# Probability and Statistics
# Topic 8. Descriptive analysis of data

- Population and sample
- Sample mean, median and mode
- Data span, sample variance, sample standard deviation (standard error)
- Percentile (quantile), quartile
- Empirical cumulative distribution function (`ecdf`)
- Distribution of the sample mean and sample variance

## Seminar 8

1. Complete the tasks from the file `Seminar_8.Rmd`.

   **Now let us recall the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT) and see how they work in practice.**

2. (i) Consider the normal distribution $\mathcal{N}(\mu_1, \sigma_1^2)$, where $\mu_1$ and $\sigma_1$ are taken from **Task 1**. Generate $N = 1000$ samples of size $n$, and for each sample compute the sample mean. For $n = 10,\ 100,\ 1000,\ 10000$, plot and overlay the histograms of these sample means. What conclusions can you draw from the results?

   (ii) For each $n = 10,\ 100,\ 1000,\ 10000$, compute the theoretical probability

   $$p_n = \mathsf{P}(|\overline{\mathbf{X}} - \mu_1| > 0.03),$$

   where $\overline{\mathbf{X}}$ is the sample mean of i.i.d. random variables $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$. Among the $N = 1000$ generated samples, count the relative frequency

   $$f_n = \frac{\#\{|\overline{\mathbf{x}} - \mu_1| > 0.03\}}{N}.$$

   Compare $p_n$ and $f_n$. Are the results as expected? What do they demonstrate about the LLN?

3. (i) Generate $N = 1000$ samples of size $n = 4$ from the exponential distribution $\mathcal{E}(1)$. Compute their sample means and plot the corresponding histogram. Overlay the plot with the probability density function of the normal distribution $\mathcal{N}(1, 1/n)$.
   Hint: In the histogram, plot probabilities (not counts).

   (ii) Standardize the sample means $\overline{\mathbf{x}}$ as

   $$\mathbf{z}_n = \sqrt{n}\,\frac{\overline{\mathbf{x}} - \mu}{\sigma}$$

   (as suggested by the CLT), and plot the empirical c.d.f. $\widehat{F}_N(t)$ of $\mathbf{z}_n$. Overlay it with the c.d.f. $\Phi(t)$ of the standard normal distribution, and compute

   $$d = \max_t |\widehat{F}_N(t) - \Phi(t)|.$$

   Compare $d$ with the upper bound predicted by the Berry–Esseen theorem.

   (iii) Repeat parts (i) and (ii) for $n = 100$. Compare the results and formulate your conclusions.