Repetition and important formulas from modul 5

1 Important terms

Independent (oberoende) A collection of random variables $X_1, \ldots, X_n$ is said to be *independent (oberoende)* if their joint distribution is the product of the single distributions:

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n).$$

(The rhs is the probability that $X_1 \leq x_1$ and $X_2 \leq x_2$ and ... and $X_n \leq x_n$.)

Correlation (korrelation) Two variables X and Y have correlation coefficient

$$\rho(X, Y) = \frac{\mathrm{C}(X, Y)}{\sqrt{\mathrm{V}(X)}\sqrt{\mathrm{V}(Y)}}.$$

Recall that $\mathrm{C}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$. As in the case of observations (R_{xy}) it holds $-1 \leq \rho(X, Y) \leq 1$ and it gives information about positive or negative dependence.

2 Sums of independent and identically distributed (iid) random variables

In this module we were mainly looking at the sum

$$S_n = X_1 + \cdots + X_n \tag{1}$$

of identically distributed random variables $X_1, \ldots, X_n$ with a common expected value $\mu = \mathrm{E}(X_i)$ and variance $\mathrm{V}(X_i) = \sigma^2 > 0$ for $i = 1, \ldots, n$.
It holds $\mathrm{E}(S_n) = n\mu$ and if the random variables X_i are also independent $\mathrm{V}(S_n) = n\sigma^2$.

We looked at the following distributions:

Normal distribution Let $X_1, \dots, X_n$ be independent and normally distributed, $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), i = 1, \dots, n$ (note that the parameters μ_i and σ_i may differ for the random variables, they only have the normal distribution in common). Let $a_1, \dots, a_n$ and b be constants and define

$$Y = \sum_{i=1}^{n} a_i X_i + b.$$

It holds

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2), \quad \text{where} \quad \mu_Y = \mathrm{E}(Y) = \sum_{i=1}^{n} a_i \mu_i + b \quad \text{and} \quad \sigma_Y^2 = \mathrm{V}(Y) = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

I the special case of $\mu_1 = \dots = \mu_n$ and $\sigma_1 = \dots = \sigma_n$:

$$S_n \sim \mathcal{N}(n\mu, n\sigma^2)$$

(since in this case S_n is a linear combination of the X_i's with $\mathrm{E}(X_i) = \mu$, $\mathrm{V}(X_i) = \sigma^2, a_i = 1$ for all $i = 1, \dots, n$ and $b = 0$).

Binomial distribution Consider $X \sim \mathrm{Bin}(1, p)$ (side remark: this special case of the binomial distribution is also called Bernoulli distribution, where we only have one try, $n = 1$). Then

$$S_n = X_1 + \dots + X_n \sim \mathrm{Bin}(n, p);$$

and thus if $X_1 \sim \mathrm{Bin}(n_1, p)$ and $X_2 \sim \mathrm{Bin}(n_2, p)$ are independent:

$$X_1 + X_2 \sim \mathrm{Bin}(n_1 + n_2, p).$$

For S_n it follows

$$\mathrm{E}(S_n) = n\mathrm{E}(X) = np \quad \text{and} \quad \mathrm{V}(X_n) = n\mathrm{V}(X) = np(1-p).$$

Poisson distribution For two independent random variables $X_1 \sim \mathrm{Po}(m_1)$ and $X_2 \sim \mathrm{Po}(m_2)$ it holds

$$X_1 + X_2 \sim \mathrm{Po}(m_1 + m_2),$$

and especially

$$S_n = X_1 + \dots + X_n \sim \mathrm{Po}(n \cdot m)$$

if $\mathrm{E}(X_i) = m$ for all $i = 1, \dots, n$.

3 Two important theorems: LLN and CLT

Law of Large Numbers (LLN) (Stora talens lag) Let $X_1, \dots, X_n$ be independent with $\mathrm{E}(X_i) = \mu$ and $\mathrm{V}(X_i) = \sigma^2$ for all $i = 1, \dots, n$. The law of large numbers gives us (almost sure) convergence of

$$\frac{S_n}{n} = \frac{X_1 + \dots + X_n}{n} \to \mu$$

as $n \to \infty$.

In words: If we add more and more random variables with the same expected value and variance, their average goes to the expected value μ, i.e. if we repeat the same experiment over and over again then the average of the results goes to μ and the more experiments we perform the closer we get to it.

Central Limit Theorem (CLT) (Centrala gränsvärdessatsen) Define

$$Z_n = \frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{V}(S_n)}} = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

(i.e. "standardize" the random variable S_n from eq. (1)), then for $n \to \infty$, Z_n converges in distribution to a standard normally distributed random variable Z,

$$Z_n \Rightarrow Z \sim \mathcal{N}(0, 1).$$

Convergence in distribution means that the distribution function of Z_n gets closer and closer to the distribution function of the standard normal distribution, i.e.

$$P(Z_n \leq x) \to \Phi(x) = P(Z \leq x)$$

for all x.

Alternative formulations of the CLT are

- For $n \to \infty$ it holds approximately

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \sim \mathcal{N}(0, 1).$$

- For $n \to \infty$ and a random variable $Z \sim \mathcal{N}(0, 1)$ it holds

$$S_n \approx n\mu + \sqrt{n}\sigma Z,$$

  i.e. approximately $S_n \sim \mathcal{N}(n\mu, n\sigma^2)$.

For examples see F7 or JR chapter 5.