Applied Statistics Computational Project 2

For the dataset, We have taken the Video game sales Data.

Conclusions from Gamma Distribution Fitting of Global_Sales Data

```
import pandas as pd
import numpy as np
from scipy.stats import gamma

# Load the Excel file
df = pd.read_excel("newgame.xlsx")

# Extract Global_Sales and filter positive values
global_sales = df['Global_Sales']
global_sales = global_sales[global_sales > 0].dropna()

# Method of Moments Estimation
sample_mean = global_sales.mean()
sample_var = global_sales.var()

# MoM formulas for Gamma(a, b)
b_mom = sample_var / sample_mean
a_mom = sample_mean / b_mom

print("Method of Moments Estimates:")
print(f"a (shape) = {a_mom:.4f}")
print(f"b (scale) = {b_mom:.4f}")

# Maximum Likelihood Estimation (fit Gamma with loc=0)
a_mle, loc_mle, b_mle = gamma.fit(global_sales, floc=0)

print("\nMaximum Likelihood Estimates:")
print(f"a (shape) = {a_mle:.4f}")
print(f"b (scale) = {b_mle:.4f}")
```

The objective of the analysis is to model the distribution of `Global_Sales` using the Gamma distribution. Two estimation methods were used:

- Method of Moments (MoM)
- Maximum Likelihood Estimation (MLE)

Estimated Parameters

Method	**Shape** (a)	**Scale** (b)
Method of Moments	2.1766	6.7504
Maximum Likelihood Estimation	3.9330	3.7359

Interpretations and Conclusions

1. Both estimation methods suggest that the Gamma distribution is a reasonable model for the `Global_Sales` data. This is justified since the sales data is non-negative and skewed, which aligns with the properties of the Gamma distribution.

2. The two methods yield different parameter values:

   - The MoM estimates indicate a lower shape parameter and a higher scale parameter, suggesting a more right-skewed distribution.
   - The MLE estimates yield a higher shape parameter and a lower scale parameter, suggesting less skewness and more concentration near the mean.

3. MLE is generally more accurate and statistically efficient than MoM, especially with larger datasets. While MoM provides a quick and intuitive estimation based on sample moments, MLE makes use of the full likelihood function.

4. For robust statistical inference or simulation purposes, the MLE-based Gamma model is preferable.

Confidence Interval for the Variance of `Global_Sales`

```python
import pandas as pd
from scipy.stats import chi2

# Sample statistics
n = len(global_sales)
sample_var = global_sales.var(ddof=1)  # unbiased estimator

# Confidence level
alpha = 0.05

# Chi-squared critical values (corrected)
chi2_lower = chi2.ppf(alpha / 2, df=n - 1)
chi2_upper = chi2.ppf(1 - alpha / 2, df=n - 1)

# Confidence interval for the variance
lower_bound = (n - 1) * sample_var / chi2_upper
upper_bound = (n - 1) * sample_var / chi2_lower

print(f"95% Confidence Interval for the Variance: ({lower_bound:.4f
    }, {upper_bound:.4f})")
```

Using the sample of `Global_Sales` (filtered to include only positive values), we compute a 95% confidence interval for the population variance based on the Chi-squared distribution.

Method

Let s^2 denote the sample variance computed from n independent observations. The confidence interval for the true variance σ^2 of a normal population is given by:

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \right)$$

where:

- n is the sample size,
- s^2 is the unbiased sample variance,
- $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ are the critical values from the Chi-squared distribution with $n-1$ degrees of freedom,
- $\alpha = 0.05$ corresponds to a 95% confidence level.

Results

For the given data:

$$\text{95\% Confidence Interval for the Variance: } (76.3680,\ 134.0711)$$

Interpretation

We are 95% confident that the true variance of global sales lies between approximately 76.37 and 134.07. This interval quantifies the uncertainty around the sample variance estimate due to sampling variability.

Confidence Interval for Difference in Mean Global Sales: `Sports` vs `Platform`

```python
import pandas as pd
import numpy as np
from scipy import stats

# Load Excel file
df = pd.read_excel('newgame.xlsx')


#  Confidence Interval for Difference in Means


# Choose two non-overlapping genres
genre1 = 'Sports'
genre2 = 'Platform'

# Filter data for each genre
sales1 = df[df['Genre'] == genre1]['Global_Sales'].dropna()
sales2 = df[df['Genre'] == genre2]['Global_Sales'].dropna()

# Calculate means, standard deviations, and sample sizes
n1, n2 = len(sales1), len(sales2)
mean1, std1 = sales1.mean(), sales1.std(ddof=1)
mean2, std2 = sales2.mean(), sales2.std(ddof=1)

# Compute standard error and margin of error
se_diff = np.sqrt(std1**2 / n1 + std2**2 / n2)
z_critical = stats.norm.ppf(0.975)  # 95% CI
margin_of_error = z_critical * se_diff

# Confidence Interval
ci_lower = (mean1 - mean2) - margin_of_error
ci_upper = (mean1 - mean2) + margin_of_error

# Output for Part 1
print("=== 95% Confidence Interval for Difference in Means ===")
print(f"Genre 1: {genre1} | Mean: {mean1:.3f}")
print(f"Genre 2: {genre2} | Mean: {mean2:.3f}")
print(f"95% CI for (Mean of {genre1} - Mean of {genre2}): ({
    ci_lower:.3f}, {ci_upper:.3f})\n")
```

Objective

To determine whether there is a statistically significant difference in the average global sales between the `Sports` and `Platform` (Non - Overlapping populations) genres, we compute a 95% confidence interval for the difference in their means.

Method

Let:

- $\bar{x}_1$, s_1, and n_1 be the sample mean, standard deviation, and size for `Sports`,
- $\bar{x}_2$, s_2, and n_2 be the same for `Platform`.

The confidence interval for the difference in means $(\mu_1 - \mu_2)$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

with $z_{\alpha/2} = 1.96$ for a 95% confidence level.

Results

- Mean Global Sales for `Sports`: $\bar{x}_1 = 29.502$
- Mean Global Sales for `Platform`: $\bar{x}_2 = 15.745$
- 95% Confidence Interval for $(\mu_{\text{Sports}} - \mu_{\text{Platform}})$:

$$(-8.942,\ 36.456)$$

Interpretation

The confidence interval includes 0, indicating that the difference in mean global sales between `Sports` and `Platform` genres is not statistically significant at the 95% confidence level. That is, we do not have sufficient evidence to conclude that one genre significantly outsells the other on average.

Hypothesis Test for Proportion of High-Selling Games

```

# Hypothesis Test for Binary Response (Bernoulli)


# Define binary response variable: success if Global_Sales > 10 million
df['Success'] = (df['Global_Sales'] > 10).astype(int)

# Sample proportion and size
p_hat = df['Success'].mean()
n = len(df['Success'])
p0 = 0.5  # Null hypothesis value

# Z-test statistic and p-value
z_stat = (p_hat - p0) / np.sqrt(p0 * (1 - p0) / n)
p_value = 1 - stats.norm.cdf(z_stat)

# Output for Part 2
print("=== Hypothesis Test for Proportion (Bernoulli) ===")
print("H0: p <= 0.5 vs H1: p > 0.5")
print(f"Sample proportion (p  ): {p_hat:.3f}")
print(f"Z-statistic: {z_stat:.3f}")
print(f"P-value: {p_value:.4f}")
if p_value < 0.05:
    print("Conclusion: Reject H0 at 5% significance level. Evidence suggests p > 0.5.")
else:
    print("Conclusion: Fail to reject H0. No strong evidence that p > 0.5.")
```

Objective

To test whether the proportion of games with global sales greater than 10 million units exceeds 0.5.

Definitions

- Define a binary variable Success such that:

$$\text{Success} = \begin{cases} 1, & \text{if Global\_Sales} > 10 million \\ 0, & \text{otherwise} \end{cases}$$

- Let $\hat{p}$ denote the sample proportion of successful games.
- Let n be the total number of observations.

Hypotheses

$$H_0 : p \leq 0.5 \quad \text{(Proportion of successful games is 50\% or less)}$$
$$H_1 : p > 0.5 \quad \text{(Proportion of successful games exceeds 50\%)}$$

Test Statistic

The test statistic is computed using the formula for a one-sample proportion Z-test:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where:

- $\hat{p} = 0.626$ (sample proportion),
- $p_0 = 0.5$ (null hypothesis proportion),
- n is the total sample size.

Results

- Z-statistic: $Z = 2.513$
- P-value: 0.0060

Conclusion

Since the p-value is less than the significance level $\alpha = 0.05$, we reject the null hypothesis H_0.

Conclusion: There is statistically significant evidence at the 5% level to suggest that the proportion of high-selling games (over 10 million units) exceeds 50%.

Done By :

- **Puli Dinesh - AI23BTECH11019**
- **Rathod sai dhanush - AI23BTECH11021**
- **Jatavath ajay- AI23BTECH11011**