

Praktikum 6

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

Praktikum püüab kirjeldada mõningaid võimalusi tunnuste vaheliste seoste uurimiseks.

Kommentaariid 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 astakorrelatsioonikordajad.
 - 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 astakorrelatsioonikordajad korrektsete astakute alusel nii valemi (*) alusel (vt teooria osa järgm. lehel) kui ka astakute vahelise lineaarse korrelatsioonikordajana. Miks tulevad piimatoodangu ja somaatiliste rakkude arvu astakorrelatsioonikordajad seemenduste arvuga erinevate arvutusmeetodite korral erinevad?
 - Arvutage seemenduste arvu ja piimatoodangu vaheline astakorrelatsioonikordaja ka internetilehe <http://faculty.vassar.edu/lowry/VassarStats.html> abil. Kirjutage tulemustest välja ka astakorrelatsioonikordajale vastav p-väärtus ja sõnastage lõppjärgeldus.
 - Miks on piimatoodangu ja somaatiliste rakkude (või piimatoodangu ja seemenduste arvu) vaheline astakorrelatsioonikordaja 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 arv-tunnuste näol on tegu pidevate ja sümmeetrilise jaotusega tunnustega, on loomulikem seosekordaja **lineaarne** (e **Pearsoni**) **korrelatsioonikordaja**, mis mõõdab kahe arv-tunnuse 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

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

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 arv-tunnuse vahelise monotoonse seose tugevust ja suunda.

Ülesande 1 tööjuhend

- Arvutage kõigi arvutunnuste 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 nõ klassikalisel viisil ja käsitsi,

- leides esmalt teststatistiku väärtuse valemist

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

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	=D22-2			
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 $n-2$ $|r\sqrt{n-2}/\sqrt{1-r^2}|$

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.

VassarStats: Web Site for Statistical Computation

- Utilities
- Clinical Research Calculators
- Probabilities
- Distributions
- Frequency Data
- Proportions
- Ordinal Data
- Correlation & Regression
- t-Tests & Procedures
- ANOVA
- ANCOVA
- Miscellanea
- HOME

Basic Linear Correlation and Regression
The following pages calculate r , r^2 , regression constant and Y , and perform a t-test for the significance of the

Data-Import Version. Allows for import of raw data from an Excel spreadsheet.
Direct-Entry Version. Values of X and Y are entered directly through the web page, though it will be rather unwieldy with samples larger than 100.

	A	B	C	D
1	Lehm	Piim	Seemravv	SRA
2	1	6598	1	371
3	2	7654	4	196
4	3	9012	3	12
5	4	6856	2	231
6	5	9453	2	65
7	6	8002	5	167
8	7	8239	1	99
9	8	9972	3	134
10	9	6534	1	252
11	10	8702	2	316
12	11	9862	2	60
13	12	7115	1	24
14	13	6853	1	3011
15	14	8651	1	150
16	15	7509	2	763
17	16	9629	3	39

Data Entry

7654	4
9012	3
6856	2
9453	2
8002	5
8239	1
9972	3
6534	1
8702	2
9862	2
7115	1
6853	1
8651	1
7509	2
9629	3

Data Report

Column 1: X
Column 2: Y
Column 3: Residual

Kopeeri ja kleebi!

Nn arvutada oskavad internetilehed, mis võimaldavad kopeerida andmeid *Excelis* tabelist, nõuavad, et taolise tegevuse tagajärjel kopeeritud tabeli lõppu tekki tühi rida (üleliigne reavahetus) oleks kustutatud!

Reset Calculate

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

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

Data Summary

$\sum X = 130641$ $\sum X^2 = 1088257679$

$\sum Y = 34$ $\sum Y^2 = 94$

$\sum XY = 285219$

	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	one-tailed	0.0916415	
1.4	14		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 arvturnuste vahelised astakkorrelatsioonikordajad.

Selleks leidke esmalt kõigile turnustele 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;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		3	7	11	3	11	3							
7	6	8002	5	167		8	11	11	8	13,5	11							
8	7	8239	1	99		1	3	10	1	4	10							
9	8	9972	3	134		16	11	5	16	13,5	16							
10	9	6534	1	252		6	6	4	6	8	6							
11	10	8702	2	316		2	6	13	2	8	2							
12	11	9862	2	60		12	11	15	12	13,5	12							
13	12	7115	1	24		14	11	1	14	13,5	14							
14	13	6853	1	3011		7	11	9	7	13,5	7							
15	14	8651	1	150		11	6	2	11	8	11							
16	15	7509	2	763		3	3	14	3	4	3							
17	16	9629	3	39														

Korrektuuri liige, mis liidetuna *Exceli* funktsiooni RANK tulemusele võimaldab arvutada korrektse võrdseid väärtusi arvesse võtva astaku; igale väärtusele vastav korrektuuri liige leitakse kujul:

$$[(\text{vaatluste arv}) + 1 - (\text{väärtuse jrk. nr. kasvavalt järjestatud reas}) - (\text{väärtuse jrk. nr. kahanevalt järjestatud reas})] / 2$$

Microsoft Excel Help

RANK
[See Also](#)

Returns the rank of a number in a list of numbers. The rank of a number is its size relative to other values in a list. (If you were to sort the list, the rank of the number would be its position.)

Syntax
RANK(number,ref,order)

Number is the number whose rank you want to find.
 Ref is an array of, or a reference to, a list of numbers. Nonnumeric values in ref are ignored.
 Order is a number specifying how to rank number.

- If order is 0 (zero) or omitted, Microsoft Excel ranks number as if ref were a list sorted in descending order.
- If order is any nonzero value, Microsoft Excel ranks number as if ref were a list sorted in ascending order.

Remarks

- RANK gives duplicate numbers the same rank. However, the presence of duplicate numbers affects the ranks of subsequent numbers. For example, in a list of integers sorted in ascending order, if the number 10 appears twice and has a rank of 5, then 11 would have a rank of 7 (no number would have a rank of 6).
- For some purposes one might want to use a definition of rank that takes ties into account. In the previous example, one would want a revised rank of 5.5 for the number 10. This can be done by adding the following correction factor to the value returned by RANK. This correction factor is appropriate both for the case where rank is computed in descending order (order = 0 or omitted) or ascending order (order = nonzero value).

Correction factor for tied ranks=[COUNT(ref) + 1 - RANK(number, ref, 0) - RANK(number, ref, 1)]/2.

In the following example, RANK(A2,A1:A5,1) equals 3. The correction factor is (5 + 1 - 2 - 3)/2 = 0.5 and the revised rank that takes ties into account is 3 + 0.5 = 3.5. If number occurs only once in ref, the correction factor will be 0, since RANK would not have to be adjusted for a tie.

4. Arvutage astakkorrelatsioonikordajad korrektsete astakute alusel nii valemi (*) alusel kui ka astakute vahelise lineaarse korrelatsioonikordajana. Miks tulevad piimatoodangu ja soomaatiliste 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	1	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	astakkorrelatsioonikordaja			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:

$$\rho = 1 - 6 \times \sum_{i=1}^n (x_{(i)} - y_{(i)})^2 / n(n^2 - 1)$$

- 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 lõppjärelendus.

VassarStats: Web Site for Statistical Computation

- Utilities
- Clinical Research Calculators
- Probabilities
- Distributions
- Frequency Data
- Proportions
- Ordinal Data
- Correlation & Regression
- t-Tests & Procedures
- ANOVA
- ANCOVA
- Miscellanea
- HOME

Data Entry

pairs	Ranks for		Raw Data for		Data Import
	X	Y	X	Y	
1					7654 4
2					9012 3
3					6856 2
4					9453 2
5					8002 5
6					8239 1
7					9972 3
8					6534 1
9					8702 2
10					9862 2
11					7115 1
12					6853 1
13					8651 1
14					7509 2
15					9629 3
16					

Import Raw Data

Calculate from Ranks Calculate from Raw Data

n	r_s	t	df

one-tailed

two-tailed

Partial Correlation

For Three Intercorrelated Variables: f

For Four Intercorrelated Variables: f

Rank Order Correlation. As the page opens, you will be prompted to enter the number of pairs of data you are starting out with raw (unranked) data, the necessary rank-ordering will be performed.

than about N=50. As the page

A	B	C
1	Lehm	Piim
2		Seemrav
3		
4		
5		
6	5	
7	6	
8	7	
9	8	
10	9	
11	10	
12	11	
13	12	
14	13	
15	14	
16	15	
17	16	
18		

Kopeeri ja kleebi Exceli tabelist!

The page at <http://faculty.vassar.edu> says:

Please enter the value of n (the number of XY pairs).

Sisesta vaatluspaaride arv

16

OK Cancel

Tulemused

n	r_s	t	df
16	0.5201	2.28	14

one-tailed 0.019396

two-tailed 0.038792

- Miks on piimatoodangu ja soomaatiliste 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.

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 *VassarStats*-lehel korrelatsioonikordajate erinevuse testimise leheküljel vastavasse 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		$r(\text{SA, SRA} \mid \text{GT4}=\text{'AA'})$		
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		$r(\text{SA, SRA} \mid \text{GT4}=\text{'AG'})$		
6	15	7509	2	763	GC	CC	CC	AA	AT		$=\text{CORREL}(C14:C17;D14:D17)$		
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				

- Ordinal Data
- **Correlation & Regression**
- t-Tests & Procedures
- ANOVA
- ANCOVA
- Miscellanea
- HOME

0.95 and

Sample A	Sample B
$r_a = -0.27413881$	$r_b = 0.795232125$
$n_a = 12$	$n_b = 4$

The Significance of the Difference Between:

Two Independent Values of r

An observed Value of r and a Hypothesized Value of r

[Both are based on the Fisher r-to-z transformation]

Sample A	Sample B
$r_a = -0.27413881$	$r_b = 0.795232125$
$n_a = 12$	$n_b = 4$

$z = -1.3$

P	one-tailed	0.0968
	two-tailed	0.1936

= $p > 0,05 \Rightarrow$ korrelatsioonikordajate vaheline erinevus ei ole statistiliselt oluline

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 mitteamarvulised, on nende vahelise seose (assotsiatsiooni) uurimisel loomulik kasutada 2-mõõtmelisi sagedustabeleid. Järgnevalt on lihtsuse ja enim kasutatavuse huvides piiratud 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/ <i>responders</i> /...)	Kontrollid (terved/mittetiined/ <i>nonresponders</i> /...)	Kokku
Eksponeeritud (marker-genotüüp/-alleel 1)	<i>a</i>	<i>b</i>	<i>a+b</i>
Mitteeksponeeritud (marker-genotüüp/-alleel 2)	<i>c</i>	<i>d</i>	<i>c+d</i>
Kokku	<i>a+c</i>	<i>b+d</i>	<i>a+b+c+d=N</i>

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

- **sagedused** (*observed frequencies*) *a, b, c, d*;
- **oodatavad sagedused** assotsiatsiooni puudumise korral (uuritavate tunnuste sõltumatuse eeldusel, st nullhüpoteesi kehtides; *expected frequencies*), arvutatavad valemist $(a+b) \times (a+c) / N, \dots$ – võimaldavad välja selgitada sõltumatuse juhust enim erinevad väärtuste kombinatsioonid;
- **(haigestumus, tiinestuvus, ...)kordaja** (IR – *incidence rate, rate*) iga grupi (marker-genotüübi/-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 riskisuhtele või šansside suhtele.

Ligikaudsed 95%-usalduspiirid riskisuhtele on leitavad valemist

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

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

Ligikaudsed 95%-usalduspiirid šansside suhtele on leitavad valemist

$$95\% \text{ CI}_{OR} \approx e^{\ln(OR) \pm 1.96 \text{se}[\ln(OR)]} = \left(\frac{OR}{e^{1.96 \text{se}[\ln(OR)]}}; OR * e^{1.96 \text{se}[\ln(OR)]} \right), \quad (**)$$

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

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 vaikumisi 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 sõltumatus) eeldusel välja arvutada iga võimaliku tabeli esinemistõenäosuse valemist

$$\pi = \frac{(\prod_{i=1}^k n_i!) (\prod_{j=1}^m n_j!)}{n! \prod_{i,j} n_{i,j}}, \quad n! = n \times (n-1) \times \dots \times 2 \times 1,$$

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 2x2-tabeli tarvis esitub toodud valem kujul

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

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 kõikkõimalike sagedustabelite tõenäosusjaotuse ühele „sabale“ sümmeetrilise „saba“ ka teiselt poolt või korrutades 1-poolsele hüpoteesile vastava *p*-väärtuse lihtsalt 2-ga.

Ülesande 2 tööjuhend

- 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

- 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 arvutada 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:

Count of Lehm	Tiin1	Grand Total
GT4	0	1
AA	2	5
AG	8	1
Grand Total	10	6

Genotüüp	Tiinestus		Tiinestuvuskordaja
	Ei	Jah	
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 kordaja
	Ei	Jah		
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 kordaja	Riskisuhe	Šansside suhe	
4			Ei	Jah		IR		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:

	G	H	I	J	K	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AI
2																			
3		Genotüüp	Tiinestus			Šansside suhe													
4			Ei	Jah		OR	95% CI _{OR}												
5		AA	2	5		20	=CONCATENATE("(",ROUND(EXP(LN(O5)-1,96*SQRT(1/I5+1/J5+1/I6+1/J6)),3);",",ROUND(EXP(LN(O5)+1,96*SQRT(1/I5+1/J5+1/I6+1/J6)),3);"))												
6		AG	8	1		1	-												

95% CI_{OR}
 =CONCATENATE("(",ROUND(EXP(LN(O5)-1,96*SQRT(1/I5+1/J5+1/I6+1/J6)),3);",",ROUND(EXP(LN(O5)+1,96*SQRT(1/I5+1/J5+1/I6+1/J6)),3);"))

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% (tiinestuvuskordaja 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 mõisteta- vama 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 nõ sündmuse toimumisena, kas tiinestumist või mitte tiinestumist.
- Šansside suhe OR = 20 (95% CI (1,416; 282,463)) näitab, et genotüübiga AA lehmadel on šans 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 = $p < 0,05 \Rightarrow$ seos *GT4* ja tiinestumise vahel on statistiliselt oluline (analoogsele järeldusele jõudsimise 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 veerusummade sagedustabelid ja arvutades valemist (***) nende esinemistõenäosused).

Antud rea- ja veerusummadele (st antud genotüübiarvude ning tiinestunud ja mittetiinestunud lehmade arvudele) vastavate kõikvõimalike 2x2-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.

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

H25	A	B	C	D	E	F	G	H	I	J	K	L	M
21													
22													
23	Genotüüp	Tiinestus		Kokku									
24		Ei	Jah										
25	AA	7	0	7				0,0104895					
26	AG	3	6	9									
27	Kokku	10	6	16									

Valem (***)

Tabeli tõenäosus

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 korraka nii konstrueeritud tabelit (koos selles sisalduvate valemitega) kui ka tabeli tõenäosuse arvutamise valemit, ja seda 7 korda (miks ??). 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).

Kopeeri ja kleebi

				Tabeli tõenäosus
Genotüüp	Tiinestus		Kokku	
	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	

Andmeteile vastav sagedustabel ja selle esinemise tõenäosus (genotüübi ja tiinestumise sõltumatuse eeldusel arvutatud).

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

Alternatiivina leitakse olulisuse tõenäosus Fisher'i täpsest testist mõnikord ka kui 2-kordne andmeteile 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.

Frequency Data

	Condition		Totals
	absent	present	
Group 1	2	5	—
Group 2	8	1	—
Totals	—	—	—

Version 2
Same as Version 1, but with provision for calculating Rates, Risk Ratio, Odds, Odds Ratio, and Log Odds.

Sisestage oma tabeli sagedused ja klõpsake nupul 'Calculate'

	Condition		Totals	Expected Cell Frequencies per Null Hypothesis	
	absent	present			
Group 1	2	5	7	4.38	2.63
Group 2	8	1	9	5.63	3.38
Totals	10	6	16		

Reset Calculate

	Rate	Risk Ratio	Odds	Odds Ratio	Log Odds
Group 1	0.7143		2.5		
Group 2	0.1111	6.4286	0.125	20	2.9957

Rate = proportion in group with condition present
 Risk Ratio = Rate[1]/Rate[2]
 Odds[1] = present[1]/absent[1]
 Odds[2] = present[2]/absent[2]
 Odds Ratio = Odds[1]/Odds[2]
 Log Odds = natural logarithm of Odds Ratio

	Observed	.95 Confidence Intervals	
		Lower Limit	Upper Limit
Risk Ratio	6.4286	0.9554	43.2557
Odds Ratio	20	1.4161	282.4627

Phi	Chi-Square	
	Yates	Pearson
+0.62		
P		

Chi-square is calculated only if all expected cell frequencies are equal to or greater than 5. The Yates value is corrected for continuity; the Pearson value is not. Both probability estimates are non-directional.

Fisher Exact Probability Test:

P	one-tailed	0.024475524475524646
	two-tailed	0.034965034965035224

Tulemuseks on tiinestuvuskordajad, tiinestumise šansid [odds], riski- ja šansside suhted koos 95%-usaldusintervalliga.

Tänu andmete vähesusele χ^2 -testi ei teostata (tulemused ei ole usaldusväärsed)

Fisheri täpse testiga leitud 1- ja 2-poolsed olulisuse tõenäosused (kontrollige, kas on samad, mis *Excelis* arvatud)

6. Uurige tiinestuvuse seost ka mõne teise genotüübiga (Fisheri täpse testi *Excelis* teostamise võib vahele jätta, sest näiteks 2x3 tabeli puhul on see juba üsna keeruline ja korra läbi tegemisest vast piisab, mõistmaks nimetatud testi olemust).