19IT601 - DATA SCIENCE & ANALYTICS

UNIT-II DESCRIPTIVE ANALYTICS USING STATISTICS

Types of Data – Mean, Median and Mode – Standard Deviation and Variance – Probability – Probability Density Function – Types of Data Distribution – Percentiles and Moments – Correlation and Covariance – Conditional Probability – Bayes' Theorem – Introduction to Univariate, Bivariate and Multivariate Analysis – Dimensionality Reduction using Principal Component Analysis and LDA – Dimensionality Reduction using Principal Component Analysis and Linear Discriminant Analysis (LDA) – Principal Component Analysis (PCA) example with Iris Data Set from UCI repository.

Types of Data

There are several flavors of data, and there are three specific types of data that we will primarily focus on.

- Numerical data
- Categorical data
- Ordinal data

Numerical data

It's probably the most common data type. Basically, it represents some quantifiable data that you can measure.

Numerical data refers to the data that is in the form of numbers, and not in any language or descriptive form. Often referred to as quantitative data, numerical data is collected in number form.

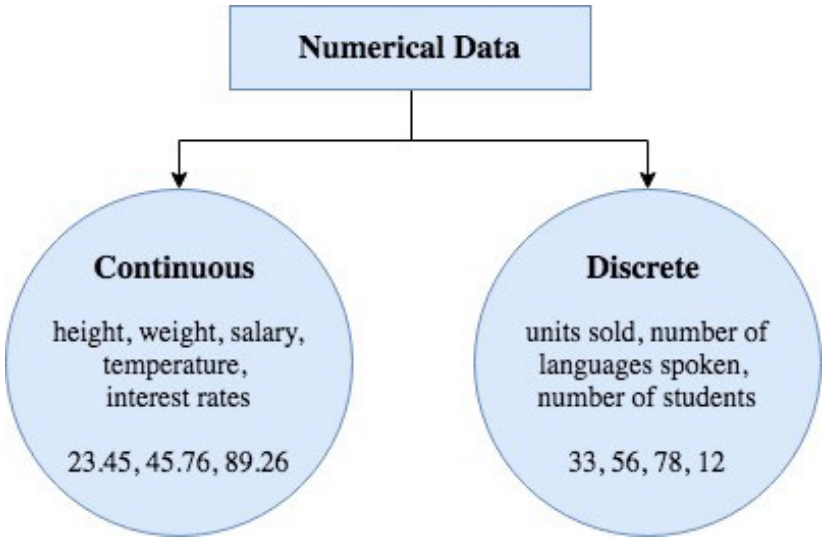
Some examples are

heights of people, page load times, stock prices, and so on.

Things that vary, things that you can measure, things that have a wide range of possibilities are numerical data.

Types of Numerical Data

- Discrete data
- Continuous data



Discrete data

Discrete data, which is integer-based and, for example, can be counts of some sort of event. Discrete data is used to represent countable items. It can take both numerical and categorical forms and groups them into a list. This list can be finite or infinite too

Some examples are

how many purchases did a customer make in a year.

They bought one thing, or they bought two things, or they bought three things. They couldn't have bought, 2.25 things or three and three-quarters things.

It's a discrete value that has an integer restriction to it.

Continuous data

Continuous data, is that has an infinite range of possibilities where you can go into fractions.

So, for example, going back to the height of people, there is an infinite number of possible heights for people. You could be five feet and 10.37625 inches tall, or the time it takes to do something like check out on a website could be any huge range of possibilities, 10.7625 seconds for all you know, or how much rainfall in a given day.

Again, there's an infinite amount of precision there. So that's an example of continuous data.

To recap, numerical data is something you can measure quantitatively with a number, and it can be either discrete, where it's integer-based like an event count, or continuous, where you can have an infinite range of precision available to that data.

Categorical data

Qualitative data that has no inherent numeric meaning. Categorical data is a collection of information that is divided into groups.

Examples are

- Gender,
- yes/no questions, True or False
- race,
- state of residence,
- product category,
- political party;

You can assign numbers to these categories, and often you will, but those numbers have no inherent meaning.

Categorical data does not have any intrinsic numerical meaning; it's just a way that you're choosing to split up a set of data based on categories

Ordinal data

Ordinal data is one that its a mixture of numerical and categorical data. It is categorical data that has mathematical meaning.

A common example is star ratings for a movie or music, or what have you.

In this case, we have categorical data in that could be 1 through 5 stars, where 1 might represent poor and 5 might represent excellent, but they do have mathematical meaning.

Mean, Median and Mode

Mean

The mean, as you probably know, is just another name for the average.

To calculate the mean of a dataset, all you have to do is sum up all the values and divide it by the number of values that you have.

Mean = Sum of samples / Number of samples.

Example

Number of children in each house on my street: 0, 2, 3, 2, 1, 0, 0, 2, 0

The mean is (0+2+3+2+1+0+0+2+0)/9 = 1.11

Median

The way you compute the median of the dataset is by sorting all the values (in either ascending or descending order), and taking the one that ends up in the middle.

So, for example, let's use the same dataset of children in my neighborhood
0, 2, 3, 2, 1, 0, 0, 2, 0
I would sort it numerically, and I can take the number that's slap dab in the middle of the data, which turns out to be 1.
0, 0, 0, 0, 1, 2, 2, 2, 3.

Mode
All mode means, is the most common value in a dataset.
Let's go back to my example of the number of kids in each house.
0, 2, 3, 2, 1, 0, 0, 2, 0
How many of each value are there:
0: 4, 1: 1, 2: 3, 3: 1
The MODE is 0

Standard Deviation and Variance

Standard deviation and variance are two fundamental quantities for a data distribution.

Variance
Variance measures how spread-out the data is. A variance is the average of the squared differences from the mean.

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

Where,
X denotes each data point
μ denotes the mean
N denotes the number of data points

Example
To compute the variance of a dataset, first figure out the mean. Lets say our data set has five values

(1,4,5,4,8)

1. The first step in computing the variance is just to find the mean, or the average, of that data.

Mean of the above dataset is (1+4+5+4+8)/5 = 4.4

2. Now the next step is to find the differences from the mean for each data point.

1-4.4 = -3.4,

4-4.4 = -0.4,

5-4.4 = 0.6,

4-4.4 = -0.4,

8-4.4 = 3.6

-3.4, -0.4, 0.6, -0.4, 3.6

3. Next is to do is find the square of these differences.

$(-3.4)^2 = 11.56$

$(-0.4)^2 = 0.16$

$(0.6)^2 = 0.36$

$(-0.4)^2 = 0.16$

$(3.6)^2 = 12.36$

4. To find the actual variance value, we just take the average of all those squared differences.

$\sigma 2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.156) / 5 = 5.04$

Standard Deviation

Standard deviation is just the square root of the variance.

Variance is $\sigma 2 = 5.04$

Standard Deviation is $\sqrt{5.04} = 2.24$

Population variance versus sample variance

If variance is calculated for complete data set then this is called population variance.

For example $\sigma 2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.156) / 5 = 5.04$

Sample Variance

If we calculate for a subset of the data then that is called sample variance.

Instead of dividing by the number of samples, you divide by the number of samples minus 1.

Sample variance, which is designated by S2, it is found by the sum of the squared variances divided by 4, that is (n - 1).

S2 = (11.56 + 0.16 + 0.36 + 0.16+ 12.156) / 4 = 6.3

Probability – Probability Density Function

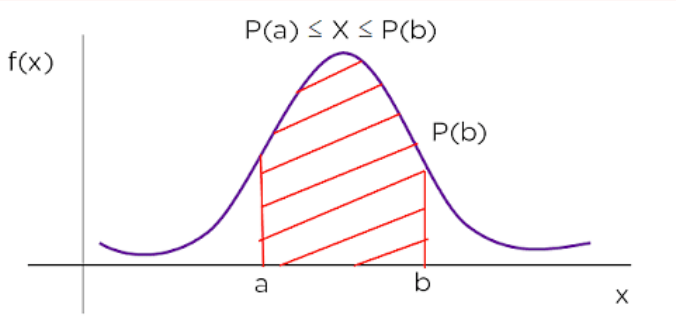
A function that defines the relationship between a random variable and its probability, such that you can find the probability of the variable using the function, is called a Probability Density Function (PDF) in statistics.

The Probability Density Function(PDF) defines the probability function representing the density of a continuous random variable lying between a specific range of values. In other words, the probability density function produces the likelihood of values of the continuous random variable.

Probability Density Function is denoted by f(x)

Continuous Variable: A continuous random variable can take on infinite different values within a range of values, e.g., amount of rainfall occurring in a month.

Now, consider a continuous random variable x, which has a probability density function, that defines the range of probabilities taken by this function as f(x)



As the probability cannot be more than P(b) and less than P(a), you can represent it as: P(a) <= X <= P(b).

Types of Data Distribution

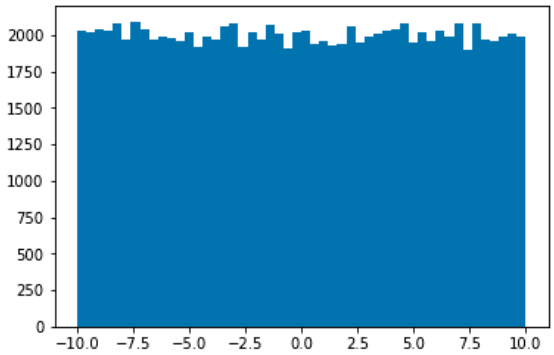
In statistical terms, a distribution function is a mathematical expression that describes the probability of different possible outcomes for an experiment.

There are many categories of data distribution

1. Uniform Distribution
2. Normal Distribution
3. Exponential Distribution
4. Binomial Distribution
5. Poisson Distribution

Uniform Distribution

- A uniform distribution just means there's a flat constant probability of a value occurring within a given range.
- Uniform distribution can either be discrete or continuous where each event is equally likely to occur. It has a constant probability constructing a rectangular distribution.
- When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution.
- All the n number of possible outcomes of a uniform distribution are equally likely. Every value, every range of values has an equal chance of appearing as any other value.



Normal Distribution

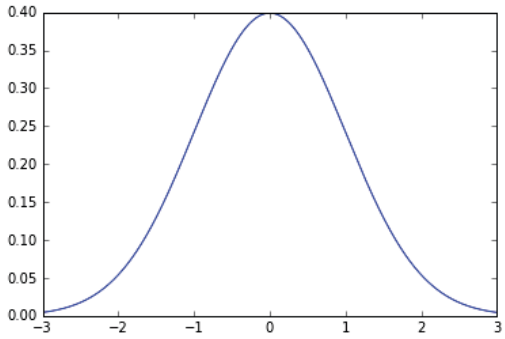
- It is otherwise known as Gaussian Distribution and Symmetric Distribution. It is a type of continuous probability distribution which is symmetric to the mean. The majority of the observations cluster around the central peak point.

- Normal distribution represents the behavior of most of the situations in the universe.
- The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application.

Any distribution is known as Normal distribution if it has the following characteristics:

- The mean, median and mode of the distribution coincide.
- The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
- The total area under the curve is 1.
- Exactly half of the values are to the left of the center and the other half to the right.

```
Out[3]: [<matplotlib.lines.Line2D at 0xde514e0>]
```



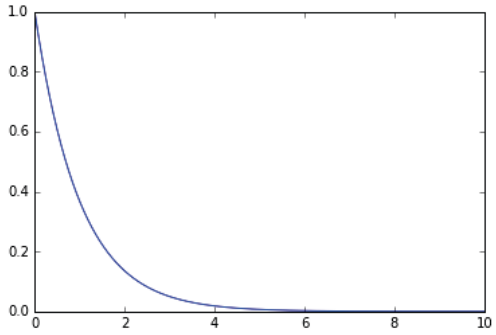
Exponential Distribution

- The exponential distribution is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate.
- It is concerned with the amount of time until some specific event occurs.

Example:

- The amount of time until an earthquake occurs has an exponential distribution
- The amount of time in business telephone calls
- The car battery lasts.
- The exponential distribution is widely used in the field of reliability.

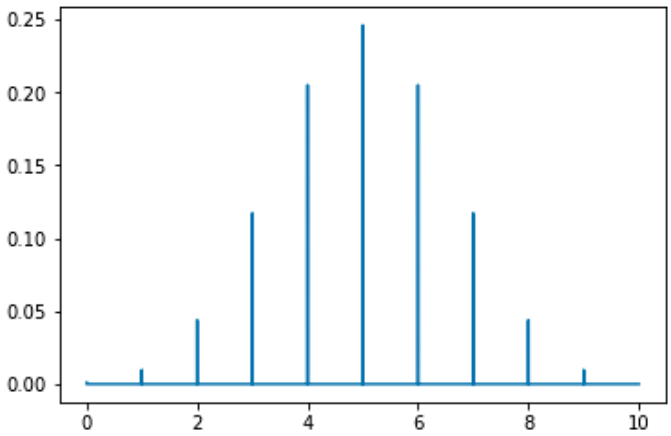
```
Out[5]: [<matplotlib.lines.Line2D at 0xe3304e0>]
```



Binomial Distribution

- A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times.
- The binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial. These outcomes are appropriately labeled "success" and "failure“
- The binomial is a type of distribution that has two possible outcomes (the prefix “bi” means two, or twice).
- For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.
- The terms p and q remain constant throughout the experiment, where p is the probability of getting a success on any one trial and q = (1 – p) is the probability of getting a failure on any one trial.

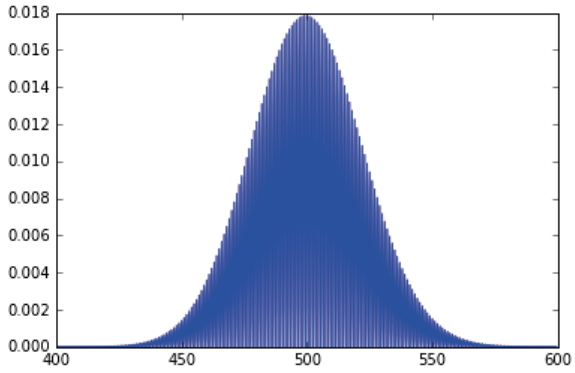
```
Out[5]: [<matplotlib.lines.Line2D at 0x966b128>]
```



Poisson Distribution

- It is the discrete probability distribution of the number of times an event is likely to occur within a specified period of time. It is used for independent events which occur at a constant rate within a given interval of time.
- The occurrences in each interval can range from zero to infinity (0 to α).
- Examples:
- How many black colours are there in a random sample of 50 cars
- No of cars arriving at a car wash during a 20 minute time interval

```
Out[7]: [<matplotlib.lines.Line2D at 0xe742e48>]
```



Percentiles and Moments

Percentiles

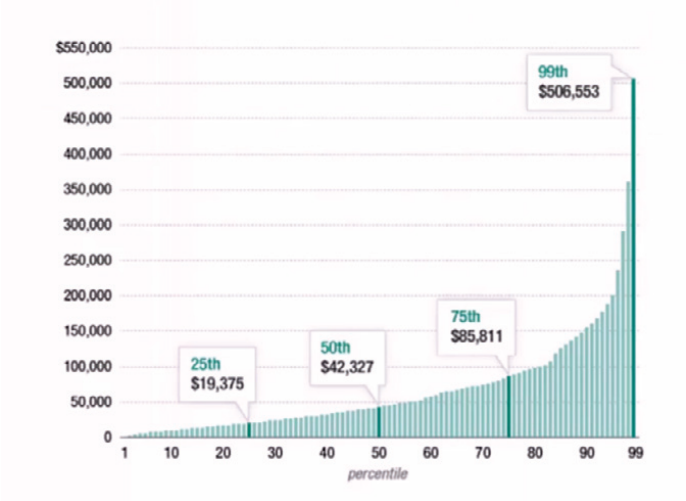
In a dataset, a percentile is the point at which x% of the values are less than the value at that point.

A percentile is a measure at which that percentage of the total values are the same as or below that Measure.

Percentiles are useful for giving the relative standing of particular data in a dataset. Percentiles are essentially normalized ranks.

Example

The 80th percentile is a value where you'll find 80% of the values lower and 20% of the values higher.



Quartiles

Quartiles divide the data into four groups, each containing an equal number of values.

Quartiles are divided by the 25th, 50th, and 75th percentile, also called the first, second and third quartile. It can be represented as Q1,Q2,Q3 and Q4 respectively.

One quarter of the values are less than or equal to the 25th percentile.

Three quarters of the values are less than or equal to the 75th percentile.

Interquartile range

The difference between the 75th (Q3) and 25th (Q1) percentile is called the interquartile range.

For example, the interquartile range (IQR), when we talk about a distribution, is the area in the middle of the distribution that contains 50% of the values.

Example

10	20	30	40	50	60	70	80	90	100

In the above dataset the minimum value is 10 and maximum value is 100.

10	20	30	40	50	**55**	60	70	80	90	100

Median is 55.

- The first quartile (Q1) is just the "median" of all the values to the left of the true median.
- We can see that 30 is the middle number of the numbers to the left of the true

median, so 30 is the 25th percentile and the first quartile (Q1).

- What if we were asked for the 75th percentile? We know that the 75th percentile is the third quartile (Q3). The third quartile (Q3) is similarly the "median" of the values to the right of the true median.
- We can see that 80 is the middle number of the numbers to the right of the true median, so 80 is the 75th percentile and the third quartile (Q3).

Minimum		**(Q_1)**			**Median**			**(Q_3)**		**Maximum**
10	20	**30**	40	50	**55**	60	70	**80**	90	100

Moments

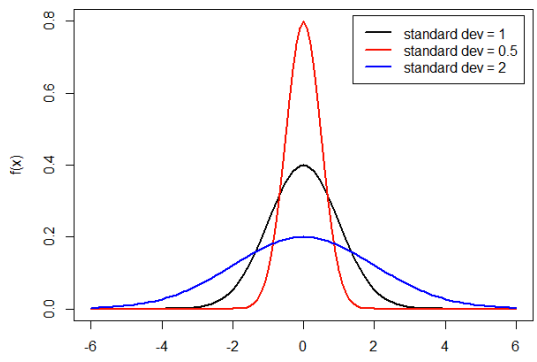
Moments can be defined as quantitative measures of the shape of a probability density function.

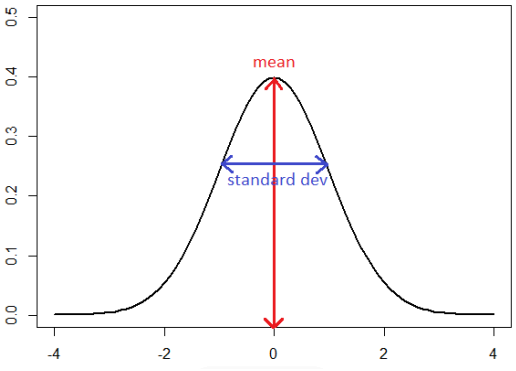
Moments in statistics are popularly used to describe the characteristic of a distribution. The shape of any distribution can be described by its various ‘moments’.

The four commonly used moments in statistics are-

- The first moment - the mean (Measure the location of the central point)
- The second moment - variance (which indicates the width or deviation)
- The third moment – skewness (which indicates any asymmetric ‘leaning’ to either left or right)
- The fouth moment - kurtosis (which indicates the degree of central ‘peakedness’ or, equivalently, the ‘fatness’ of the outer tails.)

The greater the variance/ standard deviation (e.g. blue line), the wider the spread of values around the mean. If a variance is lower, the values are cumulated closer to the mean (red line) and the peak is higher.





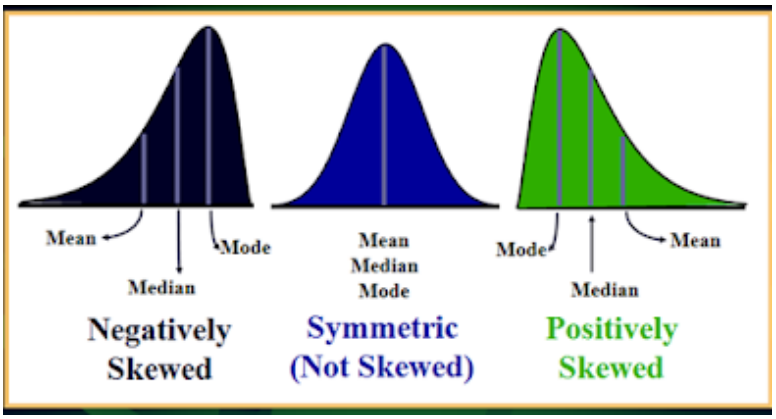
Skew – Third Moment

Skew is how lopsided the data is, how stretched out one of the tails might be.

Symmetrical distribution: as in examples above. Both tails are symmetrical and the skewness is equal to zero.

Positive skew (right-skewed, right-tailed, skewed to the right): the right tail (with larger values) is longer.

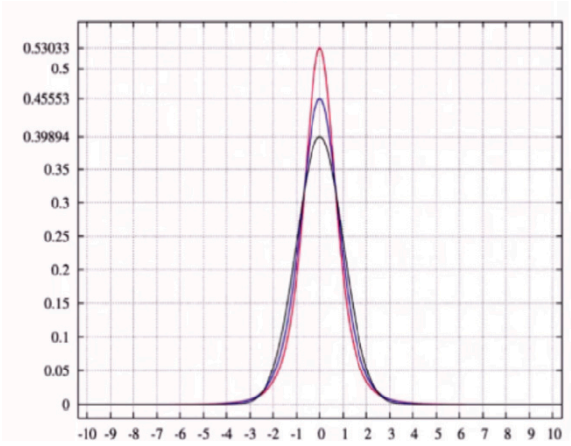
Negative skewed (left-skewed, left-tailed, skewed to the left): the left tail (with small values) is longer.



Kurtosis – Fourth Moment

Kurtosis is how peaked, how squished together the data distribution is.

It focuses on the tails of the distribution and explains whether the distribution is flat or rather with a high peak. Kurtosis informs us whether our distribution is richer in extreme values than normal distribution.



Correlation and Covariance

Covariance

In statistics, covariance is the measure of the directional relationship between two random variables.

These are ways of measuring whether two different attributes are related to each other in a set of data, which can be a very useful thing to find out.

A positive covariance indicates that both random variables tend to move upward or downward at the same time.

A negative covariance indicates that both variables tend to move away from each other — when one moves upward the other moves downward, and vice versa.

Covariance between 2 random variables is calculated by taking the product of the difference between the value of each random variable and its mean, summing all the products, and finally dividing it by the number of values in the dataset.

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Calculate covariance for the following data set:

x: 2.1, 2.5, 3.6, 4.0 (mean = 3.1)

y: 8, 10, 12, 14 (mean = 11)

Substitute the values into the formula and solve:

Cov(X,Y) = ΣE((X-μ)(Y-ν)) / n-1

= (2.1-3.1)(8-11)+(2.5-3.1)(10-11)+(3.6-3.1)(12-11)+(4.0-3.1)(14-11) /(4-1)

= (-1)(-3) + (-0.6)(-1)+(.5)(1)+(0.9)(3) / 3

= 3 + 0.6 + .5 + 2.7 / 3

= 6.8/3

= 2.267

Correlation

The correlation between two random variables measures both the strength and direction of a linear relationship that exists between them.

The Pearson Correlation Coefficient is defined to be the covariance of x and y divided by the product of each random variable's standard deviation.

$$r = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}$$

$$\frac{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}\frac{\sum(y_i - \bar{y})^2}{n-1}}}$$

Correlation of -1 means there's a perfect inverse correlation, so as one value increases, the other decreases, and vice versa.

A correlation of 0 means there's no correlation at all between these two sets of attributes.

A correlation of 1 would imply perfect correlation, where these two attributes are moving in exactly the same way as you look at different data points.

Conditional Probability

Conditional probability is a way to measure the relationship between two things happening to each other.

In mathematical notation, the way we indicate things here is that P(A,B) represents the probability of both A and B occurring independent of each other.

That is, what's the probability of both of these things happening irrespective of everything else.

Whereas this notation, P(B|A), is read as the probability of B given A. So, what is the probability of B given that event A has already occurred

$$P(B|A) = \frac{P(A,B)}{P(A)}.$$

The probability of B given A is equal to the probability of A and B occurring over the probability of A alone occurring, so this teases out the probability of B being dependent on the probability of A

A as the probability of passing the first test, and B as the probability of passing the second test. What I'm looking for is the probability of passing the second test given that you passed the first, that is, P (B|A).

$$P(B|A) = \frac{P(A,B)}{P(A)} = \frac{0.6}{0.8} = 0.75$$

Bayes' Theorem

The Bayes theorem is a mathematical formula for calculating conditional probability in probability and statistics. In other words, it's used to figure out how likely an event is based on its proximity to another. Simply put, it is a way of calculating conditional probability.

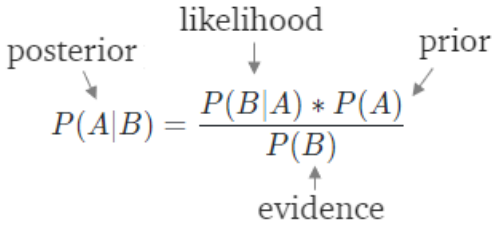
The probability of A given B is equal to the probability of A times the probability of B given A over the probability of B.

The key insight is that the probability of something that depends on B depends very much on the base probability of B and A.

We can find conditional probability using Bayes' Theorem with the following formula:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

The components has special names:



'A' is the event of interest.

P(A) represents our prior belief: probability of event A occurring.

With new evidence B, the posterior belief or updated probability is represented P(A|B): probability of event A given evidence B has occurred.

P(B | A) is the conditional probability of event B occurring, given that A is true

Example 1

John flies frequently and likes to upgrade his seat to first class. He has determined that if he checks in for his flight at least two hours early, the probability that he will get an upgrade is 0.75; otherwise, the probability that he will get an upgrade is 0.35. With his busy schedule, he checks in at least two hours before his flight only 40% of the time. Suppose John did not receive an upgrade on his most recent attempt. What is the probability that he did not arrive two hours early?

Let B = {John arrived at least two hours early}, and A = {John received an upgrade},
then ¬B = {John did not arrive two hours early}, and ¬A = {John did not receive an upgrade}.

John checked in at least two hours early only 40% of the time, or P(B)=0.4 .
Therefore P(¬B) = 1 – P(B) = 1- 0.4 = 0.6.

The probability that John received an upgrade given that he checked in early is 0.75, or P(A|B)=0.75.

The probability that John received an upgrade given that he did not arrive two hours early is 0.35, or P(A|¬B) = 0.35

Therefore
P(¬A | ¬B) = 0.65

The probability that John received an upgrade P(A) can be computed as shown
$P(A) = P(A \cap B) + P(A \cap \neg B)$

$= P(B) * P(A|B) + P(\neg B) * P(A|\neg B)$

$= 0.4 * 0.75 + 0.6 * 0.35$

$= 0.51$

Thus, the probability that John did not receive an upgrade

$P(\neg A) = 1 - 0.51 = 0.49$

Using Bayes' theorem, the probability that John did not arrive two hours early given that he did not receive his upgrade is shown
$P(\neg B|\neg A) = P(\neg A|\neg B) * P(\neg B) / P(\neg A)$
$= 0.65 * 0.6 / 0.49$
$= 0.796$

Example 2
Assume that a patient named Mary took a lab test for a certain disease and the result came back positive. The test returns a positive result in 95% of the cases in which the disease is actually present, and it returns a positive result in 6% of the cases in which the disease is not present. Furthermore, 1% of the entire population has this disease. What is the probability that Mary actually has the disease, given that the test is positive?

Let B = {having the disease} and A = {testing positive}. The goal is to solve the probability of having the disease, given that Mary has a positive test result, P(B|A). From the problem description,
$P(B) = 0.01$, $P(\neg B) = 0.99$
$P(A|B) = 0.95$ and $P(A|\neg B) = 0.06$

Bayes' theorem defines
$P(B|A) = P(A|B) * P(B) / P(A)$

The probability of testing positive, that is P(A) , needs to be computed first. That computation is shown below
$P(A) = P(A \cap B) + P(A \cap \neg B)$

$= P(B) * P(A|B) + P(\neg B) * P(A|\neg B)$

$= 0.01 * 0.95 + 0.99 * 0.06$
$= 0.0689$

According to Bayes' theorem, the probability of having the disease, given that Mary has a positive test result, is
$P(B|A) = P(A|B) * P(B) / P(A)$
$= 0.95 * 0.01 / 0.0689$
$= 0.1379$

Introduction to Univariate, Bivariate and Multivariate Analysis

Univariate Analysis

Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable.

The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately.

When the data contains only one variable and doesn't deal with a causes or effect relationships then a Univariate analysis technique is used.

Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation.

Common visual technique used for univariate analysis is a histogram, which is a frequency distribution graph.

Other visualization techniques includes Frequency Distribution Tables, Frequency Polygons, Pie Charts, Bar Charts.

Example
In a survey of a class room, the researcher may be looking to count the number of boys and girls.

In this instance, the data would simply reflect the number, i.e. a single variable and its quantity as per the below table.

The key objective of Univariate analysis is to simply describe the data to find patterns within the data.

This is be done by looking into the mean, median, mode, dispersion, variance, range, standard deviation etc.

Bivariate Analysis

This type of data involves two different variables.

The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables

Bivariate analysis is where you are comparing two variables to study their relationships.

These variables could be dependent or independent to each other. In Bivariate analysis is that

there is always a Y-value for each X-value.

Bivariate analysis is conducted using

- Correlation coefficients
- Regression analysis

Example
In a survey of a classroom, the researcher may be looking to analysis the ratio of students who scored above 85% corresponding to their genders.

In this case, there are two variables – gender = X (independent variable) and result = Y (dependent variable).

A Bivariate analysis is will measure the correlations between the two variables.

Multivariate analysis

Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set.

When the data involves three or more variables, it is categorized under multivariate.

Types of Multivariate Analysis include

- Cluster Analysis,
- Factor Analysis,
- Multiple Regression Analysis,
- Principal Component Analysis

Example
A doctor has collected data on cholesterol, blood pressure, and weight.

She also collected data on the eating habits of the subjects (e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed per week).

She wants to investigate the relationship between the three measures of health and eating habits?

In this instance, a multivariate analysis would be required to understand the relationship of each variable with each other.

Dimensionality Reduction using Principal Component Analysis and LDA

Dimensionality Reduction

- In Machine Learning, the number of attributes, features or input variables of a dataset is referred to as its dimensionality.

- The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant.

- More input features often make a predictive modeling task more challenging to model, more generally referred to as the curse of dimensionality.

- Dimensionality reduction techniques used to find a way to reduce higher dimensional information into lower dimensional information

- Dimensionality reduction refers to techniques that reduce the number of features or input variables in a dataset.

- Example : classify whether the e-mail is spam or not.

Importance of Dimensionality Reduction

- The performance of machine learning algorithms can degrade with too many input variables.

- Having a large number of dimensions in the feature space can mean that the volume of that space is very large.

- This can dramatically impact the performance of machine learning algorithms fit on data with many input features, generally referred to as the "**curse of dimensionality**."

- A lower number of dimensions in data means less training time and less computational resources and increases the overall performance of machine learning algorithms.

- Dimensionality reduction avoids the problem of overfitting.

- Dimensionality reduction is extremely useful for data visualization Therefore, it is often desirable to reduce the number of input features.

Components of Dimensionality Reduction

- There are two components of dimensionality reduction:

- **Feature selection**: Feature selection is based on omitting those features from the available measurements which do not contribute to class separability. In other words, redundant and irrelevant features are ignored: It includes Filter and Wrapper.

- **Filter** - use scoring methods, like correlation between the feature and the target variable, to select a subset of input features that are most predictive.

- **Wrapper** - fitting and evaluating the model with different subsets of input features and selecting the subset the results in the best model performance

- **Feature extraction**: considers the whole information content and maps the useful information content into a lower dimensional feature space.

- This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

Methods of Dimensionality Reduction
The various methods used for dimensionality reduction include:
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)

Principal Component Analysis (PCA)

- This method was introduced by Karl Pearson.

- PCA is a linear dimensionality reduction technique (algorithm) that transforms a set of correlated variables (p) into a smaller k (k<p) number of uncorrelated variables called principal components while retaining as much of the variation in the original dataset as possible.

- In the context of Machine Learning (ML), PCA is an unsupervised machine learning algorithm that is used for dimensionality reduction.

- It works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.

- In PCA, it takes a higher dimensional data space, and it finds planes within that data space and higher dimensions.

- These higher dimensional planes are called hyper planes, and they are defined by things called eigenvectors.

Four Steps of Principal Component Analysis
PCA implementation is quite straightforward. We can define the whole process into just four steps:

1. **Standardization**: The data has to be transformed to a common scale by taking the difference between the original dataset with the mean of the whole dataset. This will make the distribution 0 centered.

2. **Finding covariance**: Covariance will help us to understand the relationship between the mean and original data.

3. **Determining the principal components**: Principal components can be determined by calculating the **eigenvectors and eigenvalues**.

4. **Final output**: It is the dot product of the standardized matrix and the eigenvector

Eigenvectors
Eigenvectors are a list of coefficients which shows how much each input variable contributes to each new derived variable. If we square and add each Eigenvector then we get Eigenvalue.

Eigenvalue
It Represents the proportion of variance explained by each PC. Also represents the largest variance reduction. Sum of all Eigenvalues equals the sum of the variances of all input variables as variance summarization.

Goals of PCA
1.Extract the most important information from the data table.
2.Compress the size of the data set by keeping only this important information
3.Simplify the description of data set
4.Analyze the structure of the observations and variables.

In order to achieve these goals, PCA computes new variables called principal components, which are obtained as linear combinations of the original variables.

Applications of PCA
- Image compression
- Facial Recognition
- Data Visualisation

These applications are most commonly used in Healthcare and Financial Industries.

Linear Discriminant Analysis

Linear Discriminant Analysis is a dimensionality reduction technique that is commonly used for supervised classification problems. It is used for modelling differences in groups i.e. separating two or more classes.

It is used to project the features in higher dimension space into a lower dimension space**.**

In 1936, Ronald A.Fisher formulated Linear Discriminant first time and showed some practical uses as a classifier, it was described for a 2-class problem, and later generalized as ‘Multi-class Linear Discriminant Analysis’ or ‘Multiple Discriminant Analysis’ by C.R.Rao in the year 1948.

It projects the dataset into moderate dimensional-space with a genuine class of separable features that minimize overfitting and computational costs.

For example, we have two classes and we need to separate them efficiently. Classes can have multiple features. Using only a single feature to classify them may result in some overlapping. So, we will keep on increasing the number of features for proper classification.

Two criteria are used by LDA to create a new axis:

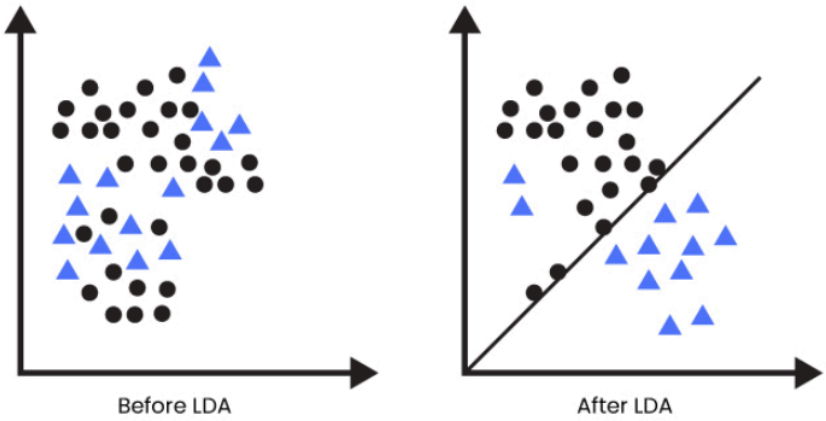
- Maximize the distance between means of the two classes.
- Minimize the variation within each class.

Steps in LDA

First step: To compute the separate ability amid various classes,i.e, the distance between the mean of different classes, that is also known as between-class variance.

Second Step: To compute the distance among the mean and sample of each class,that is also known as the within class variance.

Third step: To create the lower dimensional space that maximizes the between class variance and minimizes the within class variance.



Applications

LDA is used in Marketing, Finance, and other areas to perform a number of classification tasks such as customer profiling and fraud detection.

LDA can be used as a classification task for speech recognition, microarray data classification, face recognition, image retrieval, bioinformatics, biometrics, chemistry, etc. below are other applications of LDA.

In medical: LDA is used here to classify the state of patients' diseases as mild, moderate or severe based on the various parameters and the medical treatment the patient is going through in order to decrease the movement of treatment.