Introduction

In this report, we study the statistics of online reviews in the metropolitan area of Arizona. Then, we analyse the consistency of ratings for the *Nightlife* business in Arizona (AZ) and Nevada (NV).

Our data is discrete therefore we acknowledge that some of the continuous approximations made in this report will have shortcomings, nevertheless they will help us understand the statistical behaviour of our sample.

Methodology and results – RQ1

We study the empirical distribution of the number of reviews written by each individual user and the number of reviews received by each individual business for the metropolitan area of Arizona. We observe that:

- **Positive skewness** *(18.7 for user reviews and 8.1 for business reviews)*– implying that the distribution is asymmetric, and the tail is skewed on the right.
- **High excess kurtosis** (*712 for user reviews and 112 for business reviews)*– suggesting heavy tails.
- Based on the above, we expect a more pronounced power law behaviour (i.e., smaller tail exponent) for user reviews.

When plotting the histograms on doubly logarithmic axes, we see that the empirical distributions fall approximately on a straight line, which suggests a power law behaviour, with a scaling parameter α given by the absolute slope of the straight line.

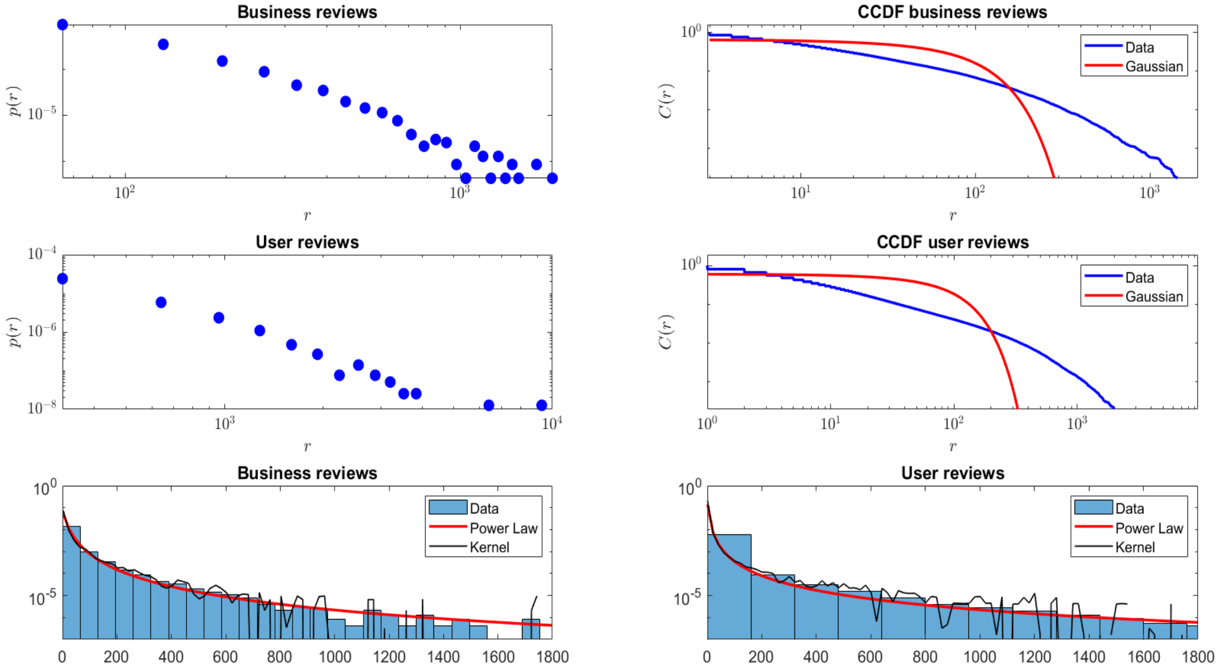
We also see that our data decays much slowly compared to a Gaussian distribution, which decays exponentially in the tails.

We try fitting different distributions to our data using Matlab’s fitdist function - which solves for the optimal parameters of the chosen pdf based on the maximum likelihood estimation.

The composite distribution of a Gaussian Kernel for the central part and a power law for the tail, appears a suitable fit to our empirical data.

Figure 1, summarises the above:

Figure 1: Empirical properties of the data and the presence of power law behaviour



Distribution fitting

1. Estimate the lower bound of the power law behaviour.

We choose a minimum value $x_{\min}$, that makes the probability distribution of the empirical data that fall above the lower bound and the best-fit power-law model as similar as possible.

For a power law pdf given by $p(x) = \frac{\alpha}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-(\alpha+1)}$, based on the maximum likelihood principle we estimate an optimal tail exponent parameter $\alpha^* = T/\sum_{i=1}^{T}\log\left(\frac{x_i}{x_{min}}\right)$.

A different estimate of α^*, is obtained for each element of the logarithmically spaced array of $x_{\min}$ thresholds.

An objective solution to choosing the lower bound is to minimize a goodness-of-fit statistic such as the Kolmogorov-Smirnov (KS) statistic, between the fitted power lower model $p(\mathrm{x}|\alpha^*)$ and the data.[1]

The optimal lower bound $x_{\min}*$, corresponds to the largest p-value for which the null hypothesis of the KS test cannot be rejected at the 5% significance level. This analysis, results in a conservative estimation of the lower bound because the KS test measures *the largest deviation* in cumulative density between the fitted model and the empirical data.

$$D_{nm} = \sup_x \left|C_{1,n}(x) - C_{2,m}(x)\right|$$

[1] Clauset et al. (2009): “Power-Law Distributions in Empirical Data

Our analysis suggests an $x_{\min}$ threshold of **145** for the business reviews sample and **322** for the user reviews. (Further details in appendix **A1 Lower bound of the power law behaviour**)

2.Calibrate a Gaussian Kernel to the central part of the data

We argue that a Kernel distribution is a suitable fit for the central part of our empirical data because it is a nonparametric representation of the pdf and we want to avoid making assumptions about the body of the distribution.

We will use a Gaussian Kernel pdf for the central part of the distribution (i.e., for x< $x_{\min}$):

$$p(x|h,\{x_i\}) = \frac{1}{N_{T\sqrt{2\pi h^2}}} \sum_{i \in T} \exp\left(-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2\right)$$

We estimate the optimal bandwidth parameter by maximising the likelihood that the empirical data below the lower bound $x_{\min}$, is generated by the Gaussian Kernel pdf evaluated on the validation set.

$$\log \mathcal{L}\,(h) = \sum_{j \in V} \log p\,(x|h,\{x_i\}_{i\varepsilon T})$$

$$h_{opt} = argmax\big(\log L\,(h)\big)$$

Our analysis suggests that h_{opt} is **3.26** for business reviews and **1** for user reviews.

3.Body-tail fitting to our empirical data

The calibration of the composite Kernel-Power Law distribution will be implemented using Matlab’s built-in *paretotails* function, which takes as arguments: the empirical dataset, the lower bound of the power law pdf, and a given distribution for the central part (i.e. the Gaussian Kernel with optimal bandwidth parameter h_{opt}).

The composite pdf fitted to our empirical data is:

$$p(x) = \begin{cases} C^{-1} \dfrac{1}{N_{T\sqrt{2\pi h^2}}} \displaystyle\sum_{i \in T} \exp\left(-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2\right) & for\, x < x_{\min} \\ C^{-1}\, \dfrac{\alpha}{x_{\min}} \left(\dfrac{x}{x_{\min}}\right)^{-(\alpha+1)} & for\, x \geq x_{\min} \end{cases}$$

where C^{-1} is a normalising constant. To ensure continuity at $x_{\min}$ the following should hold:

$$\frac{1}{N_{T\sqrt{2\pi h^2}}} \sum_{i \in T} \exp\left(-\frac{1}{2}\left(\frac{x_{\min}-x_i}{h}\right)^2\right) = \frac{\alpha}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-(\alpha+1)}$$

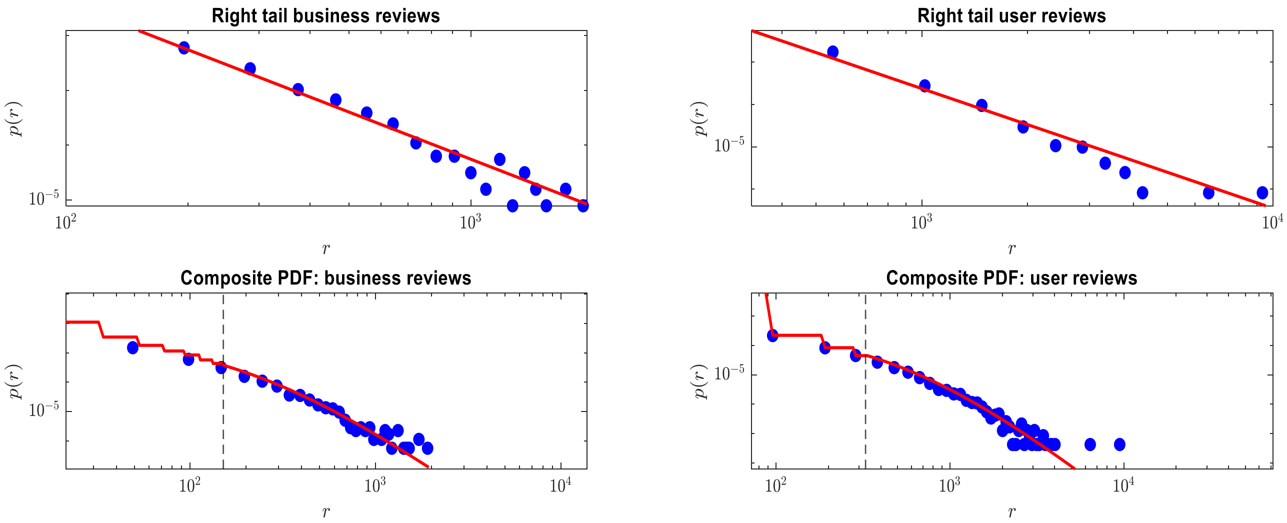
Using the $x_{\min}$ and h_{opt} parameters estimated above, the function will then calibrate the optimal tail exponent that maximises the log-likelihood of the composite pdf while penalising for discontinuities at $x_{\min}$.

The optimal tail exponent is **1.93** for business reviews and **1.84** for user reviews.

4. Estimating the goodness of fit

We will inspect how well the calibrated Kernel-Power law distribution fits our data. A visual comparison presented in Figure 2, indicates that the composite distribution provides a reasonable fit.

Figure 2: Body tail fitting: First row shows tail fitting using optimal xmin and alpha parameters for the power law pdf. The second row shows the fitting of the composite distribution.



Methodology and results – RQ2

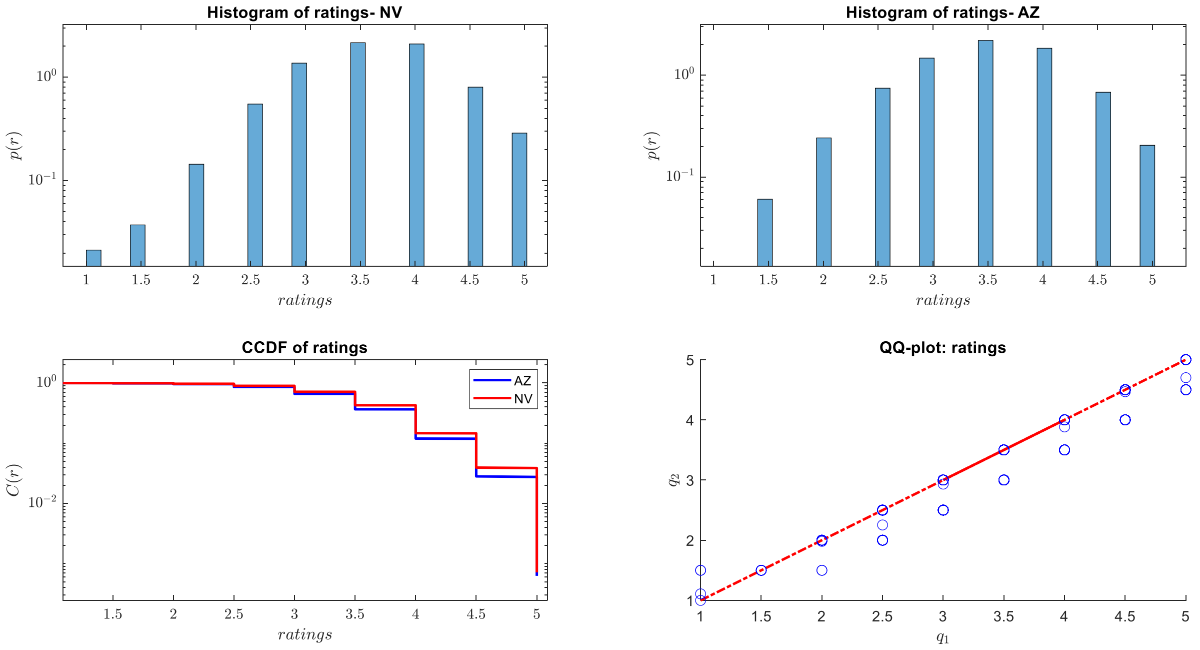
Are Yelp ratings for the Nightlife business consistent across metropolitan areas of AZ and NV?

We observe that the empirical distributions of the ratings are:

- centred around the same mean (i.e., 3.5-star ratings);
- display a negative skewness (-0.27 for AZ and -0.33 for NV) and slightly positive excess kurtosis (0.05 for AZ and 0.40 for NV).
- ratings in NV are more negatively skewed and excess kurtosis is higher compared to AZ. This implies a higher probability of larger deviations for ratings in NV.

We have plotted the empirical distributions, the corresponding CCDFs of the ratings and the QQ-plot across the two metropolitan areas.

Figure 3 Consistency of the distribution of Yelp ratings in AZ and NV



While the empirical moments and the plots indicate a high degree of similarity, this can be better quantified using a more rigorous statistical tool such as the KS test.

The KS test suggests that the null hypothesis that our two samples are generated from the same distribution is *rejected at the 5% confidence level with a p-value of 7.31e-03.*

However, the adaptation of the KS test to discrete data has some drawbacks:

- it is applied to *continuous distributions*.
- It is slightly conservative because it measures the *maximum distance* between the two empirical distributions.

An alternative option is the *Cramer-von Mises (CVM) test*, which uses the squared L2 norm of the difference between the two empirical distributions as the test statistic. The CVM test is 'better' in the sense that the *distance metric considers the whole of the two ECDFs*, rather than just picking out the largest distance. It is also applicable to discrete distributions.[2]

The CVM test also suggests that the null hypothesis is *rejected at the 5% confidence level with a p-value of 5.00e-02.* Comparing the two p-values one can see that the CVM test is less strict than KS.

[2] D.A Darling: “The Kolmogorov-Smirnov, Cramer-von Mises tests.

Appendix

A1 Lower bound of the power law behaviour

- We assume that the data follows a power-law distribution for values larger than a threshold $x_{\min}$. We explore an array of logarithmically spaced thresholds.
- For each individual threshold $x_{\min}$, we can use the following power law pdf —which is a continuous form of the power law behaviour, however it is standard practice to use it when working with discrete data.

$$p(x) = \frac{\alpha}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-(\alpha+1)}$$

- Maximum likelihood principle suggests an optimal estimation of the tail exponent parameter (α), which maximises the probability of observing the empirical data if they were generated by a power law distribution for $x > x_{\min}$. We estimate (α) for each given threshold $x_{\min}$.

$$\alpha^* = T/\sum_{i=1}^{T} \log\left(\frac{x_i}{x_{min}}\right)$$

- We generate random numbers that follow a power law distribution, by inverting the CDF.

General form: $u \in [0,1] \Rightarrow y = C^{-1}(u)$

Power-law: $u \in [0,1] \Rightarrow u \in [0,1] \Rightarrow y = x_{min} \cdot \mu^{-1/\alpha}$

- Then we estimate a goodness-of-fit statistic such as the Kolmogorov-Smirnov (KS) test, to test the null hypothesis that the power-law distributed random numbers and our empirical sample, are generated by the same distribution.

$$D_{nm} = \sup_x \left| C_{1,n}(x) - C_{2,m}(x) \right|$$

- Finally, we select the optimal lower bound that minimizes the KS statistic between the fitted model and the data (i.e., the value of $x_{\min}$ that corresponds to the largest p-value for which the null hypothesis of the KS test cannot be rejected at the 5% significance level).