Exam Statistical Data Analysis

VU University Amsterdam, Faculty of Exact Sciences

December 22, 2011

NB. Use of a basic calculator is allowed; graphical calculators, mobile phones, etc. are not allowed.

NB. The exam can be made in the language of your preference: English or Dutch.

The 8 questions below all have the same weight.

1. Are the following statements sensible/correct? Motivate your answer.

   a) For the chi-square test for goodness-of-fit it is recommended to choose the intervals such that the number of observed values in each interval is at least 5.

   b) Suppose that x is a sample from the Unif[0,8]-distribution, and y from the $N(4,2)$-distribution. The Wilcoxon two-sample test applied to x and y will find a significant difference between the underlying distributions of x and y better than the Kolmogorov-Smirnov test applied to x and y.

   c) A two-sample permutation test is in fact a bootstrap test.

   d) To test whether an explanatory variable should be included in a multiple linear regression model with uncorrelated measurement errors, a t-test can be used.

2. a) Figure 1 shows a histogram and a boxplot for two data sets x and y. 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) For each of the two data sets: will the median be larger, smaller, or approximately equal to the mean? Why?

   c) Based on your conclusions about the underlying distributions, which test would be appropriate to test the null hypothesis that the underlying distributions of x and y differ? Explain your answer briefly.

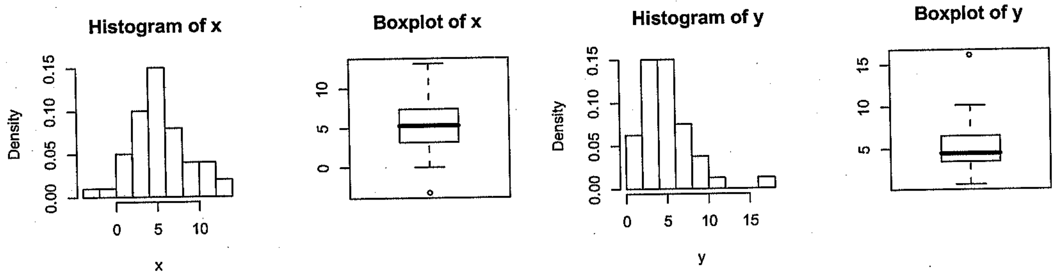

Figure 1: Histograms and boxplots of data sets x and y.

3. In Figure 2 four QQ-plots of the same data set are given: quantiles of the data against quantiles of the $N(0,1)$, $Unif(0,1)$, $Exp(1)$, and the standard χ^2_5 distribution.

   a) Which of the four location-scale families—of $N(0,1)$, $Unif(0,1)$, $Exp(1)$, or of the standard χ^2_5 distribution—suits these data best in your opinion? Why?

   b) Estimate by eye the location parameter a and the scale parameter b corresponding to the family that you chose in part a).

   c) Determine estimates of the expectation and standard deviation of the underlying distribution of the data based on your estimates of part b).
   (*Note: a $Unif(0,1)$ has expectation $1/2$ and variance $1/12$; an $Exp(1)$ distributed random variable has expectation and variance equal to 1; a standard χ^2_k distributed random variable has expectation k and variance $2k$.*)

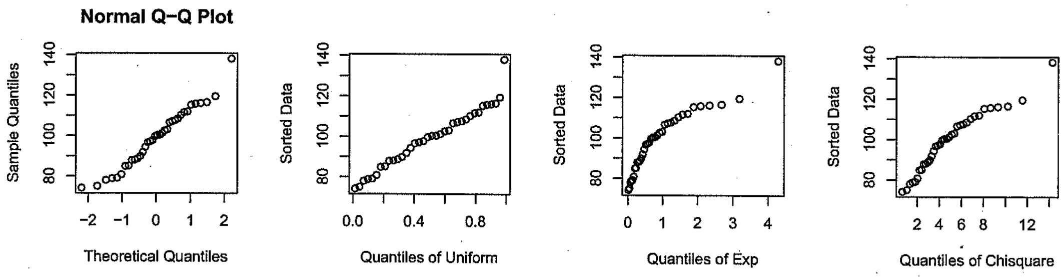

Figure 2: Four different QQ-plots of a data set.

4. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with unknown distribution P. Suppose that $T_n(X_1, \ldots, X_n) = Median(X_1, \ldots, X_n)$ 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) Describe the steps of the bootstrap scheme that you would use to find the bootstrap estimate of the standard deviation of T_n by computer simulation for the case that nothing is known about the unknown distribution P.

   b) Describe shortly which two errors are made in the procedure that you described in part a). Which of the two can be made arbitrarily small by the user.

   c) What would be different from the scheme in part a) if it would be known that the data come from an exponential distribution with unknown parameter λ?

5. a) Which estimator would you use for estimating the spread of the underlying distributions of the data sets of annual income in two countries depicted in Figure 3? Motivate your answer.

b) Suppose that the *difference* in spread of the underlying distributions of the two data sets is estimated by the difference of the estimator that you chose in part a) for the two countries, and denote this difference estimator by T_d. Suppose that the observed value of T_d is $t_d = 6.0$, and that the following quantiles of 1000 bootstrap values of the estimator T_d were obtained.

quantile	0.025	0.05	0.5	0.95	0.975
	1.5	2.9	6.1	10.7	12.2

Compute a 95% bootstrap confidence interval (as defined in the Reader) for the difference in spread of the underlying distributions.

c) Based on the confidence interval that you computed in part b), is there a difference in spread of the underlying distributions of the two data sets? Explain your answer.

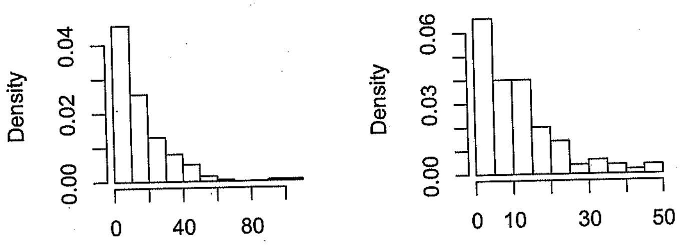

Figure 3: Histograms of annual income in $1000 in two countries.

	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.

6. One of the filling machines in a soda factory is suspected of putting the wrong amount of soda into the soda cans. To investigate this the amount of soda in 12 cans of a day's production was measured. The cans are supposed to contain 33.00 cl of soda. The following sorted amounts (in cl) were measured:
32.41, 32.46, 32.72, 32.81, 32.86, 32.91, 32.92, 32.93, 32.98, 33.05, 33.15, 33.36.

   a) The problem was investigated by performing a sign test on these data. Perform this sign test, i.e. formulate H_0 and H_1, give the formula for the test statistic and its distribution under H_0, give the p-value and the conclusion of the test. Take significance level $\alpha = 0.1$ and use Table 1.

   b) Could the following four tests also be used for investigating the problem? If yes, which assumption(s) concerning the underlying distribution of the data should be made for the test; if no, why not?
   i) a (Wilcoxon) signed rank test;
   ii) a Kolmogorov-Smirnov-test;
   iii) a t-test;
   iv) a Shapiro-Wilk test.

7. A study seeks to determine whether vitamin C has an effect on preventing colds. Among a sample of 220 people, 105 randomly selected people took a vitamin C pill daily for a period of 10 weeks and the remaining 115 people took a placebo daily for 10 weeks. At the end of the 10 weeks the number of people who got colds was recorded. The data were

	cold	no cold	total
vitamin C	45	60	**105**
placebo	75	40	**115**
total	**120**	**100**	**220**

a) Specify a suitable model and state the corresponding null and alternative hypothesis for investigating with a chi-square test whether there is a relationship between taking vitamin C and getting colds. (You may give your answer in formulas or in words.)

b) Suppose that the null hypothesis in part a) is rejected. Shortly describe one of the methods for investigating whether this is due to the fact that taking vitamin C helps preventing people from getting colds.

c) With these data the claim that taking vitamin C helps preventing people from getting colds, can also be investigated with Fisher's exact test. Formulate for this test the null and the alternative hypothesis, state the necessary assumptions on the marginals, give the test statistic and its distribution under the null hypothesis, and indicate when the null hypothesis will be rejected. (You do not need to specify parameter values of the distribution.)

8. To investigate the economic growth of a country the variable Gross domestic product (GDP), the market value of all final goods and services produced within a country in a given period, is often used. In a study of the dependence of GPD on several other economic variables, multiple linear regression analysis was used on data from 36 countries. It was found that only the variables Expenditure (EXP) and government's debt (DEBT) had significant influence on GPD.

a) Formulate the resulting final multiple linear regression model of the study, including its assumptions; explain the notation that you use in terms of the context.

b) Describe shortly how the model assumptions can be checked.

c) Give the definition of the determination coefficient in terms of sums of squares, and explain briefly what this coefficient measures within the given context.