Exam Statistical Data Analysis

VU University Amsterdam, Faculty of Exact Sciences
February 9, 2010

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) For testing whether the location of the distribution of a sample is equal to a certain value, it is better to perform a Wilcoxon signed rank test than a sign test, because the Wilcoxon signed rank test uses more information of the data than the sign test.

   b) For testing whether a sample originates from a normal distribution the Kolmogorov-Smirnov test for goodness of fit can be used.

   c) In general it is a good idea to use a 10% or 20% trimmed mean for estimating the location of a distribution.

   d) With Kendall's tau linear correlation between two samples can be tested.

2. In a study of the effects of passive smoking the level of serum cotinine, a product of nicotine, in the blood of three groups of, otherwise comparable, people was measured: smokers, passive smokers and non-smokers.

   a) In Figure 1 histograms and boxplots of the measurements are given. Describe briefly what these graphical summaries tell you about the underlying distributions 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 distributions, which test would you used to test whether or not the location of the underlying distributions of the data for the passive smokers and the non-smokers are the same? Why?

3. The data presented in Figure 2 are sales data from the "Wereldwinkel-VU" on one weekday during 1990-1991. The mean of the data is 63.4, and the variance 839.3. In the figure four QQ-plots are given: the data against quantiles of the N(0,1), Unif(0,1), Exp(1), and the standard χ^2_8 distributions.

   a) Describe the form of the distribution of the data based on these QQ-plots.

   b) Which of the 4 location-scale families—of N(0,1), Unif(0,1), Exp(1), or of the standard χ^2_8 distribution—suits these data best in your opinion? Why?

   c) Which distribution from the family that you chose in part b) would be most appropriate? (Indicate both location and scale.) (*Note: 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 $2k$.*)

4. Consider the situation where one has n=25 observations $X_1, \ldots, X_{25}$ from an unknown distribution P and suppose that two unbiased estimators $T_{1,n} = T_{1,n}(X_1, \ldots, X_{25})$ and $T_{2,n} = T_{2,n}(X_1, \ldots, X_{25})$ are proposed for estimating the location of P. To investigate which of the two estimators is a more accurate estimator in this situation, the variances of the two estimators are estimated by means of the bootstrap. Describe the steps of the bootstrap scheme that you would use to find the bootstrap estimates of the variance of the two estimators by simulation with the computer. Do this

   a) for the case that nothing is known about the distribution P.

   b) for the case that it is known that P is an exponential distribution with unknown parameter λ.

5. a) When is it advisable to use a robust estimator for location?

   b) Sketch the influence functions of the mean and the trimmed mean.

   c) Which of the two estimators under b) is more robust? Why?

6. To investigate whether the belief that first year students gain weight is true, students were measured in September and the following April. The differences (in kg) of the weights (weight in April − weight in September) of 14 students were -1, 1, 4, 3, 0, -2, 1, 5, 8, 3, 1, 0, -3, 2. 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 1. Give the p-value and the conclusion of the test.

7. A baseball team that played in a recent baseball World Series wanted to close the roof on their domed stadium, so that fans could make noise and give the team a better advantage. However, the team was not allowed to do so, unless weather conditions justified closing the roof. For a number of games in the stadium the result (won or lost) and the position of the roof during that game (open or closed), were registered.

   a) Which are the model and the corresponding null and alternative hypothesis for investigating whether there really is a relationship between the variables 'result of the game' and 'roof position' with a chi-square test? (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 whether or not closing the roof had a positive effect?

8. Concerns about global warming have led to studies of the influence of the concentration of CO_2 and of other air pollutants, like NO_2 and SO_2 on global temperature. Assume that 50 measurements at different locations of these 4 variables are available.

   a) Formulate a multivariate linear regression model, including its assumptions, with the three air pollutants as explanatory variables. Explain the notation that you use in terms of the context.

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

   c) Describe shortly what collinearity means in the context of this model and these variables. Name (not explain) at least two methods, like test, measure or other tool that can be used to search for collinearrity.

9. a) In the context of linear regression, which problem(s) can be caused by outliers?

   b) Suppose that for the air pollution data of Question 8 the 3rd observation point has a remarkably high amount of CO_2. 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 t-test that can be performed for the extended model to decide whether or not the 3rd 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

February 9, 2010

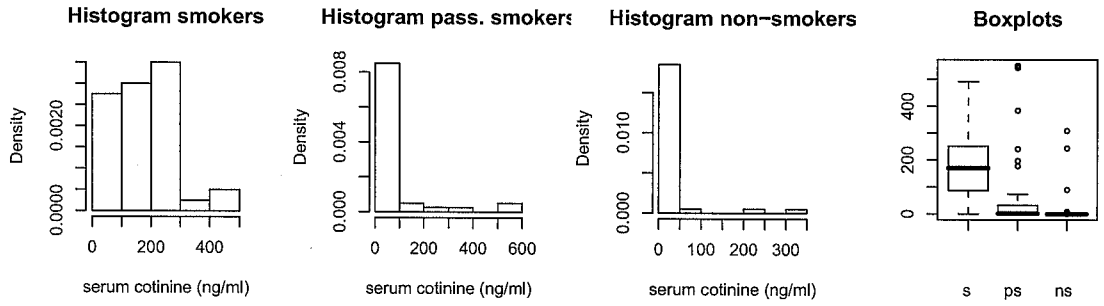

Figure 1: Histograms and boxplots of serum cotinine levels for smokers, "environmental" smokers and non-smokers.

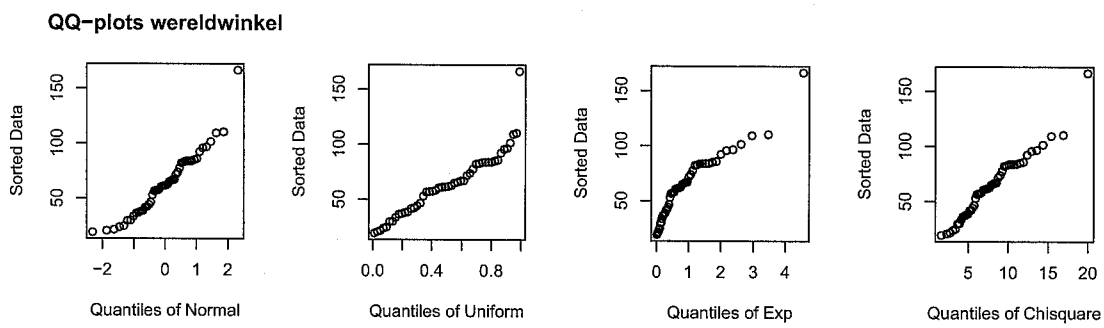

Figure 2: QQ-plots of sales.

	n			
k	4	8	12	14
0	0.062	0.004	0.000	0.000
1	0.313	0.035	0.003	0.001
2	0.688	0.145	0.019	0.006
3	0.938	0.363	0.073	0.029
4	1.000	0.637	0.194	0.090
5		0.855	0.387	0.212
6		0.965	0.613	0.395
7		0.996	0.806	0.605
8		1.000	0.927	0.788
9			0.981	0.910
10			0.997	0.971
11			1.000	0.994
12			1.000	0.999
13				1.000
14				1.000

Table 1: Probabilities $P(X \leq k)$ for binomially distributed random variable X with parameters n, as given in the table, and $p = 0.5$, for different values of k.