This assignment consists of 17 questions for a total of **100 points**. You have two weeks to work on the assignment in pairs. Remember to prepare your answers using the template that is supplied on Canvas.

Exercise 1 Suppose that you decide to move to the countryside to raise goats. Now all you need is some goats. You've been monitoring the goat market for the last few days and took note of the prices for which the last 30 goats that you saw being auctioned sold for:

```
301.20, 238.82, 252.79, 212.17, 325.43, 245.92, 200.08, 307.88, 193.33, 232.56,
243.39, 162.40, 226.75, 231.37, 208.21, 226.49, 297.49, 252.77, 289.41, 283.34,
265.80, 280.76, 240.61, 287.22, 216.95, 264.74, 232.78, 204.10, 227.01, 231.31
```

(Prices are in €.) You can assume throughout that this data is a random sample from some distribution. Below you can find some questions relating to this situation.

5 pts (a) Plot a histogram data. (Check the tutorial to see how to plot histograms.) Does it seem like a Normal model is appropriate for the data? Justify your answer.

8 pts (b) Explain succinctly how the CLT and LLN can be used to derive an approximate pivot for the expectation of the price of one goat.

8 pts (c) Derive, step-by-step, the expression for a two-sided, 95% confidence interval for the expectation of the price of one goat based on the pivot from (b).

5 pts (d) Write a Python function that takes the data as input and outputs (the bounds of) the confidence interval from (c).

5 pts (e) What is the confidence interval that you get for the dataset from above? Is it correct to say that the expectation of the price of one goat belongs to this interval with probability 0.95? Justify your answer.

5 pts (f) Suppose that you have 4000€ to spend on goats and the the seller is willing to sell you goats at the expected price but you don't know in advance what this price is. Based on the information that you got from the confidence interval from (d), how many goats can you confidently expect to be able to purchase? (Please keep in mind that, in polite society, you cannot expect to buy something like 3.62 goats.)

Exercise 2 Suppose that you always start your day by drinking a glass of freshly pressed orange juice and so you often buy oranges in bags of 12. It sometimes happens that these bags come with some rotten oranges in them. You wonder how often this is actually happening as maybe you would want to start buying your oranges somewhere else...

Here is some data about the number of rotten oranges in the last 10 bags that you bought:

```
0, 1, 1, 2, 1, 2, 1, 0, 2, 1
```

(You can think of these as a random sample from some distribution on $\{0, 1, \ldots, 11, 12\}$.) In this exercise, you'll build a statistical procedure to help you decide, based on this data, if you should start getting your oranges somewhere else.

3 pts (a) Suppose that you can think of the number of rotten oranges in a given bag as a random sample of size 12 from some *population of oranges* where an orange is rotten with some (unknown) probability p. If X_i represents the number of rotten oranges in the i-th bag, then what is the distribution of X_i?

3 pts (b) Suppose that you would find it acceptable to get *on average* 1 rotten orange per bag. How would you pick p_0 in a null hypothesis of the type $H_0 : p = p_0$ to reflect this?

6 pts (c) Which alternative hypothesis would you pick so that if you reject the null from (b) you can conclude that (at a certain significance level) there is, on average, more that one rotten orange per bag of 12?

6 pts (d) Suppose that you use the following rejection rule; you reject H_0 at significance level α if

$$T = \sum_{i=1}^{n} X_i > C_\alpha,$$

where C_α is some appropriately chosen critical value. ($n = 10$ here.) What is the distribution of T under the null hypothesis from (b)?

8 pts (e) Write a function in Python that takes in $p_0, \alpha \in (0, 1)$ and $n \in \mathbb{N}$ as input, and outputs the respective critical value C_α. Suppose that you take $\alpha = 0.05$; what is the respective C_α? What is the conclusion of the test?

8 pts (f) Write a function that takes in $p \in (0, 1)$ and $n \in \mathbb{N}$ as input, and outputs the respective power of the test. Plot the power of the test as a function of p for $n = 10$, and for $n = 20$. Answer the following: (i) why are both curves below 0.05 at $p = p_0$? and (ii) how do the two power curves compare to one another and why?

Hint: In Python, use from scipy.stats import binom and then binom.pmf($n = \cdots, p = \cdots, k = \cdots$) to compute binomial probabilities.

Exercise 3 Being able to predict sales figures is very important for online businesses, and the number of visitors to a website might be a good indicator for sales. It might make sense for a business to invest in advertisement to attract more visitors if this will result in extra sales but then it is important to understand the relation between number of visitors and volume of sales to decide how much to invest.

Below you can find some data about total number of visitors and respective sales figures on a set of $n = 24$ days that the business finds representative of typical days. (Prices are in thousands of €.)

```
Visitors (x) 288 351 332 268 289 319 300 298 295 287 284 297 302 294 284 299 298
350 308 284 295 307 338 311

Sales (y) 1.968 2.472 2.286 1.980 2.004 2.290 2.054 2.135 2.125 2.016 2.016 1.998
2.113 1.973 2.004 2.069 2.154 2.388 2.125 2.054 1.963 2.207 2.342 2.168
```

You can assume that these 24 (x, y) pairs form a random sample of size 24 from some (bivariate) distribution.

Below you are asked to perform a statistical analysis on this data using the simple linear regression (SLR) model to regress sales on number of visitors.

3 pts (a) Make a scatterplot of the (x, y) pairs.

6 pts (b) Write down the modelling equations that you use to regress the sales linearly on the number of visitors. What do you need to assume on the noise terms in the modelling equations in order to use the SLR model?

9 pts (c) Using the expressions that we derived in class, estimate the parameters of the model – i.e., the intercept α, the slope β, and the variance σ^2 – from the data. Make a plot of the data and the respective regression line.

6 pts (d) Compute the residuals and use them to perform the diagnostics of the fit. What do you conclude?

6 pts (e) Suppose that an advertisement company is asking you for 450€ to run a campaign that increases the number of visitors in one day by 20%. Based on your statistical analysis, do you think that this is worth it? In what sense is it indeed/not worth it?