UNIT - 2
Data and Sampling Distribution

SYLLABUS –

Random Sampling and Sample Bias, Selection Bias, Sampling Distribution of a Statistic, The Bootstrap, Confidence Intervals, Normal Distribution, Long-Tailed Distribution, Student's t-Distribution, Binomial Distribution, Chi-Square Distribution, F-Distribution

What is sampling?

- Statistics defines sampling as the process of gathering information about a population from a subset, like a selected individual or a small group and analyzing that information to study the whole population.
- The sample space constitutes the foundation of data which in turn is responsible for determining the accuracy of the study or research.
- Sampling, however, is not as simple as it seems. To land an accurate result, the sample size needs to be accurate, followed by implementing the right sampling methods based on the sample size.

Sampling Steps

An analyst needs to follow certain steps in order to reach conclusions from a broader perspective. The Sampling steps include the following -

Step 1: Identity and clearly define the target group/population.
Step 2: Create a specific sampling frame.
Step 3: Select the right sampling methods to be used.
Step 4: Specify the sample size.
Step 5: Collect the required sampled data.

What is Random Sampling? Define the key terms related to Random Sampling?

- A ***sample*** is a subset of data from a larger data set; statisticians call this larger data set the population. A population in statistics is not the same thing as in biology—it is a large, defined (but sometimes theoretical or imaginary) set of data.
- ***Random sampling*** is a process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw. The sample that results is called a simple random sample.
- ***Sampling*** can be done with replacement, in which observations are put back in the population after each draw for possible future reselection. Or it can be done without replacement, in which case observations, once selected, are unavailable for future draws.
- Data quality often matters more than data quantity when making an estimate or a model based on a sample.
- Data quality in data science involves completeness, consistency of format, cleanliness, and accuracy of individual data points. Statistics adds the notion of representativeness.
- Key Terms for Random Sampling are

Key Terms for Random Sampling

Sample
A subset from a larger data set.

Population
The larger data set or idea of a data set.

N (n)
The size of the population (sample).

Random sampling
Drawing elements into a sample at random.

Stratified sampling
Dividing the population into strata and randomly sampling from each strata.

Stratum (pl., strata)
A homogeneous subgroup of a population with common characteristics.

Simple random sample
The sample that results from random sampling without stratifying the population.

Bias
Systematic error.

Sample bias
A sample that misrepresents the population.

- Simple random sampling gives each member of the population an equal chance of being chosen for the sample.
- It's similar to drawing a name out of a bowl. Simple random sampling can be performed by anonymizing the population, for example, assigning a number to each object or person in the population and selecting numbers randomly.
- Simple random sampling eliminates any bias from the sampling process and is inexpensive, simple, and quick to use.
- It also provides the researcher with no means of control, increasing the likelihood that unrepresentative groupings will be chosen randomly.

Applications

- Lottery techniques,
- Split and Train in machine learning.

Advantages

- Little bias due to the random nature of the sample collection
- Given the usage of random generators, sample selection is straightforward.
- Due to representativeness, the findings can be broadly interpreted.

Disadvantage

- All responders' potential availability might be expensive and time-consuming.
- Huge sample size

Explain Bias?

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

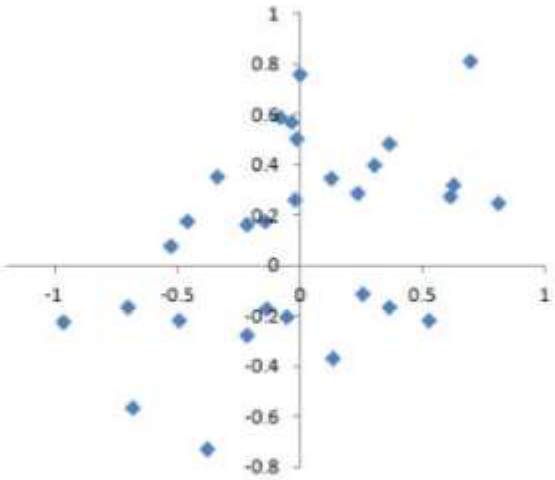


Figure. Scatterplot of shots from a gun with true aim

- The results shown in Figure show a biased process—there is still random error in both the x and y direction, but there is also a bias. Shots tend to fall in the upper-right quadrant.

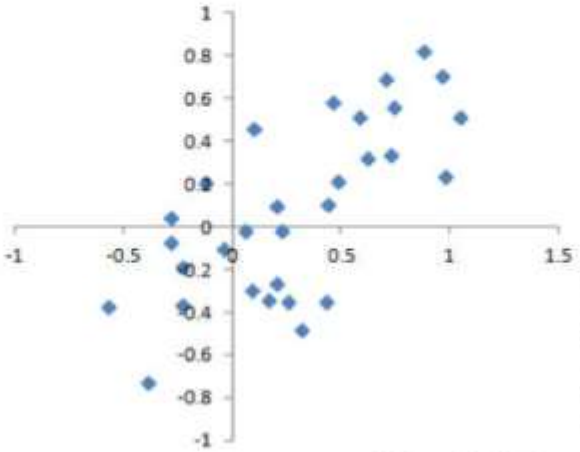


Figure- Scatterplot of shots from a gun with biased aim

- Bias comes in different forms, and may be observable or invisible. When a result does suggest bias (e.g., by reference to a benchmark or actual values), it is often an indicator that a statistical or machine learning model has been misspecified, or an important variable left out.

Explain Sample Mean Versus Population Mean?

- The symbol $\bar{x}$ (pronounced "x-bar") is used to represent the mean of a sample from a population, whereas μ is used to represent the mean of a population.
- Why make the distinction? Information about samples is observed, and information about large populations is often inferred from smaller samples.
- Statisticians like to keep the two things separate in the symbology.

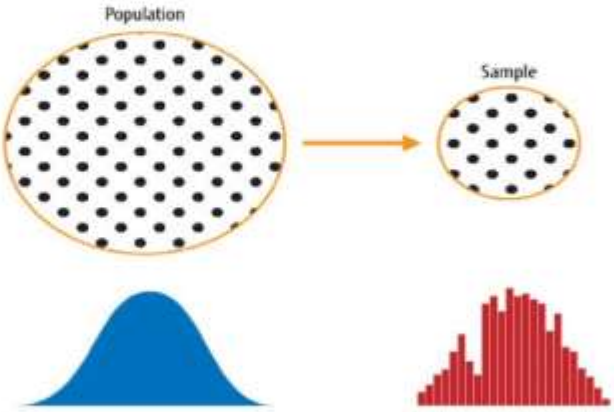


Figure. Population versus sample

What is Selection Bias? Define the key terms related to Selection Bias?

- Selection bias refers to the practice of selectively choosing data consciously or unconsciously in a way that leads to a conclusion that is misleading or ephemeral.

Key Terms for Selection Bias

Selection bias
Bias resulting from the way in which observations are selected.

Data snooping
Extensive hunting through data in search of something interesting.

Vast search effect
Bias or nonreproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables.

- If you specify a hypothesis and conduct a well-designed experiment to test it, you can have high confidence in the conclusion.
- This is frequently not what occurs, however. Often, one looks at available data and tries to discern patterns.
- But are the patterns real? Or are they just the product of data snooping—that is, extensive hunting through the data until something interesting emerges?
- There is a saying among statisticians: "If you torture the data long enough, sooner or later it will confess."

Sampling Distribution of a Statistic
Or
Define Sample statistic, Data distribution, Sampling distribution, Central limit theorem, and Standard error?

- The term sampling distribution of a statistic refers to the distribution of some sample statistic over many samples drawn from the same population.
- Much of classical statistics is concerned with making inferences from (small) samples to (very large) populations.

Key Terms for Sampling Distribution

Sample statistic
A metric calculated for a sample of data drawn from a larger population.

Data distribution
The frequency distribution of individual *values* in a data set.

Sampling distribution
The frequency distribution of a *sample statistic* over many samples or resamples.

Central limit theorem
The tendency of the sampling distribution to take on a normal shape as sample size rises.

Standard error
The variability (standard deviation) of a sample *statistic* over many samples (not to be confused with *standard deviation*, which by itself, refers to variability of individual data *values*).

- Typically, a sample is drawn with the goal of measuring something (with a sample statistic) or modeling something (with a statistical or machine learning model). Since our estimate or model is based on a sample, it might be in error; it might be different if we were to draw a different sample.
- We are therefore interested in how different it might be a key concern is sampling variability. If we had lots of data, we could draw additional samples and observe the distribution of a sample statistic directly.
- Typically, we will calculate our estimate or model using as much data as is easily available, so the option of drawing additional samples from the population is not readily available.
- *It is important to distinguish between the distribution of the individual data points, known as the data distribution, and the distribution of a sample statistic, known as the* ***sampling distribution.***
- The distribution of a sample statistic such as the mean is likely to be more regular and bell-shaped than the distribution of the data itself. The larger the sample the statistic is based on, the more this is true. Also, the larger the sample, the narrower the distribution of the sample statistic.

What is Central Limit Theorem and how bootstrap can be used to estimate the sampling distributions of a statistic?

- Central limit theorem is a statistical theory which states that when the large sample size has a finite variance, the samples will be normally distributed and the mean of samples will be approximately equal to the mean of the whole population.
- In other words, the central limit theorem states that for any population with mean and standard deviation, the distribution of the sample mean for sample size N has mean μ and standard deviation $\sigma / \sqrt{n}$.
- As the sample size gets bigger and bigger, the mean of the sample will get closer to the actual population mean.
- If the sample size is small, the actual distribution of the data may or may not be normal, but as the sample size gets bigger, it can be approximated by a normal distribution.
- This statistical theory is useful in simplifying analysis while dealing with stock indexes and many more.
- ***The central limit theorem states*** that whenever a random sample of size n is taken from any distribution with mean and variance, then the sample mean will be approximately normally distributed with mean and variance. The larger the value of the sample size, the better the approximation to the normal.

Assumptions of Central Limit Theorem

- The sample should be drawn randomly following the condition of randomization.
- The samples drawn should be independent of each other. They should not influence the other samples.
- When the sampling is done without replacement, the sample size shouldn't exceed 10% of the total population.
- The sample size should be sufficiently large.

The formula for the central limit theorem is given below:

Central Limit Theorem for Sample Means,

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

Sample size and the central limit theorem

- The **sample size** (n) is the number of observations drawn from the population for each sample.
- The sample size is the same for all samples.
- The sample size affects the sampling distribution of the mean in two ways.

1. Sample size and normality

- The larger the sample size, the more closely the sampling distribution will follow a normal distribution.
- When the sample size is small, the sampling distribution of the mean is sometimes non-normal. That's because the central limit theorem only holds true when the sample size is "sufficiently large."
- By convention, we consider a sample size of 30 to be "sufficiently large."
    - **When $n < 30$**, the central limit theorem doesn't apply. The sampling distribution will follow a similar distribution to the population. Therefore, the sampling distribution will only be normal if the population is normal.
    - **When $n \geq 30$**, the central limit theorem applies. The sampling distribution will approximately follow a normal distribution.

2. Sample size and standard deviations

The sample size affects the standard deviation of the sampling distribution. Standard deviation is a measure of the variability or spread of the distribution (i.e., how wide or narrow it is).

- **When *n* is low**, the standard deviation is high. There's a lot of spread in the samples' means because they aren't precise estimates of the population's mean.
- **When *n* is high**, the standard deviation is low. There's not much spread in the samples' means because they're precise estimates of the population's mean.

Conditions of the central limit theorem

The central limit theorem states that the sampling distribution of the mean will always follow a normal distribution under the following conditions:

1) The sample size is **sufficiently large**. This condition is usually met if the sample size is $n \geq 30$.
2) The samples are **independent and identically distributed (iid) random variables**. This condition is usually met if the sampling is random.
3) The population's distribution has **finite variance**. Central limit theorem doesn't apply to distributions with infinite variance, such as the Cauchy distribution. Most distributions have finite variance.

Explain bootstrap with suitable example?

- One easy and effective way to estimate the sampling distribution of a statistic, or of model parameters, is to draw additional samples, with replacement, from the sample itself and recalculate the statistic or model for each resample.
- This procedure is called the ***bootstrap***, and it does not necessarily involve any assumptions about the data or the sample statistic being normally distributed.
- **Bootstrapping** is a technique used to make estimations from data by taking an average of the estimates from smaller data samples.

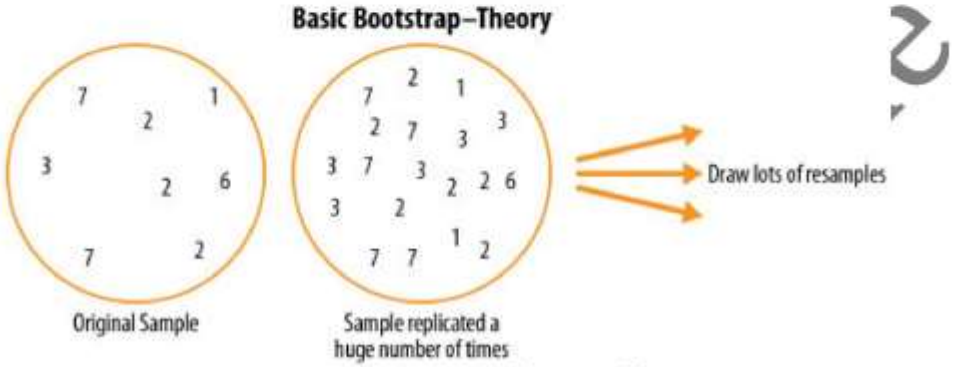
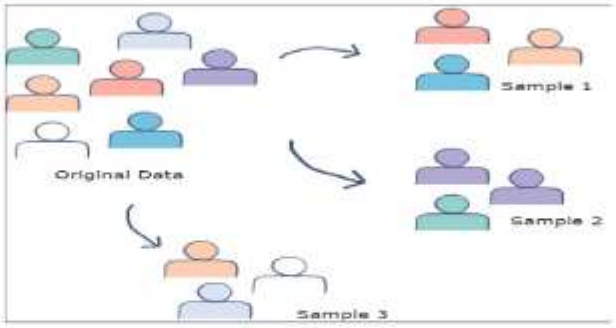


Figure - The idea of the bootstrap

- In practice, it is not necessary to actually replicate the sample a huge number of times. We simply replace each observation after each draw; that is, we sample with replacement.
- In this way we effectively create an infinite population in which the probability of an element being drawn remains unchanged from draw to draw.
- ***The algorithm for a bootstrap resampling of the mean, for a sample of size n, is as follows:***
  1) Draw a sample value, record it, and then replace it.
  2) Repeat n times.
  3) Record the mean of the n resampled values.
  4) Repeat steps 1–3 R times.
  5) Use the R results to:
     a. Calculate their standard deviation (this estimates sample mean standard error).
     b. Produce a histogram or boxplot.
     c. Find a confidence interval.

Method

- The bootstrap method involves iteratively resampling a dataset with replacement. Instead of only estimating our statistic once on the complete data, we can do it many times on a re-sampling (with replacement) of the original sample.
- Repeating this re-sampling multiple times allows us to obtain a vector of estimates. We can then compute variance, expected value, empirical distribution, and other relevant statistics of these estimates.

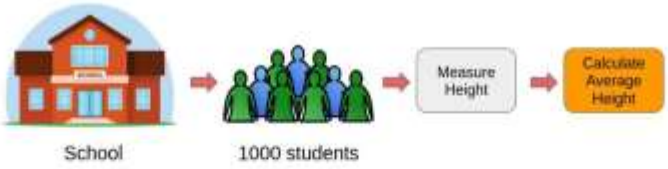


Uses

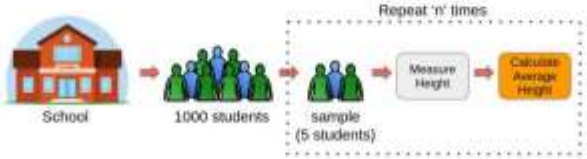
- Bootstrapping allows statistical inferences to be made about the population distribution of a small sample.
- It is used to account for distortions caused by certain sample data that could make for a bad representation of the overall data.
- In statistics, Bootstrap Sampling is a method that involves drawing of sample data repeatedly with replacement from a data source to estimate a population parameter.

Why Do We Need Bootstrap Sampling?

- This is a fundamental question I've seen machine learning enthusiasts grapple with. What is the point of Bootstrap Sampling? Where can you use it?
- Let's say we want to find the mean height of all the students in a school (which has a total population of 1,000). So, how can we perform this task?
- One approach is to measure the height of all the students and then compute the mean height. I've illustrated this process below:



- However, this would be a tedious task. Just think about it, we would have to individually measure the heights of 1,000 students and then compute the mean height. It will take days! We need a smarter approach here.
- This is where Bootstrap Sampling comes into play.
- Instead of measuring the heights of all the students, we can draw a random sample of 5 students and measure their heights. We would repeat this process 20 times and then average the collected height data of 100 students (5 x 20). This average height would be an estimate of the mean height of all the students of the school.
- Pretty straightforward, right? This is the basic idea of Bootstrap Sampling.



- Hence, when we have to estimate a parameter of a large population, we can take the help of Bootstrap Sampling.

What is Confidence Interval? Write an algorithm for a bootstrap confidence interval?

- The confidence interval is used to represent the **interval or range of values needed to match a confidence level** for **estimating** the **parameters of the entire population such as mean or proportion**.
- When there is a need to estimate about the population parameter, it is considered as a good practice to represent the estimate as a confidence interval. The population parameter generally represents the mean or median and proportion.
- And, the confidence level is represented using the number such as 98% confidence, 95% confidence etc.
- The confidence interval is associated with the confidence level represented using a number, say, N, and termed as an N% confidence interval. N can take values such as 99, 95, 90, etc.
- An N% confidence interval would mean the following – If an experiment to find the average height of male out of 100 male, is performed for, say, 50 times, the interval in which the average height will fall for 95% of times (45 times) will be between, say, 173 and 179 cm. Thus, a 95% confidence interval for average height will be 173 and 179 cm.
- In simple terms, Confidence Interval is a range where we are certain that true value exists.
- The selection of a confidence level for an interval determines the probability that the confidence interval will contain the true parameter value.
- This range of values is generally used to deal with population-based data, extracting specific, valuable information with a certain amount of confidence, hence the term 'Confidence Interval'.
- Fig. Shows how a confidence interval generally looks like.

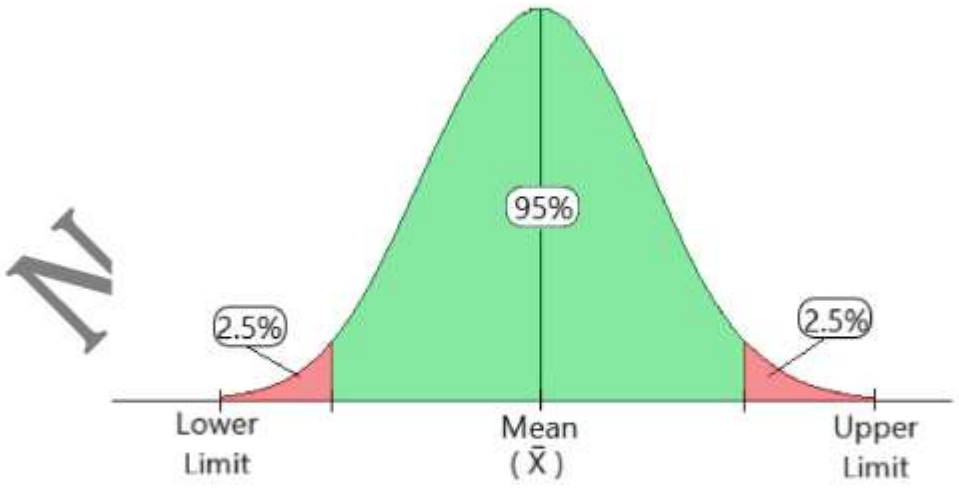


Fig 1: Confidence Interval Illustration

- Given a sample of size n, and a sample statistic of interest, the ***algorithm for a bootstrap confidence interval*** is as follows:

  1) Draw a random sample of size n with replacement from the data (a resample).
  2) Record the statistic of interest for the resample.
  3) Repeat steps 1–2 many (R) times.
  4) For an x% confidence interval, trim [(100-x) / 2] % of the R resample results from either end of the distribution.
  5) The trim points are the endpoints of an x% bootstrap confidence interval.

- The percentage associated with the confidence interval is termed the level of confidence. The higher the level of confidence, the wider the interval.
- Also, the smaller the sample, the wider the interval (i.e., the greater the uncertainty). Both make sense: the more confident you want to be, and the less data you have, the wider you must make the confidence interval to be sufficiently assured of capturing the true value.
- A confidence interval is a tool that can be used to get an idea of how variable a sample result might be.
- Confidence intervals are the typical way to present estimates as an interval range.
- The more data you have, the less variable a sample estimate will be.
- The lower the level of confidence you can tolerate, the narrower the confidence interval will be.
- The bootstrap is an effective way to construct confidence intervals.

Define the various terms related to Normal Distribution?

- A normal distribution is the continuous probability distribution with a probability density function that gives you a symmetrical bell curve.
- Simply put, it is a plot of the probability function of a variable that has maximum data concentrated around one point and a few points taper off symmetrically towards two opposite ends. The bell-shaped normal distribution is iconic in traditional statistics.
- The fact that distributions of sample statistics are often normally shaped has made it a powerful tool in the development of mathematical formulas that approximate those distributions.
- In this definition of a normal distribution, you will explore the following terms:
  i. **Continuous Probability Distribution:** A probability distribution where the random variable, X, can take any given value, e.g., amount of rainfall. You can record the rainfall received at a certain time as 9 inches. But this is not an exact value. The actual value can be 9.001234 inches or an infinite amount of other numbers. There is no definitive way to plot a point in this case, and instead, you use a continuous value.
  ii. **Probability Density Function:** An expression that is used to define the range of values that a continuous random variable can take.
- A normal distribution has a probability distribution that is centered on the mean. This means that the distribution has more data around the mean. The data distribution decreases as you move away from the center. The resulting curve is symmetrical about the mean and forms a bell-shaped distribution.
- In a normal distribution 68% of the data lies within one standard deviation of the mean, and 95% lies within two standard deviations as shown in figure.

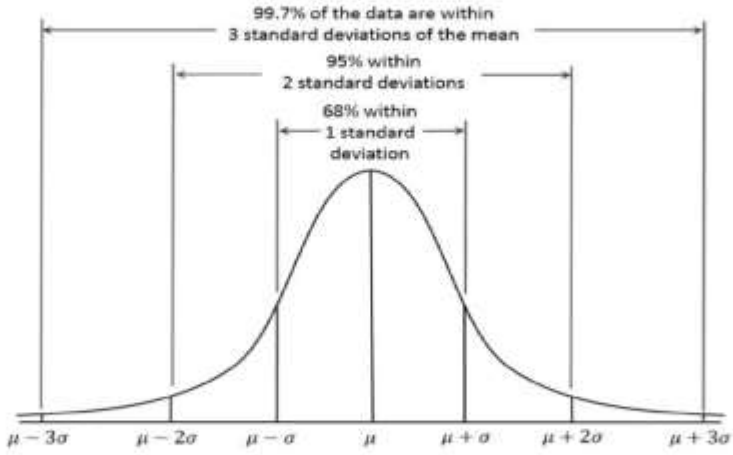


Figure 2-10. Normal curve

- The normal distribution is also referred to as a Gaussian distribution after Carl Friedrich Gauss, a prodigious German mathematician from the late 18th and early 19th centuries. Another name previously used for the normal distribution was the "error" distribution. Statistically speaking, an error is the difference between an actual value and a statistical estimate like the sample mean.
- The various key used in Normal Distribution are;

Key Terms for Normal Distribution

Error
The difference between a data point and a predicted or average value.

Standardize
Subtract the mean and divide by the standard deviation.

z-score
The result of standardizing an individual data point.

Standard normal
A normal distribution with mean = 0 and standard deviation = 1.

QQ-Plot
A plot to visualize how close a sample distribution is to a specified distribution, e.g., the normal distribution.

What is Standard Normal Distribution and QQ plots?

- A **Standard Normal Distribution** is one in which the units on the x-axis are expressed in terms of standard deviations away from the mean.
- To compare data to a standard normal distribution, you subtract the mean and then divide by the standard deviation; this is also called normalization or standardization
- Note that "standardization" in this sense is unrelated to database record standardization (conversion to a common format).
- The transformed value is termed a z-score, and the normal distribution is sometimes called the z-distribution.
- A Standard Normal Distribution is a type of normal distribution with a mean of 0 and a standard deviation of 1. This means that the normal distribution has its center at 0 and intervals that increase by 1.
- The mean and standard deviation in a normal distribution is not fixed. They can take on any value.
- However, when you standardize the normal distribution, the mean and standard deviation remain fixed and are the same for all standard normal distributions. Consider the example given below of weights of students in a class:

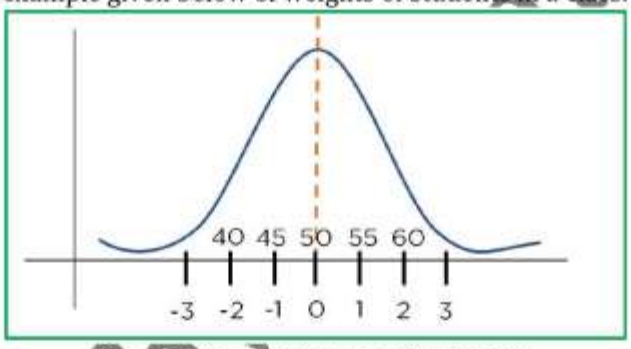
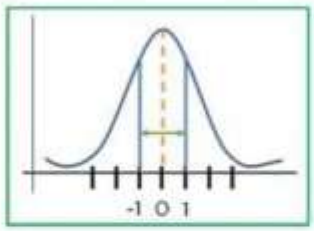
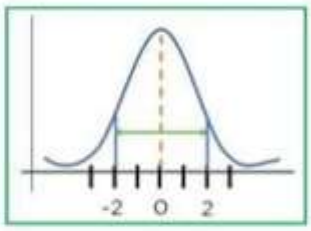


Figure: Standard Normal Distribution

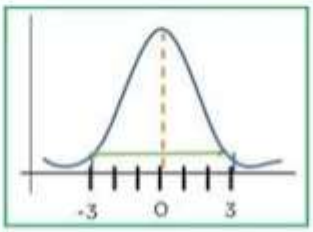
- It gives the actual weights of the students above the x-axis. But from the graph, you can see that the data points differ by 5 points.
- On finding the mean, you get it as 50, so you can take this as the 0th point. The rest of the points are equally spaced and, on standardizing, differ by 1, so you can rewrite the scale to be centered around 0 and increasing by 1.
- The points above the mean fall on positive values and below the mean fall on negative values.
- When you standardize your data, calculating the probabilities in your graph becomes easier.
- You can also easily compare different graphs with one another, as they all have the same scale. Some features of a Standard Normal Distribution are given below:



68% of values are within the first standard deviation



95% of values are within the second standard deviation



99.7% of values are within the third standard deviation

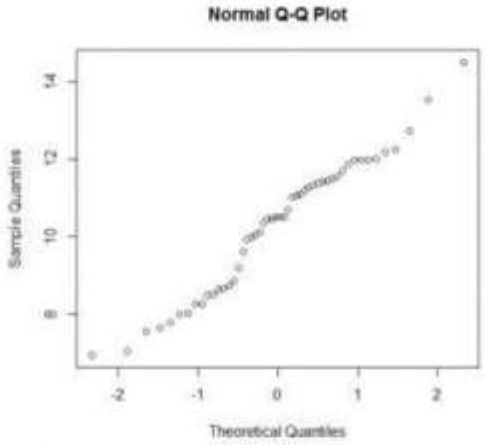
Figure: Characteristics of Standard Normal Distribution

Quantile-Quantile plot or Q-Q plot

- A **QQ-Plot** is used to visually determine how close a sample is to a specified distribution in this case, the normal distribution.
- The QQ-Plot orders the z-scores from low to high and plots each value's z-score on the y-axis; the x-axis is the corresponding quantile of a normal distribution for that value's rank.
- Since the data is normalized, the units correspond to the number of standard deviations away from the mean. If the points roughly fall on the diagonal line, then the sample distribution can be considered close to normal.
- Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other.
- The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.
- For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve.
- If two quantiles are sampled from the same distribution, they should roughly fall in a straight line. Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s).
- Below are the steps to generate a Q-Q plot for team members age to test for normality
  1) Take your variable of interest (team member age in this scenario) and sort it from smallest to largest value. Let's say you have 19 team members in this scenario.

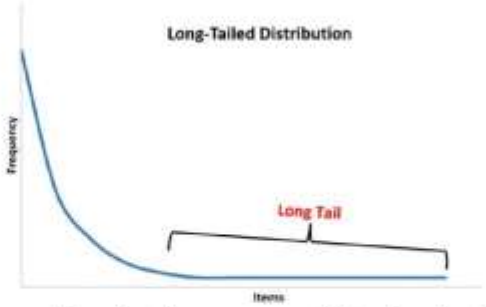
2) Take a normal curve and divide it into 20 equal segments (n+1; where n=#data points)
3) Compute z score for each of these points
4) Plot the z-score obtained against the sorted variables. Usually, the z-scores are in the x-axis (also called theoretical quantiles since we are using this as a base for comparison) and the variable quantiles are in the y-axis (also called ordered values)
5) Observe if data points align closely in a straight 45-degree line
6) If it does, the age is normally distributed. If it is not, you might want to check it against other possible distributions.

Here's an example of a Normal QQ plot when both sets of quantiles truly come from Normal distributions.



Explain Long-Tailed Distributions with example?

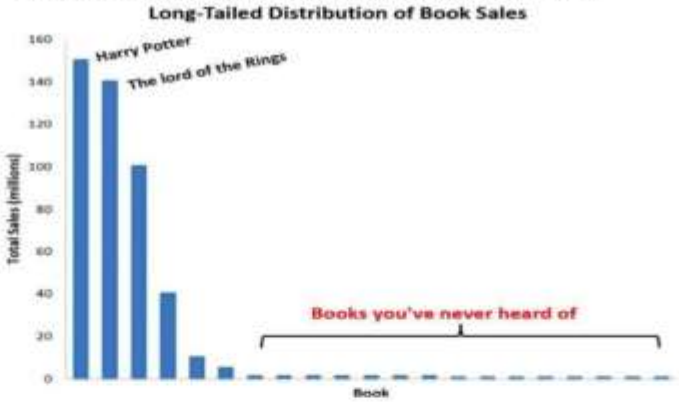
- Despite the importance of the normal distribution historically in statistics, and in contrast to what the name would suggest, data is generally not normally distributed.
- Key Terms for Long-Tailed Distributions are
  Tail – The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.
  Skew – Where one tail of a distribution is longer than the other.
- While the normal distribution is often appropriate and useful with respect to the distribution of errors and sample statistics, it typically does not characterize the distribution of raw data.
- Sometimes, the distribution is highly skewed (asymmetric), such as with income data; or the distribution can be discrete, as with binomial data. Both symmetric and asymmetric distributions may have long tails.
- The tails of a distribution correspond to the extreme values (small and large). Long tails, and guarding against them, are widely recognized in practical work.
- In statistics, a **long tail distribution** is a distribution that has a long "tail" that slowly tapers off toward the end of the distribution:



- It turns out that this type of distribution appears all the time in different real-world domains and it has interesting implications.
- Following are the several examples of long-tail distributions in the real world and shares why long-tail distributions are important.

Example: Long-Tail Distributions in Book Sales

- One of the most well-known examples of a long-tail distribution is book sales. There are a few books that have sold hundreds of millions of copies (Harry Potter, The Lord of the Rings, The Da Vinci Code, etc.) but most books sell less than one hundred copies total. If we created a bar chart to visualize the total sales of every book ever published, we would find that the chart exhibits a long-tailed distribution:



Explain Student's t-Distribution with example?

- The t-distribution is a normally shaped distribution, except that it is a bit thicker and longer on the tails. It is used extensively in depicting distributions of sample statistics.
- Distributions of sample means are typically shaped like a t-distribution, and there is a family of t-distributions that differ depending on how large the sample is. The larger the sample, the more normally shaped the t-distribution becomes.
- The various key terms of Student's t-distribution are

  n- Sample size.

  Degrees of freedom - A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and numbers of groups.
- The t-distribution is often called Student's t because it was published in 1908 in Biometrika by W. S. Gosset under the name "Student."

- Student's t-distribution or t-distribution is a probability distribution that is used to calculate population parameters when the sample size is small and when the population variance is unknown.
- Theoretical work on t-distribution was done by **W.S. Gosset**; he has published his findings under the pen name "**Student**". That's why it is called as **Student's t-test**.
- It is the sampling distribution of the t-statistic. The values of the t-statistic are given by:

```
t = [ x̄ - μ ] / [ s / sqrt( n ) ]

where,
t = t score
x̄ = sample mean,
μ = population mean,
s = standard deviation of the sample,
n = sample size
```

- Student's t Distribution is used when
  - i. The sample size must be 30 or less than 30.
  - ii. The population standard deviation(σ) is unknown.
  - iii. The population distribution must be unimodal and skewed.

Mathematical Derivation of t-Distribution:

- The t-distribution has been derived mathematically under the assumption of normally distributed population and the formula or equation will be like this

$$f(t) = c(1+(t^2/\nu))^{(-\nu+1)/2}$$

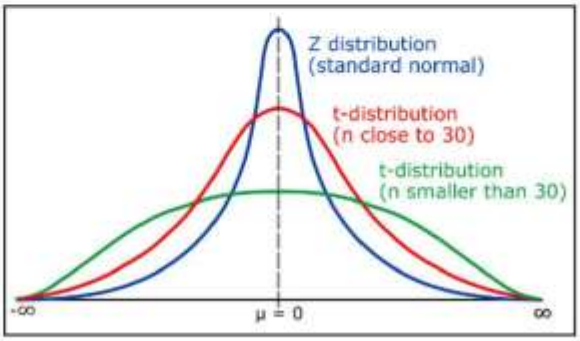
where,

c = Constant required to make the area under the curve equal to unity

ν = Degrees of freedom

- So, this above equation indicates the probability density function (pdf) of t distribution for ν degrees of freedom.

Properties of the t-Distribution:



- The above diagram indicates that the blue color curve is a standard normal distribution curve or a Z distribution curve because the sample size(n) is greater than 30. And the red color curve is a t-distribution curve because the sample size(n) is close to 30. Similarly, the green color curve is also a t-distribution curve because the sample size(n) is smaller than 30.
- Degrees of freedom refer to the number of independent observations in a set of data. When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one. Hence, the distribution of the t statistic from samples of size 10 would be described by a t distribution having 10 – 1 or 9 degrees of freedom. Similarly, a t- distribution having 15 degrees of freedom would be used with a sample of size 16.

Explain Binomial Distribution with example?

- The binomial distribution is a discrete probability distribution that represents the probabilities of binomial random variables in a binomial experiment.
- The binomial distribution is defined as a probability distribution related to a binomial experiment where the binomial random variable specifies how many successes or failures occurred within that sample space.
- It's important for data scientists and professionals in other fields to understand this concept as binomials are used often in business applications.
- Yes/no (binomial) outcomes lie at the heart of analytics since they are often the culmination of a decision or other process; buy/don't buy, click/don't click, survive/die, and so on.
- Central to understanding the binomial distribution is the idea of a set of trials, each trial having two possible outcomes with definite probabilities.
- The binomial distribution is the frequency distribution of the number of successes (x) in a given number of trials (n) with specified probability (p) of success in each trial. There is a family of binomial distributions, depending on the values of n and p.

Key Terms for Binomial Distribution

Trial
An event with a discrete outcome (e.g., a coin flip).

Success
The outcome of interest for a trial.

Synonym
"1" (as opposed to "0")

Binomial
Having two outcomes.

Synonyms
yes/no, 0/1, binary

Binomial trial
A trial with two outcomes.

Synonym
Bernoulli trial

Binomial distribution
Distribution of number of successes in *x* trials.

Synonym
Bernoulli distribution

- A binomial experiment represents a binomial random variable X which counts the number "n" of successes in N trials when each trial has only two outcomes, success, and failure.
- Thus, an experiment could consist of 1 trial, 5 trials, 10 trials, 20 trials, etc. Sighting real-world examples, an experiment could be tossing a coin 10 times (10 trials), taking 10 items for examining whether the items are defective, etc.
- If the experiment consists of just one trial that has only two outcomes such as success or failure, the trial is called a **Bernoulli trial.**
- ***For example***, the expected value of the number of heads in 100 trials of heads or tails is 50, or (100 × 0.5). Another common example of binomial distribution is estimating the chances of success for a free-throw shooter in basketball, where 1 = a basket made and 0 = a miss.

The binomial distribution function is calculated as:

$$P(x: n, p) = {}^{n}C_{x}\, p^{x}\, (1 - p)^{n - x}$$

Where:

- n is the number of trials (occurrences)
- x is the number of successful trials
- p is the probability of success in a single trial
- ${}^{n}C_{x}$ is the combination of n and x. A combination is the number of ways to choose a sample of x elements from a set of n distinct objects where order does not matter, and replacements are not allowed. Note that ${}^{n}C_{x} = n! / r! (n - r) !$, where! is factorial (so, $4! = 4 \times 3 \times 2 \times 1$).

Explain Chi-Square Distribution with example?

- The chi-square distribution is the distribution of this statistic under repeated resampled draws from the null model.
- A low chi-square value for a set of counts indicates that they closely follow the expected distribution.
- A high chisquare indicates that they differ markedly from what is expected. There is a variety of chi-square distributions associated with different degrees of freedom (e.g., number of observations.
- The Chi-Square Statistic is a number that describes the relationship between the theoretically assumed data and the actual data.
- It is usually considered as a number or statistic value that verifies the theoretical dataset with the actual dataset and gives the result in the form of a number.

$$\text{(Chi-Square Statistic) } \chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The formula for Chi-Square statistic is as shown above. Here, *Observed* is the actual observed dataset with the actual values and the *Expected* is the expected dataset that has the theoretical values.
- An important thing to note about the Chi-Square Statistic is that this value is always positive because when we have the dataset and we are evaluating the statistic, even though we will have positive values and might have negative values as well in the

dataset, those values are squared. This will always result in a positive value. Also, a Chi-Square Statistic has an approximate Chi-Squared Distribution.

- A Chi-Squared distribution is a set of values that are distributed and separated by the p-value (*P*) and Degree of Freedom (*DF*). The Chi-Squared Distribution can be used to check the probability of a result that is extreme to that value or greater than that.
- In such cases, we usually consider a significance level like for example we consider here P=10% (0.1).
- So, when we get a Chi-Square Statistic value, we check that value in the distribution by the specific Degrees of Freedom (DF).
- If the Chi-Square statistic value in the distribution has a probability (P) of 10 %(0.1) and above, we fail to reject the null hypothesis or else we reject it.
- The P in the table below is the p-value which is basically the probability which is checked in the distribution table when we get the statistic value.
- The DF is the degree of Freedom which depends on the number of values in the dataset. If there are N=4 values in the dataset in the Expected and Observed columns, the 4 values will be applied to the Chi-Square Statistic formula.
- But if we note here that when we have N=4 values in the dataset and we apply a particular formula to that N values, we basically need only the first 3 values to predict the 4th value. Therefore, DF here will be 3. DF=N-1=3.
- Let's consider ***an example*** wherein we want to know the attendance of students in a class. ***Suppose we have 5 students in a Class.*** Therefore, N = 5. So, we ask the class teacher to tell us the approximate attendance of students of the class in percentages. The teacher provides us with this data:

Expected (%): [50, 70, 75, 82, 86]

- But, then we go and verify the teacher's given data with the actual data stored of the attendance of students. We get this result:

Observed (%): [55, 65, 73, 82, 80]

- Now, we can see that there is a difference between both the datasets. So, if we want to reject or fail to reject this hypothesis (Expected (%)), we have to test this hypothesis.
- For this we will consider 2 probabilities:
  1st Probability: The teacher's distribution is correct
  2nd Probability: The teacher's distribution is incorrect

- We will do this hypothesis testing with a ***significance of 10% (0.1).*** If the Chi-Square statistic value that has been calculated on the basis of Expected and Observed has a P of 10% or above within specific Degree of Freedom (DF), only then we will fail to reject the teacher's hypothesis or else we will reject it.
- Now we will calculate the Chi-Square Statistic:

$$\text{(Chi-Square Statistic) } \chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Now, we will apply this formula to both datasets:

(Chi-Square Statistic) $\chi^2 = (\frac{(55-50)^2}{50}) + (\frac{(65-70)^2}{70}) + (\frac{(73-75)^2}{75}) + (\frac{(82-82)^2}{82}) + (\frac{(80-86)^2}{86})$

(Chi-Square Statistic) $\chi^2 = 0.5 + 0.36 + 0.05 + 0 + 0.41$

(Chi-Square Statistic) $\chi^2 = 1.32$

As we see above, we get a chi-square statistic value of 1.32.

We have the following data :

$\chi^2 = 1.32$

Degree of Freedom $(DF) = N - 1 = 5 - 1 = 4$

Significance level $(\alpha) = 10\%$ (0.1)

Now, we will check the p-value(P) for the statistic value 1.32 in the Chi-Squared Distribution table within the Degree of Freedom $DF = 4$.

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

Chi-Squared Distribution (Source)

Here, within DF = 4, we see that the statistic value 5.989 has a probability or p-value (P) of 0.2 which is 20%. So, any value less than 5.989 will have a probability of more than 20%. Our statistic value is 1.32. Therefore, this value will also have a probability of more than 20%, which is fairly higher than our significance level (10%). ***Therefore, we will fail to reject the teacher's hypothesis (NULL Hypothesis).***

Explain F-Distribution?

- A common procedure in scientific experimentation is to test multiple treatments across groups—say, different fertilizers on different blocks of a field.
- This is similar to the A/B/C test referred to in the chi-square distribution, except we are dealing with measured continuous values rather than counts.
- In this case we are interested in the extent to which differences among group means are greater than we might expect under normal random variation.
- The Fstatistic measures this and is the ratio of the variability among the group means to the variability within each group (also called residual variability). This comparison is termed an analysis of variance.
- The distribution of the Fstatistic is the frequency distribution of all the values that would be produced by randomly permuting data in which all the group means are equal (i.e., a null model).
- There are a variety of F-distributions associated with different degrees of freedom. The F-statistic is also used in linear regression to compare the variation accounted for by the regression model to the overall variation in the data.
- The F-distribution is used with experiments and linear models involving measured data. The F-statistic compares variation due to factors of interest to overall variation.
- The F distribution allows us to use an F statistic to compare two populations. For ANOVA tests we can use the F distribution to determine if the variance between the means of two populations significantly differs.
- The F distribution can also be used in regression analysis to compare the fit of different models. The F distribution allows us to obtain an F value that determines the p-value.
- A p-value is the probability of getting a result as extreme or more extreme as the one that you observed, given that the null hypothesis is true.
- An F distribution is a probability distribution that results from comparing the variances of two samples or populations using the F statistic. It is the distribution of all possible F values for a specific combination of samples sizes that are being compared.
- The F value is used in analysis of variance (ANOVA). It is calculated by dividing two mean squares. This calculation determines the ratio of explained variance to unexplained variance. The F distribution is a theoretical distribution.
- 4 characteristics of the F-distribution are
  i. The curve is not symmetrical but skewed to the right.
  ii. There is a different curve for each set of DFS.
  iii. The F statistic is greater than or equal to zero.
  iv. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.