevuse statistiliselt oluliseks?

Ülesanne 2

Uurige 1. seemendusest tiinestuvuse ja geneetiliste markerite vahelisi seoseid?

- Arvutage kõigile lehmadele uue tunnuse, tiinestuvus 1. seemendusest, väärtused (baseeruvana seemenduste arvul: kui *SMARV* = 1, siis tiinestuvus esimesest seemendusest = 1, kui *SMARV* > 1, siis tiinestuvus esimesest seemendusest = 0).
- Uurige genotüübi *GT4* ja tiinestumise vahelist seost – konstrueerige 2-mõõtmeline sagedustabel, leidke selle baasil nii tiinestumiskordajad, riskisuhted kui ka šansside suhted, arvutage šansside suhte 95%-usaldusintervall ning sõnastage järeldused.
- Testige *GT4* ja tiinestumise vahelist seost χ^2-testiga.
- Viige sama seose testimiseks läbi ka Fisheri täpne test – esmalt *Excelis* (pannes selleks kirja kõik võimalikud samade rea- ja veerusummadega sagedustabelid ja arvutades valemist (**) nende esinemistõenäosused) ning seejärel internetilehe http://faculty.vassar.edu/lowry/VassarStats.html vahendusel.
- Uurige tiinestuvuse seost ka mõne teise genotüübiga (Fisheri täpse testi *Excelis* teostamise võib vahele jätta, sest näiteks 2×3 tabeli puhul on see juba üsna keeruline ja korra läbi tegemisest vast piisab, mõistmaks nimetatud testi olemust).

Teooria ülesande 2 juurde

Kui uuritavad tunnused on diskreetsed või mittearvulised, on nende vahelise seose (assotsiatsiooni) uurimisel loomulik kasutada 2-mõõtmelisi sagedustabeleid. Järgnevalt on lihtsuse ja enim kasutatavuse huvides piirdutud üksnes nn 2×2-tabeli analüüsiga, aga kogu teooria laieneb loomulikult ka suurematele tabelitele.

Tüüpiline 2×2-tabel on statistilise analüüsi tulemuseks juhul, kui uuritav tunnus on binaarne (nn 0-1-tüüpi tunnus), väljendades mingi sündmuse toimumist (haige=1/terve=0; on parasiit=1/ ei ole parasiiti=0; tiine=1/mittetiine=0; juhud=1/kontrollid=0), ning potentsiaalsel riski- või mõjuteguril on samuti vaid 2 väärtust (genotüüp1/genotüüp2; pidamissüsteem1/pidamissüsteem2; mõjule eksponeeritud/mõjule mitteeksponeeritud):

	Juhud (haiged/tiined/*responders*/…)	Kontrollid (terved/mittetiined/*nonresponders*/…)	Kokku
Eksponeeritud (markergenotüüp/-alleel 1)	*a*	*b*	*a+b*
Mitteeksponeeritud (markergenotüüp/-alleel 2)	*c*	*d*	*c+d*
Kokku	*a+c*	*b+d*	*a+b+c+d=N*

Sagedustabelite analüüsil leitavad kirjeldavad statistikud, mille alusel võimalikke assotsiatsioone lahti rääkida:

- **sagedused** (*observed frequences*) *a*, *b*, *c*, *d*;
- **oodatavad sagedused** assotsiatsiooni puudumise korral (uuritavate tunnuste sõltumatuse eeldusel, st nullhüpoteesi kehtides; *expected frequences*), arvutatavad valemeist $(a+b)\times(a+c)/N$, … – võimaldavad välja selgitada sõltumatuse juhust enim erinevad väärtuste kombinatsioonid;
- **(haigestumus, tiinestuvus, …)kordaja** (IR – *incidence rate*, *rate*) iga grupi (markergenotüübi/-alleeli) tarvis – $a/(a+b)$, $c/(c+d)$;

- **riskisuhe** (RR – *risk ratio*, *relative risk*), mis leitakse tavaliselt vähima (haigestumus) kordajaga grupi (markergenotüübi/-alleeli) suhtes – näiteks eeldades, et $a/(a+b)>c/(c+d)$: $RR_1=[a/(a+b)]/[c/(c+d)]$, $RR_2=1$;
- **šansside suhe** (OR – ***odds ratio***) leitakse sarnaselt riskisuhtele enamasti vähima (haigestumis) tõenäosusega grupi suhtes ja on oma olemuselt RR-i hinnang – näiteks $OR_1=(a/b)/(c/d)$, $OR_2=1$. Šansside suhe näitab, kui mitu korda erineb uuritava sündmuse toimumise šanss ühes grupis võrreldes teis(te)ga.

Gruppide erinevuse statistilise olulisuse tuvastamiseks (see on samaväärne seose statistilise olulisuse testimisega) arvutatakse sageli 95%-usaldusintervall (*95% confidence interval*; 95% CI) kas riski-suhtele või šansside suhtele.

Ligikaudsed 95%-usalduspiirid riskisuhtele on leitavad valemist

$$95\%\ \mathrm{CI_{RR}} \approx e^{\ln(\mathrm{RR})\pm 1.96 se[\ln(\mathrm{RR})]} = \left(\frac{\mathrm{RR}}{e^{1.96 se[\ln(\mathrm{RR})]}}; \mathrm{RR} * e^{1.96 se[\ln(\mathrm{RR})]}\right),$$

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

Ligikaudsed 95%-usalduspiirid šansside suhtele on leitavad valemist

$$95\%\ \mathrm{CI_{OR}} \approx e^{\ln(\mathrm{OR})\pm 1.96 se[\ln(\mathrm{OR})]} = \left(\frac{\mathrm{OR}}{e^{1.96 se[\ln(\mathrm{OR})]}}; \mathrm{OR} * e^{1.96 se[\ln(\mathrm{OR})]}\right), \qquad (**)$$

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

Kui arv 1 (mis on nii riski- kui ka šansside suhte väärtuseks nö baasgrupil) jääb usalduspiiride vahelt välja, võib gruppide vahelise erinevuse (ja seega ka uuritavate tunnuste assotsieerituse) lugeda tõestatuks ($p<0,05$ – seda juhul, kui leiti 95%-usalduspiirid).

Statistilise testina, kontrollmaks sagedustabelist ilmneva seose statistilist olulisust, kasutatakse enamasti **χ^2-testi** (hii-ruut test; *chi-square test*).

Viimast püütakse siiski vältida, kui mõni oodatavatest sagedustest on <5. Põhjuseks on asjaolu, et χ^2-test (nagu väga paljud teisedki statistikas vaikimisi kasutatavad meetodid) on asümptootiline (st, et tulemused on ligikaudsed, lähenedes õigetele väärtustele andmemahu suurenedes). Alternatiivina χ^2-testile kasutatakse enamasti **Fisheri täpset testi** (*Fisher exact test*)..

Fisheri täpse testi korral leitakse esmalt kõik võimalikud sagedustabelid fikseeritud rea- ja veerusummade korral (st, et leitakse andmete põhjal (= empiiriline) sagedustabel ning pannakse kirja kõik sellised tabelid, mis erineva „sisu“ korral annavad tulemuseks ikkagi samad rea- ja veerusummad).

Hüpergeomeetriline jaotus on see matemaatiline seaduspära, mis võimaldab fikseeritud rea- ja veerusummade ning väärtuste juhusliku kombineerumise (tunnuste sõltumatuse) eeldusel välja arvutada iga võimaliku tabeli esinemistõenäosuse valemist

$$\pi = \frac{\left(\prod_{i=1}^{k} n_{i.}!\right)\left(\prod_{j=1}^{m} n_{.j}!\right)}{n!\prod_{i,j} n_{i,j}}, \quad n! = n \times (n-1) \times \ldots \times 2 \times 1,$$

siin n on vaatluste koguarv, n_{ij} sagedustabeli i. reas ja j. veerus paiknev sagedus ning $n_{i.}$ ja $n_{.j}$ vastavalt i. rea ja j. veeru summad. Eelnevalt vaadeldud 2×2-tabeli tarvis esitub toodud valem kujul

$$\pi = [(a+b)!\times(c+d)!\times(a+c)!\times(b+d)!] \,/\, [N!\times a!\times b!\times c!\times d!]. \qquad (***)$$

Järgnevalt liidetakse kokku andmeid kirjeldava ning sellest ekstreemsemate (vähemtõenäolises suunas valitud) sagedustabelite esinemistõenäosused – tulemuseks on 1-poolsele hüpoteesile vastav olulisuse tõenäosus (*p*-väärtus); 2-poolsele hüpoteesile vastava olulisuse tõenäosuse saab, liites eelnevalt leitud kõikvõimalike sagedustabelite tõenäosusjaotuse ühele „sabale“ sümmeetrilise „saba“ ka teiselt poolt või korrutades 1-poolsele hüpoteesile vastava *p*-väärtuse lihtsalt 2-ga.

Ülesande 2 tööjuhend

1. Arvutage kõigile lehmadele uue tunnuse, tiinestuvus 1. seemendusest, väärtused (baseeruvana seemenduste arvul: kui *SMARV* = 1, siis tiinestuvus esimesest seemendusest = 1, kui *SMARV* > 1, siis tiinestuvus esimesest seemendusest = 0).

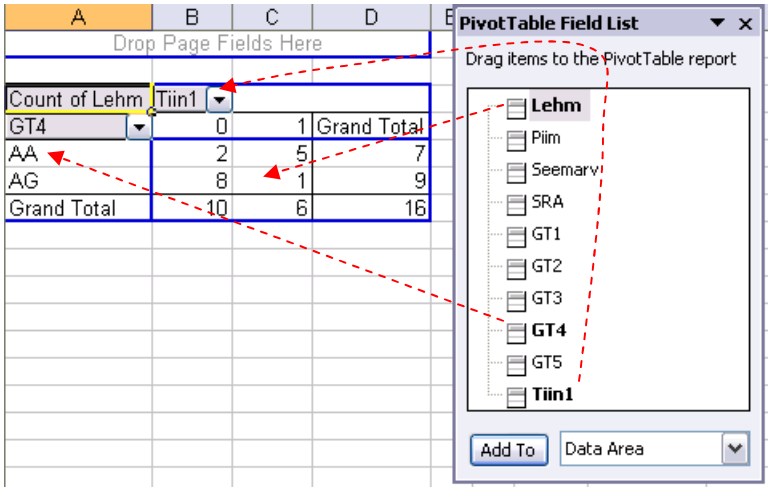
	A	B	C	D	E	F	G	H
1	**Lehm**	**Piim**	**Seemarv**	**SRA**	**Tiinestuvus**	**GT1**	**GT2**	**GT3**
2	1	6598	1		=IF(C2=1;1;0)		CT	CC
3	2	7654	4	196	IF(logical_test; [value_if_true]; [value_if_false])			
4	3	9012	3	12		CC	CT	CC
5	4	6856	2	231		GG	CC	CC

2. Uurige genotüübi GT4 ja tiinestumise vahelist seost – konstrueerige 2-mõõtmeline sagedustabel, leidke selle baasil nii tiinestuvuskordajad, riskisuhted kui ka šansside suhted, arvutage šansside suhte 95%-usaldusintervall ning sõnastage järeldused.

Esmase 2-mõõtmelise sagedustabeli leidmiseks on mugav kasutada *Exceli* vahendit *Pivot Table*:

Kõiksugu edasiste arvutuste tegemiseks võib konstrueeritud tabeli väärtustest teha koopia … ja siis arvutadagi vastavalt teooria osas toodud valemitele.

NB! Enne kõiksugu kordajate arvutamist peab panema paika, mida uuritakse – kas tiinestumist või mittetiinestumist (mis on huvi pakkuv ja modelleeritav sündmus). St, kas järeldusi soovitakse sõnastada tiinestumise šansi või mittetiinestumise šansi kohta.



Näiteks modelleerides tiinestumist, saame leida tiinestuvuskordaja:

Genotüüp	Tiinestus			Tiinestuvuskordaja
	Ei	Jah		IR
AA	2	5	5 / (2+5) =	0,714
AG	8	1	1 / (1+8) =	0,111

Aga soovides sõnastada järeldusi mittetiinestumise kohta, võime leida mittetiinestumise kordaja:

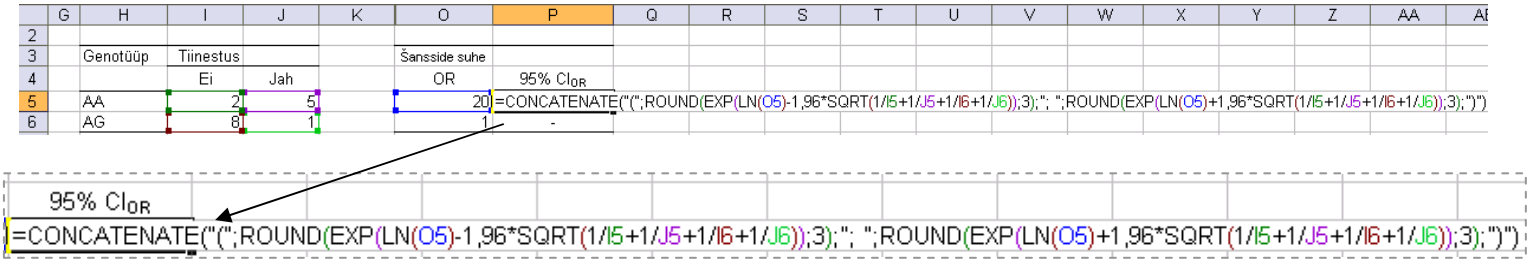
Genotüüp	Tiinestus		Tiinestuvuskordaja	Mittetiinestumise	
	Ei	Jah	IR	kordaja	
AA	2	5	0,714	0,286	= 2 / (2+5)
AG	8	1	0,111	0,889	= 8 / (8+1)

Oletame, et tahame järgnevalt sõnastada järeldusi tiinestumise kohta, seega arvutame ka riskisuhted ja/või šansside suhted tiinestumise kohta.

	G	H	I	J	K	L	M	N	O	P
2										
3		Genotüüp	Tiinestus			Tiinestuvuskordaja	Mittetiinestumise	Riskisuhe	Šansside suhe	
4			Ei	Jah		IR	kordaja	RR	OR	95% CI_{OR}
5		AA	2	5		0,714	0,286	6,429	20	(1,416; 282,463)
6		AG	8	1		0,111	0,889	1	1	-

[5 / (2+5)] / [1 / (8+1)] = (5 / 2) / (1 / 8) =

Šansside suhte ligikaudne 95% usaldusintervall on arvutatud vastavalt valemile (**) – viimase tarvis võib *Excelis* eraldi lahtritesse arvutada nii logaritmilise šansside suhte standardvea *se*[ln(OR)] kui ka alumise ja ülemise usalduspiiri. Mina realiseerisin kõik ühe korraga, pika ja kohmaka valemiga:



Panin selle siia kirja lihtsalt demonstreerimaks, kuidas saab *Excelis* erinevaid funktsioone (nii teksti kui ka matemaatika omi) ja tehteid koos kasutada. Eesrindlikumad võivad proovida aru saada (ja küsida, kui vaja).

Aga järeldused arvutatust.

- Tiinestuvuskordajad näitavad, et genotüübiga *AA* lehmadest tiinestus 71,4% (tiinestuvus-kordaja *IR* = 0,714) ja genotüübiga *AG* lehmadest 11,1%, seega tiinestusid *AA*-genotüübiga lehmad 6,43 korda paremini
- Seda, milline on võrreldavate gruppide vaheline erinevus, näitab ka riskisuhe, millest järeldused sõnastatakse enamasti kujul, et *AA*-genotüübiga lehmadel on 6,43 korda suurem risk tiinestuda võrreldes *AG*-genotüübiga lehmadega. Muidugi on riskist rääkides sisuliselt mõistetavam teha juttu „riskist mitte tiinestuda“ (mitte terveneda, haigestuda), selle asemel, et väita midagi „riskist tiinestumise“ (tervenemise, mitte haigestumise) kohta … Aga sellest lähtuvalt valitaksegi, mida käsitleda nö sündmuse toimumisena, kas tiinestumist või mitte tiinestumist.
- Šansside suhe OR = 20 (95% CI (1,416; 282,463)) näitab, et genotüübiga AA lehmadel on šanss tiinestuda 20 korda suurem, võrreldes AG-genotüübiga lehmadega, seejuures on see erinevus statistiliselt oluline (sest 95%-usaldusintervall ei sisalda arvu 1).

Lisaülesanne. Arvutage riskisuhted ja šansside suhted (+ 95%-usaldusintervall) ka mittetiinestumise tarvis. Kindlasti püüdke leitud kordajate alusel sõnastada mõned järeldused.

3. Testige *GT4* ja tiinestumise vahelist seost χ^2-testiga.

	A	B	C	D	E	F
1						
2	**Empiirilised sagedused**					
3	Count of Lehm	Tiin1				
4	GT4	0	1	Grand Total		
5	AA	2	5	7		
6	AG	8	1	9		
7	Grand Total	10	6	16		
8						
9	**Teoreetilised (0-hüpoteesile vastavad) sagedused**					
10	Tiin1	AA	AG	Grand Total		
11	0	4,375	2,625	7		
12	1	5,625	3,375	9		
13	Grand Total	10	6	16		
14						
15						
16	**Hii-ruut test**	=CHITEST(B5:C6;B11:C12)				

Hii-ruut test	**0,0134**	= p < 0,05 => seos GT4 ja tiinestumise vahel on statistiliselt oluline
		(analoogsele järeldusele jõudsime ka šansside suhte usaldusintervalli alusel)

4. Viige sama seose testimiseks läbi Fisheri täpne test.
   Tehke seda esmalt *Excelis* (pannes selleks kirja kõik võimalikud samade rea- ja veerusummadega sagedustabelid ja arvutades valemist (***) nende esinemistõenäosused).

Antud rea- ja veerusummadele (st antud genotüübiarvude ning tiinestunud ja mittetiinestunud lehmade arvudele) vastavate kõikvõimalike 2×2-tabelite kirja panekuks on lihtsaim moodus võtta vaatluse alla vähimale rea- ja veerusummale vastav lahter, muuta seal olevat arvu 0-st kuni maksimumini (= minimaalne rea- või veerusumma) ning arvutada ülejäänud 3 sagedust sellest lähtuvalt.

	A	B	C	D
22				
23	Genotüüp	Tiinestus		Kokku
24		Ei	Jah	
25	AA	=D25-C25	0	7
26	AG	=B27-B25	=C27-C25	9
27	Kokku	10	6	16

See väärtus on fikseeritud ja ülejäänud arvutatakse selle ning fikseeritud rea- ja veerusummade alusel.

Konkreetsele tabelile vastav tõenäosus tuleb leida valemist (***):

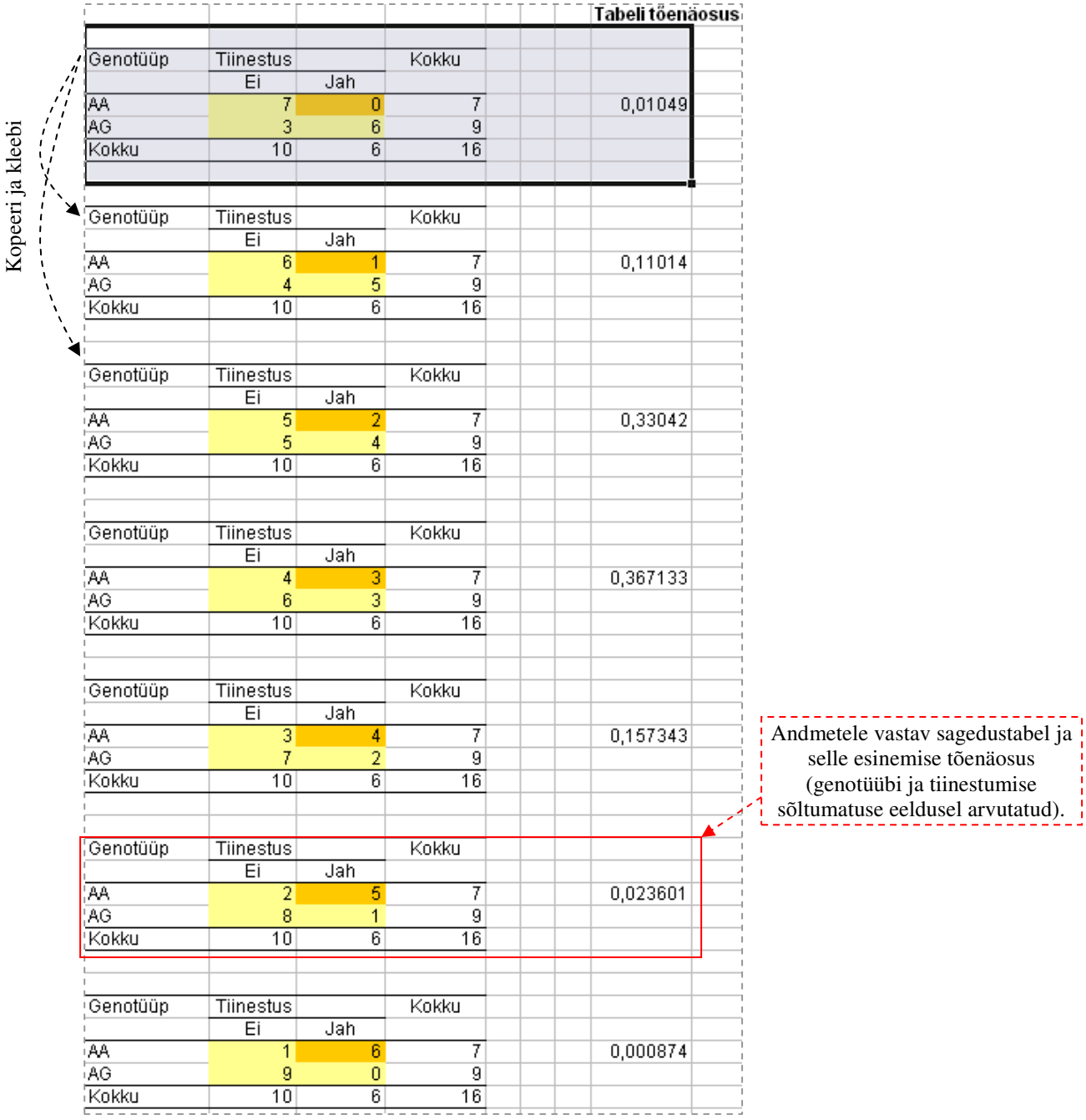
H25 =(FACT(D25)*FACT(D26)*FACT(B27)*FACT(C27))/(FACT(B25)*FACT(C25)*FACT(B26)*FACT(C26)*FACT(D27))

	A	B	C	D	H
21				Valem (***)	
22					Tabeli tõenäosus
23	Genotüüp	Tiinestus		Kokku	
24		Ei	Jah		
25	AA	7	0	7	0,0104895
26	AG	3	6	9	
27	Kokku	10	6	16	

Faktoriaali $n!$ leidmiseks on *Excelis* funktsioon `FACT(n)`.

Tõenäosus, et **7** AA- ja **9** AG-genotüüpi ning **6** tiinestumist ja **10** mittetiinestumist võinuks juhuslikult kombineeruda just antud tabelis toodu kohaselt.

Järgnevalt on mõttekas kopeerida korraga nii konstrueeritud tabelit (koos selles sisalduvate valemitega) kui ka tabeli tõenäosuse arvutamise valemit, ja seda 7 korda (miks 7**?**). Seejärel tuleb muuta üksnes algselt 0-ga võrdseks võetud lahtris paiknevat väärtust, et arvutada välja kõikvõimalikud antud rea- ja veerusummadele vastavad tabelid ja nende esinemise tõenäosused (eeldades genotüübi ja tiinestuvuse vahelise seose puudumist).



Genotüüp	Tiinestus		Kokku	Tabeli tõenäosus
	Ei	Jah		
AA	7	0	7	0,01049
AG	3	6	9	
Kokku	10	6	16	

Genotüüp	Tiinestus		Kokku	
	Ei	Jah		
AA	6	1	7	0,11014
AG	4	5	9	
Kokku	10	6	16	

Genotüüp	Tiinestus		Kokku	
	Ei	Jah		
AA	5	2	7	0,33042
AG	5	4	9	
Kokku	10	6	16	

Genotüüp	Tiinestus		Kokku	
	Ei	Jah		
AA	4	3	7	0,367133
AG	6	3	9	
Kokku	10	6	16	

Genotüüp	Tiinestus		Kokku	
	Ei	Jah		
AA	3	4	7	0,157343
AG	7	2	9	
Kokku	10	6	16	

Genotüüp	Tiinestus		Kokku	
	Ei	Jah		
AA	2	5	7	0,023601
AG	8	1	9	
Kokku	10	6	16	

Genotüüp	Tiinestus		Kokku	
	Ei	Jah		
AA	1	6	7	0,000874
AG	9	0	9	
Kokku	10	6	16	

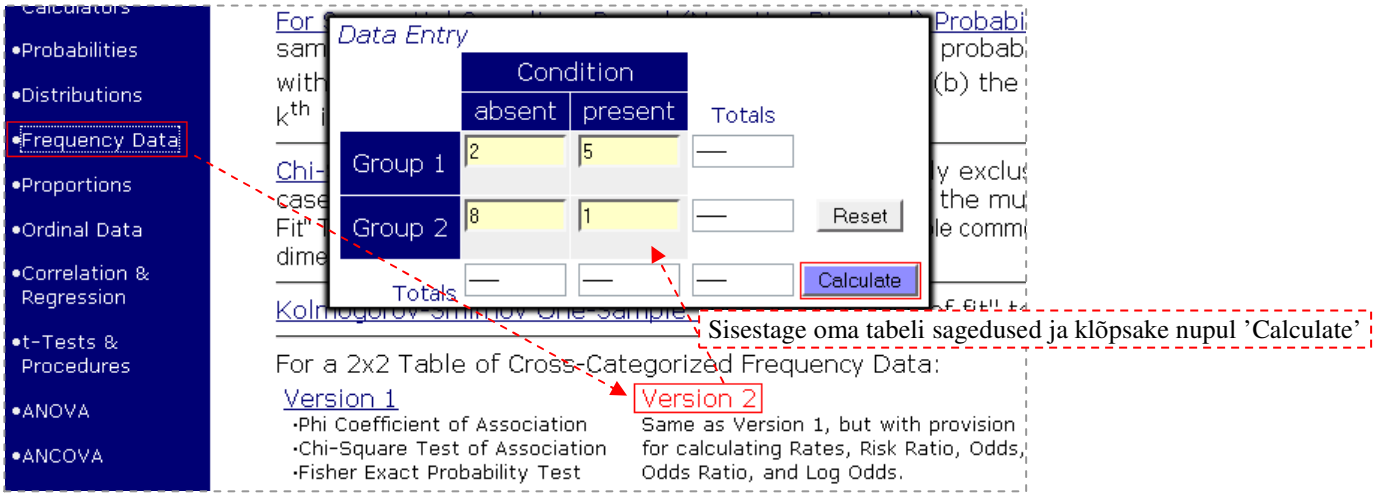
Genotüübi ja tiinestumise sõltuvuse statistilise olulisuse üle otsustamiseks vajalik olulisuse tõenäosus p on leitav, kui andmetele vastava sagedustabeli ja sellest veel ekstreemsemate tabelite esinemise tõenäosuste summa: $p = 0{,}023601 + 0{,}000874 + 0{,}01049 \approx 0{,}0350$.

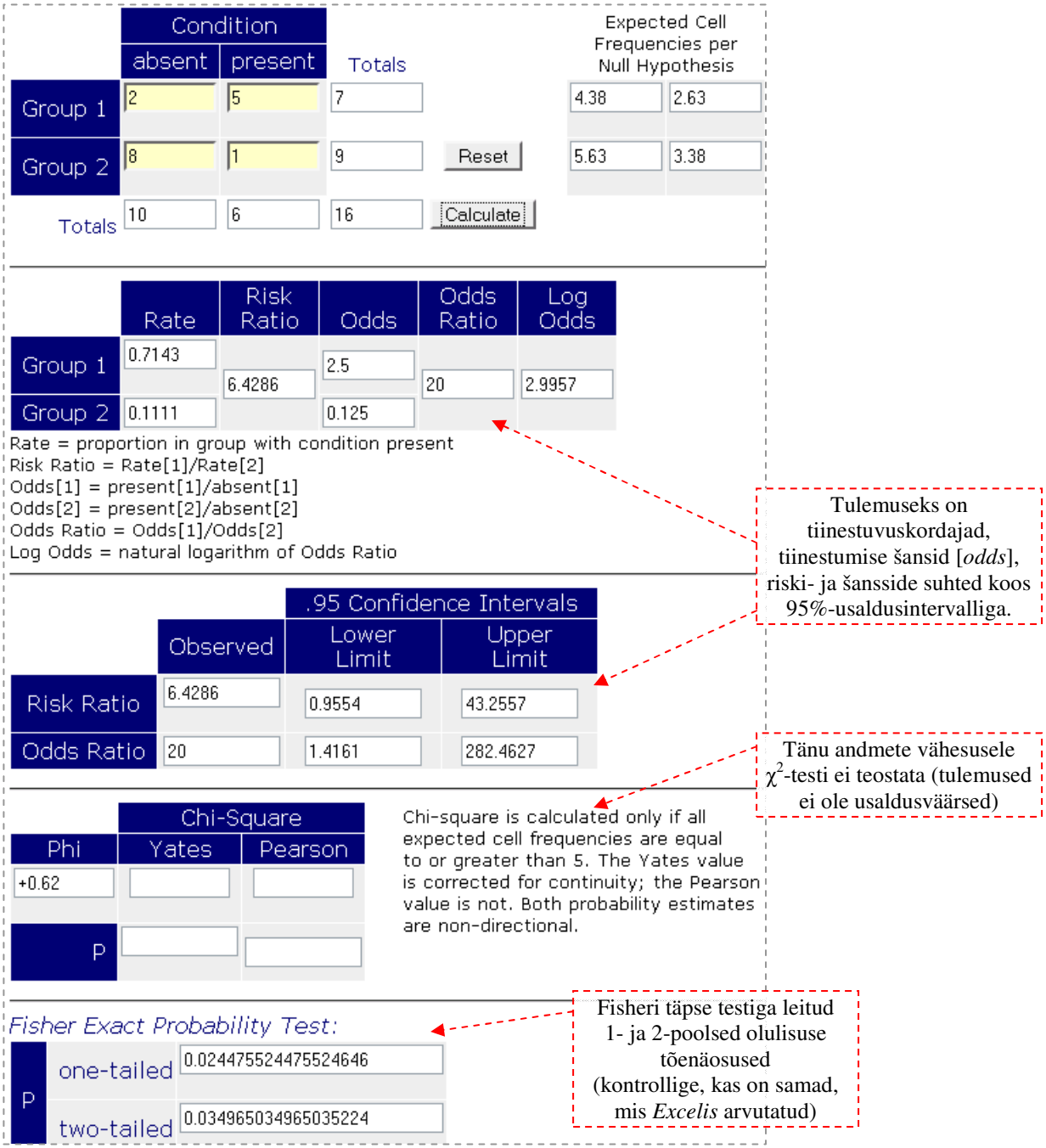
Alternatiivina leitakse olulisuse tõenäosus Fisheri täpsest testist mõnikord ka kui 2-kordne andmetele vastava tabeli ja sellest tõenäosuse vähenemise suunas valitud tabelite esinemistõenäosuste summa: $p = 2\times(0{,}023601 + 0{,}000874) \approx 0{,}0490$.

Hüpoteeside paar					
H_0: genotüüp ja tiinestumine ei ole seotud					
H_1: genotüüp ja tiinestumine on seotud					
Fisheri täpne test					
Olulisuse tõenäosus:		**0,0350**			
	või	0,0490			
Seega p < 0,05, mistõttu võime ka Fisheri täpse testi alusel teha järelduse:					
	GT4 ja tiinestumise vaheline seos on statistiliselt oluline.				

Võrreldes χ^2-testiga on saadud *p*-väärtus pisut suurem, mis tähendab, et tänu andmete vähesusele χ^2-test pisut ülehindas seose tugevust.

5. Teostage Fisheri täpne test internetilehe http://faculty.vassar.edu/lowry/VassarStats.html vahendusel.





6. Uurige tiinestuvuse seost ka mõne teise genotüübiga (Fisheri täpse testi *Excelis* teostamise võib vahele jätta, sest näiteks 2×3 tabeli puhul on see juba üsna keeruline ja korra läbi tegemisest vast piisab, mõistmaks nimetatud testi olemust).