Task 2: Fitting and Comparing Distributions

COURSE NAME

CSDS 413 Introduction to Data Analysis

Authors:

Wiam Skakri
Jacob Anderson

September 18, 2025

Contents

Context

In this task, you will explore how different types of real-world datasets may follow different distributions. You will need to develop a set of hypotheses and perform experiments to validate your own hypotheses.

1 Normal Distribution Dataset

1.1 Part A: Developing Hypotheses

Identify and collect a real-world dataset that you hypothesize follows a Normal distribution. Please be clear about the reasoning behind your hypothesis and be specific about the source of the dataset.

https://www.kaggle.com/datasets/uciml/iris

1.2 Part B: Fitting Distributions

For this exercise, we will call each of the four different theoretical distributions (normal, uniform, power law, exponential) a “model". Fit the dataset (i.e., estimate the model parameters) against each model (not just the one you hypothesized) using maximum likelihood estimation (or using any technique you think is appropriate; make sure to comment on the validity of your approach). This should result in a total of **4 parameter sets**. Report the estimated parameters in the following tabular format:

		Model			
Dataset	***# Observations***	**Normal**	**Uniform**	**Power law**	**Exponential**
Dataset 1	n_1	μ_1, σ_1	a_1, b_1	$\alpha_1, x_{\min_1}$	λ_1

Be sure to show the code you used to arrive at your final estimates clearly.

1.3 Part C: Comparing Real and Synthetic Data

For each fitted distribution (there will be 4 of them for this dataset, each corresponding to a different model), generate a synthetic sample of data points equal to the sample size of the real dataset using the respective model parameters you inferred from the real dataset.

Compare the real vs. synthetic data distributions using methods you think are the most appropriate, including visualizations. So, for this dataset, we compare the original dataset to four synthetic datasets, all with equal number of observations, but each synthetic dataset is generated using a different model.

For this dataset, identify the synthetic dataset (which corresponds to a model) that is most similar to the original data in terms of its distribution.

Now revisit your initial hypothesis. For this dataset: Did the dataset behave as expected, or was another model (assumed distribution) a better fit to the dataset? Reflect on why the observed results may differ from your expectations.

2 Uniform Distribution Dataset

2.1 Part A: Developing Hypotheses

Identify and collect a real-world dataset that you hypothesize follows a Uniform distribution. Please be clear about the reasoning behind your hypothesis and be specific about the source of the dataset.

2.2 Part B: Fitting Distributions

For this exercise, we will call each of the four different theoretical distributions (normal, uniform, power law, exponential) a “model". Fit the dataset (i.e., estimate the model parameters) against each model (not just the one you hypothesized) using maximum likelihood estimation (or using any technique you think is appropriate; make sure to comment on the validity of your approach). This should result in a total of **4 parameter sets**. Report the estimated parameters in the following tabular format:

		Model			
Dataset	***# Observations***	**Normal**	**Uniform**	**Power law**	**Exponential**
Dataset 2	n_2	μ_2, σ_2	a_2, b_2	$\alpha_2, x_{\min_2}$	λ_2

Be sure to show the code you used to arrive at your final estimates clearly.

2.3 Part C: Comparing Real and Synthetic Data

For each fitted distribution (there will be 4 of them for this dataset, each corresponding to a different model), generate a synthetic sample of data points equal to the sample size of the real dataset using the respective model parameters you inferred from the real dataset.

Compare the real vs. synthetic data distributions using methods you think are the most appropriate, including visualizations. So, for this dataset, we compare the original dataset to four synthetic datasets, all with equal number of observations, but each synthetic dataset is generated using a different model.

For this dataset, identify the synthetic dataset (which corresponds to a model) that is most similar to the original data in terms of its distribution.

Now revisit your initial hypothesis. For this dataset: Did the dataset behave as expected, or was another model (assumed distribution) a better fit to the dataset? Reflect on why the observed results may differ from your expectations.

3 Power Law Distribution Dataset

3.1 Part A: Developing Hypotheses

Identify and collect a real-world dataset that you hypothesize follows a Power Law distribution. Please be clear about the reasoning behind your hypothesis and be specific about the source of the dataset.

3.2 Part B: Fitting Distributions

For this exercise, we will call each of the four different theoretical distributions (normal, uniform, power law, exponential) a “model". Fit the dataset (i.e., estimate the model parameters) against each model (not just the one you hypothesized) using maximum likelihood estimation (or using any technique you think is appropriate; make sure to comment on the validity of your approach). This should result in a total of **4 parameter sets**. Report the estimated parameters in the following tabular format:

		Model			
Dataset	***# Observations***	**Normal**	**Uniform**	**Power law**	**Exponential**
Dataset 3	n_3	μ_3, σ_3	a_3, b_3	$\alpha_3, x_{\min_3}$	λ_3

Be sure to show the code you used to arrive at your final estimates clearly.

3.3 Part C: Comparing Real and Synthetic Data

For each fitted distribution (there will be 4 of them for this dataset, each corresponding to a different model), generate a synthetic sample of data points equal to the sample size of the real dataset using the respective model parameters you inferred from the real dataset.

Compare the real vs. synthetic data distributions using methods you think are the most appropriate, including visualizations. So, for this dataset, we compare the original dataset to four synthetic datasets, all with equal number of observations, but each synthetic dataset is generated using a different model.

For this dataset, identify the synthetic dataset (which corresponds to a model) that is most similar to the original data in terms of its distribution.

Now revisit your initial hypothesis. For this dataset: Did the dataset behave as expected, or was another model (assumed distribution) a better fit to the dataset? Reflect on why the observed results may differ from your expectations.

4 Exponential Distribution Dataset

4.1 Part A: Developing Hypotheses

Identify and collect a real-world dataset that you hypothesize follows an Exponential distribution. Please be clear about the reasoning behind your hypothesis and be specific about the source of the dataset.

4.2 Part B: Fitting Distributions

For this exercise, we will call each of the four different theoretical distributions (normal, uniform, power law, exponential) a “model". Fit the dataset (i.e., estimate the model parameters) against each model (not just the one you hypothesized) using maximum likelihood estimation (or using any technique you think is appropriate; make sure to comment on the validity of your approach). This should result in a total of **4 parameter sets**. Report the estimated parameters in the following tabular format:

		Model			
Dataset	***# Observations***	**Normal**	**Uniform**	**Power law**	**Exponential**
Dataset 4	n_4	μ_4, σ_4	a_4, b_4	$\alpha_4, x_{\min_4}$	λ_4

Be sure to show the code you used to arrive at your final estimates clearly.

4.3 Part C: Comparing Real and Synthetic Data

For each fitted distribution (there will be 4 of them for this dataset, each corresponding to a different model), generate a synthetic sample of data points equal to the sample size of the real dataset using the respective model parameters you inferred from the real dataset.

Compare the real vs. synthetic data distributions using methods you think are the most appropriate, including visualizations. So, for this dataset, we compare the original dataset to four synthetic datasets, all with equal number of observations, but each synthetic dataset is generated using a different model.

For this dataset, identify the synthetic dataset (which corresponds to a model) that is most similar to the original data in terms of its distribution.

Now revisit your initial hypothesis. For this dataset: Did the dataset behave as expected, or was another model (assumed distribution) a better fit to the dataset? Reflect on why the observed results may differ from your expectations.