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 member names here]

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

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

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

a) We start from the definition of expectation:

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

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

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

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

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

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

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

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

Show

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

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

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

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

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

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

- 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:

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

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

$\mathbb{E}[x_1] + \mathbb{E}[x_2] + \dots = \mu + \mu + \dots = n\mu$
& $\mu = \mathbb{E}[x_i]$

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

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

$$= \dots \; (*)$$

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:

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

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

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

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

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

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

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

Task 3

In [74]:

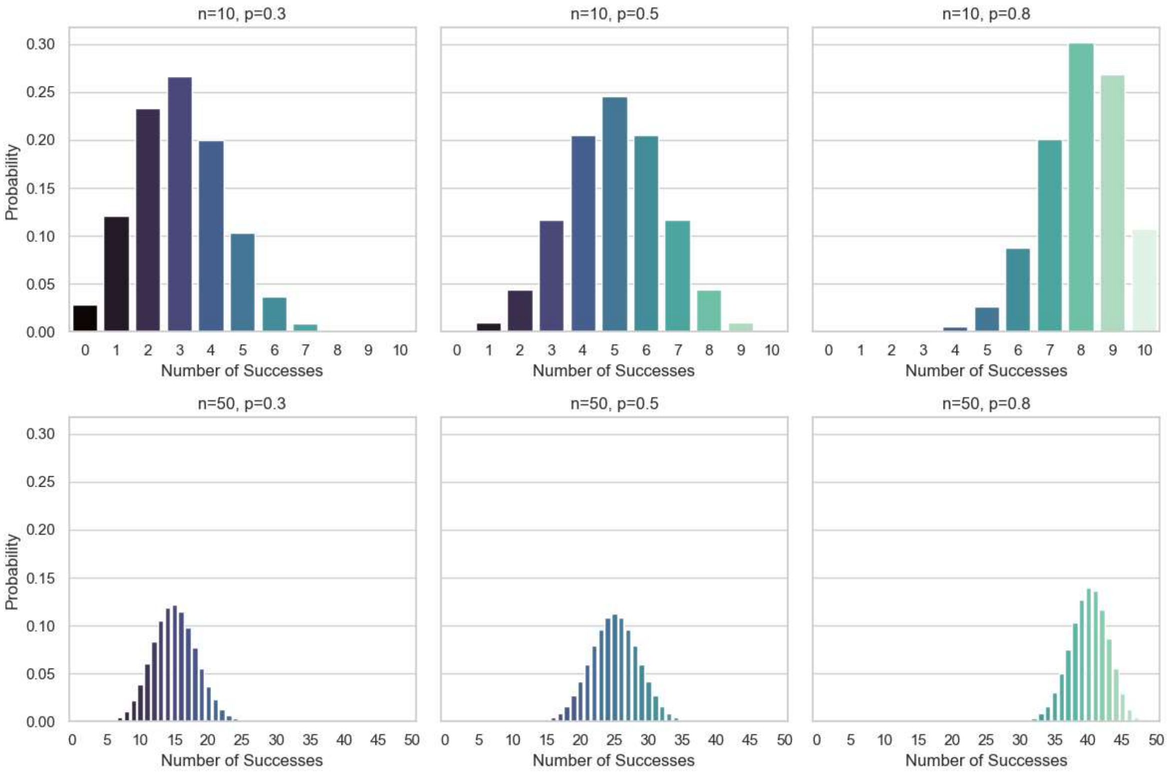
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import math
```

Plot

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

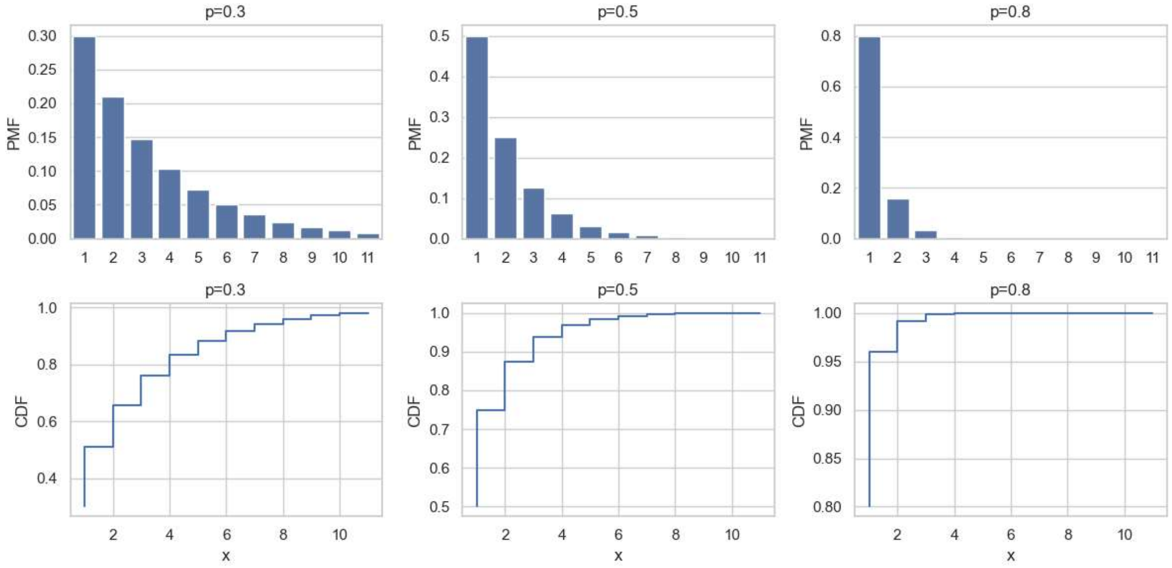
In [75]:

```
# Set Seaborn style
sns.set(style="whitegrid")

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

# Values of n and p
n_values = [10, 50]
p_values = [0.3, 0.5, 0.8]

fig, axes = plt.subplots(len(n_values), len(p_values), figsize=(12, 8), sharey=T

for i, n in enumerate(n_values):
    for j, p in enumerate(p_values):
        # Generate the possible outcomes (0 to n successes)
        if i == 0:
            x = np.arange(0, 11)
        else:
            x = np.arange(0, n + 1)

        # binomial formula
        probabilities = [math.comb(n, k) * (p**k) * ((1-p)**(n-k)) for k in x]

        # Create n*p subplots, one for each experiment using Seaborn
        sns.barplot(x=x, y=probabilities, ax=axes[i, j], palette='mako', hue=x,
        axes[i, j].set_title(f'n={n}, p={p}')
        axes[i, j].set_xlabel('Number of Successes')
        axes[i, j].set_ylabel('Probability')

        # Adjust x-axis ticks in the second row (where n=50)
        if n == 50:
            axes[i, j].set_xticks(np.arange(0, n + 1, 5))
            axes[i, j].set_xticklabels(np.arange(0, n + 1, 5))

plt.tight_layout()
plt.show()
```



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

In [72]:
```
# Different probability values
p_values = [0.3, 0.5, 0.8]

fig, axes = plt.subplots(2, len(p_values), figsize=(12, 6))

for j, p in enumerate(p_values):
    # Make a list of x values, where x <= 11
    x = np.arange(1, 12)

    # probability mass funtion for geometric distribution
    pmf = [(1-p)**(k-1) * p for k in x]
    pmf_mean = 1/p

    # cummulative distribution function for geometric distribution
    cdf = 1 - (1-p)**x

    # 3 plots for pmf, each with a different p-value
    sns.barplot(x=x, y=pmf, ax=axes[0, j], legend=False)
    axes[0, j].set_title(f'p={p}')
    axes[0, j].set_ylabel('PMF')

    # 3 plots for cdf, each with a different p-value
    sns.lineplot(x=x, y=cdf, ax=axes[1, j], drawstyle='steps-pre')
    axes[1, j].set_title(f'p={p}')
    axes[1, j].set_xlabel('x')
    axes[1, j].set_ylabel('CDF')

plt.tight_layout()
plt.show()
```



(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 [79]:
```
# Different lambda values
lambda_values = [0.3, 2, 6]

# Make a list of x values, where x<=16
x_values = np.arange(0, 17)

# Create 3 subplots, one for each  λ value
fig, axes = plt.subplots(1, len(lambda_values), figsize=(12, 4))

for i, lmbda in enumerate(lambda_values):
    # Calculate the Poisson PMF for each x value
    pmf = np.exp(-lmbda) * (lmbda**x_values) / [math.factorial(x) for x in x_val

    # Plot the PMF
    axes[i].bar(x_values, pmf, align='center', alpha=0.5, label=f'λ={lmbda}')
    axes[i].set_title(f'Poisson PMF (λ={lmbda})')
    axes[i].set_xlabel('x')
    axes[i].set_ylabel('Probability')
    axes[i].legend()

plt.tight_layout()
plt.show()
```