VU University	Statistical Data Analysis
Faculty of Sciences	2 July 2015

Use of a basic calculator is allowed. Graphical calculators are not allowed. Please write all answers in English.
The exam consists of 6 questions (45 points). Grade = $\frac{total+5}{5}$.

GOOD LUCK!

Question 1 [8 points]
Are the following statements correct? Motivate your answer by a short argument or sketch.

a. [2 points] A two sample QQ-plot is the same as a scatter plot of a bivariate sample.

b. [2 points] The mean is more robust as location estimator than a trimmed mean.

c. [2 points] M-estimators arealways robust estimators.

d. [2 points] Kendall's τ and Spearman's ρ are two measures for the rank correlation between two independent samples.

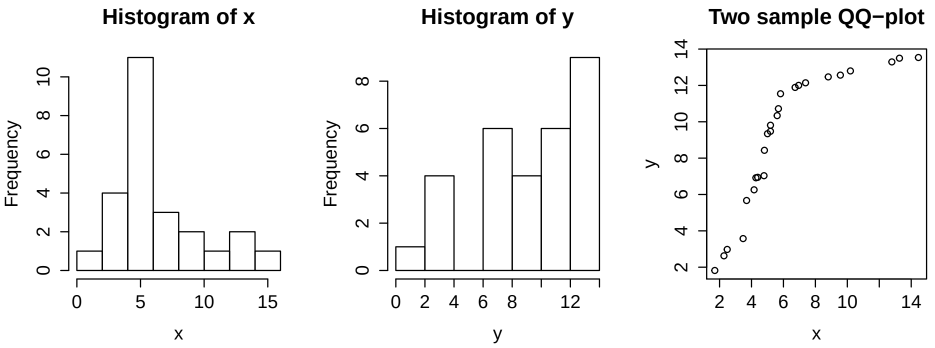

Figure 1: Histogram of data set x (left), histogram of data set y (middle) and two sample QQ-plot of data sets x and y (right). Data set x contains 25 observations, data set y contains 30 observations.

Question 2 [8 points]
In Figure 1 histograms of two independent data sets, x and y, are given, together with the two sample QQ-plot of these data sets.

a. [2 points] What can you say about the right tail of the underlying distribution of data set x compared to the right tail of the underlying distribution of data set y based on the histograms of both samples? Can you deduce it also from the QQ-plot? How?

b. [3 points] Suppose that we would like to test whether or not the distribution of data set x equals the normal distribution with

expectation 6 and variance 3. Evaluate for each of the following tests for goodness of fit whether it is suitable for testing this null hypothesis and motivate your answer:
i) Kolmogorov-Smirnov test;
ii) chi-square test for goodness of fit;
iii) Shapiro-Wilk test.

c. [3 points] Suppose we want to test whether the underlying distributions of the two samples x and y are the same. Evaluate for each of the following two-sample tests whether it is suitable for testing this null hypothesis and motivate your answer:
i) Wilcoxon rank sum test;
ii) Kendall's rank correlation test;
iii) Two sample Kolmogorov-Smirnov test.

Question 3 [7 points]
Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with unknown distribution P. Suppose that $T_n(X_1, \ldots, X_n)$ is used to estimate the location of P. To determine the accuracy of this estimator, its variance is estimated by means of the bootstrap.

a. [2 points] Describe shortly which two errors are made in this procedure. Explain your notation.

b. [2 points] Which of the two errors can be made arbitrarily small by the user? Motivate your answer.

Suppose we are given a data set $x_1, \ldots, x_n$ and we want to test the null hypothesis H_0: "the underlying distribution of $x_1, \ldots, x_n$ is a normal distribution".

c. [3 points] Is the following bootstrap scheme appropriate for testing the given null hypothesis? If you think that the scheme is not suitable, indicate where the error is, and how to fix it (do not change the test statistic).

1. Compute the mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ of the original data $x_1, \ldots, x_n$.
2. Compute for this data the value d of the Kolmogorov-Smirnov statistic $D = \sup_x |F_n(x) - F(x)|$, where F_n is the empirical distribution function of $x_1, \ldots, x_n$ and F is the distribution function of the $N(\hat{\mu}, \hat{\sigma}^2)$ distribution.
3. Generate B samples $X_1^*, \ldots, X_n^*$ from the $N(\hat{\mu}, \hat{\sigma}^2)$ distribution.
4. Compute for each sample the Kolmogorov-Smirnov statistic $D^* = \sup_x |F_n^*(x) - F(x)|$, where F_n^* is the empirical distribution function of $X_1^*, \ldots, X_n^*$ and F is again the distribution function of the $N(\hat{\mu}, \hat{\sigma}^2)$ distribution.
5. Compute the right p-value p_r of the observed d by $p_r = \#(D^* \geq d)/B$.

Question 4 [8 points]
One of the filling machines in a beer brewery is suspected of putting too much beer in the beer bottles. To investigate this the amount of beer in 12

	p						
k	0.025	0.05	0.33	0.5	0.67	0.95	0.0975
0	0.738	0.540	0.008	0.000	0.000	0.000	0.000
1	0.965	0.882	0.057	0.003	0.000	0.000	0.000
2	0.997	0.980	0.188	0.019	0.000	0.000	0.000
3	1.000	0.998	0.403	0.073	0.004	0.000	0.000
4	1.000	1.000	0.641	0.194	0.018	0.000	0.000
5	1.000	1.000	0.829	0.387	0.063	0.000	0.000
6	1.000	1.000	0.937	0.613	0.171	0.000	0.000
7	1.000	1.000	0.982	0.806	0.359	0.000	0.000
8	1.000	1.000	0.996	0.927	0.597	0.002	0.000
9	1.000	1.000	1.000	0.981	0.812	0.020	0.003
10	1.000	1.000	1.000	0.997	0.943	0.118	0.035
11	1.000	1.000	1.000	1.000	0.992	0.460	0.262

Table 1: Probabilities $P(X \leq k)$ for binomially distributed random variable X with parameters $n = 12$ and p as given in table, for different values of k.

bottles of a day's production was measured. The bottles are supposed to contain 33.00 cl of beer. The following amounts (in cl, sorted) were measured: 32.85, 32.91, 32.93, 32.98, 33.04, 33.13, 33.17, 33.30, 33.32, 33.41, 33.47, 33.52. Next, the problem was investigated by performing a sign test on these data.

a. [1 point] Formulate H_0 and H_1 for the test. Explain your notation.

b. [2 points] Give the formula for the test statistic and its distribution under H_0.

c. [3 points] Perform the test with significance level $\alpha = 0.05$ using Table 1. Give the p-value and the conclusion of the test.

d. [2 points] Explain the following statement: "*The test statistic T is distribution free over the class of distribution functions under H_0*".

Question 5 [4 points]
Suppose we are given a $r \times c$ contingency table, containing the counts N_{ij} for $i = 1, \ldots, r$, $j = 1, \ldots, c$, and we want to test the null hypothesis of indepence, $H_0 : p_{ij} = p_i p_j$ for p_{ij} the probability for cell (i, j) in the rc-nomial distribution of the full sample $(N_{11}, \ldots, N_{rc})$.

a. [3 points] Describe the chi-square test for contingency tables. Formulate the test statistic X^2 and its (approximate) distribution under the null hypothesis of independence. Explain your notation.

b. [1 point] Describe the rule of thumb that needs to be fulfilled for this approximate distribution of X^2 to be reliable.

Question 6 [10 points]
To determine the dependence of the amount of nicotine in cigarettes on the amount of tar and carbon monoxide, the amount of nicotine, tar and carbon monoxide was determined for 25 different brands of cigarettes. A multiple linear regression model is used to model the dependence.

a. [3 points] Formulate the multiple linear regression model suitable for this situation. Include the model assumptions. Explain the notation that you use in terms of the context.

b. [3 points] Give a one sentence description for each of the following three concepts that may cause problems in a multiple linear regression analysis: leverage (potential) point, influence point, collinearity. Name (not explain) for each of the concepts a test, measure or other tool(s) that can be used to search for their presence.

c. [2 points] Suppose that the 6th brand has a remarkably high amount of nicotine. To investigate whether this value is an outlier, the model in part (a) needs to be extended to a mean-shift-outlier model. Formulate this extended model.

d. [2 points] Describe the test that can be performed for the extended model to decide whether or not the 6th point is an outlier. (Formulate null and alternative hypothesis, give the test statistic and its distribution under the null hypothesis, and indicate when the null hypothesis will be rejected.)

THE END