Cheatsheets / **Statistics: Histograms**

Histograms

Matplotlib Function To Create Histogram

In Python, the `pyplot.hist()` function in the Matplotlib pyplot library can be used to plot a histogram. The function accepts a NumPy array, the range of the dataset, and the number of bins as input.

```
import numpy as np
from matplotlib import pyplot as plt

# numpy array
data_array = np.array([1,1,1,1,1,2,3,3,3,4,4,5,5,6,7])

# plot histogram
plt.hist(data_array, range = (1,7), bins = 7)
```

Mean of a Dataset

The *mean*, or average, of a dataset is calculated by adding all the values in the dataset and then dividing by the number of values in the set.
For example, for the dataset $[1,2,3]$, the mean is $1+2+3$ / 3 = 2 .

Histogram Bins

In a histogram, the range of the data is divided into sub-ranges represented by *bins*. The width of the bin is calculated by dividing the range of the dataset by the number of bins, giving each bin in a histogram the same width.

What is a Histogram?

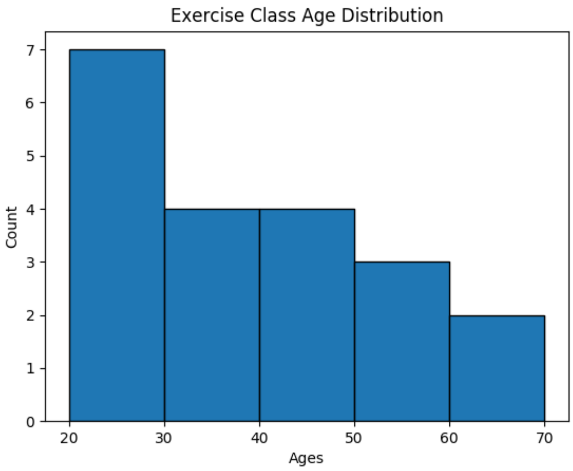
A *Histogram* is a plot that displays the spread, or distribution of a dataset. In a histogram, the data is split into intervals, called bins. Each bin shows the number of data points that are contained within that bin.

Histogram Bin Count

In a histogram, the bin *count* is the number of data points that fall within the bin's range.

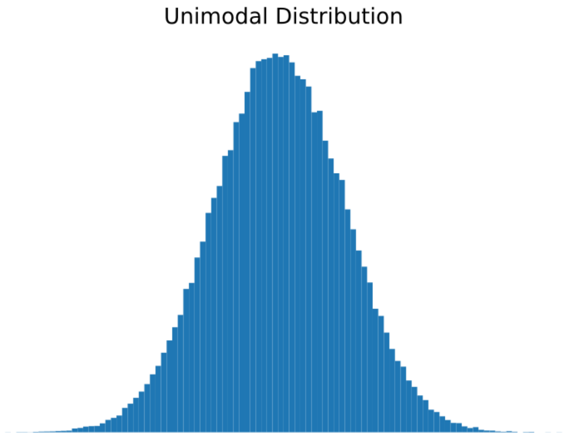
Histogram's X and Y Axis

A histogram is a graphical representation of the distribution of numerical data. In a histogram, the bin ranges are on the x-axis and the counts are on the y-axis.



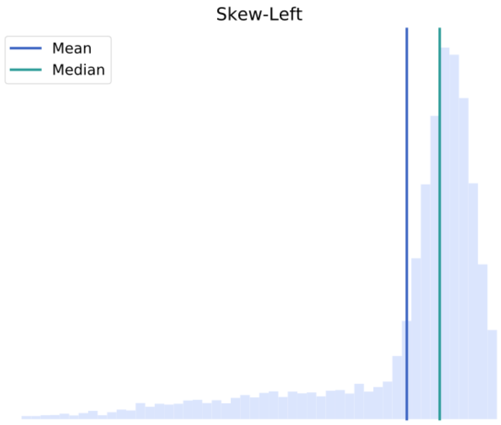
Unimodal Distribution

Modality describes the number of peaks in a dataset. A *unimodal* distribution in a histogram means there is one distinct peak indicating the most frequent value in a histogram.



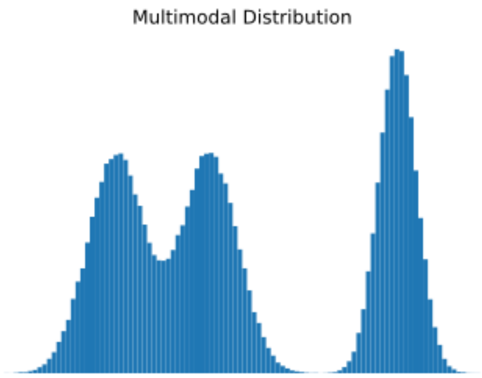
Left-Skewed Dataset

A left-skewed dataset has a long left tail with one prominent peak to the right. The median of this dataset is greater than the mean of this dataset.



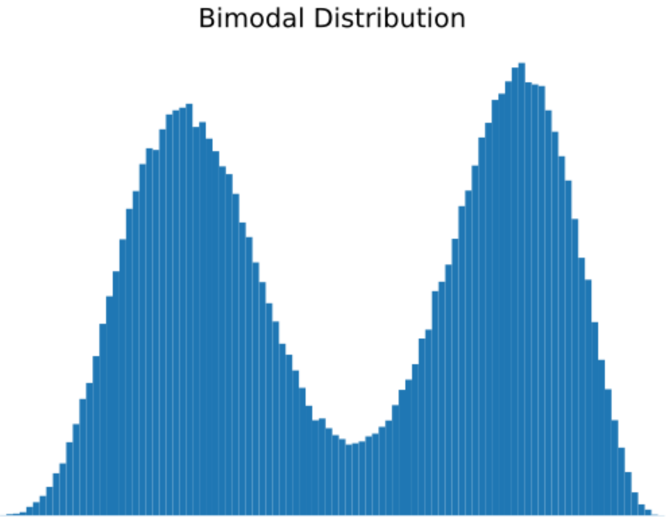
Multimodal Dataset

If a histogram has more than two peaks, then the dataset is referred to as *multimodal*.



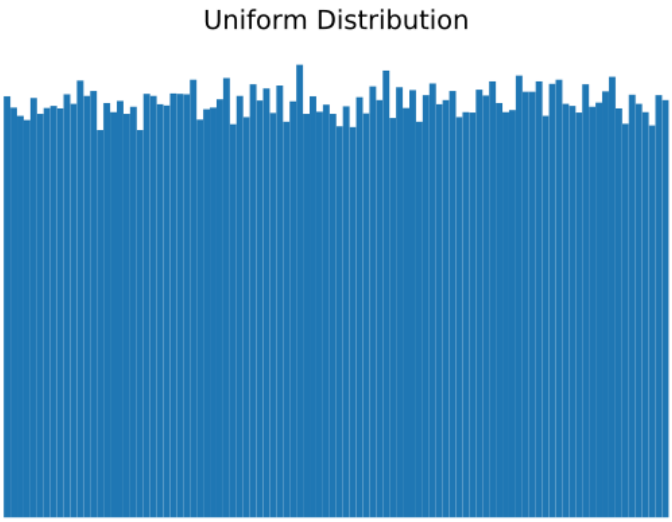
Bimodal Dataset

A bimodal dataset has two distinct peaks. This typically happens when the dataset contains two different populations.



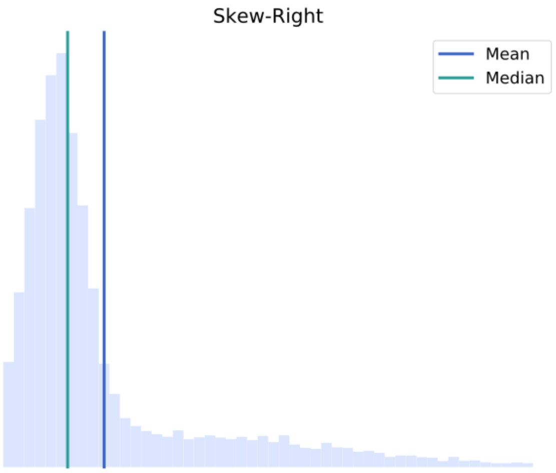
Uniform Dataset

A *uniform* dataset does not have any distinct peaks.
As seen in the histogram below, uniform datasets have approximately the same number of values in each group represented by a bar - there is no obvious clustering.



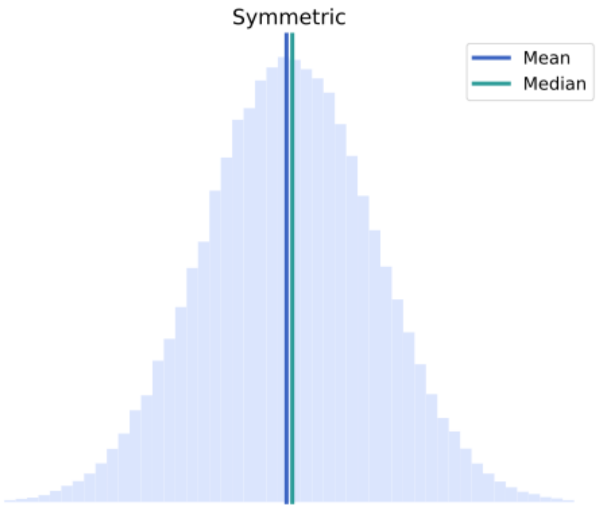
Right-skewed Dataset

In a histogram, if the prominent peak lies to the left with the tail extending to the right, then it is called a *right-skewed* dataset. In this case, the median is less than the mean of the dataset.



Symmetric Distribution in Histogram

In a histogram, the distribution of the data is symmetric if it has one prominent peak and equal tails to the left and the right. The Median and the Mean of a symmetric dataset are similar.



Dataset Outliers

An *outlier* is a data point that differs significantly from the rest of the values in a dataset.
For example, in the dataset $[1, 2, 3, 4, 100]$ the value 100 is an outlier because it lies a large distance from the rest of the data.

Spread of a Dataset

The spread of a dataset is the dispersion from the dataset's center. The descriptive statistics that describe the spread are range, variance and standard deviation.
For example, for the dataset $[1, 4, 7, 10]$, the *range* of the dataset would be the maximum value of the set - the minimum value of the set, or $10 - 1 = 9$.

codecademy

Peak of Unimodal Distribution

The center of a dataset is the peak of a unimodal distribution. The statistics that describe the center of a dataset are the mean and median.

Print Share ▾