

Statistical Data Analysis
Problem Sheet 1
(Revision and warm-up)

1. Exercise 1 (2+2+2+2 Points)

Let X and Y be random variables. Show that

- (a) $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$, where $a, b \in \mathbb{R}$.
- (b) $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
- (c) $\text{Var}(a + bX) = b^2\text{Var}(X)$, where $a, b \in \mathbb{R}$.
- (d) $\text{Var}(a) = 0$, where $a \in \mathbb{R}$.

2. Exercise 2 (2+2 Points)

Let X_1, \dots, X_n be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$ and define the empirical variance

$$S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (1)$$

Show

- that for estimator S_n^2 the following equivalence holds true

$$S_n^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right) \quad (2)$$

- that estimator S_n^2 is an unbiased estimator of the variance

$$\mathbb{E}[S_n^2] = \sigma^2 \quad (3)$$

Exercise 3 (4+5+3 Points)

Plot

- (a) the probability of a random variable that follows the Binomial distribution $\text{Bin}(n, p)$ for different $p \in \{0.3, 0.5, 0.8\}$ and $n \in \{10, 50\}$.
- (b) the probability of a random variable that follows the Geometric distribution $\text{Geom}(p)$ and the corresponding cumulative distribution function F for different $p \in \{0.3, 0.5, 0.8\}$ for all $x \leq 11$.
- (c) the probability of a random variable that follows the Poisson distribution for different $\lambda \in \{0.3, 2, 6\}$ for $x \leq 16$.

in Python. Attach the plots to your exercise submission.

Homework 1: Statistical Data Analysis

Group Members: Dhvaniben Jasoliya, Leutrim Uka, Tauqeer Rumaney, Nicola Horst, Yuvraj Dhepe

1. Exercise 1 (2+2+2+2 Points)

Let X and Y be random variables. Show that

(a) $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$, where $a, b \in \mathbb{R}$.

(b) $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

(c) $\text{Var}(a + bX) = b^2 \text{Var}(X)$, where $a, b \in \mathbb{R}$.

(d) $\text{Var}(a) = 0$, where $a \in \mathbb{R}$.

Definition of expectation:

• Discrete: $\mathbb{E}[X] = \sum_x x \cdot \overbrace{p(x)}^{\text{pmf}}$

• Continuous: $\mathbb{E}[X] = \int x \cdot \overbrace{f(x)}^{\text{pdf}} dx$

a) We start from the definition of expectation:

$$\begin{aligned}\mathbb{E}[a + bX] &= \sum_x (a + bx) \cdot p(x) = \sum_x (a \cdot p(x) + bx \cdot p(x)) = \sum_x a \cdot p(x) + \sum_x bx \cdot p(x) \\ &= a \cdot \underbrace{\sum_x p(x)}_{=1} + b \cdot \underbrace{\sum_x x \cdot p(x)}_{\mathbb{E}[X]} = a + b \cdot \mathbb{E}[X]\end{aligned}$$

(law of total prob.)

b) $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \Rightarrow$

$$\begin{aligned}&= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X \cdot \mathbb{E}[X] + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[2X \cdot \mathbb{E}[X]] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - \sum_x 2x \cdot (\sum_x x \cdot p(x)) \cdot p(x) + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - 2 \cdot \sum_x \sum_x x \cdot x \cdot p(x) \cdot p(x) + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - 2 \cdot \left(\sum_x x \cdot p(x)\right)^2 + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - 2 \cdot (\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

$$\begin{aligned}\text{c) } \text{Var}(a + bX) &= \mathbb{E}[(a + bX)^2] - (\mathbb{E}[a + bX])^2 \\ &= \mathbb{E}[a^2 + 2abX + b^2X^2] - (a + b\mathbb{E}[X])^2 \\ &= \cancel{a^2} + \cancel{2ab \cdot \mathbb{E}[X]} + b^2 \mathbb{E}[X^2] - \cancel{a^2} - \cancel{2ab \mathbb{E}[X]} - b^2 \mathbb{E}[X]^2 \\ &= b^2 (\mathbb{E}[X^2] - \mathbb{E}[X]^2) \\ &= b^2 \cdot \text{Var}(X)\end{aligned}$$

$$\begin{aligned}\text{d) } \text{Var}(a) &= \mathbb{E}[(a - \mathbb{E}[a])^2] = \mathbb{E}[(a - \sum_x a \cdot p(x))^2] \\ &= \mathbb{E}[(a - a)^2] = \mathbb{E}[0] \\ &= \sum_x 0 \cdot p(x) = 0\end{aligned}$$

2. Exercise 2 (2+2 Points)

Let X_1, \dots, X_n be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$ and define the empirical variance

$$S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (1)$$

Show

- that for estimator S_n^2 the following equivalence holds true

$$S_n^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right) \quad (2)$$

$$\begin{aligned} \bullet S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - 2x_i \bar{x}_n + \bar{x}_n^2) \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (x_i^2) - 2 \cdot \bar{x}_n \underbrace{\sum_{i=1}^n x_i}_{=n\bar{x}_n} + \sum_{i=1}^n \bar{x}_n^2 \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (x_i^2) - 2 \cdot \bar{x}_n \cdot n \cdot \bar{x}_n + n \cdot \bar{x}_n^2 \right] \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n (x_i^2) - n\bar{x}_n^2 \right) \end{aligned}$$

- that estimator S_n^2 is an unbiased estimator of the variance

• We start off by taking the expectation of the expression we just proved: $\mathbb{E}[S_n^2] = \sigma^2$ (3)

$$\begin{aligned} \mathbb{E}[S_n^2] &= \mathbb{E} \left[\frac{1}{n-1} \left(\sum_{i=1}^n (x_i^2) - n\bar{x}_n^2 \right) \right] \\ &= \frac{1}{n-1} \cdot \left[\underbrace{\sum_{i=1}^n \mathbb{E}[x_i^2]}_{\substack{\mathbb{E}[x_1] + \mathbb{E}[x_2] + \dots = \mu + \mu + \dots = n\mu \\ \times \mu = \mathbb{E}[x_i]}} - n \cdot \mathbb{E}[\bar{x}_n^2] \right] \\ &= \frac{1}{n-1} \left[n \cdot \mathbb{E}[x_i^2] - n \cdot \mathbb{E}[\bar{x}_n^2] \right] \\ &= \frac{1}{n-1} \left[n \cdot (\mathbb{E}[x_i^2] - \mathbb{E}[\bar{x}_n^2]) \right] \\ &= \dots \quad (*) \end{aligned}$$

We can express $\mathbb{E}[x_i^2]$ using the variance definition: we proved in task 1b:

$$\text{Var}(x_i) = \mathbb{E}[x_i^2] - (\mathbb{E}[x_i])^2$$

We know that $\text{Var}(x_i) = \sigma$ & $\mathbb{E}(x_i) = \mu$:

$$\sigma^2 = \mathbb{E}[x_i^2] - \mu^2$$

$$\boxed{\mathbb{E}[x_i^2] = \sigma^2 + \mu^2}$$

• Similarly, for \bar{x}_n we have:

$$\text{Var}(\bar{x}_n) = \mathbb{E}[x_n^2] - (\mathbb{E}[\bar{x}_n])^2$$

$$\frac{\sigma^2}{n} = \mathbb{E}[x_n^2] - \mu^2$$

$$\mathbb{E}[x_n^2] = \frac{\sigma^2}{n} + \mu^2$$

$$(*) \dots = \frac{1}{n-1} \left[n \cdot (\sigma^2 + \mu^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) \right]$$

$$= \frac{1}{n-1} \left[n\sigma^2 + \cancel{n\mu^2} - \sigma^2 - \cancel{n\mu^2} \right]$$

$$= \frac{1}{\cancel{n-1}} \left[\sigma^2 (\cancel{n-1}) \right]$$

$$\Rightarrow \boxed{\mathbb{E}[s_n^2] = \sigma^2}$$

Exercise_1_task_3

November 3, 2023

```
[1]: import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

```
[2]: # set seaborn style
sns.set(style="whitegrid")

# define a custom color palette
custom_palette = sns.color_palette("Set2")
```

1 Exercise 3

1.1 (a)

the probability of a random variable that follows the Binomial distribution $\text{Bin}(n, p)$ for different $p \in \{0.3, 0.5, 0.8\}$ and $n \in \{10, 50\}$.

```
[3]: num_samples: int = 100000
p: list = [0.3, 0.5, 0.8]
n: list = [10, 50]
```

```
[4]: fig, axs = plt.subplots(len(n), len(p), figsize=(12, 8), sharey=True)
for i, _n in enumerate(n):
    for j, _p in enumerate(p):
        # set up x values and empty probabilities
        values = list(range(0, _n + 1))
        probabilities = np.zeros(_n + 1)

        # generate samples
        unique_values, counts = np.unique(np.random.binomial(_n, _p,
↪num_samples), return_counts=True)

        # compute probs
        probabilities[unique_values] = counts/num_samples

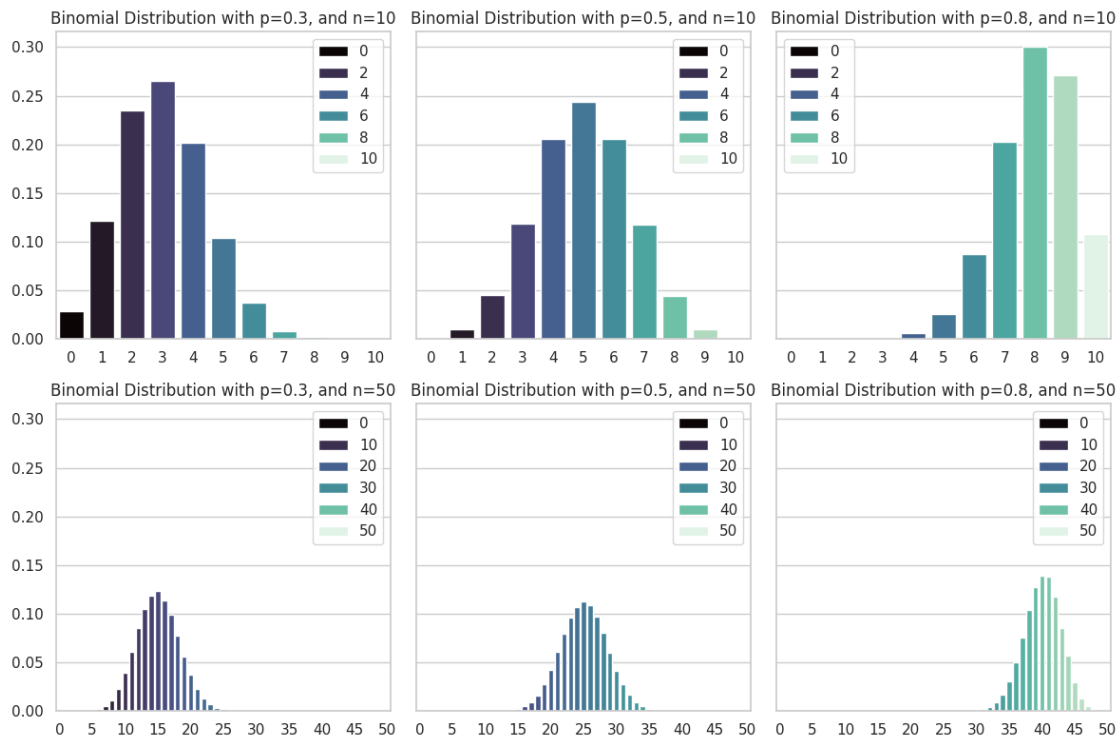
        # plot
```

```

sns.barplot(x = values, hue=values, y=probabilities, ax=axes[i, j],
↪palette='mako', legend = True)
axes[i, j].set_title(f"Binomial Distribution with p={_p}, and n={_n}")

# adjust x ticks when n = 50
if _n == 50:
    axes[i, j].set_xticks(list(range(0, _n + 1, 5)))
    axes[i, j].set_xticklabels(list(range(0, _n + 1, 5)))
plt.tight_layout()
plt.show()

```



1.2 (b)

the probability of a random variable that follows the Geometric distribution $\text{Geom}(p)$ and the corresponding cumulative distribution function F for different $p \in \{0.3, 0.5, 0.8\}$ for all $x \leq 11$.

```

[5]: p: list = [0.3, 0.5, 0.8]
size=num_samples

```

```

[6]: fig, axes = plt.subplots(2, len(p), figsize=(18, 8))
for j, _p in enumerate(p):
    values, counts = np.unique(np.random.geometric(_p, size),
↪return_counts=True)

```

```

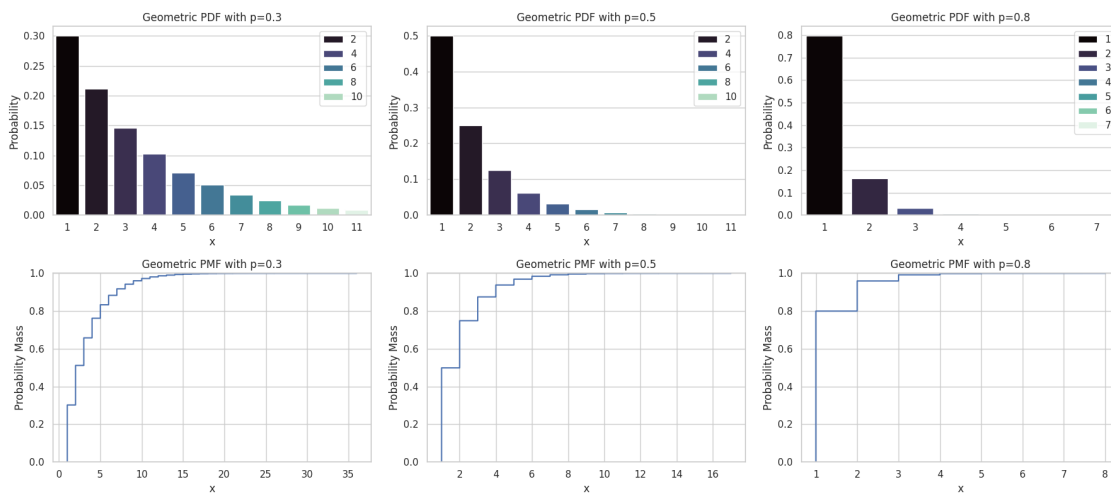
# cut to be smaller 11
counts = counts[values <= 11]
values = values[values <= 11]

sns.barplot(x=values, hue = values, y=counts/size, ax=axes[0, j], palette =
↳ 'mako')
axes[0, j].set_title(f"Geometric PDF with p={_p}")
axes[0, j].set_ylabel("Probability")
axes[0, j].set_xlabel("x")

sns.ecdfplot(np.random.geometric(_p, size), ax=axes[1, j],
↳ drawstyle='steps-pre')
axes[1, j].set_title(f"Geometric PMF with p={_p}")
axes[1, j].set_ylabel("Probability Mass")
axes[1, j].set_xlabel("x")

plt.tight_layout()

```



1.3 (c)

the probability of a random variable that follows the Poisson distribution for different λ {0.3, 2, 6} for $x = 16$.

```
[7]: lambdas: list = [0.3, 2, 6]
```

```
[8]: fig, axes = plt.subplots(1, 3, figsize=(12, 4), sharey=True)
for j, l in enumerate(lambdas):
    values, counts = np.unique(np.random.poisson(l, size), return_counts=True)
```

```
# cut x to <= 16
counts = counts[values <= 16]
values = values[values <= 16]

sns.barplot(x=values, hue = values, y=counts/size, ax=axes[j], palette='mako')
axes[j].set_title(f"poisson dist with  $\lambda = \{1\}")
axes[j].set_ylabel(f"probability")
axes[j].set_xlabel(f"x")$ 
```

