Assignment 2.1

Name: Parker Christenson

Date: 09/12/2023

For this assignment, you will refer to the textbook to solve the practice exercises. **Use Python to answer any coding problems (not R, even if indicated in your textbook).** Use Jupyter Notebook, Google Colab, or a similar software program to complete your assignment. Submit your answers as a **PDF or HTML** file. As a best practice, always label your axes and provide titles for any graphs generated on this assignment. Round all quantitative answers to 2 decimal places.

These are all of the libraries that I will use for this assignment.

In [7]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
```

Problem # 2.1.

For the rain simulation example in Section 2.1.1, but with probability of rain 0.30 on any given day, simulate the outcome (a) on the next day, (b) the next 10 days. (c) Simulate the proportion of days of rain for the next (i) 100 days, (ii) 10,000 days, (iii) 1,000,000 days. Use the simulation to explain the long-run relative frequency definition of probability.

`(a) Your answer goes here`

In [8]:
```
import random
probability_of_rain = 0.30

def calculate_probability_of_rain(num_days):
    # Simulate the outcome for a specified number of days
    rainy_days = sum(1 for _ in range(num_days) if random.random() <= probability_of_r
    probability = rainy_days / num_days
    return probability

next_day_probability = calculate_probability_of_rain(1)

print(next_day_probability)
```

```
1.0
```

`(b) Your answer goes here`

In [9]:
```
next_ten_days_probability = calculate_probability_of_rain(10)
print(next_ten_days_probability)
```

```
0.3
```

`(c) Your answer goes here`

In [10]:
```
next_100_days_probability = calculate_probability_of_rain(100)
next_1000_days_probability = calculate_probability_of_rain(1000)
next_10000_days_probability = calculate_probability_of_rain(10000)

print(next_100_days_probability)
print(next_1000_days_probability)
print(next_10000_days_probability)
```

```
0.38
0.281
0.2989
```

The long run frequenecy definition of probability is that the probability of an event is the proportion of times that the event occurs in a long run of observations. This is shown in the simulation above. As the number of days increases, the proportion of days that it rains approaches the probability of rain. This is because the probability of rain is the proportion of days that it rains in the long run.

With that being said the probability of rain is 0.30. This means that in the long run, 30% of the days will be rainy.

In Latex the probability of rain is $P(Rain) = 0.30$

The LaTex formula looks like this for the Long Run Frequency Definition of Probability:

$$P(A) = \lim_{n \to \infty} \frac{number of times A occurs in n trials}{n}$$

Problem # 2.2.

Data analysts often implement statistical inference methods by setting the probability of a correct inference equal to 0.95. Let *A* denote the event that an inference for the population about men is correct. Let *B* represent the event of a corresponding inference about women being correct. Suppose that these are independent events.

(a) Find the probability that (i) *both* inferences are correct, (ii) *neither* inference is correct.
(b) Construct the probability distribution for *Y* = number of correct inferences.
(c) With what probability would each inference need to be correct in order for the probability to be 0.95 that *both* are correct?

`(a) Your answer goes here`

(i) P(Both correct)=P(A)·P(B)

(ii) P(Neither correct)=1−P(A)·P(B)

```
(b) Your answer goes here
```

1 = correct 0 = incorrect 2 = both_correct

P(Y=0) = P(neither correct) = .0025

P(Y-1) = P(one correct) = 1 - P(both correct) - P(neither correct) = .9975

P(Y=2) = P(both_correct) = .9025

```
(c) Your answer goes here
```

P(Both Correct) = P(A) *P(B)* = *.95* .95 = .9025

P(A) = P(B) = √0.95025 ≈ 0.9745

Problem # 2.4.

A wine connoisseur is asked to match five glasses of red wine with the bottles from which they came, representing five different grape types.

(a) Set up a sample space for the five guesses.
(b) With random guessing, find the probability of getting all five correct.

```
(a) Your answer goes here
```

There are 5! ways (120 combos) to match the glasses of wine with the bottles.

```
(b) Your answer goes here
```

With the Random guesses, the probability of getting all five correct is, 1/120 = 0.0083 *100 = 0.83%

Problem # 2.15.

Each week an insurance company records *Y* = number of payments because of a home burning down. State conditions under which we would expect *Y* to approximately have a Poisson distribution.

```
Your answer goes here
```

- The numnber of events must be rare, and the probability of the event happening must be small.
- The events must be independent of each other.

- The avg number of events must be constant.

Problem # 2.16.

Each day a hospital records the number of people who come to the emergency room for treatment.

(a) In the first week, the observations from Sunday to Saturday are 10, 8, 14, 7, 21, 44, 60. Do you think that the Poisson distribution might describe the random variability of this phenomenon adequately. Why or why not?
(b) Would you expect the Poisson distribution to better describe, or more poorly describe, the number of weekly admissions to the hospital for a rare disease? Why?

`(a) Your answer goes here`

The Poisson distribution is a discrete probability distribution that describes the probability of a certain number of events occurring in a fixed interval of time or space, if the events occur independently of each other.

`(b) Your answer goes here`

The Poisson distribution would be more likely to describe the number of weekly admissions to the hospital for a rare disease than the number of daily emergency room visits

Problem # 2.17.

An instructor gives a course grade of B to students who have total score on exams and homeworks between 800 and 900, where the maximum possible is 1000. If the total scores have approximately a normal distribution with mean 830 and standard deviation 50, about what proportion of the students receive a B?

`Your answer goes here`

In [11]:
```
# Calculate the area between -0.60 and 1.40
proportion = stats.norm.cdf(1.40) - stats.norm.cdf(-0.60)

print(proportion)
```

```
0.6449902230161553
```

With that being said, about 64% of the students received a B.

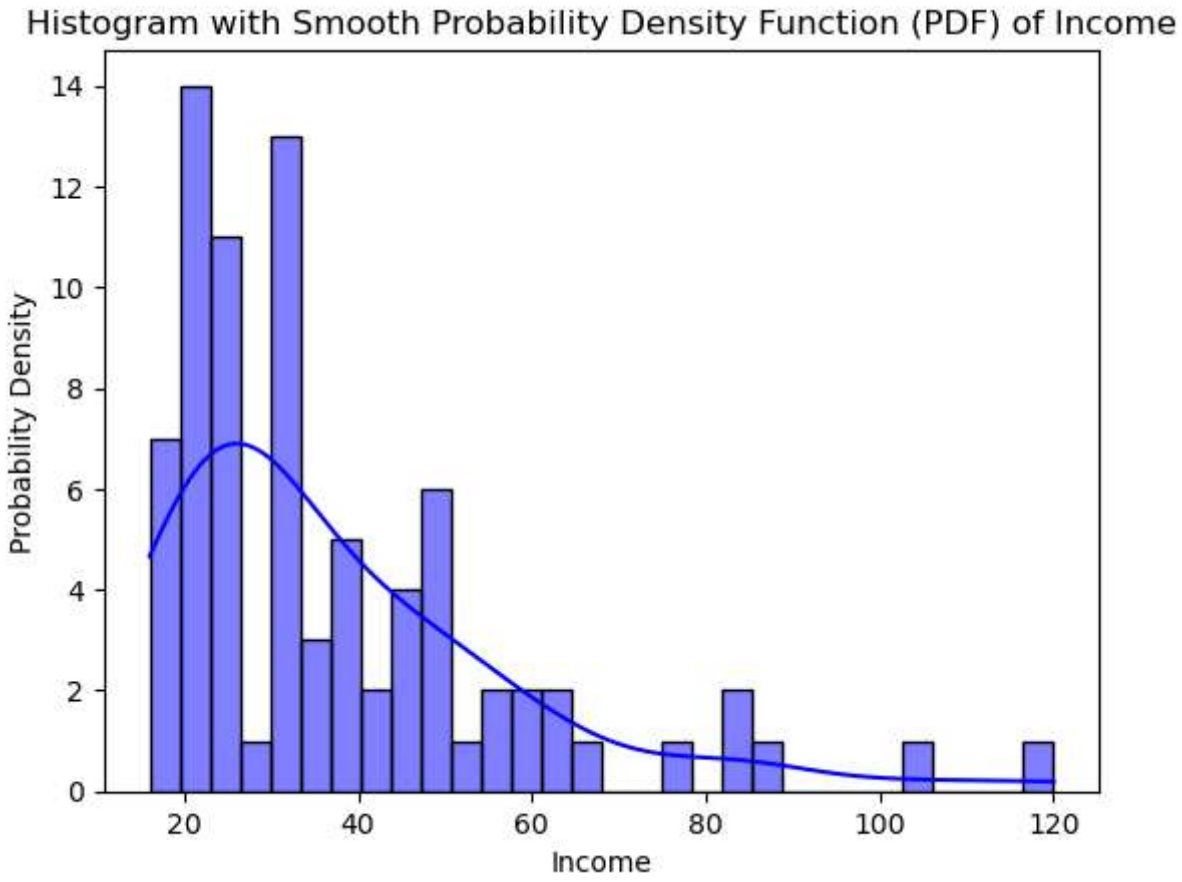
Problem # 2.20.

Create a data file with the income values in the `Income` data file at the text website.

(a) Construct a histogram or a smooth-curve approximation for the *pdf* of income in the corresponding population by plotting results using the density function in R (explained in Exercise 1.18).
(b) Of the probability distributions studied in this chapter, which do you think might be most appropriate for these data? Why? Plot the probability function of that distribution having the same mean and standard deviation as the income values. Does it seem to describe the income distribution well?

(a) Your answer goes here

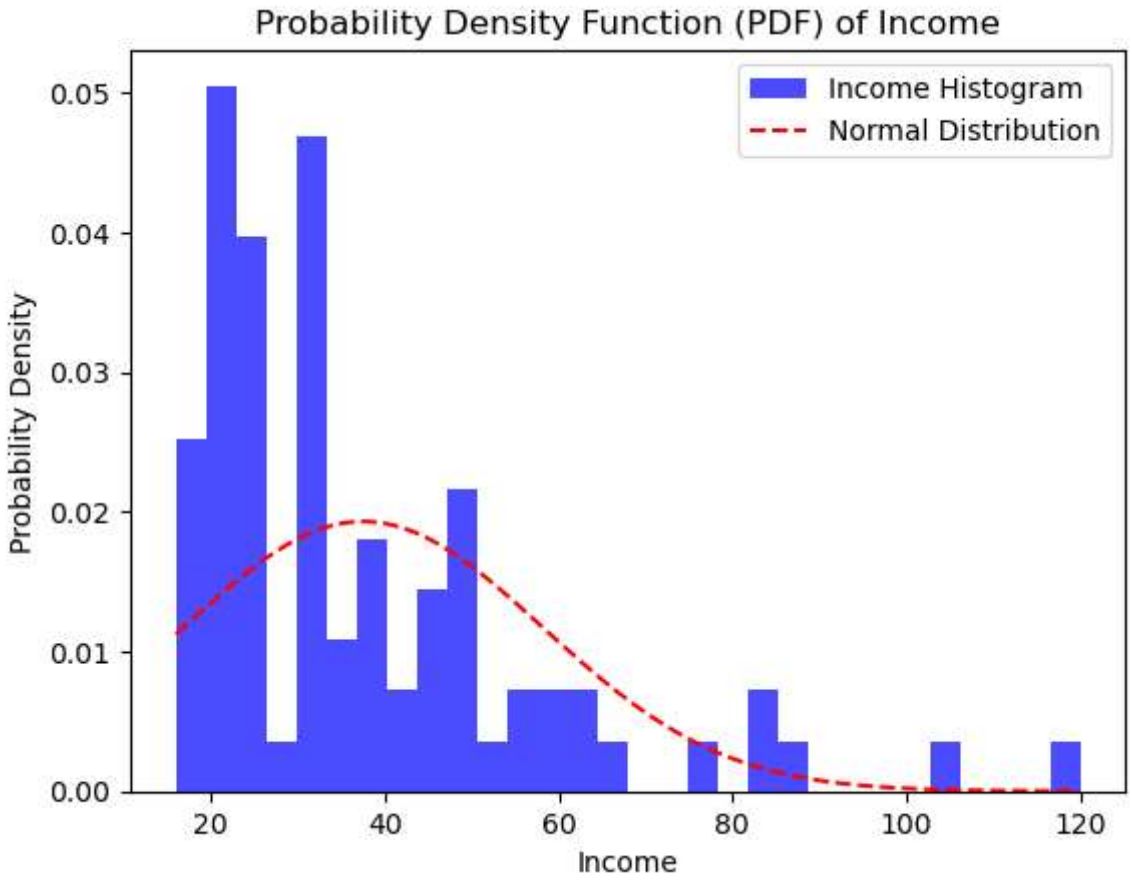
In [12]:
```
df = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_500\Assignments\Assignment_1\Inc

# Extract the income values from the DataFrame
income_values = df['income']

sns.histplot(income_values, bins=30, kde=True, color='blue')
plt.xlabel('Income')
plt.ylabel('Probability Density')
plt.title('Histogram with Smooth Probability Density Function (PDF) of Income')
plt.show()
```



(b) Your answer goes here

In [13]:
```
# Extract the income values from the DataFrame
income_values = df['income']

# Calculate the mean and standard deviation of income data
```

```
mu, std = income_values.mean(), income_values.std()

x = np.linspace(income_values.min(), income_values.max(), 100)

pdf = stats.norm.pdf(x, mu, std)
plt.hist(income_values, bins=30, density=True, alpha=0.7, color='b', label='Income His
plt.plot(x, pdf, 'r--', label='Normal Distribution')
plt.xlabel('Income')
plt.ylabel('Probability Density')
plt.title('Probability Density Function (PDF) of Income')
plt.legend()
plt.show()


# I think that the fitted distribution fits the best. I think that the reason behind t
```



Problem # 2.21.

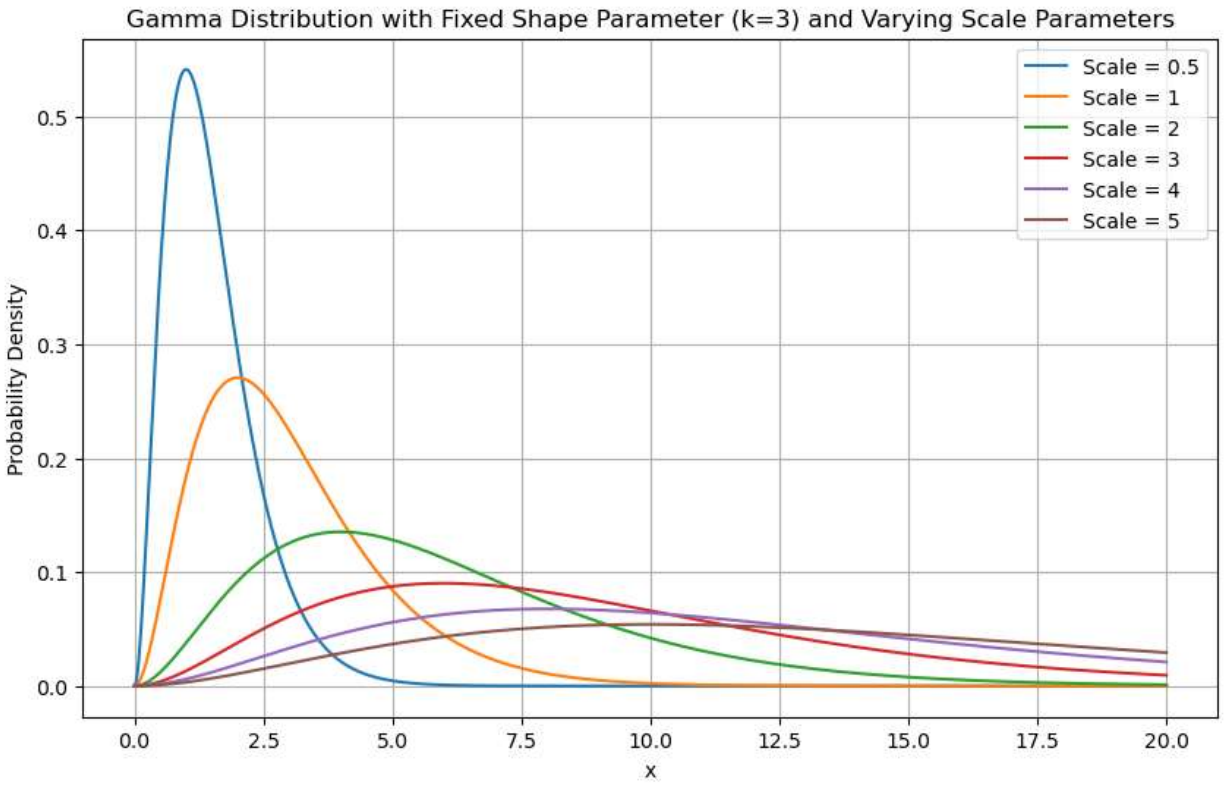
Plot the gamma distribution by fixing the shape parameter k = 3 and setting the scale parameter = 0.5, 1, 2, 3, 4, 5. What is the effect of increasing the scale parameter? (See also Exercise 2.48.)

`Your answer goes here`

In [14]:

```
# Fixed shape parameter (k)
k = 3

# Scale Params
```

```
theta_values = [0.5, 1, 2, 3, 4, 5]

# X Vals
x = np.linspace(0, 20, 1000)

# Plotting Gamma Distributions
plt.figure(figsize=(10, 6))
for theta in theta_values:
    pdf = stats.gamma.pdf(x, a=k, scale=theta)
    plt.plot(x, pdf, label=f'Scale = {theta}')

plt.xlabel('x')
plt.ylabel('Probability Density')
plt.title('Gamma Distribution with Fixed Shape Parameter (k=3) and Varying Scale Param
plt.legend()
plt.grid(True)
plt.show()
```



Problem # 2.22.

Consider the mammogram diagnostic example in Section 2.1.4.

(a) Show that the joint probability distribution of diagnosis and disease status is as shown in Table 2.6. Given that a diagnostic test result is positive, explain how this joint distribution shows that the 12% of incorrect diagnoses for the 99% of women not having breast cancer swamp the 86% of correct diagnoses for the 1% of women actually having breast cancer.
(b) The first test for detecting HIV-positive status had a sensitivity of 0.999 and specificity of 0.9999. Explain what these mean. If at that time 1 in 10,000 men were truly HIVpositive, find the positive predictive value. Based on this example, explain the potential disadvantage of routine diagnostic screening of a population for a rare disease.

TABLE 2.6 Joint probability distribution for disease status and diagnosis of breast cancer mammogram, based on conditional probabilities in Table 2.1

Disease Status	**Diagnosis from Mammogram**		
	Positive (+)	Negative (-)	**Total**
Yes (D)	0.0086	0.0014	0.01
No (D^c)	0.1188	0.8712	0.99

`Your answer goes here`

(a)

P(D) is the probability of a positive test result given that the patient has breast cancer. 1%
P(D^C) is the probability of a positive test result given that the patient does not have breast cancer. 99%

P(+|D) is the probability that the positive test result is correct. 0.0086 or 0.86% P(+|D^C) is the probability that the positive test result is incorrect. .1188 or 11.88%

P(D|+) is the probability of the patient having breast cancer given a positive test result.
P(D^C|+) is the probability of the patient not having breast cancer given a positive test result.

Using Baynes Therom, we can find the probability of the patient having breast cancer given a positive test result.

Banyes Therom is as follows:

P(A|B) = P(B|A) * P(A) / P(B)

The denominator is the probability of a positive test result. And Can be calculated using the law of total probability.

P(+) = P(+|D) *P(D)* + *P(+|D^C)* P(D^C)

After the calculations, we see that:

P(D|+) = 0.0672

and

P(D^C|+) = 0.9328

B

Sensitivity rate: The measure of how well a test can identify true positives Specificity rate: The measure of how well a test can identify true negatives Prevalence: The proportion of people in the population who have the disease

Imma Keep it Real, Im not gunna post these formulas. I will just show the solutions Dawg. This problem Has burnt my brain.

Calculating the PPV:

PPV = True positives / (True positives + False positives) = 0.999/10000 / (0.999/10000 + 0.0001*0.9999) = 0.0098

The Calculated PPV is 0.0098 or 0.98% and that means that the sensitivity and specificity indicates that a person in the polpulation who is randomly selected, will have a very low chance of having HIV.

Problems with routine diagnostic screening of a population for a rare disease:

False Alarms may arise. Over use of resources. Routine screening may not be cost effective.

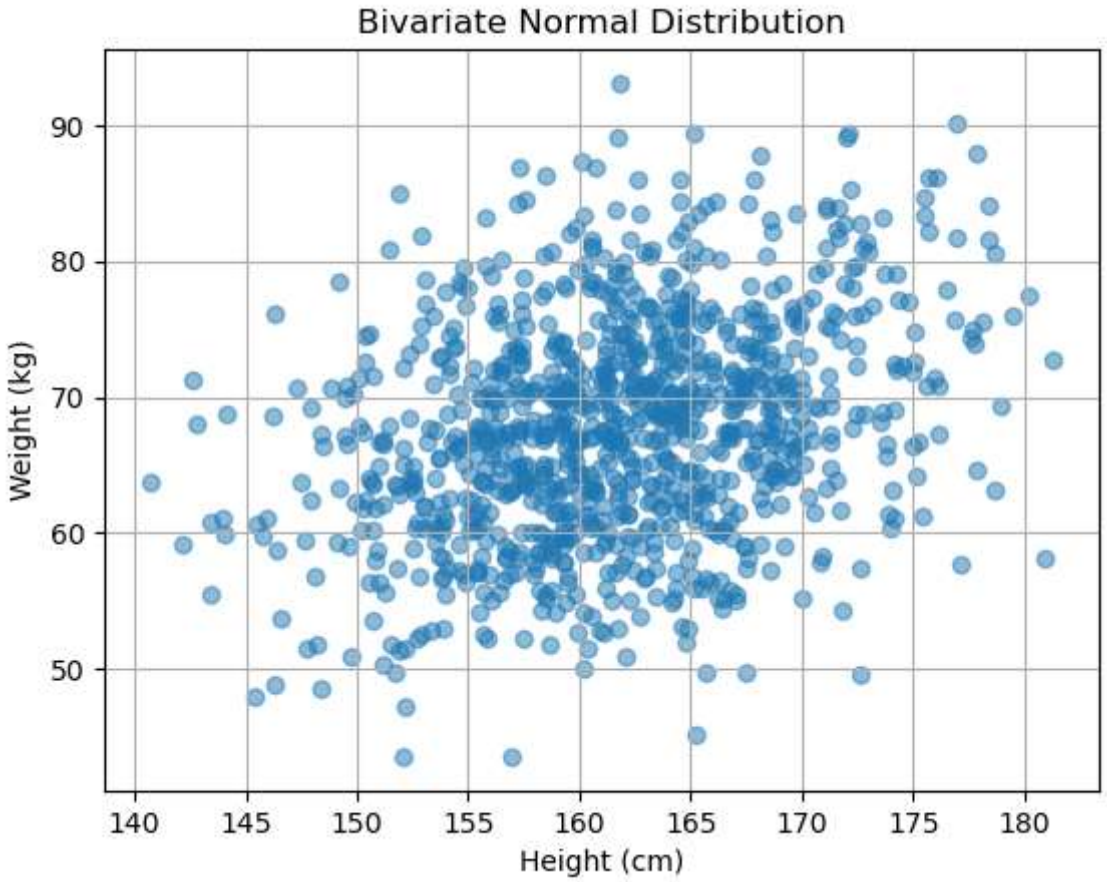
Problem # 2.27.

The distribution of X = heights *(cm)* of women in the U.K. is approximately $N(162, 7^2)$. Conditional on $X = x$, suppose Y = weight *(kg)* has a $N(3.0 + 0.40x, 8^2)$ distribution. Simulate and plot 1000 observations from this approximate bivariate normal distribution. Approximate the marginal means and standard deviations for X and Y. Approximate and interpret the correlation.

`Your answer goes here`

In [15]:
```
np.random.seed(0)

num_observations = 1000

mean_height = 162
std_dev_height = 7
heights = np.random.normal(mean_height, std_dev_height, num_observations)

mean_weight = 3.0 + 0.40 * heights
std_dev_weight = 8
weights = np.random.normal(mean_weight, std_dev_weight, num_observations)

plt.scatter(heights, weights, alpha=0.5)
plt.xlabel('Height (cm)')
plt.ylabel('Weight (kg)')
plt.title('Bivariate Normal Distribution')
plt.grid(True)
plt.show()

mean_height_approx = np.mean(heights)
```

```
std_dev_height_approx = np.std(heights)
mean_weight_approx = np.mean(weights)
std_dev_weight_approx = np.std(weights)

correlation = np.corrcoef(heights, weights)[0, 1]

print(f"Approx Marginal Mean for X: {mean_height_approx:.2f} cm")
print(f"Approx Marginal Standard Deviation for X: {std_dev_height_approx:.2f} cm")
print(f"Approx Marginal Mean for Y: {mean_weight_approx:.2f} kg")
print(f"Approx Marginal Standard Deviation for Y: {std_dev_weight_approx:.2f} kg")
print(f"Approx Correlation: {correlation:.2f}")
```




```
Approximate Marginal Mean for X (Height): 161.68 cm
Approximate Marginal Standard Deviation for X (Height): 6.91 cm
Approximate Marginal Mean for Y (Weight): 67.78 kg
Approximate Marginal Standard Deviation for Y (Weight): 8.14 kg
Approximate Correlation: 0.31
```