Exam Statistical Data Analysis

VU University Amsterdam, Faculty of Exact Sciences

December 15, 2009

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

Note: This exam has a separate Appendix with Figures and Tables.

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

The 9 questions below all have the same weight.

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

   a) A boxplot yields more information than a stem-and-leaf plot.

   b) For the chi-square test for goodness-of-fit the intervals should be selected such that there are at least 5 observed values in each interval.

   c) The influence function for the mean is bounded for $x \to \pm\infty$.

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

2. The idea is that people with diabetes are often much too heavy. According to the Voedingscentrum (Food Center) people older than 19 with a body mass index (BMI) larger than 30 are much too heavy.

   a) In Figure 1 a histogram, a boxplot and a normal QQ-plot are given of the BMI of 25 women older than 19 with diabetes. Assume that the data form a representative sample of the population of women older than 19 with diabetes. Describe briefly what these graphical summaries tell you about the underlying distribution of the data. Consider at least the aspects location, scale, shape and extreme values.

   b) Which estimator for location would you choose for these data and why?

   c) Based on your conclusions about the underlying distribution, which test would you use to test whether or not the location of the underlying distribution is larger than 30?

3. The data presented in Figure 2 are annual incomes in US dollars for 200 families in the US. In the figure six plots are given: a histogram, a boxplot and QQ-plots against quantiles of the N(0,1), Exp(1), the standard χ^2_1 and the standard χ^2_5 distributions.

a) Which location-scale family suits these data best in your opinion? Why?

b) Give an appropriate distribution for these income data. Indicate both location and scale. (*You may use that for an Exp(1) distributed random variable the expectation and variance are equal to 1, and that a standard χ^2_k distributed random variable has expectation k and variance $2k$.*)

4. a) Describe shortly the essential differences between the empirical bootstrap and the parametric bootstrap method.

b) Suppose that an estimate for the standard deviation of an estimator T for θ needs to be computed with the bootstrap. When would you use the empirical and when the parametric bootstrap method?

c) Is the following bootstrap scheme appropriate for testing the null hypothesis H_0: "the underlying distribution of a data set $x_1, \ldots, x_n$ is normally distributed"?

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 $p_r = \#(D^* \geq d)/B$.

If you think that the scheme is not suitable, indicate where the error is, and how to fix it.

5. The income data in Question 3 are actually data from 100 white families and 100 black families in the US. In Figure 3 the histograms of the separate data sets are shown (in \$1000). We want to investigate whether there is more spread in one data set than in the other. The values for the sample standard deviation are $s_w = 18824$ for the white families, and $s_b = 10811$ for the black families. To determine how accurate these estimates of spread are, for each of the two data sets 1000 bootstrap values for the sample standard deviation were generated. In Table 1 some quantiles of the two sets of bootstrap values are given.

Can you deduce from the given information whether there is more spread in one data set than in the other? If so, for which families do you find more spread in their incomes?

6. 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 bottles of a day's production was measured. The bottles are supposed to contain 33.00 cl of beer. The following, sorted, amounts (in cl) 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) Formulate H_0 and H_1 for the test.

   b) Give the formula for the test statistic.

   c) What is the distribution of the test statistic under H_0?

   d) Perform the test with significance level $\alpha = 0.05$ using Table 2. Give the p-value and the conclusion of the test.

7. In Amsterdam, Copenhagen, Berlin, London and Paris a random sample of size 1000 was taken from adults between 20 and 60 years old and each person was asked whether he or she is concerned, not concerned or has no opinion about global warming. To investigate whether or not the opinions on global warming differ in the five cities one could apply a chi-square test to these data.

   a) What are the model and the corresponding null and alternative hypothesis for investigating with a chi-square test whether or not the opinions on global warming differ in the five cities? (You may give your answer in words, instead of in formulas).

   b) If the null hypothesis of part a) is rejected, how could it be investigated what the differences are between the cities?

8. To determine the dependence of the amount of nicotine in cigarettes on the amount of tar and carbon monoxide, for 25 different brands of cigarettes the amount of nicotine, tar and carbon monoxide was determined. A multivariate linear regression model is used to model the dependence.

   a) Formulate the multivariate linear regression model, including its assumptions, for this situation; explain the notation that you use in terms of the context.

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

   c) Give the general definition in terms of sums of squares for the determination coefficient and explain briefly what this coefficient measures in terms of the cigarette context.

9. a) Give a one sentence description for each of the following four concepts that can cause problems in a multivariate regression analysis: outlier, 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.

b) Suppose that for the cigarette data of Question 8 the 6th brand has a remarkably high amount of nicotine. To investigate whether this value is an outlier, the model in part 8 a) needs to be extended to a mean-shift-outlier model. Formulate this extended model.

c) 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.)

Figures and Tables for Exam Statistical Data Analysis

VU University Amsterdam, Faculty of Exact Sciences

December 15, 2009

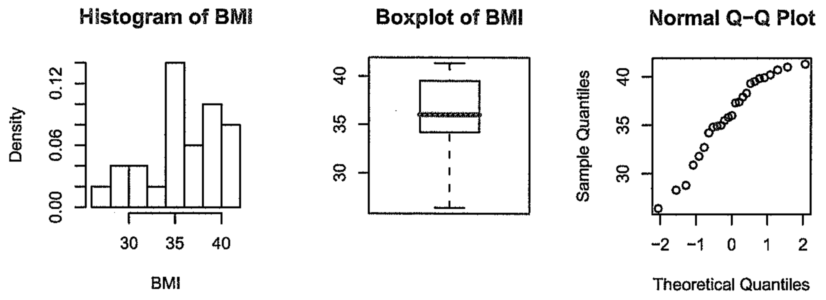

Figure 1: Histogram, boxplot and normal QQ-plot for the BMI data. (The whiskers in the boxplot connect the box with the most extreme data points that lie not more than 1.5 times the interquartile range from the edges of the box.)

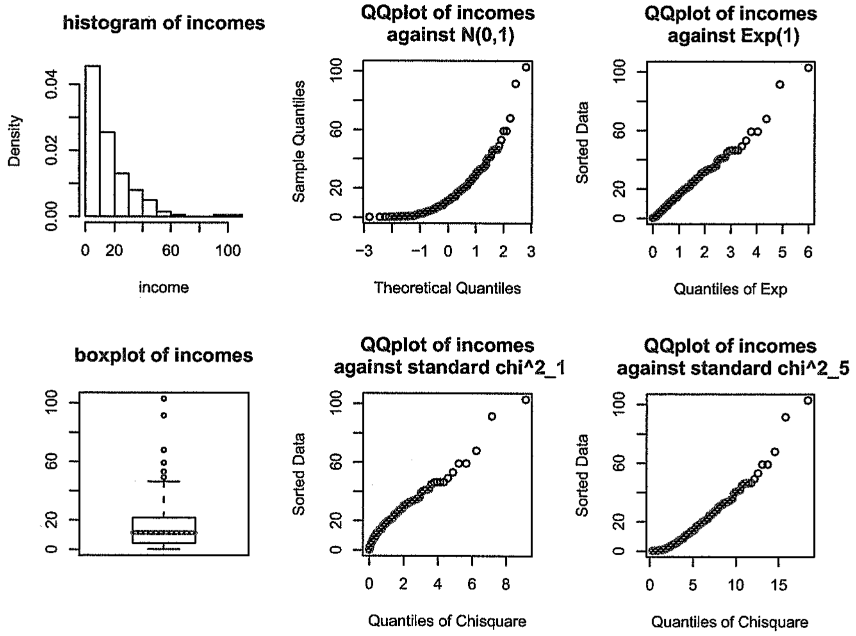

Figure 2: Plots on annual income (in $1000) for 200 families in the US.

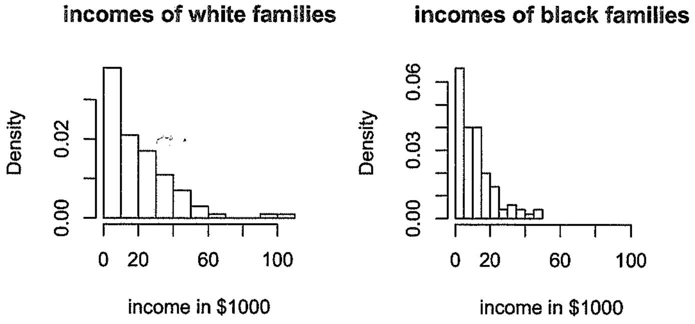

Figure 3: Plots on annual income (in $1000) for 100 white and 100 black families in the US.

	0.025	0.05	0.5	0.95	0.975
bootstrap values sd white	14263	14810	18330	22319	23110
bootstrap values sd black	8504	8886	10626	12476	12794

Table 1: Quantiles of the sets of bootstrap values for income data.

	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 2: 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.