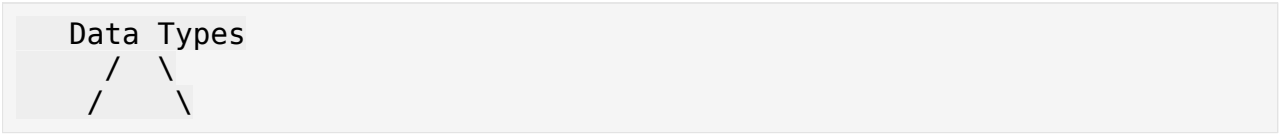
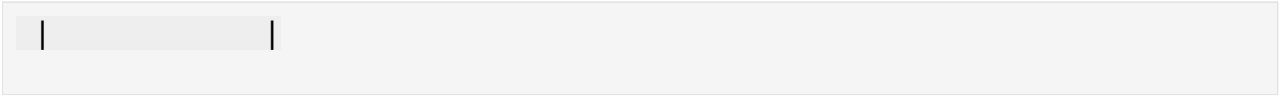
Introduction to Statistics

Statistics is a branch of mathematics focused on collecting, analyzing, interpreting, presenting, and organizing data. It is widely used in many fields to make decisions and predictions.

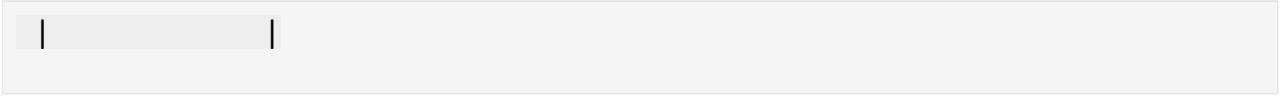
Example: All students in a university Example: 100 randomly chosen students |



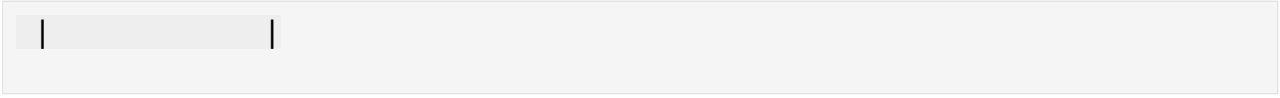
Qualitative Quantitative



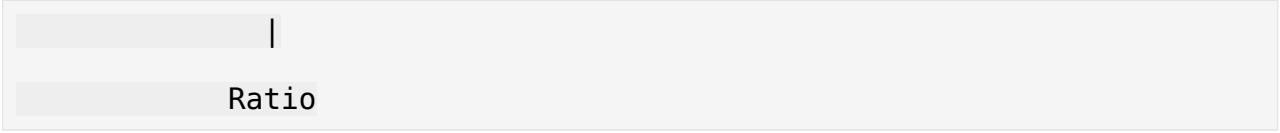
Nominal Discrete



Ordinal Continuous



Binomial Interval



Population(N)

Refers to the entire group you want to study.

Data from population = **Parameters**.

Example: All students in a university.

Sample(n)

A smaller group selected from the population.

Data from the sample = **Statistics**.

Example: 100 randomly chosen students from the university.

Characteristics of Sample:

Randomness: A sample is considered random if every individual in the population has an equal chance of being selected. This ensures that the sample is unbiased and reflects the diversity of the entire population.

Representativeness: A sample is representative if it accurately reflects the characteristics of the population from which it is drawn. This means that the sample should mirror the demographic and relevant attributes of the population, such as age, gender, income level, etc.

Sampling

Sampling is the process of selecting a subset of data from a population to estimate characteristics of the entire population.

Types of Sampling

1.**Probability Sampling**

Key Sampling Methods:

Simple Random Sampling: Every individual in the population has an equal chance of being selected.

Systematic Sampling: Selecting individuals at regular intervals from a random starting point.

Stratified Sampling: Divides the population into strata (groups) and randomly selects samples from each group. In stratified sampling, there is homogeneity within groups.

Cluster Sampling: Divides the population into clusters and randomly selects entire clusters for the sample. In cluster sampling, there is homogeneity between groups.

2.**Non-probability sampling**

Key Sampling Methods:

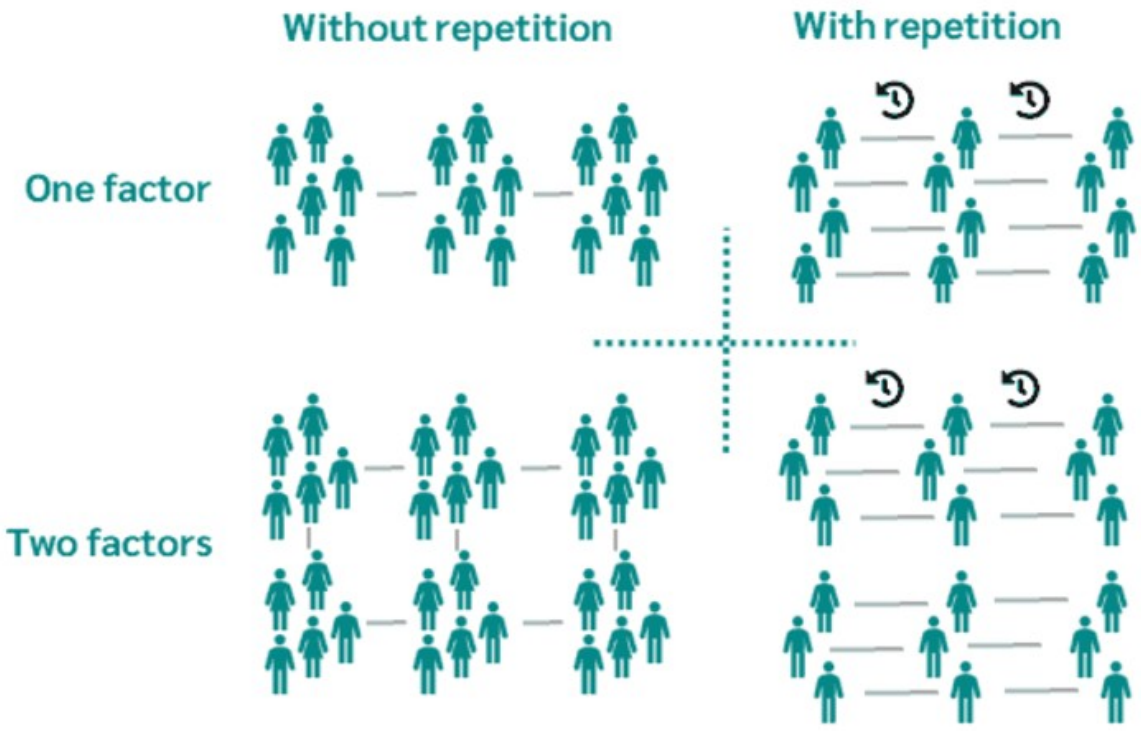
Convenience Sampling: Selecting individuals who are easiest to reach.

Judgmental (Purposive) Sampling: Selecting individuals based on the researcher's judgment.

Snowball Sampling: Participants recruit other participants from among their acquaintances.

Quota Sampling: Setting quotas for certain subgroups and selecting individuals until those quotas are met.

```
                        Sampling
                           |
          ------------------------------------
         |                                    |
 Probability Sampling                  Non-Probability Sampling
         |                                    |
 --------------------
----------------------------------------
|          |          |                    |          |          |
|
Simple   Systematic   Stratified     Convenience   Judgmental
Snowball        Quota
 Random    Sampling     Sampling         Sampling       Sampling
Sampling      Sampling
Sampling
```



Types of Statistics:

Descriptive Statistics: Summarizes and organizes data (e.g., mean, median, mode, standard deviation).

Inferential Statistics: Makes predictions or inferences about a population based on a sample.

Probability theory: Understanding uncertainty and randomness in data.

Hypothesis Testing: Assessing the validity of assumptions and conclusions.

Regression Analysis: Modeling the relationship between the variables and making predictions.

Types of data

Data can be categorized into two broad types: **Qualitative and Quantitative**.

1. **Qualitative (Categorical) Data**:

Data that represents characteristics or qualities. They can't be measured.

Nominal Data: Categories with no logical/ranking order (e.g., colors, types of fruits).

Ordinal Data: Categories with a logical/ranking order, but no fixed differences between them (e.g., rankings like 1st, 2nd, 3rd, education levels like "high school diploma," "bachelor's degree," "master's degree," "Ph.D.", ratings like good, poor, excellent).

Binomial Data: Binomial data represents outcomes of events where there are only two possible outcomes (e.g., success or failure, yes or no, pass or fail, heads or tail).

1. **Quantitative (Numerical) Data**:

Data that represents measurable quantities.

Discrete Data(Random Variable): Countable data (e.g., number of students in a class).

Continuous Data(Random Variable): Data that can take any value within a range (e.g., height, temperature).

Interval Data(Measurement Scale): Data with meaningful intervals between values but no true zero.(eg.,Temperature in Celsius (0 degrees does not mean "no temperature")).

Ratio Data(Measurement Scale): Data with all the properties of interval data, plus a true zero point.(eg.,Weight (0 kg means "no weight") or age (0 years means "not born")).

Key measures of descriptive statisitcs

Measures of Central Tendency:

Mean: The average value of a dataset. Formula: Mean = (Sum of all values) / (Number of values) Example: For the dataset {1, 2, 3, 4, 5}, the mean is (1+2+3+4+5)/5 = 3.

Median: The middle value in a dataset when the values are arranged in ascending order. Example: For the dataset {1, 2, 3, 4, 5}, the median is 3.

Mode: The most frequent value in a dataset. Example: For the dataset {1, 2, 2, 3, 4}, the mode is 2.

Measures of Dispersion (Variability)

Range: The difference between the largest and smallest values in a dataset. Formula: Range = Maximum value - Minimum value Example: For the dataset {1, 2, 3, 4, 5}, the range is 5 - 1 = 4.

Variance: The average squared deviation from the mean. It measures how far each number in the set is from the mean (average), and thus from every other number in the set.

Formula: Variance = (Sum of squared differences from the mean) / (Number of values - 1) $\sigma 2 = \sum (xi - \bar{x})2/n$ - population variance x - mean , n -population $\sigma 2 = \sum (xi - \bar{x})2/(n - 1)$ - sample variance

Example: For the dataset {1, 2, 3, 4, 5}, the variance is [(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2] / (5-1) = 2.5.

Standard Deviation: The square root of the variance. It is a statistical measurement that shows how far data points are spread out from the mean of a data set

Formula: Standard Deviation = √(Variance) Example: For the dataset with a variance of 2.5, the standard deviation is $\sqrt{2.5} \approx 1.58$.

Interquartile Range(IQR): It is a measure of statistical dispersion that represents the range within which the middle 50% of the data falls. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1).

$IQR=Q3-Q1$ Where: Q1 is the first quartile (25th percentile), which divides the lowest 25% of the data from the rest. Q3 is the third quartile (75th percentile), which divides the lowest 75% of the data from the highest 25%. Steps to Calculate IQR:

Order the Data: Sort the data in ascending order. Find Q1: The median of the lower half of the data (excluding the median if the number of data points is odd). Find Q3: The median of the upper half of the data. Subtract Q1 from Q3

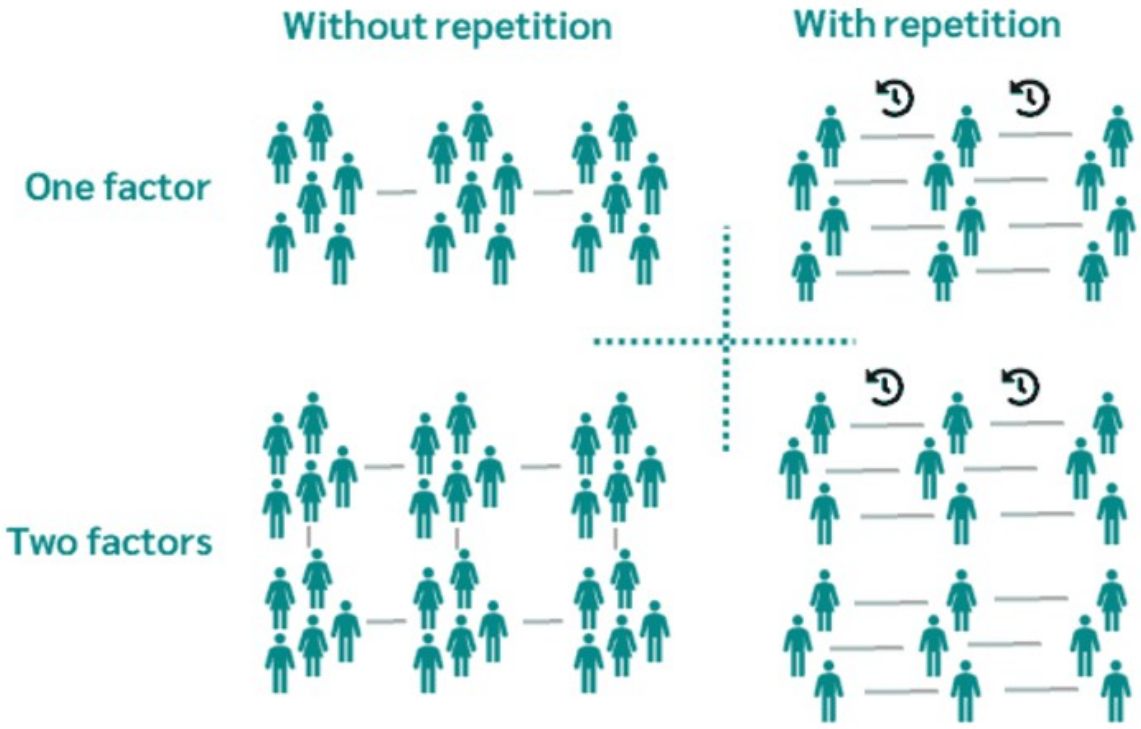
Lower bound = Q1-1.5*IQR

Upper bound = Q3+1.5*IQR

Anything beyond upper and lower bound are outliers

Empirical Rule

The empirical rule, also known as the 68-95-99.7 rule, represents the percentages of values within an interval for a normal distribution. That is, 68 percent of data is within one standard deviation of the mean; 95 percent of data is within two standard deviation of the mean and 99.7 percent of data is within three standard deviation of the mean.


Without repetition
With repetition
One factor
Two factors

QUARTILES vs PERCENTILES

Quartiles	Percentiles
Quartiles divide a dataset into four equal parts, denoted as Q1, Q2, and Q3.	Percentiles divide a dataset into 100 equal parts, ranging from the 1st percentile to the 99th percentile.
• Q1 represents the lower quartile and marks the boundary below which 25% of the data falls. • Q2 represents the median and marks the boundary below which 50% of the data falls. • Q3 represents the upper quartile and marks the boundary below which 75% of the data falls.	• The nth percentile represents the value below which n% of the data falls. • For example, the 25th percentile is the value below which 25% of the data falls, and the 75th percentile is the value below which 75% of the data falls.
Consider the following dataset of exam scores: 60,70,75,80,85,90,92,95,98,100. Quartiles: Q1 = 75 (25th percentile) Q2 = 87.5 (50th percentile) Q3 = 95 (75th percentile)	Consider the following dataset of exam scores: 60,70,75,80,85,90,92,95,98,100. Percentiles: 25th percentile = 75 50th percentile = 87.5 75th percentile = 95

Measures of shapes

Skewness : Describes asymmetry in data distribution. skewness = $\Sigma(xi-\bar{x})3/N*\sigma3$

In a perfectly symmetrical distribution (like a normal distribution), the data points are evenly spread around the mean. However, in skewed distributions, data points are concentrated more on one side of the distribution, leading to a skew.

Kurtosis : Describes the "tailedness" of data distribution. Kurtosis = $\Sigma(xi-\bar{x})4/N*\sigma4 -3$

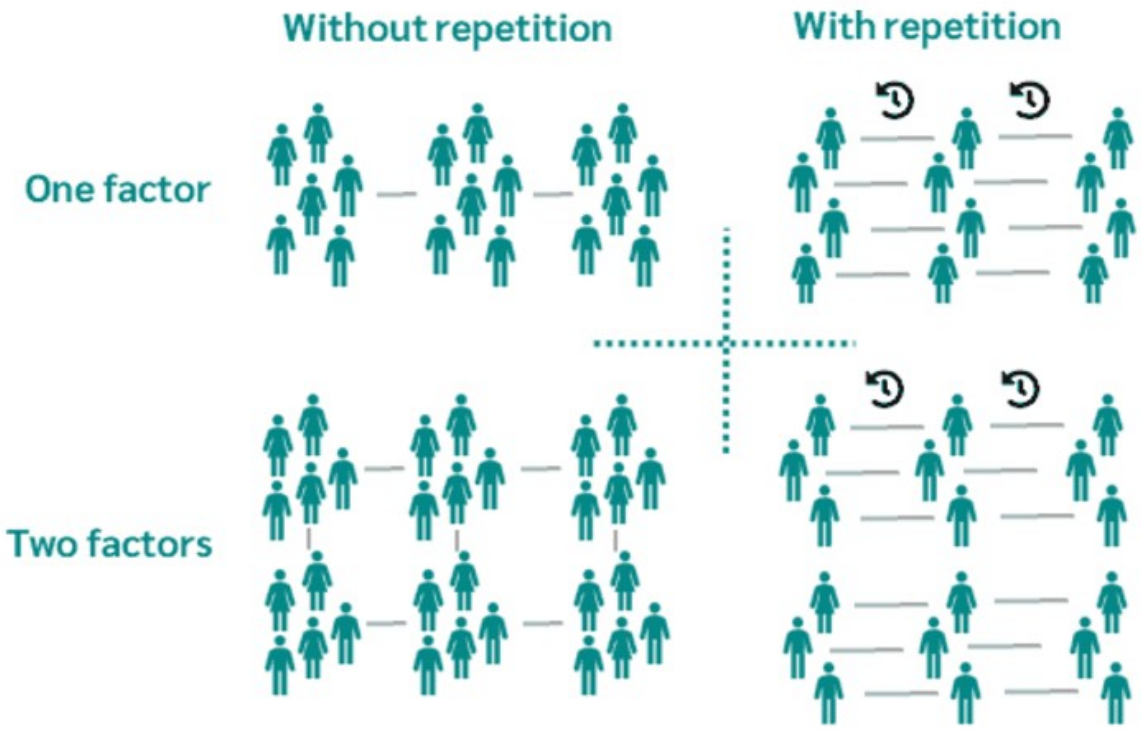
Kurtosis refers to the "tailedness" or the sharpness of the peak of a distribution compared to a normal distribution. It helps in understanding whether a distribution has heavier or lighter tails than the normal distribution.

Types of skewness

Positive Skew (Right-Skewed): Tail on the right; mean < median < mode.

Negative Skew (Left-Skewed): Tail on the left; mean > median > mode.

No Skew (Symmetrical): No tails; mean = median = mode (normal distribution).

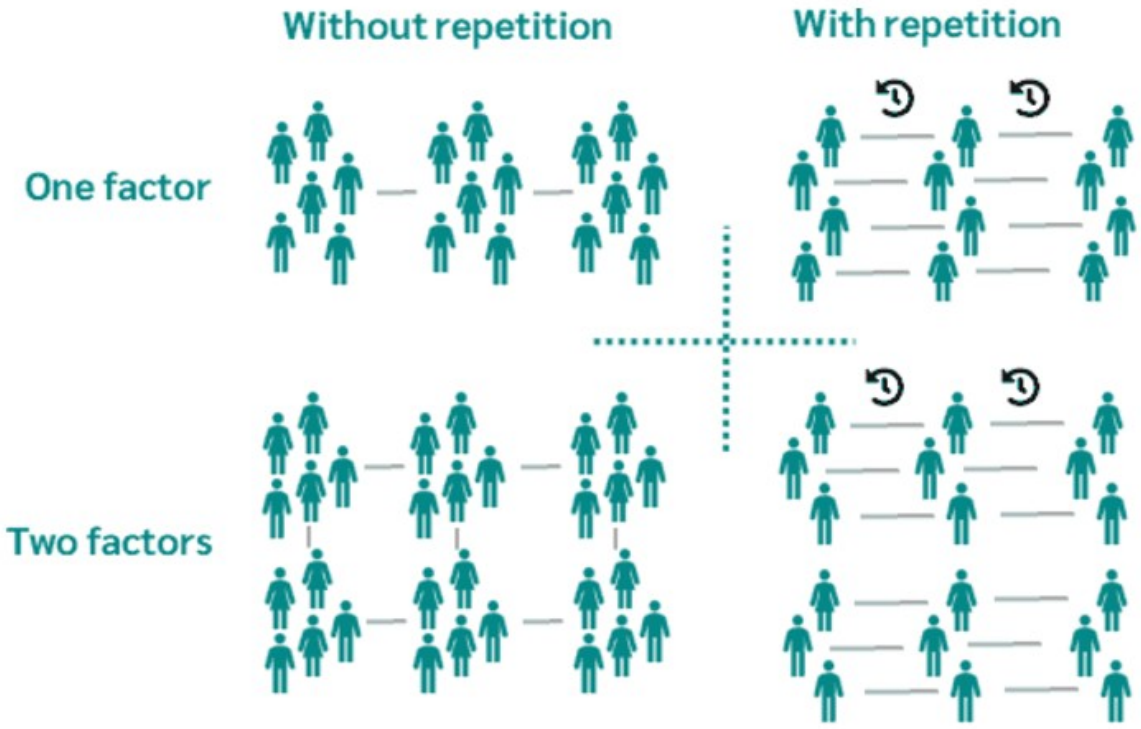


Types of Kurtosis

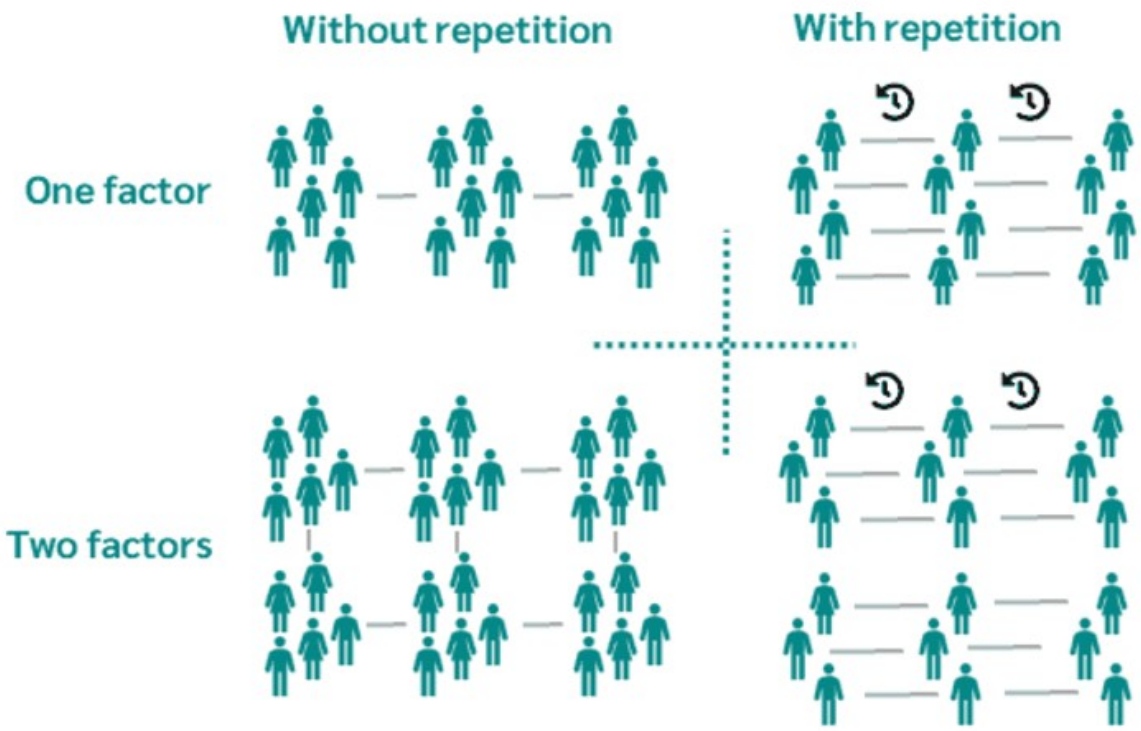
Mesokurtic(Normal): Kurtosis is close to that of a normal distribution (moderate peak and tails). Kurtosis = 3

Leptokurtic(Positive Kurtosis): Kurtosis is greater than a normal distribution (sharper peak, heavier tails, more outliers). Kurtosis >3

Platykurtic(Negative Kurtosis): Kurtosis is less than a normal distribution (flatter peak, lighter tails, fewer outliers).Kurtosis < 3


Without repetition
With repetition
One factor
Two factors

Difference between skewness and kurtosis



Modality

It refers to the number of peaks(modes) in a distribution.

Types:

Unimodal distributions have one peak.

Bimodal distributions have two peaks.

Multimodal distributions have more than two peaks

Unimodal Distribution:

A distribution with one peak.

Example: A **normal distribution (bell curve) is unimodal**, where the single peak represents the most common value (mean) in the dataset.

Visualization: A single peak in the **histogram or density plot**.

Common examples:

Heights of adults in a population.

IQ scores.

Bimodal Distribution:

A distribution with two distinct peaks.

Bimodal distributions often suggest that the data is drawn from two different populations or sources.

Visualization: Two distinct peaks in the **histogram or density plot**.

Common examples:

Exam scores of students from two different classes.

Daily temperatures measured in two different climates (e.g., summer and winter temperatures).

Multimodal Distribution:

A distribution with more than two peaks.

Multimodal distributions indicate multiple groups or clusters within the data, each with their own concentration of frequent values.

Visualization : More than two peaks in the **histogram or density plot**.

Common examples:

Sales data for different categories of products.

Population data with multiple age groups or clusters.

Uniform Distribution:

A distribution with no peaks; every value is equally likely.

Visualization: A flat line in a histogram or density plot.

Common examples:

Rolling a fair die (each outcome has equal probability).

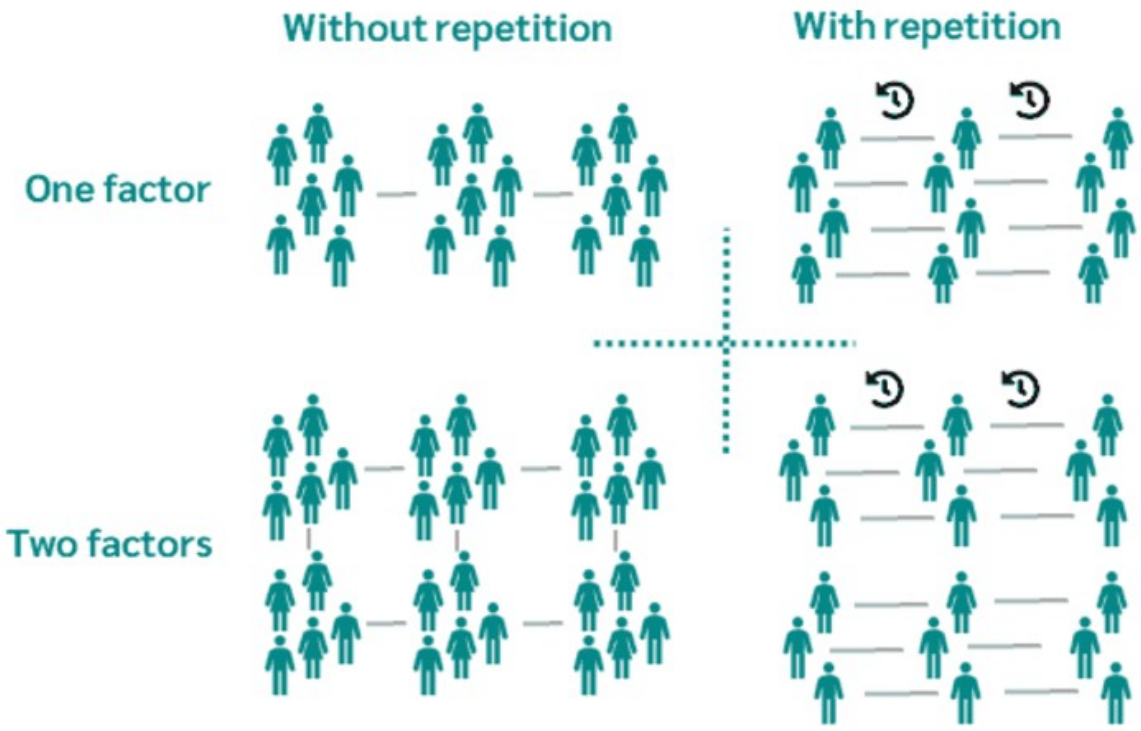
Random assignment of subjects in an experiment.

Importance of Understanding Modality

Unimodal distributions suggest a single dominant trend or feature in the data.

Bimodal or multimodal distributions may suggest that the data is a mixture of different groups or categories, leading to insights about clusters or sub-populations.

Uniform distributions indicate that no specific values are favored over others.



Types of variables

Independent Variable (X):

Also known as the predictor variable or explanatory variable.

It is the variable that you manipulate or categorize to see how it affects another variable.

In equations and graphs, it is often represented on the horizontal axis (x-axis).

Dependent Variable (Y):

Also known as the response variable or outcome variable.

It is the variable that you measure or observe to see how it responds to changes in the independent variable.

In equations and graphs, it is often represented on the vertical axis (y-axis).

Example 1: Iris Dataset The Iris dataset contains measurements of iris flowers, with the following columns:

sepal_length: The length of the sepal (independent variable).

sepal_width: The width of the sepal (dependent variable).

petal_length: The length of the petal.

petal_width: The width of the petal.

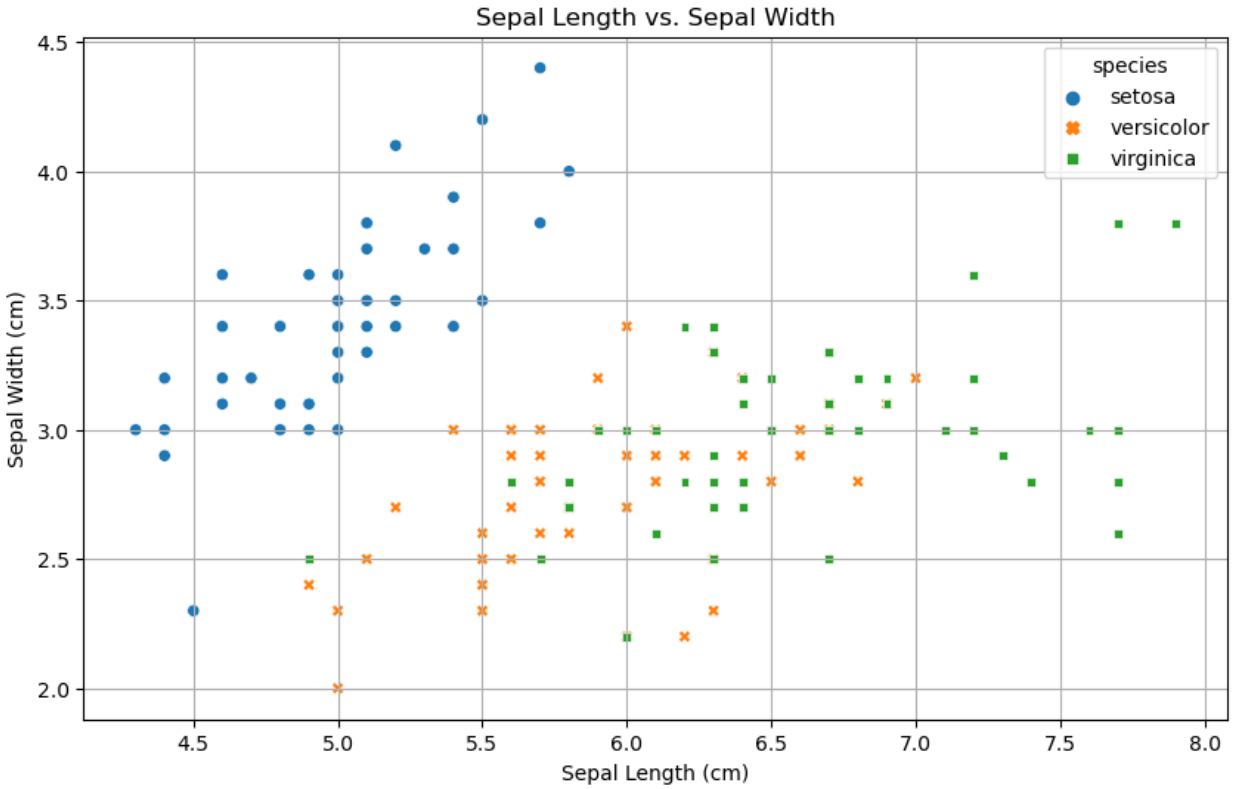
species: The species of the iris flower (categorical variable).

Example Scenario Suppose you want to study how the sepal length influences the sepal width of iris flowers. In this case:

Independent Variable: sepal_length (the length you manipulate or categorize).

Dependent Variable: sepal_width (the outcome you measure based on sepal length).

```
import seaborn as sns
import matplotlib.pyplot as plt

# Load the Iris dataset
iris = sns.load_dataset('iris')

# Create a scatter plot
plt.figure(figsize=(10, 6))
sns.scatterplot(data=iris, x='sepal_length', y='sepal_width',
hue='species', style='species')
plt.title('Sepal Length vs. Sepal Width')
plt.xlabel('Sepal Length (cm)')
plt.ylabel('Sepal Width (cm)')
plt.grid(True)
plt.show()
```



Example 2: Tips Dataset The Tips dataset contains information about restaurant tips, with the following columns:

total_bill: The total bill amount (independent variable).

tip: The tip amount (dependent variable).

sex: The sex of the person paying the bill.

smoker: Whether the person is a smoker.

day: The day of the week.

time: Lunch or dinner.

size: The size of the party.

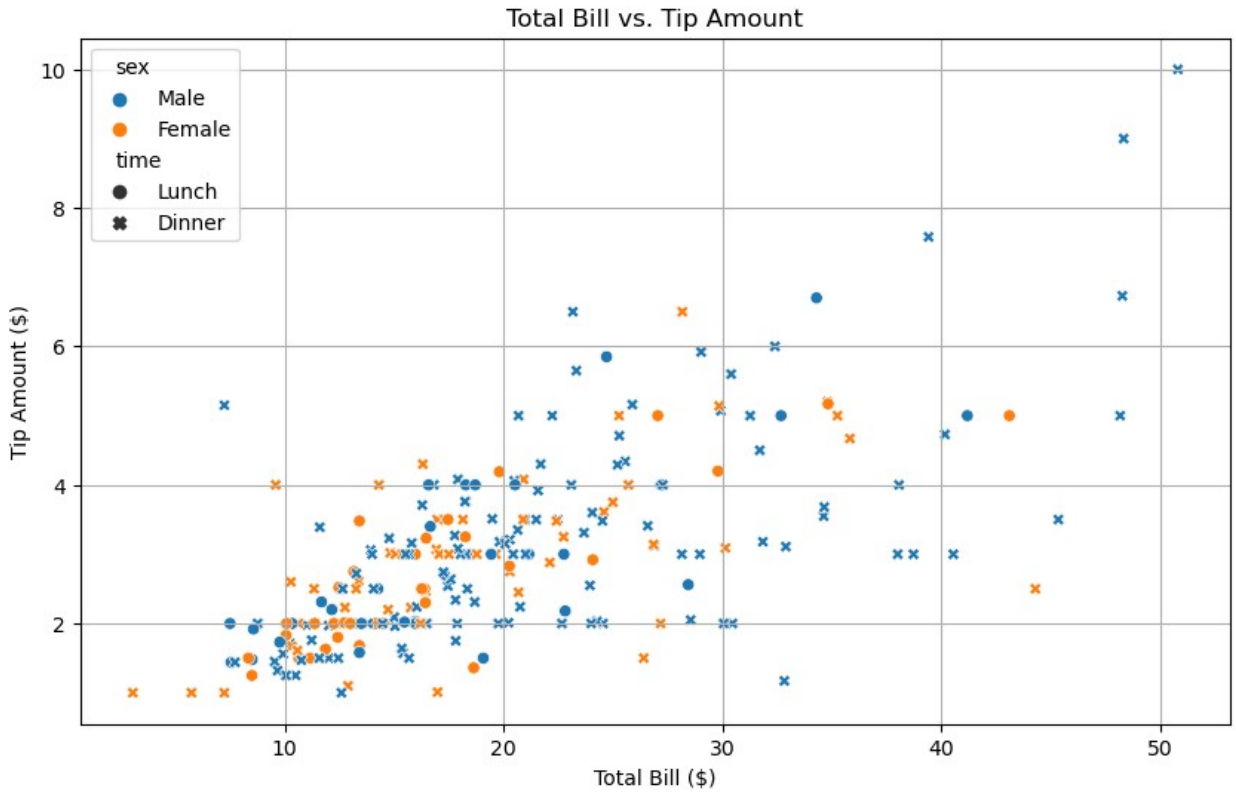
Example Scenario Suppose you want to analyze how the total bill influences the tip amount. Here:

Independent Variable: total_bill (the amount you manipulate or categorize).

Dependent Variable: tip (the amount you measure based on the total bill).

```
# Load the Tips dataset
tips = sns.load_dataset('tips')

# Create a scatter plot
```

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=tips, x='total_bill', y='tip', hue='sex',
style='time')
plt.title('Total Bill vs. Tip Amount')
plt.xlabel('Total Bill ($)')
plt.ylabel('Tip Amount ($)')
plt.grid(True)
plt.show()
```




```
import pandas as pd
data = {
    'empid':range(1,9),
    'salary(k)':[15,18,22,25,24,30,22,28]
}
df = pd.DataFrame(data)
df
```

	empid	salary(k)
0	1	15
1	2	18
2	3	22
3	4	25
4	5	24
5	6	30

```
6      7          22
7      8          28
```

```
df['salary(k)'].mean()
```

```
23.0
```

```
df['salary(k)'].median()
```

```
23.0
```

```
df['salary(k)'].mode()
```

```
0    22
Name: salary(k), dtype: int64
```

```
salary_range = df['salary(k)'].max()-df['salary(k)'].min()
salary_range
```

```
15
```

```
df['salary(k)'].var()
```

```
24.285714285714285
```

```
df['salary(k)'].std()
```

```
4.928053803045811
```

```
Q1 = df['salary(k)'].quantile(0.25)
Q3 = df['salary(k)'].quantile(0.75)
IQR = Q3-Q1
IQR
```

```
4.75
```

```
lower_bound = Q1-1.5*IQR
lower_bound
```

```
13.875
```

```
upper_bound = Q3+1.5*IQR
upper_bound
```

```
32.875
```

```
skewness= df['salary(k)'].skew()
skewness
```

```
-0.2578275899725006
```

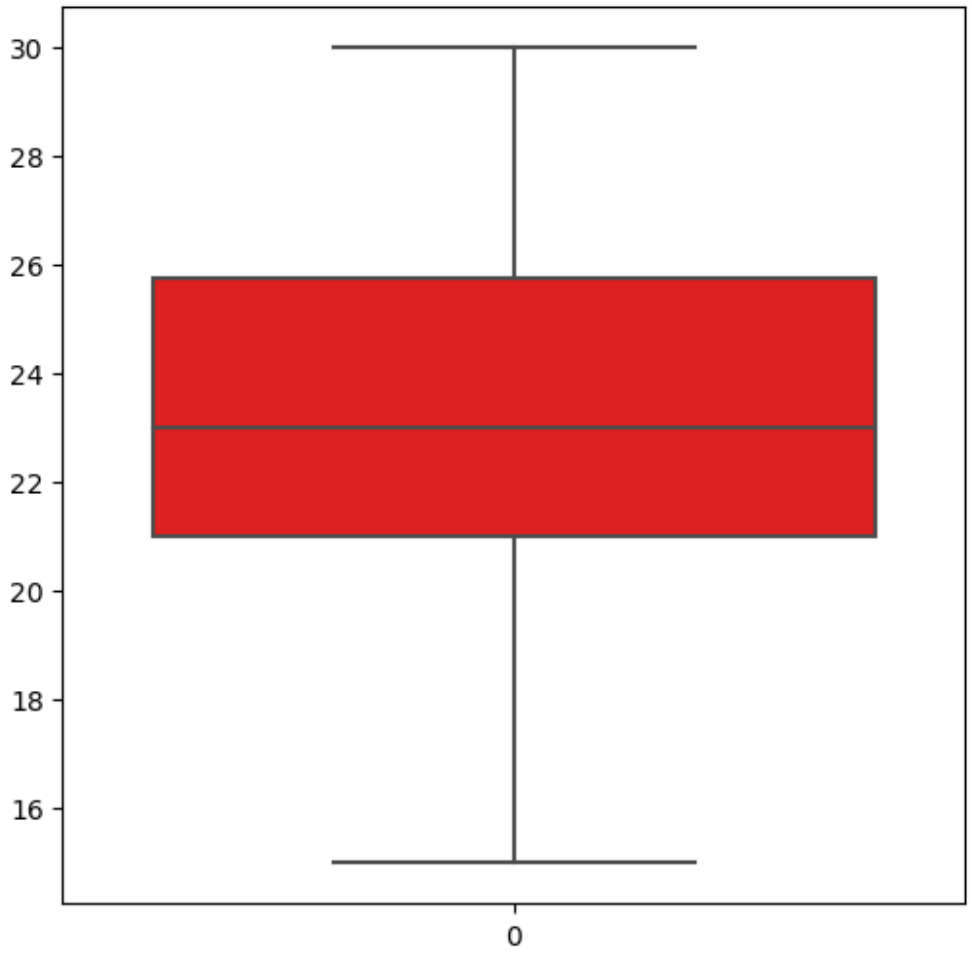
```
kurtosis =df['salary(k)'].kurtosis()
kurtosis
```

```
-0.38550865051903127
```

```
df['salary(k)'].describe()
```

```
count     8.000000
mean     23.000000
std       4.928054
min      15.000000
25%      21.000000
50%      23.000000
75%      25.750000
max      30.000000
Name: salary(k), dtype: float64
```

```
plt.figure(figsize=(6,6))
sns.boxplot(data=df['salary(k)'],color = 'red',orient = 'v')
plt.show()
```



Activity

```
iris = sns.load_dataset('iris')
iris
```

```
     sepal_length  sepal_width  petal_length  petal_width    species
0             5.1          3.5           1.4          0.2     setosa
1             4.9          3.0           1.4          0.2     setosa
2             4.7          3.2           1.3          0.2     setosa
3             4.6          3.1           1.5          0.2     setosa
4             5.0          3.6           1.4          0.2     setosa
..            ...          ...           ...          ...        ...
145           6.7          3.0           5.2          2.3  virginica
146           6.3          2.5           5.0          1.9  virginica
147           6.5          3.0           5.2          2.0  virginica
148           6.2          3.4           5.4          2.3  virginica
149           5.9          3.0           5.1          1.8  virginica

[150 rows x 5 columns]
```

```
i_c =iris.copy()
i_c
```

```
     sepal_length  sepal_width  petal_length  petal_width    species
0             5.1          3.5           1.4          0.2     setosa
1             4.9          3.0           1.4          0.2     setosa
2             4.7          3.2           1.3          0.2     setosa
3             4.6          3.1           1.5          0.2     setosa
4             5.0          3.6           1.4          0.2     setosa
..            ...          ...           ...          ...        ...
145           6.7          3.0           5.2          2.3  virginica
146           6.3          2.5           5.0          1.9  virginica
147           6.5          3.0           5.2          2.0  virginica
148           6.2          3.4           5.4          2.3  virginica
149           5.9          3.0           5.1          1.8  virginica

[150 rows x 5 columns]
```

```
df = i_c['sepal_length']
df
```

```
0      5.1
1      4.9
2      4.7
3      4.6
4      5.0
      ...
145    6.7
146    6.3
147    6.5
148    6.2
```

```
149    5.9
Name: sepal_length, Length: 150, dtype: float64
```

```
df.isnull().sum()
```

```
0
```

```
mean = df.mean()
mean
```

```
5.843333333333334
```

```
median = df.median()
median
```

```
5.8
```

```
mode = df.mode()
mode
```

```
0    5.0
Name: sepal_length, dtype: float64
```

```
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3-Q1
IQR
```

```
1.3000000000000007
```

```
lower_bound = Q1-1.5*IQR
lower_bound
```

```
3.1499999999999986
```

```
upper_bound = Q3+1.5*IQR
upper_bound
```

```
8.350000000000001
```

```
range_ = df.max()-df.min()
range_
```

```
3.6000000000000005
```

```
variance = df.var()
variance
```

```
0.6856935123042505
```

```
standard_deviation = df.std()
standard_deviation
```

```
0.8280661279778629
```

```
from scipy import stats
skewness = stats.skew(df)
skewness
```

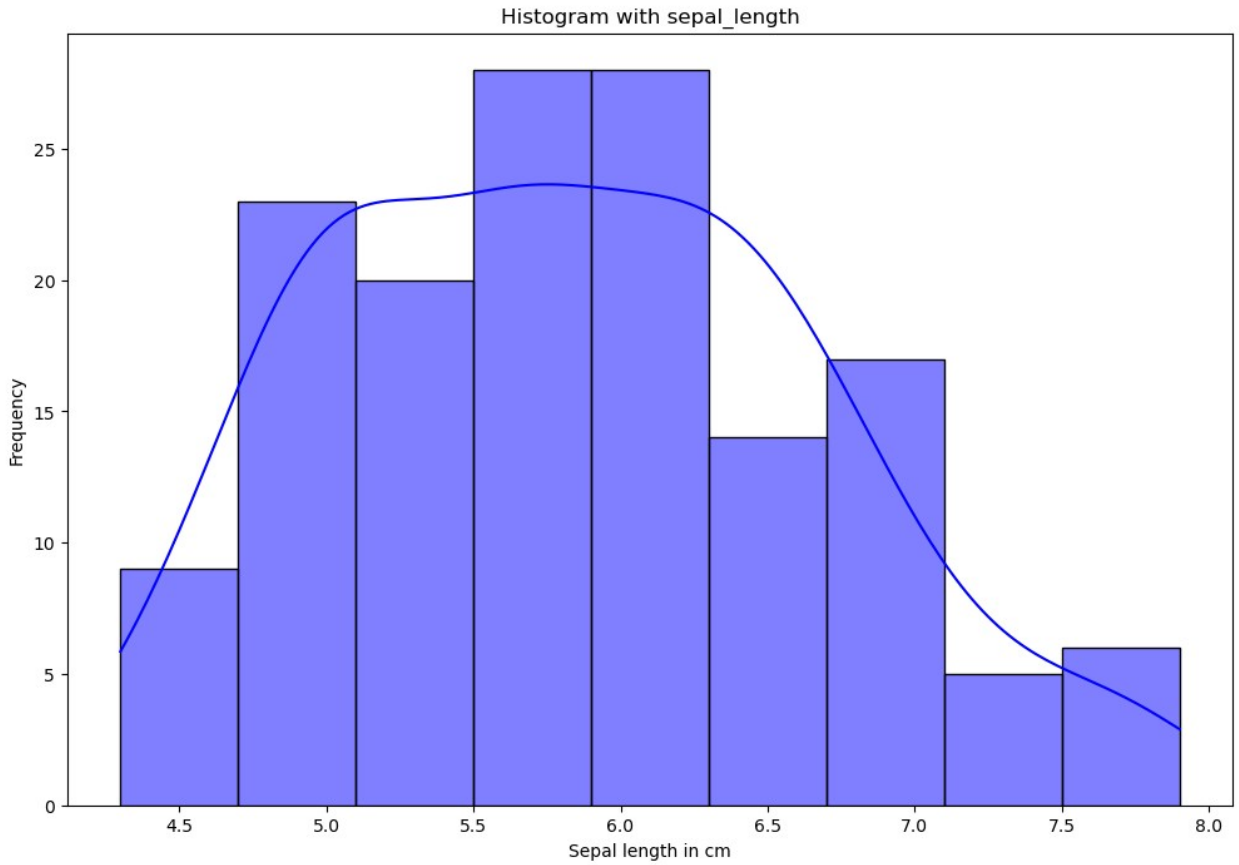
```
0.3117530585022963
```

```
kurtosis = stats.kurtosis(df)
kurtosis
```

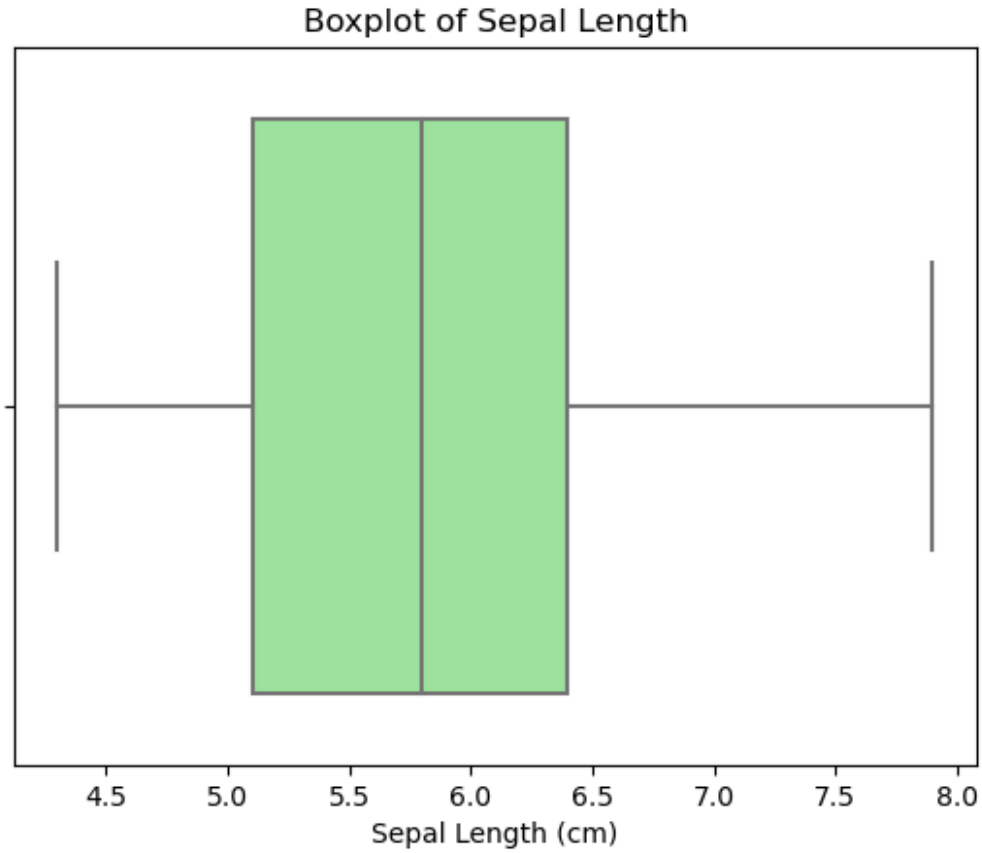
```
-0.5735679489249765
```

```
plt.figure(figsize=(12,8))
sns.histplot(df,kde = True,color = 'blue')
plt.title('Histogram with sepal_length')
plt.xlabel('Sepal length in cm')
plt.ylabel('Frequency')
plt.show()
```

```
E:\HDD1\Anaconda\envs\CV\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



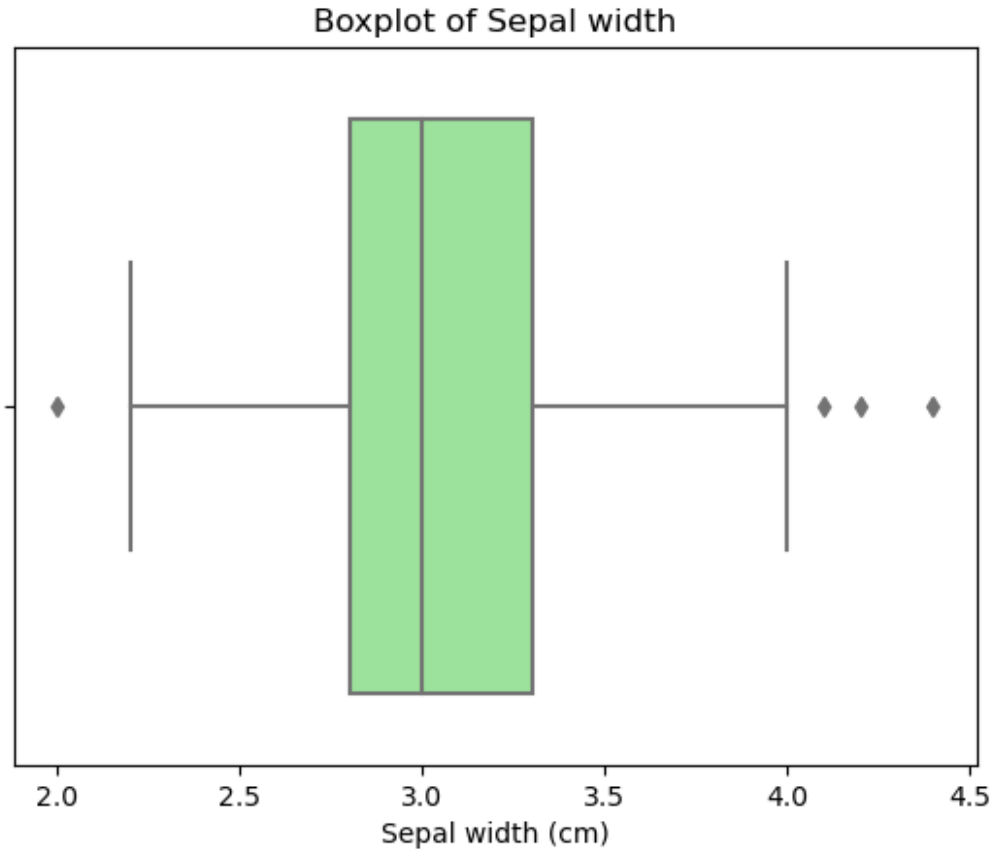
```
import matplotlib.pyplot as plt
sns.boxplot(x=df,color = 'lightgreen')
plt.title('Boxplot of Sepal Length')
plt.xlabel('Sepal Length (cm)')
plt.show()
```




```
df1 = i_c['sepal_width']
df1
```

```
0      3.5
1      3.0
2      3.2
3      3.1
4      3.6
      ...
145    3.0
146    2.5
147    3.0
148    3.4
149    3.0
Name: sepal_width, Length: 150, dtype: float64
```

```
sns.boxplot(x=df1,color = 'lightgreen')
plt.title('Boxplot of Sepal width')
plt.xlabel('Sepal width (cm)')
plt.show()
```




```
Q1 = df1.quantile(0.25)
Q3 = df1.quantile(0.75)
IQR1 = Q3-Q1
IQR1
```

```
0.5
```

```
lower_bound = Q1-1.5*IQR1
lower_bound
```

```
2.05
```

```
upper_bound = Q3+1.5*IQR1
upper_bound
```

```
4.05
```

```
df1.min()
```

```
2.0
```

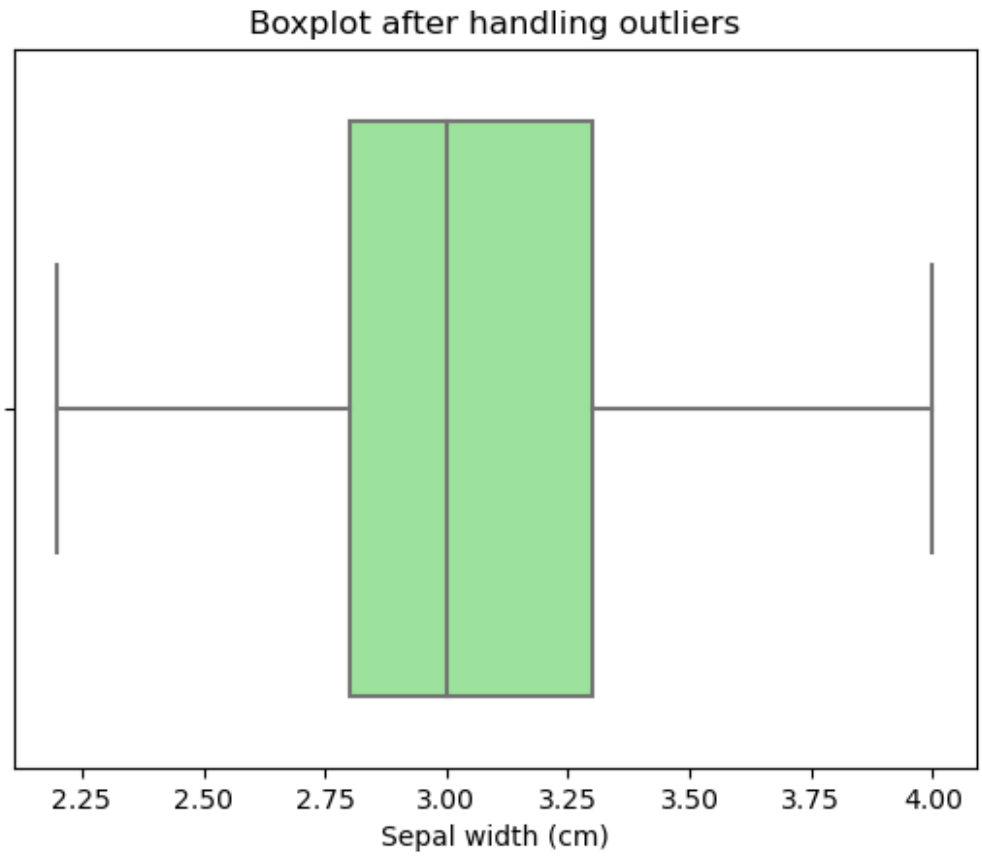
```
df1.max()
```

```
4.4
```

```
filtered_data = df1[(df1>=lower_bound) & (df1<=upper_bound)]
filtered_data
```

```
0      3.5
1      3.0
2      3.2
3      3.1
4      3.6
      ...
145    3.0
146    2.5
147    3.0
148    3.4
149    3.0
Name: sepal_width, Length: 146, dtype: float64
```

```
sns.boxplot(x=filtered_data,color = 'lightgreen')
plt.title('Boxplot after handling outliers')
plt.xlabel('Sepal width (cm)')
plt.show()
```




```
import numpy as np

d = [12, 15, 18, 22, 25, 28, 31, 34, 37, 40, 43, 46, 49, 52, 55, 58,
61, 64, 67, 70]
b = np.array(d)
b
```

```
array([12, 15, 18, 22, 25, 28, 31, 34, 37, 40, 43, 46, 49, 52, 55, 58,
61,
       64, 67, 70])
```

```
b.mean()
```

```
41.35
```

```
c =np.median(b)
c
```

```
41.5
```

```
range_d = np.ptp(b)
range_d
```

```
58
```

```
np.var(b)
```

```
307.02750000000003
```

```
np.std(b)
```

```
17.522200204312245
```

```
a =[85, 92, 78, 88, 95, 60, 72, 94, 43,100, 57, 81, 90, 84]
k = np.array(a)
k
```

```
array([ 85,  92,  78,  88,  95,  60,  72,  94,  43, 100,  57,  81,
90,
        84])
```

```
k.mean()
```

```
79.92857142857143
```

```
np.median(k)
```

```
84.5
```

```
np.ptp(k)
```

```
57
```

```
np.var(k)
```

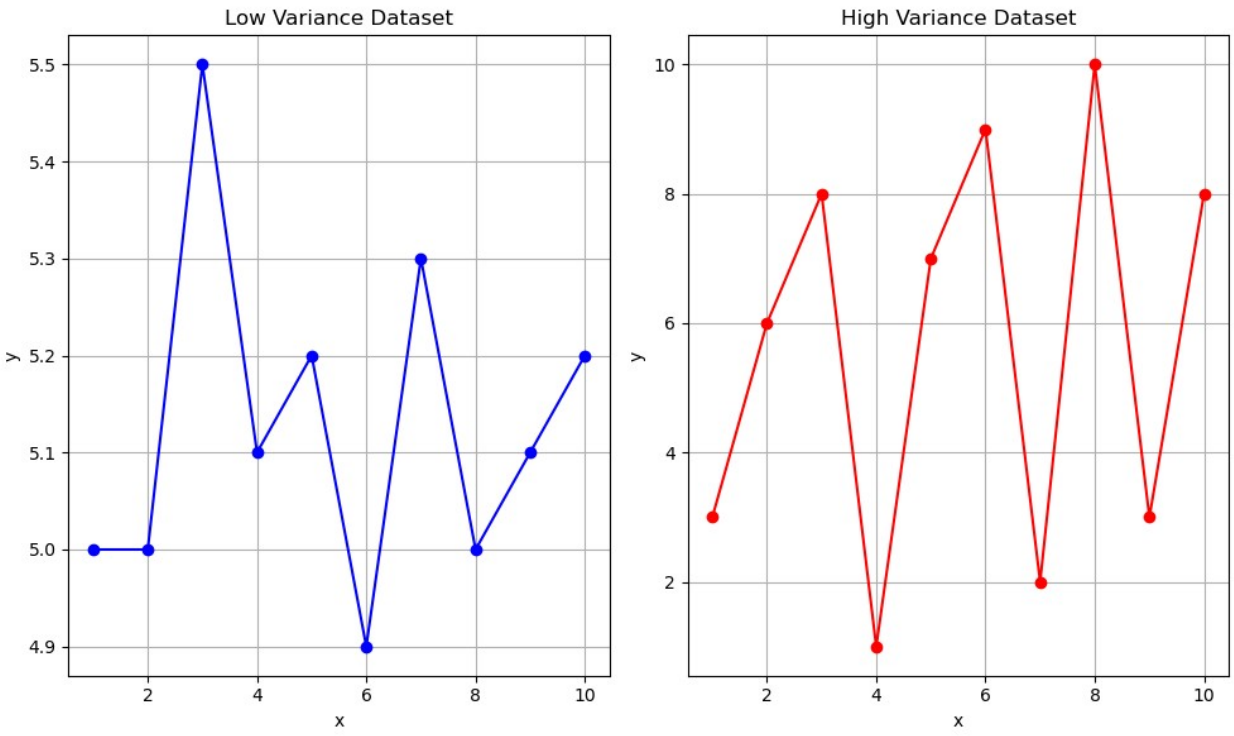
```
252.6377551020408
```

```
np.std(k)
```

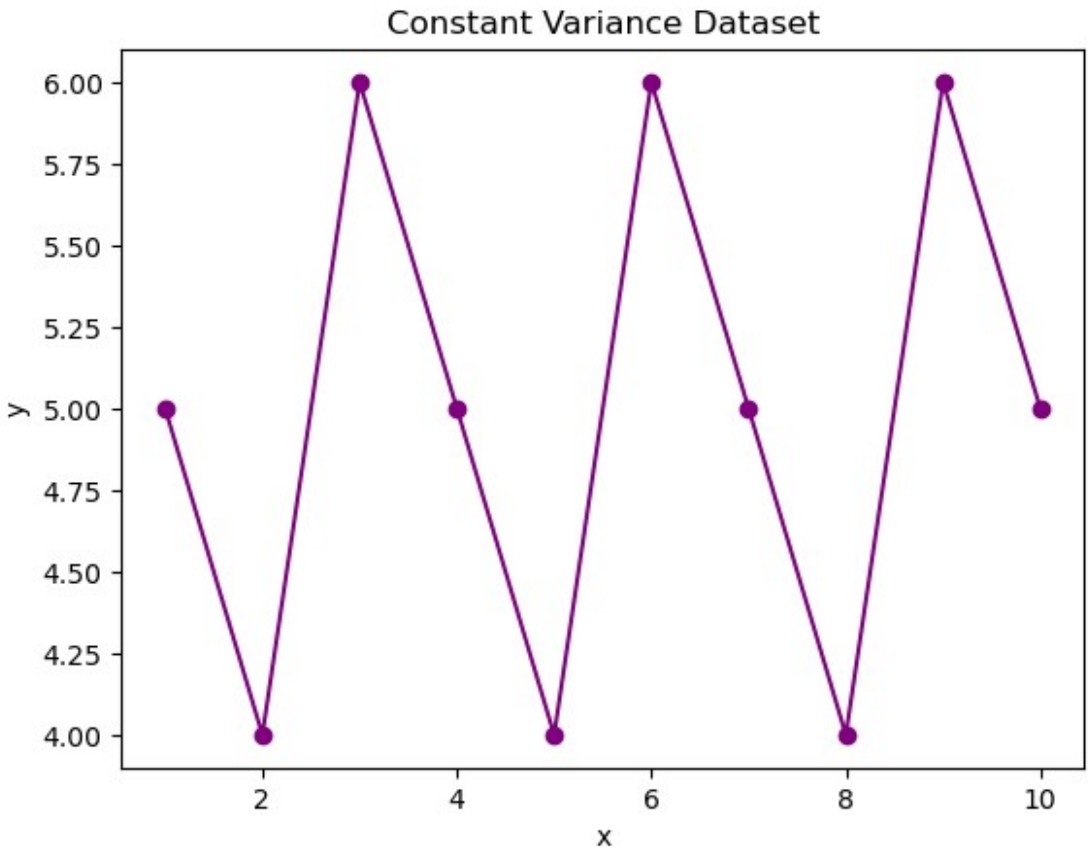
```
15.894582570864854
```

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

# Generate data with low variance and high variance
x = np.arange(1, 11)
y_low_variance = [5, 5, 5.5, 5.1, 5.2, 4.9, 5.3, 5, 5.1, 5.2]  # Low
variance
y_high_variance = [3, 6, 8, 1, 7, 9, 2, 10, 3, 8]  # High variance

# Plot both on the same graph for comparison
plt.figure(figsize=(10, 6))

# Low variance
plt.subplot(1, 2, 1)
plt.plot(x, y_low_variance, 'o-', color='blue', label='Low Variance')
```

```
plt.title('Low Variance Dataset')
plt.xlabel('x')
plt.ylabel('y')
plt.grid(True)

# High variance
plt.subplot(1, 2, 2)
plt.plot(x, y_high_variance, 'o-', color='red', label='High Variance')
plt.title('High Variance Dataset')
plt.xlabel('x')
plt.ylabel('y')
plt.grid(True)

# Show the plot
plt.tight_layout()
plt.show()
```




```
import matplotlib.pyplot as plt

x = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
y = [5, 4, 6, 5, 4, 6, 5, 4, 6, 5]

plt.plot(x, y, 'o-',color = 'purple')
plt.title('Constant Variance Dataset')
plt.xlabel('x')
plt.ylabel('y')
plt.show()
```




```
import pandas as pd
data = {
    'empid':range(1,9),
    'salary(k)':[15,18,22,25,24,30,22,28]
}
df = pd.DataFrame(data)
df
```

	empid	salary(k)
0	1	15
1	2	18
2	3	22
3	4	25
4	5	24
5	6	30
6	7	22
7	8	28

```
df['salary(k)'].mean()
```

```
23.0
```

```
df['salary(k)'].median()
```

```
23.0
```

```
df['salary(k)'].mode()
```

```
0    22
Name: salary(k), dtype: int64
```

```
df['salary(k)'].var()
```

```
24.285714285714285
```

```
df['salary(k)'].std()
```

```
4.928053803045811
```

```
df['salary(k)'].skew()
```

```
-0.2578275899725006
```

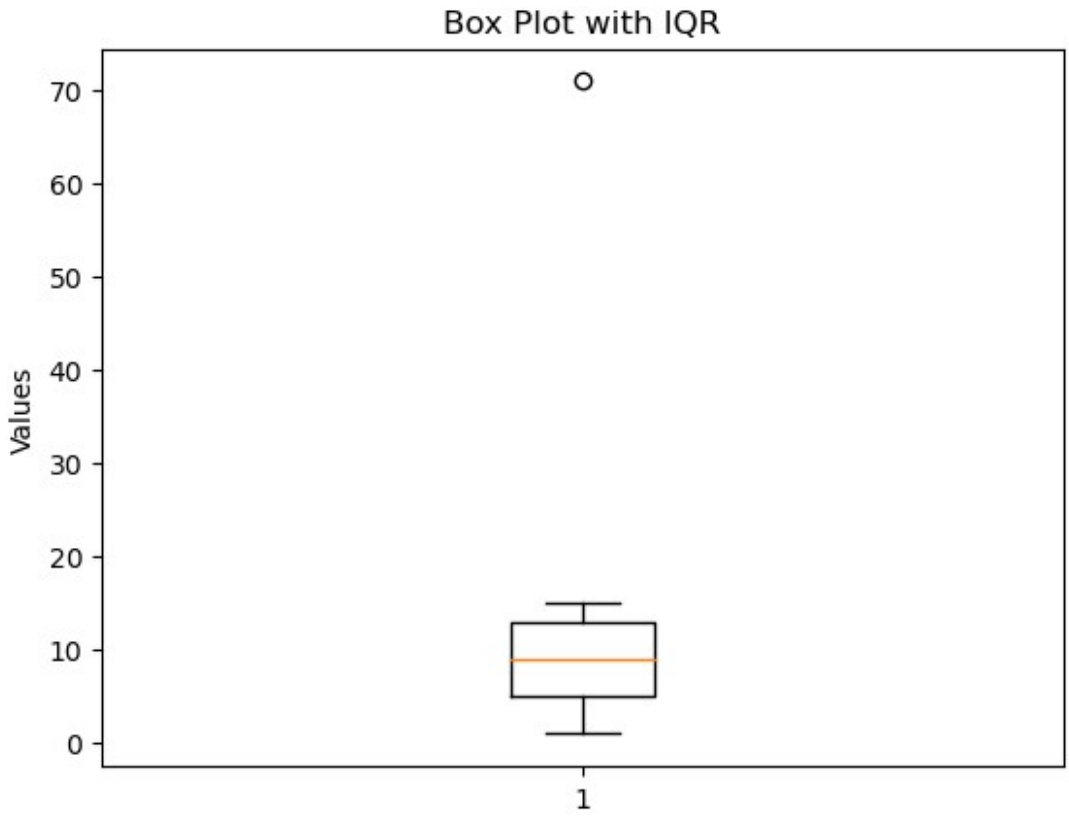
```
df['salary(k)'].kurtosis()
```

```
-0.38550865051903127
```

```
import pandas as pd

# Sample data
data = [1, 3, 5, 7, 9, 11, 13, 15, 71]
df = pd.DataFrame(data, columns=['Values'])

# Calculate IQR
Q1 = df['Values'].quantile(0.25)
Q3 = df['Values'].quantile(0.75)
IQR = Q3 - Q1

print(f"Q1: {Q1}, Q3: {Q3}, IQR: {IQR}")
```

```
Q1: 5.0, Q3: 13.0, IQR: 8.0
```

```
import matplotlib.pyplot as plt

# Box plot to visualize IQR
plt.boxplot(data)
plt.title("Box Plot with IQR")
plt.ylabel("Values")
plt.show()
```




```
import numpy as np
Q1 = np.percentile(df['Values'],25)
Q3 = np.percentile(df['Values'],75)
IQR = Q3-Q1

Q1

5.0

Q3

13.0

IQR

8.0

lower_bound = Q1 - 1.5*IQR
lower_bound

-7.0

upper_bound = Q3 + 1.5 * IQR
upper_bound

25.0
```

```
import numpy as np

# Sample data
data = [1, 3, 5, 7, 9, 11, 13, 15, 17]

# Calculate Q1, Q3, and IQR
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
IQR = Q3 - Q1

# Calculate Lower and Upper Bounds
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

print(f"Lower Bound: {lower_bound}")
print(f"Upper Bound: {upper_bound}")
```

```
Lower Bound: -7.0
Upper Bound: 25.0
```

```
Q1
```

```
5.0
```

```
Q3
```

```
13.0
```

```
IQR
```

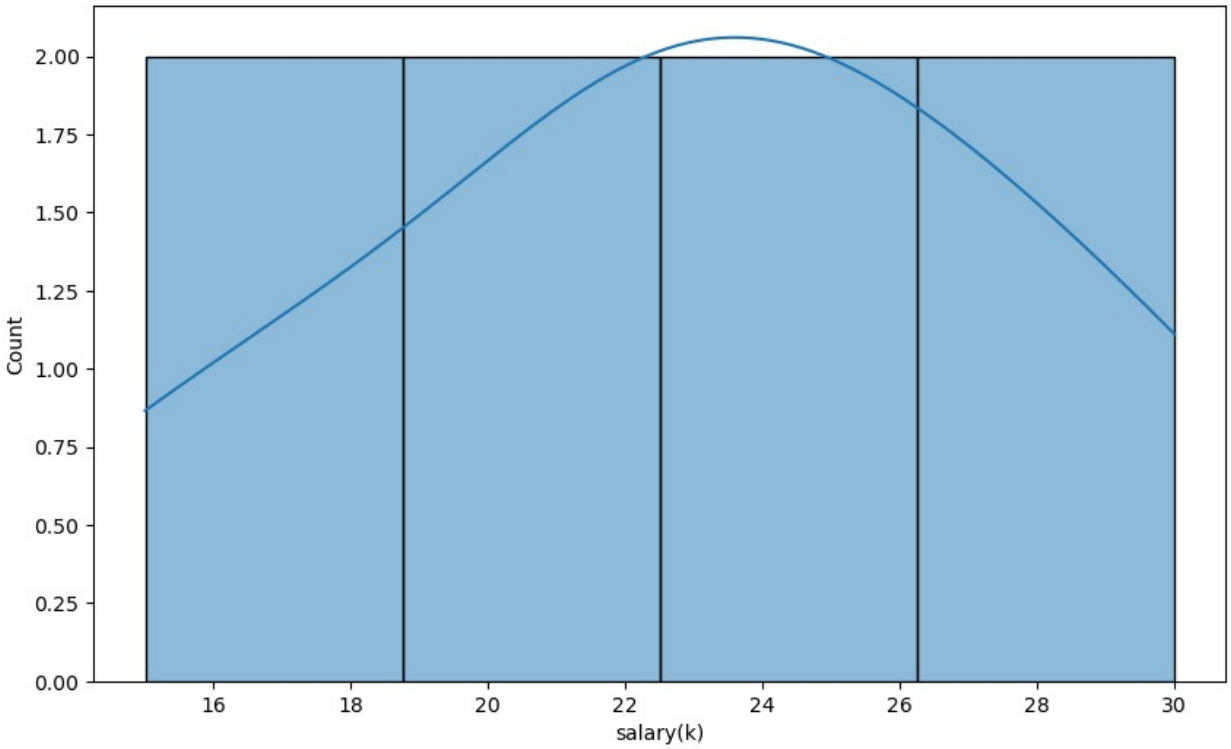
```
8.0
```

```
median_ = np.median(data)
median_
```

```
9.0
```

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize = (10,6))
sns.histplot(df['salary(k)'],kde = True)
plt.show()
```

```
E:\HDD1\Anaconda\envs\CV\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

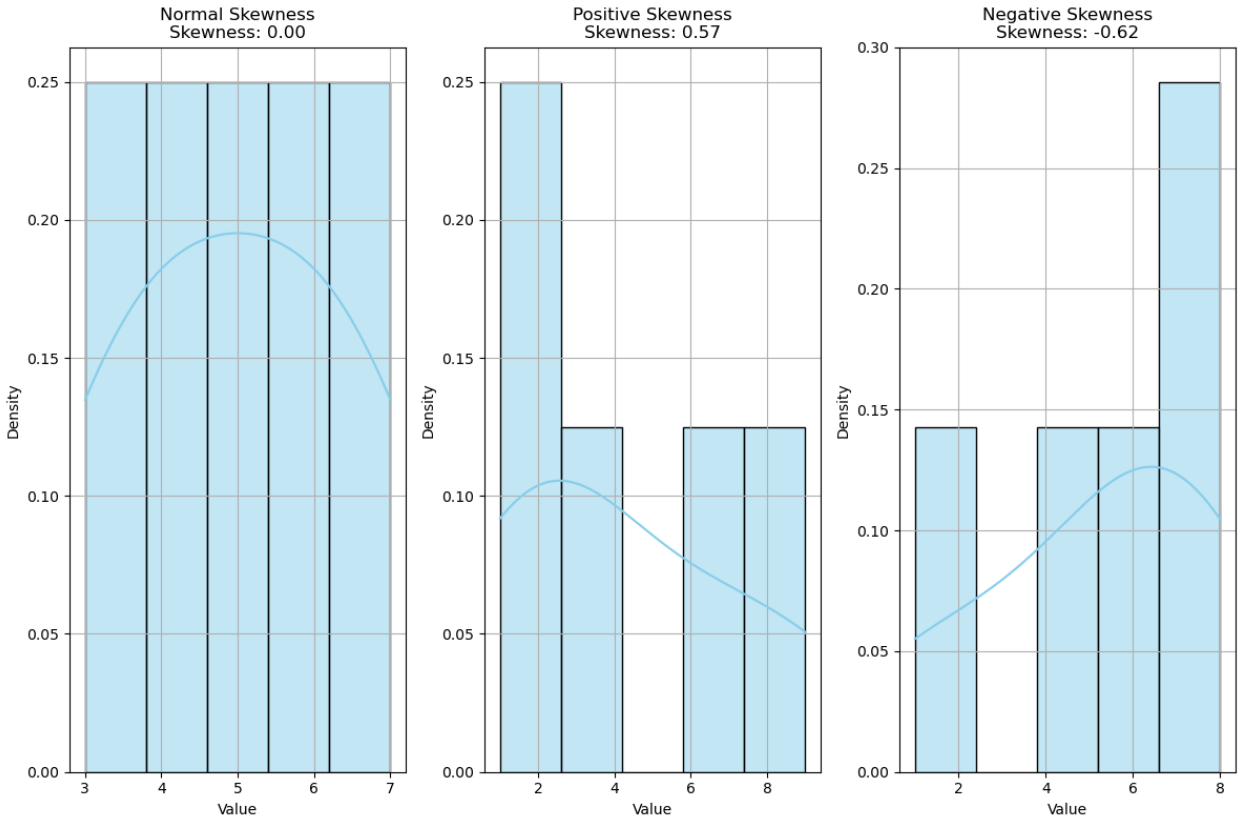



```
import numpy as np
from scipy.stats import skew

# Normal Skewness
data_normal = [3, 4, 5, 6, 7]
print("Normal Skewness:", skew(data_normal))

# Positive Skewness
data_positive = [1, 2, 3, 6, 9]
print("Positive Skewness:", skew(data_positive))

# Negative Skewness
data_negative = [1, 4, 6, 7, 8]
print("Negative Skewness:", skew(data_negative))
```

```
Normal Skewness: 0.0
Positive Skewness: 0.5692290289523131
Negative Skewness: -0.6216348579624362
```

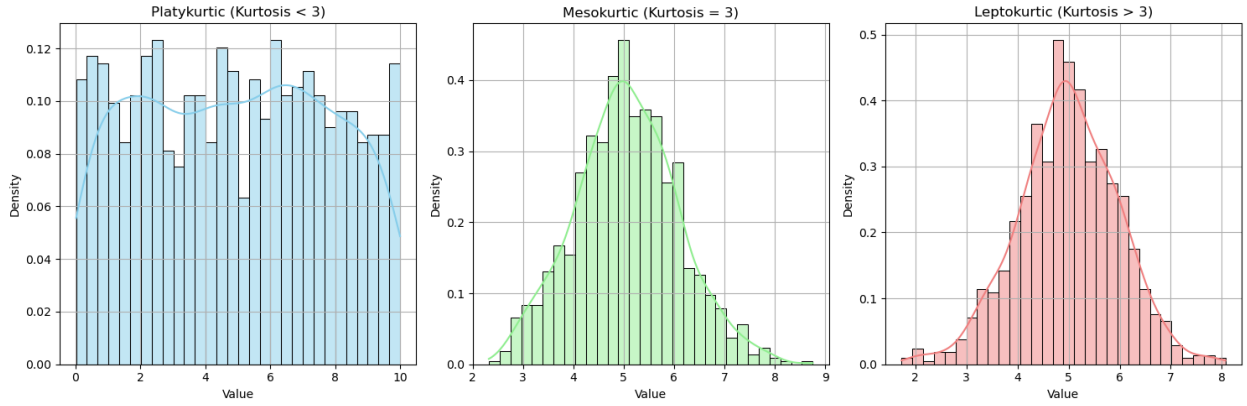
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import skew
import warnings

# Suppress specific warnings
warnings.filterwarnings("ignore", category=FutureWarning)
```

```
# Data for different types of skewness
data_normal = [3, 4, 5, 6, 7]      # Normal skewness
data_positive = [1, 2, 3, 6, 9]    # Positive skewness
data_negative = [1, 4, 6, 7, 8]    # Negative skewness

# Function to plot histogram and show skewness
def plot_histogram(data, title):
    # Ensure the data is clean (no infinite values or NaNs)
    data = np.array(data)

    # Create histogram with KDE
    sns.histplot(data, bins=5, color='skyblue', edgecolor='black',
kde=True, stat='density', alpha=0.5)

    # Calculate skewness
    skewness = skew(data)

    # Set title with skewness value
    plt.title(f"{title}\nSkewness: {skewness:.2f}")
    plt.xlabel('Value')
    plt.ylabel('Density')
    plt.grid(True)

# Create subplots for each dataset
plt.figure(figsize=(12, 8))

# Normal skewness
plt.subplot(1, 3, 1)
plot_histogram(data_normal, 'Normal Skewness')

# Positive skewness
plt.subplot(1, 3, 2)
plot_histogram(data_positive, 'Positive Skewness')

# Negative skewness
plt.subplot(1, 3, 3)
plot_histogram(data_negative, 'Negative Skewness')

# Show the histograms
plt.tight_layout()
plt.show()
```



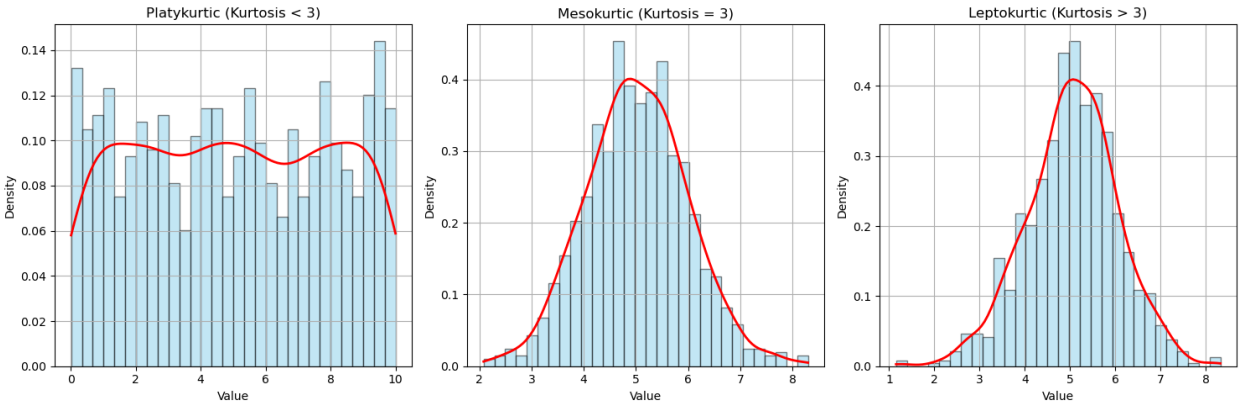

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Sample data for different types of kurtosis
data_platykurtic = np.random.uniform(0, 10, 1000)  # Platykurtic
(kurtosis < 3)
data_mesokurtic = np.random.normal(5, 1, 1000)     # Mesokurtic
(kurtosis = 3)
data_leptokurtic = np.random.normal(5, 1, 1000)     # Leptokurtic
(kurtosis > 3)

# Create a figure
plt.figure(figsize=(15, 5))

# Platykurtic
plt.subplot(1, 3, 1)
sns.histplot(data_platykurtic, bins=30, kde=True, color='skyblue',
stat='density', alpha=0.5)
plt.title('Platykurtic (Kurtosis < 3)')
plt.xlabel('Value')
plt.ylabel('Density')
plt.grid(True)
```

```
# Mesokurtic
plt.subplot(1, 3, 2)
sns.histplot(data_mesokurtic, bins=30, kde=True, color='lightgreen',
stat='density', alpha=0.5)
plt.title('Mesokurtic (Kurtosis = 3)')
plt.xlabel('Value')
plt.ylabel('Density')
plt.grid(True)

# Leptokurtic
plt.subplot(1, 3, 3)
sns.histplot(data_leptokurtic, bins=30, kde=True, color='lightcoral',
stat='density', alpha=0.5)
plt.title('Leptokurtic (Kurtosis > 3)')
plt.xlabel('Value')
plt.ylabel('Density')
plt.grid(True)

# Show the plots
plt.tight_layout()
plt.show()
```



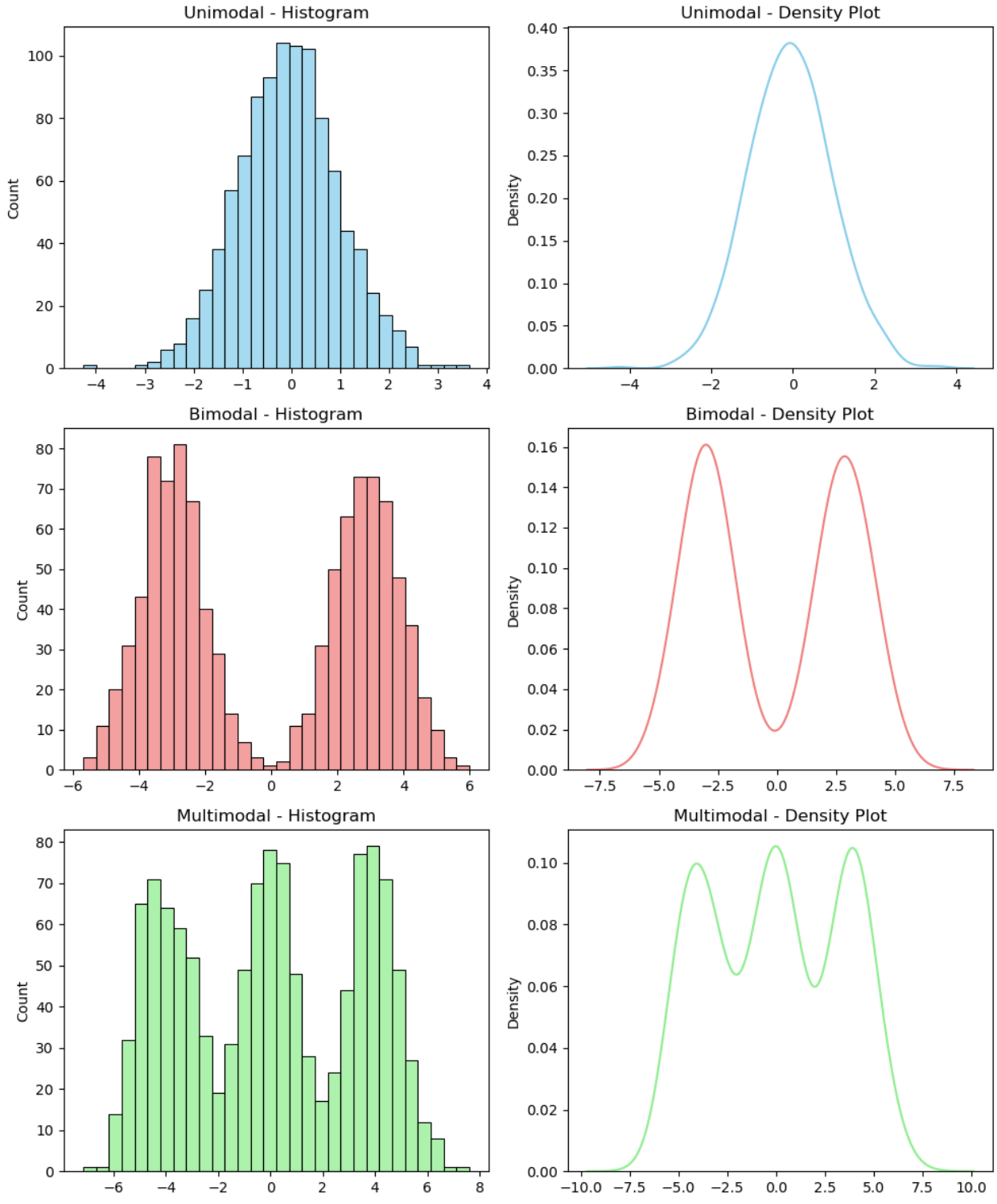

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import gaussian_kde

# Sample data for different types of kurtosis
data_platykurtic = np.random.uniform(0, 10, 1000)  # Platykurtic
(kurtosis < 3)
data_mesokurtic = np.random.normal(5, 1, 1000)     # Mesokurtic
(kurtosis = 3)
data_leptokurtic = np.random.normal(5, 1, 1000)     # Leptokurtic
(kurtosis > 3)

# Create a figure
```

```
plt.figure(figsize=(15, 5))

# Function to plot histogram and KDE
def plot_kurtosis(data, title, subplot_position):
    plt.subplot(1, 3, subplot_position)
    plt.hist(data, bins=30, density=True, color='skyblue',
edgecolor='black', alpha=0.5)
    kde = gaussian_kde(data)
    x = np.linspace(min(data), max(data), 100)
    plt.plot(x, kde(x), color='red', label='KDE', linewidth=2)
    plt.title(title)
    plt.xlabel('Value')
    plt.ylabel('Density')
    plt.grid(True)

# Plot for Platykurtic
plot_kurtosis(data_platykurtic, 'Platykurtic (Kurtosis < 3)', 1)

# Plot for Mesokurtic
plot_kurtosis(data_mesokurtic, 'Mesokurtic (Kurtosis = 3)', 2)

# Plot for Leptokurtic
plot_kurtosis(data_leptokurtic, 'Leptokurtic (Kurtosis > 3)', 3)

# Show the plots
plt.tight_layout()
plt.show()
```



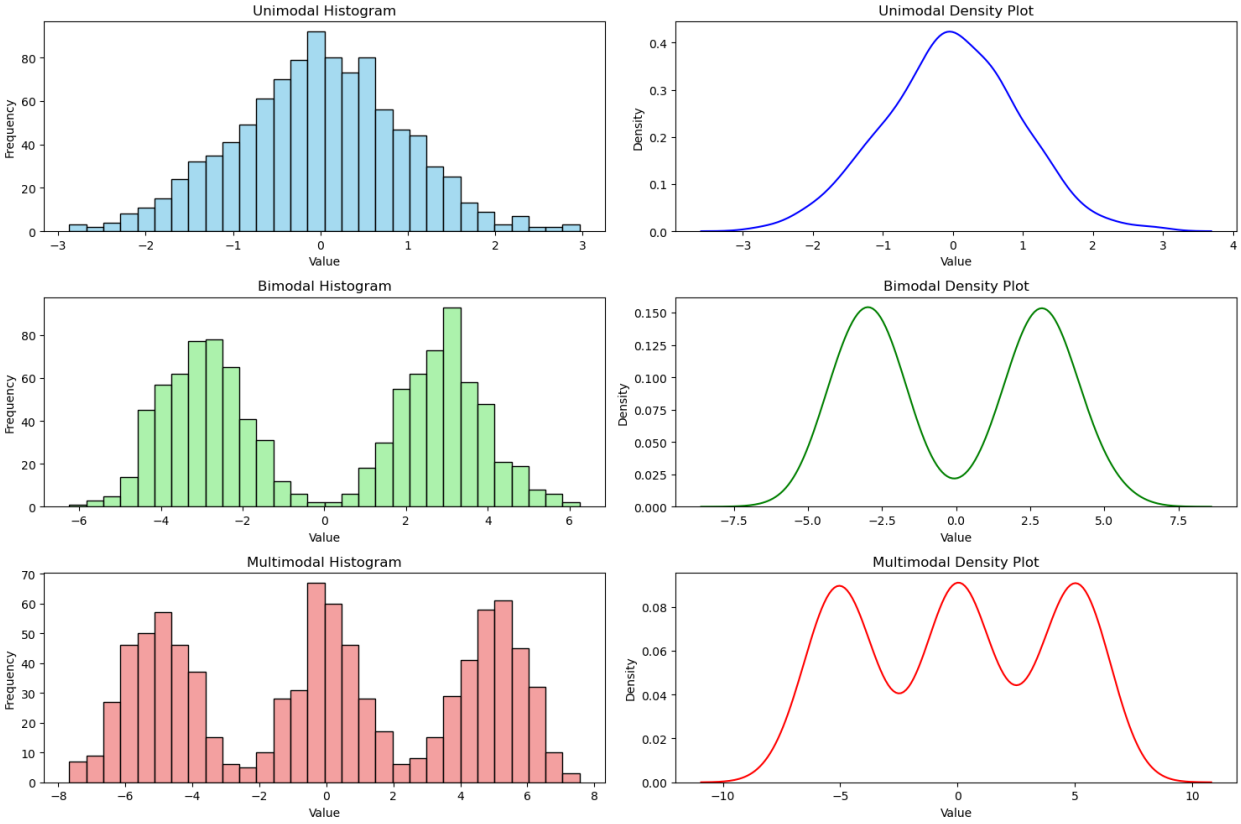

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import warnings
warnings.filterwarnings('ignore')

# Generating sample data for each modality
data_unimodal = np.random.normal(loc=0, scale=1, size=1000)
data_bimodal = np.concatenate([np.random.normal(loc=-3, scale=1,
```

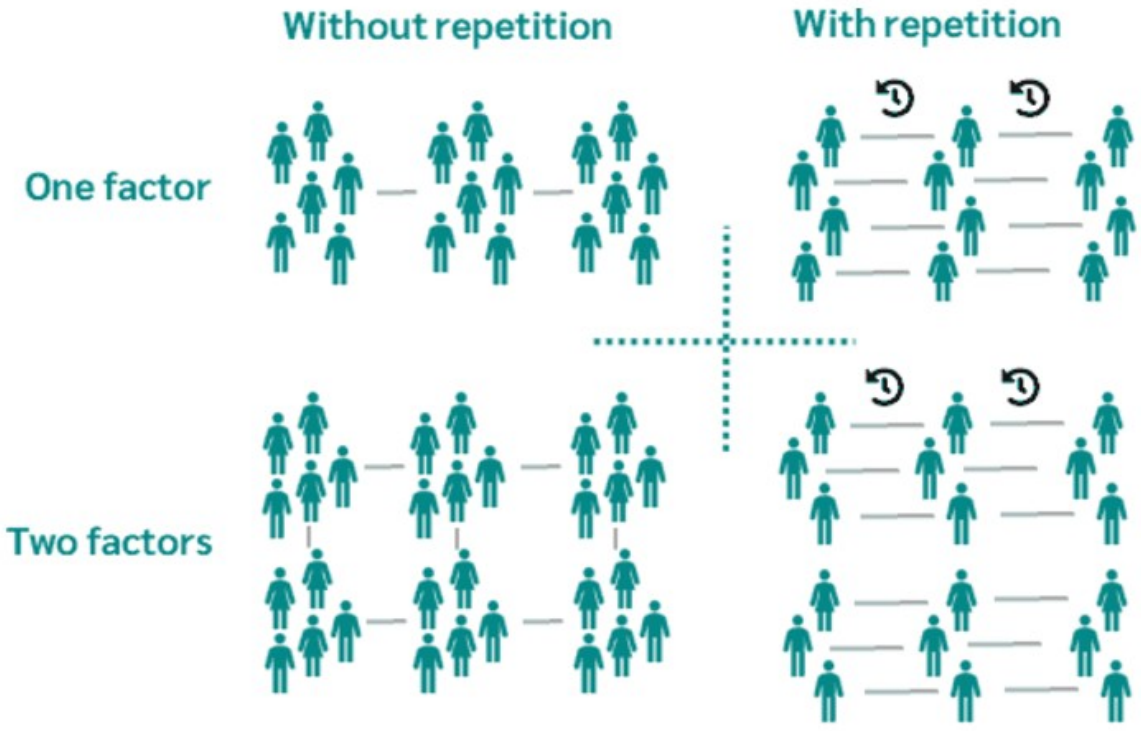
```
size=500), np.random.normal(loc=3, scale=1, size=500)])
data_multimodal = np.concatenate([np.random.normal(loc=-4, scale=1,
size=400), np.random.normal(loc=0, scale=1, size=400),
np.random.normal(loc=4, scale=1, size=400)])

# Create subplots for histogram and density plots
fig, axs = plt.subplots(3, 2, figsize=(10, 12))

# Unimodal distribution
sns.histplot(data_unimodal, bins=30, kde=False, ax=axs[0, 0],
color='skyblue')
sns.kdeplot(data_unimodal, ax=axs[0, 1], color='skyblue')
axs[0, 0].set_title('Unimodal - Histogram')
axs[0, 1].set_title('Unimodal - Density Plot')

# Bimodal distribution
sns.histplot(data_bimodal, bins=30, kde=False, ax=axs[1, 0],
color='lightcoral')
sns.kdeplot(data_bimodal, ax=axs[1, 1], color='lightcoral')
axs[1, 0].set_title('Bimodal - Histogram')
axs[1, 1].set_title('Bimodal - Density Plot')

# Multimodal distribution
sns.histplot(data_multimodal, bins=30, kde=False, ax=axs[2, 0],
color='lightgreen')
sns.kdeplot(data_multimodal, ax=axs[2, 1], color='lightgreen')
axs[2, 0].set_title('Multimodal - Histogram')
axs[2, 1].set_title('Multimodal - Density Plot')

# Adjust layout
plt.tight_layout()
plt.show()
```

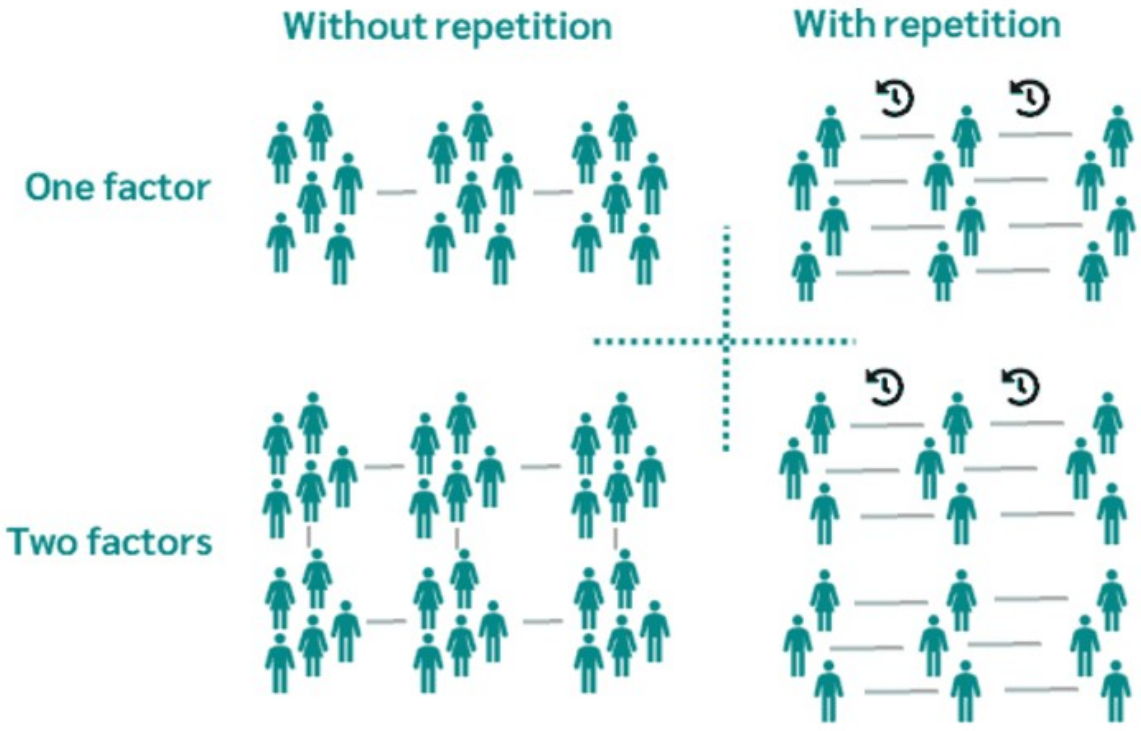


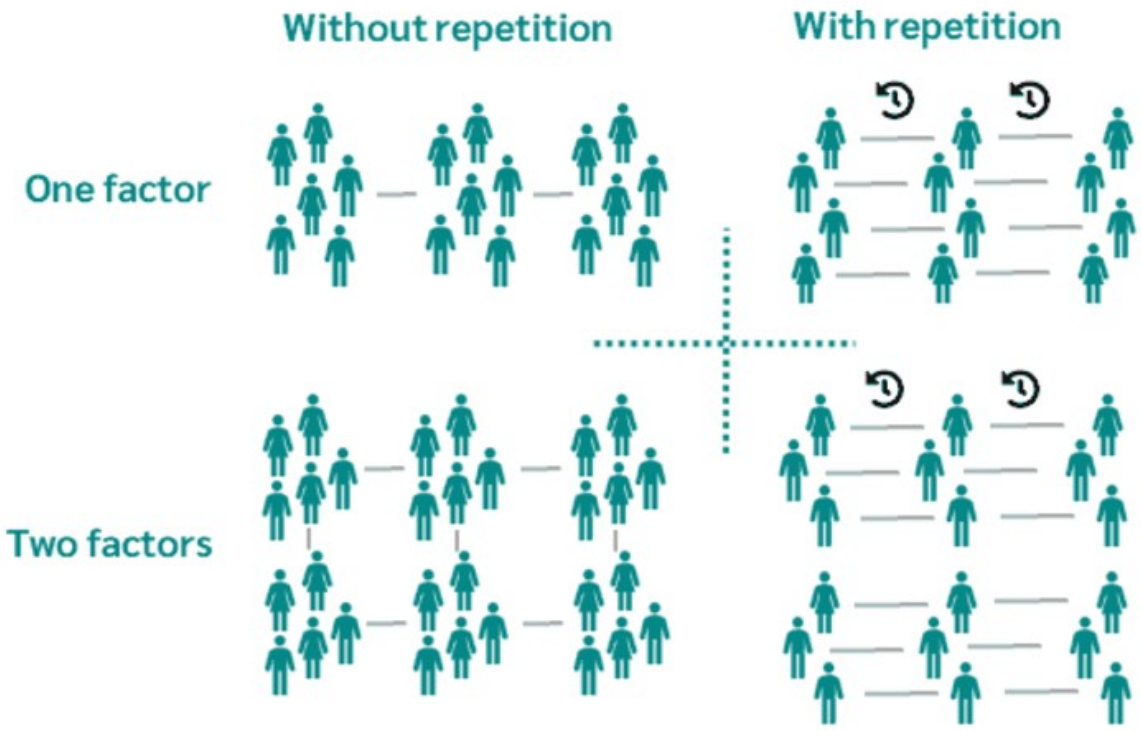

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Generating data for unimodal, bimodal, and multimodal distributions
```

```
data_unimodal = np.random.normal(loc=0, scale=1, size=1000)
data_bimodal = np.concatenate([np.random.normal(loc=-3, scale=1,
size=500), np.random.normal(loc=3, scale=1, size=500)])
data_multimodal = np.concatenate([np.random.normal(loc=-5, scale=1,
size=300), np.random.normal(loc=0, scale=1, size=300),
np.random.normal(loc=5, scale=1, size=300)])

# Create subplots for histogram and density plots
plt.figure(figsize=(15, 10))

# Unimodal
plt.subplot(3, 2, 1)
sns.histplot(data_unimodal, bins=30, kde=False, color='skyblue')
plt.title("Unimodal Histogram")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.subplot(3, 2, 2)
sns.kdeplot(data_unimodal, color='blue')
plt.title("Unimodal Density Plot")
plt.xlabel("Value")

# Bimodal
plt.subplot(3, 2, 3)
sns.histplot(data_bimodal, bins=30, kde=False, color='lightgreen')
plt.title("Bimodal Histogram")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.subplot(3, 2, 4)
sns.kdeplot(data_bimodal, color='green')
plt.title("Bimodal Density Plot")
plt.xlabel("Value")

# Multimodal
plt.subplot(3, 2, 5)
sns.histplot(data_multimodal, bins=30, kde=False, color='lightcoral')
plt.title("Multimodal Histogram")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.subplot(3, 2, 6)
sns.kdeplot(data_multimodal, color='red')
plt.title("Multimodal Density Plot")
plt.xlabel("Value")

# Adjust layout
plt.tight_layout()
plt.show()
```


Unimodal Histogram
Frequency
Value
Unimodal Density Plot
Density
Value
Bimodal Histogram
Frequency
Value
Bimodal Density Plot
Density
Value
Multimodal Histogram
Frequency
Value
Multimodal Density Plot
Density
Value

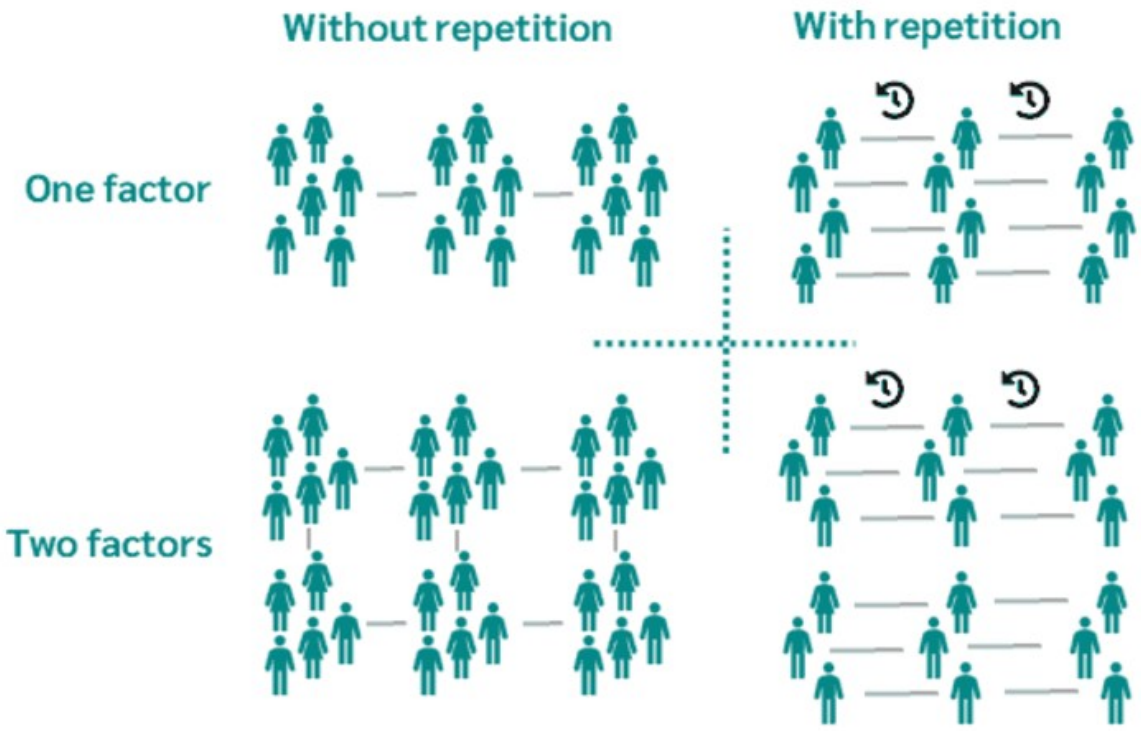

Without repetition
With repetition
One factor
Two factors


Without repetition
With repetition
One factor
Two factors



Probability

Probability is a way of measuring how likely it is that something will happen. It helps us understand and make decisions when we're not sure about the outcome of an event. For example, when flipping a coin, probability tells us there's a 50% (0.5)chance of getting heads and a 50% (0.5) chance of getting tails. It’s a way of expressing the chance of different possible outcomes.



Formula:

P(A) = Number of favourable outcomes / Total number of possible outcomes

Where:

$0 \leq P(A) \leq 1$

P(A)=0 means the event is impossible.

P(A)=1 means the event is certain.

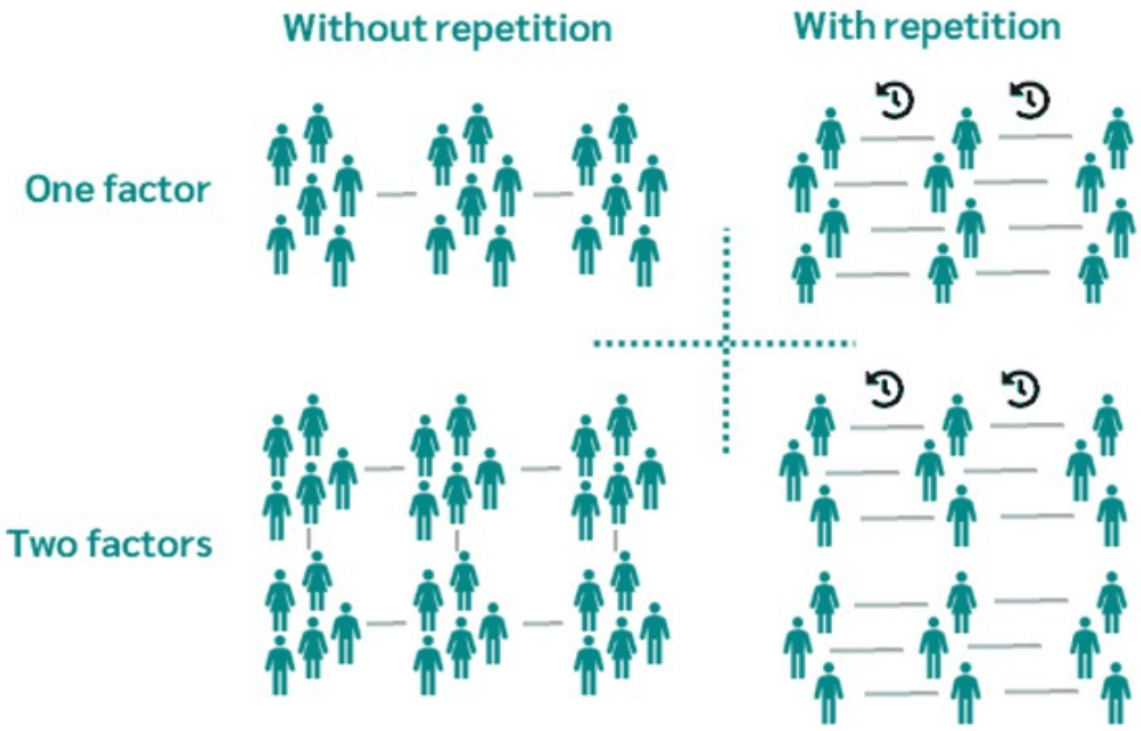
Eg1: The probability of rolling a 4 on a six-sided die is: P(4) = 1/6 (preferred outcome = 1, total possible outcomes =6)

Eg2: Divisible by 3 numbers on a die:(3,6)

preferred outcome =2

Total outcome =6

so, P(A) = 2/6 = 0.33



The probability of 2 independent events occurring at the same time is equal to the product of all the probability of individual event.

P(A and B) = P(A).P(B)

eg: P(Ace spade) = P(Ace).P(Spade)

Exercise

1.A standard die has 6 sides, numbered from 1 to 6. If you roll the die once, what is the probability of:

Rolling a 3?

Rolling an even number?

1. You flip a coin twice. What is the probability of:

Getting heads both times?

Getting one head and one tail (in any order)?

1. A standard deck of cards has 52 cards. What is the probability of drawing:

A heart?

A queen?

1. A bag contains 3 red marbles, 4 blue marbles, and 5 green marbles. If you randomly pick one marble, what is the probability that it is:

Red?

Not blue?

1. A spinner is divided into 8 equal sections, numbered 1 to 8. What is the probability of spinning:

An odd number?

A number greater than 6?

1. $1/6$, $3/6$
2. $1/2$*$1/2$ = $1/4$, $2/4$
3. $13/52$, $4/52$
4. $3/12$, $8/12$
5. $4/8$, $2/8$

Key Terms:

Experiment: An action or process that results in outcomes (e.g., rolling a die, flipping a coin).

Sample Space (S): The set of all possible outcomes of an experiment. Example: For a die, S={1,2,3,4,5,6}

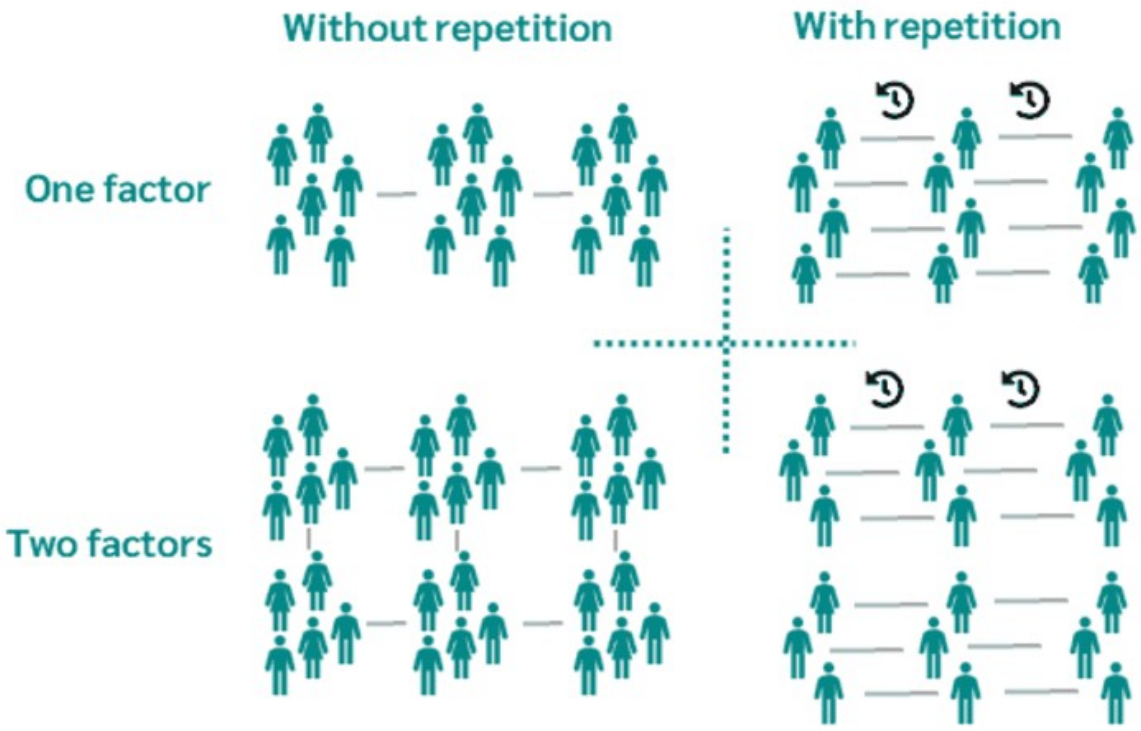
Event (E): A subset of the sample space, which represents one or more outcomes of experiment. Example: Rolling an even number E={2,4,6}

Sample Point: One of the possible results in an experiment. Example: in rolling a fair six-sided dice, sample points are 1 to 6.

Favorable Outcome:An outcome that produces the desired or expected consequence.

Probability of an Event

Probability of an event is a measure of the likelihood that the event will occur, expressed as a number between 0 and 1. An event with a probability of 1 is considered certain to happen, while an event with a probability of 0 is certain not to happen.



Experimental probabilities

```
    The probabilities, we get after conducting an experiment, are
called experimental probabilities.
```

Theoretical probabilities

The ones we declare before doing the experiment are theoretical(true) probabilities.

The experimental probabilites are not always equal to theoretical probabilities, but they're a good approximation. They're easy to compute.

$$P(A)=\frac{Successful-trials}{All-trials}$$

Experimental probabilities are useful because they provide an empirical way to estimate the likelihood of an event based on actual experiments or observations. When we conduct an experiment multiple times and record the outcomes, the **experimental probability is calculated as the ratio of the number of successful outcomes to the total number of trials**. This makes it relatively easy to compute, especially when theoretical probabilities are difficult to determine due to complex or unknown factors.

Moreover, experimental probabilities often serve as good predictors for theoretical probabilities, especially when a large number of trials are conducted. **As the number of trials increases, the Law of Large Numbers states that the experimental probability tends to converge towards the theoretical probability, assuming the trials are fair and unbiased**. Therefore, experimental probabilities provide a practical approach to estimating theoretical probabilities when direct calculation is challenging.

Expected Values

The average outcome we expect if we run an experiment many times.

Experiment is a process that results in an outcome A.

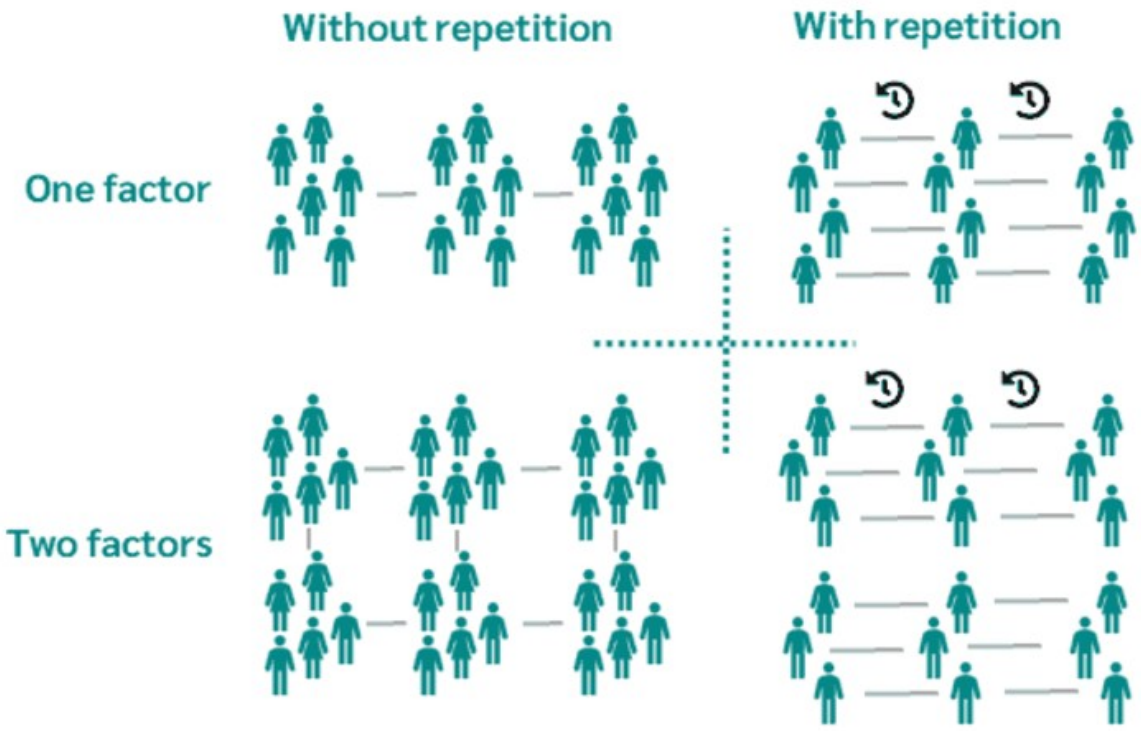
Expected value, E(A) is the outcome we expect to occur when we run an experiment.

eg:

1. Tossing coins 20 times and recording the outocme - single experiment with 20 trials
2. How many spades will we get when drawing a card 20 times?

E(A) = P(A) * n = 0.25*20 = 5(we expect to get the spade 5 times, but it can be more or less)

For Multiple events: E(A) = P(A)*A + P(B)*B +P(C)*C....+P(n)*n



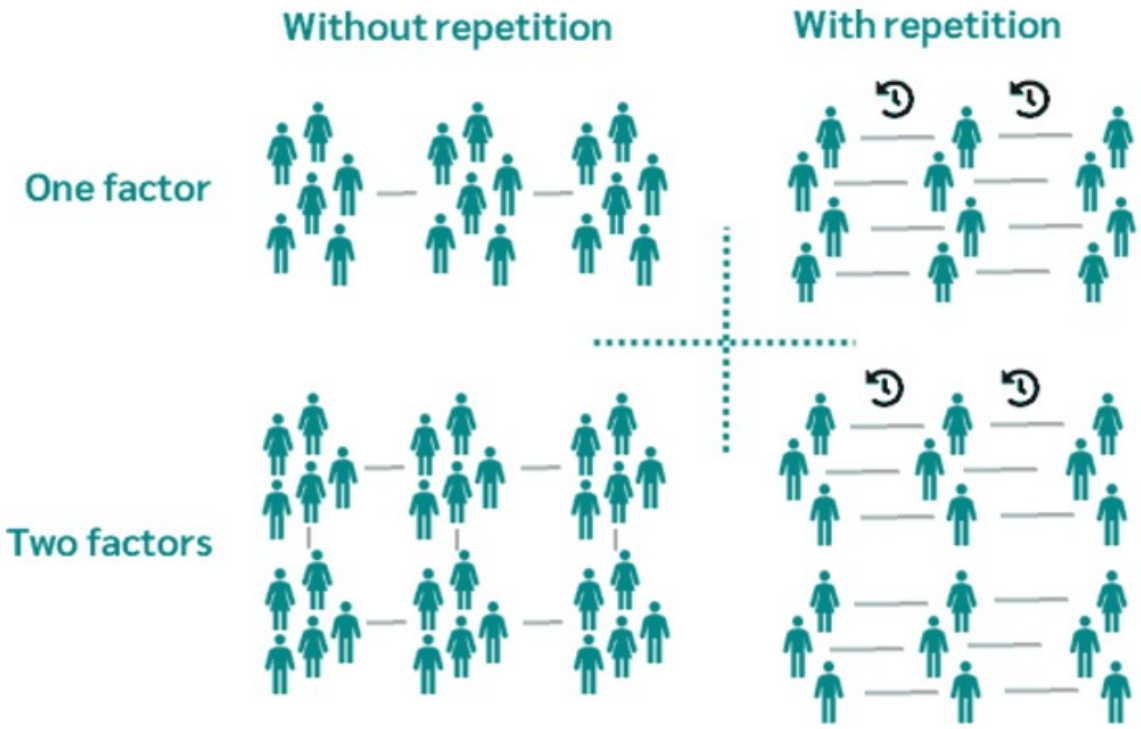
P(7) = 6/36 = 1/6

E(A) = P(2)*2 + P(3)*3* + P(4)*4 + P(5)*5* + P(6)*6* + P(7)*7* + P(8)*8* + P(9)*9* + P(10)*10* + P(11)*11* + P(12)*12

Probability Frequency Distribution

It is a collection of the probabilities for each possible outcome.

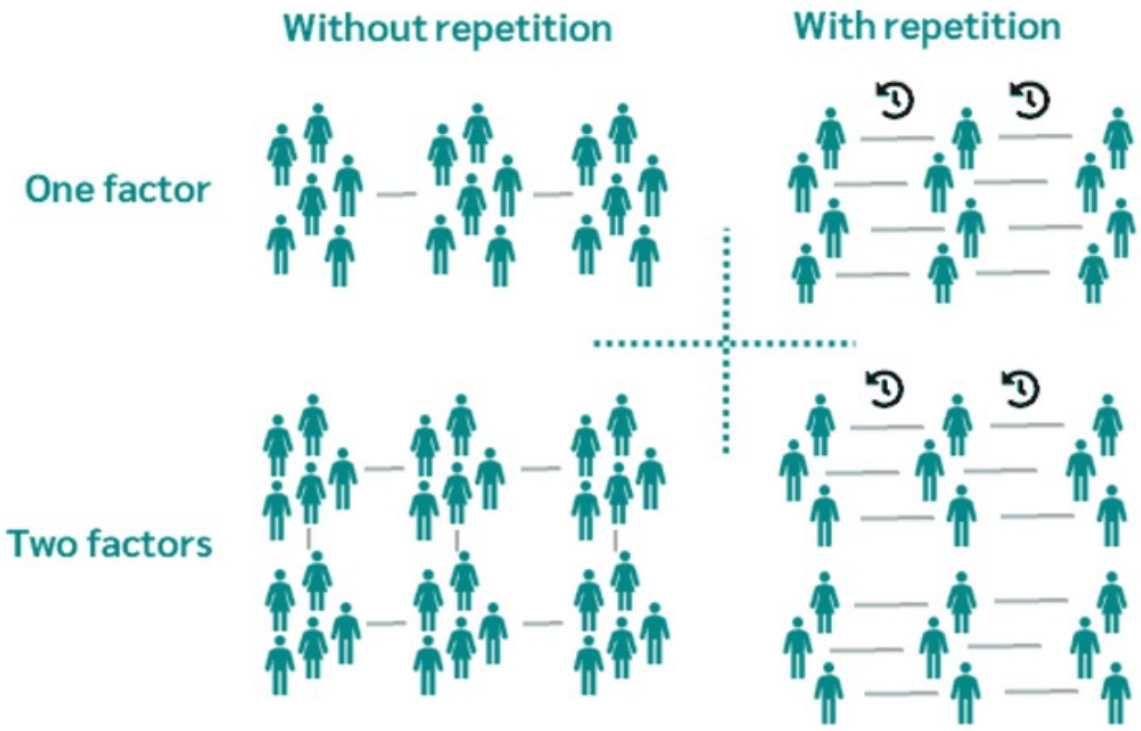
In probability theory, the frequency of a value within a sample space represents the number of times that specific outcome occurs in a set of trials. For example, if we roll a six-sided die 36 times and the number 4 appears 6 times, then the frequency of rolling a 4 is 6. Frequency is a fundamental concept in constructing a frequency distribution table, which shows how often each outcome occurs.



E(A) = 1/36*2 + 2/36*3 + 3/36 * 4* + 4/36*5* + 5/36 * 6* + 6/36*7* + 5/36 * 8* + 4/36 * 9* + 3/36*10* + 2/36*11 + 1/36*12

```
= 7 (7 is the most probable sum of 2 dice)
```

P(E(A)) = P(7) = 1/6



Highest Frequency = Highest Probability

Exercise

Problem 1: A fair six-sided die is rolled. Let X represent the outcome of the roll. What is the expected value of X?

Problem 2: A game involves drawing a card from a deck of 52 cards. If you draw an Ace, you win \$100. If you draw a King, a Queen, or a Jack, you win \$50. For any other card, you win nothing. What is the expected value of your winnings?

Problem 3: A bag contains 4 red balls, 3 blue balls, and 3 green balls. You randomly draw one ball from the bag. If you draw a red ball, you win \$10, if you draw a blue ball, you win \$20, and if you draw a green ball, you win \$30. What is the expected value of your winnings?

Answers

1. $E(X) = 1 \times 1/6 + 2 \times 1/6 + 3 \times 1/6 + 4 \times 1/6 + 5 \times 1/6 + 6 \times 1/6 = 21/6 = 3.5$

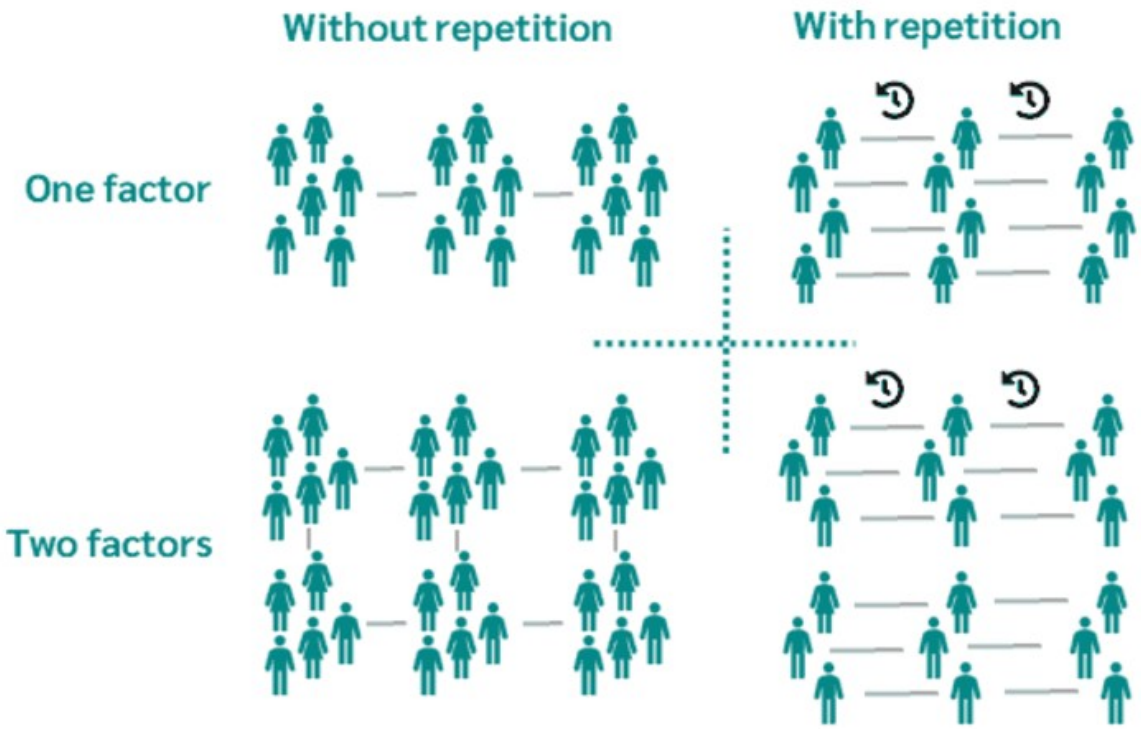
2. $E(\text{Win}) = 100 \times 4/52 + 50 \times 12/52 = 400/52 + 600/52 = 19.23$

3. $E(\text{Win}) = 10 \times 4/10 + 20 \times 3/10 + 30 \times 3/10 = 4 + 6 + 9 = 19$

Complements

Complement of an event is everything that is not an event.

A+Ac= Sample space



eg: Coin Toss A -> Heads B -> Tails P(A)+P(B) = 1 -> 100% certain

If, P(A) + P(B) >1, then we have double-counted some trials.

If, P(A) + P(B) <1, then we have missed counting some possible outcomes.

All events have complements.

P(A) + P(A') = 1

eg: Rolling a die

A -> Rolling an even number

A' -> Not rolling an even number(Rolling an odd number)

P(A') = 1 - P(A)

eg: Probability of getting numbers other than 3

P(A) = P(1) +P(2) + P(4) +P(5) + P(6) = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 5/6

P(A') = P(3) = 1/6

P(A') = 1 - 5/6 = 1/6

Combinatorics

It deals with combinations of objects from a specific finite set.

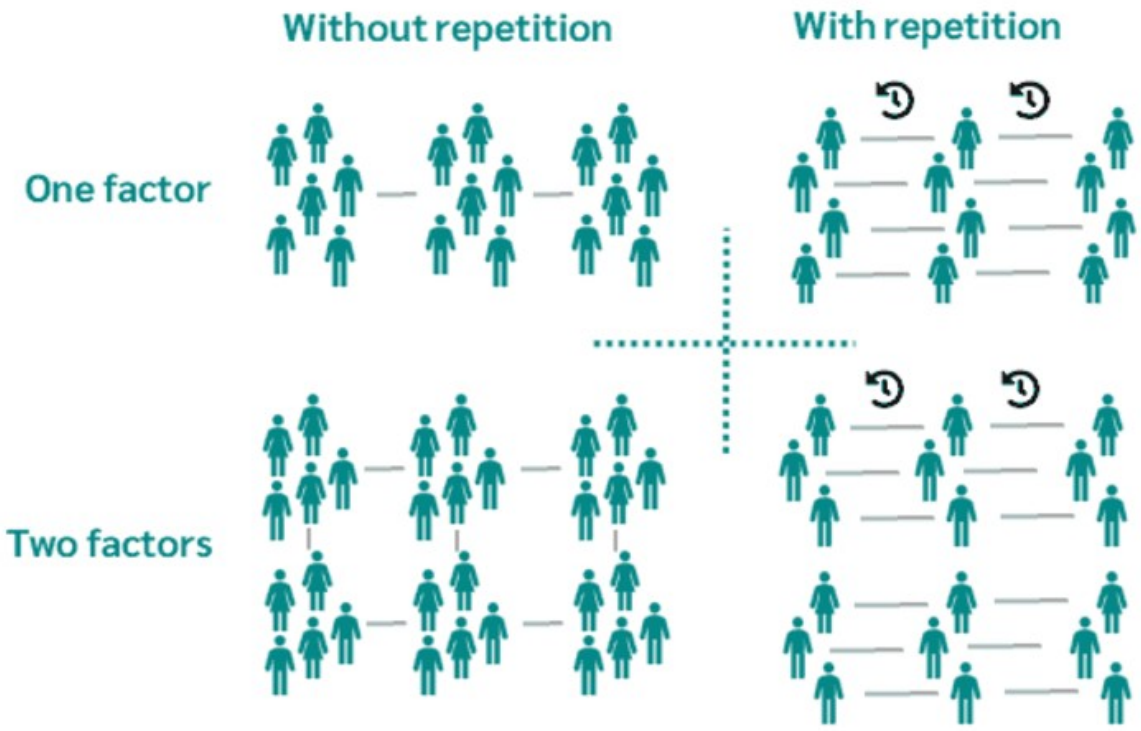
3 integral part of combinatorics:

1. Permutations
2. Variations
3. Combinations

   These 3 are used to determine the number of favourable outcomes or number of all elements in a sample space.

Permutations

The number of different possible ways we can arrange a set of elements. These elements can be digits, letters, objects or even people.



We can arrange the balls in 6 unique ways. These 6 ways are called permutations.

The number of permutations of n distinct objects is given by n! (n factorial)

$$P_n = n!$$

Variations

The total number of ways, we can pick and arrange some elements of a given set.

Variations with Repetition

eg: 2-letter code with 3 letters A,B,C

AA,AB,AC - 3 options to start with A

BA,BB,BC - 3 options to start with B

CA,CB,CC - 3 options to start with C

The total number of variations = 3*3 =9

Formula for variations(with repetition) $V_p^n = n^p$

n- total no. of elements

p - the number of positions we need to fill

For the letter code above,

$$V_2^3 = 3^2$$

Formula for Variations (without Repetition)

$$V_p^n = \frac{n!}{(n-p)!}$$

Combinations

The number of different ways, we can pick certain elements of a set.

eg: Pick 3 people for a conference from 10 people.

Formula for combination(without repetition)

$$C_p^n = \frac{V_p^n}{P_p} = \frac{n!}{p! \times (n-p)!}$$

This is important formula that we will use later for binomial distribution

Formula for combination(with repetition)

$$C(n, p)=\frac{(n+p-1)!}{(n-1)!P!}$$

Permutations are all possible arrangements of a set (order matters, all elements are used). Think of this as the most specific case.

$$P_p=p!$$

Variations are arrangements of a subset of the set (order matters, not all elements are used). This is a more general case of permutations.

$$V_p^n=\frac{n!}{(n-p)!}$$

Combinations are selections of a subset of the set (order does not matter, not all elements are used). This is the most general case.

$$C_p^n=\frac{V_p^n}{P_p}=\binom{n}{p}=\frac{n!}{p!\times(n-p)!}$$

The expression

$$\binom{n}{p}$$

is called a binomial coefficient, also known as n-choose-p.

Summary

1. **Arrange** - permutations (eg: 4 runners, 4 positions)

2. **Pick and Arrange** - Variations(eg: 6 runners, 4 positions, 4 spots)

3. **Pick** - Combinations(Order is irrelevant)(eg: 6 runners,4 spots)

Exercise for combination without repetition

Problem 1: A committee of 4 people is to be selected from a group of 10 people. How many different ways can the committee be chosen?

Problem 2: You have 8 different books, and you want to select 3 books to take on a trip. In how many different ways can you select the books?

Problem 3: A basketball team consists of 12 players. The coach needs to select 5 players to start the game. How many different groups of 5 players can be selected?

Answer

1. $$C_4^{10} = \frac{10!}{4! \times (10-4)!} = \frac{10!}{4! \times (6!)} = 210$$

2. $$C_3^8 = \frac{8!}{3! \times (8-3)!} = \frac{8!}{3! \times (5!)} = 56$$

3. $$C_5^{12} = \frac{12!}{5! \times (12-5)!} = \frac{12!}{5! \times (7!)} = 792$$

Probability Rules

1. Addition Rule

Used to find the probability of the union of two events (the probability that at least one of the events occurs).

For **mutually exclusive** events A and B: **P(A U B) = P(A) + P(B)**

For **non-mutually exclusive** events A and B: **P(A U B) = P(A) + P(B) - P(A ∩ B)**

Example: In a standard deck of cards, the probability of drawing a heart or a queen:

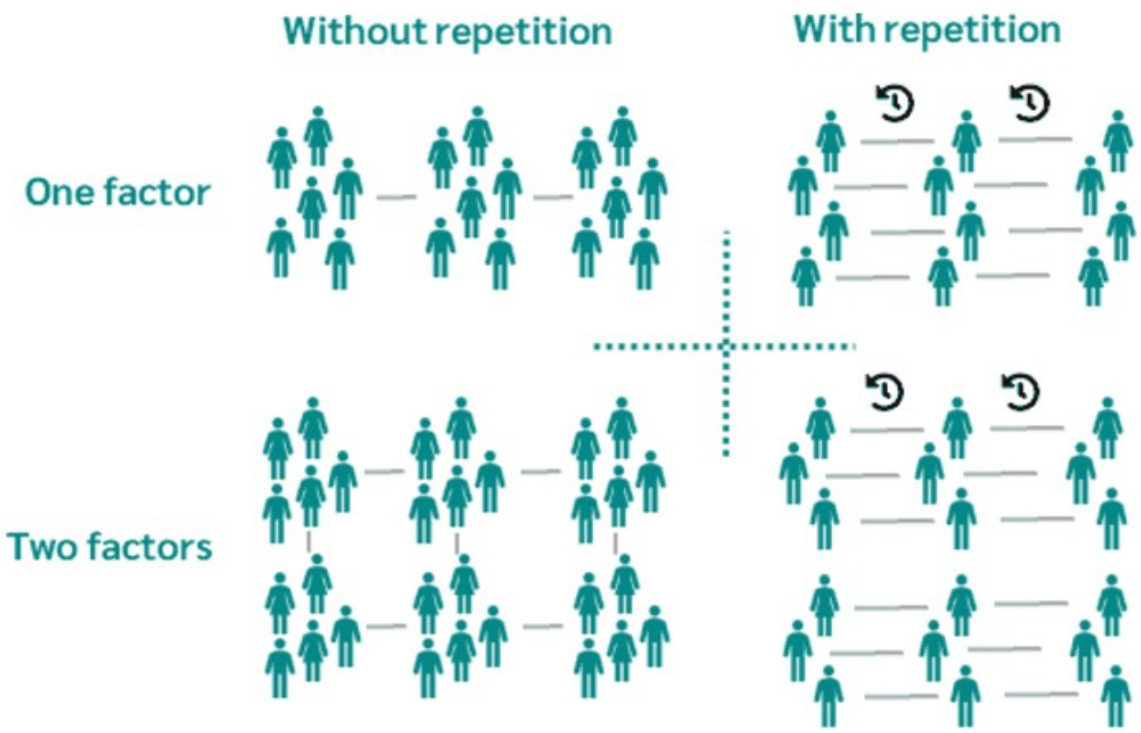
P(Heart) = 13/52

P(Queen) = 4/52

P(Queen of Hearts) = 1/52

P(Heart or Queen) = P(Heart)+P(Queen)−P(Queen of Hearts)

```
                = 13/52 + 4/52 - 1/52

                 = 4/13
```



Exercise

Problem 1: A card is drawn from a standard deck of 52 cards. What is the probability that the card is either a King or a Heart?

Problem 2: In a survey of 100 people, 40 like pizza, 30 like burgers, and 15 like both pizza and burgers. What is the probability that a randomly selected person likes either pizza or burgers?

Problem 3: A die is rolled once. What is the probability of rolling either a 3 or an even number?

Problem 4: A box contains 5 red, 4 green, and 3 blue marbles. What is the probability of drawing either a red or a blue marble?

Problem 5: In a class of 30 students, 12 students are part of the drama club, 15 are part of the music club, and 5 are part of both. What is the probability that a randomly selected student is either in the drama club or the music club?

Answer

1. P(K U H) = P(K) + P(H) - P(K ∩ H) = 4/52 + 13/52 - 1/52 = 16/52 = 4/13

2. P(P U B) = P(P) + P(B) - P(P ∩ B) = 40/100 + 30/100 - 15/100 = 55/100 = 0.55

3. P(3 U E) = P(3) + P(E) = 1/6 + 3/6 = 4/6 = 2/3

4. P(R U B) = P(R) + P(B) = 5/12 + 3/12 = 8/12 = 2/3

5. P(D U M) = P(D) + P(M) - P(D ∩ M) = 12/30 + 15/30 - 5/30 = 22/30 = 11/15

1. **Complementary Rule**

The probability of an event not occurring is equal to 1 minus the probability of the event occurring.

Whenever an event is the complement of another event, specifically, if A is an event,

then P(not A)=1−P(A) or **P(A') = 1 - P(A)**.

Example: If the probability of raining tomorrow is P(Rain) = 0.3,

then the probability of not raining is:

P(Not Rain) = 1−P(Rain) = 1−0.3 = 0.7

Exercise

Problem 1: The probability of rolling a number less than 5 on a fair six-sided die is 2/3. What is the probability of not rolling a number less than 5?

Problem 2: A bag contains 20 marbles: 7 red, 8 green, and 5 blue. If a marble is drawn randomly, what is the probability that it is not red?

Problem 3: The probability of passing an exam is 0.85. What is the probability that a student does not pass the exam?

Answer

1. P(Not less than 5) = 1 - P(Less than 5) = 1 - 2/3 = 1/3

2. P(Not Red) = 1 - P(Red) = 1 - 7/20 = 13/20

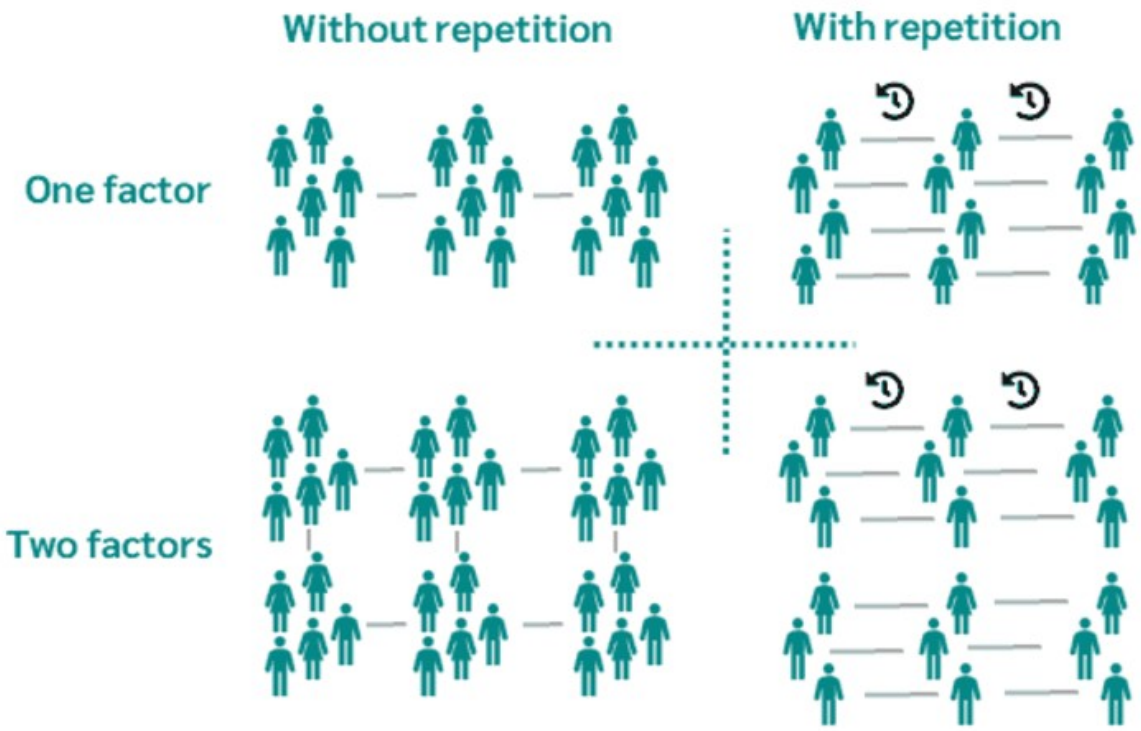
3. P(Not passing) = 1 - P(Passing) = 1 - 0.85 = 0.15

3. **Multiplication Rule**

Whenever an event is the intersection of two other events, that is, events A and B need to occur simultaneously.

Then **P(A and B) = P(A n B) = P(A) x P(B)**. (For independent events)

P (A ∩ B) = P (A) x P (B | A) (For dependent events) . P(B | A) is the probability of event B given that event A happened



Exercise

Problem 1 : The probability that it rains tomorrow is 0.6, and the probability that your friend comes to visit is 0.5. Assuming these events are independent, what is the probability that it rains and your friend visits on the same day?

Problem 2 : A box contains 5 red balls and 3 green balls. Two balls are drawn without replacement. What is the probability that the first ball drawn is red and the second ball drawn is green?

Problem 3 : A student has a 0.7 probability of passing a math test and a 0.8 probability of passing a science test. What is the probability that the student passes both the math and science tests, assuming the events are independent?

Problem 4 : A bag contains 6 blue and 4 yellow marbles. Two marbles are drawn without replacement. What is the probability that the first marble drawn is blue and the second marble drawn is yellow?

Problem 5 : A deck of 52 playing cards contains 4 Aces. Two cards are drawn without replacement. What is the probability that both cards drawn are Aces?

Answer

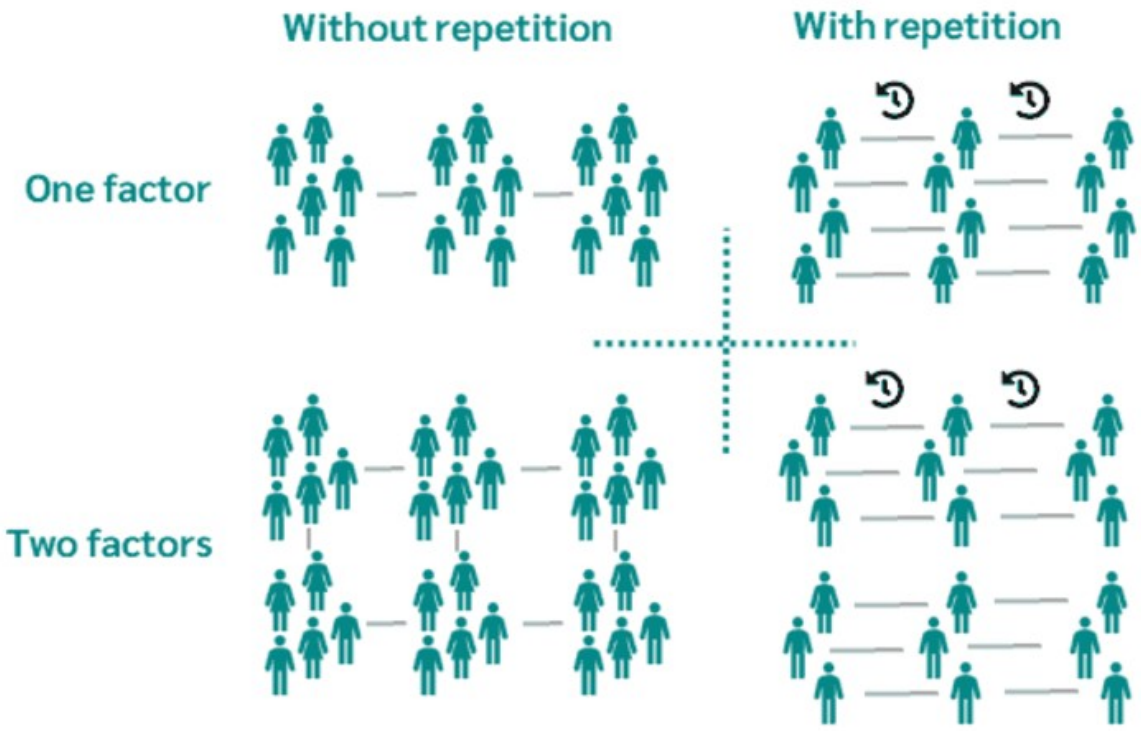
1. P(Rain n Friend) = P(Rain) x P(Friend) = 0.6*0.5 = 0.3

2. P(Red) = 5/8 , P(Green|Red) = 3/7, P(Red and Green) = P(Red) x P(Green|Red) = 5/8 * 3/7 = 15/56

3. P(Math n Science) = 0.7*0.8 = 0.56

4. P(Blue) = 6/10, P(Yellow|Blue) = 4/9 , P(B n Y) = 6/10 * 4/9 = 24/90 = 4/15

5. P(Ace first) = 4/52, P(Ace second|Ace first) = 3/51 , P(Ace n Ace) = 4/52 * 3/51 = 12/2652 = 1/221



Conditional Probability

The likelihood of an event occurring, assuming a different one has already happened.

Independent events

Eg: 2 coin flips

A -> head **P(A) = 0.5

B ->head(on the previous flip) **P(A|B) = 0.5**

The probability of getting heads now, after getting heads the last time is still 0.5.

Therefore, P(A) = P(A|B) -> two events are independent. **P(A ∩ B) = P(A) x P(B)

Dependent events

eg: Queen of spades

A -> Queen of spades P(A) = 1/52

B -> spade P(A|B) = 1/13

C -> Queen P(A|C) = 1/4

Since, P(A) is not equal P(A|B) -> two events(A and B) are dependent.

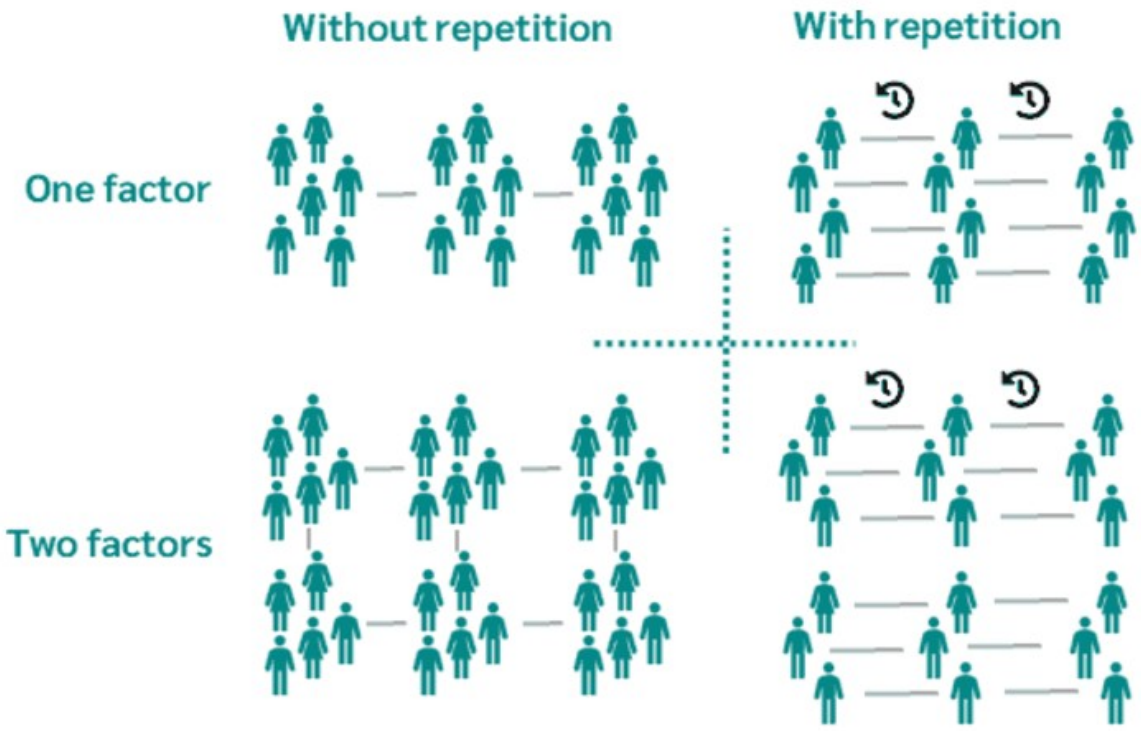
similarly, P(C) is not equal to P(A|C) -> A and C are dependent.

Formula for conditional probability

$$P(A \vee B) = \frac{P(A \cap B)}{P(B)}$$

The conditional probability requires that event B occur

P(A|B) is not equal to P(B|A)



Exercise

Problem 1: A deck of 52 playing cards contains 13 cards of each suit (hearts, diamonds, clubs, and spades). What is the probability of drawing a heart, given that the card drawn is red?

Problem 2: A survey finds that 30% of people own a smartphone, 20% own a tablet, and 15% own both a smartphone and a tablet. If a person owns a smartphone, what is the probability that they also own a tablet?

Problem 3: In a certain class, 60% of students passed the math exam, and 40% passed both the math and science exams. What is the probability that a student passed the science exam, given that they passed the math exam?

Answer

1. $P(A) = \text{heart}, P(B) = \text{red}, P(A|B) =$

$$\frac{(13/52)}{(26/52)}$$

$= 1/2 = 0.5$

2. $P(T|S) = 0.15/0.30 = 0.5$

3. $P(S|M) = 0.4/0.6 = 2/3 = 0.67$

Joint Probability

Joint probability refers to the probability of two (or more) events happening at the same time. It represents the likelihood of the simultaneous occurrence of two or more random events.

Formula for Joint probability:

1.**For Independent Events** When events A and B are independent, meaning that the occurrence of one event does not impact the other, we use the multiplication rule: $P(A \cap B) = P(A) \times P(B)$

Eg: Event A: You flip a coin, and it lands on heads.

Event B: You roll a 6-sided die, and it lands on a 3.

These two events are independent because the outcome of the coin flip does not affect the outcome of the die roll.

P(A) = 1/2

P(B) = 1/6

$P(A \cap B) = 1/2 * 1/6 = 1/12$. There is 1/12 chance that both events occur.

2.**For Dependent Events** Events are often dependent on each other, meaning that one event's occurrence influences the likelihood of the other. Here, we employ a modified formula: $P(A \cap B) = P(A) \times P(B|A)$

Eg: Event A: You draw a card from a deck, and it's an Ace.

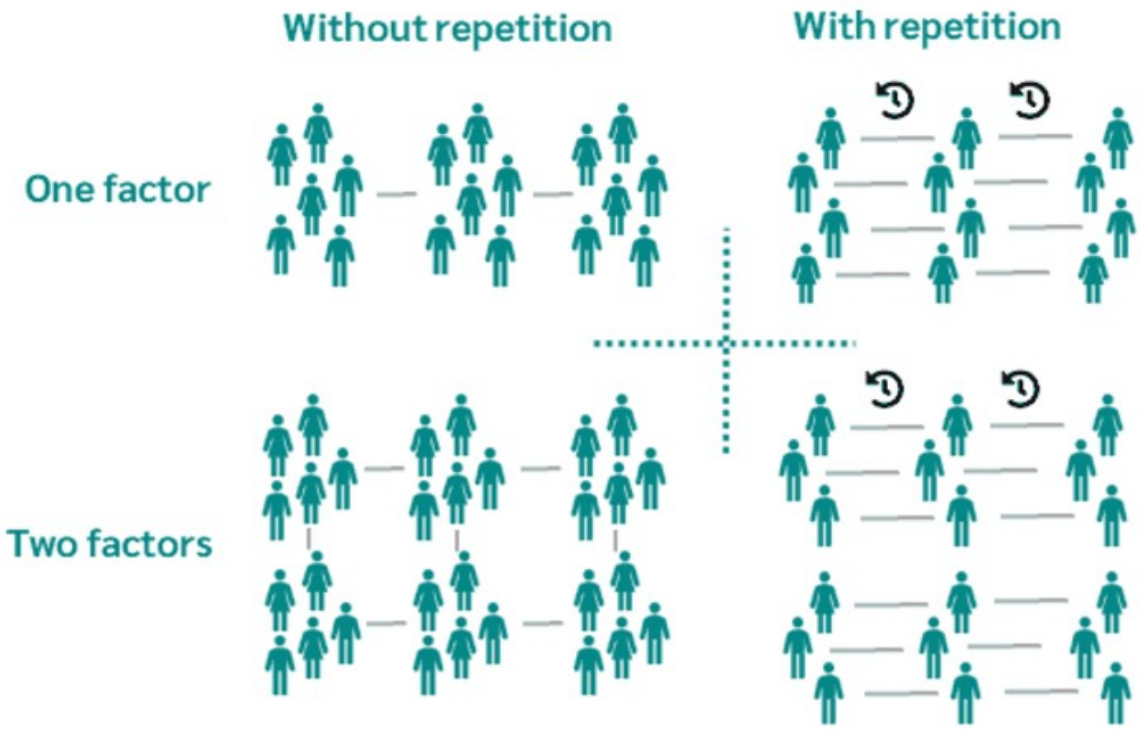
Event B: You draw a second card, and it's also an Ace (without replacement).

In this case, the events are dependent because after drawing the first card, the total number of cards and the number of Aces left in the deck both change.

P(A) = 4/52

P(B|A) = 3/51

$P(A \cap B) = 4/52 * 3/51 = 3/663 = 0.0045$. This means there is about a 0.45% chance of drawing two Aces consecutively without replacement.



Exercise

1. Suppose you are running an e-commerce platform, and you want to find the probability of a customer purchasing a red shirt (event A) and a blue hat (event B) independently. Find out the Joint Probability where

P(A): The probability of a customer buying a red shirt is 0.3.

P(B): The probability of a customer purchasing a blue hat is 0.2.

2. You roll two 6-sided dice.

Event A: The first die shows a 4.

Event B: The sum of the two dice is 7.

What is the joint probability $P(A \cap B)$ of rolling a 4 on the first die and getting a sum of 7 on both dice?

3. A deck contains 5 red cards and 5 black cards. You draw two cards without replacement.

Event A: The first card drawn is red.

Event B: The second card drawn is black.

What is the joint probability $P(A \cap B)$ of drawing a red card first and a black card second?

Answer

1.$P(A \cap B) = P(A) \times P(B)$

$P(A \cap B)$ = P(customer buying a red shirt) x P(customer buying a blue hat)

$P(A \cap B) = 0.3 \times 0.2$

$P(A \cap B) = 0.06$

2. You roll two dice:

The probability of rolling a 4 on the first die is P(A) = 1/6

For the sum of both dice to be 7 and the first die is already 4, the second die must show 3. The probability of rolling a 3 on the second die is P(B|A) = 1/6

$P(A \cap B) = 1/6 * 1/6 = 1/36$

3. You draw two cards from a deck of 5 red and 5 black cards without replacement:

The probability of drawing a red card first is P(A) = 5/10 = 1/2

After drawing a red card, there are 9 cards left (4 red, 5 black), so the probability of drawing a black card next is P(B|A) = 5/9

$P(A \cap B) = 1/2*5/9 = 5/18$

Marginal Probability

Marginal Probability can be defined as the probability of an event occurring irrespective of the outcome of another event. It's termed 'marginal' because it's derived from the margins of a probability table, summarising the probabilities of various variables. **In simple terms, marginal probability is the likelihood of a specific event happening, without considering other related events.**

For eg: In a study of students' performance in maths and English, the marginal probability might tell you the likelihood of students passing maths, ignoring their performance in English.

Theorem/Law of Total Probability

The Law of Total Probability is a fundamental rule in probability theory that helps calculate the probability of an event by breaking it down into different possible conditions or outcomes. It states that if an event A can happen under mutually exclusive and exhaustive conditions (events B1, B2, ...,Bn), the total probability of A can be found by summing the probabilities of A occurring in each condition, weighted by the probability of that condition.

Formula:

$$P(A)=P(A\vee B1)\,P(B1)+P(A\vee B2)\,P(B2)+\cdots+P(A\vee Bn)\,P(Bn)$$

where,

P(A|Bi) is the conditional probability of A given Bi

p(Bi) is the probability of condition Bi

B1 ,B2 ,…,Bn are mutually exclusive and exhaustive events.

Example 1 (Red Ball from Two Bags):

You have two bags:

Bag 1 contains 3 red and 5 black balls.

Bag 2 contains 2 red and 4 black balls.

A bag is chosen at random, and a ball is drawn. What is the probability of drawing a red ball?

Here, the event A is drawing a red ball, and the events B1and B2are choosing Bag 1 and Bag 2, respectively.

The total probability of drawing a red ball is:

P(Red Ball)=P(Red Ball∣Bag 1)⋅P(Bag 1)+P(Red Ball∣Bag 2)⋅P(Bag 2)

Example 2 (Late to Work): In a city, 60% of people use public transportation (B1), and 40% use private vehicles (B2). The probability of being late to work (A) is 20% for public transport users and 10% for private vehicle users. What is the overall probability that a person is late to work?

P (Late) = P (Late | Public) ⋅ P (Public) + P (Late | Private) ⋅ P (Private)

Connection to Bayes' Theorem:

The Law of Total Probability forms the core foundation of Bayes' Theorem. Bayes' Theorem uses the Law of Total Probability to calculate the marginal probability of the evidence (denoted as P(B)), which is essential for updating the probability of a hypothesis based on new evidence.

Exercise

1.In a locality, there are two clinics. On a particular day [100] patients visit clinic A and [50] patients visit clinic B. Further, the probability that the patient visiting is suffering from a heart disease is [0.4] for clinic A, while it is [0.3] for clinic B. A patient visiting one of these two clinics on that particular day is chosen at random. Find the probability that he or she suffers from a heart disease.

2.There are two bags. Bag [I] contains [3] red and [5] black balls, and Bag [II] contains [2] red and [4] black balls. A bag is chosen at random and a ball is taken out. Find the probability that the ball is red.

3.In a city, there are [3000] Hyundai cars, [8000] Maruti cars and [2000] Honda cars. The probabilities that there is a manufacturing defect in the car are [4 %], [6 %] and [3 %] for Hyundai, Maruti and Honda respectively. Find the probability that a car picked at random has a manufacturing defect.

Answer

1. Total Patients = 150

P(A) = 100/150 = 2/3

p(B) = 50/150 = 1/3

P(Heart Disease|A) = 0.4

P(Heart Disease|B) = 0.3

P(Heart Disease) = P(Heart Disease|A)*P(A) + P(Heart Disease|B)*P(B)

=(0.4*2/3) + (0.3*1/3) = 0.367

The probability that a randomly chosen patient suffers from heart disease is approximately **0.367 or 36.7%.**

1. P(Bag I) = 1/2

P(Bag II) = 1/2

P(Red Ball|Bag I) = 3/8

P(Red Ball|Bag II) = 2/6

P(Red Ball) = P(Red Ball|Bag I)*P(Bag I) + P(Red Ball|Bag II)*P(Bag II)

=(3/8 * 1/2) + (2/6 * 1/2) = 3/16+2/12 = 9/48 + 8/48 =17/48 = 0.354

3.P(Hyundai) = 3000/13000

P(Maruti) = 8000/13000

P(Honda) = 2000/13000

P(Defect|Hyundai) = 0.04

P(Defect|Maruti) = 0.06

P(Defect|Honda) = 0.03

P(Defect) = P(Defect|Hyundai)*P(Hyundai) + P(Defect|Maruti)*P(Maruti) + P(Defect|Honda)*P(Honda)

=(0.04 * 3000/13000)+(0.06 * 8000/13000)+(0.03 * 2000/13000)

Bayes' Theorem

Bayes' Theorem helps us update the probability of a hypothesis (event A) after observing new evidence (event B). It allows us to combine prior knowledge with new data to make more accurate predictions.

It is a mathematical formula which is used to determine the conditional probability of a given event.

Formula

$$P(A \vee B) = \frac{P(B \vee A) * P(A)}{P(B)}$$

where:

P(A|B) is the **posterior probability**, the probability of event A occurring given that event B has occurred.

P(B|A) is the **likelihood**, the probability of event B occurring given that event A has occurred.

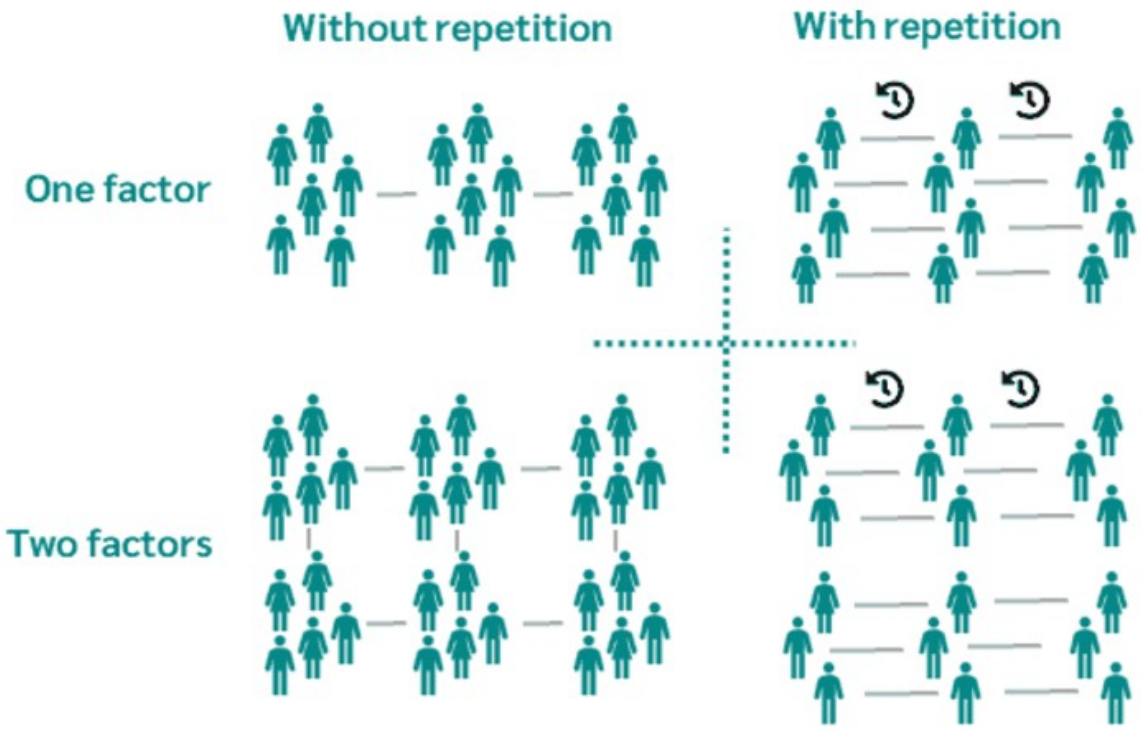
P(A) is the **prior** probability of event A (before knowing about B)

P(B) is the **marginal likelihood** the total probability of event B occurring.

In this formula,

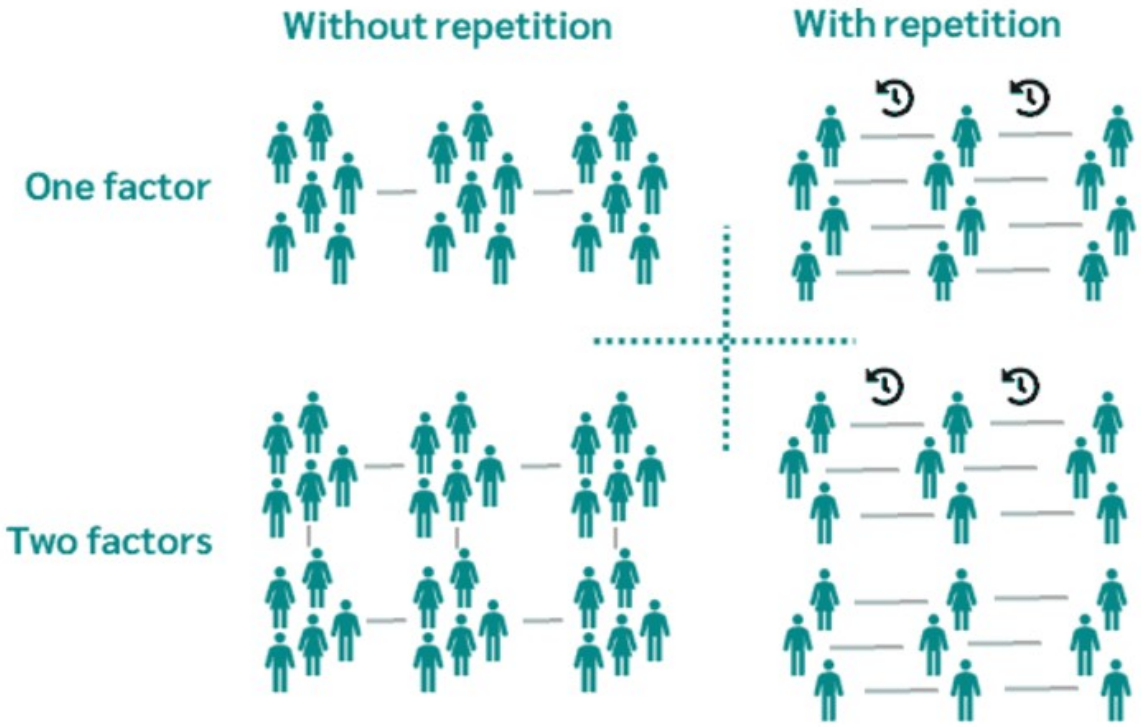
P(B) (the marginal probability of the evidence) is often calculated using the Law of Total Probability as follows:

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid A) \cdot P(A)$$



As Bayes' theorem is based on conditional probability, if we know conditional probability, we can find the reverse probability using Bayes' theorem.

Example



Task:

We want to calculate the probability that a transaction is fraudulent given that it has both high amount and foreign location indicators, or P(Fraud | High Amount and Foreign Location).

A: Transaction is fraudulent.

B: Transaction has both a high amount and foreign location.

P(A): Probability of a transaction being fraudulent.

P(B): Probability of a transaction having both a high amount and foreign location.

P(B|A): Probability that a transaction has both indicators, given that it is fraudulent.

Total transactions = 10

Fraudulent transactions = 4

P(A) = 4/10 = 0.4

Transactions with both high amount and foreign location = 3

P(B) = 3/10 = 0.3

Fraudulent transactions with both indicators = 3

P(B|A)= 3/4 = 0.75

Bayes' Theorem Fomula:

$$P(A \vee B) = \frac{P(B \vee A) * P(A)}{P(B)}$$

$$P(A \vee B) = \frac{0.75 * 0.4}{0.3}$$

= 0.1

The probability that a transaction is fraudulent given that it has both a high amount and is in a foreign location is 1.0 or 100%.

Real world applications of Bayes' theorem

Medical Diagnosis: Bayes' Theorem is used to update the probability of a disease after getting test results, considering both the test's accuracy and the disease's prevalence. It helps doctors give more accurate risk assessments after positive or negative tests.

$$P(Disease \vee PositiveTest) = \frac{P(PositiveTest \vee Disease) * P(Disease)}{P(PositiveTest)}$$

Spam Filtering: Email systems use Bayes' Theorem to calculate the probability that an email is spam based on keywords. It continuously updates as more emails are flagged, improving accuracy in identifying spam.

$$P(Spam \vee Keyword) = \frac{P(Keyword \vee Spam) * P(Spam)}{P(Keyword)}$$

Machine Learning (Naive Bayes): Bayes' Theorem is the basis for Naive Bayes classifiers, commonly used in text classification tasks like sentiment analysis. It updates predictions as new data is encountered.

$$P(Class \vee X) = \frac{P(X \vee Class) * P(Class)}{P(X)}$$

Weather Forecasting: Bayes' Theorem helps meteorologists update weather predictions (like rain) as new data (cloud patterns, wind, etc.) becomes available, allowing for more accurate forecasts.

$$P(Rain \vee CloudData) = \frac{P(CloudData \vee Rain) * P(Rain)}{P(CloudData)}$$

Finance: Investors use Bayes' Theorem to revise stock price predictions based on new information, such as earnings reports, helping them make more informed investment decisions.

$$P(StockRise \vee NewInfo) = \frac{P(NewInfo \vee StockRise) * P(StockRise)}{P(NewInfo)}$$

Legal Evidence: Bayes' Theorem is used in courts to update the likelihood of guilt as new evidence (like DNA) is presented, ensuring a more rational interpretation of multiple pieces of evidence.

$$P(Guilty \vee DNAMatch) = \frac{P(DNAMatch \vee Guilty) * P(Guilty)}{P(DNAMatch)}$$

Personalized Recommendations: Platforms like Netflix use Bayes' Theorem to update recommendations for users, predicting what they might like based on their past interactions with similar content.

$$P(LikesMovie \vee WatchedSimilarMovies) = \frac{P(WatchedSimilarMovies \vee LikesMovie) * P(Li}{P(WatchedSimilarMovies)}$$

Terms related to Bayes' Theorem

Hypotheses: Events happening in the sample space E1, E2,… En is called the hypotheses

Priori Probability: Priori Probability is the initial probability of an event occurring before any new data is taken into account. **P(Ei)** is the priori probability of hypothesis Ei.

Posterior Probability: Posterior Probability is the updated probability of an event after considering new information. Probability **P(Ei|A)** is considered as the posterior probability of hypothesis Ei.

Exercise

1.A disease affects 1% of the population. A test for this disease is 95% accurate, meaning the probability of testing positive given the person has the disease is 0.95. However, the test also

has a 5% false positive rate. If a patient tests positive, what is the probability that they actually have the disease?

2.Suppose 30% of emails are spam, and a particular word (e.g., "free") appears in 70% of spam emails but only 10% of non-spam emails. If an email contains the word "free," what is the probability that it is spam?

3.The probability of rain on a given day is 40%. If it rains, the probability that you will see clouds beforehand is 80%. If it doesn't rain, the probability of seeing clouds is 20%. If you see clouds in the sky, what is the probability it will rain?

4.In a chess tournament, there are three players: Alice, Bob, and Charlie. Each has an equal chance of reaching the final. The probabilities that Alice will defeat each opponent in the final are: Probability Alice defeats Bob: 0.5 Probability Alice defeats Charlie: 0.6 Alice wins the final. What is the probability that her opponent was Bob?

Answer

1. P(Disease) = 0.01

P(No Disease) = 0.99

P(Positive Test | Disease) = 0.95

P(Positive Test | No Disease) = 0.05

$$P(Disease \vee PositiveTest) = \frac{P(PositiveTest \vee Disease) * P(Disease)}{P(PositveTest)}$$

P(Positive Test) = P(Positive Test|Disease)*P(Disease) + P(Positive Test|No Disease)* P(No Disease)

=(0.95 * 0.01) + (0.05 * 0.99) = 0.0095 + 0.0495 = 0.059

$$P(Disease \vee PositiveTest) = \frac{0.95 * 0.01}{0.059}$$

= 0.161

So, the probability the patient has the disease given a positive test result is approximately **16.1%**

1. P(Spam) = 0.30

P(Not Spam) = 0.70

P(Free|Spam) =0.70

P(Free| Not Spam) = 0.10

$$P(Spam \vee Free) = \frac{P(Free \vee Spam) * P(Spam)}{P(Free)}$$

P(Free) = P(Free|Spam)*P(Spam) + P(Free/Not Spam)*P(Not Spam)

=(0.70 * 0.30) + (0.10 * 0.70) = 0.21 + 0.07 = 0.28

$$P(Spam \vee Free) = \frac{0.70 * 0.30}{0.28}$$

= 0.75

So, the probability that an email is spam given that it contains the word "free" is **75%.**

1. P(Rain) = 0.40

p(No Rain) = 0.60

P(Clouds|Rain) = 0.80

P(Clouds|No Rain) = 0.20

$$P(Rain \vee Clouds) = \frac{P(Clouds \vee Rain) * P(Rain)}{P(Clouds)}$$

P(Clouds) = P(Clouds|Rain)*P(Rain) + P(Clouds|No Rain)* P(No Rain)

=(0.80 * 0.40) + (0.20 * 0.60) = 0.44

$$P(Rain \vee Clouds) = \frac{0.80 * 0.40}{0.44}$$

=0.727

So, the probability of rain given that you see clouds is approximately **72.7%.**

1. Each player has equal chance to win, P(Bob) = P(Charlie) = 1/3

Probability that Alice defeats Bob: P(Win | Bob)=0.5

Probability that Alice defeats Charlie: P(Win | Charlie)=0.6

Alice has won the final, so we need P(Bob | Win).

$$P(Bob \vee Win) = \frac{P(Win \vee Bob) * P(Bob)}{P(Win)}$$

P(Win) = P(Win|Bob)*P(Bob) + P(Win|Charlie)*P(Charlie)

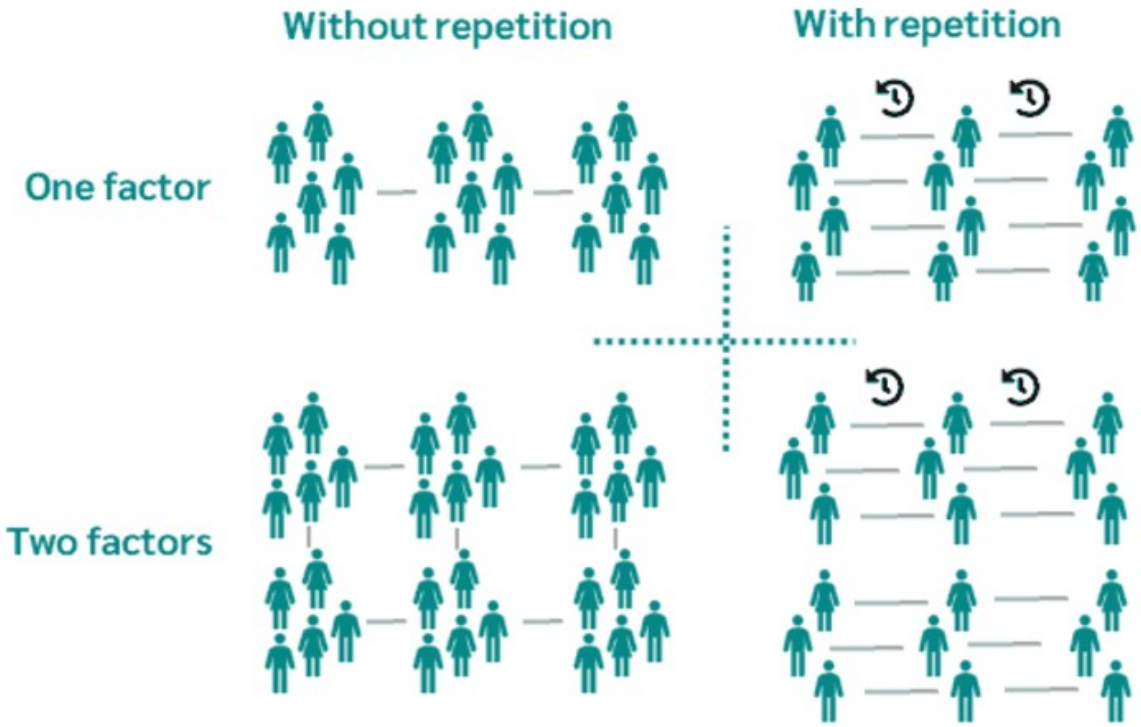
= (0.5 * 1/3) + (0.6 * 1/3) = (0.5 + 0.6)/3 = 0.3667

$$P(Bob \vee Win) = \frac{0.5 * 1/3}{0.3667}$$

= 0.5/1.1 = 0.4545

The probability that Alice's opponent in the final was Bob, given that she won, is approximately 0.4545 or **45.45%.**

Difference between conditional Probability and Bayes' Theorem



Probability Distribution

A distribution shows the possible values a variable can take and how frequently they occur in the sample space.

Y - The actual outcome of an event

y- one of the possible outcomes

P(Y = y) or P(y)

eg:

Y - The no. of red marbles we draw out of a bag.

y - A specific no. like 3 or 5

Then, we express the probability of getting exactly 5 red marbles as:

P(Y = 5) or P(5)

P(y) - expresses the probability for each distinct outcome, so, P(y) is the probability function.

Probability or probability distributions measure the likelihood of an outcome, depending on how often it features in the sample space

Probability Frequency Distribution(We did earlier)

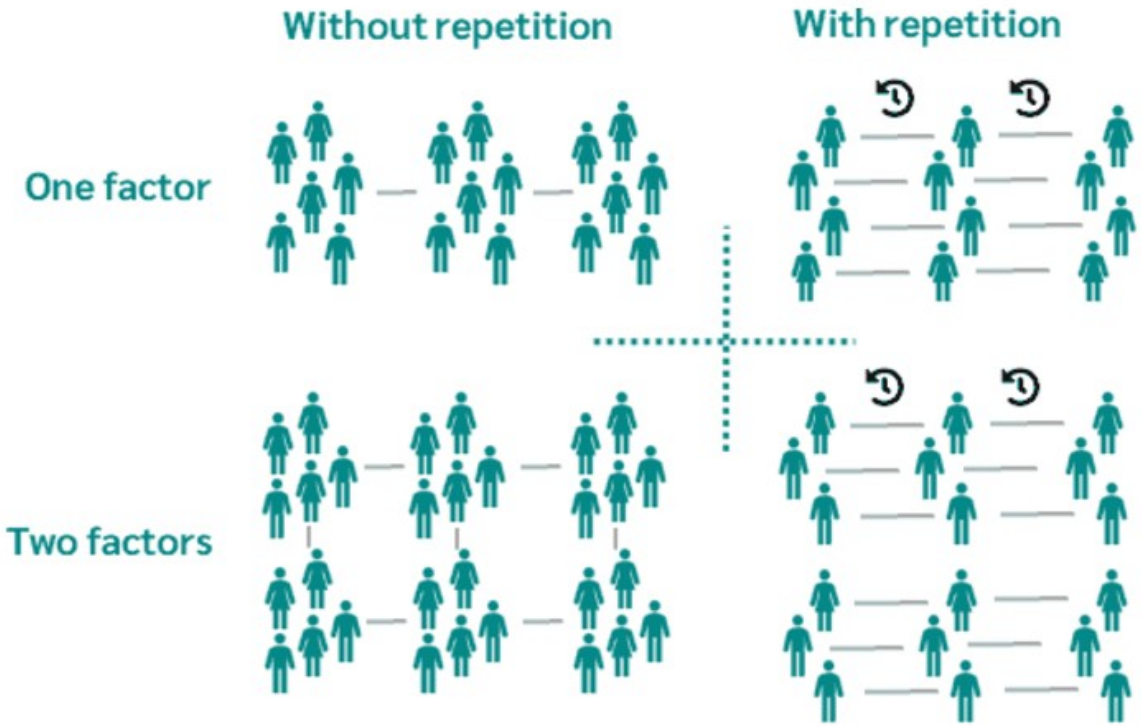
Recorded the frequency for each unique value and divide it by the total no. of elements in the sample space.

Usually, that is how we construct probabilities, when we have finite number of possible coutcomes. If we had an infinite number of possibilities, then recording the frequency for each one becomes impossible, because there are infinitely many of them.

Regardless of finite or infinite number of possibilities, we define distributions using only 2 characteristics.

1. mean - the average value
2. variance - how spread out the data is

Difference between Probability and Frequency Distribution



Random Variables

It is the numerical value determined by the outcome of a random experiment. It is denoted by X.

Types of Random Variables:

1. Discrete

2. Continuous

Discrete Random Variable: It takes on a countable(finite) number of distinct values. For example, the number of heads when flipping three coins could be 0, 1, 2, or 3. This is an example of a discrete random variable because it can only take specific, separate values. The probability function associated with it is, Probability Mass Function(PMF)

Continuous Random Variable: It takes on an infinite number of possible values within a given range. For example, the time it takes for a car to complete a lap in a race could be 12.5 seconds, 12.55 seconds, 12.555 seconds, etc. This is a continuous random variable because it can take any value within a continuous range. The probability function associated with it is, Probability Density Function(PDF)

Types of Distributions

Probability Distributions

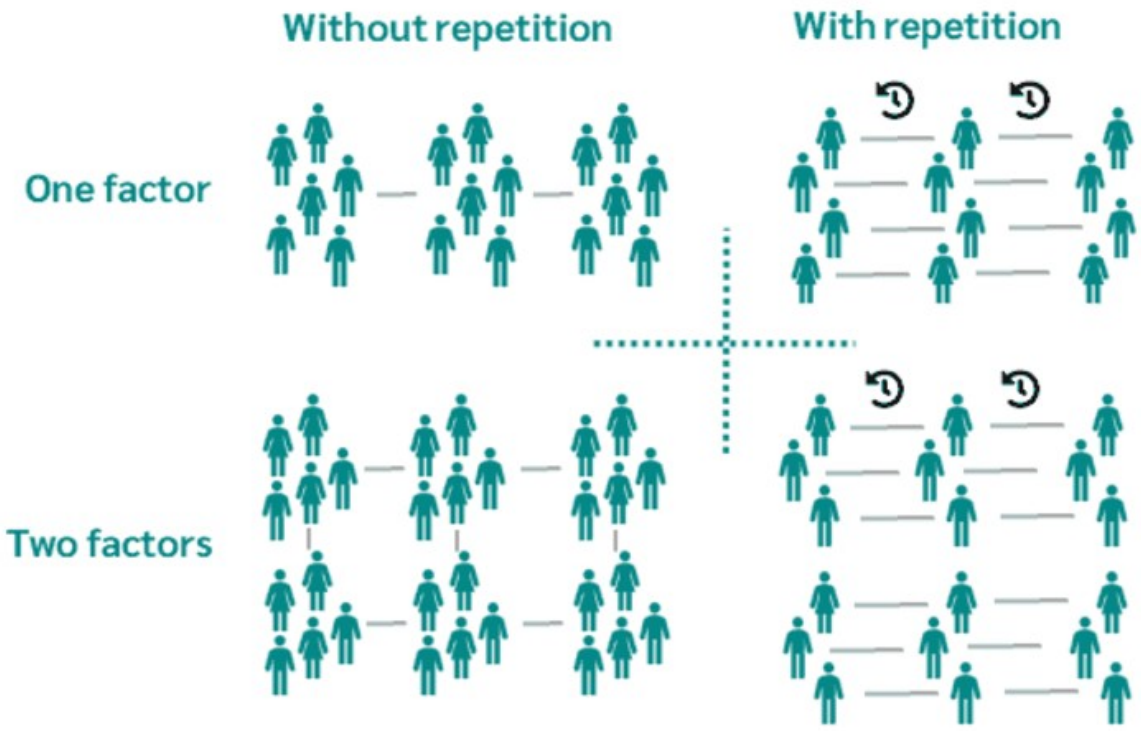
```
|
|-- Discrete Distributions
|    |-- Bernoulli
|    |-- Binomial
|    |-- Poisson
|    |-- Uniform
|
|-- Continuous Distributions
|    |-- Chi-Squared
|    |-- Exponential
|    |-- Student's T
|    |-- Normal
|    |-- Logistic
```

Discrete Distributions

A discrete distribution deals with random variables that can take on a countable number of values. These values are often integers, such as the number of heads in a coin toss or the number of cars passing a certain point in an hour.

Continuous Distributions

A continuous distribution deals with random variables that can take on any value within a specific range. Examples include height, weight, or time.



Probability Mass Function(PMF)

The Probability Mass Function (PMF) gives us the probability of each individual value for a discrete variable. In simple terms, the PMF tells us, “What is the probability that X equals a specific value?”

The notation often used is **P(X=x)**, meaning "the probability that X equals x."

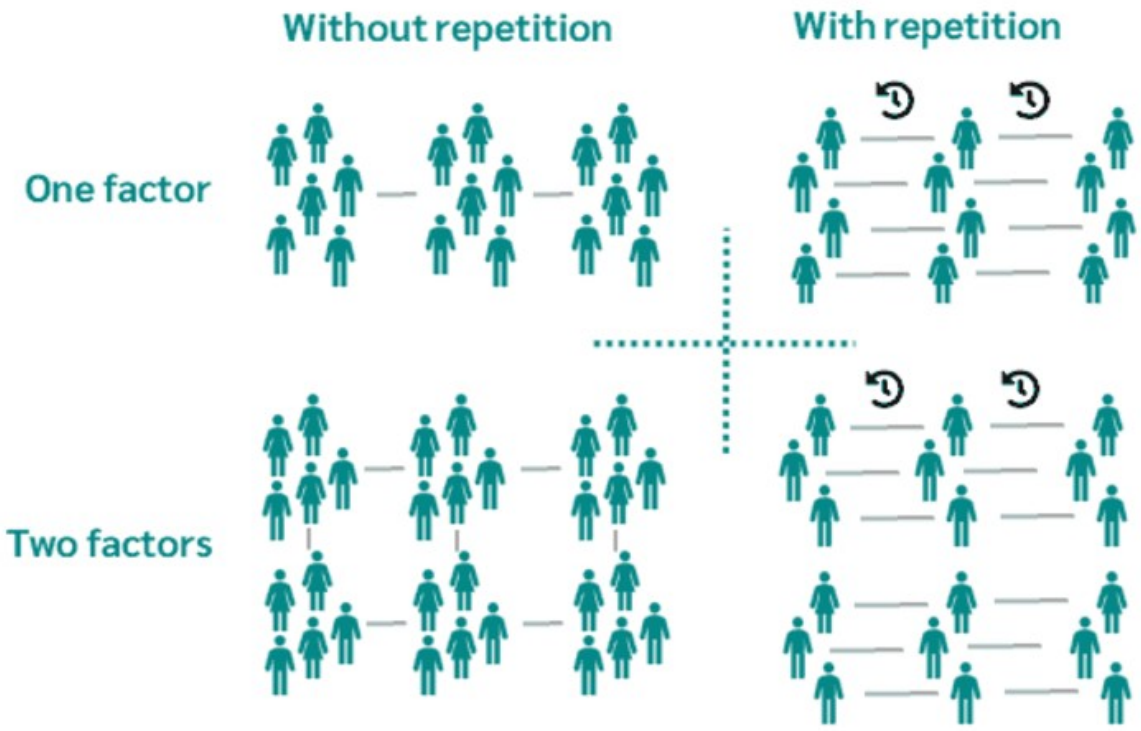
Example Using a Coin Toss: Consider tossing a fair coin. The possible outcomes are:

X = 0 (for "Tails")

X = 1 (for "Heads")

Since the coin is fair, the probability of each outcome is 0.5.

The PMF tells us that P(X=0)=0.5 and P(X=1)=0.5.



Cumulative Distribution Function (CDF) for a Discrete Variable

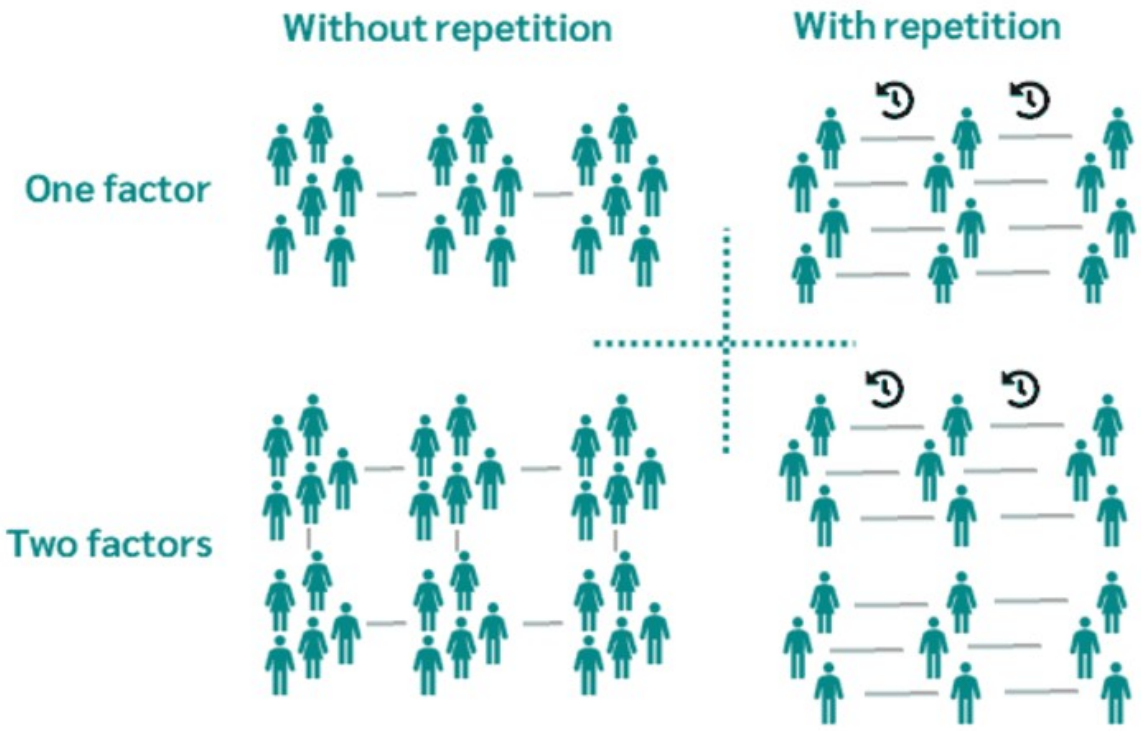
The Cumulative Distribution Function (CDF) for a discrete variable shows the probability that the variable will take a value less than or equal to a certain number. In other words, the CDF tells us the probability of getting a value up to and including a certain point.

The CDF accumulates probabilities as you go from the lowest value to the highest.

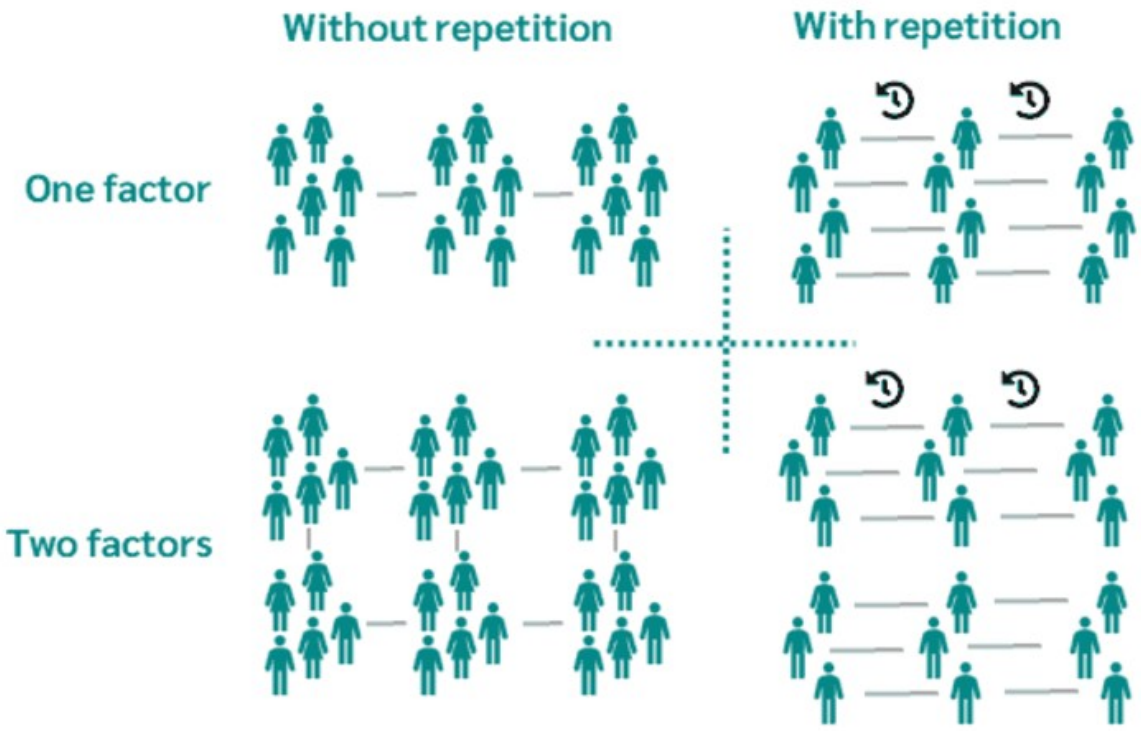
For each value x, the CDF shows **P(X≤x)**, meaning "the probability that X is less than or equal to x."

The CDF is always increasing or staying the same as you move to larger values.

Example Using a Die Roll: Let's say we want to find the CDF of rolling a fair six-sided die. The possible outcomes are 1, 2, 3, 4, 5, and 6, each with a probability of 1/6.



Think of the CDF as a "running total." As you go from one value to the next, you keep adding up the probabilities. By the time you reach the last value, your total probability should be 1 (or 100%).

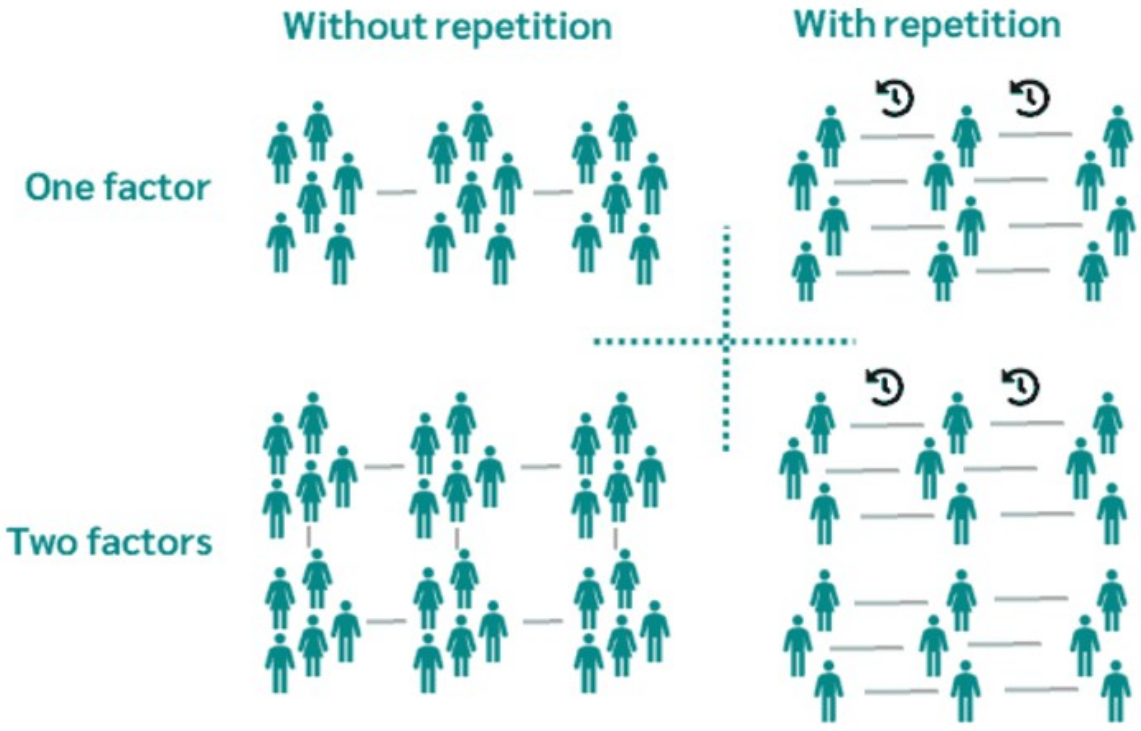


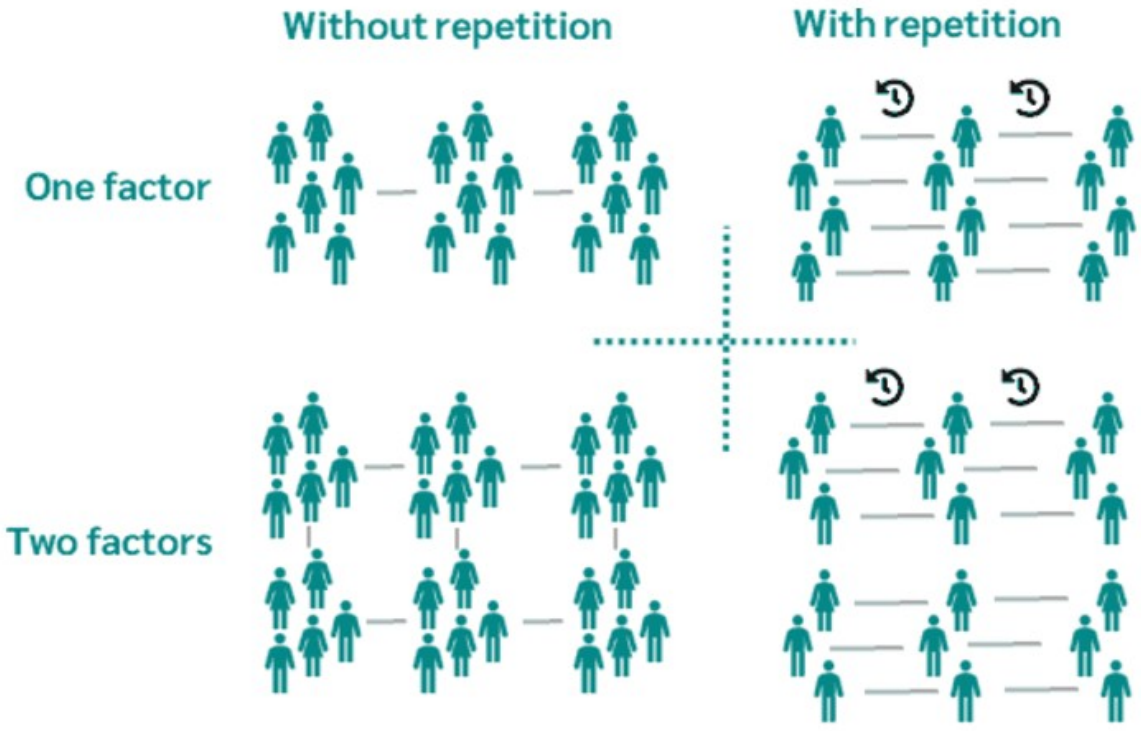
If we visualize the concepts:

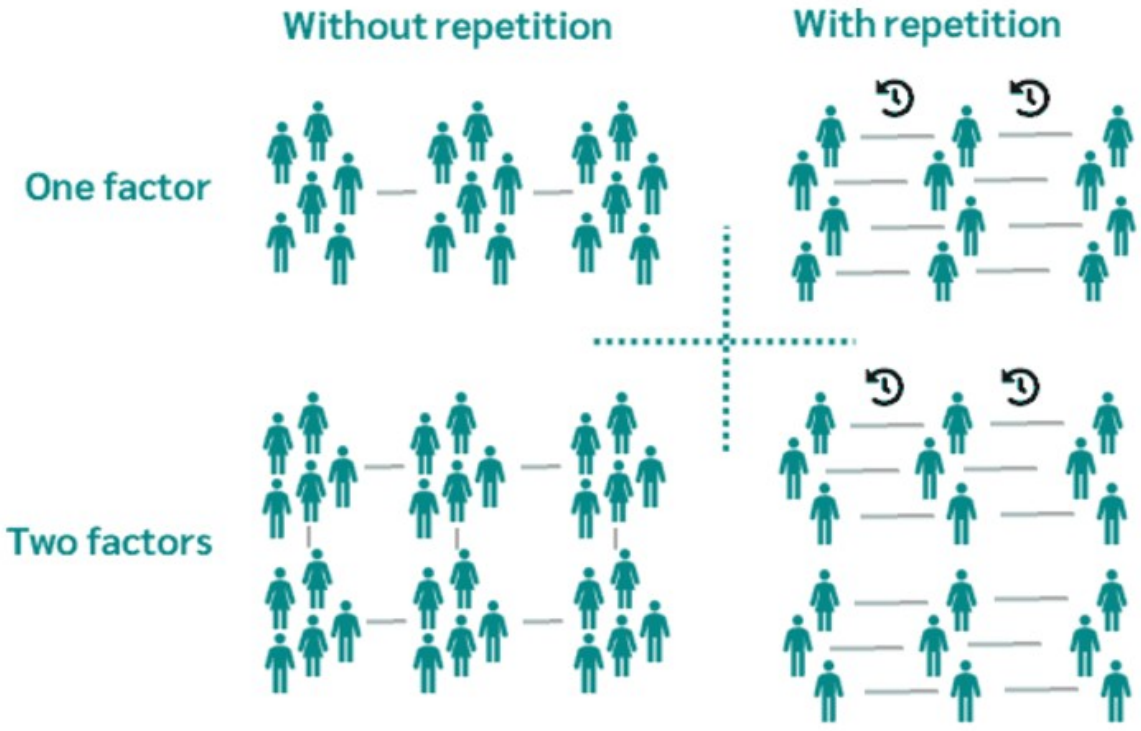
Discrete Probability Distribution is like a table listing each possible value and its probability.

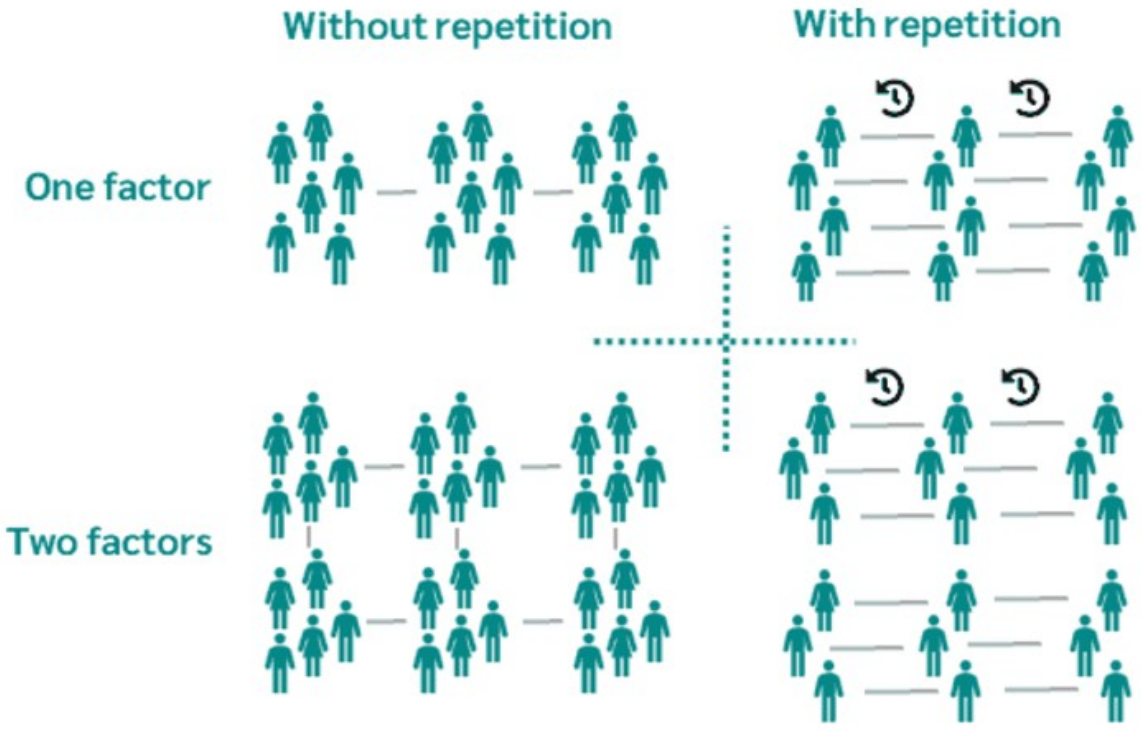
PMF is like looking at each individual value in that table and asking, "What's the probability of this exact value?"

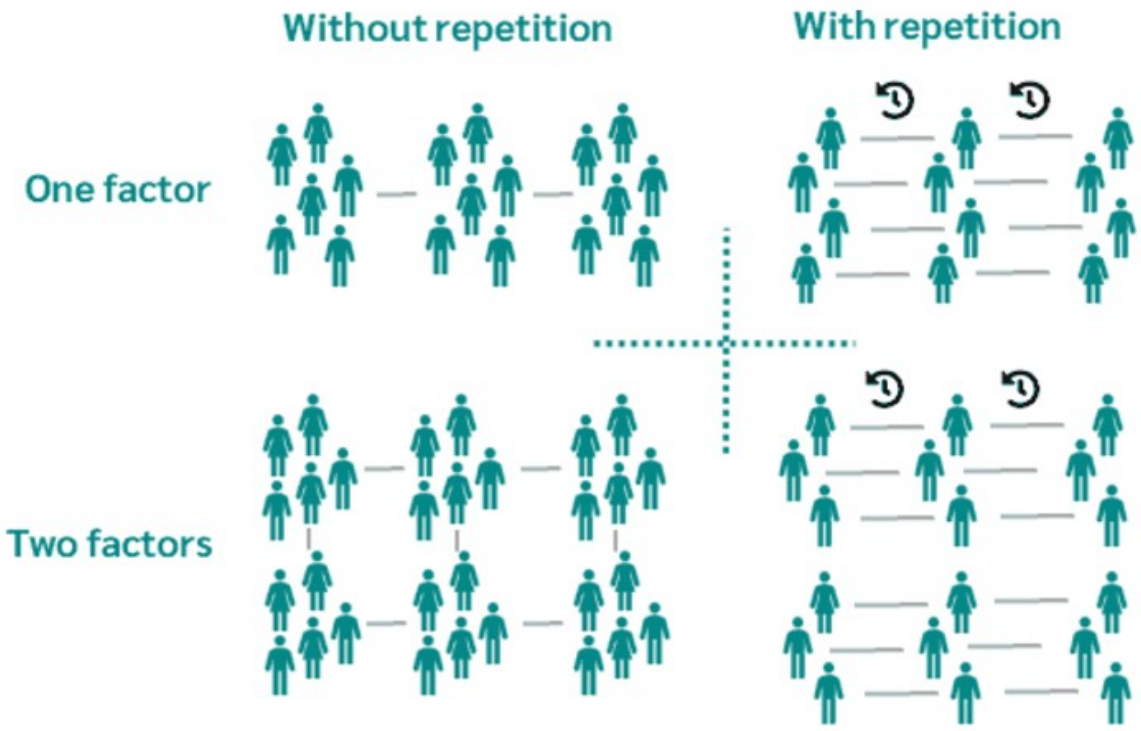
CDF is like a running total in that table, adding up probabilities as you move to larger values.


Without repetition
With repetition
One factor
Two factors


Without repetition
With repetition
One factor
Two factors


Without repetition
With repetition
One factor
Two factors


Without repetition
With repetition
One factor
Two factors



Exercise

1. Roll a fair die twice. Let X be the random variable that gives the maximum of the two numbers. Find the probability mass function describing the distribution of X.

2.An urn contains five green balls, two blue balls, and three red balls. You remove three balls at random without replacement. Let X denote the number of red balls. Find the probability mass function describing the distribution of X.

3. Consider rolling a fair six-sided die. We want to calculate the PMF for the random variable X, which represents the number of dots on the upper face of the die.

Answer

1.Sample space = 36

For each outcome (i,j), we determine X=max(i,j). The possible values of X are 1 through 6

We will calculate how many times each value of X can occur:

X=1: The only combination is (1,1).

Count: 1

X=2: The combinations are (1,2),(2,1),(2,2).

Count: 3

X=3: The combinations are (1,3),(2,3),(3,1),(3,2),(3,3).

Count: 5

X=4: The combinations are (1,4),(2,4),(3,4),(4,1),(4,2),(4,3),(4,4).

Count: 7

X=5: The combinations are (1,5),(2,5),(3,5),(4,5),(5,1),(5,2),(5,3),(5,4),(5,5).

Count: 9

X=6: The combinations are (1,6),(2,6),(3,6),(4,6),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6).

Count: 11

The total number of outcomes is 36. We can now find the PMF for X:

P(X=1)= 1/36

P(X=2) = 3/36

P(X=3) = 5/36

P(X=4) = 7/36

P(X=5) = 9/36

P(X=6) = 11/36

Sum of PMF is 1.

2. Total balls = 10

Red balls = 3

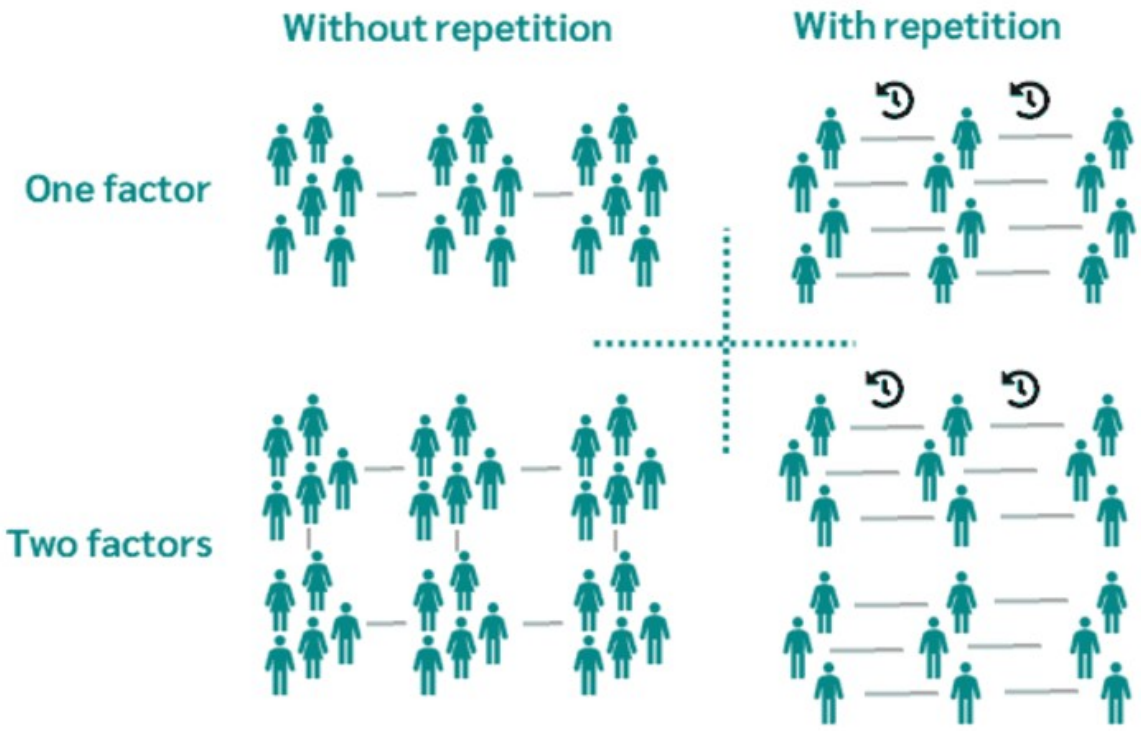
Non-red balls = 7

X = 0,1,2,3

Total ways =

$$C_3^{10} = \frac{V_3^{10}}{P_3} = \binom{10}{3} = \frac{10!}{3! \times (10-3)!}$$

=120



3. Sample space = 6

The probability of rolling any specific number (1, 2, 3, 4, 5, or 6) is 1/6

P(X=x) = 1/6

Sum of PMF is 1

Bernoulli Distribution

It is a probability distribution that describes the outcome of a single binary experiment (an experiment with only two possible outcomes, typically labeled as "success" and "failure"). It's useful in situations where there's a clear yes/no, 0/1, or true/false result.

A random variable X follows a Bernoulli distribution if it takes a value of 1 (for success) with probability p, and 0 (for failure) with probability 1−p.

The probability mass function of the Bernoulli distribution is given by:

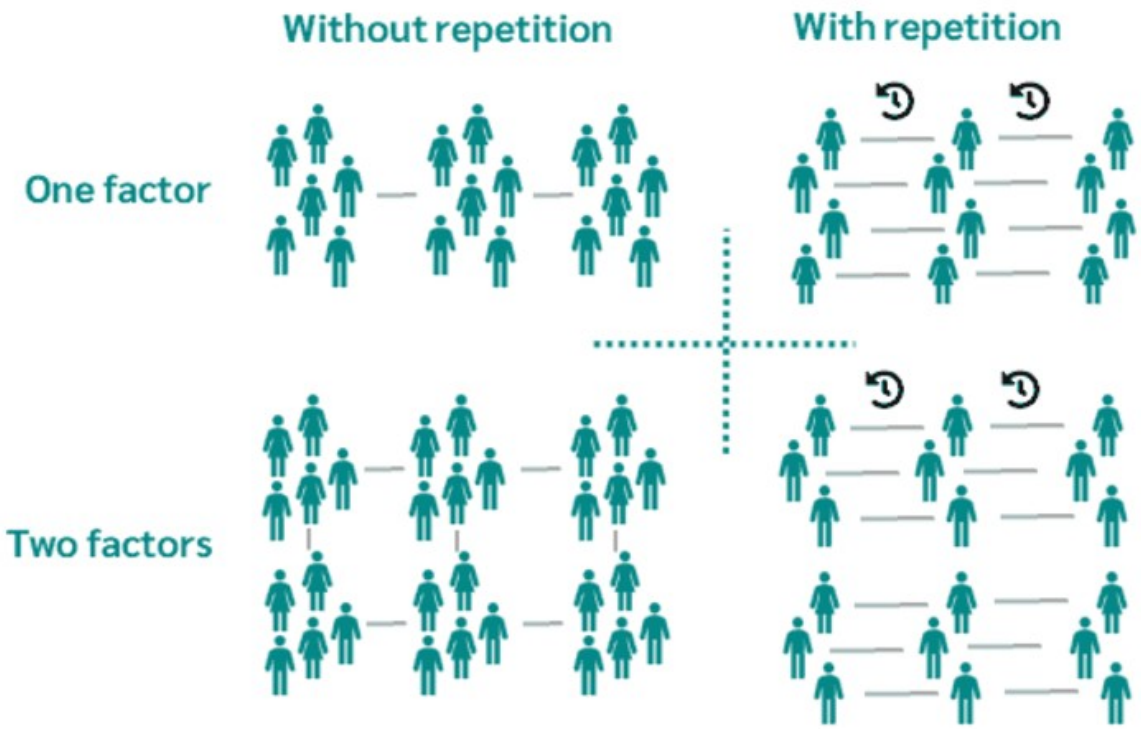
$$P(X=x)=p^{x}(1-p)^{1-x}$$

where:

x is the outcome, which can be 0 or 1,

p is the probability of success (X=1),

1−p is the probability of failure (X=0).



Properties:

Mean (Expected Value): E(X)=p

$$\left(E\left(X\right)=1\,.\,P+0\,.\left(1-p\right)=p\right)$$

Variance: Var(X)=p(1−p)

$$\left(\sum i\left(xi-\mu\right)2\,p\left(xi\right)=\left(0-p\right)^{2}.\left(1-p\right)+\left(1-p\right)^{2}.\,p=p\left(1-p\right)\right)$$

Example: Consider a scenario where a coin is tossed once, and we define:

"Heads" as a success (X=1),

"Tails" as a failure (X=0).

If the coin is fair, the probability of getting heads (p) is 0.5, and the probability of tails (1−p) is also 0.5.

For this Bernoulli random variable X:

P(X=1)=0.5

P(X=0)=0.5

This simple binary scenario represents a Bernoulli distribution where each outcome is either a success or a failure.

Exercise

1. If a light bulb has a 40% chance of lasting more than 1000 hours, and we assign the outcome of the light bulb lasting more than 1000 hours as the 'success' outcome with a value of 1, find the expected value and variance.

2. A light bulb has a 70% chance of working correctly when switched on. Let X be the random variable representing the outcome (1 = working, 0 = not working).

Find the PMF of X. Calculate E(X), the expected value of X.

3. In a survey, 45% of people say they enjoy reading books. Let X represent the reading preference of a randomly selected person.

Find P(X=1) (enjoys reading).

Find P(X=0) (doesn't enjoy reading). Calculate E(X).

4. A new medication is effective 90% of the time. Let X represent the outcome of the medication (1 = effective, 0 = ineffective).

Find P(X=1) and P(X=0). Calculate Var(X).

5. A student guesses the answer to a true-or-false question. The probability of guessing correctly is 50%. Define X as the outcome of the guess (1 = correct, 0 = incorrect).

Find P(X=1) and P(X=0). Determine the standard deviation of X.

Answer

1. E(X) = p = 0.40 , variance = 0.4(1-0.4) = 0.24

2. PMF(X) = P(X=1) = 0.7 , P(X=0) = 1-0.7 = 0.3 , E(X) = 0.7

3. P(X=1) = 0.45 , p(X=0) = 1-0.45 = 0.55, E(X) = 0.45

4. P(X=1) = 0.9, p(X=0) = 1-0.9 = 0.1, variance = p(1-p) = 0.9(1-0.9) = 0.9*0.1 = 0.09

5. P9X=1) = 0.5, p(X=0) = 0.5 , std = sqrt(p(1-p)) = sqrt(0.5*0.5) = 0.5

Binomial Distribution

It is a discrete probability distribution that models the number of successes in a fixed number of independent and identical trials, where each trial has two possible outcomes: success or failure. It is commonly used to represent situations where an event occurs a certain number of times out of a fixed number of trials, with a constant probability of success.

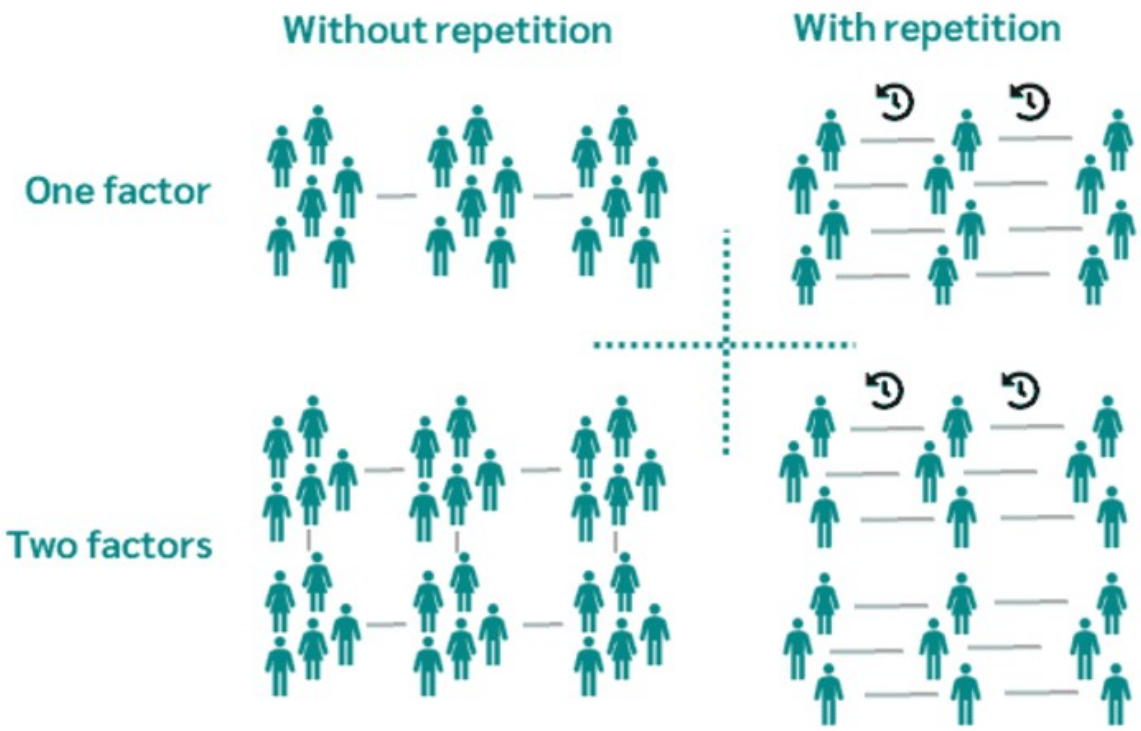
A random variable X follows a binomial distribution if it counts the number of successes in n independent trials, where:

Each trial has two outcomes: success or failure.

The probability of success in each trial is constant, denoted by p.

The probability of failure in each trial is 1−p.

The probability mass function of the Binomial distribution is given by:



where:

x = number of successess

n = total number of trials

p = probability of success in a single trial

nCx = the binomial coefficient repressenting the number of ways to choose x successes out of n trials

Properties:

Mean (Expected Value): E(X)=p.n

Variance: Var(X)=n.p(1−p)

Example:

Suppose a factory produces light bulbs, and each bulb has a 5% defect rate (probability of a defect, p=0.05). If we select 10 light bulbs at random, what is the probability that exactly 2 of them are defective?

Given:

n=10 (total bulbs),

x=2 (defective bulbs),

p=0.05 (probability of defect).

Using the formula:

P(X=2) = 10C2 *(0.05)^2* (1-0.05)^10-2 = 45 * 0.0025 * (0.95)^8 = 45 * 0.0025 * 0.6634 = 0.0746

Exercise

1.In a certain town, 60% of the population prefers tea over coffee. If 10 people are randomly surveyed, find the probability that exactly 7 of them prefer tea.

2.A fair die is rolled 8 times. Find the probability of rolling exactly three 4s.

3.A salesperson has a 30% chance of making a sale with each customer. If they approach 15 customers, find the probability that they will make exactly 5 sales.

4.A biologist observes that a type of flower has a 20% chance of being red. In a random sample of 6 flowers, what is the probability of finding exactly 2 red flowers?

5.In a quality control test, a machine produces 15 items, and there is a 12% chance that any given item is defective. Find the probability of finding exactly 3 defective items.

Answers

1. n=10, p=0.6, x=7 , P(X=7) = 10C7*(0.6)^7*(1-0.6)^3 =120×0.02799×0.064 ≈0.2150

2. n = 8, p= 1/6 , x=3, P(X=3) = 0.1043

3. n= 15, p =0.3, x = 5, P(X=5) = 0.2054

4. n = 6, p = 0.2, x = 2, P(X=2) = 0.2458

5. n = 15, p = 0.12 , x = 3 , P(X=3) = 0.1995

Poisson Distribution

It is a discrete probability distribution that models the number of events occurring within a fixed interval of time or space, given that these events occur with a known constant mean rate and independently of the time since the last event. It's commonly used for counting the number of occurrences of an event within specific intervals, like the number of phone calls at a call center in an hour, or the number of cars passing through a toll booth.

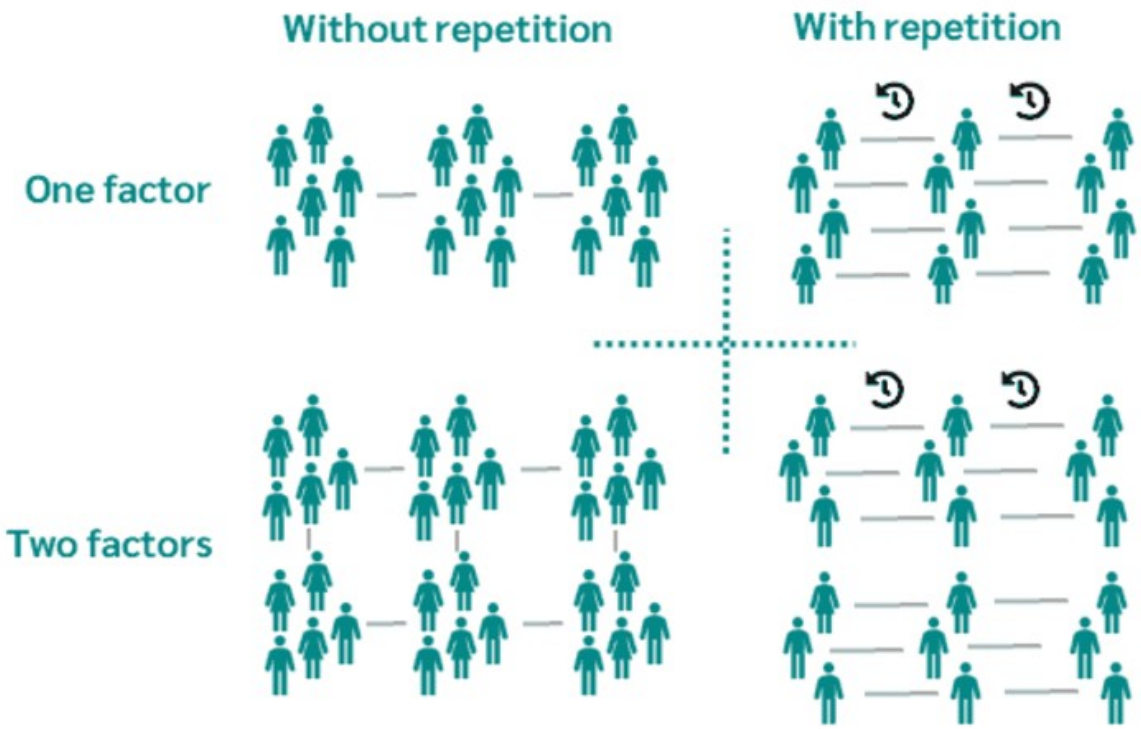
A random variable X follows a Poisson distribution if it represents the number of times an event occurs in a fixed interval, with:

The average rate (mean) of occurrences, denoted by λ (lambda).

Events occur independently.

The Poisson distribution is often used in situations where n is large and p is small, such as counting rare events.

The probability mass function of Poisson distribution is given by:



where:

λ = the mean (average) number of occurrences in the interval,

x = the actual number of occurrences,

e ≈ 2.71828, the base of the natural logarithm.

Example Suppose a bookstore sells an average of 3 books per hour (λ=3). What is the probability that exactly 5 books will be sold in the next hour?

Given:

λ=3, x=5. Using the formula:

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Thus, the probability that exactly 5 books are sold in the next hour is approximately 0.1008, or 10.08%.

Exercise

1.A library sees an average of 5 visitors every 30 minutes. What is the probability that exactly 3 visitors will enter the library in the next 30 minutes?

2.A botanist observes an average of 7 rare flowers blooming per month. Find the probability that exactly 4 rare flowers will bloom next month.

3.A traffic light on a busy road is crossed by an average of 9 cars per minute. What is the probability that exactly 12 cars will cross the light in the next minute?

4.A web server receives an average of 4 requests per second. Find the probability that it will receive exactly 3 requests in the next second.

5.A bookstore sells an average of 2 books per hour. What is the probability that no books will be sold in the next hour?

Answer

1. $\lambda=5$, $x=3$, $P(X=3) = 0.1403$

2. $\lambda=7$, $x=4$, $P(X=4) = 0.091$

3. $\lambda=9$, $x=12$, $P(X=12) = 0.0725$

4. $\lambda=4$, $x=3$, $P(X=3) = 0.1952$

5. $\lambda=2$, $x=0$, $P(X=0) = 0.1353$

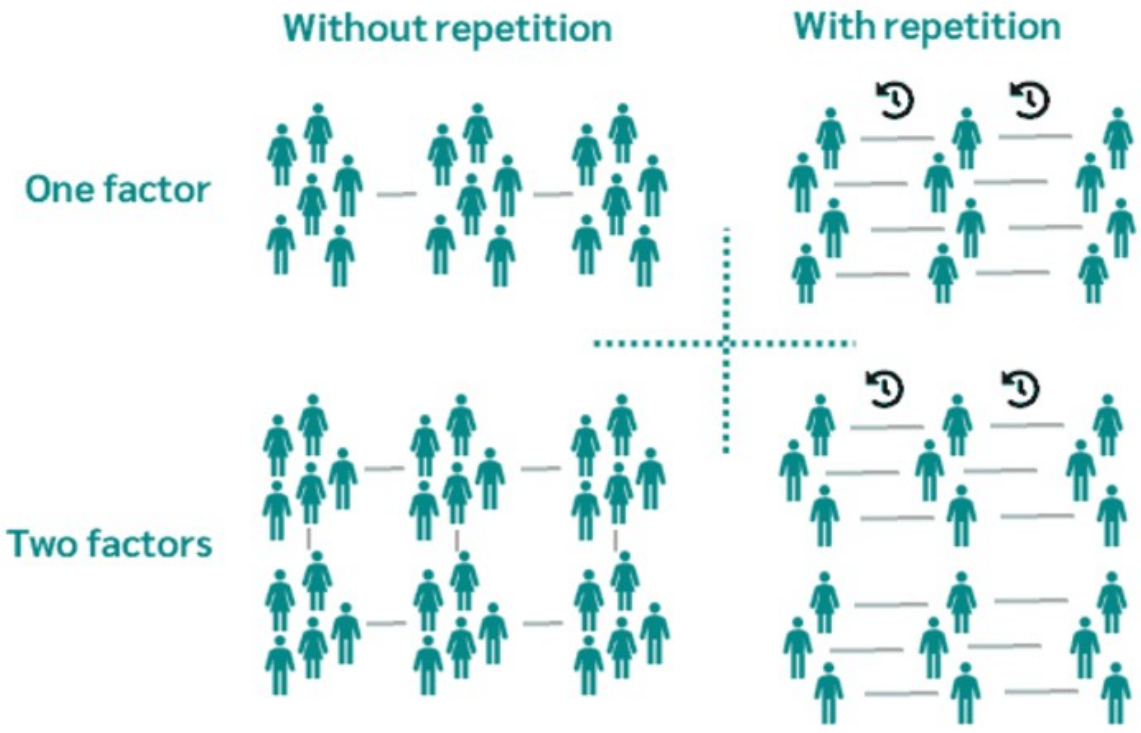
Discrete Uniform Distribution

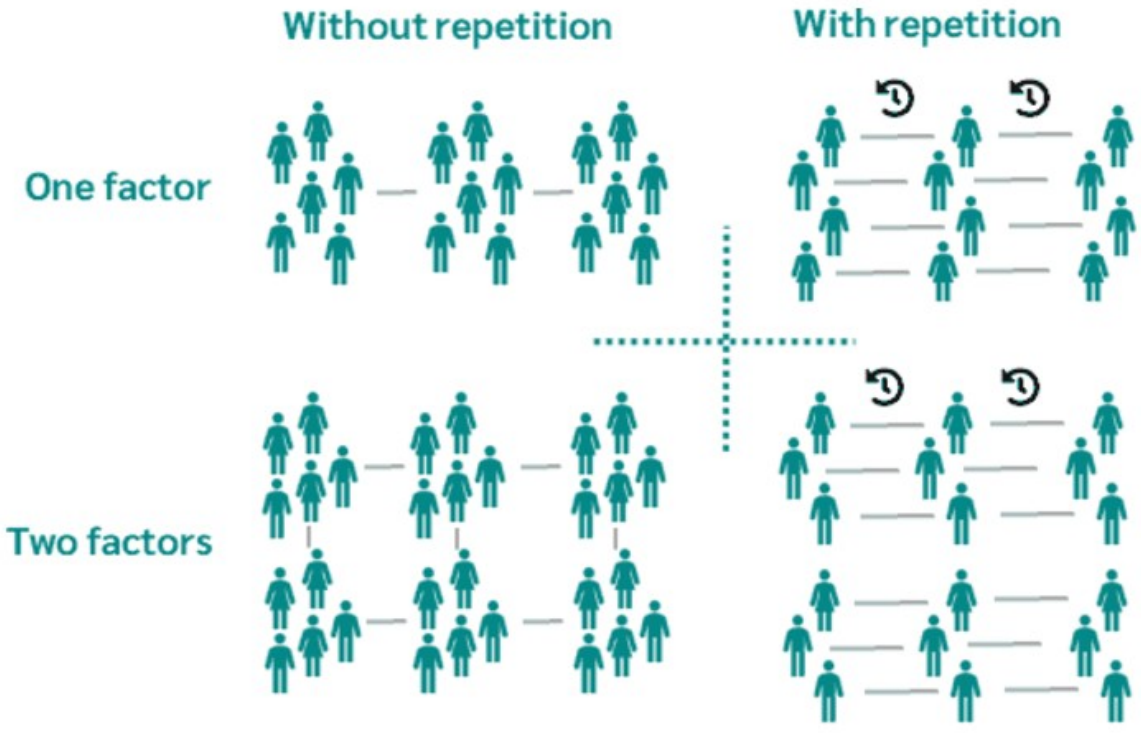
It is a probability distribution in which a finite number of outcomes are equally likely. Each outcome has the same probability of occurring. The distribution is called "uniform" because each value within the given range is equally probable.

Only discrete outcomes are considered (e.g., whole numbers).

All outcomes are equally likely.

Commonly used when each outcome in a set has an equal chance, like rolling a fair die or choosing a random card from a deck (ignoring suits or ranks).





Example:

Rolling a Fair Die

Outcomes: {1,2,3,4,5,6}

Since each face is equally likely, P(X=x)= 1/6 for each x.

Exercise

1. A fair twelve-sided die is rolled. Find the probability of rolling a 3.

2. A random integer between 1 and 15 is generated. Find the probability of getting a 10.

3. A standard deck of 52 cards is shuffled, and one card is drawn. Find the probability of getting a King.

4. A random number between 1 and 30 is generated. Find the probability of getting a multiple of 5.

5. A fair ten-sided die is rolled. Find the probability of getting a number greater than 7.

Answer

1. P(X=3) = 1/12

2. P(X=10) = 1/15

3. P(X=King) = 4/52

4. P(X=multiple of 5) = 6/30

5. P(X>7) = 3/10

Probability Density Function (PDF)

The Probability Density Function (PDF) is a function that represents the probability of a continuous random variable falling within a particular range of values. While the PDF itself does not give the probability of any exact value (since that probability is essentially zero), it helps us calculate the probability for a range of values.

PDF is only for continuous variables.

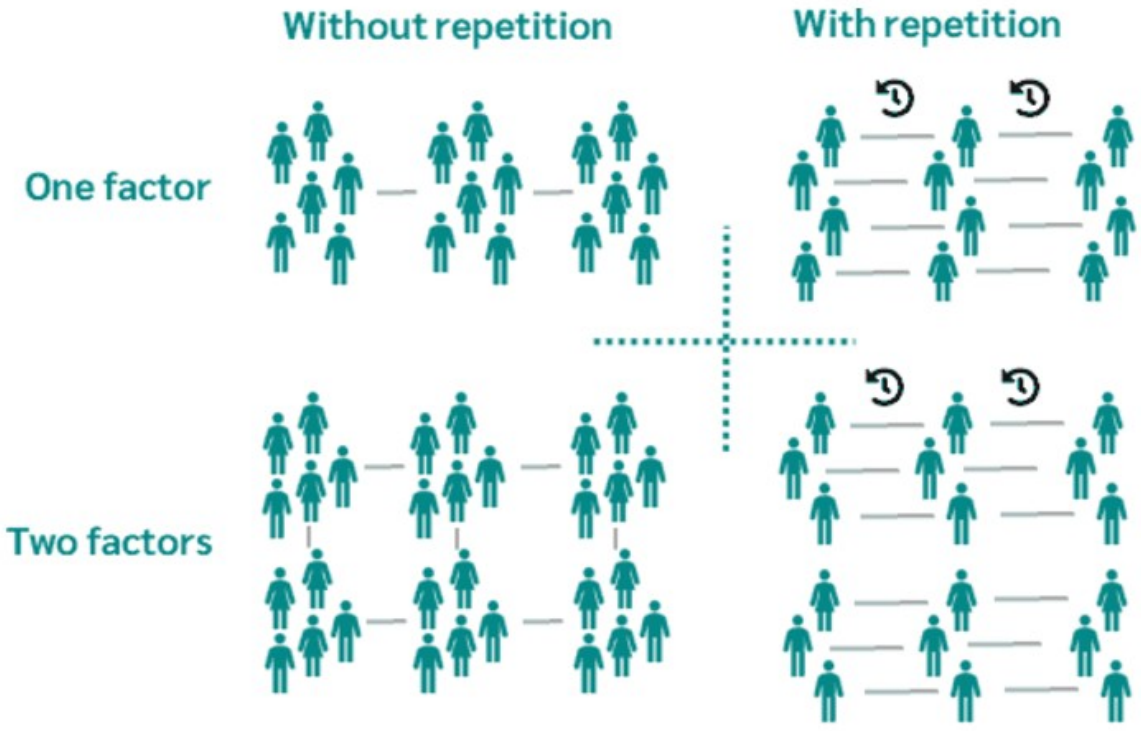
The probability of a specific, exact value in a continuous distribution is zero. We only calculate probabilities over a range of values.

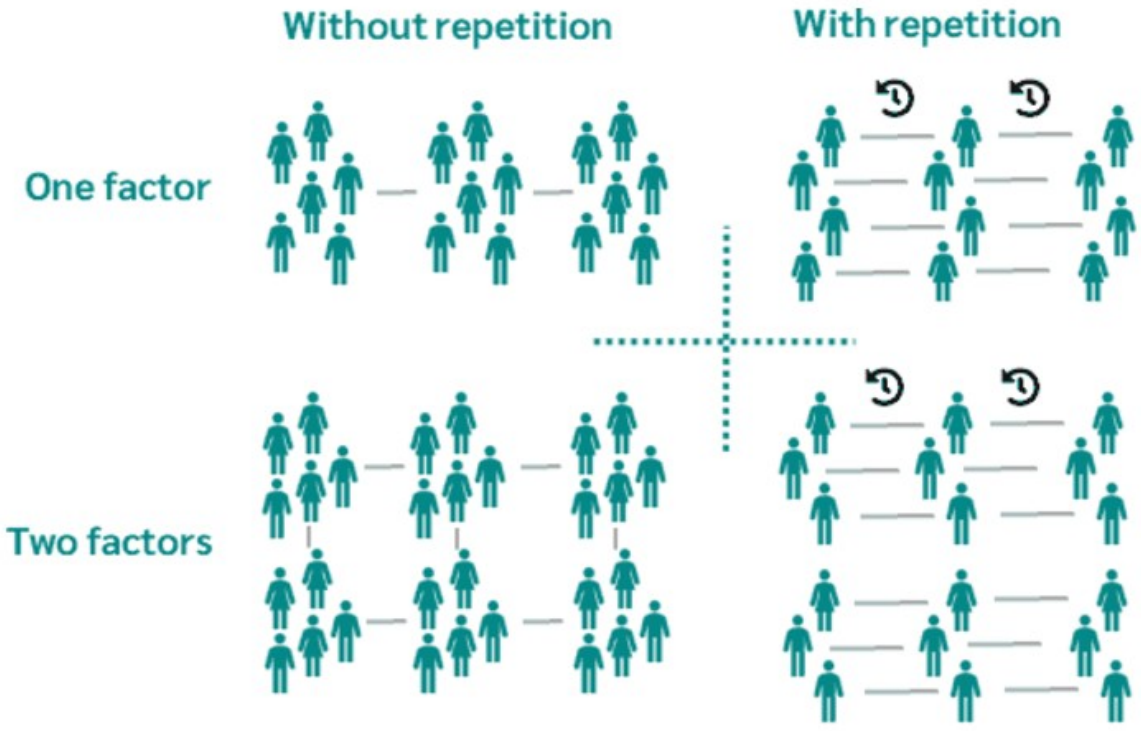
The area under the PDF curve between two points gives the probability that the variable falls within that range.

The notation often used is **f(x)**, representing the height of the curve at point x.

Eg: Imagine the heights of people in a population are normally distributed. The PDF curve (often a bell curve) shows the probability density of heights.

To find the probability that a person's height is between, say, 160 cm and 170 cm, you would find the area under the curve from 160 to 170. The area corresponds to the probability of a person's height being in that range.


Without repetition
With repetition
One factor
Two factors


Without repetition
With repetition
One factor
Two factors

Cumulative Distribution Function (CDF) for a Continuous Variable

The Cumulative Distribution Function (CDF) for a continuous variable gives the probability that the variable will take a value less than or equal to a certain number. For continuous variables, the CDF is a smooth, steadily increasing curve.

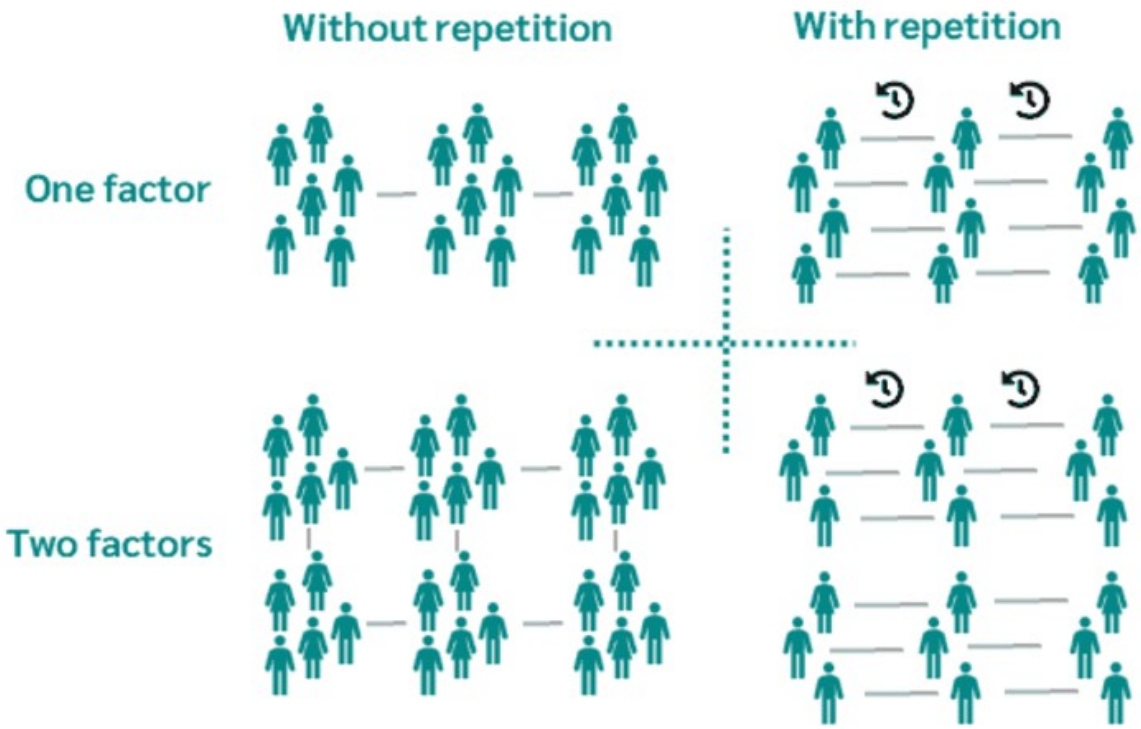
The CDF of a continuous variable shows the probability that the variable is less than or equal to a given value x.

The CDF for continuous variables is typically written as **$F(x)=P(X\leq x)$**.

The CDF always increases as you move along the range of values (it's a non-decreasing function).

When you reach the maximum value in the range, the CDF equals 1, meaning 100% probability.

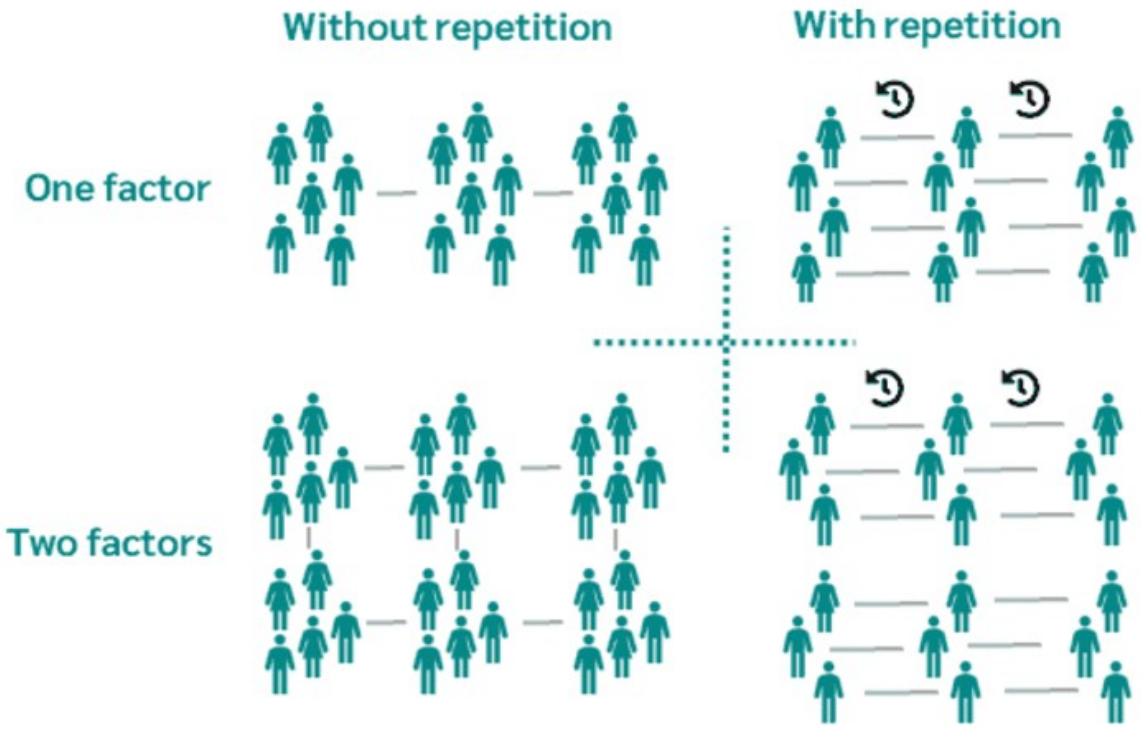
The formula for CDF is:

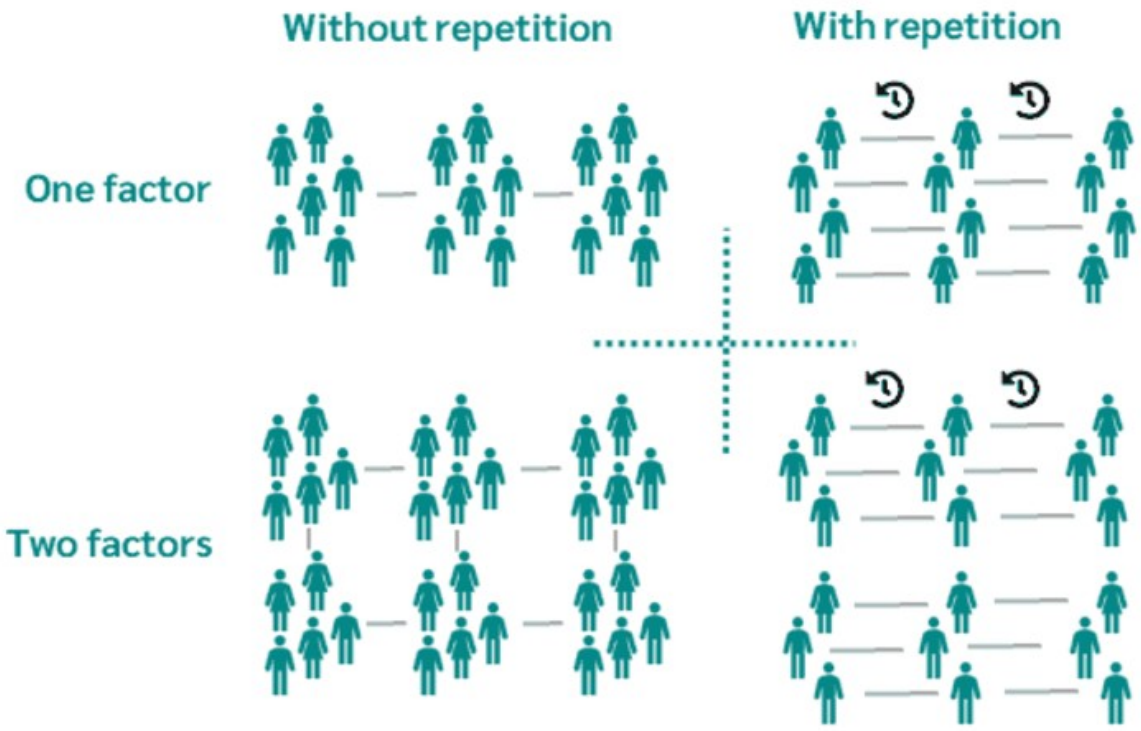


Example: Continuing with the height example, suppose you want to know the probability that a randomly selected person's height is less than or equal to 165 cm. The CDF at x=165 gives you this probability.

If F(165)=0.7, it means there's a 70% chance a randomly chosen person's height will be 165 cm or shorter.

The CDF is like a running total of probabilities. As you go from lower to higher values, the CDF shows the cumulative probability up to each point. By the time you reach the top end of the range, the CDF will reach 1, representing 100% probability.





Imagine the concepts like this:

Continuous Probability Distribution: Think of it as a smooth curve where each point represents a potential value within a range.

PDF: This is the shape of that curve and shows where the values are more "densely packed" along the range.

CDF: This is like climbing a hill along the curve. As you move up the curve, you accumulate probability, and by the time you're at the top, you've reached 100% probability.

Example to Tie It All Together

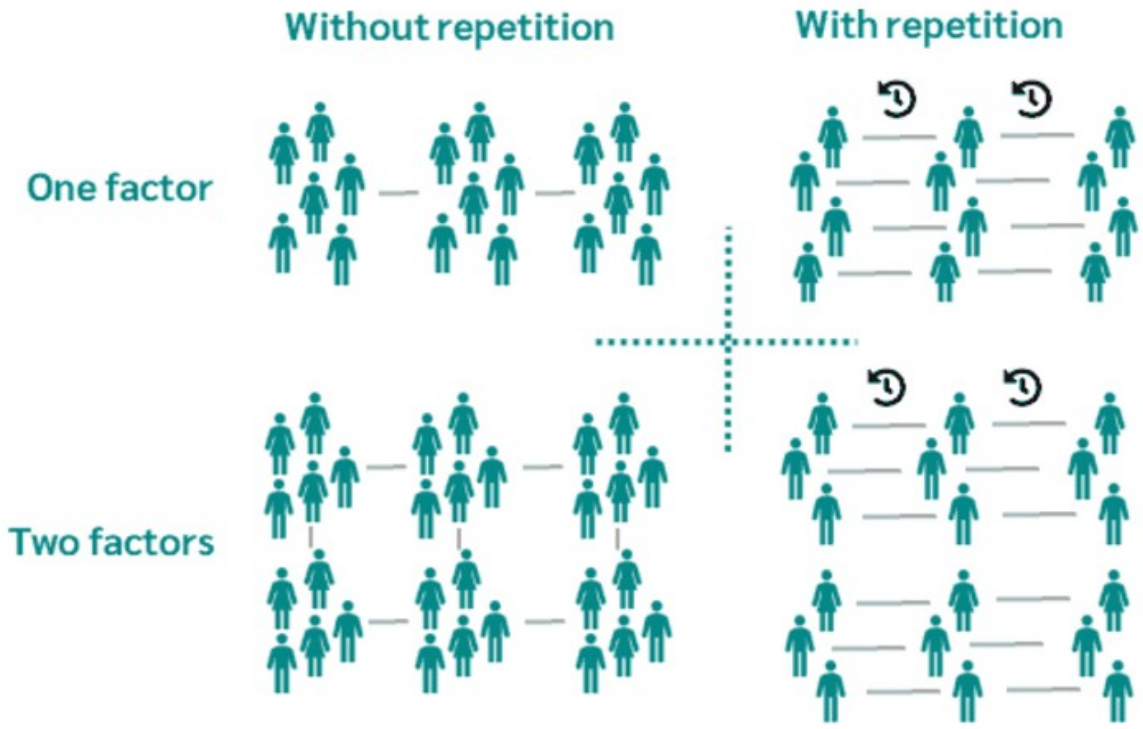
Suppose you're looking at the time it takes students to complete a test. The times might follow a normal distribution with a PDF that looks like a bell curve centered around the average time.

The PDF could show that times around the average (say, 60 minutes) are most likely, with fewer students finishing much earlier or later.

If you wanted to know the probability that a student finishes in less than 50 minutes, you'd use the CDF at 50 minutes, F(50), to get the probability of finishing in that time or less.

Continuous Uniform Distribution

A continuous uniform distribution is a type of probability distribution in which all outcomes are equally likely within a specific range or interval. In other words, the probability density function (PDF) is constant over the interval [a,b], meaning that any value within this interval has the same likelihood of occurring. The distribution is defined by two parameters, a (the minimum value) and b (the maximum value).



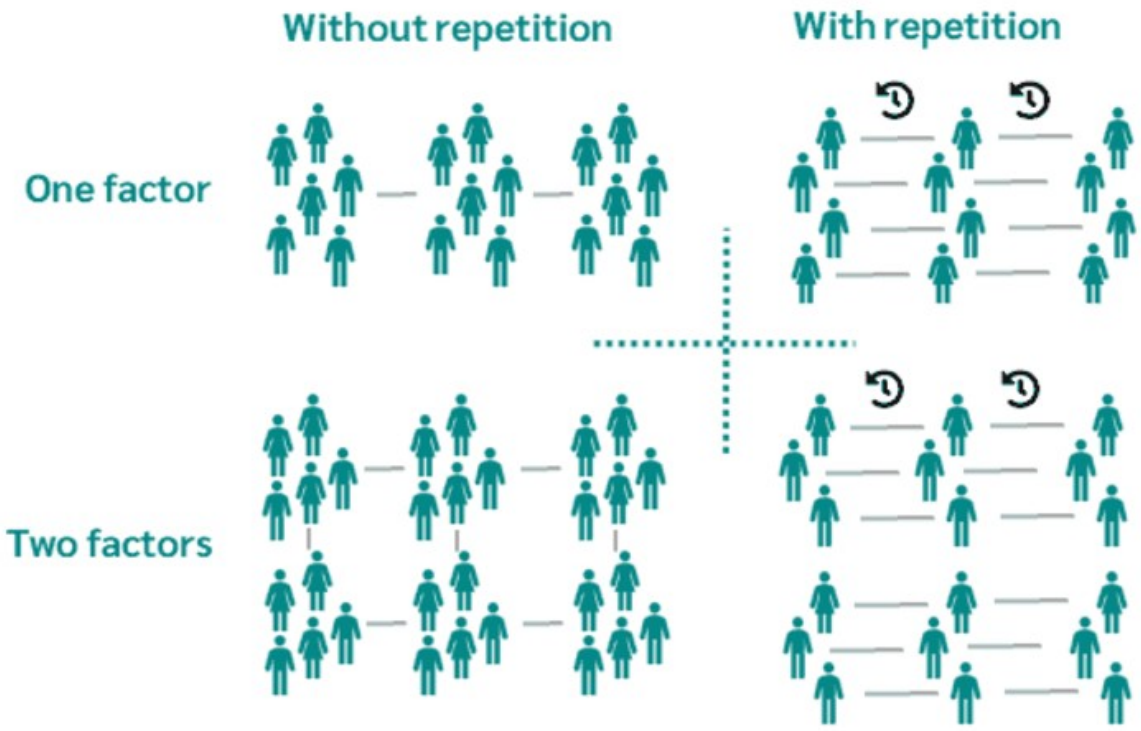
	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Example:

Bus-arrival time:

A bus arrives at a station every 15 minutes, uniformly distributed between 0 and 15 minutes.

Parameters: a=0, b=15



Exercise

1. A factory produces light bulbs, and the lifespan of these light bulbs is uniformly distributed between 500 hours and 1000 hours.Find the probability that a randomly selected light bulb will last between 600 hours and 800 hours.

2. The amount of time (in minutes) that customers spend in a coffee shop is uniformly distributed between 10 minutes and 30 minutes.What is the probability that a customer spends more than 20 minutes in the coffee shop?

3. The weight of a certain type of fruit is uniformly distributed between 150 grams and 250 grams.What is the probability that a randomly selected fruit weighs between 175 grams and 225 grams?

4. A computer program completes tasks in a time that is uniformly distributed between 5 seconds and 15 seconds.Find the probability that a task is completed in less than 10 seconds.

Answer

1. b-a = 1000-500 = 500 , d-c = 800-600 = 200 , P(600<X<800) = 200/500 = 0.4

2. b-a = 30-10 = 20 , d-c = 30-20 = 10 , P(X>20) = 10/20 = 0.5

3. b-a = 250-150 = 100 , d-c = 225-175 = 50, P(175<X<225) = 50/100 = 0.5

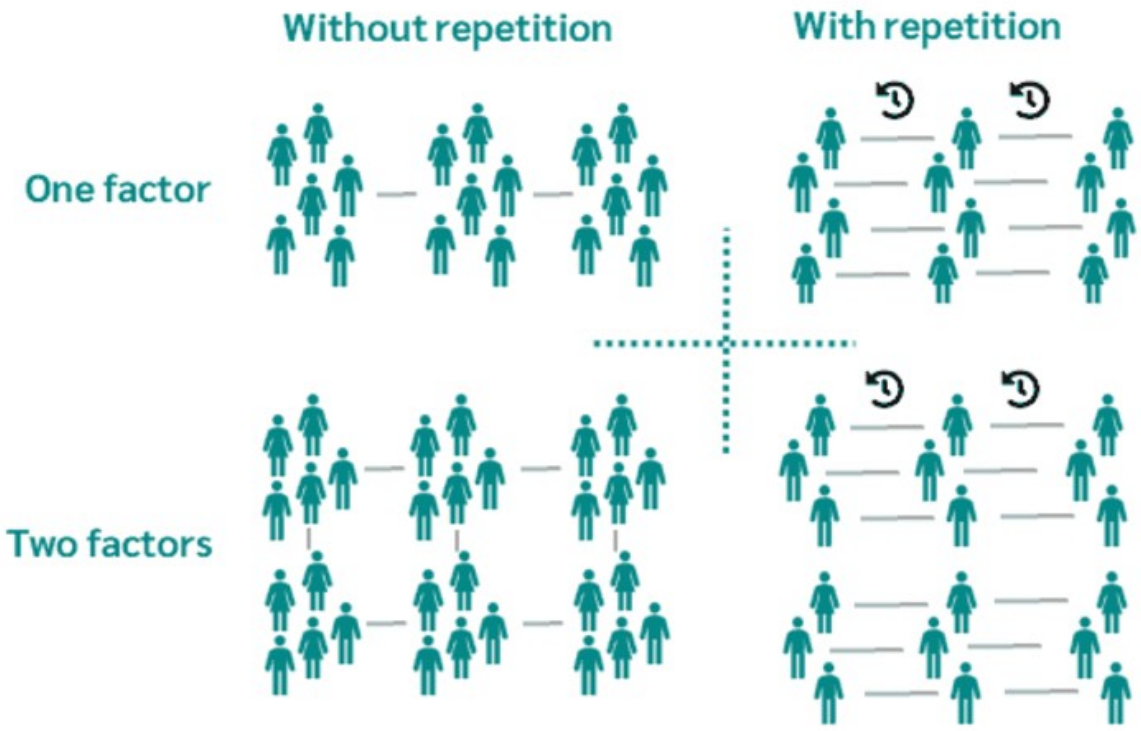
4. b-a = 15-5 = 10, d-c = 10-5 = 5, P(X<10) = 5/10 = 0.5

Normal Distribution

The Normal Distribution, sometimes referred to as the Gaussian Distribution, is a bell-shaped, symmetrical basic continuous probability distribution.

The graph of the normal distribution is characterized by two parameters: the mean, or average, which is the maximum of the graph and about which the graph is always symmetric; and the standard deviation, which determines the amount of dispersion away from the mean.

In a normal distribution, approximately 68% of the data falls within one standard deviation of the mean, about 95% within two standard deviations, and about 99.7% within three standard deviations. This is often referred to as the empirical rule or the 68-95-99.7 rule.



	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Standard Normal Distribution

The standard normal distribution is a specific type of normal distribution with:

Mean (μ) = 0

Standard Deviation (σ) = 1

Shape: Symmetrical, bell-shaped curve centered at zero.

Properties:

68% of values within 1 standard deviation (−1 to 1).

95% of values within 2 standard deviations (−2 to 2).

99.7% of values within 3 standard deviations (−3 to 3).

Z-scores: Each value in the standard normal distribution corresponds to a Z-score:

The Z-score represents how many standard deviations a value is from the mean.

For example, a Z-score of 1.5 indicates a value that is 1.5 standard deviations above the mean.

Member	X (Weight)	Y (Bench Press Strength)
1	60	70
2	65	75
3	70	78
4	75	82
5	80	85

Every Normal distribution can be standardized using the standardization formula which is Z-score.

Why standardize?

Standardization allows us to:

• compare different normally distributed datasets

• detect normality

• detect outliers

• create confidence intervals

• test hypotheses

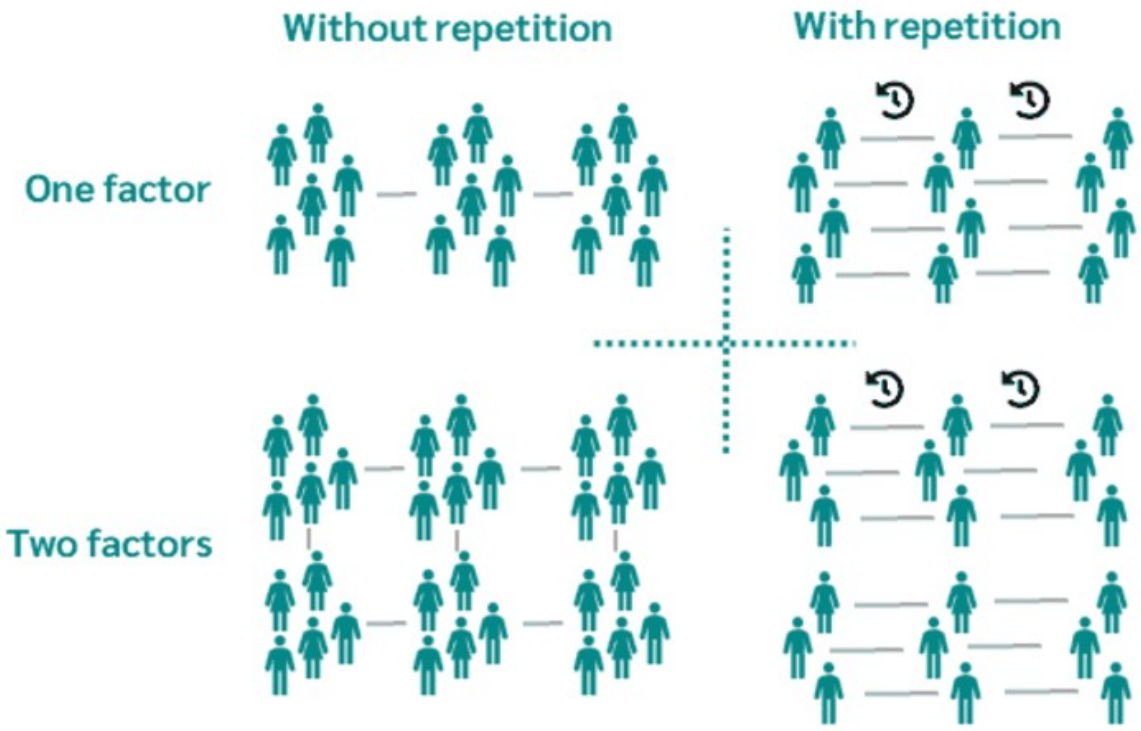
• perform regression analysis

Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

Steps to Find Probabilities:

1. Calculate the Z-score using the formula $Z = (x-\mu)/\sigma$
2. Use the Z-score to find the corresponding probability from the standard normal distribution table



Summary

If you're interested in the height of the distribution at x=120, you calculate the PDF.

If you want the probability of obtaining a value less than or greater than 120, you calculate the Z-score and find the corresponding probability from the Z-table.

Exercise

1. A normal distribution has a mean (μ) of 75 and a standard deviation (σ) of 10. What is the probability of a randomly selected value being less than 80?

2. In a manufacturing process, the weight of a product is normally distributed with a mean weight (μ) of 50 kg and a standard deviation (σ) of 5 kg. What is the probability that a randomly selected product weighs more than 48 kg?

3. The average score of students in a statistics exam follows a normal distribution with a mean (μ) of 70 and a standard deviation (σ) of 8. What is the probability that a student scores between 65 and 78?

4. A company finds that the daily sales of a product are normally distributed with a mean (μ) of 300 units and a standard deviation (σ) of 40 units. What is the probability of selling fewer than 250 units in a day?

5. The heights of adult men in a city are normally distributed with a mean (μ) of 175 cm and a standard deviation (σ) of 7 cm. What is the probability of a randomly selected man being taller than 180 cm?

6. In a quality control process, the lifespan of a light bulb is normally distributed with a mean (μ) of 1200 hours and a standard deviation (σ) of 150 hours. What is the probability that a light bulb lasts longer than 1300 hours?

7. The time taken by students to complete a test follows a normal distribution with a mean (μ) of 45 minutes and a standard deviation (σ) of 10 minutes. What is the probability that a student finishes the test in less than 40 minutes?

8. The average monthly salary of employees in a company is normally distributed with a mean (μ) of $4,000 and a standard deviation (σ) of $500. What is the probability of an employee earning more than $4,500?

9. A study finds that the IQ scores of individuals follow a normal distribution with a mean (μ) of 100 and a standard deviation (σ) of 15. What is the probability of a randomly selected individual having an IQ score less than 85?

10. The daily temperatures in a city are normally distributed with a mean (μ) of 25°C and a standard deviation (σ) of 5°C. What is the probability that the temperature on a randomly chosen day exceeds 30°C?

Answer

1. Mean (μ): 75, Standard Deviation (σ): 10, Z = 80-75/10 = 0.5 , from z-table(P(Z<0.5) = 0.6915) 69.15%

2. Mean (μ): 50, Standard Deviation (σ): 5, Z = 48-50/5 = -0.4 , from z-table(P(Z<-0.4) = 0.3446) For P(X>48) = 1-P(Z<-0.4) = 1-0.3446 = 0.6554

3. Mean (μ): 70, Standard Deviation (σ): 8

For X = 65, Z = 65-70/8 = -0.625 , P(Z<-0.625) =0.2676)

For X = 78, Z = 78-70/8 = 1, P(Z<1) = 0.8413)

P(65<X<78) = P(Z<1)-P(Z<-0.625) = 0.8413-0.2676 = 0.5737

4. Mean (μ): 300, Standard Deviation (σ): 40, Z = 250-300/40 = -1.25, P(Z<-1.25)= 0.1056

5. Mean (μ): 175 cm, Standard Deviation (σ): 7 cm, Z = 180-175/7 = 0.7143, P(Z<0.7143) = 0.7611, P(X>180) = 1-0.7611 = 0.2389

6. Mean (μ): 1200 hours, Standard Deviation (σ): 150 hours, Z = 0.6667, P= 0.7454 , P(X>1300) = 1-0.7454 = 0.2546

7. Mean (μ): 45 minutes, Standard Deviation (σ): 10 minutes, Z = -0.5 , P = 0.3085

8. Mean (μ): $4000, Standard Deviation (σ): $500 , Z = 1, P = 0.8413 , P(X>4500) = 0.1587

9. Mean (μ): 100, Standard Deviation (σ): 15, Z = -1, P = 0.1587

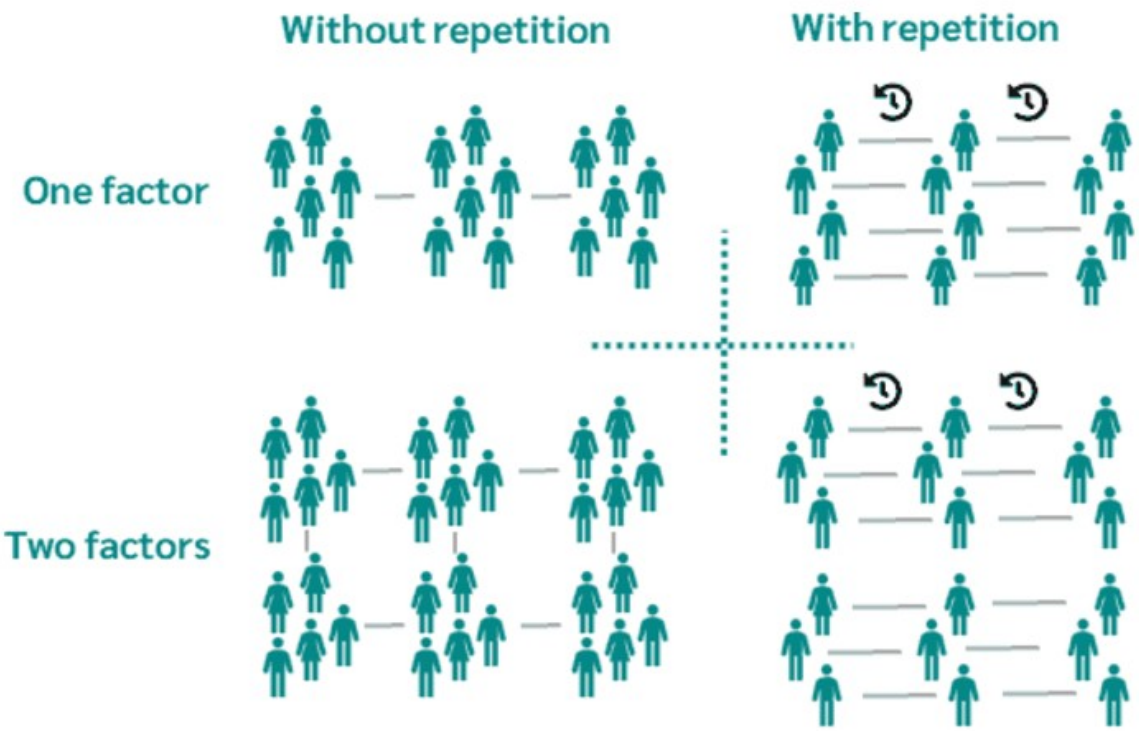
10. Mean (μ): 25°C, Standard Deviation (σ): 5°C, Z = 1, P = 0.8413, P(X>30) = 0.1587

Student's t-distribution

The Student's t-distribution, often simply called the t-distribution, is a probability distribution used when estimating the mean of a normally distributed population in situations where the sample size is small, and/or the population standard deviation is unknown. It is particularly useful when conducting hypothesis tests and constructing confidence intervals for small sample sizes.

The t-distribution is similar in shape to the standard normal distribution (bell-shaped and symmetric) but has heavier tails. This means it is more prone to producing values that fall far from the mean, which makes it better suited for handling small sample sizes where there's more variability.

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130



Key Properties of the t-Distribution:

The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails.

It approaches the normal distribution as the sample size increases. As $n \to \infty$, the t-distribution converges to the standard normal distribution.

The shape of the t-distribution depends on the degrees of freedom (df), which is equal to n−1 for a single sample. Lower degrees of freedom result in heavier tails.

Example

A sample of 16 plants has an average height of 25 cm, with a sample standard deviation of 4 cm. Calculate the t-score if the assumed population mean height is 27 cm.

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

t-score is -2

Exercise

1. A sample of 10 students has an average study time of 3 hours per day, with a sample standard deviation of 0.5 hours. Calculate the t-score for this sample mean if the population mean is 3.2 hours.

2. A study measures the resting heart rates of 15 people, finding a sample mean of 72 beats per minute and a sample standard deviation of 8 beats per minute. Calculate the t-score for a population mean of 75 beats per minute.

3. A sample of 9 people has an average height of 170 cm, with a sample standard deviation of 6 cm. Find the t-score if the population mean is assumed to be 172 cm.

4. The average weight of a sample of 8 puppies is 5 kg, with a sample standard deviation of 0.4 kg. Calculate the t-score for this sample if the population mean is 5.3 kg.

5. A sample of 7 test scores has a mean of 82 and a sample standard deviation of 3. Calculate the t-score if the population mean is assumed to be 85.

Answer

1. $t = \dfrac{3 - 3.2}{\dfrac{0.5}{\sqrt{10}}} = -1.27$

2. $t = \dfrac{72 - 75}{\dfrac{8}{\sqrt{15}}} = -1.45$

3. $t = \frac{170 - 172}{\frac{6}{\sqrt{9}}} = -1$

4. $t = \frac{5 - 5.3}{\frac{0.4}{\sqrt{8}}} = -2.13$

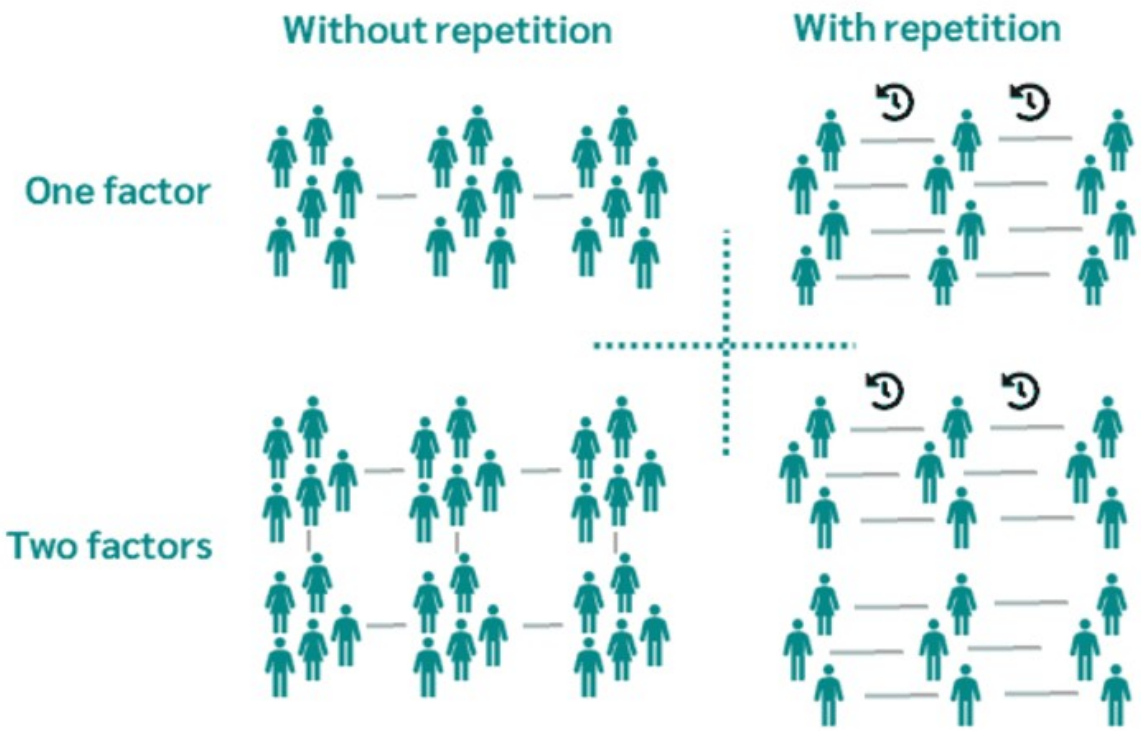
5. $t = \frac{82 - 85}{\frac{3}{\sqrt{7}}} = -2.65$

Exponential Distribution

The exponential distribution is a continuous probability distribution used to model the time elapsed before a given event occurs. In other words, it describes the probability of waiting time between two successive events that occur at a constant average rate.

Key Properties Memoryless Property: The probability of an event occurring in the future is independent of how much time has already elapsed. Parameter: λ, which represents the rate (average number of events per time unit). The mean waiting time between events is 1/λ.

The probability density function (PDF) for the Exponential Distribution is:



where:

x is the time between events,

λ is the rate parameter (how frequently events occur),

e is Euler's number, approximately equal to 2.718.

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

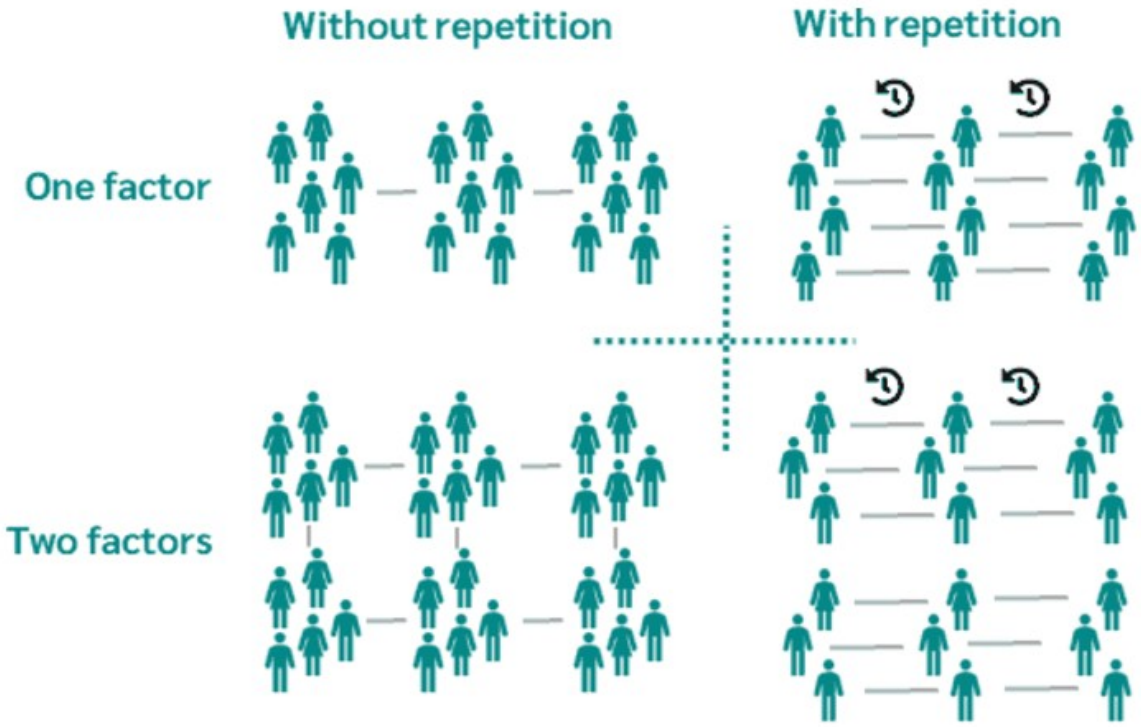
PDF of exponential distribution for several values of λ

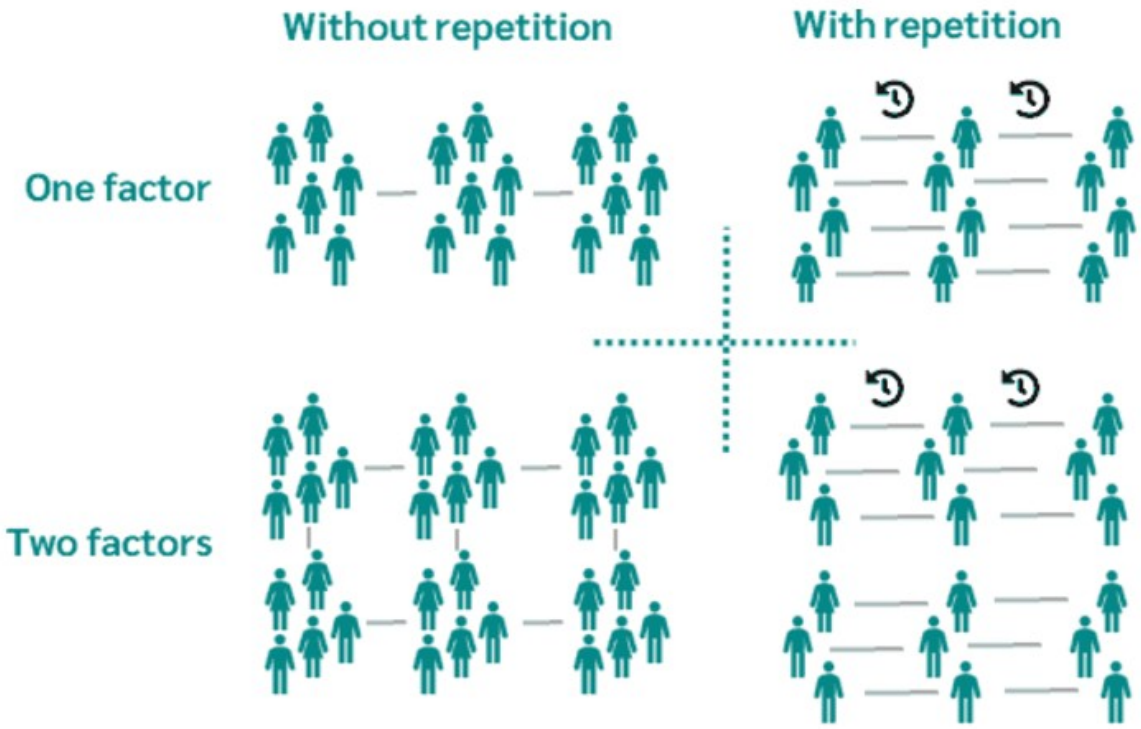
Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

Example

A factory experiences breakdowns at an average rate of 2 breakdowns per day. What is the probability density of a breakdown occurring exactly 1 hour from now?



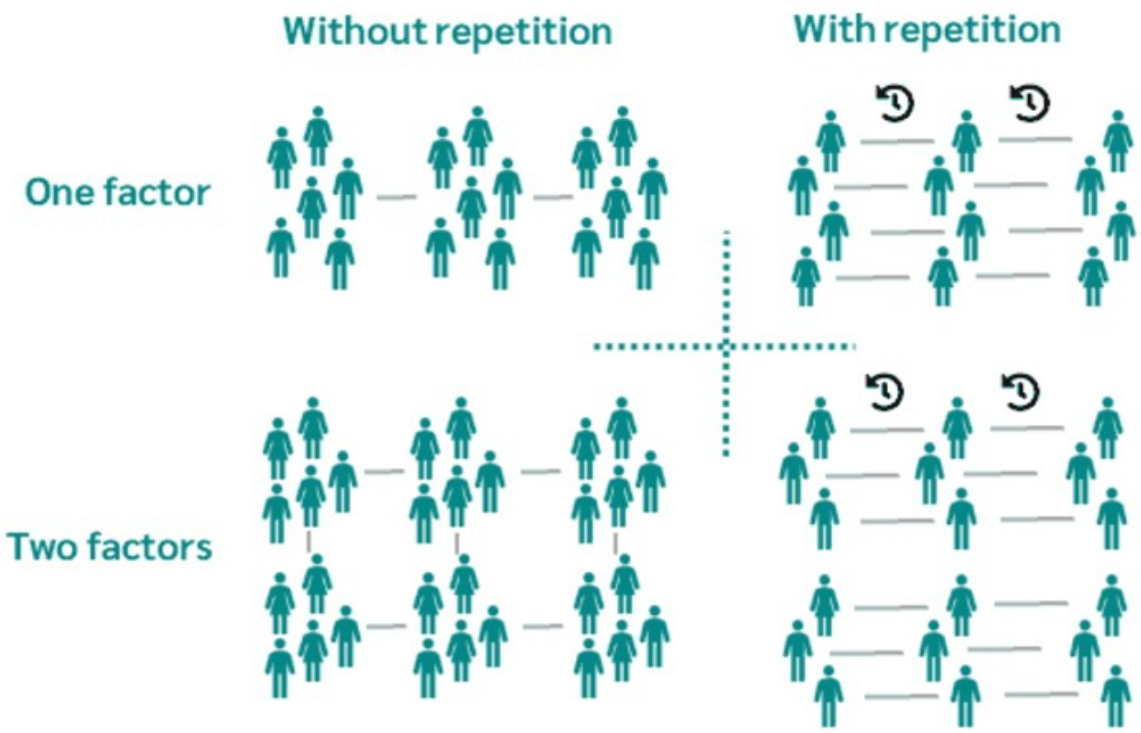


CDF of exponential distribution for several values of λ

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Example

A machine breaks down on average 3 times per day. What is the probability that the next breakdown will occur after more than 1 hour?



Exercise

1.The average waiting time between customer arrivals at a service desk is 5 minutes. What is the probability that the next customer will arrive within 2 minutes?

2.A rare event occurs on average once every 10 days. Find the probability that this event will occur within the next 5 days.

3.The average time between customer arrivals at a café is 10 minutes. Calculate the probability density of a customer arriving exactly 5 minutes from now.

Answer

1. λ = 1/5 = 0.2, x = 2

$$P(X \leq 2) = 1 - e^{-\lambda x} = 1 - e^{-0.2 \times 2} = 1 - e^{-0.4} \approx 1 - 0.6703 \approx 0.3297$$

2.λ = 1/10 =0.1, x = 5

$$P(X \leq 5) = 1 - e^{-\lambda x} = 1 - e^{-0.1 \times 5} = 1 - e^{-0.5} \approx 1 - 0.6065 \approx 0.3935$$

3.λ = 1/10 =0.1, x = 5

$$f(5; 0.1) = 0.1 \cdot e^{-0.1 \times 5} = 0.1 \cdot e^{-0.5} \approx 0.1 \cdot 0.6065 = 0.0607$$

Logistic Distribution

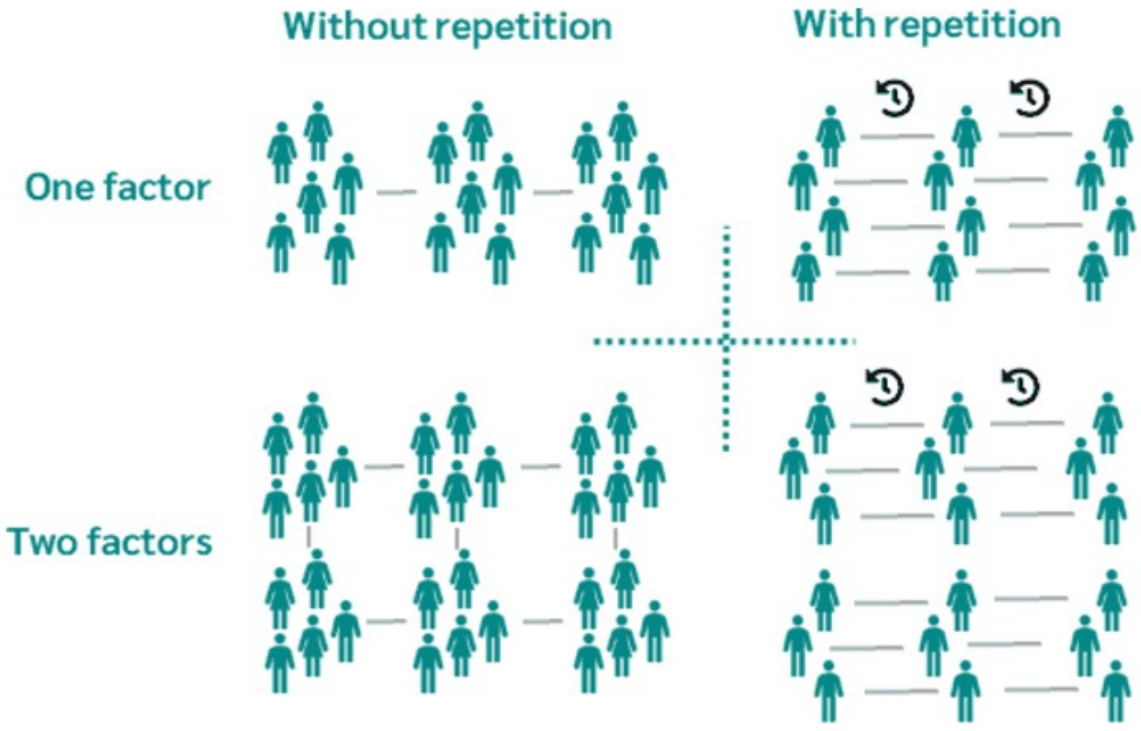
The Logistic Distribution is a continuous probability distribution often used in statistics to model growth or data with an "S-shaped" curve. It's commonly applied in logistic regression and neural networks and is also useful in analyzing growth processes and in cases where data tends to exhibit rapid growth and then plateau.

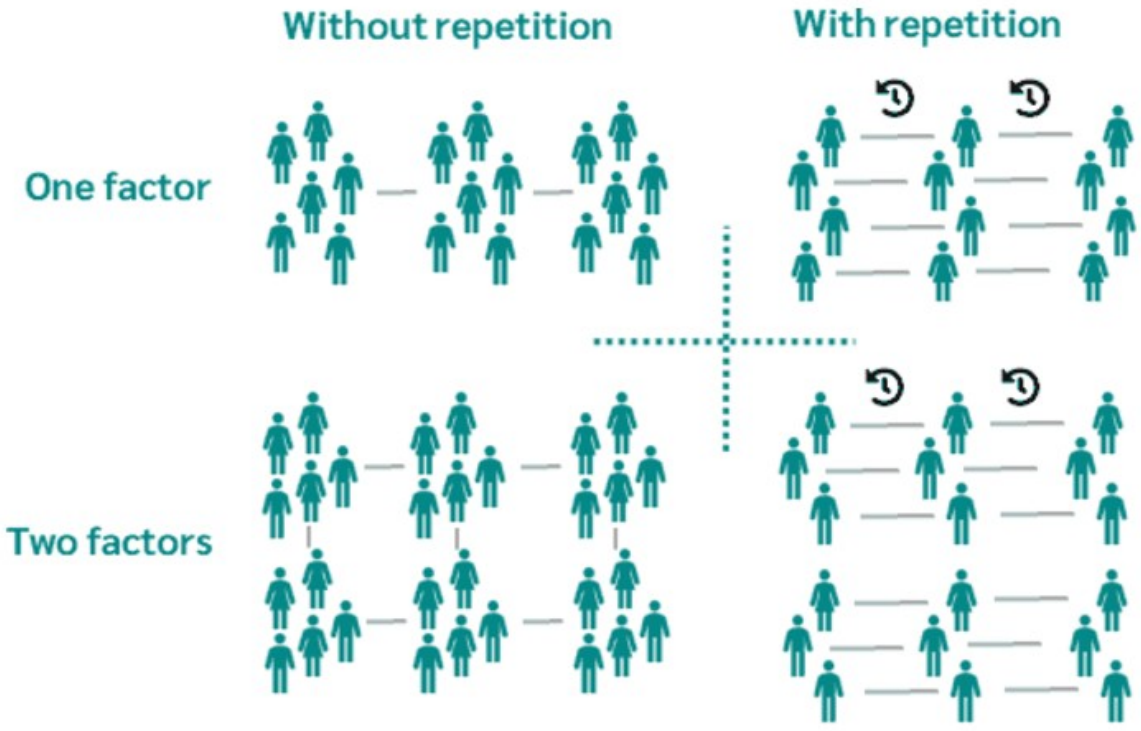
Key Characteristics of the Logistic Distribution:

Symmetry: The logistic distribution is symmetric, centered around its mean, similar to the normal distribution.

S-Shape: In cumulative form, it has an S-shaped (sigmoidal) curve, making it useful for binary classification problems (like "success/failure").

Heavier Tails: Compared to the normal distribution, it has heavier tails, which means there's a higher probability of extreme values.


Without repetition
With repetition
One factor
Two factors


Without repetition
With repetition
One factor
Two factors

Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

Example

Month	X (Company A)	Y (Company B)
1	100	150
2	110	155
3	105	148
4	115	160
5	120	165

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Member	X (Weight)	Y (Bench Press Strength)
1	60	70
2	65	75
3	70	78
4	75	82
5	80	85

Example

Using the same distribution with $\mu = 10$ and $s = 2$, what is the cumulative probability that the product will reach popularity within 12 days?

1. **Given**:
   - $x = 12$,
   - $\mu = 10$,
   - $s = 2$.
2. **Solution**:
   - Using the CDF formula:
   $$F(12; 10, 2) = \frac{1}{1 + e^{-\frac{12-10}{2}}}$$
   - Simplify:
   $$F(12; 10, 2) = \frac{1}{1 + e^{-1}} \approx \frac{1}{1 + 0.3679} \approx 0.7311$$

   Answer: There is approximately a 73.11% probability that the product will reach popularity within 12 days.

Exercise

1.The time it takes for a service request to be resolved follows a logistic distribution with a mean of μ=10 minutes and a scale parameter s=2. Calculate the probability density at exactly 8 minutes.

2.Suppose the time between customer arrivals at a store follows a logistic distribution with a mean of μ=15 minutes and s=3 minutes. What is the probability that a customer will arrive within 12 minutes?

3.The time until a machine requires maintenance is modeled with a logistic distribution with μ=20 hours and s=5. Calculate the probability that the machine requires maintenance within 18 hours.

Answer

1. μ=10, s=2, and x=8.

$$f(8;10,2)=\frac{e^{\frac{8-10}{2}}}{2\left(1+e^{\frac{8-10}{2}}\right)^2}=\frac{e^{1}}{2\left(1+e^{1}\right)^2}\approx 0.0983$$

2. μ=15, s=3, and x=12.

$$F(12;15,3)=1+e^{-\frac{3}{12-15}}=1+e^{1}\approx 1+2.718\approx 0.2689$$

3. μ=20, s=5, and x=18.

$$F(18;20,5)=1+e^{-\frac{5}{18-20}}=1+e^{0.4}\approx 1+1.4918\approx 0.4013$$

Sampling Distribution

A sampling distribution represents the distribution of a statistic, like the mean or standard deviation, which is calculated from multiple samples of a population. It shows how these statistics vary across different samples drawn from the same population.

Example: If you take multiple samples from a population and calculate the mean of each sample, the distribution of those means is the sampling distribution of the mean.

Importance of Sampling Distribution: It allows statisticians to make inferences about the population parameters based on sample statistics and provides the basis for hypothesis testing and confidence intervals. Understanding the concepts of samples and their distributions is crucial for conducting effective statistical analyses and making informed decisions based on data.

Properties of Sampling Distribution

1. Central Limit Theorem (CLT):

The sampling distribution of the sample mean will approach a normal distribution as the sample size (n) increases, regardless of the shape of the population distribution, provided that n is sufficiently large (typically n ≥ 30 is considered adequate).

2. Mean of the Sampling Distribution:

The mean of the sampling distribution ($\mu_{\bar{x}}$) is equal to the population mean (μ):

$$\large\mu_{\bar{x}} = μ $$

3.Standard Deviation of the Sampling Distribution (Standard Error):

The standard deviation of the sampling distribution ($\sigma_{\bar{x}}$) is called the standard error and is calculated as:

$$\large\sigma_{\bar{x}} = \frac{\sigma}{\sqrt n}$$

where σ is the population standard deviation and n is the sample size.

4.Shape of the Distribution:

As mentioned in the Central Limit Theorem, the shape of the sampling distribution of the mean tends to be normal if the sample size is large enough, even if the original population is not normally distributed.

5.Variance of the Sampling Distribution:

The variance of the sampling distribution is the square of the standard error:

$$\large\sigma^{2}_{\bar{x}} = \frac{\sigma ^{2}}{n}$$

6. Effect of Sample Size:

Increasing the sample size decreases the standard error, making the estimates of the population parameter more precise.

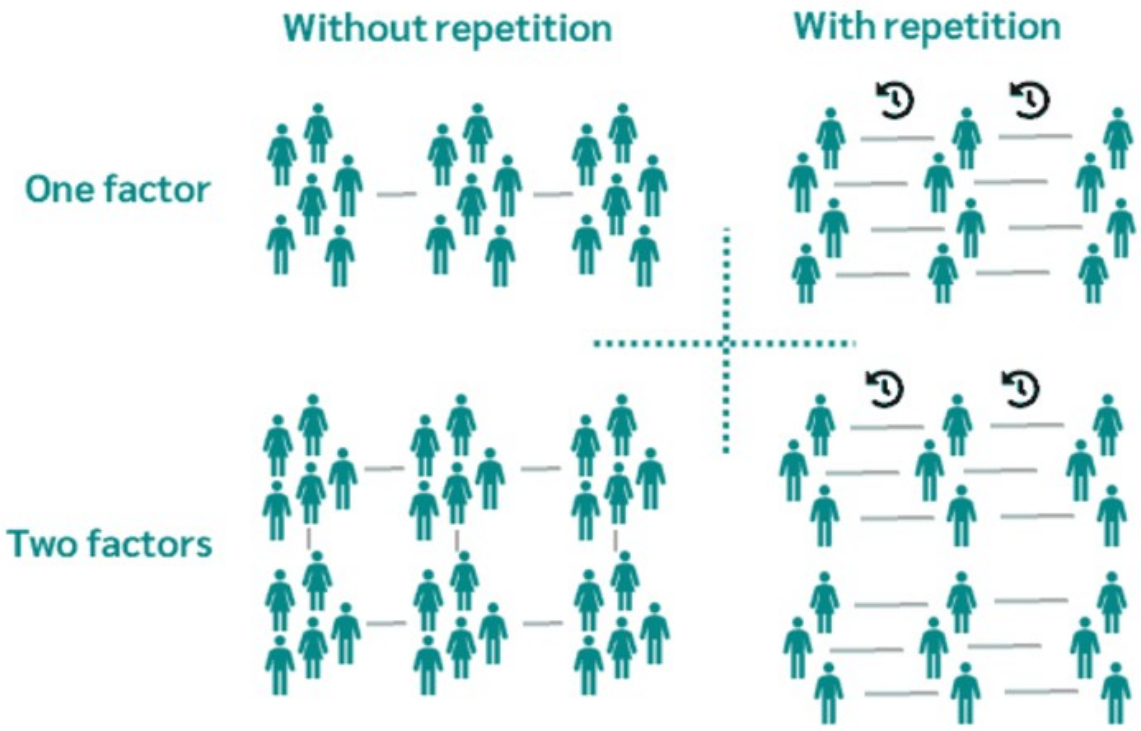
Factors Influencing Sampling Distribution

3 main factors influencing the variability of a sampling distribution are:

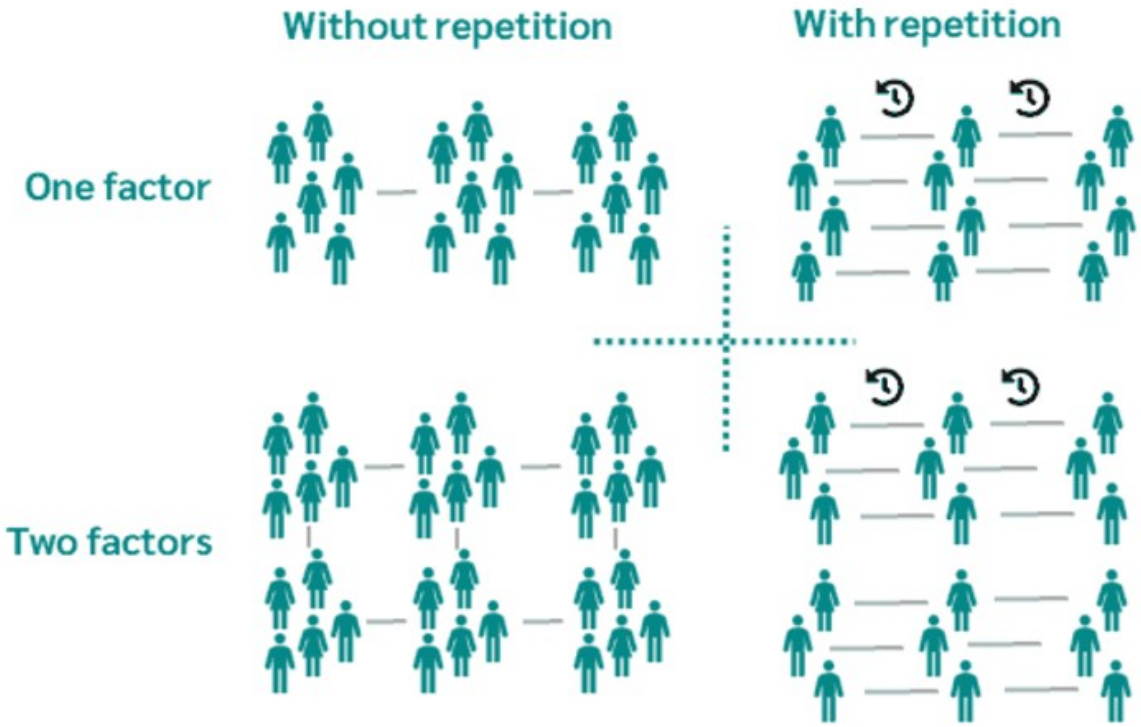
Number Observed in a Population: The symbol for this variable is "N." It is the measure of observed activity in a given group of data.

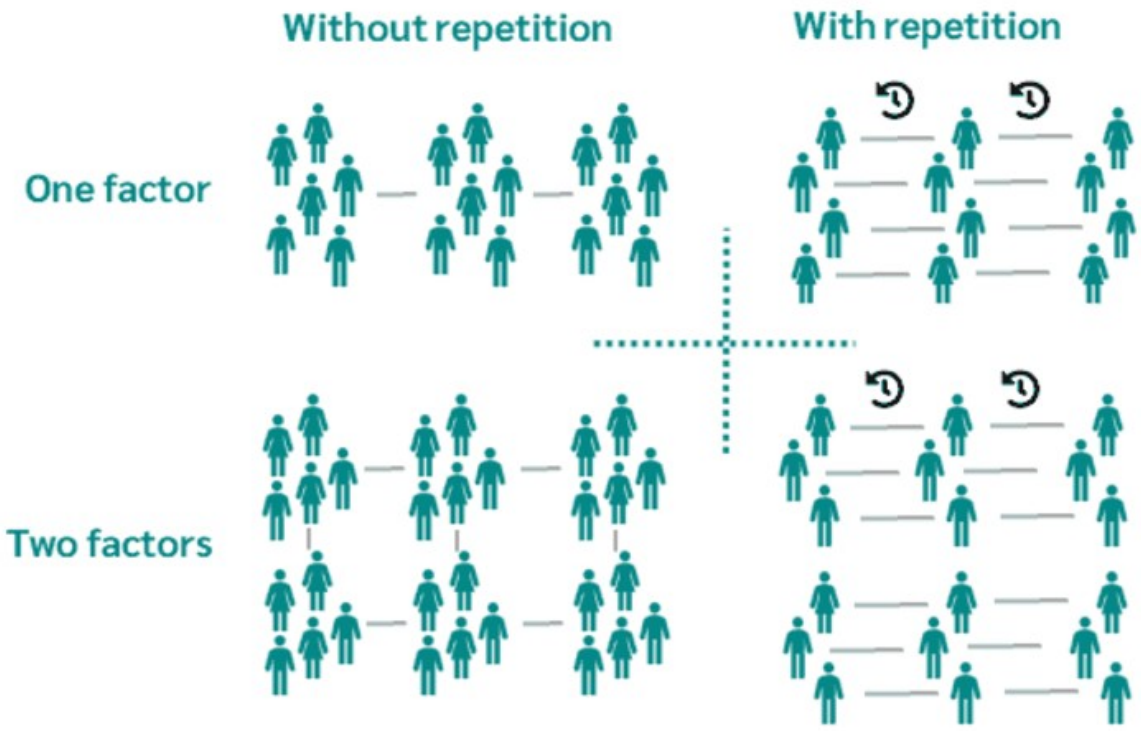
Number Observed in Sample: The symbol for this variable is "n." It is the measure of observed activity in a random sample of data that is part of the larger grouping.

Method of Choosing Sample: How you chose the samples can account for variability in some cases.


Without repetition
With repetition
One factor
Two factors

Example:


Without repetition
With repetition
One factor
Two factors



Measures of relationship between variables

1. **Covariance**
2. **Linear Correlation Coefficient**

Covariance

Covariance is a measure of how two variables change together. In simple terms, it indicates whether an increase in one variable tends to be associated with an increase (or decrease) in the other.

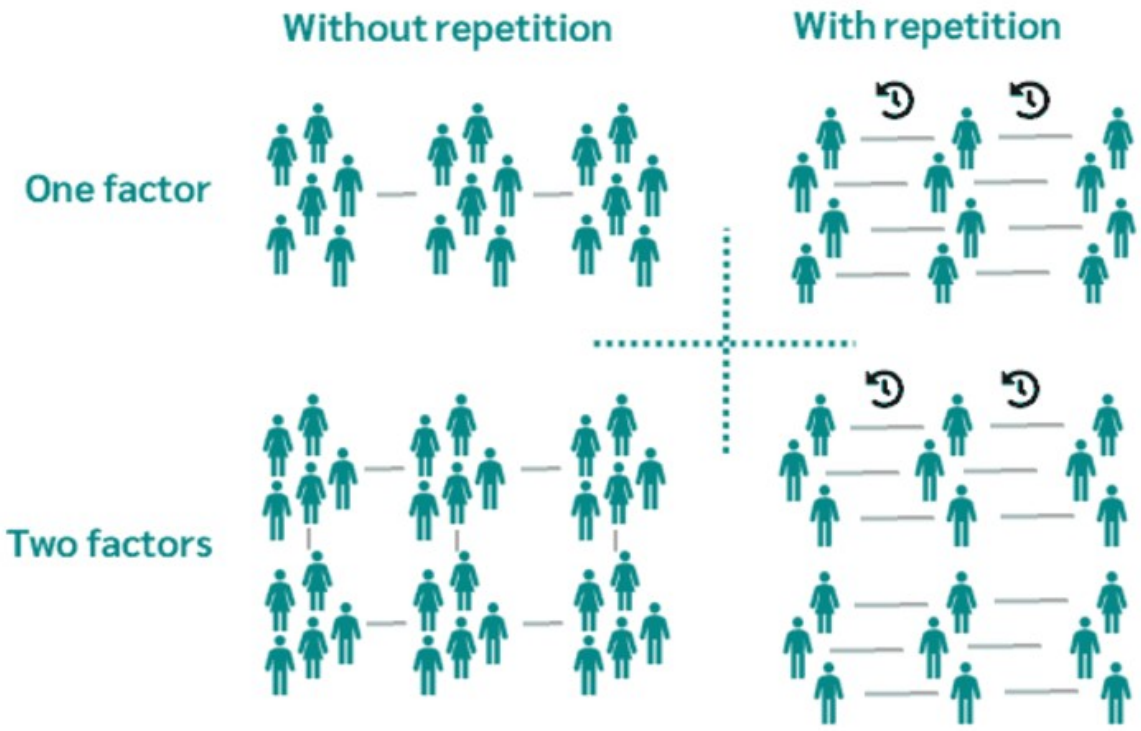
Covariance quantifies the relationship between two variables. It shows the direction of this relationship:

A **positive**covariance means that as one variable increases, the other tends to increase as well.

A **negative** covariance means that as one variable increases, the other tends to decrease.

Zero covariance suggests no predictable relationship in the variables' movements.

Covariance formula for population

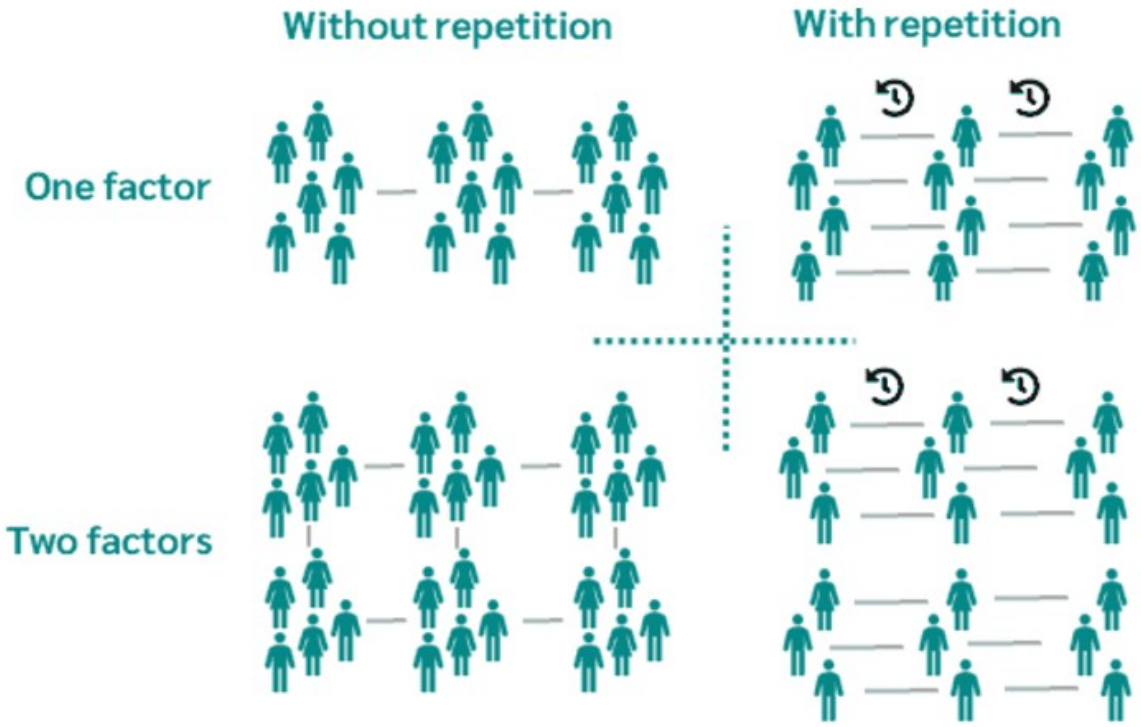


Covariance formula for sample

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Example

Consider a small dataset of two variables: hours studied (X) and test scores (Y) for 5 students.



	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

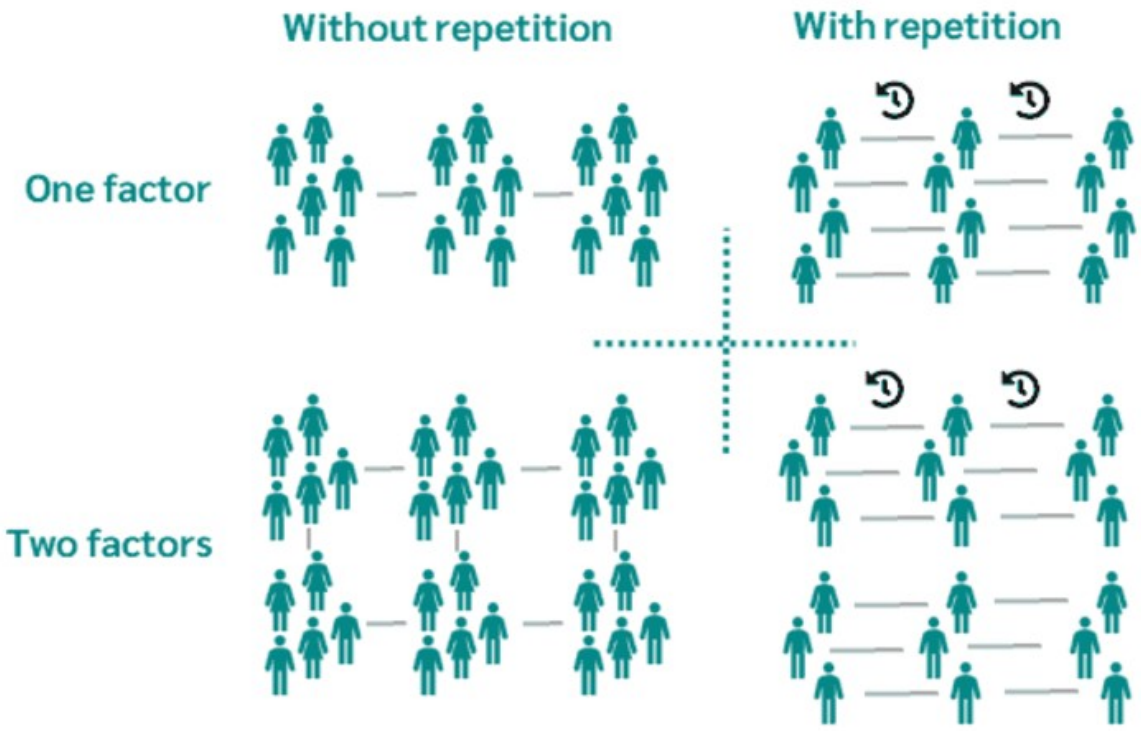
Member	X (Weight)	Y (Bench Press Strength)
1	60	70
2	65	75
3	70	78
4	75	82
5	80	85

The covariance between hours studied (X) and test score (Y) is 10.

Since the covariance is positive, it indicates that as the number of hours studied increases, the test scores tend to increase as well, which makes sense for this dataset.

Exercise

1.Two variables, X and Y, represent the number of study hours and test scores for a group of students:



2.

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

1. A researcher is studying the relationship between the age of participants X and the number of hours they exercise per week Y: Calculate the covariance between age and weekly exercise hours.

Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

4.A gym monitors the weight X (in kg) and bench press strength Y (in kg lifted) of five members:Find the covariance between weight and bench press strength.

Member	X (Weight)	Y (Bench Press Strength)
1	60	70
2	65	75
3	70	78
4	75	82
5	80	85

5.For a financial analysis, an analyst is comparing the monthly closing stock prices X of Company A and Company B's stock prices Y:Calculate the covariance between Company A and Company

B's stock prices.

Month	X (Company A)	Y (Company B)
1	100	150
2	110	155
3	105	148
4	115	160
5	120	165

Correlation

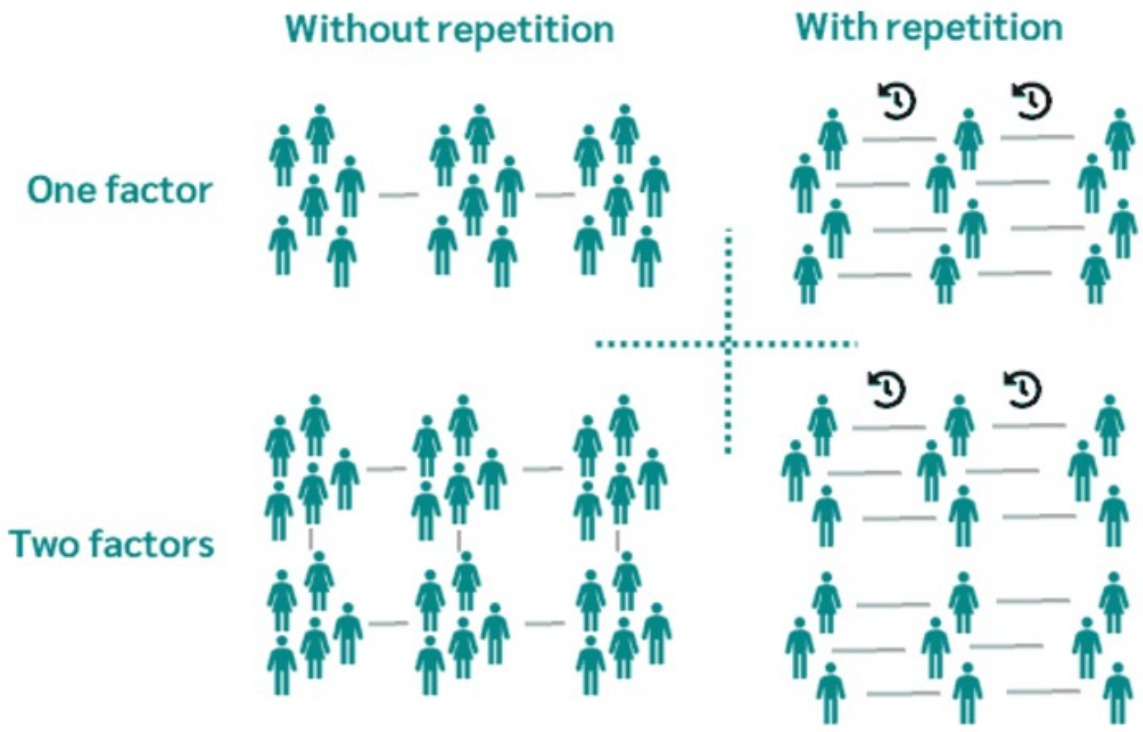
Correlation measures the strength and direction of the linear relationship between two variables. It is expressed as a value between -1 and 1, where:

+1 indicates a perfect positive correlation (as one variable increases, the other also increases).

-1 indicates a perfect negative correlation (as one variable increases, the other decreases).

0 indicates no linear correlation between the variables.

The correlation coefficient that indicates the strength of the relationship between two variables can be found using the following formula:



Where:

rxy – the correlation coefficient of the linear relationship between the variables x and y

xi – the values of the x-variable in a sample

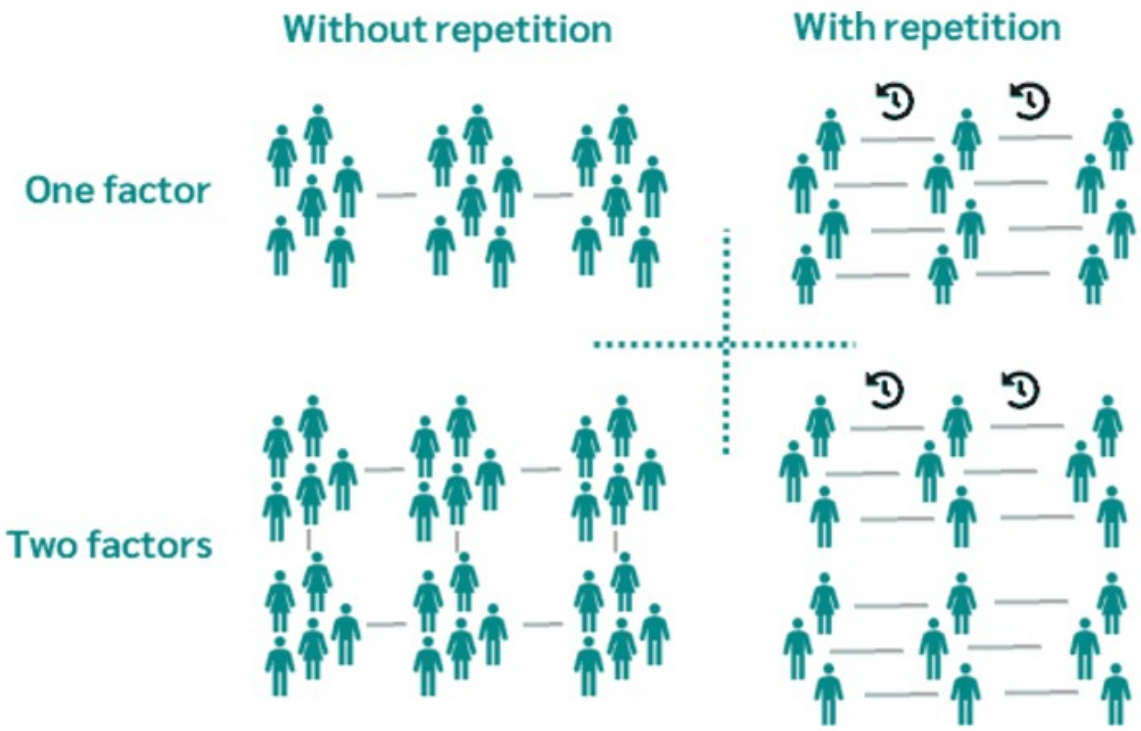
$\bar{x}$ – the mean of the values of the x-variable

yi – the values of the y-variable in a sample

$\bar{y}$ – the mean of the values of the y-variable

It can also be expressed as

$$\huge\frac{\text{Cov}_{XY}}{\sigma_X \sigma_Y}$$

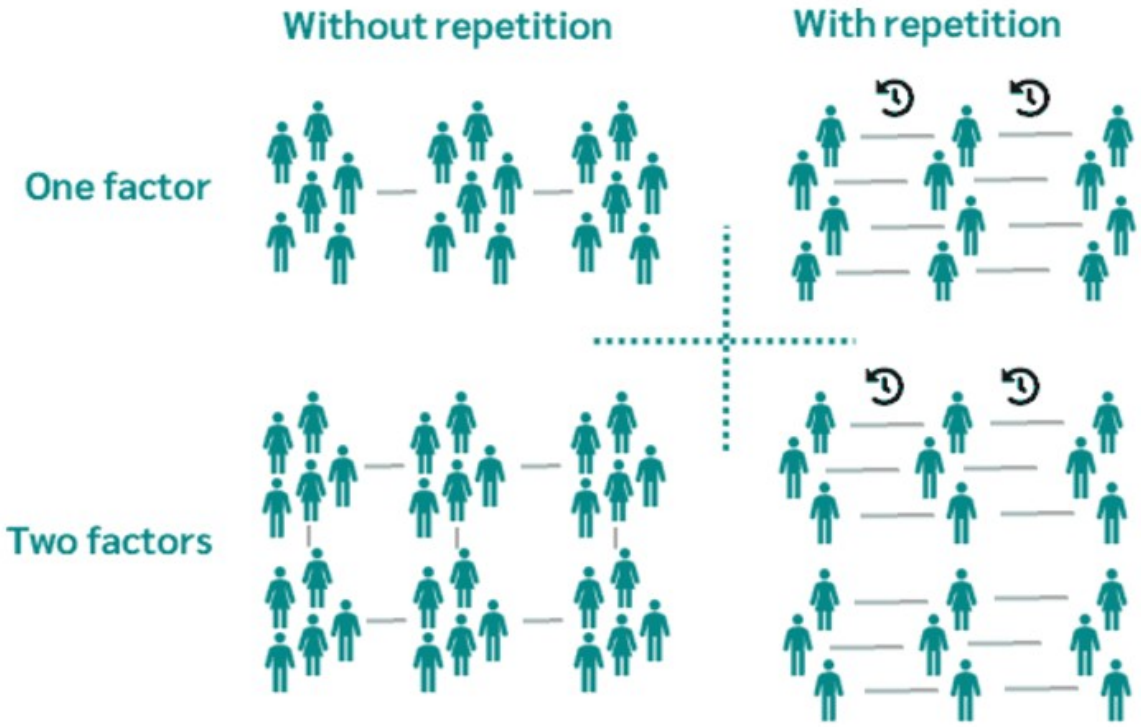


	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

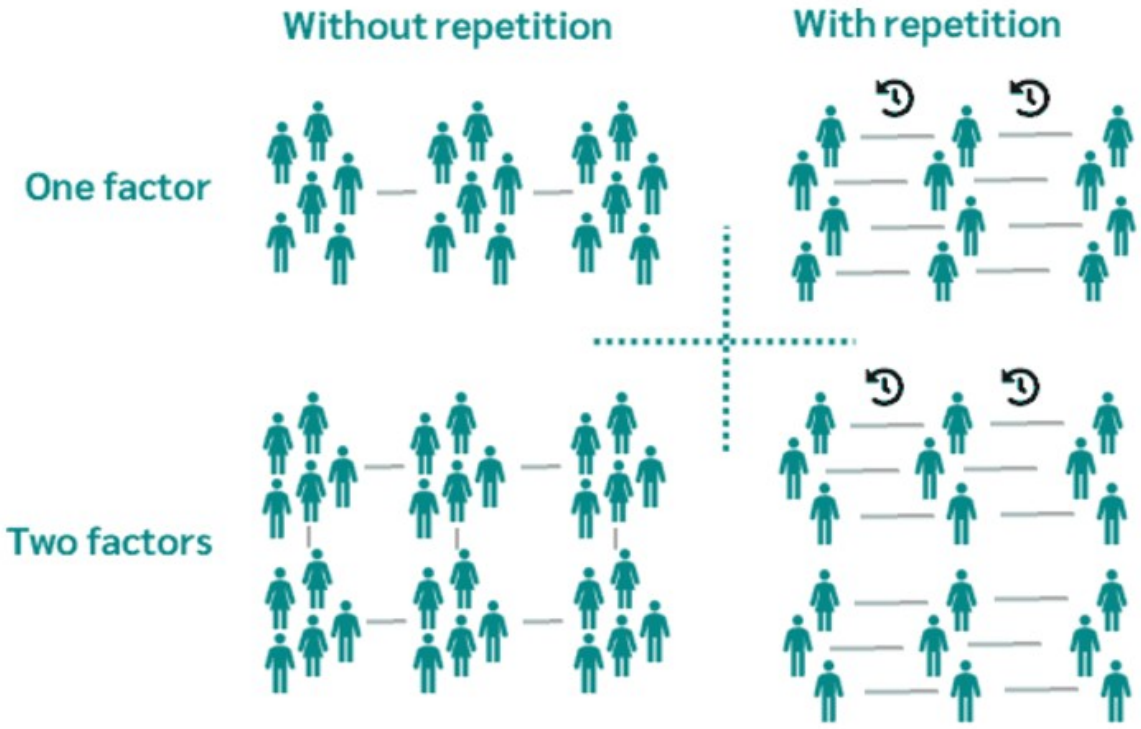
Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

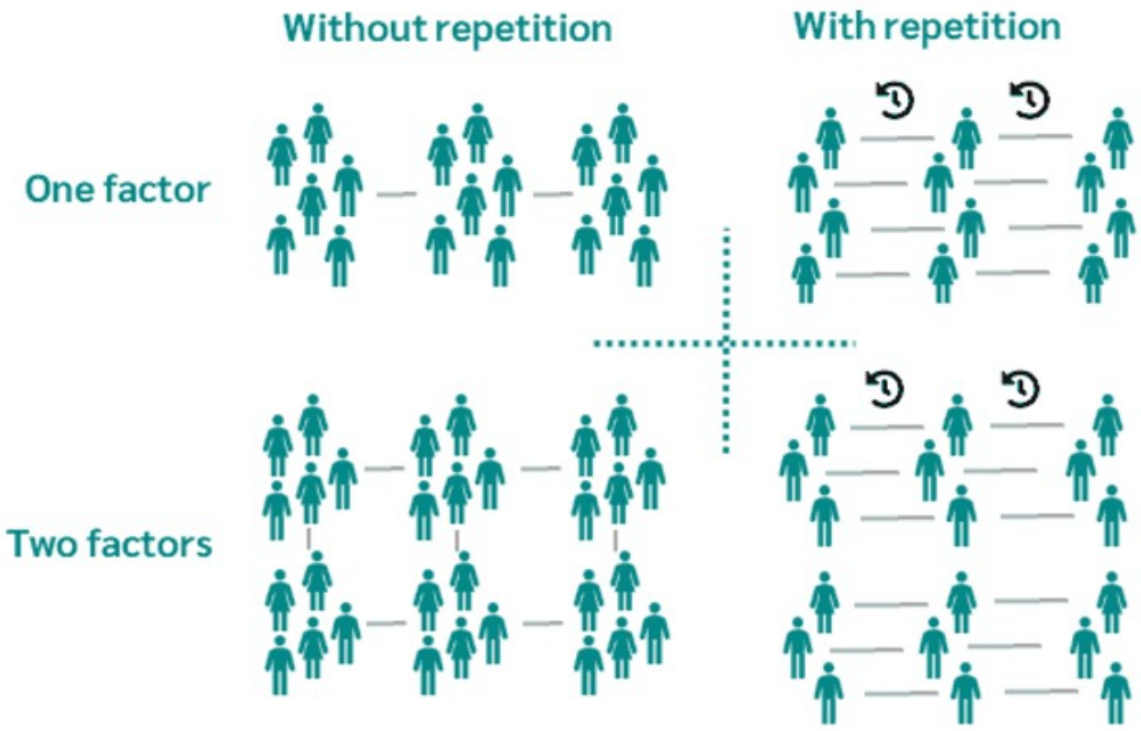
Difference between Covariance and Correlation



Difference between Correlation and Correlation Coefficient



1. A teacher wants to know if there's a relationship between the hours students study and their test scores. Calculate the correlation between study hours (X) and test scores (Y).



2. An economist wants to examine the relationship between a person's age and their average monthly spending on entertainment. Calculate the correlation between age (X) and monthly spending (Y).

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

3. A fitness coach is interested in understanding the relationship between the time spent exercising and the calories burned. Calculate the correlation between exercise time in minutes (X) and calories burned (Y).

Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

Answer

1. $\acute{X}$ = 4 , $\acute{Y}$ = 60

X - $\acute{X}$: -2, -1, 0, 1, 2

Y - $\acute{Y}$: -10, -5, 0, 5, 10

(X - $\acute{X}$)(Y - $\acute{Y}$) :20, 5, 0, 5, 20

(X - $\acute{X}$ ¿^{2} : 4, 1, 0, 1, 4

(Y - $\acute{Y}$ ¿^{2} : 100, 25, 0, 25, 100

r = $\frac{20+5+0+5+20}{\sqrt{(¿4+1+0+1+4)(100+25+0+25+100)¿}}$ = 1

The correlation coefficient is 1, indicating a perfect positive correlation.

1. $\acute{X}$ = 30, $\acute{Y}$ = 264

X - $\acute{X}$: -10, -5, 0, 5, 10

Y - $\acute{Y}$: -64, -14, 6, 26, 46

(X - $\acute{X}$)(Y - $\acute{Y}$) : 640, 70, 0, 130, 460

(X - $\acute{X}$ ¿^{2} : 100, 25, 0, 25, 100

(Y - $\acute{Y}$ ¿^{2} : 4096, 196, 36, 676, 2116

$$r = \frac{640+70+0+130+460}{\sqrt{(¿100+25+0+25+100)(4096+196+36+676+2116)¿}} \approx 0.93$$

The correlation coefficient is approximately 0.93, indicating a strong positive correlation.

1. $\acute{X}$ = 30, $\acute{Y}$ = 300

X - $\acute{X}$: -20, -10, 0, 10, 20

Y - $\acute{Y}$: -200, -100, 0, 100, 200

(X - $\acute{X}$)(Y - $\acute{Y}$) :4000, 1000, 0, 1000, 4000

(X - $\acute{X}$ ¿^{2} : 400, 100, 0, 100, 400

(Y - $\acute{Y}$ ¿^{2} : 40000, 10000, 0, 10000, 40000

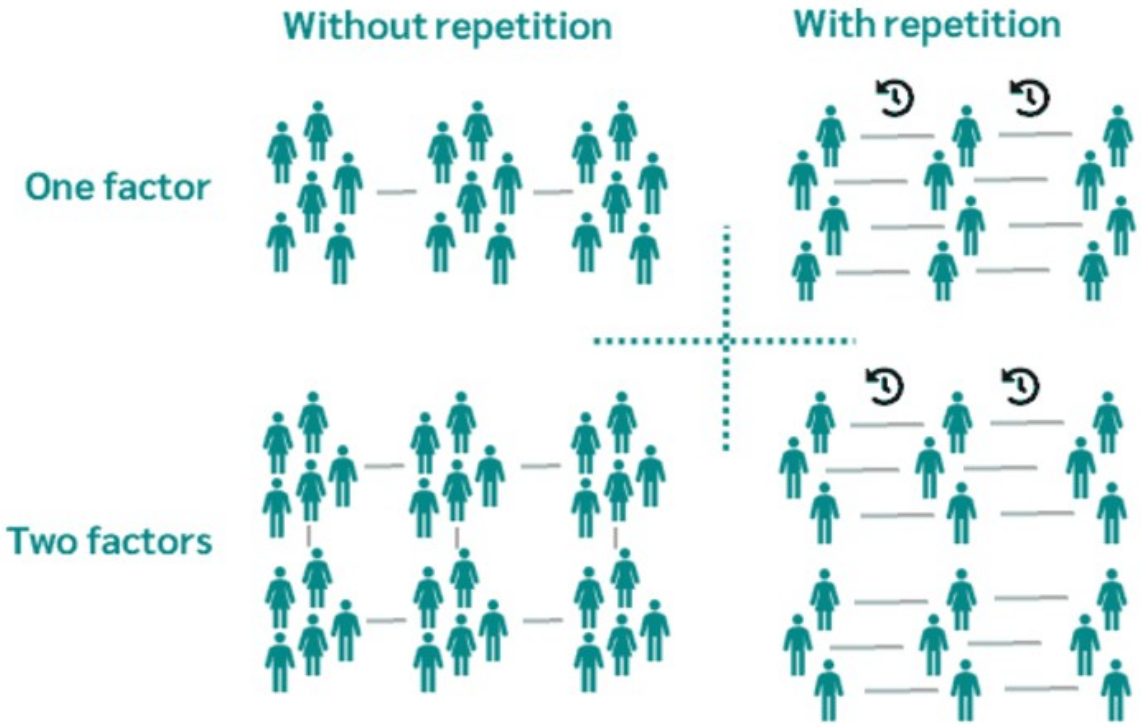
$$r = \frac{4000+1000+0+1000+4000}{\sqrt{(¿400+100+0+100+400)(40000+10000+0+10000+40000)¿}} = 1$$

The correlation coefficient is 1, indicating a perfect positive correlation

Causation

Causation describes a relationship between two variables where a change in one directly leads to a change in the other. In other words, causation implies that one variable is the reason behind the occurrence or change in the other variable. For example, if increasing study hours causes an increase in test scores, then study hours and test scores have a causal relationship.

Difference between Causation and correlation



Central Limit Theorem(CLT)

The Central Limit Theorem (CLT) states that when you take a sufficiently large number of random samples from any population, the distribution of the sample means will approach a normal distribution (a bell curve), regardless of the shape of the original population. This is true as long as the samples are independent and come from the same population.

The Central Limit Theorem can be mathematically represented as:

$$\acute{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

where:

$\acute{X}$ is the sample mean

μ is the population mean

σ is the population standard deviation

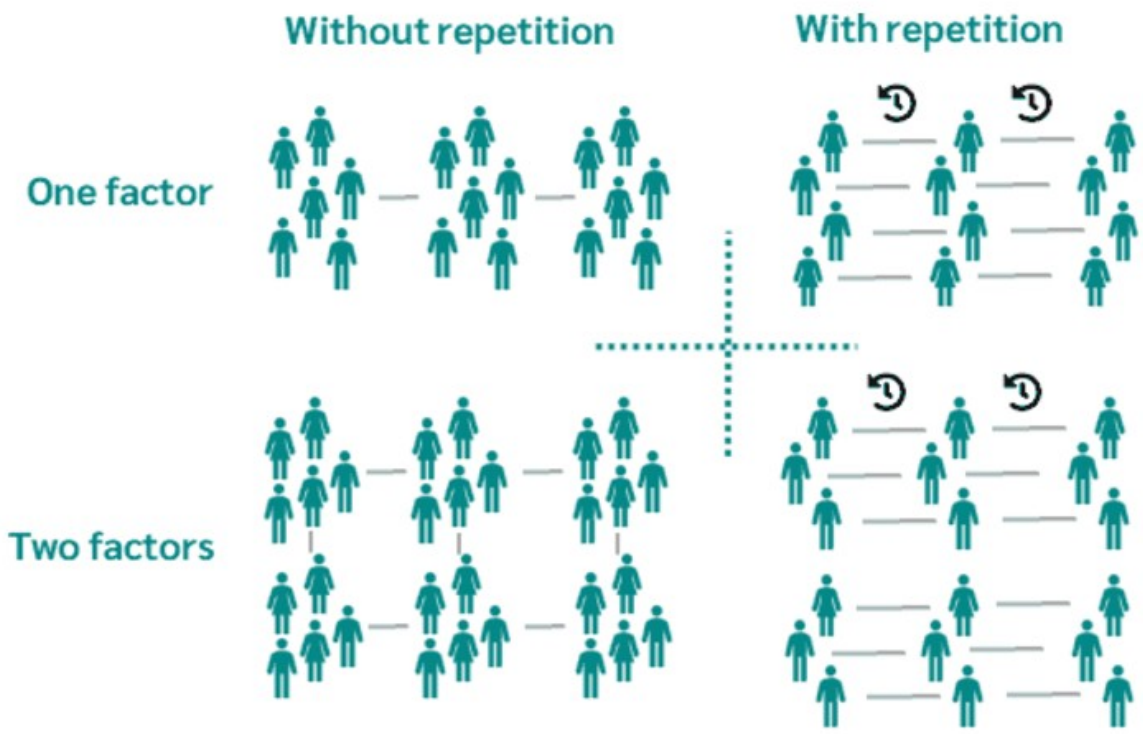
n is the sample size

$\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean, which indicates how much the sample mean is expected to vary from the true population mean.

Key points:

As the sample size (n) increases, the standard error ($\frac{\sigma}{\sqrt{n}}$) decreases, meaning the sample means become more concentrated around the population mean.

The larger the sample size, the more closely the distribution of sample means will resemble a normal distribution.



	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Example 1: Tossing a Coin (Discrete Example)

Imagine you're tossing a fair coin, which has two possible outcomes: heads or tails (probability = 0.5 for each).

The population distribution is uniform, not normal.

You take multiple random samples of, say, 30 tosses at a time, and calculate the average number of heads for each sample.

According to the CLT, if you repeat this process many times, the distribution of these averages will approach a normal distribution, even though the original coin toss distribution was uniform.

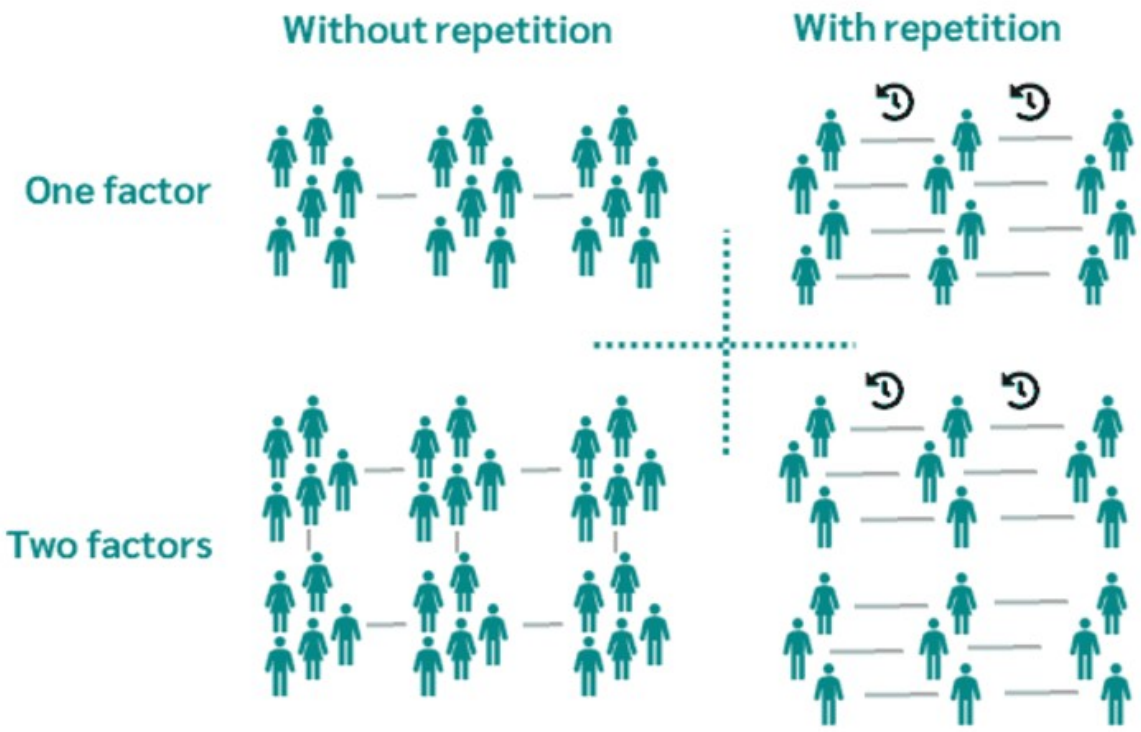
Example 2: Heights of People (Continuous Example)

Suppose you are studying the heights of adults in a city, and the distribution of heights is not normal but skewed to the right.

If you take multiple random samples (say, 50 people per sample) and calculate the average height for each sample, the distribution of these sample means will approximate a normal distribution as the number of samples increases.

CLT formula

Let us assume we have a random variable X. Let σ be its standard deviation and μ is the mean of the random variable. Now as per the Central Limit Theorem, the sample mean $\acute{X}$ will approximate to the normal distribution which is given as $\acute{X} \sim$ N(μ, σ/√n). The Z-Score of the random variable $\acute{X}$ is given as Z = $\large\frac{\bar{X} - μ}{\frac{σ}{\sqrt n}}$



Example

1. The male population's weight data follows a normal distribution. It has a mean of 70 kg and a standard deviation of 15 kg. What would the mean and standard deviation of a sample of 50 guys be if a researcher looked at their records?

Given: $\mu = 70$, $\sigma = 15$, $n = 50$. As per the Central Limit Theorem, the sample mean is equal to the population mean.

Hence, sample mean = 70, sample std = $\frac{15}{\sqrt{50}} \approx 2.1$

Exercise

1. A distribution has a mean of 69 and a standard deviation of 420. Find the mean and standard deviation if a sample of 80 is drawn from the distribution.

2. The mean age of people in a colony is 34 years. Suppose the standard deviation is 15 years. The sample of size is 50. Find the mean and standard deviation of the sample.

3. The mean time taken to read a newspaper is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of size 70. Find its mean and standard deviation.

4. A distribution has a mean of 12 and a standard deviation of 3. Find the mean and standard deviation if a sample of 36 is drawn from the distribution.

5. A distribution has a mean of 4 and a standard deviation of 5. Find the mean and standard deviation if a sample of 25 is drawn from the distribution.

Answer

1. mean = 69, std = 46.95

2. mean = 34 , std = 2.12

3. mean = 8.2, std = 0.11

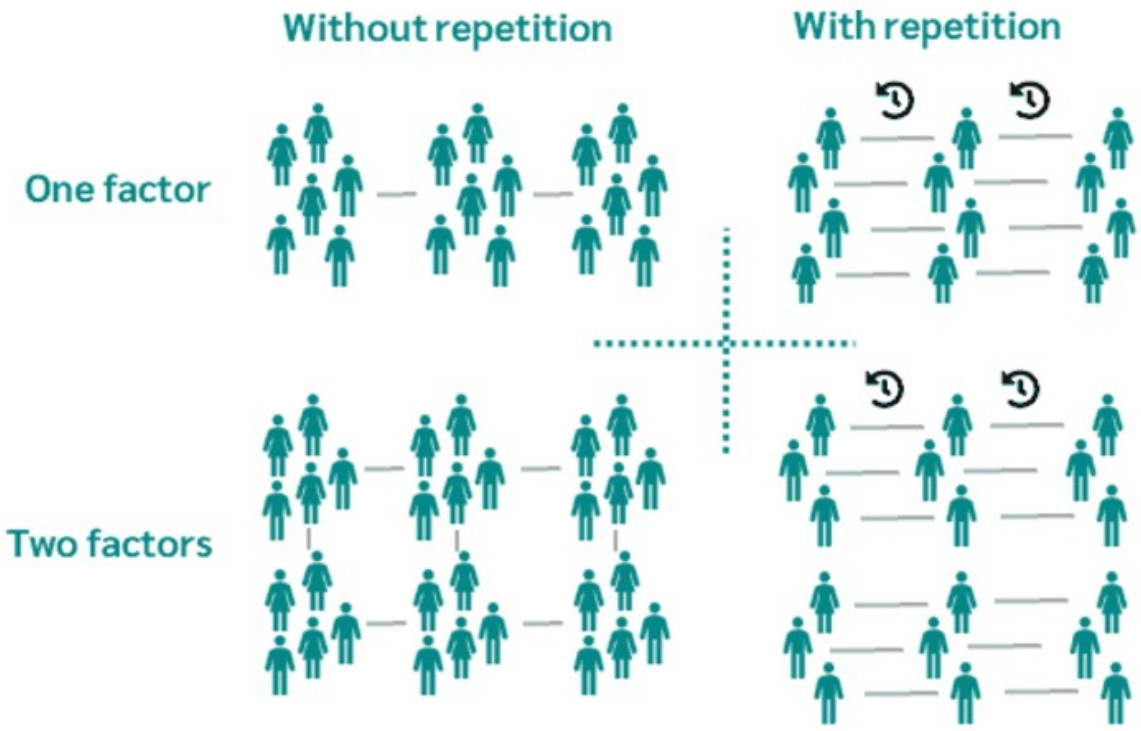
4. mean = 12, std = 0.5

5. mean = 4, std = 1

Estimators and Estimates

Estimators: An estimator is a formula or rule we use to calculate an estimate of a population parameter (like the mean or proportion) from a sample. For example, the sample mean $\acute{X}$ is an estimator of the population mean μ.

Estimates: An estimate is the specific numerical value you get when you apply the estimator to a sample. So, if you calculate a sample mean of 75, then 75 is your estimate of the population mean.



Types of Estimate

1. **Point Estimate**

A point estimate provides a single value as an estimate of a population parameter. For example, using the sample mean $\acute{X}$ as an estimate of the population mean μ.

Example: If the average height of a sample of 50 people is 170 cm, then 170 cm is the point estimate of the population mean height.

1. **Interval Estimate**

An interval estimate gives a range of values within which the population parameter is likely to fall. This range is often expressed with a confidence level, which indicates how certain we are that the parameter lies within this interval.

Example: A 95% confidence interval for the population mean height might be [167 cm, 173 cm].

Two Main Properties of Estimators:

Unbiasedness: An estimator is unbiased if its average (expected value) equals the true population parameter, ensuring no systematic over- or under-estimation.

Efficiency: An estimator is efficient if it has the lowest variance among unbiased estimators, providing consistent values close to the true parameter across samples.

These properties help ensure estimators are both accurate (unbiased) and reliable (efficient).

Confidence Level

The confidence level represents the probability that a confidence interval will contain the true population parameter if you were to repeat the sampling process many times. It's expressed as a percentage (e.g., 90%, 95%, 99%) and is used to show how certain you are that the population parameter lies within the interval.

Mathematically, the confidence level is written as $1-\alpha$, where alpha (α) is the significance level or risk level. For instance, if $\alpha=0.05$, then $1-\alpha=0.95$, meaning a 95% confidence level.

Example: A 95% confidence level (where $\alpha=0.05$) means that if you repeated the sampling process multiple times, 95% of the confidence intervals would contain the true population parameter.

Key Components:

Confidence Level (CL):

It's the percentage of times you expect the true parameter to lie within the confidence interval if you were to repeat the sampling process many times. Example: A 95% confidence level means that 95% of the intervals created from repeated samples will contain the true population parameter.

Z-score or Critical Value:

The confidence level corresponds to a Z-score (for normal distributions) or a t-score (for small sample sizes or when the population standard deviation is unknown). A higher confidence level (e.g., 99%) will result in a wider interval because you want to be more confident that the parameter lies within that interval. Example: For a 95% confidence level, the Z-score is typically 1.96 (from the standard normal distribution).

How Confidence Level Affects the Interval:

Higher Confidence Level: Results in a wider confidence interval. This is because a higher confidence level means you want to be more certain that the true population parameter is within the interval, so the range of possible values is larger.

Example: 99% CI gives a wider interval than a 95% CI.

Lower Confidence Level: Results in a narrower confidence interval. The trade-off is lower confidence that the true population parameter is inside the interval.

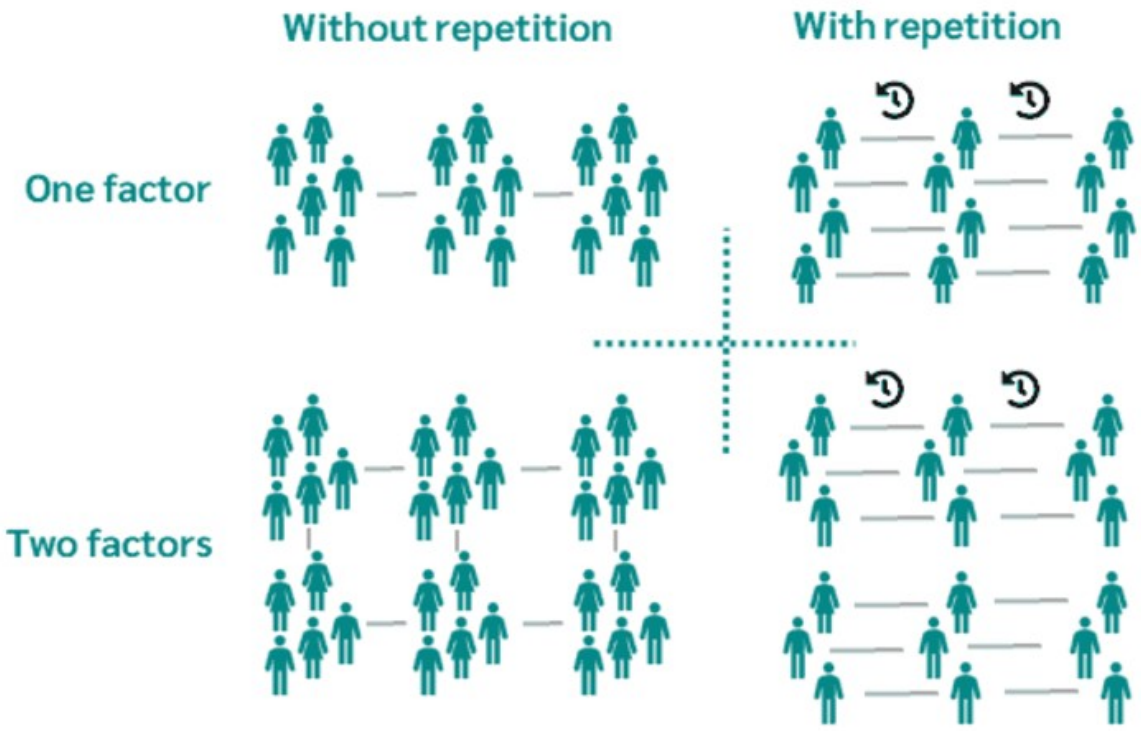
Example: 90% CI gives a narrower interval than a 95% CI.

Common Confidence Levels:

90% Confidence Level: Z-value ≈ 1.645

95% Confidence Level: Z-value ≈ 1.96

99% Confidence Level: Z-value ≈ 2.576



Margin of Error

The margin of error (MOE) represents the range within which the true population parameter is likely to fall, relative to the sample estimate. It shows how much you can expect your sample results to differ from the actual population parameter.

Formula for Margin of Error (for the Mean):

$$\text{Margin of Error} = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Where:

$Z_{\alpha/2}$: Z-score corresponding to the chosen confidence level (e.g., 1.96 for 95% confidence)

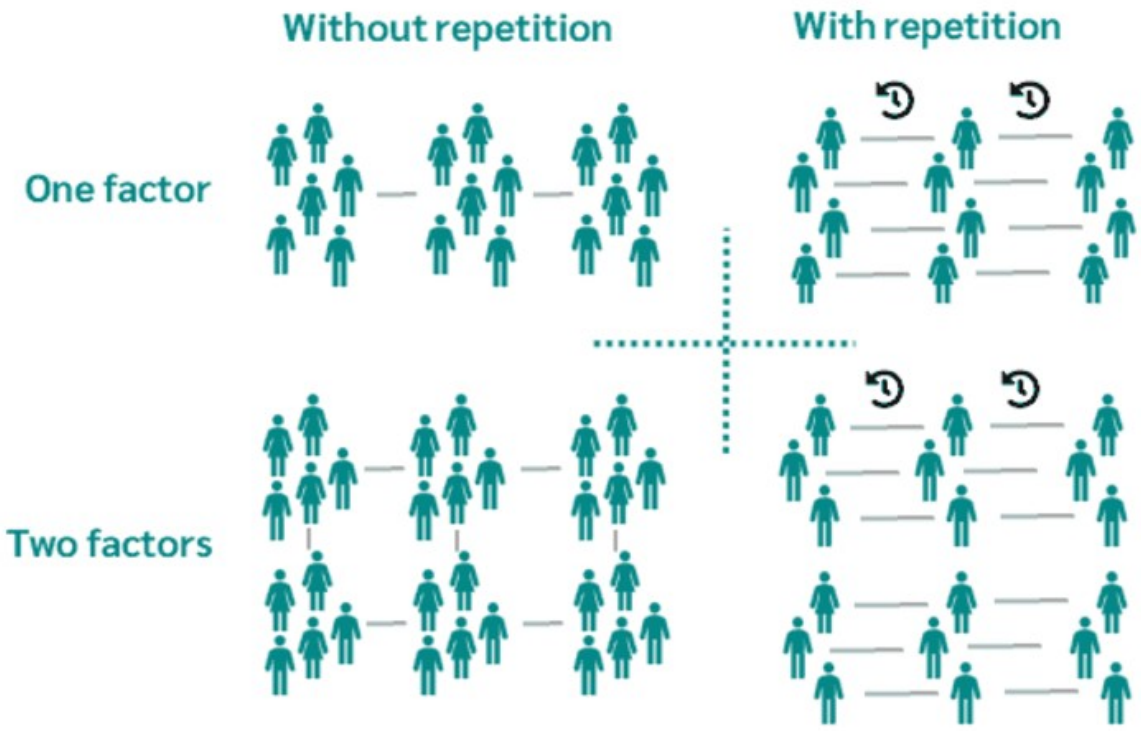
σ: Population standard deviation (or an estimate if unknown)

$\sqrt{n}$: Sample size

Key Points

Confidence Level: Higher confidence levels increase the margin of error, making the interval wider.

Sample Size: Larger sample sizes decrease the margin of error, making the interval narrower.



Confidence Interval

A confidence interval is a range of values derived from a sample that, if the sampling were repeated many times, would contain the true population parameter a certain percentage of the time

Key Components of Confidence Intervals

Point Estimate: The sample statistic (e.g., sample mean) used as the center of the interval.

Confidence Level: The probability that the interval includes the population parameter. Common levels are 90%, 95%, and 99%.

Margin of Error: The range around the point estimate, influenced by the confidence level and sample size.

Formula for CI:

$$CI = \acute{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Where:

$\acute{X}$ = sample mean

$Z_{\alpha/2}$ = Z-score based on the confidence level (e.g., 1.96 for 95%)

σ = population standard deviation (or an estimate if unknown)

n = sample size

Example:

A researcher conducted a study on the average weight of apples in an orchard. A sample of 50 apples had an average weight of 150 grams with a standard deviation of 20 grams. Calculate the 95% confidence interval for the average weight of all apples in the orchard.

Sample Mean ($\acute{X}$) = 150 grams

Standard Deviation (σ) = 20 grams

Sample Size (n) = 50

Confidence Level = 95% (Z-value ≈ 1.96)

$$CI = \acute{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

$$CI = 150 \pm 1.96 \times \frac{20}{\sqrt{50}}$$

$\frac{20}{\sqrt{50}}$ = 2.83

1.96*2.83 = 5.55

CI = 150 ± 5.55 = (144.45,155.55)

The 95% confidence interval for the average weight of apples is (144.45, 155.55) grams.

Exercise

1.A sample of 25 students' exam scores has a mean score of 75, with a sample standard deviation of 8. Calculate the 95% confidence interval for the mean score of all students. Assume that the population is normally distributed.

2.A factory produces light bulbs, and a sample of 40 bulbs is tested for their lifespan. The average lifespan of the bulbs in the sample is found to be 1,200 hours, with a standard deviation of 100 hours. Calculate the 95% confidence interval for the average lifespan of all light bulbs produced by the factory.

3.A study is conducted to estimate the average height of adult men in a city. A random sample of 36 men is taken, and the sample mean height is found to be 175 cm with a standard deviation of 8 cm. Calculate the 99% confidence interval for the average height of all adult men in the city.

Answer

1. n = 25, $\acute{X} = 75$, σ = 8, confidence level = 95%, Z-value = 1.96, CI = 75 ± 3.136 (71.864,78.136)

2. n = 40, $\acute{X} = 1200$, σ = 100, confidence level = 95%, Z-value = 1.96, CI = 1200 ± 30.989 (1169.010,1230.989)

3. n = 36, $\acute{X} = 175$, σ = 8, confidence level = 99%, Z-value = 2.576, CI = 175 ± 3.434 (171.566,178.434)

Hypothesis Testing

Hypothesis testing is a statistical method used to make decisions or inferences about a population based on sample data. It involves formulating a hypothesis, collecting and analyzing sample data, and using probability to determine whether there is enough evidence to support or reject a hypothesis about the population.

Null Hypothesis (H_0): This is the hypothesis of no effect, no difference, or no association. It assumes that any observed effect in the sample data is due to random chance.

Alternative Hypothesis (H_1): This is the hypothesis that indicates a difference, effect, or association exists. It is what the researcher aims to support.

The hypotheses are structured as mutually exclusive statements; if one is true, the other must be false.

Framing Null and Alternate Hypotheses

1. Null Hypothesis(H_0): Represents the default or no-effect statement about the population. It is assumed true unless evidence suggests otherwise.
2. Alternate Hypothesis(H_1): Represents a statement that indicates a change, effect, or difference.

Example:

A researcher wants to test if the mean lifespan of a battery is 100 hours.

Null Hypothesis: H_0: $\mu = 100$

Alternative Hypothesis : H_1 : $\mu \neq 100$

Types of Alternate Hypothesis:

1. One-tailed test
2. Two-tailed test

In hypothesis testing, the choice between a one-tailed and a two-tailed test depends on the direction of the research question or hypothesis.

One-Tailed Test

Purpose: Used when the research hypothesis predicts a specific direction of effect or difference.

Types:

Right-Tailed Test: Tests if the sample mean or statistic is significantly greater than the hypothesized population parameter.

Example: Suppose a manufacturer claims a machine produces at least 500 units per day.

The hypothesis is:

Null Hypothesis: H_0: $\mu \leq 500$

Alternative Hypothesis : H_1 : $\mu > 500$

Left-Tailed Test: Tests if the sample mean or statistic is significantly less than the hypothesized population parameter.

Example: A teacher claims that the average score on a test is less than 70.

The hypothesis is:

Null Hypothesis: H_0: $\mu \geq 70$

Alternative Hypothesis : H_1 : $\mu < 70$

Two-Tailed Test

Purpose: Used when the research hypothesis does not predict a specific direction, only a difference.

Hypothesis:

Null Hypothesis (H_0): The parameter is equal to a specific value (e.g., $\mu=\mu_0$).

Alternative Hypothesis (H_1): The parameter is different from the hypothesized value (e.g.,$\mu \neq \mu_0$).

Example: A company wants to test if the mean weight of products is different from 250 grams.

The hypothesis is:

Null Hypothesis: H_0: $\mu = 250$

Alternative Hypothesis : H_1 : $\mu \neq 250$

Decision Making in Hypothesis Testing

Significance Level (α): A threshold probability, often 0.05, to determine the strength of evidence needed to reject the null hypothesis.

Test Statistic: A value calculated from the sample data used to assess the hypotheses.

Selecting the Appropriate Test

The choice of test depends on the data type, sample size, and whether assumptions like normality or equal variances are met.

Common tests include:

t-tests for comparing means

Chi-square tests for categorical data

ANOVA for comparing means across multiple groups

Z-tests for large samples or known population variance

Formula for Test Statistic:

Z-test (for large samples or known population standard deviation): Z = $\large\frac{\bar{X} - μ}{\frac{σ}{\sqrt n}}$

t-test (for small samples or unknown population standard deviation): t = $\large\frac{\bar{X} - μ}{\frac{s}{\sqrt n}}$

P-value and Decision Making

The p-value is the probability of observing the test results under the null hypothesis.

If the p-value is less than α (e.g., 0.05), we reject the null hypothesis, concluding that there is a statistically significant effect.

If the p-value is greater than α, we fail to reject the null hypothesis, indicating insufficient evidence to support a statistically significant effect.

Example:

A company tests whether their light bulbs last on average 1,000 hours (μ=1000).

Null Hypothesis: :μ=1000

Significance Level: α = 0.05

If the p-value of the test is 0.03, reject H_0 ; if it's 0.07, fail to reject H_0

Critical Value:

A point on the distribution beyond which we reject H_0 at the chosen significance level. It defines the boundaries of the acceptance and rejection regions.

Critical Value Formulas:

For a Z-test at significance level α:

Two-tailed: $Z_{critical} = \pm Z_{\alpha/2}$

Right-tailed: $Z_{critical} = Z_{\alpha}$

Left-tailed: $Z_{critical} = -Z_{\alpha}$

For a t-test: Use t-values from a t-distribution table based on degrees of freedom.

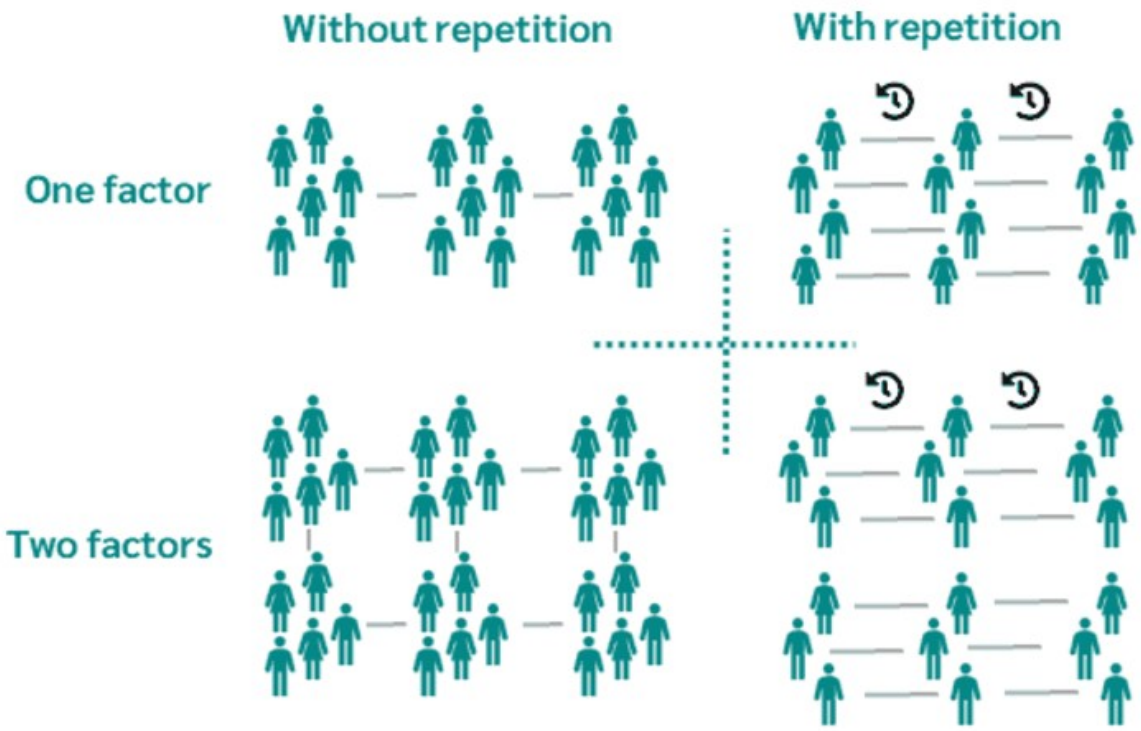
Degrees of Freedom

Degrees of freedom are the maximum number of logically independent values, which may vary in a data sample. Degrees of freedom are calculated by subtracting one from the number of items within the data sample. df = n-1

Types of Errors

In hypothesis testing, Type I and Type II errors are two possible errors that researchers can make when drawing conclusions about a population based on a sample of data. These errors are associated with the decisions made regarding the null hypothesis and the alternative hypothesis.

Type I Error: Rejecting the null hypothesis when it's actually true (false positive). Probability = α.
Type II Error: Failing to reject the null hypothesis when it's actually false (false negative). Probability = β.



Steps for Hypothesis testing

1.**Define Null and Alternative Hypothesis:**

State the null hypothesis (H_0), representing no effect, and the alternative hypothesis (H_1), suggesting an effect or difference.

2.**Choose the Significance Level (α):**

Select a significance level (α), typically 0.05, to determine the threshold for rejecting the null hypothesis.

3.**Collect and Analyze data:**

Gather relevant data through observation or experimentation. Analyze the data using appropriate statistical methods to obtain a test statistic.

4.**Calculate Test Statistic:**

The data for the tests are evaluated in this step we look for various scores based on the characteristics of data. The choice of the test statistic depends on the type of hypothesis test being conducted.

There are various hypothesis tests, each appropriate for various goal to calculate our test. This could be a Z-test, Chi-square, T-test, and so on.

Z-test: If population means and standard deviations are known. Z-statistic is commonly used.

t-test: If population standard deviations are unknown. and sample size is small than t-test statistic is more appropriate.

Chi-square test: Chi-square test is used for categorical data or for testing independence in contingency tables

F-test: F-test is often used in analysis of variance (ANOVA) to compare variances or test the equality of means across multiple groups.

5.**Comparing Test Statistic:**

There are two ways to decide where we should accept or reject the null hypothesis.

1.**Using Crtical values**

Comparing the test statistic and tabulated critical value we have,

If Test Statistic>Critical Value: Reject the null hypothesis.

If Test Statistic≤Critical Value: Fail to reject the null hypothesis.

2.**Using p-values**

If the p-value ≤ α, reject H_0.

If the p-value > α, fail to reject H_0.

6.**Interpret the results:**

Summarize the findings in the context of the research question. Example: If you reject H_0, conclude that there is significant evidence to suggest a difference or effect as stated in H_1.

Calculating test statistics

1.**One-sample proportion test:**

A one-sample proportion test is used to determine whether the proportion of a certain outcome in a single sample differs significantly from a known or hypothesized population proportion.

Purpose

It helps test if the observed proportion in a sample is significantly different from a specified population proportion,po.

Hypotheses

Null Hypothesis (H_0): The sample proportion is equal to the population proportion.

$$H_0 : p = p_0$$

Alternative Hypothesis (H_1): The sample proportion is not equal to, greater than, or less than the population proportion, depending on the test direction.

Two-tailed test: **H_1: $p \neq p_0$**

Right-tailed test: **H_1: $p > p_0$**

Left-tailed test: **H_1: $p < p_0$**

Test Statistic

The test statistic for a one-sample proportion test is based on the Z-score, calculated as:

$$Z = \large\frac{\hat{p} - p_0}{\sqrt{\frac{p_0 (1 - p_0)}{n}}}$$

where:

$\hat{p}$ = observed sample proportion

p_0 = hypothesized population proportion

n = sample size

1. **t-test:**

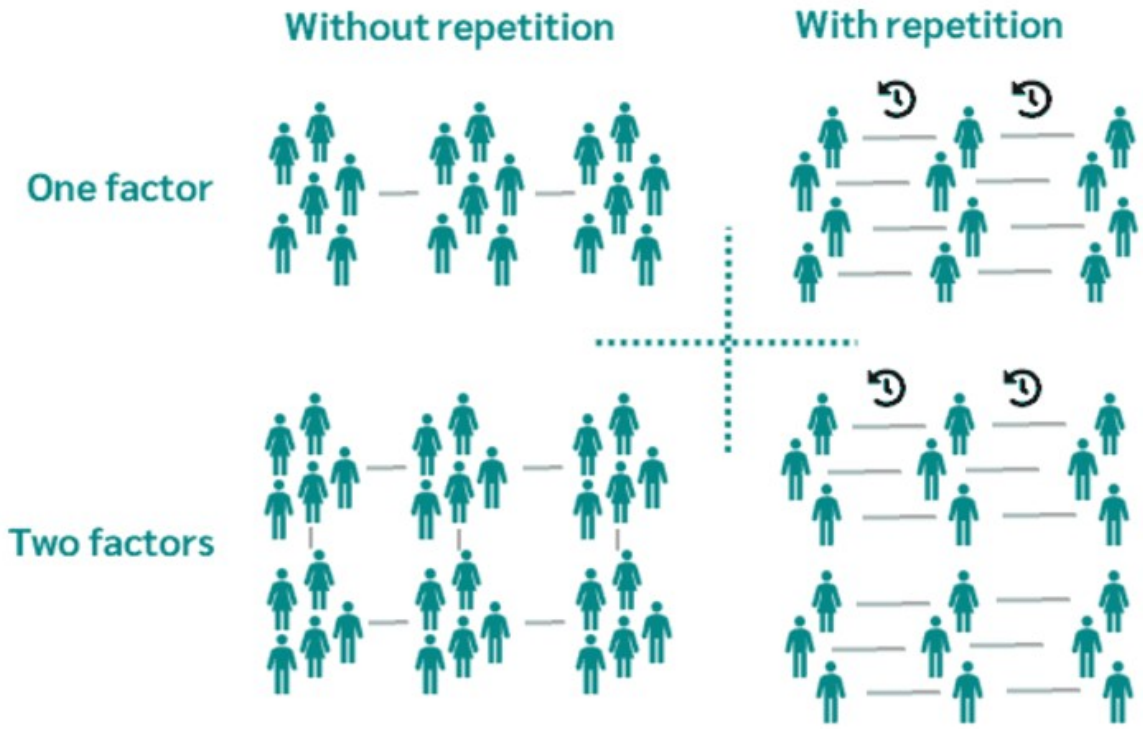
The t-test is a statistical test used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It's commonly used when the sample size is small and when the population standard deviation is unknown. There are different types of t-tests, each suited to different scenarios:

One-sample t-test: Compares the mean of a single sample to a known or hypothesized population mean. Example: Check if average height of a sample differs from national average.

Independent (two-sample) t-test: Compares the means of two independent groups to see if there is statistical evidence that the associated population means are significantly different. This

test assumes the two samples have similar variances. Example: Compare exam scores of students from two different schools.

Paired (dependent) t-test: Used when comparing the means of two related groups. Example: Analyze weight loss before and after a diet



One-sample t-test:

The one-sample t-test is used to determine if the mean of a single sample differs significantly from a known or hypothesized population mean. This test is commonly used when the population standard deviation is unknown and the sample size is relatively small.

When to use one-sample t-test

Use a one-sample t-test when:

You have one sample, and you want to test if its mean is significantly different from a known or assumed population mean. The sample data is approximately normally distributed (important when the sample size is small).

one-tailed test or two-tailed test

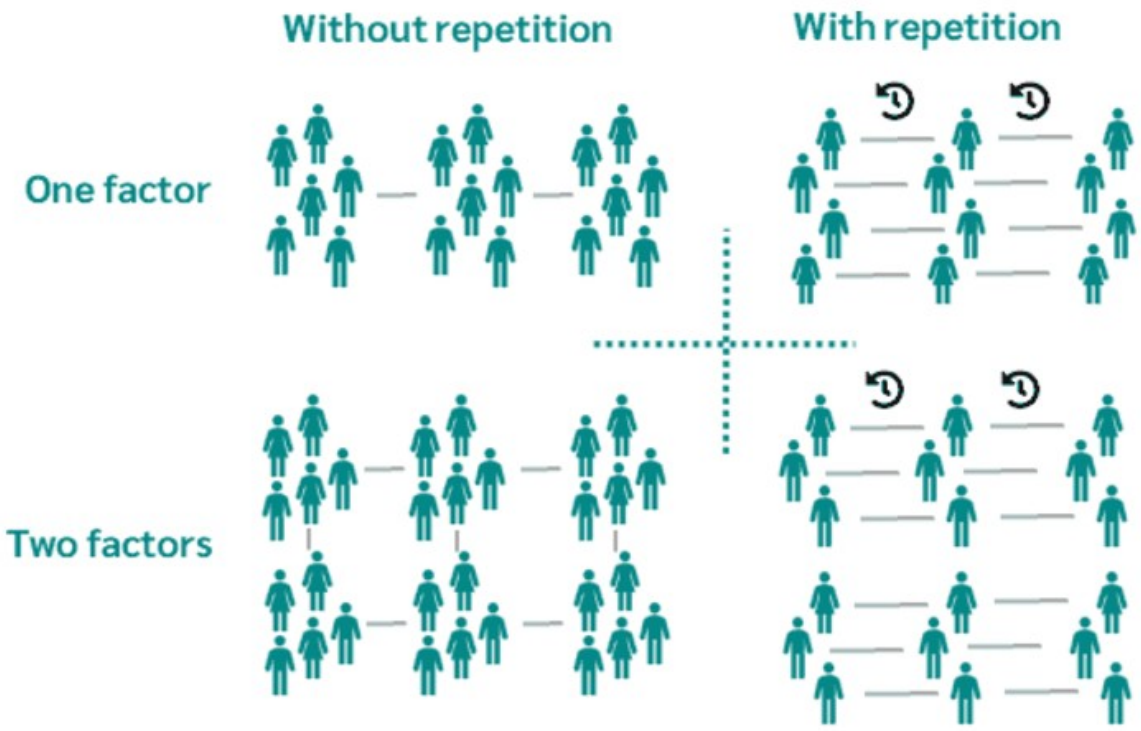
1. If you only care whether the two populations are different from one another, perform a two-tailed t test.

2. If you want to know whether one population mean is greater than or less than the other, perform a one-tailed t test.

Left-Tailed T-Test

Null Hypothesis (H0): $\bar{x} = \mu$

Alternative Hypothesis (H1): $\bar{x} < \mu$



Right-Tailed T-Test

Null Hypothesis (H0): $\bar{x} = \mu$

Alternative Hypothesis (H1): $\bar{x} > \mu$

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Two-Tailed T-Test Null Hypothesis (H0): $\bar{x} = \mu$

Alternative Hypothesis (H1): $\bar{x} \neq \mu$

Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

Key Points: H0: Sample mean equals population mean.

H1: Sample mean differs from population mean.

Hypotheses in a One-Sample t-Test

Null hypothesis (H_0): The sample mean is equal to the population mean, $\mu = \mu_0$

Alternative hypothesis (H_1): The sample mean is not equal to the population mean, $\mu \neq \mu$ 0 (two-tailed test).

Formula for one-sample t-test

$$t = \frac{\acute{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where:

$\acute{X}$ = sample mean

μ_0 = Population mean (the value you're comparing against)

s = sample standard deviation

n = sample size

Steps to Perform a One-Sample t-Test

1. Define the hypotheses: Set up H_0 and H_1.
2. Set the significance level (α): Common choices are 0.05 or 0.01.
3. Calculate the t-statistic using the formula above.
4. Determine the degrees of freedom (df): For a one-sample t-test, df=n−1.
5. Find the critical t-value from a t-table or use a statistical tool to find the p-value.
6. Draw conclusions:

If the absolute value of t is greater than the critical value (or p-value < α), reject H_0.

If not, fail to reject H_0.

Example

Suppose you have a sample of 10 students' scores on a test, with a sample mean score of 78 and a sample standard deviation of 10. You want to test if the average score is significantly different from a hypothesized mean of 75.

Hypotheses:

H_0 :μ=75

H_1 :μ≠75

Significance Level: α=0.05(A significance level of 0.05 indicates that the researcher is willing to accept a 5% chance of making a Type I error (incorrectly rejecting a true null hypothesis))

Calculate t: $t = \frac{78 - 75}{\frac{10}{\sqrt{10}}} = \frac{3}{3.16}$ = 0.949

Degrees of Freedom: df=10−1=9

Compare t-statistic to critical t-value (or calculate p-value): Using a t-table or calculator, find the critical value for α=0.05 and df=9, which is approximately ±2.262 for a two-tailed test

Conclusion:

Since |t|=0.95 is less than 2.262, you fail to reject H_0.

This means there is no statistically significant difference between the sample mean and the hypothesized population mean of 75 at the 5% level.

Two-tailed t-test exercise

1. A company claims its employees work an average of 40 hours per week. A random sample of 20 employees shows an average of 42 hours with a standard deviation of 5 hours.

2. A farmer's average wheat yield is 500 bushels per acre. A sample of 15 acres yields an average of 520 bushels with a standard deviation of 20 bushels.

3. The average height of adults in a city is 175 cm. A sample of 30 adults shows an average height of 178 cm with a standard deviation of 5 cm.

4. A manufacturing process targets a mean product weight of 200 grams. A sample of 25 products weighs an average of 205 grams with a standard deviation of 10 grams.

5. The average score of students on a standardized test is 80. A sample of 40 students scores an average of 85 with a standard deviation of 10.

left-tailed t-test

1. A company claims its employees work an average of 40 hours per week. A random sample of 20 employees shows an average of 38 hours with a standard deviation of 5 hours.

2. The average height of adults in a city is 175 cm. A sample of 30 adults shows an average height of 172 cm with a standard deviation of 5 cm.

3. A manufacturing process targets a mean product weight of 200 grams. A sample of 25 products weighs an average of 195 grams with a standard deviation of 10 grams.

4. The average score of students on a standardized test is 80. A sample of 40 students scores an average of 75 with a standard deviation of 10.

5. A farmer's average wheat yield is 500 bushels per acre. A sample of 15 acres yields an average of 480 bushels with a standard deviation of 20 bushels.

right-tailed t-test

1. A gym claims its members lose an average of 5 kg in 3 months. A random sample of 25 members shows an average weight loss of 6 kg with a standard deviation of 2 kg.

2. The average lifespan of a battery is 100 hours. A sample of 30 batteries lasts an average of 105 hours with a standard deviation of 10 hours.

3. A school's average student GPA is 3.0. A sample of 40 students has an average GPA of 3.2 with a standard deviation of 0.5.

4. The average response time for emergency services is 10 minutes. A sample of 50 responses shows an average time of 12 minutes with a standard deviation of 3 minutes.

5. A company's average sales per day is `$1000`. A sample of 20 days shows an average sales of `$1100` with a standard deviation of `$200`.

```
import numpy as np
from scipy import stats
```

```
# Sample data (weight loss in kg)
sample_data = [6, 7, 5, 4, 6, 5, 7, 5, 6, 5, 4, 6, 7, 5, 6, 5, 7, 6,
6, 5, 4, 6, 5, 6, 5]

# Known population mean
population_mean = 5

# Perform the one-sample t-test
t_statistic, p_value = stats.ttest_1samp(sample_data, population_mean)

# Output the results
print(f"T-statistic: {t_statistic}")
print(f"P-value: {p_value}")

# Determine the result
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: The sample mean is
significantly different from the population mean.")
else:
    print("Fail to reject the null hypothesis: The sample mean is not
significantly different from the population mean.")
```

```
T-statistic: 3.055050463303891
P-value: 0.0054428900133787015
Reject the null hypothesis: The sample mean is significantly different
from the population mean.
```

```
import pandas as pd
import numpy as np
from scipy.stats import ttest_1samp
df = pd.read_csv(r'C:\Users\ASUS\Downloads\diabetes.csv')
df
```

```
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI
\
0              6      148             72             35        0  33.6

1              1       85             66             29        0  26.6

2              8      183             64              0        0  23.3

3              1       89             66             23       94  28.1

4              0      137             40             35      168  43.1

..           ...      ...            ...            ...      ...   ...

763           10      101             76             48      180  32.9

764            2      122             70             27        0  36.8
```

```
765             5      121              72              23      112  26.2

766             1      126              60               0        0  30.1

767             1       93              70              31        0  30.4
```

```
     DiabetesPedigreeFunction  Age  Outcome
0                       0.627   50        1
1                       0.351   31        0
2                       0.672   32        1
3                       0.167   21        0
4                       2.288   33        1
..                        ...  ...      ...
763                     0.171   63        0
764                     0.340   27        0
765                     0.245   30        0
766                     0.349   47        1
767                     0.315   23        0

[768 rows x 9 columns]
```

```
age = df['Age']
mean_age = np.mean(age)
print(mean_age)
```

```
33.240885416666664
```

```
np.random.seed(11)
sample_size = 10
sample_age = np.random.choice(age,sample_size)
sample_age
```

```
array([41, 22, 34, 44, 28, 41, 22, 42, 28, 36], dtype=int64)
```

```
_,p_value = stats.ttest_1samp(a=sample_age,popmean=mean_age)
```

```
p_value
```

```
0.8367793985047209
```

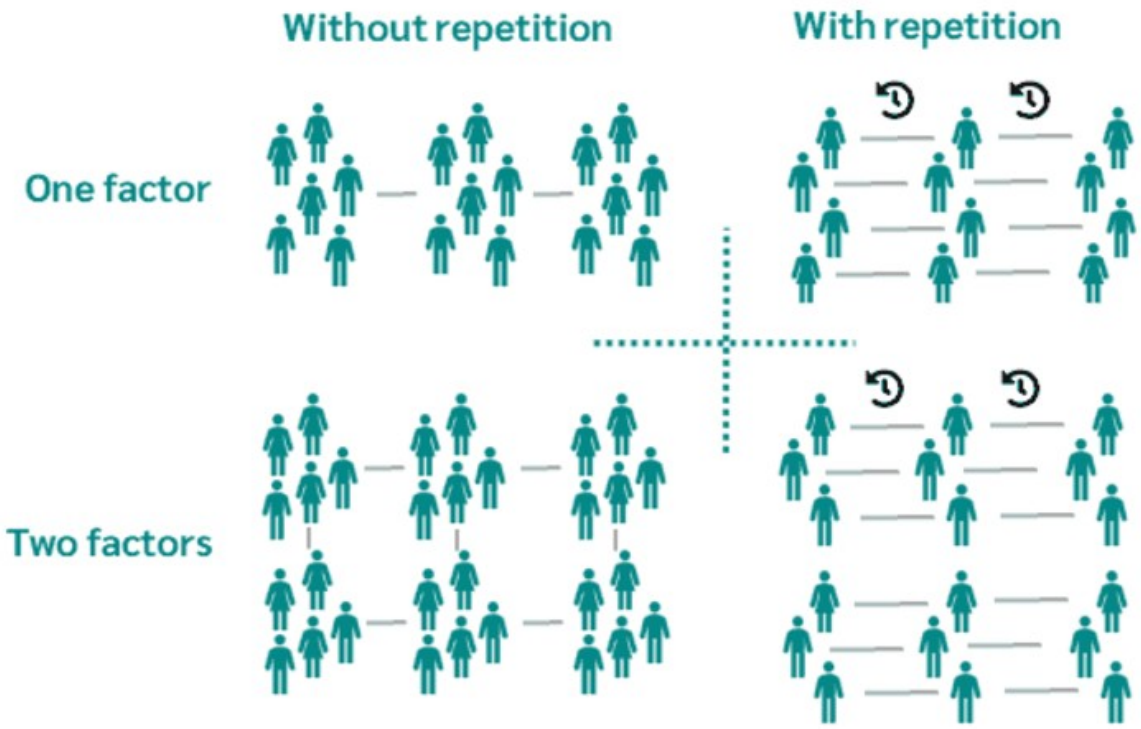
```
if p_value<0.05:
    print("reject the null hypothesis")
else:
    print("Accept the null hypothesis")
```

```
Accept the null hypothesis
```

Independent sample t-test

Also known as two-sample t-test or unpaired t-test

It is used to determine whether there is a statistically significant difference between the means of two independent groups. This test is often applied when comparing the means from two different groups, such as testing the effectiveness of two treatments, comparing scores from two different classes, or examining gender differences on a particular metric.



Key Assumptions

Continuous Dependent Variable: The dependent variable should be continuous (e.g., height, score).

Categorical Independent Variable: The independent variable should have two independent, categorical groups.

Independence of Observations: Observations in each group should be independent of each other.

No Significant Outliers: There should be no significant outliers in the data.

Normality: The dependent variable should be approximately normally distributed within each group.

Homogeneity of Variances: Both groups should have similar variances in the dependent variable (equal spread).

Steps to Perform an Independent Samples t-test

1.State Hypotheses: Null Hypothesis (H_0): There is no difference in means between the two groups ($\mu 1 = \mu 2$).

Alternative Hypothesis (H_1): There is a difference in means ($\mu 1 \neq \mu 2$) for a two-tailed test, or that one mean is greater than or less than the other for a one-tailed test.

2.Calculate the Test Statistic:

The test statistic for the independent t-test can be calculated as follows:

$$t=\large\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

where:

$\acute{X}_1 and \acute{X}_2$ are the sample means

$s_1^2 and s_2^2$ are the sample variances

$n_1 and n_2$ are the sample sizes

1. Degrees of freedom: df = $n_1 + n_2 - 2$

4.Determine the Critical Value or p-Value:

Compare the calculated t-statistic to the critical t-value from a t-distribution table, or calculate the p-value.

If |t| is greater than the critical value or p is less than the significance level (typically 0.05), reject the null hypothesis.

5.Interpret the Results:

A significant result ($p < 0.05$) suggests that the group means are significantly different.

A non-significant result suggests there is not enough evidence to conclude a difference between the group means.

Example

Suppose a researcher wants to test whether there is a difference in test scores between two independent groups of students who used different study methods.

Group 1(Method A): $\acute{X}_1 = 78, s_1^2 = 64, n_1 = 25$

Group 2(Method B): $\acute{X}_2 = 85, s_2^2 = 81, n_2 = 30$

Hypotheses:

Null Hypothesis (H_0): μ_1 = μ_2

Alternative Hypothesis (H_1): μ_1 ≠ μ_2

$$t = \frac{78 - 85}{\sqrt{\frac{64}{25} + \frac{81}{30}}} = -3.05$$

Degrees of freedom

Using the formula for equal variances, df=25+30−2=53.

Find the critical value or p-value

For df = 53 and α =0.05 for a two-tailed test, the critical t-value is approximately ± 2.0057

Using p-value calculator, the p-value is approximately 0.0036

Since our calculated t-value of -3.05 is less than -2.006, it lies in the critical region.

Interpret the result

since |t| = -3.05 is greater than the critical value of -2.006, and the p-value is less than 0.05, we reject the null hypothesis.

This result suggests there is a statistically significant difference in the mean test scores between the two study methods, indicating that one method may have had a greater effect on the test scores than the other.

Exercise

1.Two different tutoring programs were used to help students prepare for an exam. To test if there's a significant difference in scores between the two programs, scores from two independent groups of students are collected.

Group 1 (Program A): 62, 75, 68, 80, 72, 78, 66, 70, 76, 74

Group 2 (Program B): 85, 81, 89, 83, 87, 78, 88, 90, 86, 84

Calculate the mean and variance for each group.(variance = $\frac{\sum \left(X_i - \dot{X}\right)^2}{n-1}$)

Using the calculated values, test at the 0.05 significance level if there is a statistically significant difference in the average scores between the two programs.

2.A nutritionist wants to determine if two different diet plans have an effect on weight loss. Weight loss data (in pounds) for two groups of participants who followed each diet plan for six weeks is given below:

Group 1 (Diet A): 8, 10, 12, 11, 9, 10, 14, 13, 7

Group 2 (Diet B): 15, 14, 17, 16, 15, 18, 12, 13, 14

Calculate the mean and variance for each group.

Test at a 0.01 significance level if there is a significant difference in weight loss between the two diets.

3.A researcher is interested in whether exercise influences the amount of sleep people get. Two independent groups of people are studied: one group that exercises regularly and one that doesn't. The average hours of sleep per night is recorded. Test at the 0.05 significance level if there is a difference in sleep hours between those who exercise and those who do not.

Group 1 (Exercisers):

Mean ($\acute{X}_1$) = 7.5 hours

Variance (s_1^2) = 9

Sample size (n_1) = 25

Group 2 (Non-exercisers):

Mean ($\acute{X}_2$) = 6.8 hours

Variance (s_2^2) = 12

Sample size (n_2) = 20

```
import pandas as pd
import scipy.stats as stats

# Example data for two groups
group_a = [25, 30, 35, 40, 45, 50, 55, 60, 65]  # Group A
group_b = [22, 28, 34, 40, 46, 52, 58, 64, 70]  # Group B

# Create a DataFrame to organize the data
df = pd.DataFrame({'Group A': group_a, 'Group B': group_b})
print(df)

# Calculate and display the mean for each group
print("Mean of Group A:", df['Group A'].mean())
print("Mean of Group B:", df['Group B'].mean())

# Calculate and display the variance for each group
print("Variance of Group A:", df['Group A'].var())
print("Variance of Group B:", df['Group B'].var())

# Perform the independent t-test on the two groups
t_statistic, p_value = stats.ttest_ind(group_a, group_b)

# Define significance level (alpha)
alpha = 0.05

# Make a decision based on the p-value
if p_value < alpha:
    print("Reject the null hypothesis: Means are significantly
different.")
else:
```

```
    print("Fail to reject the null hypothesis: Means are not
significantly different.")
```

```
   Group A  Group B
0       25       22
1       30       28
2       35       34
3       40       40
4       45       46
5       50       52
6       55       58
7       60       64
8       65       70
Mean of Group A: 45.0
Mean of Group B: 46.0
Variance of Group A: 187.5
Variance of Group B: 270.0
Fail to reject the null hypothesis: Means are not significantly
different.
```

```
import pandas as pd
import numpy as np
from scipy.stats import ttest_ind

df = pd.read_csv(r'C:\Users\ASUS\Downloads\diabetes.csv')

group1 = df[df['Outcome'] == 0]['Glucose']  # Non-diabetic group
group2 = df[df['Outcome'] == 1]['Glucose']  # Diabetic group

# Perform the two-sample t-test
t_stat, p_val = ttest_ind(group1, group2, equal_var=True)  # Use
equal_var=False if variances differ

print("t-statistic:", t_stat)
print("p-value:", p_val)
```

```
t-statistic: -14.600060005973894
p-value: 8.935431645289913e-43
```

```
import numpy as np
from scipy.stats import ttest_ind

# Define the two groups for Glucose levels
group1 = df[df['Outcome'] == 0]['Glucose']  # Non-diabetic group
group2 = df[df['Outcome'] == 1]['Glucose']  # Diabetic group

# Calculate and display variances
var_group1 = np.var(group1, ddof=1)  # Using ddof=1 for sample
variance
var_group2 = np.var(group2, ddof=1)  # Using ddof=1 for sample
variance
```

```
print("Variance of Glucose for non-diabetic group:", var_group1)
print("Variance of Glucose for diabetic group:", var_group2)

# Check if variances are approximately equal
alpha = 0.05
if abs(var_group1 - var_group2) / max(var_group1, var_group2) < 0.1:
# Allowing a 10% tolerance
    print("Variances are approximately equal; setting
equal_var=True.")
    equal_var = True
else:
    print("Variances are not approximately equal; setting
equal_var=False.")
    equal_var = False

# Perform the two-sample t-test with the appropriate equal_var setting
t_stat, p_val = ttest_ind(group1, group2, equal_var=equal_var)
print("t-statistic:", t_stat)
print("p-value:", p_val)
if p_val < alpha:
    print("Since p-value < alpha, we reject the null hypothesis: there
is a significant difference between the groups.")
else:
    print("Since p-value >= alpha, we fail to reject the null
hypothesis: there is no significant difference between the groups.")
```

```
Variance of Glucose for non-diabetic group: 683.3623246492986
Variance of Glucose for diabetic group: 1020.1394572083399
Variances are not approximately equal; setting equal_var=False.
t-statistic: -13.751537067396413
p-value: 2.6441613495403223e-36
Since p-value < alpha, we reject the null hypothesis: there is a
significant difference between the groups.
```

```
group1 = df[df['Outcome'] == 0]['Age']
group2 = df[df['Outcome'] == 1]['Age']
print("Non-diabetic group length:", len(group1))
print("Diabetic group length:", len(group2))
print("Non-diabetic group variance (Age):", np.var(group1, ddof=1))
print("Diabetic group variance (Age):", np.var(group2, ddof=1))
t_stat, p_val = ttest_ind(group1, group2)
print("T-Statistic:", t_stat)
print("P-Value:", p_val)
alpha = 0.05
if p_val < alpha:
    print("Reject Null Hypothesis: There is a significant difference
in mean age between diabetic and non-diabetic individuals.")
else:
    print("Fail to Reject Null Hypothesis: There is no significant
```

```
difference in mean age between diabetic and non-diabetic
individuals.")

Non-diabetic group length: 500
Diabetic group length: 268
Non-diabetic group variance (Age): 136.13416833667347
Diabetic group variance (Age): 120.30258818268211
T-Statistic: -6.792688071649956
P-Value: 2.2099754606654358e-11
Reject Null Hypothesis: There is a significant difference in mean age
between diabetic and non-diabetic individuals.
```

Paired sample t-test

A paired sample t-test (also known as the dependent t-test) is used to determine if there is a statistically significant difference between the means of two related groups. These groups are "paired" because the data points in the two groups are linked or matched in some way (e.g., before and after measurements for the same individuals, or matched subjects in an experiment)

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Assumptions of a Paired Sample T-test:

The data are paired: Each observation in the first group has a corresponding observation in the second group. The differences are normally distributed: The differences between the paired observations should be approximately normally distributed. This can be checked with a histogram or normality tests like the Shapiro-Wilk test.

Hypotheses:

Null hypothesis (H_0): The mean difference between the two groups is zero (no effect).

Alternative hypothesis (H_1): The mean difference between the two groups is not zero (there is an effect).

Formula: The test statistic for a paired t-test is calculated as:

$$t = \frac{\acute{d}}{\frac{s_d}{\sqrt{n}}}$$

where:

$\acute{d}$ = mean of the differences between paired observations

s_d = standard deviation of the differences between paired observations

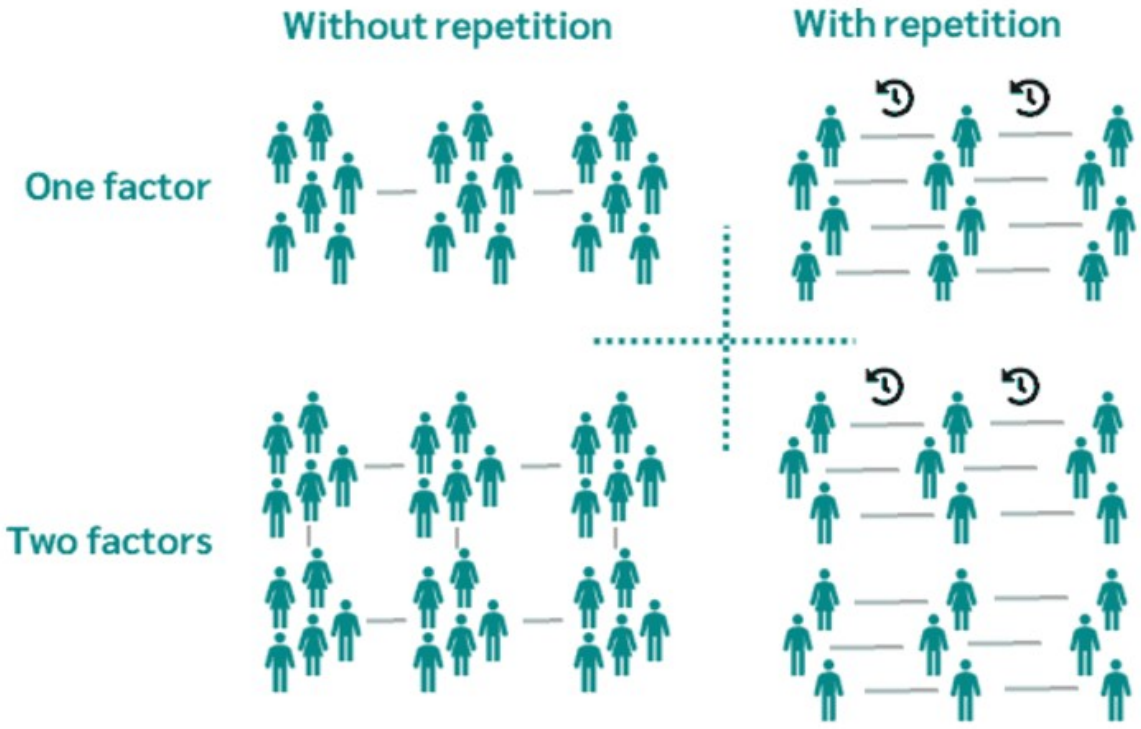
n = number of pairs

Steps for performing the test:

1. Calculate the differences between the paired observations.
2. Calculate the mean and standard deviation of the differences.
3. Compute the t-statistic using the formula.
4. Determine the degrees of freedom: df=n−1, where n is the number of pairs.
5. Find the critical t-value from the t-distribution table based on your chosen significance level (usually 0.05) and degrees of freedom.
6. Compare the calculated t-value with the critical t-value to determine whether to reject the null hypothesis.

Example:

Suppose you want to test if a new study method improves students' scores. You have test scores for the same group of students before and after the study method.



You calculate the differences (After - Before) and then find the mean and standard deviation of these differences. Then, apply the t-test formula to determine whether the mean difference is significantly different from zero.

Mean of the differences:

$$\acute{d} = \frac{\sum Difference}{n} = \frac{10+2+5+5+3}{5} = \frac{25}{5} = 5$$

Standard Deviation of the differences:

To find the standard deviation, we first calculate the squared deviations from the mean difference:

(10-5)^2 = 25

(2-5)^2 = 9

(5-5)^2 = 0

(5-5)^2 = 0

(3-5)^2 = 4

sum of squared deviations = 25+9+0+0+4 = 38

$$s_d = \sqrt{\frac{\sum \left(Difference - \acute{d}\right)^2}{n-1}} = \sqrt{\frac{38}{5-1}} = \sqrt{\frac{38}{4}} = \sqrt{9}.5 \approx 3.08$$

Calculate the t-statistic

$$t = \frac{\acute{d}}{\frac{s_d}{\sqrt{n}}} = \frac{5}{\frac{3.08}{\sqrt{5}}} = \frac{5}{1.377} = 3.6299$$

Degrees of Freedom

For a paired t-test, df = n-1 = 5-1 = 4

Determine the critical t-value

Assuming a significance level of α=0.05 for a two-tailed test, we look up the critical t-value for 4 degrees of freedom. From the t-distribution table, the critical value is approximately 2.776.

Compare the calculated t-value with the critical t-value

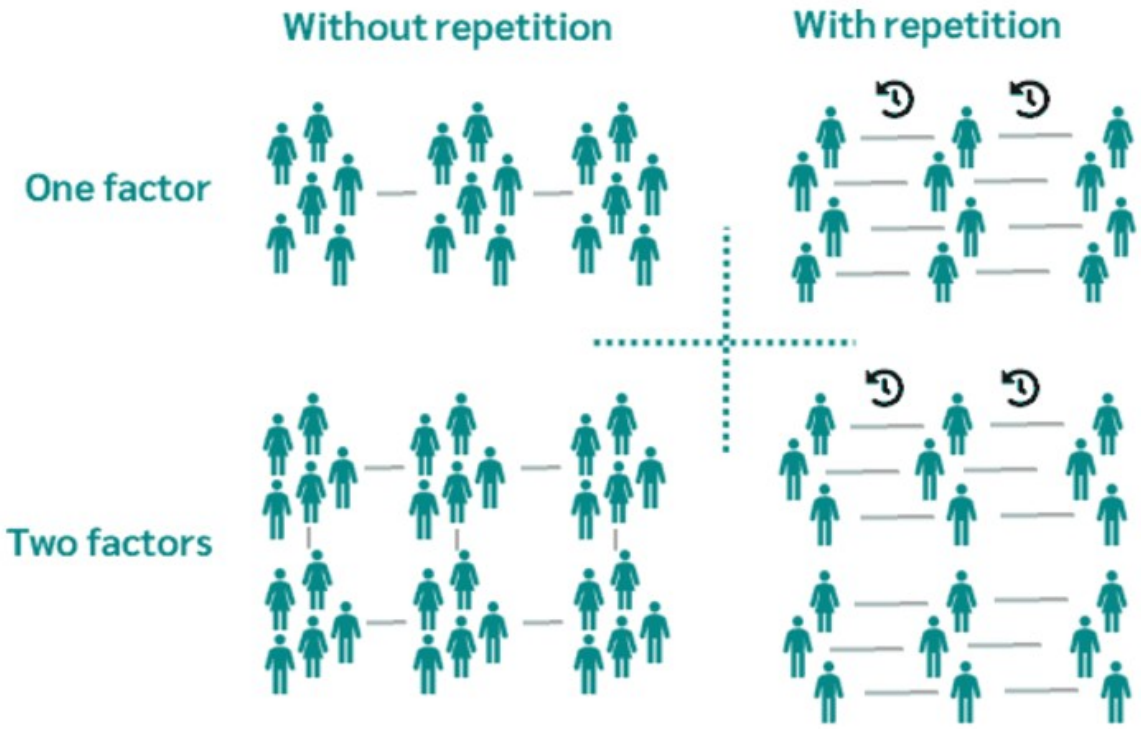
Since our calculated t=3.63 is greater than the critical value $t_{crit} = 2.776$, we reject the null hypothesis.

Conclusion

There is enough evidence to suggest a statistically significant difference between the "Before" and "After" scores, implying that the study method had an effect on scores.

Exercise

1. A doctor wants to check if a new medication helps reduce blood pressure. She measures the blood pressure of 6 patients before and after a 2-week treatment.Calculate the t-statistic and determine whether the change in blood pressure is significant at the 0.05 level.



2. A fitness trainer wants to assess the effectiveness of a 4-week weight loss program. He records the weights of 8 clients before and after the program.Calculate the t-statistic and use a 0.05 significance level to test if the weight loss program is effective.

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

3.A psychologist is testing the effect of a memory training course. She administers a memory test to 7 students before and after they complete the course.Compute the t-statistic and check if

there is a significant improvement at the 0.05 level.

Steps for Performing a Chi-Square Test

1. **State the hypotheses:**
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies:**
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic:**
   - Use the formula above.
4. **Determine the degrees of freedom (df):**
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value:**
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision:**
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

```
import numpy as np
from scipy import stats

before = np.array([78, 82, 85, 90, 76, 80, 78, 74])
after = np.array([85, 84, 88, 92, 79, 82, 81, 78])

mean_diff = np.mean(after - before)
t_stat, p_value = stats.ttest_1samp(after - before, 0)

print(f"Mean Difference: {mean_diff}")
print(f"t-statistic: {t_stat}")
print(f"p-value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant
effect.")
else:
```

```
    print("Fail to reject the null hypothesis: No significant
effect.")
```

```
Mean Difference: 3.25
t-statistic: 5.507570547286102
p-value: 0.0008993006369287485
Reject the null hypothesis: There is a significant effect.
```

```
import numpy as np
from scipy import stats

before = np.array([65, 70, 72, 68, 60, 75, 70, 67])
after = np.array([72, 75, 78, 72, 65, 80, 74, 70])

mean_diff = np.mean(after - before)
t_stat, p_value = stats.ttest_1samp(after - before, 0)

print(f"Mean Difference: {mean_diff}")
print(f"t-statistic: {t_stat}")
print(f"p-value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant
effect.")
else:
    print("Fail to reject the null hypothesis: No significant
effect.")
```

```
Mean Difference: 4.875
t-statistic: 11.062518264157921
p-value: 1.0954691572370328e-05
Reject the null hypothesis: There is a significant effect.
```

```
import numpy as np
from scipy import stats

before = np.array([78, 82, 79, 81, 77, 80, 79, 78])
after = np.array([79, 83, 80, 82, 78, 81, 80, 79])

mean_diff = np.mean(after - before)
t_stat, p_value = stats.ttest_1samp(after - before, 0)

print(f"Mean Difference: {mean_diff}")
print(f"t-statistic: {t_stat}")
print(f"p-value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant
effect.")
```

```
else:
    print("Fail to reject the null hypothesis: No significant
effect.")
```

```
Mean Difference: 1.0
t-statistic: inf
p-value: 0.0
Reject the null hypothesis: There is a significant effect.
```

```
import numpy as np
from scipy import stats

before = np.array([78, 82, 79, 81, 77, 80, 79, 78])
after = np.array([78, 82, 79, 81, 77, 80, 79, 78])

mean_diff = np.mean(after - before)
t_stat, p_value = stats.ttest_1samp(after - before, 0)

print(f"Mean Difference: {mean_diff}")
print(f"t-statistic: {t_stat}")
print(f"p-value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant
effect.")
else:
    print("Fail to reject the null hypothesis: No significant
effect.")
```

```
Mean Difference: 0.0
t-statistic: nan
p-value: nan
Fail to reject the null hypothesis: No significant effect.
```

```
import numpy as np
from scipy import stats

# Example BMI values before and after treatment
before_bmi = np.array([30.5, 32.1, 28.7, 29.9, 31.4, 30.2, 29.1, 28.4,
32.5, 30.8])
after_bmi = np.array([30.2, 32.0, 28.5, 29.8, 31.2, 30.1, 29.0, 28.3,
32.4, 30.7])

# Calculate the mean difference and perform paired sample t-test
mean_diff = np.mean(after_bmi - before_bmi)
t_stat, p_value = stats.ttest_1samp(after_bmi - before_bmi, 0)

print(f"Mean Difference: {mean_diff}")
print(f"t-statistic: {t_stat}")
```

```
print(f"p-value: {p_value}")

# Set significance level
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant
effect.")
else:
    print("Fail to reject the null hypothesis: No significant
effect.")
```

```
Mean Difference: -0.13999999999999985
t-statistic: -6.331738236133021
p-value: 0.00013577665393919158
Reject the null hypothesis: There is a significant effect.
```

3. Z-test

A z-test is a statistical test used to determine if there is a significant difference between the sample mean and the population mean (or between two sample means) when the population variance is known, or the sample size is large enough to assume that the sample variance is a good estimate of the population variance.

Assumptions:

The sample is randomly selected from the population.

The sample size is sufficiently large (usually n ≥ 30).

The population is normally distributed or approximately normally distributed.

Types of Z-test

1. One-Sample Z-test: Compares the mean of a sample to a known population mean when the population variance is known. Example: Test if the average lifespan of a sample of light bulbs differs from the factory's known population average of 1,200 hours.

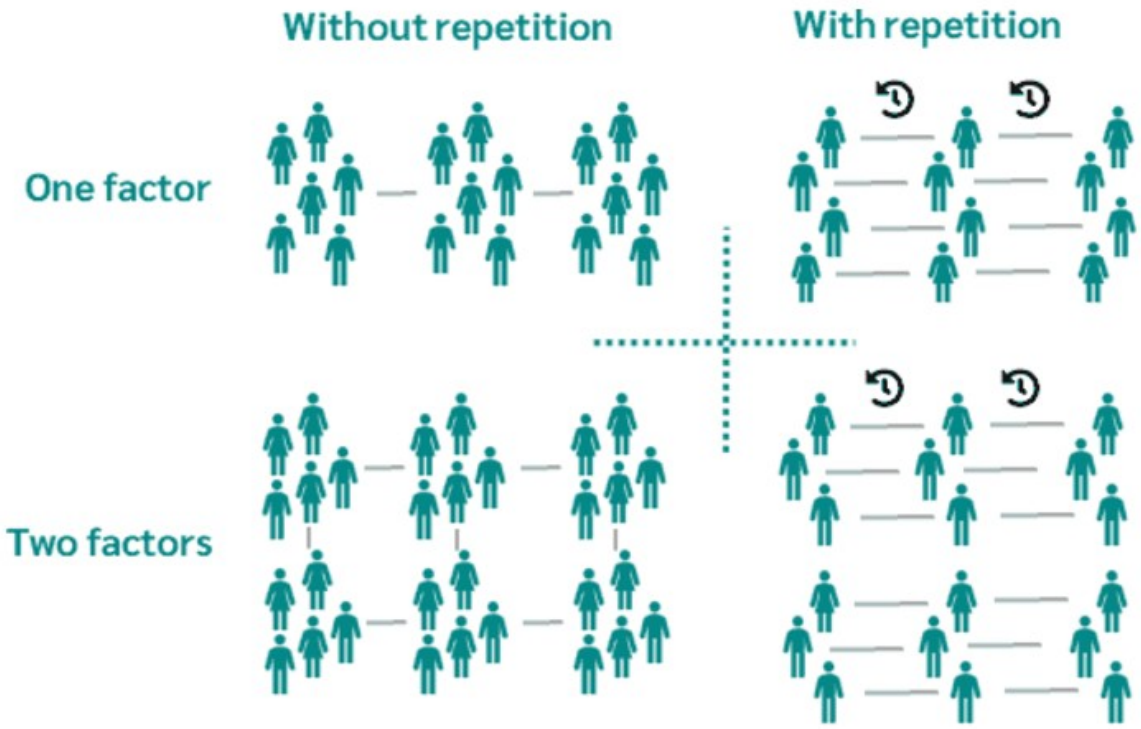
2. Two-Sample Z-test: Compares the means of two independent samples to see if there is statistical evidence that the population means are significantly different. Example: Compare the average salary of employees in two different departments to see if one department's mean salary differs from the other.

3. Paired Z-test: Compares the means of two related groups, often for before-and-after measurements. Example: Analyze the change in blood pressure before and after medication in a group of patients to see if the treatment has a significant effect.

One-sample Z-test

The One-Sample Z-test can be either left-tailed, right-tailed, or two-tailed depending on the direction of the hypothesis you're testing.

1. Left-tailed Z-test: In this test, our region of rejection is located to the extreme left of the distribution. Here our null hypothesis is that the claimed value is less than or equal to the mean population value.



1. Right-tailed Z-test: In this test, our region of rejection is located to the extreme right of the distribution. Here our null hypothesis is that the claimed value is less than or equal to the mean population value.

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

1. Two-tailed Z-test: In this test, our region of rejection is located to both extremes of the distribution. Here our null hypothesis is that the claimed value is equal to the mean population value.

Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

Formula for a One-Sample Z-test: For a sample mean $\acute{x}$, population mean μ, population standard deviation σ, and sample size n:

$$Z=\large\frac{\bar x - μ}{\frac{\sigma}{\sqrt n}}$$

where:

$\acute{x}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

Steps to Perform a Z-test:

1.**State the hypotheses**:

Null hypothesis H_0: The sample mean is equal to the population mean. Alternative hypothesis H_1: The sample mean is not equal to the population mean (two-tailed) or is greater/less than the population mean (one-tailed).

2.**Select the significance level** (usually $\alpha=0.05$).

3.**Compute the Z-statistic using the formula.**

4.**Determine the critical value** based on the significance level (you can look up the Z-value in a Z-table for a given significance level).

5.**Compare the Z-statistic to the critical value:**

If the Z-statistic is beyond the critical value, reject the null hypothesis.

If the Z-statistic is within the critical value range, do not reject the null hypothesis.

Example:

You are testing whether the average height of a population of adults is 65 inches. A sample of 50 adults has a mean height of 66 inches. The population standard deviation is known to be 4 inches. Test at the 5% significance level.

Null Hypothesis (H_0): μ_1 = 65

Alternative Hypothesis (H_1): μ_1 ≠ 65

Z-statistic: $Z=\frac{66-65}{\frac{4}{\sqrt{50}}}=\frac{1}{0.5657}=1.77$

Determine the critical Z-value for a two-tailed test at $\alpha=0.05$, which is approximately ±1.96.

Since 1.77 is less than 1.96, we fail to reject the null hypothesis and conclude that the sample mean is not significantly different from 65 inches.

Exercise

Left-tailed Z-test

1.A machine produces screws with an average length of 5 cm. A sample of 40 screws has an average length of 4.8 cm, and the population standard deviation is 0.2 cm. Test if the machine is producing screws shorter than the desired length at the 1% significance level.

2.The average lifespan of a type of light bulb is advertised as 1,000 hours. A sample of 50 bulbs has an average lifespan of 980 hours, with a standard deviation of 30 hours. Test if the actual lifespan is less than advertised at the 5% significance level.

3.A factory claims that its packets of chips weigh 500 grams. A sample of 25 packets shows a mean weight of 495 grams with a standard deviation of 10 grams. Test if the packets weigh less than claimed at the 10% significance level.

4.A company states that the average delivery time for its service is 2 days. A sample of 60 deliveries shows a mean of 1.8 days with a standard deviation of 0.4 days. Test if the average delivery time is shorter than stated at the 5% significance level.

5.A battery manufacturer claims its batteries last 100 hours. A sample of 36 batteries shows a mean lifespan of 98 hours, with a standard deviation of 5 hours. Test if the batteries last less than the claimed time at the 1% significance level.

Right-tailed Z-test

1.The average score on a standardized test is 75. A sample of 30 students scored an average of 78, with a population standard deviation of 8. Test if the sample performed better than average at the 5% significance level.

2.A factory claims its machines produce 100 items per hour on average. A sample of 50 hours shows an average production rate of 102 items, with a standard deviation of 3 items. Test if the machines are producing more than claimed at the 1% significance level.

3.A company states that the average monthly revenue is $50,000. A sample of 20 months shows an average revenue of $52,000 with a standard deviation of $5,000. Test if the revenue is higher than claimed at the 10% significance level.

4.A university claims its students spend an average of 20 hours per week studying. A sample of 35 students reports an average of 22 hours with a standard deviation of 4 hours. Test if students study more than claimed at the 5% significance level.

5.A health club claims that its members lose an average of 2 kg in the first month. A sample of 25 members reports an average loss of 2.5 kg with a standard deviation of 0.8 kg. Test if the average weight loss is greater than claimed at the 1% significance level.

Two-tailed Z-test

1.The average lifespan of a brand of tires is 40,000 miles. A sample of 50 tires shows an average lifespan of 39,000 miles, with a standard deviation of 2,500 miles. Test if the actual lifespan differs from the claimed value at the 5% significance level.

2.A hospital reports an average wait time of 15 minutes. A sample of 60 patients shows a mean wait time of 17 minutes, with a standard deviation of 3 minutes. Test if the actual wait time differs from the reported time at the 1% significance level.

3.The average temperature in a city during summer is 30°C. A sample of 100 days has an average temperature of 29.5°C with a standard deviation of 1.5°C. Test if the average temperature is different from the expected value at the 5% significance level.

4.A study claims the average height of adult men is 175 cm. A sample of 40 men has a mean height of 178 cm with a standard deviation of 6 cm. Test if the average height differs from the claimed value at the 10% significance level.

5.A company claims that its employees work an average of 40 hours per week. A sample of 45 employees shows a mean of 38 hours with a standard deviation of 5 hours. Test if the average working hours differ from the claim at the 5% significance level.

```
'''A company claims that their new product’s battery life is at least
12 hours.
You test a sample of 50 units, and the average battery life is 11.5
hours with a standard deviation of 1.2 hours.
You want to test if the battery life is significantly less than the
claimed 12 hours.'''
from scipy import stats
import numpy as np

x = 11.5  # Sample mean
mu = 12  # Population mean (claimed)
std = 1.2
n = 50

# Calculate Z-statistic
se = std / np.sqrt(n)  # Standard error
z = (x - mu) / se  # Z-statistic

# Calculate p-value (Left-Tailed)
p_left = stats.norm.cdf(z)

print("Left-Tailed One-Sample Z-Test")
print("Z-Statistic:", z)
print("P-Value:", p_left)

alpha = 0.05
if p_left < alpha:
    print("Reject H0: The battery life is significantly less than 12
hours.")
else:
    print("Fail to reject H0: No significant evidence that the battery
life is less than 12 hours.")
```

```
Left-Tailed One-Sample Z-Test
Z-Statistic: -2.946278254943948
P-Value: 0.0016081146550637245
Reject H0: The battery life is significantly less than 12 hours.
```

```
from statsmodels.stats.weightstats import ztest

data = [4.8, 4.9, 5.1, 4.7, 4.6]
pop_mean = 5
alpha = 0.05

z_stat, p_value = ztest(data, value=pop_mean, alternative='smaller')

if p_value < alpha:
    decision = "Reject H0"
else:
    decision = "Fail to reject H0"
```

```
print("Left-Tailed One-Sample Z-Test")
print(f"Z-Statistic: {z_stat}")
print(f"P-Value: {p_value}")
print(f"Decision: {decision}")
```

```
Left-Tailed One-Sample Z-Test
Z-Statistic: -2.092457497388744
P-Value: 0.018198805001832602
Decision: Reject H0
```

```
'''A factory claims that their machine produces bottles that weigh on
average 500 grams.
You test a sample of 100 bottles and find that the average weight is
505 grams with a standard deviation of 8 grams.
You want to test if the bottles are significantly heavier than the
claimed 500 grams.'''
from scipy import stats
import numpy as np


x = 505
mu = 500
std = 8
n = 100

se = std / np.sqrt(n)
z = (x - mu) / se

# Calculate p-value (Right-Tailed)
p_right = 1 - stats.norm.cdf(z)

print("\nRight-Tailed One-Sample Z-Test")
print("Z-Statistic:", z)
print("P-Value:", p_right)

alpha = 0.05
if p_right < alpha:
    print("Reject H0: The bottle weight is significantly greater than
500 grams.")
else:
    print("Fail to reject H0: No significant evidence that the bottle
weight is greater than 500 grams.")
```

```
Right-Tailed One-Sample Z-Test
Z-Statistic: 6.25
P-Value: 2.0522639143649712e-10
Reject H0: The bottle weight is significantly greater than 500 grams.
```

```
x = 178
mu = 175
```

```
std = 5
n = 36

se = std / np.sqrt(n)
z = (x - mu) / se

p_right = 1 - stats.norm.cdf(z)

print("z-statistic:", z)
print("p-value (right-tailed):", p_right)

alpha = 0.05
if p_right < alpha:
    print("Reject H0 (right-tailed): Average height is greater than
175 cm")
else:
    print("Fail to reject H0 (right-tailed): Average height is not
greater than 175 cm")
```

```
z-statistic: 3.5999999999999996
p-value (right-tailed): 0.00015910859015755285
Reject H0 (right-tailed): Average height is greater than 175 cm
```

```
data = [78, 80, 76, 79, 81]
pop_mean = 75

# Perform Right-Tailed Z-Test
z_stat, p_value = ztest(data, value=pop_mean, alternative='larger')  #
'larger' for right-tailed
print("Right-Tailed One-Sample Z-Test")
print(f"Z-Statistic: {z_stat}")
print(f"P-Value: {p_value}")
```

```
Right-Tailed One-Sample Z-Test
Z-Statistic: 4.41741027226513
P-Value: 4.994525793448572e-06
```

```
'''A school claims that their students have an average score of 75 on
a standardized test.
You test a sample of 30 students, and their average score is 77 with a
standard deviation of 10.
You want to check if the average score is significantly different from
75.'''

from scipy import stats
import numpy as np

x = 77
mu = 75
std = 10
n = 30
```

```
se = std / np.sqrt(n)
z = (x - mu) / se

# Calculate p-value (Two-Tailed)
p_two = 2 * (1 - stats.norm.cdf(abs(z)))

print("\nTwo-Tailed One-Sample Z-Test")
print("Z-Statistic:", z)
print("P-Value:", p_two)

alpha = 0.05
if p_two < alpha:
    print("Reject H0: The average score is significantly different
from 75.")
else:
    print("Fail to reject H0: No significant evidence that the average
score is different from 75.")
```

```
Two-Tailed One-Sample Z-Test
Z-Statistic: 1.0954451150103321
P-Value: 0.27332167829229803
Fail to reject H0: No significant evidence that the average score is
different from 75.
```

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from scipy import stats
from statsmodels.stats.weightstats import ztest
#from scipy.stats import ztest

titanic_df= sns.load_dataset('titanic')

titanic_df1 = titanic_df.dropna(subset=['age']).sample(100,
replace=True)

pop_mean = 30

x = titanic_df1['age'].mean()
print("Sample Mean:", x)

se = titanic_df1['age'].std()
print("Sample Standard Deviation:", se)
zstat, p_val = ztest(titanic_df1['age'], value=pop_mean)
print("z-statistic:", zstat)
print("p-value:", p_val)

alpha = 0.05
```

```
if p_val < alpha:

    print("Reject Null Hypothesis , sample doesn't represent
population")
else:

    print("Fail to Reject Null Hypothesis, sample represent
population")
```

```
Sample Mean: 29.5575
Sample Standard Deviation: 15.026065526345219
z-statistic: -0.2944882672208259
p-value: 0.7683848275505132
Fail to Reject Null Hypothesis, sample represent population
```

Two-sample Z-test

A Two-Sample Z-Test is used to compare the means of two independent samples to determine if there is a significant difference between the two population means. This test assumes that both populations are normally distributed with known population variances (or sufficiently large sample sizes).

Hypothesis Setup for Two-Sample Z-Test:

Null Hypothesis (H_0): The two population means are equal ($\mu_1 = \mu_2$).

Alternative Hypothesis (H_1): The two population means are not equal ($\mu_1 \neq \mu_2$) in a two-tailed test. If you are doing a one-tailed test, you might test if one mean is greater than or less than the other.

Formula for the Two-Sample Z-Test: The test statistic for a two-sample Z-test is calculated as:

$$Z=\large\frac{\bar X_1 - \bar X_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

where:

$\acute{X}_1, \acute{X}_2$ = sample means of the two samples

σ_1, σ_2 = population standard deviations of the two samples

n_1, n_2 = sample sizes of the two groups

Key Assumptions for Two-Sample Z-Test:

Independence:

The two samples should be independent of each other. For example, the measurements in one group should not affect those in the other group.

Normality:

Both populations should be normally distributed. If the sample sizes are large enough (usually n > 30), the Central Limit Theorem (CLT) allows us to approximate the sampling distribution of the mean as normal, even if the populations themselves are not perfectly normal.

Known Population Variances:

The population standard deviations ($\sigma 1, \sigma 2$) should be known. If they are unknown, a two-sample t-test is generally more appropriate.

Large Sample Size:

While not a strict requirement, large sample sizes (n > 30 for each group) are preferred to better approximate the normal distribution for the sample means.

Steps to Perform an Independent Samples Z-test

1.State Hypotheses: Null Hypothesis (H_0): There is no difference in means between the two groups ($\mu 1 = \mu 2$).

Alternative Hypothesis (H_1): There is a difference in means ($\mu 1 \neq \mu 2$) for a two-tailed test, or that one mean is greater than or less than the other for a one-tailed test.

2.Calculate the Z-Statistic:

The Z-statistic for the independent Z-test can be calculated as follows:

$$Z=\large\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

where:

$\acute{X}_1 and \acute{X}_2$ are the sample means

$\sigma_1^2 and \sigma_2^2$ are the sample variances

$n_1 and n_2$ are the sample sizes

3.Calculate the p-value:

```
Use the Z-statistic to calculate the p-value based on the normal
distribution. For a two-tailed test, you will double the area on one
side of the Z-statistic.
```

4.Decision Rule:

Compare the p-value with the significance level (α, typically 0.05).

If $p \leq \alpha$, reject H_0 (indicating a significant difference between the groups).

If $p > \alpha$, fail to reject H_0 (indicating no significant difference).

5.Conclusion: Based on the p-value, conclude whether the sample means are significantly different or not.

Example

A researcher wants to compare the average test scores of two classes to determine if there is a significant difference between their performances. The following data is collected:

Class 1: Sample mean ($\acute{X}_1$) = 75, Population standard deviation (σ_1) = 10, Sample size (n_1) = 30.

Class 2: Sample mean ($\acute{X}_2$) = 70, Population standard deviation (σ_2) = 12, Sample size (n_2) = 35.

The researcher wants to test at a 5% significance level (α=0.05) whether there is a significant difference between the two classes (two-tailed test).

Steps to Solve:

1. Set up the hypotheses:

Null Hypothesis (H_0):μ1=μ2 (The population means are equal).

Alternative Hypothesis (H_1):μ1≠μ2 (The population means are not equal).

1. Calculate the test statistic(Z):

$$Z=\large\frac{75-70}{\sqrt{\frac{10^2}{30}+\frac{12^2}{35}}}$$

$$Z=\large\frac{5}{\sqrt{{3.33}+{4.11}}}$$

$$Z=\large\frac{5}{2.73} ≈1.83 $$

1. Determine the critical value:

For a two-tailed test at α=0.05: The critical values are Z=±1.96.

1. Make a decision:

If |Z|>1.96, reject the null hypothesis.

If |Z|≤1.96, fail to reject the null hypothesis.

Here: |Z|=1.83

Since 1.83<1.96, we fail to reject the null hypothesis.

Conclusion: There is no significant difference between the average test scores of the two classes at the 5% significance level.

Exercise

1.A company wants to compare the average weekly hours worked by employees in two departments.

Department 1:

Sample mean ($\acute{X}_1$) = 42 hours,

Population standard deviation (σ_1) = 6 hours,

Sample size (n_1) = 40.

Department 2:

Sample mean ($\acute{X}_2$) = 38 hours,

Population standard deviation (σ_2) = 5 hours,

Sample size (n_2) = 35.

Test at a 1% significance level (α=0.01) whether there is a significant difference in the average weekly hours worked between the two departments.

2.A study compares the average daily water intake of men and women in liters.

Men:

Sample mean ($\acute{X}_1$) = 2.8 liters,

Population standard deviation (σ_1) = 0.5 liters,

Sample size (n_1) = 50.

Women:

Sample mean ($\acute{X}_2$) = 2.5 liters,

Population standard deviation (σ_2) = 0.6 liters,

Sample size (n_2) = 60.

Test at a 5% significance level (α=0.05) whether men consume more water than women (right-tailed test).

```
from scipy import stats
import numpy as np

# Sample Data
group_A = np.random.normal(loc=25, scale=5, size=50)  # Sample A: mean
= 25, std = 5, n = 50
group_B = np.random.normal(loc=30, scale=6, size=50)  # Sample B: mean
= 30, std = 6, n = 50

# Known population standard deviations (assumed for this example)
sigma_A = 5
sigma_B = 6

# Sample means
x_bar_A = group_A.mean()
x_bar_B = group_B.mean()

# Sample sizes
n_A = len(group_A)
n_B = len(group_B)
```

```
# Z-Test statistic
z = (x_bar_A - x_bar_B) / np.sqrt((sigma_A**2 / n_A) + (sigma_B**2 /
n_B))

# P-value (two-tailed test)
p_value = 2 * (1 - stats.norm.cdf(abs(z)))

print("Two-Sample Z-Test")
print(f"Z-Statistic: {z}")
print(f"P-Value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print("Reject H0: The means of the two groups are significantly
different.")
else:
    print("Fail to reject H0: No significant difference in the means
of the two groups.")
```

```
Two-Sample Z-Test
Z-Statistic: -3.1151912080541546
P-Value: 0.0018382570654031927
Reject H0: The means of the two groups are significantly different.
```

```
import seaborn as sns

titanic = sns.load_dataset("titanic")

titanic = titanic.dropna(subset=["fare", "sex"])

male_fares = titanic[titanic["sex"] == "male"]["fare"]
female_fares = titanic[titanic["sex"] == "female"]["fare"]

male_mean = male_fares.mean()
female_mean = female_fares.mean()
male_std = male_fares.std(ddof=1)
female_std = female_fares.std(ddof=1)

male_n = len(male_fares)
female_n = len(female_fares)

# Z = (x¯1 - x¯2) / √((s1^2 / n1) + (s2^2 / n2))
z = (male_mean - female_mean) / np.sqrt((male_std**2 / male_n) +
(female_std**2 / female_n))

print("Male mean fare:", male_mean)
print("Female mean fare:", female_mean)
print("Z-score (Formula 1):", z)

p1 = 2 * (1 - stats.norm.cdf(abs(z)))
```

```
print("p-value (Formula 1):", p1)

alpha = 0.05
if p1 < alpha:
    print("Reject H0 : Mean fares are significantly different between
males and females.")
else:
    print("Fail to reject H0: Mean fares are not significantly
different between males and females.")
```

```
Male mean fare: 25.523893414211443
Female mean fare: 44.47981783439491
Z-score (Formula 1): -5.07749901345891
p-value (Formula 1): 3.824354921633244e-07
Reject H0 : Mean fares are significantly different between males and
females.
```

```
import pandas as pd
import numpy as np
from scipy import stats

diabetes = pd.read_csv(r"C:\Users\ASUS\Downloads\diabetes.csv")

diabetes_positive = diabetes[diabetes["Outcome"] == 1]["Glucose"]
diabetes_negative = diabetes[diabetes["Outcome"] == 0]["Glucose"]

positive_mean = diabetes_positive.mean()
negative_mean = diabetes_negative.mean()
positive_std = diabetes_positive.std(ddof=1)
negative_std = diabetes_negative.std(ddof=1)

# Sample sizes
positive_n = len(diabetes_positive)
negative_n = len(diabetes_negative)

# Z = (x¯1 - x¯2) / √((s1^2 / n1) + (s2^2 / n2))
z = (positive_mean - negative_mean) / np.sqrt((positive_std**2 /
positive_n) + (negative_std**2 / negative_n))

# Calculate p-value for two-tailed test
p = 2 * (1 - stats.norm.cdf(abs(z)))

print("Mean Glucose (Diabetes Positive):", positive_mean)
print("Mean Glucose (Diabetes Negative):", negative_mean)
print("Z-score:", z)
print("p-value:", p)

alpha = 0.05
if p < alpha:
    print("Reject H0: Mean glucose levels are significantly different
```

```
between the two groups.")
else:
    print("Fail to reject H0: Mean glucose levels are not
significantly different between the two groups.")
```

```
Mean Glucose (Diabetes Positive): 141.25746268656715
Mean Glucose (Diabetes Negative): 109.98
Z-score: 13.751537067396413
p-value: 0.0
Reject H0: Mean glucose levels are significantly different between the
two groups.
```

Paired Z-test

A paired z-test is used to compare the means of two related groups when the sample size is large (n>30) and the population standard deviation is known. It is used in cases where data points are dependent or paired, such as measurements taken before and after an intervention on the same group.

Formula

The paired z-test statistic is calculated using:

$$z = \large\frac{\bar{d} - \mu_d}{\frac{\sigma_d}{\sqrt{n}}}$$

where:

$\acute{d}$ = mean of the differences between the paired observations

μ_d = hypothesized mean difference (usually 0 for a null hypothesis of no difference)

σ_d = standard deviation of the differences

n = number of pairs

Steps for a Paired Z-Test:

1.State the hypotheses:

Null Hypothesis (H_0): $\mu_d=0$

Alternative Hypothesis (H_1): $\mu_d \neq 0$(two-tailed),$\mu_d > 0$ (right-tailed), $\mu_d < 0$ (left-tailed)

2. Compute the difference(d_i):

Calculate the difference for each pair: $d_i = X_i1 - X_i2$

3.Calculate the mean ($\acute{d}$) and standard deviation (σ_d) of the differences.

4.Compute the z-statistic using the formula above.

5.Determine the critical value or p-value based on the standard normal distribution.

6.Make a decision:

Reject H_0 if |z| exceeds the critical value for a given significance level (α) or if the p-value is less than α.

Example:

A company wants to check if a new training program improves employee productivity. Productivity scores before and after the training are measured for 40 employees. The mean of the differences ($\acute{d}$) is 4, the population standard deviation of differences (σ_d) is 3, and the significance level is 0.05.

1. Hypothesis:

Null Hypothesis (H_0): μ_d=0 (no improvement)

Alternative Hypothesis (H_1):μ_d > 0 (improvement)

Given:

$\acute{d}$ = 4

μ_d = 0

σ_d = 3

n = 40

$$z = \frac{4 - 0}{\frac{3}{\sqrt{40}}} = 8.44$$

Determine critical value: For a one-tailed test at α=0.05, the critical value is 1.645.

Decision: Since z=8.44>1.645, reject H_0 . The training program significantly improved productivity.

Exercise

1.A teacher wants to evaluate if a study program improves test scores. Ten students take a pre-test and post-test. The differences in scores (d=Post-test−Pre-test) are:

d={4,5,3,6,7,4,3,5,6,4}

The population standard deviation of the differences is known to be σ_d=1.5. Test if there is a significant improvement in test scores at α=0.05.

2.A weight-loss clinic tests a new diet program on 50 participants. Their weights (in kg) are measured before and after the program. The mean of the differences ($\acute{d}$) is 2.5, and the population standard deviation of the differences is σ_d=1.8. Is the program effective in reducing weight at α=0.01?

```
import numpy as np
from scipy import stats
```

```
before = np.array([78, 82, 85, 90, 76, 80, 78, 74,
                   88, 92, 89, 85, 80, 75, 82, 81,
                   79, 87, 91, 86])

after = np.array([85, 84, 88, 92, 79, 82, 81, 78,
                  90, 94, 91, 87, 83, 80, 85, 84,
                  82, 90, 95, 89])

before_mean = np.mean(before)
after_mean = np.mean(after)
before_std = np.std(before, ddof=1)
after_std = np.std(after, ddof=1)

diff = after - before

diff_mean = np.mean(diff)
diff_std = np.std(diff, ddof=1)

z_score = diff_mean / (diff_std / np.sqrt(len(diff)))

p_value = 2 * (1 - stats.norm.cdf(abs(z_score)))

print("Before mean:", before_mean)
print("After mean:", after_mean)
print("Difference mean:", diff_mean)
print("Z-score:", z_score)
print("p-value:", p_value)

alpha = 0.05
if p_value < alpha:
    print("Reject null hypothesis: The means are significantly
different.")
else:
    print("Fail to reject null hypothesis: The means are not
significantly different.")
```

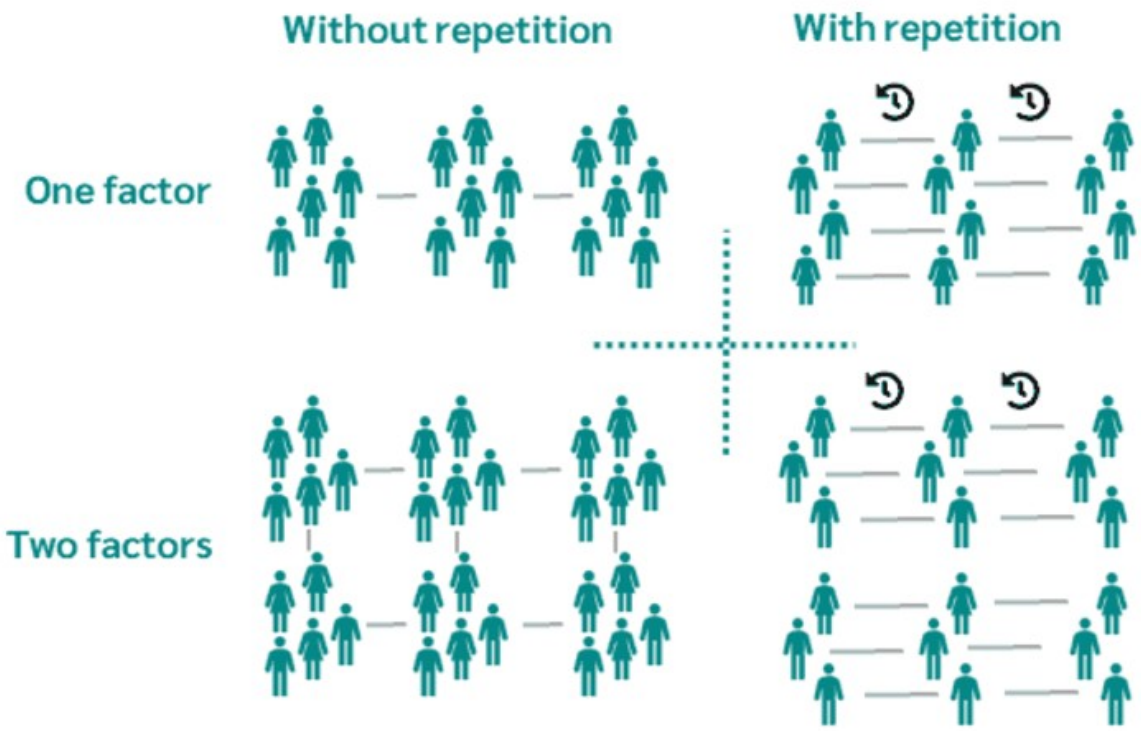
```
Before mean: 82.9
After mean: 85.95
Difference mean: 3.05
Z-score: 11.050129142173947
p-value: 0.0
Reject null hypothesis: The means are significantly different.
```

```
import pandas as pd
import numpy as np
from scipy.stats import norm
data = {
    "Glucose_Before": [120, 135, 140, 125, 150, 130, 128, 115, 145,
160],
    "Glucose_After": [110, 130, 138, 120, 145, 125, 126, 112, 140,
```

```
155]
}

df = pd.DataFrame(data)
df['Difference'] = df['Glucose_After'] - df['Glucose_Before']
mean_diff = df['Difference'].mean()
std_diff = df['Difference'].std()
sigma_d = 5
n = len(df)
z_score = mean_diff / (sigma_d / np.sqrt(n))
p_value = 2 * (1 - norm.cdf(abs(z_score)))
print(f"Mean of Differences: {mean_diff:.2f}")
print(f"Z-Score: {z_score:.2f}")
print(f"P-Value: {p_value:.4f}")
alpha = 0.05
if p_value < alpha:
    print("Reject null hypothesis: Significant difference in glucose
levels before and after.")
else:
    print("Fail to reject null hypothesis: No significant difference
in glucose levels before and after.")
```

```
Mean of Differences: -4.70
Z-Score: -2.97
P-Value: 0.0030
Reject null hypothesis: Significant difference in glucose levels
before and after.
```

A guide to choose the right test



Chi-square test

The Chi-Square Test is a statistical method used to determine whether there is a significant association between two categorical variables or to check the goodness of fit between observed and expected frequencies. It is a non-parametric test, meaning it does not rely on assumptions of a specific population distribution.

Types of Chi-Square Tests

1.Chi-Square Test for Independence:

Used to test if two categorical variables are independent.

Example: Testing if gender and preference for a product are independent.

2.Chi-Square Goodness of Fit Test:

Used to test if an observed frequency distribution fits an expected distribution.

Example: Checking if a die is fair by comparing observed outcomes to the expected probabilities.

Formula:

The formula for the Chi-Square statistic is:

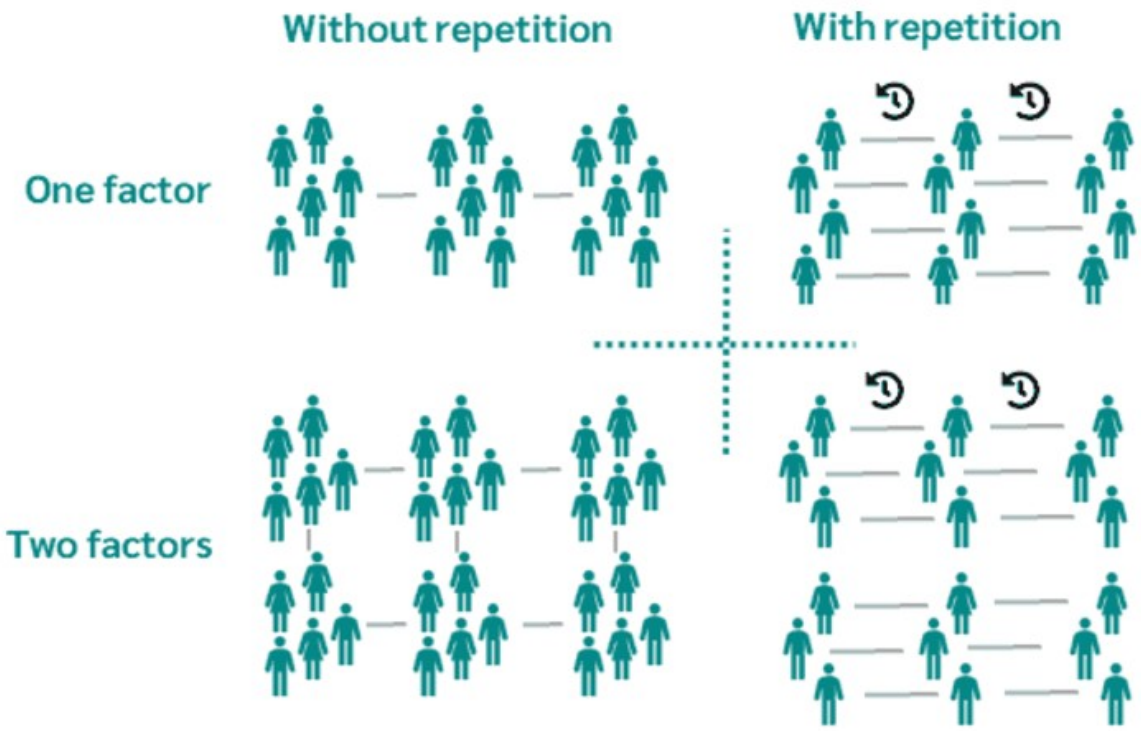
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

O_i = Observed Frequency

E_i = Expected Frequency

Summation is over all categories.

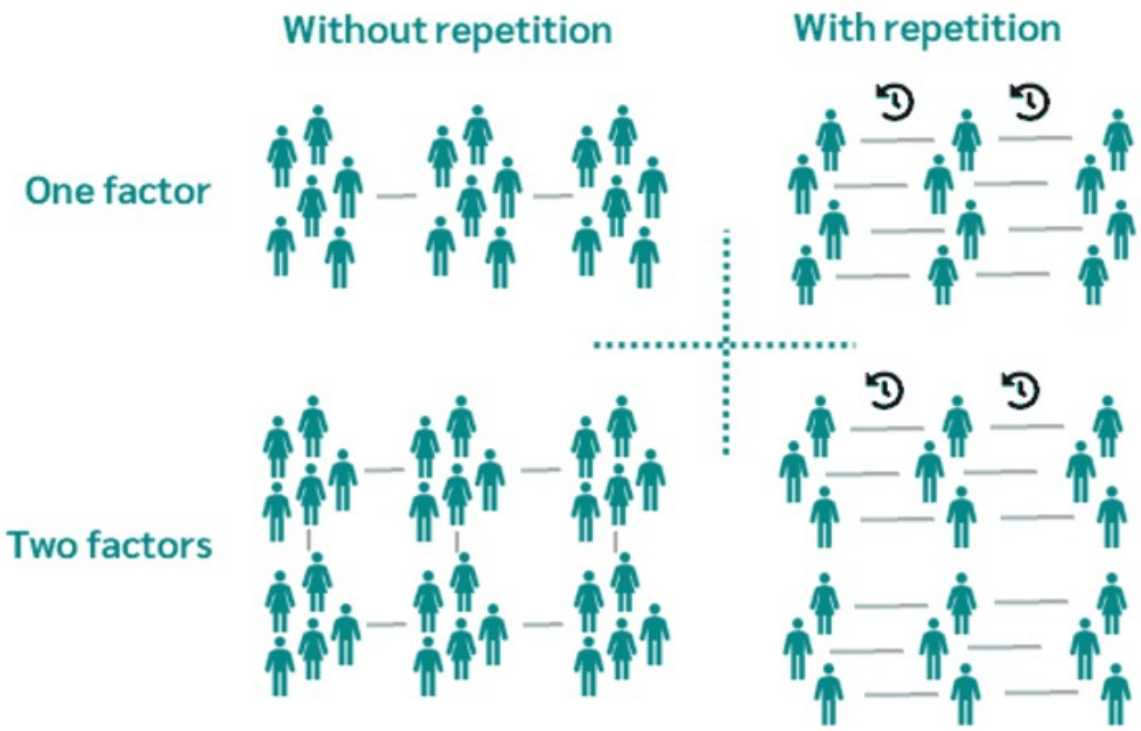


	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Steps for Performing a Chi-Square Test

1. **State the hypotheses**:
   - Null hypothesis (H_0): Assumes no relationship or difference.
   - Alternative hypothesis (H_a): Assumes there is a relationship or difference.
2. **Calculate the expected frequencies**:
   - For independence: $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
   - For goodness of fit: Use the expected proportions or theoretical distribution.
3. **Compute the Chi-Square statistic**:
   - Use the formula above.
4. **Determine the degrees of freedom (df)**:
   - For independence: $df = (\text{number of rows} - 1)(\text{number of columns} - 1)$
   - For goodness of fit: $df = \text{number of categories} - 1$
5. **Find the p-value**:
   - Compare the calculated χ^2 value to the critical value from the Chi-Square table or compute the p-value.
6. **Make a decision**:
   - Reject H_0 if p-value < significance level (α), otherwise fail to reject H_0.

Example



1. Define Hypothesis

Null Hypothesis (H_0): Gender and beverage preference are independent.

Alternative Hypothesis (H_1): Gender and beverage preference are not independent.

1. Calculate Expected Frequencies

The formula for expected frequencies (E) in a contingency table is:

$$E = \frac{RowTotal * ColumnTotal}{GrandTotal}$$

Calculate the expected frequencies for each cell:

For Males preferring Tea:

$$E = \frac{60 * 30}{100} = 18$$

For Males preferring Coffee:

$$E = \frac{60 * 50}{100} = 30$$

For Males preferring Juice:

$$E = \frac{60 * 20}{100} = 12$$

For Females preferring Tea:

$$E = \frac{40 * 30}{100} = 12$$

For Females preferring Coffee:

$$E = \frac{40 * 50}{100} = 20$$

For Females preferring Juice:

$$E = \frac{40 * 20}{100} = 8$$

The expected frequency table is:

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

1. Calculate the Chi-Square Statistic

For Males preferring Tea:

$$\chi^2 = \sum \frac{(20 - 18)^2}{18} = 0.222$$

For Males preferring Coffee:

$$\chi^2 = \sum \frac{(30 - 30)^2}{30} = 0$$

For Males preferring Juice:

$$\chi^2 = \sum \frac{(10 - 12)^2}{12} = 0.333$$

For Females preferring Tea:

$$\chi^2 = \sum \frac{(10 - 12)^2}{12} = 0.333$$

For Females preferring Coffee:

$$\chi^2 = \sum \frac{(20-20)^2}{20} = 0$$

For Females preferring Juice:

$$\chi^2 = \sum \frac{(10-8)^2}{8} = 0.5$$

$$\chi^2 = 0.222 + 0 + 0.333 + 0.333 + 0 + 0.5 = 1.388$$

1. Degrees of Freedom

The degrees of freedom (df) for a contingency table is calculated as:

df=(number of rows−1)(number of columns−1)

In this example:

df=(2−1)(3−1)=1×2=2

5.Critical Value and Decision At a significance level (α) of 0.05 and df=2, the critical value from the Chi-Square table is 5.991.

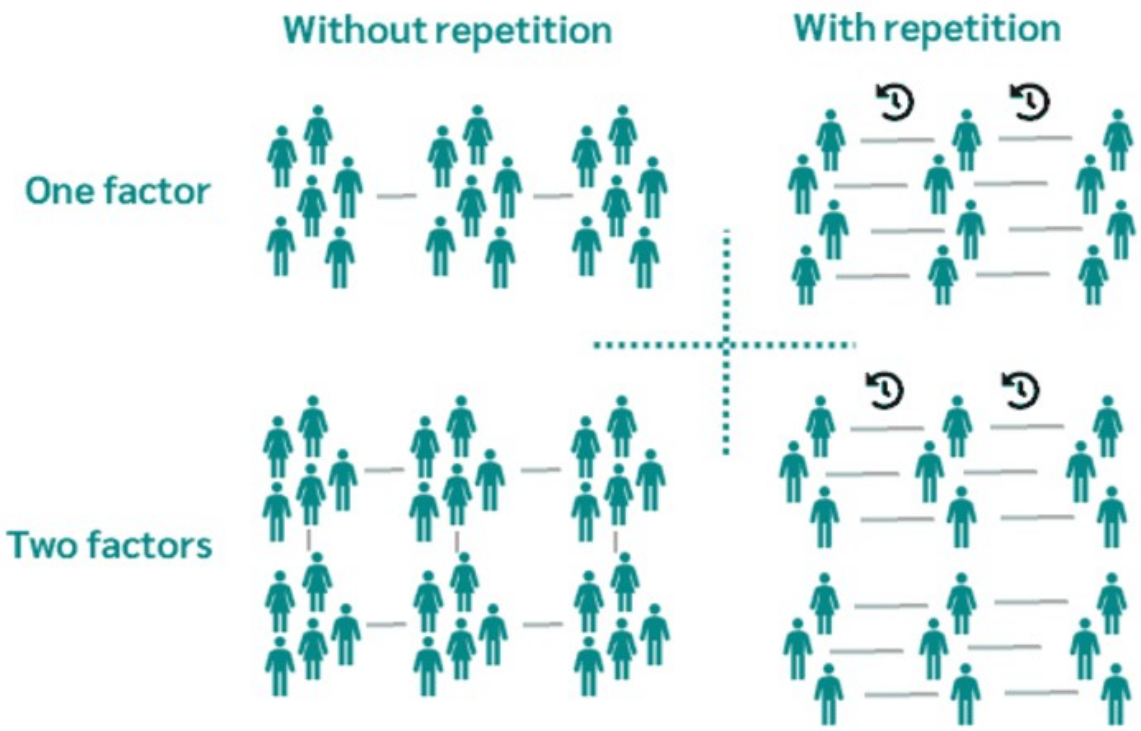
Decision Rule: If $\chi 2 > 5.991$, reject H_o.

1. Conclusion The calculated $\chi 2 = 1.388$ is less than the critical value 5.991.

Fail to reject H_o : There is no significant evidence to conclude that gender and beverage preference are dependent.

Exercise

1.A survey of 150 people records their favorite sport based on gender. The observed data is:



Test if gender and favorite sport are independent at a significance level of 0.05.

2.A survey of 130 people records their favorite ice cream flavor based on age group. The observed data is:

	Chocolate	Vanilla	Strawberry	Total
Children (under 18)	30	10	10	50
Adults (18–60)	20	25	15	60
Seniors (60+)	10	5	5	20
Total	60	40	30	130

Test if age group and ice cream flavor preference are independent at a significance level of 0.05.

```
import pandas as pd
from scipy.stats import chi2_contingency
import seaborn as sns

titanic = sns.load_dataset('titanic')

contingency_table = pd.crosstab(titanic['survived'], titanic['sex'])

chi2, p, dof, expected = chi2_contingency(contingency_table)

print("contingency_table", contingency_table)
```

```
print("Chi-square statistic:", chi2)
print("p-value:", p)
print("Degrees of freedom:", dof)
print('expected frequencie',expected)

if p < 0.05:
    print("Reject the null hypothesis - variables are not
independent")
else:
    print("Fail to reject the null hypothesis - variables are
independent")
```

```
contingency_table sex       female  male
survived
0             81   468
1            233   109
Chi-square statistic: 260.71702016732104
p-value: 1.1973570627755645e-58
Degrees of freedom: 1
expected frequencie [[193.47474747 355.52525253]
 [120.52525253 221.47474747]]
Reject the null hypothesis - variables are not independent
```

```
import pandas as pd
from scipy.stats import chi2_contingency

diabetes = pd.read_csv(r"C:\Users\ASUS\Downloads\diabetes.csv")

diabetes['age_group'] = diabetes['Age'].apply(lambda x: 'Under 30' if
x < 30 else '30 and above')

contingency_table_diabetes = pd.crosstab(diabetes['age_group'],
diabetes['Outcome'])

chi2, p, dof, expected = chi2_contingency(contingency_table_diabetes)

print("Contingency Table:\n", contingency_table_diabetes)
print("Chi-Square Statistic:", chi2)
print("P-value:", p)
print("Degrees of Freedom:", dof)
print("Expected Frequencies:\n", expected)

if p < 0.05:
    print("Reject the null hypothesis - age group and diabetes outcome
are not independent")
else:
    print("Fail to reject the null hypothesis - age group and diabetes
outcome are independent")
```

```
Contingency Table:
 Outcome         0    1
age_group
30 and above  188  184
Under 30      312   84
Chi-Square Statistic: 66.14348877314308
P-value: 4.1926238130017757e-16
Degrees of Freedom: 1
Expected Frequencies:
 [[242.1875 129.8125]
 [257.8125 138.1875]]
Reject the null hypothesis - age group and diabetes outcome are not
independent
```

ANOVA(Analysis of Variance)

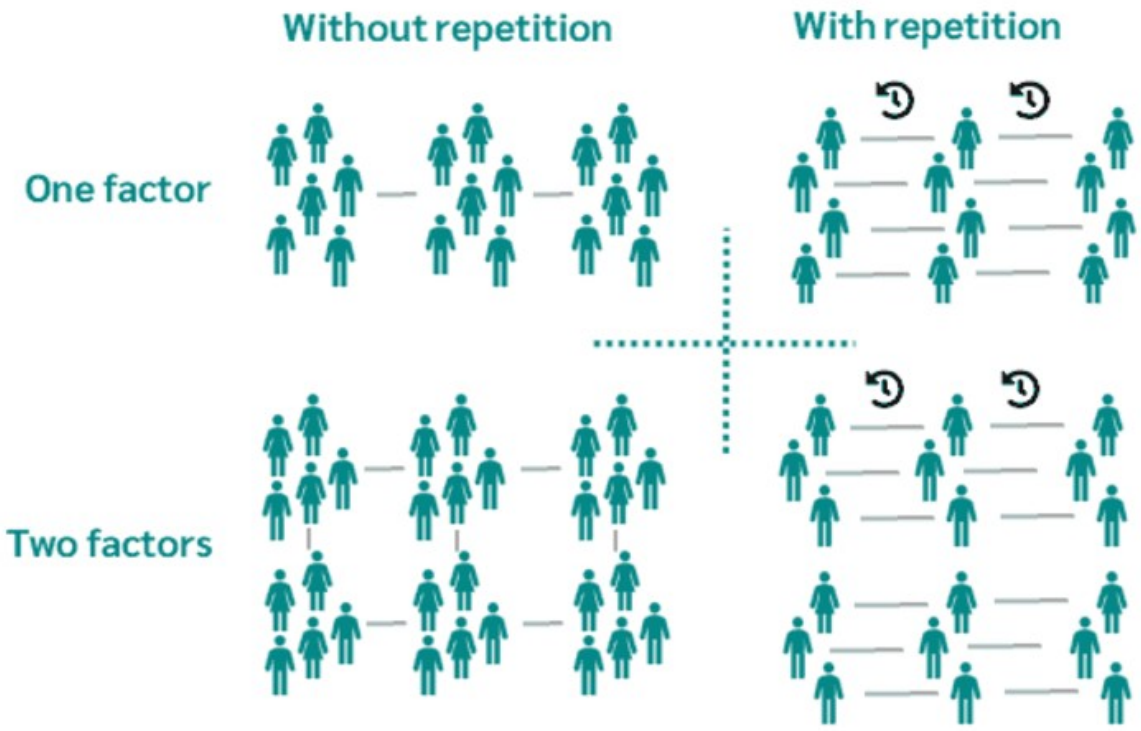
An ANOVA (Analysis of Variance) test is used to compare means across three or more groups to determine if at least one group mean is significantly different from the others. It is an extension of the t-test, which is used for comparing two means.

Types of ANOVA:

One-Way ANOVA: Used when you have one independent variable with three or more levels (groups) to compare.

Example: Comparing the average test scores of students from three different teaching methods.

Two-Way ANOVA: Used when there are two independent variables. This test can also examine the interaction effect between the two factors on the dependent variable.



Example: Comparing the average test scores based on teaching method and student gender.

Hypotheses:

Null hypothesis (H_0): All group means are equal.

Alternative hypothesis (H_1): At least one group mean is different.

Steps in Conducting a One-Way ANOVA: State the hypotheses:

H_0: $\mu_1 = \mu_2 = \ldots = \mu k$ (all group means are equal) H_1: At least one $\mu \neq \mu$ (at least one group mean is different) Set significance level (α): Typically 0.05.

Calculate the F-statistic: The F-statistic is the ratio of the variance between the groups to the variance within the groups.

$$F = \frac{VarianceBetweenGroups}{VariancewithinGroups}$$

Determine the critical value: Using an F-distribution table, find the critical F-value based on the degrees of freedom and the significance level.

Compare the F-statistic to the critical value:

If F > critical value, reject H_0 (there is a significant difference between group means).

If F ≤ critical value, fail to reject H_0 (no significant difference).