Statistics

Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. It provides tools and methods for understanding and making inferences from data to aid in decision-making and understanding patterns and relationships in various fields.

Data

Data comes from many sources: sensor measurements, events, text, images, and videos. The Internet of Things (IoT) is spewing out streams of information. Much of this data is unstructured: images are a collection of pixels with each pixel containing RGB (red, green, blue) color information. Texts are sequences of words and nonword characters, often organized by sections, subsections, and so on. Clickstreams are sequences of actions by a user interacting with an app or web page. In fact, a major challenge of data science is to harness this torrent of raw data into actionable information. To apply the statistical concepts covered in this book, unstructured raw data must be processed and manipulated into a structured form — as it might emerge from a relational database — or be collected for a study

Here are some types of data:

Numerical Data:

Numerical data consists of numerical values and can be further classified into two subtypes:

- a. Continuous Data: Continuous data can take on any value within a certain range. For example, the height of individuals, temperature readings, or stock prices.

Examples

heights = [165.2, 170.5, 162.7, 175.9, 168.4, 173.1, 169.8, 166.3, 171.6, 167.2]

- b. Discrete Data: Discrete data can only take on specific, separate values. Examples include the number of students in a classroom, the number of cars in a parking lot, or the number of goals scored in a soccer game.

Examples

numberOfchildren= [2, 1, 0, 3, 1, 2, 0, 2, 1, 1, 4, 2, 3, 1, 0]

Categorical Data:

Categorical data represents qualities or characteristics and is divided into different categories or groups. It can be further classified into two subtypes:

- a. Nominal Data: Nominal data consists of categories with no inherent order or ranking. Examples include gender (male/female), marital status (single/married/divorced), or types of cars (sedan/sports/utility).

Examples

color = [Red, Blue, Green, Blue, Red, Yellow, Green, Red, Blue, Green, Red]

- b. Ordinal Data: Ordinal data represents categories with a natural order or ranking. Examples include educational levels (elementary school/middle school/high school/college), customer satisfaction ratings (poor/fair/good/excellent), or income brackets (low/middle/high).

Examples

rating = [4, 3, 5, 2, 4, 3, 2, 5, 3, 4]

#

Time Series Data:

Time series data is collected at regular intervals over time. It enables the analysis of trends, patterns, and changes over a specific period. Examples include daily stock prices, monthly sales figures, or annual rainfall measurements.

Examples

monthly_profit = {("January", 10000),("February", 12000),("March", 15000), ("April", 13500), ("May", 11800),("June", 14200), ("July", 16500),("August", 18000),("September", 14500), ("October", 16200),("November", 19500),("December", 22000)}

#

Cross-Sectional Data:

Cross-sectional data is collected from different individuals, subjects, or items at a single point in time. It provides a snapshot view of a population or sample at a specific moment. Examples include survey responses from different participants, demographic data from a particular year, or data collected from multiple products.

Examples

review1 = {"Customer1", 5, "Excellent"} review2 = {"Customer2", 3, "Average"} review3 = {"Customer3", 4, "Good"} review4 = {"Customer4", 2, "Poor"} review5 = {"Customer5", 4, "Good"}

#

Geospatial Data:

Geospatial data refers to information that is associated with specific geographic locations. It can include coordinates, addresses, or boundaries. Examples include GPS coordinates, maps, or satellite images.

Examples

City1 = { "Latitude": "40.7128° N", "Longitude": "74.0060° W", "Population Density": "10,000 people/km²" }

City2 ={ "Latitude": "34.0522° N", "Longitude": "118.2437° W", "Population Density": "8,000 people/km²" }

City3 = { "Latitude": "51.5074° N", "Longitude": "0.1278° W", "Population Density": "12,000 people/km²" }

#

Textual Data:

Textual data comprises unstructured text, such as emails, social media posts, articles, or customer reviews. Analyzing textual data involves techniques like natural language processing (NLP) to extract insights and sentiment analysis.

Examples

image = { "Image1" : "A photo of a cat", "Image2" : "A photo of a dog", "Image3" : "A photo of a bird", "Image4" : "A photo of a horse" }

Descriptive statistics

Descriptive statistics involves organizing, summarizing, and describing the main features of a dataset. It provides a way to understand and present data in a meaningful and concise manner.

Measures of Central Tendency(First dimension):

An estimate of where most of the data is located.

- Mean: The arithmetic average of a set of numerical values. The most basic estimate of location is the mean, or average value. The mean is the sum of all the values divided by the number of values. Consider the following set of numbers: {5, 7, 8, 7, 10, 12}. The mean is ([5 + 7 + 8 + 7 + 10 + 12]) / 6 = 8.167. The formula to compute the mean of a sample from a population.

$$\text{Mean} = \bar{x} = \frac{\sum_i^n x_i}{n}$$

Where

- x_bar = sample mean
- x_i = data point
- n = number of data point #

- Median: The median is the middle number on a sorted list of the data. If there is an even number of data values, the middle value is one that is not actually in the data set, but rather the average of the two values that divide the sorted data into upper and lower halves.
  - For example, the median of [5, 7, 8, 7, 10, 12] is 7.5.

Compared to the mean, which uses all observations, the median depends only on the values in the center of the sorted data. While this might seem to be a disadvantage, since the mean is much more sensitive to the data, there are many instances in which the median is a better metric for location. there is advantage, median is robust to outlier

- Mode: The mode is the most frequently occurring value(s) in a dataset. It represents the value(s) with the highest frequency. A dataset can have no mode, one mode (unimodal), or multiple modes (multimodal).
  - For example, the mode of [5, 7, 8, 7, 10, 12] is 7.

In [1]:
```
# code implementation
import statistics

# Example data
data = [5, 7, 8, 7, 10, 12]

# Mean
mean = statistics.mean(data)
print("Mean:", round(mean,3))

# Median
median = statistics.median(data)
print("Median:", median)

# Mode
mode = statistics.mode(data)
print("Mode:", mode)
```

```
Mean: 8.167
Median: 7.5
Mode: 7
```

Measures of Variability(Second dimension):

- Deviations The most widely used estimates of variation are based on the differences, or deviations, between the estimate of location and the observed data. For a set of data {5, 7, 8, 7, 10, 12}, the mean is 8.167 and the median is 7.5. The deviations from the mean are the differences:
  - 5 - 8.167 = -3.167
  - 7 - 8.167 = -1.167
  - 8 - 8.167 = -0.167
  - 7 - 8.167 = -1.167
  - 10 - 8.167 = 1.833
  - 12 - 8.167 = 3.833 These deviations tell us how dispersed the data is around the central value.

#

- Variance: A measure of how spread out the values in a dataset are from the mean. It quantifies the average squared deviation from the mean.

$$\text{Variance} = s^2 = \frac{\sum (x - \overline{x})^2}{n - 1}$$

where

```
* x = data_point
* x_bar = sample_mean
* n-1 = degree of freedom


* For example, the Variance of [5, 7, 8, 7, 10, 12] is 6.167.
```

- Standard Deviation: The square root of the variance. It measures the dispersion of values around the mean.
  - For example, the standard deviation of [5, 7, 8, 7, 10, 12] is 2.483.
- Range: The difference between the maximum and minimum values in a dataset.
  - For example, the range of [5, 7, 8, 7, 10, 12] is 12 - 5 = 7. #
- Percentiles:

Percentiles represent the value below which a given percentage of the data falls. For example, the 75th percentile is the value below which 75% of the data falls. Percentiles help understand the relative position of a particular value within a dataset.

- Interquartile range:

The difference between the 75th percentile and the 25th percentile.

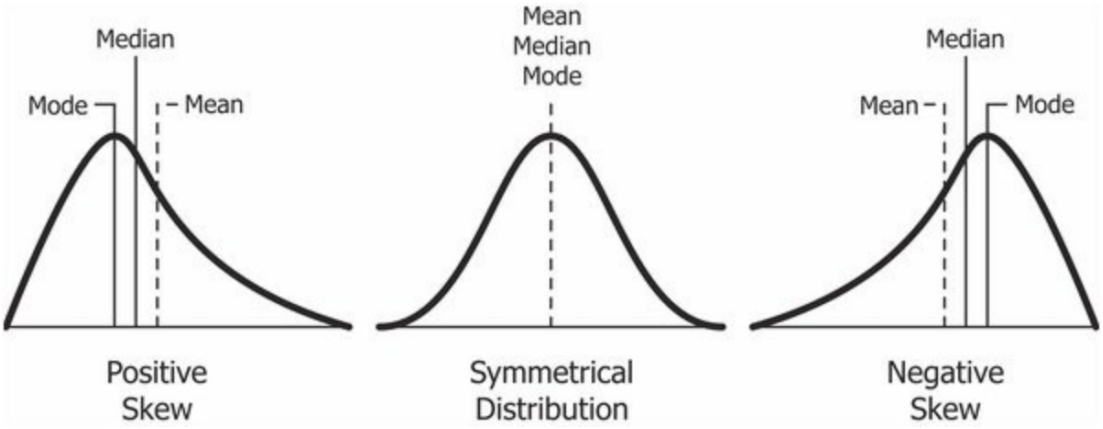
In [2]:
```
import numpy as np
# Range
data_range = max(data) - min(data)
print("Range:", data_range)

# Variance
variance = statistics.variance(data)
print("Variance:", round(variance,3))

# Standard Deviation
std_deviation = statistics.stdev(data)
print("Standard Deviation:", round(std_deviation,3))

# Calculate percentiles
p25 = np.percentile(data, 25)
p50 = np.percentile(data, 50)
p75 = np.percentile(data, 75)

print("25th percentile:", p25)
print("50th percentile:", p50)
print("75th percentile:", p75)

# Calculate Interquartile range
interquartile_range = p75-p25
print("Interquartile range:",interquartile_range)
```

```
Range: 7
Variance: 6.167
Standard Deviation: 2.483
25th percentile: 7.0
50th percentile: 7.5
75th percentile: 9.5
Interquartile range: 2.5
```

Skewness (Third dimension)

Skewness is a statistical measure that describes the asymmetry or lack of symmetry in the distribution of a dataset. It provides insights into the shape and nature of the distribution.

- Positive Skewness: Also known as right skewness, it occurs when the tail of the distribution extends towards the right, and the majority of the data is concentrated towards the left. The mean is usually greater than the median, and the skewness value is positive.
- Negative Skewness: Also known as left skewness, it occurs when the tail of the distribution extends towards the left, and the majority of the data is concentrated towards the right. The mean is usually less than the median, and the skewness value is negative.
- Zero Skewness: In a perfectly symmetric distribution, the skewness is zero, indicating a balanced distribution where the mean and median are equal, and the data is evenly spread around the central tendency.

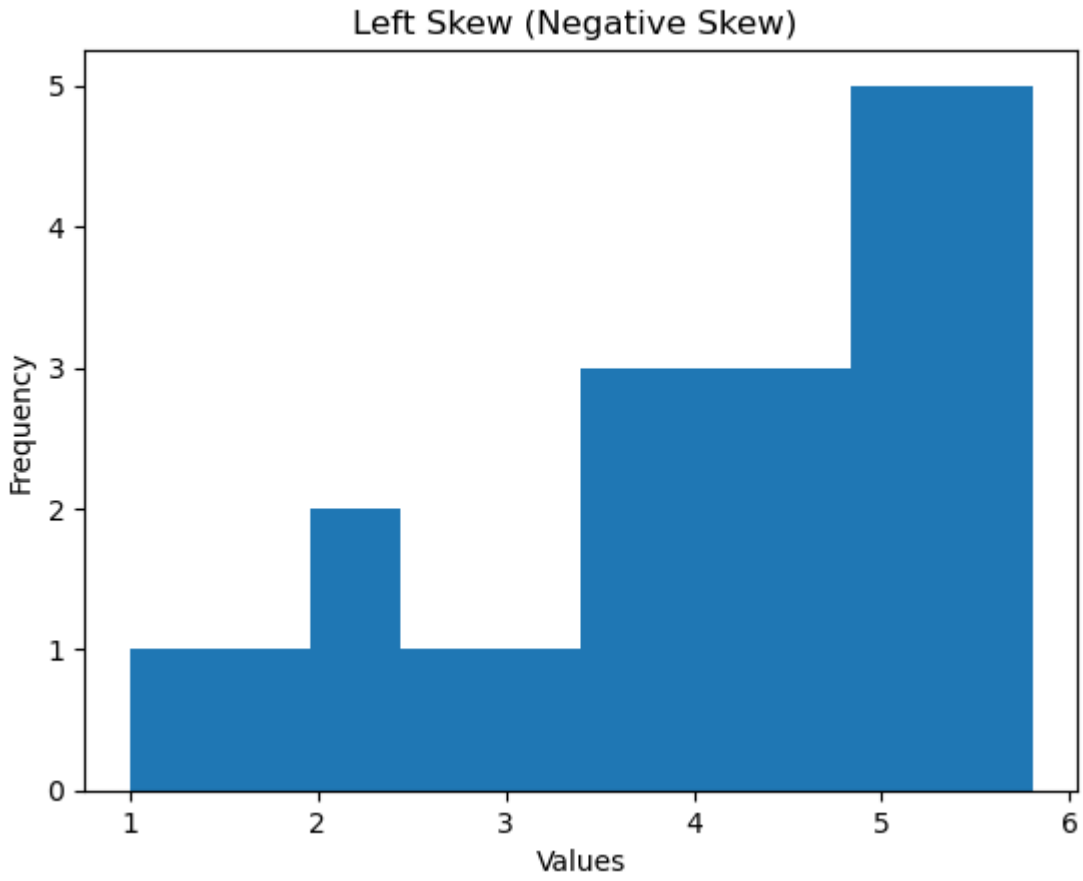


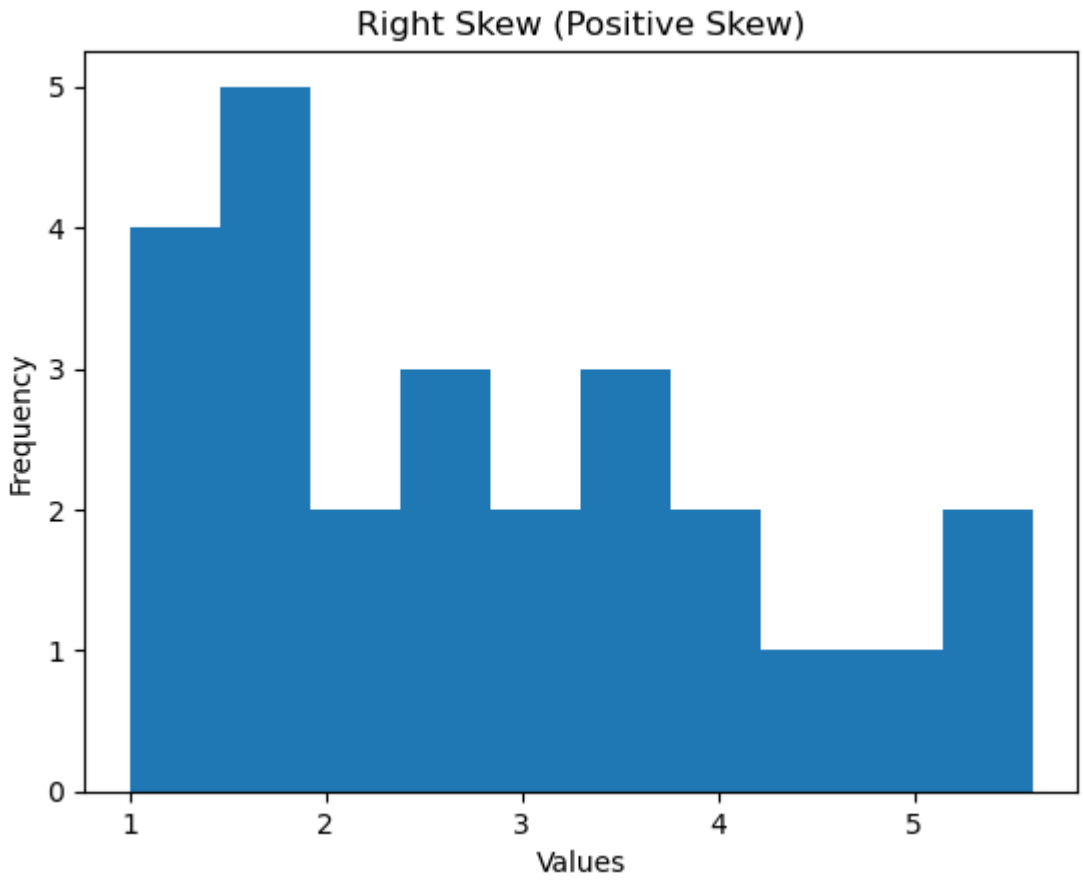
https://en.wikipedia.org/wiki/Skewness

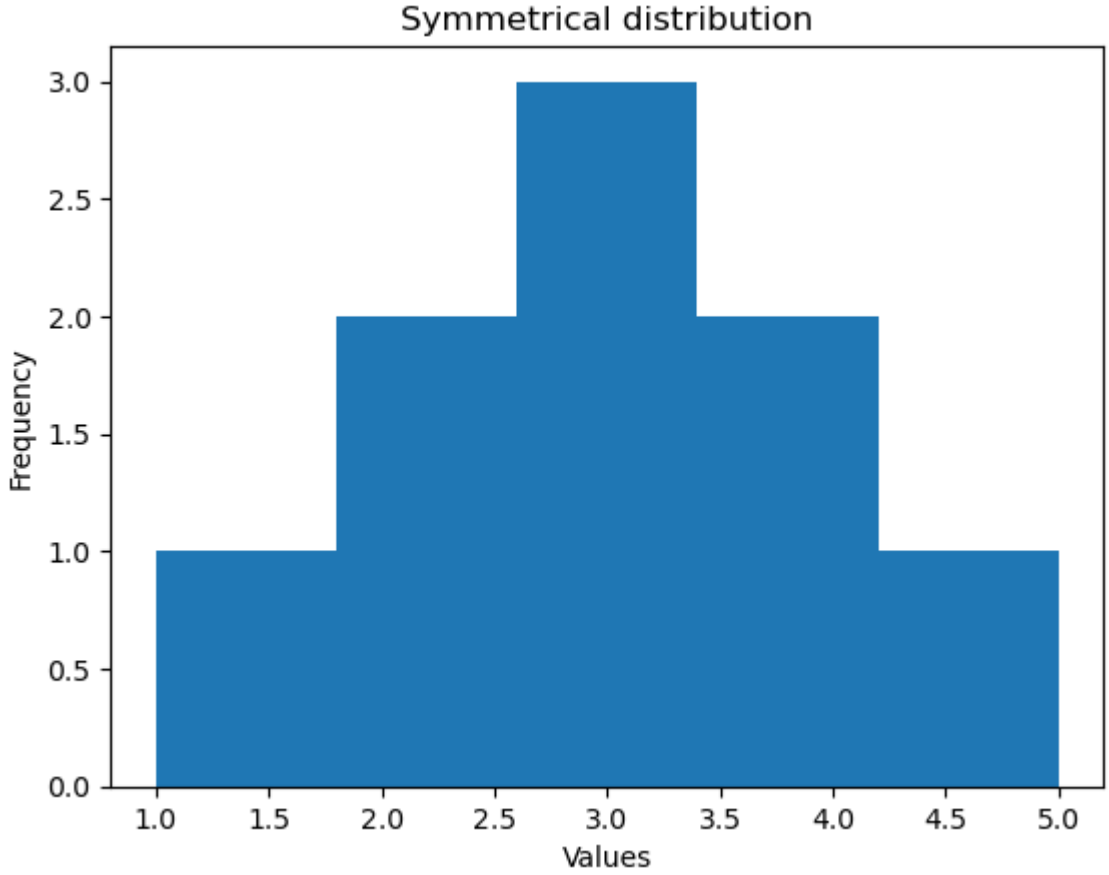
In [3]:
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import skew
# Calculate skewness
def skewkness(data):
    data_skewness = skew(data)
    print("Skewness:", data_skewness)

    if data_skewness< 0:
        print("Left skew distribution")
    elif data_skewness > 0 :
        print("Right skew distribution")
    else:
        print("Symmetrical distribution")
```

In [4]:
```
# Example data
data1 = np.array([1.0, 1.5, 2.0, 2.4, 2.8, 3.2, 3.4, 3.6, 3.8, 4.0, 4.15, 4.3,
                  4.45, 4.6, 4.75,4.9,5.0, 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8])

data2 = np.array([1.0, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7,1.8, 1.9,2.0, 2.2, 2.4, 2.6,
                  2.8, 2.9, 3.1, 3.3, 3.5, 3.7, 3.9, 4.1, 4.4, 4.8, 5.2, 5.6])

data3 = np.array([1, 2, 2, 3, 3, 3, 4, 4, 5])

# Right Skew (Positive Skew)
plt.hist(data1)
plt.xlabel("Values")
plt.ylabel("Frequency")
plt.title("Left Skew (Negative Skew) ")#
plt.show()

# Left Skew (Negative Skew)
plt.hist(data2)
plt.xlabel("Values")
plt.ylabel("Frequency")
plt.title("Right Skew (Positive Skew)")#
plt.show()

# Symmetrical distribution
plt.hist(data3,bins=5)
plt.xlabel("Values")
```

```
plt.ylabel("Frequency")
plt.title("Symmetrical distribution")#
plt.show()


result1 = skewkness(data1)
result2 = skewkness(data2)
result3 = skewkness(data3)
```




Right Skew (Positive Skew)
Frequency
Values
0
1
2
3
4
5

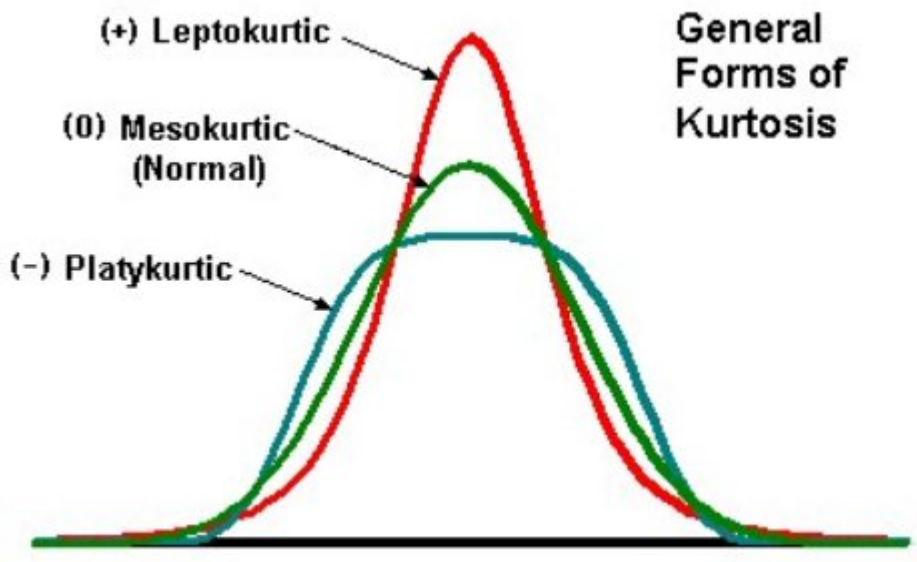

Symmetrical distribution
Frequency
Values
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

```
Skewness: -0.796938685841155
Left skew distribution
Skewness: 0.5368500349438866
Right skew distribution
Skewness: 0.0
Symmetrical distribution
```

Kurtosis (Fourth dimension):

Kurtosis is a statistical measure that describes the shape and peakedness of a probability distribution. It provides insights into whether the distribution is relatively flat or peaked compared to the normal distribution.

- Mesokurtic: A mesokurtic distribution has kurtosis equal to 0. This means that the distribution has a shape similar to the normal distribution, with moderate peakness and tail behavior.
- Leptokurtic: A leptokurtic distribution has positive kurtosis. It indicates a distribution that is more peaked and has heavier tails compared to the normal distribution. This suggests that the distribution has more extreme values or outliers.
- Platykurtic: A platykurtic distribution has negative kurtosis. It indicates a distribution that is less peaked and has lighter tails compared to the normal distribution. This suggests that the distribution has fewer extreme values and is more dispersed.



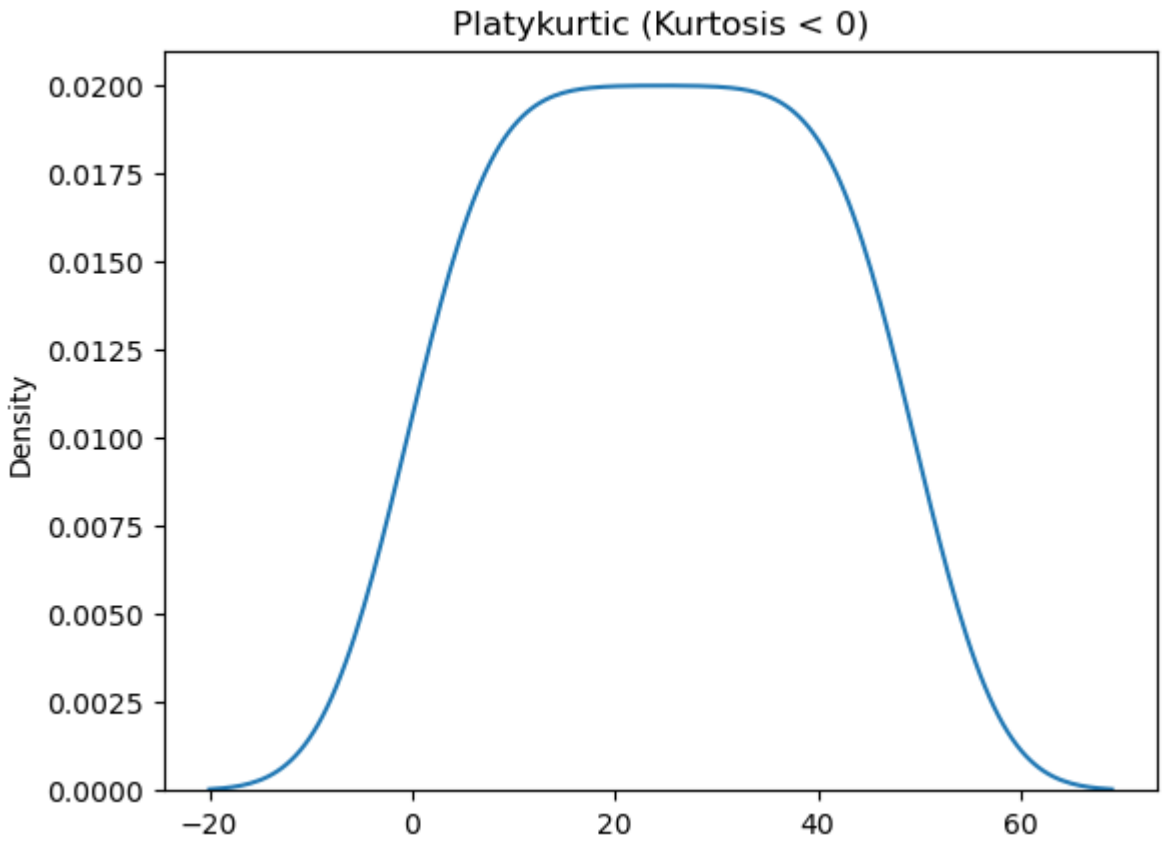
https://en.wikipedia.org/wiki/Kurtosis

In [5]:
```python
from scipy.stats import kurtosis
# Calculate kurtosis
def calculate_kurtosis(data):
    data_kurtosis = kurtosis(data)
    print("Kurtosis:", data_kurtosis)

    if data_kurtosis < -1:
```

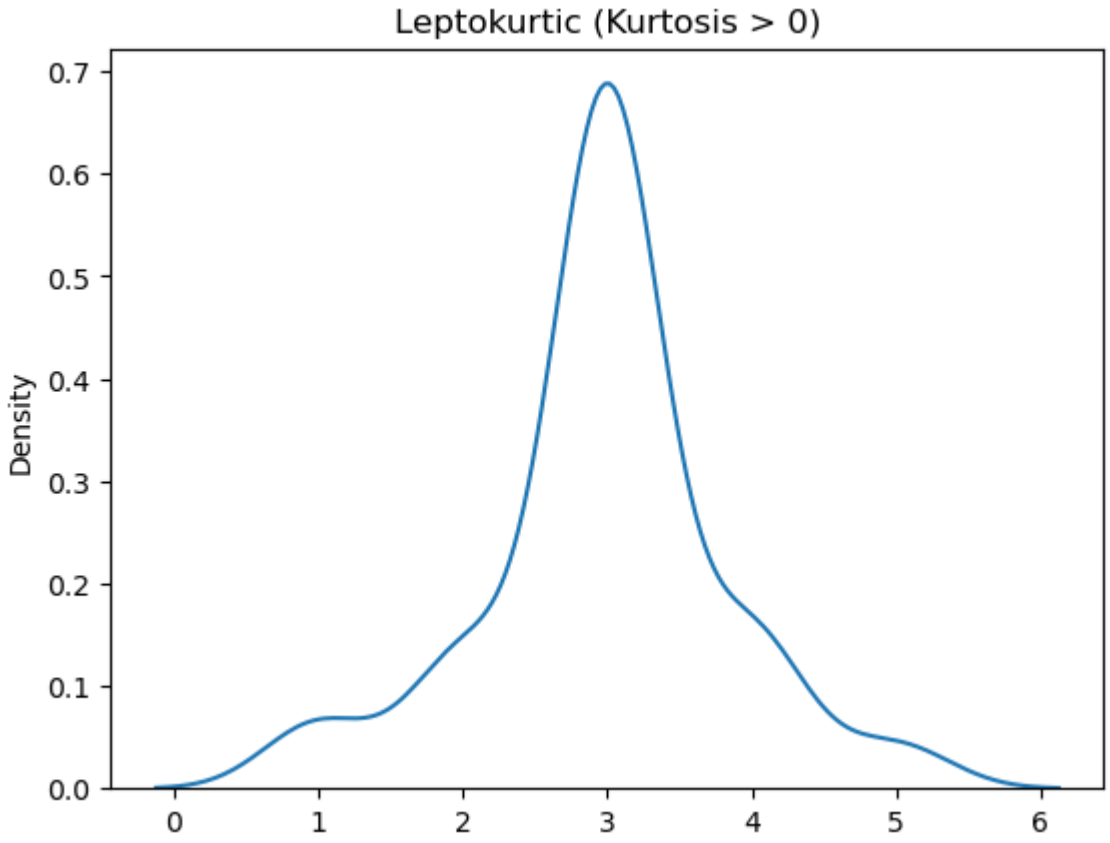
```
        print("platykurtic (lighter tail) ")

    elif data_kurtosis > 1 :
        print("leptokurtic (heavier tail)")
    else:
        print("Mesokurtic distribution")

data1 = np.array(list(range(50)))
data2 = np.array([1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3,
                  3, 3,3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
                  3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5])

result1 = calculate_kurtosis(data1)
result2 = calculate_kurtosis(data2)
result3 = calculate_kurtosis(data3)
```

```
Kurtosis: -1.2009603841536614
platykurtic (lighter tail)
Kurtosis: 1.2530497009967272
leptokurtic (heavier tail)
Kurtosis: -0.75
Mesokurtic distribution
```

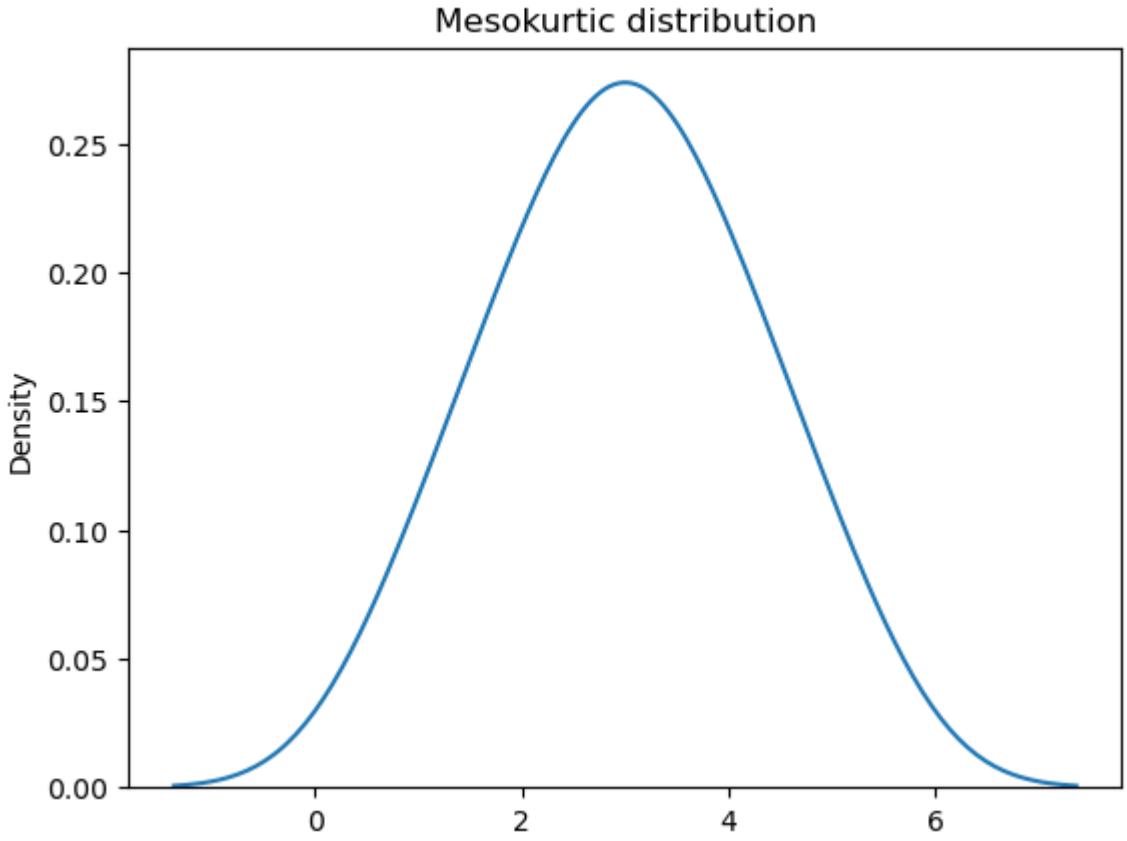
In [6]:
```
sns.kdeplot(data1)
plt.title('Platykurtic (Kurtosis < 0)')
plt.show()
```



In [7]:
```
sns.kdeplot(data2)
plt.title('Leptokurtic (Kurtosis > 0)')
plt.show()
```



In [8]:
```
sns.kdeplot(data3)
plt.title('Mesokurtic distribution')
plt.show()
```



Box plot

A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It provides a summary of the data's central tendency, variability, and any potential outliers. A box plot consists of several key elements:

Median (Q2): The line inside the box represents the median, which is the middle value of the dataset when it is sorted in ascending order.

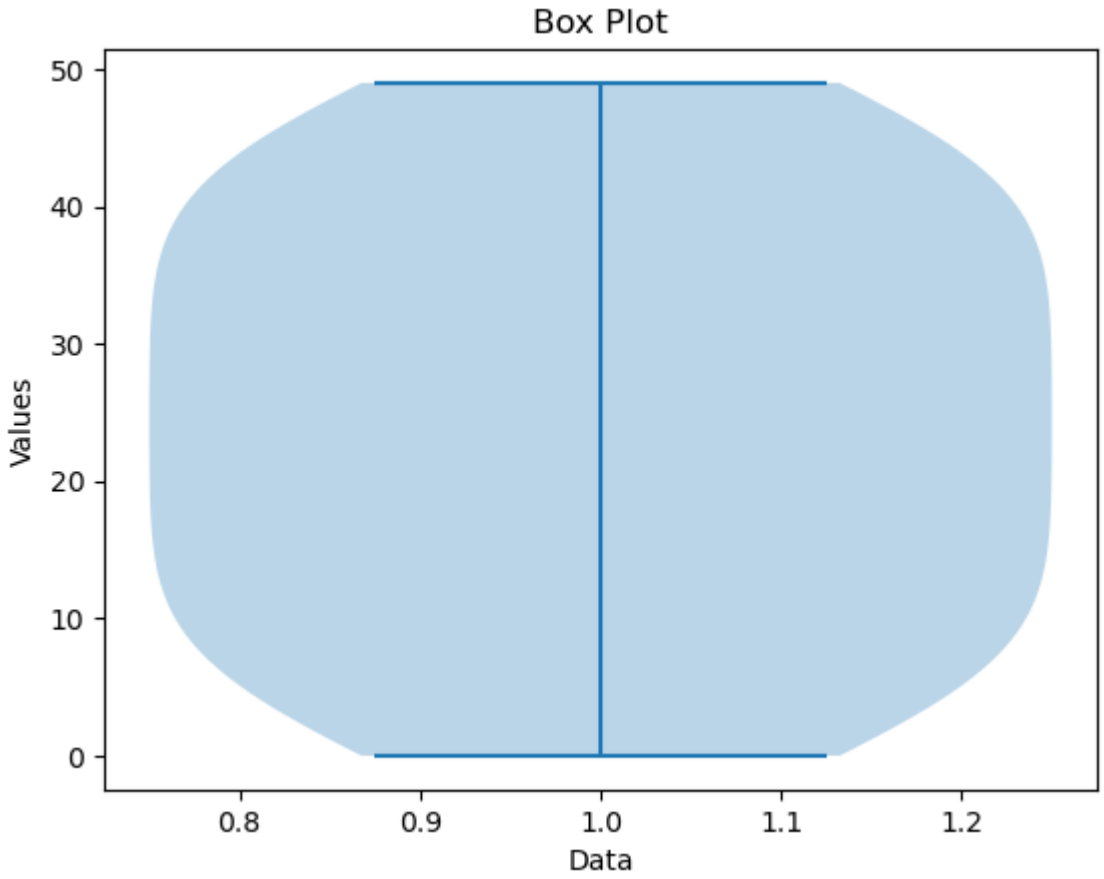
Quartiles (Q1 and Q3): The box is divided into two parts by a horizontal line at the median. The lower boundary of the box represents the first quartile (Q1), which is the median of the lower half of the data. The upper boundary represents the third quartile (Q3), which is the median of the upper half of the data.

Interquartile Range (IQR): The IQR is the range between the first and third quartiles (IQR = Q3 - Q1). It provides a measure of the spread of the middle 50% of the data.

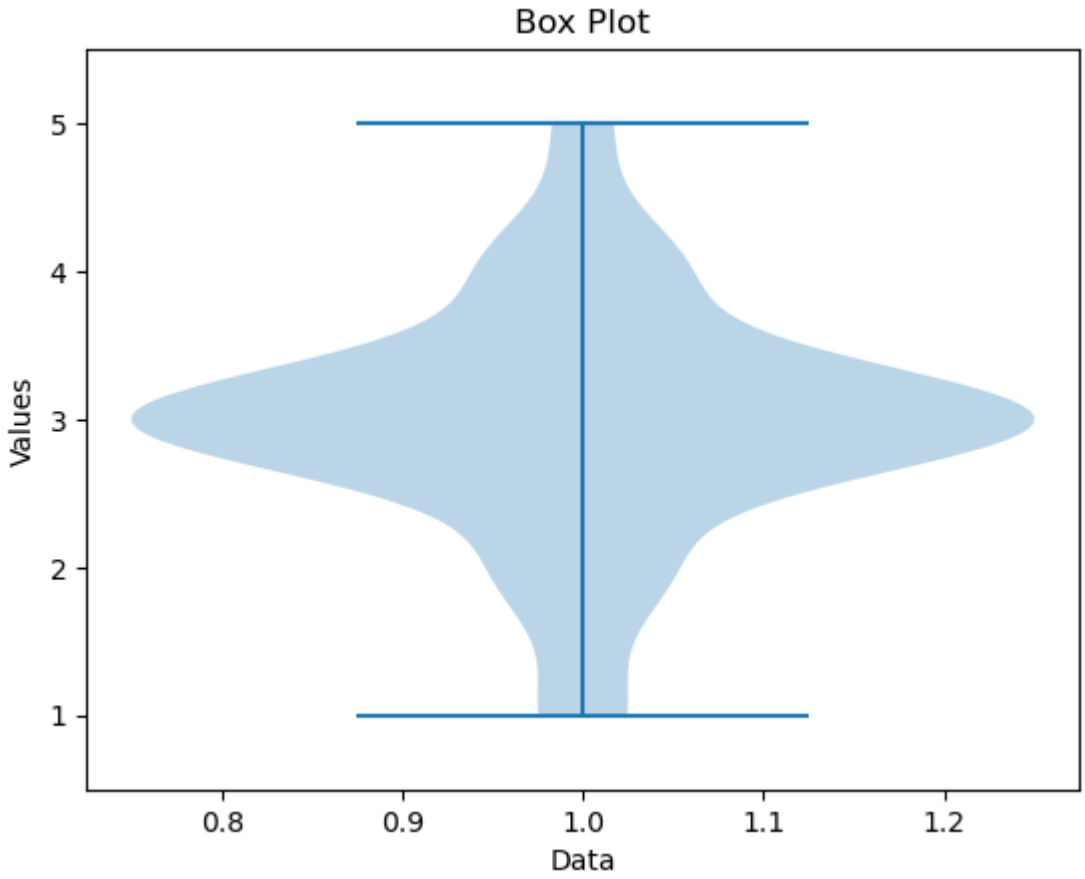
Whiskers: The lines extending from the box represent the minimum and maximum values within 1.5 times the IQR from the first and third quartiles, respectively. Any data points outside this range are considered outliers and displayed as individual points.

Outliers: Outliers are data points that fall outside the whiskers and are represented as individual points on the plot.

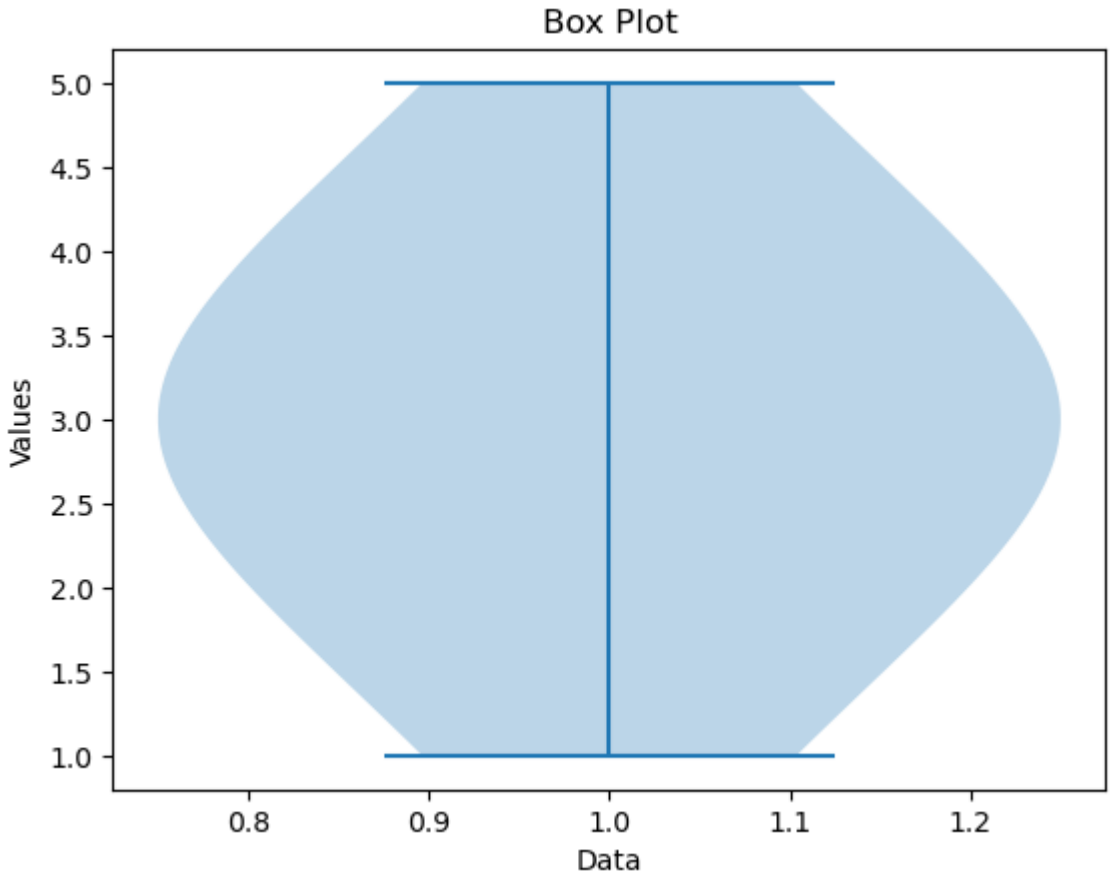
In [9]:
```
# Example 1 :
plt.violinplot(data1)

# Add labels and title
plt.xlabel('Data')
plt.ylabel('Values')
plt.title('Box Plot')

# Display the plot
plt.show()
```



In [10]:
```
# Example 2 :

plt.violinplot(data2)

# Add labels and title
plt.xlabel('Data')
plt.ylabel('Values')
plt.title('Box Plot')

plt.ylim(np.min(data2) - 0.5, np.max(data2) + 0.5)

# Display the plot
plt.show()
```



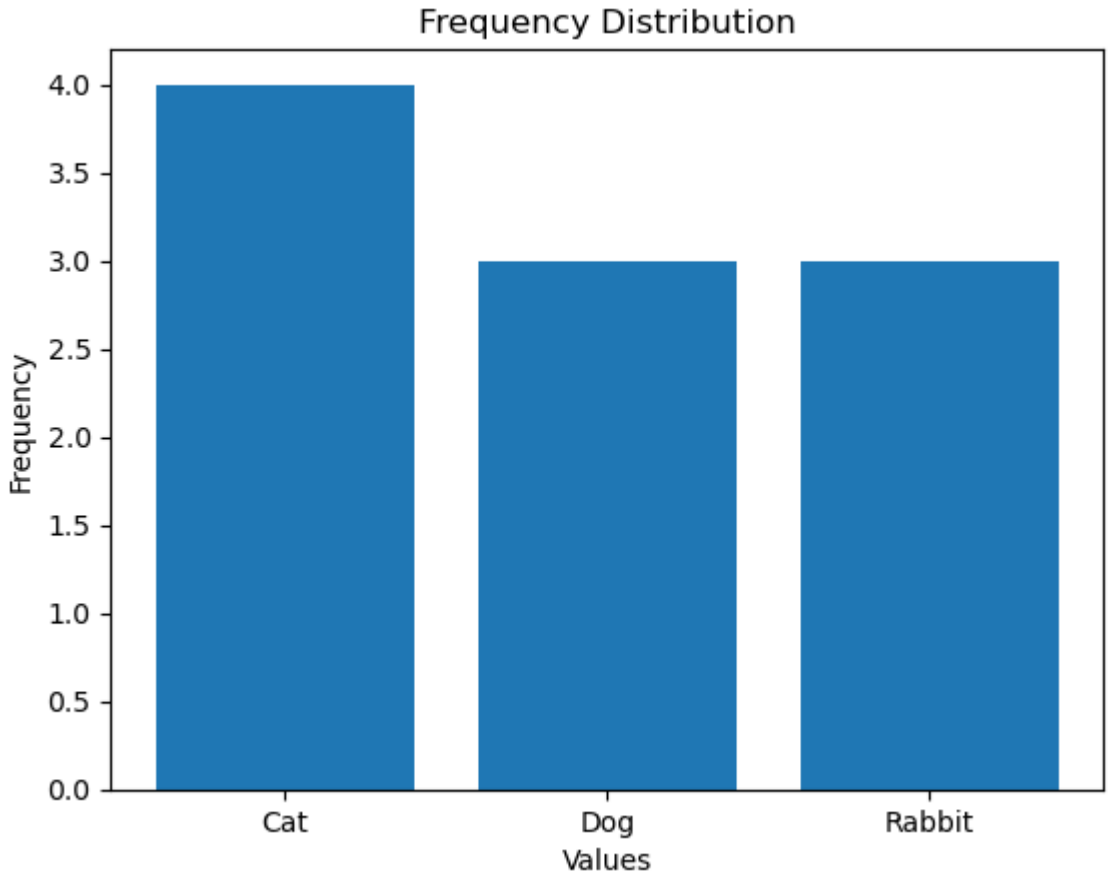
In [11]:
```
# Example 3 :
plt.violinplot(data3)

# Add labels and title
plt.xlabel('Data')
plt.ylabel('Values')
plt.title('Box Plot')

# Display the plot
plt.show()
```



Frequency Distribution:

Frequency distribution represents the count or proportion of each value in a dataset. It helps identify patterns and the distribution of values. For example, a frequency distribution table for the scores of students in a class may show the number of students who scored in different ranges, such as 0-10, 11-20, and so on.

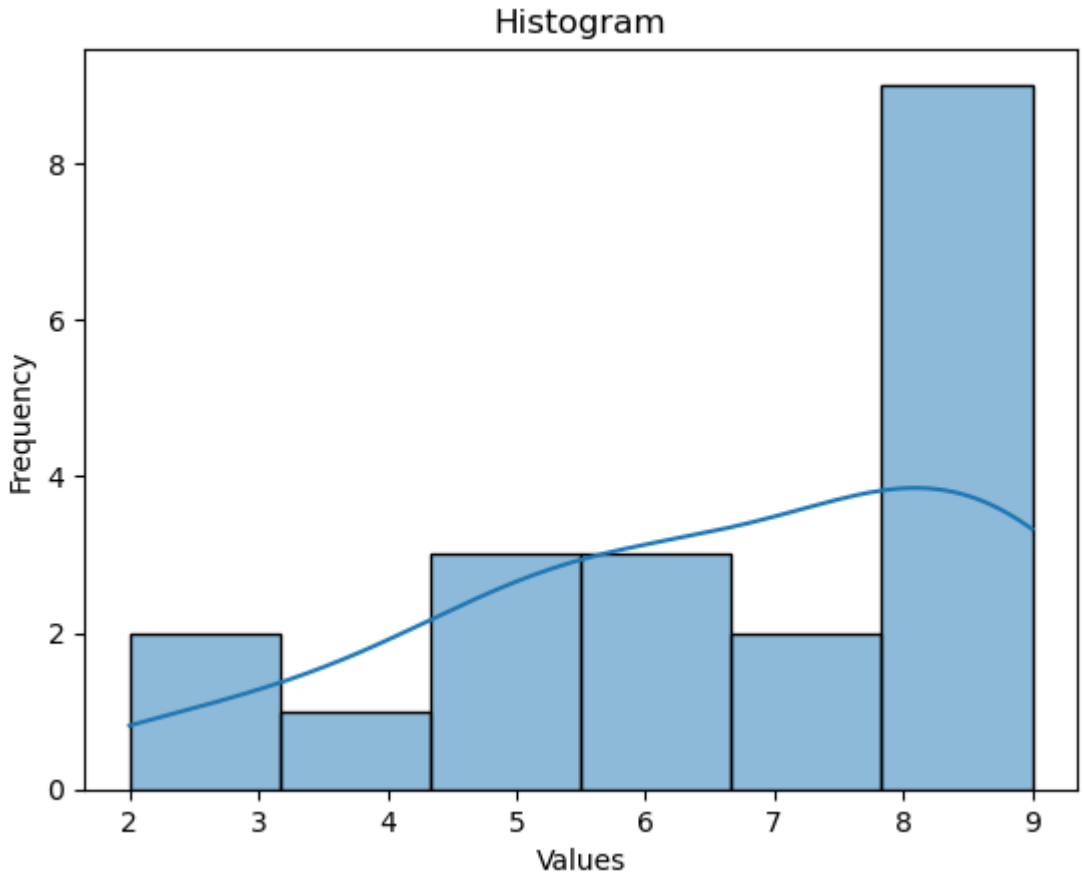
In [12]:
```
# sample data
data = np.array(['Cat', 'Dog', 'Cat', 'Dog', 'Cat', 'Rabbit', 'Dog', 'Rabbit', 'Rabbit'

# Calculate frequency counts
unique_values, counts = np.unique(data, return_counts=True)

# Plotting the frequency distribution
plt.bar(unique_values, counts)
plt.xlabel('Values')
plt.ylabel('Frequency')
plt.title('Frequency Distribution')
plt.show()
```

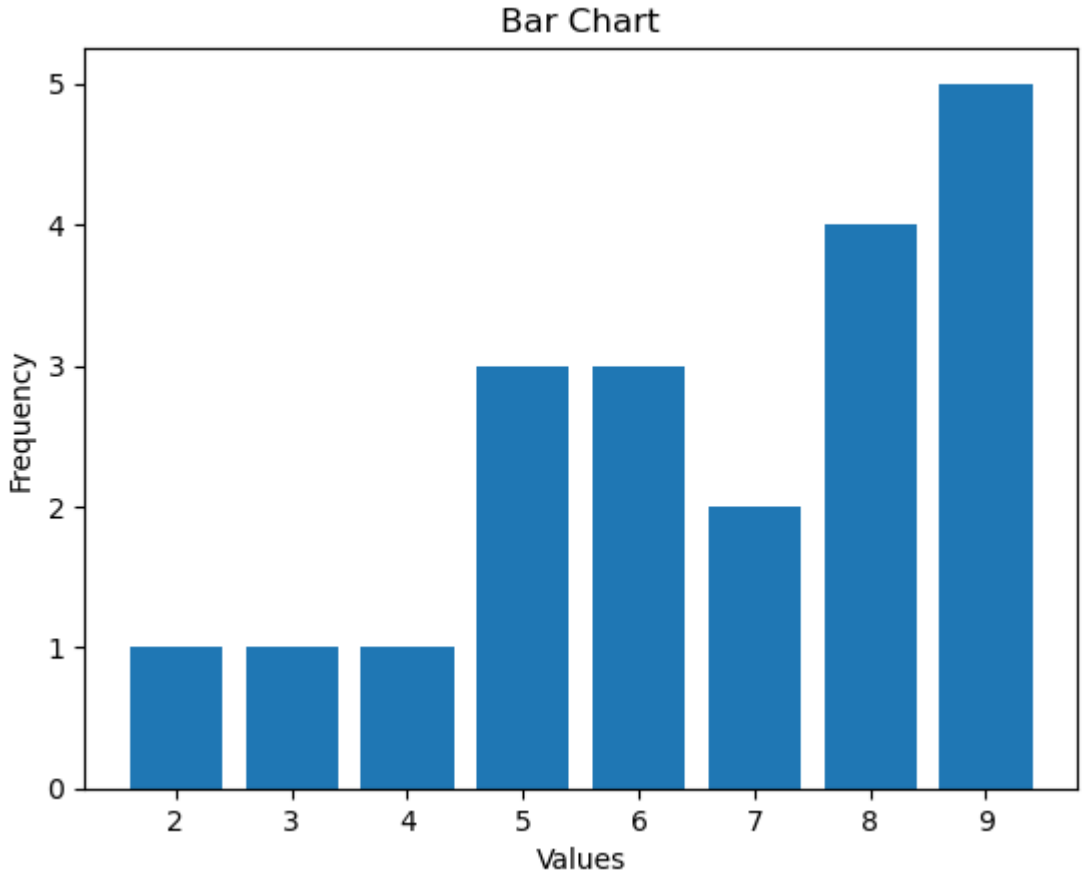


Histograms and Bar Charts:

Histograms and bar charts are graphical representations of frequency distributions. Histograms are used for numerical data, while bar charts are used for categorical data. They provide visual insights into the distribution and shape of the data.

In [13]:
```
# sample data
data = np.array([2, 3, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, 9])

# Plotting a histogram
sns.histplot(data, kde=True)
plt.xlabel('Values')
plt.ylabel('Frequency')
plt.title('Histogram')
plt.show()

# Calculate frequency counts
unique_values, counts = np.unique(data, return_counts=True)

# Plotting a bar chart
plt.bar(unique_values, counts)
plt.xlabel('Values')
plt.ylabel('Frequency')
plt.title('Bar Chart')
plt.show()
```





Cross-Tabulations and Contingency Tables:

Cross-tabulations and contingency tables are used to summarize and compare categorical data. They show the distribution of variables across different categories and help identify relationships and dependencies.

In [14]:
```
import pandas as pd

# Example data
data = {
    'Gender': ['Male', 'Female', 'Male', 'Female', 'Male', 'Male', 'Female', 'Female'],
    'Smoker': ['Yes', 'No', 'No', 'Yes', 'Yes', 'No', 'No', 'Yes'],
    'Count': [10, 20, 15, 25, 30, 35, 40, 45]
}

# Create a DataFrame
df = pd.DataFrame(data)

# Cross-tabulation
cross_tab = pd.crosstab(df['Gender'], df['Smoker'], values=df['Count'], aggfunc='sum')

print(cross_tab)
```

```
Smoker  No  Yes
Gender
Female  60   70
Male    50   40
```

Exploring Two or More Variables

Covariance

Covariance is a statistical measure that quantifies the relationship between two variables. It describes how changes in one variable correspond to changes in another variable. Covariance indicates the direction and magnitude of the linear relationship between two variables.

The formula for calculating the covariance between two variables X and Y is as follows:

$cov(X, Y) = \Sigma[(x_i - \mu_x) * (y_i - \mu_y)] / n$

Where:

- x_i and y_i are the individual data points in X and Y, respectively.
- μ_x and μ_y are the means (averages) of X and Y, respectively.
- n is the number of data points.

The covariance can take on positive or negative values. A positive covariance indicates a positive relationship, meaning that as one variable increases, the other tends to increase as well. A negative covariance indicates a negative relationship, meaning that as one variable increases, the other tends to decrease.

However, it is important to note that the magnitude of the covariance value alone does not provide a standardized measure of the strength of the relationship. It is influenced by the scales

of the variables, making it difficult to compare across different datasets.

To address this issue, the correlation coefficient is often used as it provides a standardized measure of the linear relationship between two variables, ranging from -1 to 1.

Correlation

Correlation is a statistical measure that quantifies the relationship between two variables. It indicates how changes in one variable are associated with changes in another variable.

Correlation is a statistical measure that quantifies the relationship between two variables. It indicates how changes in one variable are associated with changes in another variable. Correlation values range from -1 to 1, where:

A correlation of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other decreases in a perfectly linear manner.

A correlation of 0 indicates no correlation, meaning that there is no linear relationship between the variables.

A correlation of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other also increases in a perfectly linear manner.

Correlation is often measured using the Pearson correlation coefficient, also known as Pearson's r. It is commonly used for continuous variables and assumes a linear relationship between the variables. The formula for calculating Pearson's correlation coefficient is:

r = (Σ((xi - x̄)(yi - ȳ))) / (n *σx* σy)

Where:

- xi and yi are the individual data points in the two variables
- x̄ and ȳ are the means of the two variables
- σx and σy are the standard deviations of the two variables
- n is the number of data points

Positive values of r indicate a positive correlation, while negative values indicate a negative correlation. The magnitude of r represents the strength of the correlation, with values closer to -1 or 1 indicating a stronger relationship.

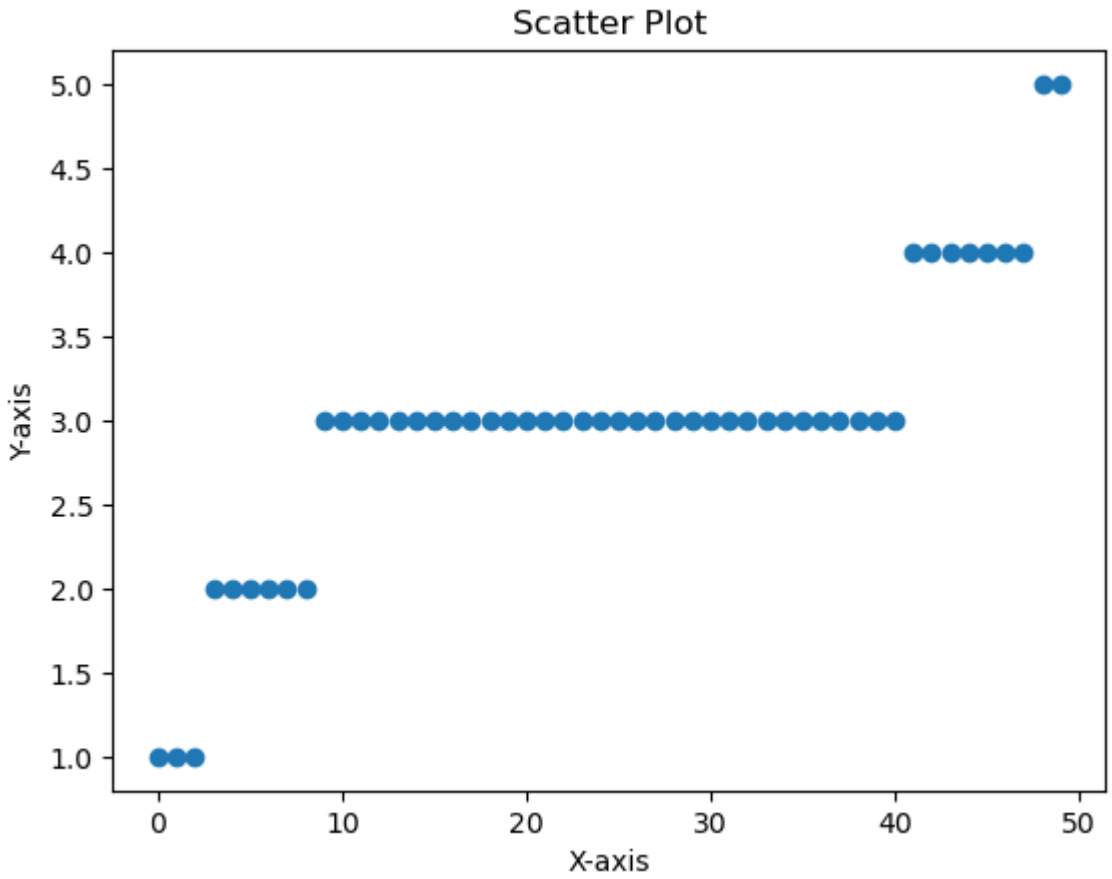
In [15]:
```
import numpy as np

# Calculate covariance
covariance_matrix = np.cov(data1, data2)
covariance = covariance_matrix[0, 1]

# Calculate correlation coefficient
correlation_matrix = np.corrcoef(data1, data2)
correlation = correlation_matrix[0, 1]
```

```
# Print the covariance and correlation coefficient
print("Covariance:", covariance)
print("Correlation Coefficient:", correlation)

# Print the covariance and correlation matrix
print("Covariance Matrix:\n", covariance_matrix)
print("Correlation Matrix:\n", correlation_matrix)
```

```
Covariance: 9.948979591836734
Correlation Coefficient: 0.831901269752335
Covariance Matrix:
 [[212.5          9.94897959]
 [  9.94897959   0.67306122]]
Correlation Matrix:
 [[1.         0.83190127]
 [0.83190127 1.        ]]
```

In [16]:

```
# Create a scatter plot
plt.scatter(data1, data2)

# Add labels and title
plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.title('Scatter Plot')

# Display the plot
plt.show()
```



Sampling and Sampling Distribution

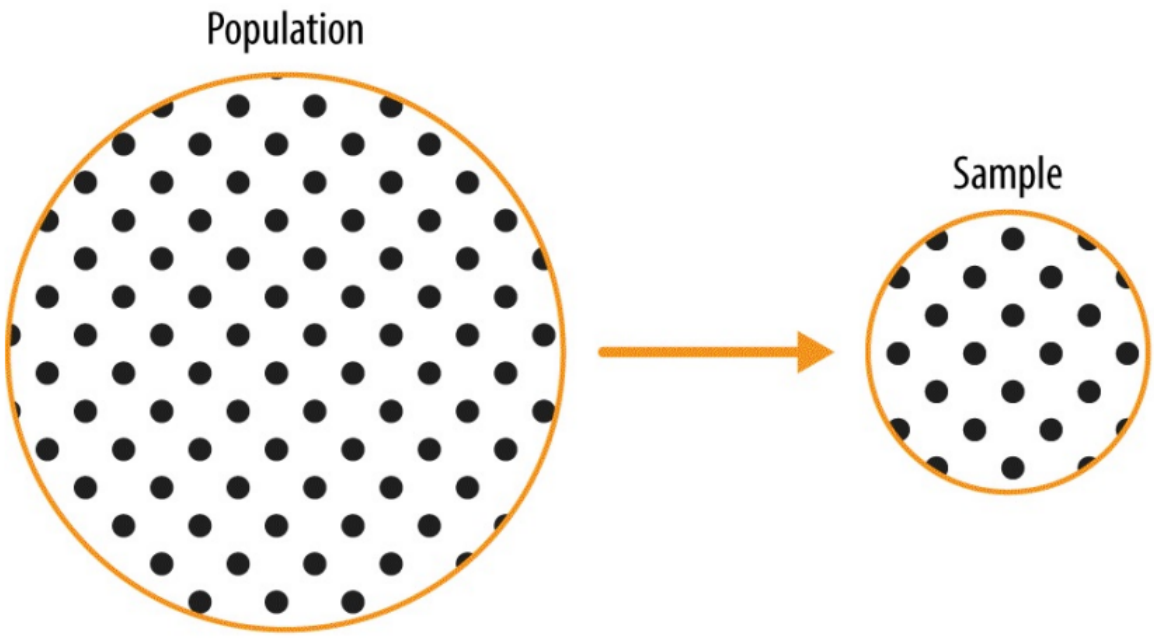
Sampling in statistics refers to the process of selecting a subset of individuals or items from a larger population to gather data and make inferences about the entire population. Instead of collecting data from every individual in the population, sampling allows statisticians to study a representative sample, which is a smaller and more manageable subset of the population. By studying the sample, one can make inferences about the population as a whole.

Population:

- The population refers to the entire group of individuals or items that the researcher is interested in studying. It can be finite or infinite.

Sample:

- A sample is a subset of the population that is selected for data collection and analysis. The goal is for the sample to be representative of the population, so that conclusions drawn from the sample can be generalized to the larger population.



Sampling Frame:

- A sampling frame is a list or representation of all the individuals or items in the population from which the sample will be selected. It provides a basis for selecting the sample and should ideally include all members of the population.

Sampling Methods:

Simple Random Sampling:

- Every member of the population has an equal chance of being selected for the sample. This can be done with or without replacement.

Stratified Sampling:

- The population is divided into homogeneous subgroups called strata, and a sample is selected from each stratum proportionate to its size or importance.

Cluster Sampling:

- The population is divided into clusters or groups, and a subset of clusters is randomly selected. Then, all members within the selected clusters are included in the sample.

Systematic Sampling:

- The population is ordered, and individuals or items are selected at regular intervals, such as every 10th person or every 5th item.

Convenience Sampling:

- Selecting individuals or items based on convenience or accessibility. This method may introduce bias and is generally not considered representative.

Bias

Bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process. An important distinction should be made between errors due to random chance, and errors due to bias. Consider the physical process of a gun shooting at a target. It will not hit the absolute center of the target every time, or even much at all. An unbiased process will produce error, but it is random and does not tend strongly in any direction

Inferential statistics

Inferential statistics involves making inferences and drawing conclusions about a population based on a sample of data. It uses probability theory and statistical techniques to estimate parameters, test hypotheses, and make predictions. Here are some common techniques used in inferential statistics:

Central Limit Theorem

The Central Limit Theorem (CLT) is a fundamental concept in statistics that states that the sampling distribution of the mean of a sufficiently large number of independent and identically distributed (i.i.d.) random variables will approximate a normal distribution, regardless of the shape of the original population distribution.

Key points about the Central Limit Theorem:

Sampling Distribution: The CLT applies to the distribution of sample means (or sums) rather than the original population distribution. It describes how the means of different samples from the

same population behave.

Independent and Identically Distributed: The random variables in the sample should be independent of each other and have the same distribution. This assumption allows for the random variables to be added together, which forms the basis of the CLT.

Sample Size: As the sample size increases, the sampling distribution of the mean approaches a normal distribution. Typically, a sample size of 30 or greater is considered sufficient for the CLT to hold, although the exact threshold depends on the population distribution.

Normal Approximation: The CLT states that, regardless of the shape of the population distribution, the sampling distribution of the mean tends to follow a normal distribution as the sample size increases. The mean of the sampling distribution will be equal to the population mean, and the standard deviation will be the population standard deviation divided by the square root of the sample size.

The practical implication of the Central Limit Theorem is that it allows us to make inferences about population parameters based on sample statistics. It forms the foundation for many statistical techniques, such as hypothesis testing and confidence intervals.

Standard Error

The standard error is a single metric that sums up the variability in the sampling distribution for a statistic. The standard error can be estimated using a statistic based on the standard deviation s of the sample values, and the sample size n

$$\text{Standard error} = SE = \frac{s}{\sqrt{n}}$$

Confidence Intervals:

A confidence interval provides a range of values within which the true population parameter is estimated to lie with a certain level of confidence.

- For example, a 95% confidence interval for the population mean is a range of values that is expected to contain the true mean with 95% confidence.

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

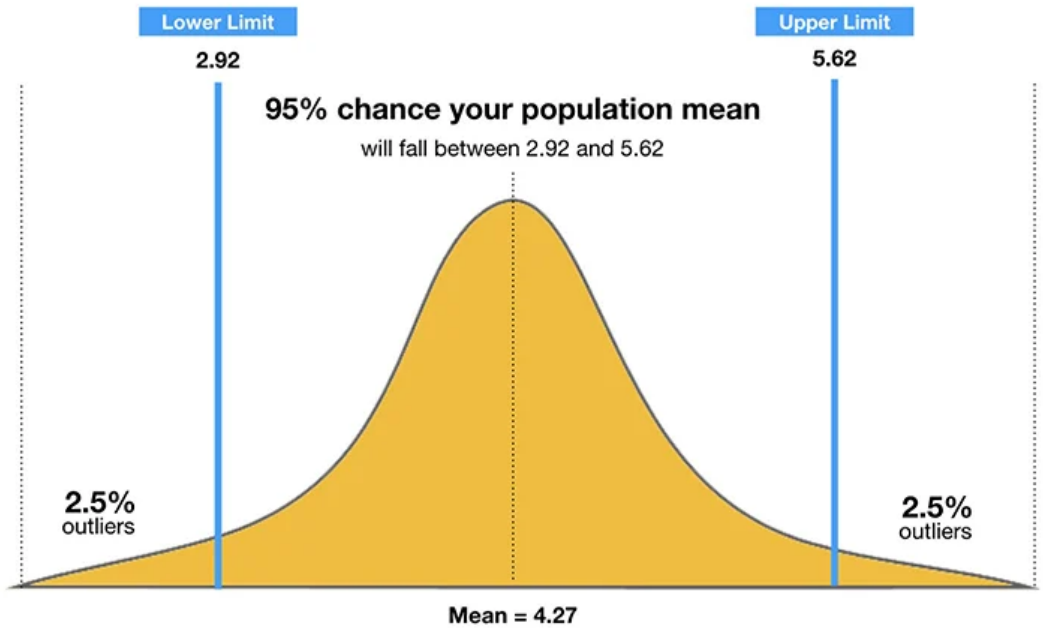
CI = confidence interval

$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size



1. Confidence Interval for the Population Mean (σ Known): For a known population standard deviation (σ) and a sample mean ($\bar{x}$) based on a sample of size n, the confidence interval for the population mean is given by:

$\bar{x} \pm z * (\sigma / \sqrt{n})$

Where:

- $\bar{x}$ is the sample mean
- z is the z-score corresponding to the desired level of confidence (e.g., for a 95% confidence interval, z ≈ 1.96)
- σ is the population standard deviation
- n is the sample size

1. Confidence Interval for the Population Mean (σ Unknown): When the population standard deviation (σ) is unknown, you can use the sample standard deviation (s) as an estimate. In this case, the confidence interval formula for the population mean is given by:

$\bar{x} \pm t * (s / \sqrt{n})$

Where:

- $\bar{x}$ is the sample mean
- t is the t-score corresponding to the desired level of confidence and degrees of freedom (df)
- s is the sample standard deviation
- n is the sample size

1. Confidence Interval for the Population Proportion: When estimating the population proportion (p) based on a binomial distribution, the confidence interval formula is given by:

$\hat{p} \pm z \sqrt{((\hat{p} (1 - \hat{p})) / n)}$

Where:

- $\hat{p}$ is the sample proportion (number of successes divided by sample size)
- z is the z-score corresponding to the desired level of confidence (e.g., for a 95% confidence interval, z ≈ 1.96)
- n is the sample size

In [17]:
```
import scipy.stats as stats
# sample data
data = np.array([25, 30, 35, 40, 45, 50, 55, 60, 65, 70])

# Calculate sample mean and standard deviation
sample_mean = np.mean(data)
sample_std = np.std(data, ddof=1)  # ddof=1 for sample standard deviation

# Set confidence level and alpha
confidence_level = 0.95
alpha = 1 - confidence_level

# Calculate the critical value (two-tailed)
critical_value = stats.t.ppf(1 - alpha / 2, df=len(data) - 1)

# Calculate the margin of error
margin_of_error = critical_value * sample_std / np.sqrt(len(data))

# Calculate the confidence interval
lower_bound = sample_mean - margin_of_error
upper_bound = sample_mean + margin_of_error

# Print the results
print("Sample Mean:", sample_mean)
print("Margin of Error:", margin_of_error)
print("Confidence Interval:", lower_bound, "-", upper_bound)
```

```
Sample Mean: 47.5
Margin of Error: 10.829252948066955
Confidence Interval: 36.67074705193305 - 58.32925294806695
```

Distribution

In statistics, a distribution refers to the way in which values are spread or distributed across a dataset or a population. It describes the probability of observing different outcomes or values. Understanding the distribution of data is crucial for making inferences, performing statistical analyses, and making predictions. There are various types of distributions commonly used in statistics, including:

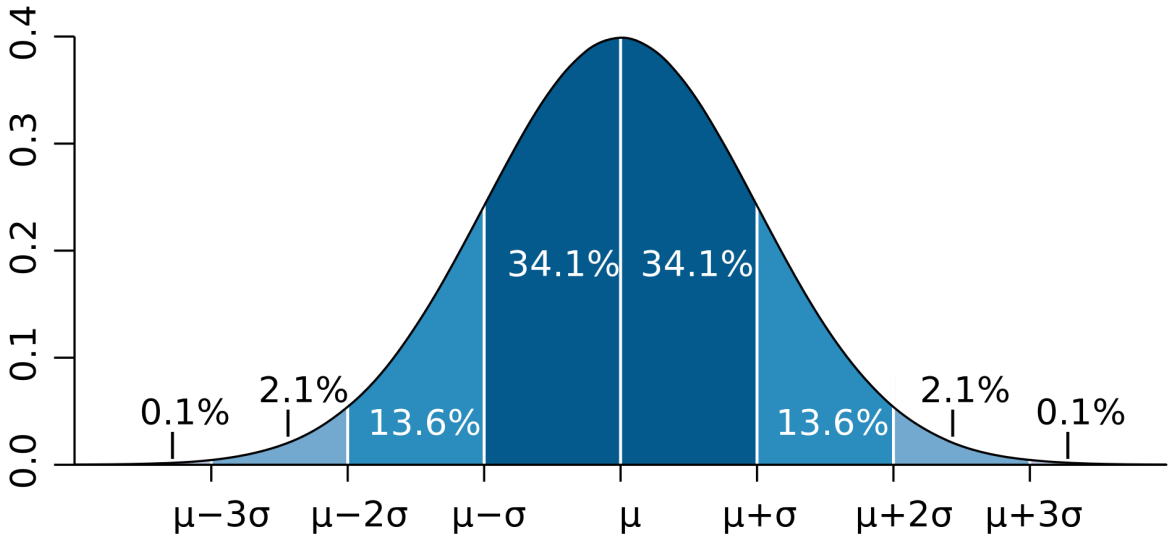
Normal Distribution (Gaussian Distribution):

The normal distribution is the most well-known and frequently encountered distribution. It is symmetrical and bell-shaped, with the mean, median, and mode all located at the center of the distribution. Many real-world phenomena follow a normal distribution, such as heights, weights, and IQ scores. It is characterized by its mean (μ) and standard deviation (σ).

Probability density function (PDF)

$f(x) = (1 / (\sigma \sqrt{(2\pi)})) e^{(-((x - \mu)^2) / (2\sigma^2))}$

- x represents a random variable that follows a normal distribution.
- μ is the mean of the distribution, which determines the center or average value.
- σ is the standard deviation of the distribution, which determines the spread or variability of the data.
- π is a mathematical constant (approximately 3.14159).
- e is the base of the natural logarithm (approximately 2.71828).



https://en.wikipedia.org/wiki/Normal_distribution

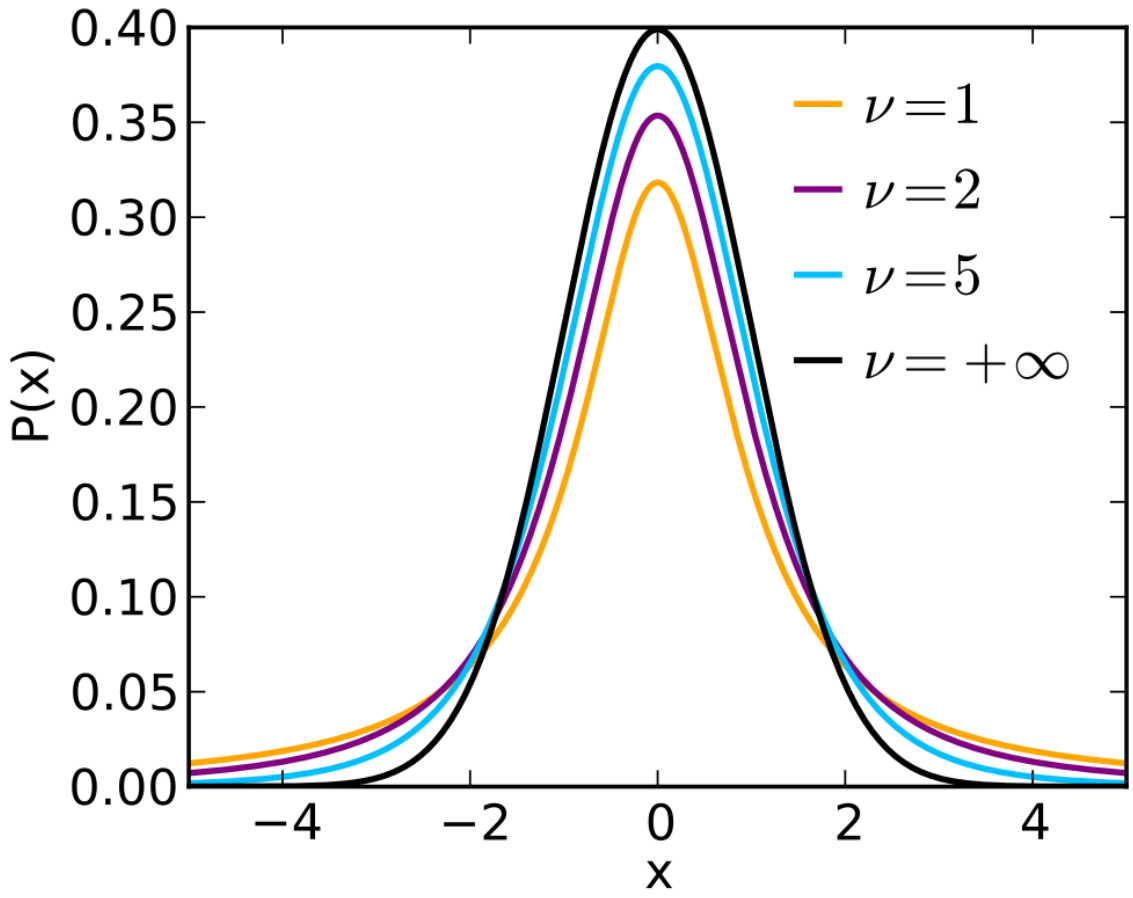
Student's t-Distribution:

The t-distribution is used when the sample size is small, and the population standard deviation is unknown. It is commonly used in hypothesis testing and constructing confidence intervals.

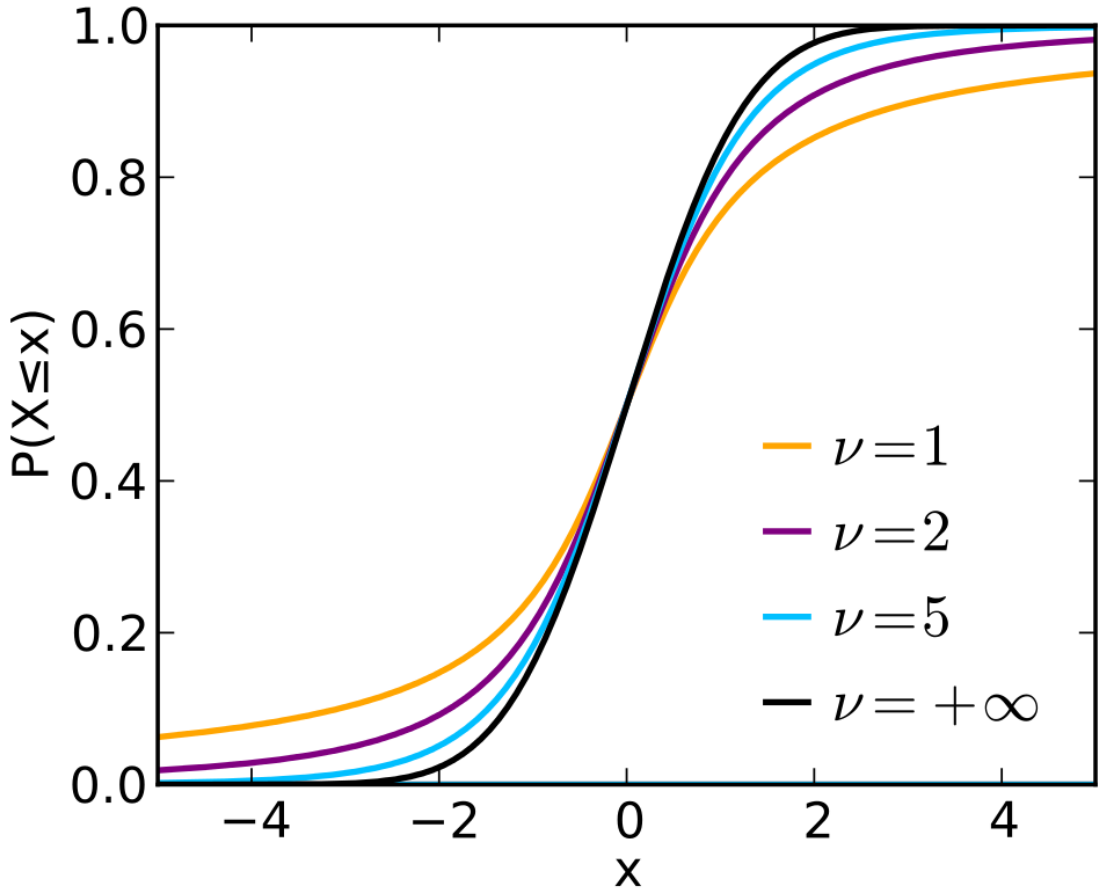
Probability density function (PDF)

f(x) = [Γ((df + 1) / 2) / (sqrt(df *π*) Γ(df / 2))] * [(1 + (x^2 / df))^(-(df + 1) / 2)]

- f(x) is the probability density function at the value x.
- Γ is the gamma function, which is a generalization of the factorial function.
- df is the degrees of freedom.



,



https://en.wikipedia.org/wiki/Student%27s_t-distribution

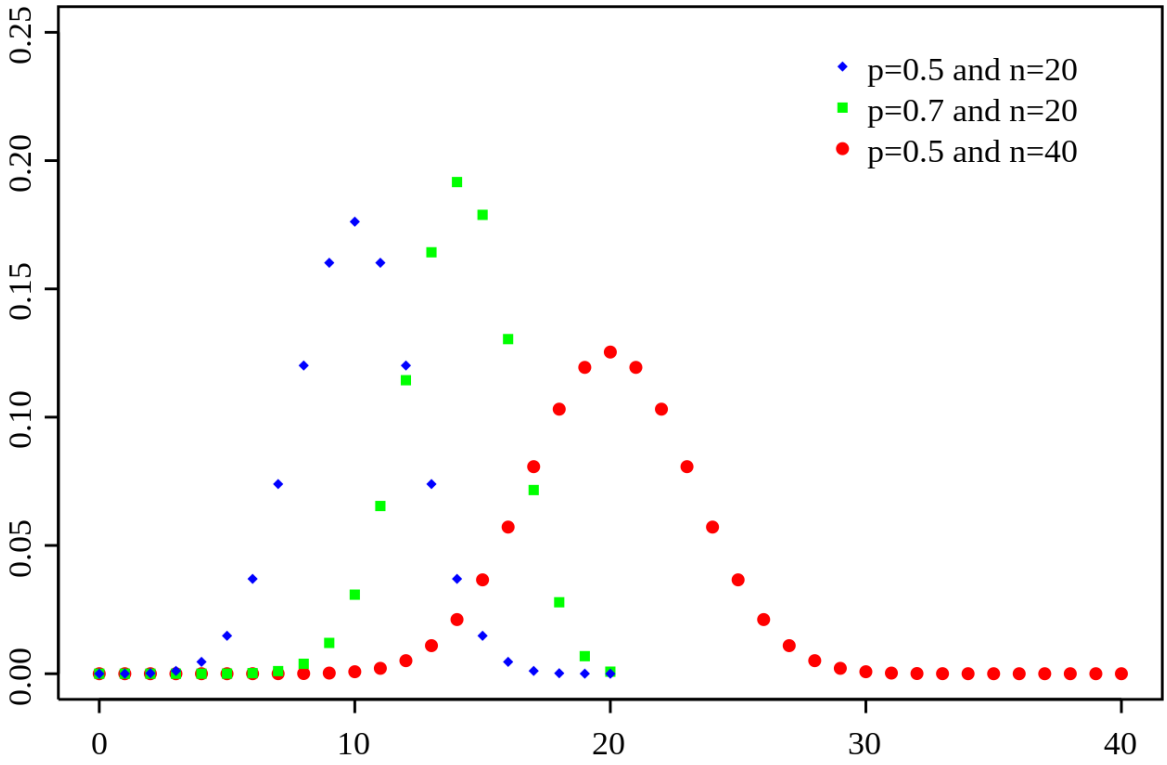
Binomial Distribution:

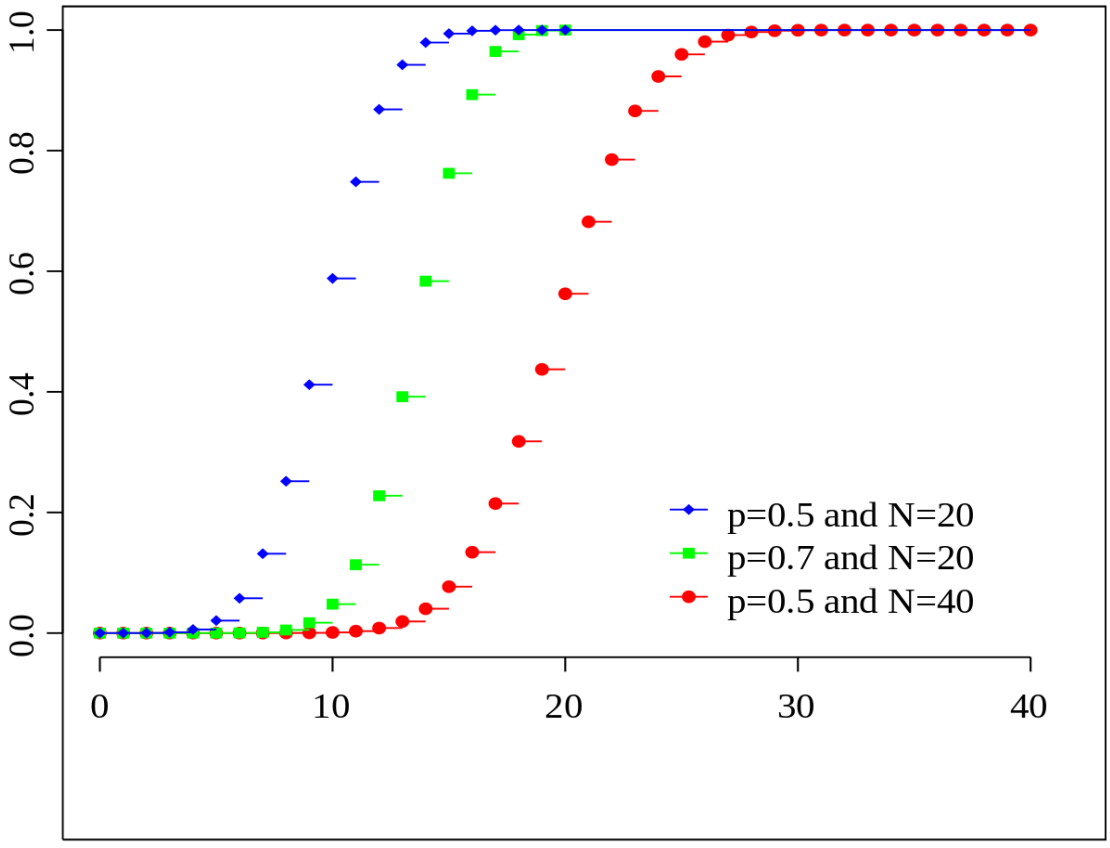
The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials (experiments with two possible outcomes: success or failure). It is characterized by two parameters: the number of trials (n) and the probability of success (p) in each trial. Examples include the number of heads obtained in a series of coin flips or the number of defective items in a production line.

Probability mass function (PMF)

P(X = k) = C(n, k) *p^k* q^(n-k)

- P(X = k) is the probability of exactly k successes in n trials.
- C(n, k) represents the number of combinations, also known as binomial coefficients, of choosing k successes from n trials. It is calculated as: C(n, k) = n! / (k! * (n-k)!).
- p^k represents the probability of k successes occurring.
- q^(n-k) represents the probability of (n-k) failures occurring.





https://en.wikipedia.org/wiki/Binomial_distribution

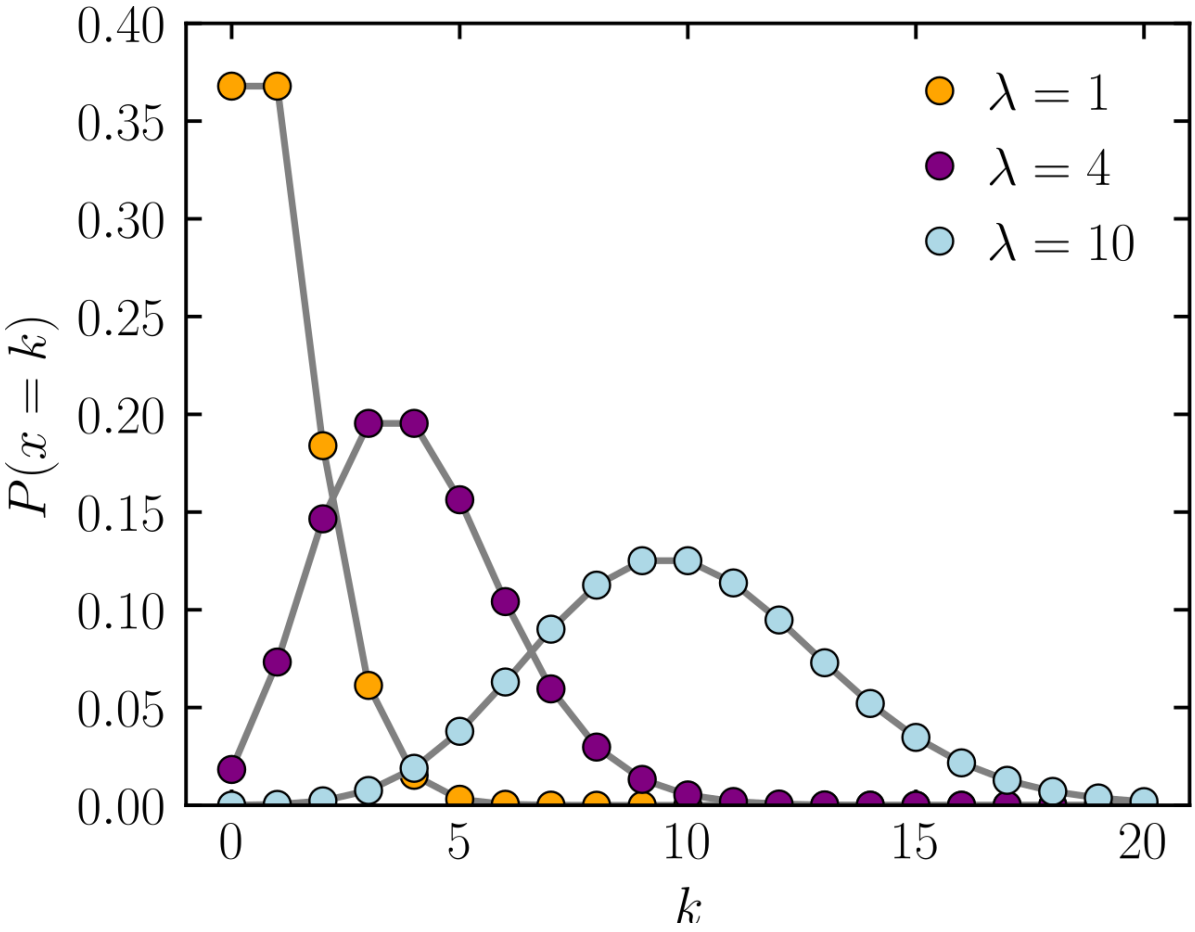
Poisson Distribution:

The Poisson distribution models the number of events occurring within a fixed interval of time or space. It is used when the events are rare and occur independently of each other. The Poisson distribution is characterized by a single parameter, λ (lambda), which represents the average rate of events occurring in the interval. It is often applied to model occurrences of traffic accidents, phone calls, or defects in a product.

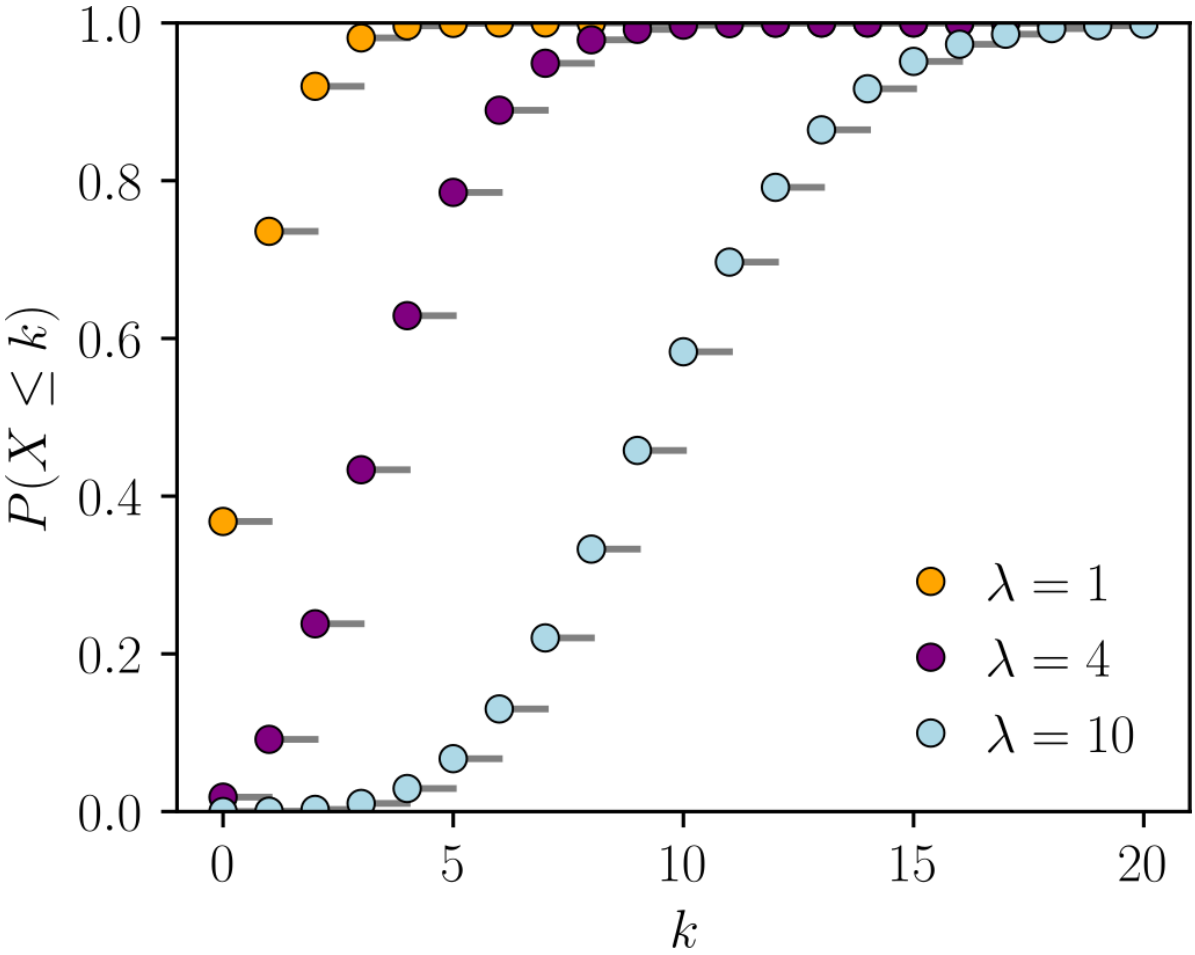
Probability mass function (PMF)

P(X = k) = (e^(-λ) * λ^k) / k!

- P(X = k) is the probability of observing exactly k events within the interval.
- e is the base of the natural logarithm (approximately 2.71828).
- λ is the average rate of occurrence of events within the interval.
- k represents the number of events (ranging from 0 to infinity) to be observed.



'



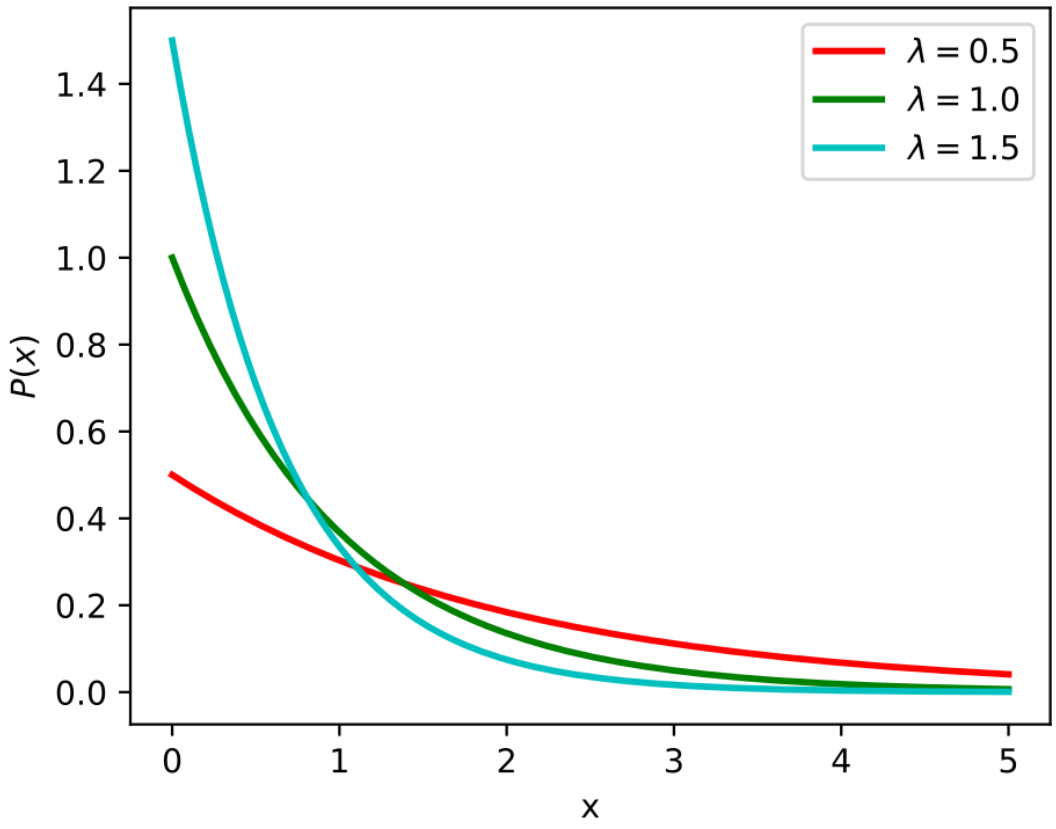
https://en.wikipedia.org/wiki/Poisson_distribution

Exponential Distribution:

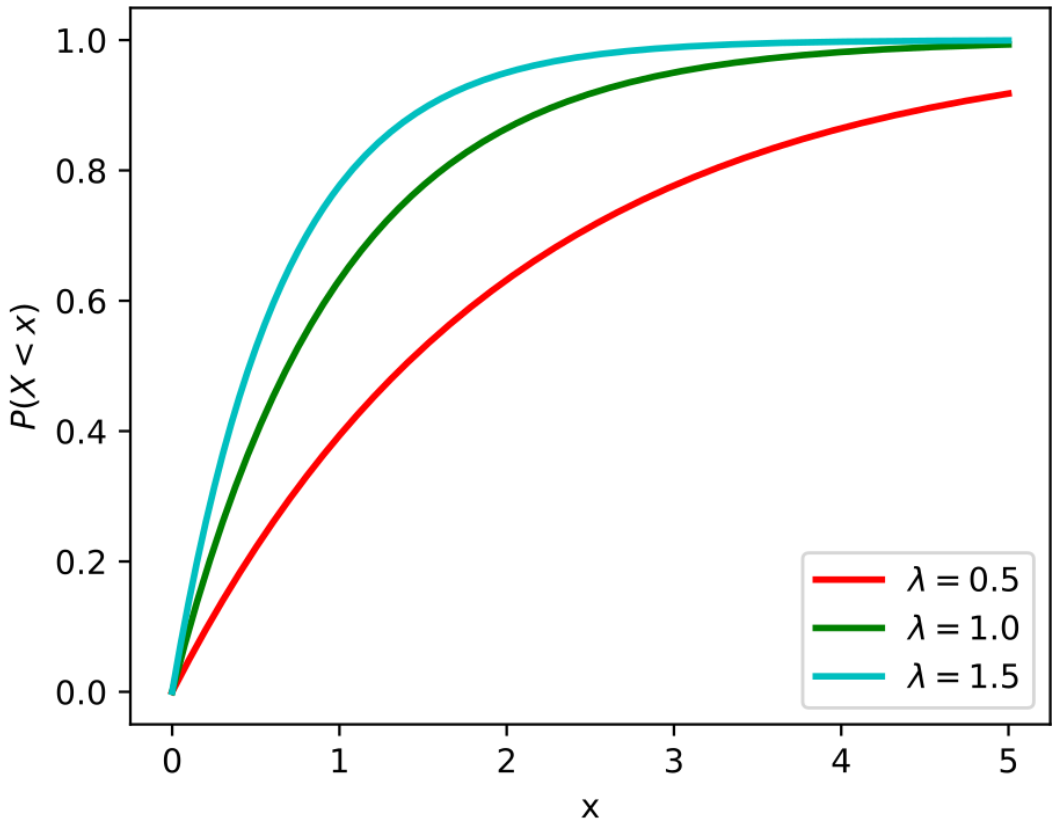
The exponential distribution models the time between events occurring in a Poisson process. It is characterized by a single parameter, λ (lambda), which represents the average rate at which events occur. The exponential distribution is commonly used in reliability analysis and queuing theory.

Probability density function (PDF) f(x) = λ * e^(-λx)

- f(x) is the probability density function at the value x.
- λ is the rate parameter.
- e is the base of the natural logarithm (approximately 2.71828).
- x is the time or the duration until the event occurs.


λ = 0.5
λ = 1.0
λ = 1.5
P(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
1
2
3
4
5

,



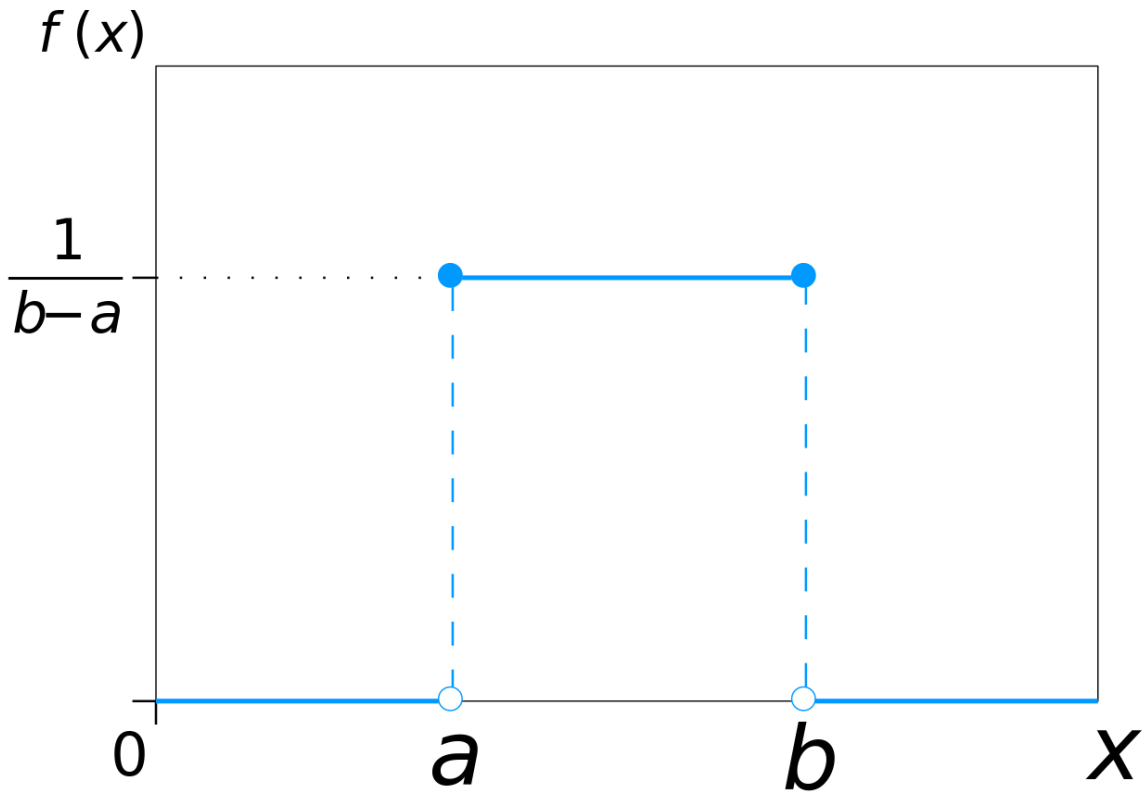
https://en.wikipedia.org/wiki/Exponential_distribution

Uniform Distribution:

The uniform distribution, also known as the rectangular distribution, is characterized by a constant probability for all values within a specified range. It is a symmetric distribution, and each value has an equal chance of being observed. An example is rolling a fair die, where each side has an equal probability of occurring.

Probability density function (PDF) f(x) = 1 / (b - a) for a ≤ x ≤ b

- x represents a random variable that follows a uniform distribution.
- a is the lower bound of the distribution.
- b is the upper bound of the distribution.



https://en.wikipedia.org/wiki/Continuous_uniform_distribution

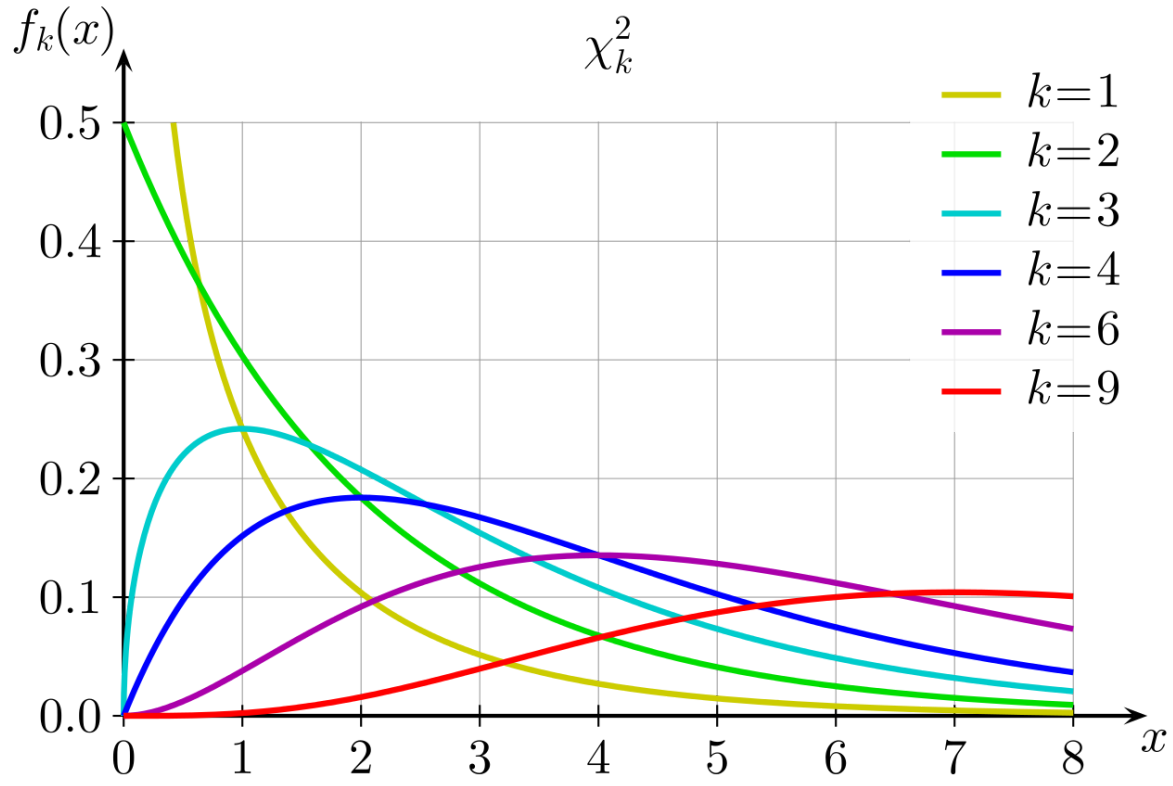
Chi-Square Distribution:

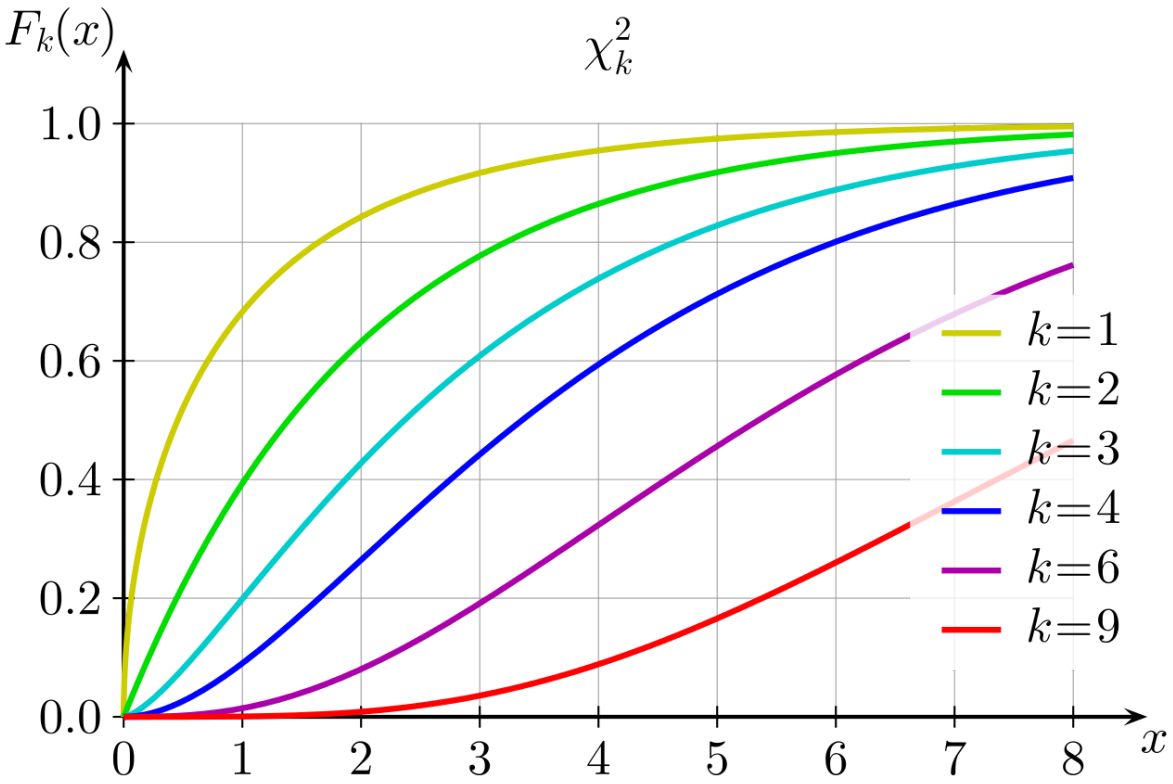
The chi-square distribution arises in various statistical tests, such as the chi-square test of independence or the chi-square test of goodness of fit. It is a right-skewed distribution and its shape depends on the degrees of freedom.

Probability density function (PDF)

f(x) = (1 / (2^(df/2) *Γ(df/2))*) x^((df/2)-1) * e^(-x/2)

- f(x) is the probability density function at the value x.
- df is the degrees of freedom.
- Γ is the gamma function, which is a generalization of the factorial function.





https://en.wikipedia.org/wiki/Chi-squared_distribution

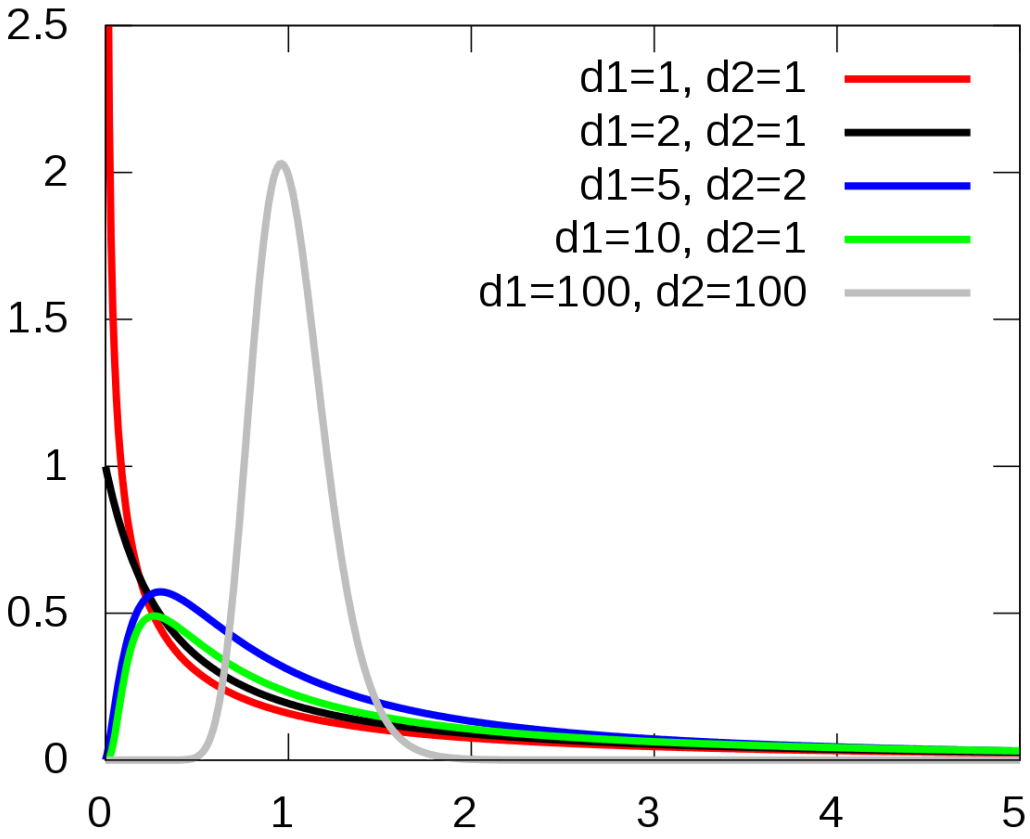
F-Distribution:

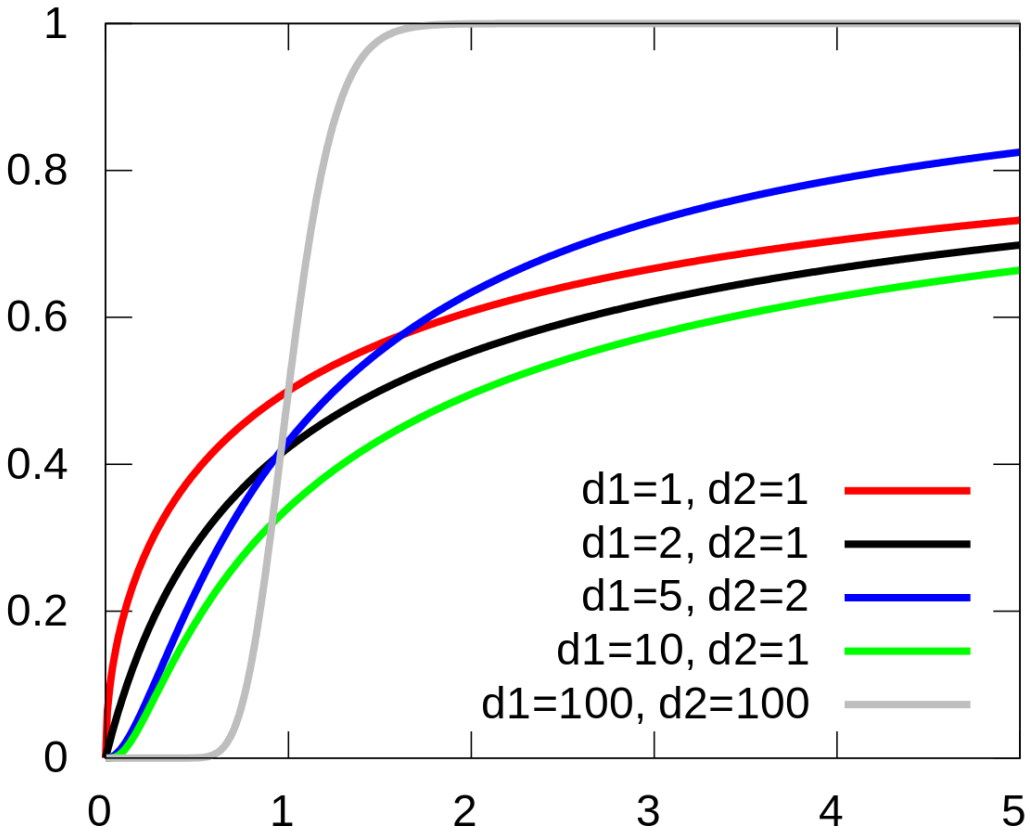
The F-distribution is used in the analysis of variance (ANOVA) and in comparing variances between groups. It arises when comparing the variances of two or more samples.

Probability density function (PDF)

f(x) = [((df1 / 2)^(df1 / 2) *(df2 / 2)^(df2 / 2)) / (B(df1 / 2, df2 / 2))]* [(x^(df1 / 2 - 1)) / ((df1 / 2) *(x* df1 / df2 + 1)^(df1 / 2 + df2 / 2))]

- f(x) is the probability density function at the value x.
- B(a, b) represents the beta function.





https://en.wikipedia.org/wiki/F-distribution

P-value:

P-Value The p-value is a measure of the evidence against the null hypothesis in hypothesis testing. It represents the probability of obtaining the observed data or more extreme results if the null hypothesis is true. A p-value below a pre-determined significance level (e.g., 0.05) indicates statistical significance.

Simply looking at the graph is not a very precise way to measure statistical significance, so of more interest is the p-value. This is the frequency with which the chance model produces a result more extreme than the observed result. We can estimate a p-value from our permutation test by taking the proportion of times that the permutation test produces a difference equal to or greater than the observed difference:

- Example: If the p-value is less than 0.05, it suggests strong evidence against the null hypothesis, leading to its rejection.

Type 1 and Type 2 Errors

In assessing statistical significance, two types of error are possible:

- Type 1 error, in which you mistakenly conclude an effect is real, when it is really just due to chance
- Type 2 error, in which you mistakenly conclude that an effect is not real (i.e.,due to chance), when it really is real

A/B testing

A/B testing, also known as split testing or bucket testing, is a method used in statistics and data analysis to compare the performance of two or more variants of a webpage, app, or marketing campaign. It is commonly used in the field of data science and digital marketing to make data-driven decisions and optimize conversions.

Suppose an e-commerce website wants to test the impact of a simplified checkout process on their conversion rate. The current checkout process is considered the control group (Group A), while the simplified checkout process is the experimental group (Group B).

- Define the Objective: The objective is to determine if the simplified checkout process improves the conversion rate.

- Formulate the Hypothesis: The hypothesis is that the simplified checkout process will lead to a higher conversion rate compared to the current checkout process.

- Design the Experiment: Randomly assign visitors to either the control group (Group A) or the experimental group (Group B). Visitors in Group A will experience the current checkout process, while visitors in Group B will experience the simplified checkout process.

- Collect Data: Track the number of visitors and the number of conversions (e.g., successful purchases) for each group during a specified time period.

- Analyze Results: Perform statistical analysis to compare the conversion rates between Group A and Group B. We will use a hypothesis test, such as a two-sample proportion test, to determine if there is a statistically significant difference between the two groups. We'll calculate the p-value associated with the test to evaluate the significance of the results.

In [18]:
```
import statsmodels.stats.proportion as smprop

# significence level
significance_level = 0.05

# Control group (Group A) data
num_visitors_A = 1000  # number of visitors in Group A
num_conversions_A = 50  # number of conversions in Group A

# Experimental group (Group B) data
num_visitors_B = 1000  # number of visitors in Group B
num_conversions_B = 70  # number of conversions in Group B

# Perform two-sample proportion test
successes = [num_conversions_A, num_conversions_B]
visitors = [num_visitors_A, num_visitors_B]
z_stat, p_value = smprop.proportions_ztest(successes, visitors)

# Print the p-value
print("p-value:", p_value)

if p_value < significance_level:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

```
p-value: 0.059685605532426224
Fail to reject the null hypothesis
```

- Draw Conclusions:

We fail to reject the null hypothesis. This means that we do not have enough evidence to conclude that the simplified checkout process significantly improves the conversion rate compared to the current checkout process.

Hypothesis Testing:

Hypothesis testing involves making decisions and drawing conclusions about a population based on sample data. It typically involves testing a null hypothesis against an alternative hypothesis.

- Common tests include t-tests, z-tests, chi-square tests, and ANOVA.
- Example: Testing whether a new drug has a statistically significant effect on a certain medical condition by comparing the outcomes of a treatment group (sample) and a control group (sample).

Example: Suppose a pharmaceutical company develops a new drug to treat a specific medical condition and wants to determine if the new drug is more effective than the current standard

treatment. They conduct a clinical trial with two groups: Group A receives the standard treatment, and Group B receives the new drug treatment.

- Hypotheses:
  - Null Hypothesis (H_0): The new drug treatment is not more effective than the standard treatment.
  - Alternative Hypothesis (H_1): The new drug treatment is more effective than the standard treatment.
- Data: Let's say the trial includes 100 patients in each group. We record the number of patients who show improvement after the treatment in each group.
- Group A (Standard Treatment): Number of patients showing improvement = 60
- Group B (New Drug Treatment): Number of patients showing improvement = 75
- Assumptions: We assume that the patients in each group are independent, and the data follows a binomial distribution.
- Test Statistic and Significance Level: Since we have two independent groups and are comparing proportions (success rates), we can use a two-sample proportion test. Let's choose a significance level of 0.05 ($\alpha = 0.05$).
- Performing the Hypothesis Test:

In [19]:
```
import statsmodels.api as sm

# Data
successes = [60, 75]
nobs = [100, 100]

# Set significance level
alpha = 0.05

# Perform two-sample proportion z-test
z_stat, p_value = sm.stats.proportions_ztest(successes, nobs, alternative='larger')

# Print the p-value
print("p-value:", p_value)

if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

```
p-value: 0.988229970869411
Fail to reject the null hypothesis
```

- Conclusion There is no evidence to conclude that the new drug treatment is more effective than the standard treatment for the given medical condition.

t-Tests

t-Tests are statistical tests used to compare the means of two groups and determine if they are significantly different from each other. There are three main types of t-tests: independent samples t-test, paired samples t-test, and one-sample t-test. Each type of t-test is used in different scenarios based on the nature of the data and the research question.

- Independent Samples t-Test: The independent samples t-test is used when comparing the means of two independent groups. It determines if there is a significant difference between the means of two populations or groups.
- Paired Samples t-Test: The paired samples t-test is used when comparing the means of two related or paired groups. It analyzes the differences between paired observations to determine if there is a significant difference between the means.
- One-Sample t-Test: The one-sample t-test is used when comparing the mean of a single sample to a known or hypothesized population mean. It determines if there is a significant difference between the sample mean and the hypothesized population mean.

Example: Suppose we have data from two departments: Department A and Department B. We want to determine if there is a significant difference in the average salaries between the two departments.

- Hypotheses:
  - Null Hypothesis (H_0): There is no significant difference in the average salaries between Department A and Department B.
  - Alternative Hypothesis (H_1): There is a significant difference in the average salaries between Department A and Department B.
- Data: Let's say we have the following average salary data for each department:

  Department A: [50000, 52000, 48000, 51000, 49000] Department B: [55000, 53000, 52000, 56000, 54000]
- Assumptions: We assume that the data follows a normal distribution, the variances of the two groups are equal, and the samples are independent.
- Significance Level: Let's choose a significance level of 0.05 ($\alpha = 0.05$).
- Performing the t-Test: We calculate the test statistic (t-statistic) and p-value using the independent samples t-test.

In [20]:
```
import scipy.stats as stats

# Data
department_A = [50000, 52000, 48000, 51000, 49000]
department_B = [55000, 53000, 52000, 56000, 54000]

# Perform independent samples t-test
t_stat, p_value = stats.ttest_ind(department_A, department_B)
```

```
# Print the p-value
print("p-value:", p_value)

if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

```
p-value: 0.003949772803445322
Reject the null hypothesis
```

- Conclusion: There is sufficient evidence to conclude that there is a significant difference in the average salaries between Department A and Department B.

Analysis of Variance (ANOVA):

ANOVA is used to compare means across two or more groups or treatments. It determines if there are statistically significant differences among the means and which groups are significantly different from each other. Example: Comparing the mean test scores of students from different schools to determine if there are significant differences in performance.

ANOVA Test

- Hypotheses:
  - Null Hypothesis (H_0): The means of all groups are equal.
  - Alternative Hypothesis (H_1): At least one group mean is different from the others.
- Data: Collect data from three or more groups, with each group having numerical measurements or observations.

Assumptions:

ANOVA assumes the following:

- Independence: Observations within each group are independent.
- Normality: The data within each group follows a normal distribution.
- Homogeneity of variances: The variance within each group is approximately equal.

Test Statistic and Significance Level:

- ANOVA calculates the F-statistic, which represents the ratio of the between-group variability to the within-group variability. The significance level (α) determines the threshold for rejecting the null hypothesis. Commonly used significance levels are 0.05 (5%) or 0.01 (1%).

- Performing ANOVA: In Python, you can use the scipy.stats module to perform ANOVA. The f_oneway function is used for one-way ANOVA, which compares the means of multiple groups. Here's an example code implementation:

In [21]:
```
import scipy.stats as stats

# Data
group1 = [10, 12, 14, 16, 18]
group2 = [8, 9, 11, 13, 15]
group3 = [5, 7, 9, 11, 13]

# Perform one-way ANOVA
f_stat, p_value = stats.f_oneway(group1, group2, group3)

# Print the p-value
print("p-value:", p_value)

if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

```
p-value: 0.07025990018324225
Fail to reject the null hypothesis
```

- Conclusion: This means that we do not have enough evidence to conclude that there is a significant difference among the group means.

Chi-square test

The chi-square test is a statistical test used to determine if there is a significant association between categorical variables. It helps assess whether the observed frequencies of categorical data differ significantly from the expected frequencies. The chi-square test is commonly used in fields such as social sciences, market research, and genetics.

chi-square test procedure

- Hypotheses:
  - Null Hypothesis (H_0): There is no association between the variables.
  - Alternative Hypothesis (H_1): There is an association between the variables.
- Data: Collect data in the form of a contingency table or cross-tabulation, which displays the frequencies or counts of observations for each combination of categories.
- Assumptions: The chi-square test assumes the following:
- The observations are independent.
- The expected frequency for each cell in the contingency table is at least 5.
- Test Statistic and Significance Level: The chi-square test calculates the chi-square statistic (χ^2), which measures the difference between observed and expected frequencies. The significance level (α) determines the threshold for rejecting the null hypothesis. Commonly used significance levels are 0.05 (5%) or 0.01 (1%).

Performing the Chi-Square Test:

In [22]:
```
import scipy.stats as stats

# Data
observed = [[30, 40, 20], [15, 20, 10]]

# Perform chi-square test
chi2_stat, p_value, dof, expected = stats.chi2_contingency(observed)

# Print the p-value
print("p-value:", p_value)

if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

```
p-value: 1.0
Fail to reject the null hypothesis
```

- Conclusion : There is no evidence to conclude that there is an association between the categorical variables.

Regression Analysis:

Regression analysis is used to model the relationship between a dependent variable and one or more independent variables. It estimates the coefficients and provides insights into the strength and direction of the relationship. Example: Analyzing the relationship between income (dependent variable) and education level (independent variable) to determine if higher education is associated with higher income.

In [23]:
```
import statsmodels.api as sm

# sample data
X = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
y = np.array([2, 4, 5, 7, 8, 11, 15, 16, 12, 20])

# Add a constant term to the X data
X = sm.add_constant(X)

# Fit the linear regression model
model = sm.OLS(y, X)
results = model.fit()

# Print the regression coefficients and summary
print("Regression Coefficients:")
print(results.params)
print("\nRegression Summary:")
print(results.summary())
```

```
Regression Coefficients:
[8.88178420e-16 1.81818182e+00]

Regression Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.884
Method:                 Least Squares   F-statistic:                     69.77
Date:                Sun, 09 Jul 2023   Prob (F-statistic):           3.20e-05
Time:                        21:50:17   Log-Likelihood:                -19.890
No. Observations:                  10   AIC:                             43.78
Df Residuals:                       8   BIC:                             44.39
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       8.882e-16      1.351   6.58e-16      1.000      -3.115       3.115
x1             1.8182      0.218      8.353      0.000       1.316       2.320
==============================================================================
Omnibus:                        7.317   Durbin-Watson:                   2.568
Prob(Omnibus):                  0.026   Jarque-Bera (JB):                2.759
Skew:                          -1.150   Prob(JB):                        0.252
Kurtosis:                       4.153   Cond. No.                         13.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specif
ied.
C:\Users\RACHIT\anaconda3\lib\site-packages\scipy\stats\_stats_py.py:1736: UserWarning:
kurtosistest only valid for n>=20 ... continuing anyway, n=10
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

Probability

Probability plays a fundamental role in statistics as it provides a mathematical framework for quantifying uncertainty and making predictions based on data. In statistics, probability refers to the likelihood of an event occurring or a specific outcome happening. Here are some key concepts related to probability in statistics:

- Theoretical Probability: Theoretical probability is based on mathematical principles and assumptions. It is calculated by dividing the number of favorable outcomes by the total number of possible outcomes. Theoretical probability is used when all outcomes are equally likely.

For example, if you roll a fair six-sided die, the theoretical probability of rolling a 4 is 1/6 because there is one favorable outcome (rolling a 4) out of six possible outcomes (rolling any number from 1 to 6).

Experimental probability, also known as empirical probability, is a type of probability that is based on observed data or experimentation. It involves conducting experiments or gathering data to

determine the likelihood of an event occurring.

- Conduct an Experiment: Perform a series of trials or observations related to the event of interest. Each trial should be independent and have the same conditions.
- Count the Favorable Outcomes: Determine the number of times the event of interest occurs during the experiments or observations. These are the favorable outcomes.
- Count the Total Outcomes: Count the total number of trials or observations conducted.
- Calculate the Experimental Probability: Divide the number of favorable outcomes by the total number of trials or observations.
- Experimental Probability = Number of Favorable Outcomes / Total Number of Trials or Observations
- Interpret the Results:

The experimental probability provides an estimate of the likelihood of the event occurring based on the observed data.

Example:

Suppose you want to determine the experimental probability of rolling a 6 on a fair six-sided die. You roll the die 100 times and record the number of times it lands on a 6, which is 20.

Experimental Probability = Number of Times Rolling a 6 / Total Number of Rolls = 20 / 100 = 0.2 or 20%

In this case, the experimental probability of rolling a 6 is 0.2 or 20%, based on the observed data from the 100 rolls.

Bayes' Theorem:

Bayes' Theorem is a fundamental concept in probability theory that allows us to update our beliefs or probabilities based on new evidence or data. Bayesian inference is an approach that uses Bayes' theorem to combine prior knowledge or beliefs with observed data to obtain posterior probabilities.

P(A|B) = (P(B|A) * P(A)) / P(B)

- P(A|B) represents the posterior probability of event A given event B (the probability of A given the evidence B).
- P(B|A) is the likelihood of event B given event A (the probability of B given A).
- P(A) is the prior probability of event A (the probability of A before considering any evidence).
- P(B) is the prior probability of event B (the probability of B before considering any evidence).

https://en.wikipedia.org/wiki/Bayes%27_theorem

Suppose there is a rare disease that affects 1% of the population. A medical test has been developed to detect this disease, and it is known that the test has a 95% accuracy rate, meaning it correctly identifies the disease in 95% of the cases and gives a false negative (misses the disease) in 5% of the cases. For individuals without the disease, the test correctly gives a negative result in 98% of the cases and gives a false positive (incorrectly indicates the presence of the disease) in 2% of the cases.

Now, suppose a person takes the test and tests positive for the disease. We want to determine the probability that this person actually has the disease.

Let's define the events: A: The person has the disease. B: The person tests positive for the disease.

We are interested in finding P(A|B), the probability that the person has the disease given that they tested positive.

- Using Bayes' Theorem: P(A|B) = (P(B|A) * P(A)) / P(B)
- Given information: P(A) = 0.01 (prior probability of having the disease) P(B|A) = 0.95 (likelihood of testing positive given having the disease) P(B|not A) = 0.02 (likelihood of testing positive given not having the disease) P(not A) = 1 - P(A) = 0.99 (prior probability of not having the disease)
- Calculating P(B): P(B) = P(B|A) *P(A) + P(B|not A)* P(not A) = 0.95 *0.01 + 0.02* 0.99 = 0.0297
- Now we can substitute the values into Bayes' Theorem: P(A|B) = (P(B|A) *P(A)) / P(B) = (0.95* 0.01) / 0.0297 ≈ 0.320

The probability that the person actually has the disease given that they tested positive is approximately 0.320 or 32.0%.