Universität Bielefeld
Tanja Pasurek
Ehsan Abedi

WiSe 2024/25

Foundations of Statistics

Homework 6

Exercise 1 (A universal random number generator, cf. Ch. 1.8).

Let $X : \Omega \to \mathbb{R}$ be a continuous random variable. Let F_X denote its CDF, which is a continuous function. Suppose that F_X is strictly increasing from 0 to 1 over some interval $\mathcal{I} \subseteq \mathbb{R}$. In this case, F_X has an inverse function $F_X^{-1} : [0,1] \to \mathcal{I}$.

a Define $Y := F_X(X)$, i.e., you *plug a continuous random variable into its own CDF*. Show that $Y \sim \mathrm{Unif}(0,1)$. This is called the **probability integral transform**.

b Let now $U \sim \mathrm{Unif}(0,1)$ and define $Z := F_X^{-1}(U)$, i.e., you *plug a uniform random variable into an inverse CDF*. Show that Z and X have the same distribution, i.e., $F_Z = F_X$.

▶ **Conclusion:** *Any continuous real-valued random variable can be transformed into a uniform random variable and back by using its CDF.*

c Write an `R` code to simulate continuous random variables from the density

$$f(x) = \frac{2}{(x+1)^3}, \quad x > 0.$$

Make a histogram of $n = 10^5$ simulated values and superimpose the density function to check the work.

Hint: The distribution is heavy-tailed, so in order to make a nice histogram, plot only the values less than 10 (which is about 99% of the values).

Exercise 2 (Order statistic, part I). Let $X_1, X_2, ..., X_n$ be i.i.d. real-valued random variables with CDF $x \mapsto F(x) \in [0, 1]$. Let us consider their maxium and minimum:

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

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? Comment on your observation.

Exercise 3 (Order statistic, part I, examples).

a Let $U, V \sim \mathrm{Unif}(0, 1)$ 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 $U, V \sim \mathrm{Unif}(0, 1)$ be independent and $p \in (0, 1)$ be a constant. In HW3, Exercise 4, we studied indicator random variables and showed that $\mathbb{I}_{\{U \leq p\}} \sim \mathrm{Ber}(p)$. Now, use order statistic to find the distribution of the random variables $\mathbb{I}_{\{U \leq p\}}\mathbb{I}_{\{V \leq p\}}$ and $\mathbb{I}_{\{U > p\}}\mathbb{I}_{\{V > p\}}$.

c Let $X_1, \cdots X_n \sim \mathrm{Unif}(a, b)$ 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 $n \to \infty$?

d Let $X_1, \cdots X_n \sim \exp(\lambda)$ 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 $n \to \infty$?

Exercise 4 (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 $x_1, ..., x_n$ the z-score is defined by

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

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 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, $n = 100$ amounts? To study this we model the rounding errors by 100 independent Unif$(-0.5, 0.5)$ random variables $X_1, ..., X_{100}$.

a Compute the expectation and the variance of each X_i..

b Use Chebyshev's inequality to estimate the probability that the cumulative rounding error exceeds €10:

$$p := \mathbb{P}\left(|X_1 + X_2 + ... + X_{100}| > 10\right) \leq 1/12.$$

c A manager wants to know what happens to the mean absolute error $\frac{1}{n}\sum_{i=1}^{n} |X_i|$ as n becomes large. What can you say about this, applying the Law of Large Numbers?

d In order to know the exact value of p in (b) one has to determine the distribution of the sum $X_1 + X_2 + ... + X_{100}$. This is difficult, but the Central Limit Theorem is a handy tool to get an approximation of p. Check that $\mathbb{P}\left(|X_1 + X_2 + ... + X_{100}| > 10\right) \approx 0.0006$.