INTRODUCTORY STATISTICS

Alandra Kahl

Bentham Books

Introductory Statistics

Authored By

Alandra Kahl
Department of Environmental Engineering, Penn State
Greater Allegheny, McKeesport, Pennsylvania
USA

Introductory Statistics

Author: Alandra Kahl

ISBN (Online): 978-981-5123-13-5

ISBN (Print): 978-981-5123-14-2

ISBN (Paperback): 978-981-5123-15-9

First published in 2023.

Bentham Science Publishers Pte. Ltd.
80 Robinson Road #02-00
Singapore 068898
Singapore
Email: subscriptions@benthamscience.net

CONTENTS

PREFACE

Statistics is a complex and multi-faceted field that is relevant to many disciplines including business, science, technology, engineering and mathematics. Statistical analysis and research is critical to understanding data sets, compiling and analyzing scientific results and presenting findings. Without statistics, research would grind to a halt for lack of support and discourse regarding presentation of results. We rely on statistics and analysis to make sense of patterns, nuances and trends in all aspects of science.

This volume presents a brief but thorough overview of common statistical measurements, techniques and aspects. It discusses methods as well as areas of presentation and discourse. Chapter 1 presents an introduction to the field and relevant data types and sample data. Chapter 2 highlights summarizing and graphing, including relevant charts such as histograms, box plots, and pie charts. Chapter 3 discusses the basic concepts of probability by discourse on sample events, sample spaces, intersections, unions and complements. Chapter 3 also encompasses conditional probability and independent events as well as basic principles and rules. Chapter 4 targets random variables, including discrete values and binomial distributions. Chapter 5 summarizes continuous random variables as well as the normal distribution. Chapter 6 surveys sampling distributions, the sample mean and the central limit theorem. Chapter 7 holds forth on estimation, including intervals of confidence and the margin of error. Chapter 8 covers hypothesis testing as well as the t-test and z- test. Chapter 9 speaks about the important topics of correlation and regression. Chapter 10 briefly examines the ethics associated with statistics, including the tenets of ethical conduct for those in the discipline.

In short, this book presents a brief scholarly introduction to the chief topics of interest in statistics. It is hoped that this volume will provide a better understanding and reference for those interested in the field as well as the greater scientific community.

I am grateful for the timely efforts of the editorial personnel, particularly Mrs. Humaira Hashmi (Editorial Manager Publications) and Mrs. Fariya Zulfiqar (Manager Publications).

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The author declares no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENT

Declared none.

Alandra Kahl
Department of Environmental Engineering
Penn State Greater Allegheny
McKeesport, Pennsylvania
USA

CHAPTER 1

Introduction to Statistics

Abstract: The field of statistics is vast and utilized by professionals in many disciplines. Statistics has a place in science, technology, engineering, medicine, psychology and many other fields. Results from statistical analysis underlying both scientific and heuristic reasoning, and therefore, it is important for everyone to grasp basic statistical methods and operations. A brief overview of common statistical methods and analytical techniques is provided herein to be used as a reference and reminder material for professionals in a broad array of disciplines.

Keywords: Analysis, Heuristic reasoning, Scientific reasoning, Statistical methods.

INTRODUCTION

The field of statistics deals with the collection, presentation, analysis and use of data to make decisions and solve problems. Statistics is important for decision-making, cost-benefit analysis and many other fields. A good grasp of statistics and statistical methods can be beneficial to both practicing engineering as well as practicing businessmen. Specifically, statistical techniques can be a powerful aid in designing new products and systems, improving existing designs and developing and improving production processes. Statistical methods are used to help decide and understand variability. Any phenomenon or operation which does not produce the same result every time experiences variability. Individuals encounter variability in their everyday lives, and statistical thinking and methods can be a valuable aid to interpret and utilize variability for human benefit. For example, consider the gas mileage of the average consumer vehicle. Drivers encounter variability in their gas mileage driven by the routes they take, the type of gas they put in their gas tanks, and the performance of the car itself as examples. There are many more areas in which variability is introduced, all of which drive variability related to the gas mileage of the individuals' car. Each of these are examples of potential sources of variability in the system of the car. Statistics gives us a framework for describing this variability as well as for learning which potential sources of variability are the most important or have the greatest impacts on performance. Statistics are numerical facts or figures that are observed or obtained from experimental data.

Alandra Kahl

Data is typically collected in one of two ways, either observational study or designed experiments. Data can also be obtained *via* random sampling or randomized experiments, but it is difficult to discern whether the data has any statistical significance- that is, is the difference found in the sample strictly related to a specific factor [1]. Simply put, is there a cause-and-effect relationship between the observed phenomena and the result? It is far more useful to collect data using observational study or designed experiments for statistics, as researchers can better narrow, understand and discard confounding factors within the gathered data set.

The first way that data can be collected is by observational study. In an observational study, the researcher does not make any impact on the collection of the data to be used for statistics; rather, they are taking data from the process as it occurs and then trying to ascertain if there are specific trends or results within that data [1]. For example, imagine that the interested researcher was curious about whether high iron levels in the body were associated with an increased risk of heart attacks in men. They could look at the levels of iron and other minerals within a group of men over the course of five years and see if, in those individuals who displayed high iron levels, there were more heart attacks. By simply tracking the subjects over time, the researchers are performing an observational study [1]. It is difficult in an observational study to identify causality as the observed statistical difference could be due to factors other than those the researchers are interested in, such as stress or diet in our heart attack example. This is because the underlying factor or factors that may increase the risk of heart attack was not equalized by randomization or by controlling for other factors during the study period, such as smoking or cholesterol levels [2]. Another way that observational data is obtained to by data mining, or gleaning information from previously collected data such as historical data [1]. This type of observational study is particularly useful in engineering or manufacturing, where it is common to keep records on batches or processes. Observational engineering data can be used to improve efficiency or identify shortcomings within a process by allowing a researcher to track a trend over time and make conclusions about process variables that may have positively or negatively caused a change in the final product.

The second way that data can be obtained for statistical work is through a designed experiment. In a designed experiment, the researcher makes deliberate or purposeful changes in the controllable variables of a system, scenario or process, observes the resultant data following these changes and then makes an inference or conclusion about the observed changes. Referring to the heart attack study, the research could design an experiment in which healthy, non-smoking males were given an iron supplement or a placebo and then observe which group had more heart attacks during a five-year period. The design of the experiment now controls for underlying factors, such as smoking, allowing the researchers to make a stronger conclusion or inference about the obtained data set. Designed experiments play an important role in science, manufacturing, health studies and engineering as they help researchers eliminate confounding factors and come to strong conclusions [1]. Generally, when products, guidelines or processes are designed or developed with this framework, the resulting work has better performance, reliability and lower overall costs or impacts. An important part of the designed experiments framework is hypothesis testing. A hypothesis is an idea about a factor or process that a researcher would like to accept or reject based on data. This decision-making procedure about the hypothesis is called hypothesis testing. Hypothesis testing is one of the most useful ways to obtain data during a designed experiment, as it allows the researcher to articulate precisely the factors which the researcher would like to prove or disprove as part of the designed experiment [1].

Modelling also plays an important role in statistics. Researchers interested in statistics can use models to both interpret data as well as to construct data sets to answer hypotheses. One type of model is called a mechanistic model. Mechanistic models are built from underlying knowledge about physical mechanisms. For example, Ohm's law is a mechanistic model which relates current to voltage and resistance from knowledge of physics that relates those variables [1]. Another type of model is an empirical model. Empirical models rely on our knowledge of a phenomenon but are not specifically developed from theoretical or first principles understanding of the underlying mechanism [3]. As an example, to illustrate the difference between mechanistic models and empirical models, consider the bonding of a wire to a circuit board as part of a manufacturing process. As part of this process, data is collected about the length of the wire needed, the strength of the bond of the wire to the circuit and the amount of solder needed to bond the wire. If a researcher would like to model the amount of solder needed to bond the wire related to the amount of force required to break the bond, they would likely use an empirical model as there is no easily applied physical mechanism to describe this scenario. Rather, the researcher determines the relationship between the two factors by creating a plot that compares them. This type of empirical model is called a regression model [1]. By estimating the parameters in regression models, a researcher can determine where there is a link between the cause and effect of the observed phenomena.

Another type of designed experiment is factorial experiments. Factorial experiments are common in both engineering and biology as they are experiments in which several factors are varied together to study the joint effects of several factors. Returning to the circuit board manufacturing example, an interested

CHAPTER 2

Summarizing and Graphing

Abstract: Data captured during analysis can be easily summarized using visual techniques such as graphing. Graphing is utilized to both summarize and convey information in a clear and readable format. Graphing techniques discussed in this chapter include frequency distributions, histograms, box plots and Pareto charts.

Keywords: Box plot, Frequency distribution, Histogram.

INTRODUCTION

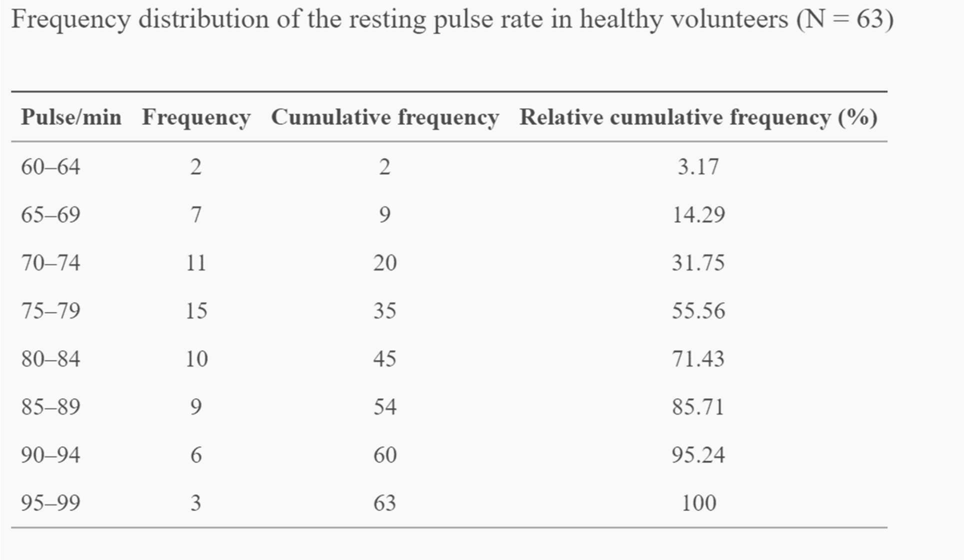
Visualization of data *via* graphical means is a common way to convey statistical information. Data is easily summarized by several graphical techniques, including frequency distributions, histograms, box plots, Pareto charts, and dot plots, among others. By showing data in graphical form, researchers can pinpoint trends, isolate outliers and approve or disprove hypotheses. Showing data in graphs allows for broader distribution of the dataset as well as simple interpretations of data. Differentiations of data can be accomplished by outline or color difference, or other charting techniques may be used. Graphing is important to convey information as well as support further data analysis.

FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

Once data has been collected, the next step for data analysis is to organize the data to determine if there are meaningful trends or patterns present within the dataset. A common method for organizing data is to construct a frequency distribution or frequency table. This type of data organization is a tabulation of the number of individuals in each category organized graphically with respect to the scale of measurement [9]. Frequency distribution is useful because it allows a researcher to understand overarching trends in the dataset briefly. For example, by graphing data this way, it can be seen if observations are high or low, or concentrated around one area of the scale [10]. By showing how individual distributions relate to the entire dataset, trends can be seen.

Alandra Kahl

Frequency distributions or frequency tables, are constructed to show the different categories of measurement within the dataset as well as the number of observations within each set. An example of a frequency table is shown in Fig. (**1**).



Frequency distribution of the resting pulse rate in healthy volunteers (N = 63)

Pulse/min	Frequency	Cumulative frequency	Relative cumulative frequency (%)
60–64	2	2	3.17
65–69	7	9	14.29
70–74	11	20	31.75
75–79	15	35	55.56
80–84	10	45	71.43
85–89	9	54	85.71
90–94	6	60	95.24
95–99	3	63	100

Fig. (1). Frequency table.

It is important to have an idea of the maximum and minimum ranges of values on the scale of the dataset before beginning to construct a frequency distribution so that the measurement scale is chosen appropriately. The selected data range is then divided into intervals called class intervals. These class intervals occur arbitrarily in order to minimize bias within the data sorting. It is also important to choose the correct amount of class intervals for the dataset. If there are too few class intervals, the dataset is too bulky, and it is difficult to see deviations. If there are too many class intervals, the dataset divisions are too small and small deviations are magnified. Generally, between six and fourteen class intervals are sufficient. It is also important to know the width of the class. The class width can be calculated by dividing the range of observations by the number of classes [9]. It is desirable to have equal class widths. Unequal class widths should only be used when there are large gaps in the data. Class intervals should also be mutually exclusive and nonoverlapping [11]. Therefore, data within one class is not repeated within another class. Finally, for greater precision and data sorting, open-ended classes at the lower and high ends of the data set should be avoided. For example, classes <10 or >100.

There are many ways to represent frequency data graphically. A commonly used way to show frequency distribution is called a histogram. A histogram is shown in Fig. (**2**).

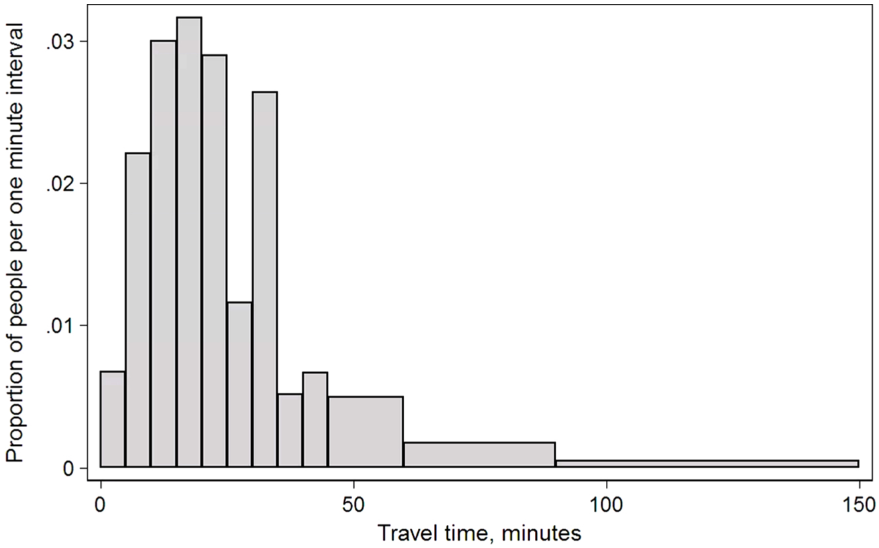


Fig. (2). A histogram [Source: WFP / Wikimedia Commons / CC BY 3.0].

A histogram shows the variable of interest in the x-axis and the frequency of the occurrence of that variable or the number of observations in the y-axis. Percentages can also be used if the objective is to compare two histograms having a different number of subjects. A histogram is used to depict the frequency when data are measured on an interval or ratio scale [10]. While a histogram may appear at first glance to look like a bar chart, there are three major differences between these graphical representations of data. In a histogram, there is no gap between the bars as the variable is continuous. Only if there is a large gap in the data scale will a gap occur in the histogram. In a histogram, the width of the bars is dependent on the class interval. If there are unequal class intervals, the widths of the bars of the histogram will reflect this inequality. Thirdly, in a histogram, the area of each bar corresponds to the frequency, where in a bar chart, it is the height that shows the frequency [10].

Histograms can also be used to generate other representations of the dataset. For example, a frequency polygon can be constructed by connecting the midpoints of the tops of the bars of a histogram using straight lines. A frequency polygon is shown in Fig. (**3**).

CHAPTER 3

Basic Concepts of Probability

Abstract: A commonly used statistical measure is the measurement of probability. Probability is governed by both additive and multiplicative rules. These rules determine if events are independent of one another or dependent on previous or related outcomes. Conditional probability governs events that are not independent of one another and helps researchers to better make predictions about future datasets.

Keywords: Conditional probability, Dependent, Independent, Probability.

INTRODUCTION

Probability is a commonly encountered statistical measure that helps to determine the likelihood of the outcome of a specific event. By understanding what the odds are of an event occurring, researchers can make further predictions about future datasets as well as to better understand collected data. The rules of probability govern the way that odds are generated as well as their interpretation. Conditional probability is the analysis of events that are not independent of one another and is frequently utilized to better understand collected datasets and make predictions about future outcomes.

A probability is a number that expresses the possibility or likelihood of an event occurring. Probabilities may be stated as proportions ranging from 0 to 1 and percentages ranging from 0% to 100%. A probability of 0 implies that an event cannot happen, while a probability of 1 suggests that an event is very likely to happen. A probability of 0.45 (45%) means that the event has a 45 percent chance of happening [26].

A study of obesity in children aged 5 to 10 seeking medical treatment at a specific pediatric clinic may be used to demonstrate the notion of likelihood. All children seen in the practice in the previous 12 months are included in the population (sample frame) described below [27].

Assume a polling company asks 1,200 people a series of questions to assess the percentage of all voters who support a certain bond issue. We would anticipate the percentage of supporters among the 1,200 polled to be similar to the percentage

Alandra Kahl

among all voters, but this does not have to be the case [28]. The survey result has a certain amount of unpredictability to it. We have faith in the survey result if the result is extremely likely to be close to the real percentage. If the percentage isn't likely to be near the population proportion, we shouldn't take the survey results too seriously. Our confidence in the survey result is determined by the possibility that the survey percentage is close to the population proportion [28]. As a result, we'd want to be able to calculate that probability. The problem of calculating it falls within the probability category, which we will look at in this chapter [28].

SAMPLES EVENTS AND THEIR PROBABILITIES

Sample Spaces

A common example of a random experiment is the rolling of a six-sided die; although all potential outcomes may be stated, the actual result on any specific trial of the experiment cannot be predicted with confidence [28]. When dealing with a situation like this, it is preferable to give a numerical number to each outcome (such as rolling a two) that indicates how often the event will occur. A probability would be assigned to any event or collection of outcomes, such as rolling an even number that shows the likelihood that the event will occur if the experiment is carried out similarly [28].

A random phenomenon's sample space is just the collection of all conceivable (basic) outcomes. Outcomes are the most fundamental things that can happen. When you roll a dice, for example, the potential outcomes are 1, 2, 3, 4, 5, and 6--resulting in a sample space of {1,2,3,4,5,6} [28].

Following the specification of the sample space, a set of probabilities is given to it, either by repeated testing or through common sense. The outcomes are usually given the same probability; however, this is not always the case [28].

You commonly use letters to indicate outcomes, such as x or c, and then P to represent the probability of the occurrence (x). The following are the two rules that the probability assignments must follow: It must be true for any result x that

$0 < P(x) < 1$ (it is allowed for $P(x)$ to equal 0 or 1 -- if $P(x) = 0$, it indicates that x virtually never occurs, and if $P(x) = 1$, it means that x practically always happens.) [28]

When all the probability of all the eventualities are added together, you get 1. (this means that your list of outcomes includes everything that can happen) [28].

Event

A random experiment is a method that generates a specific result that cannot be anticipated with confidence. A random experiment's sample space is the collection of all conceivable results. A subset of the sample space is an event [28].

If the result observed is an element of the set E, an event E is said to occur on a certain experiment trial [28].

Examples

Experiments are a part of life in almost every field of study. Special sorts of experiments are also addressed in probability and statistics. Consider the examples below.

Example 1

A coin is thrown. If the coin does not fall on the side, the experiment may have two alternative outcomes: heads or tails. It is impossible to indicate the result of this experiment at any time. You may throw the coin as many times as you like [29].

Example 2

A roulette wheel is a circular disc with 38 identical sections numbered 0 through 36, plus 00. The wheel is rolled in the opposite direction after a ball is rolled on its edge. Any of the 38 numbers, or a combination of them, maybe gambled on. You may also wager on a certain hue, red or black. If the ball falls in the number 32 sector, for example, everyone who bet on 32 or a combination of 32 wins, and so on. In this experiment, all potential results are known in advance, namely 00, 0, 1, 2,..., 36, yet the outcome is undetermined on every execution of the experiment, provided, of course, that the wheel is not fixed in any way. The wheel can be rolled an infinite number of times [29].

Example 3

A company makes 12-inch rulers. The experiment aims to measure the length of a ruler manufactured by the manufacturer as precisely as feasible. Because of manufacturing mistakes, it is impossible to tell the real length of the chosen ruler. However, the length will be between 11 and 13 inches, or between 6 and 18 inches if one wants to be careful [29].

CHAPTER 4

Discrete Random Variables

Abstract: Discrete random variables are variables that can only take on a countable number of discrete or distinct variables. Only a finite number of values can be obtained for the result to be a discrete random variable. Discrete variables can be visualized or understood as a binomial distribution. The outcome of a binomial distribution is how often a particular event occurs in a fixed number of times or trials.

Keywords: Binomial distribution, Outcome, Random.

INTRODUCTION

Random Variables

Unknown variables and functions that assign values to each of an experiment's results are referred to as random variables and functions, respectively. For the most part, random variables are denoted by letters, and they may be divided into two categories: discrete variables, which are variables with specified values, and continuous variables, which are variables that can take on either value within a certain range of values. Statistical connections between random variables are often seen in econometric and regression analysis, where they are utilized to discover statistical links between variables [30].

Understanding Random Variables

Random variables are exploited in probability and statistics to measure the results of a random event, and as a result, they might have a wide range of possible values. Random variables must be measured to be used, and real numbers commonly represent them. After three dice are rolled, the letter X may be selected to symbolize the number that comes up as a consequence. In this scenario, X may be three (1+1+1), eighteen (6+6+6), or anything in between, since the greatest number on a die is six, the lowest number is one, and the highest number on a die is six [30].

A random variable differs from an algebraic variable because it is not predictable. A parameter in an algebraic equation is an unknown value but may be determined

Alandra Kahl

by calculation. With the equation 10 + x = 13, we can see that we can compute the precise number for x, which happens to be 3. On the other contrary, a random variable has a set of possible values, and any of those values might result in the desired outcome, as shown by the dice in the preceding example [30].

Various features, such as the average price of an asset over a specific period, the return on investment after several years, the projected turnover rate at a firm during the next six months, and so on, may be assigned random variables in the corporate world. Whenever a risk analyst wants to evaluate the likelihood of an undesirable event happening, they must include random variables in their risk models. These factors are provided *via* scenario and sensitivity analysis tables, which risk managers utilize to make judgments on risk mitigation strategies [30].

Types of Random Variables

A random variable can be a discrete variable or a continuous variable. Discrete random variables have a finite number of different values that may be calculated. Consider the following experiment: a coin is thrown three times, and the results are recorded. If X represents the number of times the coin has shown up heads, then X is a discrete random variable that can have the quantities 0, 1, 2, and 3 when the coin comes up heads (from no heads in three successive coins tosses to all heads). There is no other potential value for X [30].

A continuous random variable may represent any value inside a particular range or period, and the variable can take on an endless number of potential values. An experiment in which the quantity of rainfall in a city is measured over a year, or the height of a random group of 25 individuals, would be an example of a continuous random variable [30].

As an illustration of the latter, consider the case in which Y represents a random variable representing the average height of a random group of 25 persons. You will discover that the final output is a continuous number since height maybe 5 feet or 5.01 feet or even 5.0001 feet in height. To be sure, there is an endless number of different heights to choose from [30]!

Random variables have probability distributions, which describe the possibility that any possible values will be seen for that variable. Assume that the random variable, Z, is the number that appears on the top face after it has been rolled once in a row. As a result, the potential values for Z will be 1, 2, 3, 4, 5, and 6, respectively. Each of these numbers has a 1/6 chance of being the value of Z since they are all reasonably plausible to be the value of Z [30].

In the case of a dice throw, the chance of obtaining a 3 or P (Z=3) is 1/6, as is the likelihood of receiving a 4 (Z=2), or any other number, on all six sides of the die. It is important to note that the total of all probability equals one [30].

Example of Random Variable

A coin flip is an excellent demonstration of a random variable in statistical analysis. For example, suppose you had a probability distribution in which the outcomes of a random event are not equally likely to occur. If the random variable, Y, represents the number of heads we get from tossing two coins, then Y might be one of three values: 0, one, or two. We may get no heads, one head, or both heads, depending on how the dice fall on a two-coin toss [30].

On the other extreme, the two coins land in four possible configurations: TT, HT, TH, and HH. As a result, the probability of receiving no heads is 1/4 since we only have one opportunity of getting no heads (*i.e.,* two tails [TT] when the coins are tossed). In a similar vein, the chance of receiving two heads (HH) is one in every four. It is important to note that receiving one head can occur twice: in HT and TH. As a result, P (Y=1) = 2/4 = 1/2 in this example [30].

Discrete Random Variables

When it comes to discrete variables, they are variables that can "only" be represented by specified integers on the number line [30].

Discrete variables are often denoted by capital letters: X, Y, Z, *etc*.

X is a random variable, and the probabilities in the probability distribution of X must meet the following two scenarios [30]:

- Each probability P(x) must be between 0 and 1 to be valid: $0 \leq P(x) \leq 1$.
- In this case, the total of all probability is one: P(x) =1.

Example:

A variable that can only contain integers or a variable that can only contain positive whole numbers are both examples of this kind of variable [30].

In the case of discrete variables, they may either take on an unlimited number of values or be restricted to a fixed number of possible values [30].

For example, the number we get when rolling a die is a discrete variable, and it can only have one of these values: 1, 2, 3, 4, 5, or 6 [30].

CHAPTER 5

Continuous Random Variables

Abstract: Distributions of data give important information about the dataset, both to researchers and to analysts. Continuous random variables are those variables whose outcomes are measured instead of random variables that are counted. These variables can be handled using probability density functions and cumulative distribution functions. Both have appropriate times and usage cases.

Keywords: Binomial, Continuous, Normal.

INTRODUCTION

The distribution of data within a dataset is important information as adequately understanding the shape of a dataset assists researchers to both draw better conclusions and visualize their results during analysis. A binomial distribution is used when researchers want to understand the consequences of two independent outcomes, whereas a Poisson distribution is used to better understand the results of independent trials with the additional inclusion of time within the dataset. Distributions are powerful analyses of data that researchers can use to add depth and understanding to the analysis of their datasets.

A random variable X is seen as being continuous if and only if the probability that its recognition will fall within the interval [a,b] can be represented as an integral [36]:

$$\mathrm{P}(X \in [a, b]) = \int_a^b f_X(x)dx$$

Whereas integrated part

$$f_X : \mathbb{R} \to [0, \infty)$$

Is known as probability density of function X.

Alandra Kahl

It is yet critical to note that an integral is used to compute the area under a curve. If you want to define a continuous variable, you may consider its integral as the surface under the probability density function in the period between a and b (Fig. **1**) [36].

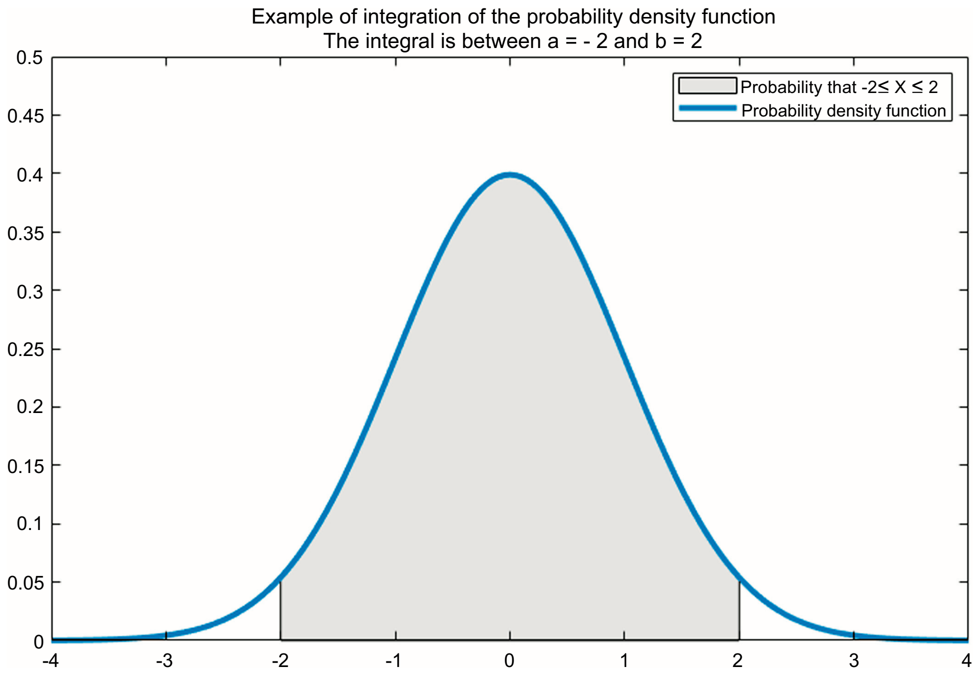


Fig. (1). Example of integration of the probability density function [36].

An illustration of a continuous random variable is one that has two primary characteristics [36]:

- The set of possible values it may take is not countable, and the cumulative distribution function of the variable is determined by integrating a probability density function [36].

Intervals are given probabilities based on the data.

The first thing to observe about the definition preceding is that the assignment of probabilities describes the distribution of a continuous variable to intervals of integers rather than the assignment of frequencies. Consider the fact that the distribution of a discrete integer is defined by assigning probabilities to single numbers, as opposed to the former [36, 37].

The uniform distribution (Fig. **2**) is a continuous probability distribution that deals with occurrences with an equal chance of occurring. When solving issues with a uniform distribution, keep in mind whether the data includes or excludes endpoints.

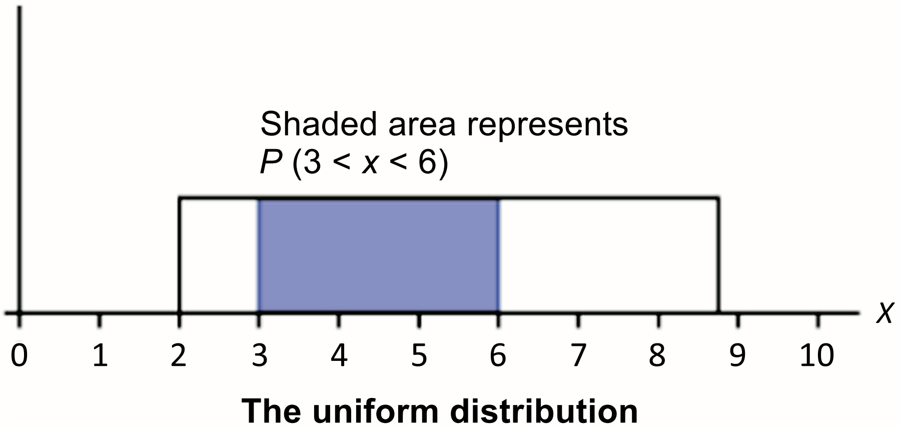


Fig. (2). The uniform distribution [37].

The exponential distribution (Fig. **3**) is often used to calculate the length of time before a certain event happens. The length of time (starting now) before an earthquake happens, for example, has an exponential distribution. Other examples are the duration of long-distance business phone conversations in minutes, and the time a vehicle battery lasts in months. It may also be shown that the value of the change in your pocket or handbag follows a roughly exponential distribution.

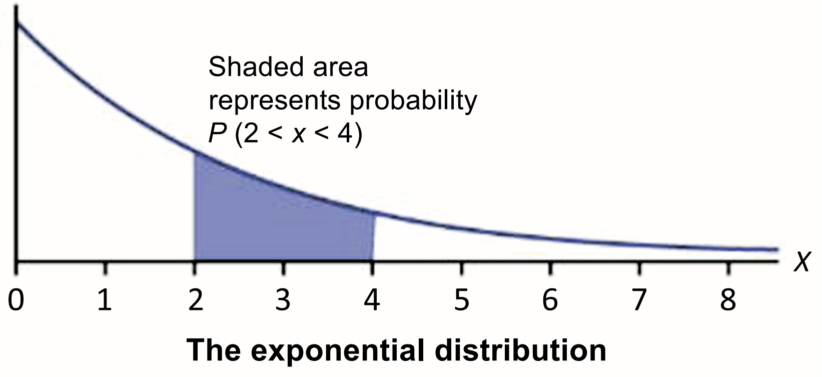


Fig. (3). The exponential distribution [37].

Exponential distributions are widely employed in product dependability estimates or determining how long a product will survive.

CHAPTER 6

Sampling Distributions

Abstract: Sampling of a distribution and as well as how to treat the sampling distribution of the sample mean is an important topic for understanding statistics. The central limit theorem, when applied to this type of dataset can allow researchers to make predictions about future datasets as well as to better understand current datasets. Normally distributed populations are important to correctly treating the data occurring within the dataset.

Keywords: Bell-curve, Central limit theorem, Normal distribution.

INTRODUCTION

The mean and standard deviation from the sample mean are important information from the dataset that helps researchers to understand the distribution of the data as well as its shape and values. A small sample proportion will show the errors as well as the distance from mean to assist researchers in better understanding the character of their data by allowing researchers to drill down into the nuances of the data.

A sampling distribution is a probability distribution of a statistic derived by randomly selecting a population sample. A finite-sample distribution, also known as a finite-sample distribution, depicts the frequency distribution of how far apart distinct events will be for a given population (Fig. **1**) [43].

The sampling distribution is influenced by several variables, including the statistic, sample size, sampling method, and general population. It is used to determine statistics for a given sample, such as means, ranges, variances, and standard deviations (SD) [43].

A statistic is a number calculated from a sample, such as a sample mean or standard deviation. Every statistic is a random variable since a sample is random: it varies from sample to sample in ways that can't be predicted confidently [44]. It has a mean, a standard deviation (SD), and a probability distribution as a random variable. The sampling distribution of a statistic is the probability distribution of that statistic. Sample statistics are often performed to estimate the corresponding population parameters rather than their goal [44].

Alandra Kahl

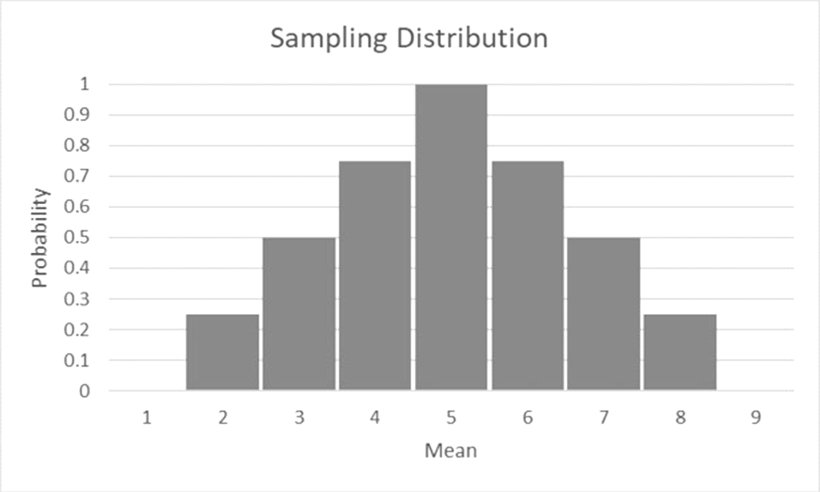


Fig. (1). Sampling Distribution [43].

The mean, standard deviation (SD), and sampling distribution of a sample statistic are introduced in this chapter, with a focus on the sample mean x¯ [44].

THE MEAN AND STANDARD DEVIATION (SD) OF THE SAMPLE MEAN

Let's say we want to calculate a population's mean μ. In actuality, we would usually collect one sample. Consider the case where we collect sample after sample of the same size n and calculate the sample mean x¯. For each. Each time, we'll most likely obtain a different x¯ value [44]. The sample mean x¯ is a random variable, meaning that it fluctuates from sample to sample in unpredictable ways. When we think of the sample mean as a random variable, we'll write X¯ for the values it takes. The random variable X¯, has a mean denoted as μ x and a standard deviation (SD) indicated by the letters σ x. Here's an example where the population is so tiny, and the sample size is so small that we can write down every sample [44].

Examples

Example 1

Four rowers, weighing 152, 156, 160, and 164 pounds, make up a rowing team. Calculate the sample mean for each available random sample with size two replacement. Use these to calculate the sample mean X¯, probability distribution, mean, and standard deviation (SD) [44].

Solution

The following table lists all potential samples with size two replacement, as well as the mean of each [44]:

Sample	Mean		Sample	Mean		Sample	Mean		Sample	Mean
152, 152	152	-	156, 152	154	-	160, 152	156	-	164, 152	158
152, 156	154	-	156, 156	156	-	160, 156	158	-	164, 156	160
152, 160	156	-	156, 160	158	-	160, 160	160	-	164, 160	162
152, 164	158	-	156, 164	160	-	160, 164	162	-	164, 164	164

The sample mean X¯ has seven different values, as shown in the table. The number x¯=152, like the value x¯=164, occurs just once (the rower weighing 152 pounds must be picked both times), while the other values occur several times and are therefore more likely to be detected than 152 and 164. Because the 16 samples are all equally probable, we can count to get the sample mean's probability distribution [44]:

x¯	152	154	156	158	160	162	164
P(x¯)	1/16	2/16	3/16	4/16	3/16	2/16	1/16

We use the formulae for the mean, and standard deviation of a discrete random variable from Section 4.3.1 "The Mean and Standard Deviation of a Discrete Random Variable" in Chapter 4 "Discrete Random Variables", get a result for μX¯ [44].

μX¯.= Σx- P(-)=152(116)+154(216}+156(316)+158(416)+160(316}+162(216)+ 164(116)=158

For σX- we first compute Σx-2P(x-):

1522(116)+1542(216)+1562(316)+1582(416)+1602(316)+1622(216)+1642(116)

which is 24,974, so that

σX =Σx-2P (x-) - μx-2=24,974-1582=10

In the example, the mean and standard deviation (SD) of the population {152,156,160,164} are μ = 158 and σ= 20 respectively. The mean of the sample mean X¯ that we just calculated is identical to the population mean. We just calculated the standard deviation (SD) of the sample mean X¯, which is the population standard deviation (SD) divided by the square root of the sample size:

CHAPTER 7

Estimation

Abstract: It is critically important for researchers to select the correct sample size to determine the population means. By understanding the difference between small and large sample sets, researchers can then construct intervals of confidence that assist in determining the population means. Confidence intervals are the margins of the error present within the dataset and are included to show the confidence of the researcher in the integrity of their dataset. Common confidence intervals are 90%, 95%, and 99%. The Z confidence level is calculated to show where the mean is likely to fall, and the T confidence level is used only when the sample size is smaller than 30 samples.

Keywords: Interval, Population mean, Sample estimation.

INTRODUCTION

By including sample sizes in their data analysis, researchers can simply describe the uncertainty associated with their dataset as well as to show the error margins present within the set. The distinctions between large and small sample sizes are typically set to encompass both the upper and lower ends of the dataset as well as the margins of error present within the data set. Confidence levels are also included with confidence intervals to show agreement with the assumptions of the dataset. Common confidence levels are 90%, 95%, and 99%. These levels are used to show the surety of the researchers in their measurements as well as to assist in predicting the characteristics of future datasets based on current understandings to ensure that the correct size sample is being used to determine the population means.

Construction of Confidence Intervals

The statistical inference technique refers to obtaining conclusions from data using statistical methods. As a result, the most crucial things to consider are testing hypotheses and concluding. As a branch of statistics, estimate theory is responsible for extracting parameters from data that have been contaminated by noise [46].

Calculating the values of parameters using measured and observed empirical data is a subfield of statistics and signal processing used to calculate the values of par-

Alandra Kahl

ameters in mathematical models. To measure and diagnose the real value of a function or a certain group of populations, the process of estimating must be carried out. It is carried out based on observations made on samples that provide a composite representation of the target population or function. When performing the estimate process, a variety of statistics are used [46].

In statistics, one of the most common applications is the estimation of population parameters using sample statistics. For example, a survey may be conducted to determine the percentage of adult citizens of a city who favor a proposal to construct a new sports stadium. A random sample of 200 persons was used to determine whether or not they endorsed the concept. As a result, 0.53 (106/200) of the persons in the sample agreed with the notion. The population percentage point estimate is defined as 0.53 (or 53 percent) divided by the total population. In this case, the estimate is a point estimate since it is composed of a single number or point [47].

It is very unusual for the actual population parameter to be the same sample value. When considering the hypothetical situation in which we surveyed the whole city's adult population, it is very implausible that precisely 53 percent of the population would support the notion. As an alternative, we may present a range of possible values for the parameter by using confidence intervals [47].

As a result, point estimates are often augmented with interval estimates or confidence ranges to provide a complete picture. Constructed by utilizing a technique that includes the population parameter for a set fraction of the time, confidence intervals include the population parameter. The pollster would arrive at the following 95 percent confidence interval, for example, if he or she utilized a procedure that included the parameter 95% of the time it was used. The pollster would thus conclude that anywhere between 46 percent and 60 percent of the public favors the initiative. In most cases, the media will publish this result by stating that 53 percent of the population supports the idea with a margin of error of 7 percent or less [47].

Interval Estimate *vs.* Point Estimate

To estimate population parameters, statisticians employ sample statistics. Sample means, for example, are used to get population means, whereas sample proportions are used to calculate population proportions [48].

A population parameter estimate may be stated in two ways:

1. An educated guess. A single value of a statistic is a point a parameter for a population is estimated. The sample mean x, for example, is a point estimate of

the population mean. The sample percentage p is a point estimate of the population proportion P, in the same way.

2. Estimated interval. A population parameter is between two values specified by an interval estimate. An interval estimate of the population mean, for example, is an x b. It denotes that the population mean exceeds a but falls short of b.

Intervals of Confidence

Statisticians use a confidence interval to represent the accuracy and uncertainty associated with a sampling process. There are three elements to a confidence interval.

- The interval itself.
- The confidence level.
- The parameter being estimated.

A sampling method's confidence level expresses the method's degree of uncertainty. The statistic and the margin of error together provide an interval estimate of the method's accuracy [48].

Consider the situation in which you are attempting to calculate an interval a parameter for a population is estimated. For the purpose of defining this interval estimate, a 95 percent confidence interval might be employed. It shows that, if we used the same sampling approach to pick other samples and produce different interval estimates, the true population parameter would fall inside the range indicated by the sample statistic + margin of error 95 percent of the time [48].

Because confidence intervals represent (a) the accuracy of the estimate and (b) the estimate's uncertainty, they are favored over point estimates [48].

Confidence Level

A confidence level is the probability portion of a confidence interval. The confidence level indicates how likely a sampling technique yields a confidence interval containing the genuine population parameter [49].

Learn how to interpret what a confidence level signifies in this article. Consider the following scenario: we collected all possible samples from a population and generated confidence intervals for each of them. The true population parameter would be included in certain confidence intervals but not in others, depending on the confidence interval [49]. It is implied that the genuine population parameter is present in 95 percent of the intervals if the confidence level is 95 percent; if the

CHAPTER 8

Hypothesis Testing

Abstract: Hypothesis testing is a simple way to construct information about a dataset and to determine if that information is true or false. Another method to think about hypothesis testing is to use it as a way to decide whether an assumption about a dataset succeeds and is supported or fails and is not supported. It is used frequently in science and medicine to better understand the effects of drugs and treatments.

Keywords: Alternative, Hypothesis, Null.

INTRODUCTION

Hypothesis testing evaluates two statements about a population to determine which is supported by the data. The two hypotheses are called the null hypothesis and the alternative hypothesis [54]. The null hypothesis states that there is no effect or the effect is equal to zero, while the alternative hypothesis states that there is an effect or the effect is not equal to zero. Hypothesis testing is used frequently in both the medical industry as well as in scientific research to determine if the results of an experiment within a sample population are statistically significant or have an effect [54]. It is important to understand whether the desired outcome of an experiment is for the null hypothesis to be true or for the alternative hypothesis to be true. For example, if researchers were testing a new drug for harmful side effects, they would want the null hypothesis for their results to be true, or that the side effects would be equal to zero. If the researchers were instead curious about whether the drug had an effect on the disease they were treating, they would want the alternative hypothesis to be true, or the drug to have an effect on the disease. Hypothesis tests use a random sample to draw conclusions about entire populations, so they are not 100% accurate [55]. There are two types of errors that can occur in hypothesis testing; false positives and false negatives [55]. False positives occur when the null hypothesis is rejected, but it is true, while for false negatives, the null hypothesis should be rejected and is not. Hypothesis tests rely on measurements of significance to determine how strongly the sample results must contradict the null hypothesis. The measurement of significance is called alpha and is set before the study begins. It can be thought of as the probability that the research says there is an ef-

Alandra Kahl

fect when there is no effect. Lower levels of significance require stronger evidence to be true. For example, a 0.05 level of significance states that there is a 5% chance of deciding an effect exists when there is not an effect, while a 0.01 level of significance lowers this threshold to a 1% chance. Significance levels can also be visualized as critical regions on a normal curve. For example, a 0.05 significance level would result in shading in the far ends of the normal curve, 0.025 from each end. Both ends are used when it is a two-tailed hypothesis test, indicating that there could be potentially positive or negative effects shown in the test [56]. The term tail refers to the ends of the normal distribution curve. One-tailed hypothesis tests only test for effects in one direction, either positive effects above the mean or negative effects below the mean. Two-tailed tests are used when researchers are curious about any effect in a population or sample, while one-tailed tests are used either when researchers only care about an effect in one direction, or the effect can only occur in one direction [56]. These classic methods are provided to the reader as an introduction to the subject. There are other methods, which the reader is invited to seek out and explore on their own, should they find the subject interesting.

Z-Test

The Z-test is a type of statistical test that is used to determine if two population means are different when their variances are known, and the sample size is large [57]. Z-scores are the resultant of the test and are used in hypothesis testing. When a z-score is used, the data also follows a normal distribution. The standard deviation should also be known when a z-test is to be performed. In order for a Z-test to be conducted, the sample should also be randomized and all of the sample values independent from one another [57]. The Z-score derives from how many standard deviations above or below the mean the Z-test result is. Types of tests that can be performed using a Z-test include one and two-sample tests, paired difference tests and a maximum likelihood estimate.

Z -tests take a few steps to calculate: the alternative and null hypothesis should be stated, the alpha level and critical value of z should be found, and then the result of the z-test is compared to the critical z value to determine if the null hypothesis should be supported or rejected [57].

A one-sample Z-test is performed to compare the sample mean with the population mean. In order to complete a one-sample z-test, the following formula is used [57]:

$$\text{z score} = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

where x bar represents the sample mean, mu represents the population mean, alpha is the population standard deviation, and n is the number of samples. If the calculated Z-score is greater than the critical value, then the null hypothesis can be rejected, meaning that there is an effect or the effect of the experiment is not zero.

A two-sample Z-test is used to compare the mean of two samples. It is represented by the following formula, where 1 and 2 are used to represent each of the samples [57]:

$$\text{z score} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

the x bar calculation is used to represent the differences between the sample means, the mu calculation represents the difference between the population means, each alpha reflects the population standard deviation in each set, and each n represents the number of samples in each set. As before, if the Z-score result is greater than the critical value, the null hypothesis is rejected [57].

A paired difference Z-test is used to determine if the mean difference between two populations is greater than, less than or equal to zero. For a paired sample Z-test, the usual Z-test population requirements of a large data set that is normally distributed and randomly sampled apply as well as two additional requirements that the data must be continuous and the two sets, should show a similar spread between groups. Two calculations are made for a paired different Z-test. First, the paired differences in each sample are calculated using the following equation [57]:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

where Xi represents each sample up to the total number of samples present, and n is the total number of sample. This X bar value is then input into the following formula to find the Z-score [57]:

$$z = \frac{\bar{X}}{\sigma/\sqrt{n}}$$

where alpha is the standard deviation, as we have seen before. The result is then compared to the p-value to reject or accept the hypothesis.

CHAPTER 9

Correlation and Regression

Abstract: Researchers use correlation and regressions to show the relationships between factors or scenarios in a dataset. These types of analyses are important to all aspects of statistical analysis and are a common way to report connections between factors in a data set. This section will address basic linear correlation and regression techniques.

Keywords: Correlation, Linear, Negative, Positive, Regression.

INTRODUCTION

Correlation and regression are complex and powerful techniques that are used in data analysis. Correlation and regression are used to analyze the relationship between two continuous variables. In general, the dependent or outcome variable is referred to as Y, and the independent or predictor variable is referred to as X [61]. This type of analysis is used in various disciplines, but is most common in science and technology, when researchers are trying to understand how aspects of the data affect one another and to make predictions about the future dataset that include those parameters.

CORRELATION

Correlation quantifies the direction and strength of the relationship between two or more numeric variables. It lies between +1.0 and -1.0. When the correlation is negative, the slope of the regression line will also be negative and vice versa [62]. The correlation squared, which is written as R2, has a special meaning in simple linear regression. The R2 value represents the proportion of variation in Y as explained by X. For correlation, X and Y data can be used interchangeably. Also, in correlation, X and Y are random variables. Correlation is a more precise summary of the relationship between two variables [63]. The result of a pairwise correlation can be gathered together in a table to summarize relationships within the data set. Variables are considered to be uncorrelated when a change in one variable does not result in a change in the other variable. When both variables move in the same direction, or an increase in X results in an increase in Y, this is considered a positive correlation [64]. For example, the demand and price of a

Alandra Kahl

product are often positively correlated. As the demand for the item increases, so does the price of the item. When the variables are moving in opposite directions, or an increase in X causes a decrease in Y, this is a negative correlation [62]. The example of price and demand can also work here, as an increase in the price of the product will result in a decrease in demand as less people are willing to pay the increased price. Correlations allow for the association or absence of a relationship between two variables [65]. If we can show two variables are associated or correlated, then the strength of the relationship between the two can be measured. This allows for predictions using one factor to be applied to additional factors as we know how a change in one aspect will positively or negatively affect the other and by what degree, based on the direction and strength of the correlation [62]. Correlations are visualized using scatter plots, as shown in Fig. (**1**).

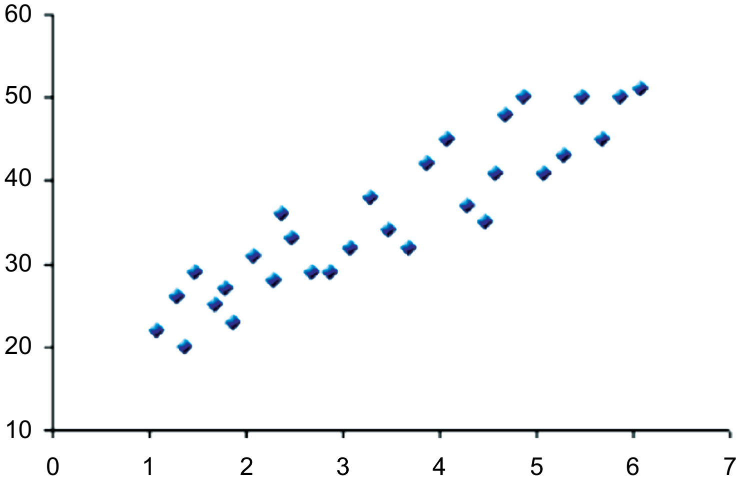


Fig. (1). Correlation scatters plot. [Source: Wikimedia Commons, By Jsmura - Own work, CC BY-SA 4.0].

REGRESSION

Regression is the relationship between the independent and dependent variables to each other. In simple linear regression, this relationship is the relationship of X to Y and is described as the equation of the line Y = a +bX . Regression attempts to explain how X causes Y to change and will generate different results if X and Y are reversed [62]. Regression assumes that X is a fixed value with no error. As regression analysis generates a linear equation, this analysis can be used for prediction or optimization and used to make assumptions about similar data sets. Regression can be described as how one variable affects another, or how changes in a variable trigger change in another variable, essentially cause and effect [66]. We can use a simple example of agriculture to understand regression. If there is ample rainfall, then seeds planted in a field will grow. If there is a drought and no rain falls, then the seeds will not grow. Regression analysis is used to determine the functional relationship between two variables, X and Y to make predictions about an unknown variable Z [62]. The value of this unknown variable Z can be estimated based on the values of the fixed variables X and Y. For simple linear regression, the best fit of the line through the data points is used to generate an equation that can then be used with the unknown Z variable [67]. A regression plot takes the correlation plot one step further by fitting a line through the scattered data points. An example regression is shown in Fig. (**2**).

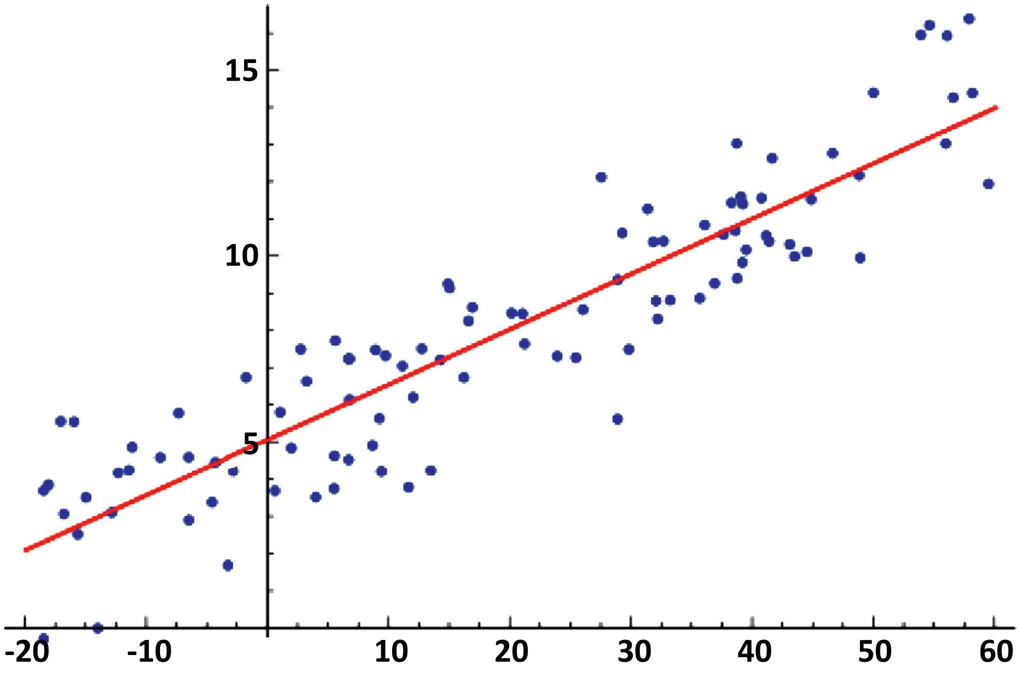


Fig. (2). Random data points and their linear regression. [Source: Wikimedia Commons Created with the following Sage (http://sagemath.org) commands: X = RealDistribution('uniform', [-20, 60]) Y = RealDistribution('gaussian', 1.5) f(x) = 3*x/20 + 5 xvals].

CONCLUSION

In conclusion, linear regression and correlation are one of the simplest and most common ways to show the relationship between independent and dependent variables to one another. It is a very powerful tool in statistics. Linear regressions help researchers to visualize these relationships and thereby assist in understanding the strength of similarities or differences within a dataset. By using these tools, researchers are able to better elucidate the outcomes of the inclusion of new variables as well as the relationships between relevant datasets.

CHAPTER 10

Ethics

Abstract: Ethical behavior by researchers, analysts and statisticians is paramount to the creation and reporting of data in all fields. Without ethics, data obtained and analyzed cannot be relied upon. A short discussion of ethical standards for all involved in data handling, research and statistical analysis is included, as this important aspect of the field cannot be understated.

Keywords: Ethics, Integrity, Standards.

INTRODUCTION

An overview of common statistics would be incomplete without the inclusion of a discussion on ethics within the field. The field of statistics must also include ethics as a way to govern both the integrity of researchers and the data set that those researchers analyze and handle. Grounding in ethics allows for confidence in reported results as well as integrity for future works and analysis of those experiments and investigations. The ethical standards associated with the field of statistics are laid out by the American Statistical Association.

ETHICS

Statistics is the science of uncertainty and variation, but the paradox is that in scientific research, statistics are used to justify claims or solidify determinations. Statistics are meant to learn from data, and gather insights about future trends and plan further investigations based on the current datasets. There are necessary considerations statisticians must wrestle with in order to effectively communicate the uncertainty and variation in their analyses [68]. To assist researchers with these considerations, the American Statistical Association has set out eight guidelines for ethical practice. First, the guideline of professional integrity and accountability. For statisticians to adhere to this guideline, they agree to use the methodology that is relevant and appropriate to produce valid, interpretable and reproducible results [69]. The statistician also agrees to accept responsibility for their professional work, credit others whose work is used and not perform work beyond their scope of expertise [71]. The second guideline follows from the first

Alandra Kahl

and specifies that the statistician is candid and upfront about any deficiencies, limitations or biases in the data [72]. This includes acknowledgement of assumptions, outlining methods to ensure the integrity of the dataset, and addressing potential confounding variables that are not present in the study [69]. The ethical statistician also has a responsibility to the funder, client and scientific community as outlined in the third guideline [70]. This means that the statistician strives to make new knowledge available to the scientific community, applies analyses scientifically, and protects the use and disclose of data appropriately [69]. The fourth guideline of responsibility to the research subjects follows directly from the previous guideline. As part of this guideline, the statistician protects and preserves the rights and data of all studied subjects to the best of their ability as well as providing participants with all appropriate disclosures, releases and research approvals [69]. The fifth guideline addresses the responsibility of the statistician to their research colleagues by promoting transparency in study design and analysis as well as effectively communicating and reviewing findings with other researchers [69]. The sixth guideline addresses the responsibility to other statisticians or statistics practitioners. Out of respect for other statisticians, the leader of the study agrees to treat others with respect and focus on scientific principles, methodology and the substance of data interpretations [69]. The seventh guideline notes the responsibility of the statistician regarding scientific misconduct. The ethical statistician avoids misconduct and does not condone questionable practices with regard to scientific, professional or statistical misconduct [69]. The final guideline for ethical statisticians addresses the responsibility of employers to statisticians. This guideline says that those employing statisticians agree to respect their professional expertise, support sound analysis and promote a safe and ethical working environment [71]. By adhering to these eight guidelines, statisticians ensure that they are acting in an ethical manner and best representing the science of statisticians and their professional colleagues in the discipline.

CONCLUSION

Ethics in statistical analysis are paramount. Researchers and statisticians must show integrity in their work and analyses. It is the gravest responsibility of the person doing statistics to upload the ethical standards and practices of the profession as well as to address any potential conflicts of interest or biases within their analysis. Upholding ethical standards is important for the public trust of data analysts with respect to their professional reputation of themselves as well as their colleagues.

REFERENCES

[1] D. Montgomery, G. Runger, and N. Hubele, "Engineering Statistics, 5th Edition", 2010. Available at:https://www.wiley.com/en-us/Engineering+Statistics%2C+5th+Edition-p-9780470631478

[2] Population *vs* Sample Data – Math Bits Notebook (A1 - CCSS Math),” Available at:, https://mathbitsnotebook.com/Algebra1/StatisticsData/STPopSample.html

[3] Sample Dataset - an overview | ScienceDirect Topics, Available at:, https://www.sciencedirect.com/topics/computer-science/sample-dataset

[4] L. Daniels, and N. Minot, "An Introduction to statistics and data analysis using stata®: from research design to final report", *Sage Publications 2019.*

[5] R.A. Johnson, and G.K. Bhattacharyya, "Statistics: principles and methods", *John Wiley & Sons, 2019.*

[6] J.E. Kolassa, "An introduction to nonparametric statistics", *Chapman and Hall/CRC, 2020.*

[7] C. Chatfield, "Statistics for technology: a course in applied statistics", *Routledge, 2018.*

[8] S.M. Ross, "Introduction to probability and statistics for engineers and scientists", *Academic press 2020.*

[9] S. Manikandan, "Frequency distribution", *J. Pharmacol. Pharmacother. 2011,* vol. 2, no. 1, pp. 54-56.

[10] “Pie Charts, Histograms, and Other Graphs Used in Statistics, Available at: https://www.thoughtco.com/frequently-used-statistics-graphs-4158380

[11] D. Russell, *What Is a Bar Graph?*. https://www.thoughtco.com/definition-of-bar-graph-231238

[12] “A Complete Guide to Pie Charts,”, https://chartio.com/learn/charts/pie-chart-complete-guide/

[13] 13.2 - Stem-and-Leaf Plots | STAT 414, Available at: https://online.stat.psu.edu/stat414/lesson/13/13.2

[14] Statistics: Power from Data! Organizing data: Stem and leaf plots, Available at: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch8/5214816-eng.htm

[15] “IXL | Dot plots,” Available at:, https://www.ixl.com/math/lessons/dot-plots

[16] S. Foster, “The Federal Reserve’s Dot Plot Explained – And What It Says About Interest Rates,” Available at:, https://www.bankrate.com/banking/federal-reserve/federal-reserve-dot-plot-expl-ined-how-to-read-interest-rates/

[17] “A Complete Guide to Scatter Plots,” Available at:, https://chartio.com/learn/charts/what-is-a-scat-er-plot/

[18] “Create and use a time series graph—ArcGIS Insights | Documentation,” Available at:, https://doc.arcgis.com/en/insights/latest/create/time-series.htm

[19] E.G.E. Kyonka, S.H. Mitchell, and L.A. Bizo, "Beyond inference by eye: Statistical and graphing practices in JEAB, 1992-2017", *J. Exp. Anal. Behav.,* vol. 111, no. 2, pp. 155-165. *2019.*

[20] N. Barnett, "Graphing causation: getting a clearer picture or fuzzy logic? Comment on Br J Anaesth 2020: 125: 393–97", *Br. J. Anaesth.,* vol. 126, no. 3, pp. e100-e101.

[21] D.M. Finkelstein, and D.A. Schoenfeld, "Graphing the win ratio and its components over time", *Stat. Med. 2019,* vol. 38, no. 1, pp. 53-61.

[22] Ungrouped Data, "Introduction to Statistics 2019",

[23] C.E.L. Kinney, J.C. Begeny, S.A. Stage, S. Patterson, A. Johnson, "Three alternatives for graphing behavioral data: A comparison of usability and acceptability", Behavior Modification 46.1, 2022: 3-35.

Alandra Kahl

[24] M. Malloy, J. Koller, and A. Cahn, "Graphing crumbling cookies", *ACM, 2019*.

[25] J.O. Aldrich, "Using IBM SPSS statistics: An interactive hands-on approach", *Sage Publications 2018*.

[26] L. Sullivan, *The role of probability.,* .https://sphweb.bumc.bu.edu

[27] F.V. Kuhlmann, *A simple explanation of probability. Licensed under Creative Commons by-nc-sa, 2021*.

[28] A. Schmitz, *Basic Concepts of Probability. Licensed under Creative Commons by-nc-sa 3.0*. 2013, Available at: https://2012books.lardbucket.org/books/beginning-statistics/s07-basic-concept--of-probability.html

[29] An Introduction to Probability and Statistics (Wiley Series in Probability and Statistics). New York, 2021.

[30] W. Kenton, *"Random Variable." Investopedia: published by Dot dash Meredith Group. 2022*. Available at: https://www.investopedia.com/terms/r/random-variable.asp

[31] A. Schmitz, *"Discrete Random Variables." Saylor Academy Texts, 2021. Available at:*.https://saylordotorg.github.io/text_introductory-statistics/s08-discrete-random-variables.html

[32] J.M. Russell, *"Introduction to Discrete Random Variables and Notation." Pressbooks, Licensed under Creative Commons cc-by-sa 4.0*. 2022.

[33] A. Barone, *"Binomial Distribution." Investopedia: published by Dot dash Meredith Group. 2021*. Available at: https://www.investopedia.com/terms/b/binomialdistribution.asp

[34] C.F.I. Team, *Binomial Distribution. 2022*. https://corporatefinanceinstitute.com/resources/knowledge/other/ binomial-distribution/

[35] H.B. Berman, *Negative Binomial Distribution*. 2022, Available at: https://stattrek.com/probability-distributions/negative-binomial

[36] M. Taboga, *"Continuous random variable", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing, 2021. Available at:*.https://www.statlect.com/glossary/continuous-random-variable

[37] M. Taboga, Absolutely Continuous random variable", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. 2021, Available at: https://www.statlect.com/glossary/absolutely-continuous-random-variable

[38] Maple Tech International. Continuous Random Variable., Available at: https://www.math.net/continuous-random-variable

[39] J. Chen, Normal Distribution, Available at: https://www.investopedia.com/terms/n/normaldistribution.asp

[40] M. Taboga, "Central Limit Theorem", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. Online appendix, 2021, https://www.statlect.com/asymptotic-theory/centra--limit-theorem

[41] P. Bhandari, The Standard Normal Distribution, Available at: https://www.scribbr.com/statistics/standard-normal-distribution/#:~:text=The%20standard%20normal %20distribution%2C%20also, the%20mean%20each%20value%20lies

[42] Saylor Academy. Areas of tails of distributions, Available at: https://saylordotorg.github.io/text_introductory-statistics/s09-04-areas-of-tails-of-distribution.html

[43] Corporate Finance Institute, "Sampling Distribution", Available at: https://corporatefinance institute.com/resources/ knowledge/other/sampling-distribution/

[44] A. Schmitz, Beginning Statistics. Sampling Distributions, Available at: https://2012books.lardbucket.org/books/beginning-statistics/s10-sampling-distributions.html

[45] Penn State Open Education. Sampling Distribution of the Sample Mean, Available at: https://online.stat.psu.edu/stat500/lesson/4/4.1

[46] Lani, James. Statistics Solutions. “Estimation.”, Available at: https://www.statisticssolutions.com/estimation/

[47] D. Lane, *Introduction to Estimation. September 17, 2013. Provided by: OpenStax CNX. Available at:.* https://cnx.org/contents/5530cbcc-820d-4f48-83c7-fe03ef5823be@4 Available at:, https://courses.lumenlearning.com/boundless-statistics/chapter/estimation/

[48] H.B. Berman, *Estimation in Statistics*. https://stattrek.com/estimation/estimation-in-statistics.aspx

[49] H.B. Berman, *What is a Confidence Interval*. https://stattrek.com/estimation/confidence-interval https://stattrek.com/estimation/confidence-interval.aspx?tutorial=AP

[50] G. Stephanie. "Estimator: Simple Definition and Examples" From StatisticsHowTo.com: Elementary Statistics for the rest of us!, Available at: https://www.statisticshowto.com/estimator/

[51] H.B. Berman, *What is the Standard Error?*.https://stattrek.com/estimation/standard-error.aspx?tutorial=AP

[52] H.B. Berman, *Margin of Error*. https://stattrek.com/estimation/margin-of-error.aspx?tutorial=AP

[53] S. Andy. Beginning Statistics. “ Estimation.”, Available at: https://2012books.lardbucket.org/books/beginning-statistics/s11-estimation.html

[54] Frost, Jim. Statistics by Jim. “One-Tailed and Two-Tailed Hypothesis Tests Explained,”, Available at: https://statisticsbyjim.com/hypothesis-testing/

[55] Frost, Jim. Statistics by Jim. “Null hypothesis,” Available at: , https://statisticsbyjim.com/?s=null+hypothesis

[56] Frost, Jim. Statistics by Jim. “Hypothesis Tests Explained,” Available at:, https://statisticsbyjim.com/hypothesis-testing/

[57] J. Chen, Z-Test Definition: Its Uses in Statistics Simply Explained With Example, https://www.investopedia.com/terms/z/z-test.asp

[58] S. Mina. Analytics Vidya. “Hypothesis Testing | Difference between Z-Test and T-Test” Available at:, https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test

[59] R.C. Sprinthall, *(2011). Basic Statistical Analysis (9th ed.). Pearson Education. “Z-test,” Available at:* .https://en.wikipedia.org/wiki/Z-test

[60] Stats Test Team. “Paired Samples Z-Test,” Available at:, https://www.statstest.com/paired-samples-z-test/

[61] https://www.g2.com/articles/correlation-vs-regression

[62] B. Gerstman, *Correlation and Regression*.https://www2.sjsu.edu/faculty/gerstman/StatPrimer/cont-cont.htm

[63] J.M. Bland, and D.G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement", *Lancet,* vol. i, pp. 307-310. 1986.

[64] V. Bewick, L. Cheek, and J. Ball, "Statistics review 7: Correlation and regression", *Crit. Care,* vol. 7, no. 6, pp. 451-459. 2003.[http://dx.doi.org/10.1186/cc2401]

[65] P. Vadapalli, *Correlation vs Regression: Difference Between Correlation and Regression*. https://www.upgrad.com/blog/correlation-vs-regression/

[66] K. Kozak, *Regression and Correlation.*. https://www.coconino.edu/resources/files/pdfs/academics/sabbatical-reports/kate-kozak/chapter_10.pdf

[67] S. Crawford, "Circulation. 2006;114: 2083–2088", [http://dx.doi.org/10.1161/circulationaha.105.586495]

[68] A. Gelman, "Ethics in statistical practice and communication: Five recommendations", *Significance,* vol. 15, no. 5, pp. 40-43. 2018.

[69] “Ethical Guidelines for Statistical Practice,” https://www.amstat.org/ASA/Your-Career/Ethica--Guidelines-for-Statistical-Practice.aspx

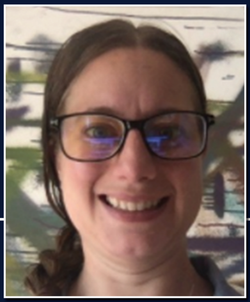
[70] L.M. Lesser, and E. Nordenhaug, "Ethical Statistics and Statistical Ethics: Making an Interdisciplinary Module", *J. Stat. Educ.,* vol. 12, p. 3. 2004. [http://dx.doi.org/10.1080/10691898.2004.11910630]

[71] J.S. Gardenier, "Making Statistical Ethics Work for You, of Course", *Amstat News,* no. 296, pp. 21-22. 2002.

[72] L.M. Lesser, *2001 Proceedings of the American Statistical Association Section on Statistical Education [CD-ROM], Alexandria, VA: American Statistical Association, 2001.*

SUBJECT INDEX

Alandra Kahl



Alandra Kahl

Dr. Alandra Kahl currently teaches engineering design and sustainable systems at The Pennsylvania State University, Greater Allegheny campus. She received her doctorate in environmental engineering from the University of Arizona in 2013, where her dissertation was focused on the fate and transport of contaminants of emerging concern in an arid region. Dr. Kahl's research interests include engineering of sustainable systems, treatment of emerging contaminants via natural systems and engineering education. She is the author of several technical papers and conference proceedings centered on environmental engineering. Her professional affiliations include the Society of Toxicology and Chemistry, the American Chemical Society, and the American Society for Engineering Education.