Welcome to Online Engineering at George Washington University

Class will begin shortly

Audio: To eliminate background noise, please be sure your audio is muted. To speak, please click the hand icon at the bottom of your screen (**Raise Hand**). When instructor calls on you, click microphone icon to unmute. When you've finished speaking, ***be sure to mute yourself again***.

Chat: Please type your questions in Chat.

Recordings: Please note the recording of this class meeting will be available to download later today. The class recordings are to be used exclusively by registered students in this particular class. Releasing these recordings is strictly prohibited.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

SEAS 8510
Analytical Methods for Machine Learning

Lecture 8

Dr. Zachary Dennis

Adapted from Rodriguez, Etemadi, Fall 2023, George Washington University

Agenda

9:00 – 9:15		Discussion Group
9:15 – 9:45		Continuous Variables Review
9:45 – 10:30		Data Statistics and Metrics
10:30 – 10:40		*BREAK (10 min)*
10:40 – 11:45		Covariance and Correlation
11:45 – 12:00		Homework and Discussion Look Ahead

Assignments

Last week: Homework 6 and Discussion 6 due on 5/11 at 9 AM Eastern

This week: Homework 7 and Discussion 7 due on 5/18 at 9 AM Eastern

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Continuous Distributions

Continuous Distributions

Continuous random variables take on values in a continuum

Examples:

- Car Speed [0, 150]
- Gas Price [$3, $4]

Probability Density Function

- For a continuous random variable P(X=x*) = 0 for any specific value x*
- The pdf, f(x), is used to express probability of intervals

$$P(a \leq X \leq b) = \int_a^b f(x\_dx$$

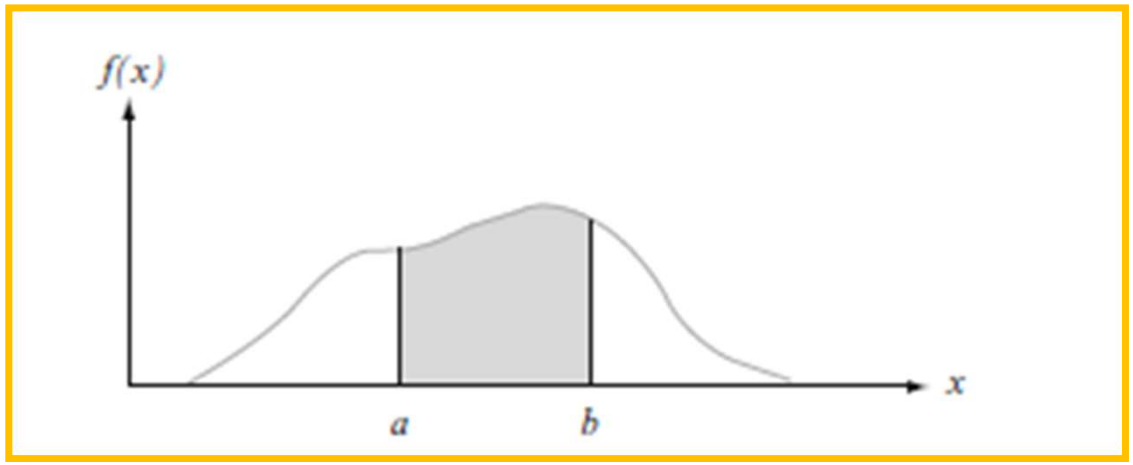
- Note that for a continuous random variable

  P(a < X < b) = P(a <= X < b) = P(a < X <= b) = P(a <= X <= b)

(Devore & Berk, 2018, p.160)

Continuous Distributions

That is, the probability that X takes on a value in the interval [a, b] is the area above this interval and under the graph of the density function (illustrated below).



The graph of f(x) is often referred to as the **density curve**.

(Devore & Berk, 2018, p.160)

Probability Distributions for Continuous Variables

For f(x) to be a legitimate pdf, it must satisfy the following two conditions:

1. $f(x) \geq 0$ for all x

2. $\int_{-\infty}^{\infty} f(x)dx = [\text{area under the entire graph of f(x)}] = 1$

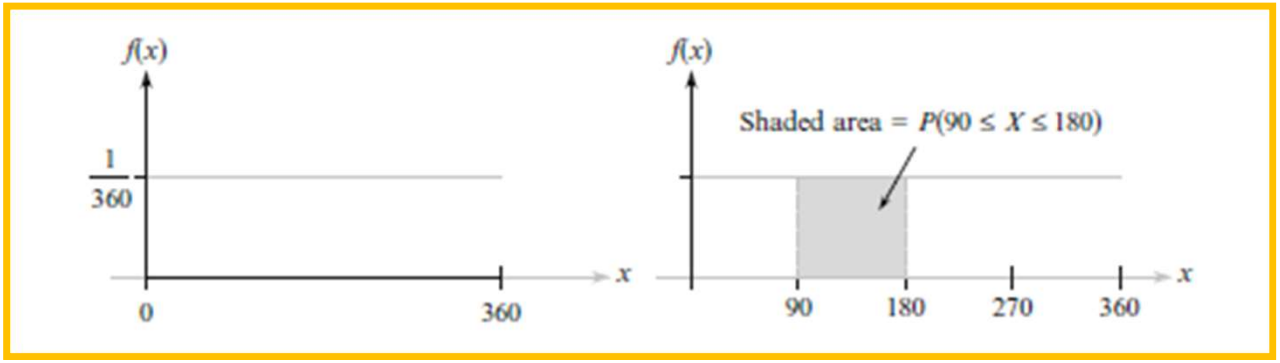
(Devore & Berk, 2018, p.161)

Continuous Uniform Distribution

Definition

A continuous rv X is said to have a uniform distribution on the interval [A, B] if the pdf of X is:

$$f(x; A, B) = \begin{cases} \dfrac{1}{B - A} & A \le X \le B \\ 0 & Otherwise \end{cases}$$



(Devore & Berk, 2018, p.161)

Normal Distribution

The **normal distribution** is the most important one in all of probability and statistics. Many numerical populations have distributions that can be fit very closely by an appropriate normal curve.

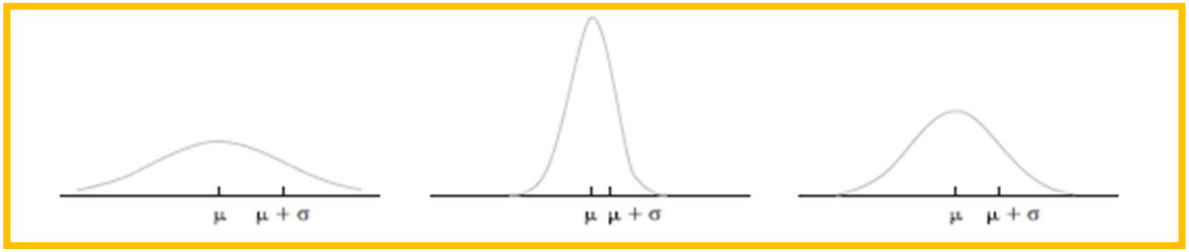
Definition

A continuous rv X is said to have a normal distribution with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is

$$f(x;\ \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty$$

Mean and Variance:

$$E(X) = \mu$$
$$V(X) = \sigma^2$$



(Devore & Berk, 2018, p.179)

Normal Distribution Applications

- Often an assumption in ML algorithms
  - Gaussian Naïve Bayes – assumes likelihood of features is normal
  - Linear Regression – assumes residuals normal
- Often an assumption in Statistical Tests
  - ANOVA (Analysis of Variance) – assumes residuals normal
  - T-tests – assumes populations samples from normal
- Featuring scaling - standardization

Exponential Distribution

Definition

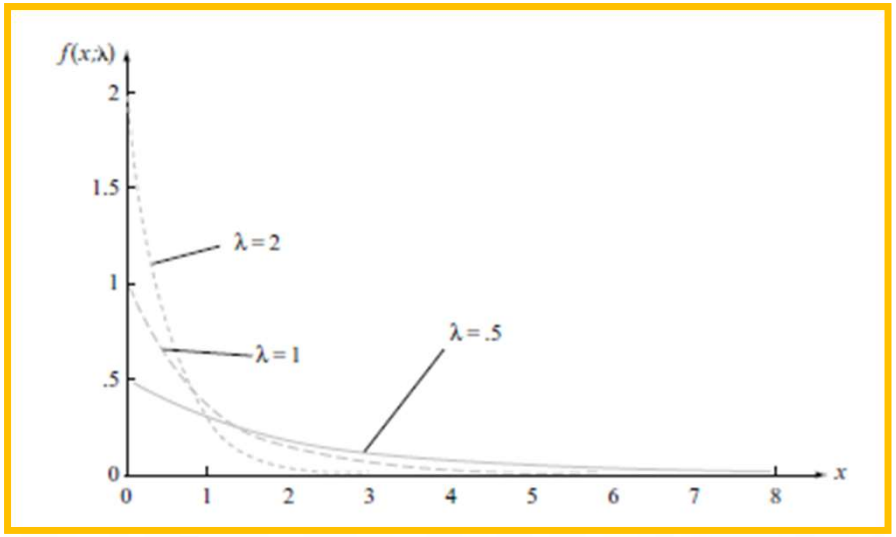
X is said to have an **exponential distribution** with parameter λ $(\lambda > 0)$ if the pdf of X is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & otherwise \end{cases}$$

$$F(x;\lambda) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

$$\mu = \frac{1}{\lambda}$$

$$\sigma^2 = \frac{1}{\lambda^2}$$



(Devore & Berk, 2018, p.198)

Exponential Distribution

The exponential distribution is frequently used as a model for the distribution of **times between the occurrence of successive events**, such as customers arriving at a service facility or calls coming into a switchboard.

The reason for this is that the exponential distribution is closely related to the Poisson process

(Devore & Berk, 2018, p.198)

Gamma Distribution

Definition

A continuous random variable X is said to have a **gamma distribution** if the pdf of X is

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where the parameters $\alpha \; and \; \beta \; \alpha > 0, \; \beta > 0.$

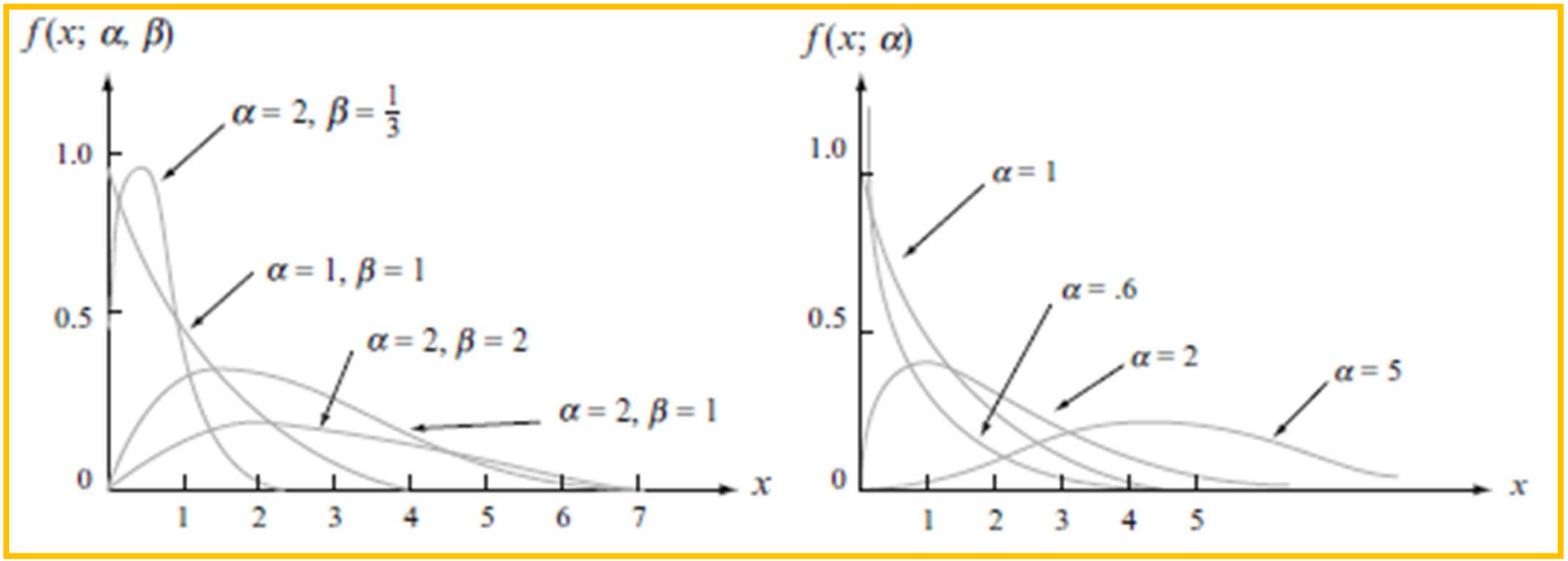
Mean and Variance:

$$E(X) = \mu = \alpha\beta$$
$$V(X) = \sigma^2 = \alpha\beta^2$$

The **standard gamma distribution** has β=1

(Devore & Berk, 2018, p.195)

Gamma Distribution



Gamma Density Curves

Standard Gamma Density Curves

The **exponential distribution** results from taking $\alpha = 1 \; and \; \beta = 1/\lambda$

(Devore & Berk, 2018, p.196)

Continuous Random Variables

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

$$Var(X) = E[(X - \mu_X)^2] = EX^2 - (EX)^2 = \int_{-\infty}^{\infty} x^2 f_X(x) dx - \mu_X{}^2$$

Probability Distribution Percentile

- Let X be some continuous r.v., and p be a probability of interest.
- Sometimes we are interested in finding q_p such that

$$F_X(q_p) = P(X \leq q_p) = p$$

  where the smallest value of q_p for which this is true is the p-th quantile (or 100p-th percentile) of the distribution for X. The median of a distribution is its 50th percentile

- **Example**: If exam scores are distributed normally with mean and std. dev. of 80 and 5, what is the 90th percentile score?

```
1 norm.ppf(0.9, loc=80, scale=5)
```

```
86.407757827723
```

Understanding Data

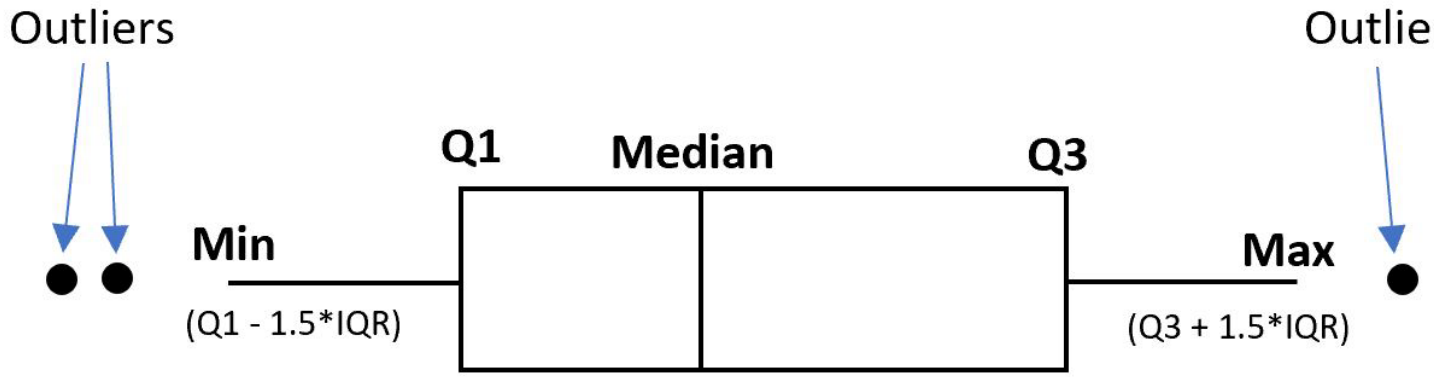
- In machine learning, understanding your data is the first step toward building effective models. Let's explore key data statistics and metrics.
- **Mean**: The mean is a measure of central tendency. It represents the average value of a dataset.
- **Median:** Another measure of central tendency. It represents the middle value when data is sorted (or the average of two middle values). Less affected by outliers compared to the mean
- **Variance**: Variance measures the spread or dispersion of data points. A higher variance indicates greater data variability.

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Standard Deviation**: Standard deviation is the square root of the variance, s_x. It provides a standardized measure of data dispersion.

Understanding Data

- **Quantiles:** Divide data into equal-sized subsets. Common quantiles include quartiles (25th, 50th, 75th percentiles). Useful for understanding data distribution and identifying outliers.
- **Interquartile Range (IQR):** IQR measures the spread of data around the median. It is the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Useful for identifying outliers and assessing data variability.
- **Skewness:** Quantifies the asymmetry of the data distribution.
  - Positive skew: Data is skewed to the right (tail on the right).
  - Negative skew: Data is skewed to the left (tail on the left).



https://www.statology.org/how-to-read-box-plot-with-outliers/

Understanding Data

- **Covariance:** Covariance measures the degree to which two variables change together.
  - Positive covariance: Variables move in the same direction.
  - Negative covariance: Variables move in opposite directions.

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- **Correlation:** Correlation is a standardized measure of covariance. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). Helps assess the linear relationship between two variables.

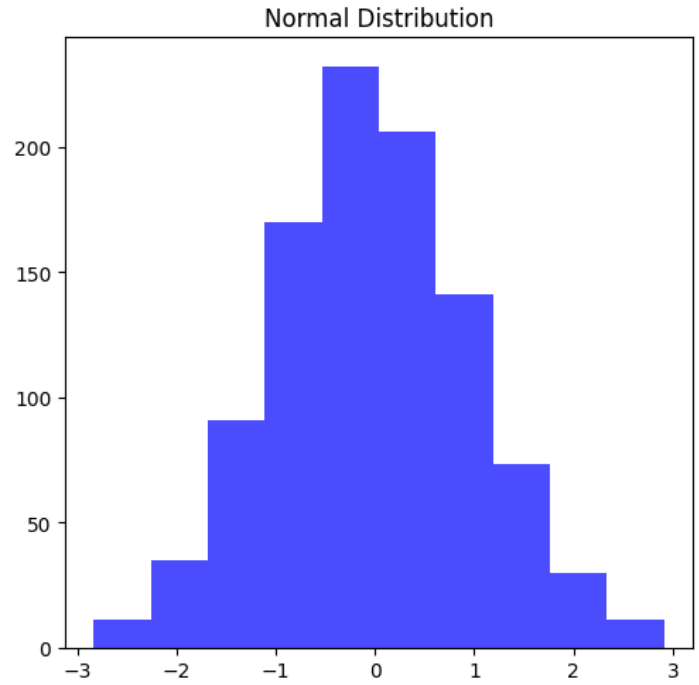
$$r = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y}$$

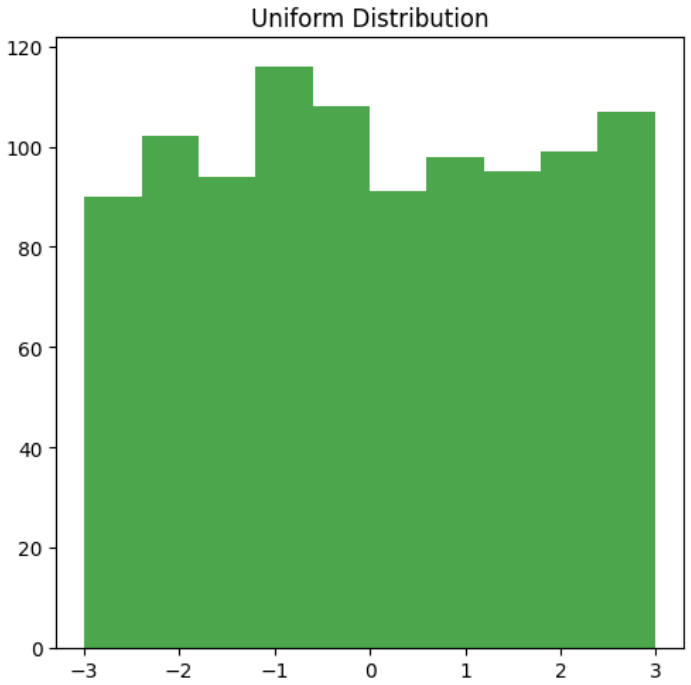
Understanding Data

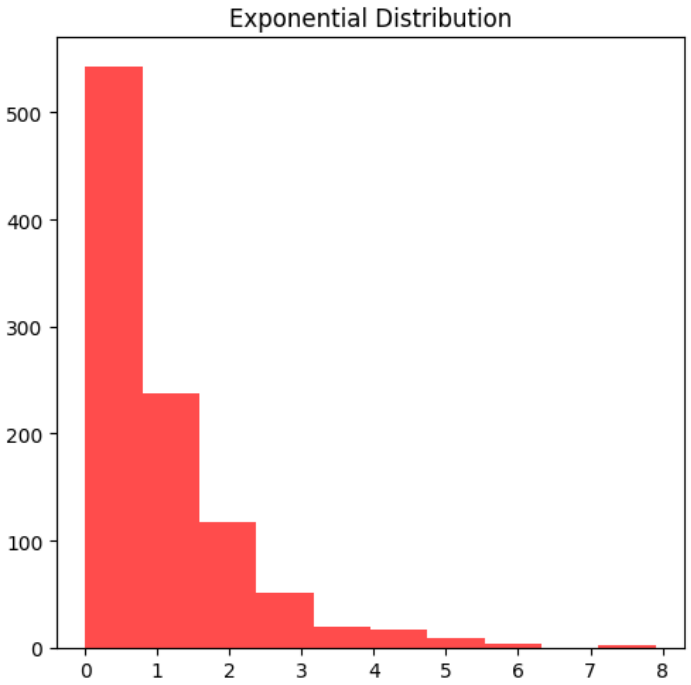
```
mean = np.mean(data)
variance = np.var(data, ddof = 1) #Denominator of N-1, sample var
std_dev = np.std(data)
median = np.median(data)
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
cov_matrix = np.cov(data_X, data_Y)
corr_matrix = np.corrcoef(data_X, data_Y)

from scipy.stats import iqr, skew
iqr(data)
skew(data)
```

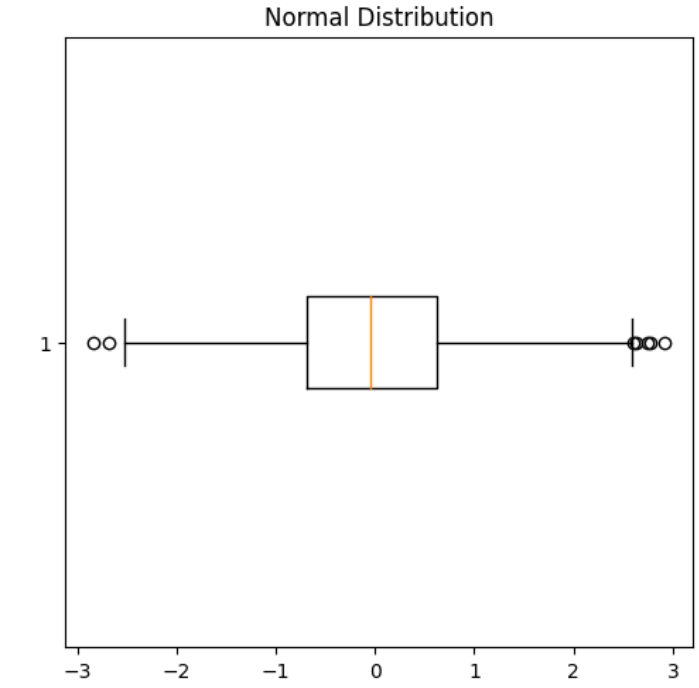
- `import matplotlib.pyplot as plt`
- `plt.hist(data)`
- `plt.boxplot(data)`

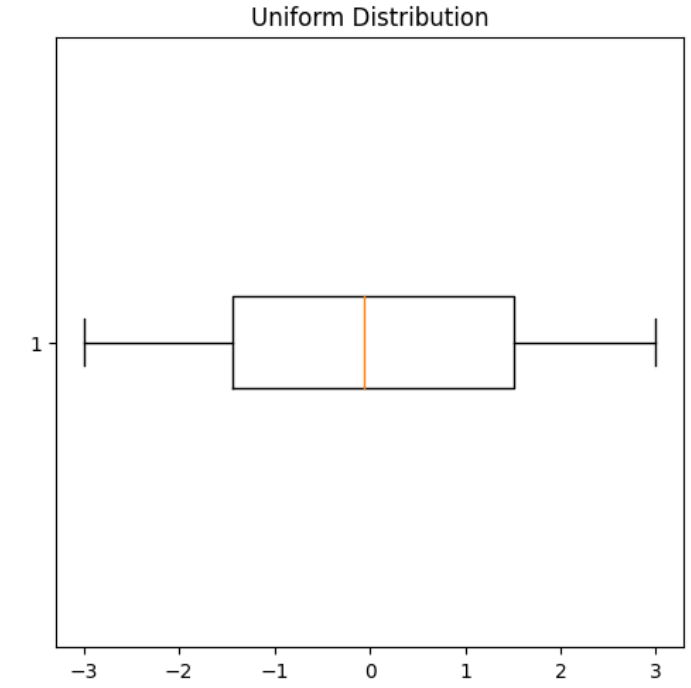
Understanding Data

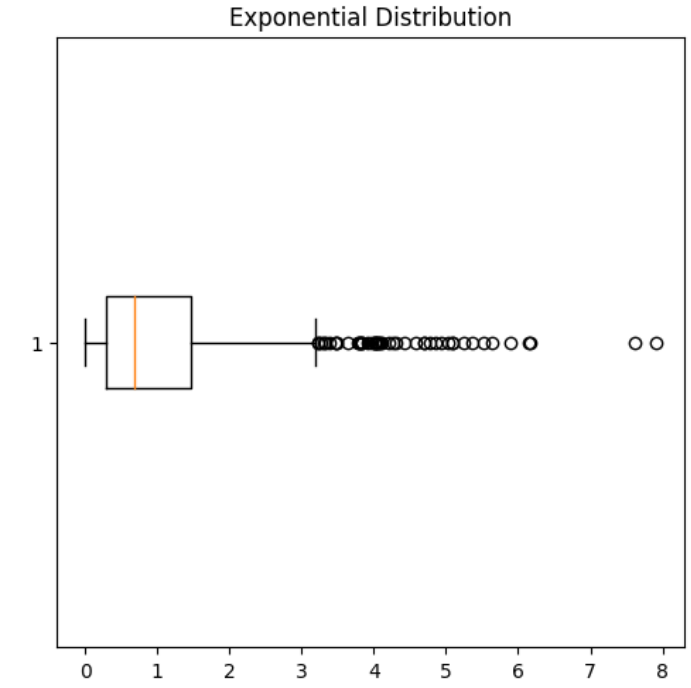












THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Back-Up

Resources

- Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for Machine Learning.
- Modern Mathematical Statistics with Applications Second Edition by Jay L. Devore and Kenneth N. Berk
- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, Aurelien Geron, 2019
- Introduction to Machine Learning (2014) by Ethem Alpayadin