VU University	Statistical Data Analysis
Faculty of Sciences	29 May 2015

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!

PART 1

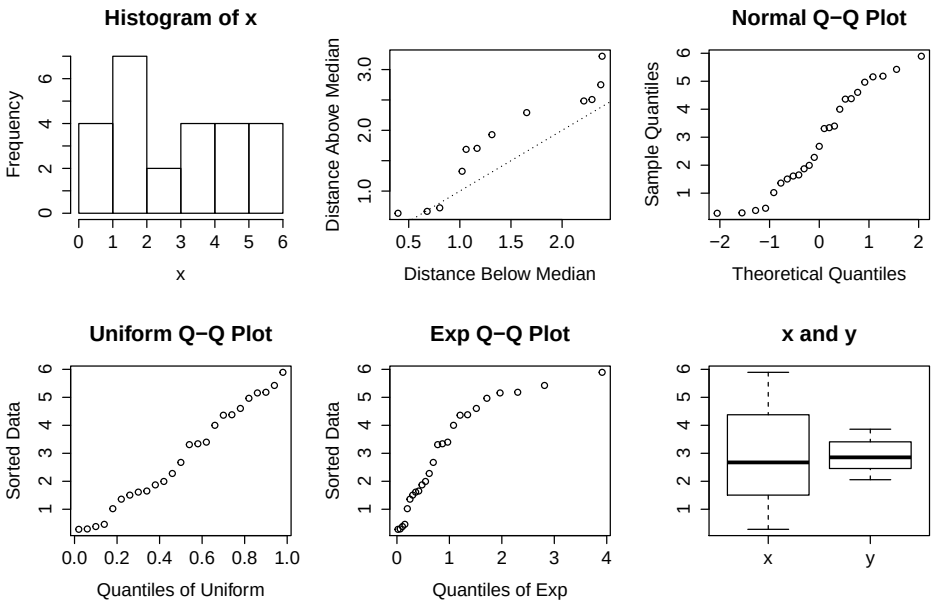

Figure 1: Histogram, symplot and QQ-plots of a data set x against the standard normal, standard uniform and standard exponential distributions. The bottom right plot is only necessary for Question 2 and shows boxplots of data sets x and y.

Question 1 [6 points]
In Figure 1 a histogram, symplot and several QQ-plots of 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 3 location-scale families do you think is most appropriate for 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.

Question 2 [4 points]
Consider the bottom right plot in Figure 1.

a. [2 points] Sketch the empirical distribution functions of data sets x and y in one figure.

b. [2 points] Could one use a chi-square goodness of fit test to test the null hypothesis that the underlying distributions of x and y are the same? Motivate your answer.

Question 3 [8 points]
Suppose we have $n = 30$ observations $X_1, \ldots, X_{30}$ from an unknown distribution P and suppose that we have two unbiased estimators $S = S(X_1, \ldots, X_{30})$ and $T = T(X_1, \ldots, X_{30})$ for the location of P. In order to investigate which of the two estimators S and T is more accurate, the variances of the two estimators are estimated by means of bootstrap.

a. Describe the steps of the bootstrap scheme that you would use to find the bootstrap estimates of the variances of S and T. Specify your answer

   i. [3 points] for the case that nothing is known about P

   ii. [3 points] for the case that we know that P is an exponential distribution with unknown parameter λ.

b. [2 points] Indicate which errors are made in the bootstrap procedure in either case. One of the errors is different in nature between the two cases. Explain this difference.

Part 2 starts on the next page

PART 2

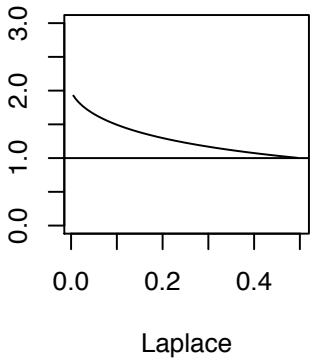

Figure 2: Asymptotic variance of α-trimmed mean as a function of α for the Laplace distribution. The horizontal line indicates the asymptotic Cramér-Rao lower bound.

Question 4 [6 points]

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

Are the following statements correct/sensible? Motivate your answer by a short argument.

b. [2 points] A linear regression model is linear in the explanatory variables. Hence, it cannot contain quadratic terms as in $Y = \beta_0 + \beta_1 x_1^2$.

c. [2 points] The sign test is preferred over Kendall's rank correlation test.

Question 5 is on the next page

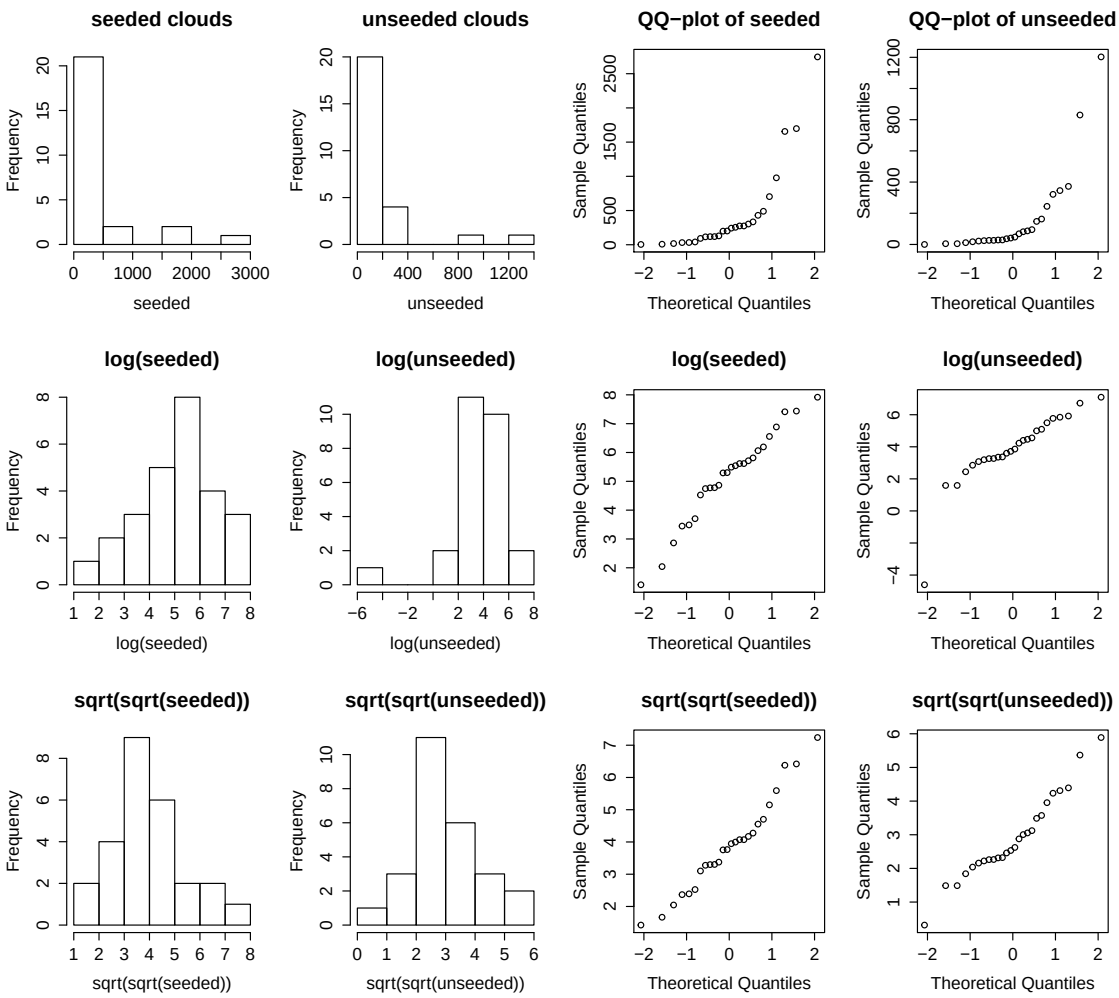

Figure 3: Graphical representations of the clouds data, and its transformations. The QQ-plots are against the $N(0,1)$ distribution.

Question 5 [6 points]
Consider the data on seeded clouds and unseeded clouds in Figure 3. To test the null hypothesis that these two samples come from the same distribution we have performed the two-sample t-test and the Wilcoxon two-sample test on the original data, and on the log transformed data and fourth root of the data. Some of the p-values of these tests are given in the table below.

data	t-test	Wilcoxon test
seeded vs. unseeded	0.0538	0.0138
log(seeded) vs. log(unseeded)	0.0181	p_1
$\sqrt[4]{\text{seeded}}$ vs. $\sqrt[4]{\text{unseeded}}$	0.0124	p_2

a. [2 points] Which p-values in the column under t-test do you trust? Motivate your answer.

b. [2 points] Indicate whether p_1 and p_2 are bigger, equal or smaller than 0.0138. Motivate your answer. (*Hint: consider the test statistic of the Wilcoxon two-sample test.*)

c. [2 points] State two other tests that could be applied to test the null

hypothesis above. Motivate the choice of these tests. Indicate the assumptions of your proposed tests.

Question 6 [6 points]
In a study on the relationship between the colors of helmets worn by motorcyclists and whether they are injured or killed in a crash, the following results were found:

	black	white	yellow/orange	red	blue	total
not injured	491	377	31	170	55	1124
injured or killed	213	112	8	70	26	429
total	704	489	39	240	81	1553

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 the colors of helmets and whether they are injured/killed. (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] What is the approximate distribution of the chi-square test statistic for these data under the null hypothesis?

d. [1 point] It appears that the chi-square test rejects the null hypothesis for these data with $p = 0.04$. How would you investigate in what way the data differ from what is expected under the null hypothesis?

Question 7 [9 points]

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

The new ICON E-Flyer is a brand new electrical bike, developed in the US, that can reach a speed of 60 km/h. We investigate how this maximum speed depends on the bodyweight of the biker and the windspeed in the opposite direction. We use the linear regression model:

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

where Y is the maximum speed (in km/h), x_1 the bodyweight of the biker (in kg) and x_2 the windspeed in the opposite direction (in Bft). Some scatter plots of data of 25 observations are shown in Figure 4. The model estimated by least squares using these $n = 25$ observations is

$$Y = 60.6 - 0.16x_1 - 1.91x_2 + e.$$

Furthermore: $R^2 = 0.825$, $se(\hat{\beta}_1) = 0.094$ and $se(\hat{\beta}_2) = 0.201$. ($se(\hat{\beta}_j)$ denotes estimated standard error of $\hat{\beta}_j$).

b. [3 points] Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at a level of 0.05 (see Table 1). State explicitly your assumptions about the distribution of the errors.

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

d. [2 points] Which assumption from parts (a) and (b) can be checked in the QQ-plot in Figure 4? Do you think this assumption is appropriate for these data? Motivate your answer and explain why this check is necessary.

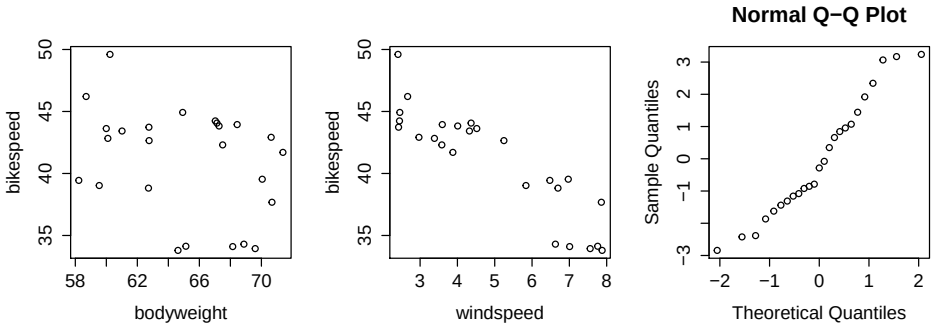

Figure 4: Two scatter plots of the bike data and a QQ-plot of the residuals of the stated model against the standard normal distribution.

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