MAT20306 - Advanced Statistics

Lecture 12: Overview Advanced Statistics

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Some EXAM announcements

Zaal C0004

Adres (Plattegrond Wageningen)
Bornsesteeg 2
Gebouw 130 (Sporthal de Bongerd)
6708PE Wageningen

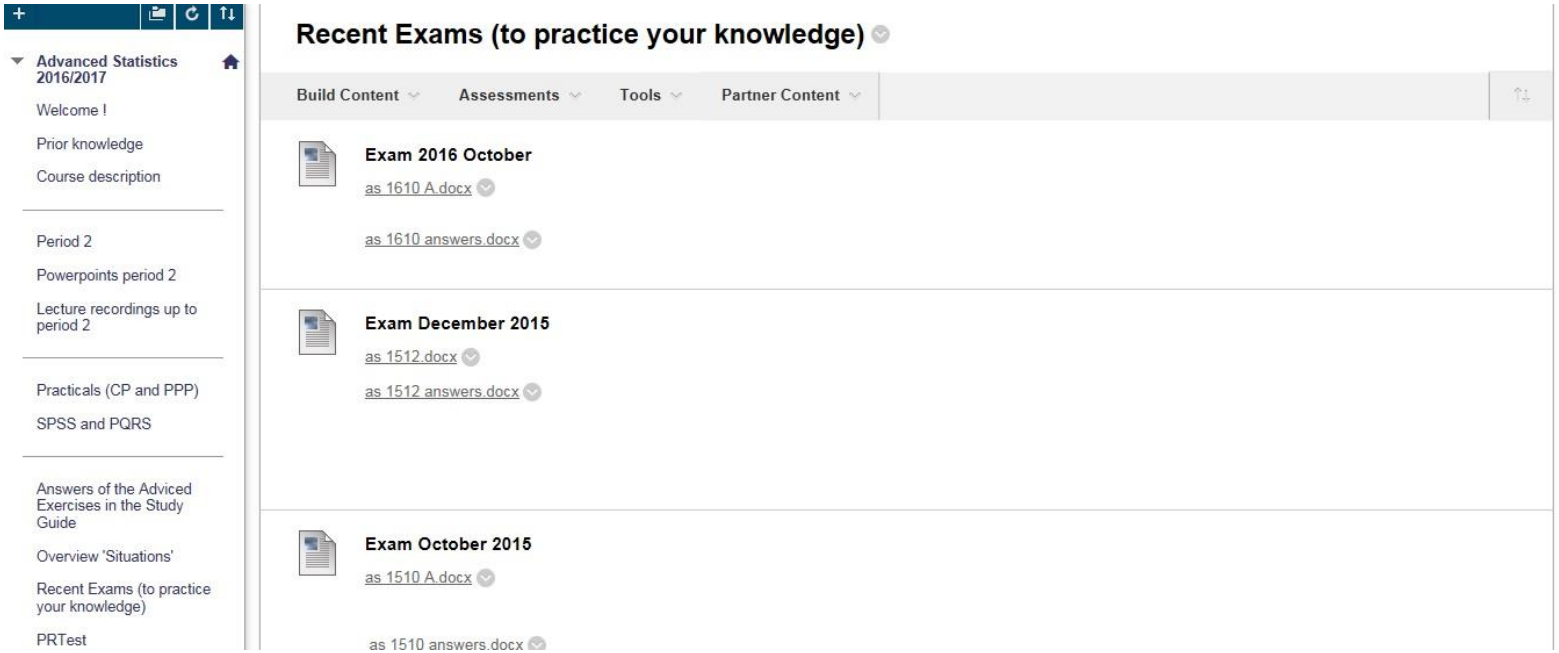
Capaciteit	
Cap. normaal	0
Cap. examen	420

Faciliteiten
Access for disabled people
Furniture Moveable
Microphone

July 5th ** 8.30 – 11.30 ** De Bongerd New Sport Hall

Question hour: July 4th, 11.30 – C3033 Orion

Some EXAM announcements



There will be multiple choice questions (45 points) and open questions (45 points)
Good news: you already have 10 points. In total: 100 points:
Grade: nr. of points divided by 10.

Practice exams can be found on Blackboard.

N correct	≤4	5	6	7	…	…	…	17	18	19	20
Score	0	3	6	9	3 per correct answer			39	42	45	45

Exam

You can bring to the exam:

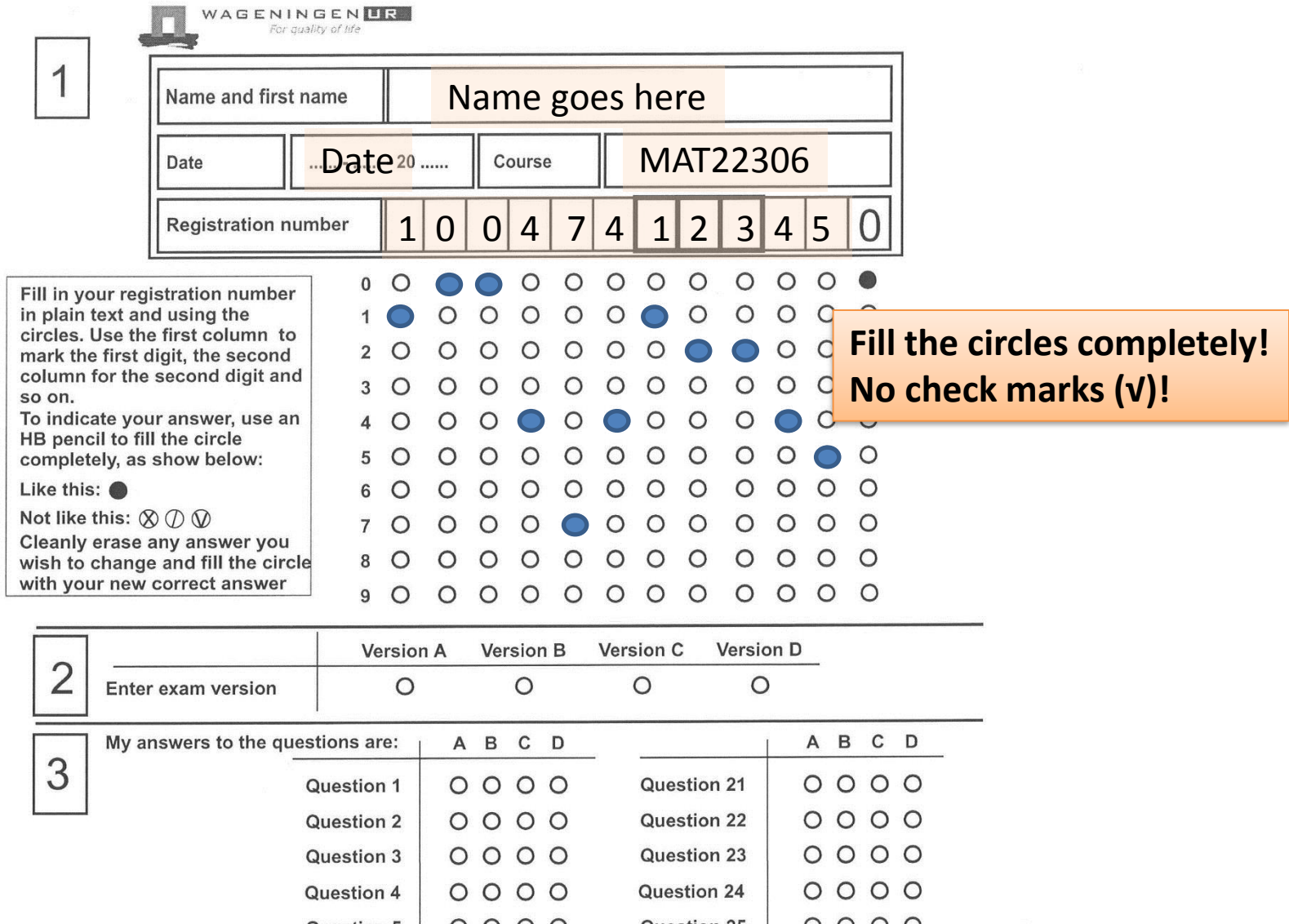
- the book of Ott & Longnecker (O&L),
- the study guide (green) + course scheme from Lecture 12
- a calculator
- a summary (one A4, two sides), in your own hand writing, no printed or photocopied summaries
- a dictionary

Exam

You cannot bring to the exam:

- A telephone / tablet (also not as calculator)
- Handouts of powerpoint slides

Multiple choice form



WAGENINGEN UR
For quality of life

1

Name and first name	Name goes here
Date	Date ..... 20 .....
Course	MAT22306
Registration number	1 0 0 4 7 4 1 2 3 4 5 0

Fill in your registration number in plain text and using the circles. Use the first column to mark the first digit, the second column for the second digit and so on.
To indicate your answer, use an HB pencil to fill the circle completely, as show below:

Like this: ●

Not like this: ⊗ ⊘ Ⓥ

Cleanly erase any answer you wish to change and fill the circle with your new correct answer

	0	1	2	3	4	5	6	7	8	9
0	○	●	●	○	○	○	○	○	○	○	○	●
1	●	○	○	○	○	○	●	○	○	○	○	
2	○	○	○	○	○	○	○	●	●	○	○	
3	○	○	○	○	○	○	○	○	○	○	○	
4	○	○	○	●	○	●	○	○	○	●	○	
5	○	○	○	○	○	○	○	○	○	○	●	○
6	○	○	○	○	○	○	○	○	○	○	○	○
7	○	○	○	○	●	○	○	○	○	○	○	○
8	○	○	○	○	○	○	○	○	○	○	○	○
9	○	○	○	○	○	○	○	○	○	○	○	○

Fill the circles completely!
No check marks (v)!

2

	Version A	Version B	Version C	Version D
Enter exam version	○	○	○	○

3

My answers to the questions are:

	A	B	C	D
Question 1	○	○	○	○
Question 2	○	○	○	○
Question 3	○	○	○	○
Question 4	○	○	○	○

	A	B	C	D
Question 21	○	○	○	○
Question 22	○	○	○	○
Question 23	○	○	○	○
Question 24	○	○	○	○

LET'S NOT DO THIS!

WAGENINGEN UR
For quality of life

1

Name and first name	[redacted]
Date	03-05-2012
Course	MAT 22306
Registration number	[redacted]

Fill in your registration number in plain text and using the circles. Use the first column to mark the first digit, the second column for the second digit and so on.
To indicate your answer, use an HB pencil to fill the circle completely, as show below:
Like this: ●
Not like this: ⊗ ⊘ ⓥ
Cleanly erase any answer you wish to change and fill the circle with your new correct answer

[redacted]

2 Enter exam version: Version A ⊗ | Version B ○ | Version C ○ | Version D ○

3 My answers to the questions are:

	A	B	C	D		A	B	C	D
Question 1	○	○	⊗	○	Question 21	○	○	○	⊗
Question 2	○	⊗	○	○	Question 22	⊗	○	○	○
Question 3	○	⊗	○	○	Question 23	○	⊗	○	○
Question 4	○	⊗	○	○	Question 24	○	⊗	○	○
Question 5	○	⊗	○	○	Question 25	○	○	○	○
Question 6	⊗	○	○	○	Question 26	○	○	○	○
Question 7	○	⊗	○	○	Question 27	○	○	○	○
Question 8	⊗	○	○	○	Question 28	○	○	○	○
Question 9	⊗	●	○	○	Question 29	○	○	○	○
Question 10	⊗	○	○	○	Question 30	○	○	○	○
Question 11	○	○	⊗	○	Question 31	○	○	○	○
Question 12	○	⊗	○	○	Question 32	○	○	○	○
Question 13	○	○	○	⊗	Question 33	○	○	○	○
Question 14	○	○	⊗	○	Question 34	○	○	○	○
Question 15	○	○	○	⊗	Question 35	○	○	○	○
Question 16	○	⊗	○	○	Question 36	○	○	○	○
Question 17	○	○	⊗	○	Question 37	○	○	○	○
Question 18	⊗	○	○	⊗	Question 38	○	○	○	○
Question 19	⊗	○	○	⊗	Question 39	○	○	○	○
Question 20	⊗	○	○	○	Question 40	○	○	○	○

Wrong, Right, right, wrong, right, wrong

Did you fill in all 3 sections?

Stastistics_V4

Good luck at your exam!

After the exam: Applied Statistics

After the exam: Applied Statistics

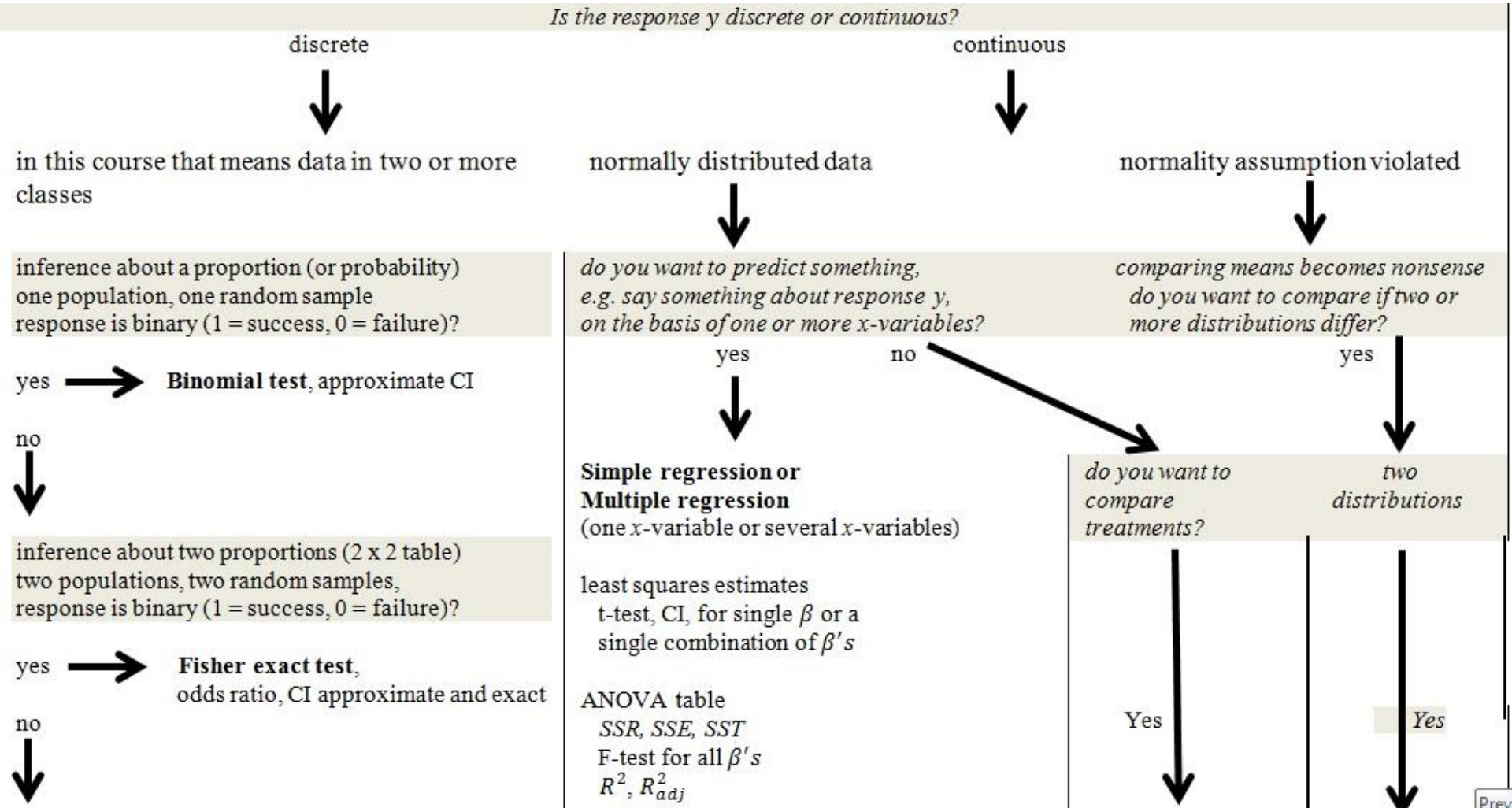
Sir Ronald Aylmer Fisher
(1890 - 1962)

To call in the statistician after the experiment is done may be no more than asking him [or her] to perform a postmortem examination: He [or she] may be able to say what the experiment died of.

Sankhya, the Indian Journal of Statistics, 1938.

https://www.youtube.com/watch?v=reMMNFwJfGc

Overview MAT 20306



inference about more than two proportions
one or several populations, one or several samples, response in classes?

yes

↓

table with one entry, three or more classes K, one population, one sample?

yes → **Chi-square goodness of fit test**

no

↓

table with two entries (contingency table or $r \times c$ table), one population, one sample?

yes → **Chi-square test for independence**

no

↓

table with two entries (contingency table or $r \times c$ table), several populations, several samples?

yes → **Chi-square test for homogeneity**

chi-square distribution for approximate P-value with $df = K - 1$ or $df = (r-1)(c-1)$
often also exact P-value,
in table with one entry from multinomial distribution,
in table with two entries by conditioning on margins

model that is quadratic in x is special case of multiple linear regression: $x_1 = x, x_2 = x^2$

model with dummies is equivalent to one-way ANOVA on the right

F-test for subset of β's;
extra SS, comparing full and reduced models

type I, II *SS*
extra *SS* = type II is safest

checking assumptions:
QQ plot,
residuals against fitted values
also checking for
outliers, high leverage points, influential points

do you want to see how the relation between response y and variable x is affected by different treatments or conditions?

Yes

↓

several lines "hidden" in a single multiple regression model, by use of dummies and products with dummies

parallel lines model same as ANCOVA on the right (but here emphasis is on relationships and

ANOVA
(or ANCOVA)

one factor =
one-way ANOVA

two factors =
two way ANOVA,
with main effects and interaction
(factorial experiment)

use F-tests to decide which table to use for pairwise comparisons between means (Fisher LSD, Tukey)

type I, II, III *SS*

blocking to increase accuracy of comparison between treatments (RCBD vs. CRD)
relative efficiency

do you want to correct for a covariate?

Yes

↓

ANCOVA,
parallel lines,

Wilcoxon sum rank test (indep. data)

OR

Wilcoxon Signed rank test (paired data)

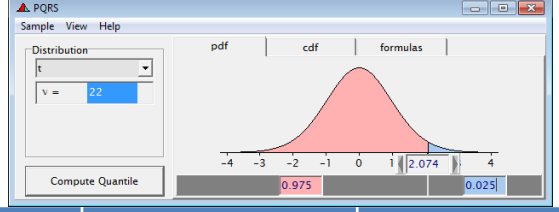
More than two distributions

Yes

↓

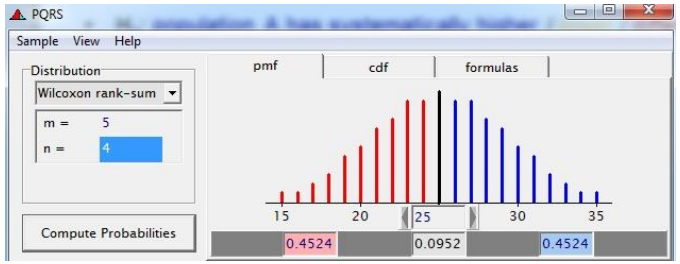
Kruskal–Wallis test

Normality assumption valid !!!



# samples & # variables	We have a research question about:	H_0:	Note:	TS:	Distribution when H_0 is true	1-α c.i.
1 sample 1 variable	Population expected value	$\mu=\mu_0$	σ is known	$z=\frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$	$z \sim z(0,1)$	$\bar{y}\pm z_{\alpha/2}*\sigma/\sqrt{n}$
1 sample 1 variable	Population expected valued	$\mu=\mu_0$	σ is unknown	$t=\frac{\bar{y}-\mu_0}{s/\sqrt{n}}$	$t \sim t(n-1)$	$\bar{y}\pm t_{\alpha/2}*s/\sqrt{n}$
2 samples 1 variable	Difference between two population expected values	$\mu_1-\mu_2=D_0$	$\sigma_1=\sigma_2$ OR $\sigma_1\neq\sigma_1$	$t=\frac{\bar{y}_1-\bar{y}_2-0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $t'=\frac{\bar{y}_1-\bar{y}_2-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	$t \sim t(n_1+n_2-2)$ $t' \sim t(df)$ from SPSS output	$\bar{y}_1-\bar{y}_2\pm t_{\alpha/2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $(\bar{y}_1-\bar{y}_2)\pm t_{\alpha/2}*\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
1 sample 2 variable	Population expected difference	$\mu_d=D_0$	Observations are paired	$t=\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}$	$t \sim t(n-1)$	$\bar{d}\pm t_{\alpha/2}*s_d/\sqrt{n}$

Probability distributions in AS



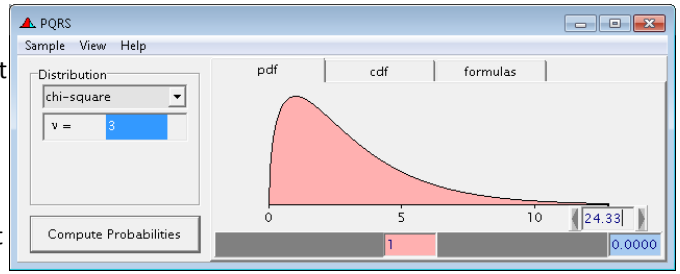
Wilcoxon Rank Sum Test

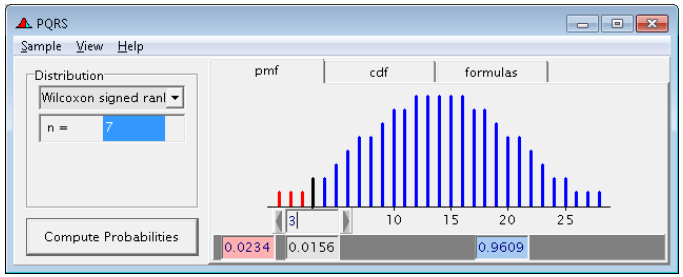
Goodness of Fit Test

Independence Test

Homogeneity Test

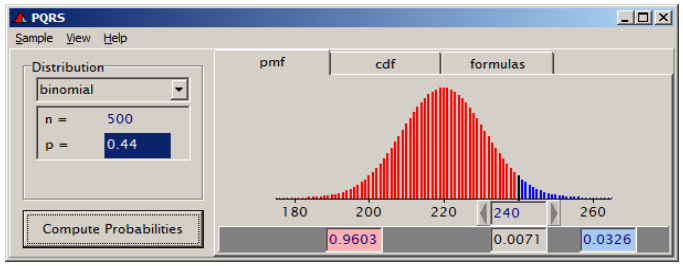
Kruskal-Wallis Test





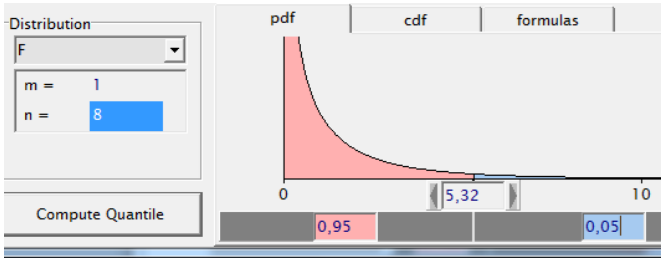
Wilcoxon Signed Rank Test

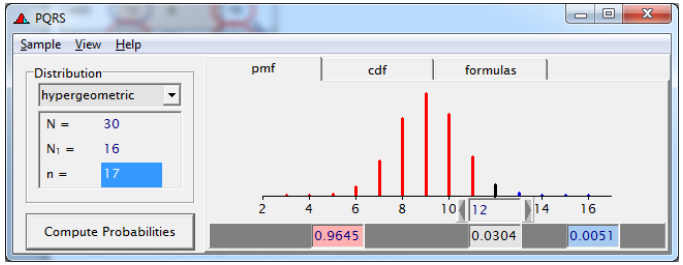
Embedded in GLM model



Binomial Test

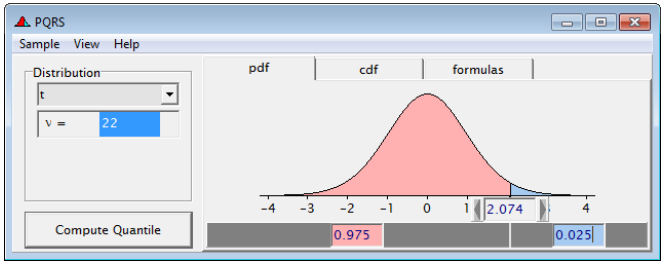
Test for ratio's of Mean Squares to asses the predictive value of a given model





Fisher Test

Test for difference between population / treatment means



Summary terminology

Factor:	controlled variable (qualitative or quantitative),
Factor levels:	the levels of the factor in the experiment.
Treatments:	combinations of the factor levels used in the experiment.
Blocks:	groups of similar experimental units, entered in the model to reduce the error (residual) variance.
Experimental units:	units to which treatments are (randomly) assigned.
Measurement units:	parts of experimental units where response is measured.
Replications:	repetitions of treatments, i.e. several experimental units receiving the same treatment.
Control treatment:	(1) to monitor experimental conditions, or (2) standard method for comparison, or (3) placebo.
Covariate:	quantitative variable *x* measured along with *y*, e.g. in the model to reduce the error variance σ_ϵ^2.

Overview GeneralLinearModel in AS

Multiple linear regression: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$

One way ANOVA: $y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Factorial ANOVA: $y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \epsilon_{ijk}$

ANCOVA: $y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \epsilon_{ij}$

GLM with nonparallel lines: $y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \lambda_i x_{ij} + \varepsilon_{ij}$

Follow up on this course

ABG-30806 Modern Statistics for the Life Sciences

Studiegids

Vakrooster

Week	Start - Eind	Opleiding (Allemaal)	Type	Zaal
2016/2017 February re-exams				
24	2017-02-15 8:30 - 11:30	MAS MBF MPS WUPBR	Resit Exam	
2016/2017 Period 5				
29	2017-03-20 13:30 - 15:15	MAS MBF MPS WUPBR	Lecture	C0525
29	2017-03-20 15:30 - 17:15	MAS MBF MPS WUPBR	Practical	PC0713 combi
29	2017-03-21 13:30 - 15:15	MAS MBF MPS WUPBR	Lecture	C0525
29	2017-03-21 15:30 - 17:15	MAS MBF MPS WUPBR	Practical	PC0713 combi
29	2017-03-22 13:30 - 15:15	MAS MBF MPS WUPBR	Lecture	C0525
29	2017-03-22 15:30 - 17:15	MAS MBF MPS WUPBR	Practical	PC0713 combi
29	2017-03-23 13:30 - 15:15	MAS MBF MPS WUPBR	Lecture	C0525
29	2017-03-24 13:30 - 15:15	MAS MBF MPS WUPBR	Lecture	C0525
30	2017-03-27 13:30 - 15:15	MAS MBF MPS WUPBR	Lecture	C0525
30	2017-03-27 15:30 - 17:15	MAS MBF MPS WUPBR	Practical	PC0713 combi
30	2017-03-28 13:30 - 15:15	MAS MBF MPS WUPBR	Lecture	C0525
30	2017-03-28 15:30 - 17:15	MAS MBF MPS WUPBR	Practical	PC0713 combi

MAT-50303 R for Statistics

Studiegids

Vakrooster

Week	Start - Eind	Opleiding	Type	Zaal
2016/2017 Period 4				
25	2017-02-20 8:30 - 12:15		Practical	PC0089
25	2017-02-21 8:30 - 12:15		Practical	PC0089
25	2017-02-22 8:30 - 12:15		Practical	PC0089
25	2017-02-23 8:30 - 12:15		Practical	PC0089
25	2017-02-24 8:30 - 12:15		Practical	PC0089
26	2017-02-27 8:30 - 12:15		Practical	PC0089
26	2017-02-28 8:30 - 12:15		Practical	PC0089
26	2017-03-01 8:30 - 12:15		Practical	PC0089
26	2017-03-02 8:30 - 12:15		Practical	PC0089
26	2017-03-03 8:30 - 12:15		Practical	PC0089
27	2017-03-06 8:30 - 12:15		Practical	PC0089
27	2017-03-07 8:30 - 12:15		Practical	PC0089
27	2017-03-08 8:30 - 12:15		Practical	PC0089
27	2017-03-09 8:30 - 12:15		Practical	PC0089
27	2017-03-10 8:30 - 12:15		Practical	PC0089

Universiteit Leiden

Statistical Science

for the Life and Behavioural Sciences
and
Data Science

Leiden University Mathematical Institute
Leiden University Medical Center
Leiden University Institute of Psychology
Leiden University Institute of Advanced Computer Science
Wageningen UR
VU Medical Center

Statistics Quote
The best thing about being a statistician is that you get to play in everyone's backyard. - John Tukey

Home

The master program *Statistical Science for the Life and Behavioural Sciences* was initiated by a group of statisticians from different institutes at Leiden University together with applied statisticians from Wageningen University and Research Center. The program was meant to fill the educational gap in Statistics caused by the increasing demand of applied statisticians and data scientists, and the scarcity of applied statistics master programs in the Netherlands.

News and Events

Videos!
We recently filmed some of our alumni to talk about how they look back on the the master program and ...

more...

Information Session Data Science 26th of May
If you are interested in Data Science, then find us on the 26th of May for an information session about ...

more...

Spinoza Prize for Leiden statistician Aad van der Vaart!

Structure MSc Statistical Science

The Statistical Science: Life & Behavioural Sciences specialization is structured as follows:

Core Programme	EC
Statistics and Probability	9
Mathematics for Statisticians	3
Statistical Computing with R	3
Linear & Generalized Linear Models and Linear Algebra	9
Multivariate and Multidimensional Data Analysis	6
Bayesian Statistics	6
Mixed and Longitudinal Modelling	6
Statistical Consulting	5
Advanced Statistical Computing	3

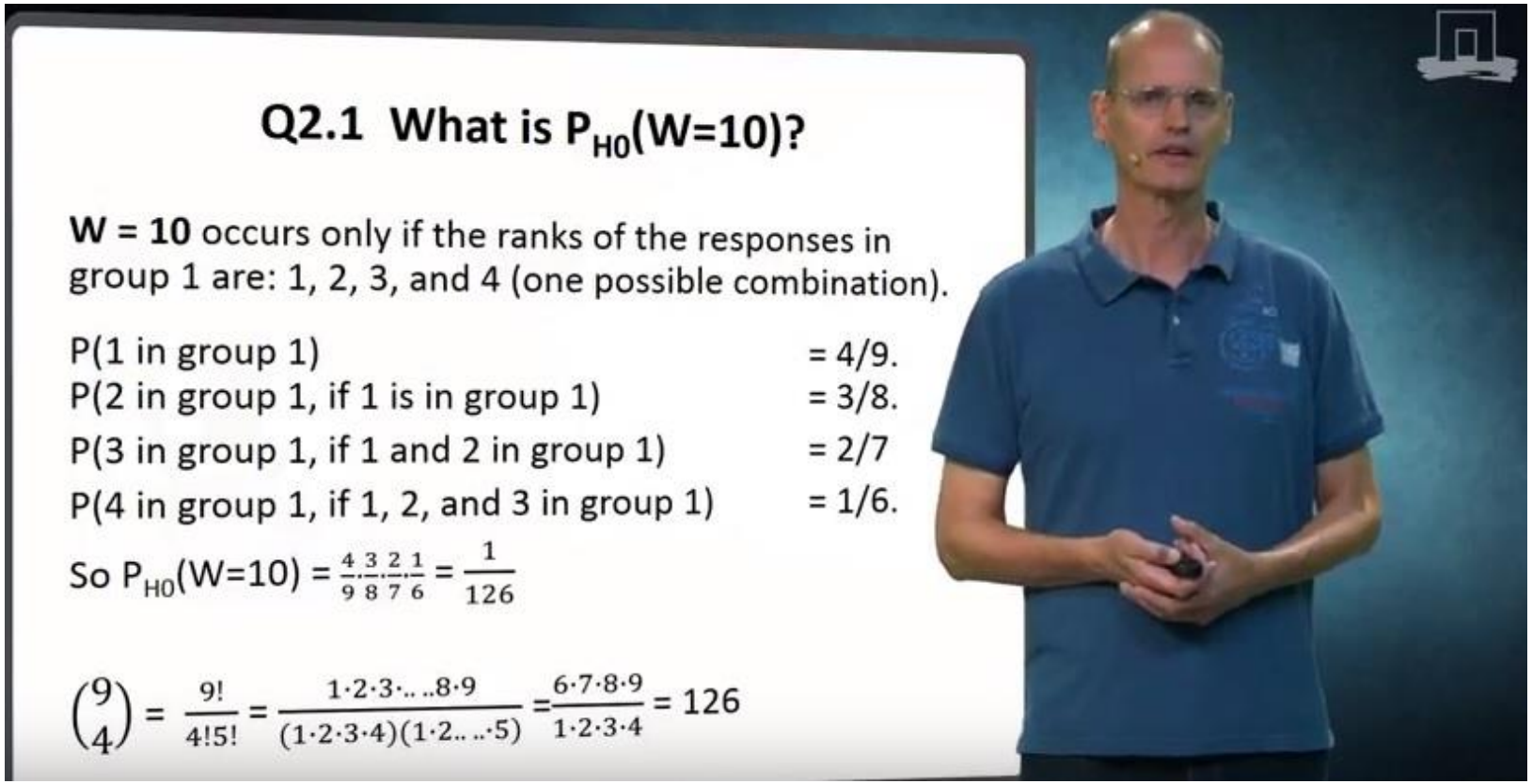
Specialisation Courses	EC
Study Designs in the Life and Behavioural Sciences	6
Introduction to the Life and Behavioural Sciences	3
R for the Life and Behavioural Sciences	3

Elective Courses (choose 4 out of 7 courses)	EC
Statistical Learning Theory	6
Psychometrics and SEM	6
High-Dimensional Data Analysis	6
Statistical Genetics	6
Survival Analysis	6
Optional Course 1	6
Optional Course 2	6

Internship, Master Thesis	EC
Internship	10
Master Thesis	24

Together this makes a total of 120 ECTS. Find more information about our other specialization here.

Evert Jan presents...



Inferential Statistics:

a love song

By Evert Jan Bakker

Biometris, Wageningen UR

Inferential Statistics, required everywhere.

How to test hypotheses, that knowledge is quite rare

But Biometris' statisticians like to share the news

Until all students understand which test they have to choose

And then go through the testing steps, what fairly easy is:

A small p-value proves, - the research-hypothesis

Statistics is disliked, contempt the science's fate
"With Stats you can prove anything", explains the people's hate
But rules for design and subsequent analysis
Makes that Stats a guardian of objectivity is

T-procedures at the start: four elements are key
The **population parameter** you wanna know, you see
And then its **estimator** : the method that you choose
to guess for the parameter which value we can use
The **standard error** and **df**, both needed for t-test
This beauty through simplicity is whyyyy we like it best

Statistics is a science field, that generates much fear
Students and researchers both get nervous when they hear:
You must redo statistics - for article or test
Biometris' help is then appreciated best

ANOVA splits df and - total sum of squares
Into factor components and F-tests make us aware
Which factor associates or influences the response -
Then **pair-wise means comparisons**, all with their pro's and con's
Above Tukey, and Bonferroni, LSD 's liked 'best' ,
Because it shows significance more eeeeeasy than the rest

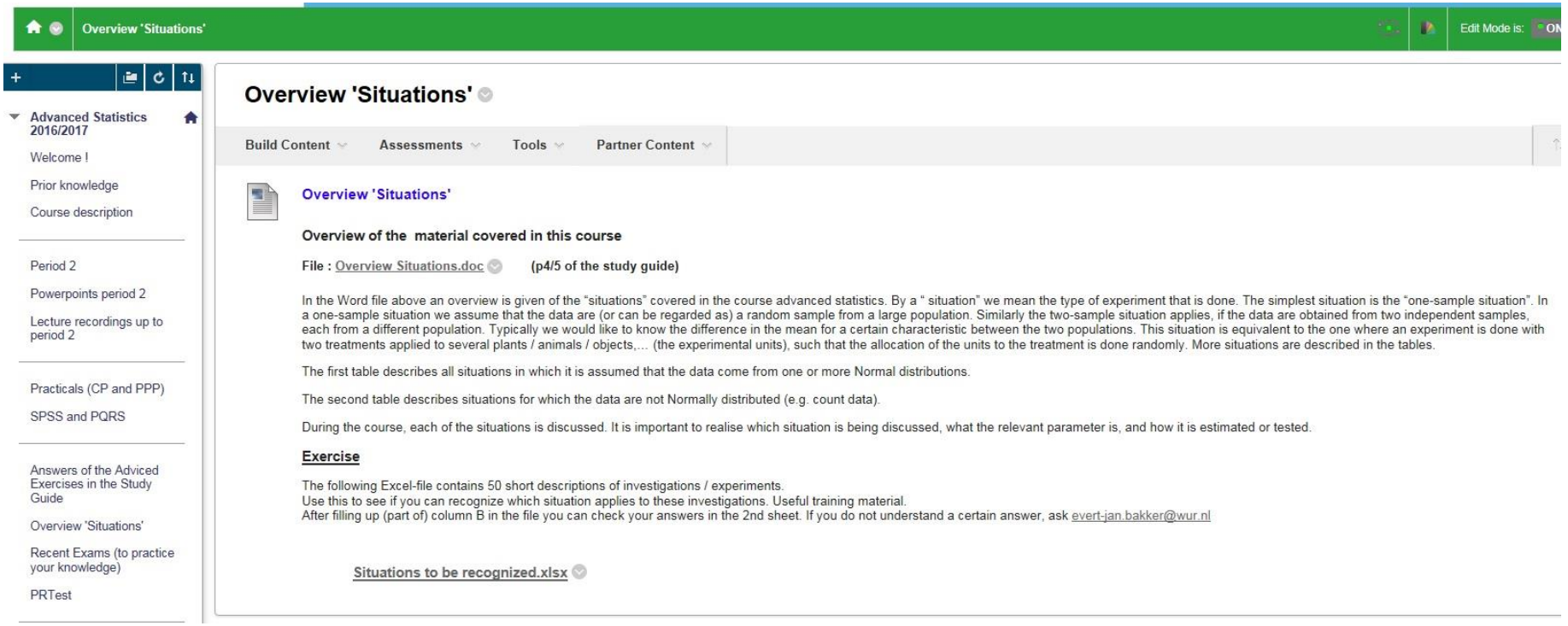
A song on Stats without **Least Squares**: a flowerless vase
This estimation principle stays often in the haze
But from one-mean to GLM it makes us arrive at
the estimators that we know, like y-bar , beta-hat.

Quantitative factor, dose-response relationship,
Or simply correlation: “use regression” is the tip
These models lead to formula’s for beta´s and se´s
Too difficult to handle without modern-day pc’s.
Can they, too, describe the means for treatments, more than two?
Yes: with the use of dummies they incluuuude ANOVA too

To know the why of methods is the statisticians' brew,
Apply with understanding is what students have to do :
when to use which model, - and know the weaker spot(s)
write up the results and what one may conclude and not

Inferential statistics, required everywhere
There are more complex models - to be applied out there
Pseudo replicates can in Mixed models find a place
In Generalized LM's - is Non-Normal a normal case
For all these diff´rent models there's a course that can be done.
They're taught by Biometris and should beeee a lot of fun

Blackboard material



Exercises

Below there is a short description of experiment / investigation

Influence on the rate of reproduction for catfish of 4 different water temperature regimes

The difference in expected weight of calves at birth between cows that are outside in summer and those that remain in the stable all year.

Difference in fraction of women that is unemployed between US and Netherlands

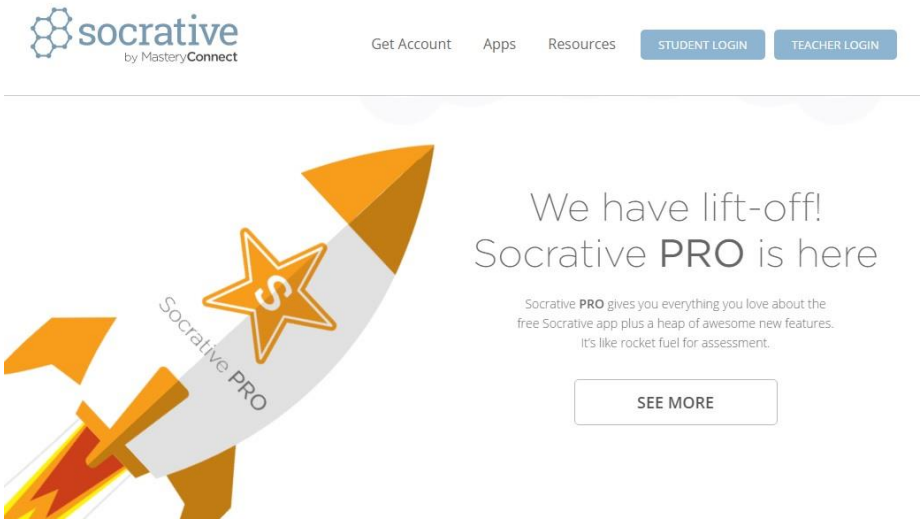
Percentage damaged tomatoes and the amount of pesticide used during the growth period.

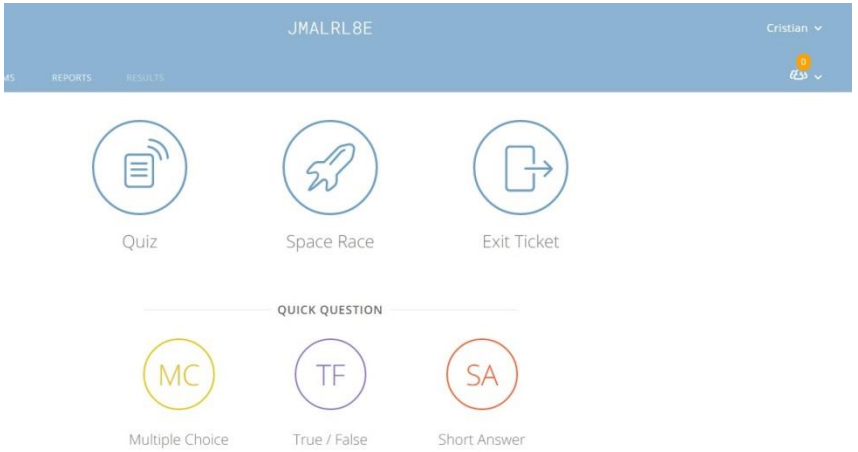
Differences in mean income b/w salesmen in NL, Belgium, Luxemburg and 4 types of trade

Growth of young trees treated with different quantities of manure either protected or not protected from goats by wire

Test if the fraction of alfa, beta and gamma students in Groningen, Amsterdam, Utrecht are similar, based on a random sample of 100 students from each city.

Solutions to this exercises you can find in the previously indicated excel file on the blackboard





1. Influence on the rate of reproduction for catfish of 4 different water temperature regimes.

A: 4 way (factorial) ANOVA
B: one way ANOVA
C: ANCOVA
D: simple linear regression
E: can not decide

2. The difference in expected weight of calves at birth between cows that are outside in summer and those that remain in the stable all year.

A: paired t-test
B: Wilcoxon signed rank test
C: 2 independent samples t-test
D: Wilcoxon sum rank test
E: can not decide

3. Difference in fraction of women that is unemployed between US and Netherlands

A: binomial test
B: paired t-test
C: one sample t-test
D: Fisher Test
E: can not decide

4.Percentage damaged tomatoes and the amount of pesticide used during the growth period

A: chi squared for goodness of fit
B: multiple linear regression
C: one way ANOVA
D: simple linear regression
E: can not decide

5. Differences in mean income between salesmen in NL, Belgium, Luxemburg and 4 types of trade

A: two way ANOVA
B: 2 independent samples t-test
C: one way ANOVA
D: ANCOVA
E: can not decide

6. Growth of young trees treated with different quantities of manure either protected or not protected from goats by wire

A: two way ANOVA
B: simple linear regression
C: one way ANOVA
D: ANCOVA
E: can not decide

7. Test if the fraction of alfa, beta and gamma students in Groningen, Amsterdam, Utrecht are similar, based on a random sample of 100 students from each city

A: chi squared for homogeneity
B: chi squared for independence
C: chi squared for goodness of fit
D: Fisher test
E: can not decide

Do not forget: It is just another language !

Time to say good bye ☺