

Foundations of Statistics

Homework 7

Exercise 1 (Order statistic, part I). Let X_1, X_2, \dots, X_n be i.i.d. real-valued random variables with CDF $x \mapsto F(x) \in [0, 1]$. Let us consider their maximum and minimum:

$$Y := \max\{X_1, \dots, X_n\}, \quad Z := \min\{X_1, \dots, X_n\}.$$

The random variables Y and Z are called *largest order statistic* and *smallest order statistic*, respectively.

- a Prove that the distribution function of Y is given by

$$F_Y(y) = F(y)^n \quad \forall y \in \mathbb{R}.$$

If, in particular, F has density function f , find the density function f_Y of the random variable Y .

- b Prove that the distribution function of Z is given by

$$F_Z(z) = 1 - [1 - F(z)]^n \quad \forall z \in \mathbb{R}.$$

If, in particular, F has density function f , find the density function f_Z of the random variable Z .

- c Find the joint CDF of the random vector $\mathbf{U} := (Z, Y)^\top$.
- d If, in particular, F has density function f , find the joint density function of \mathbf{U} . Are Z and Y independent?

Exercise 2 (Order statistic, part I, examples).

- a Let $U, V \sim \text{Unif}(0, 1)$ be independent. Based on the previous exercise, find density function of $\max\{U, V\}$ and $\min\{U, V\}$. Compare your result with a simulation in R. Generate random samples from the uniform distribution, and for each pair, record both the maximum and minimum values. Finally, plot the histogram of these values.

- b Let $U, V \sim \text{Unif}(0, 1)$ be independent and $p \in (0, 1)$ be a constant. In HW4, Exercise 1, we studied indicator random variables and showed that $\mathbb{I}_{\{U \leq p\}} \sim \text{Ber}(p)$. Now, use order statistic to find the distribution of the random variables $\mathbb{I}_{\{U \leq p\}} \mathbb{I}_{\{V \leq p\}}$ and $\mathbb{I}_{\{U \leq p\}} + \mathbb{I}_{\{V \leq p\}}$.
- c Let $X_1, \dots, X_n \sim \text{Unif}(a, b)$ be i.i.d random variables. Find CDF F_Y and F_Z of the largest and smallest order statistic Y and Z , respectively. Do random variables Y and Z converge in distribution as $n \rightarrow \infty$?
- d Let $X_1, \dots, X_n \sim \exp(\lambda)$ be i.i.d random variables. Find CDF F_Y and F_Z of the largest and smallest order statistic Y and Z , respectively. Do random variables Y and Z converge in distribution as $n \rightarrow \infty$?

Exercise 3 (Sample skewness and sample kurtosis).

- a Show with Chebyshev's inequality that for any random variable not more than about 11% of the data can be more than three standard deviations away from the mean.
- b Show that for a $N(\mu, \sigma^2)$ -distributed random variable the proportion calculated in a) is now 0.3%.
- c For a sample x_1, \dots, x_n the z -score is defined by

$$z_i := \frac{1}{\tilde{s}}(x_i - \bar{x}), \quad i = 1, \dots, n.$$

Here \bar{x} and \tilde{s} are the sample mean and standard deviation (with denominator n not $n - 1$), respectively. Explain what $z_i = 3$ means.

- d Install the package `UsingR` with the commands `install.packages("UsingR")` and `require("UsingR")`. The dataset `exec.pay` contains direct compensation data for 199 United States CEOs. Compare the mean, median and quantiles by using the function `summary(exec.pay)`. Draw the boxplot and determine the outliers.
- e Calculate with R the z -score of the data to find out what proportion of the data are more than 3 standard deviations from the mean. Compare your result with the results in a) and b).
- f The *sample skewness* is defined by

$$\sqrt{n} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{[\sum_{i=1}^n (x_i - \bar{x})^2]^{3/2}}.$$

Show that this is equal to

$$\frac{1}{n} \sum_{i=1}^n z_i^3.$$

- g Calculate the sample skewness of the `exec.pay` dataset.
- h The *sample kurtosis* is the measure of the tails in a data set. Long tails will lead to larger values, while “normal” data will have kurtosis close to 0. It is defined by the formula

$$n \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} - 3.$$

Show that this is equal to

$$\frac{1}{n} \sum_{i=1}^n z_i^4 - 3.$$

Guess why we are taking out number 3 here.

- i Calculate the kurtosis of the `exec.pay` dataset.

Exercise 4. Suppose we have a computer program consisting of $n = 100$ pages of code. Let X_i be the number of errors on the i^{th} page of code. Suppose that the X_i ’s are Poisson with mean 1 and that they are independent. Let $Y = \sum_{i=1}^n X_i$ be the total number of errors. Use the Central Limit Theorem to approximate $\mathbb{P}(Y \leq 90)$. Check your answer with the exact value of $\mathbb{P}(Y \leq 90)$. (*Hint:* recall HW5, Exercise 4(c)).

Exercise 5. An accountant wants to simplify his bookkeeping by rounding amounts to the nearest integer, for example, rounding €99.53 and €100.46 both to €100. What is the cumulative effect of this if there are, say, 100 amounts? To study this we model the rounding errors by $n = 100$ independent $U(-0.5, 0.5)$ random variables X_1, \dots, X_{100} .

- a Compute the expectation and the variance of each X_i .
- b Use Chebyshev’s inequality to compute an upper bound for the probability $\mathbb{P}(|X_1 + X_2 + \dots + X_{100}| > 10)$ that the cumulative rounding error $X_1 + X_2 + \dots + X_{100}$ exceeds €10.
- c What can you say about the mean of the absolute error $\frac{1}{n} \sum_{i=1}^n |X_i|$, applying the Law of Large Numbers?