Statistics

Onkar Pandit

03 February 2022

Last updated on September 2, 2022.

Contents

Statistics is a branch of mathematics that deals with data. It consists of thorough techniques to first acquire useful data so that we can answer questions which we intend to solve and then derive meaningful insights from the data. This was the earlier science branch which dealt with data, thus the modern day, predictive modelling algorithms developed in machine learning and new branch which is based on machine learning algorithms: data science, extensively used statistical techniques. Hence, learning statistics is useful for understanding concepts of machine learning and data science.

1 Introduction

Various branches of mathematics are designed to solve different real world examples. Calculus was developed when humans wanted to understand the speed of certain object, force required to move object, or to measure volume of objects. On the similar lines, to answer questions that raised due to data, statistics was developed. Suppose, a government body wants to design a response for a pandemic in a certain area. Then for that we should first understand how pandemic will affect such the people, how people will react to the pandemic? In another scenario, suppose world health organization wants to know the effect of poverty on the health. These cases demand careful collection of relevant data and accurate analysis of the data. To answer the first question, we obtain the historical data to see how severely similar disease has affected the people as well how people handled a pandemic. If such a data is unavailable from the specific region we can obtain the data from nearby region with the assumption that the people will behave similarly. By doing such a study, a government can make a educated guess and design appropriate response. Similarly, WHO can obtain the data historical data to see how the disease has affected based on the income of people. Both the problems require, thorough investigation of the data with the statistical techniques.

Broadly statistical techniques can be categorised into two types–descriptive statistics and inferential statistics. Descriptive statistics provide meaningful insights about the data at hand with different tools whereas inferential statistical approaches provide information about the unseen data. The difference between descriptive and inferential statistics lies in the process as much as it does in the reported statistics.

Descriptive statistics mainly measure the central tendency and dispersion of the data which provide meaningful summary of the data. The central tendency measures like, mean, mode, and median of the data tell where the most values of the data fall. Whereas absolute mean deviations, variance, and coefficient of deviation inform the spread of the data. Likewise *skewdness* and *kurtoisis* give more idea about the distribution of the data. In addition to these quantitative measures, valuable insights can be derived by graphing the data with different plots such as histograms, box plots, and whisker plots.

The descriptive statistical methods only produce insights about the seen data. However, most of the time in the real world problems, we want to get

information about the unseen larger data of which the seen data is subset. In these scenarios, inferential statistical techniques are employed. For inferential statistics, we need to define the population and then devise a sampling plan to produce a representative sample. For selection of such a representative samples, different strategies are applied such as simple random sampling, stratified sampling, cluster sampling, and systematic sampling. From the obtained sample, various parameters of the actual population are estimated such as mean, standard deviation, etc. To make the guess of such parameters as accurate as possible different *estimation* approaches are proposed, which predict a single value or the range of the estimation with the use of concept of *confidence intervals*. *Regression analysis* is another types of inferential statistics which reveals the relation between different features of the data. One of the common problems in statistics is to understand the effectiveness of the newly proposed drug, treatment, or approach. Such problems arise regularly in healthcare, manufacturing industry, web industry, etc. Different *hypothesis tests* are proposed to address such problems.

In comparison to descriptive statistics, inferential statistics is harder to perform. However, it gives insights about the overall understanding of the population data as it is not limited only over the sample data.

Statistics vs Probability We are going to see which are the different techniques are used to get the important insights from the data. One of the important attribute of the data is chance associated with occurrence of it i.e. probability distribution. Thus, statistics and probability are closely studied but in my opinion probability is one of the tool in the arsenal of statistics which measures the certain aspect of data —chance of occurrence. In other words, The relationship between those two is that in statistics, we apply probability (probability theory) to draw conclusions from data. Probability is a branch of pure mathematics–probability questions can be posed and solved using axiomatic reasoning, and therefore there is one correct answer to any probability question.

Statistical questions can be converted to probability questions by the use of probability models. Once we make certain assumptions about the mechanism generating the data, we can answer statistical questions using probability theory. HOWEVER, the proper formulation and checking of these probability models is just as important, or even more important, than the subsequent analysis of the problem using these models.

One could say that statistics comprises of two parts. The first part is the question of how to formulate and evaluate probabilistic models for the problem; this endeavour lies within the domain of "philosophy of science". The second part is the question of obtaining answers after a certain model has been assumed. This part of statistics is indeed a matter of applied probability theory, and in practice, contains a fair deal of numerical analysis as well.

2 Design of experiments

In some scenarios statisticians have control over the process of acquiring data. For example, a private firm wants to predict the next president, then they need to carefully gather a data which can be used to analyze and predict the results of the elections. In these kind of scenarios, the careful design of experiments is needed.

2.1 Controlled Experiments

In these kind of experiments, typically, scientists have control over the experiments and can decide the procedures. For example, to determine the effectiveness of certain drug in treating a disease, scientists can control the whole experiment. To accurately carry out the experiments requires careful design. To do that it is important to take following points into consideration:

- *Similar distributions in both groups*: Suppose we want to understand the effectiveness of certain drug, for that we devise a strattegy where we give a previous treatment to one group and newly developed drug to other group. Then we measure the effectiveness on each group by seeing the effectiveness of the treatment. Then while selecting these two groups we have to insure that all factors (financial condition, gender, age, etc.) are same except the kind of treatment given to both the groups. Thus we can confidently say that the difference in the curability is only because of the drug and not any other factor.

- *Randomized controlled experiments* To insure that both the groups–treatment group on which the real drug is administered, and control group with placebo or no treatment–are selected randomly from the pool of people.

- *Placebo* To make sure that the response is obtained for the actual drug not to the idea of the drug, instead of administering no dose, placebo is given to the control group.

- *Double Blind* To reduce any bias introduced because of the knowledge of which group is given what kind of treatment, generally, those people who administer the treatment do know what kind of treatment they are giving–actual drug or placebo. Similarly, the patients also do not know whether they got actual treatment or not.

2.2 Observational Studies

Observational studies are different than controlled experiments. These studies are done when it is not possible to do controlled experiments. Suppose we want to understand the risk of smoking, then it is not possible to do a controlled experiment as no one in sane mind will smoke for ten years so as to please a statistician. In these kind of problems, observational studies are done where scientists observe smokers and non-smokers so as to gauge the risk.

As there is no way to ensure same factors between these two groups before the experiments, there is a high risk of *confounding factors* which may affect results i.e. apart from actual effect there can be other issues which may affect the study. In case of smokers vs non-smokers, gender of people can be confounding factor as there are generally more male smokers than females and also in general males have higher risk of heart diseases. Thus, if all smokers and non-smokers are considered then it will show worsened effects of smoking than actually it is. Therefore, a better strategy would be to compare male non-smokers to male smokers and female non-smokers to female non-smokers. This shows, while doing observational studies, it makes sense to compare more homogeneous set of groups as possible to get the correct observation and effective insight.

Observational studies can establish association: one thing is linked to another. Association may point to causation: more people exposed to smoking are higher at risk of contracting hart diseases but generally it does not prove that it is a cause. There can be some other factors like maybe the reason behind people smoke (e.g. stress) is also cause of heat disease. These confounding factors are weakness of observational studies which are minimized in randomized controlled experiments.

Few terminologies:

- *Controls*: Those subjects which were not given a treatment and were generally given neutral placebo.
- *Controlled experiments*: where experts decide who can be in experiment and who are not.
- *Controlling* for confounding effects is a technique where small homogeneous groups are compared to minimize the effect of other factors.
- *Historical controls* who are subjects from the past which had undergone different experiments. In case of drug trial, past patients who have been given old drugs.
- *Contemporaneous controls* are opposite of historical controls i.e. those controls which are present at the study.
- *Cross-sectional studies* are done over different objects at the one-point in time, e.g. measurement of hieght of different people.
- *Longitudinal studies* where objects are followed over different point of times, e.g. smokers vs non-smokers where same subjects are studied over time.

3 Descriptive Statistics

When we obtain a sample, it is important to understand few basic properties so as to make appropriate plans for the further studies. For example, in data science if we first understand where the data is clustered and where it is spread, what is the degree of skewness then we can further design experiments for better results. In this section we note down different techniques to understand the data easily.

3.1 Central Tendency of the Data

These attributes describe the central behaviour of the data.

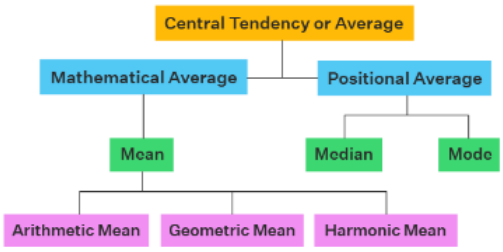

Figure 1: Measure of central tendency.

3.1.1 Arithmetic Mean

value is calculated by summing all values and dividing it by total number of items. Let $x_1, x_2, x_3, \cdots, x_n$ be n observations. We can find the arithmetic mean (average[1]) using the formula:

$$\mu = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

3.1.2 Geometric Means

It is a another type of mean where instead of taking sum of all the data points, product is taken and its n^{th}-root is considered as a geometric mean.

$$g_m = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{2}$$

Suppose we are given a 9×4 rectangle and we want to know what is the length of side of square having same area as the rectangle, then we take a geometric

[1] Mean is a specific type of average but their can be other averages as well such as weighted mean, harmonic mean, geometric mean, etc.

mean to answer this. In nutshell, what we are doing is taking something which is not evenly distributed but making it evenly distributed. This seems really non-trivial example but let see its application in finance. Suppose from a Tesla stocks, a company made a profit of 10% in the first year, 20%, and 30% in following years. What is the rate of return per year? This example is similar to what we did in the last example. Each year we got the return of 1.1, 1.2, and 1.3, and we take geometric mean: $\sqrt[3]{1.1 \times 1.2 \times 1.3}$ which comes to be 1.1972, which means yearly return of 19.72%.

3.1.3 Harmonic Means

Harmonic mean is calculated as follows:

$$h_m = \frac{n}{\sum_{i=1}^{n}(\frac{1}{x_i})} \tag{3}$$

3.1.4 Trimmed mean

Mean is suseptible to the outliers, as it can drive the value upwords or downwords, therefore trimmed mean is considered. The values are arranged in ascending order and then few values from the up and down are deleted. The trimmed mean is calculated on the remaining values which will reduce the effect of extreme values.

$$\bar{x}_{trimmed} = \frac{1}{n-2p} \sum_{i=1+p}^{n-p} x_i \tag{4}$$

3.1.5 Weighted mean

is calculated by multiplying each data value x_i by a user-specified weight w_i and dividing their sum by the sum of the weights. The formula for a weighted mean is:

$$\bar{x}_{weighted} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i x_i \tag{5}$$

3.1.6 Mode

The value which appears most often in the given data i.e. the observation with the highest frequency is called a mode of data.

For example in the data: 6, 8, 9, 3, 4, 6, 7, 6, 3 the value 6 appears the most number of times. Thus, mode = 6. An easy way to remember mode is: Most Often Data Entered. Note: A data may have no mode, 1 mode, or more than 1 mode. Depending upon the number of modes the data has, it can be called unimodal, bimodal, trimodal, or multimodal. The example discussed above has only 1 mode, so it is unimodal.

Suppose we are given probability density function or mass function instead of actual observations of the data, then the data point which has the highest probability is considered as a mode. For example, in normal distribution, value at which the peak is attained is called mode of the data.

3.1.7 Median

The value of the middle-most observation, obtained after arranging the data in ascending order, is called the median of the data. For n observation $\frac{n}{2}$th point if n is even otherwise $\frac{n+1}{2}$th point after arranging observations in ascending order.

For example, consider the data: 4, 4, 6, 3, 2. Let's arrange this data in ascending order: 2, 3, 4, 4, 6. There are 5 observations. Thus, median = middle value i.e. 4. We can see here: 2, 3, 4, 4 , 6 (Thus, 4 is the median)

Relation Between Mean, Median and Mode The three measures of central values i.e. mean, median, and mode are closely connected by the following relations (called an empirical relationship).

$$2\text{Mean} + \text{Mode} = 3\text{Median} \tag{6}$$

For instance, if we are asked to calculate the mean, median, and mode of continuous grouped data, then we can calculate mean and median using the formulas as discussed in the previous sections and then find mode using the empirical relation.

3.1.8 Percentiles, Quartiles, Deciles

Percentiles divide data into 100 parts, so 1 percentile is the value above which lie 99% of the data. Similarly, *quartiles* divide the data into 4 parts whereas the *deciles* divide data into 10 parts.

Before obtaining any of the *-iles*, the data is arranged in ascending order, so that we can correctly say which values are above it and which are before.

3.1.9 Range and IQR

Suppose we have an ordered dataset, then the *range* is nothing but the difference between maximum and minimum value.

$$\text{range} = x_{max} - x_{min} \tag{7}$$

With the range value we can get an idea about the spread of the data, however it is susceptible to outliers. If there is extreme outlier at the end of the ordered list then the range will not give a true picture of the data. To tackel that, Inter Quartile Range (IQR) is used. IQR is the difference between the third quartile and the first quartile of the data, which tells where the fifty percent of the data lies.

$$IQR = Q_3 - Q_1 \tag{8}$$

3.2 Dispersion of data

After seeing the attributes to get the central tendency of the data, let us see some statistics to understand the spread of the data. The most widely used estimates of variation are based on the differences, or deviations, between the mean value and the observed data.

3.2.1 Absolute mean deviation

It is computed as:

$$\text{Mean absolute deviation} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{9}$$

where $\bar{x}$ is a sample mean.

3.2.2 Variance and Standard Deviation

It is a measure of how far a set of data are dispersed out from their mean or average value. It is denoted as σ^2. It is always non-negative since each term in the variance sum is squared and therefore the result is either positive or zero.

Variance for entire population is calculated as

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x-\mu)^2}{n} \tag{10}$$

where μ is a population mean. Similarly, sample variance is calculated as:

$$S^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1} \tag{11}$$

where $\bar{x}$ is a sample mean.

Standard Deviation is a square root of variance, generally denoted as σ.

$$\text{Standard deviation} = \sqrt{\text{Variance}} = S \tag{12}$$

3.2.3 Median absolute deviation

Neither the variance, the standard deviation, nor the mean absolute deviation is robust to outliers and extreme values. The variance and standard deviation are especially sensitive to outliers since they are based on the squared deviations.

A robust estimate of variability is the median absolute deviation from the median or MAD which is calculated from median m as:

$$\text{Median absolute deviation} = \text{Median}(|x_1 - m|, |x_2 - m|, \cdots, |x_N - m|) \tag{13}$$

The variance, the standard deviation, the mean absolute deviation, and the median absolute deviation from the median are not equivalent estimates, even in the case where the data comes from a normal distribution. In fact, the standard

deviation is always greater than the mean absolute deviation, which itself is greater than the median absolute deviation. Sometimes, the median absolute deviation is multiplied by a constant scaling factor to put the MAD on the same scale as the standard deviation in the case of a normal distribution. The commonly used factor of 1.4826 means that 50% of the normal distribution fall within the range ±MAD.

3.2.4 Coefficient of variation

Suppose we are given two different samples: $s_1 = \{1, 2, 3\}$ and $s_1 = \{101, 102, 103\}$. In both of these samples, the standard deviation is same, but if we look closely the magnitude of variation is different. In the first sample, the last data point is three times of the first one, which is hugely different than the difference of only 1 in the second one. In these scenarios, standard deviation is not a good measure of dispersion of the data. Instead of that *coefficient of variation* is used to gauge the spread. Is calculated as the ratio of standard deviation (s) and the mean ($\bar{x}$).

$$cv = \frac{s}{\bar{x}} \tag{14}$$

If we calculate, coefficient of variation for the above two examples, it comes out to be 0.5 and 0.0098, which shows that the first sample is having more variation than the second one, which was failed with only standard deviation.

3.3 Moments

The concept of moments come from physics, where distance from the points is measured. Similarly, different orders of moments give notion of dispersion of data differently.

Order of Moment	Formula	Standardised formula	Sampling Adjustments
First order (Mean)	$\frac{\sum_{i=1}^{n} x_i}{n}$	–	–
Second order (Variance)	$\frac{\sum_{i=1}^{n} x_i^2}{n}$	$\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
Third order (Skewness)	$\frac{\sum_{i=1}^{n} x_i^3}{n}$	$\frac{\sum_{i=1}^{n}(x_i-\mu)^3}{n\sigma^3}$	$\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-1)(n-2)s^3}$
Fourth order (Kurtosis)	$\frac{\sum_{i=1}^{n} x_i^4}{n}$	$\frac{\sum_{i=1}^{n}(x_i-\mu)^4}{n\sigma^4}$	$\frac{n(n+1)\sum_{i=1}^{n}(x_i-\bar{x})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$

In Table ??, we noted four orders of moments and the general names for them. We have already explained mean and variance, in the following paragraphs we elaborate more on skewness and kurtosis.

Figure 2: Skewness of data.

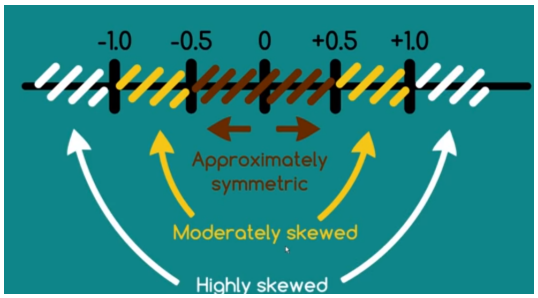

Figure 3: As the absolute value of skewness increases, data becomes more skewed.

Skewness There is a possibility that the distribution is skewed positively or negatively. It is shown in Figure 2.

This skewness of the data is quantified with the formula at row 3 in Table ??. Note that this is an one of the ways of obtaining skewness there have been different proposals such as pearson methods for finding skewness, however moment based calculation is more robust. As the absolute value of skewness increases, skewness of the data also increases, as depicted in Figure 3.

Kurtosis The earlier definition of kurtosis is said to be a value which measures the the "peakedness" of the distribution. But, recent research has clarified that in fact the formula for the kurtosis measures the spread of the data. The fourth line in Table ?? shows the formula for calculating kurtosis, which means, we taking the 4^{th}-power of the distance between value and mean, which means at the median (value where the distribution peaks), the numerator will be very small as the difference between that high value and mean will be lower. However, this will be higher at the edges of the graph, which means, in fact the formula for kurtosis is measuring the spread of the data and the flatness of the curve, because as the flatness increases the kurtosis will be higher.

3.4 Covariance and Correlation

Covariance or Correlation is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense corre-

lation is any statistical association, though it actually refers to the degree to which a pair of variables are linearly related. It is a measure of the strength of a linear relationship between two quantitative variables (e.g., height, weight). Familiar examples of dependent phenomena include the correlation between the height of parents and their offspring, and the correlation between the price of a good and the quantity the consumers are willing to purchase, as it is depicted in the so-called demand curve.

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. [2]

Let us first see the formula to calculate covariance:

$$\text{Covariance} = \sigma_{xy} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{(n-1)} \tag{15}$$

Covariance gives a value, if the absolute value is higher then there is higher correlation between the variables. But, it is difficult to understand the strength of the association between them. To understand this strength, covariance is normalized to get correlation.

To compute Pearson's correlation coefficient, we multiply deviations from the mean for variable 1 times those for variable 2, and divide by the product of the standard deviations:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{(n-1)S_x S_y} \tag{16}$$

Correlation matrix presents the correlation factors between n variables in a compact way in $n \times n$ matrix. If the rows and columns are labled with variable names in a same order, then each entry in the matrix indicates the correlation between the row variable and column variable.

Like the mean and standard deviation, the correlation coefficient is sensitive to outliers in the data. Software packages offer *robust* (robust means which is not affected by outliers) alternatives to the classical correlation coefficient. For example, the R package robust uses the function *covRob* to compute a robust estimate of correlation. The methods in the scikit-learn module *sklearn.covariance* implement a variety of approaches.

Statisticians long ago proposed other types of correlation coefficients, such as Spearman's ρ or Kendall's τ. These are correlation coefficients based on the rank of the data. Since they work with ranks rather than values, these estimates are robust to outliers and can handle certain types of nonlinearities. However, data scientists can generally stick to Pearson's correlation coefficient, and its robust alternatives, for exploratory analysis. The appeal of rank-based estimates is mostly for smaller data sets and specific hypothesis tests.

[2] In this example, there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling. However, in general, the presence of a correlation is not sufficient to infer the presence of a causal relationship (i.e., correlation does not imply causation).

3.5 Visualization

We saw different quantifying attributes which can give more information about the data at hand. In addition to these tools, we can leverage graphs which produces way for data visualization that can convey information easily. Following are two commonly used plotting techniques that are used for visualization.

3.5.1 Histogram

It is a plot to visualize the data as looking at the data graphically conveys information easily than looking only at the numbers. In a histogram, ranges are plotted on the x-axis whereas their corresponding frequencies are presented on y-axis. With this set-up it is clear to see the difference in the occurrence of ranges. Suppose we analyse the income of each home and plot the ranges of income from 0-1000, 1000-2000, 2000-3000, and so on, on the x-axis and either absolute number of homes in this ranges or percentages we can see the blocks which can convey in which income range most of the homes fall. We can get the good idea about the distribution of income.

There different factors to consider while plotting the graphs. In our example of income, we can take into consideration gender, education, or health etc. of the people and can plot different histograms to see how it affects the income. Plotting these different histograms each considering these factors can given better insights.

3.5.2 Box plots

In addition to these quantitative techniques, visually also we can guage the spread of the data. For that the values are plotted with box plots. Let us see a boxplot to get the idea.

The box plot presented in Figure 4 depicts the median population of states with the thick line in the box. The borders of the box show the 25%tile and 75%tile values, which means that the 50% values lie in that box. The whiskers of the box which are extended from both the ends, show the range of the most of the values.

Discussion: In general samples are differently distributed from the population and samples are the only available data most of the time. Thus, statistics are calculated over samples and based on these statistics population parameters are inferred. Intuitively, we can see that sample distribution will be more skewed than the actual population. Imagine we have population with normal distribution and mean μ and σ standard deviation. Then if we sample some data points from this distribution, then because of normal distribution, there is high chance that the selected samples will be nearer to μ. So we can visualize that the graph of distribution of population will be more spread than the

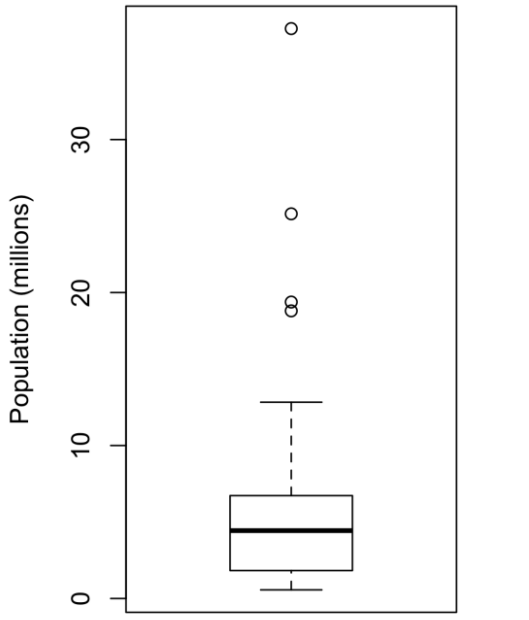

Figure 4: Box Plot over population of states.

distribution graph of samples but both will be mostly centred around μ. In a way, the mean and standard deviation values (in general all the statistics) calculated over samples will be different than the actual population values, therefore, statisticians distinguish them by denoting them differently. The greek letters like μ, σ are reserved for population mean, s.d. and $\bar{x}, S$ are used to denote sample statistics. There is also a slight deviation, in the calculation of sample mean whole sum is divided by $n-1$ and not n. However, these distinctions are not significant if the value of n is large which is the case in most of the data science scenarios.

There is a small difference between bar chart and histograms. Histograms are always binned and plotted where their buildings touch each other whereas this is not the limitations for bar charts.

3.6 Regression

Regression – as fancy as it sounds can be thought of as relationship between any two things. For example, imagine you stay on the ground and the temperature is 70°F. You start climbing a hill and as you climb, you realize that you are feeling colder and the temperature is dropping. When you reach the hilltop

which is 500 meters above ground level and you measure the temperature is 60°F. We can conclude that the height above sea level influences temperature. Hence, there is a relationship between height and temperature. This is termed *regression* in statistics. The temperature depends on height and hence is the *dependent* variable, whereas height is the *independent* variable. There may be various factors influencing the temperature such as humidity, pressure, even air pollution levels etc. All such factors have a relationship with the temperature which can be written mathematically as an equation.

Formally, any equation, that is a function of the dependent variables and a set of weights is called a regression function. $y\ f(x; w)$ where y is the dependent variable (in the above example, temperature), x are the independent variables (humidity, pressure etc) and w are the weights of the equation (co-efficients of x terms).

The question now is to figure out how to learn the weights of the equation. Why are we even doing this? Yes, we are doing this to make predictions going forward. Once we know the relationship between the dependent and the independent variables, we can predict the dependent variable beforehand. To learn the regression equation, we need to have some true data collected from the field. We humans, learn from real world experiences. Similarly, regression function can be learnt only by having an initial real data – termed as *training* data. We see detailed presentation of it again in machine learning.

4 Sampling

when we want to draw some conclusions about the data, it is not possible to get the whole data which is known as *population*. But, therefore people work smaller portion of the data which is representative of the actual data. This smaller portion is called as *sample*. The process of selecting such samples is known is sampling. It is important to select samples in unbiased way so that the conclusions drawn from them can be correct. If there is a selection bias in sampling then even though the data is larger, the conclusion will not be correct. In fact, the in-correctness will be increased with the increase in data. The wrong conclusion will become more and more confident. The sample is only part of the population, so the percentage composition of the sample usually differs by some amount from the percentage composition of the whole population.

A popular misconception holds that the era of big data means the end of a need for sampling. In fact, the proliferation of data of varying quality and relevance reinforces the need for sampling as a tool to work efficiently with a variety of data and to minimize bias. Even in a big data project, predictive models are typically developed and piloted with samples. Samples are also used in tests of various sorts (e.g., comparing the effect of web page designs on clicks).

4.1 Sampling Distribution of Statistic

Sample statistic is the attribute of the data which we are concentrating on. For example, suppose we want to measure the mean value of the observed data or samples, then in this case we call mean as a sample statistic. Then *sampling distribution* is the distribution of sample statistic obtained by considering different samples of the population data [3]. Consider a scenario, suppose mean is a sample statistic, now we sample the data suppose 100 times, this produces 100 different means over different samples of the data. This produces distribution of means which we call as sampling distribution. Note that this distribution will always tend to be normal distribution as stated by central limit theorem. *Central limit theorem* says that the sampling distribution tends to take on a normal shape as sample size rises, even if the source population is not normally distributed. Standard error is measured over these different samples for a given sample statistic. It is a variability (standard deviation) of a sample statistic over many samples [4].

4.2 Bootstrap

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.

Conceptually, you can imagine the bootstrap as replicating the original sample thousands or millions of times so that you have a hypothetical population that embodies all the knowledge from your original sample (it's just larger). You can then draw samples from this hypothetical population for the purpose of estimating a sampling distribution; see Figure 5.

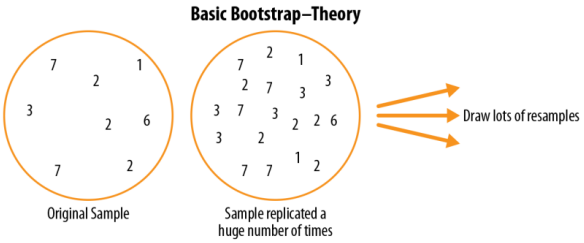

Figure 5: Bootstrap approach.

[3] Statisticians refer to whole dataset as a sample whereas in machine learning community sample is referred to a single data point.

[4] Standard error should not be confused with standard deviation, which by itself, refers to variability of individual data values

4.3 A/B test

An A/B test is an experiment with two groups to establish which of two treatments, products, procedures, or the like is superior. Often one of the two treatments is the standard existing treatment, or no treatment. If a standard (or no) treatment is used, it is called the control. A typical hypothesis is that a new treatment is better than the control.

A proper A/B test has subjects that can be assigned to one treatment or another. The subject might be a person, a plant seed, a web visitor; the key is that the subject is exposed to the treatment. Ideally, subjects are randomized (assigned randomly) to treatments. In this way, you know that any difference between the treatment groups is due to one of two things:

- The effect of the different treatments.
- Luck of the draw in which subjects are assigned to which treatments (i.e., the random assignment may have resulted in the naturally better-performing subjects being concentrated in A or B)

A blind study is one in which the subjects are unaware of whether they are getting treatment A or treatment B. Awareness of receiving a particular treatment can affect response. A double-blind study is one in which the investigators and facilitators (e.g., doctors and nurses in a medical study) also are unaware which subjects are getting which treatment. Blinding is not possible when the nature of the treatment is transparent, for example, cognitive therapy from a computer versus a psychologist.

A/B testing in data science is typically used in a web context. Treatments might be the design of a web page, the price of a product, the wording of a headline, or some other item. Some thought is required to preserve the principles of randomization. Typically the subject in the experiment is the web visitor, and the outcomes we are interested in measuring are clicks, purchases, visit duration, number of pages visited, whether a particular page is visited, and the like. In a standard A/B experiment, you need to decide on one metric ahead of time. Multiple behaviour metrics might be collected and be of interest, but if the experiment is expected to lead to a decision between treatment A and treatment B, a single metric, or test statistic, needs to be established beforehand. Selecting a test statistic after the experiment is conducted opens the door to researcher bias.

4.4 ANOVA

Suppose instead of comparison between two groups of effects, we want to test multiple groups suppose four. In these cases, we use ANOVA. The statistical procedure that tests for a statistically significant difference among the groups is called *analysis of variance*, or ANOVA.

The procedure used to test this is ANOVA. The basis for it can be seen in the following resampling procedure (specified here for the A/B/C/D test of web page stickiness):

1. Combine all the data together in a single box.
2. Shuffle and draw out four resamples of five values each.
3. Record the mean of each of the four groups.
4. Record the variance among the four group means.
5. Repeat steps 2-4 many (say, 1,000) times.

What proportion of the time did the resampled variance exceed the observed var- iance? This is the p-value.

4.5 Box model

In box model, it is considered that the elements are either numbered 1 or 0. Then from this box the numbers are drawn without replacement. Then the total number of 1s in the sample can be calculated by just summing the samples.

4.6 Chance error

Sampling faces a lot of challenges if we are taking particular section of the population. Suppose, we are a polling agency then, the sampling would involve selecting some people from the whole population and then understanding their inclination towards the electoral candidates. These surveys are faced with many obstacles, such as people absent at the time of survey, people not revealing their real choices, selection bias of surveyors, etc. Apart from these problems, there is an inherent problem of *chance error* even in the ideal conditions. Because, the samples are not real population they are always some part of the population and the sample is chosen at random, the amount off is governed by chance which can induce some error.

Chance error is often called *sampling error*: the error comes from the fact that the sample is only part of the whole. Similarly, bias is called *non-sampling error* – the error from other sources, like non-response. Bias is often a more serious problem than chance error, but methods for assessing bias are not well developed. Usually, *bias* means prejudice. However, statistics is a dry subject. For a statistician, bias just means any kind of systematic error in an estimate. Non-sampling error is a more neutral term, and may be better for that reason.

In simple words, sampling or chance error is nothing but the different distribution of the sample than population because of pure randomness. Suppose we have a population of 6672 people out of that 46% are male. Now, if we select 100 people randomly, then the distribution in the samples is not going to be 46% all the time. It can vary, sometimes it can be more and sometimes it can be less, few times it also can be exactly 46%. The variability in the resulting distribution in the samples is known as chance error.

4.7 Expected Value and Standard error

With a simple random sample, the expected value for the sample percentage equals the population percentage.

The standard error of the mean and the standard deviation is such that, for a given sample size, the standard error of the mean equals the standard deviation divided by the square root of the sample size. In other words, the standard error of the mean is a measure of the dispersion of sample means around the population mean.

Standard error for N number of samples $(e_s) = \frac{\sigma}{\sqrt{N}}$ where is σ is standard deviation.

4.8 Correction factor

Suppose there are 1.5 million people in New Mexico and 15 million in Texas and we want to poll in these two states. Now, two agencies select 2500 people each from these two states and give some predictions. Intuitively it seems that the prediction from New Mexico should be closer to the accurate result than Texas as we know that the population is higher in that state and we are selecting same number of people of even though with the difference in the population.

Though, it seems intuitively correct our statistical knowledge till now says otherwise. Suppose, SD in both the states is 0.5. Then the SE will be $\sqrt{2500}\times 0.5$ for both the states. Which means eventhough with the difference in the actual population our theory is saying that standard error will be same. This seems bit off, so we introduce a correcting factor. When drawing without replacement, to get the exact SE you have to multiply by the correction factor:

$$\frac{\text{Number of tickets in the box } - \text{ number of draws}}{\text{Number of tickets in the box } - 1} \tag{17}$$

The correcting factor is also close to 1 for both the states as the difference between population is not that big according to this factor!

This shows that the likely size of the chance error in sample percentages depends mainly on the absolute size of the sample, and hardly at all on the size of the population.

4.9 Inference

So far, we were knowing the percentages of 1s in the box and we were interested in seeing the chance error with difference size of the samples. Now, we want to go other way around. We want to *infer* about the population from the sample and then want to check by what margin we are off.

Suppose we want to see how many people vote for democrates. Then we take the sample of 2500 and see that 1328 people favor them which is 53%. Then are we sure that democrates are going to win. No! because there is chance error which we want to calculate to be sure. The standard error (SE) is calculated from number of samples and standard deviation (SD). But, we do not SD, as

those are unkonwns. So the neat trick is to calculate it from samples and use as the actual Sd. In the above example, $\sqrt{0.53 \times .47} \approx 0.5$ **(we have been writing sd as square root of multiplication of probabilities. where does this formula come?)**. This sd produces se as: $\sqrt{2500} \times 0.5$ which is 25 which is 1%. Then, with this understanding we can say that democrates are going to win as the estimation can not be wrong by 3 SEs.

Confidence interval We have seen that sample statistics from the previous example tell us that there is high chance that 79% of students from the university live at home. But, how can be sure? and with how much assurity. Confidence intervals tell us that. Suppose we are within ±1SEs, then there is a 68% chance that we are right. If we increase the range to ±2SEs then we are sure 95% that the population statistics is within that range. From the above example, we can say that there is 95% probability that the 77%-81% students from the university live at home.

A confidence interval is used when estimating an unknown parameter from sample data. The interval gives a range for the parameter, and a confidence level that the range covers the true value. A sample percentage will be off the population percentage, due to chance error. The SE tells you the likely size of the amount off. Confidence levels were introduced to make this idea more quantitative. The standard error depends on the sample taken for the inference as with different samples the sample probability changes in turn changing the standard deviation. Thus, the percentages we estimate for the population changes which solely depends on the sample at hand. Therefore, there is a possibility in the above example that we may get the estimation as 70% albeit rarely.

The above methods were developed for simple random samples. They may not apply to other kinds of samples. Many survey organizations use fairly complicated probability methods to draw their samples. As a result, they have to use more complicated methods for estimating their standard errors. Some survey organizations do not bother to use probability methods at all.

Employment surveys
We looked at the sampling approach where samples are selected randomly. But, in real world, this simple strategy of sampling is replaced with well designed sample selection. This also means that the standard error calculations which were presented here will not be directly applicable. In unemployment rate calculation, roughly the US government department divides the counties and selects some samples from those counties over the age of 16. Then the questions like their employment status, hours of work, last job, etc. are asked. Based on these surveys, it is observed that the black teenagers are mostly hit. Also, the subjects are chosen in cluster, like from the same locality, four houses in a row. This is also done to reduce the cost of interviewing, and this is the main reason behind not taking random samples from the whole population as they can be sparsely distributed throughout the country. Cluster samples are less informative than simple random samples of the same size. So the simple random sample formulas for the standard error do not apply. Instead of that from samples, samples are divided in half and difference between their estimates is measured to understand the accuracy.

In addition to that the chosen samples are weighted differently, depending on their low or high representation. Any group which is over-represented in the sample gets proportionately smaller weights, and under-presented groups get higher weights, bringing the sample back into line with the population. For example, suppose there are too many white males age 16–19 in the sample, relative to their share in the total population. Adjusting the weights this way helps to correct imbalances caused by chance variation. That reduces sampling error.

4.10 Accuracy of averages

Suppose there are 1 to 7 numbers and 25 numbers chosen from them with replacement. Then by how much the average will vary? We know that the box average is 4 and standard deviation is 2. Then the expected sum of the sample is 25×4 and standard error is $\sqrt{25} \times 2 = 10$. which means the average of 100 will be off by ± 10.

When drawing at random from a box:

1. Expected Value for average of draws = average of box.
2. SE for average of draws = $\frac{\text{SE for sum}}{\text{number of draws}}$

Sample average So far, the numbers in the box were known, and the problem was to say something about the average of the draws. This section reasons in the opposite and more practical direction. A random sample is taken from a box of unknown composition, and the problem is to estimate the average of the box. Naturally, the average of the draws is used as the estimate. And the SE

for the sample average can be used with the normal curve to gauge the accuracy of the estimate which is similar to the same techniques used for percentages in the previous section.

Let us see this with an example, suppose we want to estimate the average income of the household of 25,000 families. For that, an agency takes 1000 samples. It turns out the 62400 is an average of the sample which we know would be off by some number. Now we calculate SE for averages. We know the SE for percentages is $\sqrt{\text{number of drws}} \times SD$, this formula is similar for the averages as well. We calculate SD of the income from the sample, which comes out to be 53,000. Therefore the SE=$\sqrt{1000} \times 53,000 \approx 1700000$. Then the average SE for 1000 samples is 1700. Therefore we say that, the average income will be \$62,400 $\pm$1700. Now, if we discuss the confidence intervals, then the above number is 1SE, so we are 68% confident about it. Further, \$62,400 $\pm$ 2$\times$1700, average income lies in the 95% confidence interval. The confidence intervals always make statements about the samples. In about 95% of all samples, if you go 2 SEs either way from the sample average, your confidence interval will cover the average for the whole town; in the other 5%, your interval will miss. The word *confidence* is to remind you that the chances are in the sampling procedure; the average of the box is not moving around.

4.11 Model for measurement error

So far we have seen a box model, where samples can be taken from the large pool and then estimation is done over them. This kind of scenario is always not applicable. Suppose, we want to measure the error incurred by the weighing pan. Then in this case, the analogous to box model approach will be take repeated measurements with some known weights and to check the weight shown by the machine. For this kind of problems, different models have been proposed.

Gauss model The basic situation is that a series of repeated measurements are made on some quantity. According to the model, each measurement differs from the exact value by a chance error; this error is like a draw made at random from a box of tickets: the error box. Successive measurements are done independently and under the same conditions, so the draws from the error box are made with replacement. To capture the idea that the chance errors are not systematically positive or systematically negative, it is assumed that the average of the numbers in the error box equals 0.

In the Gauss model, each time a measurement is made, a ticket is drawn at random with replacement from the error box. The number on the ticket is the chance error. It is added to the exact value to give the actual measurement. The average of the error box is equal to 0. When the Gauss model applies, the SD of a series of repeated measurements can be used to estimate the SD of the error box. The estimate is good when there are enough measurements.

5 Statistical Inference

Statistical inference, or "learning" as it is called in computer science, is the process of using data to infer the distribution that generated the data. A typical statistical inference question is: "*Given a sequence of observations* $X_1, X_2, \cdots, X_n \approx F$*, estimate probability distribution* F." In some cases, we may want to infer only some feature of F such as its mean.

In other words, statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population. Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.

In machine learning, the term inference is sometimes used instead to mean "make a prediction, by evaluating an already trained model"; in this context inferring properties of the model is referred to as training or learning (rather than inference), and using a model for prediction is referred to as inference (instead of prediction).

Any statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of the observed data and similar data. Descriptions of statistical models usually emphasize the role of population quantities of interest, about which we wish to draw inference. Descriptive statistics are typically used as a preliminary step before more formal inferences are drawn.

Statisticians distinguish between three levels of modelling assumptions:

Parametric models The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters. A parametric model is a set $\mathfrak{F}$ that can be parametrized by a finite number of parameters. For example, if we assume that the data come from a Normal distribution, then the model is

$$\mathfrak{F} = \left\{ f(x, \mu, \sigma) : \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \right\} \tag{18}$$

This is a two-parameter model. We have written the density as $f(x, \mu, \sigma)$ to show that x is a value of the random variable whereas μ and σ are parameters. In general, parametric models take the following form

$$\mathfrak{F} = \{ f(x, \theta) : \theta \in \Theta \} \tag{19}$$

where parameters θ are selected from the parameter space Θ.

Non-parametric models A non-parametric model is a set F that cannot be parametrized by a finite number of parameters. The assumptions made

about the process generating the data are much less than in parametric statistics and may be minimal.[8] For example, every continuous probability distribution has a median, which may be estimated using the sample median or the Hodges–Lehmann–Sen estimator, which has good properties when the data arise from simple random sampling.

Semi-parametric : This term typically implies assumptions 'in between' fully and non-parametric approaches. For example, one may assume that a population distribution has a finite mean. Furthermore, one may assume that the mean response level in the population depends in a truly linear manner on some covariate (a parametric assumption) but not make any parametric assumption describing the variance around that mean (i.e. about the presence or possible form of any heteroscedasticity). More generally, semi-parametric models can often be separated into 'structural' and 'random variation' components. One component is treated parametrically and the other non-parametrically. The well-known Cox model is a set of semi-parametric assumptions.

There are many approaches to statistical inference. The two dominant approaches are called *frequentist inference* and *Bayesian inference*. We will cover both but we will start with frequentist inference.

5.1 Test of significance

The whole idea behind employing different tests of significance is to understand if the result we have obtained is by fluke or it is the real phenomena. Statistical significance is how statisticians measure whether an experiment (or even a study of existing data) yields a result more extreme than what chance might produce. If the result is beyond the realm of chance variation, it is said to be statistically significant.

Null and alternative hypothesis The null hypothesis corresponds to the idea that an observed difference is due to chance. To make a test of significance, the null hypothesis has to be set up as a box model for the data. The alternative hypothesis is another statement about the box, corresponding to the idea that the observed difference is real. A null hypothesis is a logical construct embodying the notion that nothing special has happened, and any effect you observe is due to random chance.

Significance tests
There are many tests for significance and they are applied based on the type of requirement. The overall steps for them are similar:

- set up the null hypothesis, in terms of a box model for the data;
- pick a test statistic, to measure the difference between the data and what is expected on the null hypothesis;
- compute the observed significance level P. The p-value indicates the probability of seeing the test statistics which we observed or more extreme value deviation from the expected test statistic given that the Null hypothesis is true. If this value is very less then it becomes easy to reject the Null hypothesis. In general a threshold is set and p-value is compared to that. Most of the researchers consider 5% as a good threshold, if the p-value is less than 5% then the result is considered to be statistically significant. If it is less than 1% then it is considered to be highly significant and in both cases the null hypothesis is rejected. Given a chance model that embodies the null hypothesis, the p-value is the probability of obtaining results as unusual or extreme as the observed results.

I have seen p-value mentioned in the results to show that the results obtained with them is statistically significant. Generally, McNeymar significance test is done on the obtained results. The p-value in these cases are really low, which is $p < 0.00001$, which shows the results are significant. In those cases, I guess, their null-hypothesis is that the result is obtained by fluke. But, our p-value is very small which shows there is so less chance of getting such result by fluke. Hence, the null hypothesis is rejected and the observed result is considered to be by actual improvement in the results. Small values of P are evidence against the null hypothesis: they indicate something besides chance was operating to make the difference.

5.1.1 Test Statistics

A test statistic is used to measure the difference between the data and what is expected on the null hypothesis.

$$\mathcal{Z} = \frac{observed - expected}{SE} \tag{20}$$

Tests using the $\mathcal{Z}$-statistic are called z-tests. $\mathcal{Z}$ says how many SEs away an observed value is from its expected value, where the expected value is calculated using the null hypothesis. As the number for $\mathcal{Z}$ increases it becomes less probable that the null hypothesis is true. For example, if the z-statistics is 3 which means there is 1 in 1000 chance that the observed value is expected value. This chance of 1 in 1,000 is called an *observed significance level*. The observed significance level is often denoted P, for probability, and referred to as a P-value.

The observed significance level is the chance of getting a test statistic as extreme as, or more extreme than, the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller this chance is, the stronger the evidence against the null. The P-value of a test is the chance of getting a big test statistic–assuming the null hypothesis to be right. P is not the chance of the null hypothesis being right.

5.1.2 Student t-test

In calculations of test-statistics $\mathcal{Z}$ in the previous section, we assumed that the drawn samples are sufficiently large. So the assumption that, the observed statistics, i.e. those values derived from samples are similar to populations holds true. But, there are some cases where observed samples are so small (if less than 30), this assumption breaks. For this extra steps are proposed in *student's t-test* which are as follows:

- When the number of measurements is small, the SD of the population should not be estimated by the SD of the measurements. Instead, SD^+ is used:

$$SD^+ = \sqrt{\frac{\text{number of measurements}}{\text{number of measurements} - 1}} \times SD \tag{21}$$

  Then based on this SD^+ SE is calculated in usual fashion.

$$t = \frac{observed - expected}{SE} \tag{22}$$

- The next step is to find the P-value. With a large number of measurements, this can be done using the normal curve. But with a small number of measurements, a different curve must be used, called Student's curve. As it turns out, the P-value from Student's curve is about 5%. That is quite a bit more than the 1% from the normal curve.

  Using Student's curve takes some work. Actually, there is one of these curves for each number of degrees of freedom. In the present context,

$$\text{degrees of freedom} = \text{number of measurements} - 1 \tag{23}$$

  From this a particular curve of degree is considered and from that with *t-statistics* p-value is calculated. It is obtained by considering area under the curve of the graph from the right of *t-statistics*.

5.1.3 Two-sample z-tests

Suppose we want to compare two samples to understand the difference between their averages. Here the assumption is that the samples are independent and chosen randomly. Then the z-test can be applied with little change.

The expected value for the difference of two quantities equals the difference of the expected values. Then, the standard error for the difference of two independent quantities is $a^2 + b^2$, where a is the SE for the first quantity; b is the SE for the second quantity. For dependent quantities, this formula is usually wrong.

Then the whole procedure for two-sample z-test is given as follows. Suppose that two independent and reasonably large simple random samples are taken from two separate boxes. The null hypothesis is about the difference between the averages of the two boxes. The appropriate test statistic is

$$\mathcal{Z} = \frac{\text{observed difference} - \text{expected difference}}{\text{SE for difference}}$$

In the formula, the *difference* is between the averages of the two samples. (If the null hypothesis says that the two boxes have the same average, the expected difference between the sample averages is 0.)

The two-sample z-test can also be used to compare treatment and control averages or rates in an experiment. Suppose there is a box of tickets. Each ticket has two numbers: one shows what the response would be to treatment A; the other, to treatment B. For each ticket, only one of the two numbers can be observed. Some tickets are drawn at random without replacement from the box, and the responses to treatment A are observed. Then, a second sample is drawn at random without replacement from the remaining tickets. In the second sample, the responses to treatment B are observed. The SE for the difference between the two sample averages can be conservatively estimated as follows:

1. compute the SEs for the averages as if drawing with replacement;
2. combine the SEs as if the two samples were independent.

5.1.4 χ^2-test

So far, the tests worked on the samples which had binary labels– either 1 or 0. However, there are some scenarios where there are more than two types of outputs. For these samples, χ^2-test is devised.

The χ^2-statistic can be used to test the hypothesis that data were generated according to a particular chance model.

$$\chi^2 = \sum_{\text{number of types}} \frac{\text{observed frequency} - \text{expected frequency}}{\text{expected frequency}} \tag{24}$$

When the model is fully specified (no parameters to estimate from the data), degrees of freedom = number of terms-1. The observed significance level P can be approximated as the area under the χ^2-curve to the right of the observed value for χ^2. The significance level gives the chance of the model producing observed frequencies as far from the expected frequencies as those at hand, or even further, distance being measured by χ^2.

5.2 Estimation

Common modeling problem involves how to estimate a joint probability distribution for a dataset. For example, given a sample of observation (X) from a domain $(x_1, x_2, x_3, \cdots, x_n)$, where each observation is drawn independently from the domain with the same probability distribution (so-called independent and identically distributed, i.i.d., or close to it).

Density estimation involves selecting a probability distribution function and the parameters of that distribution that best explains the joint probability distribution of the observed data (X). Often estimating the density is too challenging; instead, we are happy with a point estimate from the target distribution, such as the mean. There are many techniques for solving this problem, although two common approaches are:

- Maximum a Posteriori (MAP), a Bayesian method.
- Maximum Likelihood Estimation (MLE), a frequentist method.

Both approaches frame the problem as optimization and involve searching for a distribution and set of parameters for the distribution that best describes the observed data.

In Maximum Likelihood Estimation, we wish to maximize the probability of observing the data from the joint probability distribution given a specific probability distribution and its parameters, stated formally as: $P(X; \theta)$ or $P(x_1, x_2, x_3, \cdots, x_n; \theta)$. This resulting conditional probability is referred to as the likelihood of observing the data given the model parameters.

The objective of Maximum Likelihood Estimation is to find the set of parameters (theta) that maximize the likelihood function, e.g. result in the largest likelihood value.

$$\text{maximize } P(x_1, x_2, x_3, \cdots, x_n; \theta) \tag{25}$$

An alternative and closely related approach is to consider the optimization problem from the perspective of Bayesian probability.

5.3 Likelihood

The likelihood is the extent of support provided by the samples for any parameter value. It is nothing but a probability distribution over the parameters and observed samples. If there is an higher support for certain values, which means that there is a higher probability of that parameter being true given the samples.

$$\mathcal{L}(\theta_0; y) = P(Y = y | \theta = \theta_0) \tag{26}$$

From this definition, we can understand the meaning behind the likelihood. It states, given a *population* parameter θ_0, the probability of observing the

sample statistic y. This is the likelihood being measured here, and not the other way round.

In practice, the exact value of likelihood is meaningless, however, comparing it with other parameters gives useful information. So as to compare two likelihoods, likelihood ration is considered:

$$\frac{\mathcal{L}(\theta_0; y)}{\mathcal{L}(\theta_1; y)} \tag{27}$$

Likelihood function We saw a likelihood of getting sample statistic for a particular value of population parameters, now we extend it to all values of parameters. The likelihood function gives this support value for the observed value for all population parameters which is given as:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; y) = f_Y(y; \theta) \tag{28}$$

Technically, in the sample size of n, it takes the form:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_i(y_i; \theta) \tag{29}$$

From this, loglikelihood function is given as:

$$l(\theta) = \sum_{i=1}^{n} \log f_i(y_i; \theta) \tag{30}$$

5.3.1 Sufficient Statistic

A statistical value which can be calculated purely from the samples without depending on the value of parameter (θ) is called sufficient statistic. And a statistic is sufficient if it conveys all the information about the population parameter.

Let us see this with an example: A sample of 5 women have systolic blood pressure in pregnancy (SBP), which is normally distributed with standard deviation $\sigma = 10$, $\mathcal{S} = \{135, 123, 102, 110, 105\}$.

The likelihood is given as $\mathcal{L}(\theta) = \prod_{i=1}^{5} f_i(y_i; \theta)$. We know that, SBP is normally distributed with s.d. $\sigma = 10$. Thus we write likelihood as,

$$
\begin{aligned}
\mathcal{L}(\theta) &= \prod_{i=1}^{5} f_i(y_i;\theta) \\
&= \prod_{i=1}^{5} \frac{1}{10\sqrt{2\pi}} \exp^{-\frac{(y_i-\mu)^2}{200}} \\
&= (\frac{1}{10\sqrt{2\pi}})^5 \exp^{-\sum_{i=1}^{5} \frac{(y_i-\mu)^2}{200}} \\
&= (\frac{1}{10\sqrt{2\pi}})^5 \exp^{-\sum_{i=1}^{5} \frac{y_i^2}{200}} \exp^{-\sum_{i=1}^{5} \frac{\mu^2}{200}} \exp^{\mu \frac{\sum_{i=1}^{5} y_i}{100}}
\end{aligned}
$$

Now, the term with summation over y_i is called sufficient statistics, as if we have information over them we do not need actual sample information.

The example is good to understand the concept however, we still need to find out summation over square terms, for that we need sample information. So in true sense we can not get rid of the samples, if we have to, then we have to consider square of summation as well as a sufficient statistic.

This is indeed true, when we deal with two parameters, we need to have two values as sufficient statistics, in fact sufficient statistic $T(y)$ is a vector:

$$
T(y) = (T_1(y), T_2(y)) = (\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i^2) \tag{31}
$$

Nuisance Parameter In general, we can be interested in a single parameter instead of multiple parameters. For example, in the previous case with a normal distribution, we were interested in a mean value rather than a standard deviation, in that case the standard deviation is considered as a nuisance parameter. In this cases, we still need to their values to get the likelihood function, so instead of dealing with these parameter values, these values are replaced with sample statistic of that parameter. So in above case, we calculate the standard deviation from the samples and put them in the equation to get the likelihood function. Now, the likelihood function is called profile likelihood rather than the actual likelihood function.

5.3.2 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) provides the parameter values for which the observed samples becomes the most probable sample among all the samples. Formally, $\hat{\theta}$ is called maximum likelihood estimation of parameters if

$$
\mathcal{L}(\hat{\theta}) > \mathcal{L}(\theta_0) \qquad \forall \theta_0 \tag{32}
$$

Score and Information To get the MLE, we take the log-likelihood function, differentiate it w.r.t θ and put this value equal to 0. **(when we put derivative equal to 0, we either get minimum or maximum, but in this cases we are confident that we will get maximum. Does it mean that the function is concave so that whenever we get 0 the value is maximum? Does logarithm produce concave functions?)** This is considered as score.

$$\frac{d}{d\theta}l(\theta) = 0$$

Information is calculated as negative of the second derivative of the log-likelihood.

$$I_O(\theta) = -\frac{d^2}{d\theta^2}l(\theta)$$

Note that, the second derivative is always negative so the information will be always positive. This also gives the confidence of MLE, and called as observed information, if the observed information is higher, which means that the MLE is more confident.

There is an another concept called expected information, which is calculated

$$I(\theta) = E(I_O(\theta; Y))$$

which gives the variance of the estimate

$$V(\hat{\Theta}) = I^{-1}(\theta)$$

If the information is higher, which means MLE is more confident which means the variance is lower.

6 Baysian statistics

In a frequentist approach of statistical inference, the unknown parameter is considered to be fixed. In other words, suppose we are estimating mean of a population, in frequentist approach we consider it to be true and try to estimate it from the sample. Then we can apply any of estimation methods such as maximum likelihood estimation. Also, we can use confidence intervals to predict that the true value of the population mean will always lie in some interval if we take multiple samples from the population. In other words, 95% confidence interval says that repeated sampling of the population will produce mean value as $\bar{\mu}$.

Bayesian statistics considers a parameter to be estimated as a random variable which means instead of considering a fixed value it says that the parameter is distributed with certain probability. In the process of that, it also considers the apriori distribution of the parameter. So in general, from the background knowledge we have some intuition about the distribution of a parameter, this can be encoded in bayesian way of estimation. The confidenc interval equivalent in bayesian statistics, i.e. credible intervals also have better interpretation which says that there is a 95% chance that the value $\bar{\theta}$ will lie in the interval.

Bayes' theorem over data and parameter θ is given as

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})} \tag{33}$$

In this equation, $P(\theta)$ is called a prior probability of the parameter which are our prior intuitions about the parameters, and the probability $P(\theta|\text{data})$ is called posterior, as we are changing the estimate after seeing the data. The probability of $P(\text{data}|\theta)$ is called likelihood of seeing the data if we consider the parameters to be true. Finally, the $P(\text{data})$ is called normalizing factor, as we want the numerator to be probability distribution.

Because, the denominatore term is just normalizing factor and it does not depend on the value of θ, it is ignored in most of the calculations, to produce following proprtionality:

$$P(\theta|\text{data}) \propto P(\text{data}|\theta)P(\theta) \tag{34}$$

6.1 Bayesian mean inference for normal distribution

Consider that θ be the inferred mean of normal distribution with known variance σ. Then the likelihood of seeing the data with θ is given as:

$$p(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(y-\theta)^2}{2\sigma^2}} \tag{35}$$

Now, we want to select a prior distribution for θ, we can choose any distribution we want, but have to consider the intuition behind the choice of distribution. Suppose $\theta \sim N(\mu_0, \tau_0^2)$, then the prior distribution is:

$$P(\theta) \propto \exp^{-\frac{(\theta-\mu_0)^2}{2\tau_0^2}}$$

Therefore, the posterior will be

$$p(\theta|y) \propto exp\{-0.5\Big[\frac{(y-\theta)^2}{2\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\Big]\} \tag{36}$$

$$\propto exp\{-\frac{(\theta-\mu_1)^2}{2\tau_1^2}\} \tag{37}$$

where $\mu_1 = \frac{(\mu_0/\tau_0^2)+(y/\sigma^2)}{(1/\tau_0^2)+(1/\sigma^2)}$ and $\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$.

6.2 Conjugate Priors

Consider the bayes' equation again,

$$P(\theta|\text{data}) \propto P(\text{data}|\theta)P(\theta) \tag{38}$$

We always choose a managable priori distribution so that we can get nice expression for our posterior distribution. When the parameter's parametric form is retained in the posterior then it is called a conjugate prior. Following are some distribution pairs, if they are used as a likelihood and prior distribution produce managable posterior distributions.

Likelihood	Conjugate prior
Normal	Normal
Binomial or Bernoulli	Beta
Exponential or Poisson	Gamma
Uniform	Pareto
Multinomial	Drichlet

A Discussion

A.1 Statistical Models

A statistical model is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population). A statistical model represents, often in considerably idealized form, the data-generating process. A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variables. All statistical hypothesis tests and all statistical estimators are derived via statistical models. More generally, statistical models are part of the foundation of statistical inference.

In mathematical terms, a statistical model is usually thought of as a pair $(\mathcal{S}, \mathcal{P})$, where $\mathcal{S}$ is the set of possible observations, i.e. the sample space, and $\mathcal{P}$ is a set of probability distributions on $\mathcal{S}$. The intuition behind this definition is as follows. It is assumed that there is a "true" probability distribution induced by the process that generates the observed data. We choose $\mathcal{P}$ to represent a set (of distributions) which contains a distribution that adequately approximates the true distribution.

Note that we do not require that $\mathcal{P}$ contains the true distribution, and in practice that is rarely the case. Indeed, as Burnham & Anderson state, "A model is a simplification or approximation of reality and hence will not reflect all of reality"–hence the saying "all models are wrong".

The set $\mathcal{P}$ is almost always parameterized: $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. The set Θ defines the parameters of the model. A parameterization is generally required to have distinct parameter values give rise to distinct distributions, i.e. $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$ must hold (in other words, it must be injective). A parametrization that meets the requirement is said to be identifiable. Note that these parameters of statistical models are different than the notions of parameters we have in ML. We explicitly distinguish them in the next section.

Example **Suppose that we have a population of children, with the ages of the children distributed uniformly, in the population. The height of a child will be stochastically related to the age: e.g. when we know that a child is of age 7, this influences the chance of the child being 1.5 meters tall. We could formalize that relationship in a linear regression model, like this: $height_i = b_0 + b_1 \times age_i + \epsilon_i$, where b_0 is the intercept, b_1 is a parameter that age is multiplied by to obtain**

a prediction of height, ϵ_i is the error term, and i identifies the child. This implies that height is predicted by age, with some error.

To do statistical inference, we would first need to assume some probability distributions for the ϵ_i. For instance, we might assume that the ϵ_i distributions are i.i.d. Gaussian, with zero mean. In this instance, the model would have 3 parameters: b_0, b_1, and the variance of the Gaussian distribution.

We can formally specify the model in the form $(\mathcal{S}, \mathcal{P})$ as follows. The sample space, $\mathcal{S}$, of our model comprises the set of all possible pairs (age, height). Each possible value of $\theta = (b_0, b_1, \sigma^2)$ determines a distribution on $\mathcal{S}$; denote that distribution by P_θ. If Θ is the set of all possible values of θ, then $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. (The parametrization is identifiable, and this is easy to check.)

Historically, a primary use of regression was to illuminate a supposed linear relationship between predictor variables and an outcome variable. The goal has been to understand a relationship and explain it using the data that the regression was fit to. In this case, the primary focus is on the estimated slope of the regression equation, b . Economists want to know the relationship between consumer spending and GDP growth. Public health officials might want to understand whether a public information campaign is effective in promoting safe sex practices. In such cases, the focus is not on predicting individual cases but rather on understanding the overall relationship among variables. With the advent of big data, regression is widely used to form a model to predict individual outcomes for new data (i.e., a predictive model) rather than explain data in hand. In this instance, the main items of interest are the fitted values Y. In marketing, regression can be used to predict the change in revenue in response to the size of an ad campaign. Universities use regression to predict students' GPA based on their SAT scores. A regression model that fits the data well is set up such that changes in X lead to changes in Y. However, by itself, the regression equation does not prove the direction of causation. Conclusions about causation must come from a broader understanding about the relationship. For example, a regression equation might show a definite relationship between number of clicks on a web ad and number of conversions. It is our knowledge of the marketing process, not the regression equation, that leads us to the conclusion that clicks on the ad lead to sales, and not vice versa.

A.2 Statistical Parameters vs Machine learning parameters vs hyperparameters

A.2.1 Statistical Parameter

In statistics, as opposed to its general use in mathematics, a parameter is any measured quantity of a statistical population that summarises or describes an aspect of the population, such as a mean or a standard deviation. If a population exactly follows a known and defined distribution, for example the normal distribution, then a small set of parameters can be measured which completely describes the population, and can be considered to define a probability distribution for the purposes of extracting samples from this population. A parameter is to a population as a statistic is to a sample; that is to say, a parameter describes the true value calculated from the full population, whereas a statistic is an estimated measurement of the parameter based on a sub-sample. Thus a "statistical parameter" can be more specifically referred to as a population parameter. Inference is done for parameters from the given samples.

Parameterised Distributions **Suppose that we have an indexed family of distributions. If the index is also a parameter of the members of the family, then the family is a parametrized family. Among parametrized families of distributions are the normal distributions, the Poisson distributions, the binomial distributions, and the exponential family of distributions. For example, the family of normal distributions has two parameters, the mean and the variance: if those are specified, the distribution is known exactly. The family of chi-squared distributions can be indexed by the number of degrees of freedom: the number of degrees of freedom is a parameter for the distributions, and so the family is thereby parameterized.**

Measurement of Parameters **In statistical inference, parameters are sometimes taken to be unobservable, and in this case the statistician's task is to estimate or infer what they can about the parameter based on a random sample of observations taken from the full population. Estimators of a set of parameters of a specific distribution are often measured for a population, under the assumption that the population is (at least approximately) distributed according to that specific probability distribution. In other situations, parameters may be fixed by the nature of the sampling procedure used or the kind of statistical procedure being carried out (for example, the number of degrees of freedom in a Pearson's chi-squared test). Even if a family of distributions is not specified, quantities such as the mean and variance can generally still be regarded as statistical parameters of the population, and statistical procedures can still attempt to make inferences about such population parameters.**

A.2.2 ML parameters and hyperparameters

Crudely, we can say, machine learning model parameters are those which would be learned by the machine like Weights and Biases whereas hyper-parameters are those which we supply to the model, for example: number of hidden nodes and layers, input features, learning rate, activation function, etc.

We can further elaborate on that, in machine learning, a model M with parameters and hyper-parameters looks like,

$$Y \approx M_H(\Phi|D)$$

where Φ are parameters and H are hyper-parameters, D is training data and Y is output data (class labels in case of classification task). The objective during training is to find estimate of *parameters* $\hat{\Phi}$ that optimizes some loss function $\mathcal{L}$ we have specified. Since, model M and loss-function $\mathcal{L}$ are based on H, then the consequent parameters $\hat{\Phi}$ are also dependent on *hyper-parameters* H.

The hyper-parameters H are not 'learnt' during training, but does not mean their values are immutable. Typically, the hyper-parameters are fixed and we think simply of the model M, instead of M_H. Herein, the hyper-parameters can also be considers as a-priori parameters. The source of confusion stems from the use of M_H and modification of hyper-parameters H during training routine in addition to, obviously, the parameters $\hat{\Phi}$. There are potentially several motivations to modify H during training. An example would be to change the learning-rate during training to improve speed and/or stability of the optimization routine. The important point of distinction is that, the result, say label prediction, Y_{pred} is based on model parameters Φ and not the hyper-parameters H.

Statistical inference involves making educated guesses about the population given a sample which is drawn from the population. There are two important sections in inference: first is estimation of parameters, and the second one is hypothesis testing. In the first part, parameters are estimated given a sample and assuming certain statistical model. A statistical model is a probability distribution function with parameters. In point estimation we estimate these values with different methods of estimation: methods of moments, maximum likelihood, and maximum a posterior. In hypothesis testing, two hypothesis are considered: null and alternate, and evidence are obtained for alternate hypothesis. If there is a less chance of observing such an sample given the null hypothesis then null hypothesis is rejected. Significance tests are subset of the hypothesis testing where, there is no alternate hypothesis. The null hypothesis says that the observed result is due to chance and there is no effect of new

method or treatment. When there is a less chance of observing such a sample ($p < \alpha$, where α is significance level) then we reject the null hypothesis, which roughly means the observation is because of the improvement in the method.