Introduction to Probability and Statistics

Session 1 Exercises

Israel Leyva-Mayorga

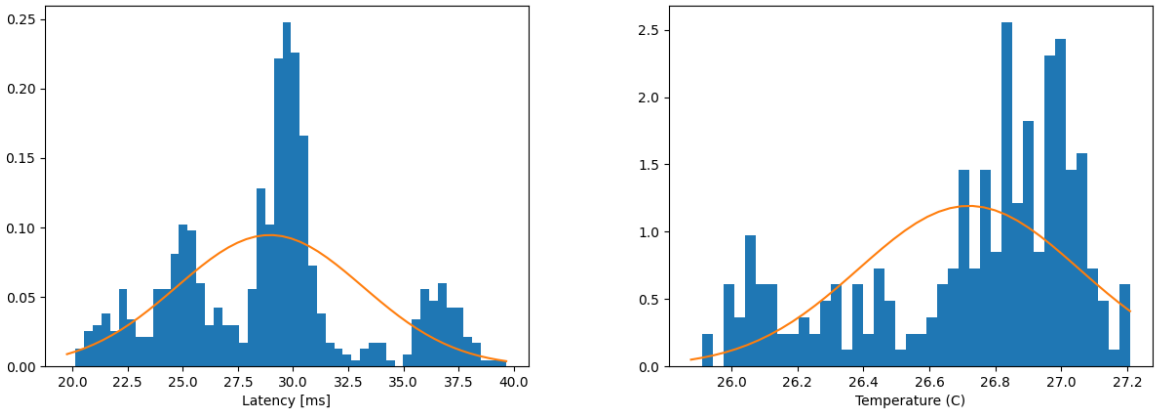
Exercise 1: Use the data in the moodle materials folder to plot the absolute, relative, and normalized histograms of

a) The latency for communication in our Starlink setup (file: Starlink_latency.csv)

b) The temperature measured by a sensor during a PhD course at Oulu University (file: temperature.csv)

Are these datasets approximately normal?

Solution: The normalized histograms are shown in Fig. 2, along with with the probability density function (PDF) of the Gaussian distributions with appropriate means and variances calculated from the sample (i.e., sample mean and variance). These plots show that neither the latency nor temperature measurements follow a Gaussian distribution.



(a) (b)

Fig. 1. Normalized histograms for (a) the latency with Starlink and (b) the temperature at the PhD course, along with the PDF of the Gaussian distributions with appropriate means and variances.

Exercise 2: Simulate rolling a fair dice n times.

a) Calculate the sample mean $\overline{X}_i$ and variance S_i^2 for all $i = 1, 2, \ldots, n$.

b) Plot the sample mean $\overline{X}_i$ and $\overline{X}_i \pm S_i$.

c) Plot a normalized histogram with the n outcomes. How many times do you need to roll the dice so the histogram resembles the pmf of a uniform RV?

Tip: You can use the code dice_loln.py in the moodle page as a base

Solution: The sample mean $\overline{X}_i$ and variance S_i^2 for $i = \{1, 2, \ldots, 1000$ rolls of a 6-sided dice ($a = 1$ and $b = 6$) and the normalized histogram of the outcomes. As it can be seen, the sample mean $\overline{X}_i$ tends to the theoretical mean of the distribution $\mu = (a + b)/2 = 3.5$ as i increases and so does S_i to $((b - a + 1)^2 - 1)/12 = 35/12$. Consequently, $\overline{X}_i \pm S_i^2$ tends to 3.5 ± 1.707. Nevertheless, even with 1000 experiments, the normalized histogram is not flat.

Exercise 3: Consider the case of flipping a fair coin n times and let H_n be the random variable (RV) of the number of *heads*. Also recall that the sum of n Bernoulli RVs is a Binomial random variable with mean np and variance $np(2-p)$. Calculate and compare the probability of observing $H_{10} \geq 7$ and $H_{100} \geq 70$ using each of the following methods.

a) Summing over the formula for the probability mass function (pmf) of Binomial RVs

$$P(H_n \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tag{1}$$

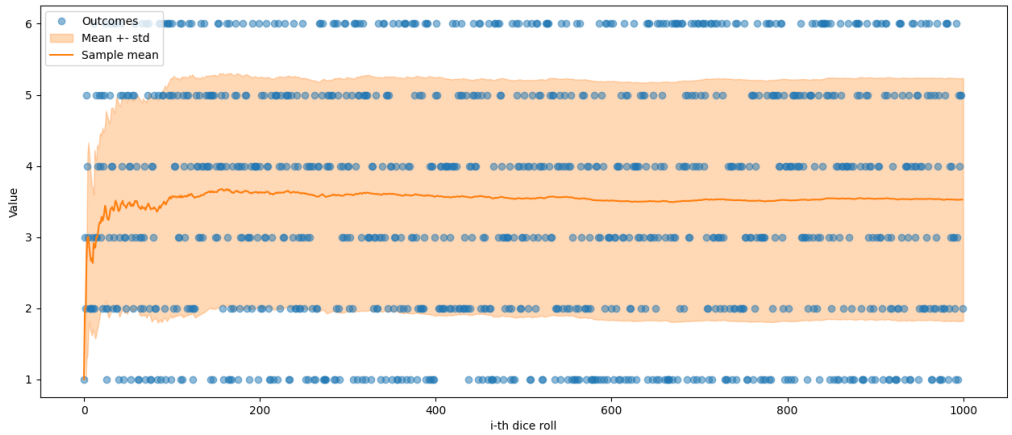
b) Using the central limit theorem

Solution:

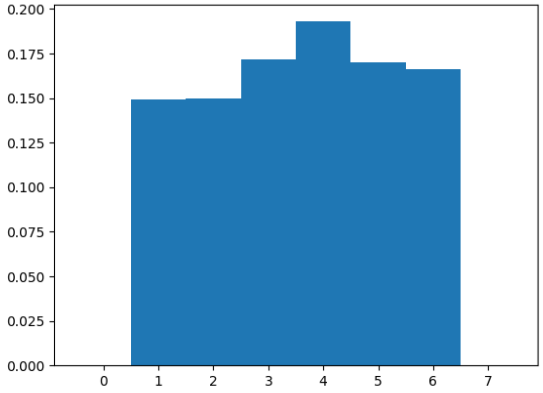
a) By using the formula for the Binomial distribution is $P(H_{10} \geq 7) = 0.1718$ and $P(H_{100} \geq 70) = 0.00003925$. This shows that having 70% of the tosses being heads is likely with 10 coin tosses but is extremely rare with 100 coin tosses.

b) By using the normal approximation, we get $P(H_{10} \geq 7) = 0.1029$ and $P(H_{100} \geq 70) = 0.00003167$.

By comparing the results with these two methods, the absolute error is greater for H_{10} than for H_{100}. This suggests that the normal approximation becomes more accurate as n increases.



(a)



(b)

Fig. 2. (a) Sample mean and variances for $i = \{1, 2, \ldots, 1000\}$ and (b) normalized histogram for $n = 1000$.

Exercise 4: A football team will play 60 games this year. Thirty-two of these games are against teams playing the champions league, denoted as class A teams, and 28 are against other teams, denoted as class B teams. The outcomes of the games are independent. The team will win each game against a class A team with probability 0.5, and it will win each game against a class B team with probability 0.7. Let X denote its total number of victories in the year.

a) Is X a Binomial RV?

b) Let X_A and X_B denote, respectively, the number of victories against class A and class B teams. What are their distributions?

c) What is the relationship between X, X_A and X_B?

d) Approximate the probability that the team wins 40 or more games this year

Solution:

a) No, it is not, since we have two success probabilities $p_A = 0.5$ and $p_B = 0.7$.

b) The distributions are $X_A \sim \text{Binom}(n_A, p_A)$ and $X_B \sim \text{Binom}(n_B, p_B)$.

c) The relationship is $X = X_A + X_B$

d) The variance of a binomial distributions with parameters n and p is $np(1-p)$. Then, we assume that the distribution of X_A and X_B can be approximated by normal distributions $N_A \sim N(n_A p_A, n_A p_A(1-p_A))$ and $N_B \sim N(n_B p_B, n_B p_B(1-p_B))$. Therefore, the sum of N_A and N_B can be approximated $X \sim N\left(\mu_X, \sigma_X^2\right)$, where

$$\mu_X = n_A p_A + n_B p_B$$

$$\sigma_X^2 = n_A p_A(1-p_A) + n_B p_B(1-p_B).$$

With these, we approximate

$$P(X \geq 40) = 1 - \frac{1}{2}\left[1 + \text{erf}\left(\frac{40 - \mu_X}{\sigma_X\sqrt{2}}\right)\right] = 0.1187$$

Therefore, the probability that this team wins 40 or more gains is quite low.

Exercise 5: The room temperature during a PhD course was recorded using a Raspberry Pi and a temperature sensor. The collected values are in the file temperature.csv (the temperature is the second column).

a) Calculate the sample mean $\overline{X}_n$ and variance S_n^2 of the temperature with all the $n = 253$ measurements

b) Assume that $\overline{X}_n$ is the true temperature μ and that S_n^2 is the real variance of the Gaussian noise that affects each measurement of the sensor, denoted by σ^2. Consequently, assume that the temperature measurements have a distribution $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Plot the likelihood or log-likelihood function for the first 10 and with the first 100 measurements

c) Calculate the Maximum Likelihood Estimator (MLE) for μ with the first 10 and with the first 100 measurements under the previous assumption.

Solution:

a) Using all the $n = 253$ samples, these give $\overline{X}_n = 26.720$ and variance $S_n^2 = 0.112$

b) Given that the PDF of the Gaussian RV is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}},$$

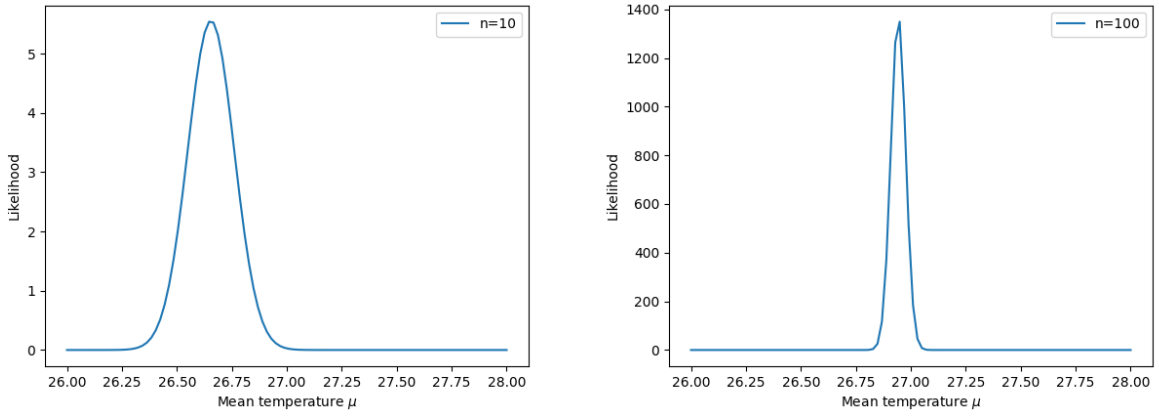
we calculate the likelihood function as

$$\mathcal{L}_n(\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} = \frac{e^{\left(\sum_{i=1}^{n} \frac{-(X_i-\mu)^2}{2\sigma^2}\right)}}{\left(\sigma\sqrt{2\pi}\right)^n}$$

and the log-likelihood as

$$\ell_n(\mu) = \log\left(\mathcal{L}_n(\mu)\right) = -n(\sigma + \sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2\sigma^2}.$$

By plotting the likelihood and log-likelihood functions with the first $n = \{10, 100\}$ measurements, we get the plots shown in Fig. 3 and Fig. 4.



(a) (b)

Fig. 3. Likelihood for the temperature with (a) $n = 10$ and (b) $n = 100$, given that it is a Gaussian RV.

c) The MLE is obtained by setting $X = \sum_{i=1}^{n} X_i$ and finding the arg max of

$$\frac{\partial \ell_n(\mu)}{\partial \mu} = \frac{2X - 2n\mu}{2\sigma^2} = 0, \longrightarrow \hat{\mu}_n = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n$$

Therefore, the MLE for $n = 10$ gives $\hat{\mu}_{10} = 26.655$ and for $n = 100$ gives $\hat{\mu}_{100} = 26.943$.

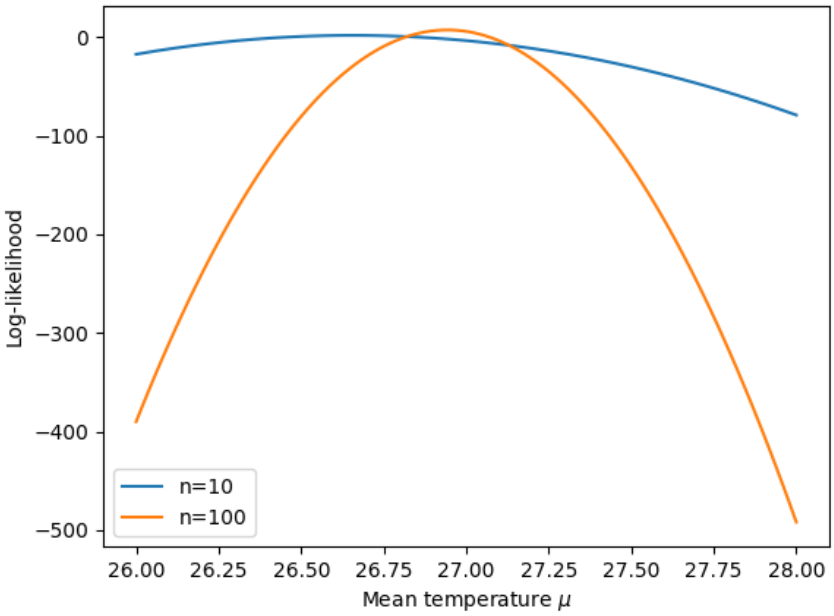


Fig. 4. Log-likelihood for the temperature, given that it is a Gaussian RV.