

# POWER-LAW DISTRIBUTIONS IN BINNED DATA

## SDS2024

27 June 2024



created by


**XINYI GU**

**YIAN LI**



# Table of content

01	Abstract
02	Introduction
03	Definition of binned power-law distribution
04	Power-law fitting
05	Fitting modell evaluation
06	Alternative distributions



# Abstract

Power-law distributions are prevalent in phenomena like earthquakes, city populations, and wars. Accurately identifying these patterns is essential for understanding complex systems, but it's challenging due to data fluctuations, especially in binned data.

We replicated the study by Yogesh Virkar and Aaron Clauset on power-law distributions in binned empirical data.

The paper implements statistical methods to fit power-law distributions to binned data, validate the effectiveness of the power-law model, and compare the applicability of power-law distributions with other heavy-tailed distributions in real-world datasets.



# INTRODUCTION: POWER-LAW

A quantity  $x$  obeys a power-law if it is drawn from a probability distribution with a density of the form:

$$p(x) \propto x^{-\alpha}$$

$$\text{for } x > x_{min} > 0 \quad \wedge \quad \alpha > 0$$

**Scale invariance:** common, small events are not qualitatively distinct from rare, large events (e.g.: diameter of the tree branches).

**Fluctuations:** large fluctuations in the empirical distribution's upper tail.



# INTRODUCTION: TASKS

Apply a set of statistically principled methods for fitting and testing the power-law hypothesis for binned data:

- Power-Law **fitting**: maximum-likelihood (for all binning scheme)
- Power-Law **plausibility test**: hypothesis test (KS goodness-of-fit)
- **Alternative distribution** comparison: likelihood ratio test
- **Information loss** due to binning
- Real data **applying**

We mainly focus on the first three parts.

# BINNED POWER-LAW DISTRIBUTIONS: DEFINITION

## Power-law Distributions:

- Can be continuous or discrete.
- For continuous values, the probability density function (p.d.f.) is defined as:

$$P = Cx^{-\alpha}$$

- for  $x > x_{\min} > 0$ , where  $C$  is the normalization constant.
- For discrete values, the probability mass function is defined similarly for  $x > x_{\min}$ , where  $x$  is an integer.
- The paper focuses on continuous distributions but the methods are adaptable to discrete cases.

# BINNED POWER-LAW DISTRIBUTIONS: DEFINITION

Binned data consists of counts of observations over a set of non overlapping ranges.

Bin boundaries  $B$  are denoted as:

$$B = (b_1, b_2, \dots, b_k) \text{ for } b_1, k > 0$$

Bin counts  $H$  are denoted as:

$$H = (h_1, h_2, \dots, h_k) \quad h_i = \# \{b_i \leq x < b_{i+1}\}$$

# Probability Calculation for Bins

The probability that an observation falls within the  $i$ th bin is the fraction of total density in the interval  $[b_i, b_{i+1})$ :

$$\begin{aligned}\Pr(b_i \leq x < b_{i+1}) &= \int_{b_i}^{b_{i+1}} p(x) dx \\ &= \frac{C}{\alpha - 1} [b_i^{1-\alpha} - b_{i+1}^{1-\alpha}]\end{aligned}$$

The binning scheme  **$B$**  is assumed to be fixed by an external source.

If raw data were available, direct methods could be used to test the power-law hypothesis.



How to estimate  
 $\alpha$  and  $b_{min}$

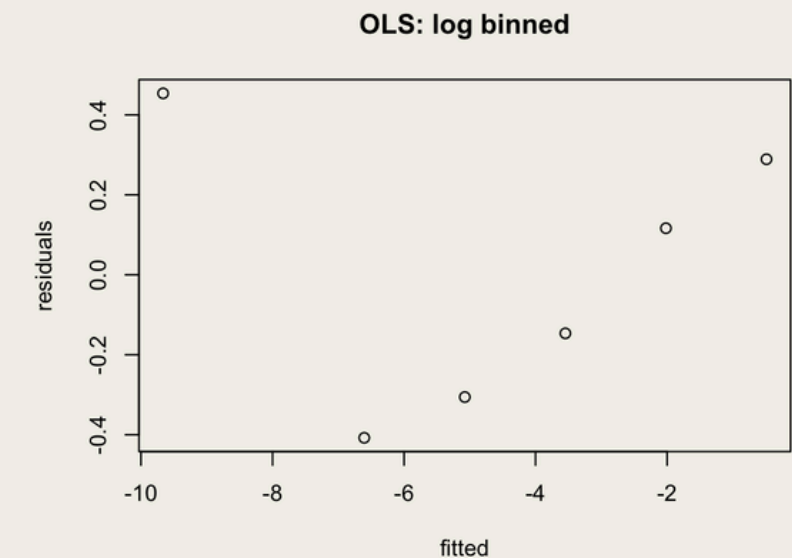
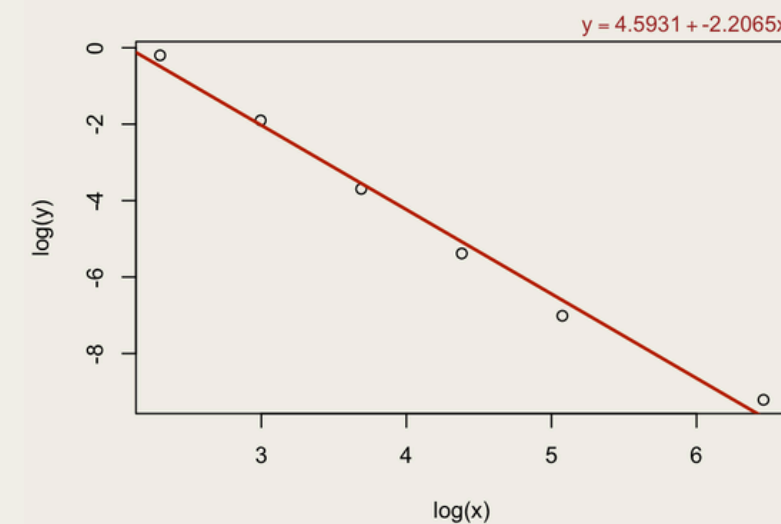
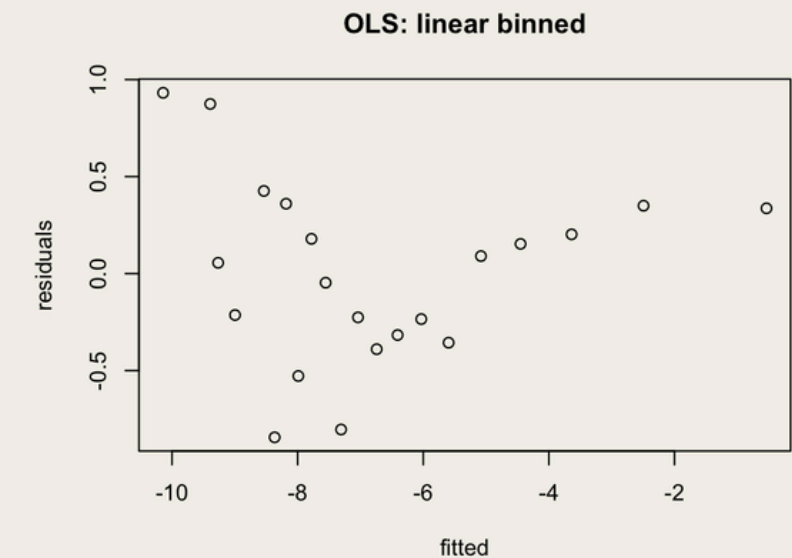
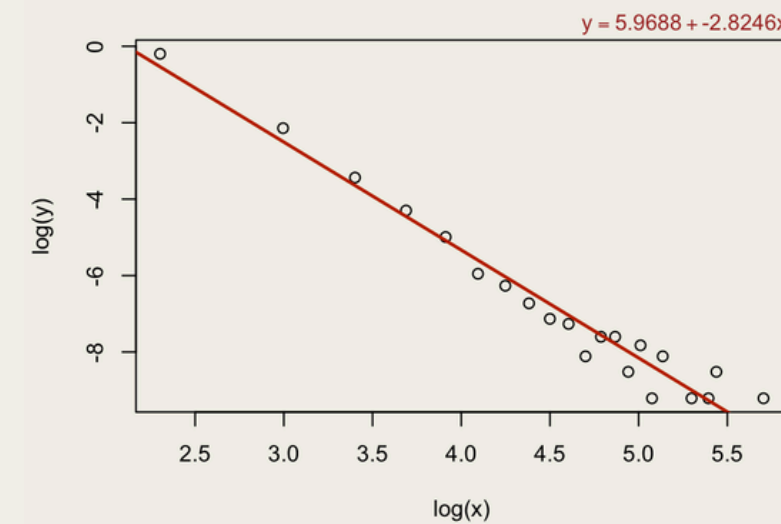


# POWER-LAW FITTING: LINEAR REGRESSION MODEL

Common approach: linear regression

$$p(x) = \ln C - \alpha \ln x$$

**Problems:** noises not normally distributed,  $R^2$  not reliable, etc..



# POWER-LAW FITTING: MODEL

- For each possible  $b_{min} \in (b_1, b_2, \dots, b_{k-1})$ , estimate  $\hat{\alpha}$  using **MLE**.
  - **Log-likelihood function**: maximize the equation given  $\alpha$

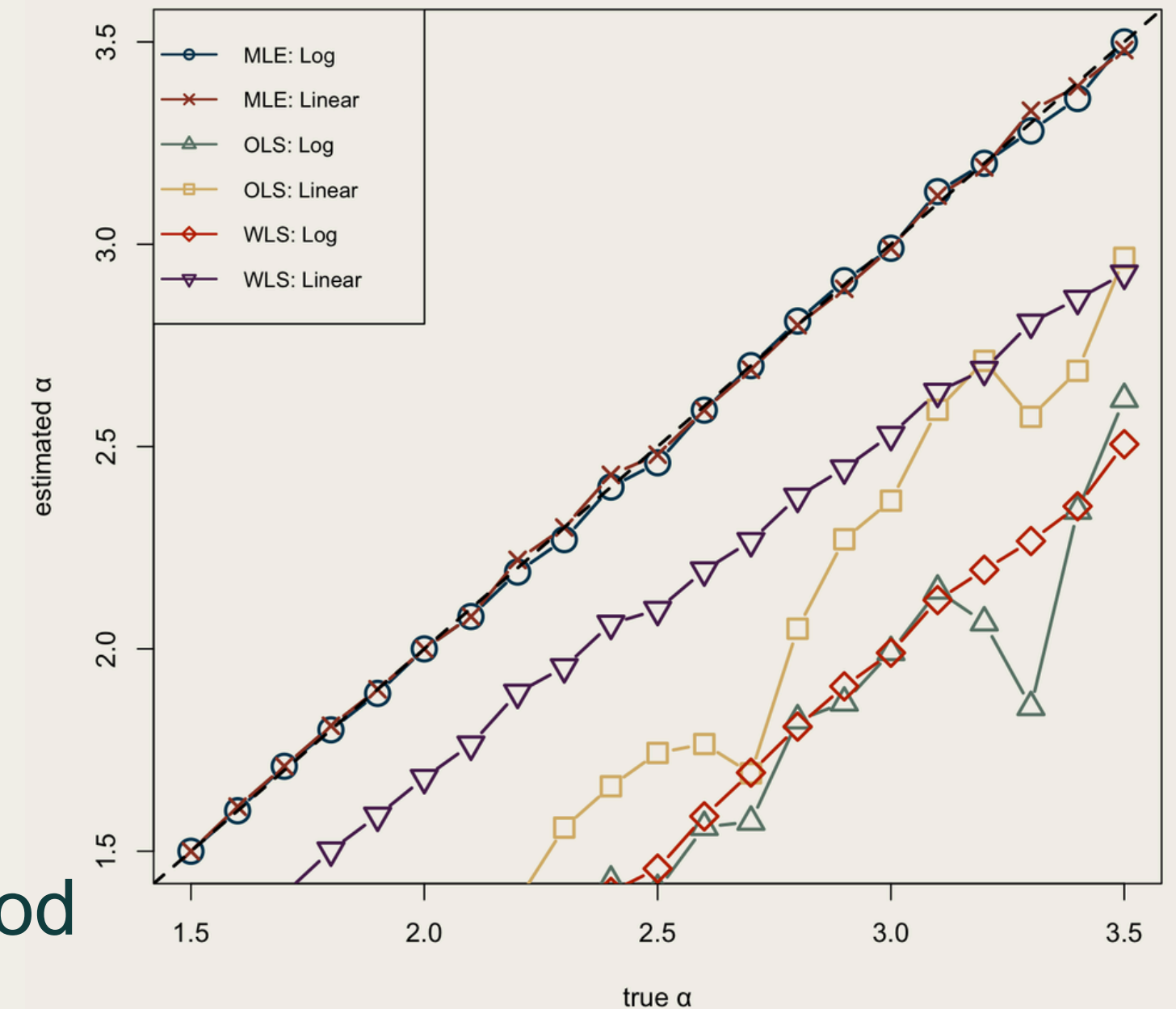
$$L = n(\alpha - 1) \ln b_{min} + \sum_{i=min}^k h_i \ln [b_i^{(1-\alpha)} - b_{i+1}^{(1-\alpha)}]$$

- Compute the Kolmogorov-Smirnov (KS) goodness-of-fit statistic between the fitted c.d.f. and the empirical one.
  - **KS distance**:  $D = \max_{b \geq b_{min}} |S(b) - P(b|\alpha, b_{min})|$
- Choose as  $\hat{b}_{min}$  the bin boundary with the **smallest** KS statistic.

# POWER-LAW FITTING: PERFORMANCE $\hat{a}$

## Performance of MLE:

- carry out the **accuracy** test on randomly generated datasets ( $N = 10,000$ )
  - fix  $x_{min} = 10$
  - $\alpha$  from 1.5 to 3.5 with step 0.1
- **bin** the data with 2 different scheme:
  - **linear**:  $b_i = 10$
  - **logarithmic**:  $b_i = 10 \times 2^{(i-1)}$
- Use the **MLE** and **Linear regression** method to estimate



# POWER-LAW FITTING: PERFORMANCE $\hat{b}_{min}$

## Performance of KS goodness-of-fit:

- evaluate the accuracy using a synthetic data that follows the power-law distribution above but some other distribution. The form of the density is:

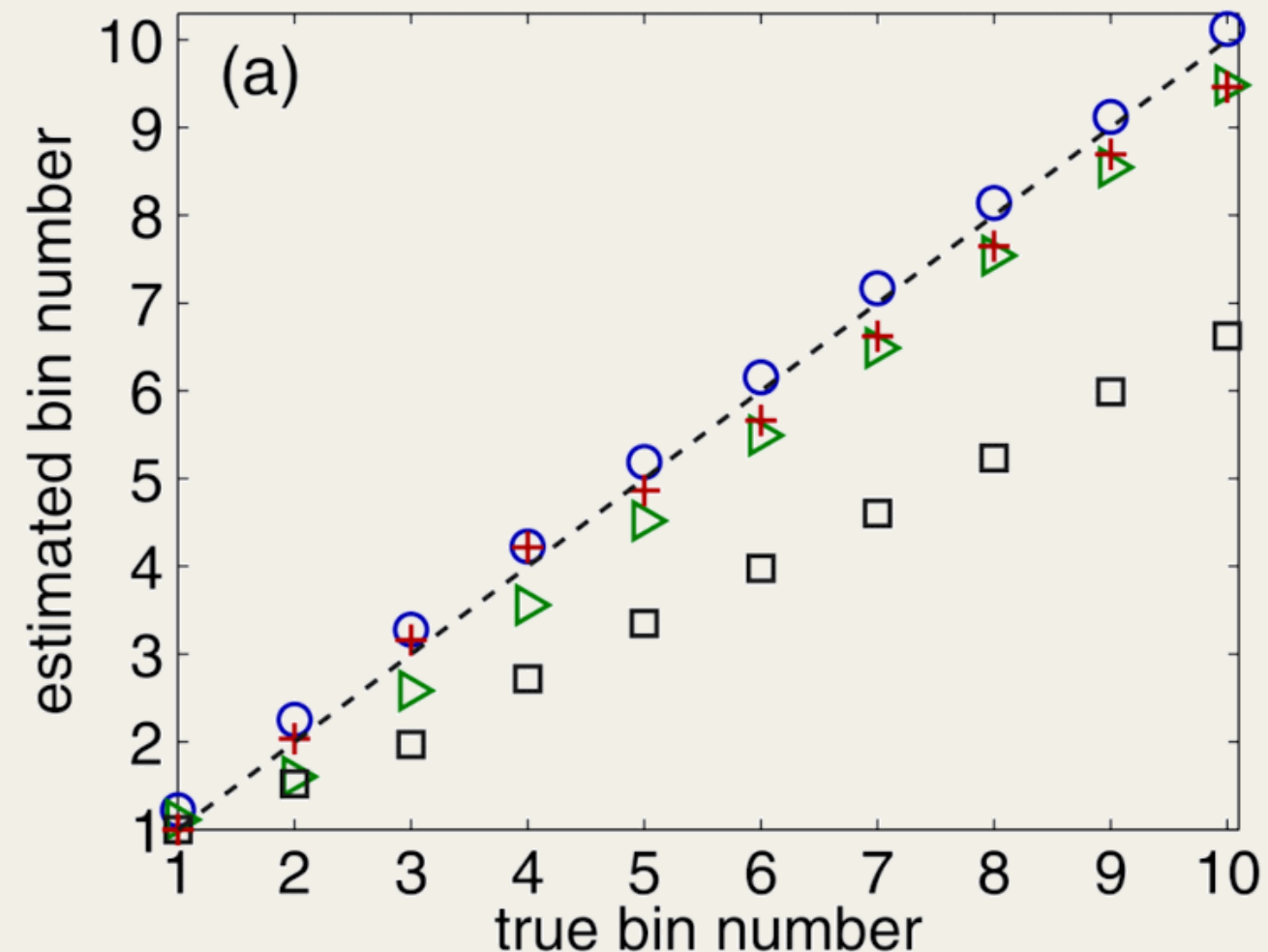
$$p(x) = \begin{cases} C e^{-\alpha \left( \frac{x}{b_{min}} - 1 \right)}, & b_1 \leq x < b_{min} \\ C \left( \frac{x}{b_{min}} \right)^{-\alpha}, & otherwise \end{cases}$$

- compare the **KS statistic** and the **RT method** by Reiss and Thomas (2017)
  - selects the bin boundary that minimizes the following formula, aiming at **minimize the asymptotic mean squared error** in estimated  $\alpha$ .

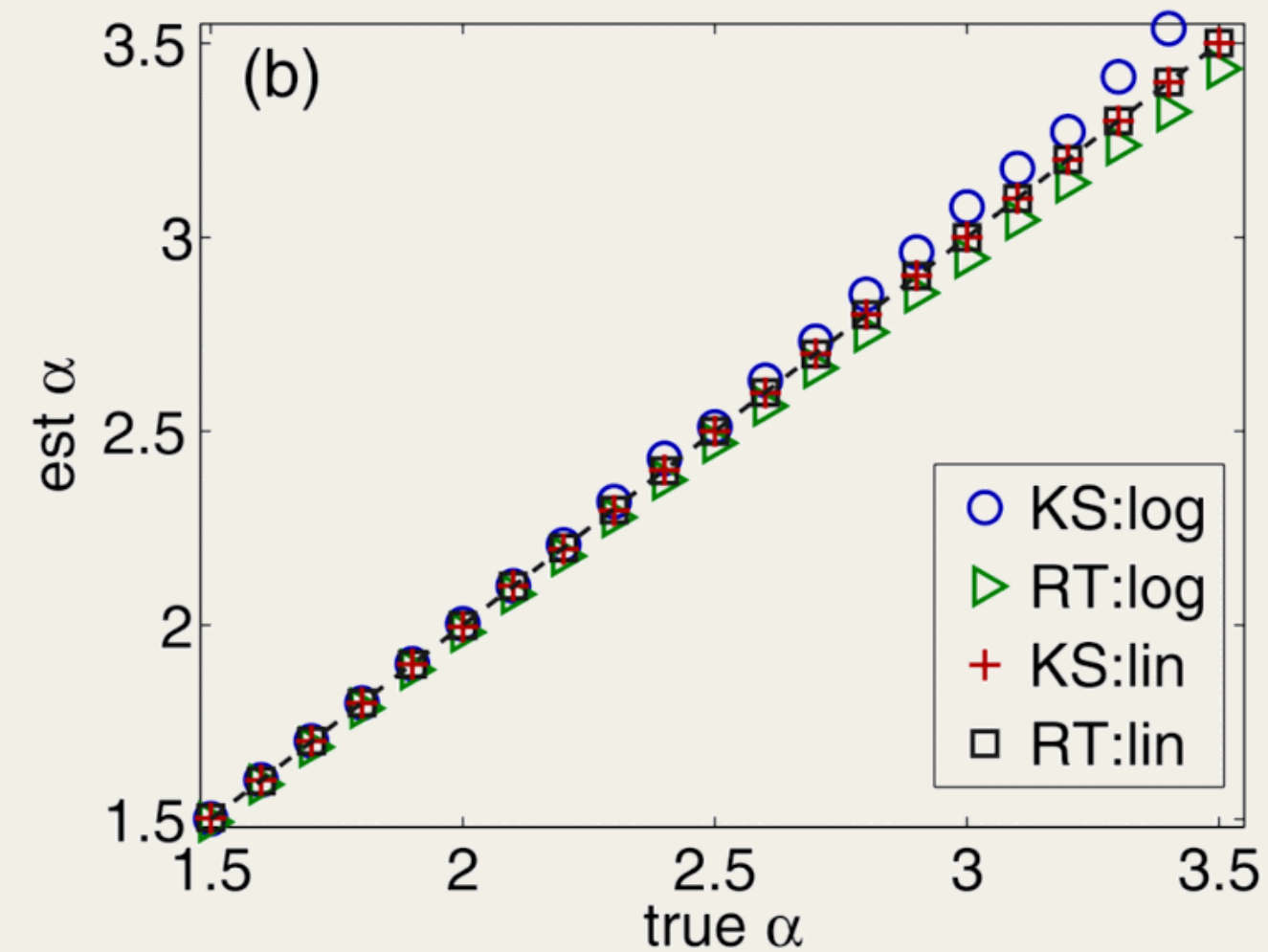
$$\frac{1}{\sum_{j=min}^k h_j} \sum_{i=min}^k \left( |\hat{a}_i - median(\hat{a}_{min}, \dots, \hat{a}_k)| \left( \sum_{j=i}^k h_j \right)^\beta \right), \quad 0 \leq \beta < \frac{1}{2}$$

# POWER-LAW FITTING: PERFORMANCE $\hat{b}_{min}$

fixed  $\alpha=3$



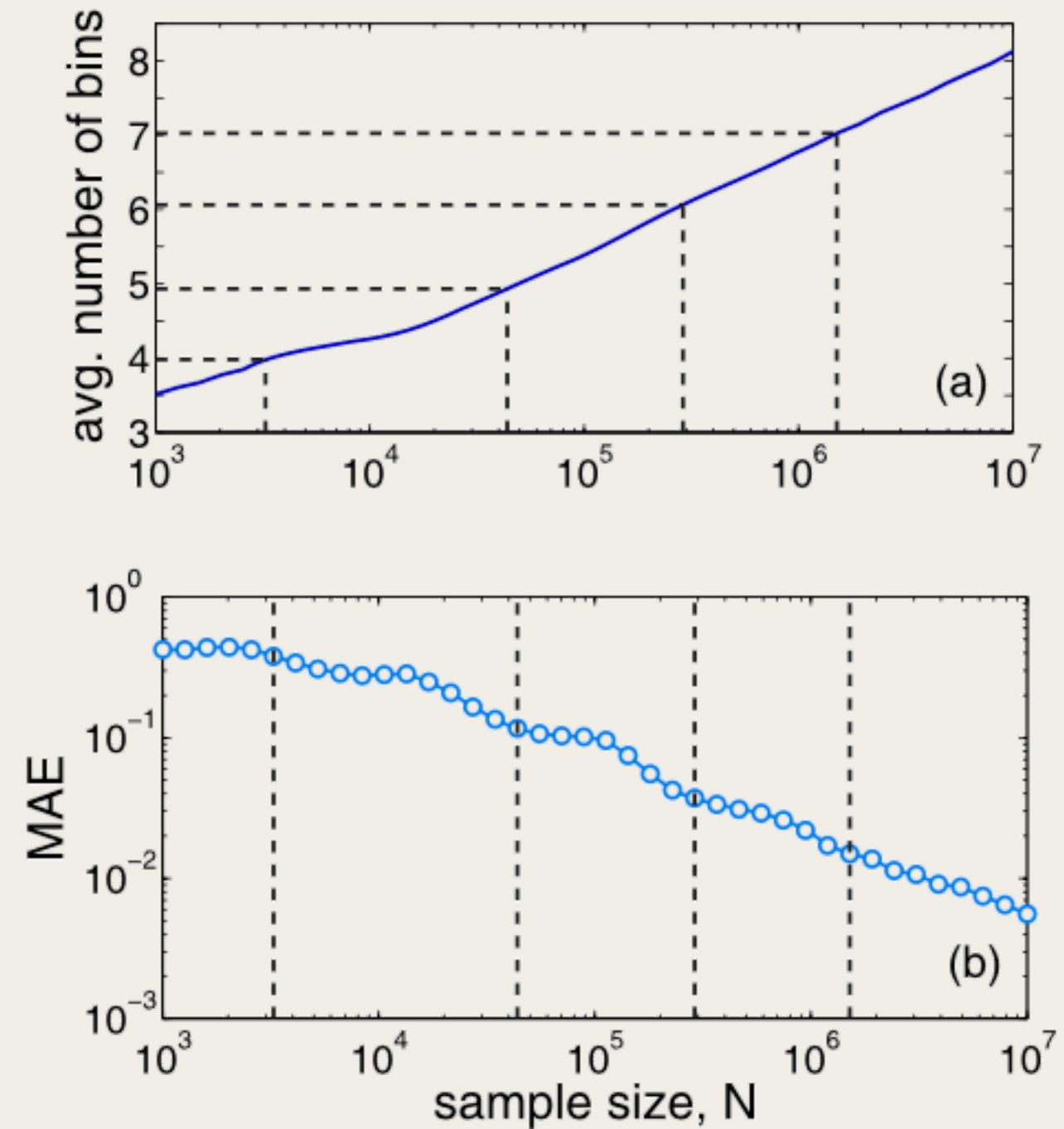
fixed  $b_{min}$  at the 10th boundary



- Experiment conducted on both binning scheme (linear and logarithmic)
- KS method for linear scheme shows slightly underestimation whereas the bias does not affect the estimation of scaling parameter

# POWER-LAW FITTING: BIAS

- **small sample bias:**
  - small # of observations
  - small # of bins
- experiment fixing parameters while varying the sample size





# POWER-LAW FITTING: APPLICATION

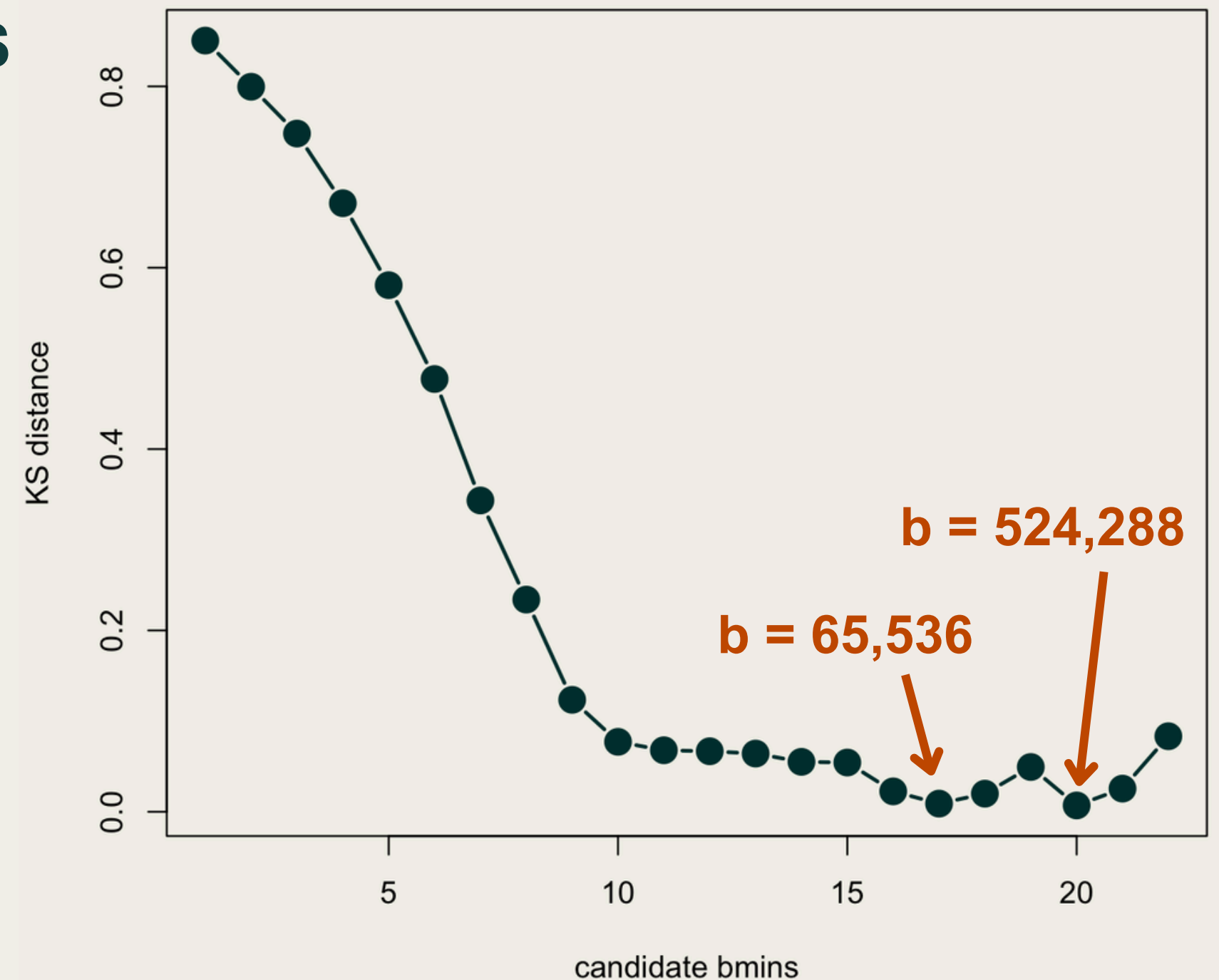
We tried to apply the fitting method on the **US cities population** dataset:

- best  $b_{min}$  : 524,288
- best  $\alpha$ : 2.66

**Result in paper:**

- best : 65,536
- best  $\alpha$ : 2.38

(We failed to find out why such a deviation exists.)





**How to evaluate if the fitting  
model is plausible**





# Testing the power-law hypothesis

## Problem Statement:

- Our methods can accurately fit a power-law tail model to binned empirical data, but they do not indicate if the fitted model is a good representation of the data.

## Challenge:

- Many heavy-tailed distributions, such as log-normal and stretched exponential (Weibull), can produce samples that resemble power-law distributions.

## Solution:

- To address this, we adapt the goodness-of-fit test by Clauset, Shalizi, and Newman (2009) for binned data.
- This test provides a p-value to determine if the power-law model is plausible.

# Goodness-of-Fit Test: Detailed Steps

- **Fit the Power-Law Model:**
  - Estimate the parameters  $\hat{\alpha}$  and  $\hat{b}_{min}$  using the methods described in the previous section.
  - Construct the hypothesized model  $M$  based on these parameters.
- **Compute Distance  $D^*$ :**
  - Calculate the KS statistic  $D^*$ , which measures the distance between the empirical data and the hypothesized power-law model.
- **Generate Synthetic Data:**
  - Use a semi-parametric bootstrap to create synthetic data sets that follow the power-law distribution above  $\hat{b}_{min}$  and the empirical distribution below  $\hat{b}_{min}$ .
  - Generate synthetic bin counts  $H'$ .
- **Fit the Power-Law Model to Synthetic Data:**
  - Fit the power-law model to the synthetic data  $H'$ , resulting in a new model  $M'$  with parameters  $\hat{\alpha}$  and  $\hat{b}_{min}$ .
- **Compute Distance  $D$ :**
  - Calculate the KS distance  $D$  between the synthetic data  $H'$  and the new model  $M'$ .
- **Repeat and Calculate p-value:**
  - Repeat steps 2-5 many times to build a distribution of the distances  $D$ .
  - Calculate the p-value  $p = P(D \geq D^*)$ , which is the fraction of synthetic distances  $D$  that are at least as large as  $D^*$ .

# Generating Synthetic Data Sets

To evaluate the performance of estimation methods, we first generate synthetic data from a power-law distribution. This approach allows us to control the underlying distribution and ensures that the true parameters are known.

## Semi-Parametric Bootstrap:

- $n$ : Number of observations in the power-law region from data  $H$ .
- Probability  $n/N$ : Generate a non-binned power-law random deviate from  $M$  and increment the corresponding bin count in the synthetic data set.
- Probability  $1 - n/N$ : Increment the count of a bin  $i$  below  $bmin$  chosen with probability proportional to its empirical count  $hi$ .

# Generating Synthetic Data Sets

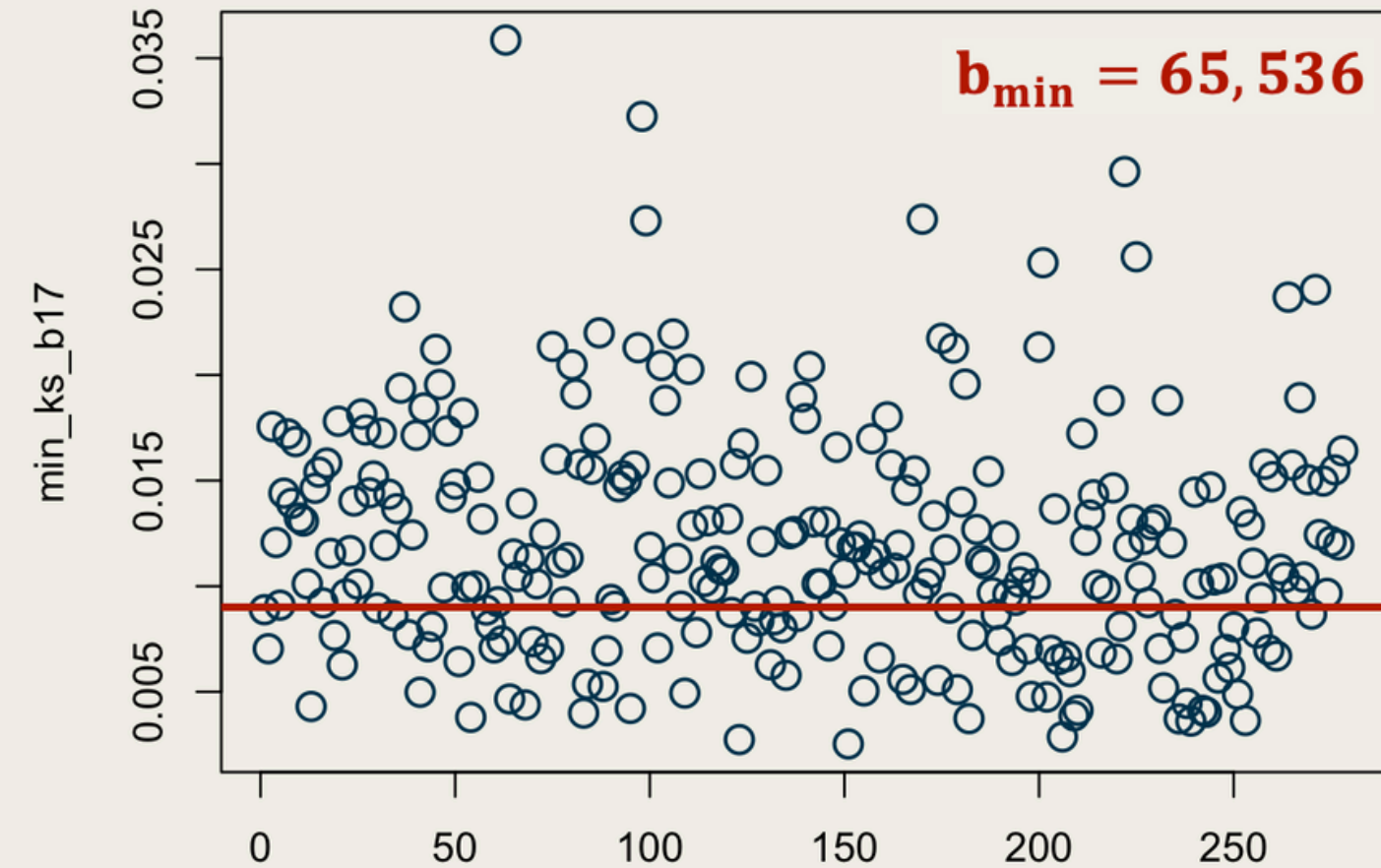
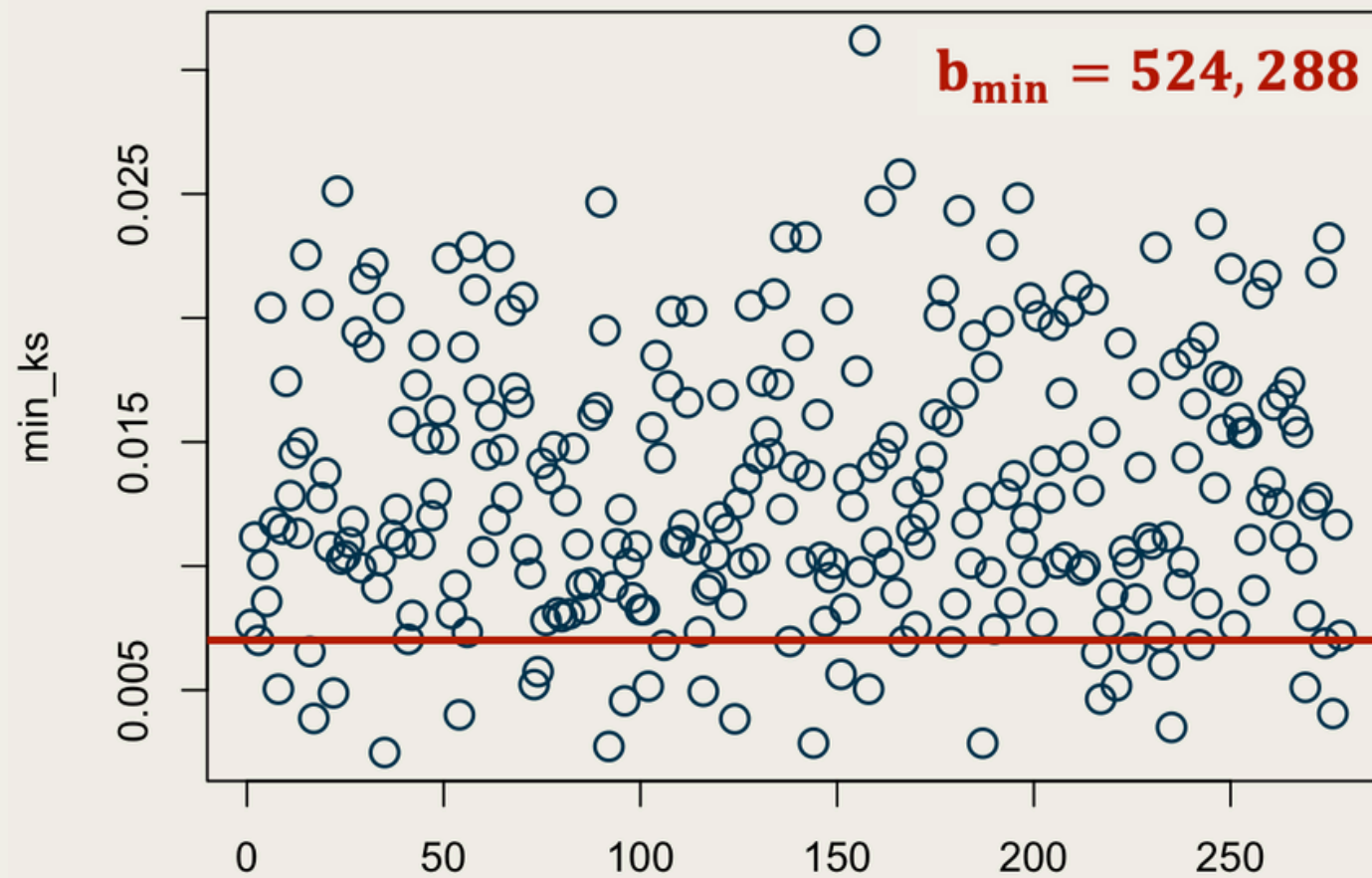
## Number of Synthetic Data Sets:

- Clauset, Shalizi, and Newman's (2009) recommendation: generate at least  $1/\varepsilon^2$  synthetic data sets for accuracy.
- Example: To know  $p$  within  $\varepsilon = 0.01$ , generate about **2500** synthetic data sets.

## Interpreting p-value:

- Conservative choice: reject power-law hypothesis if  $p < 0.1$ .
- Large p-value does not imply correctness of power-law model.
- Large p-value can arise due to:
  - Alternative distributions fitting the data as well or better.
  - Small  $n$  or few bins above  **$bmin$**  making it hard to rule out power-law shape.
- Interpret large p-value cautiously if  **$n$**  or number of bins is small.

# Goodness-of-Fit Test

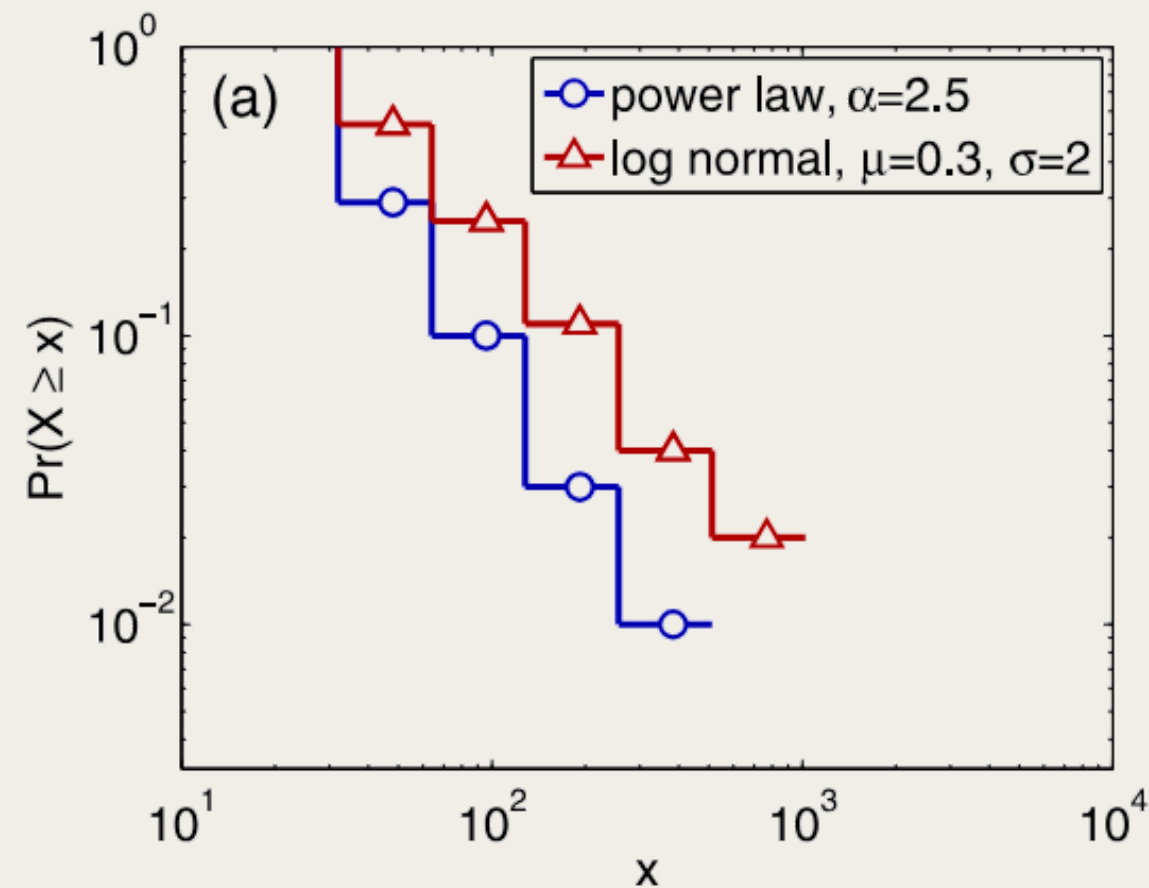


- Points: Each point represents the KS statistic for one of the bootstrap samples.
- Line: The red line represents the KS statistic  $D^*$  for the best  $\hat{b}_{min}$  determined from the empirical data.

The analysis using the KS statistic and bootstrap method helps in determining the robustness of the power-law fit.



# Performance of the Goodness-of-Fit Test



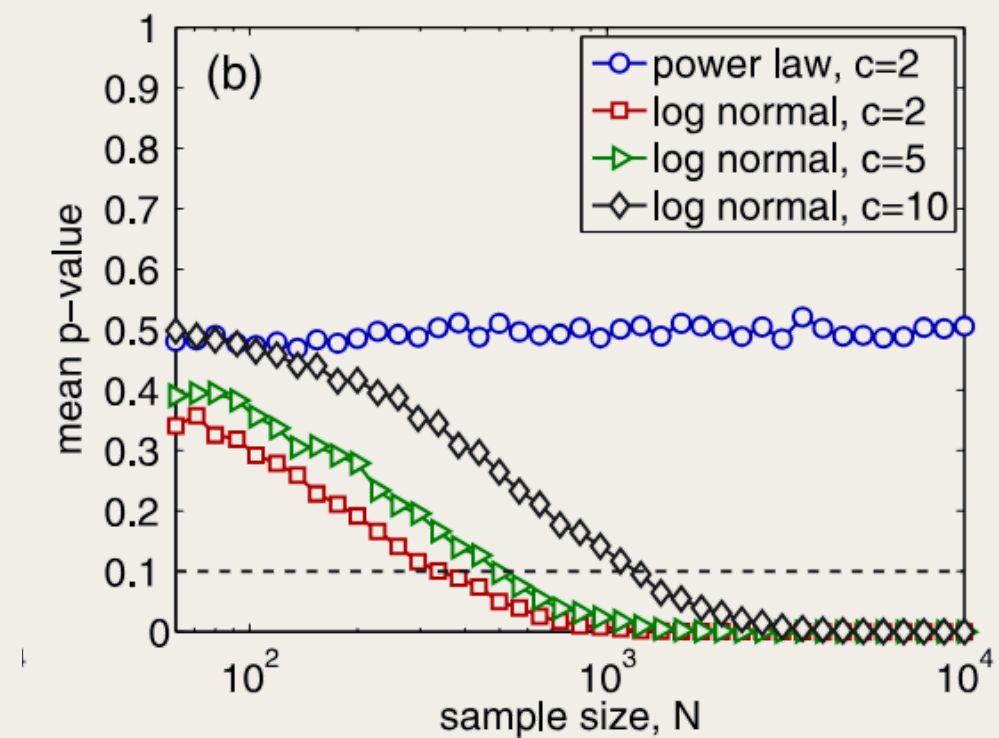
**Figure(a)** showing logarithmic histograms of synthetic data from power-law and log-normal distributions.

- The log-normal distribution provides a strong test due to its similarity to power-law distributions on log-log axes for various sample sizes.

# Performance of the Goodness-of-Fit Test

**Figure(b)** Showing the average p-value as a function of sample size  $N$  for the power-law hypothesis.

- Correct Model: When fitting the correct model, the p-value is uniformly distributed with a mean of **0.5**.
- Log-Normal Data:
  - P-value remains above the rejection threshold for small samples ( $N \leq 300$ ).
  - Correctly rejects the power law for larger samples.
- Impact of Binning Scheme:
- The sample size required for correct rejection depends on the binning scheme.
- Coarser binning schemes (larger  $c$ ) require larger sample sizes.





# What about alternative distributions



# Alternative attributes

## Log-likelihood ratio test:

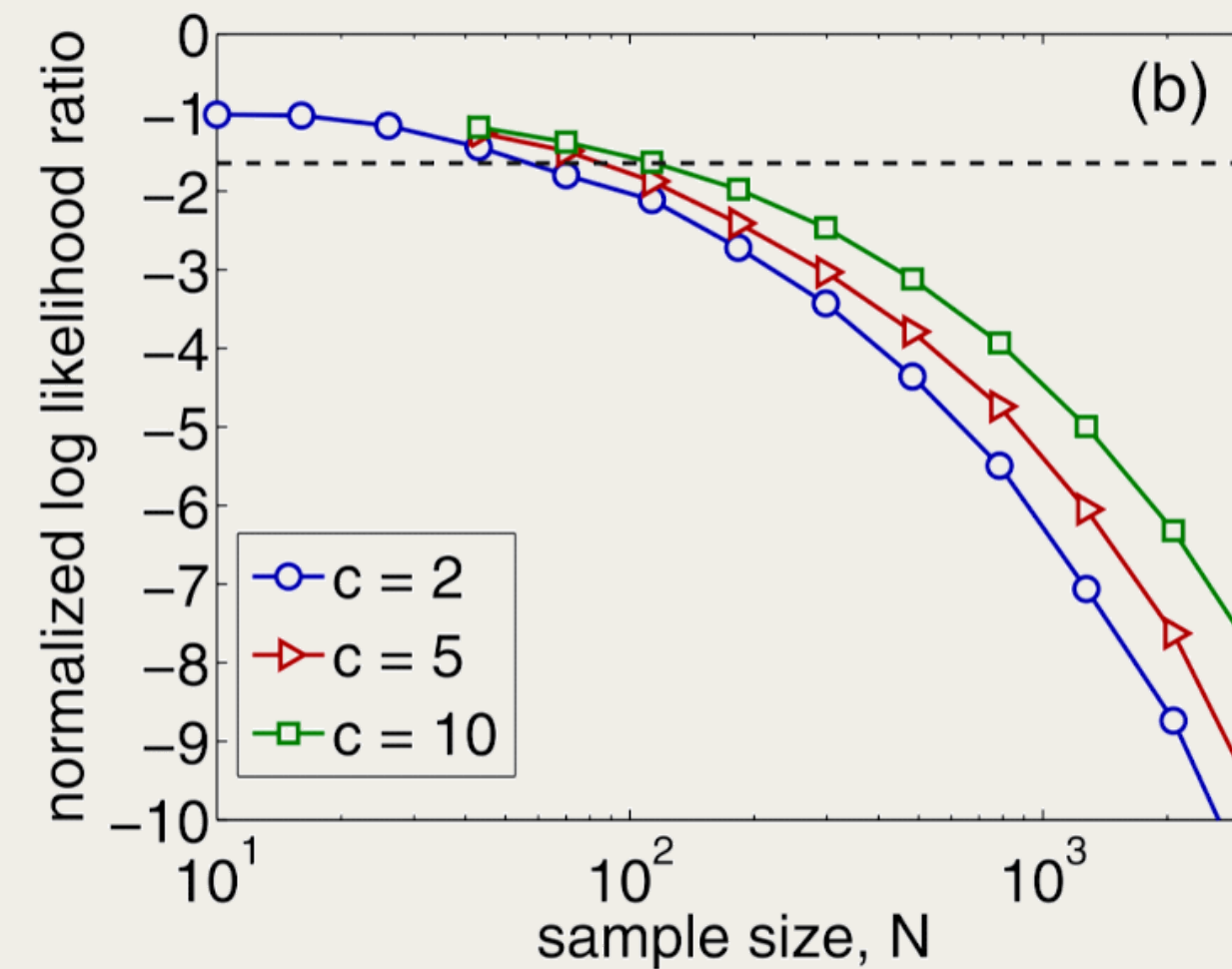
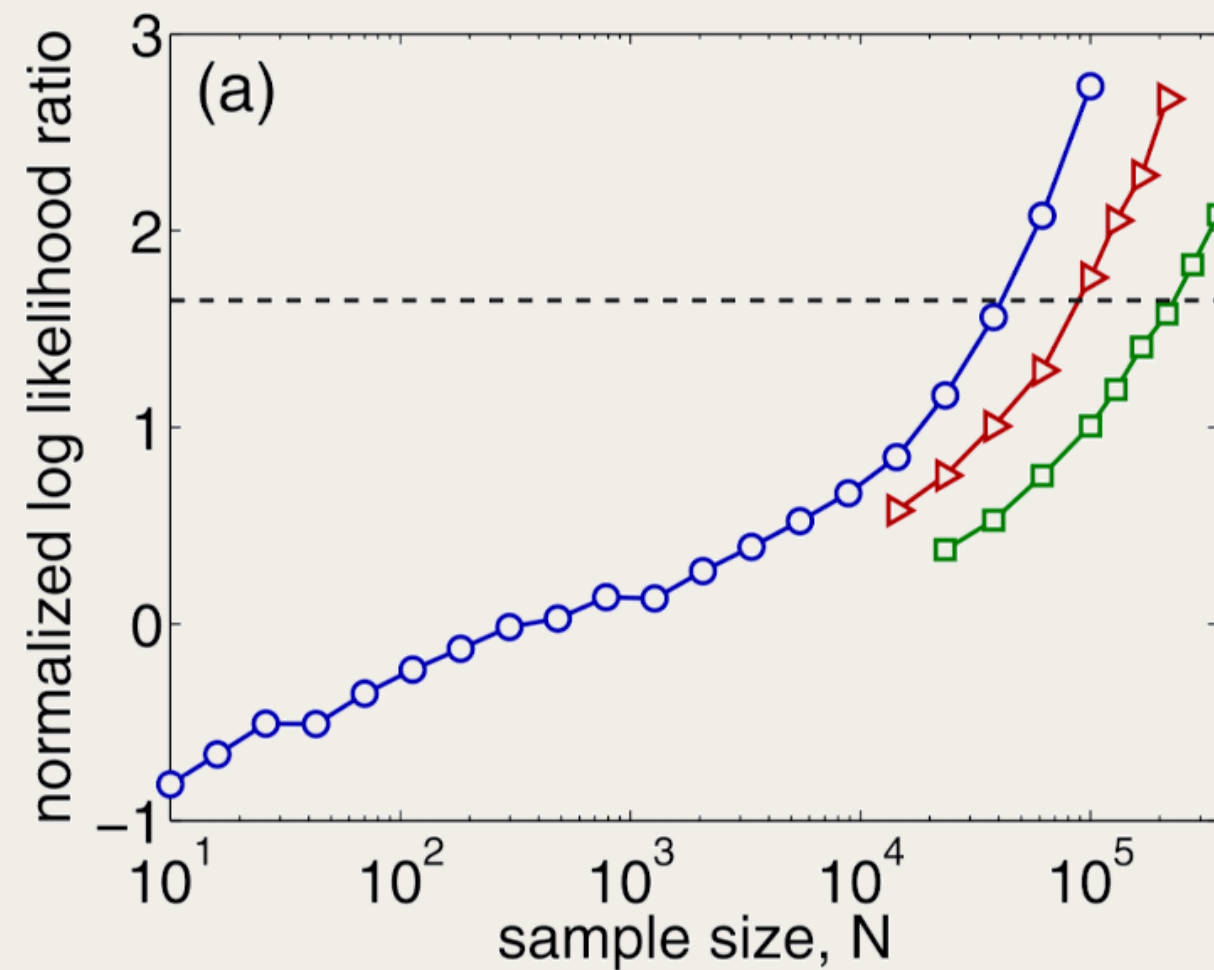
$$R = \ln \left( \frac{L_A(H|\hat{\theta}_A)}{L_B(H|\hat{\theta}_B)} \right)$$

- Compared with 4 distribution families: power-law with cutoff, exponential, stretched exponential and log-normal
- The distribution should be fitted to the same bin counts chosen by the power law model
- Test the statistic significance of the sign of R to increase the reliability of the likelihood ratio test

# Alternative attributes

## Performance of LRT:

- Power-law hypothesis vs. log-normal hypothesis on synthetic data
- Experiment with different coarseness of the binning scheme





# Thank You

----