Importance of Statistical Tools and Methods in Data Science

Krish Bajaj
DPS Faridabad & Intern with SaveLife Foundation

Abstract

This paper aims to highlight the prominent position of statistics as a foundational pillar for descriptive and inferential statistical analysis to deduce underlying patterns in a population by looking at a sample drawn from the population. It focusses on the intuitive aspects of the statistical tools and its relevance and applicability .The paper concludes by highlighting some common misconceptions and misuse of statistics.

“In God we trust,all other must bring data.” W. Edwards Deming

MENTORED BY | BHARAT SACHDEVA -Sr MATHEMATICS FACULTY Delhi Public School, Faridabad.
| Dr. ROSHAN JOSE-IIT Kharagpur & Head Data Science Division @ SaveLife foundation.

Introduction to statistics and Its relevance:

Scientists seek to answer questions using rigorous methods and careful observations.

These observations are often collected in the form of field notes, surveys, and experiments and are called as "data". This data forms the backbone of a statistical investigation. Statistics is the study of how best to collect, analyze, and draw conclusions from data.

In general, any statistical investigation involves the following steps:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Researchers from different fields nowadays depend heavily on statistical data collection, analysis and inference making tools to aid them in decision making or validating if a given belief or "hypothesis" is true or false .This paper aims to provide an intuitive feel [3] of the concepts behind these tools and relate them to some real-life examples to drive home the relevance and significance of using them

Importance of quality of data sets and sampling:

Any statistical study always begins with the collection of data .Data is often collected in a matrix form like an "excel sheet" where the rows represent the "observational sets" and the columns represent "variables" or characteristics of interest for which the data is collected. Data is collected for a "sample" which is representative of the entire "population". The statistical parameter like mean, median, standard deviation etc. that is calculated for the sample is called "sample statistic" and the same for the population is called "population parameter". Researchers always collect samples and from the sample statistics they make inferences for the underlying population from which the sample is drawn. There are two methods of data collection: through an observational study in which researchers collect data about things that occur or arise naturally or through randomized experiment in which case the researchers collect a sample of individuals that are assigned to groups The individuals in each group are assigned a treatment. When individuals are randomly assigned to a group, the experiment is called a randomized experiment.

Almost all statistical methods for observational data rely on a sample being random and unbiased. When a sample is collected in a biased way, these statistical methods will not generally produce reliable information about the population. The idea of a simple random sample and systematic random sampling is shown graphically below.

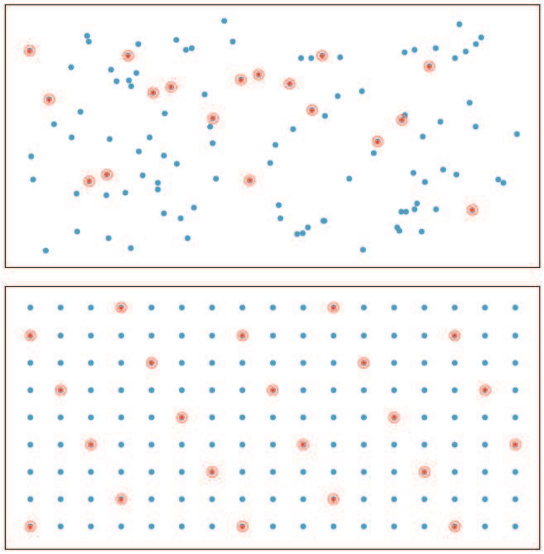


Fig 1: Representation of simple random and systematic random sampling [15]

The fig 2 shows graphically some other random sampling methods that are used like stratified, cluster, and multistage sampling. In a simple random sample, every individual as well as every group of individuals has the same probability of being in the sample. A systematic random sample involves choosing from of a population using a random starting point, and then selecting members according to a fixed, periodic interval (such as every 10th member).

A stratified random sample involves randomly sampling from every stratum, where the strata should correspond to a variable thought to be associated with the variable of interest. A cluster random sample involves randomly selecting a set of clusters, or groups, and then collecting data on all individuals in the selected clusters. This can be useful when sampling clusters is more convenient and less expensive than sampling individuals, and it is an effective strategy when each cluster is approximately representative of the population.

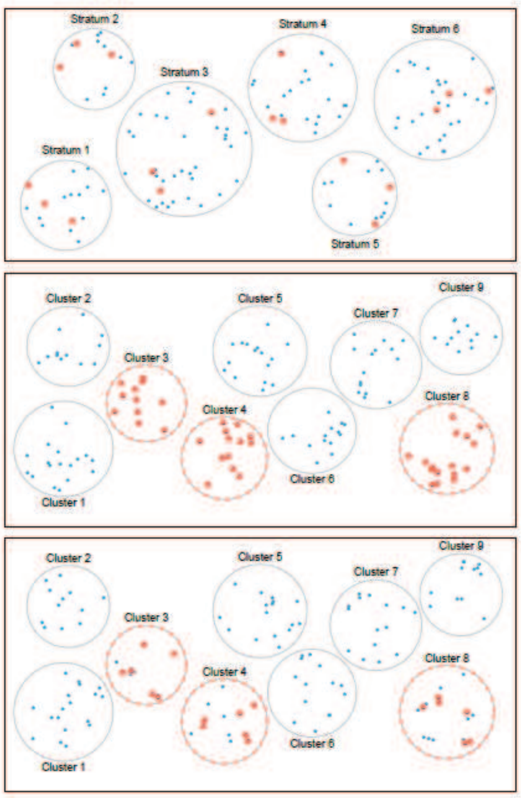


Fig 2: Representation of stratified random and cluster random sampling[15]

Using graphs and descriptive statistics to unfold the truth:

Once data has been collected via a chosen sampling method, the next step is to analyze the data set through graphical tools or descriptive statistics parameters [15,12] that help describe the data.

A scatterplot is a bivariate display illustrating the relationship between two numerical variables. The observations must be paired, which is to say that they correspond to the same case or individual. The linear association between two variables can be positive or negative, or there can be no association. Positive association means that larger values of the first variable are associated with larger values of the second variable. Negative association means that larger values of the first variable are associated with smaller values of the second variable. Additionally, the association can follow a linear trend or a curved (nonlinear) trend.

When looking at a univariate display, researchers want to understand the distribution of the variable. The term distribution refers to the values that a variable takes and the frequency of those values. When looking at a distribution we should note the presence of clusters, gaps, and outliers. Distributions may be symmetric, or they may have a long tail. If a distribution has a long left tail (with greater density over the higher numbers), it is left skewed. If a distribution has a long right tail (with greater density over the smaller numbers), it is right skewed. In addition, distributions may be unimodal, bimodal, or multimodal if they have one, two or more peaks

Two graphs that are useful for showing the distribution of a small number of observations are the stem-and-leaf plot and dot plot. These graphs are ideal for displaying data from small samples because they show the exact values of the observations and how frequently they occur. However, they are impractical for larger data sets. For larger data sets it is common to use a frequency histogram or a relative frequency histogram to display the distribution of a variable. This requires choosing bins of an appropriate width. To see cumulative amounts, use a cumulative frequency histogram. A cumulative relative frequency histogram is ideal for showing percentiles.

Descriptive statistics describes or summarizes data, by calculating and summarizing key parameters like measures of center, measures of spread and measures of shape. In general, for univariate distributions we consider two measures of center (mean and median) and three measures of spread: Standard deviation (SD), Interquartile range (IQR) and Range.

When summarizing or comparing distributions, researchers always comment on center, spread, and shape. Also, it is a practice to mention outliers or gaps if applicable

• Mean and median are measures of center. (A common mistake is to report mode as a measure of center. However, a mode can appear anywhere in a distribution. The mean is the sum of all the observations divided by the number of observations, n.

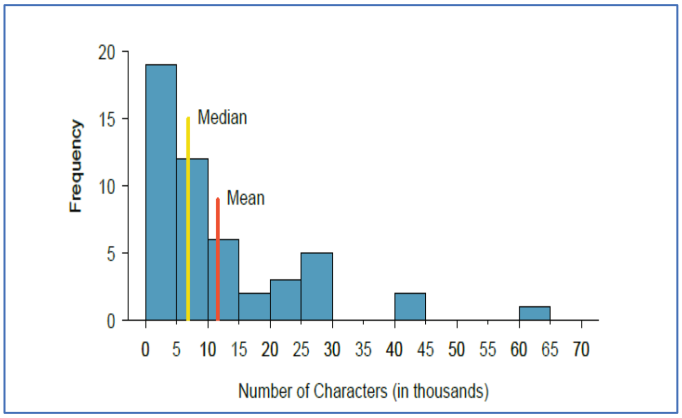


Fig 3: Representation of Mean Vs Median[15]

In an ordered data set, the median is the middle number when n is odd. When n is even, the median is the average of the two middle numbers. Because large values exert more "pull" on the mean, large values on the high end tend to increase the mean more than they increase the median. In a right skewed distribution, therefore, the mean is greater than the median. Analogously, in a left skewed distribution, the mean is less than the median.

•. For measures of spread SD (standard deviation) measures the typical spread from the mean, whereas IQR (Interquartile range) measures the spread of the middle 50% of the data. Range is also sometimes used as a measure of spread.

• Outliers are observations that are extreme relative to the rest of the data. Two rules of thumb
for identifying observations as outliers are:

> more than 2 standard deviations above or below the mean
> more than 1.5X IQR below Q1 or above Q3

• Mean and SD are sensitive to outliers. Median and IQR are more robust and less sensitive to outliers.

• The empirical rule states that for normal distributions, about 68% of the data will be within one standard deviation of the mean, about 95% will be within two standard deviations of the mean, and about 99.7% will be within three standard deviations of the mean.

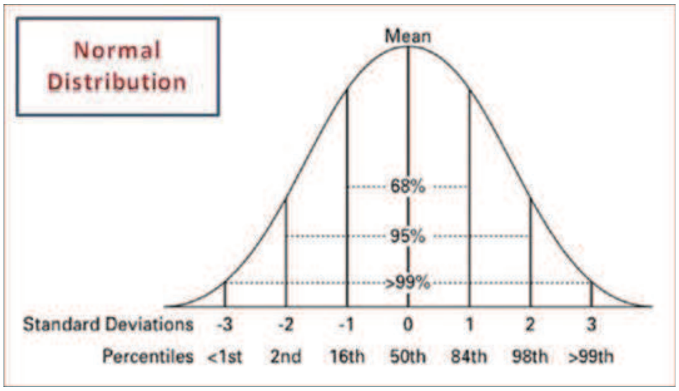


Fig. 4 Normally distributed data [7]

• When analyzing a distribution or comparing 2 distributions all three parameters of center, spread and shape must be looked at. For eg in the figure below there are three different distributions with the same mean (zero) and standard deviation(one)

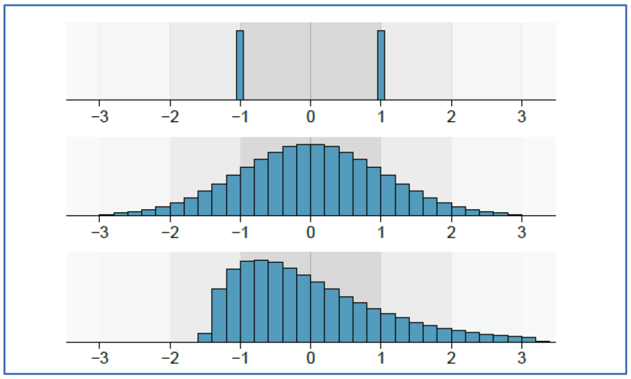


Fig. 5 Different distributions with same means and standard distributions

By looking at the graphs and descriptive statistics together for the sample data we can find out about the nature of distribution of the underlying population that the sample represents and thus from the sample data we can infer or make decisions for the entire population.

Importance of probability in decision making [2]:

A discrete probability distribution can be summarized in a table that consists of all possible outcomes of a random variable and the probabilities of those outcomes. The outcomes must be disjoint, and the sum of the probabilities must equal 1.

A probability distribution can be represented with a histogram and, like the distributions of data and can be summarized by its center, spread, and shape. When given a probability distribution table, we can calculate the mean (expected value) and standard deviation of a random variable using the following formulas.

$$\begin{aligned} E(X) = \mu_X &= \sum x_i \cdot P(x_i) \\ &= x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \cdots + x_n \cdot P(x_n) \\ Var(X) = \sigma_x^2 &= \sum (x_i - \mu_X)^2 \cdot P(x_i) \\ SD(X) = \sigma_X &= \sqrt{\sum (x_i - \mu_X)^2 \cdot P(x_i)} \\ &= \sqrt{(x_1 - \mu_X)^2 \cdot P(x_1) + (x_2 - \mu_X)^2 \cdot P(x_2) + \cdots + (x_n - \mu_X)^2 \cdot P(x_n)} \end{aligned}$$

We can think of P(xi) as the weight, and each term is weighted its appropriate amount. The mean of a probability distribution does not need to be a value in the distribution. It represents the average of many, many repetitions of a random process. The standard deviation represents the typical variation of the outcomes from the mean, when the random process is repeated over and over.

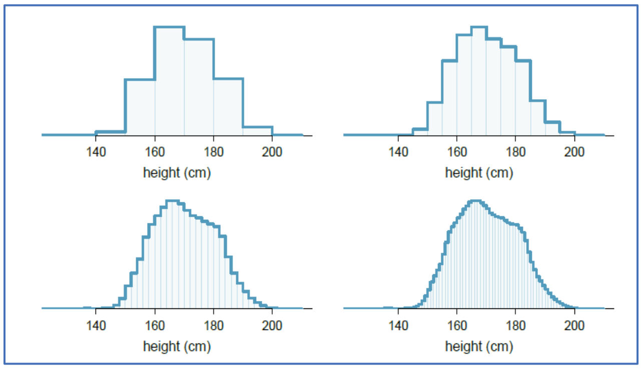


Fig. 6: Discrete to continuous distributions [7]

Histograms use bins with a specific width to display the distribution of a variable. When there is enough data and the data does not have gaps, as the bin width gets smaller and smaller, the histogram begins to resemble a smooth curve, or a continuous distribution.

Continuous distributions are often used to approximate relative frequencies and probabilities. In a continuous distribution, the area under the curve corresponds to relative frequency or probability. The total area under a continuous probability distribution must equal 1. Because the area under the curve for a single point is zero, the probability of any specific value is zero. This implies that, for example, P(X < 5) = P(X _ 5) for a continuous probability distribution.

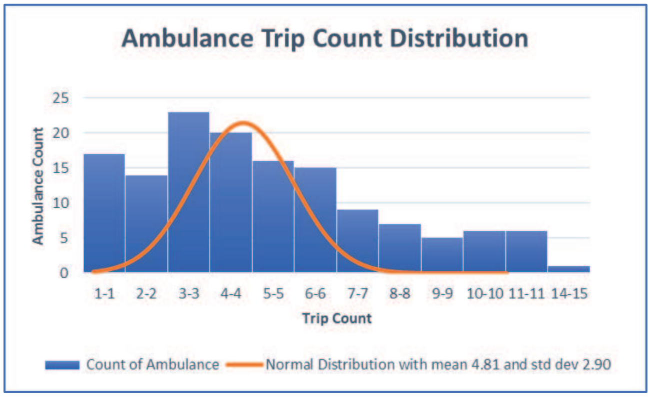


Fig. 7: Representation of Normal Distribution of Ambulance Trip Count, Courtesy SaveLife foundation[4]

The study of probability is useful for measuring uncertainty and assessing risk. In addition, probability serves as the foundation for inference, providing a framework for evaluating when an outcome falls outside of the range of what would be expected by chance alone.

Central Limit Theorem and its significance in drawing inferences:

If the population is normal, the sampling distribution of sample means will be normal for any sample size.The less normal the population, the larger n needs to be for the sampling distribution of the sampling means to be nearly normal. However, a good rule of thumb is that for almost all populations, the sampling distribution of sample means will be approximately normal if n >= 30.This is what forms the basis of central limit theorem [6]-which is at core of all statistical inference making. No matter what the nature of underlying population distribution is, the distribution of sample means will always approximate a normal distribution specially as sample size is greater than 30

The standard deviation of population of sample means describes the typical error or distance of the sample mean from the population mean. It also tells us how much the sample mean is likely to vary from one random sample to another. The standard deviation of sample means will be smaller than the standard deviation of the population by a factor of square root of n. The larger the sample, the better the estimate tends to be.

Three important facts about the sampling distribution of the sample average:

a. The mean of a sample means is is equal to mean of the population
b. The SD of a sample mean is given by standard deviation of population divided by square root of n
c. When the population is normal or when n >= 30, the sample mean closely follows a normal distribution. (shape)

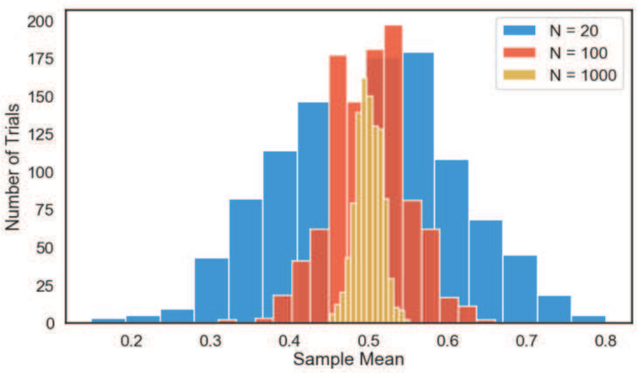
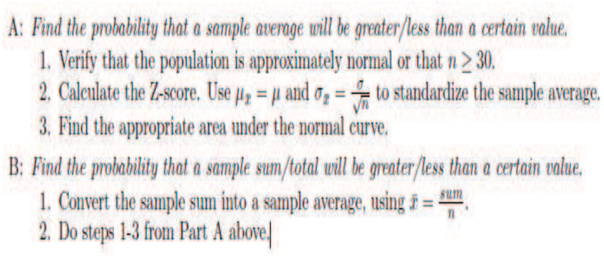


Fig. 8: Central Limit Theorem: Graphical representation of sampling means distribution [6]

These facts are used in the following scenarios



A: *Find the probability that a sample average will be greater/less than a certain value.*

1. Verify that the population is approximately normal or that $n \geq 30$.
2. Calculate the Z-score. Use $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to standardize the sample average.
3. Find the appropriate area under the normal curve.

B: *Find the probability that a sample sum/total will be greater/less than a certain value.*

1. Convert the sample sum into a sample average, using $\bar{x} = \frac{sum}{n}$.
2. Do steps 1-3 from Part A above.

When our sampling method produces estimates in an unbiased way, the sampling distribution will be centered on the true value and we call the method accurate. When the sampling method produces estimates that have low variability, the sampling distribution will have a low standard error, and we call the method precise.

But sample statistics calculated for a sample are point estimates that provides a single plausible value for a population parameter. However, a point estimate isn't perfect and will have some standard error associated with it. When estimating a parameter, it is better practice to provide a plausible range of values instead of supplying just the point estimate.

A plausible range of values for the population parameter is called a confidence interval. Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish. If we report a point estimate, we probably will not hit the exact population parameter. On the other hand, if we report a range of plausible values- a confidence interval -we have a good shot at capturing the parameter.

Interpreting confidence intervals and confidence levels:

68% and 95% are examples of confidence levels. The confidence level tells us the capture rate with repeated sampling. For example, a correct interpretation of a 95% confidence level is that if many samples of the same size were taken from the population, about 95% of the resulting confidence intervals would capture the true population parameter (assuming the conditions are met and the probability model is true). Note that this is a relative frequency interpretation.

We cannot use the language of probability to interpret an individual confidence interval once it has been calculated. The confidence level tells us what percent of the intervals will capture the population parameter, not the probability that a calculated interval captures the population parameter. Each calculated interval either does or does not capture the population parameter.

Confidence intervals are often reported as:

Confidence interval: point estimate (+-) margin of error.

The margin of error (ME) = critical value x SE of estimate

This tells us, with a confidence, how much we expect our point estimate to deviate from the true population value due to chance. The margin of error depends on the confidence level; the standard error does not. Other things being constant, a higher confidence level leads to a larger margin of error. For a fixed confidence level, increasing the sample size decreases the margin of error. This assumes a random sample. The margin of error formula only applies if a sample is random. Moreover, the margin of error measures only sampling error; it does not account for additional error introduced by response bias and non-response bias. Even with a perfectly random sample, the actual error in a poll is likely higher than the reported margin of error.

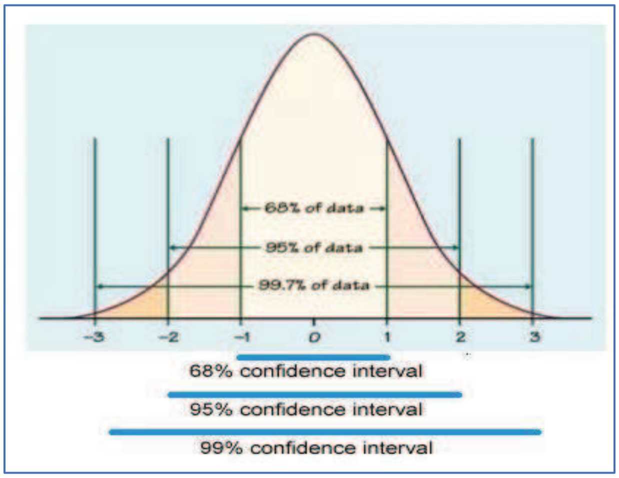


Fig 9: Representation of Confidence Intervals [5]

Hypothesis testing in inference making:

A hypothesis test [8] is a statistical technique used to evaluate competing claims based on data. The competing claims are called hypotheses and are often about population parameters (e.g. Population mean and p); they are never about sample statistics. The null hypothesis is abbreviated H0. It represents a skeptical perspective or a perspective of no difference or no change.

The alternative hypothesis is abbreviated HA. It represents a new perspective or a perspective of a real difference or change. Because the alternative hypothesis is the stronger claim, it bears the burden of proof.

The logic of a hypothesis testis that we begin by assuming that the null hypothesis is true. Then, we calculate how unlikely it would be to get a sample value as extreme as we actually got in our sample, assuming that the null value is correct. If this likelihood is too small, it casts doubt on the null hypothesis and provides evidence for the alternative hypothesis. We set a significance level, alpha, which represents the threshold below which we will reject the null hypothesis. The most common significance level is alpha = 0:05. If we require more evidence to reject the null hypothesis, we use a smaller alpha. After verifying that the relevant conditions are met, we can calculate the test statistic. The test statistic tells us how many standard errors the point estimate sample value is fromthe null value (i.e. the value hypothesized for the parameter in the null hypothesis). When investigating a single mean or proportion or a difference of means or proportions, the test statistic is calculated as: (point estimate -null value)/SE of estimate. After the test statistic, we calculate the p-value. We find and interpret the p-value according to the nature of the alternative hypothesis. The three possibilities are:

HA: parameter > null value: The p-value corresponds to the area in the upper tail.

HA: parameter < null value: The p-value corresponds to the area in the lower tail.

HA: parameter = null value: The p-value corresponds to the area in both tails.

The p-value is the probability of getting a test statistic as extreme or more extreme than the observed test statistic in the direction of HA if the null hypothesis is true and the probability model is accurate.

The conclusion or decision of a hypothesis test is based on whether the p-value is smaller or larger than the preset significance level alpha. When the p-value < alpha, we say the results are statistically significant at the alpha level and we have evidence of a real difference or change. The observed difference is beyond what would have been expected from chance variation alone. This leads us to reject H0 and gives us evidence for HA.When the p-value > alpha, we say the results are not statistically significant at the alpha level and we do not have evidence of a real difference or change. The observed difference was within the realm of expected chance variation. This leads us to not reject H0 and does not give us evidence for HA.

Decision errors in a hypothesis test refer to two types of decision errors that could be made. These are called Type I and Type II Errors. A Type I Error is rejecting H0, when H0 is true. We commit a Type I Error if we call a result significant when there is no real difference or

P(Type I error) =alpha.

A Type II Error is not rejecting H0 when HA is true. We commit a Type II Error if we call a result not significant when there is a real difference or

P(Type II error) = Beta.

The probability of a Type I Error (alpha) and a Type II Error (beta) are inversely related. Decreasing alpha makes beta larger; increasing beta makes alpha smaller. Once a decision is made, only one of the two types of errors is possible. If the test rejects H0, for example, only a Type I Error is possible.

Another concept in hypothesis testing is power of test [11]. When a HA is true, the probability of not making a Type II Error is called power.

Power = 1 -Beta.

The power of a test is the probability of detecting an effect of a size when it is present. Increasing the significance level decreases the probability of a Type II Error and increases power. For a fixed alpha, increasing the sample size n makes it easier to detect an effect and therefore decreases the probability of a Type II Error and increases power.

Correlation and regression analysis:

A bivariate display called a scatterplot, shows the relationship between two numerical variables. When we use x to predict y, we call x the explanatory variable or predictor variable, and we call y the response variable. A linear model for bivariate numerical data can be useful for prediction when the association between the variables follows a constant, linear trend. This is called Linear regression and is a powerful statistical tool. Linear models [10] should not be used if the trend between the variables is curved. When we write a linear model, we use ^y to indicate that it is the predicted value according to the model. The residual is the error between the true value and the modeled value, computed as y -^y. The order of the difference matters, and the sign of the residual will tell us if the model overpredicted or underpredicted a particular data point. The symbol s in a linear model is used to denote the standard deviation of the residuals, and it measures the typical prediction error by the model. A residual plot is a scatterplot with the residuals on the vertical axis. The residuals are often plotted against x on the horizontal axis, but they can also be plotted against y, ^y, or other variables. Two important uses of a residual plot are the following:

- Residual plots help us see patterns in the data that may not have been apparent in the scatterplot.
- The standard deviation of the residuals is easier to estimate from a residual plot than from the original scatterplot.
- In Linear regression, the line that minimizes the sum of the squared residuals and is represented as the solid line as shown in Figure 10. This is commonly called the least squares line.

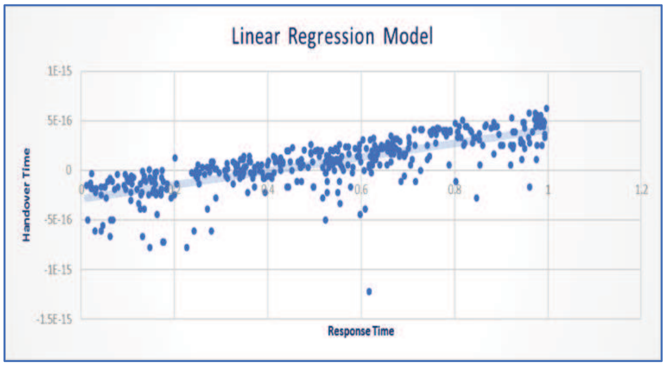


Fig 10: Representative Linear Regression Model (Handover Time Vs Response Times, courtesy SaveLife foundation) [4]

Correlation, denoted with the letter r, measures the strength and direction of a linear relationship. The value of r is always between -1 and 1, inclusive, with an r = -1 indicating a perfect negative relationship (points fall exactly along a line that has negative slope) and r = 1 indicating a perfect positive relationship (points fall exactly along a line that has positive slope). An r = 0 indicates no linear association between the variables, though there may well exist a quadratic or other type of association. Just like Z-scores, the correlation has no units. Changing the units in which x or y are measured does not affect the correlation. Correlation is sensitive to outliers. Adding or removing a single point can have a big effect on the correlation. As we learned previously, correlation is not causation. Even a very strong correlation cannot prove causation; only a well-designed, controlled, randomized experiment can prove causation.

Common misunderstanding's in relation with statistics

Most of us have heard the famous quote: "There are three kinds of lies: lies, damned lies and statistics" [16] which has been attributed to many different people since the 1800s. It has also been stated that 86% of statistics are made up! However, If as Serge Benhayon factually states: "numbers cannot lie for they are exact" then it follows that any lies must be coming from somewhere other than the numbers. Furthermore it has been proven that lies come not only from the people or authors of statistics but also from the misconceptions in ignoring and misunderstanding both true statistics when they are presented, and also our lived experiences which provide another form of research and data.

Misuse seen in Data Collection

> Falsifying data
> Choosing the wrong data-– choosing something that looks like a valid support for an argument, but really isn't
> "Cherry-picking" the data-– selecting a subset of the data collected such that the data supports a "statistically significant" result.
> Sometimes they even make their data conveniently unavailable to other researchers so that others are unable to verify the results.

Misuse seen in In Analysis of Data

> Reporting differences that are not "statistically significant" as if they were. If the results are not statistically significant, as far as statistics is concerned, the differences are not "real".
> Unjustifiably generalizing to a population based on a sample that's not representative of the population (for example, using a sample consisting of just American women to generalize to the total population of the U.S.).
> Implying links between factors when such a conclusion is not statistically justified by the research

data. If you read such keywords as "suggests"; "may" as in "may be linked to", "may help explain"; "some" as "in some people", "in some cases" be wary,

> Never mentioning the risks associated with "false positives"(getting a positive result when the result is really negative) or "false negatives" (getting a negative result when the result is really positive) when recommending a course of action based on the conclusions of the study done (medical research studies all too often do this).
> False positives and false negatives are always possible in any test procedure. So the standard procedure is to do additional testing if a test comes in positive. If those subsequent tests arrive at the same diagnosis consistently, all's well and good.
> However, if they don't and the positive result turns out to be a misdiagnosis ("false positive") inevitably you'll have been put through tests that involve additional time, money, and sometimes additional pain or complications. You really need to know about all of this up front before the first test is performed.

In the presentation/visualization of results

> Showing no zero point on graphs– making small changes seem bigger than they really are
> Without appropriate warning, by breaking the y-axis such that a zero point can be presented This is better than no zero point, but still makes small changes seem bigger than they are. The reader, if they don't register the break, can be deceived in the same fashion that they're deceived by no-zero-point graphs
> Omitting labels on one or more axes, leaving the reader to conclude the wrong thing
> Uneven/deceitful plot scaling
> Reporting results as percentages without indicating the actual numerical bases (i.e., "detached statistics")
> Quoting %'s of %'s — usually purposely used to confuse the reader and make results seem more "dramatic" than they really are.
> Selectively choosing to report statistics that support a desired result rather than those that wouldn't
> Tacking on adjectives and adverbs when describing numbers and differences between numbers (e.g., a meager 4%, a huge 4% difference) to influence your interpretation of the results, These are strategies used in marketing, politics, and propaganda. Numbers are just numbers; they don't have attribute.

Misconceptions about Statistics:

Statistics is just "Math's" and "Formulas"

Statistics uses numbers but numbers are not the primary focus of statistics, at least to most practitioners. Applied statistics is a form of inductive reasoning that uses math as one of its tools. It also uses sorting for ranks, filtering for classification, and all kinds of graphics. The point of using statistics is to discover new knowledge and solve problems using inductive reasoning involving numbers.

"Statistics requires a lot of data"

This is not true as the whole premise about statistical inference is using representative samples to predict population parameters What is more important than the number of data points is the quality of the data points. In statistics, the quality of a set of data point is how well the data points represent the population from which they are drawn.

"Data are dependable"

Data are messy. Most newly generated datasets have errors, missing observations, and unrepresentative samples. Some population properties may be under-represented or over-represented. There may be samples that should not be included in the analysis, like replicates, QA samples, and metadata. All these problems with data require a lot of processing before an analysis can begin

"Statistics always should provide unique solutions to a problem"

Even if two statisticians start with identical data sets, they may not come to identical results, and sometimes, even identical conclusions. This is because they may make different assumptions and scrub the data differently. Furthermore, there may be more than one way, even many ways, to approach a problem.

Conclusion:

In summary, statistical tools and methods are one of the the foundation pillar of the field of "data science". A data scientist has to have a good understanding and intuitive understanding of the principles of statistics right from data collection and sampling, descriptive data statistics and visualization , data analysis with the application of Central Limit Theorem and confidence intervals to sample statistics ,hypothesis testing and inference making for the population from the sample statistics and data visualization ,correlation and regression models to check for association between variables .

Nowadays the mathematical part of statistics is a highly automated affair with several tools available as depicted below, that the data scientists leverage to automate the application of the statistical processes and methods to a given data set.

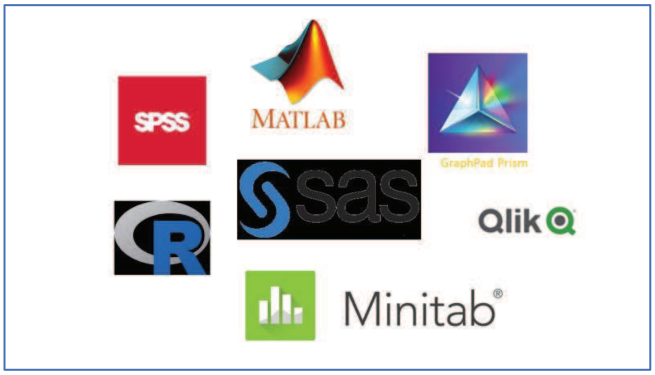


Figure11: Leading statistical analysis software tools [14]

However, the intuitive understanding of the tools and how they are to be applied and what precautions must be taken to avoid statistical bias or statistical hubris is something that separates a “full stack data scientist” from others who have only partial view of the entire data science workflow.

xxxxxxxxxxxxxx

“In God we trust,all other must bring data.”

W. Edwards Deming

References Used

[1] Statistics for data science, https://blog.floydhub.com/statistics-for-data-science/

[2] Seeing Theory, A visual introduction to probability and statistics https://seeing-theory.brown.edu/

[3] Naked Statistics, Charles Wheelan

[4]Internship data ,Courtesy SaveLife foundation

[5] Confidence Intervals, Dr Renju S Ravi https://www.slideshare.net/renjusravi/confidence-interval-50257759

[6] A Primer on A/B Testing Part 1: The Central Limit Theorem, Dimitri Theoharatos ,Jan 2019

[7] Explaining the 68-95-99.7 rule for a Normal Distribution, Michael Galarnyk, Jan 2018

[8] Hypothesis Testing-Means, Roger B. Davis and Kenneth J. Mukamal, Sep 2006,ahajournals circulations

[9] Inferential statistics, Onlineststatbook.com, Mikki Hebl and David Lane

[10] Linear Regression — Detailed View, Saishruthi Swaminathan, Feb 2018, Towards Data science

[11]Using Power of Test for good hypothesis testing, Chew Jian Chieh, isixsigma,

[12] Intuitive introductory statsitics, Douglas A. Wolfe and Grant Schneider

[13] Art of Data Analysis, Kristin H Jarman.

[14] Statistical analysis software tools https://www.capterra.com/statistical-analysis-software/

[15] Advanced High School Statistics, David Diez, Mine Cetinkaya-Rundel, Leah Dorazio, Christopher D Barr

[16] Statistics don’t lie, people do, jmvais@ https://steemit.com/trending/science