University of Texas Arlington

Probability and Statistics

Project-2

Rohan Kurma

1002192255

A report submitted in partial fulfillment
of the requirements for the degree
of *MSDS*

UNIVERSITY OF
TEXAS
ARLINGTON

"I, Rohan Kurma, did not give or receive any assistance on this project, and the report submitted is wholly my own."

Date: 11/26/2024

1 Introduction

The Chi-Square Goodness-of-Fit Test is a statistical method used to determine whether a sample data set fits a specific theoretical probability distribution [Pearson, 1900]. This non-parametric test compares the observed frequencies in the data to the expected frequencies derived from the hypothesized distribution.

The null hypothesis (H_0) assumes that the data follow the hypothesized distribution. The alternative hypothesis (H_a) suggests that the data do not follow the distribution. The test statistic is computed as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{1}$$

where:

- O_i: Observed frequency in the i^{th} class.
- E_i: Expected frequency in the i^{th} class.
- k: Number of classes.

The χ^2 statistic is compared to a critical value from the Chi-Square distribution table based on the degrees of freedom ($df = k - 1 - p$), where p is the number of parameters estimated from the data. If the calculated χ^2 exceeds the critical value, the null hypothesis is rejected [Freund and Perles, 2010].

Why Use the Chi-Square Test to Assess Distributions?

The Chi-Square Goodness-of-Fit Test is ideal for testing distributions because:

- It is distribution-free and does not assume the sample data follow a specific distribution.
- It can be applied to any type of theoretical distribution (e.g., Normal, Exponential).
- It is straightforward to implement by binning data into frequency classes and comparing observed counts to expected counts.
- It quantifies deviations between observed and expected frequencies into a single test statistic [Cochran, 1952].

Calculating Class Probabilities and Expected Frequencies

To apply the Chi-Square test, the following steps are crucial [Mood et al., 1974]:

1. Class Probabilities: For each bin, calculate the probability of the data falling within that bin using the cumulative distribution function (CDF) of the hypothesized distribution. For example:

   - For a Normal Distribution:

     $$P_i = F(x_{\text{upper}}; \mu, \sigma) - F(x_{\text{lower}}; \mu, \sigma),$$

     where $F(x; \mu, \sigma)$ is the CDF of the Normal Distribution.

   - For an Exponential Distribution:

     $$P_i = F(x_{\text{upper}}; \lambda) - F(x_{\text{lower}}; \lambda),$$

     where $F(x; \lambda)$ is the CDF of the Exponential Distribution.

2. Expected Frequencies: Multiply the class probability (P_i) by the total number of observations (n):

   $$E_i = n \cdot P_i.$$

3. Degrees of Freedom: The degrees of freedom for the Chi-Square test are given by:

   $$df = k - 1 - p,$$

   where k is the number of classes and p is the number of parameters estimated (e.g., mean and standard deviation).

—

2 Dataset and Problem Description

2.1 Dataset: NYC Flights Data

The dataset used in this study is the `nycflights13` dataset, which contains detailed information about all flights departing from New York City airports (JFK, LGA, and EWR) in 2013. This dataset includes variables such as departure and arrival times, flight delays, flight

durations, carriers, and destinations. The dataset has 336,776 observations and 19 variables, making it an excellent source for statistical analysis of air traffic patterns.

The dataset used for this analysis is the `nycflights13` dataset is acquired from [Wickham, 2014],

2.2 Set 1: Arrival Delays (`arr_delay`)

For Set 1, we focus on the variable `arr_delay`, which records the arrival delay of each flight in minutes. Positive values indicate delays, while negative values indicate flights arriving earlier than scheduled. The goal is to test whether the distribution of `arr_delay` follows a Normal Distribution using the Chi-Square Goodness-of-Fit Test. This analysis is important for understanding if arrival delays conform to a predictable pattern.

Preprocessing Steps for Set 1

- Filtered out missing or extreme values to ensure data quality.
- Computed the sample mean (μ) and standard deviation (σ) for `arr_delay`.
- Binned the data into intervals for calculating observed and expected frequencies.

2.3 Set 2: Inter-Arrival Times of Flights

Set 2 focuses on the inter-arrival times of flights, which represent the time difference (in minutes) between the departure times of consecutive flights. This variable was derived from the `dep_time` column of the dataset. The goal is to test whether the inter-arrival times follow an Exponential Distribution, which is commonly used to model time between random events in a Poisson process.

Preprocessing Steps for Set 2

- Sorted the dataset by `dep_time` to ensure consecutive flights are considered.
- Calculated inter-arrival times by taking the difference between successive departure times.
- Removed negative or missing values resulting from data anomalies.
- Computed the sample mean and rate parameter ($\lambda = 1/\text{mean}$).
- Binned the data into intervals for calculating observed and expected frequencies.

These two datasets form the basis of our hypothesis tests, where we analyze whether the observed distributions of Set 1 and Set 2 align with their respective theoretical distributions.

All the pre processing codes used in this project are available in [Kurma, 2024]

–

3 Set 1: Hypothesis Testing for Normal Distribution

Hypotheses

For `arr_delay`, the hypotheses are:

- Null Hypothesis (H_0): The arrival delays follow a Normal Distribution with:

$$\mu = \text{mean of } \texttt{arr\_delay}, \quad \sigma = \text{standard deviation of } \texttt{arr\_delay}.$$

- Alternative Hypothesis (H_a): The arrival delays do not follow the specified Normal Distribution.

Methodology

The observed data (O_i) are binned into classes, and the expected frequencies (E_i) are computed using the Normal Distribution's cumulative distribution function (CDF) for each class. The χ^2 statistic is calculated and compared to the critical value with degrees of freedom:

$$df = \text{Number of bins} - 1 - \text{Number of parameters estimated (2)}.$$

Results

Table: Observed and Expected Frequencies for Set 1 (Normal Distribution)

Table 1: Chi-Square Test Results for Set 1 (Normal Distribution)

Class	Observed Frequency	Class Probability	Expected Frequency	Chi^2 Class Component
$X \leq -20$	64,916	0.2696	88,224.15	6,157.83
$-20 < X \leq 0$	129,426	0.1687	55,204.44	99,789.79
$0 < X \leq 20$	66,146	0.1809	59,211.47	812.13
$20 < X \leq 60$	39,069	0.2697	88,285.21	27,436.48
$X > 60$	27,738	0.1111	36,369.73	2,048.59
Total	**327,295**	**1.0000**	**327,295.00**	**136,244.83**

Conclusion and Observations

Based on the Chi-Square Goodness-of-Fit Test conducted on the arrival delays (`arr_delay`) in Set 1, the following observations and conclusions can be drawn:

- **Chi-Square Statistic:** The total Chi-Square statistic is $\chi^2 = 136,244.83$, which quantifies the deviation between the observed frequencies and the expected frequencies under the assumption of a Normal Distribution.
- **Degrees of Freedom:** The degrees of freedom (df) for this test are calculated as:

$$df = \text{Number of bins} - 1 - \text{Number of estimated parameters (2)} = 5 - 1 - 2 = 2.$$

- **Critical Value:** At a significance level of $\alpha = 0.05$, the critical value for $df = 2$ is approximately $\chi^2_{\text{critical}} = 5.991$ (based on the Chi-Square distribution table).
- **Comparison:** The calculated Chi-Square statistic ($\chi^2 = 136,244.83$) far exceeds the critical value ($\chi^2_{\text{critical}} = 5.991$).
- **Interpretation:** Since the test statistic is significantly larger than the critical value, the null hypothesis (H_0) that the arrival delays follow a Normal Distribution is **rejected**.
- **Conclusion:** The observed distribution of arrival delays does not align with the theoretical Normal Distribution. The large deviations, particularly in the bins $-20 < X \leq 0$ and $20 < X \leq 60$, suggest that the data's distribution is heavily skewed or exhibits characteristics inconsistent with normality.

This conclusion highlights that arrival delays may have non-normal characteristics such as skewness or heavy tails, making them unsuitable for normality-based modeling in this context.

—

4 Set 2: Hypothesis Testing for Exponential Distribution

Hypotheses

For the inter-arrival times, the hypotheses are:

- Null Hypothesis (H_0): The inter-arrival times follow an Exponential Distribution with:

$$\lambda = \frac{1}{\text{mean of inter-arrival times}}.$$

- Alternative Hypothesis (H_a): The inter-arrival times do not follow the specified Exponential Distribution.

Methodology

The observed data (O_i) are binned into classes, and the expected frequencies (E_i) are computed using the Exponential Distribution's cumulative distribution function (CDF) for each class. The χ^2 statistic is calculated and compared to the critical value with degrees of freedom:

$$df = \text{Number of bins} - 1 - \text{Number of parameters estimated (1)}.$$

Results

Table: Observed and Expected Frequencies for Set 2 (Exponential Distribution)

Table 2: Chi-Square Test Results for Set 2 (Exponential Distribution)

Class	Observed Frequency	Class Probability	Expected Frequency	Chi^2 Class Component
$X \leq 1$	1,272	0.4222	555.64	923.56
$1 < X \leq 2$	10	0.2439	321.04	301.35
$2 < X \leq 5$	8	0.2694	354.58	338.76
$5 < X \leq 10$	1	0.0602	79.28	77.29
$X > 10$	25	0.0041	5.46	70.01
Total	**1,316**	**1.0000**	**1,316.00**	**1,710.98**

Conclusion and Observations

Based on the Chi-Square Goodness-of-Fit Test conducted on the inter-arrival times of flights in Set 2, the following observations and conclusions can be drawn:

- **Chi-Square Statistic:** The total Chi-Square statistic is $\chi^2 = 1,710.98$, which quantifies the deviation between the observed frequencies and the expected frequencies under the assumption of an Exponential Distribution.
- **Degrees of Freedom:** The degrees of freedom (df) for this test are calculated as:

$$df = \text{Number of bins} - 1 - \text{Number of estimated parameters } (1) = 5 - 1 - 1 = 3.$$

- **Critical Value:** At a significance level of $\alpha = 0.05$, the critical value for $df = 3$ is approximately $\chi^2_{\text{critical}} = 7.815$ (based on the Chi-Square distribution table).
- **Comparison:** The calculated Chi-Square statistic ($\chi^2 = 1,710.98$) far exceeds the critical value ($\chi^2_{\text{critical}} = 7.815$).
- **Interpretation:** Since the test statistic is significantly larger than the critical value, the null hypothesis (H_0) that the inter-arrival times follow an Exponential Distribution is **rejected**.
- **Conclusion:** The observed distribution of inter-arrival times does not align with the theoretical Exponential Distribution. Significant deviations are evident, particularly in the first bin ($X \leq 1$) and subsequent bins, indicating that the inter-arrival times exhibit characteristics inconsistent with a memoryless exponential process.

—

5 Summary and Overall Observations

In this study, we applied the Chi-Square Goodness-of-Fit Test to assess the conformity of observed data with theoretical distributions. The results for Set 1 and Set 2 highlight:

- Whether arrival delays in NYC flights follow a Normal Distribution.
- Whether inter-arrival times between flights follow an Exponential Distribution.

These findings provide insights into the distributional characteristics of delays and inter-arrival times in the aviation context.

References

William G Cochran. The chi-square test of goodness of fit. *The Annals of Mathematical Statistics*, 23(3):315–345, 1952.

John Freund and Benjamin Perles. *Modern Elementary Statistics*. Pearson, 2010.

Rohan Kurma. Source code for chi-square goodness-of-fit test, 2024. URL `https://github.com/RohanKurma/Probability`$_{S}$*tatistics*$_{P}$*roject*2*.git.Accessed* : 2024−11 − 26.

Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, 1974.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175, 1900.

Hadley Wickham. *nycflights13: Airline On-Time Data for All Flights Departing NYC in 2013*, 2014. URL `https://CRAN.R-project.org/package=nycflights13`. R package version 1.0.2.

Appendix: R Formulas and Methods Used

Set 1: Arrival Delays (Normal Distribution)

The following steps and R formulas were used to perform the Chi-Square Goodness-of-Fit Test for Set 1:

1. **Define Bins for Observed Frequencies:**

```
breaks_arr <- c(-Inf, -20, 0, 20, 60, Inf)
observed_arr <- hist(flights_clean$arr_delay, breaks = breaks_
    arr, plot = FALSE)$counts
```

Listing 1: Define bins for arrival delays

2. **Calculate Class Probabilities:**

```
expected_probs_arr <- pnorm(breaks_arr[-1], mean = mean_arr,
    sd = sd_arr) -
                      pnorm(breaks_arr[-length(breaks_arr)],
                         mean = mean_arr, sd = sd_arr)
```

Listing 2: Class probabilities using Normal CDF

3. **Calculate Expected Frequencies:**

```
expected_arr <- sum(observed_arr) * expected_probs_arr
```

Listing 3: Expected frequencies based on probabilities

4. **Compute Chi-Square Components:**

```
chisq_components_arr <- (observed_arr - expected_arr)^2 /
    expected_arr
chisq_stat_arr <- sum(chisq_components_arr)
```

Listing 4: Chi-Square components for each bin

—

Set 2: Inter-Arrival Times (Exponential Distribution)

The following steps and R formulas were used to perform the Chi-Square Goodness-of-Fit Test for Set 2:

1. **Compute Inter-Arrival Times:**

```
flights_clean <- flights_clean %>%
                 arrange(dep_time) %>%
                 mutate(inter_arrival = c(NA, diff(dep_time)))
```

Listing 5: Compute inter-arrival times from departure times

2. **Define Bins for Observed Frequencies:**

```
breaks_inter <- c(-Inf, 1, 2, 5, 10, Inf)
observed_inter <- hist(flights_clean$inter_arrival, breaks =
    breaks_inter, plot = FALSE)$counts
```

Listing 6: Define bins for inter-arrival times

3. **Calculate Class Probabilities:**

```
expected_probs_inter <- pexp(breaks_inter[-1], rate = rate_
    inter) -
                        pexp(breaks_inter[-length(breaks_inter
                            )], rate = rate_inter)
```

Listing 7: Class probabilities using Exponential CDF

4. **Calculate Expected Frequencies:**

```
expected_inter <- sum(observed_inter) * expected_probs_inter
```

Listing 8: Expected frequencies based on probabilities

5. **Compute Chi-Square Components:**

```
chisq_components_inter <- (observed_inter - expected_inter)^2
    / expected_inter
chisq_stat_inter <- sum(chisq_components_inter)
```

Listing 9: Chi-Square components for each bin

—

General Observations

- The formulas used for class probabilities and expected frequencies depend on the assumed theoretical distribution (Normal or Exponential).
- The Chi-Square statistic was calculated as the sum of all class components:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed frequency, E_i is the expected frequency, and k is the number of bins.