Test Mathematical Statistics (M5-201800139), October 25, 2019, 8.45-11.45 h. Instructors Johannes Schmidt-Hieber and Dick Meijer, module coordinator Pranab Mandal

A formula sheet is added. A regular scientific calculator is allowed, a programmable calculator ("GR") is not. No tables of probability distributions are needed for the test.

Part A: Basic concepts

(a) [1 point] What is a statistic?

(b) [1 point] Give the formula for the multiple linear regression model.

(c) [1 point] What is the difference between an explanatory and a response variables?

(d) [2 points] Describe a hypothesis test and two tools from descriptive statistics that can be used to check normality of data.

(e) [1 point] Describe the testing problem underlying Wilcoxon's rank sum test.

(f) [1 point] Give the definitions for the type II error and the power of a test.

(g) [3 points] Wilcoxon's rank sum test is based on an analysis of the ranks. Given a sample of independent and identically distributed random variables $(Z_1, \ldots, Z_N)$, denote by $R(Z_1), \ldots, R(Z_N)$ the ranks of $Z_1, \ldots, Z_N$. Show that

$$\operatorname{Cov}\big(R(Z_1), R(Z_2)\big) = -\frac{(N+1)}{12}.$$

Part B: Visual interpretation of data

1. [1 point] Which boxplot corresponds to which histogram?

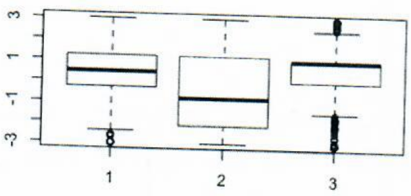


Fig. A

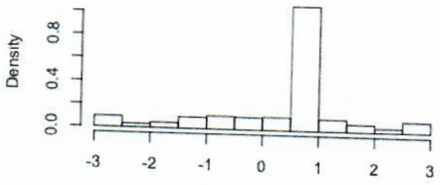


Fig. B

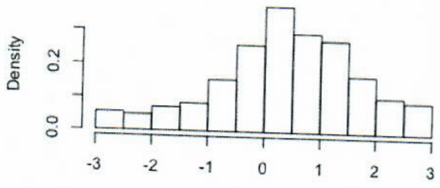
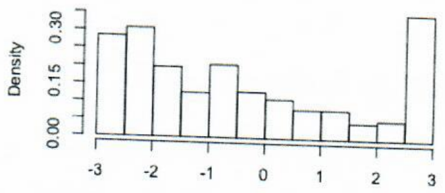


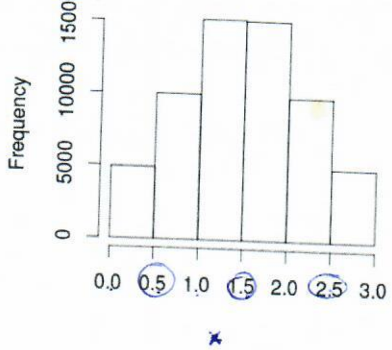
Fig. C



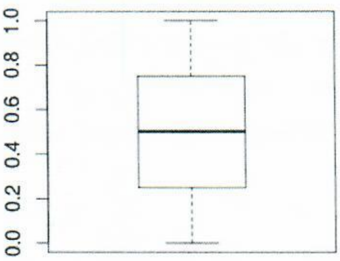
2. A distribution can be visualized using histograms or boxplots.

(a) [2 points] Given the histogram below, draw the corresponding boxplot (ignoring potential outliers).

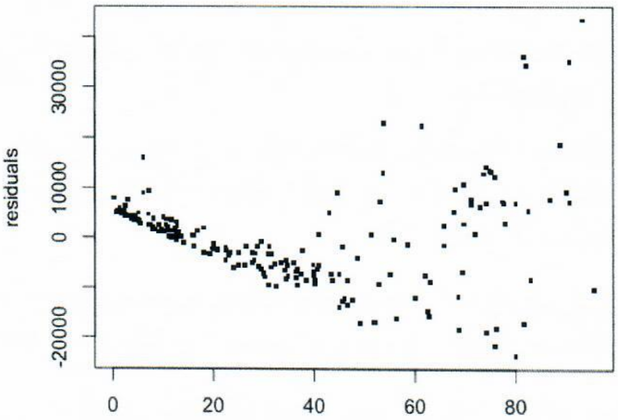
Histogram



(b) [2 points] Given the boxplot below, draw a histogram representing the same data [Histograms are more informative than boxplots. Therefore, given a boxplot there could be many different histograms. Your task is to draw one possible histogram with at least four bins.]



3. [1 point] The plot below shows the residuals of a linear regression fit. Discuss whether the linear model describes the data well.



Part C: Theory

4. Suppose we observe two independent random variables $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(-\mu, \sigma^2)$.

   (a) [3 points] Derive the MLE for (μ, σ^2).

   (b) [1 point] Compute the bias for the MLE of the variance σ^2.

5. [2 points] Suppose we observe independent random variables $X_1, \ldots, X_n$ ~Exp(λ). Show that the likelihood ratio test for $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda = \lambda_1$ with $0 < \lambda_0 < \lambda_1 < \infty$ rejects the null hypothesis if and only if

$$\sum_{i=1}^{n} X_i \leq c_\alpha$$

for a constant c_α that is independent of the data.

Hint: The p.d.f. of X ~Exp(λ) is $f_X(x) = 0$ for $x < 0$ and

$$f_X(x) = \lambda e^{-\lambda x}, \quad \text{for } x \geq 0.$$

6. In a representative survey with sample size 100, 58% of the persons answered that they will vote for person X in the next election. Suppose we are interested in testing whether this is significantly more than half of the population.

   (a) [2 points] State the null and alternative hypothesis and say which statistical test should be applied here.

   (b) [3 points] Derive the test statistic for this testing problem. Can we reject the null hypothesis at level $\alpha = 0.05$? *Remark: It is fine to do the computations without continuity correction.*

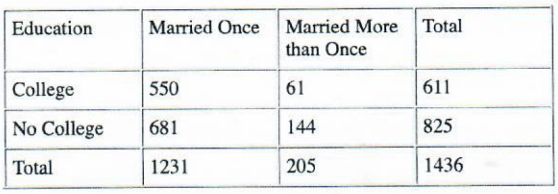
Hint: For the computations, you are allowed to approximate the binomial distribution by the normal distribution. You can also use the fact that $\Phi^{(-1)}(0.95) \approx 1.64$.

7. Suppose we observe $(x_1, Y_1), \ldots, (x_n, Y_n)$ for independent $Y_i \sim \mathcal{N}(\theta x_i^2, 1)$, $i = 1, \ldots, n$ and $x_1, \ldots, x_n$ deterministic.

   (a) [2 points] Compute the maximum likelihood estimator $\widehat{\theta}$ for the parameter θ.

   (b) [2 points] Find the sampling distribution of $\widehat{\theta}$.

8. We want to study the relationship between marital status and educational level. From a study of 1436 married women we obtain the following table:



Education	Married Once	Married More than Once	Total
College	550	61	611
No College	681	144	825
Total	1231	205	1436

   (a) [1 point] How can we decide using statistics whether women who did not attend college have a significantly higher chance to be married more than once if compared to women with college attendance?

   (b) [Bonus +2 points] Write down the test statistic and the rejection region for given significance level α.

9. The mathematical statistics course instructor thinks about designing the final test as a multiple choice quiz with n questions, where for each question four possible answers are provided and exactly one of them is correct whereas the other three are wrong. The FTS (final test score) is computed by counting the number of correctly answered questions divided by n (the total number of questions). Clearly, the FTS will always be a number between 0 and 1. Suppose there is a student RNC (which stands for really no clue). This student does not know anything, so RNCs strategy is to just randomly select one of the possible answers of each question.

   (a) [2 points] Suppose that in total there are 48 questions in the test (so $n = 48$). By making a link to sampling distributions and normal approximations, compute explicitly an interval which contains the FTS of this student with probability approximately 0.95. You can use that $\Phi^{-1}(0.025) = -\Phi^{-1}(0.975) \approx$ ~~2~~ -2

   (b) [2 points] Show that at least 300 questions have to be posed if one wants to make sure that with probability 97.5% student RNC has FTS≤ 0.3 (at most 30% correct answers).

   (c) [1 point] Suppose that there are 40 students in the mathematical statistics class, none of them paying attention. Therefore, all of them follow the same strategy as student RNC and select all answers randomly. If the multiple choice test has 300 questions, could it still happen that one of the students is lucky and has more than 30% of the correct answers? Discuss.

Now, let us introduce another student, in the following called AHL (attends half of the lectures). Because this student only attends half of all lectures, for every question, AHL can exclude correctly two of the possible four answers. Consequently, for every question AHL has to choose between two remaining answers and AHL does so by tossing a fair coin.

(d) [2 points] Suppose that the final test has $n = 63$ multiple choice questions. Compute explicitly an interval which contains the difference of the FTS for student RNC and AHL with probability approximately 0.95. Interpret the result.