Exam 1 – Statistical Data Analysis

March 22, 2024

Exam Info

- The use of a basic calculator is allowed
- Graphical calculators and mobile phones are *not* allowed
- Please write all of your answers in English
- This exam consists of 3 questions on 3 pages
- This exam is worth a total of 27 points
- Denoting by τ the number of achieved points, the exam grade is $\frac{\tau+3}{3}$
- You have 120 minutes to write the exam

Good Luck ☺

Question 1 [10 points]

Mark each of the following statements as *True* or *False* and shortly motivate your answers (even for the statements you deem to be true). *Note*: a correct assessment without explanation will grant you 0.5 points.

Consider a sample $\boldsymbol{x} = x_1, \ldots, x_n$, which originates from an unknown distribution F. Figure 1 below consists of a normal QQ-plot for sample $\boldsymbol{x}$ and of two histograms, where *one of the two* histograms depicts sample $\boldsymbol{x}$.

a. [2 points] The QQ-plot shows that the sample originates from $\mathcal{N}(0, 1)$ (that is, from a normal distribution with mean zero and variance 1).

b. [2 points] Histogram A is the one which depicts sample $\boldsymbol{x}$.

c. [2 points] Assuming that the vertical lines in each histogram depict the mean and the median of the underlying data sets, one can deduce that the solid line corresponds to the median and that the dashed line corresponds to the mean.

d. [2 points] The Kolmogorov-Smirnov test with test statistic $D_n = \sup_{-\infty<x<\infty} |\widehat{F}_n(x) - F_0(x)|$ is used to test the (composite) null that F is a member of the location-scale family of normal distributions. Sorry for the typo ☺

e. [2 points] If the Shapiro-Wilk test applied to sample $\boldsymbol{x}$ results in a p-value of 0.98, one would *reject* the null hypothesis that the sample comes from a normal distribution (significance level $\alpha = 5\%$).

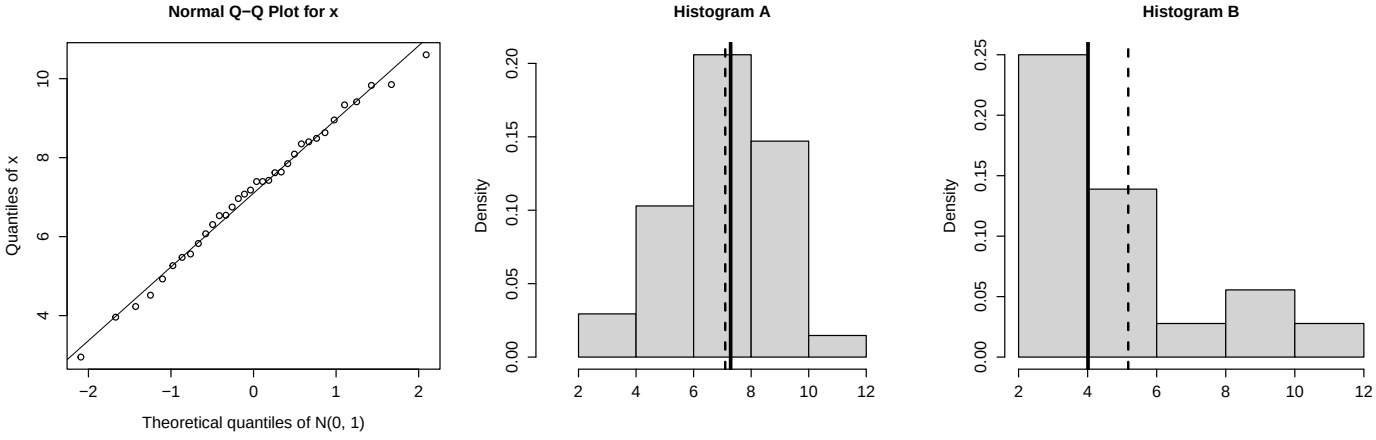

Figure 1: Normal QQ-plot for sample $\boldsymbol{x}$ (leftmost panel) and two histograms (only *one* of which depicts sample $\boldsymbol{x}$).

Sample solution for Question 1

a. **FALSE**, because the straight line in the QQ-plot is *not* the $y = x$ line.

b. **TRUE**, because the QQ-plot shows that sample $\boldsymbol{x}$ originates from a normal distribution, which is a symmetric distribution. This is consistent with Histogram A and *not* with Histogram B.

c. **TRUE**, because Histogram B depicts a right-skewed distribution, for which the mean should be greater than the median (rule of thumb). Alternatively, one can estimate the median for Histogram B to be 4 (location of solid line) as the histogram is scaled to density and the first gray rectangle has area $(4 - 2) \cdot 0.25 = 0.5$.

d. **FALSE**, the Kolmogorov-Smirnov test with the given test statistic is used to test a *simple* null hypothesis of the form $H_0 : F = F_0$ (against the alternative $H_1 : F \neq F_0$).

e. **FALSE**: as the p-value is $0.98 > 0.05 = \alpha$, where α is the significance level, one would *fail to reject* the null hypothesis.

Question 2 [8 points]

Let $x_1, \ldots, x_n$ be realizations of i.i.d. random variables $X_1, \ldots, X_n$ from a (continuous) distribution F with density f. Given a kernel K (density with expectation 0 and variance 1) and a positive bandwidth h, a *kernel density estimator* for the density f is the stochastic function $t \mapsto \widehat{f}(t)$, where

$$\widehat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{t - x_i}{h}\right).$$

a. [1.5 points] In your own words, explain the idea behind kernel density estimation.

b. [1.5 points] Explain the influence of the bandwidth on kernel density estimators. In particular, how does the smoothness of the estimator change when a small or a large bandwidth is chosen?

c. [0.5 points] Among the four kernels we have seen in the lecture (rectangular, triangular, Epanechnikov and normal), and assuming the optimal bandwidth is used, for which one is $\int K(x)^2 dx$ (and thus the MISE) minimized?

d. [1.5 points] Which choice (bandwidth or kernel) influences the quality of the estimator $\widehat{f}$ the most? Motivate your answer.

e. [3 points] Assume you want to estimate the density of a *non-negative* random variable. In the lecture, we have discussed different approaches to construct a kernel density estimator $\widehat{f}$ such that $\widehat{f}(t) = 0$ for $t < 0$. Two of these approaches were deemed reasonable; choose one of them and summarize it in your own words.

Sample solution for Question 2

a. The kernel density estimator $\widehat{f}$ allocates mass $1/n$ smoothly around each observation x_i, $i = 1, \ldots, n$, according to kernel K and with spread of mass specified by bandwidth $h > 0$.

b. For a given kernel type, the bandwidth choice strongly influences the smoothness of the resulting kernel density estimator, with a smaller bandwidth corresponding to a 'peakier' estimator and a larger one to a 'smoother' estimator.

c. Epanechnikov.

d. The choice of bandwidth has more influence than the choice of kernel on the kernel density estimator, as the bandwidth choice controls the smoothness of the resulting estimator (see b.) and can potentially lead to misleading impressions: in particular, a too small bandwidth results in an estimator with several spurious modes, while a too large bandwidth results in an estimator which excessively hides any underlying structure. Conversely, different kernel choices for a given bandwidth do not influence the quality of the estimator as radically in practice (although they determine the 'shape' of the estimator around the modes and influence its continuity and differentiability).

e. The following two choices were deemed reasonable:

   i.) *Data transformation*: for each $i = 1, \ldots, n$, set $y_i = \log(x_i)$. Find the kernel density estimator $\widehat{f}_y$ for

the transformed sample and transform back to obtain $\widehat{f}_x(t) = \frac{1}{t}\widehat{f}_y(\log(t))$;

ii.) *Symmetrization*: derive the kernel density estimator $\widehat{f}_S$ based on the 'symmetrized' sample $x_1, -x_1, \ldots, x_n, -x_n$. Then use as an estimator for f the function

$$\widehat{f}_x(t) = \begin{cases} 2\,\widehat{f}_S(t) & \text{if } t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Question 3 [9 points]

Let $x_1, \ldots, x_n$ be realizations of independent random variables $X_1, X_2, \ldots, X_n$ ($n \geq 2$) following a lognormal distribution with unknown parameters $\mu \in \mathbb{R}$ and $\sigma > 0$; that is, for $i = 1, \ldots, n$, one has $\log(X_i) \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. For lognormal distributions, $\exp(\mu)$ can be considered as a scale parameter. We are interested in an estimator of $\exp(\mu)$ and our choice is the statistic

$$T_n(X_1, \ldots, X_n) = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log(X_i)\right).$$

a. [3 points] We wish to use a *parametric* bootstrap to determine the standard deviation of the statistic T_n. Describe the steps of a reasonable parametric bootstrap scheme for estimating the standard deviation of T_n. In particular, what is necessary in order to perform the sampling step of this parametric bootstrap?

b. [3 points] In the lectures, we have seen another bootstrap method to accomplish such task. *i)* How is this other bootstrap method called? *ii)* When using this other bootstrap method, how are bootstrap samples generated? *iii)* Finally, in the specific situation considered here, would you prefer this other bootstrap method or the parametric bootstrap method of a.? Motivate your answer.

c. [3 points] Explain the two types of errors that are usually involved when employing a bootstrap procedure and comment briefly on their different nature. Which of the two can be made arbitrarily small and how?

Sample solution of Question 3

a.
- Using $x_1, \ldots, x_n$, obtain the necessary estimates $\hat{\mu}$ and $\hat{\sigma}$ for the unknown parameters of the lognormal distribution.
- For each $i = 1, \ldots, B$, where B large, generate a bootstrap sample $X^*_{1,i}, \ldots, X^*_{n,i} \overset{i.i.d.}{\sim} \text{Lognormal}\,(\hat{\mu}, \hat{\sigma})$ and compute $T^*_{n,i} = T_n\left(X^*_{1,i}, \ldots, X^*_{n,i}\right) = \exp\left(\frac{1}{n}\sum_{j=1}^{n} \log(X^*_{j,i})\right)$.
- As an estimator of the standard deviation of the statistic T_n, we use the sample standard deviation of bootstrap values; i.e. we calculate $\text{sd}\left(T^*_{n,1}, \ldots, T^*_{n,B}\right)$.

b. *i)* Empirical bootstrap.

ii) By resampling with replacement from realizations.

iii) As in this case we have additional information about the parametric family of which the underlying distribution is a member, the parametric bootstrap making use of this information should (at least marginally) outperform the empirical bootstrap.

c. The first type of error is due to the estimation of P via $\tilde{P}_n$ (and of Q_P via $Q_{\tilde{P}_n}$) and it is – in some sense – unavoidable. The second type of error is due to the estimation of $Q_{\tilde{P}_n}$ via the empirical distribution of $T^*_{n,1}, \ldots, T^*_{n,B}$. The latter can be made arbitrarily small by increasing the number B of bootstrap samples.