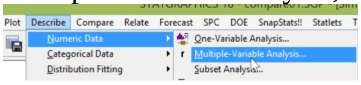
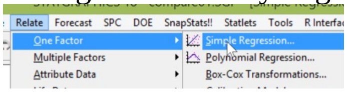
Homelearning Course: Statistics SommerSemester 2020

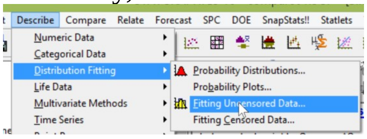
Dear colleagues
What remains to be done is to explain the intercorrelation patterns in the separate soils and compile everything (intercorrelations and differences) to conceptual models suitable for discussion and prae-modelling..
As we have allready learned, any correlations must be performed inside, within, the clearly distinct groups of samples as evaluated by anova, the so called "homogenuous groups", in order to avoid fals true correlations. You may see how this is done in the statgraphics comp1.SGP. The are named: multiple variable analysis, and you find them in this part of the menu:



the regression lines you get here:



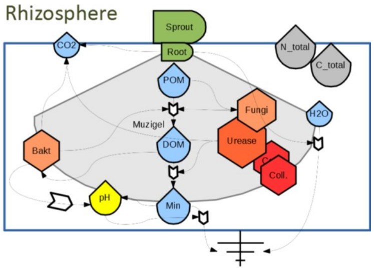
If you have not a sound reason, stick to the linear model. If in doubt, partition a potentially multi-part function in its linear parts, and avoid the data at the changing points (broken stick sections, knees).
Please read up in textbook chapter: 4. and 5.
So far we have walked the paved lane of straightforward "parametric and linear" statistics and complete datasets, but now for the "swamp":
Alas, data sets are rarely complete and seldomly evenly distributed. If you want to test for normality, do it here:



As far as normality and homogeneity of variance are concerned, there is curiously enough, due to tradition only, just a single LS, 95%. Because of this, we realise that most datasets are not suitable for parametric test, and all test we used so far are parametric tests, de lege artis requiering normality, variance homogeneity and equal dataset length.
What people unfortunately tend to advertise is
1. heavy logarithmic data transformation
2. data smoothing with splines,
3. t-Test filtering that kicks out badly behaing variables
4. filling missing values
5. using discriminant funtions to enhance separation

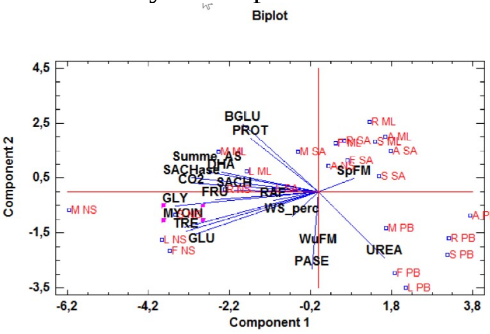
All this should be avoided.
If you have unlinear incomplete and unnormal data, resort to nonparametric tests. Concerning anova, the method of choice is the Kruskal Wallis Test that you find in the one way anova tables, and for multivariate situations, like metabolome or species diversity analysis, resort to the non Parametric dimensional scaling (NMDS, MDS), which is available in R and in Primer. If you managed to install primer 6, which I will not include in this course, open the compare1.pwk in it. I have displayed an example for this method on p.11 in the using02_ pdf. Please read up on that in chapter 3.4.1, unbiased description.
The last task and homework I want you to try: draft an Odum model for the 4 soils in the compare01 example including: WS_perc, WM, SG, PC1_Carbon, PC2_Remin, and Summe_ENZ(for faunal/microorganismal activity). Use just 3 Symbol sizes: small, middle, large,

according to the anova results in compare01.SGP and translate correlation strengthes in 3 line thicknesses (weights).
You find the template in Odum_conc_mod01.odp on the course websites root directory. Use it with LibreOffice only, it will missbehave in Powerpoint!



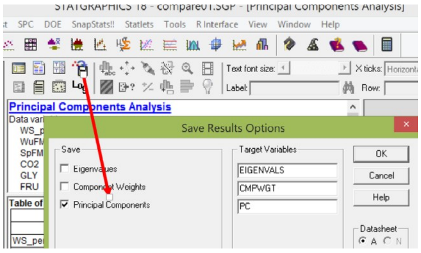
So this is it, we shall wrap it up here, there were priorities to set in these times, I appologise for this. Well, thank you for joining me in this homelearning course, I hope you learned something, you guys take good care!
Of course I am available for any questions you may have now, and also for data evaluation consulting if you need that later on. Please leave an evaluation of my booklet in amazon if you feel like it.
Sincerely yours
Gert

Am 26.03.2020 um 09:55 schrieb Gert Bachmann:
Dear Participants,
Let us proceed in the PCA. Double dlick on the 2D biplot to enlarge it. You may click/drag some labels if they overlap.



We have allready seen that a centering and scaling is necessary to include many variables with different maxima in one graph or in a comparative calculation. In the PCA graph, we have to get familiar with the fact that all variables are plotted as vectors in a 3D space, and some vectors bundles that have a high correlation on each other, are directed to the left, right, top or buttom, regardless oft the fact that their real variable values are never negative. The negative values in the vector space simply express negative (indirectly propotional) correlations!
Due to the scaling, all vectors feature a similar length as mesured from the central zero point to their end. But, as they point in different directions, they may seem to be shorter or longer in a 2D projection. There are also 3D graphs, and you may rotate them, but they are not very useful in presentations as they tend to play tricks on our eyes: if looked at them for a bit, the corners tend to "jump" from front to back and vice versa.

It has to be emphasised that the correlations (small angles between vectots) and the clustering (vicinity of data points) are useful for exploration of what might correlate and where differences may be found due to which variable, but: all this has to be quantified for proper publication.
The correlations are done with intercorrelation matrices, plese see the multivariable analysis below the PCA procedures folder, and by simple regression lines for detailed bivariate observations.
In order to quantify the clustering, the principal components themselves need to be stored in the datasheet A:


STATGRAPHICS 18 - compare01.SGP - [Principal Components Analysis]
SPC DOE SnapStats!! Statlets Tools R Interface View Window Help
Text font size:
X ticks: Horizont
Label:
Row:
Principal Components Analysis
Data vari
WS_p
WuFM
SpFM
CO2
GLY
FRU
Table of
WS_pe

Save Results Options

Save
□ Eigenvalues
□ Component Weights
☑ Principal Components

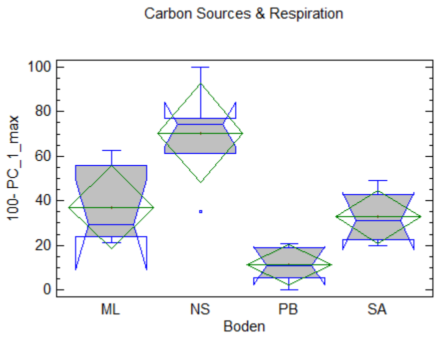
Target Variables
EIGENVALS
CMPWGT
PC

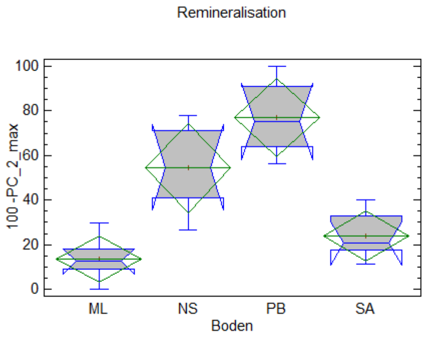
OK
Cancel
Help

Datasheet
● A ○ N

For anova purposes, the centering needs to be reversed by:
- adding the minima and adding a very small number (again to avoid zero), and
- the pca coordinates need a restandardisation for the maxima, as they are arbitrary anyhow and percentual changes in the values need to be discussed.

Please open up the compare1.ods, tab PC_Eval, to see how this is done in two steps, employing semi-fixed adressations in the formulas.
These decentered and rescaled PCs are then imported in the statgrahics datasheet (p.) and may be processed by means of one way anova (see using02 p. 7-9), as we did it previously in the frog pond example.
Now it is a good moment to give the PC appropriate names! In this example it is quite straightforeward: the PC1 we shall name: Carbon Sources & Respiration, whereas PC2 is aptly named Remineralisation. Now we may discuss these ecosystem processes. So we manged to pinpoint and calculate synergetic processes in four quite diverse soil ecosystems.





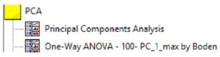
see you in an hour!
best regards Gert

Am 26.03.2020 um 08:14 schrieb Gert Bachmann:
Dear Participants,

Again I have actualised all files in the folder "Boden", so please download it again, as well as the equally actualised presentation using02_statgraphics.pdf. Also I corected the most horrible typos in yesterdays mail below .

Today, I would like to start by teaching you how to work with principal components and how to quantify cluster separation by means of anova, so please

- open comp01.SGP in Statgraphics (the actualised one!) and open the PCA procedure in the PCA folder



PCA
Principal Components Analysis
One-Way ANOVA - 100- PC_1_max by Boden

- open using02_statgraphics.pdf
  refer to p.3

See you in a minute!
Gert

Am 25.03.2020 um 17:49 schrieb Gert Bachmann:
Dear Collegues,

I have actualised the using02_statgraphics.pdf as well as the comp01.SGP. For any multivariate method, we need to select the variables and subject them to a prea-treatment in order to make them displayable in one single graph. This does not mean to transform them, which I woul strongly caution against - in a first exploration.

In the first page of using02, I show you how to select the procedure from the menu and in p.2 you see the steps of selcting variables and options.

Now, the necessary prea-treatment is also called Z-transformation, and as you see in the help section, this corresponds to standardizing each input variable before calculating the covariances, by subtracting its mean and dividing by its standard deviation. This procedure moves the mean to the zero line, and makes all original values fitting between -10 and +10. Multiply that result by a factor of 10, this will be a centering and % of the max standardization.

In the tables and graphs choice dialogue, we concentrate on the so called biplot, that plots the variable vectors in the correlation space above the scatterplot of the new xyz coordinates obtained by the covariance matrix space. In p.5, you may see how the labelling of the points is done. Please note in p.1 that I have combined the factors for soils and plants in order to make the recognition of the single datapoints easier.

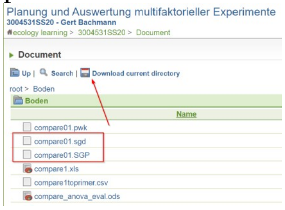
The datapoint groups are also called clusters, and the density of the clustering gives you a picture of the multivariate variance of datapoints belonging to a sample. The centers of these clusters are called centroids. As these Clusters are distanced from each other, we want to now the reason for it. This is where the variable vectors come in. The closer their vicinity, the higher their correlation - which has to be quantified by - you will allready now it - a regressison analysis. The distance of the centroids may also quantified, by anova. To that end the principal components need to be stored. See p.4 how to do that. You find these PC stored in the end of our data sheet in the comp01.sgd data file.

Please observe that all PB-soil datapoints, PB (Purkersdorf, Vienna Forest) beeing an acidic brown soils with very little nitrogen content, is separated from the other soils by the urease activity, meaning the urea-deamination activity (you may also call it NH4 remineralistion process). Makes sense, does n´t it? Also the rich black soil from Nickelsdorf (NS) is set apart by its high content on sugars and the high CO2 development, as to be expected...

A discriminant analysis (DA) modifies the original dataset in order to enhance the separation of the clusters due to the variables which produce the highest separation of the clusters. Such a discrimint function you may see on p.9 of the using02 pdf. As only the most discriminating variables determine the clustering, this is also the name for the Analysis, and therefore this is a method for diagnostics, for differential diagnostics, in fact, to find so called markers. It is not a tool for system similarity description!
Well, this wraps it up for today!
best regards, Gert
Am 25.03.2020 um 16:10 schrieb Gert Bachmann:
Dear Participants,
Let us proceed now. In the datamatrix for comp1, the anova revealed a very dynamic situation that is not easy to discuss. In such a dataset situation, a top down approach may serve. Clicking trough the anova windows, we may observe that CO2 development of the soil, sugar conten of the soil an overall enzme acticity, regardless of the obvious trophic connections, exhibit very different results. So this is a good moment to employ the next to the boxplot second very common exploratory approach that is called PCA.
Principal component analysis aims to combine all originally meaured parameters (variables) to 2 ore more metavariables according to a vector addition based on the covariance and on the correlation pattern. These metavariables are called principal components.
Please open the compt.SGP file and also
http://131.130.57.230/clarotest190/claroline/backends/download.php?url=L3VzaW5nMDJfc3RhdGdyYXBoaWNzMTgucGRm&cidReset=true&cidReq=3004531SS20
In this second pdf you find the sequence of how to select the method, select the numeric variables, the options incl. Z-standardisation, choosing the appropriate graphs, mainly the biplot consisting of scatterplot (factor scores) and component plot (factor weights) and the method to label the datapoints.
back in an hour
Gert

Am 25.03.2020 um 11:40 schrieb Gert Bachmann:
Dear All,
The minutes got a bit more, sorry folks! Okay then: please download again: compare1.ods as I completed the documentation in the tab COMP1_Documentation.
Some of the variables are named Summe_x and we shall use them to demonstrate that a mathematival sum is not the synergismal effect of its single components.
The first tab COMP1_Urdat has two parts. Whereas the upper one is not modified, the lower one has no more zero values, I replaced them bz verz small value, as the zero values would make correlations impossible. Please read in the textbook Chapter 3.2.2. about missing data filling The method to replace them,e.g for cell K25, is to use a small formula:
=WENN(K25>0,000001;K25;0,000001)
This so modified part of the Table I imported into Statgraphics. So without further ado, please make a double klick on compare01.sgp. This takes a bit longer, as several procedures are included. Most of them are one way anova by the factor soil (Boden). This is followed by a PCA (principal component analysis), our second most important exploratory procedure. Laterin the list come correlation matrices, and a discriminat analysis (DA).
What is a PCA? Please read up in: textbook chapter 3.4. This is yout task for today. I shall walk you through the PCA in the early afternoon!
Gert

Am 25.03.2020 um 09:23 schrieb Gert Bachmann:
Good Morning Everybody,
Today, I want to introduce you to the world of multivariate data Analysis. To that end, we need another small dataset. In this case, it is not compiled / fabricated from different sources, but an aggregated version of the data matrix of my ancient phD thesis:
https://www.sciencedirect.com/science/article/abs/pii/003807179290079D
please download the data here: again it is best to download the current directory as a zipfile.



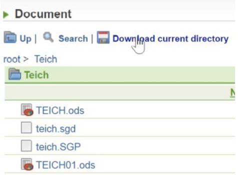
Planung und Auswertung multifaktorieller Experimente
3004531SS20 - Gert Bachmann
ecology learning > 3004531SS20 > Document

Document

root > Boden
Boden

Name
compare01.pwk
compare01.sgd
compare01.SGP
compare1.xls
compare1toprimer.csv
compare_anova_eval.ods

Open the file compare1.xls to have a look at the data. They represent results for soil respiration, metabolomics and enzymatics of 4 soils and 6 plants in factorial combination, with the aim of assessing plant/ soil preaadaptations and modifications.
see you in some minutes!
Gert
Am 24.03.2020 um 15:38 schrieb Gert Bachmann:
Dear All,
Hope you could reproduce the steps I suggested. However , please do download the actualised files again:

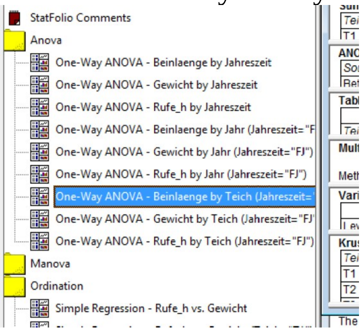


Document

root > Teich
Teich

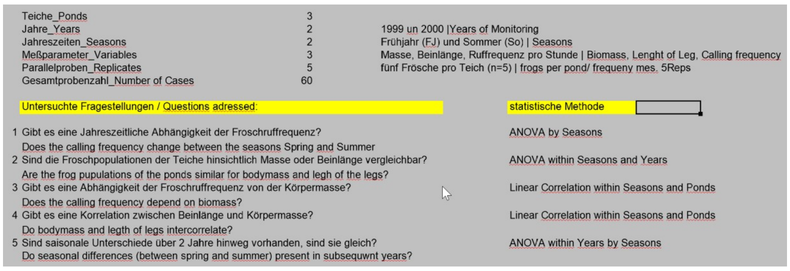
N
TEICH.ods
teich.sgd
teich.SGP
TEICH01.ods

You can see that you may organise your many anylyses in Folders.



StatFolio Comments
Anova
One-Way ANOVA - Beinlaenge by Jahreszeit
One-Way ANOVA - Gewicht by Jahreszeit
One-Way ANOVA - Rufe_h by Jahreszeit
One-Way ANOVA - Beinlaenge by Jahr (Jahreszeit="F
One-Way ANOVA - Gewicht by Jahr (Jahreszeit="FJ")
One-Way ANOVA - Rufe_h by Jahr (Jahreszeit="FJ")
One-Way ANOVA - Beinlaenge by Teich (Jahreszeit="
One-Way ANOVA - Gewicht by Teich (Jahreszeit="FJ"
One-Way ANOVA - Rufe_h by Teich (Jahreszeit="FJ")
Manova
Ordination
Simple Regression - Rufe_h vs. Gewicht

Before we continue, it is a good moment to remember what we set out to ask and learn from the experiment! If memory serves, we did jot that down in the first tab of the original datafile TEICH.ods:

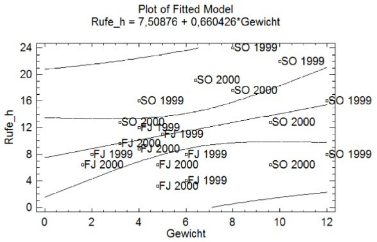


Teiche_Ponds	3	
Jahre_Years	2	1999 un 2000 |Years of Monitoring
Jahreszeiten_Seasons	2	Frühjahr (FJ) und Sommer (So) | Seasons
Meßparameter_Variables	3	Masse, Beinlänge, Ruffrequenz pro Stunde | Biomass, Lenght of Leg, Calling frequency
Parallelproben_Replicates	5	fünf Frösche pro Teich (n=5) | frogs per pond/ frequeny mes. 5Reps
Gesamtprobenzahl_Number of Cases	60	

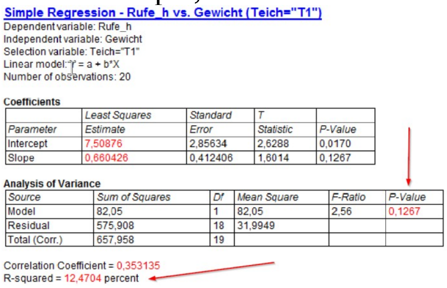
Untersuchte Fragestellungen / Questions adressed:	statistische Methode
1 Gibt es eine Jahreszeitliche Abhängigkeit der Froschruffrequenz? Does the calling frequency change between the seasons Spring and Summer	ANOVA by Seasons
2 Sind die Froschpopulationen der Teiche hinsichtlich Masse oder Beinlänge vergleichbar? Are the frog pupulations of the ponds similar for bodymass and legh of the legs?	ANOVA within Seasons and Years
3 Gibt es eine Abhängigkeit der Froschruffrequenz von der Körpermasse? Does the calling frequency depend on biomass?	Linear Correlation within Seasons and Ponds
4 Gibt es eine Korrelation zwischen Beinlänge und Körpermasse? Do bodymass and legth of legs intercorrelate?	Linear Correlation within Seasons and Ponds
5 Sind saisonale Unterschiede über 2 Jahre hinweg vorhanden, sind sie gleich? Do seasonal differences (between spring and summer) present in subsequwnt years?	ANOVA within Years by Seasons

so far, we can answer questons 1, 2, 5 and we have still to elucidate 3 and 4.

Now, as we allready know, there are huge differences between spring and summer. In such a case, any correlation would be highlighting only the smaller values for spring and the hig values for summer, connecting these two clusters by a line, and giving a false positive information. Th following correlation graph for Rufe on Gewicht, Reserch Question No 3, (for Teich1/pond 1 only) shows just that:



Thus, we just learned a very important rule for ordinstion (correlation analysis): If differences exist between parts of a data matrix, any correlation analysis must be done within that part of the matrix.
the table beside graph contains two essential informtions:
R-squared, the coefficient of determination, highlights the STRENGTH of the correlation in %
in this case P is 0.1267, meaning the Level of Signifivance is 87.4%
P-Value highlights the error probability of the correlation.
in our example, this is 12.4%



Simple Regression - Rufe_h vs. Gewicht (Teich="T1")

Dependent variable: Rufe_h
Independent variable: Gewicht
Selection variable: Teich="T1"
Linear model: Y = a + b*X
Number of observations: 20

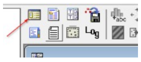
Coefficients

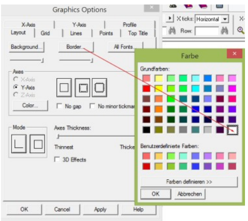
	Least Squares	Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
Intercept	7,50876	2,85634	2,6288	0,0170
Slope	0,660426	0,412406	1,6014	0,1267

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	82,05	1	82,05	2,56	0,1267
Residual	575,908	18	31,9949		
Total (Corr.)	657,958	19			

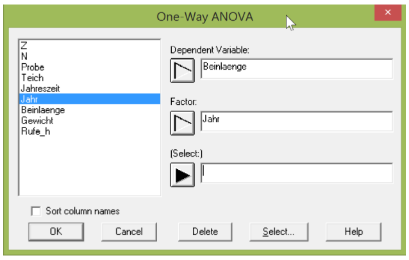
Correlation Coefficient = 0,353135
R-squared = 12,4704 percent

We learn a second rule: any statistical information is only RELEVANT, IF the LS is appropriately high (depending on the field of reearch between 50 and 99.999%) AND if either the correlation strenghth is appreciable (> 10 % r2) OR the difference between medians is reasonably high, eg 10% for matabolomics data.
Now, as a home work No 4, look at all the results and answer the Research questions 1-5 in a concise text: half a page wthout graphs, if any (not required, but recommended).
see you later!
Gert
Am 24.03.2020 um 11:43 schrieb Gert Bachmann:
Dear Students,
I am a bit concerned that many of you did not reply as yet. Please let me know how I may help?
I dare say you successfully completed the first steps in Statgraphics and have tried to do this "one way anova" for the two remaining variables by the factor Jahreszeit (seasons). If you have the same issue like me with the dysfunctional "repeat for" command, use copy instead and use the button "Procedure input":
 in order to switch to the two remaining variables.
Next make yourselfes familiar with the graphics formatting options (remove the gray background) and the analysis modifications (see using01_statgraphics.pdf, p10-13). You obtain them by making a right click on the respective sub windows/subprocedures/graphs.



Once you are happy with a graph (to meet e.g. the requirements of a journal), use the right click and "copy", and you will be able to paste the graph e.g. in ms-powerpoint, where you may ungroup it into its single parts and edit it as need be.
If you copy a procedure in the left part /list of statgraphics, all formating is copied as well, so this is how you make templates for later.
So far, we tackled the complete table, but since the differences between the seasons are that big, we need to select parts of the matrix to answer the remaining research questions. Why? because these big seasonal differences would make the smaller differences between the years and between the three ponds invisible, as they feature much smaller variation. Thus we need to select either the Frühjahr (spring) or the summer part of the matrix.

Make a copy of the first anova, click on the procedure input button and click in the select section: as soon as you do that, the transform button will be exchanged by a select button!
We woul like to look for annual differences in spring now, so we do modify the input of our copy and choose:

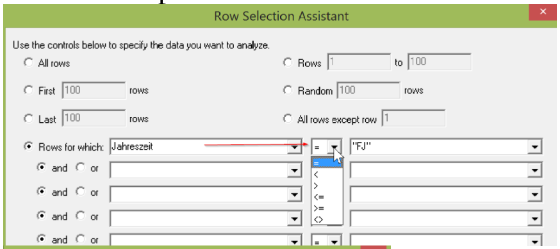

One-Way ANOVA

Dependent Variable: Beinlaenge

Factor: Jahr

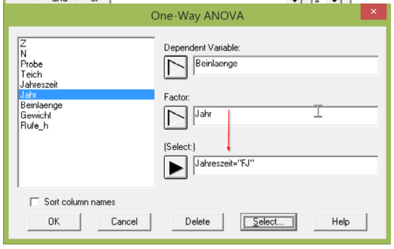
(Select:)

now clicking on that select button opens the database filter section of Statgraphics, allowing you to choose very specific parts of the matrix.
In our example we choose all rows for which the factor Jahreszeit equals FJ (spring).


Row Selection Assistant

Use the controls below to specify the data you want to analyze.

○ All rows
○ Rows 1 to 100
○ First 100 rows
○ Random 100 rows
○ Last 100 rows
○ All rows except row 1
● Rows for which: Jahreszeit = "FJ"
● and ○ or
● and ○ or
● and ○ or
● and ○ or


One-Way ANOVA

Dependent Variable: Beinlaenge

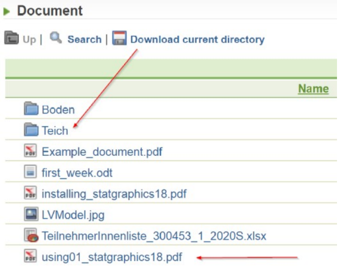
Factor: Jahr

(Select:) Jahreszeit="FJ"

Make the same graph/procedure for Summer. Then, after having evaluated the differences for the years, do it for the ponds in spring and in summer.
catch you in two hours!
Gert

Homelearning Course: Statistics SommerSemester 2020

Am 23.03.2020 um 11:49 schrieb Gert Bachmann:
http://131.130.57.230/clarotest190/claroline/document/document.php?cmd=exChDir&file=L1RlaWNo&cidReset=true&cidReq=3004531SS20



Document

Up | Search | Download current directory

Name

Boden

Teich

Example_document.pdf

first_week.odt

installing_statgraphics18.pdf

LVModel.jpg

TeilnehmerInnenliste_300453_1_2020S.xlsx

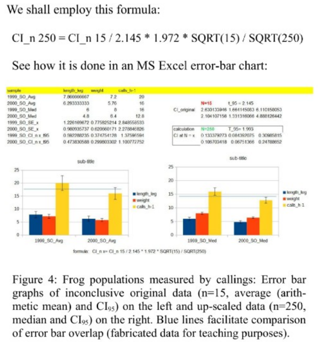
using01_statgraphics18.pdf

what to download

Am 23.03.2020 um 11:45 schrieb Gert Bachmann:
Dear Participants
Good morning and I wish you a good start in the second half of our home learning course.
Excel and libreoffice are all very well for ascertaining the basic requirements (sufficient n, differences occuring where they were part of the research question) , but after basic procedures it makes much more sense to switch to a better toolset in order to
- explore
- quantify
- deduce a mechanistic model
in our data matrix.
- please download the complete folder Teich again: click/open drectory, choose the button "Download current directory" (you get a file named: ecology learning.3004531SS20.Teich.zip) and extract all to one and the same folder - TEICH
- download using01_statgraphics18.pdf and have a thorough look at it
1. start your pc and excel after having changed the and number settings tu US english
2. start Libreoffice, then Statgraphics, first close all other apps excl. LO to avoid clipboard issues
3. start Statgraphics
4. mark all relevant content (the data matrix) in LO: go to A1, next use shift |crtl| end (umsch |Strg | Ende), and copy it to the clipboard
5. In Statgraphics, go to the left and upmost cell, make a right click there and choose paste, next select column names.
check weter all your variables are labeled: numeric and wether all your numbers behing the comma are visible - if not, you still have a language settings issue
6. perform the first Anova as given in the pdf: using01_statgraphics18 on page 8 (I give you details later)
I will be with you again in the afternoon.
best regards
Gert

Am 23.03.2020 um 09:59 schrieb Gert Bachmann:
Dear All,
To answer the questions that arose concerning last weeks homework (that just 3 of you tackled so far....)
a) which formula should be used to upscale from an insufficient number of replicates to an appropriate one **and/o**r a higher level of significance (e.g. 95% to 99%)?
- referring to the textbook p.29 (see below), this is the correct formula, if upscaling from n=15 to

n=50:
CI_n 50 = CI_n 15 / 2.145 * 2.011 * SQRT(15) / SQRT(50)
where 2.145 is the two-tailed t value for LS_95, DF 14, and 2.011 the t-value for LS_95, DF 49.
- there are, as the variance (within *and* between) the samples is presumed to remain constant ! just two variables to be included (and exchanged in the formula): the t-value and the se_x = s/sqrt(n).
We can do this because the se_x we employ is not hypothetical but coming from actual data obtained by a preliminary experiment. Thus, the variance between the samples may be kept constant as well.
b) is this graphical approach the best /only one?
- for all practical purposes, imho the best one, yes, if you have the data from a small scale preliminary experiment, because you can see what you are doing, and there is not the risk of wrong assessment of the variance *between* the samples. You just have to extrapolate and interpolate to find the correct n with a 10% safety margin and to fuzz around with the error bars: In some cases you need to delete them and make them again
- obviously, not the only one! today, I shall introduce you to the approach taken in Statgraphics today, and, as a reference, I attach to this mail the page with the original approach from Sachs (1992) - you see why I recommend the graphical approach, but since you wanted to know... ;=))
c) please look up again teich01.ods | Graph02_FJ_Jahre to see
- the solution for your homework, where you upscale for the appropriate n on the same LS of 95,
- whereas on the example on tab Graph01_Seasons, we had to find the right n if LS is to be changed from 95 to 99.
- please observe: the formula in the textbook refers to an other question concerning the length of the legs, hence the difference in numbers...
I have to emphasise that mastering this procedure is one of the most important skills needed to plan an experiment economically in terms of YOUR time and money!
Please ask again if in any doubt! I presume everybody ha installed Statgraphics now.
I will be with you within an hour!
best regards Gert



We shall employ this formula:

CI_n 250 = CI_n 15 / 2.145 * 1.972 * SQRT(15) / SQRT(250)

See how it is done in an MS Excel error-bar chart:

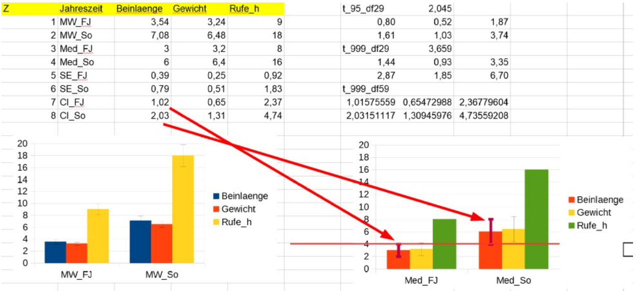
sample	length_leg	weight	calls_h-1
1999_SO_Avg	7.866666667	7.2	20
2000_SO_Avg	6.293333333	5.76	16
1999_SO_Med	6	8	16
2000_SO_Med	4.8	6.4	12.8
1999_SO_SE_x	1.226169672	0.775825214	2.848558533
2000_SO_SE_x	0.980935737	0.620660171	2.278846826
1999_SO_CI_n x_t95	0.592288235	0.374754128	1.37596594
2000_SO_CI_n x_t95	0.473830588	0.299803302	1.100772752

	N=15	t_95 = 2.145	
CI_original	2.630133946	1.664145083	6.110158053
	2.104107156	1.331316066	4.888126442
calculation	N=250	T_95= 1.993	
CI at N = x	0.133379273	0.084392075	0.30985815
	0.106703418	0.06751366	0.24788652

formula: CI_n x= CI_n 15 / 2.145 * 1.972 * SQRT(15) / SQRT(250)

Figure 4: Frog populations measured by callings: Error bar graphs of inconclusive original data (n=15, average (arithmetic mean) and CI_{95}) on the left and up-scaled data (n=250, median and CI_{95}) on the right. Blue lines facilitate comparison of error bar overlap (fabricated data for teaching purposes).

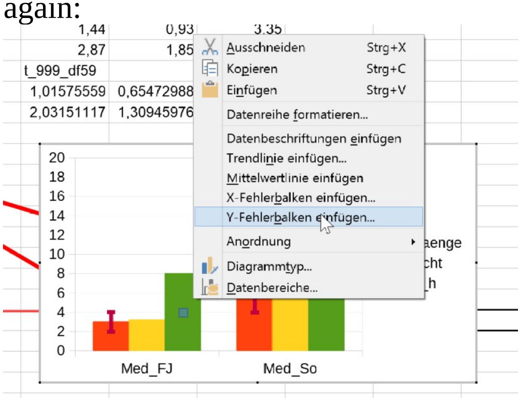
p.29 from the textbook
Am 21.03.2020 um 15:24 schrieb Gert Bachmann:
Dear all,
please have a look at this snapshot of teich01| graph01_seasons this, so it becomes clear what I meant with the last homework: CI and recalculated CI are ploted over medians. next weak, I will show you how to do that replicate number assessment in statgraphics, the pdf is allready available for download.



Z	Jahreszeit	Beinlaenge	Gewicht	Rufe_h
1	MW_FJ	3,54	3,24	9
2	MW_So	7,08	6,48	18
3	Med_FJ	3	3,2	8
4	Med_So	6	6,4	16
5	SE_FJ	0,39	0,25	0,92
6	SE_So	0,79	0,51	1,83
7	CI_FJ	1,02	0,65	2,37
8	CI_So	2,03	1,31	4,74

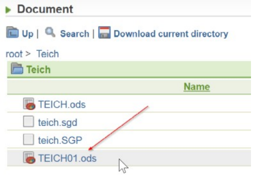
t_95_df29	2,045	
0,80	0,52	1,87
1,61	1,03	3,74
t_999_df29	3,659	
1,44	0,93	3,35
2,87	1,85	6,70
t_999_df59		
1,01575559	0,65472988	2,36779604
2,03151117	1,30945976	4,73559208

also I use this version of LO Version: 6.3.5.2 (x64), please install the same, just in case. often the errorbars do not readjust when the respective input cells are edited, so you may have to insert them again:

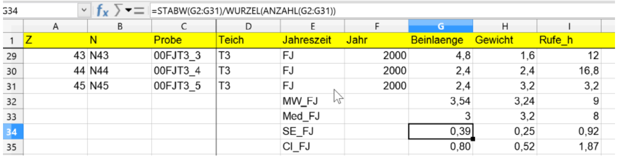


best regards
Gert

Am 19.03.2020 um 16:02 schrieb Gert Bachmann:
Dear participants,
Hoping that you are with me so far and you have all been able to access files and install Statgraphics, I want to "walk" you through basic statistics within the Teich example with the aid of this apt tool for screenshots:
https://getgreenshot.org/downloads/
To that end, please download the next data file: teich01.ods from the ecology learning server:



Open it, and proceed to the 3rd table named "Calculate_Seasons_allYears". Scrolling to the middle part as well as to the end, you see that 4 rows have been inserted, respectively:

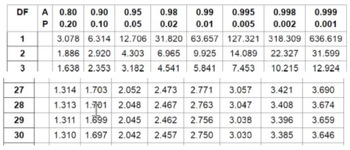


G34 =STABW(G2:G31)/WURZEL(ANZAHL(G2:G31))

	A	B	C	D	E	F	G	H	I
1	Z	N	Probe	Teich	Jahreszeit	Jahr	Beinlaenge	Gewicht	Rufe_h
29	43	N43	00FJT3_3	T3	FJ	2000	4,8	1,6	12
30	44	N44	00FJT3_4	T3	FJ	2000	2,4	2,4	16,8
31	45	N45	00FJT3_5	T3	FJ	2000	2,4	3,2	3,2
32					MW_FJ		3,54	3,24	9
33					Med_FJ		3	3,2	8
34					SE_FJ		0,39	0,25	0,92
35					CI_FJ		0,80	0,52	1,87

in the editing area you may see how the formula for the standard error of the mean is put together. We would like to compare results for arithmetic mean with standard error of the mean errorbars to median and confidence interval errorbars. These later values should be utilized as a general rule, as for evenly distributed datasets the average does have the same value as the median, and only not overlapping CI (confidence intervals) let you deduce that there is a significant difference for e.g. the

95% level of significance. Of course, as CI_95%=SEmean *t_95%, we need that student-t value (2.045) from a table for two tailed distributions:
https://www.medcalc.org/manual/t-distribution.php



DF	A P	0.80 0.20	0.90 0.10	0.95 0.05	0.98 0.02	0.99 0.01	0.995 0.005	0.998 0.002	0.999 0.001
1		3.078	6.314	12.706	31.820	63.657	127.321	318.309	636.619
2		1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3		1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
27		1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28		1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29		1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30		1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646

For a total of 30 cases, as we have it here, the DF (degrees of freedom) is 29.
Please klick into the cells G32:G35 and learn the syntax of the formula for mittelwert (average), median, SE_x and CI. You may see immediately that the median differs greatly from the average, if, as in this table and most data matrices you will ever see, is not exactly normally distributed. also, SE_x is nearly half as small as CI! Still, most errorbar graphs employ average and SE_x, for traditional und quite unfathomable reasons.
Next you need to copy values without formulas and get rid of all that you do not need for the graphs. this is done in the table: Graph01_Seasons, where you may also find the two graphs for Avg/SE_x and Median/CI. Compare them to see that the second graph lets you assess significant differences, but not the first one!
Why plot the seasons first without caring about annual diferences? Please look at table1:Dokumentation: It is the frst priority reserch question, so we need that info asap! On the right hand graph, use the red line to move it to see which CI actually overlap. Obviously, the No of replicates ist just big enough in this case for LS(level of significance)_95%, but not with a 10% safety margin that would be needed, and certainly not for LS_99%.
So, we have to assess the appropriate no of replicates in an practical way, that I show you in the same table just above the right hand graph. Remember how SE is calculate: SE_x = STDEV/SQRT(No_replicates), and CI = SE_x*t_LSxx%. So, we correct CI for 99% in this way by retrieving the appropriate t-value and calculate as this : CI_99% = G2/2,045*3,659 .
Please utilise these new CI and see the consequence it has on the overlap of the errorbars by copying the obtained values in G5:I6 to C8:E9.
To find the appropriate No of reps and "make the error bars smaller again" to avoid overlapping, there is just one way: mathematically enlarging (upscaling) the No of replicates! This was done in G9:I9. Again, copy the new CI to C9:E9 and observe the impact on sample separation.
Now you know how to upscale from an insuffient No of replicates in a preliminary experimentto a main experiment with enough reps. And you will never again stammer at a progress report something like " there is no significancebut a slight trend may be there" but you will confidently put forward: "By upscaling from n=x to n=y significanve may be obtained for this and that...!"
Try, as a homework to apply this to the assessment of appropriate No reps (n) in the last table of this file, Graph01_FJ_Jahre. What will be the appropriate n for not overlapping CI_95 for all 3 variables between those two years? Please send me the file! Also, please read up on chapters 2.2 incl 2.3 in the textbook!
Tomorrow, a brief intro in statgraphics will be due, and btw, there is a good oneline help available:
http://www.statgraphics.com/how-to-guides
best regards
Gert
p.s.

Am 19.03.2020 um 10:29 schrieb Gert Bachmann:
Dear Students,

Homelearning Course: Statistics SommerSemester 2020

In the past units we have talked a lot about research questions, broad and narrow, and the hypotheses (that can be proven right or wrong, i.e. verified or falsified). Once the hypotheses have been agreed on, the eperimental approach (Versuchsansatz) may be drafted: what experiments employing which materials, locations, analysis methods may be employed to evaluate (veri- or-falsify) the hypotheses? Once this is in the clear, the experimental setup (sequence of experiments) may be designed.

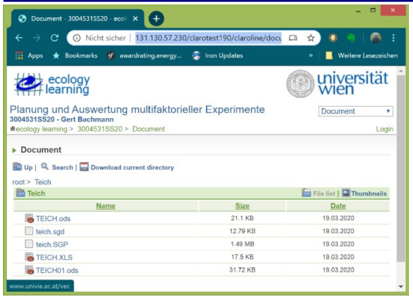
The experimental setup includes allready all factors (locations, seasons, species, treatments), variables (measuring parameters), as well as the necessessary number of replicates, i.e. repetitions (biological and analytical) of the samples.

The experimental setup mayand should be utilised to draft a template for the resulting data matrix (table of measurement results). This data matrix (I am not using n:n databases here), that my be later on imported in any statistical or graphing tool (like statgraphics), must obviosly be compatible to the repective import methods:

the basic rules are: names of factors and variables in the first row, sample names in the leftmost column, no empty rows or columns, no mathematical expressions, blancs or german umlauts or any special characters in the variable names-

The sample matrix I want to utilize to introduce you to the first steps is the teich example. Since I modified and translated it a bit, please download again here, please use the open database spreadsheat format .ods for all following procedures

http://131.130.57.230/clarotest190/claroline/backends/download.php?url=L1RlaWNoL1RFSUNILm9kcw%3D%3D&cidReset=true&cidReq=3004531SS20



Planung und Auswertung multifaktorieller Experimente
3004531SS20 - Gert Bachmann
ecology learning > 3004531SS20 > Document

Document

root > Teich

Teich

Name	Size	Date
TEICH.ods	21.1 KB	19.03.2020
teich.sgd	12.79 KB	19.03.2020
teich.SGP	1.49 MB	19.03.2020
TEICH.XLS	17.5 KB	19.03.2020
TEICH01.ods	31.72 KB	19.03.2020

Again look at the first page, that qualifies as a dokomentation and repetition of hypotheses.

The next steps are: doing basic statistics (averaging, assessing variation) and graphing them, and later: assess weather the number of replicates are sufficient, and, if not, assessing the appropriate number of replicates. You should learn to be able to do that in libre office. This will be done in teich01.ods, you my download it allready and have a look at it.

After that, it is best to switch to the Statgraphics. Only after having mastered basic statistics in Excel/Libreoffice, you may savely proceed to Statgraphics, where you can play araund a lot and faniliarize yourself with methods and their applcability. Only after having accquired some experience in such a "sand box", you may venture to optimize your work in e.g. R or Mathlab, because there you allready need to be focuse and know exactly what you are doing.

I will get back to you in the afternoon.

best regards

Gert

Am 19.03.2020 um 07:44 schrieb Gert Bachmann:

Dear All,

Two questions arose concerning the Statgraphics intallation:

1: which email adress shoul I employ, if the primary adress spelled without "unet." is not accepted?

Homelearning Course: Statistics SommerSemester 2020

please refer to this resource: https://www.univie.ac.at/ZID/uaccount-aktivierung/ to get a u:account along with the email adress in the proper format "name.surname@univie.ac.at" as required for the Statgraphics activation. It is not necessary to register in the Statgraphics Website.
2: is it necessary to install R right now? no, it is not, I shall explain later how to to that
cu later
Gert
Am 18.03.2020 um 17:02 schrieb Gert Bachmann:
if in doubt, this is the serial no
BHB0-DB0A-00E8-YK0E-4EM0
Am 18.03.2020 um 17:00 schrieb Gert Bachmann:
Dear Participants,
In order to make first steps easier and to meet data safety demands, I attached a short installation guide for statgraphics - please give it a try and mail me once it was done or ask if in need of further instruction.
we shall work with http://131.130.57.239/scripten/EDV/table/TEICH.XLS next. Look at its first table and review the similarities to homework No1.
best regards Gert
...eagerly awaiting your hw

Am 17.03.2020 um 12:13 schrieb Gert Bachmann:
Dear Students,
Good Morning to you on a still chaotic Tuesday! I trust you found all the links and I shall now write a few words concerning the appropriate usage of basic and advanced Statistics, following up on http://131.130.57.239/scripten/EDV/Ausw_excel_pckursE.pdf
The first page highlights the common sequence of gathering data, storing them in spread sheet tables or databases, and performing priliminary testing. as the saying goes, you may start
- "bottom up" by comparing mean/average and variation of samples for the most impotant parameters, referred to as variables that are different due to factors that were changed in the experiment. The most common approach is a bar graph with errorbars, emloying the arithmetic mean (average), and the standard error of the mean. In essence, this is allready explorative statistics on a very basic level. Obviously, there are better test at our disposal, eg. the box plot (slide 11, 12).
- Or, you may start "top down", performing a multivariate ordination: PCA (principal component analysis), CA (cluster analysis, DA(discriminant analysis) and decide later wich variables need or merit quantification by means of ANOVA (Analysis of Variance). I want to state quite early on that *this is very risky, and subject to a lot of pitfalls* to the uncritical user.
Anyhow, both approaches are part of **exploratory statistics**, apt to explore waht differences, intercorrelations and clustering may be found in your dataset. As a rule, it is recommandable to start with Anova, with a boxplot, to be precise, in order to assess which factors separate the data matrix, and which variables have been changed in a relevant way (There is much more to be said about relevance in contrast to significance).
Do not skip that step because / if you have measured/obtained a lot of variables, do not simply or blindly apply any filters as you may be tempted to rule out the wrong parameters. Also, refrain from applying any transformation, just because supperiors prompt you to do it because "it is allways done".
Please have a look at the table on the slide No 6, see how it is organised - that is a template for easy importation in any statistics package.
We shall start with ANOVA tomorrow, so please download and install Statgraphics 18
http://131.130.57.239/scripten/EDV/Software/StatgCent18/

in the process oft installing it, you need to provide sensible info along with the swerial number found in the link above (this is a campus license) and an email as follows: userid@univie.ac.at, meaning to skip the unet. part of your adress.Do not close the activation window and check you email for the activation key which is applicable for your PC only, valid for one yar, after which you can extend the activation for yet another year. Paste the activation key and you will be in the clear. Additionally download the first data table to play around with and learn all about anova and appropriate number of reeplicates:
http://131.130.57.239/scripten/EDV/table/TEICH.XLS
best regards
Gert
Am 16.03.2020 um 16:30 schrieb Gert Bachmann:
Dear Participants,

As this morning proved to be rather exeptional for all of us, I come back to you just now to proceed with our course. I trust you have done as suggested in the last mail, read up on project planning strategies and had a look at the ppt sysecol0. Some of you have allready given a try at a short summery of an actual or hypothetical or future research project as suggested in slide No 1. It would like to emphasise: this is not an idle excersise: you are going to need that for talks, project process reports, and defensio purposes.
The next slides in sysecol0 are graphical presentations and project descriptions for different purposes, and slide 6 introduces a symbol language introduced by Howard tom Odum allready in 1971. I highly recommend to obtain this book: https://www.amazon.de/Environment-Power-Society-Twenty-First-Century/dp/0231128878/

slide 2: a simple summary of a research question for intruductory or project submission purposes

slide 3: summary of all interactions in a soil system, fit for textbooks utterly unsuitable for a project intro

slide 4: a concise graph to highlite a research programme to get a grant for funding, as later pulished and cited a a few times: https://scholar.google.com/citations?user=7kPaN-4AAAAJ&hl=de#d=gs_md_cita-d&u=%2Fcitations%3Fview_op%3Dview_citation%26hl%3Dde%26user%3D7kPaN-4AAAAJ%26citation_for_view%3D7kPaN-4AAAAJ%3Au5HHmVD_uO8C%26tzom%3D-60

slide 5: same project, in the process of determining the actual layers of complexity in order to focus on an actual experimental design

slide 7: basic interaction scheme, also called a conceptual model, in order to include the crucial parameters or key parameters: key pools, key organisms, key processes and key flows in order to come of with the concept for a mechanistic model

slide 8: model including the methods envisioned to cover all these parameters

9, 10: the summarizing model of some research on elevated athmospheric CO2, featuring the impact of two different plants: ryegrass, building up a lot of roots, clover, interacting with soil organisms by root exudation of sugars and more

11: basic effects of elevated CO2 for specific soil pH: acidic: loss of minerals via leaching, alcaline: immobilisation of minerals by precipitation as carbonate and neutral: relative decrease of N and thus decrease of proteins, resulting in poor biomass quality

12: the comparison of conventional land use: recycling system with soil formation, and modern, petrochemistry bases land use: degradation of soils and loss of soil biodiversity

that is it for today!

Now, as a further homework, try to come up with a graph of your own, employing odum symbols if applicable and also send it as an email. You may just photograph a hand drawing or employ any drawing tool of your preference.

Please use the keyword: statss20 in the Betreff/subject

As from tomorrow, we shall have a systematic approach to statistics by emphasising the respective purposes: Evaluation of the experimental design, exploratory statistics, quantifying statistics, preparation of modelling. Please download

http://131.130.57.239/scripten/EDV/Ausw_excel_pckursE.pdf

and read up in the textbook:all icluding chapter 2.2: Organisation of data matrices

best regards

Gert

p.s. you may ask any questions in german or in english, but please du submit homeworks in english

-

Am 13.03.2020 um 08:49 schrieb Gert Bachmann:
Dear Participants,

Please download

http://131.130.57.239/scripten/EDV/OEKO_Exp_Mod/sysecol0.ppt

and read chapters 1. to 2.1 in the Textbook "Critique of pure statistics"". After having done that, please give it at try and do homework No 1, as described in chapter2, strategy plan, and on the first slide of the ppt sysecol0.

I shall send more on Monday morning

best regards

Gert

Homelearning Course: Statistics SommerSemester 2020

Am 3/11/2020 um 9:45 AM schrieb gert.bachmann@univie.ac.at:
Home learning 300453-1 VO+UE Planung und Auswertung multifaktorieller Experimente (2020S)

Dear participants,

Due to the closing of Vienna University in order to prevent further COVID19 spreading, I shall also have to modify the practical class that was to start next monday morning. please do refer to

1. the course web site

http://131.130.57.230/clarotest190/claroline/course/index.php?cid=3004531SS20

2. purchase / download the booklet that covers the content

https://www.amazon.de/Critique-Pure-Statistics-Sciences-English-ebook/dp/B076351TF4/ref=tmm_kin_swatch_0?_encoding=UTF8&qid=&sr=

there is a lot of online reading options without owning any kindle hardware

https://www.amazon.de/Amazon-Digital-Services-LLC-Download/dp/B07N82TR9Q/ref=dp_ob_title_vg

3. await furter information on Friday

best regards

Gert