VU University	Statistical Data Analysis
Faculty of Sciences	25 May 2016

Use of a basic calculator is allowed. Graphical calculators are not allowed. Please write all answers in English.

The **complete exam** consists of 7 questions (45 points). Grade = $\frac{total+5}{5}$.

The **exam on part 2** consists of 4 questions (27 points). Grade = $\frac{total+3}{3}$.

GOOD LUCK!

In this exam all data sets are assumed to be i.i.d.

PART 1

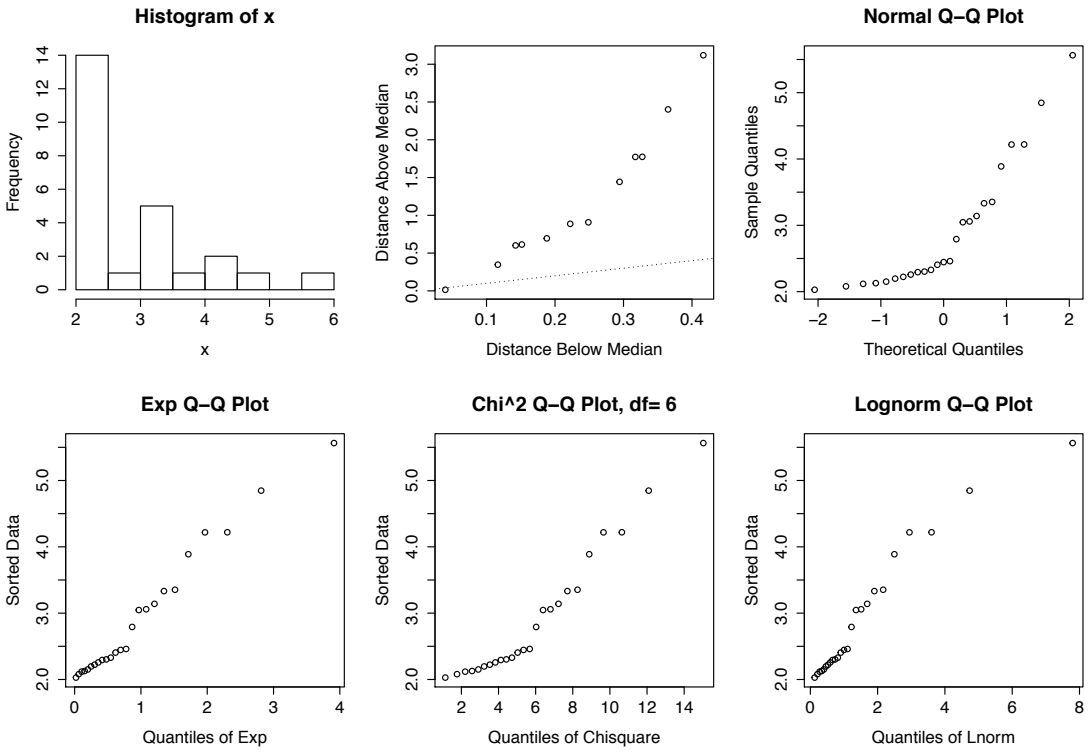

Figure 1: Histogram, symplot and QQ-plots of the data set x against the standard normal, standard exponential, χ^2_6-distribution and the standard lognormal.

Question 1 [6 points]
In Figure 1 a histogram, symplot and several QQ-plots of the data set x with sample size 25 are presented.

a. [2 points] Describe briefly what these graphical summaries tell you about the underlying distribution of data set x. Consider at least the following: location, scale, shape and extreme values.

b. [2 points] Which of the 4 location-scale families do you think is the most appropriate for the dataset x? Explain your answer.

c. [2 points] Using the QQ-plot of the location-scale family that you have selected under part (b) determine the location a and scale b approximately.

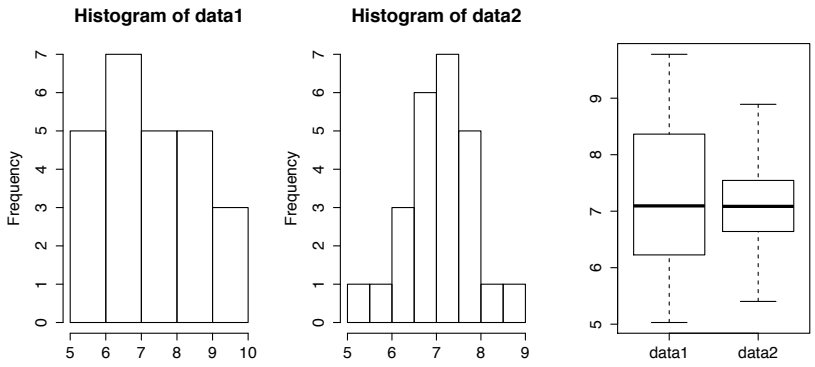

Figure 2: Histograms and boxplots of two datasets, data1 and data2.

Question 2 [5 points]
Consider the data in Figure 2 on two independent samples data1 and data2.

a. [2 points] Suppose we want to test whether the underlying distribution of data1 is a normal distribution. Which goodness-of-fit test would you use to test this hypothesis? Motivate your answer.

b. [3 points] Suppose we want to test whether the underlying distributions of these two samples, data1 and data2, are identical. Indicate for each of the following tests, whether you think this test is applicable for testing this hypothesis. Which test do you think is best (i.e. most powerful) here? Motivate your answer.

   - median test
   - Wilcoxon rank sum test (i.e. Wilcoxon two-sample test)
   - Kolmogorov-Smirnov two-sample test

Question 3 [7 points]
Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with unknown distribution P. We assume that P is an exponential distribution with unknown parameter λ. Suppose that $T_n(X_1, \ldots, X_n) = \text{median}(X_1, \ldots, X_n)$ is used to estimate the location of P. To determine the accuracy of this estimator, its standard deviation is estimated by means of the bootstrap.

a. [3 points] Describe the steps of the parametric bootstrap scheme that you would use to find the bootstrap estimate of the standard deviation of T_n.

b. [2 points] Describe shortly which two errors are (necessarily) made in a parametric bootstrap procedure.

c. [2 points] Give the formula for a bootstrap $(1-\alpha)$-confidence interval for the population median, based on T_n. Explain your notation.

PART 2

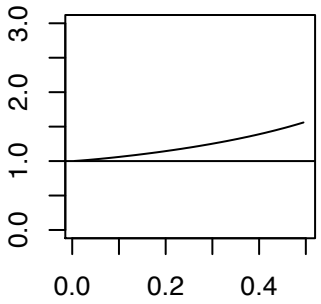

Figure 3: Asymptotic variance of α-trimmed mean as a function of α for a normal distribution. The horizontal line indicates the asymptotic Cramér-Rao lower bound.

Question 4 [6 points]

a. [2 points] Figure 3 shows the asympotic variance for the α-trimmed mean as a function of α for samples from a normal distribution. Which trim percentage is optimal for normal samples? Motivate your answer based on the figure.

Are the following statements correct, incorrect or nonsensical (i.e. makes no sense)? Choose one of the three options for each statement. Motivate your answer by a short argument.

b. [2 points] In a linear regression model: if the overall F-test yields a significant result, all parameters β in the model are significantly different from 0.

c. [2 points] The Wilcoxon signed rank test is preferred over the Wilcoxon rank sum test.

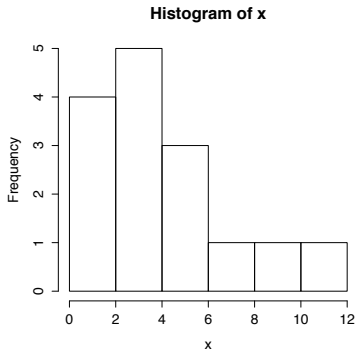

Figure 4: Histogram of a data set x.

	p						
k	0.025	0.05	0.33	0.5	0.67	0.95	0.975
0	0.684	0.463	0.002	0.000	0.000	0.000	0.000
1	0.947	0.829	0.021	0.000	0.000	0.000	0.000
2	0.994	0.964	0.083	0.004	0.000	0.000	0.000
3	1.000	0.995	0.217	0.018	0.000	0.000	0.000
4	1.000	0.999	0.415	0.059	0.002	0.000	0.000
5	1.000	1.000	0.629	0.151	0.008	0.000	0.000
6	1.000	1.000	0.805	0.304	0.029	0.000	0.000
7	1.000	1.000	0.916	0.500	0.084	0.000	0.000
8	1.000	1.000	0.971	0.696	0.195	0.000	0.000
9	1.000	1.000	0.992	0.849	0.371	0.000	0.000
10	1.000	1.000	0.998	0.941	0.585	0.001	0.000
11	1.000	1.000	1.000	0.982	0.783	0.005	0.000
12	1.000	1.000	1.000	0.996	0.917	0.036	0.006
13	1.000	1.000	1.000	1.000	0.979	0.171	0.053
14	1.000	1.000	1.000	1.000	0.998	0.537	0.316
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 1: Probabilities $P(X \leq k)$ for binomially distributed random variable X with parameters $n = 15$ and p as given in table, for different values of k.

Question 5 [6 points]
The data presented in Figure 4 contains the following 15 numbers: 0.77, 1.15, 1.79, 1.85, 2.15, 2.34, 2.36, 2.81, 2.91, 4.47, 4.86, 5.09, 7.94, 9.05, 10.69. We conjecture that the median m of the underlying distribution is smaller than 4, and want to test $H_0 : m \geq 4$ against $H_1 : m < 4$, using the sign test.

a. [2 points] Formulate the test statistic for the sign test, and give its distribution under the assumption $m = 4$.

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

c. [1 point] Is it appropriate to use the Kolmogorov-Smirnov test in stead of the sign test for testing this null hypothesis? Motivate your answer.

Question 6 [7 points]
Consider the following experiment on drug allergy. Three drugs are compared with respect to the types of allergic reaction that they cause to patients. A group of $n = 150$ patients is randomly split into three groups of 50 patients, each of which is given one of the three drugs. The patients are then categorized as being *hypoallergic*, *allergic*, *mildly allergic* or as having *no allergy*. The obtained counts are represented in the following table:

	hypoallergic	allergic	mildly allergic	no allergy	total
drug A	5	15	18	12	50
drug B	4	16	13	17	50
drug C	7	13	14	16	50
total	16	44	45	45	150

k	α						
	0.025	0.05	0.33	0.5	0.67	0.95	0.975
1	0.00	0.00	0.18	0.45	0.95	3.84	5.02
2	0.05	0.10	0.80	1.39	2.22	5.99	7.38
3	0.22	0.35	1.55	2.37	3.43	7.81	9.35
4	0.48	0.71	2.36	3.36	4.61	9.49	11.14
5	0.83	1.15	3.19	4.35	5.76	11.07	12.83
6	1.24	1.64	4.05	5.35	6.90	12.59	14.45
7	1.69	2.17	4.92	6.35	8.03	14.07	16.01
8	2.18	2.73	5.80	7.34	9.15	15.51	17.53
9	2.70	3.33	6.68	8.34	10.26	16.92	19.02
10	3.25	3.94	7.58	9.34	11.36	18.31	20.48
11	3.82	4.57	8.48	10.34	12.46	19.68	21.92
12	4.40	5.23	9.38	11.34	13.56	21.03	23.34
13	5.01	5.89	10.29	12.34	14.65	22.36	24.74
14	5.63	6.57	11.20	13.34	15.73	23.68	26.12
15	6.26	7.26	12.12	14.34	16.82	25.00	27.49

Table 2: α-quantiles of χ^2_k-distribution for indicated values of α and k.

a. [3 points] Formulate a suitable model of multinomial distribution(s) and state the corresponding null and alternative hypotheses for investigating whether there is a relationship between drug type and type of allergic reaction. (You may formulate your hypotheses either in words or in formulas.)

b. [1 point] Check whether the rule of thumb for applying the chi-square test for contingency tables is fulfilled for these data.

c. [1 point] The value of the chi-square test statistic equals 3.06 for these data. Perform the chi-square test, using Table 2 at significance level 0.05.

d. [1 point] Does the p-value increase, decrease or stay the same if all numbers in the data table are mulitplied by 10?

e. [1 point] Does the number of degrees of freedom of the chisquare distribution of the test statistic increase, decrease or stay the same if alle numbers in the data table are mulitplied by 10?

Question 7 [8 points]

a. [2 points] Formulate the general multiple linear regression model including its assumptions.

In a study on the relation between systolic blood pressure, birthweight and age of babies, the following linear regression model is fitted:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

with Y the systolic blood pressure (mm Hg), x_1 the birth weight (gram) and x_2 age (days). The estimated model, based on $n = 16$ babies is

$$Y = 53.45 + 0.0044 x_1 + 5.89 x_2 + e.$$

Other properties of the estimated model are: $R^2 = 0.88$, $se(\hat{\beta}_1) = 0.0012$ and $se(\hat{\beta}_2) = 0.68$ ($se(\hat{\beta}_j)$ denotes estimated standard error of $\hat{\beta}_j$).

b. [2 points] Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at the level of 0.05 (see Table 3). State explicitly your assumptions about the distribution of the errors.

c. [2 points] Suppose we remove the variable birth weight from the model. Indicate whether R^2 will increase/decrease or remain the same. Motivate your answer.

d. [2 points] Describe two different plots that can be used to check the model assumptions. Indicate for each of the plots which assumption can be checked with that plot.

THE END

Table of t-quantiles

α df	0.6	0.7	0.75	0.8	0.85	0.9	0.925	0.95	0.975	0.98	0.99	0.999
1	0.32	0.73	1	1.38	1.96	3.08	4.17	6.31	12.71	15.89	31.82	318.31
2	0.29	0.62	0.82	1.06	1.39	1.89	2.28	2.92	4.3	4.85	6.96	22.33
3	0.28	0.58	0.76	0.98	1.25	1.64	1.92	2.35	3.18	3.48	4.54	10.21
4	0.27	0.57	0.74	0.94	1.19	1.53	1.78	2.13	2.78	3	3.75	7.17
5	0.27	0.56	0.73	0.92	1.16	1.48	1.7	2.02	2.57	2.76	3.36	5.89
6	0.26	0.55	0.72	0.91	1.13	1.44	1.65	1.94	2.45	2.61	3.14	5.21
7	0.26	0.55	0.71	0.9	1.12	1.41	1.62	1.89	2.36	2.52	3	4.79
8	0.26	0.55	0.71	0.89	1.11	1.4	1.59	1.86	2.31	2.45	2.9	4.5
9	0.26	0.54	0.7	0.88	1.1	1.38	1.57	1.83	2.26	2.4	2.82	4.3
10	0.26	0.54	0.7	0.88	1.09	1.37	1.56	1.81	2.23	2.36	2.76	4.14
11	0.26	0.54	0.7	0.88	1.09	1.36	1.55	1.8	2.2	2.33	2.72	4.02
12	0.26	0.54	0.7	0.87	1.08	1.36	1.54	1.78	2.18	2.3	2.68	3.93
13	0.26	0.54	0.69	0.87	1.08	1.35	1.53	1.77	2.16	2.28	2.65	3.85
14	0.26	0.54	0.69	0.87	1.08	1.35	1.52	1.76	2.14	2.26	2.62	3.79
15	0.26	0.54	0.69	0.87	1.07	1.34	1.52	1.75	2.13	2.25	2.6	3.73
16	0.26	0.54	0.69	0.86	1.07	1.34	1.51	1.75	2.12	2.24	2.58	3.69
17	0.26	0.53	0.69	0.86	1.07	1.33	1.51	1.74	2.11	2.22	2.57	3.65
18	0.26	0.53	0.69	0.86	1.07	1.33	1.5	1.73	2.1	2.21	2.55	3.61
19	0.26	0.53	0.69	0.86	1.07	1.33	1.5	1.73	2.09	2.2	2.54	3.58
20	0.26	0.53	0.69	0.86	1.06	1.33	1.5	1.72	2.09	2.2	2.53	3.55
21	0.26	0.53	0.69	0.86	1.06	1.32	1.49	1.72	2.08	2.19	2.52	3.53
22	0.26	0.53	0.69	0.86	1.06	1.32	1.49	1.72	2.07	2.18	2.51	3.5
23	0.26	0.53	0.69	0.86	1.06	1.32	1.49	1.71	2.07	2.18	2.5	3.48
24	0.26	0.53	0.68	0.86	1.06	1.32	1.49	1.71	2.06	2.17	2.49	3.47
25	0.26	0.53	0.68	0.86	1.06	1.32	1.49	1.71	2.06	2.17	2.49	3.45
26	0.26	0.53	0.68	0.86	1.06	1.31	1.48	1.71	2.06	2.16	2.48	3.43
27	0.26	0.53	0.68	0.86	1.06	1.31	1.48	1.7	2.05	2.16	2.47	3.42
28	0.26	0.53	0.68	0.85	1.06	1.31	1.48	1.7	2.05	2.15	2.47	3.41
29	0.26	0.53	0.68	0.85	1.06	1.31	1.48	1.7	2.05	2.15	2.46	3.4
30	0.26	0.53	0.68	0.85	1.05	1.31	1.48	1.7	2.04	2.15	2.46	3.39
31	0.26	0.53	0.68	0.85	1.05	1.31	1.48	1.7	2.04	2.14	2.45	3.37
32	0.26	0.53	0.68	0.85	1.05	1.31	1.47	1.69	2.04	2.14	2.45	3.37
33	0.26	0.53	0.68	0.85	1.05	1.31	1.47	1.69	2.03	2.14	2.44	3.36
34	0.26	0.53	0.68	0.85	1.05	1.31	1.47	1.69	2.03	2.14	2.44	3.35
35	0.26	0.53	0.68	0.85	1.05	1.31	1.47	1.69	2.03	2.13	2.44	3.34
36	0.26	0.53	0.68	0.85	1.05	1.31	1.47	1.69	2.03	2.13	2.43	3.33
37	0.26	0.53	0.68	0.85	1.05	1.3	1.47	1.69	2.03	2.13	2.43	3.33
38	0.26	0.53	0.68	0.85	1.05	1.3	1.47	1.69	2.02	2.13	2.43	3.32
39	0.26	0.53	0.68	0.85	1.05	1.3	1.47	1.68	2.02	2.12	2.43	3.31
40	0.26	0.53	0.68	0.85	1.05	1.3	1.47	1.68	2.02	2.12	2.42	3.31

Table 3: Quantiles of t-distributions with 1 to 40 degrees of freedom (df).