Data Science I

Overall Points: 100 / 100

Exercise 4: Statistical Analysis & Mathematical Models

Submission Deadline: December 01 2025, 07:00 UTC
University of Oldenburg
Winter 2025/2026
Instructors: Jannik Schröder, Wolfram "Wolle" Wingerath

Submitted by: Cansu Horata, Charleen Owiti, Suzine Ngiedom

Part 1: Statistical Distributions

/ 25

1.) Assuming that the relevant distribution is normal, estimate the probability of the following events:

a) That there will be 70 or more heads in the next hundred flips of a fair coin?

Find the Mean (μ) and Standard Deviation (σ) of the Binomial Distribution:

$$\text{Mean: } \mu = n \times p = 100 \times 0.5 = 50$$

$$\text{Standard Deviation: } \sigma = \sqrt{n \times p \times (1-p)} = \sqrt{100 \times 0.5 \times 0.5} = \sqrt{25} = 5$$

Apply Continuity Correction:When approximating a binomial distribution with a

We want "70 or more heads". In the binomial case, this means 70, 71, 72, ... So, fo

$$\approx P(X_{normal} \geq 6$$

Calculate the Z-score:The Z-score tells us how many standard deviations away frc

Z>=4

Find the Probability:We want $P(Z \geq 3.9)$. Standard Normal Distribution tables

$$P(Z < 3.9) \approx 0.99995$$

$$\text{Therefore, } P(Z \geq 3.9) = 1 - P(Z < 3.9) = 1 - 0.99995 = 0.00005$$

Conclusion: The probability of getting 70 or more heads in 100 flips of a fair coin i

b) That a randomly selected person will weight over 150 kg, assuming mean and standard deviation of 55/9 kg for women and 70/11 kg for men?

We have two separate populations: women and men. A person is selected random

We assume weights are normally distributed:$X \sim N(\mu, \sigma^2)$

For Women:

$$\text{Mean } (\mu_w) = 55 \text{ kg}$$

$$\text{Standard Deviation } (\sigma_w) = 9 \text{ kg}$$

$$\text{We want } P(X_w > 150)$$

$$\text{Z-score: } Z_w = \frac{150 - 55}{9} = \frac{95}{9} \approx 10.56$$

$P(Z > 10.56)$ is virtually 0. It's so far in the tail that it's effectively impossible un

For Men:

$$\text{Mean } (\mu_m) = 70 \text{ kg}$$

$$\text{Standard Deviation } (\sigma_m) = 11 \text{ kg}$$

$$\text{We want } P(X_m > 150)$$

$$\text{Z-score: } Z_m = \frac{150 - 70}{11} = \frac{80}{11} \approx 7.27$$

$P(Z > 7.27)$ is also virtually 0.

Conclusion: Regardless of whether the randomly selected person is a man or a w

weighing over 150 kg is extremely close to zero. This suggests that 150 kg is an ext

2.) The average on a history exam was 85 out of 100 points, with a standard deviation of 15. Was the distribution of the scores on this exam symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

The distribution would not be symmetric but rather would be left-skewed(negatively skewed). For a distribution to be symmetric, data values are spread evenly in both direction. In this case, majoriy of students are at the 85 points and the standard deviation shows that they are only 15 points away from the maximum points of 100. They cannot score more than 100. However, the few students who score badly, pull the average down making the left tail longer. The low performing students are far away from the 85 points if they scored 0 and even if a student scored 100, the 15 extra points from 85 is not enough to cover the 85 points of the student who scored zero or even the 55 points of a student who scored 30

3.) Facebook data shows that 50% of Facebook users have a hundred or more friends. Further, the average user's friend count is 190. What do these findings say about the shape of the distribution of number of friends of Facebook users?

The distribution is right skewed(Positive). If 50 % of fcebook Users have 100 or more friends, then we can say the median number of Friends for users is 100. The mean is however, 190. The rule of thumb states that if mean is greater than median, then the distribution is positively skewed. The increase in mean is due to the possibility of a few users having way more than 100 friends on Facebook which affects the mean but does not affect the median as such.

4.) Assume that the European electricity prices for non-household consumers in the first half of 2025 are above €0.20 per kWh on 50% of all days. Furthermore, assume that the average electricity price per day is €0.25 per kWh in the same time period. What do these results tell us about the way electricity prices are distributed?

The distribution is right skewed since the median price is 0.20 Euros and the average, which is the mean is 0.25 euros. The mean is greater than the median and the reasons above apply.

Part 2: Significance & Permutation Tests / 30

4.) Which of the following events are likely independent and which are not? Explain your reasoning.

a) Coin tosses

Coin tosses are independent events. A coin has no memory previous toss doesn't change next toss.

b) Goals in soccer

Goals in soccer is not independent. If a team scores once, momentum, psychology, strategy all change. The probability of scoring again increases or decreases.

c) Party success rates in presidential elections

Party success rates in presidential elections are not independent. Historical success influences future success. Loyal voters, political climate.

d) Electricity price fluctuations on consecutive days

It is not independent. Prices follow trends seasonality, and market conditions. Today's price affects tomorrow's.

5.) The 2010 American Community Survey estimates that 47.1% of women aged 15 years and over are married. (Assume that marriages happen independently from one another.)

a) Randomly select three women between these ages. What is the probability that the third woman selected is the only one that is married?

In [1]:
```
#Given probability  that a randomly selected woman is married:
p=0.471
prob_only_third_married = (1 - p) * (1 - p) * p
print(prob_only_third_married)
```

```
0.131805111
```

b) What is the probability that all three women are married?

In [2]:
```
p=0.471
prob_all_three_married = p ** 3
print(prob_all_three_married)
```

```
0.10448711099999998
```

6.) Obtain data on the heights of m men and w women.

a) Use a t-test to establish the significance of whether the men are on average taller than the women.

In [12…
```
import pandas as pd
df=pd.read_csv("height_weight.csv")
df.head()
```

Out[12]:

	Height	Weight	Sex
0	146.323241	59.861065	Female
1	175.695412	77.863687	Male
2	183.216164	72.131992	Male
3	184.245269	77.546000	Male
4	132.302261	55.188496	Female

In [15…
```
men = df[df["Sex"] == "Male"]["Height"]
women = df[df["Sex"] == "Female"]["Height"]
```

In [19…
```
from scipy.stats import ttest_ind

t_stat, p_value = ttest_ind(men, women, equal_var=False)
```

```
print("T-statistic:", t_stat)
print("P-value:", p_value)
```

```
T-statistic: 16.067551597775196
P-value: 6.251107796682894e-38
```

A t-statistic of 16 is extremely large. This means:

The difference between male and female average heights is very large

Compared to the variation (standard deviation) inside each group

So the two groups are very different

Large |t| value → strong evidence the means are different.

The p-value written in normal form:

p=0.000000000000000000000000000000000000625

This is almost zero.

✓ Interpretation:

This p-value is far below any typical significance level (0.05, 0.01, even 0.0001)

Therefore, the difference in height between men and women is statistically significant

The probability that this height difference happens by pure chance is essentially zero

There is a highly statistically significant difference in average height between men and women (t = 16.07, p < 0.0001). Men are on average much taller than women.

b) Perform a permutation test to establish the same thing: whether the men are on average taller than the women.

In [18…

```
import numpy as np

# Extract heights
men_heights = df[df["Sex"] == "Male"]["Height"].values
women_heights = df[df["Sex"] == "Female"]["Height"].values

# Observed difference in mean height
obs_diff = abs(men_heights.mean() - women_heights.mean())

# Combine all heights
all_heights = df["Height"].values

# Number of permutations
num_permutations = 5000
permutation_diffs = []

# Permutation testing
```

```
for _ in range(num_permutations):
    # Shuffle all heights
    shuffled = np.random.permutation(all_heights)

    # Assign new groups with same sizes
    new_men = shuffled[:len(men_heights)]
    new_women = shuffled[len(men_heights):]

    # Compute difference in this permutation
    diff = abs(new_men.mean() - new_women.mean())
    permutation_diffs.append(diff)

# Compute p-value (proportion of permuted diffs >= observed diff)
p_perm = np.mean(np.array(permutation_diffs) >= obs_diff)

print("Observed difference in means:", obs_diff)
print("Permutation p-value:", p_perm)
```

```
Observed difference in means: 22.667343917962597
Permutation p-value: 0.0
```

The permutation test compares the real difference in mean height to the differences produced by randomly shuffling gender labels. The resulting p-value is extremely small (close to 0), meaning that it is very unlikely to observe such a large difference if gender had no effect.

Therefore, the permutation test confirms that men are on average significantly taller than women.

Part 3: Building Models

/ 25

7.) Quantum physics is much more complicated than Newtonian physics. Which model is preferable according to the Occam's Razor, and why?

Occam's Razor favors the simplest model that fits the facts. For everyday objects, Newtons physics is simpler and Adequate. In the case of everyday objects, Occam's Razor would discourges using Quantum phsyics since it complex and Newtons phsics is already adequate for this. However, for subatomic particles, the simple Netwons physics does the fit the fact and the complex quantum phyics in this case, would be recommended since its more accurate and fit the facts of subatomic particles.

8.) Name 2 models that predict something that you are personally interested in. For each of these, decide (and briefly explain) which properties these models have:

a) Are they discrete or continuous?

Stock Price prediction model- These trade in discrete since there are increments in figures but they are continous as well since the returns are in the form of perscentages

Electricity price forecasting models- Prices of electricity are mostly in continous values

b) Are they linear or non-linear?

Stock prices are non linear. As it is famously stated, the market is a complex adaptive system. Key changes lead to disproportional effects. Electricity prices are also non linear as it is characterized with seasons, spikes in prices e.t.c

c) Are they blackbox or descriptive?

Stock prices are mostly black box both in the datascience sector like deeplearning and hedge funds in the Finance sector which are highly blackbox

Electricity prices are maybe hybrid but lean more towards blackbox. Descriptive feature are like the Economics part of it involving supply and demand but there is the black box part where Machine learning is involved in finding complex patterns from in historical data that aren't easily interpretable

d) Are they data driven or first principle?

Stock prices are data driven and uses historical prices Electricity prices are also data driven and uses also historical prices, weather patterns, demands etc

9.) Give at least 1 example of a first-principle and at least 1 example of a data-driven model used in practice.

First Principle example: Black-Scholes Option Pricing Model. This is mathematical model for estimating the price of European-style options. It is based on the fundamental economic principles of no-arbitrage pricing and Uses partial differential equations to derive a theoretical price. It Doesn't require historical option price data to build the model

formula:

$$C = S * N(d1) - K * e^{(} - rt) * N(d2)$$

Data driven model: Credit Score model. It analyzes historical data from millions of borrowers to identify patterns associated with creditworthiness and uses machine learning algorithms to train the model. There is no theoretical economic principle that determines why 30% credit utilization is optimal. This was determined through data analysis.

Part 4: Evaluating Models

/ 20

10.) Suppose you build a classifier that answers yes on every possible input. What precision and recall will this classifier achieve?

Your classifier always says YES

So:

If the true label is YES, you predict YES → True Positive (TP) If the true label is NO, you still predict YES → False Positive (FP) If the true label is NO, you predict NO → True Negative (TN) If the true label is YES, you still predict NO → False Negative (FN)

Because you never predict NO, you will have: TP > 0 FP > 0 FN = 0 TN = 0

Recall=TP/TP+FN But since FN = 0 Recall=1

Precision = (number of real positives) / (total predictions)

Precision is low

Recall = 1 (100%) Precision = TP / (TP + FP) → very low, because the classifier says "yes" to everything.

11.) Is it better to have too many false positives, or too many false negatives? Explain.

If the true label is NO, you still predict YES → False Positive (FP) If the true label is YES, you still predict NO → False Negative (FN) It is usually worse to have too many false negatives.

A false positive (FP) means the model says “YES” even though the true answer is NO. This is usually just annoying. For example: → A healthy person gets a warning that they might be sick. Doctors can check again and correct the mistake.

A false negative (FN) means the model says “NO” even though the true answer is YES. This is usually much more dangerous. For example: → A sick person is told they are healthy. The problem is missed and no treatment is given.

Conclusion:

False negatives are usually worse because they cause real problems to be missed, while false positives only cause extra checks or inconvenience.

Finally: Submission

Save your notebook and submit it (as both **notebook and PDF file**). And please don't forget to ...

- ... choose a **file name** according to convention (see Exercise Sheet 1, but please **add your group name as a suffix** like `_group01`) and to
- ... include the **execution output** in your submission!

In [ ]:
```
!jupyter nbconvert --to webpdf --allow-chromium-download exercise_3_owiti_r
```