HOMEWORK 2 REPORT

Submitted by Group 13

Members: Aman Sanwal, Arihant Sathpathy, Lipika Bagai, Ankita Singh and Gowthamy S

Question 1

Introduction

This section presents statistical analysis using data obtained from Kaggle. The dataset is normalized and subjected to goodness-of-fit tests, including the Kolmogorov-Smirnov (KS) test and the Cramér-von Mises (CVM) test.

Data Normalization

The dataset is sourced from Kaggle and preprocessed to ensure consistency. The data is normalized using mean and variance adjustments, which helps in standardizing the data for better statistical interpretation and comparison with theoretical distributions.

Normalization is performed as follows:

$$X' = \frac{X - \mu}{\sigma}, \tag{1}$$

where X' is the normalized data, X is the original data, μ is the mean, and σ is the standard deviation.

Kolmogorov-Smirnov Test (Discrete Version)

The Kolmogorov-Smirnov (KS) test is used to compare the empirical cumulative distribution function (ECDF) of the dataset with the cumulative distribution function (CDF) of a theoretical normal distribution. Since the dataset consists of discrete values, the discrete version of the KS test is applied, accounting for ties and stepwise changes in the CDF.

The KS test statistic is given by:

$$D_n = \sqrt{n} \sup_x |F_n(x) - F(x)|, \tag{2}$$

where $F_n(x)$ is the empirical CDF, $F(x)$ is the theoretical CDF, and n is the sample size.

Bootstrap Resampling and P-Value Estimation

Bootstrap resampling is employed to estimate the p-value of the Kolmogorov-Smirnov (KS) test statistic. This technique involves generating multiple resampled datasets (with replacement) and computing test statistics for each sample. By comparing these statistics with the observed KS statistic, we can estimate the probability of obtaining a similar or more extreme result under the null hypothesis.

The bootstrap estimate for the p-value is given by:

$$p = \frac{1}{B} \sum_{i=1}^{B} I(T_i \geq T_{obs}), \tag{3}$$

where B is the number of bootstrap samples, T_i is the test statistic for sample i, and T_{obs} is the observed test statistic. Since the test statistic is right-tailed, we apply the condition. This approach aligns with the logic that larger values of the KS statistic provide evidence against the null hypothesis. For the null hypothesis to be true, we expect the p-value to be higher, indicating that the observed test statistic is not unusually large compared to the bootstrap distribution.

Another justification for using the bootstrap p-value estimation is that it empirically approximates the sampling distribution of the test statistic under the null hypothesis. Since the exact distribution of the KS test statistic for discrete data may not be straightforward to derive, bootstrapping enables us to construct an empirical distribution by resampling from the observed data. This approach provides a robust and data-driven method to estimate the probability of obtaining an observed test statistic as extreme as the one computed, ensuring a reliable approximation of the true p-value.

Cramer-von Mises Test

The Cramér-von Mises (CVM) test is conducted to assess the goodness-of-fit of the dataset to the normal distribution by evaluating the squared differences between the empirical cumulative distribution function (ECDF) and the theoretical cumulative distribution function (CDF).

The CvM test statistic is computed as:

$$T_n = n \sum_{i=1}^{n} \left(F_n(X_i) - F(X_i) \right)^2 . \tag{4}$$

Conclusion

The Kolmogorov-Smirnov (Discrete Version) and Cramér-von Mises tests were performed to assess the normality of the dataset. The results indicate that the dataset exhibits statistical properties closely aligned with the expected theoretical distribution. This analysis confirms the effectiveness of normalization techniques in improving statistical validity.

Question 2

Introduction

This section involves analyzing weight data from the "SOCR-HeightWeight.csv" dataset. The dataset undergoes normalization using mean-variance, median-MAD, and mode-based approaches, followed by statistical tests.

Data Normalization Techniques

Three normalization techniques are applied:

- **Mean-Variance Normalization:**

$$X' = \frac{X - \mu}{\sigma} \tag{5}$$

  This transformation ensures the dataset has a mean of zero and a standard deviation of one, allowing better comparability across different distributions.

- **Median-MAD Normalization:**

$$X' = \frac{X - \text{median}(X)}{\text{MAD}(X)} \tag{6}$$

  This approach is more robust to outliers and is particularly useful when the data contains extreme values that could skew the mean.

- **Mode-Based Normalization:**

$$X' = \frac{X - \text{mode}(X)}{\text{MoAD}(X)} \tag{7}$$

  This method is particularly useful for skewed distributions where the mode better represents central tendency than the mean.

Kernel Density Estimation

Kernel Density Estimation (KDE) is applied using a uniform kernel to estimate the mode of the dataset. KDE is a non-parametric method for estimating the probability density function of a dataset, offering a smoother representation of the distribution than histograms.

The kernel density estimator is given by:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{8}$$

where K is the kernel function and h is the bandwidth parameter.

Conclusion

The analysis of the weight dataset demonstrated the effectiveness of different normalization techniques in improving statistical consistency. Kernel Density Estimation successfully identified the mode of the dataset, highlighting key statistical characteristics. Future work could explore alternative normalization methods and more advanced goodness-of-fit tests to enhance data analysis.