Statistics Assignment Solutions

Assignment No. 1

1 Question 1: Z-Phone Smartphone Life Distribution

Problem: A manufacturer says the Z-Phone smartphone has a mean consumer life of 42 months with a standard deviation of 8 months. Assuming a normal distribution, what is the probability that a given random Z-Phone will last between 20 and 30 months?

Solution:

Given:

$$\mu = 42 \text{ months} \tag{1}$$
$$\sigma = 8 \text{ months} \tag{2}$$
$$X \sim N(42, 8^2) \tag{3}$$

We need to find $P(20 \leq X \leq 30)$.
First, standardize using $Z = \frac{X-\mu}{\sigma}$:

$$Z_1 = \frac{20-42}{8} = \frac{-22}{8} = -2.75 \tag{4}$$
$$Z_2 = \frac{30-42}{8} = \frac{-12}{8} = -1.5 \tag{5}$$

$$P(20 \leq X \leq 30) = P(-2.75 \leq Z \leq -1.5) = \Phi(-1.5) - \Phi(-2.75) \tag{6}$$

Using standard normal table:

$$\Phi(-1.5) = 0.0668 \tag{7}$$
$$\Phi(-2.75) = 0.0030 \tag{8}$$

Answer: $P(20 \leq X \leq 30) = 0.0668 - 0.0030 = 0.0638$ or 6.38%

2 Question 2: Electronic Component Survival Time

Problem: Eight components were tested with failure times: 75, 63, 100+, 36, 51, 45, 80, 90. The observation 100+ indicates that the unit still functioned at 100 hours. Is there any meaningful measure of location that can be calculated for these data?

Solution:

Since we have censored data (100+ means the component lasted at least 100 hours but we don't know the exact failure time), we cannot calculate the mean accurately.

However, we can calculate the **median** as a meaningful measure of location.

Ordered data: 36, 45, 51, 63, 75, 80, 90, 100+

Since we have 8 observations, the median is the average of the 4th and 5th values:

$$\text{4th value} = 63 \tag{9}$$
$$\text{5th value} = 75 \tag{10}$$

Answer: The median can be calculated and equals $\frac{63+75}{2} = 69$ hours.

This is meaningful because even with censored data, we know that at least 50% of components failed by 69 hours.

3 Question 3: Age vs Weight Linear Regression

Problem: Based on a dataset of 250 samples, calculate least squares estimates and make predictions for the relationship between age (x) and weight (y).

Note: The summary statistics appear to be missing from the provided document. I'll demonstrate the solution method assuming typical values.

Assumed Summary Statistics:

$$n = 250 \tag{11}$$
$$\sum x = 10,000 \text{ (mean age } \approx 40) \tag{12}$$
$$\sum y = 42,500 \text{ (mean weight } \approx 170 \text{ lbs)} \tag{13}$$
$$\sum x^2 = 425,000 \tag{14}$$
$$\sum y^2 = 7,500,000 \tag{15}$$
$$\sum xy = 1,750,000 \tag{16}$$

Solution:

Part (a): Least Squares Estimates

First, calculate means:

$$\bar{x} = \frac{\sum x}{n} = \frac{10,000}{250} = 40 \tag{17}$$
$$\bar{y} = \frac{\sum y}{n} = \frac{42,500}{250} = 170 \tag{18}$$

Calculate slope (β_1):

$$\beta_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \tag{19}$$
$$= \frac{1,750,000 - 250(40)(170)}{425,000 - 250(40)^2} \tag{20}$$
$$= \frac{1,750,000 - 1,700,000}{425,000 - 400,000} \tag{21}$$
$$= \frac{50,000}{25,000} = 2 \tag{22}$$

Calculate intercept (β_0):

$$\beta_0 = \bar{y} - \beta_1\bar{x} = 170 - 2(40) = 90 \tag{23}$$

Fitted equation: $\hat{y} = 90 + 2x$
Part (b): Prediction for 25-year-old

$$\hat{y} = 90 + 2(25) = 90 + 50 = 140 \text{ lbs} \tag{24}$$

Part (c): Residual calculation
Given: actual weight = 170 lbs, predicted weight = 140 lbs

$$\text{Residual} = \text{actual} - \text{predicted} = 170 - 140 = 30 \text{ lbs} \tag{25}$$

Part (d): Over/underestimate
The prediction was an **underestimate** because the residual is positive (actual > predicted).

4 Question 4: Cold Start Ignition Time Analysis

Problem: Analyze cold start ignition times for two gasoline formulations.
First Formulation Data: 1.75, 1.92, 2.62, 2.35, 3.09, 3.15, 2.53, 1.91
Solution:
Sample Statistics:

$$n = 8 \tag{26}$$
$$\bar{x} = \frac{1.75 + 1.92 + 2.62 + 2.35 + 3.09 + 3.15 + 2.53 + 1.91}{8} = \frac{18.32}{8} = 2.29 \text{ seconds} \tag{27}$$

Sample variance (s^2):

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{28}$$
$$\sum(x_i - \bar{x})^2 = (1.75 - 2.29)^2 + (1.92 - 2.29)^2 + \ldots + (1.91 - 2.29)^2 = 2.087 \tag{29}$$
$$s^2 = \frac{2.087}{7} = 0.298 \tag{30}$$

Sample standard deviation:

$$s = \sqrt{0.298} = 0.546 \text{ seconds} \tag{31}$$

Box Plot Summary (First Formulation):

$$\text{Minimum} = 1.75 \tag{32}$$
$$Q_1 = 1.915 \tag{33}$$
$$\text{Median} = 2.435 \tag{34}$$
$$Q_3 = 2.875 \tag{35}$$
$$\text{Maximum} = 3.15 \tag{36}$$

Second Formulation Data: 1.83, 1.99, 3.13, 3.29, 2.65, 2.87, 3.40, 2.46, 1.89, 3.35
Sample Statistics for Second Formulation:

$$n = 10 \tag{37}$$
$$\bar{x} = \frac{25.86}{10} = 2.586 \text{ seconds} \tag{38}$$
$$s \approx 0.662 \text{ seconds} \tag{39}$$

Box Plot Summary (Second Formulation):

$$\text{Minimum} = 1.83 \tag{40}$$
$$Q_1 = 1.97 \tag{41}$$
$$\text{Median} = 2.76 \tag{42}$$
$$Q_3 = 3.21 \tag{43}$$
$$\text{Maximum} = 3.40 \tag{44}$$

Interpretation:

1. The second formulation has a higher median ignition time (2.76 vs 2.435 seconds)
2. The second formulation shows greater variability (larger IQR and standard deviation)
3. Both formulations have similar minimum values, but the second has a higher maximum
4. The second formulation appears to have slightly worse performance with longer ignition times

5 Question 5: Linear Regression with Python

Problem: Generate synthetic data, apply linear regression, and compare polynomial degrees.

Python Code Solution:

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split

# Generate synthetic data
np.random.seed(42)
n_samples = 100
X = np.linspace(0, 10, n_samples).reshape(-1, 1)
true_y = 2 * X.ravel() + 3 + np.random.normal(0, 1, n_samples)

# Split data (80:20)
X_train, X_test, y_train, y_test = train_test_split(
    X, true_y, test_size=0.2, random_state=42)

# Function to fit and evaluate polynomial regression
def fit_polynomial(degree):
    poly_features = PolynomialFeatures(degree=degree)
    X_train_poly = poly_features.fit_transform(X_train)
    X_test_poly = poly_features.transform(X_test)

    model = LinearRegression()
    model.fit(X_train_poly, y_train)

    y_pred = model.predict(X_test_poly)
    mse = mean_squared_error(y_test, y_pred)

    return model, poly_features, mse

# Fit models for degrees 1, 2, and 3
results = {}
for degree in [1, 2, 3]:
    model, poly_features, mse = fit_polynomial(degree)
    results[degree] = {
        'model': model,
        'poly_features': poly_features,
        'mse': mse
    }
    print(f"Degree {degree} - MSE: {mse:.4f}")

# Plotting code
plt.figure(figsize=(15, 5))
X_plot = np.linspace(0, 10, 100).reshape(-1, 1)

for i, degree in enumerate([1, 2, 3], 1):
    plt.subplot(1, 3, i)

    # Plot training data
    plt.scatter(X_train, y_train, alpha=0.6, label='Training Data')
    plt.scatter(X_test, y_test, alpha=0.6,
                label='Testing Data', color='red')
```

```

    # Plot fitted line
    X_plot_poly = results[degree]['poly_features'].transform(X_plot)
    y_plot = results[degree]['model'].predict(X_plot_poly)
    plt.plot(X_plot, y_plot, 'g-', label=f'Degree {degree} Fit')

    plt.xlabel('X')
    plt.ylabel('Y')
    plt.title(f'Polynomial Degree {degree}\nMSE: {results[degree]["mse"]:.4f}')
    plt.legend()
    plt.grid(True, alpha=0.3)

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

Expected Results:

- **Degree 1 (Linear):** Should have moderate MSE, good fit for the underlying linear relationship
- **Degree 2 (Quadratic):** May have slightly lower MSE but risk of overfitting
- **Degree 3 (Cubic):** Likely to have the lowest training error but may overfit to noise

Key Insights:

1. Linear regression (degree 1) is often the best choice for truly linear relationships
2. Higher-degree polynomials can overfit, especially with limited data
3. Compare both training and testing MSE to assess generalization
4. The model with the best test MSE is typically the best choice

6 Summary

This assignment covers fundamental statistical concepts including:

- Normal distribution probability calculations
- Handling censored data
- Linear regression analysis
- Descriptive statistics and box plots
- Polynomial regression and model comparison

Each solution demonstrates both theoretical understanding and practical application of statistical methods.