#1: Measure of Central Tendency

A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. It can bethought as the tendency of data to cluster around a middle value.

https://www.linkedin.com/in/ravi-ranjan-prasad-karn/

The following are the various measures of central tendency:

1. Arithmetic Mean
2. Weighted Mean
3. Median
4. Mode
5. Geometric Mean
6. Harmonic Mean

hashtag#statistics hashtag#statisticsfordatascience

#2: Some Symbols used in Statistics

hashtag#statistics hashtag#statisticsfordatascience

#3: Significance of the Measure of Central Tendency

---

1. To get a single representative value (Summary Statistics)
2. To condense data
3. To facilitate comparison
4. Helpful in further statistical analysis

hashtag#statistics hashtag#statisticsfordatascience

#4:Properties of a Good Average

---

1. It should be simple to understand
2. It should be easy to calculate
3. It should be rigidly defined
4. It should be liable for algebraic manipulations
5. It should be least affected by sampling fluctuations
6. It should be based on all the observations
7. It should be possible to calculate even for open-end class intervals
8. It should not be affected by extremely small or extremely large observation

hashtag#statistics hashtag#statisticsfordatascience

#5:Properties of Arithmetic Mean

Property 1: Sum of deviations of observations from their mean is zero.

$\Sigma(x - \text{mean}) = 0$

Property 2: Sum of squares of deviations taken from mean is least in comparison to the same taken from any other average.

Property 3: Arithmetic mean is affected by both the change of origin and scale.

hashtag#statistics hashtag#statisticsfordatascience

#6: Merits and Demerits of Arithmetic Mean

Merits of Arithmetic Mean

1. It utilizes all the observations

2. It is rigidly defined

3. It is easy to understand and compute

4. It can be used for further mathematical treatments.

Demerits of Arithmetic Mean

1. It is badly affected by extremely small or extremely large values

2. It cannot be calculated for open end class intervals

3. It is generally not preferred for highly skewed distributions

hashtag#statistics hashtag#statisticsfordatascience

#7:Median

Median is that value of the variable which divides the whole distribution into two equal parts. Data should be arranged in ascending or descending order of magnitude. For odd number of observations, the median is the middle value of the data. For even

number of observations, there will be two middle values. So we take the arithmetic mean of these two middle values. Number of the observations below and above the median, are same.

Merits and Demerits of Median

Merits

---

1. It is rigidly defined;

2. It is easy to understand and compute

3. It is not affected by extremely small or extremely large values

Demerits

---

1. In case of even number of observations we get only an estimate of the median by taking the mean of the two middle values. We don't get its exact value

2. It does not utilize all the observations. The median of 1, 2, 3 is 2. If the observation 3 is replaced by any number higher than or equal to 2 and if the

number 1 is replaced by any number lower than or equal to 2, the median value will be unaffected. This means 1 and 3 are not being utilized

3. It is not amenable to algebraic treatment

4. It is affected by sampling fluctuations

hashtag#statistics hashtag#statisticsfordatascience

#8: Mode

Highest frequent observation in the distribution is known as mode.

Merits and Demerits of Mode

Merits

---

1. Mode is the easiest average to understand and also easy to calculate

2. It is not affected by extreme values

3. It can be calculated for open end classes

4. As far as the modal class is confirmed the pre-modal class and the post modal class are of equal width

5. Mode can be calculated even if the other classes are of unequal width

Demerits

---

1. It is not rigidly defined. A distribution can have more than one mode

2. It does not utilize all the observations

3. It is not amenable to algebraic treatment

4. It is greatly affected by sampling fluctuations

hashtag#statistics hashtag#statisticsfordatascience

#9:Relationship between Mean, Median and Mode

For a symmetrical distribution the mean, median and mode coincide. But if the distribution is moderately asymmetrical, there is an empirical relationship between them. The relationship is

Mean – Mode = 3 (Mean – Median)

Mode = 3 Median – 2 Mean

Using this formula, we can calculate mean/median/mode if other two of them are known.

hashtag#statistics hashtag#statisticsfordatascience

#10: Geometric Mean Special

---

1. All the observations for which we want to find the Geometric Mean should be non-zero positive values.

2. if GM1 and GM2 are Geometric Means of two series-Series of sizes n and m respectively, then the Geometric Mean of the combined series is given by the formula

Log GM = (n logGM1 + m logGM2) / (n + m)

hashtag#statistics hashtag#statisticsfordatascience

#11: Geometric Mean

Geometric Mean (GM) is used for averaging ratios or proportions. It is used when each item has multiple properties that have different numeric ranges. It gives high weight to lower values. It normalizes the differently-ranged values. Geometrically, GM of two numbers, a and b, is the length of one side of a square whose area is equal to the area of a rectangle with sides of lengths a and b. Similarly, GM of three numbers, a, b, and c, is the length of one edge of a cube whose volume is the same as that of a cuboid with sides whose lengths are equal to the three given numbers and so on.

Example of Scenario where GM is useful: In film and video to choose aspect ratios (the proportion of the width to the height of a screen or image). It's used to find a compromise between two aspect ratios, distorting or cropping both ratios equally.

hashtag#statistics hashtag#statisticsfordatascience

#12: Relation between Arithmetic Mean, Geometric Mean and Harmonic Mean

---

1. AM ≥ GM ≥ HM

2. GM = sqr(AM.HM) (for two variables)

3. For any n, there exists c>0 such that the following holds for any n-tuple of positive reals:

AM+cHM>=(1+c)GM.

hashtag#statistics hashtag#statisticsfordatascience

#13: Partition Values, Quartiles, Deciles ans Percentiles

Partition values: Partition values are those values of variable which divide the distribution into a certain number of equal parts. The data should be arranged in ascending or descending order of magnitude. Commonly used partition values are quartiles, deciles and percentiles.

Quartiles: Quartiles divide whole distribution in to four equal parts. There are three quartiles.

Deciles: Deciles divide whole distribution in to ten equal parts. There are nine deciles.

Percentiles divide whole distribution in to 100 equal parts. There are ninety nine percentiles.

hashtag#statistics hashtag#statisticsfordatascience

#14: Measure of Dispersion/Variation

According to Spiegel, the degree to which numerical data tend to spread about an average value is called the variation or dispersion of data. This points out as to how far an average is representative of the entire data. When variation is less, the average closely represents the individual values of the data and when variation is large; the average may not closely represent all the units and be quite unreliable.

Following are the different measures of dispersion:

1. Range

2. Quartile Deviation

3. Mean Deviation

4. Standard Deviation and Variance

hashtag#statistics hashtag#statisticsfordatascience

#15: Significance of Measures of Dispersion

Measures of dispersion are needed for the following four basic purposes:

1. Measures of dispersion determine the reliability of an average value means to how far an average is representative of the entire data. When variation is less, the average closely represents the individual values of the data and when variation is large; the average may not closely represent that value.

2. Measuring variation helps determine the nature and causes of variations in order to control the variation itself.

3. The measures of dispersion enable us to compare two or more series with regard to their variability. The relative measures of dispersion may also

determine the uniformity or consistency. Smaller value of relative measure of dispersion implies greater uniformity or consistency in the data.

4. Measures of dispersion facilitate the use of other statistical methods. In other words, many powerful statistical tools in statistics such as correlation analysis, the testing of hypothesis, the analysis of variance, techniques of quality control, etc. are based on different measures of dispersion.

hashtag#statistics hashtag#statisticsfordatascience

#16: Range

Range is the simplest measure of dispersion. It is defined as the difference between the maximum value of the variable and the minimum value of the

variable in the distribution. Range has unit of the variable and is not a pure number.

Its merit lies in its simplicity.

The demerit is that it is a crude measure because it is using only the maximum and the minimum observations of variable. If a single value lower than the minimum or higher than the maximum is added or if the maximum or minimum value is deleted range is seriously affected.

However, it still finds applications in Order Statistics and Statistical Quality Control.

It can be defined as

R = Xmax - Xmin

where, Xmax: Maximum value of variable and

Xmin : Minimum value of variable.

hashtag#statistics hashtag#statisticsfordatascience

#17: Coefficient of Range

Coefficient of Range is defined as the relative measure of the dispersion of the range. It is the ratio of difference of highest value and smallest value of the distribution to their sum. It is a pure number as it does not have unit.

Coefficient of Range is zero when Range is zero.

Formula for Coefficient of Range: (Xmax - Xmin)/(Xmax + Xmin)

hashtag#statistics hashtag#statisticsfordatascience

#18: Quartile Deviation

Let Q1 and Q3 are the first quartile and the third quartile respectively. (Q3 – Q1) gives the inter quartile range. The semi inter quartile range which is also known as Quartile Deviation (QD) is given by

Quartile Deviation (QD) = (Q3 – Q1) / 2

Relative measure of Q.D. known as Coefficient of Q.D. and is defined as

Cofficient of QD = (Q3 – Q1) /(Q3 + Q1)

The quartile deviation is a slightly better measure of dispersion than the range, but it ignores the observations on the tails of distribution.

Coefficient of quartile deviation is a pure number without unit.

For symmetric distribution (such as normal distribution where mean and mode are same), coefficient of quartile deviation is equal to the ration of quartile deviation and mean value of distribution.

hashtag#statistics hashtag#statisticsfordatascience

#19: Mean Deviation

Mean deviation is defined as average the absolute values of deviation from any arbitrary value viz. mean, median, mode, etc. It is often suggested to calculate it from the median because it gives least value when measured from the median.

The deviation of an observation xi from the assumed mean A is defined as

(xi – A).

#20: Variance

Variance is the average of the square of deviations of the values taken from mean. Taking a square of the deviation is a better technique to get

rid of negative deviations.

Variance is defined as

#21: Variance of Combined Series

If there are two or more populations and the information about the means and variances of those populations are available then we can obtain the combined variance of several populations. If n1, n2,..., nk are the sizes, x1 bar , x2 bar ,…, xk bar are the means and σ1 square , σ2 square ,…,σk

square are the variances of k populations, then the combined variance is given by

#22: Standard Deviation

Standard Deviation is a statistic that measures the dispersion of a data set relative to its mean and is calculated as the square root of the variance.

It is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.

#23: Root Mean Square Deviation

Root Mean Square Root Mean Square Deviation (RMSD) is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of mean deviation(deviation from Assumed Mean). if assumed mean is equal to mean, then RMSD is called standard deviation.

Root Mean Square Deviation is defined as

#24: Coefficient of Variation

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

#25: Measure of Central Tendency and Measure of Dispersion

While average (measure of central tendency) gives the value around which a distribution is scattered, measure of dispersion tells how it is scattered. So if one suitable measure of average and one suitable measure of dispersion is calculated (say mean and SD), we get a good idea of the distribution even if the distribution is large.

hashtag#statistics hashtag#statisticsfordatascience

#26: Moments

Moments are a set of statistical parameters to measure a distribution. They are the arithmetic means of first, second, third and so on, i.e. rth power of the deviation taken from either mean or an arbitrary point of a distribution. They represent a convenient and unifying method for summarizing many of the most commonly used statistical measures such as measures of tendency, variation, skewness and kurtosis.

Moments can be classified in raw and central moment. Raw moments are measured about any arbitrary point A (say). If A is taken to be zero then raw

moments are called moments about origin. When A is taken to be Arithmetic mean we get central moments. The first raw moment about origin is mean whereas the first central moment is zero. The second raw and central moments are mean square deviation and variance,

respectively. The third and fourth moments are useful in measuring skewness and kurtosis.

#27: Skewness

Lack of symmetry is called skewness for a frequency distribution. It is a measure of asymmetry of the frequency distribution of a real-valued random variable

If the distribution is not symmetric, the frequencies will not be uniformly distributed about the centre of the distribution. In Statistics, a frequency distribution is called symmetric if mean, median and mode coincide. Otherwise, the distribution becomes asymmetric. If the right tail is longer, we get a positively skewed distribution for which mean > median > mode while if the left tail is longer, we get a negatively skewed distribution for which mean < median < mode.

The example of the Symmetrical curve, Positive skewed curve and Negative skewed curve are given as follows:

#28: Difference between Variance and Skewness

- Variance tells us about the amount of variability while skewness gives the direction of variability.

- In business and economic series, measures of variation (e.g.variance) have greater practical application than measures of skewness. However, in medical and life science field measures of skewness have greater practical applications than the variance.

hashtag#statistics hashtag#statisticsfordatascience

#29: Why skweness occurs? How to overcome skewness?

When data is not distributed normally/symmetrically from mean, skewness occurs. The reason of occurance of skewness is occurance of excess of low values or high values in the distribution.

Skewness can be overcomed by using transformation techniques such as log transformation or standarising(scaling). The transformed distribution would be normally distributed or nearly normally distributed.

In data science paradigm, skewness is related to imbalanced class and existance of outliers. For undoing the effect of skewness, we can apply normalization techniques (such as transformation), resampling and outliers removal etc.

hashtag#statistics hashtag#statisticsfordatascience

#30: Absolute Measures of Skewness

Measures of skewness can be both absolute as well as relative. Since in a symmetrical distribution mean, median and mode are identical more the mean

moves away from the mode, the larger the asymmetry or skewness. An absolute measure of skewness can not be used for purposes of comparison

because of the same amount of skewness has different meanings in distribution with small variation and in distribution with large variation.

Following are the absolute measures of skewness:

1. Skewness (Sk) = Mean – Median

2. Skewness (Sk) = Mean – Mode

3. Skewness (Sk) = (Q3 - Q2) - (Q2 - Q1)

In general, we do not calculate these absolute measrues but we calculate the relative measures which are called coefficient of skewness.

Coefficient of skewness are pure numbers independent of units of measurements.

hashtag#statistics hashtag#statisticsfordatascience

#31: Relative Measures of Skewness

In order to make valid comparison between the skewness of two or more distributions we have to eliminate the distributing influence of variation. Such

elimination can be done by dividing the absolute skewness by standard deviation. The following are the important methods of measuring relative

skewness:

1. Beta and Gamma Coefficient of Skewness (based on second and third central moment)

2. Karl Pearson's Coefficient of Skewness (based on first and second central moment)

3. Bowleys's Coefficient of Skewness (based on quartiles)

4. Kelly's Coefficient of Skewness (based on percentile/decile)

hashtag#statistics hashtag#statisticsfordatascience

#32: Some facts about Skewness

- If the value of mean, median and mode are same in any distribution, then the skewness does not exist in that distribution. Larger the difference in

these values, larger the skewness.

- If sum of the frequencies are equal on the both sides of mode then skewness does not exist.

- If the distance of first quartile and third quartile are same from the median then a skewness does not exist. Similarly if deciles (first and ninth) and percentiles (first and ninety nine) are at equal distance from the median, then there is no asymmetry.

- If the sums of positive and negative deviations obtained from mean, median or mode are equal then there is no asymmetry.

- If a graph of a data become a normal curve and when it is folded at middle and one part overlap fully on the other one then there is no asymmetry.

hashtag#statistics hashtag#statisticsfordatascience

#33: Concept of Kurtosis

If we have the knowledge of the measures of central tendency, dispersion and skewness, even then we cannot get a complete idea of a distribution. In

addition to these measures, we need to know another measure to get the complete idea about the shape of the distribution which can be studied with

the help of Kurtosis. Prof. Karl Pearson has called it the “Convexity of a Curve”. Kurtosis gives a measure of flatness of distribution.

The degree of kurtosis of a distribution is measured relative to that of a normal curve. The curves with greater peakedness than the normal curve are called “Leptokurtic”. The curves which are more flat than the normal curve are called

“Platykurtic”. The normal curve is called “Mesokurtic.” The following describes the three different curves mentioned above:

#34: Measure of Kurtosis

#35: Scale of Measurement

#36: Types of Data

N.B.: Last basis can also be read as "On the basis of source of data"

#37: Census vs Sampling on Population

#38: Statistics and Statistic

There is a very common misconception and confusion about the word in singular sense- “STATISTIC” and in plural sense “STATISTICS”

The characteristic of population is called parameter and the characteristic of sample is called STATISTIC. It is a single measure of some attribute of a sample. For example, Xbar, which is sample mean. It is used to estimate the parameter (such as mue, population mean) for the population.

Statistics is a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation.

hashtag#statistics hashtag#statisticsfordatascience

#40: Different Types of Population

Based on Number:

- Finite Population: Population containing finite number of units or observations. E.g., the population of students in a class, the population of bolts produced in a factory in a day, the population of books in a library, etc. In these examples the number of units in the population is finite in numbers.

- Infinite Population: A population containing infinite (uncountable) number of units or observations. E.g., the population of particles in a salt bag, the population of stars in the sky, etc. In these examples the number of units in the population is not finite.

But theoretically sometimes, populations of too large in size are assumed infinite.

Based on subject:

-Real Population: A population comprising the items or units which are all physically present. All of the examples given above are examples of a real population.

-Hypothetical Population:A population consisting the items or units which are not physically present but the existence of them can only be imagined or conceptualized. E.g., the population of heads or tails in successive tosses of a coin a large number of times is considered as hypothetical population.

hashtag#statistics hashtag#statisticsfordatascience

#41: Sampling Distribution of Sample Mean

We draw all possible samples of same size from the population and calculate the sample mean for each sample. After calculating the value of sample mean for each sample we observed that the values of sample mean vary from sample to sample. Then the sample mean is treated as random variable and a probability distribution is constructed for the values of sample mean. This probability distribution is known as sampling distribution of sample mean. Therefore, the sampling distribution of sample mean can be defined as: “The probability distribution of all possible values of sample mean that would be obtained by drawing all possible samples of the same size from the population is called sampling distribution of sample mean or simply says sampling distribution of mean.”

hashtag#statistics hashtag#statisticsfordatascience

#42: Asymptotic Theory

In Statistics, asymptotic theory, or large sample theory (LST), is a generic framework for assessment of properties of estimators and statistical tests. Within this framework, it is typically assumed that the sample size n grows indefinitely, and the properties of statistical procedures are evaluated in the limit as n tends to infinity. In practice, a limit evaluation is treated as being approximately valid for large finite sample sizes, as well. The importance of the asymptotic theory is that it often makes possible to carry out the analysis and state many results which cannot be obtained within the standard “finite-sample theory”.

hashtag#statistics hashtag#statisticsfordatascience

#43:Independent and identically distributed (IID)random variables

---

Identically Distributed means that there are no overall trends–the distribution doesn’t fluctuate and all items in the sample are taken from the same probability distribution. Independent means that the sample items are all independent events. In other words, they aren’t connected to each other in any way.

In probability theory and statistics, a collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent. This property is usually called Independent and identically distributed (IID)random variables.

hashtag#statistics hashtag#statisticsfordatascience

#44: Random Variable

When the value of a variable is determined by a random event, that variable is called a random variable. It gives numbers to outcomes of random events. Random variables can be discrete or continuous. A random variable that may assume only a finite number or an infinite sequence of values is said to be discrete; one that may assume any value in some interval on the real number line is said to be continuous. For instance, a random variable representing the number of automobiles sold at a particular dealership on one day would be discrete, while a random variable representing the weight of a person in kilograms (or pounds) would be continuous.

hashtag#statistics hashtag#statisticsfordatascience

#45: Frequency Distribution

Frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval(for continuous variable) or the number of observations per distinct value(categorical or discrete variable) . The interval size depends on the data being analyzed and the goals of the analyst. The intervals must be mutually exclusive and exhaustive. Frequency distribution is mostly used for summarizing categorical data.

As a statistical tool, a frequency distribution provides a visual representation for the distribution of observations within a particular test. Analysts often use frequency distribution to visualize or illustrate the data collected in a sample. For example, the height of children can be split into several different categories or ranges. In measuring the height of 50 children, some are tall, and some are short, but there is a high probability of a higher frequency or concentration in the middle range. The most important factors for gathering data are that the intervals used must not overlap and must contain all of the possible observations.

hashtag#statistics hashtag#statisticsfordatascience

#46: Cumulative distribution functions (c.d.f.)

Cumulative distribution functions describe real random variables. Suppose that X is a random variable that takes as its values real numbers. Then the cumulative distribution function F for X is the function whose value at a real number x is the probability that X takes on a value less than or equal to x.

$F(x)=P(X\leq x)$

Each c.d.f.F has the following four properties:

- F is a non decreasing function

- F is right continuous

- $\lim x\to\infty F(x)=1$

- $\lim x\to-\infty F(x)=0$

Conversely, any function F with above four properties is the c.d.f. of a real random variable.

For a discrete random variable, the c.d.f. is a step function.

For a continuous random variable the c.d.f. is differentiable function.

hashtag#statistics hashtag#statisticsfordatascience

#47: Probability Distribution

A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability values. There are many different classifications of probability distributions. Some of them include the normal distribution, chi square distribution, binomial distribution, and Poisson distribution. The different probability distributions serve different purposes and represent different data generation processes. The binomial distribution, for example, evaluates the probability of an event occurring several times over a given number of trials and given the event's probability in each trial,and may be generated by keeping track of how many free throws a basketball player makes in a game, where 1 = a basket and 0 = a miss. Another typical example would be to use a fair coin and figuring the probability of that coin coming up heads in 10 straight flips. A binomial distribution is discrete, as opposed to continuous, since only 1 or 0 is a valid response.

hashtag#statistics hashtag#statisticsfordatascience

#48: Probability Distribution Function

A probability distribution function is some function that may be used to define a particular probability distribution. Depending upon the type of variable and problem in hand, we can have following types of probability distribution functions:

- cumulative distribution function (c.d.f.)

- probability mass function (p.m.f.)

- probability density function (p.d.f.)

hashtag#statistics hashtag#statisticsfordatascience

#49: Probability Mass Function(PMF) & Probability Density Function (PDF)

The probability that a discrete random variable X takes on a particular value x, i.e., P(X = x), denoted as f(x), is called the probability mass function. This function gives the probability that a discrete random variable is exactly equal to some value. It is often the primary means of defining a discrete probability distribution, and such functions exist for either scalar or multivariate random variables whose domain is discrete.

The probability that a continuous random variable X takes on a range (a,b), i.e., f(a,b)=P(a<X<b) is called the probability density function. It is a statistical expression that defines a probability distribution for a continuous random variable. When the PDF is graphically portrayed, the area under the curve will indicate the interval in which the variable will fall. The total area in this interval of the graph equals the probability of a continuous random variable occurring.

E.g. the probability that today's temperature will be 80 degrees (80 degrees exactly) is measured by PMF and the probability that the temperature will be between 80 and 85 degrees is measured by PDF.

hashtag#statistics hashtag#statisticsfordatascience

#50: Sampling with replacement and without replacement

Sampling 'with replacement' means that when a unit selected at random from the population, it is returned to the population (replaced), and then a second element is selected at random. Here, the two sample values are independent i.e. what we get on the first one doesn't affect what we get on the second. The covariance between the two is zero.

Sampling 'without replacement' means when a unit selected at random from the population, it is not returned to the population (replaced), and then a second element is selected at random. Each sample unit of the population has only one chance to be selected in the sample. Here, the two sample values aren't independent i.e. what we got on the for the first one affects what we can get for the second one. The covariance between the two isn't zero. Here, the covariance depends on the population size. If the population is very large, this covariance is very close to zero. In that case, sampling with replacement isn't much different from sampling without replacement. sometimes, this difference is described as sampling from an infinite population (sampling with replacement) vs sampling from a finite population (without replacement).

hashtag#statistics hashtag#statisticsfordatascience

#51: Central Limit Theorem

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed. This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually n > 30). If the population is normal, then the theorem holds true even for samples smaller than 30. In fact, this also holds true even if the population is binomial, provided that min(np, n(1-p))> 5, where n is the sample size and p is the probability of success in the population. This means that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.

For the random samples we take from the population, we can compute the mean of the sample means:

$\mu(\bar{x})=\mu$

and the standard deviation of the sample means:

$\sigma(\bar{x})=\sigma/\sqrt{n}$

hashtag#statistics hashtag#statisticsfordatascience

#52: The Law of Large Numbers (LLN)

If you repeat a trial many times, then the average of the observed values tend to be close to the expected value. (In general, the more trials you run, the better the estimates.). This is called the law of large numbers(LLN). For example, you toss a

fair die many times and compute the average of the numbers that appear. The average should converge to 3.5, which is the expected value of the roll because (1+2+3+4+5+6)/6 = 3.5. The same theorem ensures that about one-sixth of the faces are 1s, one-sixth are 2s, and so forth.

hashtag#statistics hashtag#statisticsfordatascience

#53:Z-score

A Z-score is a numerical measurement used in statistics of a value's relationship to the mean (average) of a group of values, measured in terms of standard deviations from the mean.

Simply put, a z-score is the number of standard deviations from the mean a data point is.

If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

The basic z score formula for a sample is:

$z = (x – \mu) / \sigma$

For example, let’s say you have a IIT entrance test score of 210. The test has a mean (μ) of 170 and a standard deviation (σ) of 15. Assuming a normal distribution, your z score would be:

$z = (x – \mu) / \sigma$

$= (210 – 170) / 15 = 2.67$

This means that your score is 2.67 standard deviation away from mean. This illustrates that you have very high score and you are one of the toppers in the test. This again depends on the population size. This also has to be kept in mind.

hashtag#statistics hashtag#statisticsfordatascience

#54: Standard Error

The standard error (SE) is a measure of the variability of a statistic. It is an estimate of the standard deviation of a sampling distribution.

If the statistic is the mean, it is called the standard error of the mean (SEM).

The standard error is inversely proportional to the sample size. Larger the sample size, smaller the standard error because the statistic will approach the actual value(parameter).

When we calculate the sample mean we are usually interested not in the mean of this particular sample, but in the mean the population from which the sample comes. We usually collect data in order to generalise from them and so use the sample mean as an estimate of the mean for the whole population. Now the sample mean will vary from sample to sample; the way this variation occurs is described by the “sampling distribution” of the mean. We can estimate how much sample means will vary from the standard deviation of this sampling distribution, which we call the standard error (SE) of the estimate of the mean. As the standard error is a type of standard deviation, confusion may arise. Another way of considering the standard error is as a measure of the precision of the sample mean.

hashtag#statistics hashtag#statisticsfordatascience

#55: Application of Central Limit Theorem (CLT)

Central Limit Theorem helps us find the mean, standard deviation and other sample statistics of sample means. These values help us estimate population parameters. It helps us to get around the problem of data from populations that are not normal, provided the sample size is large enough (usually at least 30) and all samples have the same size. Hypothesis testing can be used on non normal data with the help of CLT. CLT says that static of sample means estimates the population parameter. E.g. Mean of sample means is said to be an estimate of Population mean.

hashtag#statistics hashtag#statisticsfordatascience

#56: Standard Deviation and Standard Error

The standard deviation (SD) measures the amount of variability, or dispersion, for a data point from the mean. If it is calculated for population, then it is called population SD and when it is calculated for sample, then it is called sample SD. Whereas the standard error is a measure of the variability of a statistic. If the statistic is the mean, it is called the standard error of the mean (SEM). SEM measures how far the sample mean of the data is likely to be from the true population mean. The difference between standard deviation and standard error is based on the difference between the description of data and its inference.

The SEM is always smaller than the SD.

hashtag#statistics hashtag#statisticsfordatascience

#57: Symmetry in Statistics

In Statistics, symmetry is an attribute used to describe the shape of a data distribution. When it is graphed, the distribution can be divided at the center so that each half is a mirror image of the other. This is called symmetric distribution. A symmetric distribution is never a skewed distribution. In unimodal (single peak) symmetric distribution, mean, median and mode are same.

hashtag#statistics hashtag#statisticsfordatascience

#58: Residual and Error of the Estimate

The error of estimate(or disturbance) of an observed value is the deviation of the observed value from the (unobservable) true value of a quantity of interest (for example, a population mean μ, which is estimated as $\bar{x}$, then error of estimate is μ-$\bar{x}$), and the residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, a sample mean $\bar{x}$, which is estimated as $\bar{x}'$ then residual is $\bar{x}$-$\bar{x}'$). These terms are used in regression analysis very often.

Sometimes people use residual and error of estimation synonymously. But we should know the difference between them.

hashtag#statistics hashtag#statisticsfordatascience

#59: Principle of Least Square

The least squares principle states that the Sample Regression Function (SRF) should be constructed (with the constant and slope values such as Y=aX+b) so that the sum of the squared distance between the observed values of the dependent variable and the values estimated from SRF is minimized (the smallest possible value).

Let (xi,yi) be the observed data points and (x'i,y'i) is estimated data point(data point on regression line corresponding to xi,yi i.e. vertical deviation from data point to the regression line), then least square principal says that ∑ (xi-x'i)^2+(yi-y'i)^2 has to be minimized to get best fit regression line.

Least squares principle is a widely used method for obtaining the estimates of the parameters in a statistical model based on observed data.

Other techniques, including generalized method of moments (GMM) and maximum likelihood (ML) estimation, can be used to estimate regression functions, but they require more mathematical sophistication and more computing power.

Least squares is sensitive to outliers. A strange value will pull the line towards it.

hashtag#statistics hashtag#statisticsfordatascience

#60: Correlation Analysis

Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a – sign indicates a negative relationship. Usually, in statistics, we measure four types of correlations: Pearson correlation, Kendall rank correlation, Spearman correlation, and the Point-Biserial correlation.

Correlation can be visualized by scatter diagram. If the scatter forms a line, it is correlated. Sharpness of line show the extent of association. If we cannot make out line in scatter diagram, the association does not exist. If the line has non-negative slope, then association is positive and if the slope is negative, association is negative.

hashtag#statistics hashtag#statisticsfordatascience

#61:Regression Analysis

Regression Analysis is a statistical technique used to investigate the relationship between variables. It is a reliable method of identifying which variables have impact on a topic of interest. The process of performing a regression allows you to determine which factors matter most, which can be ignored, and how these

factors influence each other. Dependent and Independent variables are the factors in Regression Analysis. Dependent Variable is the main factor that you are trying to understand or predict and independent Variables are the factors that you hypothesize have an impact on dependent variable. For conducting a regression analysis, you will need to define a dependent variable that you hypothesize is being influenced by one or several independent variables, have data set collected from source. Applying loss method such as least square method , regression line is established. The regression line represents the relationship between independent variable and dependent variables. When the dependent variable is modeled as a linear function of model parameters, it is called linear regression and when modeled as a non- linear function of model parameters, it is non-linear regression.

hashtag#statistics hashtag#statisticsfordatascience

#62: Linear Regression

Linear regression (LR) attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered as an explanatory variable, and the others as dependent variables. E.g., relating the weights of individuals to their heights using a LR model.

An LR line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept.

We should determine a relationship between the variables before attempting to fit an LR model to observed data. Significant association may not necessarily imply that one variable causes the other (e.g., higher SAT scores do not cause higher college grades). A scatterplot helps determine the strength of the association between two variables. If there appears to be no association between the proposed explanatory and dependent variables, then fitting a linear regression

model to the data probably will not provide a useful model. A useful numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

hashtag#statistics hashtag#statisticsfordatascience

#63: Regression Coefficient

Regression coefficients are estimates of the unknown population parameters and describe the relationship between a predictor (independent) variable and the response(dependendt variable). In linear regression, coefficients are the values that multiply the predictor values. Suppose you have the following regression equation: y = 3X + 5. In this equation, +3 is the coefficient, X is the predictor, and +5 is the constant.

The coefficient value represents the mean change in the response given a one unit change in the predictor. For example, if a coefficient is +3, the mean response value increases by 3 for every one unit change in the predictor.

hashtag#statistics hashtag#statisticsfordatascience

#64: Non-Linear Regression

Nonlinear regression is a regression in which the dependent variables are modeled as a non-linear function of model parameters and one or more independent

variables. The parameters can take the form of an exponential, trigonometric, power, or any other nonlinear function. There are several common models, such as Asymptotic Regression/Growth Model, as:

y= b1 + b2 * exp(b3 * x)

Logistic Population Growth Model,as:

y= b1 / (1 + exp(b2 + b3 * x)), and

Asymptotic Regression/Decay Model,as:

y= b1 – (b2 * (b3 * x)) etc.

The reason that these models are called nonlinear regression is because the relationships between the dependent and independent parameters are not linear.

hashtag#statistics hashtag#statisticsfordatascience

#65:Multiple Linear Regression

Multiple linear regression (also known as multiple regression), is a statistical technique that uses more than one explanatory (independent) variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory variables and response (dependent) variable. Actually, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.Example of Multiple Linear regression model is:

$y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon$

where,

y=dependent variable

xi=explanatory variables

β0=y-intercept (constant term)

βi=slope coefficients for each explanatory variable

ϵ=the model's residue term

Multiple Linear Regression is regressed to form a hyper plane with as many dimensions as number of explanatory variables.

hashtag#statistics hashtag#statisticsfordatascience

#66:Polynomial Regression

Polynomial regression is a regression in which the relationship between dependent and the independent variable is modeled such that the dependent variable Y is an nth degree function of independent variable X. The polynomial regression fits into a non-linear relationship between the value of X and the value of Y.

For example Y=β0+β1X+β2X^2 is polynomial model.

But still the parameters are of linear degree. So, it is a special case of Multiple linear regression.

hashtag#statistics hashtag#statisticsfordatascience

#67: Misunderstanding about Linear and Non-linear Regression

Many people think that the difference between linear and nonlinear regression is that linear regression involves lines (straight lines) and non-linear regression

involves curves. Some also think that the relation between variables (dependent and independent) are non linear. This is partly true. Linear regression equations can sometimes produce curves.

Actually, linearity or non-linearity is based on the parameters (which is the change in the mean response corresponding to a unit change in the corresponding independent variable term, also known as partial regression coefficients). If the parameter is linear in regression equation, then the regression is linear else non-linear. Linearity of independent variable is not considered.

E.g.

$Y=\beta0+\beta1X1+\beta2X2+...+\beta nxn+\epsilon$ and $Y=\beta0+\beta1X+\beta2X^2$ are linear whereas $Y=\beta0+\beta1\beta2X$ is non-linear.

Here $\beta0,\beta1,\beta2,\beta n$ are parameters.

hashtag#statistics hashtag#statisticsfordatascience

#68:Statistical Hypothesis

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. The null hypothesis, denoted by Ho, is usually the hypothesis that sample observations result purely from chance. It is the hypothesis the analyst believes to be true.

The alternative hypothesis, denoted by H1 or Ha, is the hypothesis that sample observations are influenced by some non-random cause and analyst believes it not to be true. Null and Alternative hypothesis are mutually exclusive. By conducting statistical experiments, (called hypothesis testing) we can accept(fail to reject) or reject the null hypothesis.

hashtag#statistics hashtag#statisticsfordatascience

#69:Hypothesis Testing

Hypothesis testing is the statistical process to make decision based on whether to reject a null hypothesis based on sample data. It consists of four steps:

1)State the hypotheses (Null Ho and Alternative H1 or Ha)

The hypotheses are stated in such a way that they are mutually exclusive i.e. if one is true, the other must be false.

2)Formulate an analysis plan

The analysis plan describes how to use sample data to evaluate the null hypothesis.

3)Analyze sample data

Find the value of the test statistic (mean score, proportion, t statistic, z-score, etc.) described in the analysis plan.

4)Interpret results

Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

E.g. Determine whether a coin was fair and balanced.

Ho: Half the flips would result in Heads and half, in Tails.

Ha: the number of Heads and Tails would be very different.

Mathematically,

Ho: $P = 0.5$

Ha: $P \neq 0.5$

Suppose a coin is flipped 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis. We would conclude, based on the evidence, that the coin was probably not fair and balanced.

hashtag#statistics hashtag#statisticsfordatascience

#70: Terms in Hypothesis Testing

Null Hypothesis(Ho): A statement about the population & sample data used to decide whether to reject that statement or not. Typically the statement is that there is no difference between groups or association between variables.

Alternative Hypothesis(H1 or Ha) is often the research question and varies depending on whether the test is one or two tailed.

Statistically Significance Level: The probability of rejecting the null hypothesis when it is true, (also known as a type 1 error).

Test Statistic: It is a value calculated from a sample to decide whether to reject or not reject the null (Ho) and varies between tests.

p-Value: The probability of obtaining a test statistic at least as extreme as ours if the null is true and there really is no difference or association in the population. A significant result is when the p-value is less than the chosen level of statistical significance (usually 0.05).

Type I error: Error occurred when we reject a null hypothesis when it is true.

Type II error:Error occurred when we fail to reject a null hypothesis that is false. The probability of committing a Type II error is called β. The probability of not committing a Type II error is called the Power of the test.

hashtag#statistics hashtag#statisticsfordatascience

#71:Covariance

Covariance is a measure of the relationship between two random variables. It evaluates how much(to what extent) the variables change together. In other words, it essentially measures the degree to which two variables are linearly associated. A positive covariance means the variables are positively related (they increase and decrease together), while a negative covariance means the variables are inversely related (when one increases, other decreases).

The formula of covariance is

$Cov(X,Y) = \Sigma (X-\bar{X})(Y-\bar{Y}) / n$

n = the number of items in the data set

The magnitude of the covariance depends on the magnitudes of the variables. Covarience has units. Its unit is product of units of the variables for which we are finding the covariance.

If two variables are independent, their covariance is 0. But, having a covariance of 0 does not imply the variables are independent.

hashtag#statistics hashtag#statisticsfordatascience

#72: Correlation Coefficient

The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. This is unitless quantity. It is a scaled version of covariance. The magnitude of this quantity signifies the extent of association and the sign signifies the direction of association. Positive correlation means the values both the variables increase or decrease together. Negative correlation means when the value one variable increases, the value of other decreases.

A value of ±1 implies perfect correlation.

A value of 0 implies that there is no linear correlation between the variables.

There are different types of correlation that can be used for different kinds of data. Pearson's Correlation Coefficient is one of those.

Formula for Pearson's Correlation Coefficient:

$r=\rho(X,Y)=COV(X,Y)/(\sigma x.\sigma y)$

hashtag#statistics hashtag#statisticsfordatascience

#73: Critical Region or Rejection Region in Hypothesis Testing

---

In order to test a hypothesis, the entire sample space is partitioned into two disjoint sub-spaces, say, ω and $S-\omega=\bar{\omega}$, where S is the sample space. If calculated value of the test statistic lies in ω, then we reject the null hypothesis and if it lies in $\bar{\omega}$, then we fail to reject the null hypothesis. The region ω is called a "rejection region or critical region" and the region $\bar{\omega}$ is called a "non-rejection region".

Therefore, we can say that the region in the sample space in which if the calculated value of the test statistic lies, we reject the null hypothesis then it is called critical region or rejection region. And critical value is the value or values that separate the region of rejection from the non-rejection region.

hashtag#statistics hashtag#statisticsfordatascience

#74: Level of Statistical Significance

The Statistical Significance level is a measure of the strength of the evidence that must be present in sample before we reject the null hypothesis and conclude that the effect is statistically significant. It is the probability of rejecting the null hypothesis when it is true(type-I error). It is denoted by alpha or α.

E.g., an α of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference. Lower α indicate that you require stronger evidence before you will reject the null hypothesis.

The researcher determines the significance level before conducting the experiment based on his prior knowledge and wisdom. α helps to find critical value(s) or cut-off value(s) for a known test statistic which is used to reject or fail to reject null hypothesis.

hashtag#statistics hashtag#statisticsfordatascience

#75: p-value and its interpretation

The p-value is the probability of getting a result at least as extreme as the one that is observed if H0 is true.

p-value depends on our critical region and test statistic.

α is used to refer to a pre-chosen probability and the "p-value" is used to indicate a probability that we calculate after a given study.

If the p-value< α, then we reject H0.

E.g. A drug company claims that a drug is effective in 90% case (i.e. the drug cures 90% of the patients who take the drug). The doctor tries to test company's claim and conducts a test on 15 patients. 11 patients are cured and 4 not. We need to test whether at least 90% of patients are cured by the drug, so this means that the null hypothesis is that p = 90%.

Here, test static is X ~ B(15, .9) , where X is the distribution for the number of people cured in the sample and B(15,.9) is binomial distribution of 15 patients with probability of 0.9.

H0: p=0.9

H1: p<0.9

Let's take α=0.05

p-value=P(X≤11)=1 - P(X ≥ 12)

= 1 - (C(15,12)x.1^3x.9^12 + C(15,13)x.1^2x.9^13 + C(15,14)x.1x.9^14 + C(15,15)x.9^15)

= 1 - (0.1285 + 0.2669 + 0.3432 + 0.2059)

= 1 - 0.9445

= 0.0555

As p-value(calulated value)>α (pre-chosen probability), we fail to reject the null hypothesis.

hashtag#statistics hashtag#statisticsfordatascience

#76:Population Proportion and Sample Proportion

A population proportion is a fraction of the population that has a certain characteristic that we want to study. It is part of a population with a particular attribute, expressed as a fraction or percentage of the whole population.

p = (number of favorable outcomes) / (number of outcomes in the population)

E.g., Suppose there are 1000 people in the city and 411 of those people have grey hair. The fraction of people who have grey hair is 411/1000.

Sample proportion, denoted as p^, is the fraction of sample size that has favorable outcome. It is the proportion of a sample of the population as opposed to the proportion of the whole population.

p^=(number of favorable outcome in sample)/(sample size)

In many real situations we can't check the whole population for the presence of an attribute. Here, the value of the population proportion is unknown which is estimated by taking a random sample from the population, calculating the sample proportion and using this value as an estimate of the population proportion.

The population proportion is part of a population so it is a population parameter.

A population proportion can be estimated through the usage of a confidence interval.

hashtag#statistics hashtag#statisticsfordatascience

#77: Different Types of Distributions

There are many types of distribution in Statistics classified based on the type of data. Below is the list of some distribution based on discrete and continuous data.

Discrete

- Benford
- Bernoulli
- Beta-binomial
- Beta-Pascal
- Binomial
- Discrete uniform
- Discrete Weibull
- Gamma-Poisson
- Geometric
- Hypergeometric
- Logarithm
- Negative hypergeometric
- Pascal

- Poisson
- Polya
- Power series
- Rectangular
- Zeta
- Zipf

Continuous

- Arcsin
- Arctangent
- Beta
- Cauchy
- Chi
- Chi-square
- Doubly noncentral F
- Doubly noncentral t
- Erlang
- Error
- Exponential
- Exponential power
- Extreme value
- F
- Gamma
- Gamma-normal
- Generalized gamma

- Generalized Pareto
- Gompertz
- Hyperbolic-secant
- Hyperexponential
- Hypoexponential
- IDB
- Inverse Gaussian
- Inverted beta
- Inverted gamma
- Kolmogorov-Smirnov
- Laplace
- Log gamma
- Log logistic
- Log normal
- Logistic
- Logistic-exponential
- Lomax
- Makeham
- Minimax
- Muth
- Noncentral beta
- Noncentral chi-square
- Noncentral F
- Noncentral t
- Normal

- Pareto
- Power
- Rayleigh
- Standard Cauchy
- Standard normal
- Standard power
- Standard triangular
- Standard uniform
- Standard Wald
- T
- Triangular
- TSP
- Uniform
- von Mises
- Weibull

hashtag#statistics hashtag#statisticsfordatascience

#78: Confidence Interval and Confidence Level

A confidence interval (CI) is a type of interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter whereas the confidence level represents the frequency (i.e. the proportion) of possible confidence intervals that contain the true value of the unknown population parameter. Simply put, a confidence interval is a range of

values that is likely to contain an unknown population parameter. If you draw a random sample many times, a certain percentage of the confidence intervals will contain the population mean. This percentage is the confidence level.Confidence interval is calculated using the confidence level required by the user with the help of z table/t table/chi-square table based on the distribution.

E.g. In an experiment, an athlete runs and his average performance varies. Say, mostly his performance lies in the range of 21 seconds to 25 seconds. This term 'Mostly' is very subjective. For some it might be 99% of the times, and for some other it may be 80% of the times and so on. This specified range (21s to 25s) is the Confidence Interval and 99% or 80% is confidence level.

hashtag#statistics hashtag#statisticsfordatascience

#79: Bernoulli Distribution

The Bernoulli distribution is a discrete distribution having two possible outcomes namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

E.g. Think of race between Sports car and a bicycle with p as probability of bicycle winning the race and q as probability of losing. It looks obvious that car will win. So, the probability of bicycle winning may be 0.1 and losing 0.9.

The probability mass function is given by:

p(x)= 1-p, x=0

p, x=1

The expected value of a random variable X from a Bernoulli distribution is found as follows:

$E(X) = 1*p + 0*(1-p) = p$

The variance of a random variable from a Bernoulli distribution is:

$V(X) = E(X^2) – [E(X)]^2 = p – p^2 = p(1-p)^2$

In our example,

$E(X) = p = 0.1$

$V(X) = p(1-p)=).1*0.9=0.09$

There are many examples of Bernoulli distribution such as whether it's going to rain tomorrow or not where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game.

hashtag#statistics hashtag#statisticsfordatascience

#80:Uniform Distribution (Continuous)

——

The uniform distribution is a probability distribution and is concerned with events that are equally likely to occur.

A variable X is said to be uniformly distributed if the density function is:

$f(x) = 1/(b-a)$ for $a \leq x \leq b$ where a and b are the parameters.

The shape of the Uniform distribution curve is rectangular, the reason why Uniform distribution is also called rectangular distribution.

E.g. You arrive into a building and are about to take an elevator to the your floor. Once you call the elevator, it will take between 0 and 40 seconds to arrive to you. We will assume that the elevator arrives uniformly between 10 and 40 seconds after you press the button. In this case a = 10 and b = 40

Let's try calculating the probability that the wait time will fall between 15 and 30.

The probability that wait time will fall between 15 and 30 is (30-15)*(1/(40-10)) = 0.5

Similarly, the probability that wait time is greater than 20 is = 0.667

The mean and variance of X following a uniform distribution is:

Mean -> E(X) = (a+b)/2

Variance -> V(X) = $(b-a)^2/12$

The standard uniform density has parameters a = 0 and b = 1, so the PDF for standard uniform density is given by:

$f(x)=1, 0\leq x \leq 1$

=0, otherwise

hashtag#statistics hashtag#statisticsfordatascience

#81:Binomial Distribution

——

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

The outcomes need not be equally likely. Each trial is independent since the outcome of the previous toss doesn't determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number

of times is called binomial experiment. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

The properties of a Binomial Distribution are:

-Each trial is independent.

-There are only two possible outcomes in a trial- either a success or a failure.

-A total number of n identical trials are conducted.

-The probability of success and failure is same for all trials. (Trials are identical).

The mathematical representation of binomial distribution is:

P(x)=C(n,x)(p^x)(q^(n-x))

The mean and variance are:

μ = n*p

Var(X) = n*p*q

A Bernoulli distribution is a special case of binomial distribution. Specifically, when n=1 the binomial distribution becomes Bernoulli distribution.

hashtag#statistics hashtag#statisticsfordatascience

#82:Normal Distribution

—

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.

It represents the behavior of most of the situations in the universe and used in widespread application.

It is highly different from Binomial Distribution. However, if the number of trials approaches infinity then the shapes will be quite similar.

The PDF of a random variable X following a normal distribution is:

f(x)=(1/(√2πσ))*e^{-1/2((x-μ)/σ)²)} for -∞<x<∞

The mean and variance of a normally distributed random variable X is:

E(X) = μ

Var(X) = σ²

Here, μ and σ are the parameters.

The distribution is denoted as as X ~ N (μ, σ).

A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1.

The PDF becomes:

f(x)=(1/(√(2π)))*e^(-x²/2) for -∞<x<∞

In data science, most of the times the distribution is either assumed to be normally distributed or the distribution is transformed to be normally distributed.

hashtag#statistics hashtag#statisticsfordatascience

#83: Memorylessness

The memorylessness property is the property that a given probability distribution (PD) is independent of its history. If a probability distribution has the memorylessness property the likelihood of something happening in the future has no relation to whether or not it has happened in the past. The history of the function is irrelevant to the future. Any time may be marked down as time zero.

It is also called the forgetfulness property.

E.g. Tossing a fair coin is an example of probability distribution that is memoryless. Every time you toss the coin, you have a 50 percent chance of it coming up heads. It doesn't matter whether or not the last five times you tossed the coin and it came up consistently tails which indicates that the probability of heads in the next toss would be zero. It would still be 50% chance only.

Memorylessness can be observed in discrete as well as continuous distributions.

- Discrete

P(X > a + b | x > a) = P (x > b)

E.g. let a be 5 and b 10. If PD is memoryless, the probability X>15 if we know X>5 is exactly the same as the probability of X>10.

- Continuous

P(X>t+s|x>t)=P(x>s) for every real s,t

Poisson process has memorylessness property.

hashtag#statistics hashtag#statisticsfordatascience

#84: Stochastic Process

Stochastic process is something which develops randomly in time. It describes the movement of a random variable through time. So, it is also called random process. The word “stochastic” is derived from the Greek word stokhastikos, which means "to aim at; guess."

Mathematically, it can be defined as a series of random variables, {X(t):t€T} where t usually denotes time i.e. at every time t in the set T, a random number X(t) is observed. Stochastic process can be discrete-time (set T has discrete time values) or continuous-time process (set T has continuous time interval [0,∞) or [0,k] for some k).

E.g. The random variable could be the closing price of a stock.Even if we know certain stock closed yesterday at ₹ 857.71, we cannot know for sure what value this variable will take at some future time t, as the process involves some degree of randomness.

In stochastic process, we don’t know exactly where our variable will be at next point in time. But if we know the position of an object now, and we know something of the possible moves it can make, and the probabilities for each possible move, we can draw some conclusions about the process, and determine which future positions are most probable.

hashtag#statistics hashtag#statisticsfordatascience

#85:Markov Process

Markov Process is a stochastic process whose future probabilities are determined by its most recent value(state). Mathematically, a stochastic process x(t) is called markov process if for every n and t1<t2...<tn, we have

P(x(tn)<=xn|x(t(n-1)),...,x(t1)) = P(x(tn)<=xn|x(t(n-1)))

This is equivalent to

P(x(tn)<=xn|x(t) for all t<=t(n-1)) = P(x(tn)<=xn|x(t(n-1)))

E.g. Movement of a drunkard. A drunkard moves randomly. We can ascertain the probability of his next move just on the current move not on any of his previous moves.

Markov process is a memoryless process. It is independent of the history but only dependent on last state. It is used to model Markov chain.

hashtag#statistics hashtag#statisticsfordatascience

#86:Poisson Process

Poisson Process is a counting process for a series of discrete events where the average time between events is known, but the exact timing of events is random.The occurance of an event is independent of the event before that means waiting time between events is memoryless. The average time between events but they are randomly spaced meaning they are stochastic.

Criteria of process be called Poisson Process:

- Events are independent of each other. The occurrence of one event does not affect the probability another event will occur.

- The average rate (events per time period) is constant.

- Two events cannot occur at the same time.

E.g.Arrival of customers at teller counter can be described by a Poisson process. Arrival of one particular customer is not dependent of arrival of any other customer. Also, average rate of footfall is constant. We know that on an average how many customer visit teller. Two customer cannot reach teller counter at same time.

Few more examples are occurring of earthquake in an area, visitor of a website, requesting for a document on website etc.

hashtag#statistics hashtag#statisticsfordatascience

#87:Poisson Distribution

Poisson Distribution is discrete probability distribution which follows following assumptions to be valid:

- Any successful event should not influence the outcome of another successful event.

- The probability of success over a short interval must equal the probability of success over a longer interval.

- The probability of success in an interval approaches zero as the interval becomes smaller.

In short, Poisson Distribution follows Poisson process.

is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

For Poisson Random Variable X, Let μ denote the mean number of events in an interval of length t. Then, μ = λ*t, where λ is the rate at which an event occurs and t is the length of a time interval

The PMF of this Poisson distribution will be:

P(X=x)=e^(-μ)*(μ^x)/(x!) for x=0,1,2,3,...

or,

P(x event in time period) =e^(-(events/time)*(time period))*(((events/time)*(time period))^x)/(x!)

The mean μ is the parameter of this distribution.

The mean and variance of X following a Poisson distribution:

E(X) = μ

Var(X) = σ^2 = μ

hashtag#statistics hashtag#statisticsfordatascience

#88: Exponential Distribution

Exponential Distribution is a continuous probability distribution used to model the time we need to wait before a given event occurs.

A random variable X is said to have an exponential distribution with PDF:

$f(x) = \lambda e^{(-\lambda x)}, \ x \geq 0$

and parameter λ>0 which is also called the rate.

Exponential distribution is widely used for survival analysis such as expected life of bulb, expected life of a human etc.

For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.

The break down of probability distribution is,

$P(X \leq x) = 1 - e^{(-\lambda x)}$, corresponds to the area under the density curve to the left of x.

$P(X > x) = e^{(-\lambda x)}$, corresponds to the area under the density curve to the right of x.

$P(x1 < X \leq x2) = e^{(-\lambda x1)} - e^{(-\lambda x2)}$, corresponds to the area under the density curve between x1 and x2.

Mean and Variance of a random variable X following an exponential distribution:

$E(X) = 1/\lambda$

$Var(X) = (1/\lambda)^2$

hashtag#statistics hashtag#statisticsfordatascience

#89:Geometric Distribution

Geometric Distribution is a special case of the negative binomial distribution. It represents the number of failures before you get a success in a series of Bernoulli trials.

The geometric distribution has three assumptions:

- There are two possible outcomes for each trial (success or failure).

- The trials are independent.

- The probability of success is the same for each trial.

This discrete probability distribution is represented by the probability density function:

f(x) = ((1 − p)^(x − 1))*p where X is a random variable which denote number of trial to get 1st success, p is the probability of success.

Theoretically, there are an infinite number of geometric distributions. The value of any specific distribution depends on the value of the probability p.

It is the only discrete memoryless random distribution.

E.g. Ordinary die thrown repeatedly until the first time a "1" appears. The probability distribution of the number of times it is thrown is supported on the infinite set { 1, 2, 3, ... } and is a geometric distribution with p = 1/6.

Mean and Variance of a random variable X following an geometric distribution:

$E(X)= 1/p$

$Var(X)=(1-p)/p^2$

hashtag#statistics hashtag#statisticsfordatascience

#90: Relations between Probability Distributions

The distributions are interrelated. Change in parameters changes the distribution.

Relation between the distributions are:

Bernoulli and Binomial:

- Bernoulli Distribution is a special case of Binomial Distribution with a single trial
- Only two possible outcomes of a Bernoulli and Binomial distribution, success and failure
- Both Bernoulli and Binomial Distributions have independent trails

Poisson and Binomial:

Poisson Distribution is a limiting case of binomial distribution with the conditions:

- the number of trials is indefinitely large ($n \to \infty$)
- the probability of success for each trial is same and indefinitely small or $p \to 0$, $np = \lambda$, is finite

Normal and Binomial:

Normal distribution is another limiting form of binomial distribution with the conditions:

- The number of trials is indefinitely large ($n \to \infty$)
- Both p and q are not indefinitely small

Normal and Poisson:

- The normal distribution is also a limiting case of Poisson distribution with the parameter $\lambda \to \infty$

Exponential and Poisson:

○ If the times between random events follow exponential distribution with rate λ, then the total number of events in a time period of length t follows the Poisson distribution with parameter λt

hashtag#statistics& hashtag#statisticsfordatascience

#91: Factors Affecting Confidence Intervals

There are 3 factors that determine the size of the confidence interval for a given confidence level. These are:

Sample Size

The larger the sample, the more sure one can be that their answers truly reflect the population (smaller the confidence interval). However, the relationship is not linear (i.e., doubling the sample size does not halve the confidence interval).

Percentage

Accuracy also depends on the percentage of sample that picks a particular answer. If 99% of sample said “Yes” and 1% said “No” the chances of error are remote, irrespective of sample size. However, if the percentages are 51% and 49% the chances of error are more. When determining the sample size needed for a given level of accuracy we must use worst case percentage (50%).

Population Size

The size of the population is irrelevant, unless the size of the sample exceeds a few percent of the total population. This means that a sample of 500 is equally useful in examining the opinions of a state of 15,000,000 as it would a city of 100,000.

That's why, the sample calculator ignores the population size when it is large or unknown. Population size is only likely to be a factor when we work with a relatively small data.

hashtag#statistics hashtag#statisticsfordatascience

#92: Sample Size based on Confidence Interval

Sample is the part of the population that helps us to draw inferences about the population. Collecting the complete information about the population is not possible and it is time consuming and expensive. Thus, we need an appropriate sample size so that we can make inferences about the population based on that sample.

Variables affecting sample size:

- Population Size

- Margin of Error (Confidence Interval)

- Confidence Level: confidence level corresponds to a Z-score.

- Standard of Deviation

Necessary Sample Size = $(\text{Z-score})^2 * \text{StdDev}*(1-\text{StdDev}) / (\text{margin of error})^2$

E.g. How many samples are required to find out estimate of population parameter with 95% confidence interval, std dev = 0.5 and margin of error 5%?

We can use the above formula to find necessary sample size.

hashtag#statistics hashtag#statisticsfordatascience

#93:Parametric Data and Non-parametric Data

Data that is assumed to have been drawn from a particular distribution is called parametric data and it is used in a parametric test.

Data that does not assume anything about underlying distribution is called non-parametric data and it is used in a non-parametric test. By non-parametric data we usually mean that the population data does not have a normal distribution.

E.g., one assumption for the one way ANOVA is that the data comes from a normal distribution. If your data isn't normally distributed, we can't run an ANOVA, but we can run the non-parametric alternative—the Kruskal-Wallis test.

Whenever possible, we should apply parametric tests, as they tend to be more accurate. Parametric tests have greater statistical power, which means they are likely to find a true significant effect. We should apply non-parametric tests only if we have to (i.e. we know that assumptions like normality are being violated). Non-parametric tests can perform well with non-normal continuous data if we have a sufficiently large sample size (generally 15-20 items in each group).

hashtag#statistics hashtag#statisticsfordatascience

Illustrating Central Limit Theorem

We have studied Central Limit Theorem that "if you have a population and we take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally

distributed regardless of whether the source population is normal or skewed." But we have not realized it actually. Sharing a python program to illustrate Central Limit Theorem.

https://lnkd.in/fMh_2ww

hashtag#statistics hashtag#statisticsfordatascience

#94: Parametric and Non-parametric Tests

Parametric tests are the hypothesis tests that provide generalization for making statements about the mean of the population and make assumptions about the parameters of the population from which the sample is drawn. This is often assumed that the population data are normally distributed. It is done upon interval or ratio type of data. Mostly, mean is the measure of central tendency. Information about population is known.

E.g. Student's t-test.

Non-parametric tests are the hypothesis tests which are independent of population distribution and also applicable when the observations are not measured in numerical scale that is in ordinal scale or nominal scale. It is also called "distribution-free" test and can be used for non-Normal variables. These test is mainly based on differences in medians.

E.g. $\chi 2$-test.

hashtag#statistics hashtag#statisticsfordatascience

#95: Key Differences Between Parametric and Non-parametric Tests

---

Main fundamental differences between parametric and non-parametric test:

- Parametric test is a statistical test, in which specific assumptions are made about the population parameter. Non-parametric test is a statistical test applied in non-metric independent variable.

- In parametric test, the test statistic is based on distribution. Whereas, in non-parametric test, the test statistic is arbitrary.

- In parametric test, we assume that the measurement of variables of interest is on interval or ratio scale. Whereas in non-parametric test, the variables of interest are measured on nominal or ordinal scale.

- In general, the measure of central tendency in the parametric test is mean, while in the non-parametric test it is median.

- In parametric test, complete information about the population is known. But, in the non-parametric test, there is no information about the population.

- The parametric test is applicable for variables only, but non-parametric test applies to both variables and attributes.

- Degree of association in two quantitative variables in parametric test is measured as Pearson's coefficient, while spearman's rank correlation is used in non-parametric test.

hashtag#statistics hashtag#statisticsfordatascience

#96: Z-test

A z-test is a statistical hypothetical test used to determine whether two population means are different when the variances are known and the sample size is large.

The test statistic is assumed to have a normal distribution. A z-statistic, or z-score, is a number representing the result from the z-test. Z-statistic or z-score is a number representing how many standard deviations above or below the mean population a score derived from a z-test is. Z-table is used for calculating z-score.

Z-test can be used for many purposes such that:

- z-test for single proportion is used to test a hypothesis on a specific value of the population proportion.

- z-test for difference of proportions is used to test the hypothesis that two populations have the same proportion.

- z-test for single mean is used to test a hypothesis on a specific value of the population mean.

- z-test for single variance is used to test a hypothesis on a specific value of the population variance.

- z-test for testing equality of variance is used to test the hypothesis of equality of two population variances when the sample size of each sample is 30 or larger.

hashtag#statistics hashtag#statisticsfordatascience

#97: Student's t-test

The Student’s t-test or t-test is a statistical hypothesis test in which the test statistic follows a Student’s t-distribution under the null hypothesis. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. T-test uses means and standard deviations of two samples to make a comparison. T-test is performed when sample size is small (generally less than 30). T-table is used for calculating t-score.

There are three different types of t-tests:

- An Independent Samples t-test compares the means for two groups.

- A Paired sample t-test compares means from the same group at different times (say, one year apart).

- A One sample t-test tests the mean of a single group against a known mean.

hashtag#statistics hashtag#statisticsfordatascience

#98: Analysis of variance (ANOVA)

--

ANOVA is a statistical analysis tool that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. ANOVA is used to determine the influence of independent variables on the dependent variable in a regression study. It is the extension of the t- and z-tests. It is also called Fisher analysis of variance.

The Formula for ANOVA is:

F=MST/MSE

where:

F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

Types of ANOVA:

- One-Way ANOVA

It has just one independent variable. E.g., difference in IQ can be assessed by country, and county can have 2, 20, or more different categories to compare.

- Two-Way ANOVA

It refers to an ANOVA using two independent variables which is used to examine the interaction between the two independent variables. E.g., IQ based on country and gender.

- N-Way ANOVA

It is ANOVA using n(>2) independent variables to examine interaction between those variables. E.g., potential differences in IQ scores can be examined by Country, Gender, Age group, Ethnicity, etc, simultaneously.

hashtag#statistic hashtag#statisticsfordatascience

#99: Analysis of Covariance (ANCOVA)

ANCOVA is a statistical analysis tool that is a blend of ANOVA and regression. It is similar to factorial ANOVA. It can tell us what additional information we can get by considering one independent variable (factor) at a time, without the influence of the others. It is used in examining the differences in the mean values of the dependent variables that are related to the effect of the controlled independent variables while taking into account the influence of the uncontrolled independent variables. It can be used as:

-An extension of multiple regression to compare multiple regression lines

- An extension of ANOVA

General steps for ANCOVA:

- Run a regression between the independent and dependent variables

- Identify the residual values from the results

- Run an ANOVA on the residuals.

Assumptions for ANCOVA:

- Basically the same as the ANOVA assumptions

- independent variables (minimum of two) should be categorical variables

- dependent variable and covariate should be continuous variables (measured on an interval scale or ratio scale.)

- observations are independent

- dependent variable should be roughly normal for each of category of independent variables

- data should show homogeneity of variance

hashtag#statistics hashtag#statisticsfordatascience

#100: Rules Governing Probability Distributions

There are two rules that tell where most of the data values lie in a probability distribution.

1. The empirical rule for normal distributions

- About 68% of data lie within 1 std. dev. of the mean

- About 95% of data lie within 2 std. dev. of the mean

- About 99.7% of data lie within 3 std. dev. of the mean

only knowing the number of standard deviations from the mean can give a rough idea about the probability.

2. Chebyshev's rule for any distribution

A similar rule applies to any set of data called Chebyshev's rule, or Chebyshev's inequality. It states that for any distribution

- At least 75% of data lie within 2 std. dev. of the mean
- At least 89% of data lie within 3 std. dev. of the mean
- At least 94% of data lie within 4 std. dev. of the mean

Chebychev's rule is not as precise as the empirical rule, as it only gives the minimum percentages, but it still gives a rough idea of where values fall in the probability distribution. The advantage of Chebyshev's rule is that it applies to any distribution, while the empirical rule just applies to the normal distribution.

hashtag#statistics hashtag#statisticsfordatascience