Open questions

Question	Answer
A binomial distribution is based on data values which have only two states	T
A correlation of 0.95 for 4 pairs of data is significant.	F
A high F-value of a regression model implies that the model is correct.	F
A leptokurtic distribution is a flat distribution (compared to normal distribution)	F
A Mann-Whitney U-Test can be applied to ordinal data	T
A negative correlation coefficient indicates that data are uncorrelated	F
A parabolic curve fit can be obtained by linear regression	T
A representative random sample exhibits the same distribution as the population	T
A t-distribution quickly becomes normal if the sample size increase.	T
According to “DEStatis” the mean net income of German employee in 2008 was 1772 Euros (per month). This implies that 50% of all employees earned less than 1772 Euro.	F
According to “Statistic Austria” the mean income of the approx. 4 mill. employees in 2009 was 23600 Euros (gross per year). This implies that 50% of all employees earned more than 23600 Euros.	F
An excellent quadratic relationship of two variables always Pearson correlation coefficient close to 1.0	F
ANOVA tests for equal variance	F
Assume that in the course of a year you measure your blood pressure every morning. You are calculation the weekly and monthly means form the daily measurements. Please tick off which of the following statements are correct of wrong: The standard deviation of the daily measurements is less than the standard deviation of the weekly averages The daily measurements show a uniform distribution The monthly means are approx. normally distributed	 F F T
Blood groups define a variable on a nominal scale	T
Categorical variables exhibit either an ordinal or an interval scale	F
Cook’s distance is a measure which allows to detect outliers.	T
Data at a nominal scale can be sorted	F
Frequency polygons show the connecting line of the top of histogram bars	T
From the fact of a strong correlation one can conclude that there is a causal relationship	F
Heteroscedastic data show a variance which depends on time *… which depends on x-axis. But x-axis is not always time*	(?)
Higher order polynomials always fit better to some data points than lower order ones.	T
Homoscedasticity means evenly increasing variance	F
i.i.d. means “identical and independent deviation”	F
i.i.d. means “independent and identically distributed”	T

If the level of significance is set to 0.0 the test delivers only wrong results.	F
Logistic regression models can be calculated by the same methods as linear regression models	F
Logistic regression operates on the probabilities of certain values to occur.	T
Ordinal and nominal scales also called metric scales	F
Outliers never have an influence of [sic] Person's correlation coefficient.	F
Paired observation of school grades can be tested by the Wilcoxon test.	T
Persons's correlation coefficient can be zero despite there is a strong relationship between two samples.	T
Putting values measured on an interval scale into proportion is nonsense	T
Quartile change when sorting the data *you have to sort data before determining quartile ??*	(?)
Random errors showing a standard normal distribution cause a shift of the mean of the measurements	F
Rank based test can be applied to nominal data.	F
Rank variance analysis can be applied to analyse ordinal variables	T
Replacing a missing value by the mean of the corresponding features is better than replacing it by a zero value	T
Residuals are defined as the perpendicular distance of the measured data points from the estimated line.	F
School grades are based on an ordinal scale	T
Serial correlation of residuals is of no importance to regression models	F
Standardized data always have a median of 0.0	F
Standardized data always have a standard deviation of 1.0	T
Systematic errors cause an increase of the variance of measurements	F
The cells of a contingency table contain the correlation coefficient of the corresponding variables	F
The coefficient of determination equals the square of the correlation coefficient	T
The confidence interval of a regression estimate provides a probability of the location of the true value	F
The contingency table contains the counts of particular classes of observations	T
The correlation coefficient has a range from 0.0 to 1.0	F
The critical value of a z test is independent of the sample size	T
The deviation of a sample mean from a fixed reference value can be tested by the χ^2-test	F
The error probability of a statistical test is probability that the null hypothesis will be rejected in error	F
The exact test acc. To Fisher is computationally expensive and thus cannot be applied to large samples	T
The F distribution is named after R.A. Fletcher	F
The form of distribution can be compared by a 2-sample F-test	T
The form of the χ^2 distribution depends on the number of samples	T
The F-test can be used to test variances against a fixed value	F

The grading of school is based on an interval scale variable	F
The hair colour is an ordinal variable	F
The interquartile range contains 25% of the data	F
The interquartile range contains 50% of the data	T
The level of significance of a statistical test is the probability that the null hypothesis will not be rejected	F
The leverage effect prevents the logistic regression models form being calculated by the methods used in linear regression models	T
The logit function is a non-linear function	T
The mean splits any distribution into two sections having the same are(a?) in the probability function	T
The means of matched experiments may be compared by a two-sample t-test	T
The mode of a distribution is the most abundant frequency	T
The null hypothesis and the alternative hypothesis must not overlap	T
The odds for a value outside x ± 3s is always less than 15%	T
The overall reliability of a regression model can be tested by ANOVA	F
The power of a test is another name for the probability of making a type 1 error	F
The prerequisite of a two-sample t-test is a two-sample χ^2-test	T
The probabilities for type 1 and type 2 errors are proportional to each other	F
The probability of an event can be obtained from the integral of the probability density curve	T
The residuals of a regression model have to be normally distributed for a correct model	T
The Shapiro- Wilk test is used to test for uniform distributions.	F
The Spearman rank correlation coefficient is comparable to Pearson’s correlation coefficient	T
The speed measured in km/h is a ratio scale variable	T
The standard deviation of a sample decreases by 50 percent if the size of the sample is doubled.	F
The terms “feature”, “descriptor” and “variable” are synonymous as far as statistics is concerned.	T
The terms “observation”, “description” and “feature” are synonymous as far as statistics is concerned	F
The t-test for matched pairs requires both distributions together (seen as a joint distribution) to be normally distributed	T
The t-test for matched pairs requires each of the two distributions to be normally distributed	F
The type 1 error is the erroneous rejection of the null hypothesis.	T
The Weich test is a t-test which can be applied if variances are not equal	T
The weight of a body is a ratio scale variable	T
The Wilcoxon test is corresponding to the 2-sample F-test	F
The z-transform results in a distribution having a standard deviation of 1.0	T
The χ^2-test can be used to test for equal variances	T

The χ²-test can be used to test for normal distributions	F
t-tests are used to compare medians	F
Two sample means of independent samples can be compared by a two-sample t-test	T
Uniformly distributed residuals are a prerequisite of simple linear regression	F

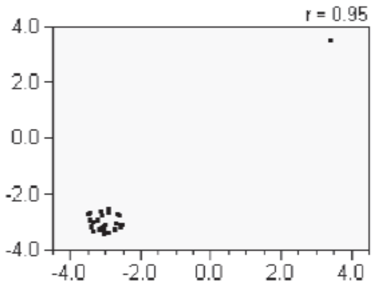
Open questions

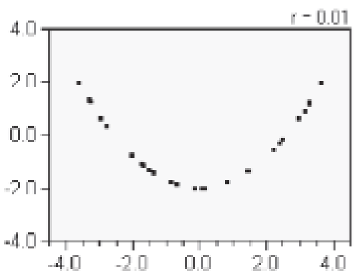
1. Describe how you would solve the following problem:

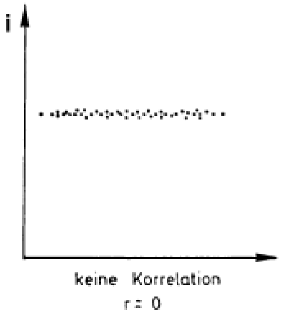
You are Measuring the blood pressure of 25 clients both in the morning and in the evening. Which statistical test(s) would you apply in order to check whether the average blood pressure in the morning is significantly different from the average pressure in the evening? Which tests are required to check the assumptions of the selected test(s)?

Paired samples → either paired t-test (if differences are normally distributed) or Wilcoxon-test (if not). Testing for normal distribution of differences: e.g. Shapiro-Wilk test

2. Please estimate the correlation coefficient (acc. To Pearson) for the following examples:





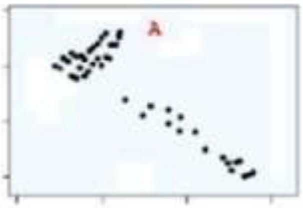


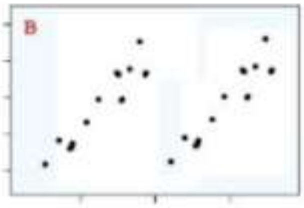
1: 0,95 **2: 0,01** **3: 0**

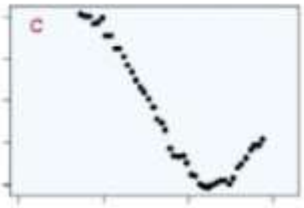
3. What are the assumptions for a linear regression model?

- **Linear relationship**
- **Independent measurements**
- **For each x-value, the y-values are normally distributed**
- **Variance of y-values is independent of ox (homoscedasticity)**

4. Please estimate the correlation coefficient (acc. To Pearson) for the following examples:

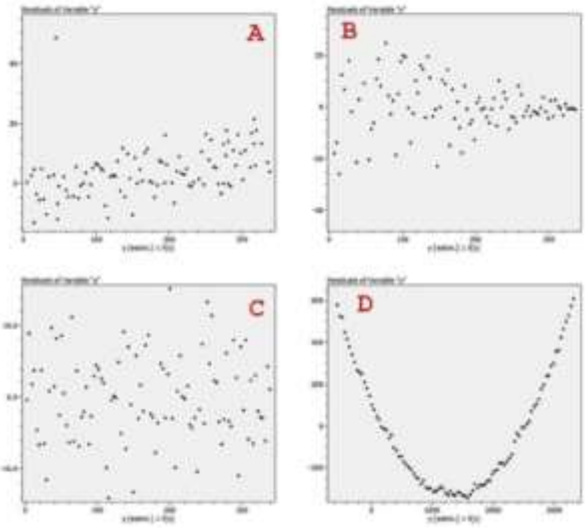






A: -0,8ish **B: 0,3 (?)** **C: -0,6**

5. The following images show residual plots of a regression model. Please explain for each scenario A to D whether the residuals indicate any problem, and if so, which kind of problem?



A: One clear outlier (and not independent on x (?));

B: Not homoscedastic

C: Looks fine

D: Wrong regression model

6. Suppose you have to make a statistically sound decision whether the crop obtained form 14 different orchards differs between 2018 and 2019. All 14 lots were treated the same, the only influence should therefore be the weather. Which statistical test(s) would you apply to decide whether there is an influence of the weather (including any tests for checking required assumptions)?

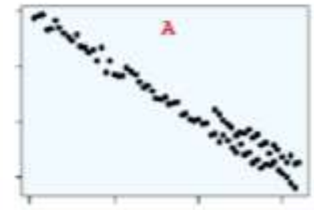
Again, paired t-test or Wilcoxon test

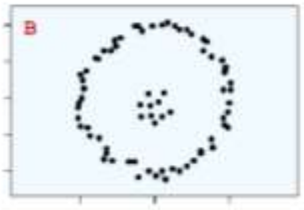
Prerequesites: testing, if differences are normally distributed (via e.g. Shapiro-Wilk test)

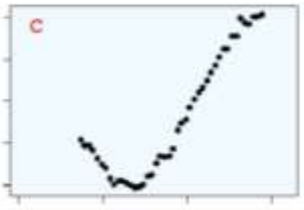
7. Describe how to solve the following problem: you have two drugs for lowering the blood pressure and want to know one has the greater effect. How would you select the clients and which statistical tests would you use? Which tests are required to check the assumptions of the selected test(s)?

Paired t-test or Wilcoxon test

8. Please estimate the correlation (acc. To Pearson) for the following examples:







A: -0,6 **B: 0,0** **C: 0,5**