Potential questions for oral exam selected from quiz sessions

Session 3

Question (True/False, provide explanation (if applicable) or additional information related to the question)
The sum of all the probabilities in a probability distribution for a discrete random variable must equal 1.
For a discrete random variable, under certain circumstances p(x) could be less than 0.
The mean of a discrete random variable must be a possible value of the random variable.
The mean of a discrete random variable can be negative.
The standard deviation of a random variable is always nonnegative.
The graph of a probability density function/curve may extend below the x-axis.
For a continuous random variable there is no (positive) probability associated with a single value.

Session 4

Question (True/False, provide explanation (if applicable) or additional information related to the question)
The density curve for any normal distribution is symmetrical and bell-shaped.
The two parameters of normal distributions are the mean and the median.
Normal distributions have positive probability for intervals below zero.
We obtain a standard normal distribution by converting the variance to a standard deviation.
The normal distribution can be used for all variable types.
A z-score is useful for computing probabilities in a normal distribution from a table, but is not necessary when using statistical software.
Because weights (of people, animals etc) are by definition non-negative, it never makes sense to approximate weights by a normal distribution.

For a normal distribution, the points in a normal plot should either follow a straight line or a bell-shaped curve.
If a P-value for a normality test is low, the data do not match a normal distribution well.

Question (True/False, provide explanation (if applicable) or additional information related to the question)
We only need to know the probability of an event in order to use a binomial distribution for the count of such events.
The binomial distribution has two parameters, of which one is (almost) always known.
The two most crucial assumptions behind a binomial distribution are independence and equal probabilities in all trials.
A binomial distribution can be used also for counts of three or more different outcomes in each trial.
We can check the validity of a binomial distribution by a normal plot.
It is valid to use a binomial distribution for samples drawn from a finite population when the proportion of the event is less than 5%.
Probabilities in a binomial distribution are most easily computed by hand (calculator).

Session 5

Question (True/False, provide explanation (if applicable) or additional information related to the question)
A confidence interval for a population mean (mu) is guaranteed to contain the true value (mu).
A confidence interval for a population mean always includes the sample mean.
As the confidence level increases, the confidence interval becomes longer.
Values of z* (z-star) for confidence intervals are calculated from a standard normal distribution.
It is equally informative to know a confidence interval (for a population mean) as to know the sample mean and a margin of error.

Question (True/False, provide explanation (if applicable) or additional information related to the question)
A statistical hypothesis is usually stated in terms of a population parameter.
A statistical hypothesis is never stated in terms of a sample estimate.
The P-value is the probability that the null hypothesis is true.

A low P-value means that the data are incompatible with the null hypothesis (or other assumptions).
A high P-value means that the null hypothesis is most likely true.
In a test of a statistical hypothesis, we attempt to find evidence in favour of the null hypothesis.

Session 6

Question (True/False, provide explanation (if applicable) or additional information related to the question)
Inference about a population mean, when the standard deviation is unknown, is based on a suitable t-distribution.
The degrees of freedom of t-distributions are calculated from sample statistics, such as the sample mean.
A t-distribution with 1 degree of freedom is essentially the same as a standard normal distribution.
Tail probabilities and percentiles used for confidence intervals are larger in any t-distribution than in a standard normal distribution.
The t* (t-star) value used for a confidence interval for the population mean is the value of t-test statistic.
If we have 95% confidence intervals for two population means (from independent samples), we can get a 95% confidence interval for their difference by subtracting the confidence intervals from each other.
It is only valid to form differences between the observations in two samples if the samples are paired.

Session 7

Question (True/False, provide explanation (if applicable) or additional information related to the question)
An approximate 95% confidence interval for an unknown proportion p is the sample mean plus or minus its standard error.
The classical (large-sample) significance test for a population proportion is based on a t distribution.
If two sample proportions are equal, then the sample counts must be equal.
When the sample proportion equals 0, there is no real confidence interval (it consists of the value 0 only).
For proportions (1-sample and 2-sample) we cannot rely on the software defaults to always give us the most appropriate answer.
Adding extra observations for the plus four method is only valid for confidence intervals.
A confidence interval for p1-p2 can be used to draw a conclusion about the equality of two population proportions (p1 and p2).

Question (True/False, provide explanation (if applicable) or additional information related to the question)
The non-parametric tests do not assume the sample mean to have an (approximately) normal distribution.
The non-parametric tests never make assumptions about the distribution of the data.
Rank-based test do not use the full information in the data.
P-values for rank-based tests are preferably computed by statistical software.
Non-parametric tests are more sensitive to outliers than tests based on the normal distribution.
The closest parametric analog to a Wilcoxon signed rank test is a t-test for one sample.
The closest parametric analog to a Mann-Whitney-Wilcoxon test is a t-test for two paired samples.
The sign test can be calculated from a binomial distribution.

Session 8

Question (True/False, provide explanation (if applicable) or additional information related to the question)
If we reject the null hypothesis in the test for homogeneity of populations, there is evidence to suggest the category proportions are all the same.
If we reject the null hypothesis in the test for independence of two categorical variables, there is evidence to suggest the two variables are dependent.
For the Pearson X2 statistic, we reject the null hypothesis only for large values of the test statistic (no matter the alternative hypothesis).
The degrees of freedom in the chi-square distribution is determined from the total sample size.
If a two-way table has an observed count of zero, the conditions for use of the chi-square distribution for the Pearson X2 can never be met.
If a two-way table has a row (or column) where all observed counts are zero except for a single count of one (1), the conditions for use of the chi-square distribution for the Pearson X2 can never be met.
We have to use different test statistics to assess whether proportions are the same across columns, and whether there is independence between row and column variables.
A chi-square distribution is similar to a standard normal distribution, in particular when the degrees of freedom is large.
It does not matter for the calculation of the Pearson X2 statistic which variable (out of two) we put on the rows and which variable we put on the columns.

Question (True/False, provide explanation (if applicable) or additional information related to the question)
As long as the statistical analysis is correct, it does not matter how it is reported.
It is better to report actual P-values instead of ranges such as $P<0.05$.
In a journal not focused on statistics, the description of statistical methods should always be kept to a few sentences.
It is appropriate to refer to the original paper by "Student" from 1908 (see media page) when using t-tests in the statistical analysis.
In addition to a P-value, it is useful to report a confidence interval.
The number of observations used for any specific analysis should always be retrievable from the information provided.

Other study guidelines, such as CONSORT for randomised trials and STROBE for observational studies, override the statistical reporting guidelines.
It is recommended to present results both in tabular and graphic format.

Session 9

Question (True/False, provide explanation (if applicable) or additional information related to the question)
In a one-way ANOVA, the within-group standard deviations are assumed equal across all groups.
In a one-way ANOVA, all group sizes must be the same in order for the F-statistic to be valid.
A balanced design means that all groups have the same number of observations.
If we reject the null hypothesis in a one-way ANOVA, there is evidence to suggest the population variances differ between groups.
If we reject the null hypothesis in a one-way ANOVA, there is evidence to suggest all the population means of the groups are different.
In a one-way ANOVA, if the value of the F-statistic is much greater than 1, then the variation between groups is larger than the variation within groups.
The Bonferroni multiple comparisons procedure can be used (also) if we are only interested in some of the comparisons.
For the ANOVA F-statistic, we reject the null hypothesis only for large values of the test statistic (no matter the alternative hypothesis).
An F-distribution has two parameters, but their order does not matter (e.g., F(1,4) is the same as F(4,1)).
The pooled variance estimate can be found in the ANOVA as the value for SSE.

Session 10

Question (True/False, provide explanation (if applicable) or additional information related to the question)
In a simple linear regression, all the points lie on the estimated regression line.
In a simple linear regression, the slope being zero corresponds to no linear association between y and x.
The hypothesis of no linear association (against a two-sided alternative) can be tested both by a t-test and an F-test, with the same P-value.
The correlation is always between 0 and 1.
When predicting y from x in a simple linear regression, the standard error and prediction error do not depend on x.
When predicting y from x in a simple linear regression, the confidence interval and prediction interval are centered on the same value.
Residuals always add up to zero (except for round-off error).
The normality assumption of the regression model can be assessed by a normal plot for the y-variable.
A simple linear regression model assumes the dispersion about the line to be constant.
A simple linear regression or ANOVA model fits the data well if all residuals are numerically less than 1.

Session 11

Question (True/False, provide explanation (if applicable) or additional information related to the question)
In a two-way ANOVA, a significant interaction means the relationship between the two factors is not additive.
Parallel lines in an interaction plot for two factors means that the relationship between the factors is additive.
A significant F-test for a multiple linear regression model means that every predictor contributes significantly to the prediction equation for the outcome.
In a two-way ANOVA, if there is evidence of a significant interaction, the analysis stops.
If there is no evidence of a significant interaction, the two-way ANOVA needs to be followed up by one-way ANOVAs for each of the factors.

The degrees of freedom for the interaction is never smaller than the degrees of freedom for (the main effects of) the factors involved.
In a two-way ANOVA, the interaction can only be tested if there is replication.
Replication in ANOVA and multiple regression models means that by these models some observations are predicted to have the same mean.
When the data have replication, the two-way ANOVA model is equivalent to a one-way ANOVA model for the factor formed by combining the two factors.
Model checking in ANOVA and multiple regression primarily uses the residuals.
Both ANOVA and multiple regression are special cases of a (General) Linear model.
Each of the x-variables in multiple regression need to be normally distributed.
The normality assumption in ANOVA and multiple regression is assessed by a normal plot and normality test for the outcome.