VU university — Statistical Data Analysis

Faculty of Sciences — 31 May 2013

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

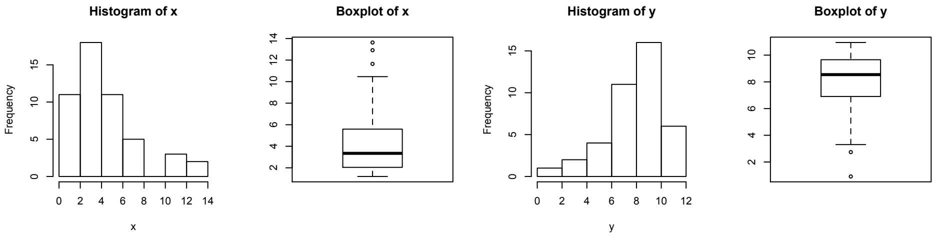

Figure 1: Histograms and boxplots of data sets x and y.

Question 1 [6 points]

Figure 1 shows a histogram and a boxplot for two data sets x and y.

a. [2 points] Describe briefly what these graphical summaries tell you about the underlying distributions of the two data sets. Consider (at least) the aspects location, scale, shape and extreme values.

b. [2 points] Decide for each of the two data sets whether the median will be larger, smaller, or approximately equal to the mean? Why?

c. [2 points] Give a sketch of the two-sample QQ-plot of x and y. Motivate your sketched plot.

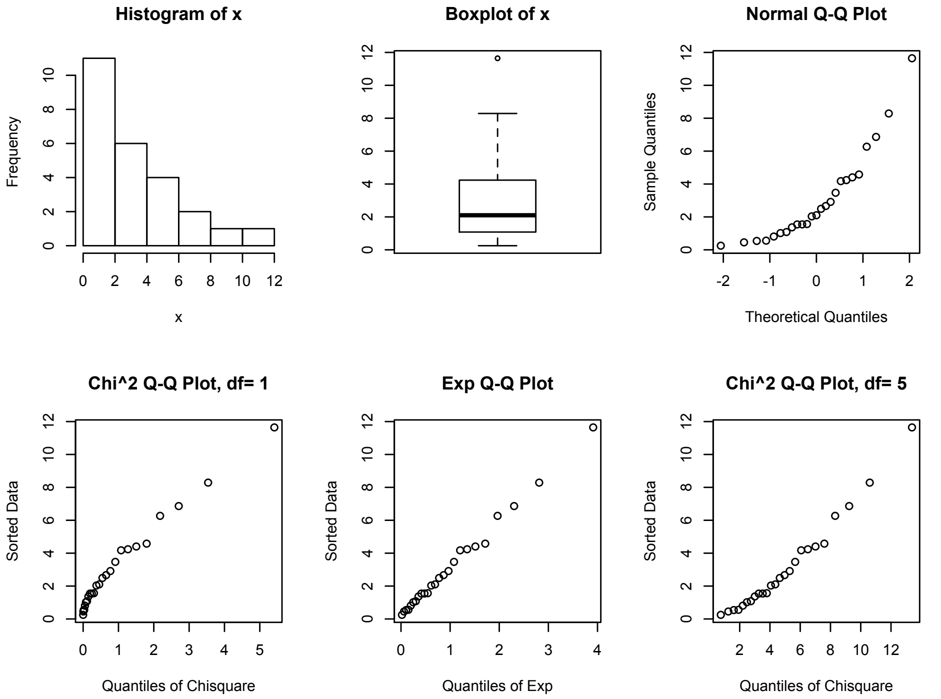

Figure 2: Histogram, boxplot and QQ-plots of data set x against the standard normal, χ_1^2, exponential and χ_5^2 distributions.

Question 2 [6 points]
In Figure 2 a histogram, boxplot and several QQ-plots of a data set x are presented.

a. [2 points] Which of the four location-scale families do you think is most appropriate for these data? Explain your answer.

b. [1 point] Using the QQ-plot of the location-scale family that you have selected under part (a) determine the location a and scale b approximately.

c. [3 points] Suppose we want to test the null hypothesis that the data $X_1, \ldots, X_n$ in Figure 2 are exponentially distributed, i.e. $H_0 : X_1, \ldots, X_n \sim \exp(\lambda)$ for some unknown $\lambda > 0$, versus the alternative hypothesis that the data do not follow an exponential distribution. Design a bootstrap test for this null hypothesis based on the test statistic $T = median(X_1, \ldots, X_n)/\overline{X}_n$. Describe the procedure to generate the bootstrap values and the p-value.

Question 3 [6 points]

a. [2 points] Give the general formula for a $(1-\alpha)$ bootstrap confidence interval for a parameter θ based on an estimator T and a data sample $X_1, \ldots, X_n$. Explain your notation.

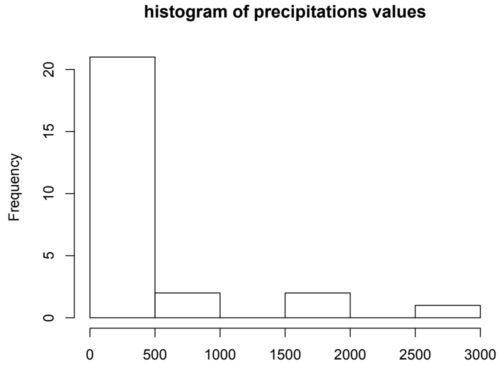

Figure 3: Histogram of precipitation values of the seeded clouds.

Consider the data in Figure 3 about precipitation values of seeded clouds. As estimates for spread we computed the sample MAD and the sample standard deviation. To assess the accuracy of these estimators, we determined 90% bootstrap confidence intervals for the spread based on both estimators. We found the following two intervals: [425, 976] and [122, 333].

b. [2 points] Which of the two intervals given above is for the MAD? Motivate your answer.

c. [2 points] Which estimator for spread is more appropriate for these data? Motivate your answer.

PART 2

Question 4 [6 points]
Are the following statements correct? Motivate your answer by a short argument.

a. [2 points] The asymptotic relative efficiency of one test with respect to another test depends on the underlying distribution of the data.

b. [2 points] Variance inflation factors are more informative for detecting influence points than hat values.

c. [2 points] The test statistic for the chisquare test for $k \times r$ contingency tables is

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{(N_{ij} - n\hat{p}_{ij})^2}{N_{ij}},$$

where N_{ij} is the observed cell count in cell (i, j), $\hat{p}_{ij}$ is the estimated cell probability for cell (i, j) and n is the total count.

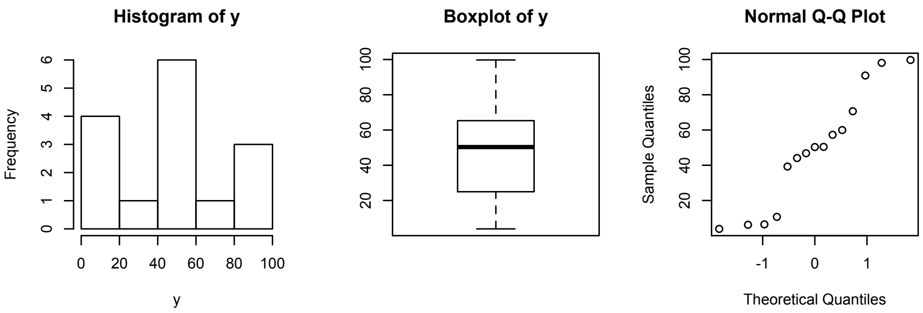

Figure 4: Histogram, boxplot and QQ-plots against the normal distribution of a data set.

Question 5 [6 points]
The data set presented in Figure 4 contains the following 15 numbers: 3.9, 6.2, 6.5, 10.7, 39.2, 44.1, 46.8, 50.3, 50.4, 57.3, 60.0, 70.6, 90.9, 98.1, 99.7. We test the null hypothesis $H_0 : m = 25$ that the median m of the underlying distribution is equal to 25 using the sign test.

a. [2 points] Formulate the test statistic for the sign test, and give its distribution under the assumption $m = 25$.

b. [2 points] Perform the sign test at significance level $\alpha = 0.05$ using Table 1. Give the p-value and the conclusion.

c. [2 points] Give an alternative test for testing the null hypothesis, which is appropriate for this data set. Motivate your answer and describe the assumptions of your proposed test.

	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 6 [6 points]

We visit 55 ice cream sellers and measure whether the number of bacteria in their vanilla ice cream complies with legal requirements on food safety. Moreover we ask each ice cream seller whether he/she has taken a training on food safety. The results we find are presented in the following table

	training	no training
number of bacteria		
in accordance	11	33
too high	5	6

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 fullfilling the legal requirements and having taken the training. (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 is fulfilled.

c. [1 point] Is the chi-square test applicable for these data? If not, what test would you use to test the null hypothesis stated in part (a)?

d. [1 point] Suppose that the null hypothesis of independence in a general $k \times r$ contingency table is rejected by the chi-square test. How would you investigate in what way the data differ from what is expected under the null hypothesis?

Question 7 [9 points]

a. [3 points] Formulate the general multiple linear regression model and its assumptions.

b. [3 points] For each assumption shortly describe a method to verify the validity of that assumption for a given data set.

c. [3 points] Consider the data shown in Figure 5 on the number of plant species on the Galápagos Islands. For these data we assume a linear regression model where the response variable is the number of plant species on an island (`Species`) and the available explanatory variables are `Area` (the area of the island), `Elevation` (the highest elevation of the island), `Scruz` (the distance to the Santa Cruz island) and `Adjacent` (the area of the adjacent island). There are 30 observations.
What problem(s) do you expect when the full model is fitted to these data? For each problem indicate at least one way you would investigate that problem.

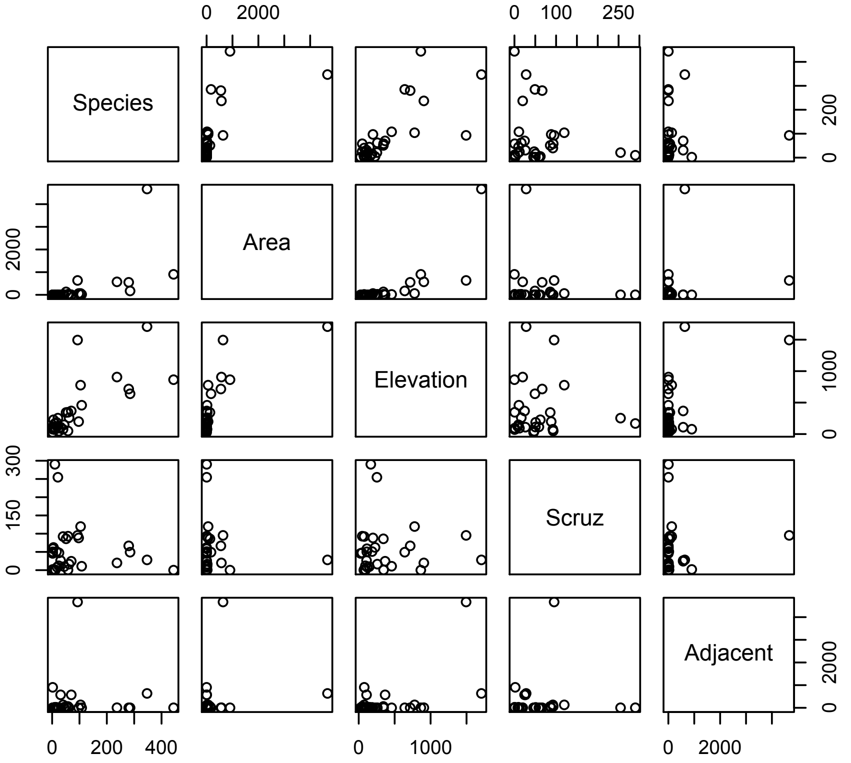

Figure 5: Scatter plots of the data of the Galápagos Islands.

THE END