

Machine Learning to build Intelligent Systems

Manas Dasgupta



Fundamentals of Statistics



Structure of this Module

Fundamentals of Statistics

TOPICS

Descriptive Statistics

Probability Theories

Inferential Statistics

Tests of Significance

Introduction to Statistics

Statistics is the science of finding numerical information about data, especially in terms of finding proportions, probabilities, distributions, correlations, patterns and also making decisions about a data population by conducting tests on samples.

Using Statistics, we can answer some of the following questions:

- How does the Square Foot measure of a Property influence the Price of the Property.
- What is the probability of a particular subscriber of a streaming video service to view a particular Series at a given time of his/her visit on the site.
- What is the probability of a positive effect of a particular medicine on a person with a given ailment given that the medicine contains a particular salt.
- Which segment of mortgage borrowers are more prone to default.

Statistics is used in Data Science to find out Patterns, Relationships and Probabilities within Data.
Statistical methods are used as part of Exploratory Data Analysis to find out various insights from Data.
Statistical methods are also built-in various Machine Learning Algorithms to make Predictions from the Data.

Topics to be covered in Statistics

- **Descriptive Statistics:** we will learn the art of summarizing and representing data. We will learn metrics like average, median, standard deviation, root mean square, histograms, etc.
- **Probability Theories:** We will then move onto Probability theories. We will look at Chance models, probability histograms and introduce ourselves to Normal Distribution.
- **Inferential Statistics:** We will understand generalizations that can be made from samples.
- **Tests of Significance:** We will understand the concepts of making inferences making hypothesis about the population from data from the samples and testing those

Statistical Experiments

- **Controlled Experiments**

- Treatment Group
- Control Group
- Double Blind
- Confounding Factors
- Randomization
- Randomized Controlled Experiments
- Association vs Causation
- Predictors and Target Variables

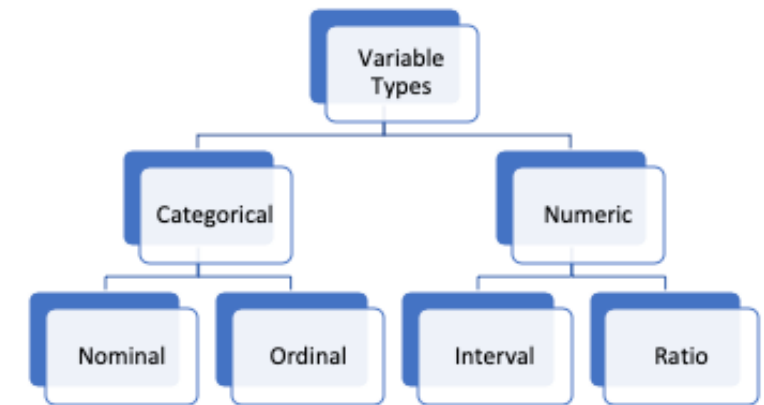
	Treatment Group	Control Group
Sample Size	20000	5000
Improved/Cured	4000	300
%age improved/cured	20.0%	6.0%

- **Observational Studies**

- Effect of smoking on incidences of Heart and Lung ailments

Types of Data/Variables

Nominal	Ordinal	Interval	Ratio
Nominal data are more names or labels.	Ordinal data are also labels however they represent a perceivable degree of measurement.	Interval data are numeric and has a fixed distance between two levels or degrees (Intervals). One key characteristic of interval data is that they do not have an absolute zero.	Ratio are numeric variables that have an equal distancing between measurement units and also an absolute zero.
E.g. Name, Mobile Number, State, Gender	E.g. a CGPA rating of A+, A, B+, B, C or feedback rating of Excellent, Very Good, Good, Not Satisfactory.	E.g. Temperature – 40° Celsius. Mind you there is no absolute zero – i.e. there is nothing like no temperature. Time is also an Interval variable. There is equal spacing in time measurements but no absolute zero.	E.g. Length measured in meters can be absolute zero. Weight in kilograms can be of absolute zero. Monthly expenses in dollars can also be absolute zero.



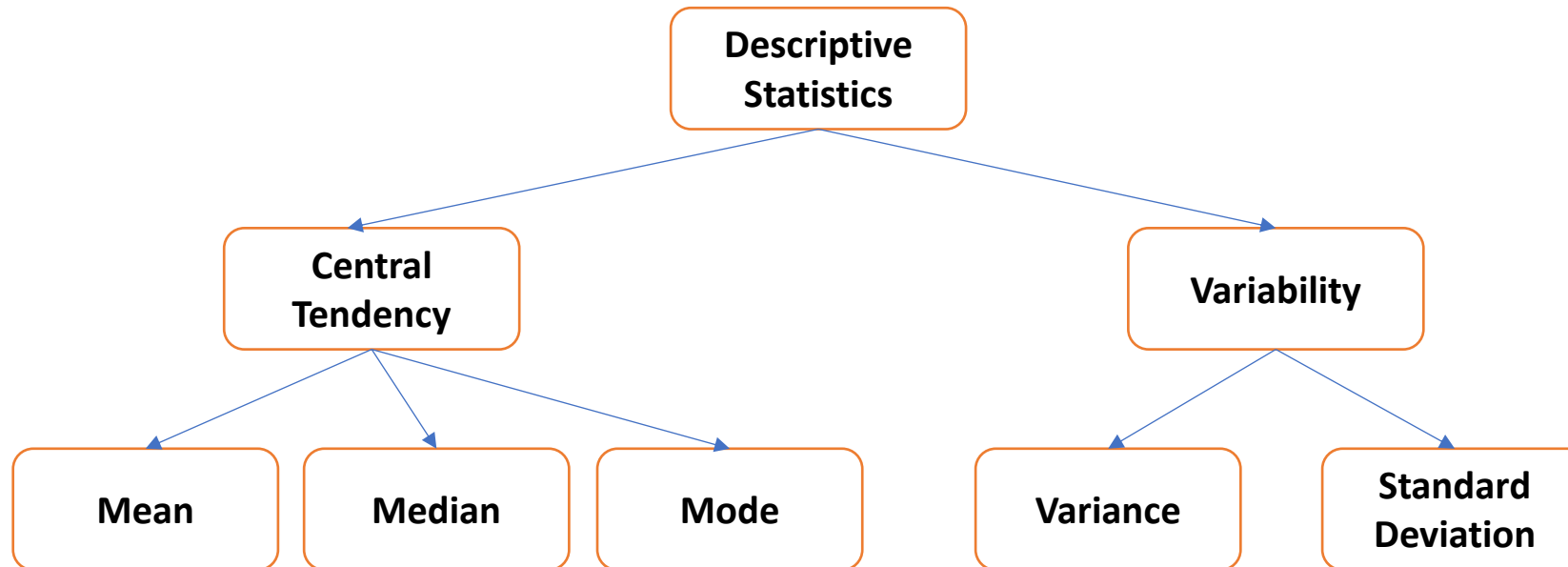
Continuous and Discrete Data

Discrete	Continuous
Discrete variables are numeric variable that have fixed intervals.	Continuous variables are those which are continuous numeric.
E.g. One bedroom, 5 persons. Note that these quantities cannot be further broken down into 1.2 bedrooms or 5.3 persons. Hence, they are termed as discrete.	E.g. Length or Weight. The length or weight can be in terms of the infinitesimally micro units such as 3.45678 mm length or 4.56789 kilograms. Hence these variables are termed as continuous.

Introduction to Descriptive Statistics

Descriptive Statistics is the field of describing or summarizing data. As statistics deal with data, it is important that we understand the various aspects of data to interpret it better.

Some of the common metrics that Descriptive Statistics use are Mean (Average), Median, Mode, Standard Deviation, etc.



Mean, Median and Mode

	Science Marks	Maths Marks	Average	CGPA
Student 1	67	87	77	A
Student 2	78	67	72.5	B+
Student 3	76	94	85	A+
Student 4	89	68	78.5	A
Student 5	56	85	70.5	B+
Student 6	78	70	74	B+
Student 7	92	55	73.5	B+
Student 8	88	79	83.5	A+
Student 9	69	59	64	B
Student 10	75	78	76.5	A
Median	77	74	75.25	
			Mode	B+

Standard Deviation

Standard Deviation: While mean, median and mode are metrics used in finding central tendency of the data, there is another metric called Standard deviation that is used to calculate the spread of the data. In other words, standard deviation measures how far away, the data is (deviation) on an average from the central point.

Standard Deviation is calculated as *root over of sum of the squares of differences between individual data points from the average of the overall data divided by the number of data points.*

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

σ = Population Standard Deviation

N = Size of the Population

x_i = Individual Values of the Population

μ = Population Mean

	Marks	Square of Diff
Student 1	67	96.04
Student 2	78	1.44
Student 3	76	0.64
Student 4	89	148.84
Student 5	56	432.64
Student 6	78	1.44
Student 7	92	231.04
Student 8	88	125.44
Student 9	69	60.84
Student 10	75	3.24
AVERAGE	76.8	110.16
Standard Deviation		10.50

Population SD vs Sample SD

- There is a slight difference between the population Standard Deviation and Sample Standard Deviation. The individual formulae are below:
- We notice that the denominator of the sample standard deviation is $N - 1$ rather than N which is the case for Population Standard Deviation (wherein N is the number of data points).
- This means that the Sample standard deviation will be slightly higher than Population standard deviation

Population standard deviation:

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

Sample standard deviation:

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

Random Variables

- **Random Variables** or **Stochastic Variables** in statistical parlance are variables whose values are outcomes of a random event.
- For example, the “*100-meter sprint timings*” of world’s fastest man, Usain Bolt from his last 10 competitions. Each of these timings are random and unpredictable. Random Numbers are important in Statistics in relation to Probability Theory. From the example of Usain Bolt’s sprint timings, you may guess that although the timings of each sprint are different, most scores are around a middle value or mean.

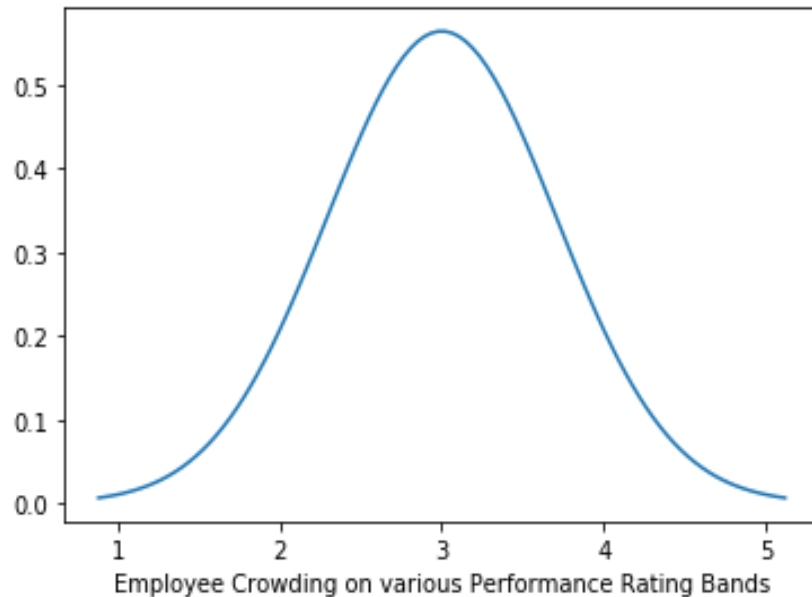
Most of the data attributes we normally deal with are Random Variables.

Sprint Number	100 Meter Timings (in Seconds)
Sprint 1	9.8
Sprint 2	10.2
Sprint 3	10.3
Sprint 4	11.1
Sprint 5	11
Sprint 6	10.8
Sprint 7	9.8
Sprint 8	10.4
Sprint 9	11.1
Sprint 10	11

Normal Distribution

Normal Distribution or Gaussian Distribution or Laplace-Gauss Distribution is a pattern of random data based on their occurrence.

The normal distribution looks like a 'bell' and is also popularly known as the 'bell-curve'. One of the most known usage of the bell-curve in the industry is possibly the fitment of employees into various performance bands at the time of appraisal.



The Normal-Curve pattern is a naturally occurring phenomena across the industry, observed among many types of data. Some of the situations where we observe a Normal Distribution among data patterns are:

- Quality Control of products (e.g., average lead content in soft drinks).
- Income levels of people in a state.
- People's buying patterns in an e-Commerce site (in terms of how much people spend on an average per month – the pattern will show a small number of people spending very high and a small number of people spending very less, and data showing a central tendency towards an average).

Histograms

The idea of the normal curve has come from the concept of Histograms. Histograms are plots of frequency (number of occurrences of data in particular range of values).

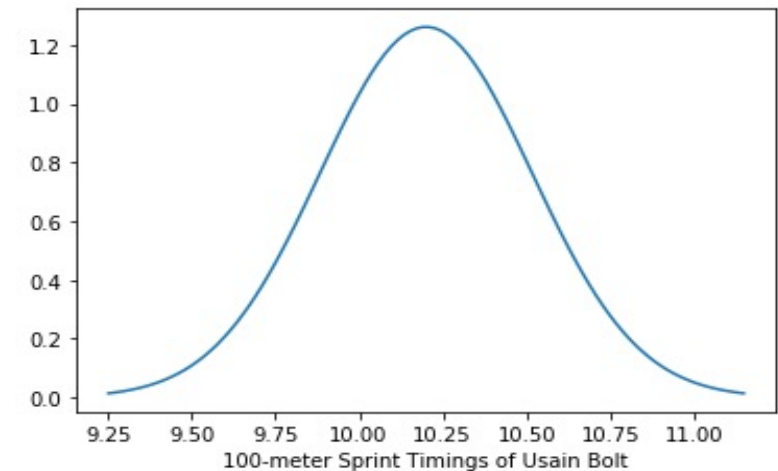
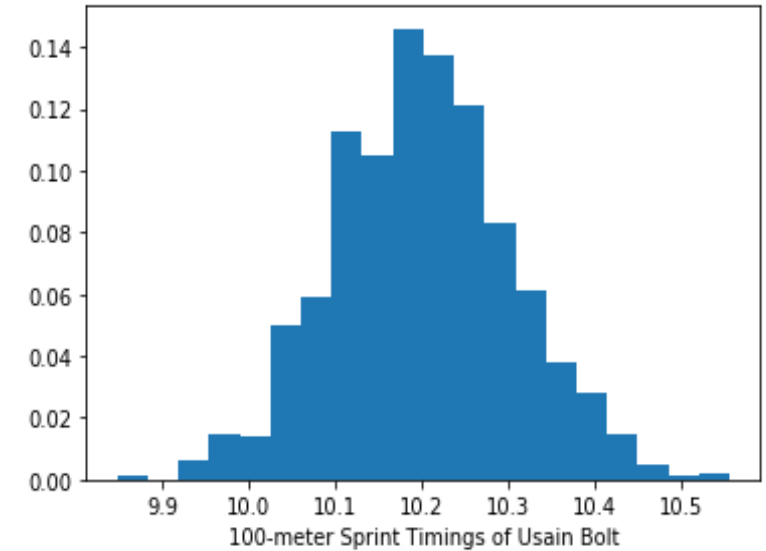
Taking the Sprint timings of Usain Bolt (fictitious values), we plot a Histogram to see how it looks like.

We have taken 1000 samples of his sprint timings and have plotted the number of times he has clocked those timings. The frequencies or the number of occurrences are taken in a range of values (timing). This range of value is called **class intervals**.

We observe that the **frequencies of data are centered symmetrically around the average** of ~10.2 seconds.

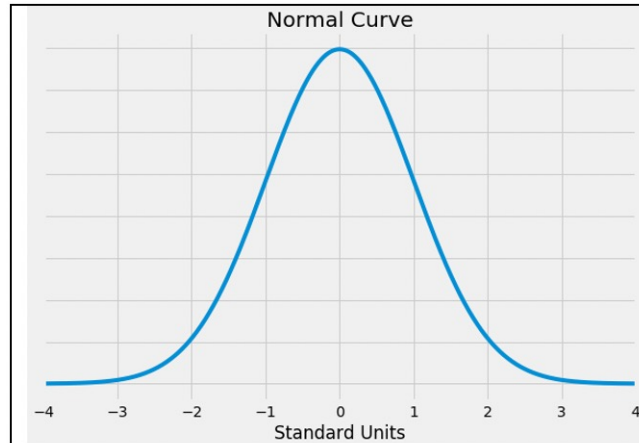
Each of the bars in the histogram reflect the number of times values in the corresponding class interval has occurred.

The Histogram above is a good example of a Normal Curve, when it is transformed into a smooth curve to represent the same data, it looks like this >>>



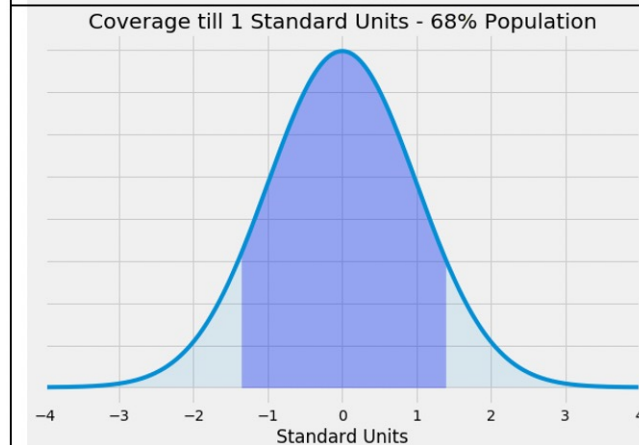
Normal Approximation of Data

- There are several characteristics of the Normal Curve that leads to its statistics usefulness using probability theories.
- Let us remember that while Mean is the measure of the central tendency of the data, the variance and standard deviation are the measures of the spread.
- We introduce the concept of Standard Units here. Standard Units are the SD value of data in terms of how many Standard Deviations the data is lower or higher from the center or mean.



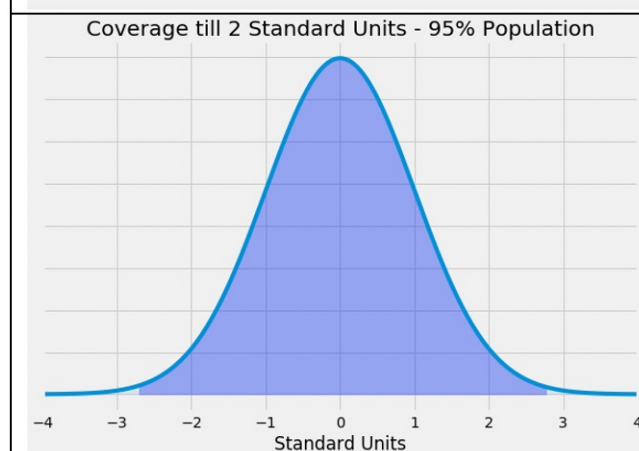
In this Normal curve, the x-axis labels are standard units, i.e. they denote how many Standard Deviations away that point in the curve is from the mid-point.

We know already that the area under the bell covers 100% of the data, and the normal curve follows a symmetrical pattern around the mean.



The nature of the Normal Curve is such that, we will find 68% of the population within the area under the bell between -1 and 1 Standard Units (or Standard Deviations), 68%.

The dark blue shaded region in this example represent this 68% population lying between -1 and 1 Standard Units.



In similar lines, 95% of the population lie between the area between -2 and 2 Standard Units under the bell.

The dark blue shaded area in the normal curve in this example cover this 95% of the population.

Normal Approximation of Data

The distribution of the data in a normal curve will be symmetrical / equal on both sides of the mean line (SD = 0).

These rules can now be used to calculate the percentage of entries/data in an interval (between given SD / SU values).

Now that you understand the basic features of a normal curve, let us note the following points:

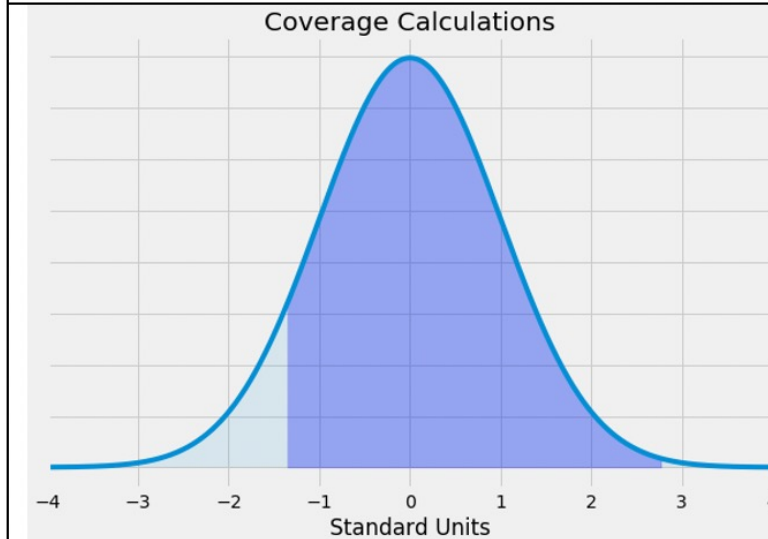
- We create frequency distributions using Histograms. Histograms capture the number of entries or occurrences of data in each class interval.
- The class intervals are in the scale of the data being represented. In relation to Usain Bolt's timing example, the class interval has the scale unit of seconds which is along the x-axis.
- The y-axis in a frequency histogram capture the number of occurrences in each interval.
- Once we have a frequency histogram, we can convert the scale of the x-axis from the scale/unit of the original data, to Standard Deviations or Standard Units.
- After the scale conversion to SD is done, we smoothen the histogram into a normal curve.

The overall process above is called '**Normal Approximation**'.

- **68% of data will lie between Standard Deviations -1 to 1**
- **95% of the data will lie between SD -2 to 2**
- **99.7% of the data will lie between SD -3 to 3**

Normal Approximation of Data

Relationship between the %age of data under the bell and SD value ranges >>>



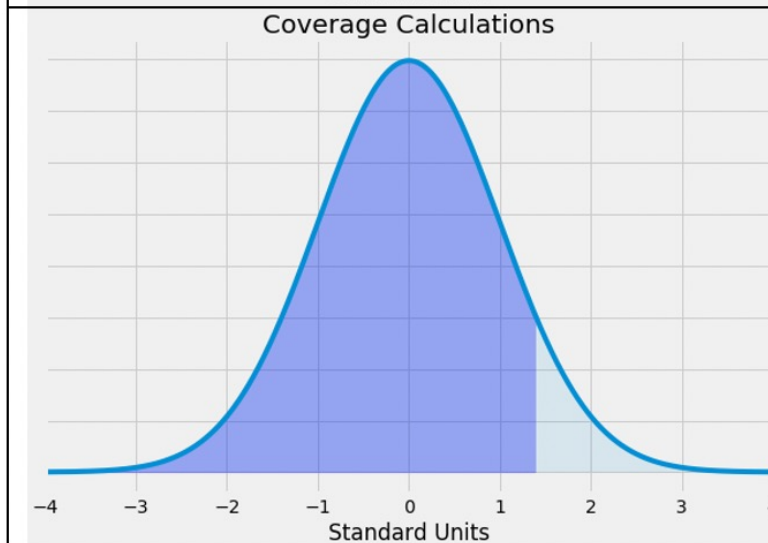
We are to calculate what %age of the data lie between SD -1 to SD 2 as reflected by the shaded region in this example.

To do this, we have to add the %age of data lying between SD 0 to 2 and SD -1 to 0.

We know that 95% data lie between SD -2 and 2. Hence, 47.5% of the data lie between SD 0 and 2.

Likewise, 68% data lie between -1 to 1. Hence, 34% of the data lie between SD -1 to 0.

Hence %age of data lying between SD -1 to 2 is 81.5%.



We are to calculate what %age of the data lie between SD -3 to SD 1 as reflected by the shaded region in this example.

To do this, we have to add the %age of data lying between SD 0 to 1 and SD -3 to 0.

We know that 68% data lie between SD -1 and 1. Hence, 34% of the data lie between SD 0 and 1.

Likewise, 99.7% data lie between -3 to 3. Hence, 49.85% of the data lie between SD -3 to 0.

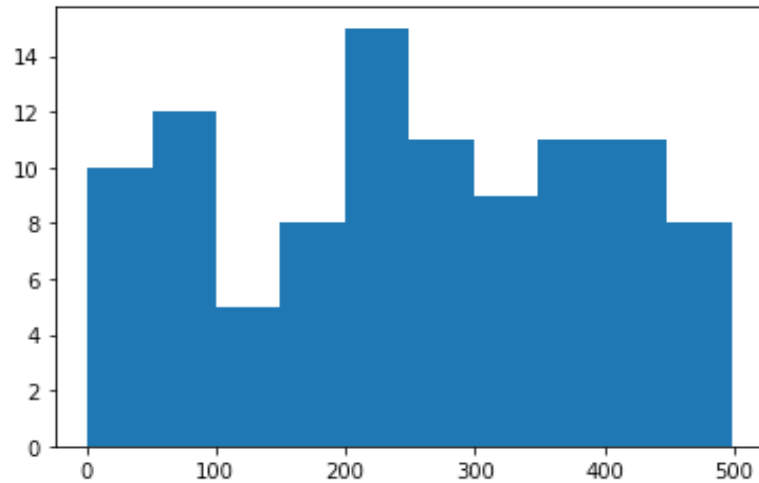
Hence %age of data lying between SD -3 to 1 is 83.85%.

Non-normal Distribution

Understandably, ***not all data populations are normal distributions.***

When the population is not normal, then we cannot apply the 68%, 95%, 99.7% rules on the data population.

Example of a non-normal distribution.



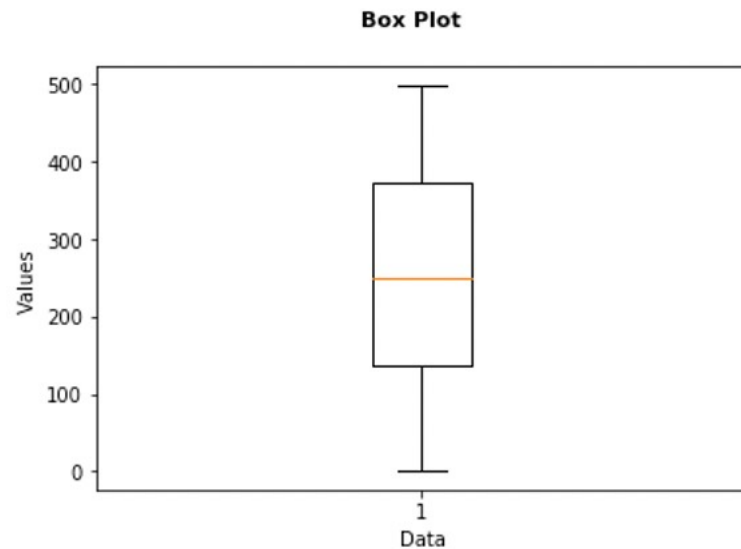
Percentiles and Box Plots

Percentiles convey ***what %age of data are covered within the specified percentile value.***

E.g., Data in 10th percentile means the range of data that have values less than or equal to the bottom 10% of the data values. 25th, 50th and 75th percentiles in data populations have special significance in statistics as they are generally accepted benchmark ranges of data that are often used for comparisons.

The range between 25th to 75th percentiles is called the ***interquartile range.***

Normal Distributions are visually represented using the Normal Curve. ***Percentiles of data are demonstrated by using Boxplots.***



The Boxplot in this example is to be interpreted as –

- 25% of the data values are between the values of 0 – 125 (approx.)
- The interquartile range, i.e. the middle 50% of the data lie between the values of 125 – 375 (approx.)

The top 25%. Of the data lie between the values of 125 – 500 (approx.)

Central Limit Theorem

Another key concept in relation to distributions of data is the **Central Limit Theorem**. Central Limit Theorem states that if we have a population of data, and we draw samples from this population, then the population of the means of all samples drawn will be approximately Normally Distributed.

- The original population need not be Normally Distributed
- The number of samples have to be sufficiently large
- Individual sample sizes should be > 30
- Drawing of Samples is with “replacement”, i.e. after a sample is drawn and mean / SD calculated, the sample is to be placed back into the original population

When σ is the Standard Deviation of the Original Population, the Standard Deviation of the Sample Means is calculated as, considering n being the sample size, in a random distribution:

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

E.g. if the standard deviation of the population is 2.5 and the sample size is 40, then the Standard Deviation of the Population means will be 0.4.

Central Limit theorem comes into use in a variety of scenarios in statistics, because of its treatment of the data when the distribution is not normal and do not follow the normal distribution rules. Central Limit Theorem helps draw inferences in Hypothesis Testing and Confidence Intervals because we can work with the Sampling Distribution rather than the original population itself.

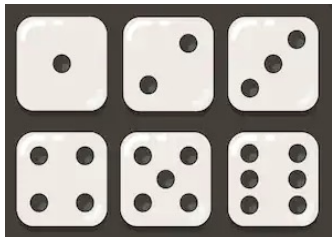
Probability Theory

We talk about “**chance**” all the time. What are the chances of rain tomorrow or the chances of person X becoming the president of our nation. Although naïve it may sound, there are mathematical interpretations of chances using which we can deduce probabilities of certain events to take place.

Roll of a Dice: A common example in the calculation of **probabilities** is the roll of a dice event.

A dice has six sides, each of the sides are dotted with 1 – 6 spots. When you roll the dice, there is equal possibility or probability of any of the sides turning up.

Hence, if you are to find the probability of getting the side with 3 dots, it would be $1/6$ or 16.7% or 0.17. Note that, it's the same for any of the sides.



Note: The probability of something is denoted by the percentage of time it is expected to happen given that the experiment or process is repeated keeping the conditions of the experiment same.

Probability is a value between 0 (not possible) or 1 (certainty). It can also be denoted as values between 0% to 100%. Probability is often denoted by the letter 'P'.

The chance of an opposite event happening is calculated as one minus the probability of the original outcome.

Conditional Probability

Conditional Probability is defined as the *probability of an event happening given that another event has occurred.*

E.g., We have a bag with 5 green balls and 4 red balls in it. We are to find out, if we are to draw two balls in succession, what is the probability of the second ball being red.

There could be various scenarios:

- **Scenario 1:** The first ball is drawn, the color is noted and placed back in the bag (this is called 'replacement'). The bag still has the same number of balls in the same combination of colors. The number of red balls being 4 out of total 9 balls, the probability of the second ball being red is calculated as $\frac{4}{9}$ or 0.44 or 44%.
- **Scenario 2:** The first ball is drawn, and it is green and is not placed back in the bag. Now let's ask what is the probability of the second ball being red given that the first ball is green. We now have 4 green balls and 4 red balls in the bag. The probability of the second ball being red will be 50%.
- **Scenario 3:** The first ball is drawn, and it is red and is not placed back in the bag. Now let's ask what is the probability of the second ball being red given that the first ball is also red. We now have 5 green balls and 3 red balls in the bag. The probability of the second ball being red will be $\frac{3}{8}$ or 0.375 or 37.5%.

Note that if we are tweak the original question as “what is the probability of the second ball being red given the first ball of green”, the calculation differs. The outcome will be dependent on the outcome of the first event.

Whereas in the first scenario, the probability of the second event happening did not depend on the outcome of the first event because of the replacement.

This is known as “Conditional Probability”.



The Multiplication Rule

If we are to calculate the probability of two events happening, we use the **Multiplication Rule**.

Examples:

Ex1: The probability of a customer buying an apparel from Amazon in a particular month is 32% and the probability of him/her buying an electronic item is 17%.





What is the probability that the customer will buy both an apparel and an electronic item in the particular month? The result is calculated as the multiplication of the two probabilities, in this case it is $32\% \times 17\% = 0.054$ or 5.4%.

Ex2: Calculate the probability of getting a  followed by a  in two consecutive rolls of dice. Using the multiplication rule and using the assumption that each side has an equal probability of appearing in a roll, the result of the question is $1/6 \times 1/6 = 0.0277$ or 2.77%.

The Addition Rule

If we are to find the probability of any one of two mutually exclusive events to happen, we add the individual probabilities.

Two events are called **Mutually Exclusive**, when the occurrence of one prevents the other, i.e., they cannot both happen together. E.g. In the toss of a coin, heads and tails are mutually exclusive. If a bag has three balls of colors green, red and white, then drawing of any one ball is mutually exclusive with the others because any one of the colors will get drawn.

Let's take another example to understand the **Addition Rule** better. We are to find out the probability of drawing a  or a  this in a roll of dice. We know there are 6 equal possibilities in a roll of dice. Individually, each of these possibilities have probabilities of $1/6$ or 16.66%. Applying the Addition Rule, we find that the total probability of getting a  or a  will be $16.66\% + 16.66\% = 33.33\%$.

*However, the important point to keep in mind that the Addition Rule to apply, the events in question have to be **Mutually Exclusive**.*

E.g., If we are to calculate the probability of a customer buying an apparel item or an electronic item from Amazon, whose individual probabilities are 17% and 6% respectively, we cannot apply the Addition Rule. The reason is that these two events are not mutually exclusive, the customer may well buy both.

The Binomial Formula

We introduce a new range of problems such as:

- **A ball is drawn 4 times from a bag containing 5 green balls and 4 red balls. What is the probability that exactly 2 balls will be red.**
- **A dice is rolled 4 times. What is the probability that we will be exactly 2 numbers of 5's.**

This kind of problems are solved using what is called the ***Binomial Coefficient*** and the ***Binomial Formula***.

There are three steps to find out the solution:

- Find out the number of possible patterns of outcomes, i.e., the Binomial Coefficient.
- Find out the probability of the desired combination of events occurring.
- Multiply the Binomial Coefficient by the above probability.

The Binomial Formula

Step 1: The number of possible outcome patterns from the experiment are obtained using the Binomial coefficient, i.e.

$$\binom{n}{k} = \frac{n!}{k! \cdot (n - k)!}$$

Where ***n*** is the number of trials (in this case, $n = 4$) and ***k*** is the number of times the event is to occur (here, $k = 2$).

Hence the number of patterns in which there will be 2 red balls and 3 green balls using the Binomial Coefficient will be calculated as $\Rightarrow (4!) / \{(2!) \times (4-2)!\} = 6$.

Binomial Formula

Step 2: Find out the total probability of the events occurring:

The individual probability of a red ball getting drawn is 4/9 and that of a green ball is 5/9. For exactly two red balls to be drawn in 4 total draws, the combination will comprise of 2 red balls and 2 green balls. To find out the probability of this combination of 2 reds and 2 greens, we use the multiplication rule.

This is calculated as $(4/9)^2 \times (5/9)^2 = 0.0609$ or **6.09%**.

Step 3: Multiply the number of patterns (Step 1) with the desired events probability (Step 2) to obtain the probability of exactly 2 red balls in 5 total draws from a bag containing 4 red balls and 5 green balls = $0.0609 \times 6 = 0.3657$ or **36.57%**.

This is the Binomial Formula:

$$\frac{n!}{(n-x)!x!} p^x q^{n-x}$$

n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

$q = 1 - p$ = the probability of getting a failure in one trial

Expected Value

Each of the chance processes produce a quantifiable outcome. A series of chance processes produce a cumulative outcome which is called the Expected Value for those many processes.

E.g., In a Raffles draw, there are 1000 tickets issued, each worth \$10. The winner gets \$1000. If you buy 10 tickets, how much can you expect to win?

Solution: A ticket is a chance event in this case. Each ticket being 1 out of total thousand will have a possibility of winning \$1 (\$1000 winning amount divided by 1000 number of tickets). So, there is a possibility of winning \$10 if you buy 10 tickets. ***Here \$10 is the Expected Value.***

Expected Value (EV) = Number of Events x Average Possible Value on each Event

Standard Error

Standard Error is the amount by which the Observed Value is off from the Expected Value.

Let's take the example of drawing a numbered card from a box with replacement. There are 5 cards in the box with numbers 1, 2, 3, 4, 5. In each draw, each card would have equal probability of getting drawn. With each draw, we note the numbers and place it back. In total, we make 25 draws.

Let's calculate the Expected Value first. Each of the numbered cards are expected to be drawn 5 times.

Total Expected Value (EV) = $(1 \times 5) + (2 \times 5) + (3 \times 5) + (4 \times 5) + (5 \times 5) = 75$.

However, in reality, the number of times each card gets drawn in this experiment may not be same as in the calculation of Expected Value. Let say, it looks like this – 1 was drawn 4 times, 2 was drawn 8 times, 3 was drawn 5 times, 4 was drawn 4 times and 5 was drawn 4 times.

Hence, the **Observed Value will be** = $(1 \times 4) + (2 \times 8) + (3 \times 5) + (4 \times 4) + (5 \times 4) = 71$.

This difference of (-4) is the Chance Error and is also called the Standard Error.

Observed Value = Expected Value + Standard Error

There is a calculation called the Square Root Law to **find the likely Standard Error for a process.**

$$\text{Standard Error} = \sqrt{\text{Number of Draws} * (\text{SD of the Values})}$$

Hypothesis Testing / Tests of Significance

“Was it due to Chance or Real?”

--- *Hypothesis Testing or Tests of Significance* are devised to answer this sort of questions

Example

Let us consider the example of testing the amount of Paracetamol in the Analgesic Medicine that a distributor has received. There should be 250mg Paracetamol in each tablet. Sample testing is carried out to ascertain the levels. It is observed from a sample of 500 tablets, that the Paracetamol level is 245mg on an average in the sample.

Question:

Is the difference between the ‘**Expected Value**’ and ‘**Observed Value**’ significant or by chance?

Hypothesis Testing

Steps to answer the question:

1. Find out the Average of the Sample	Found out to be 245mg
2. Find the Standard Deviation (SD) of the Sample	Found to be 37
3. Find the Standard Error (SE) for the Sum of the Sample	$\sqrt{500} * 37 = 827.34$
4. Find out the Standard Error for the Average of the Sample	$827.34 / 500 = 1.65$
5. Number of SE, the Sample Average is away from the Expected Average	$(250 - 245) / 1.65 = \mathbf{3\ SE}$

Given that the 'observed value' is -3 Standard Errors (SE) away from the 'expected value', the difference is statistically significant.

$$\text{Number of SE} = (\text{Observed Value} - \text{Expected Value}) / \text{SD}$$

Hypothesis Testing

The results indicate that the **difference is -3 Standard Errors**. That means that the **observed value** has a large deviation from the expected value and hence difficult to explain it as a chance. In another way, the **-3 Standard Errors indicate a 99.7% Confidence Level** to say that the Difference is Statistically Significant (Normal Distribution rules).

The above test is called a “**Test of Significance**”. Tests of Significance are used to validate a **Hypothesis** around the average value or a proportion of a population applying statistical calculations on the sample.

In Tests of Significance, we employ two types of Hypothesis – one complimentary to the other.

Null Hypothesis: The status quo – in the above example, the Null Hypothesis is that the average of the population is 250mg (known information). Null Hypothesis is also denoted as H_0 .

Alternate Hypothesis: The alternate hypothesis is the compliment of the Null Hypothesis. In the above example, the alternate hypothesis is that the average is not 250mg. Null Hypothesis is also denoted as H_1 .

There is a certain convention of framing the Null and Alternate Hypothesis. The Null Hypothesis is always the theory that is the status quo or the known or commonly believed information.

Hypothesis Testing

Examples of Null Hypothesis:

- In a Criminal Trial: A person is not a criminal.
- Quality Control: The average Paracetamol content of the Analgesic tablet is 250mg (i.e., the expected value).
- The proportion of defective items in two samples of a population are same.
- That a particular attribute in a dataset doesn't have influence on a response variable.

The Null Hypothesis corresponds to the idea that the idea that the observed difference is due to chance.

The **Alternate Hypothesis** is what **contradicts the Null Hypothesis**.

Examples of Alternate Hypothesis:

Against the above examples, we can frame the alternate hypothesis statements as:

- The person is a criminal.
- The average Paracetamol content of the Analgesic tablet is not 250mg or less than 250mg.
- The proportion of defective items in two samples of a population are not same.

The Alternate Hypothesis corresponds to the idea that the idea that the observed difference is real.

The Null and the Alternate Hypothesis are statements about the '**Population**' and not the '**Sample**'.

The Hypothesis Testing is carried out by data collected from the 'Sample'.

The hypothesis tests have one of two goals or outcomes by convention:

- **Reject the Null Hypothesis**
- **Fail to Reject the Null Hypothesis**

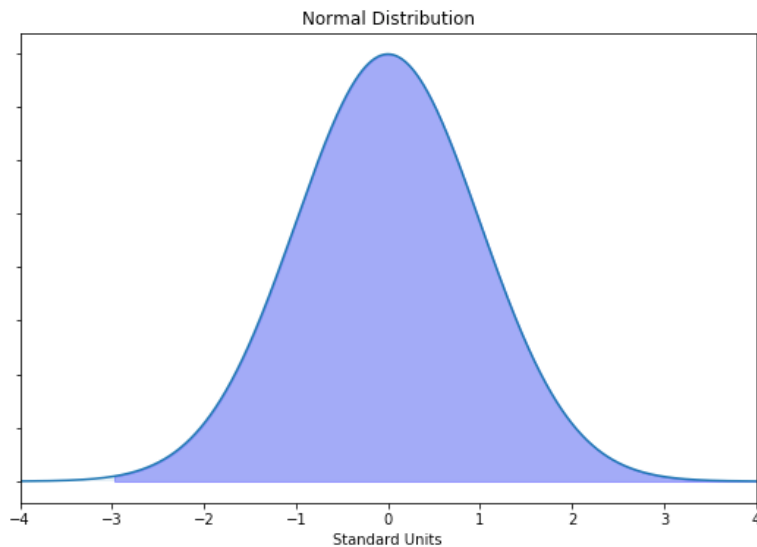
Test Statistic and Significance Level

In the previous example, **-3 is the 'test-statistic'**.

This test which is also known as the '**z-test**', uses this value as the **test-statistic**.

The z-value test-statistic (-3 here) indicates as to how many standard errors (SE) away an observed value is from an expected value where the expected value is part of the null hypothesis.

A 'test-statistic' is the measure of the difference between 'expected' and 'observed' values.



Now, let's relate the test-statistic or z-statistic in this case to the normal distribution curve.

Given that the deviation of the observed value is -3 SE further from the expected value, we plot the coverage on the Normal Curve. The area on the left beyond the value of -3 SE is 0.0013 or 0.13%.

This is called the **Significance Level** or **P-value**.

The interpretation of the P-value is that it is the percentage of probability of the Null Hypothesis to be True.

Using Z-Table

Now there is one question, how did we get the value 0.0013 corresponding to the SE value of -3?

This is from the **z-table**. **Z-table** or the **Standard Normal Table** that tell us the percentage of area covered at the left of a given z-score in a Normal Distribution.

Below are a partial snapshot of the '**z-table**' for -ve and +ve values.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064

In the z-table displayed here, the value corresponding to -3 is 0.0013.
This is the area at the left of -3 SE in the standard normal curve.

Hypothesis Testing: 'cut-off' value

In hypothesis tests, a '*cut-off value of the test is stated to decide whether to reject the Alternate Hypothesis and Accept the Null Hypothesis.*

As a standard, it is 5% or in some cases a value of 1% is taken.

- The '**cut-off value**' is compared against the **P-Value**.
- In the above example, we have got a P-value of 0.13% which is less than 5%.
- If the P-Value is less than the cut-off, we reject the Null Hypothesis. In the above case of the Paracetamol content check, we reject the null hypothesis of "The average Paracetamol content of the population of tablets is 250mg".
- P-Value of $< 1\%$ indicates a highly significant difference.

Summary: Steps in Hypothesis Testing

In summary of the above section, we can say that we have to perform the following steps to conduct a Hypothesis Testing experiment:

- Step 1: State your null hypothesis and alternate hypothesis - H_0 and H_1
- Step 2: Decide on what test statistic. Apart from the z-score that we have used here, there are a number of other kinds of Hypothesis tests possible and some of them are detailed in the following sections.
- Step 3: Decide on the cut-off %age
- Step 4: Compute the test-statistic
- Step 5: Determine Acceptance/Rejection regions in the Normal Curve
- Step 5: Based on steps 3, 4 and 5, draw conclusion on H_0 – Null hypothesis

**Hope you have liked this Video.
Please help us by providing your Ratings and Comments for this
Course!**

**Thank You!!
Manas Dasgupta**

Happy Learning!!

