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 refers to any set of information that can be collected, observed, or recorded. It can be in various forms, such as numbers, text, images, audio, or video. Data is the foundation of statistical analysis and provides the raw material for generating insights and making informed decisions.

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:

- Mean: The arithmetic average of a set of numerical values.
  - For example, the mean of [5, 7, 8, 7, 10, 12] is (5+7+8+7+10+12)/6 = 8.167.
- Median: The middle value in a sorted set of numerical values.
  - For example, the median of [5, 7, 8, 7, 10, 12] is 7.5.
- Mode: The most frequently occurring value(s) in a dataset.
  - 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:

- 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.
- 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.
  - 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.

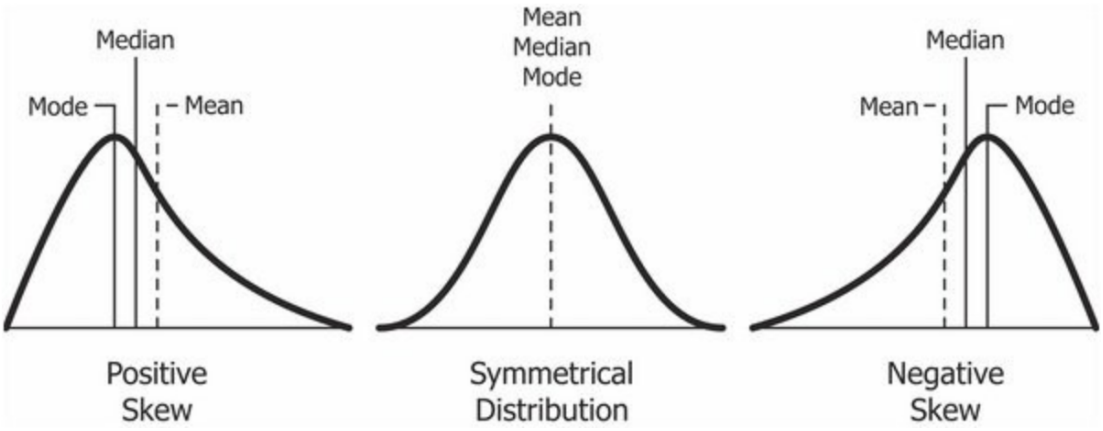
In [2]:
```
# Range
data_range = max(data) - min(data)
print("Range:", data_range)
```

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

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

```
Range: 7
Variance: 6.167
Standard Deviation: 2.483
```

Skewness

Skewness measures the asymmetry of the distribution. Positive skewness indicates a longer tail on the right side, while negative skewness indicates a longer tail on the left side.

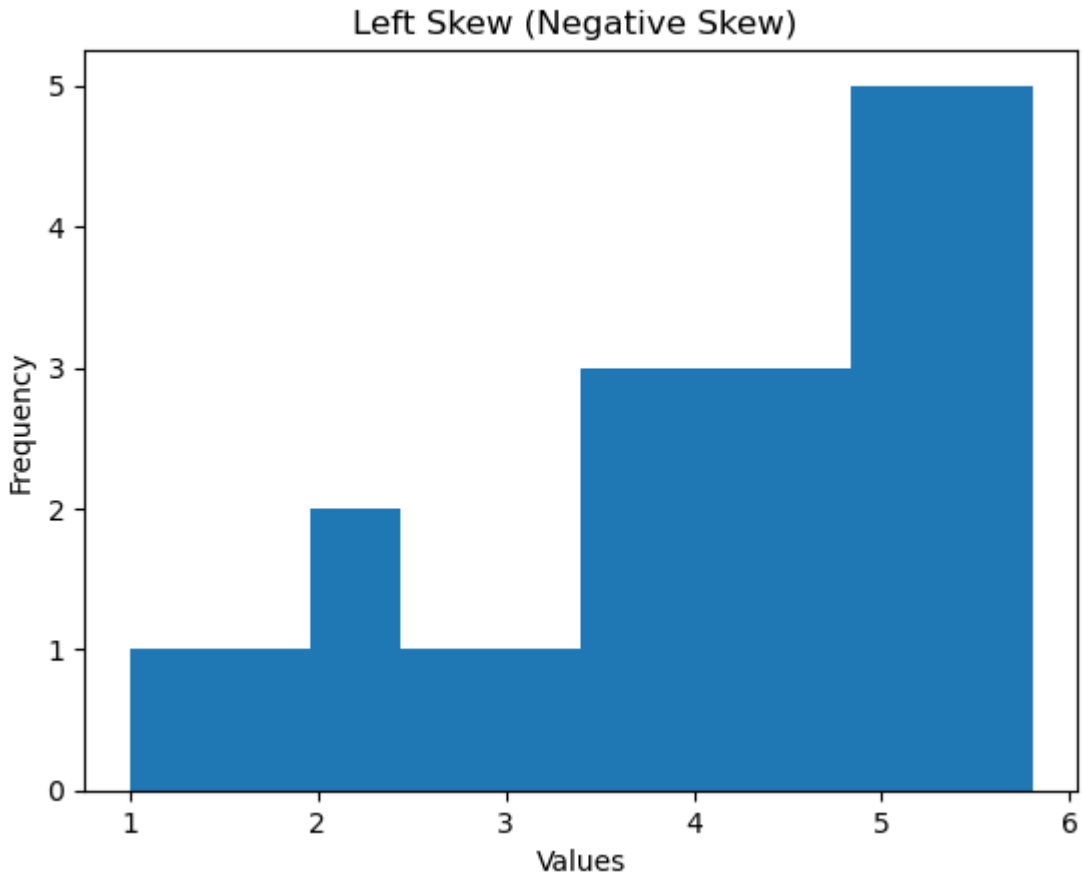


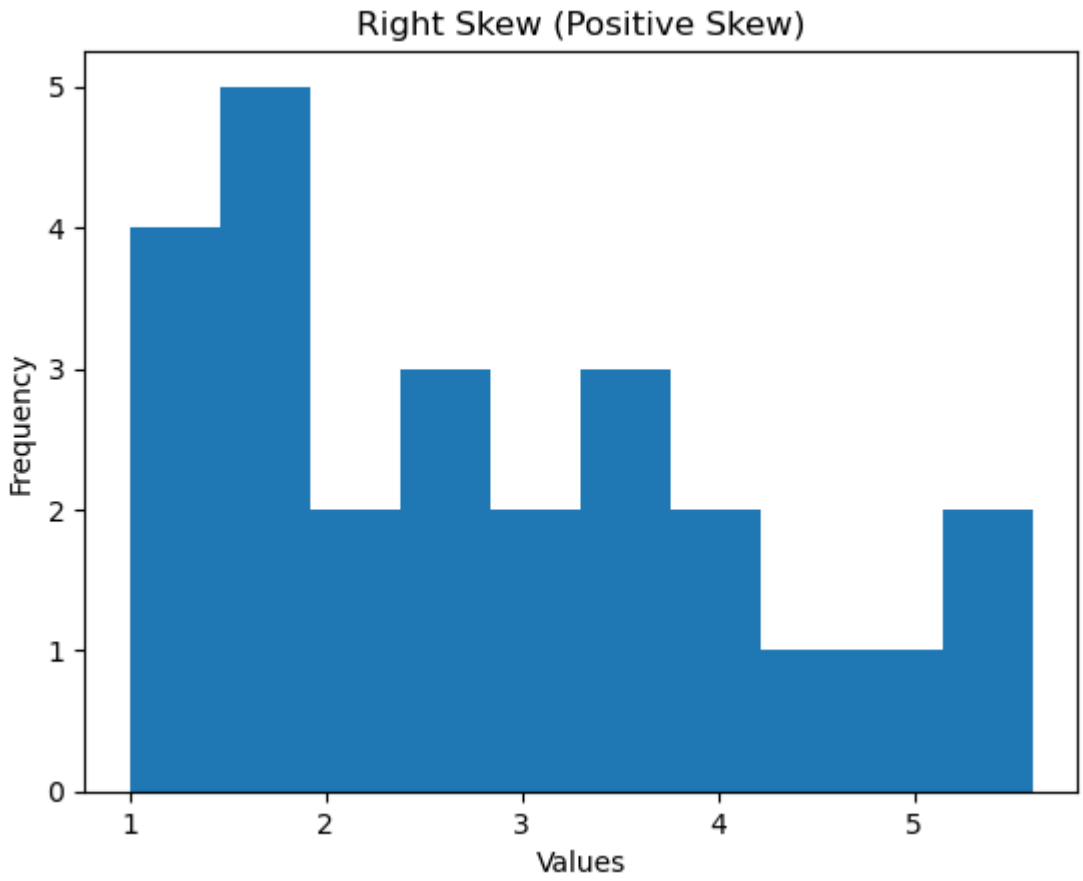
https://en.wikipedia.org/wiki/Skewness

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])
```

```
# 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()


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

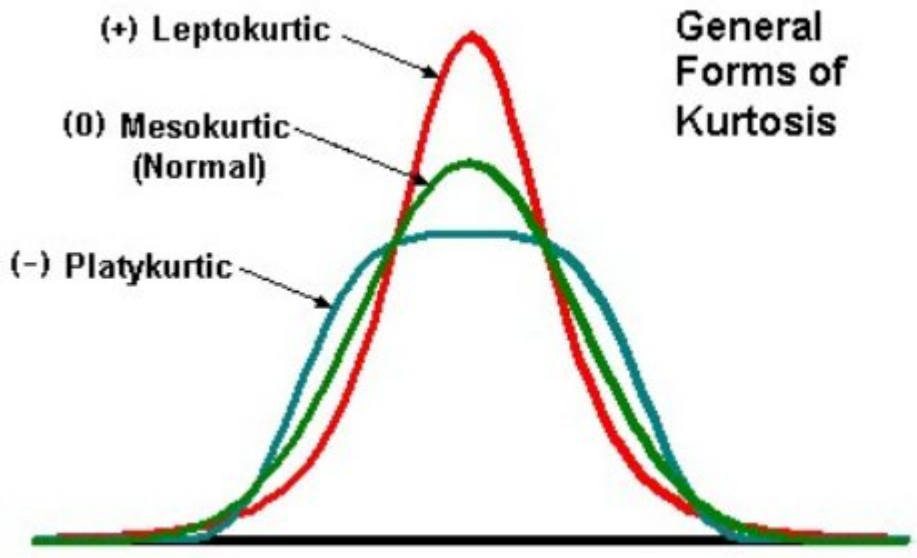





```
Skewness: -0.796938685841155
Left skew distribution
Skewness: 0.5368500349438866
Right skew distribution
```

Kurtosis:

Kurtosis measures the peakedness or flatness of a distribution. High kurtosis indicates a more peaked distribution, while low kurtosis indicates a flatter distribution.

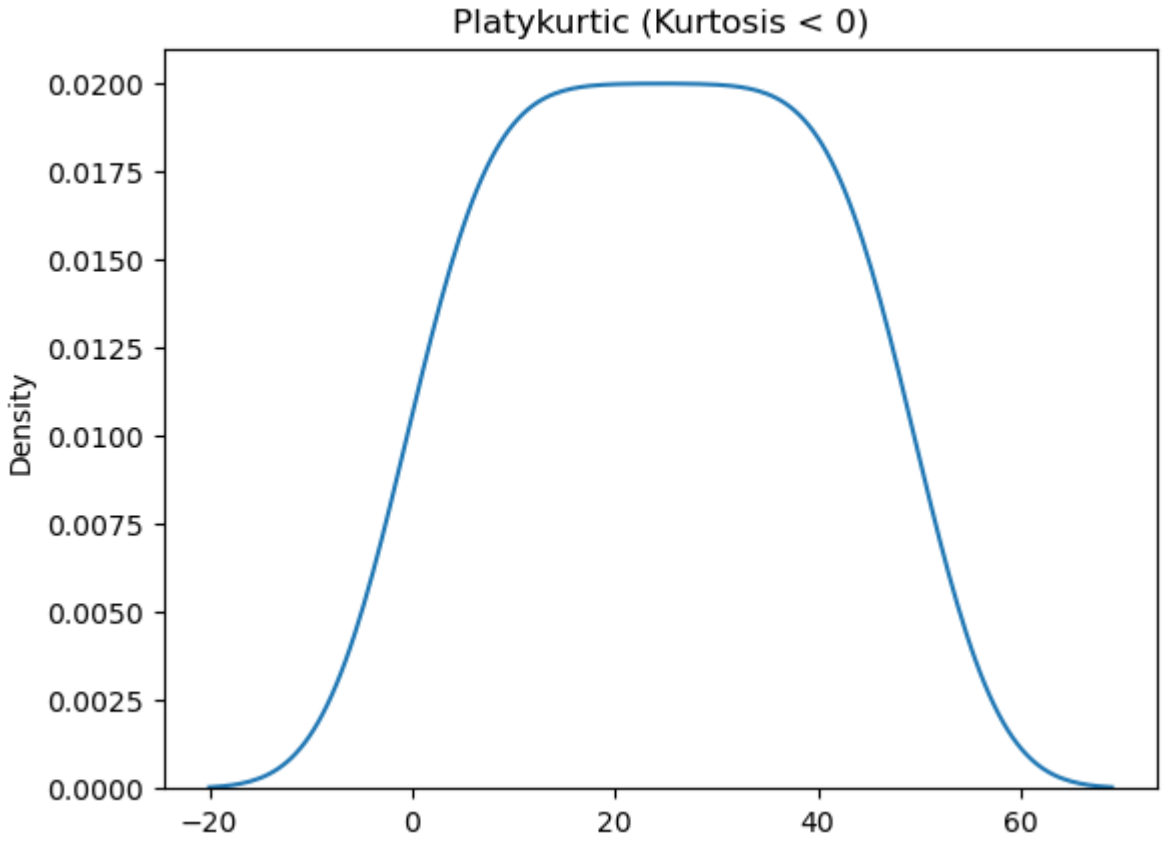


https://en.wikipedia.org/wiki/Kurtosis

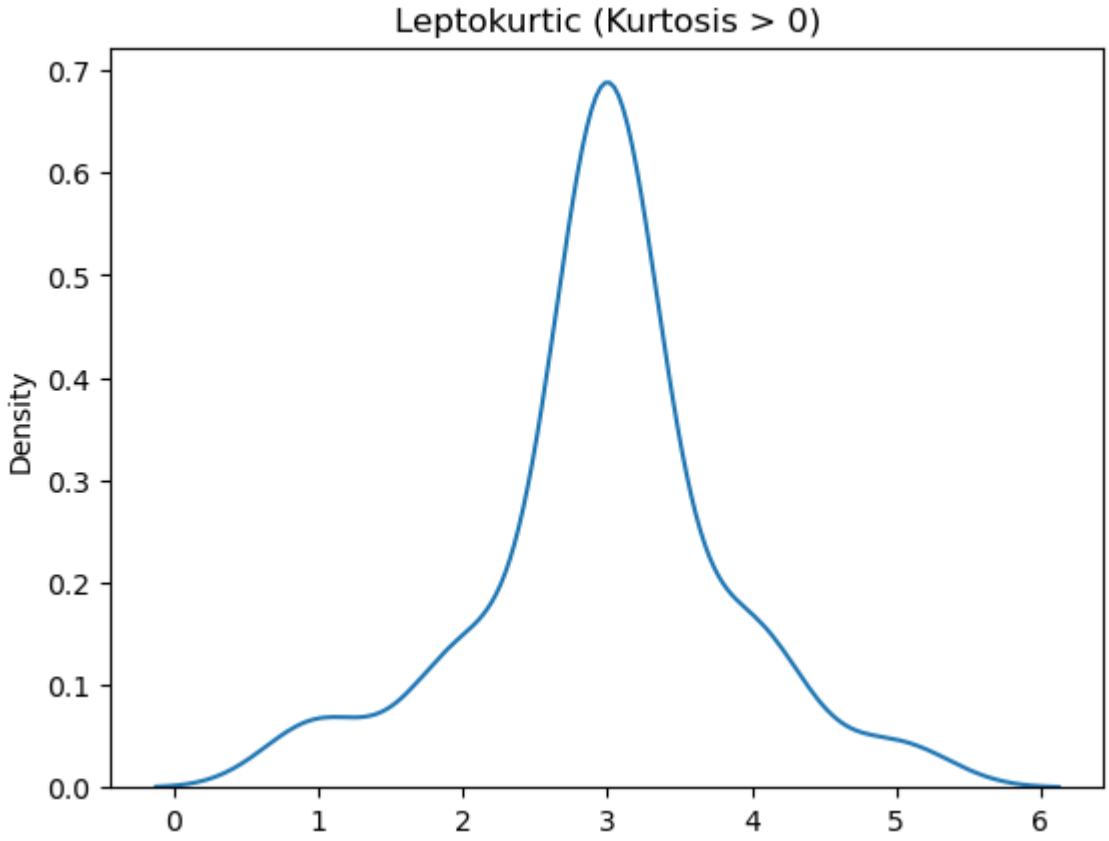
In [5]:
```
from scipy.stats import kurtosis
# Calculate kurtosis
def calculate_kurtosis(data):
    data_kurtosis = kurtosis(data)
    print("Kurtosis:", data_kurtosis)

    if data_kurtosis < 0:
        print("platykurtic (lighter tail) ")

    elif data_kurtosis > 0 :
        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)
```

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

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



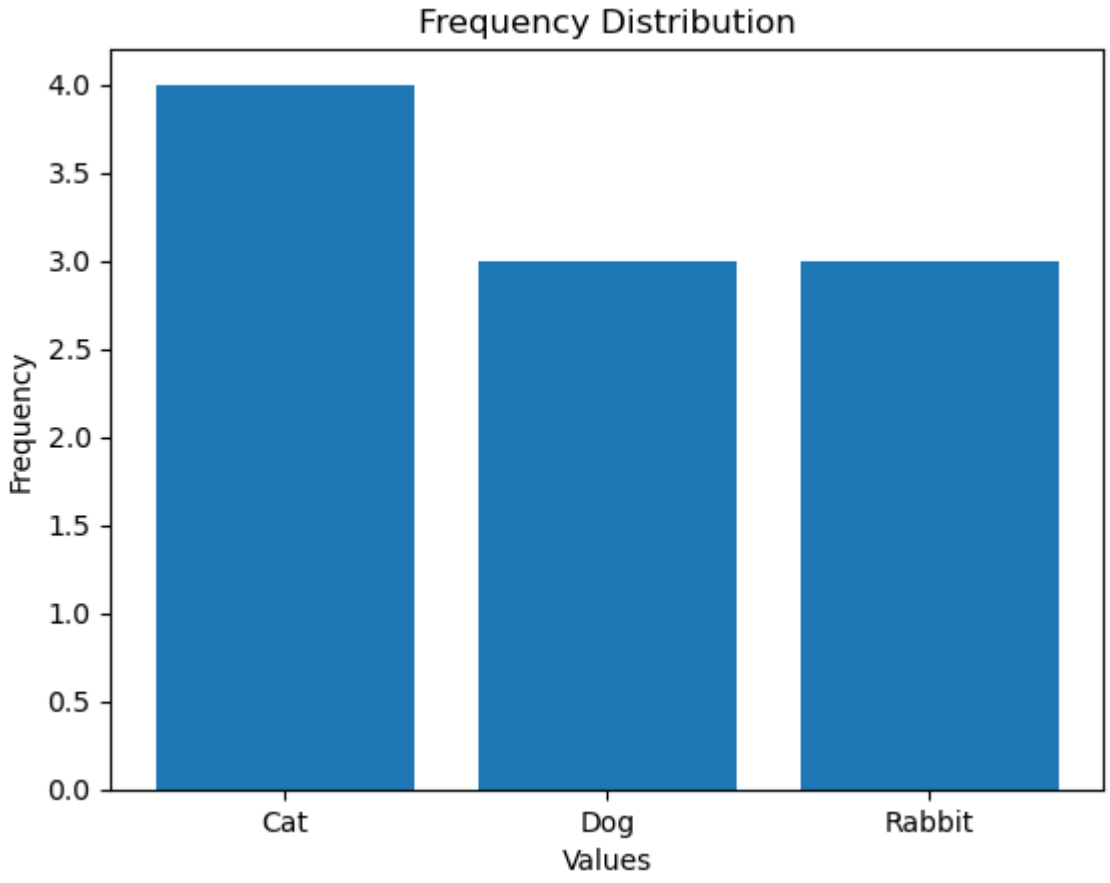
In [7]:
```
sns.kdeplot(data2)
plt.title('Leptokurtic (Kurtosis > 0)')
plt.show()
```



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.

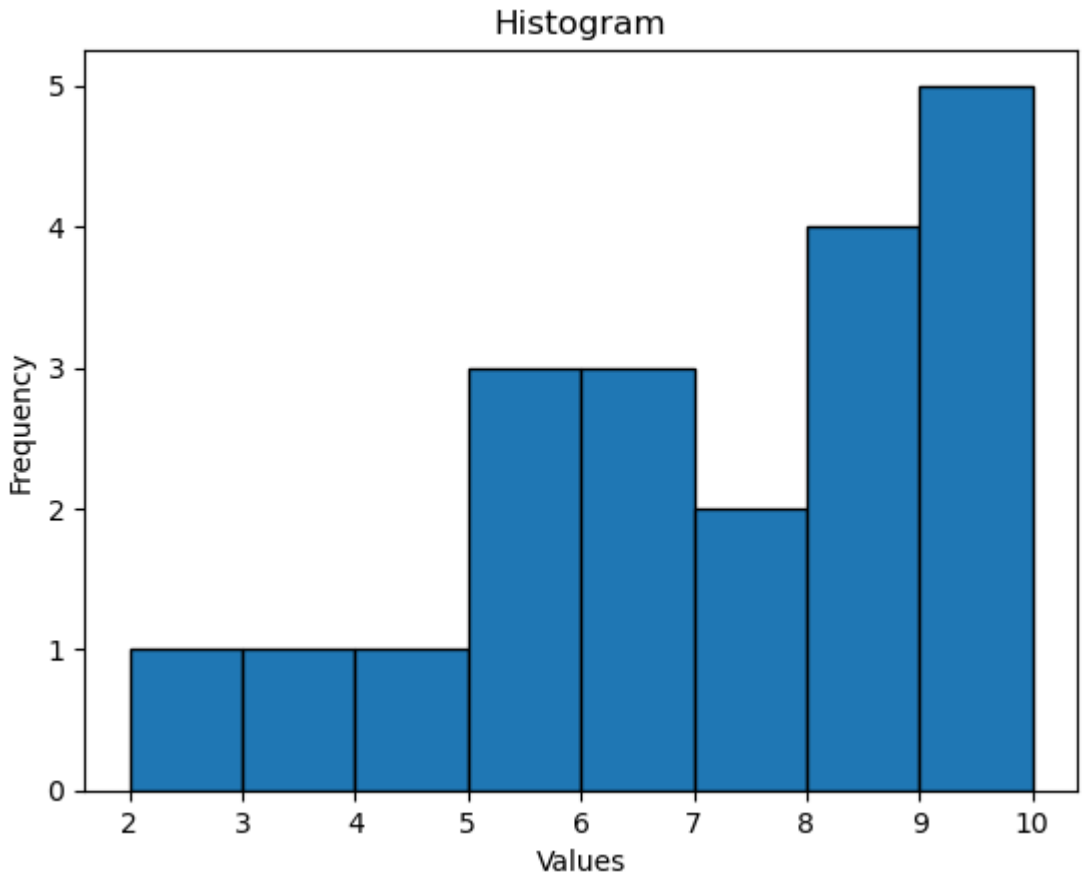
In [8]:
```
# sample data
data = np.array(['Cat', 'Dog', 'Cat', 'Dog', 'Cat', 'Rabbit', 'Dog', 'Rabbit', 'Rabbit',

# 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()
```

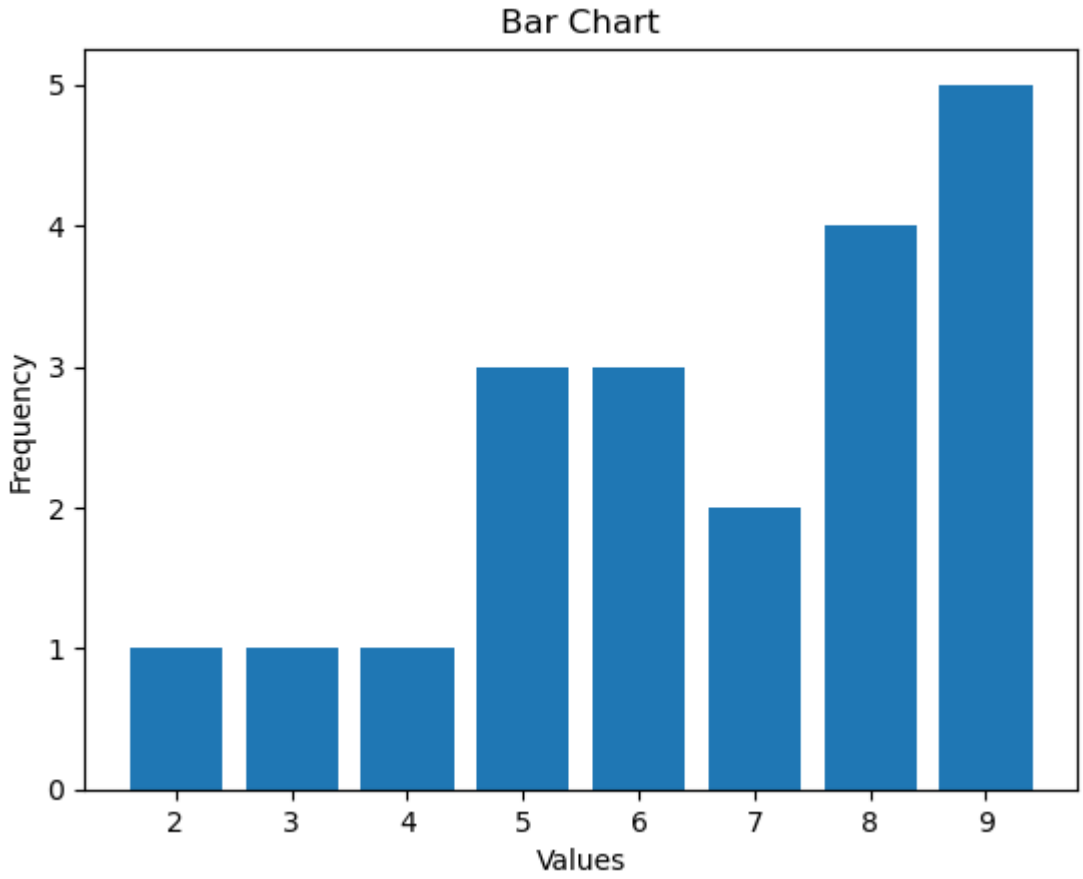


Histograms and Bar Charts:

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 [9]:
```
# 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
plt.hist(data, bins=range(min(data), max(data) + 2), edgecolor='black')
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()
```





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.

In [10]:
```
# Example data
data = np.array([2, 4, 6, 8, 10, 12, 14, 16, 18, 20])

# 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)
```

```
25th percentile: 6.5
50th percentile: 11.0
75th percentile: 15.5
```

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 [11]:
```
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
```

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:

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.

In [12]:
```
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
```

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.

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

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).

In [13]:
```
from scipy.stats import ttest_1samp

# sample data
data = np.array([10, 12, 15, 18, 20, 22, 25, 28, 30, 32, 35])

# Set the null hypothesis value
null_value = 25

# Perform one-sample t-test
t_statistic, p_value = ttest_1samp(data, null_value)

# Set significance level
```

```
alpha = 0.05

# Print the results
print("Null Hypothesis Value:", null_value)
print("Sample Mean:", np.mean(data))
print("T-Statistic:", t_statistic)
print("P-Value:", p_value)

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

```
Null Hypothesis Value: 25
Sample Mean: 22.454545454545453
T-Statistic: -1.0172742427190833
P-Value: 0.33300313911565127
Fail to reject the null hypothesis
```

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 [14]:
```
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:                Thu, 06 Jul 2023   Prob (F-statistic):           3.20e-05
Time:                        21:18:48   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 "
```

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.

In [15]:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

# sample
data = {
    'Group': ['A', 'A', 'A', 'B', 'B', 'B', 'C', 'C', 'C'],
    'Value': [10, 15, 20, 12, 18, 22, 8, 14, 16]
}

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

# Fit the ANOVA model
model = ols('Value ~ Group', data=df).fit()
anova_table = sm.stats.anova_lm(model)
```

```
# Print the ANOVA table
print(anova_table)
```

```
             df      sum_sq    mean_sq         F    PR(>F)
Group       2.0   32.666667  16.333333  0.724138  0.522741
Residual    6.0  135.333333  22.555556       NaN       NaN
```

Correlation Analysis:

Correlation analysis measures the strength and direction of the relationship between two variables. It provides insights into the extent to which changes in one variable are associated with changes in another variable. Example: Studying the correlation between hours of study and exam scores to determine if there is a relationship.

In [16]:
```
import pandas as pd

# sample data
data = {
    'X': [1, 2, 3, 4, 5],
    'Y': [2, 4, 6, 8, 10]
}

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

# Calculate the correlation coefficient
correlation = df['X'].corr(df['Y'])

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

```
Correlation Coefficient: 0.9999999999999999
```

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:

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

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:

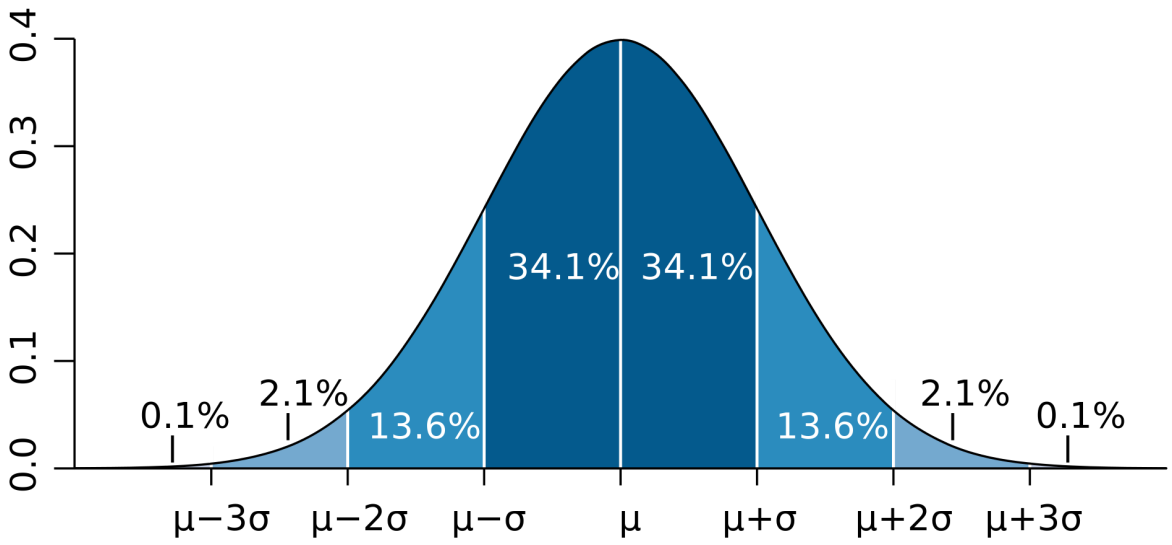
Normal Distribution (Gaussian Distribution):

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 / (σ √*(2π))*) e^(-((x - μ)^2) / (2σ^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).



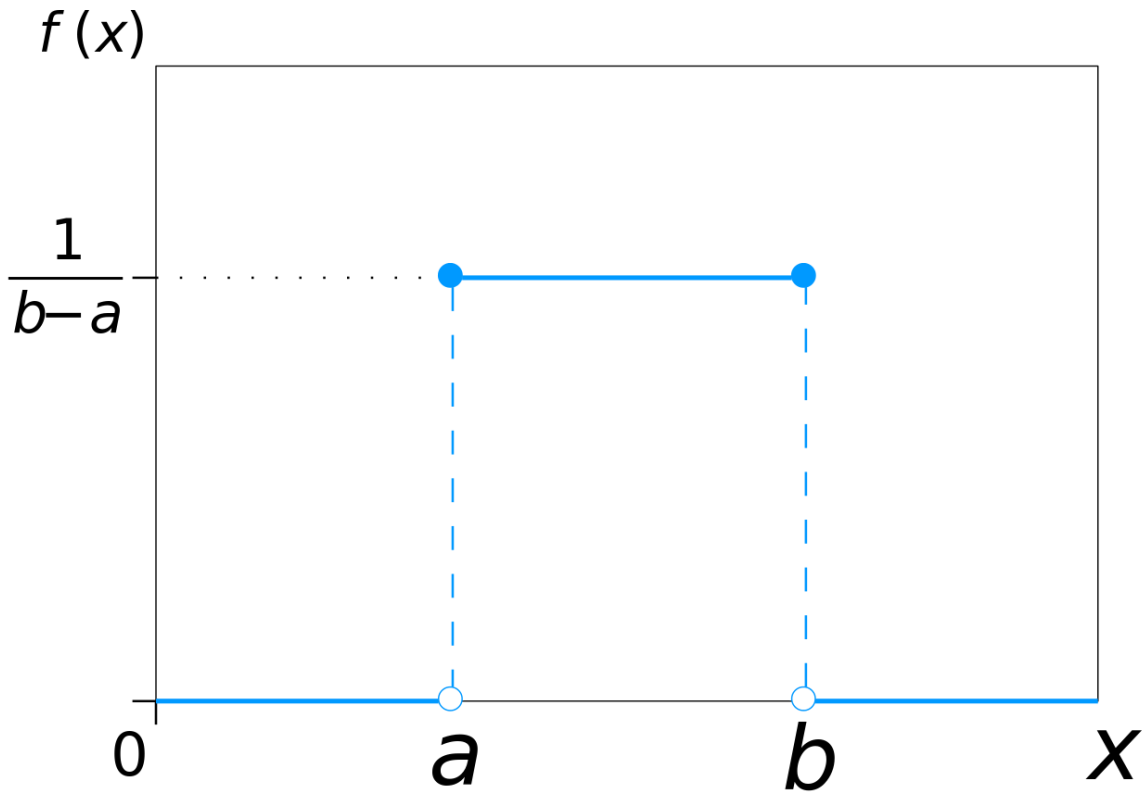
https://en.wikipedia.org/wiki/Normal_distribution

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 $a \leq x \leq b$

- 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

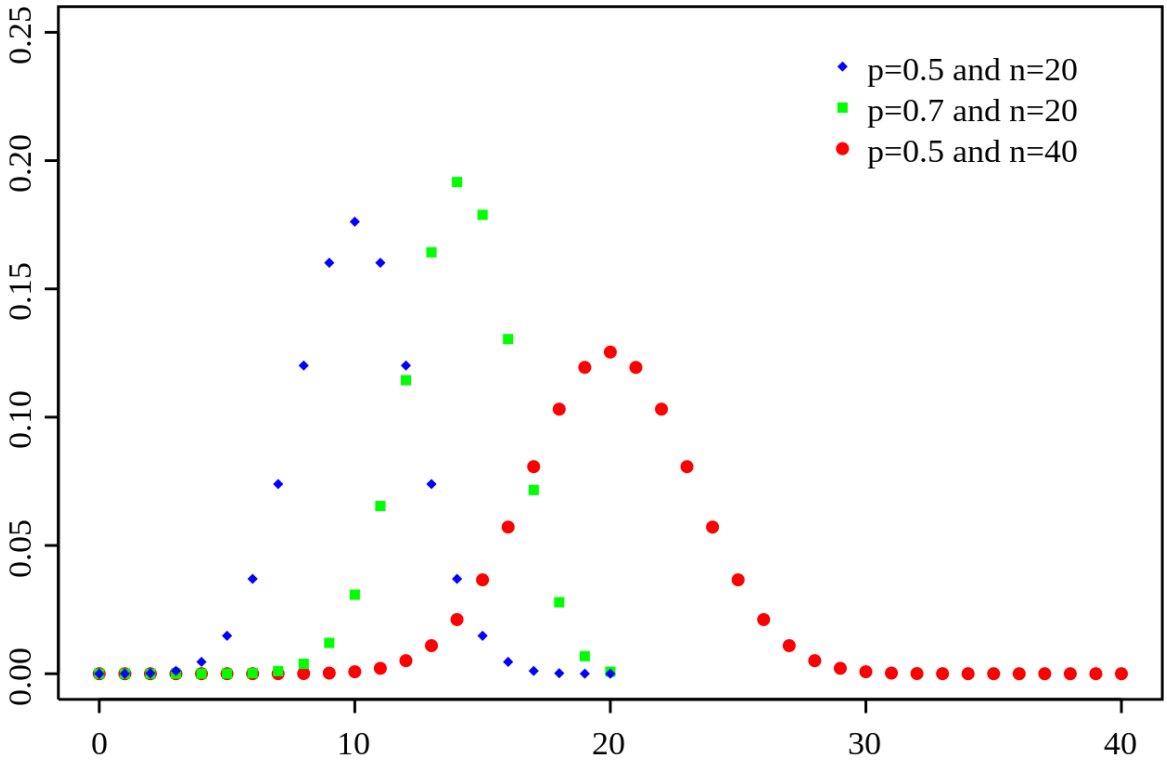
Binomial Distribution:

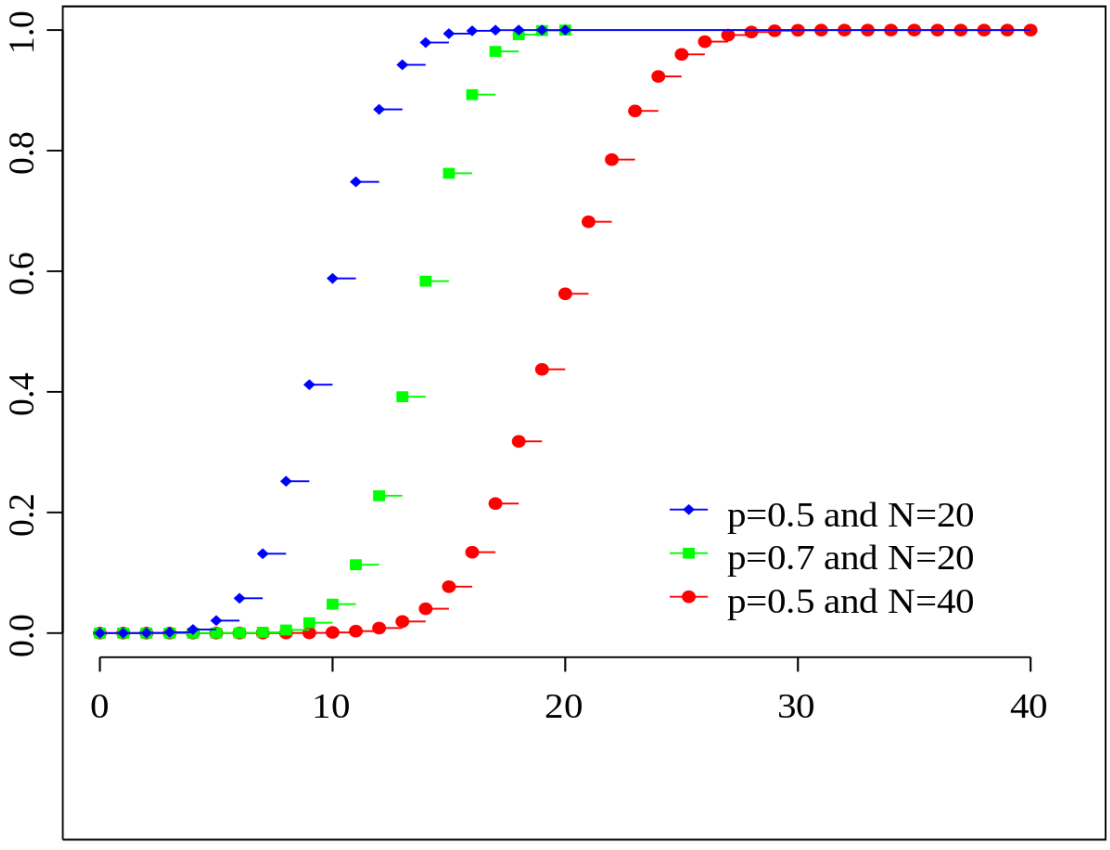
The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials (experiments with two possible outcomes: success or failure). It is characterized by two parameters: the number of trials (n) and the probability of success (p) in each trial. Examples include the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

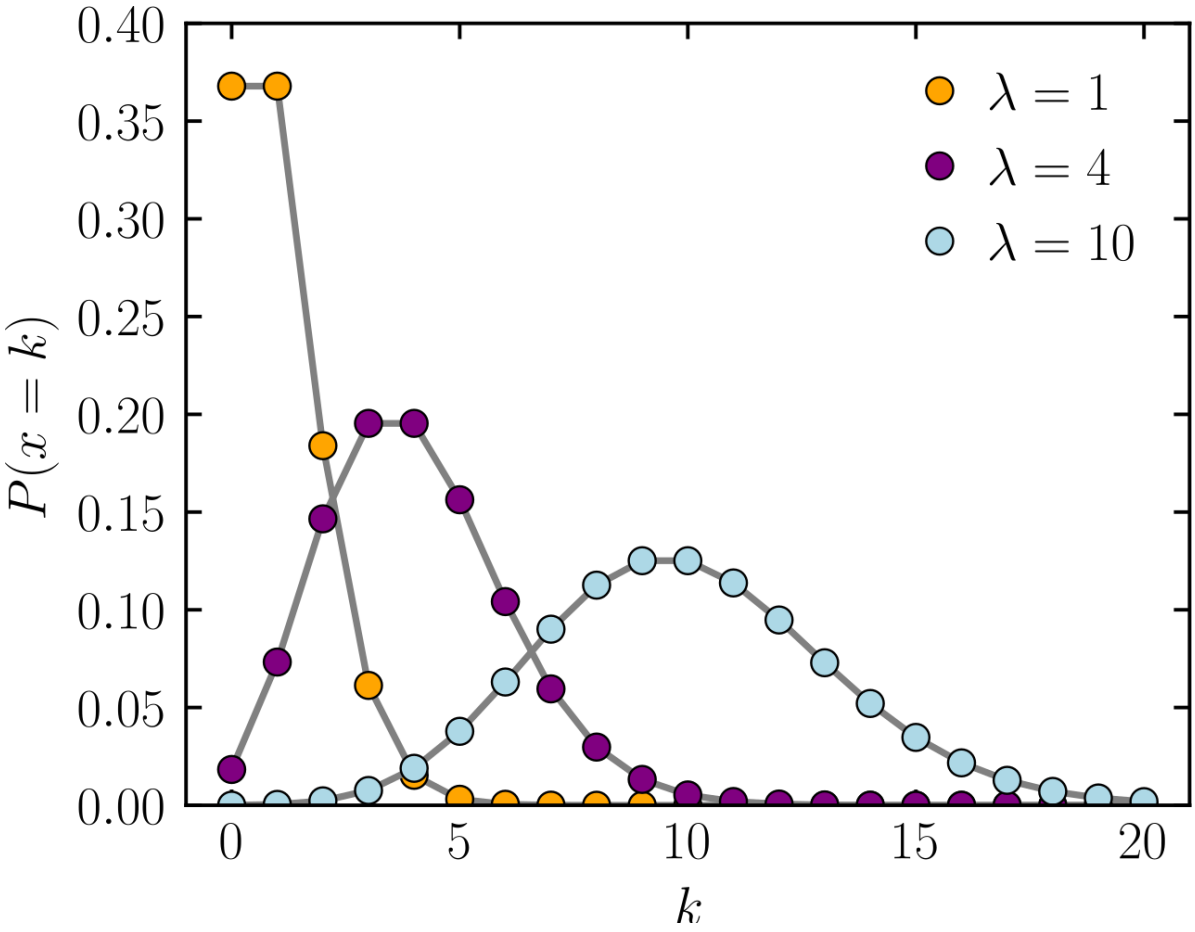
Poisson Distribution:

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.

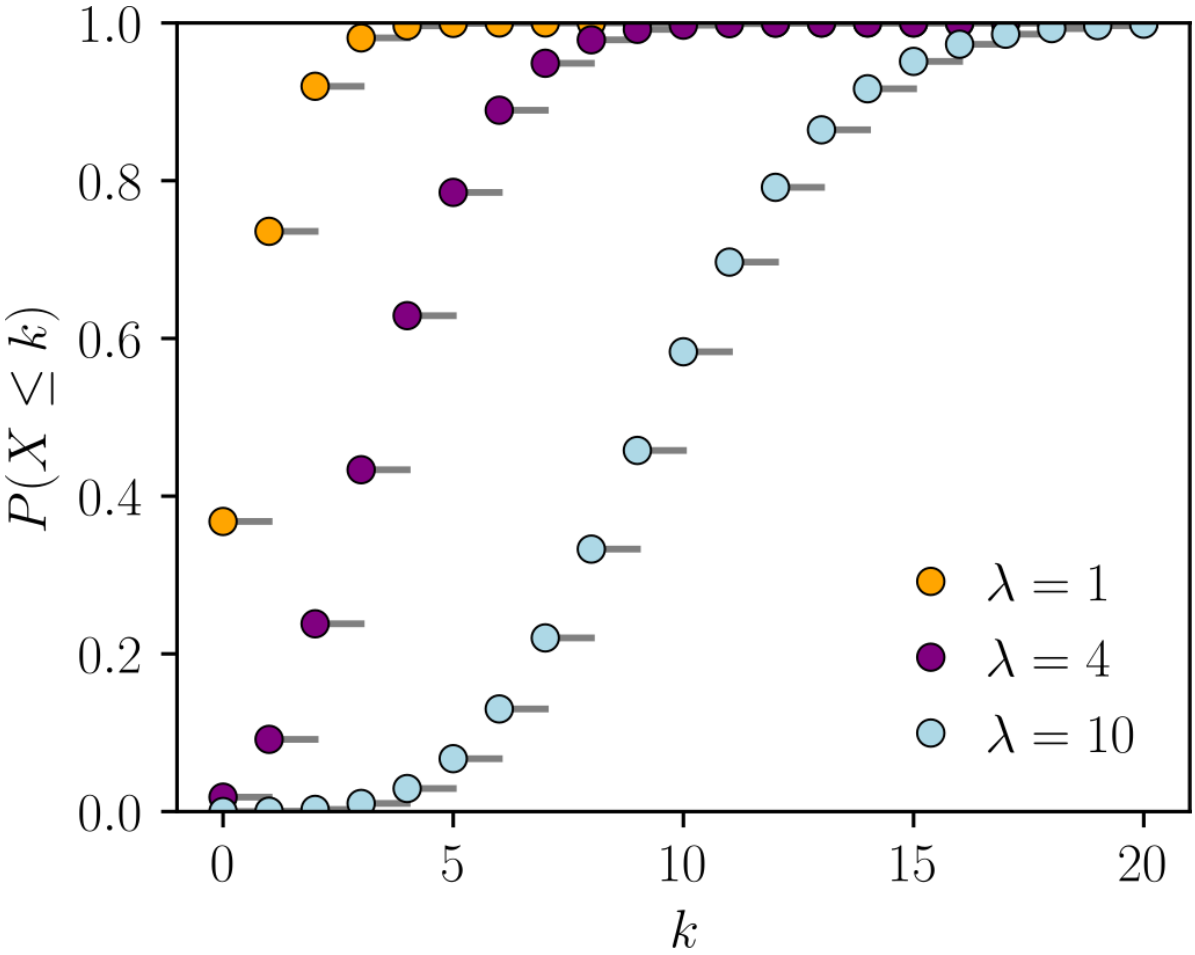
Probability mass function (PMF)

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.



'



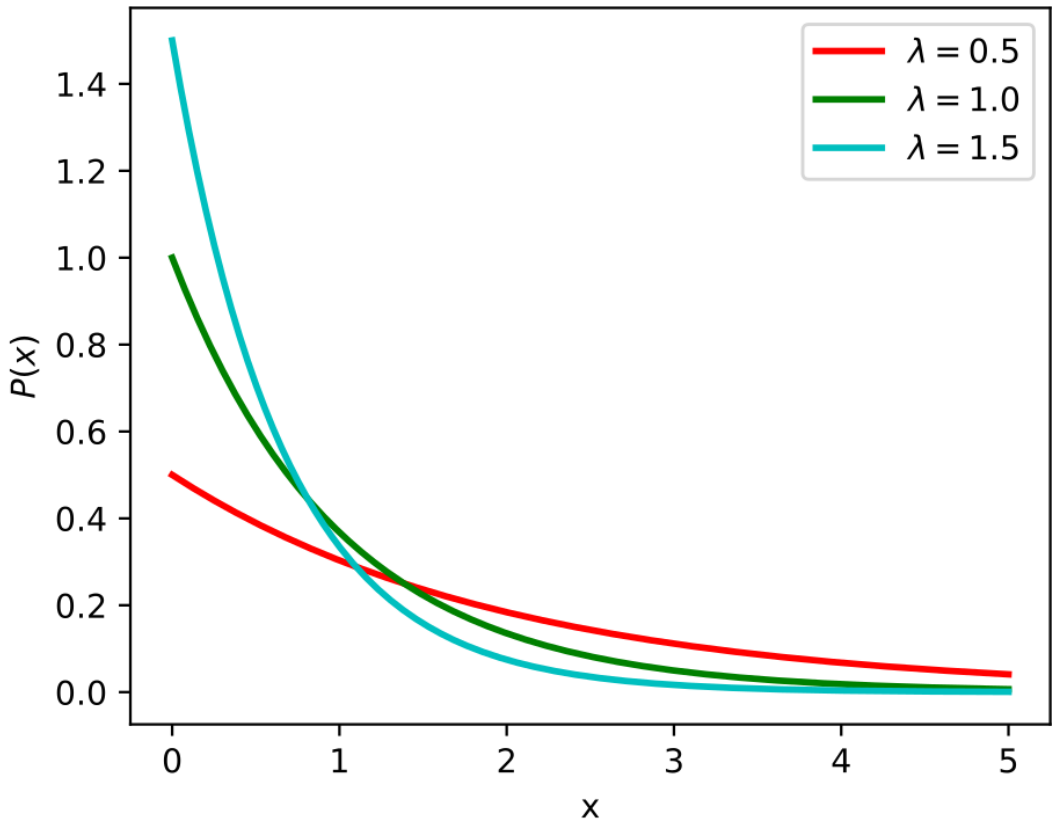
https://en.wikipedia.org/wiki/Poisson_distribution

Exponential Distribution:

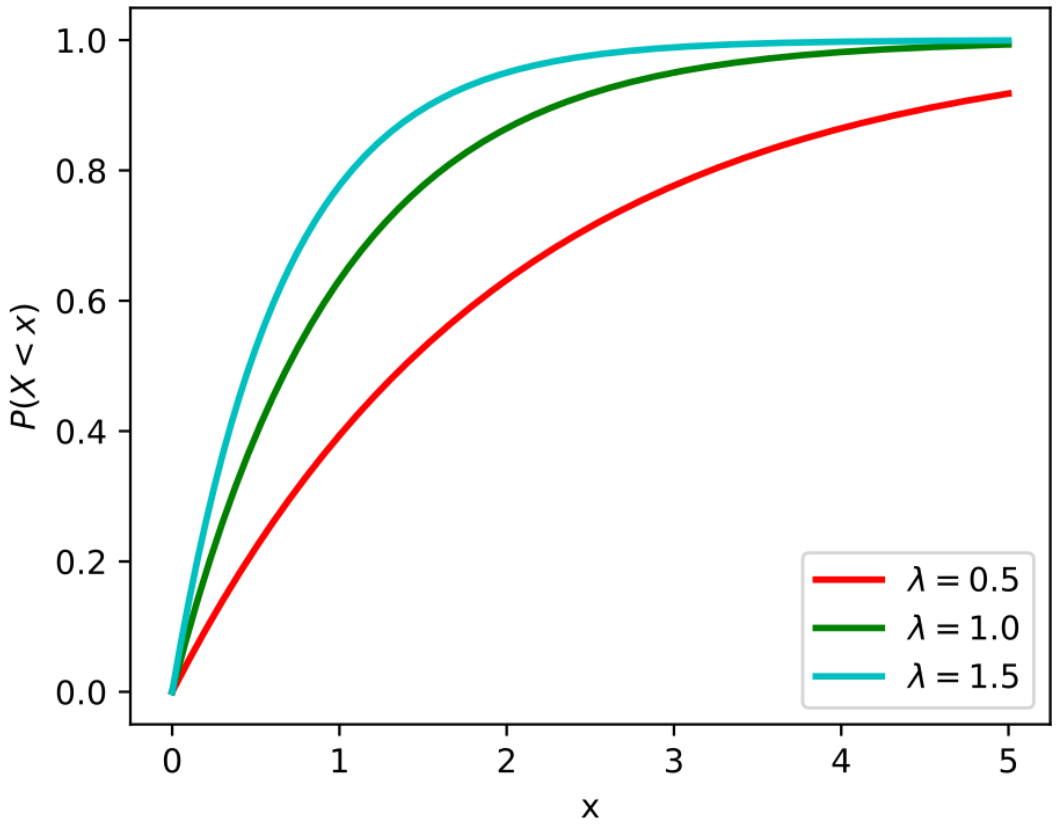
The exponential distribution models the time between events occurring in a Poisson process. It is characterized by a single parameter, λ (lambda), which represents the average rate at which events occur. The exponential distribution is commonly used in reliability analysis and queuing theory.

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

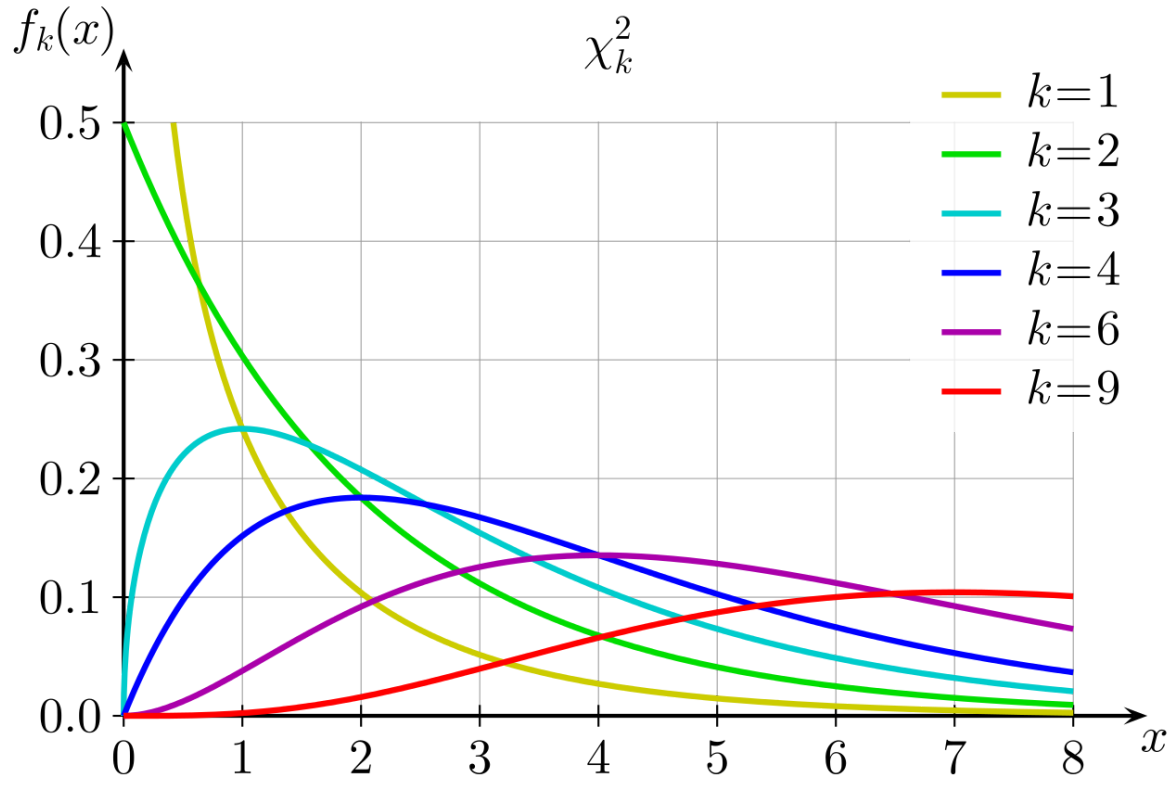
Chi-Square Distribution:

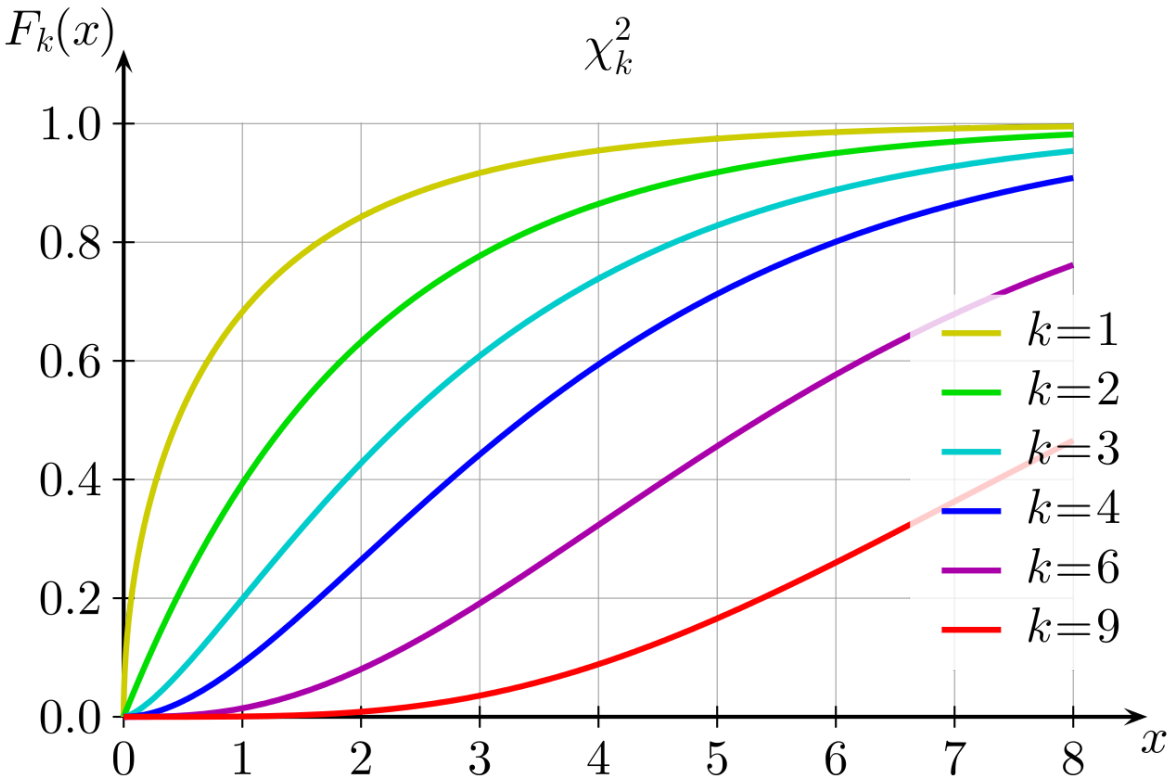
The chi-square distribution arises in various statistical tests, such as the chi-square test of independence or the chi-square test of goodness of fit. It is a right-skewed distribution and its shape depends on the degrees of freedom.

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

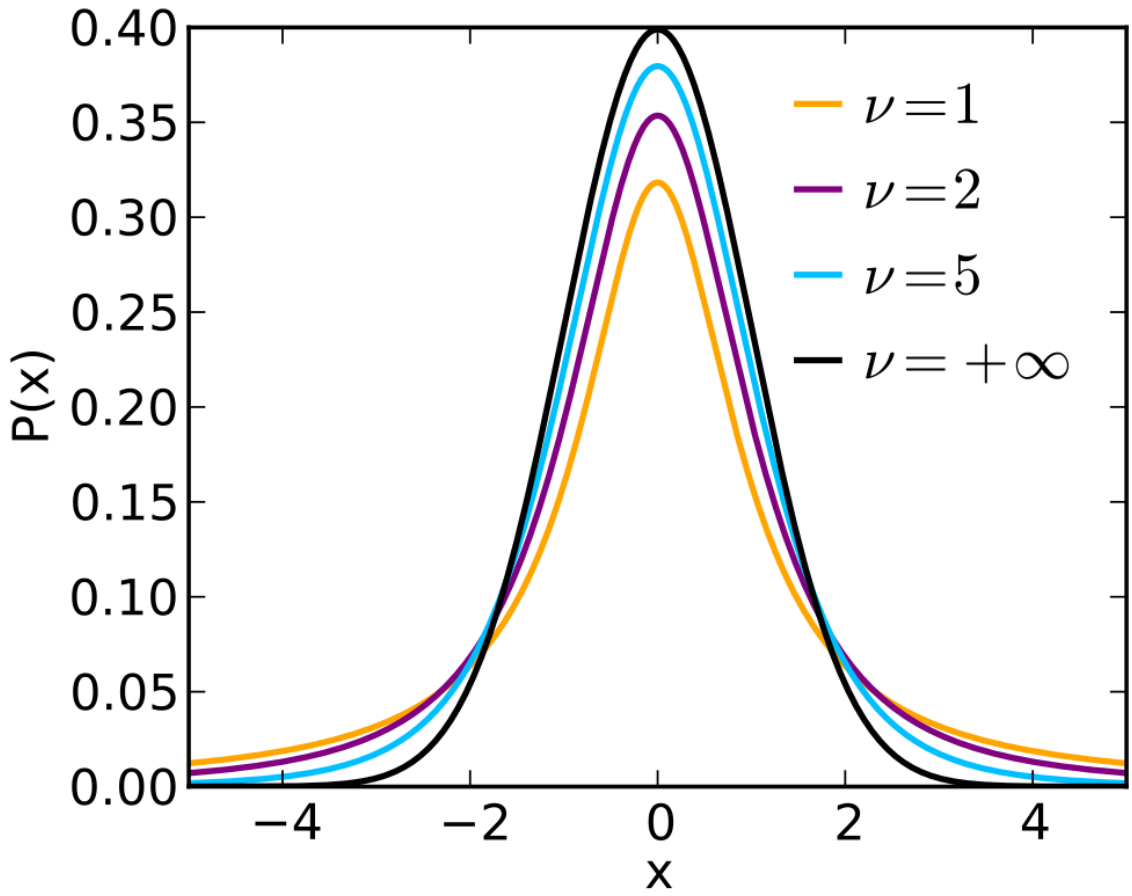
Student's t-Distribution:

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.

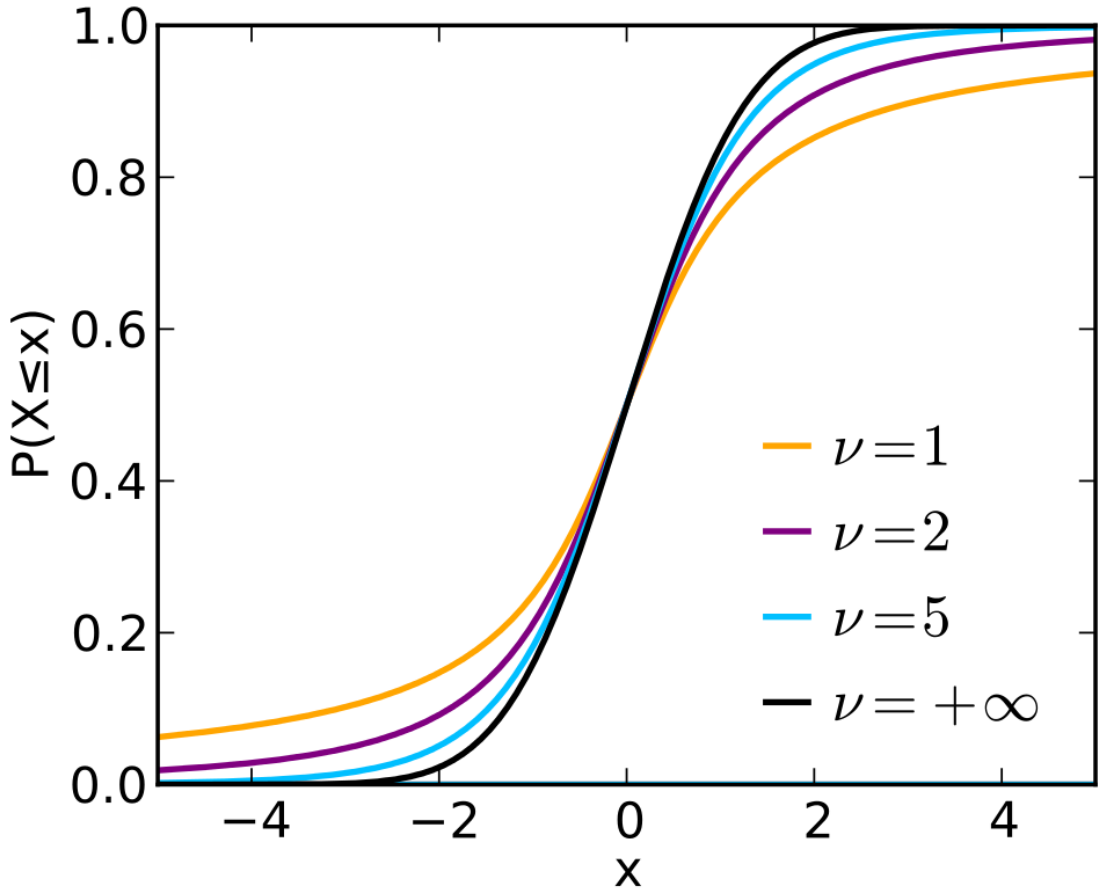
Probability density function (PDF)

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

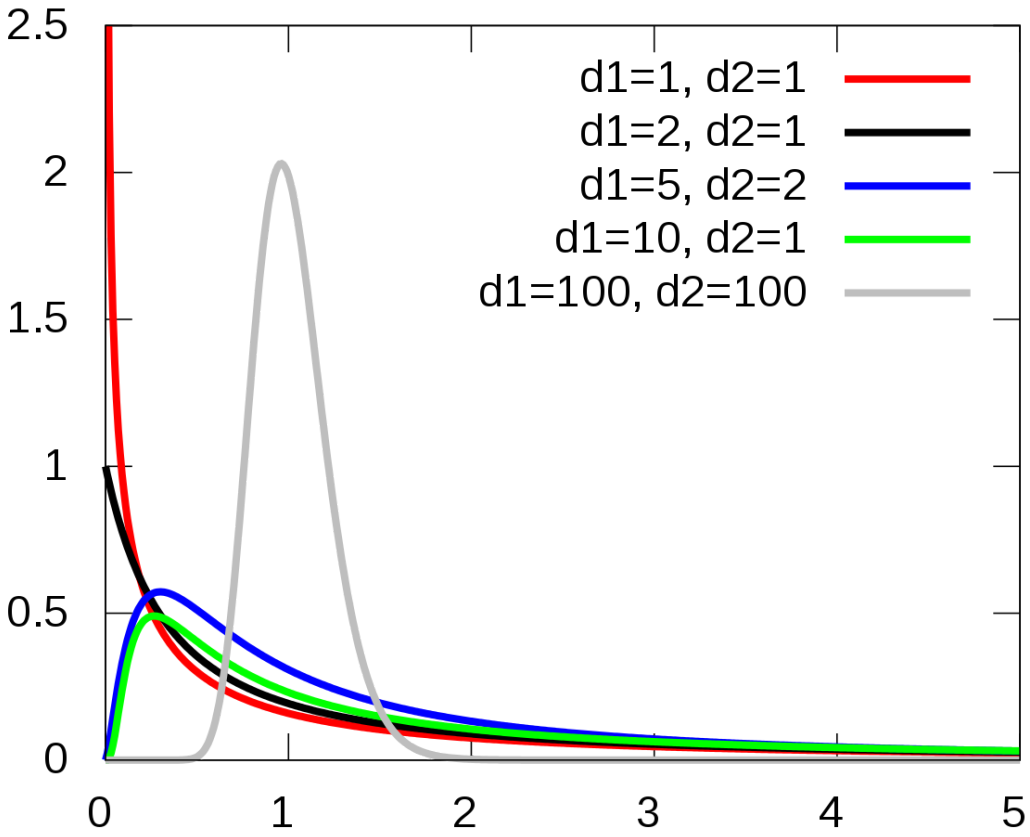
F-Distribution:

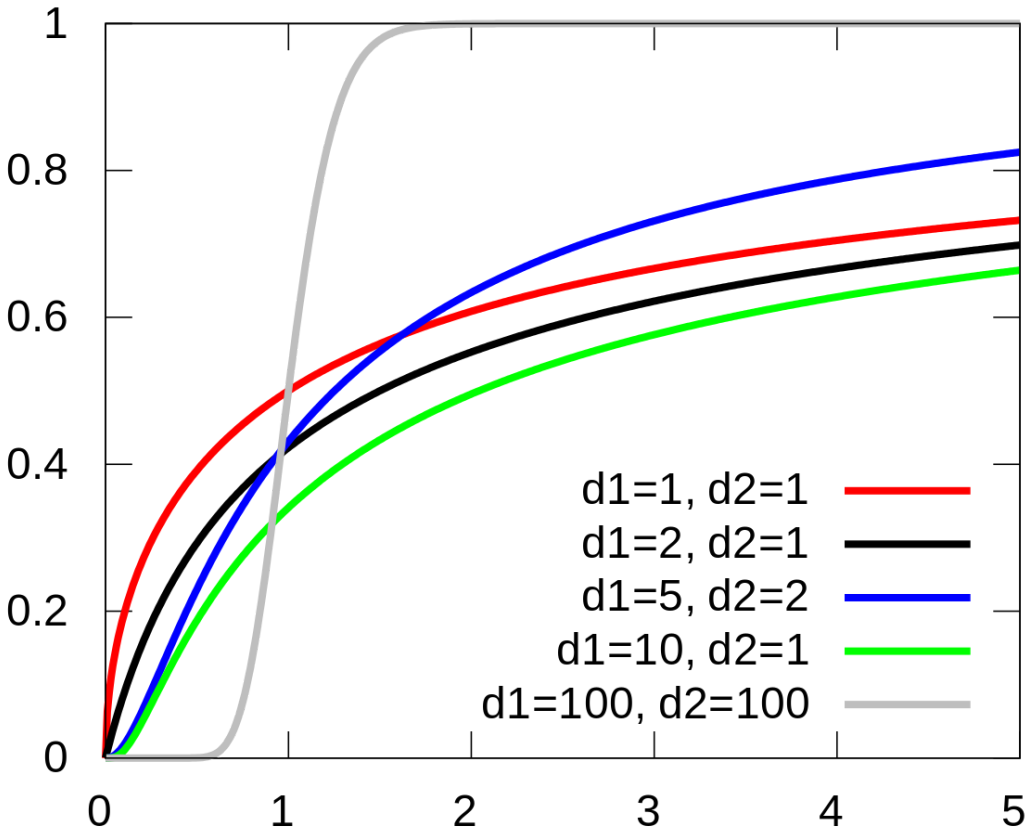
The F-distribution is used in the analysis of variance (ANOVA) and in comparing variances between groups. It arises when comparing the variances of two or more samples.

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

Sampling

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.