Statistical analysis of Yelp dataset

Word count: 970

1 Introduction

This report will summarise a statistical analysis conducted on the *Yelp* data-set focusing on two main research problems: fitting a theoretical distribution to the number of reviews in a metropolitan area (both for users and businesses), and comparing the distributions of star ratings of a business category in two different metropolitan areas.

2 The statistics of online reviews

The first section will focus on the metropolitan area of Pittsburgh, Pennsylvania (PA), containing 32133 users sending reviews and 4086 businesses receiving reviews. The following hypothesis will be tested:

$H_0 =$ The empirical data follows a Pareto (type I) distribution
$H_1 =$ The empirical data does not follow a Pareto (type I) distribution

2.1 Methodology

The data-set will be fitted with a Pareto (type I) distribution, a power-law distribution with the following theoretical PDF and CDF:

$$f(x) = \frac{\alpha {x_m}^\alpha}{x^{\alpha+1}} \quad \forall x \geq x_m$$

$$F(x) = 1 - \left(\frac{x_m}{x}\right)^\alpha \quad \forall x \geq x_m$$

where x_m is the minimum value present in the observations, α is the decay parameter and x is the vector of observations. The value of α will be retrieved from its Maximum-Likelihood estimate (MLE):

$$\alpha^* = \frac{n}{\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_m}\right)}$$

where n is the number of observations in vector x.
The empirical PDF for the data will be generated through a normalised histogram, where the bin heights will be plotted as points in a scatter-plot .The empirical CDF will be generated through a rank-frequency plot, computed as follows:

$$\hat{F}(x) = \frac{1}{n+1} \sum_{i=1}^{n} 1_{X_i \leq x}$$

The validity of the estimate we obtain for α will be tested through a bootstrap analysis at a 95% confidence level, in which portions of the data (80% of our data-set for 1000 iterations) will be re-sampled with replacement, in order to obtain a range of valid α estimates with 95% certainty.

Lastly, in order to quantitatively analyse whether the null-hypothesis should be rejected, we will run a Kolmogorov-Smirnov (KS) test between the empirical and theoretical CDF. The KS test is a non-parametric test used on uni-variate continuous data to determine whether two data-sets come from the same distribution.

2.2 Results

Observing results from plots (1.1) and (1.2), both the user and business data-set exhibit power-law behaviour, where in the majority of cases we observe a low amount of reviews. Computing key moments suggest both data-sets have a high positive skewness (13.21 for users and 6.63 for businesses) and excess kurtosis (318.81 for users and 75.87 for businesses), showing non-Gaussianity. This justifies testing the null-hypothesis that the data-sets follow a Pareto distribution.

Plots (1.3) and (1.4) show results from fitting the data-sets with the theoretical distribution. It appears to be a better fit for the business data-set compared to the users one, possibly due to the lower x_m in users, which could distort its Maximum-Likelihood estimate for α. Analysing the effectiveness of the estimated parameters with bootstrap analysis, we obtain that at a 95% confidence level, α is in the range $[0.478149 - 0.4823289]$ for users and $[0.681538 - 0.699020]$ for businesses. The range is narrow, proving we found a good fairly good estimate for the parameter.

Plots (1.5) and (1.6) compare the theoretical and empirical CDFs. The shapes of the functions look similar, although some differences can be spotted, with the theoretical CDF growing faster than the empirical one in both figures. This can be further analysed with a 2-sample KS-test (between vector of the empirical CDF and the vector of theoretical CDF we computed), which returns a p-value of 0.0 for users and a p-value of 0.0 for businesses. This means we can clearly reject the null hypothesis of both data-sets following a Pareto distribution.

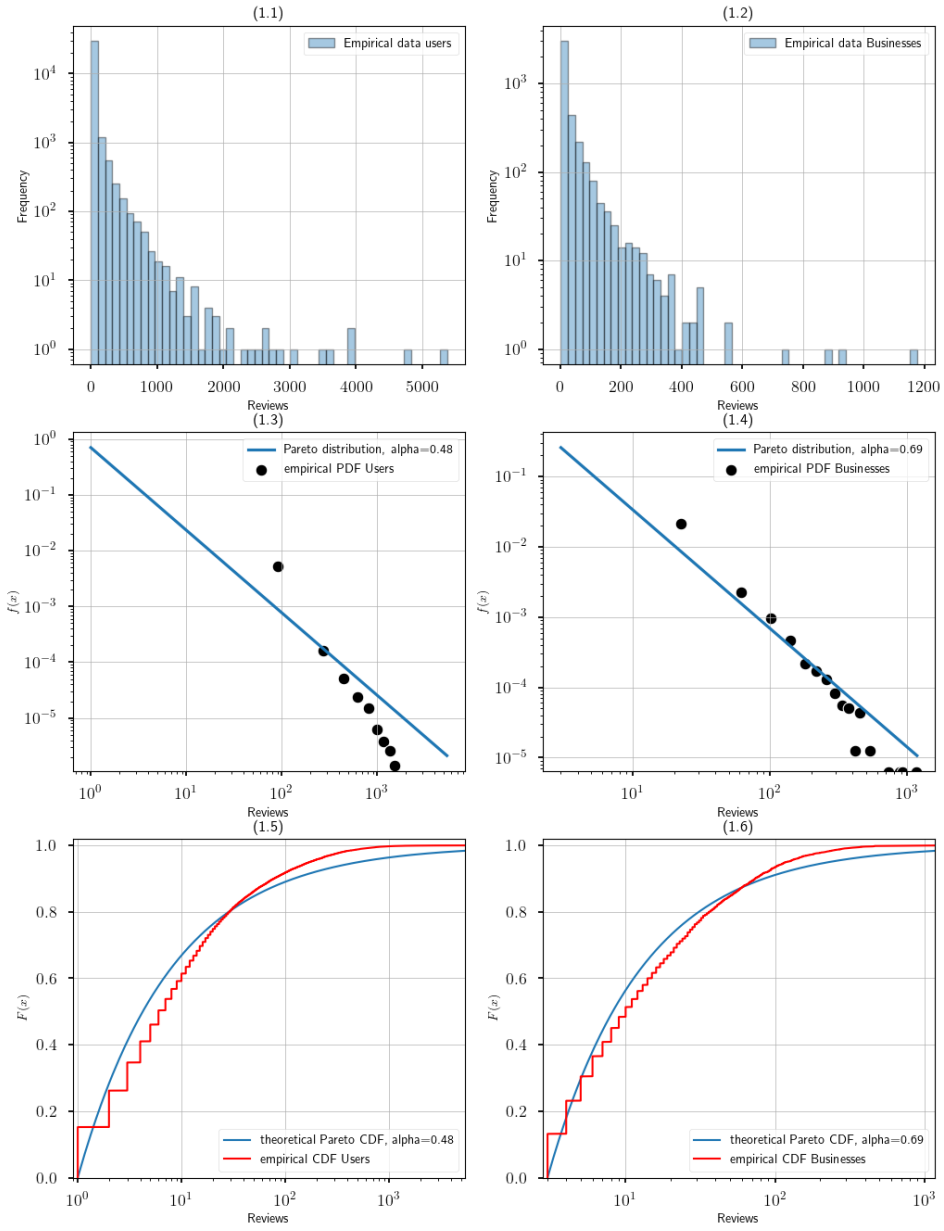

Figure 1: Histogram of user and business reviews (1.1) and (1.2), Comparison between empirical and theoretical Pareto PDF for users (1.3) and businesses (1.4), comparison between empirical and theoretical Pareto CDF for users (1.5) and businesses (1.6)

3 Are Yelp users consistent across different areas

The second section will research the similarities and differences between the distributions of star ratings of bars in two different metropolitan areas, Nevada (NV) and Arizona (AZ). The following hypothesis will be tested:

$H_0 =$ The two data-sets follow the same distribution
$H_1 =$ The two data-sets do not follow the same distribution

3.1 Methodology

The two data-sets will be analysed through descriptive statistics and the main moments will be compared. Similarly to the previous research question, empirical PDFs and CDFs will be computed (constructed through normalised histograms and rank-frequency plots). Comparison will be done through the use of a boxplot, which highlights the main quartiles of the distributions and key outliers, and a Q-Q plot, which plots the quantiles of a data-set against the quantiles of a second data-set. The closer the points lie on a 45 degree line, the more similarly distributed the two data-sets will be.

3.2 Results

Nevada (NV)	Arizona (AZ)
Descriptive Statistics: μ (mean) $= 3.62$ σ (standard deviation) $= 0.65$ ξ (skewness) $= -0.41$ κ (excess kurtosis) $= 0.25$	*Descriptive Statistics*: μ (mean) $= 3.55$ σ (standard deviation) $= 0.66$ ξ (skewness) $= -0.39$ κ (excess kurtosis) $= 0.19$
Results from Kolmogorov-Smirnov test	Statistic: 0.0499 p-value: 0.0300

Table 1: Table showing the four main moments computed for NV ratings and AZ ratings, as well as results from a 2-sample Kolmogorov-Smirnov (KS) test conducted on the two data-sets

Comparing the main moments of the two data-sets, we can observe a clear similarity, with both being slightly positively skewed and with a low amount of excess kurtosis. This can be highlighted in figure (2.1), where the PDFs of the two distributions are almost overlapping, and in figure (2.2), which shows almost identical box-plots.

The Q-Q plot in figure (2.3) shows most points lying on, or close to the 45 degree line, further supporting the idea that the two data-sets have similar distributions. The empirical CDF plotted in figure (2.4) also shows similarity, although some gaps are present between the two, which could potentially highlight divergence between the two samples.

However, the KS-test returns a p-value of 0.03, meaning we can reject the null hypothesis of the two data-sets following the same statistical distribution at a 95% confidence level. This outcome is different to what we visually observed from the four plots, but can be understandable due to the high sensitivity of the test to gaps in the CDFs (which we observed). It could also be argued that the KS-test did not perform optimally as it was ran on discrete data instead of continuous.

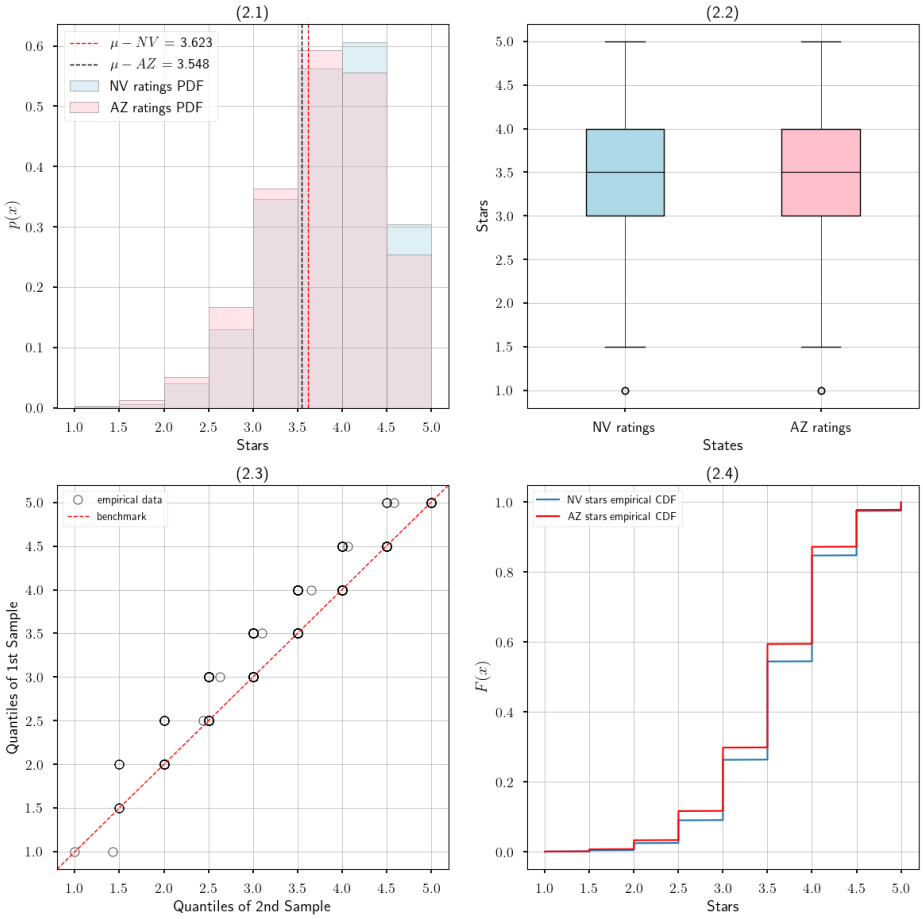

Figure 2: A comparison of the two PDFs for NV and AZ (2.1), box-plots for NV and AZ (2.2), a 2-sample Q-Q plot for the two metropolitan areas (2.3) and the comparison between NV and AZ empirical CDFs (2.4)

4 Conclusion

In conclusion, in the first research question, we were clearly able to reject the null-hypothesis of the data following a Pareto distribution, both from a visual analysis and through the KS-test. Although slight similarities could be observed visually, we could clearly see a difference, which is expected when fitting a theoretical distribution to real data. When analysing similarities between the ratings in the NV and AZ metropolitan areas, we were also able to reject the null-hypothesis with a KS-test, however it could be argued that the results were less conclusive, due to the test being intended for continuous data and our observations being discrete.