I. 7/8
II. 4/4
III.12/12

23/24

Institut für Mathematik
Lehrstuhl Datenassimilation

Winter Semester 2023/2024
Dr. Tim Jahn

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

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

   Show

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

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

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

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

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

Definition of expectation:

- Discrete: $\mathbb{E}[X] = \sum_x x \cdot \overbrace{p(x)}^{\text{pmf}}$
- Continuous: $\mathbb{E}[X] = \int x \cdot \overbrace{f(x)}^{\text{pdf}}\, dx$

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

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

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

a) We start from the definition of expectation:

$$\mathbb{E}[a+bX] = \sum_x (a+bx)p(x) = \sum_x (a\cdot p(x) + bx\,p(x)) = \sum_x a\cdot p(x) + \sum_x bx\cdot p(x)$$

$$= a\cdot \underbrace{\sum_x p(x)}_{\substack{=1 \\ \text{(law of total prob.)}}} + b\cdot \underbrace{\sum_x x\cdot p(x)}_{\mathbb{E}[x]} = a + b\cdot \mathbb{E}[x]$$

How about the continuous case ?!

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

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

✓

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

✓

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

✓

2. **Exercise 2 (2+2 Points)**

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

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

Show

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

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

- $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2\cdot x_i\cdot\bar{X}_n + \bar{X}_n^2)$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i)^2 - 2\cdot\bar{X}_n\underbrace{\sum_{i=1}^{n}x_i}_{\text{IID}} + \sum_{i=1}^{n}\bar{X}_n^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i)^2 - 2\cdot\bar{X}_n\cdot n\cdot\bar{X}_n + n\cdot\bar{X}_n^2\right]$$

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

✓

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

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

· We start off by taking the expectation of the expression we just proved:

$$\mathbb{E}[S_n^2] = \mathbb{E}\left[\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i)^2 - n\cdot\bar{X}_n^2\right)\right]$$

$$= \frac{1}{n-1}\cdot\left[\boxed{\sum_{i=1}^{n}\mathbb{E}[x_i^2]} - n\cdot\mathbb{E}[\bar{X}_n^2]\right]$$

$\mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots = \mu + \mu + \ldots = n\cdot\mu$

$\& \ \mu = \mathbb{E}[X_i]$

$$= \frac{1}{n-1}\left[n\cdot\mathbb{E}[x_i^2] - n\cdot\mathbb{E}[\bar{X}_n^2]\right]$$

$$= \frac{1}{n-1}\left[n\cdot\left(\mathbb{E}[x_i^2] - \mathbb{E}[\bar{X}_n^2]\right)\right]$$

$$= \ldots \ (*)$$

✓

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

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

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

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

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

- Similarly, for $\bar{X}_n$ we have:

$$Var(\bar{X}_n) = \mathbb{E}[X_n^2] - (\mathbb{E}[\bar{X}_n])^2$$

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

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

✓

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

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

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

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

✓

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 Bin(n, p) for different p {0.3, 0.5, 0.8} and n {10, 50}.

[3]:

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

[4]:

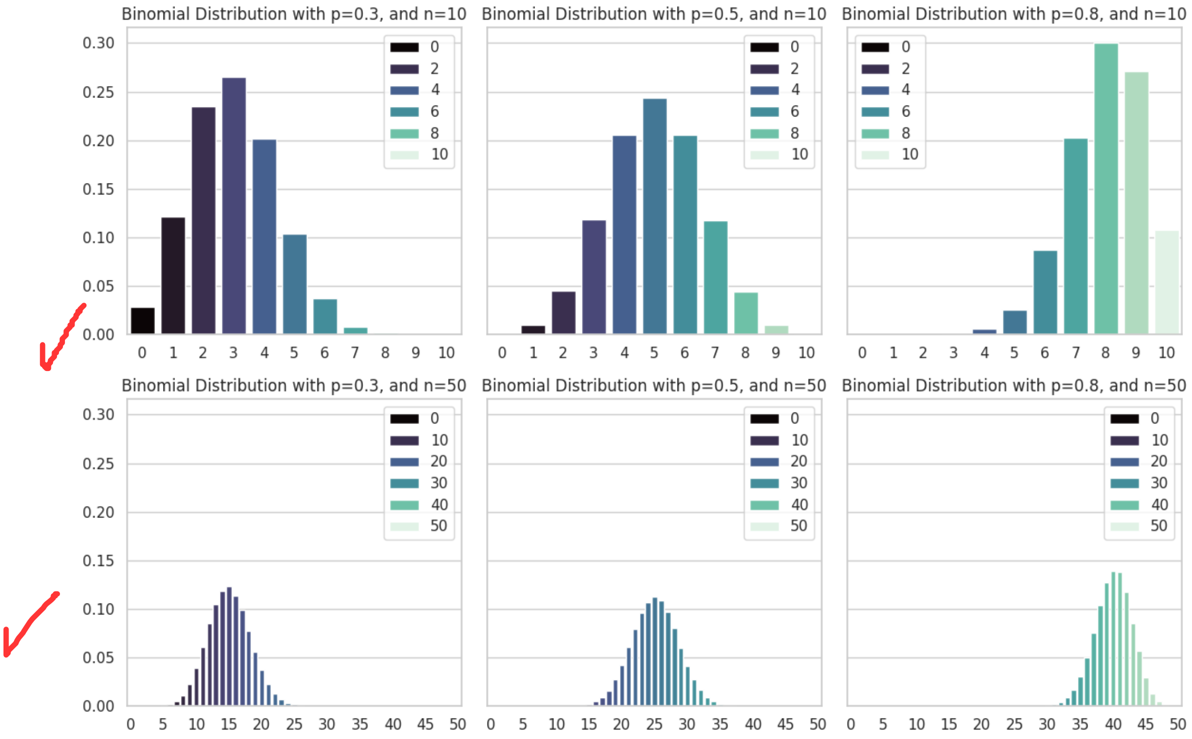
```
fig, axs = plt.subplots(len(n), len(p), figsize=(12, 8), sharey=True)
for i, _n in enumerate(n):
    for j, _p in enumerate(p):
        # set up x values and empty probabilities
        values = list(range(0, _n + 1))
        probabilities = np.zeros(_n + 1)

        # 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=axs[i, j],␣
 ↪palette='mako', legend = True)
        axs[i, j].set_title(f"Binomial Distribution with p={_p}, and n={_n}")

        # adjust x ticks when n = 50
        if _n == 50:
            axs[i, j].set_xticks(list(range(0, _n + 1, 5)))
            axs[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 Geom(p) and the corresponding cumulative distribution function F for different p {0.3, 0.5, 0.8} for all x 11.

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

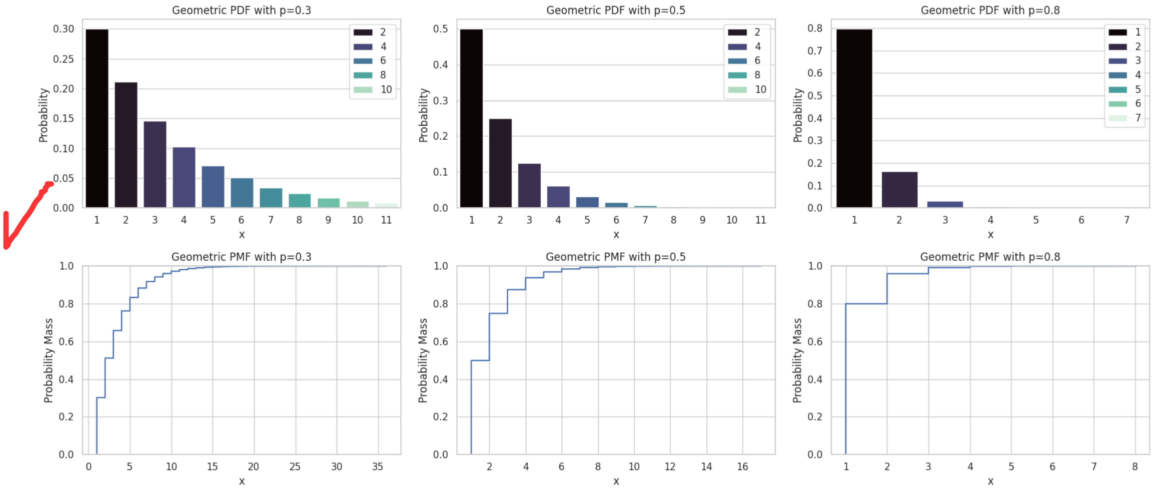
```
[6]: fig, axs = plt.subplots(2, len(p), figsize=(18, 8))
     for j, _p in enumerate(p):
         values, counts = np.unique(np.random.geometric(_p, size),␣
      ↪return_counts=True)
```

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

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

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

plt.tight_layout()
```



why the values of x=0 are assigned to x=1 ?

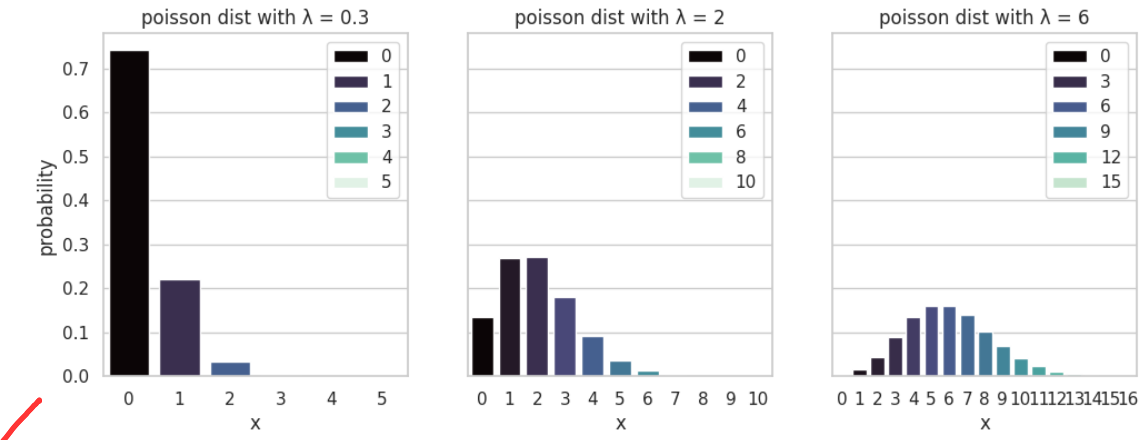
1.3 (c)

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

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

[8]:
```
fig, axs = 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=axs[j], palette='mako')
    axs[j].set_title(f"poisson dist with   = {l}")
    axs[j].set_ylabel(f"probability")
    axs[j].set_xlabel(f"x")
```



✓ ✓ ✓