Biostatistics 2018

Theory (MCQ) Compilation with Notes.

Midterm I (Introduction, Descriptive Statistics, Probability Calculus, Frequently Used Distribution, Estimation and Confidence.)

Notes: The theory MCQ questions for semifinal are mostly new so these are probably not helpful at all. Gonna make new comprehensive notes instead of finishing up these old questions. Answers in red, use at own risk because like stats and everything in life there lies uncertainty no matter how small.

What are the types of variables for which the cumulative frequencies can be determined?
Only for quantitative variables.
For ordinal and quantitative variables.
Only for discrete quantitative variables.
For nominal, ordinal and quantitative variables.

In meteorology, the intensity of UV-B radiation is classified as follows: weak, moderate, strong, very strong, extreme. What is the type of this data?
Categorical ordinal.
Discrete numerical.
Categorical nominal.
Exponential.

Celsius degree temperature scale...
...is a ratio scale, because it is possible to determine the order, the difference, and the sum of the values, and the zero point is determined on the basis of a convention.
...is an interval scale, because it is possible to determine the order, the difference, and the sum of the values, and the zero point is determined on the basis of a convention.
...is an interval scale, because it is possible to determine the order, difference, and the sum of the values, and instead of a conventional zero value, it is based on a natural zero value.
...is a ratio scale, because it is possible to determine the order, difference, and the sum of the values, and instead of a conventional zero value, it is based on a natural zero value.
Which central values can be used in the case of nominal categorical variables?
- Only the mean and median can be used.
- Mode, median, and mean can be used.
- Only the mode and the median can be used.
  **Only the mode can be used.**

Which central tendency fits for characterizing random variables measured on ordinal scale?
- only the mean and the median
- only the median
- **only the mode and the median**
- only the mode and the mean

Quantitative data may be either ...
- ... counted or discrete.
- ... measured or continuous.
- **... continuous or discrete.**
- ... numerical or discrete.

Which of the following is a measure of spread?
- Mean.
- Mode.
- Median.
- **Range.**

Which parameter is a central tendency?
- Variance
- **Arithmetic mean**
- Square of the standard deviation
- Standard deviation
Choose the correct statement.

The fourth central moment describes the extent how much heavier one tail of a probability density function is relative to the other.

The third central moment is related to skewness.

The first central moment is one.

The second central moment is standard deviation.

Note – “zeroth” is 1, first is 0, second is variance, third is skewness, and fourth is kurtosis.

In a dataset, the mean of quadratic differences from a certain value ...

... is always minimal for the mean.
... is always minimal for the median.
... is always minimal for the standard error.
... is always minimal for the mode.

Which one of the following is the characteristic of the (arithmetic) mean?

The average of the absolute deviation is minimal if it is measured from the mean.

The mean is sometimes equal to both the mode and the median.

A sample does not necessarily have a mean.

Half of the data of a sample is less than the mean of the sample.

With the help corresponding percentile curve (given in the formula collection), determine the percentage of the 9-year old boys whose body mass index (BMI) is lower than 21 kg/m².

90%.
3%.
95%.
75%.

Give the interquartile range of the BMI index of the 10 years old girls. (Use the given formula collection.)

3.1kg/m²
16.8kg/m2
5.4kg/m2
4.6kg/m2
We tossed a fair coin 3 times and the outcome was always heads. What is the probability that the outcome of the 4th toss is heads?

It is \((1/2) \div 4 = 1/8\).

Not enough data to tell.  
It is **50%**.

It is \((1/2)^4 = 1/16\).

The probability that the patient coming to our office has viral infection is 0.51. The probability of the occurrence of flu infection at our office is 0.1. What is the probability that the patient has flu, **IF** we know that this patient has viral infection?

19.61%.

5.1%.

0.459

0.049.

Note – 0.1/0.51.

The probability of occurrence of vascular stenosis among smokers is 0.26. The probability of being a smoker in a given population is 0.34. What is the probability in this population that someone is a smoker with vascular stenosis?

0.765

~1

Data is too few to answer.

0.088

Note – 0.34*0.26.

The probability of occurrence of an event is 0.21. What is the probability that it does not occur?

-0.79

0

0.11

0.79

Note – 1 - 0.21.
The consequence of the law of large numbers is ...

... that the standard error of a sample goes to the standard deviation of the corresponding theoretical distribution if the sample size goes to infinity.

... that the median of the sample goes to the mean of the theoretical distribution, if the sample size goes to infinity.

... that the absolute frequency of an event goes to its probability if the sample size goes to infinity.

... that the relative frequency of an event goes to its probability if the sample size goes to infinity.

What are the conditions of the validity of the following formula: \( p(A|B) = \frac{p(A \cap B)}{p(B)} \)?

It is always true.
It is true only if “A” and “B” are mutually exclusive events.
It is sometimes true, sometimes not, depending on the probability value.
It is true only if “A” and “B” are independent event.

Which of the following is the fundamental theorem (definition) of conditional probability (a.k.a. Bayes' theorem)?

\[ p(A \cap B) = p(A) \cdot p(B) \]
\[ p(A \cup B) = p(A) \cdot p(B) \]
\[ p(A|B) = \frac{p(A \cap B)}{p(B)} \]
\[ p(A|B) = \frac{p(A \cap B)}{p(A)} \]

In which case is the following equation valid: \( p(A + B) = p(A) + p(B) \)

If “A” and “B” are independent events.
If “A” and “B” are mutually exclusive events.
For any “A” and “B” events.
If “A” and “B” are dependent events.

When are events A and B called as mutually exclusive events?

If the probability of occurrence of events A and B in the same time is 0.
If events A and B occur only at the occurrence of an event C.
If the occurrences of events A and B are independent from each other.
If the probability of occurrence of events A and B in the same time is 1.
"A" and "B" are mutually exclusive events. The probability of event "A" is 0,42. What is the probability of event 'B'?

- **Maximum 0,58.**
- Minimum 0,42.
- Not more than -0,42.
- Exactly 0,58.

Which of the following is a stochastic event?
- Snow thaws at atmospheric pressure if heat goes above 0 °C.
- Total eclipse of the Sun (visible from Budapest).
- **Brownian motion.**
- Damage to the finger if immersed in boiling water.

In a calculation the value of the odds was found as -5. What is your conclusion according to this calculation?
- The occurrence of the event is one fifth of the non-occurrence of this event.
- **The calculation is incorrect, it should be repeated, because the value of the odds cannot be a negative number.**
- The probability of the event is higher than 0.5.
- The occurrence of the event is 5 times higher than the non-occurrence of this event.

What is NOT the requirement for a good estimation?
- A good estimation should be consistent.
- A good estimation should be efficient.
- A good estimation should be unbiased.
- **A good estimation should be unpredictable.**

Which statement is true for the standard normal distribution?
- Its density function has a maximum at 1.
- Its expected value and theoretical standard deviation are equal to each other.
- Its expected value is 1.
- **Its expected value is 0.**
Give the expected value and the theoretical variance of standard normal distribution.

- **Expected value: 0; theoretical variance: 1.**
- Expected value: 1; theoretical variance: 0.
- Expected value: 1; theoretical variance: 1.
- Expected value: 0; theoretical variance: 0.

The expected value...

- ... of the random variable $x$ is the integral of the $f(x)$ function from negative infinity to positive infinity.
- ... of a random variable can be estimated by the arithmetic mean of a sample of its values.
- ... of a sample is equal to the average of the data.
- ... is the most probably value of the random variable.

Choose the definition of frequency in statistics.

- The frequency of an event is equal to the count of its occurrence in the population.
- The frequency of an event is equal to the count of observations.
- The frequency of an event is equal to the count of its occurrence per unit time.
- **The frequency of an event is equal to the count of its occurrence in a series of observations.**

If the sample size goes to infinity ...

- ... the mean goes to the average.
- ... the standard deviation goes to zero.
- ... the mode goes to the expected value.
- ... the relative frequency of an event goes to its probability.

Choose the true statement(s).

- If the number of elements of a sample is increased, the standard deviation goes to zero.
- Confidence level is given with the mean and the sample standard deviation.
- **Standard error is a parameter used in the estimation of expected value.**
- If the number of elements of a sample is increased, the standard error goes to the theoretical standard deviation.
What is the relation between the cumulative distribution and probability density functions of a random variable?

The cumulative distribution function and the probability density function cannot be given by each other. *The probability density function is the derivative of the cumulative distribution function.*

The cumulative distribution function is the inverse of the probability density function.

The probability density function is the integral of the cumulative distribution function.

What is the reason of the fact that most physiological variables are characterized by normal distribution?

They are actually transformed so that they would follow normal distribution which makes them easier to be dealt with.

They are all related to each other so they must have the same (or at least similar) distribution.

In fact, most of them does not follow normal distribution, and we rather deal with those few which do. *They influenced by many independent events, and this, according to the central limit theorem, yields normal distribution.*

What is reference interval?

*Any interval that contains roughly 95% of the data.*

An interval determined from a (big) sample: the mean plus/minus the standard deviation.

*Any interval that contains roughly 68% of the data.*

An interval determined from a (big) sample: the mean plus/minus twice the standard deviation.

The following statements are about the reference range. Pick the true one.

The measured units falling out of the reference range come from ill people.

*There is roughly 2.5% probability that a measured value of a normally distributed variable falls below the reference range.*

The reference range contains approx. 65% of the elements of the population.

The reference range is used for both normally and non-normally distributed variables.

Select the statement which is NOT true for lognormal distribution.

The survival time of malignant tumors usually follows lognormal distribution.

*The lognormal distribution has no significance in medical practice.*

The body height in childhood follows lognormal distribution.

The body weight in childhood follows lognormal distribution.
Pick the true statement.

The columns of a histogram are called bins, if their areas are proportional to the respective relative frequencies.

The columns of a histogram are called bins, if their heights are proportional to the respective relative frequencies.

Histogram is a useful way of data representation, since the area under the curve of a given interval is proportional to the frequency of data in that interval.

Only discrete numerical data can be represented with a histogram.

There is a linear function between $y$ and $x$, if...

... the change in $y$ is proportional to the change in $x$.

... $y = a \times x$, where $a$ is a constant (the increment.)

... $y / x$ is constant.

... $x = a \times y$, where $a$ is a constant (the increment.)

Select the correct statement.

In a series of measurements, frequencies can always be summed without further conditions.

In a series of measurements, the sum of all conditional relative frequencies yields always 1.

In a series of measurements, conditional relative frequencies can always be summed without further conditions.

In a series of measurements, relative frequencies can always be summed without further conditions.

Choose the correct statement(s).

In medical practice, confidence level usually corresponds to 95% probability.

The confidence level can be chosen freely.

The confidence interval contains 95% of the data.

The confidence interval contains 68% of the data.

Midterm II (Hypothesis Testing except Chi-square Test.)

In a one-sample t-test the calculated t value is 1.897 and the t value that belongs to the significance level is 2.013.

What should be your decision?

I accept the null hypothesis.

I repeat my calculation, because this situation cannot happen in one-sample t-test.

I cannot say anything without knowing the probabilities.

I reject the null hypothesis.
The sample p-value calculated during a Wilcoxon test is 0.035. The critical p-value is 5%. Choose the correct statement.

The null hypothesis is rejected, i.e. a significant difference can be assumed between the values measured before and after the treatment.

The difference between the medians is significant in 97% of all cases.

The null hypothesis is accepted, i.e. there is no significant difference between the values measured before and after the treatment.

There is no significant difference between the medians in 97% of all cases.

What is our conclusion if we obtain \( p = 3.5 \) in a two-samples t-test?

- Nothing, we do the calculation again.
- We reject the null hypothesis considering 5% level of significance.
- Since \( p > 1 \), there is no significant different between the two groups.
- We accept the null hypothesis considering 5% level of significance.

We are executing a two-tailed t-test. Choose the possible outcome.

- \( P = 0 \) and \( t = 1 \).
- \( T = 0 \) and \( p = 1 \).
- \( T = 100 \) and \( p = 1 \).
- \( T = 0 \) and \( p = 0 \).

We tested an anticancer medicine. The result of the paired t-test is \( t=0 \), so

- We reject the null-hypothesis, so there is no effect.

We measured practically the same before and after the treatment.

- Because \( t = 0 \) then \( p \) is zero, so the drug is effective.
- We accept the null-hypothesis, so the drug has effect.

Choose the false statement(s).

- Significance level gives the probability that the rejected null hypothesis is true.

- Significance level gives the probability that the accepted null hypothesis is false.

- Significance level is equivalent with the type I error.

- Significance level gives the percentage of data outside the normal range.
In which case(s) is it appropriate to use Wilcoxon sign test?

To test the change of a non-parametric variable in two paired samples.
To compare the parametric variable of two samples with different numbers of elements.
To test normally distributed numerical variables in one sample.
To compare the non-parametric variable of two samples with different numbers of elements.

What is the aim of a test for independence?
It tests whether the means are independent from the choice of group.
Tests the effect of risk factors.
Tests whether the probabilities of possible outcomes of a quality are independent of the presence of another effect.
Tests whether two random variables are independent.

Choose the correct statement(s).
The size of a sample is proper if the calculated F-value is approx. 1.
In the case of proper sampling, the type I error is less than 5%.
The size of the sample is proper if its relative frequency distribution is not significantly different from the probability distribution of the population.
In the case of proper sampling both the type I and type II errors are less than 5%.

Which method is usually used to find the "best fitted line" in linear regression?
Least-square method.
Maximal entropy method.
Minimal mean distance method.
Minimal absolute difference method.

Choose the right statement. The y axis intercept of the regression line...
... cannot be zero.
... cannot be more than one.
... cannot be less than negative one.
... can be any real number.
What is the reason of using an ANOVA instead of several t-test on the same sample?

With ANOVA we can reduce the multiplicity (the increase in the first type error).

Variance analysis has higher power than t-tests.

Comparing variances reduces higher the second type error than comparing means.

Normality could not be interpreted for multiple comparison.

When is (parametric one way) ANOVA applicable?

If the samples to be compared are independent and normally distributed.

If the variance of the samples is equal and all samples are normally distributed.

When it yields less alpha and beta error than a series of paired t-tests.

If more than half of samples to be compared are normally distributed.

We want to check whether a die is fair or loaded. Choose the correct null hypothesis.

The die is loaded.

The outcome of a set of rolls with the die does not deviate significantly from uniform distribution.

The outcome of a set of rolls with the die deviates significantly from uniform distribution.

The die is loaded significantly.

A good null hypothesis in general:

We state there is no relation, but there is a relation in fact.

It denies the effect – if we reject it the treatment is not effective.

If null-hypothesis is true, we have a normal distribution.

We state the two samples come from a same population.

Which statement is correct?

The coefficient of determination and the intercept of the straight line always have the same sign.

The slope and the intercept of the straight line always have the same sign.

The correlation coefficient and the intercept of the straight line always have the same sign.

The correlation coefficient and the slope of the straight line always have the same sign.
The correlation coefficient is close to one if …

... the relationship between the variables is linear.
... the relationship between the variables is not significant.
... the relationship between the variables is linear with positive increment.
... the relationship between the variables is function-like.

What distribution do the data sets follow if we used Wilcoxon test to compare them?
Can be anything different from standard normal distribution.
The distribution is either unknown or non-normal.
It has a t-distribution with number of degrees of freedom equal to one minus the sample count.
Normal distribution.

What is your conclusion when the calculated p-value of an F-test is higher than the significance level?
I have to use Mann-Whitney U-test.
Variances are not equal.
Variances are equal.
I can accept the null hypothesis of two-sample t-test.

Which statement is correct?
The probability of type II error cannot be decreased.
The probability of type II error can be decreased by increasing the size of the sample.
The probability of type I error cannot be decreased.
Type I error cannot occur in the case of large samples.

How can you decrease the chance to do a second type error?
Increase the significance level.
Decrease the first type error.
Increase the sample size.
We can’t decrease the second type error, only the first type error.
Choose the right statement on type II error.
Gives the error of wrong decision.

**The alternative hypothesis is accepted, although it is false.**
Gives the error of the right decision.
The null hypothesis is accepted, although it is false.

What does the odds ratio show? (Regarding to an illness and risk factor.)
It shows how many times higher the probability of an illness in the presence of the risk factor than in the absence.
It shows how many times higher the probability of the occurrence of the illness than the non occurrence of the illness.

**It shows how many times higher the odds of a disease in the presence of the risk factor than its absence.**
It shows how many times the sampling error is more likely than the correct sampling.

Estimate the probability that 21 hernia surgery will be successful out of 26 if the success rate is 93%.

0.024084
0.032079
0.975916

We are studying a numerical, continuous variable in two groups having equal number of elements. Both groups show normal distribution, the variances can be considered equal. Can we use Mann-Whitney U-test in this case?

Yes, we can because the sample fulfills the preconditions of the test.
No, we cannot, in the described conditions we have to use Kruskal-Wallis-test.
No, we cannot, we have to use t-test, since the conditions fulfill the requirements for t-test.
No, we cannot because non-parametric test can be used only for non-numerical data.

Choose the correct statement.
The slope of the regression line cannot be less than negative one.

**The slope of the regression line can be any real number.**
The slope of the regression line cannot be zero.
The slope of the regression line cannot be more than one.
Pick the correct statement.

T-test can only be applied for variables following standard normal distribution.

**The condition of t-test for one sample is that the variable to be analyzed follows normal distribution.**

The condition of t-test for one sample is that the variable to be analyzed follows Student’s t distribution.

The condition of t-test for one sample is that the standard deviation is one.

We would like to compare the efficacy of an original drug and a generic one. Choose the correct null hypothesis.

**The efficacy of the generic drug is not significantly different from the original.**

The difference between the efficacy of the two drugs is not because of incidence.

Incidence has no role in the efficacy difference between the two drugs.

The efficacy of the generic drug is not identical to that of the original one.

The pain killer effect of Aspirin and a new drug called “Novanopain” is compared. Choose the optimal test method.

Examining correlation between the two groups with t-test.

**Mann-Whitney U-test.**

Two sample t-test, since we want to compare two independent groups.

Simply compare the confidence intervals.

What test can we use if we have 1 numerical, continuous variable in 2 not paired groups and the groups are not normally distributed?

ANOVA

Wilcoxon rank

**Mann-Whitney U**

Kolmogorov-Smirnov

What test can we use if we have 1 numerical, continuous variable in 3 (not paired) groups and the groups are normally distributed?

ANOVA

Mann-Whitney U

Kruskal-Wallis

Kolmogorov-Smirnov
In which case should one reject the null hypothesis?
If the absolute value of the statistical parameter calculated from the sample is greater than 5%.
If the p-value calculated from the sample is greater than the critical p-value.
If the p-value calculated from the sample is greater than 5%.
If the absolute value of the statistical parameter calculated from the sample is greater than the absolute value of the critical statistical parameter.

Null hypothesis is rejected if...
... the sample statistical parameter is less than the critical statistical parameter.
... the significance level (considering a two tailed test) is more than 5%.
... the significance level is less than 5%.
... the sample statistical parameter is greater than the critical statistical parameter.

We want to test with a t-test whether the 42 patients in the hematology ward have the same red cell count as the 42 patients in the contagious ward. What is the number of degrees of freedom?
21
42
82
41

What is the number of degrees of freedom in case of a correlation t-test, if the sample size is 17?
17
16
15
32

Which one of the following is an example of statistical inference?
Calculating the sample t-value.
Rejection the null hypothesis as a result of hypothesis testing.
Calculating the second central moment of the sample.
Calculating the second moment of the sample.
What does the term “tied rank” (or “linked rank”) stand for in statistics?
Ranks not exchangeable between samples.
**Identical ranks that are assigned to values of equal magnitude.**
Ranks not exchangeable within a sample.
Ranks that are assigned to data in a two-sample t-test for data pairs.

Standard normal distribution is identical to...
... Student’s t-distribution with infinity degrees of freedom.
... uniform distribution.
... the Gaussian distribution.
... lognormal distribution.

---

**Others in Semifinal (Chi-square Test, Diagnostic Tests, Information Theory, Evidence-based Medicine and Clinical Studies.)**

The diagnostic segregation shows
the frequency of diseased people in the examined population.
the probability that the test is positive for diseased person.
**the probability of healthiness if test is negative.**
the probability that the test is negative for healthy person.

Which test may be used if the conditions of the chi-square test are not true?
Kruskall-Wallis test.
**Fisher’s exact test.**
Correlation t-test.
ANOVA.

What is the number of degrees of freedom in case of chi square test for homogeneity, if we study whether the distribution of male and female patients in our 5 hospitals is uniform?

4
1
8
3
A test for homogeneity is to be conducted. Which method shall be used?

Student’s t-test for two samples.

**Chi-square test.**

Student’s t-test for one sample.

Mann-Whitney U-test.

Choose the condition of a chi-square test for independency.

Each value in the expected contingency table must be greater than 5n (where n is the count of the dataset.)

*Each value in the observed contingency table must greater than 2 and at least 50% of the values must be greater than 5.*

Each value in the observed frequency table must be greater than 5n (where n is the number of cells in the table.)

At least four-fifth of the values in the expected frequency table must be larger than 5.

Which of the following methods is part of evidence based medicine?

Naturopathy.

Traditional Chinese medicine.

**None of the other three answers is correct.**

Acupuncture.

We are studying the applicability of a diagnostic test. What is the name of the parameter given by the ratio of true positive tests and all ill people?

**Sensitivity**

Prevalence

Specificity

Negative predictive value

We are studying the applicability of a diagnostic test. What is the name of the parameter given by the ratio of true positive tests and all positive tests?

**Positive predictive value**

Sensitivity

Specificity

Negative predictive value
Select the correct statement.
The less the probability of occurrence of the signal, the less its information content.
There is no relation between the probability of occurrence of the signal and its information content.
**The less the probability of occurrence of the signal, the more its information content.**
The more the probability of occurrence of the signal, the more its information content.

Choose the right statement on information content.
Its unit is selby.
**The information content of a “message” consisting of a single sign is equal to log2(1).**
It has no unit.
The less frequent event has less information content.

What is the power gain level, if the ratio of the output and input power is 1?

10 dB
1 dB
20 dB
0 dB

Calculate the information entropy for two symptoms where the probability of the symptoms are 0.02 and 0.35.

0.1936
0.007
**0.643**
7.1584

Note: \(-0.02 \cdot \log_2(0.02) - 0.35 \cdot \log_2(0.35)\)