Statistical Inference 1 Winter Term 2024/2025

Lecture: Prof. Dr. Thomas Augustin, Prof. Dr. Christian Heumann
Exercises: Sergio Buttazzo, Jan Anders
Tutorial, Corrections: Sophie Hopp, Michael Speckbacher **Suggested Solution**

Exercise Sheet 1

Exercise 1.1 - Random Variables

Random variables are a key concept in probability theory. Let us explore some of the basics.

(a) Give the definition of a probability space as well as the definition of a random variable.

A probability space is a triplet (Ω, F, P). Here Ω denotes a sample space, F a sigma algebra (collection of measurable sets) and P a probability function.
Given a probability space, a random variable X is a measurable function from Ω on some space of outcomes, e.g. the $\mathbb{R}^n$.

(b) Random variables are usually identified with certain functions: Give the definition of a (cumulative) distribution function and a probability (density) function for a univariate real valued random variable. How are they connected to each other? Distinguish the discrete and continuous case.

The (cumulative) distribution function of a random variable X is defined as

$$F(x) = P(X \leq x), \text{ with } x \in \mathbb{R}.$$

The probability function in the discrete case is defined as $f(x) = P(x)$. In the continuous case, it is instead defined as a density $f(x)$, which satisfies the equation $P(X \in B) = \int_B f(x)dx$ for all measurable sets B. In the discrete case, one gets the distribution function from the probability function via summation, in the continuous case via integration.

(c) Often random variables are defined as functions of other random variables. This is sometimes called variable change. The following theorem provides a relationship between densities:

Let g be an invertible and differentiable function that transforms the random variable X to the random variable $Y = g(X)$, then:

$$f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X(g^{-1}(y))$$

Given $X \sim N(\mu, \sigma^2)$ compute the density of $Y = e^x$, commonly referred to as log-normally distributed.

The transformation function $g(x) = e^x$ is differentiable and invertible. Moreover, the inverse is (remembering that $y > 0$) $(g^{-1}(y))) = ln(y)$, with the derivative $\frac{d}{dy} g^{-1}(y) = \frac{1}{y}$. Using the given formula, this results in:

$$f_Y(y) = \frac{1}{y} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(ln(y)-\mu)^2}{2\sigma^2}}$$

Exercise 1.2 - Central Limit Theorem

An employee stays at the office a little bit longer than his contract requires for all 225 working days of the year. His additional working time on each day, Y, is described with an exponentially distributed random variable with an expected value of 5 minutes. This means its density is $f_Y(y) = \lambda e^{-\lambda y}$ with some λ (*hint: remember the properties of the exponential distribution*). Moreover, we assume that different days are independent.

(a) Use the central limit theorem to compute the approximate distribution of the employee's extra working time for a whole year.

Central Limit Theorem: $X_1, \ldots, X_n$ are i.i.d. random variables with

$$E(X_i) = \mu \quad \text{and} \quad Var(X_i) = \sigma^2 > 0,$$

then the distribution function $F_n(z) = P(Z_n \leq z)$ of the standardized sum,

$$Z_n = \frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma},$$

converges for $n \to \infty$, for every point $z \in \mathbb{R}$ to $\Phi(z)$, i.e. the standard normal distribution function. This is called convergence in density:

$$Z_n \xrightarrow{d} N(0,1)$$

One can also write Z_n has an asymptotic normal distribution:

$$Z_n \overset{a}{\sim} N(0,1)$$

Let X_i denote the additional working time on day i with $i = 1, \ldots, 225$. Then the $X_i \sim Exp(\lambda)$ are identically and independently distributed. The parameter λ can be computed with the help of the given expectation:

$$\begin{aligned} E(X_i) = 5[\text{min}] &= \frac{1}{\lambda} \Rightarrow \lambda = 0.2 \\ Var(X_i) = \frac{1}{\lambda^2} &= \frac{1}{0.04} = 25 \end{aligned}$$

Let now $X := \sum_{i=1}^{225} X_i$ be the annual additional working time.

Using the central limit theorem it holds:

$$Z_n := \frac{\sum_{i=1}^{225} X_i - 225 \cdot 5}{\sqrt{225}\sqrt{25}} \overset{a}{\sim} N(0,1)$$

And as a result:

$$\sum_{i=1}^{225} X_i \overset{a}{\sim} N(225 \cdot 5, 225 \cdot 25)$$

(b) Compute the probability that the employee works more than 16 additional hours.

It holds $16[\text{h}] = 16 \cdot 60[\text{min}] = 960[\text{min}]$. As a result we are interested in $P(X > 960)$.

$$\begin{aligned} P(X > 960) &= 1 - P(X \leq 960) \\ &= 1 - P\left(\frac{X - 225 \cdot 5}{15 \cdot 5} \leq \frac{960 - 225 \cdot 5}{15 \cdot 5}\right) \\ &= 1 - \Phi(-2.2) \\ &= \Phi(2.2) \\ &\approx 0.9861 \end{aligned}$$

The probability of the employee working more than 16 additional hours over one year is approximately 0.9861.

Exercise 1.3 - Speed of convergence in the Central Limit Theorem (R-exercise)

Let's now explore the central limit theorem using R. The goal of this exercise is to perform simulations and plot the results to gain a graphical intuition of the theorem.

(a) Simulate a $Bin(100, 0.75)$ variable 1000 times and plot a simple histogram of the resulting vector. Overlay the corresponding expected density function given by the Central Limit Theorem. Why does this work, even though we made a histogram of the raw simulated values of a binomial variable instead of a statistic derived from multiple samples?
Hint : Have a look at the R-commands `rbinom, curve, dnorm`.

See `Ex_1_3.R`.

(b) Write a function with parameters n, p, and r that plots a histogram of r standardised sample means Z_n (i.e. mean of 0 and variance of 1) of n draws from a Bernoulli distribution with parameter p. Again, overlay the expected distribution according to the central limit theorem (the standard normal density). You should fix the domain of the histogram (e.g. to $[-5, 5]$) so that comparisons are easier.

Play around with the parameters n, p and r. What can you say about the speed and rate of convergence for different values of p?

Tip : Have a look at the function `manipulate` from the manipulate package for interactively changing the parameters you want to adjust in the plot, for example:
`manipulate(your_function_name(n = n, p, r), n = slider(1, 100, step = 1))`

See `Ex_1_3.R`.

(c) Adjust your function from (b), so instead of simulating a binomial variable, it simulates r samples of size n from $Exp(\lambda)$. The parameters n and r remain as function parameters, p is replaced with *lambda*. The function is supposed to compute the mean of each sample, standardise them and again plot the histogram against the standard normal. What can you say about the speed of convergence now?

See `Ex_1_3.R`.