Test Mathematical Statistics (Module 5),
October 29, 2021, 8.45-11.45 h.

A formula sheet is added. A regular scientific calculator is allowed, a programmable calculator with graphical interface is not.

Part A: Basic concepts

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

(b) [1 point] During the lecture, we worked with two estimators for the variance of i.i.d. observations. One estimator has a normalization factor $1/n$ and the other has a normalization factor $1/(n-1)$. Provide a statistical reason why both estimators are of interest.

(c) [2 points] Provide a real world example for a multiple regression model. Suppose that for none of the slopes it can be shown that they are significantly different from zero. What does this then mean for the example that you picked?

(d) [2 points] For a random variable X with mean μ and variance σ^2, the kurtosis is defined as $E(X-\mu)^4/\sigma^4$. What does the kurtosis measure? Provide a distribution for which the kurtosis is zero.

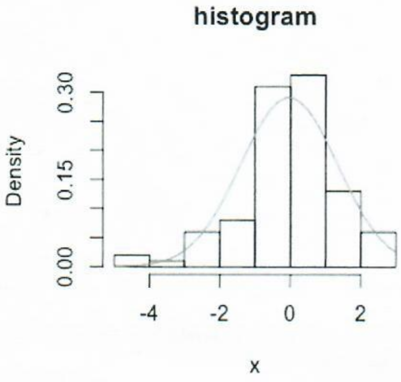
(e) [1 point] Which hypothesis test can be used to check Benford's law?

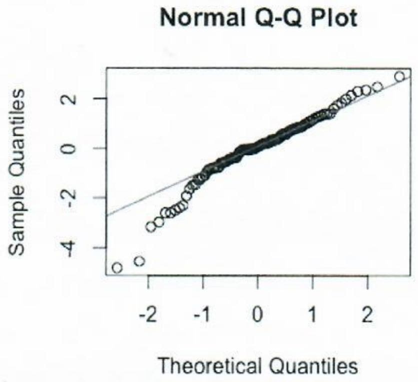
(f) [1 point] Consider a hypothesis testing problem with simple alternative. Give the definitions for the type II error and the power of a test.

(g) [2 points] Explain in at most two sentences what a permutation test is.

Part B: Visual interpretation of data

1. [1 point] To check whether the observation of a dataset follow a normal distribution, we compare the histogram with the probability density function (p.d.f.) of the normal distribution and create a normal qq-plot. These plots are given below. Moreover we run Shapiro-Wilks test and obtain the value 0.94 for the test statistic and a p-value of 0.008. What can be concluded from all these informations regarding the original question whether the observations follow a normal distribution?





2. [2 points] Draw a scatter plot for a sample of two negatively correlated variables. In a second step add one outlier to the dataset, such that the regression line has positive slope.

Part C: Theory

3. Recall that a geometric distribution with parameter $p \in [0, 1]$ has p.m.f. $x \mapsto (1-p)^{x-1}p$ for $x = 1, 2, 3, \ldots$ Assuming that we observe n independent random variables $X_1, \ldots, X_n$ drawn from a geometric distribution with unknown parameter p,

   (a) [1 point] derive the likelihood function for this model

   (b) [3 point] show that $n/\sum_{i=1}^n X_i$ is the MLE for p

   (c) [2 points] show that for $n = 1$ the MLE is biased

4. [2 points] Denote by P the uniform distribution on the interval $[-1, 1]$ and denote by Q the distribution with p.d.f. $q(x) = \max(0, 1 - |x|)$. Given one observation X, show that the most powerful level-α test for $H_0 : X \sim P$ versus $H_1 : X \sim Q$ rejects the null hypotheses if and only if $|X| \leq c_\alpha$, for a constant c_α that is independent of the data.

5. Has the success rate to complete a mathematics master changed significantly? A study finds that in 2010, 23 out of 35 students completed the master mathematics while in 2020, 32 out of 40 students completed the master. We wonder whether there is a significant difference.

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

   (b) [2 points] To construct a test statistic, we follow the strategy outlined in the lecture. For that we first need an estimator. Which estimator should we use and what is the sampling distribution of the estimator?

   (c) [3 points] Derive the test statistic. For the data above, what is the value of the test statistic?

   Hint: For the computations, you are allowed to approximate the binomial distribution $B(n, p)$ by the normal distribution $N(np, np(1-p))$.

6. Suppose we observe $(x_1, Y_1), \ldots, (x_n, Y_n)$ for independent $Y_i \sim N(\theta x_i^2, x_i^4)$, $i = 1, \ldots, n$ and deterministic and non-zero $x_1, \ldots, x_n$.

   (a) [2 points] Rephrase the problem such that the data are i.i.d. and the method of moments becomes applicable. Derive the moment estimator $\widehat{\theta}$ for the parameter θ.

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

7. Study participants are sometimes reluctant to share sensitive information as they are concerned about their privacy. Suppose we conduct a study asking study participants whether they feel addicted to alcohol. For each individual, the possible outcome is either yes or no. Suppose that the outcomes are jointly independent and that yes occurs with probability p and no with probability $1-p$, with p the unknown proportion parameter of alcohol addicts in the population. To protect the privacy, we propose the following scheme: Roll a die once. If you get "1" or "2", report then the wrong answer (if you feel addicted to alcohol choose no and if you do not feel addicted to alcohol choose yes). If you get to see a number larger than "2" report the right answer. Based on a sample $X_1, \ldots, X_n$ generated from this mechanism,

   (a) [2 points] show that $P(X_i = \text{yes}) = (1+p)/3$ and $P(X_i = \text{no}) = (2-p)/3$

   (b) [2 points] propose an unbiased estimator for the true proportion of alcohol addiction p

   (c) [2 points] for this estimator compute the variance and the MSE

   (d) [2 points] suppose every study participant would have given the correct answer. Then we would have simply taken the relative frequency as estimator for p. Compare the MSE obtained for the scheme above with the MSE of the relative frequency estimator.

[Similar mechanisms are applied nowadays to secure the privacy of customers and internet users. In this case, the additional randomness is generated by a computer.]