# Workshop exercises for statistics

Israel Leyva-Mayorga

## Session 1: Introduction to statistics and parameter estimation

1) Generate $n = 30$ data distributed as a Gaussian random variable with mean $\mu = 2$ and variance $\sigma^2 = 16$ using the software you prefer.
    a) Plot the absolute, relative, and normalized (i.e., empirical pmf) histograms of the data and compare these with the Gaussian probability density function (PDF). Which one of these is in the same scale as the Gaussian PDF?
    b) Compute the empirical cumulative density function (CDF) and compare it with the Gaussian CDF. Can you show in the plot what represent the 0.2-quantile of the distribution in both the empirical and analytical CDF?
    c) Re-do point a) and b) using $n = 1000$. Do you note any difference?
2) A company is testing 3 different implementations of a basic mathematical operation: vector-matrix multiplications. The results of $n = 1000$ executions of each of these implementations are given in the file called "execution_times.txt" in the folder "Material for workshop" (see moodle page for the workshop). In this file, the first column is the experiment number $i = 1, 2, \ldots, n$, the second column contains the execution times for each execution of implementation X, namely $X_i$, the third column contains the execution times for each execution of implementation Y, namely $Y_i$, and the fourth column contains the execution times for each execution of implementation Z, namely $Z_i$.
    a) Plot the sample average $\overline{X}_i$ and the sample variance $S_i^2$ for each value of $i$ up to $i = n$. Which algorithm is better, on average, after $i = 10$ executions? Which one is better after $i = n$ executions? Why are results different?
    b) Plot the normalized (i.e., empirical pmf) histograms of the execution times and compare these with the Gaussian PDF. Are the execution times approximately normal?
3) We want to estimate parameter $\lambda$ for a sample of $X_1, X_2, \ldots, X_n \sim \text{Exp}(\lambda)$ RVs.
    a) Calculate the Maximum Likelihood Estimator (MLE) for $\lambda$, denoted as $\hat{\lambda}_n$.

   b) **Difficult. Optional. Try at your own risk:** Calculate the bias, variance, and mean squared error (MSE) for the estimator $\hat{\lambda}_n$.
   c) **Difficult. Optional. Try at your own risk:** Is $\hat{\lambda}_n$ a consistent estimator for $\lambda$?
4) Let $X_1, X_2, \ldots, X_n \sim \text{Poisson}(\lambda)$ and let $\hat{\lambda}_n = \left(\sum_{i=1}^{n} X_i\right)/n$. Find the bias, standard error, and MSE of the estimator.
5) Consider a packet-based transmission system. We want to estimate the probability $p$ of correct transmission for each of the packets, observing the number of correct packet detections $k$ over the total independent packet transmissions $n$.
   a) Which is the random variable that represents the experiment?
   b) Evaluate the MLE of $p$, denoted as $\hat{p}_n$.
   c) Compute the MSE of $\hat{p}_n$. Is the estimator consistent?

## SESSION 2: CONFIDENCE INTERVALS AND REGRESSION

1) A researcher wants to estimate the average height of college students in a particular university. She randomly selects a sample of 200 students and measures their heights. The mean height of the sample is 175 centimeters with a standard deviation of 8 centimeters. Assume that the population of heights is normally distributed.
   a) Calculate a 95% confidence interval for the true mean height of the population.
   b) What is the margin of error for the confidence interval calculated in part a)?
   c) Suppose the researcher wants to increase the precision of her estimate by decreasing the margin of error. What could she do to achieve this goal?
   d) Now suppose the sample of students available is only 30 students. Would you change something in the evaluation of point a)? If yes, please recompute the 95% confidence interval and comment the new result.
2) A flight company estimates the probability that all the passenger show up at the gate for the flight as $\hat{p} = 42\%$.
   a) Assuming accurate estimation, how many measurements $n_1$ are needed to have a 95% confidence interval with error margin of 2%?
   b) Assuming accurate estimation, how many measurements $n_2$ are needed to have a 99% confidence interval with error margin of 2%?
3) Table I contains information about the weight and height of a group of individuals. You want to use linear regression to predict someone's weight based on their height.

TABLE I

WEIGHT AND HEIGHT TABLE OF EXERCISE 3) FOR SESSION 2 PT. 2.

| Weight [kg] | Height [cm] |
|---|---|
| 73.84 | 241.89 |
| 68.78 | 162.31 |
| 74.11 | 212.74 |
| 71.73 | 220.04 |
| 69.88 | 206.34 |
| 67.25 | 152.21 |
| 68.78 | 183.92 |
| 68.34 | 167.97 |
| 67.01 | 175.92 |
| 63.45 | 156.39 |

a) Plot the data using a scatter plot to see if there is a linear relationship between weight and height.

b) Calculate the Pearson correlation coefficient between weight and height to see how strong the linear relationship is. Pearson's correlation coefficient is

$$r_{x,y} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2}} \tag{1}$$

c) Use linear regression to fit a line to the data. Interpret the slope and intercept of the line in the context of the problem. (hint: plot the fitted line on the scatter to help you visualize)

d) Use the fitted line to predict the weight of someone who is 2 meters tall.

e) Calculate the mean squared error (MSE) of your predictions. This measures how well your model fits the data.

4) Use the file called with the life expectancy and life satisfaction for several countries taken from Our world in Data's website. Assume that the underlying model is a linear one and that the noise is Gaussian.

a) Make a scatter plot of the data

b) Find the estimates for linear regression $\hat{\beta}_0$ and $\hat{\beta}_1$ with the whole dataset and plot the model over the scatter plot

c) Evaluate the goodness of fit, first by calculating the standardized residuals $Z_i$ and, then, by calculating the coefficient of determination, which is given as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \epsilon_i^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}_n\right)^2} \tag{2}$$

d) Remove all the entries for countries starting with letter A from the dataset and recalculate $\hat{\beta}_0$ and $\hat{\beta}_1$. Are these different than before? What conclusions can you draw from this result?

e) After performing regression with the countries starting with A removed from the dataset, predict the life expectancy of people in Argentina.

## Session 3 pt. 1: Hypothesis testing 1

1) The following sample $X_1, X_2, \ldots, X_5$ was collected by measuring a process that is affected by noise with known $\sigma^2 = 1$.

$$4.5832 \quad 4.9437 \quad 2.8638 \quad 6.6402 \quad 3.2065$$

a) Perform the tests for the null hypotheses $\mathrm{H}_0 : \mu = \mu_0$ for each value of $\mu_0 = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ with level of significance $\alpha = 0.05$. For which values of $\mu_0$ the null hypothesis cannot be rejected?

b) You're given five more measurements that you add to the previous sample:

$$4.1582 \quad 5.5028 \quad 3.7547 \quad 3.9420 \quad 4.0909$$

After adding these measurements, repeat the tests for the null hypotheses $\mathrm{H}_0 : \mu = \mu_0$ for each value of $\mu_0 = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$. For which values of $\mu_0$ the null hypothesis cannot be rejected? Are these results different than with five measurements?

2) It is known that if a signal of value $\mu$ is sent from location A, then the received value at location B is normally distributed with mean $\mu$ and $\sigma = 2$. This means that Gaussian noise that is added to the signal is a RV with distribution $N(0, 4)$. Calculate and make a plot for the Type II error probability with different values of $n = 1, 2, \ldots, 20$, the number

of transmitted signals, given $\alpha = 0.05$, $\mathrm{H}_0 : \mu = 8$ and that the real value of the signal is $\mu = 9.5$? Recall that the probability of Type II error is

$$\begin{aligned}\beta(\mu) &= P(\text{accept } \mathrm{H}_0 \mid \mu) \\ &= P\left(\frac{|\hat{\mu}_n - \mu_0|\sqrt{n}}{\sigma} \leq Z_{\alpha/2}\right) \\ &= \Phi\left(\frac{(\mu_0 - \mu)\sqrt{n}}{\sigma} + Z_{\alpha/2}\right) - \Phi\left(\frac{(\mu_0 - \mu)\sqrt{n}}{\sigma} - Z_{\alpha/2}\right)\end{aligned}$$

## Session 3 pt. 2: Hypothesis testing 2

1) (From probabilitycourse.com) Let $X_1, X_2, X_3, X_4$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where $\mu$ and $\sigma^2$ are unknown. Suppose that we have observed the following values

$$3.58 \quad 10.03 \quad 4.77 \quad 14.66$$

For $\mu_0 = 10$, would like to decide between

$$\begin{aligned}\mathrm{H}_0 &: \mu \geq \mu_0, \\ \mathrm{H}_1 &: \mu \leq \mu_0,\end{aligned}$$

2) The file called `life_expectancy\_vs\_satisfaction\_2021.csv` that you'll find in the Material folder for the workshop contains the life expectancy, life satisfaction, and population for several countries in 2021. It was taken from Our world in Data's website[1]. Assume that the life expectancy across countries in each continent has a normal distribution with parameters $\mu_{\text{continent}}$ and $\sigma^2_{\text{continent}}$ where

$$\text{continent} \in \{\text{Africa}, \text{America}, \text{Asia}, \text{Europe}, \text{Oceania}\}$$

   a) Select 5 countries from America and 7 from Europe and estimate $\sigma^2_{\text{America}}$ and $\sigma^2_{\text{Europe}}$
   b) With the selected countries, $\mu_0 = 75$, and $\alpha = 0.05$, test the null hypothesis

$$\begin{aligned}\mathrm{H}_0 &: \mu = \mu_0, \\ \mathrm{H}_1 &: \mu \neq \mu_0,\end{aligned}$$

   first for America and then for Europe.

[1] https://ourworldindata.org/grapher/life-satisfaction-vs-life-expectancy

TABLE II

DATA FOR EXERCISE

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 2.68 | 2.86 | 1.31 | 2.45 | 2.1 | 0 2.83 | 2.54 | 2.90 | 1.87 | 1.85 |
| $y$ | 3.01 | -0.12 | 0 1.50 | 0.73 | 1.62 | 2.65 | 2.83 | 0.99 | 1.53 | 2.14 |

c) With the selected countries and $\alpha = 0.05$, test the null hypothesis

$$\mathrm{H}_0: \; \mu_{\text{America}} = \mu_{\text{Europe}}$$

3) Suppose you are working for a company that sells two different types of products, A and B. You want to determine if there is a significant difference between the average sales of these two products. You decide to collect sales data for a random sample of 10 days for each product, and you obtain the following data:

- Product A: 5, 8, 7, 6, 9, 10, 6, 7, 8, 9
- Product B: 4, 6, 5, 5, 7, 8, 4, 6, 6, 7

To test the hypothesis that there is a significant difference between the average sales of these two products, you can use a two-population t-test.

a) Define the null hypothesis of the above test. Is it a one side or two side test?

b) Using a significance level of 0.05, conduct the hypothesis test previously defined.

4) Table II represents a set of $x$ and $y$ values that might have a relationship between each other.

a) How can you make a test to check if there is no linear relation between $x$ and $y$? (Hint: write an hypothesis test where the null hypothesis should represent no linear relation)

b) Perform the test considering a level of significance $\alpha = 0.1$. Is the hypotheses accepted or rejected?