STATISTICS THROUGH AN EQUITY LENS (ANTHONY)

Yvonne Anthony
Massachusetts Bay Community College

LibreTexts™

Statistics Through an Equity Lens (Anthony)

Yvonne Anthony

Massachusetts Bay Community College

TABLE OF CONTENTS

Licensing

A detailed breakdown of this resource's licensing can be found in Back Matter/Detailed Licensing.

Introduction

1

This Open Educational Resource (OER) carries a significant responsibility by presenting statistics through an *equity lens*. The metaphor of a lens is used intentionally–as the glasses one wears can have a profound effect on what one sees. "Your eyes come into contact with the world—but what do they really see? Your perception is affected by the outer environment and by the ecology of your inner world."[1]

As such, the book encourages further inspection of the ways in which data is collected, interpreted, and analyzed on a variety of social justice issues, such as *health disparities, hunger, and food insecurity, homelessness, behavioral health (mental health and substance use), and incarceration of males of color*. The book endeavors to heighten awareness of how data can close disparities for marginalized or underserved communities. It also attempts to reveal how the misuse of data can reinforce inequities, for example, by stigmatizing people and labeling neighborhoods as high poverty, violent, and having poor educational opportunities. Whether an intended or unintended consequence, irresponsible data use can contribute to racist impressions of people and communities.

Whether you are a student taking this Introductory Statistics course, a seasoned statistician, or a policymaker working for a state or federal government, it is our *dharma* to use and manage data responsibly and ethically.[2] Adopting an "equity talk or walk" means that you critically examine data through a lens that questions *how and why* inequalities exist for those who have been historically and continuously marginalized in society, and perhaps how to help envision and construct a more equitable future.

This OER book is intended to get you started (or continue) on your "equity-mindfulness" journey. As you go through this journey, I encourage you to think critically about diversity-related issues, power, and oppression and gain an understanding of people from racial/ethnic groups different from your own. Hopefully, you will reflect on your own position of power in society.

I had a Black Father and a Filipino Mother and grew up culturally Black in the low-income neighborhoods of North Philadelphia. I attended Catholic private schools (K-12) and was a National Merit Commended student in the 12th grade. My entire post-secondary education has been fully funded through merit scholarships and fellowships, stipends for living expenses/textbooks, and PhD Dissertation Awards from highly selective colleges/universities. Still, I have never forgotten my roots and where my life all started in a marginalized, under-resourced community rife with social, economic, and political disadvantages. This social location and positionality have given me both knowledge and courage to apply an equity lens to the discipline of Statistics, which is the intention of this OER.

The equity framework within mathematics education is quite narrow (Gutstein 2006). Applying an equity lens to statistics has been rewarding personally as a researcher yet disheartening as a person of color, given the stubborn and persistent disparities that continue to exist in certain sectors of our society. However, if the book has enlightened you about how knowledge of the concepts and practices of statistics can help to better understand social justice issues, then its purpose has been met. When viewed through an equity lens, statistics has such a humanitarian and compassionate side to it that hopefully opens your heart to a greater understanding of socially disadvantaged populations and communities.

Warm regards,
Dr. Yvonne E. Anthony, Author

---

1. Chidvilasananda, S (1996). *The Yoga of Discipline*. New York: A Siddha Yoga Publication. 05: Introduction#return-footnote-4-1
2. [Dharma is regarded in Hinduism as a cosmic or universal law underlying right behavior and social order] 05: Introduction#return-footnote-4-2

CHAPTER OVERVIEW

1: Chapters

This page titled 1: Chapters is shared under a not declared license and was authored, remixed, and/or curated by Yvonne Anthony (Remixing Open Textbooks with an Equity Lens (ROTEL)) .

1.1: An Overview of Statistics

Welcome to your first destination along the equity-mindfulness journey!

This first destination was designed to set the tone for you as a learning community member—introducing you to statistics by "data diving" into a real-world social justice issue in health care, specifically, the public health sector. You will be introduced to fundamental concepts, definitions, and terminology of statistics along your journey—still with equity embedded in each step and not a secondary consideration.

Before we start "data diving", I want to share that statisticians have the coolest job! We can contribute to society in so many ways —from protecting endangered species and managing the impacts of climate change to making medicines more effective and reducing hunger and disease. This is because **statistics is a science that involves asking questions about the world and finding answers to them in a scientific way.**

Now, we are ready to begin, so please put on your metaphorical equity glasses as we discuss the following topics:

> **Learning Objectives**
>
> - The Statistics-Data Alignment
> - Data Equity
> - Health Equity Framework
> - Social Determinants of Health
> - The World of Statistics: Basic Concepts and Terminologies
> - Population (Parameter) vs. Sample (Statistics)
> - Two Major Branches of Statistics: Descriptive and Inferential
> - Data Classification: Quantitative and Qualitative Data
> - Quantitative Variables in Community-Based Research
> - Levels of Measurement
> - Designing a Statistical Study

1.1 The Statistics-Data Alignment

Statistics is a branch of applied mathematics that involves the collection, description, and analysis of quantitative data. The relationship between data and statistics is a symbiotic one. Understanding data is impossible without statistics. Statistics presents a rigorous scientific method for gaining insight into data. Data and statistics are used side-by-side in this OER to better understand a variety of social justice issues, such as *health disparities, incarceration, hunger, and food insecurity, income inequality (gender pay gaps and race), poverty, gun violence, racial injustice, gender identity/expression, gender-based violence, global climate change, reproductive rights, disparities in maternal and infant health, violence and discrimination experienced by LGBTQ+ people, and gaps in cultural wealth.*

Social Justice
/soh-shuhl juhs-tis / noun

Social Justice refers to the view that everyone deserves equal rights, opportunities, and treatment regardless of their race, economic status, sexuality, or gender identity.

Photo Credit: Unsplash Tim Mossholder

1.2. Data Equity at the Core of Health Outcomes

Statistical concepts and practices are best understood when applied to data. Even before this, we must think critically about how data is collected, analyzed, and interpreted for the betterment of an organization, community, or group of people. This is where **data equity** comes into play. According to Moser Technology, "Data equity frameworks apply an equity-centered lens and mindset to ensure data is collected, analyzed, interpreted, and shared with diverse stakeholders without bias or exclusion." For example, if clinical trials on new medications are only performed on white males, females, and intersex, the final product might not prove to be effective for other racial and ethnic groups and could actually produce harmful side effects.
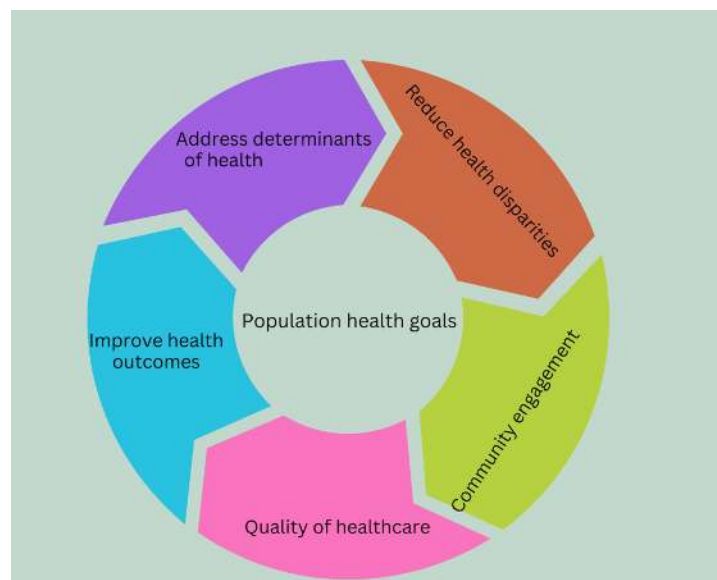
A **population health equity framework** aims to achieve optimal health for all by targeting social and structural determinants of health (Trinh-Shevrin et al. 2015)[1]. This framework is population-based, focusing on the entire *population* within the United States achieving optimal health. In theory, it moves towards the vision of improving total population health and reducing health inequities in underserved communities of color. However, in actuality, since strategies and interventions in research are conducted

on majority dominant populations—largely white and middle class—they have made little impact on eliminating health disparities on a community level. In fact, gaps in health disparities have widened for people of color. *People of color* is a term used to refer to African Americans, American Indians/Alaskan Natives, Asian Americans, Latinos/Hispanics, and Native Hawaiians/Other Pacific Islanders. *Systemic* and *structural racism* consists of the systems and structures of a society that expose people of color to health-harming conditions and that impose and sustain barriers to opportunities that promote good health and well-being (Braveman et al. 2022)[2].

Contrasting the term population is **sample**—a subset, or part, of the population which is the whole (see Figure 1 below). An example is affluent suburban communities achieving optimal health outcomes.

Figure 1: Sample vs. Population

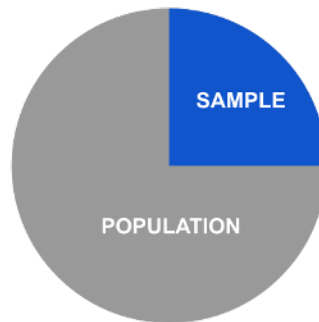A parameter is a numerical description of a population statistic. A statistic is a numerical description of a sample characteristic.

Data Equity Example #1:
The City of Boston

According to the Boston Planning and Development Agency[3], in 2022, the total population of the city of Boston is 689,326, with a median household income of $76,298 (a parameter). The city of Boston comprises many Census Block Groups and Neighborhoods that can be described as subsets of the total population of Boston. Those neighborhoods with higher proportions of ethnic/racial groups are presented with their total populations respective median household incomes[4]: Chinatown (pop. 7,143; $32,735) Dorchester (pop.122,191; $59,379), Mattapan (pop.23,834; $58,633), Mission Hill (pop.17,886; $45,392), and Roxbury (pop.54,905; $33,322). Since these neighborhoods are subsets of the city of Boston, their median household incomes are considered a statistic, ranging from $32,735 to $59,379. This range is well below the median household income of $76,298 for the city of Boston's total number of residents.

Now, let's examine neighborhoods that have lower proportions of ethnic/racial groups and higher proportions of the White population. What are their median household incomes? A sample of these neighborhoods include Back Bay (pop.19,588; $111,141), Beacon Hill (pop.9,336; $116,505), East Boston (pop.43,066; $63,721), and South Boston (pop.37,917; $122,635). Their range of median household incomes is $63,721 to $122,635 and, for the most part, are higher than the incomes of Boston's residents in total.

Discussion Questions: *Wearing your equity lens as a statistician, what associations or correlations do you see in this example? What other data would you like to see to tell the story?*

1.3. The Social Determinants of Health Approach

Different groups of people can have markedly different levels of health. The **Social Determinants of Health** approach asserts that the conditions in which people live, work, and play are primary drivers of one's health status. Thus, one's socioeconomic position–along with one's ability to access housing, transportation, political environment, cultural beliefs, and norms, as well as experience with racism and discrimination—are factors that influence the health of a population or a community neighborhood. A person's social determinants of health can contribute more to one's health than genetic code or medical care.

As part of its Health Equity Framework, the Boston Public Health Commission (BPHC) includes several factors when applying the Social Determinants of Health approach to city residents. These factors fall within three categories: *economic, environmental, and social*.

- Access to healthcare
- Access to health resources
- Access to healthy food
- Education
- Employment and occupational safety
- Environmental safety
- Exposure to violence
- Housing conditions
- Income
- Insurance coverage
- Racism and discrimination
- Transportation

BHPC assessment concludes that "many health-promoting resources, such as income, employment, education, and homeownership, are unevenly distributed within our city among those of differing races and ethnicities, socioeconomic statuses, and geographic locations."[5]

Historically speaking, zip code data is the most widely used geographic data to help understand the needs of a population. The assumption is that it truly represents the community or neighborhood of interest. According to Health IT Analytics, "geographic data is most useful for identifying hot spot areas where the population is at high risk for contracting a disease or in high need of interventions to minimize disease impact. These are areas where the average income might be well below the federal poverty line."

Cultural knowledge developed through family (immediate, extended, or chosen)

Familial Capital

Maintaining hopes and dreams in the face of barriers

Aspirational Capital

Networks within your community that can provide resources and support

Social Capital

Cultural Capital

Community Cultural Wealth

Linguistic Capital

Intellectual & social skills attained through communication in more than one language

Resistant Capital

Knowledge & skills developed through challenging inequality

Navigational Capital

Skills of maneuvering through institutions (I.e. college)

Data Equity Example #2:
Social Determinants of Health (Zip Codes)

For the city of Boston, residents living in zip codes representing the neighborhoods of *Roxbury* (02119), *Mission Hill* (02120), *Dorchester* (02121, 02124, 02125), and *Mattapan* (02126) have lower median household incomes compared to Boston overall[6]. They also have higher percentages of people living below the poverty line. Moreover, according to 2015 BPHC data, a higher percentage of Boston adult residents with incomes less than $25,000 have asthma, diabetes, hypertension, obesity, persistent anxiety, and persistent sadness compared with those with a household income of $50,000 or more.

Discussion Questions: *Wearing your equity lens as a statistician, what associations or correlations do you see in this example? What other data would you like to see to tell the story?*

1.4.(a) The World of Statistics: Basic Concepts and Terminology

Sta-tis-tics
[stuh-tis-tiks], *noun*

1. the only science wherein two recognized experts, using exactly the same set of data, may come to completely opposite conclusions

1.4(a). Definition of Statistics

Welcome to the beautiful world of statistics! We live in a world of information where much of it is determined mathematically with the help of statistics. Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions. As shown in the previous sections, there are two types of data sets you will use when studying statistics—population and samples. In statistics, population is the entire set of items (individuals, organizations, events, widgets, etc.) about whom you wish to draw conclusions. A sample is used to gain information about a population. An example of a population is the entire student body at a community college in Massachusetts. A sample is students who take an Introductory Statistics course in the evenings.

Samples are used by statisticians when the population is large and scattered or it is difficult to collect data at this level. Samples

should be randomly selected and represent the entire population and every group within it. Statistical methods are used to collect random samples to reduce sampling bias and increase validity when answering research questions. Drawing inferences from samples to populations is referred to as decision-making in the data analysis stage of social research.[7]

A parameter is a numerical description of a population statistic. It is any summary number (e.g., the mean or percentage) that describes a population. A statistic is a numerical description of a sample characteristic. For example, the median household income in the United States is a population parameter. Conversely, the median household income for a sample drawn from the United States, such as the city of Philadelphia in Pennsylvania, is a sample statistic.

Study Tip: To remember the terms parameter and statistics, match the first letters of *population parameter* and the first letters in *sample statistics.*

1.4(b). Two Major Branches of Statistics

Descriptive statistics summarizes the characteristics of a data set. **Inferential statistics** uses a sample to draw conclusions about the population. You test a hypothesis or assess whether the data is generalizable to the population. Know that you can never be 100% sure about inferences. A major theme of this OER is how to use sample statistics to make inferences about unknown population parameters.

Now Try It Yourself

1.4(c). Data Classification: Quantitative and Qualitative Variables

A prerequisite for exploring data more deeply is coming to grips with the different types of data you can encounter. A research study can consist of two types of data: **quantitative** and **qualitative**. Quantitative data is data about numerical variables (e.g., how many, how much, or how often) that are measurements or counts. Qualitative data is data about non-numerical variables (e.g., what type) that are attributes or labels.

Quantitative Variables in Community-Based Research

Variables are elements, entities, or factors that can change under *different* conditions or circumstances for *different* people. Statisticians are often interested in **demographic variables** when trying to understand the characteristics or attributes of a community or neighborhood. Examples are:

- Population Size
- Age
- Race/Ethnicity
- Gender
- Educational Attainment
- Marital Status
- Average or Median Income
- Employment Status
- Home Ownership vs. Renting
- Access to Public Transportation
- Access to Healthy and Affordable Food

1.4(d). Levels of Measurement

Another characteristic of data is its level of measurement. It is the first or perhaps the most important piece of descriptive information about a research variable. There are four hierarchical levels of measurement (lowest to highest): **nominal** (qualitative data**), ordinal** (qualitative or quantitative data), **interval** (quantitative), and **ratio** (quantitative).

Levels of Measurement
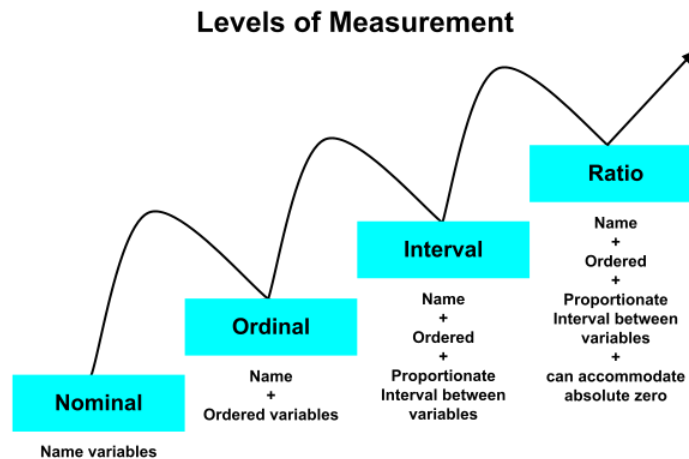
The word "nominal" means name. **A nominal level of measurement** names the attribute, characteristic, or identity we are interested in, has no numerical value and is not rank ordered. Examples of nominal variables in social science research are race, gender, marital status, religious affiliation, and voting behavior. Measurement of the quality of pain is a nominal variable. Pain might be throbbing, constant, dull, sharp, stinging, achy, or burning.[8] It may surprise you that zip code is a nominal variable. A **categorical variable** (sometimes called a nominal variable) is one that has two or more categories, but there is no intrinsic ordering to the categories. For example, there is no implied order or hierarchy to the categories "Hispanic or Latino" and "Not Hispanic or Latino".

An ordinal level of measurement includes a rank ordering of variable values, such as being greater than or less than, making them more complex than nominal variables. In ordinal measurement, the numbers represent categories, but they function much more like labels. Likert scales are often used to represent ordinal data: *strongly disagree, disagree, neither agree nor disagree, agree, strongly agree*. Ordinal scales do not have equal intervals. Thus, the difference between 1 and 2 may not be the same as that between 3 and 4. Socioeconomic standing is an ordinal variable: *lower, middle, and upper class.* Another example is the performance of a government policy on a social justice issue: *superior, effective, minimal, or inadequate.* Qualitative or quantitative data can have an ordinal level of measurement.

The interval level of measurement is continuous, can be rank ordered, exhaustive (all possible attributes are listed), and mutually exclusive (a person cannot identify with two different attributes simultaneously). The intervals are equal; there is no absolute zero. The attributes are numbers rather than categories. IQ scores are interval level, and so is temperature. Only quantitative data can have an interval level of measurement.

The **ratio level of measurement** is the highest of the four hierarchical levels of measurement. Quantitative data, such as age, income, unemployment rate, the rate of infant mortality in a particular country, or recidivism rate (reoffending and reentering the prison system) are examples of ratio variables. Unlike interval data, a distinguishing part of ratio data is that it has a "true zero". This basically means that zero is an absolute with no meaningful values below it, such as a negative number. Age is a good example—you cannot be -25 years old.

Oftentimes, ratio data is the most desirable type of data since it can perform the widest possible range of analyses, improving our ability to test hypotheses with more accurate insights. Many variables in the social sciences have ratio scales. Like interval variables, ratio-scaled variables can be **discrete** (expressed only in countable numbers) or **continuous** (can potentially take on an infinite number of values. You count discrete variables, and the results are integers, such as the number of residents in a neighborhood who are immigrants. On the other hand, if you measure continuous variables—they can take on any numeric value, including fractional and decimal values. Both types of variables are essential in statistics.

Now Try It Yourself

1.5 Designing a Statistical Study
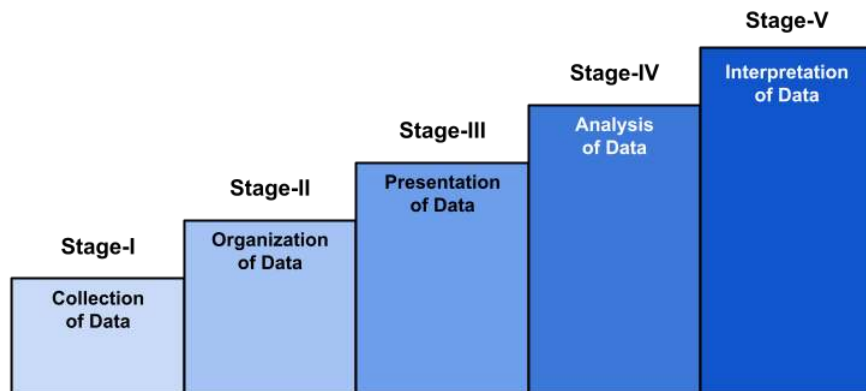
1.5(a) Purpose of the Study

A statistical study has purpose, scope, and direction, namely, why the study is being conducted. The goal of the study might be to describe a concept, predict a situation, or examine the magnitude (strength) and direction (positive or negative/inverse) of the relationship between certain variables (e.g., race/ethnicity and health outcomes). The purpose statement identifies the variables of the study. For example, the purpose of this study is to "examine the characteristics and attributes of single mothers who are first-generation college-bound, their study habits, and work commitments in relation to quality time spent with their child(ren)."

After the variables of interest have been identified, Larson & Farber (2019)[9] suggest five sequential steps to be followed:

- Develop a detailed plan for collecting data. Make sure the sample is representative of the population.
- Collect the data.
- Describe the data using descriptive statistical techniques.
- Interpret the data and make decisions about the population using inferential statistics.
- Identify any possible errors.

1.5(b) Choosing a Sample Size

When conducting a study, choosing the correct sample size can have a huge impact on the results since sample data can help approximate findings about a population. The fundamental question is, "What sample size do I need for successful inference?" The steps for calculating a sample size are:

1. Determine the population size, which is the total number of the target demographic.
2. Decide on a margin of error. This is the difference you will allow between the sample mean (or average) and the mean of the total population.
3. Choose a confidence level that indicates how assured you are that the actual mean will fall within your chosen margin of error. Most statisticians choose confidence levels that are 90%, 95%, or 99% confident.
4. Pick a standard of deviation or the level of variance you are expecting in the information gathered. Choosing 0.5 is typically a safe choice that will ensure a large enough sample.
5. Complete the calculation. The sample size formula helps calculate or determine the minimum sample size.

Sample Size $n = \frac{N*[Z2 * p * \frac{(1-p)}{e2}]]}{N-1(Z2 * p * \frac{(1-p)}{e2}]]}$ where,

- N = Population Size
- Z = Critical Value of the Normal Distribution at the Required Confidence Level
- p = Sample Proportion
- e = Margin of Error

1.5(c) Sampling Techniques

Appropriate sampling techniques are used to ensure that inferences about the population are valid. A sampling design describes exactly how to choose a sample from a population. Probability sampling is a sampling technique which chooses samples from a

larger population using a method based on the theory of probability. For example, in a population of 500 members, every member will have 1/500 chance of being selected to be a part of the sample.

The most important condition for sound conclusions from statistical inference is that the data is a random sample from the population of interest. A **simple random sample** is a likelihood of being selected. Consider a study of the number of people who live in the poor and working-class urban community of North Philadelphia (Philly), Pennsylvania.[10]
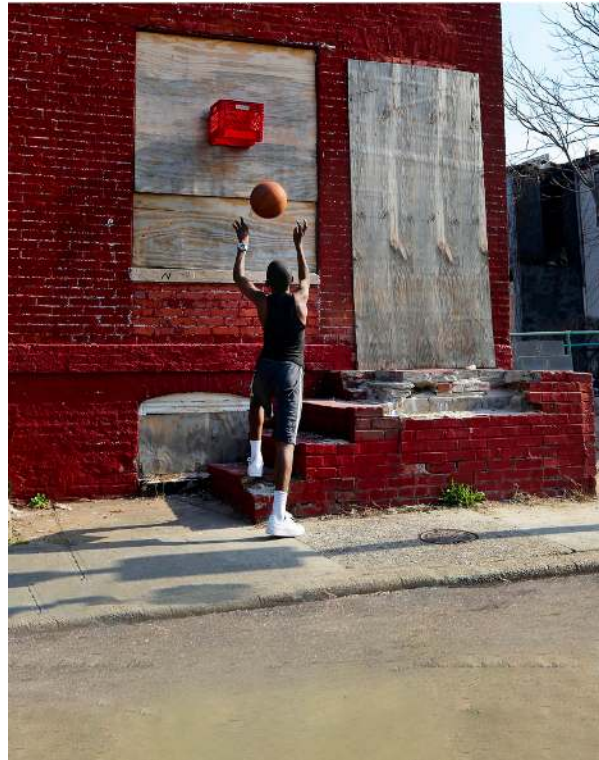
To use a simple random sample to count the number of people who live in North Philly's households, you can assign a different number to each household, use a table of random numbers to generate a sample of numbers, and then count the number of people living in each selected household.

When it is important for the sample to have members from each segment of the population, then you will use a **stratified sample.** It selects a *sample of members* from *all* strata. For instance, to make a stratified sample of the number of people who live in North Philly, you can divide the households into socioeconomic levels and then randomly select households from each level. If 70% of the people in North Philly belong to the low-income group, then the proportion of the sample should have been 70% for this group.

In **systematic sampling,** each member of the population is assigned a number. To collect a systematic sampling in North Philly, a different number can be assigned to each household, randomly choose a starting number, select every 25th household, and count the number of people living in each household. **Cluster sampling** uses *all* members from a randomly selected sample of clusters (but not all, some clusters will not be part of the sample). For instance, to collect a cluster sample of the number of people who live in North Philly households, divide the households into groups according to zip codes. Then, select all of the households in one or more, but not all, zip codes and count the number of people living in each household.

Now Try It Yourself

The American Statistical Association (ASA) has a minority membership directory that is accessible to the public. The ASA defines "minority" as being of Spanish/Hispanic/Latino ethnicity or having a race other than white. You are a new Statistician who is interested in learning about members' experiences while working in a predominantly white profession.

Discussion Questions: *What questions would you ask? How would you go about implementing this study? What specific steps would you take in securing their opinions? What information (or assumptions) do you need to know (or make) to get started?*

> **Conclusion: Reflective Essay** Write an essay about the importance of statistics in learning about a particular social justice issue of your choice (e.g., Affordable Healthcare, Behavioral Health, Climate Change, Criminal Justice, Homelessness, LGBTQ+ Rights, Policing, Racial Equality, and Voting Rights).

Chapter 1: Summary

In this Chapter, you were introduced to several new concepts and terminologies fundamental to statistics. You learned that statistics, statistical practice, and data—all three can speak together on issues of social justice and equity. The compassion of statistics is best summed by MacGillivray (2021), "statistics is one of the most unselfish sciences" and "improves human welfare not by its own ends, but by its contributions in all fields."[11]

To continue your journey in learning statistics from a data-equity perspective, Chapter 2 will help you to acquire basic data literacy skills. Data literacy can be seen as competence in making sense of data, including the interpretation and presentation of data—numerically [quantitative] or non-numerically [qualitative]. By learning how to categorize and summarize data, we are in a better position to identify and answer critical questions concerning equity-minded issues: a global focus in the "ecosystem of policymaking".

From a technical perspective, in Chapter 2, you will learn ways to organize and describe data sets, such as through graphs and numeric measures. The following three measures will be discussed in Chapter 2:

- Measures of central tendency (mean, median, and mode).
- Measures of dispersion (range, interquartile range, variance, and standard deviation).
- Measures of position (decile, percentile, and quartile).

My personal viewpoint is that building and sustaining a healthy culture of inquiry, especially one that considers evidence of equity or inequity, is an attribute of a culturally responsive statistician.

Enjoy the journey with your favorite beverage!

---

1. Trinh-Shevrin C, Nadkarni S, Park R, Islam N, and Kwon S (2015). Defining an integrative approach for health promotion and disease prevention: A population health equity framework. *J Health Care Poor Underserved, 2015 May; 26 (2 0): 146-163*.
2. Braveman PA, Arkin E, Proctor D, Kauh T, & Holm N (2022). Systemic and Structural Racism: Definitions, Examples, Health Damages, and Approaches to Dismantling. *Health Affairs*, Vol 41, No. 2: Racism and Health.
3. Boston Planning and Redevelopment Authority (2022). Research Publications: Boston At a Glance 2022. Note: Based on the 2016-2020 American Community Survey.

4. Boston Planning and Redevelopment Authority. Research Publications: Boston in Context-Neighborhoods. Note: Based on 2020 Decennial Census Redistricting Data and the 2016-2020 American Community Survey.

5. Boston Public Health Commission (2022). Health of Boston Reports. Chapter 2: Social Determinants of Health, page 108.

6. Ibid, page 136.

7. Levin J & Fox JA (2006). Elementary Statistics in Social Research. Boston, MA: Pearson Education, Inc.

8. McHugh, ML (2003). Descriptive Statistics, Part I: Levels of Measurement. JSPN, Vol. 8, No. 1, January-March 2003.

9. Larson R & Farber B (2019). Elementary Statistics: Picturing the World. Boston, MA: Pearson Education, Inc.

10. The author was born and raised in North Philly which continues to be highly segregated with predominantly Black populations living to the west of Broad Street and predominantly Hispanic populations to the east of 5th avenue. Overall, the city of Philadelphia has a high poverty rate of 23% (2019-2020), and has been described as the poorest big city in America. The economic disruptions of the COVID pandemic have hurt people of color living in Philly in large proportions.

11. MacGillivray, H (2021). Editorial: Statistics and data science must speak together.

1.2: Descriptive Statistics

"If you never collect data, you will never know the specifics of a problem."

James Bell, W. Haywood Burns Institute

> **◕ Learning Objectives**
>
> - Statistics and Data Analytics
> - Descriptive Statistics
> - Inferential Statistics
> - Quantitative vs. Qualitative Data
> - Measures of Central Tendency
> - Measures of Variability
> - The Normal Distribution: A Symmetric Curve
> - The Skewed Distribution: A Non-Symmetric Curve

This chapter focuses on the first step in any data-equity project: exploring the data set using descriptive statistics. **Descriptive statistics** are specific methods used to calculate, describe, and summarize collected data, such as equity-minded data, in a logical, meaningful, and efficient way (Vetter 2017)[1]. Descriptive statistics are reported numerically, and/or graphically. According to Vetter, valid and reliable descriptive statistics can answer basic yet important questions about a data set, namely: *Who, What, Why, When, Where, How, and How Much?*

In contrast, **inferential statistics** involve making an inference or decision about a population based on results obtained from a sample of that population (De Muth 2008)[2]. This chapter focuses solely on descriptive statistics, while inferential statistics will be introduced in another chapter. Both descriptive and inferential statistics form the two broad arms of statistics.

Descriptive statistics help facilitate "data visualization" and, as such, are a starting point for analyzing a data set. A **data set** can consist of a number of observations on a range of variables. In statistical research, a **variable** is defined as a characteristic or attribute of an object of study. **Data** is a specific measurement of a variable. Data is generally divided into two categories: **quantitative or numerical data**, and **qualitative or non-numerical data**. Data can be presented in one of four different measurement scales: **nominal, ordinal, interval, or ratio**.
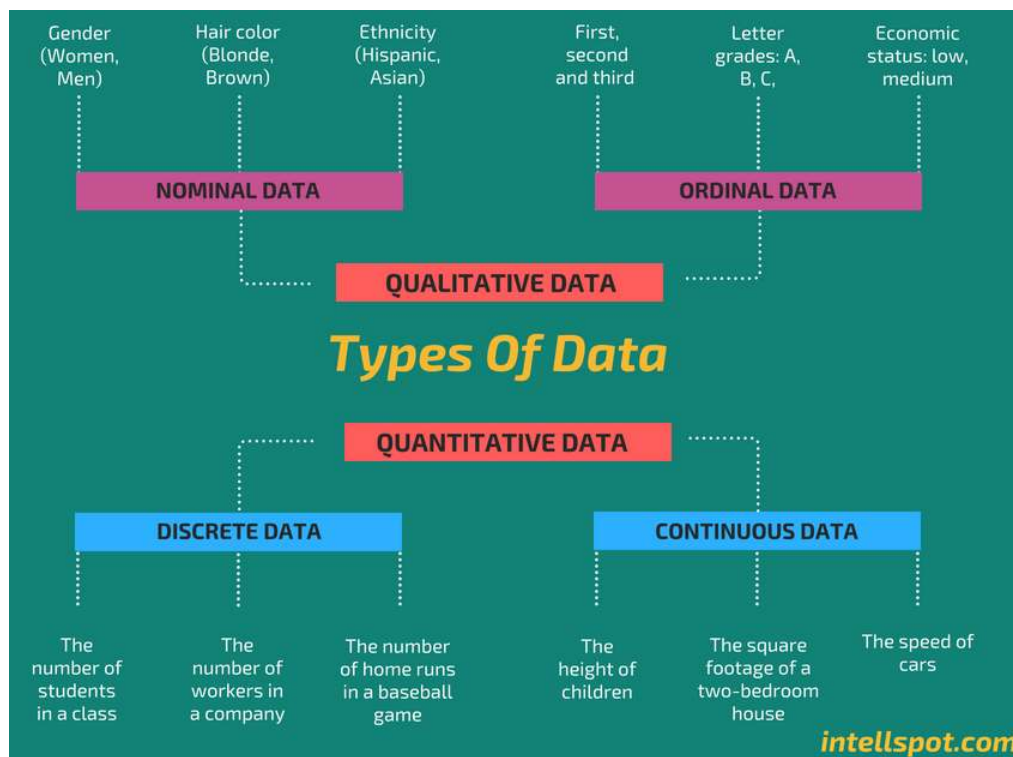
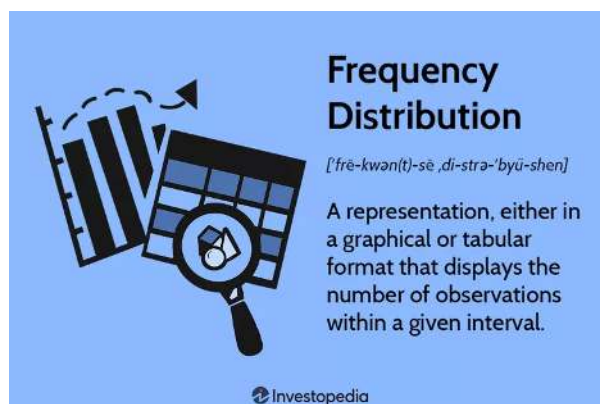There are a variety of ways that statisticians use to describe a data set:

- **Exploratory data analysis** uses graphs and numerical summaries to describe the variables and their relationship to each other. In Section 2.1, you will learn how to organize a data set by grouping the data into intervals called classes and forming a *frequency distribution*.
- **Center of the Data** uses *measures of central tendency* to locate the middle or center of a distribution where most of the data tends to be concentrated. The three best known measures of central tendency used by statisticians are discussed in Section 2.2: the mean, median and mode.
- **Spread of the Data** uses *measures of variability* to describe how far apart data points lie from each other and from the center of the distribution. Section 2.3 discusses the four most common measures of variability: range, interquartile range, variance, and standard deviation.
- **Shape of the Data** can be either *symmetrical* (the normal distribution) or *skewed* (positive or negative). Section 2.4 reveals how the shape of the distribution can easily be discernible through graphs.
- **Fractiles or Quantiles of the Data** uses *measures of position* that partition, or divide, an ordered data set into equal parts: first, second, and third quartile; percentiles; and the standard score (z-score). A fractile is a point where a specified proportion of the data lies below that point. Section 2.5 discusses how fractiles are used to specify the position of a data point within a data set.

In addition to using the above statistical procedures, the data disaggregation process can be applied to identify equity gaps. Disaggregation means breaking down data into smaller groupings, such as income, gender, and racial/ethnic groupings. Disaggregating data can reveal deprivations or inequalities that may not be fully reflected in aggregated data.

The National Center for Mental Health Promotion and Youth Violence Prevention provides two examples of the importance of disaggregating data into smaller subpopulations (National Center Brief 2012)[3]. One area where data disaggregation is commonly used is to show disproportionate minority contact, such as the number of times a minority youth is involved with the court system. In fact, the Office of Juvenile Justice and Delinquency Prevention (OJJDP) uses a specific indicator (Relative Rate Index) to show if there are differences in arrest rates or court sentences, for example, between racial/ethnic groups that are not explained by simple differences in population numbers. A similar step was taken by the Department of Health and Human Services (HHS) as part of the Affordable Care Act. Disaggregated data can be used to see if there are meaningful differences by subpopulations in who is accessing mental services and what treatments are successful.

In descriptive statistics, a particular group of interest (or target group) can be disaggregated by certain characteristics, such as, race, ethnicity, gender, age, socio-economic status, disability, education level, employment in different sectors (e.g., health care, biotechnology, cybersecurity), salary levels, and other different factors. Disaggregating data is viewed as a critical first step while embarking on the equity-minded journey. According to the Annie E. Casey Foundation (2020), "disaggregating data and presenting it in a meaningful way can help bring attention and commitment to the solving of social and racial equity problems."[4]

2.1 Frequency Distribution

Frequency Distribution

['frē-kwən(t)-sē ,di-strə-'byü-shen]

A representation, either in a graphical or tabular format that displays the number of observations within a given interval.

Investopedia

After data has been collected, the next step is to organize the data in a meaningful, intelligible way. One of the most common methods to organize data is to construct a **frequency distribution**–a graphical representation of the number of observations in each category on the measurement scale. It is the organization of **raw data** in table form, using classes and frequency.

At a minimum, a frequency table contains two columns: one listing categories on the scale of measurement (x) and another for frequency *(f)*. The sum of the frequencies should equal n.

A third column can be used for the proportion (p) where p=f/n. The sum of the p column should equal 1.00. A fourth column can display the percentage of the distribution corresponding to each x value. The percentage is found by multiplying p times 100. The sum of the percentage column is 100%. Table 1 below is an example of a four-column frequency distribution of data provided by the Federal Bureau of Prisons, as of May 13, 2023:

Race	Frequency (Number of Inmates)	Proportion (p)	Percentage (%)
White	91,302	.58	58%
Black	61,151	.39	39%
Native American	4,165	.02	2%
Asian	2,272	.01	1%
Total	**158,890**	**1.00**	**100%**

When a frequency distribution table lists all of the individual categories (x values), it is called a **regular frequency distribution** (see Table 1). When a set of data covers a wide range of values, **a group frequency distribution** is used, as shown in Table 2.

Table 2: *Age* of Inmates (n=158,890)

Age Range	Frequency (Number of Inmates)	Proportion (p)	Percentage (%)
Under 18	5	.00	0%
18-21	1,548	.01	1%
22-25	7,444	.047	4.7%
26-30	18,617	.117	11.7%
31-35	26,905	.169	16.9%
36-40	28,129	.177	17.7%
41-45	26,576	.167	16.7%
46-50	18,826	.118	11.8%
51-55	12,954	.082	8.2%
56-60	8,436	.053	5.3%
61-65	5,035	.032	3.2%
Over 65	4,415	.028	2.8%
Total	**158,890**	**1.00**	**100%**

Steps for constructing a frequency table for grouped quantitative data:

Step 1: Sort the data in ascending order to calculate the **range** (minimum and maximum values) for the particular variable of interest.

Step 2: Divide the range or group of values into **class intervals**. Intervals should cover the range of observations without gaps or overlaps.

Step 3: Create **class width** to create groups. All class intervals must be the same width.

Step 4: Determine the **frequency** for each group.

After the data have been organized into a frequency distribution, they can be presented in graphical form. The visual picturization can be used to discuss an equity issue, reinforce a critical point of view, or summarize a data set.

A frequency distribution can be plotted on the x-y plane. On the x-axis is the class intervals of the variable (attribute or characteristics) and on the y-axis is the frequency—the number of observations in a class interval. A plotted frequency distribution conveys the shape of the distribution whether it is the expected standard distribution like the normal distribution or some other known distribution.

2.2 Center of the Data: Measures of Central Tendency

A variable may have several distinct values. A basic step in exploring data is getting a "typical value" for each variable: an *estimate* of where most of the data is located, that is, its **central tendency.** Statisticians often use the term, *estimate*, to draw a distinction between what is seen from the data and the theoretical or true exact state of affairs (Bruce, Bruce & Gedeck 2020)[5].
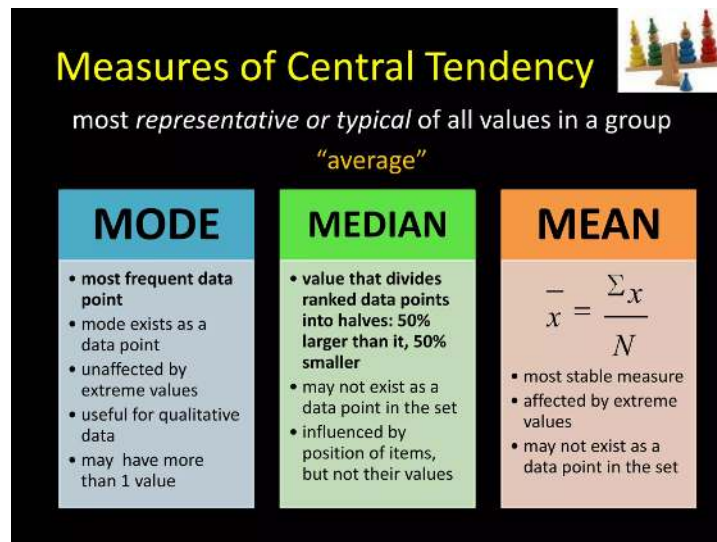
The three most commonly used measures of central tendency are:

Mean: The average, or typical value of the data set.

Median: The number in the middle of the data set.

Mode: The number most frequently found in the data set.

2.2(a) Mean: The most basic estimate of location is the mean, or average. It is generally considered the best measure of central tendency and the most frequently used one. Each individual point exerts a separate and independent effect on the value of the mean. The mean can be used to calculate the average of quantitative variables and not qualitative (categorical) variables.

There are two commonly used methods to calculate the average: simple mean and the weighted mean. The **simple mean** is the sum of values (such as total number of ages of an inmate prison population) divided by the number of values (total number of inmates). There are two steps for calculating the mean ($\bar{x}$):

$$\bar{x} = \left(\sum x_i\right)/n \tag{1.2.1}$$

Step 1: Add all of the values in the data set together ($\sum x_{i}$).

Step 2: Divide the sum by the number of values (n).

Example of How the Mean is Used for Justice-Involved Populations : According to the United States Sentencing Commission, as of January 2022, there are 153,079 offenders incarcerated in the federal bureau of prisons. Their average age is 41 years: 21.6% are 50 years or older; 6.7% are 60 years or older.

The Incarceration-Health Relationship for Justice-Involved Populations: Applying an Equity Lens

The United States has the highest incarceration rate in the world (565 per 100,000 residents)[6]—the highest in its imprisonment history. About two out of three offenders are people of color, African American (34.6%) and Hispanic (31.8%), even though individuals of color make up only one-quarter (26%) of the total population (U.S. National Supplement Prison Study 2014)[7].

From a health equity perspective, "incarceration is viewed as a structural determinant of one's health that also worsens population health. People who are incarcerated are more likely than the general population to experience a chronic condition or acquire an infectious disease. Communities with high rates of incarceration are more likely to experience poor mental health outcomes. Families of people who are incarcerated experience community fragmentation and disruption of social ties that negatively impact mental and familial health" (Peterson & Brinkley-Rubinstein 2021)[8].

Sawyer & Wagner (2023) adds to this health equity perspective, "At least 1 in 4 people who go to jail will be arrested again within the same year — often those dealing with poverty, mental illness, and substance use disorders, whose problems only worsen with incarceration. Most importantly, jail and prison environments are in many ways harmful to mental and physical health. Decades of research show that many of the defining features of incarceration are stressors linked to negative mental health outcomes: disconnection from family, loss of autonomy, boredom and lack of purpose, and unpredictable surroundings. Inhumane conditions, such as overcrowding, solitary confinement, and experiences of violence also contribute to the lasting psychological effects of incarceration, including the PTSD-like Post-Incarceration Syndrome."

The second type of mean is the weighted mean which you calculate by multiplying each data value x by a specified weight w and dividing their sum by the sum of the weights. A weighted mean is a kind of average. Instead of each data point contributing equally to the final mean, some data points contribute more "weight" than others. Often, these weights are percentages but must be converted to decimals when multiplying to the relevant data point, x.

2.2(b) Median: The median is the value that occupies the middle position once data has been sorted from smallest to largest. It divides the frequency distribution exactly into two halves. Fifty percent of observations in a distribution have scores at or below the median; thus, the median is the 50th percentile. Depending on whether the number of observations (n) is odd or even, the median can be calculated in two ways. If the number of scores, n, is odd the median is the value of the (n + l)/2 score. If n is even, there is no single middle value: the median is the average of the scores n/2 and (n/2) + 1. The median can be calculated for ratio, interval, and ordinal variables. It is great for data sets with outliers.

There are three steps for calculating the median:

Step 1: Arrange the number of values from smallest to largest (i.e., ascending order).

Step 2: Determine if n, the number of scores, is odd or even. If n is odd the median is the value of the (n + l)/2 score. If n is even, there is no single middle value: the median is the average of the scores n/2 and (n/2) + 1.

Example: The median for a data set that is even (n=4): 3, 5, 7, 9 is calculated as (5+7)/2=6.

Below is an example of how the median is used to compare the economic or income-based wealth of subpopulations.

> **Applying the Equity LensEconomic Justice: A Component of Social Justice**

According to the 2021 Racism and Racial Inequities in Health Report by the Blue Cross Blue Shield of Massachusetts Foundation[9], "Among Boston-area residents, White households have a median net worth of $247,500, while the **median net worth** for Black and Hispanic households is $12,000 or less, with U.S.-born Black households having a median net worth of $8 and Dominican households having a median net worth of $0. Also, Hispanic and Black people in Massachusetts are less likely to own their homes than White and Asian people. Homeownership is a key source of household stability, as well as a primary pathway for building wealth." (page 1)

Research shows that countries/states/metropolitan areas with a greater degree of socioeconomic inequality show greater inequality in health status. According to the Boston Review (2000)[10], "We have long known that the more affluent and better educated members in our society tend to live longer and have healthier lives. Inequality, in short, seems to be bad for our health… a study (1998) across U.S. metropolitan areas found that areas with high income inequality had an excess of death compared to areas with low inequality. This excess was very large, equivalent in magnitude to all deaths due to heart disease."

Furthermore, a more recent 2023 study published in the Journal of the American Medical Association (JAMA), showed that "because so many Black people die young — with many years of life ahead of them — their higher mortality rate from 1999 to 2020 resulted in a cumulative loss of more than 80 million years of life compared with the White population. Although the nation

made progress in closing the gap between White and Black mortality rates from 1999 to 2011, that advance stalled from 2011 to 2019. In 2020, the enormous number of deaths from COVID-19–which hit Black Americans particularly hard–erased two decades of progress." (Szabo 2023)[11]

Economic wealth is different from *cultural wealth,* which is the reservoir of personal and community resources an individual may have beyond their income or accumulated financial wealth. Challenging traditional definitions of wealth, Dr. Tara J. Yosso (2005) [12] coined the term *community cultural wealth* as "an array of knowledge, skills, abilities and contacts possessed and used by communities of color to survive and resist racism." Yosso's model includes six forms of cultural capital:

1. *Aspirational* capital (ability to maintain hope despite barriers of inequality).
2. *Linguistic* capital (communication skills: facial effect, tone, volume, rhythm).
3. *Familial* capital (wisdom and stories drawn from families in communities).
4. *Navigational* capital (skills and abilities to navigate throughout social institutions).
5. *Resistance* capital (historical legacy in securing equal rights and collective freedom).
6. *Social* capital (using contacts like community-based organizations to gain access and navigate other social institutions).
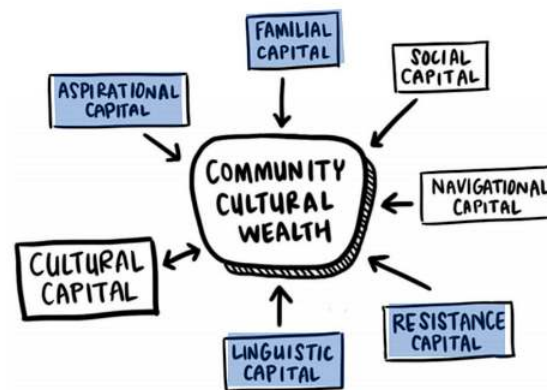
2.2(c) Mode: The mode is the value that occurs most frequently in the data. Some data sets may not have a mode, while others may have more than one mode. A distribution of scores can be unimodal, bimodal, or even polymodal. In a bimodal distribution, the taller peak is called the major mode and the shorter one is the minor mode (Manikandan 2011)[13]. The mode is the preferred measure when data are measured in a nominal scale.

2.2(d) Comparing the Mean, Median and Mode

The relative position of the three measures of central tendency (mean, median, and mode) depends on the shape of the distribution. All three measures are identical in a normal distribution (a symmetrical bell-shaped curve). The mean is most representative of the scores in a symmetrical distribution. As the distribution becomes more asymmetrical (skewed), the mean becomes less representative of the scores it supposedly represents.
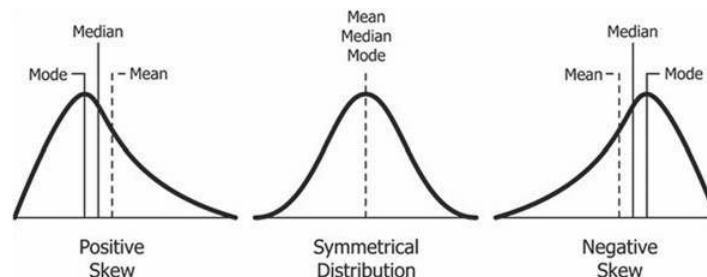
Compared to the mean, which uses all observations, the median depends only on the values in the center of the sorted data. The median is seen as a robust estimate of location since it is not influenced by **outliers (extreme cases)**; in contrast, the mean is sensitive to outliers.

A useful test of skewness is to calculate the values of the mean, median, and mode. If their values are quite different, the use of the mean is not recommended since the distribution is skewed. The median will often be more representative of such scores (Kelly &
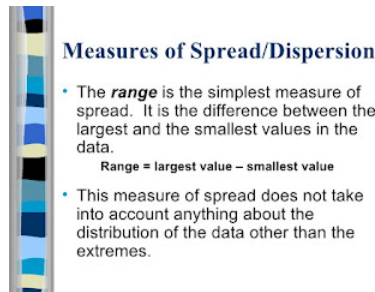
Beamer 1986).

2.3 Spread of the Data: Measures of Variability

The simplest useful numerical description of a distribution requires both a measure of center (mean, median and mode) and a measure of spread. Variability refers to how "spread out" the data points or values are in a distribution. *Variability, spread, and dispersion* are synonymous terms.

There are four main measures of variability: range, interquartile range (IQR), standard deviation and variance.

2.3(a) Range is the simplest measure of variability to calculate. It is the difference between the highest and lowest values of a data set; the result of subtracting the minimum value from the maximum value.

Measures of Spread/Dispersion

- The *range* is the simplest measure of spread. It is the difference between the largest and the smallest values in the data.
  Range = largest value – smallest value
- This measure of spread does not take into account anything about the distribution of the data other than the extremes.
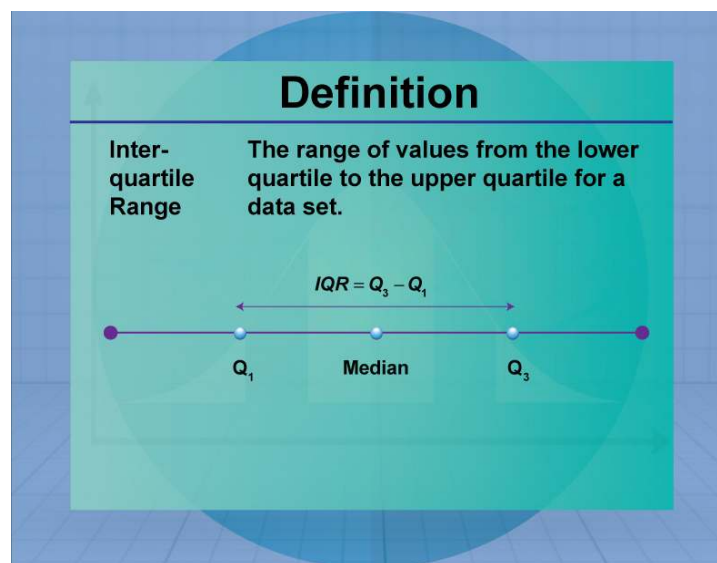

Problem

The term "alternative medicine" means any form of medicine that is outside the mainstream of western or conventional medicine as practiced by the majority of physicians, in hospitals, etc. Well-known examples of alternative medicine are homeopathy, osteopathy, and acupuncture. During the 11-year period (2013-2023), the alternative medicine industry revenue (in billions) was 21.7 (2013), 22.4 (2014), 23.6 (2015), 24.5 (2016), 25.7 (2017), 27.2 (2018), 28.8 (2019), 28.5 (2020), 30 (2021), 30.1 (2022), and 30.6 (2023). What is the range of the industry revenue?

Answer

30.6 – 21.7 = 8.9 billion dollars.

2.3(b) Interquartile Range (IQR) is the range of the middle half (50%) of the scores in the distribution. It is computed as: IQR = 75th percentile – 25th percentile, that is, the region between the 75th and 25th percentile (50% of the data).

Definition

| Inter-quartile Range | The range of values from the lower quartile to the upper quartile for a data set. |

$$IQR = Q_3 - Q_1$$

Q_1 Median Q_3

Using the previous data on Alternative Medicine Industry Revenue, the IQR can be found in three steps:

Step 1: Order the data from least to greatest.

Step 2: Find the median for the entire data set. Answer: 27.5

Step 3: Find the median for the upper (30) and lower (23.6) portions of the data.

Now, to get a quick summary of both center and spread, combine all five numbers. The five-number summary of a distribution consists of the smallest observation (minimum), the first quartile (25th percentile), the median (50th percentile), the third quartile (75th percentile), and the largest observation (maximum), written from smallest to largest. Applying the five-number summary to the previous example on the Alternative Medicine Industry Revenue:

Minimum	Q1	Median	Q3	Maximum
$21.7B	$23.6B	$27.5B	$30B	$30.6B

Study Tip: When the data set is small, the distribution may be asymmetric or the data set may include extreme values. In this case, it is better to use the interquartile range rather than the standard deviation.

2.3(c) Standard Deviation (SD) is the average amount of variability of the data set, informing how dispersed the data is from the center, specifically, the mean. It tells, on average, how far each value lies from the mean. The standard deviation is always greater than or equal to 0. A high standard deviation implies that the values are generally far from the mean, while a low standard deviation indicates that the values are clustered closer to the mean. When the standard deviation is equal to zero, the data set has no variation, and all data points have the same values.

A standard deviation close to zero indicates that data points are close to the mean, whereas a high or low standard deviation indicates data points are respectively above or below the mean.

Steps for calculating the sample standard deviation for ungrouped data are:

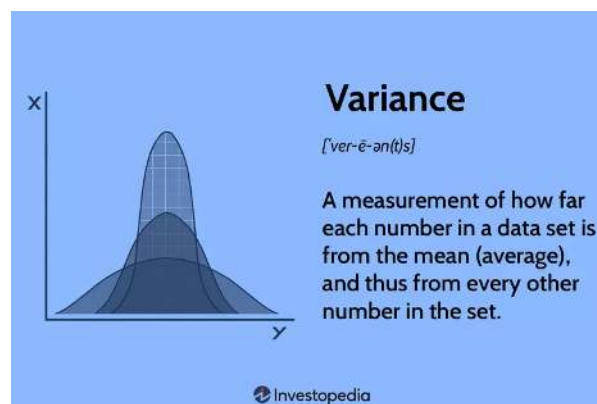Step 1: Find the mean of the sample data set. $\bar{x}=(\sum x_{i})/n$

Step 2: Find the deviation of each data point. $(x-\bar{x})$

Step 2: Square each deviation. $(x-\bar{x})^{2}$

Step 3: Add in order to get the sum of squares. $SS_{x}=(x-\bar{x})^{2}$

Step 4: Divide by n-1 to get the sample variance. $s^{2}_{x}=(x-\bar{x})^{2}/(n-1)$

Step 5: Find the square root of the variance to get the sample standard deviation. **Take the Square root of the formula in Step 4.**

Variance

['ver-ē-ən(t)s]

A measurement of how far each number in a data set is from the mean (average), and thus from every other number in the set.

Investopedia

2.3(d) Variance is the measure of dispersion of the observations or scores around the sample mean. It is the average squared deviations from the mean. The **sample variance** is used to calculate the variability in a given sample. As mentioned in chapter 1, a sample is a set of observations/scores that are pulled from a population and can completely represent it. The reason dividing by n-1 corrects the bias is because we are using the sample mean, instead of the population mean, to calculate the variance. The sample variance, on average, is equal to the population variance.

The variance and its square root, the standard deviation, are two reliable interval-level measures of variability that are employed by statisticians in generalizing from a sample to a population.

Study Tip: Know that: (1) the greater the variability around the mean of the distribution, the larger the mean deviation, range, and variance; (2) the variance is usually larger than the standard deviation; (3) the mean deviation, standard deviation, and variance all assume interval data; and (4) the so-called normal range within which approximately two-thirds of all scores fall is within one standard deviation above and below the mean.

Now Try It Yourself

2.4 Shape of the Data: Normal Distribution and Skewed Distributions

Measures of central tendency speak to the matter of distribution shapes. The **normal distribution** (or the normal curve) is a set of observations that clusters in the middle and tapers off to the left (negative) and to the right (positive). The decreasing amounts are evenly distributed to the left and right of the curve. When a variable follows a normal distribution, the histogram is bell-shaped and symmetric, and the best measures of central tendency and dispersion are the mean and the standard deviation.
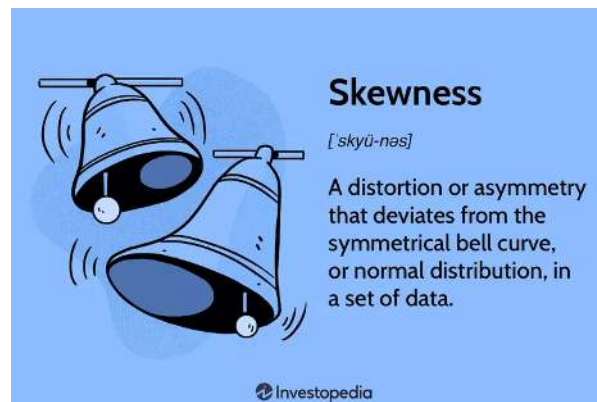
Contrasting with the normal distribution are **skewed distributions** where data is not evenly distributed and the mean deviates from the peak of the distribution**.** Skew type (positive or negative) is determined by the elongated tail of a skewed distribution. *Positively skewed distributions* have data clustering on the left side with extreme values on the right side that pull the tail out to the positive side of the number line. Positively skewed distributions are common in criminal justice and criminology research. *Negatively skewed distributions* are signaled by a tail extending to negative infinity. Scores are sparse on the left-hand side, and they increase in frequency on the right-side of the distribution.

To sum it up, a positive skew distribution is one in which there are many values of a low magnitude and a few values of extremely high magnitude, while a negative skew distribution is one in which there are many values of a high magnitude with a few values of very low magnitude.

> ### Skewed Distributions: Applying the Equity Lens
>
> **Right Skewed Variables**: Income and monetary wealth are classic examples of right skewed distributions. Most people earn a modest amount, but some millionaires and billionaires extend the right tail into very high values. Another example of a right skewed distribution is the intersectionality of race-age-incarceration rates: Black males who are in their twenties and thirties have the highest incarceration rate of any age and racial/ethnic group of any justice-involved population.
>
> **Left Skewed Variables**: According to the Council on Social Work Education, the purpose of social work is actualized through its quest for social and economic justice, the prevention of conditions that limit human rights, the elimination of poverty, and the enhancement of the quality of life for all persons, locally and globally. In the past, the Social Work Values Inventory (SWVI) was available to baccalaureate programs to evaluate students' changes in adherence to basic social work values over time as they complete academic studies. The scales of the SWVI measures three domains: confidentiality, self-determination, and social justice. The scales approximate a normal distribution, with the exception of the social justice scale which is skewed to the left. A left skew on this scale means that proportionately more students tend to have liberal views on social justice rather than conservative views.

2.5 Fractiles and Quantiles of Data (Measures of Position)

When data is arranged in ascending or descending order, it can be divided into various parts by different values. The median, quartiles, deciles, percentiles, and other partition values are collectively called quantiles or fractiles—terms sometimes used interchangeably. However, there are some subtle differences between the two. **Fractile** refers to any of the equal parts into which a population can be divided according to the value of a particular variable. **Quantile**, on the other hand, refers to any of the equal divisions of a frequency distribution that represent the values of the variable. All quantiles are percentages. A fractile is a type of quantile that divides a dataset into equal parts based on a given percentage.

In visual terms, a fractile is the point on a probability density curve so that the area under the curve between that point and the origin (i.e., zero) is equal to a specified fraction. For example, a fractile of .5 cuts the sample in one-half or .75 cuts off the bottom three-quarters of a sample. A fractile xp for p greater than .5 is called an **upper fractile**, and a fractile xp for p less than 0.5 is called a **lower fractile.**
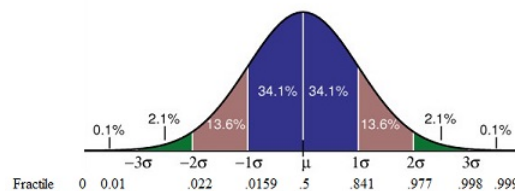

Figure 2: A Fractile on a Normal Distribution

Besides fractiles (any percentage), the other types of quantiles are:

Percentiles (100%)

Values that divide a distribution into a hundred equal parts. There are 99 percentiles denoted as $P_1, P_2...,P_{99}$.
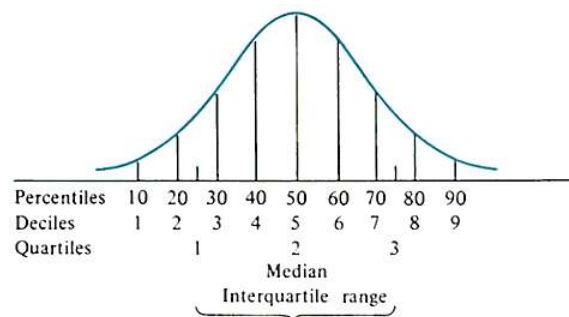
Deciles (10%)

Values that divide a distribution into ten equal parts. There are nine deciles $D_1, D_2...,D_9$.

Quartiles (25%)

Values that divide a distribution into four equal parts. There are three quartiles denoted by Q_1, Q_2, and Q_3.

Median (50%)

Values that divide a distribution into two equal parts, that is, the top 50% from the bottom 50%.

Equivalents: Percentiles, Deciles and Quartiles

The median can be expressed as $Q_2 = D_5 = P_{50}$.

$Q_1 = D_{2.5} = P_{25}$.

$Q_3 = D_{7.5} = P_{75}$.

Study Tips: Fractiles are useful for understanding the distribution of data, particularly when the data is skewed. By comparing the fractiles of two different data sets, one can see how they differ in terms of central tendency and spread. The 95th fractile of the data set is often used as a threshold for identifying outliers. The top 10% of earners in the United States are in the 90th fractiles of the income distribution.

Quantiles are used in statistics to analyze probability distributions (chapter 3). The bottom 5% of earners in the United States are in the first quantile of the income distribution.

Chapter 2: Summary

The main point of this entire OER book is that you cannot achieve equity without investing in data disaggregation. We have started in this chapter to learn how to "peel the onion" so to speak by seeing beyond the surface and splitting large, general categories into more specific groups. As we shift information from broad categories to reflect people's actual experiences, we are helping to ensure that populations who have been historically excluded are more visible. Without the process of disaggregation of data, policies are dramatically misinformed and mask disparities. *"If we continue to fail in our data methods by lumping together racial and ethnic groups, we will continue to erase the experiences of communities and mask how are they faring and, in turn, negatively affect how government and philanthropic resources are allocated, how services are provided, and how groups are perceived or stigmatized."* (Kauh 2021)[14]

In the next chapter, we will learn about probability distributions to expose you more to social issues through a mathematical lens.

Thank you for your persistence and openness to exploring real data and critically thinking about the injustices that certain groups face in their everyday lives.

1. Vetter, T (2017). Descriptive Statistics: Reporting the Answers to the 5 Basic Questions of Who, What, Why, When, Where and a Sixth, So What? Anesthesia and Analgesia, 2017 (Nov); 125(5): 1797-1802.
2. De Muth, JE (2008). Preparing for the First Meeting with a Statistician. American Journal of Health-System Pharmacy. Dec 15; 65(24): 2358-66.
3. National Center Brief (2012). The Importance of Disaggregating Student Data. The National Center for Mental Health Promotion and Youth Violence Prevention, April.
4. Annie E. Casey Foundation (2020). By the Numbers: A Race for Results Case Study-Using Disaggregated Data to Inform Policies, Practices, and Decision-Making. Baltimore, MD.
5. Bruce P, Bruce A, & Gedeck P (2020). Practical Statistics for Data Scientists. O'Reilly Media, Inc., Sebastopol, CA.
6. Sawyer W & Wagner P (2023). Mass Incarceration: The Whole Pie 2023. Prison Policy Initiative. https://www.prisonpolicy.org/reports/pie2023.html. Retrieved on May 27, 2023.
7. New America Analysis of the U.S. Department of Education (2014). National Supplement Prison Study: 2014. National Center for Education Statistics, U.S. Program for the International Assessment of Adult Competencies PIAAC 2012/2014 Household Survey (public use file).
8. Peterson M & Brinkley-Rubinstein L (2021). Incarceration is a Health Threat. Why Isn't It Monitored Like One? Health Affairs Forefront. https://www.healthaffairs.org/conten...monitored-like one. Retrieved on May 22, 2023.
9. Anthony S, Boozang P, Elam L, McAvey K, & Striar A (2021). Racism and Racial Inequalities in Health: A Data-Informed Primer of Health Disparities in Massachusetts. Manatt Health Strategies, LLC.
10. Daniels N, Kennedy B, & Kawachi I (2020). Social Justice is Good for Our Health: How Greater Economic Equality Would Promote Public Health. Boston Review. https://www.bostonreview.net/forum/n...od-our-health/ Retrieved on May 29, 2023.
11. Szabo, L (2023). Study Reveals Staggering Toll of Being Black in America: 1.6 Million Excess Deaths Over 22 years. https://www.nbcnews.com/health/healt...hs-2-rcna84627 Retrieved on May 29, 2023.
12. Yosso TJ (2005). Whose Culture Has Capital? A Critical Race Theory Discussion of Community Cultural Wealth. Race Ethnicity and Education, Volume 8, Issue 1, pages 69-91.
13. Manikandan S (2011). Measures of Central Tendency: Median and Mode. Postgraduate Corner. Journal of Pharmacology and Pharmacotherapeutics, July-September 2011, Vol 2, Issue 3.
14. Kauh TJ, Read JG & Scheitler (2021). The Critical Role of Racial/Ethnic Data Disaggregation for Health Equity. Population Research and Policy Review, 40, 1-7(2021).

1.3: Probability

Up until now, interpretations of data have pretty much come from "the seen", or what can be observed. In this chapter, we will deal with the "unseen", the unobserved, which are **probabilities**. Probability can be a difficult concept to grasp, yet we use it every day. We ask such questions as *"What are the chances it is going to rain today?", "How likely is it that this relationship will last?"*, or *"What is the chance that I will get an A out of this statistics class?"*. We answer these questions with "fair chance", "likely", or "highly likely". A pertinent question is, *"How likely is it that you will learn about the two topics of statistics and social justice issues at the same time?"* This question is for you to answer.

> **◑ Learning Objectives**
>
> There are many aspects of probabilities that will be covered in this chapter:
>
> - The Definition of Probability in Statistics
> - Rules of Probability
> - Probability Distributions
> - The Normal Curve as a Probability Distribution
> - The Standard Normal Distribution
> - Blending Probability with Social Justice Issues

Now, let's begin our journey to understand this "mystical" concept.

3.1. Probability Definition in Statistics

Simply put, probability means possibility—the extent to which something is likely to happen. As shown in Figure 1, probability (P) can range from 0 to 1, where 0 means the event is an *uncertain* (or impossible) one and 1 indicates a *certain event*. Probabilities of .05 or .10 imply very unlikely circumstances, and high probabilities such as .90, .95, or .99 signify very probable or likely outcomes.
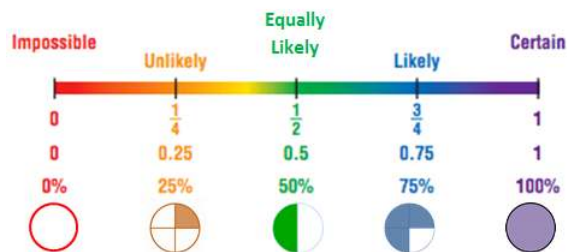
Probability is also a mathematical tool to study randomness. A phenomenon is **random** if individual outcomes are uncertain. A **random variable** represents a value associated with each outcome of a probability experiment. For a random variable, x, the word random indicates that the value of x is determined by chance. The expected value of a random variable can be positive, negative, or zero.

A **probability model** is a mathematical description of a random phenomenon consisting of a **sample space**, **S**, which is the set of all possible outcomes. If S is the sample space in a probability model, the *P(S)=1*. A probability model with a finite sample space is called **finite**. Finite probability models are often called **discrete** probability models.

An event is an outcome or a set of outcomes of a random phenomenon; it is the subset of the sample space.

Now Try It Yourself

Elliott et al. (2006) created a hybrid method (geocoding and surname analyses) for estimating race/ethnicity and associated disparities where self-reported race/ethnicity data is unavailable. Geocoding uses an individual's address to link individuals to census data about the geographic areas where they live. For example, knowing that a person lives in a Census Block Group (a small neighborhood of approximately 1,000 residents) where 90 percent of the residents are African American provides useful information for estimating that person's race. Surname analysis infers race/ethnicity from surnames (last names). Insofar as a

particular surname belongs almost exclusively to a particular group (as defined by race, ethnicity, or national origin), the researchers used well-formulated surname dictionaries to identify a probable membership in a group.

Problem #1: Verify whether the researchers' findings in Table 1 below is a legitimate assignment of probabilities:

Table 1: Probabilities of a Male Individual Living in a Census Block Group in Dorchester, MA

Surname	Asian	Hispanic	African American	White/Other
Wang	.937	.008	.008	.048
Martinez	.010	.845	.021	.125
Jones	.061	.022	.129	.787

Answer:

To be a legitimate probability distribution, the sum of the probabilities for all possible outcomes must equal to 1. For the surname (n=3), the probabilities are 1.00 for Wang, 1.00 for Martinez, and .999 for Jones. Of the surname group, only two have a legitimate probability distribution.

For the ethnic groups, the probabilities are 1.00 for Asian, .875 for Hispanic, .158 for African American, and .96 for White/Other. The Asian Group is the only ethnic group that has a legitimate assignment of probabilities.

Thus, of the seven outcomes, only three outcomes add up to 1 and have legitimate probability distributions.

3.2 Rules of Probability

Rule #1: *The probability associated with an event is the number of times that event can occur relative to the total number of times any event can occur. This is known as the **classical or theoretical probability**. The **empirical or statistical probability** is based on observations obtained from probability experiments. The empirical probability of an event E is the relative frequency of event E.*

Rule #2: *The complement of event E is the set of all outcomes in a sample space that are not included in E and is denoted as E', pronounced as E prime. The* complement or converse rule of probability is 1 minus the probability of that event occurring.

$P(E') = 1 – P(E)$

Based on the example in the box below, the recidivism rate in the United States is 77%. In other words, a person released from prison has a 77% probability of being rearrested. The converse is non-recidivism. The probability that a discharged inmate does not recidivate is $1 – .77 = .23$.

> ### Recidivism Rate:
> ### Applying the Equity Lens

Recidivism is the tendency of a convicted criminal to repeat or re-offend a crime after already receiving punishment or serving their sentence. The term is often used in conjunction with substance abuse as a synonym for "relapse" but is specifically used for criminal behavior. The United States has some of the highest recidivism rates in the world. Norway has one of the lowest recidivism rates in the world at 20%. **The U.S. has one of the highest: 76.6% of prisoners are rearrested within five years**. Among Norway's prison population that was unemployed prior to their arrests, they saw a 40% increase in their employment rates once released. The country attributes this to its mission of rehabilitation and reemergence into society through its accepting and empathetic approach (Benecchi 2021)[1].

Rule #3: *The* addition rule of probability states that the probability of obtaining any one of several different and distinct outcomes *equals the sum of their separate probabilities*. The addition rule always assumes that the outcomes being considered are **mutually exclusive *or disjointed**—that is, no two outcomes can occur simultaneously* (Levin & Fox 2006)[2].

$$P (A \text{ or } B \text{ or } C) = P (A) + P (B) + P(C) \text{ Mutually exclusive events}$$

$$P (A \text{ or } B \text{ or } C) = P (A) + P (B) + P(C) – P(A \text{ and } B) – P(A \text{ and } C) – P(B \text{ and } C) + P (A \text{ and } B \text{ and } C)$$

Example of Adding Probabilities: Suppose that a defendant in the January 6th United States Capitol Attack has a .55 probability of being convicted as charged, a .26 probability of being convicted of a lesser degree, and a .19 chance of being found not guilty.

The chance of a conviction on any charge is .55 + .26 = .81. Note also that this answer agrees with the converse rule by which the probability of being found not guilty is 1 – .19 = .81.

Now Try It Yourself

Rule #4: *The* **multiplication rule of probability** *states that the probability of obtaining a combination of independent outcomes* <u>*equals the product of their separate probabilities*</u> (Levin & Fox 2006).

P(A and B) = P(A) x P(B/A) Dependent events

[Note: *P(B/A) is a conditional probability of event B occurring given that event A has occurred.*]

P (A and B) = P(A) x P(B) Independent events

Example of Multiplying Probabilities: A prosecutor is working on two cases, a gender-based violence trial and a race-based murder trial. From previous experience, she feels that she has a .80 chance of getting a conviction on the violence against women trial and a .60 chance of a conviction on the murder trial. Thus, the probability that she will get convictions on *both* trials is (.80)(.60) = .48 (slightly less than one-half).

3.3 Probability Distributions

Probability distributions are a fundamental concept in Statistics. They are used both on a theoretical and practical level. A **probability distribution** simply shows the probabilities of getting different outcomes. For example, the distribution of flipping a coin is .5 for heads and .5 for tails. There are a multitude of probability distributions that are used in statistics, economics, finance, and engineering to model all sorts of real-life phenomena.

The **Uniform Distribution** is the probability distribution in which all outcomes have an equal probability or a constant probability. The uniform probability distribution is often used in situations where there is no clear "favorite" outcome, and all outcomes are equally likely.

Probability distributions are divided into two classes: discrete and continuous. A **Discrete Probability Distribution** is a mathematical function that calculates the probability of outcomes of discrete random variables. The most common type of discrete probability distribution is the **Binomial distribution**, which is used to model events with two possible outcomes, such as success and failure.

A **Continuous Probability Distribution** deals with random variables that can take on any continuous value within a certain range. They are often used to model physical phenomena, such as height, weight, and volume. The most common continuous probability distribution is the **normal distribution,** which is discussed in the next section.

3.4 The Normal Curve as a Probability Distribution

Because it is a probability distribution, the normal curve is a theoretical ideal. It is a type of symmetric distribution that is bell-shaped and has one peak (unimodal) or point of maximum frequency in the middle of the curve. That point is where the mean, median, and mode coincide. If you recall, having the mean, median, and mode at different points reveals a skewed distribution, as shown below in Figures 2 and 3.
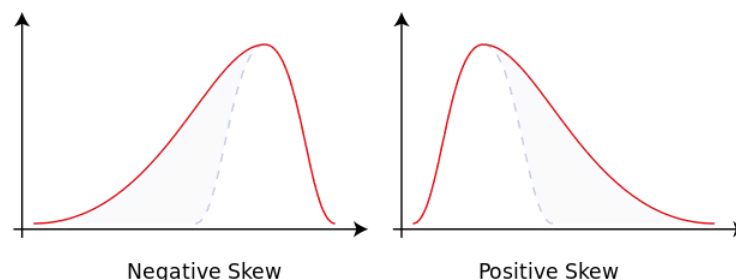

Negative Skew   Positive Skew

Figure 2 (left): Skewed to the Left: Smallest Extreme Value Distribution / Figure 3 (right): Skewed to the Right: Largest Extreme Value Distribution

Figures 4 and 5 below are sample normal distribution curves. The area under the normal curve contains 100% of all the data. The area is segmented by the **Empirical Rule**, which states that all observed data for a normal distribution fall within three standard deviations from the mean:

- 68% of the data falls between -1 and +1 standard deviations from the mean.
- 95% of the data falls between -2 and +2 standard deviations from the mean.
- 99.7% percent of the data falls between -3 and +3 standard deviations from the mean.
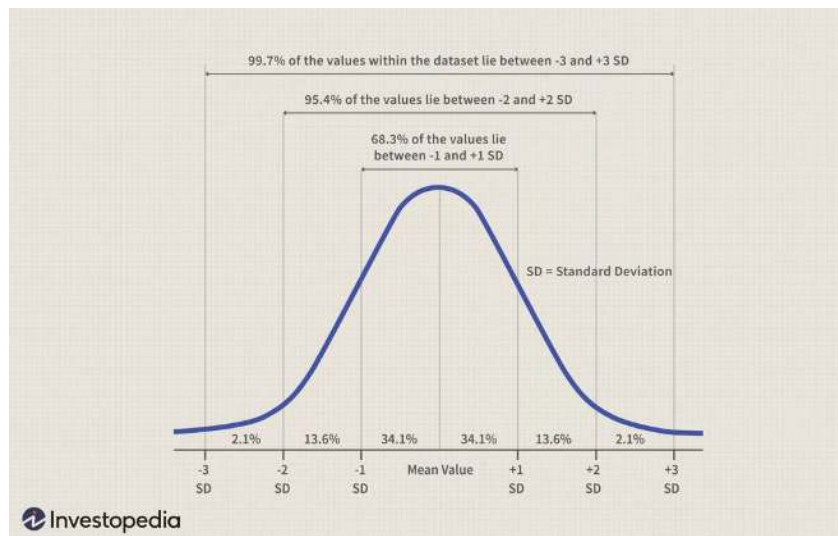

Figure 4: The Empirical Rule

As shown in Figure 5 below, a constant proportion of the total area under the normal curve will lie between the mean and any given distance from the mean as measured in deviation units. Thus, the area under the normal curve between the mean and the point 1 above the mean always turns out to include 34% of the total cases, regardless of the variable of interest. The symmetrical shape of the normal curve means that there is an identical proportion of cases (34%) below the mean. But "*What do we do to determine the percent of cases for distances lying between any two score values that do not fit precisely as one, two, three standard deviations from the mean?*" To determine, for example, the exact percentage of 1.4 standard deviations from the mean, we would use Percentage Breakdown of Area Under the Normal Curve. Corresponding to the 1.4 standard deviations from the mean includes 41.92% of the total area under the curve.
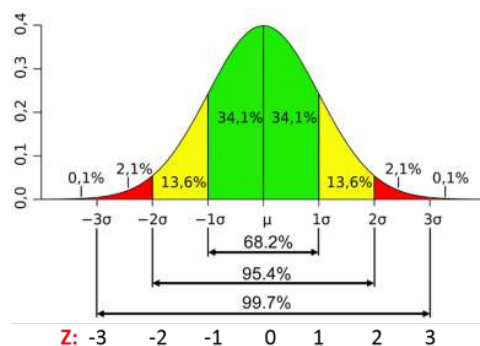

Figure 5: Percentage Breakdown of the Area Under the Normal Curve

Now Try It Yourself

3.5 The Standard Normal Distribution

There are infinitely many normal distributions, each with its own mean and standard deviation. The **standard normal distribution** has a mean of 0 and a standard deviation of 1. We can use the **standard score**, or **z-score,** to represent the number of standard deviations a value, x, lies from the mean, μ. To find the z-score for a value, the formula below is used:

$$z = \frac{Value - Mean}{Standard\ Deviation} = \frac{x - \mu}{\sigma} \tag{1.3.1}$$

A *z*-score can be positive, negative, or zero. If positive, the corresponding x-value is greater than the mean. When *z* is negative, the corresponding *x*-value is less than the mean. If zero, the corresponding x-value is equal to the mean. *z*-scores can be used as descriptive statistics and as inferential statistics. As descriptive statistics, *z*-scores describe exactly where each individual is located.

As inferential statistics, *z*-scores determine whether a specific sample is representative of its population or is extreme and unrepresentative.

The Role of Z-Scores in Measuring Malnutrition in Children: Applying the Equity Lens

The World Health Organization (WHO) defines malnutrition as deficiencies or excesses in nutrient intake, imbalance of essential nutrients, or impaired nutrient utilization. In the area of child health and nutrition, the *z*-score is the positive or negative standard deviation of a particular child with respect to the median of a carefully selected sample or a predetermined population. WHO categorizes malnutrition in children as *severe* if the z-score (weight for height and height for age) is less than –3 standard deviations, *moderate* if the z-score (weight for height and height for age) is between –2 and –3 standard deviations, and *mild malnutrition* if values are between –2 standard deviations to –1 standard deviation. The criteria for a malnutrition diagnosis are weight loss, low body mass index (BMI), reduced muscle mass, reduced food intake or assimilation, and disease burden/inflammation.

Malnutrition in children is a public health problem in many developing countries such as India. According to Narayan et al. (2019) [3], reports of the National Health & Family Survey, United Nations International Children's Emergency Fund, and WHO have highlighted that rates of malnutrition among adolescent girls, pregnant and lactating women, and children are alarmingly high in India. Factors responsible for malnutrition in the country include the mother's nutritional status, lactation behavior, women's education, and sanitation. These affect children in several ways, including stunting, childhood illness, and retarded growth.

Commentary by the Author: Several years ago, I was invited by The PRASAD Project[4] I worked in India to provide hospital planning-administration consulting support for a new hospital in the Tansa Valley of Maharashtra, India. The Shree Muktananda Mobile Hospital, which began in 1978, is one of PRASAD's first healthcare initiatives in the Tansa Valley. I rode on the mobile hospital bus every day and saw first-hand how doctors and nurses treated infectious diseases, chronic diseases (such as diabetes and hypertension), skin illnesses, and general health care. Since 1978, over 1,000,000 people have received screenings, medical care, and health education from the Shree Muktananda Mobile Hospital. In addition, PRASAD started its Milk and Nutrition Program for children in 1980. In 1990, volunteers organized the first Eye Camp in India. In 2002, the Maternal and Child Health Program began. In 2004, the HIV Program started, and since then, it has brought the prevalence rate below India's national level. Before PRASAD began offering programs in the Tansa Valley, most children in the area were malnourished, adults battled untreated tuberculosis and heart disease, and many endured lives of blindness caused by cataracts.

The Anukampaa Health Center is home to PRASAD's Tuberculosis Program. The disease accounts for more casualties than any other infectious disease in India, claiming a life every minute. Doctors at the Health Center have achieved a tuberculosis cure rate of 95 percent, surpassing the Indian government's benchmark of 85 percent. In 2010, the World Health Organization and the Indian government's Revised National TB Control Program recognized the PRASAD program as a Designated Microscopy and Treatment Center.

Now Try It Yourself

Table 2: The Monsoons
India's Regional Actual Monthly Rainfall (Millimeters) in 2019

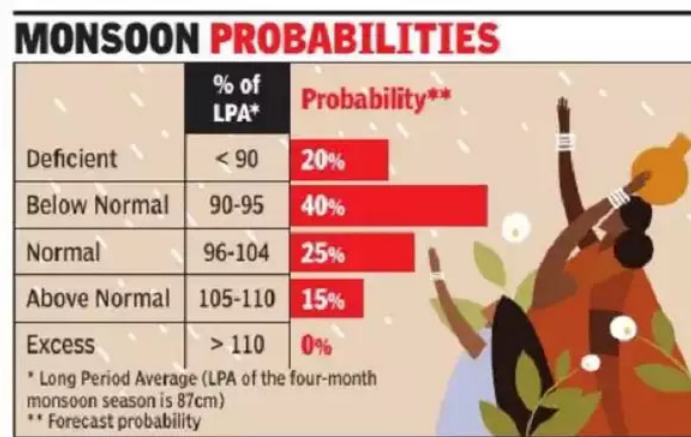| Regions	June	July	August	September
Northwest India	53.1	213.8	207.2	121.0
Central India	117.4	350.8	427.7	367.3
South Peninsula	112.3	194.3	296.1	238.2
East & Northeast India	223.1	481.9	213.6	325.1
Total	**505.9**	**1240.8**	**1144.6**	**1051.6**

Figure 6: India's 2023 Monsoon Season Probabilities by Private Forecaster. Source: Times of India, India's Meteorological Department predicts 'normal' rainfall for this year's monsoon. April 11, 2023

3.6 Blending Probability with Social Justice Issues

A powerful example of social justice is the role of race in the conviction of innocent people. According to Gross et al. (2022)[5], "as of August 8, 2022, the National Registry of Exonerations listed 3,200 defendants who were convicted of crimes in the United States and later exonerated because they were innocent; 53% of them were Black, nearly four times their proportion of the population, which is now about 13.6%." The National Registry of Exonerations tracks all known wrongful convictions in the United States since 1989.

The Report was made through the joint efforts of the University of California Irvine Newkirk Center for Science and Society, the University of Michigan Law School, and Michigan State University College of Law. The Registry's first Report on race and wrongful convictions was released in 2017. The September 2022 Report contains more information and detail than the 2017 Report, along with improved data.

When hearing this information, our reaction could be, "Well, given the extent of structural racism in our society and how race is a proxy for criminality in the criminal legal system, this seems quite probable." We engage in statistical probabilistic thinking *unconsciously* when we also hear about racial profiling or "driving while Black or Brown". Black and Hispanic males have complained, filed suit, and organized against what they believe are racist police practices: being stopped, searched, harassed, and sometimes arrested solely because they "fit" a racial profile. So, when we hear on the news about another killing of a young black male by a police officer or a neighbor, we are not surprised because such events have become normalized or "highly likely" in the United States.

Putting emotions aside, there are concepts in statistics that we can apply as we blend probabilistic thinking with social justice issues. They are *randomness, experimental and theoretical probability, simulation, sample size, and the law of large numbers*. Let's take, for example, the principle of **randomness,** which is the heart of statistics that underpins much of our knowledge. It is the apparent or actual lack of pattern or predictability of information or event. *Randomness in probability* describes a phenomenon in which the outcome is uncertain, but there is a regular distribution of relative frequencies in a large number of repetitions. In the instance of racial profiling, such occurrences do not seem to be random but have more of a definite plan, purpose, or pattern. We have seen in the multitude of racial profiling cases that there is both a predictable short- and long-term pattern that can be described by the distribution of outcomes, namely, being stopped, searched, harassed, and sometimes arrested solely because Black and Brown males "fit" a racial profile.

Chapter 3: Summary

In this chapter, we learned that probability is a statistical term used to express the *likelihood* that an event will happen. The probability of an event can be calculated by the probability formula by simply dividing the favorable number of outcomes by the total number of possible outcomes. The closer the probability is to zero, the less likely it is to happen, and the closer the probability is to one, the more likely it is to happen. The total of all the probabilities for an event is equal to one.

We also learned that the normal curve is a *probability distribution* in which the total area under the curve equals 100%. It contains a central area surrounding the mean, where scores or observations occur most frequently, and smaller areas toward either end, where

there is a gradual flattening out and, thus, a smaller proportion of extremely high and low scores. From a probability perspective, we have seen in previous graphs that probability decreases as we travel away from the mean in either direction. Thus, 68% of the cases falling within -1 and +1 standard deviations is like saying that the probability is approximately 68 in 100 that any given raw score will fall within this interval.

In Chapter 4, we will move from *descriptive statistics,* where we acquired tools to describe a data set, to *inferential statistics,* where we make inferences based on a data set. The goal is to discover a general pattern about a large group while studying a smaller group of people in the hope that results can be generalized to the larger group. The most common methodologies in inferential statistics are *hypothesis testing, confidence intervals, and regression anal*yses, which will be covered in subsequent chapters.

Remember: It is not solely about the destination. It is also about the journey of gaining a critical statistical perspective that incorporates and facilitates awareness of social justice issues. I hope you are enjoying your journey and seeing how statistics is a tool to help affect social change in society.

1. Benecchi L (2021). Recidivism Imprisons American Progress. Harvard Political Review. https://harvardpolitics.com/recidivi...ican-progress/ Retrieved on June 8, 2023.
2. Levin J & Fox JA (2006). Elementary Statistics in Social Research. Pearson, Inc., Boston, MA; Tenth Edition.
3. Narayan J, John D & Ramadas N (2019). Malnutrition in India: Status and Government Initiatives. Journal of Public Health Policy, 40, pages126–141.
4. . PRASAD is a philanthropic expression of the Siddha Yoga mission. Gurumayi Chidvilasananda, the spiritual head of the Siddha Yoga path, started the PRASAD Project in 1992. I had the honor and privilege to serve Gurumayi on this project, which has had the most amazing impact on my life—both personally and professionally. PRASAD is a leading humanitarian organization that effectively addresses health inequalities and the social determinants of health among the poorest of the poor in India. My experience with PRASAD is the genesis and motivation for this book.
5. Gross SR, Possley M, Otterbourg K, Stephens K, Paredes JW & O'Brien B (2022). Race and Wrongful Convictions in the United States 2022. National Registry of Exonerations, September.

1.4: Inferential Statistics- Sampling Methods

> **Learning Objectives**
>
> - Sampling Methods
> - Sampling Error
> - Sampling Distribution of Sample Means
> - The Central Limit Theorem
> - Sampling Variability

Descriptive statistics describes a sample, using summary statistics and graphs to present the group properties. Chapters 1 and 2 discussed the three major types of descriptive statistics: frequency distribution, central tendency, and variability/dispersion. With descriptive statistics there is no uncertainty as you are summarizing the characteristics of a data set. *Measures of central tendency* (mean, median, and mode*)* and *measures of dispersion or variation* (range, interquartile range, variance, and standard deviation) are used to understand descriptive statistical results.

Steps for Describing Descriptive Statistics Results

1. Describe the sample based on demographics of interest, such as, race/ethnicity, gender, age, zip code, education, etc.
2. Describe the center of the data.
3. Describe the spread and shape of the data distribution.
4. Describe the data using tables, charts, and graphs to summarize the frequency of values of variables, rendered in percentages or numbers.
5. Describe the data using univariate descriptive statistics (examining one variable at a time) or bivariate descriptive statistics (two variables are analyzed or compared concurrently) to see if they are correlated.

In chapter 3 on probability, you were introduced to another branch of Statistics—**inferential statistics.** *Descriptive statistics are used to effectively summarize and describe the main features of a data set, while statistical inferential methods make predictions or inferences about a larger population which goes beyond descriptive statistics involving estimating parameters and testing hypotheses.*

Probability is the underlying concept of inferential statistics and forms a direct link between a sample and the population it comes from. A random sample of data from a portion of the population is used to make *inferences* or *generalizations* about the entire population. For example, Wooditch et al. (2021)[1] reports that there are "many topics within criminology and criminal justice research where conclusions are drawn from data that are generalizable to wider populations."
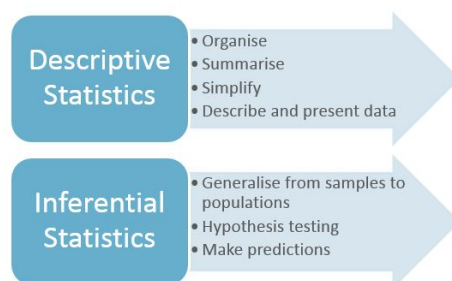
Inferential statistics plays an important role in fields such as science, medicine, social science, and business. These are fields that also have a greater opportunity to adopt social justice principles: *access, equity, participation, and human rights.* Social equity includes equal opportunities and obligations and, therefore, involves the whole of society. It is about access, mutuality, and concern for *justice* and *fairness* in social policy.[2] Social equity negates discrimination on the basis of race, gender, sexual orientation, class, income, language, religion, convictions, opinions, or disabilities (Augusty & Dizon 2020)[3].

Social equity, in the context of statistics, uses information and data to provide opportunities for success to individuals based on their right of access and specific needs. Some statisticians, therefore, add social group dimensions (variables) to data collection such as *race and gender,* whereas others may also include qualitative stories, life experiences, and realities of indigenous lives to contextualize the quantitative data, promoting social equity. Like the author, they passionately believe that the invisibility of

vulnerable groups in data collection results in the marginalization of groups. This OER book intentionally puts a "spotlight" on historically marginalized individuals, families, groups, and communities not only because it is morally right but also because not to do so is flawed analysis.

For the author, adopting a social justice framework to statistics starts with seeing *all* students as capable of learning statistics and transforming into "statistically literate citizens" who can identify and help solve equity-related problems for families and communities. This includes using data to capture the magnitude of inequities and to track the progress of equity over time.

Social equity involves the analysis of different subgroups (i.e., the sample) relative to a majority group (i.e., the population). This chapter discusses various methods that can be applied during this inferential statistical process. Specific topics on sampling will be covered in this chapter:

- Sampling Methods
- Sampling Error
    - Sampling Distribution of Sample Means
    - The Central Limit Theorem

4.1 Sampling Methods

The purpose of a sample is to give information about a larger population. The process of drawing conclusions about a population on the basis of sample data is called **inference**. To make inferences, you have to select an appropriate **sampling method.** Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences to estimate characteristics of the whole population. **Sampling design** describes exactly how to choose a sample from a population.
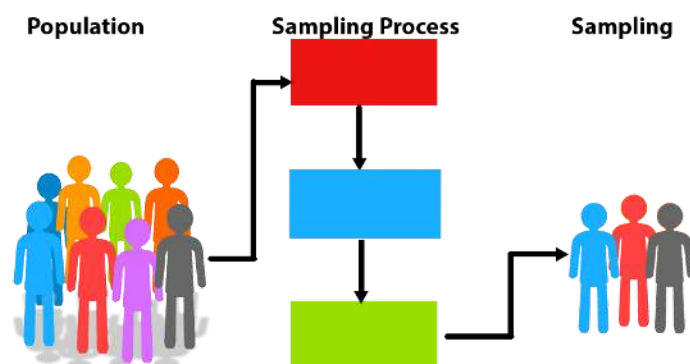
Figure 1: The Tripartite Relationship: Sample-Sampling-Population

The first question to ask about a sample is whether it was chosen at random. Moore, Notz & Fligner (2013)[4] provides two reasons to choose random sampling. The first reason is to eliminate bias in selecting samples from the list of available individuals. The second reason to use random sampling is that the laws of probability allow trustworthy inference about the population.

If every population member is given an equal chance of sample selection, a **random sampling method** is used. This is known as **probability sampling.** Otherwise, a **nonrandom or sampling with non-probability** is employed.

In *probability sampling,* a total sample has an equal chance of being chosen. The probability is that each segment of the population has a probability of being selected and gives the likelihood of the sample being representative of the population. There are four types of *probability sampling* techniques:

4.1(a). Simple Random Sampling (SRS)

Simple Random Sampling (SRS): a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each person has the same probability of being chosen to be a part of a sample. An SRS of size *n* individuals from the population is chosen in such a way that every set of *n* individuals has an equal chance to be the sample actually selected. A specific advantage of simple random sampling is that it is the most straightforward method of probability sampling. A disadvantage is that you may not find enough individuals with your characteristic of interest, especially if that characteristic is uncommon.

4.1(b). Cluster Sampling

Cluster Sampling: a method where statisticians divide the entire population into clusters or sections representing a population. Demographic characteristics, such as race/ethnicity, gender, age, and zip code can be used to identify a cluster. Cluster sampling can be more efficient than simple random sampling, especially where a study takes place over a wide geographic region.

An extended version of cluster sampling is **multi-stage sampling,** where, in the first stage, the population is divided into clusters, and clusters are selected. At each subsequent stage, the selected clusters are further divided into smaller clusters. The process is completed until you get to the last step, where some members of each cluster are selected for the sample. Multi-stage sampling involves a combination of cluster and stratified sampling.

The U.S. Census Bureau uses multistage sampling by first taking a simple random sample of counties in each state, then taking another simple random sample of households in each county and collecting data on those households.

An Example of a Multi-Stage Cluster Sampling: Applying the Equity Lens

The National Survey of American Life (NSAL) is the most comprehensive and detailed study of mental disorders and the mental health of Americans of African descent. The study was conducted by the Program for Research on Black Americans (PRBA) within the Institute for Social Research at the University of Michigan.[5] According to Jackson et al. (2004), the study includes "a large, *nationally* representative sample of African Americans, permitting an examination of the heterogeneity of experience across groups within this segment of the Black American population. Most prior research on Black Americans mental health has lacked adequate sample sizes to systematically address this within-race variation."

Special emphasis of the study is given to the nature of race and ethnicity within the Black population by selecting and interviewing national samples of African American (N=3,570) and Afro-Caribbean (N=1,623) immigrants, and second and older generation populations. National multi-stage probability methods were used in generating the samples and race/ethnic matching of interviewers and respondents were used in the face-to-face interview, which lasted on average 2 hours and 20 minutes. For generalizability, probability sampling methods were used in the study for stronger statistical inferences.

4.1(c). Systematic Sampling

Systematic Sampling: a method where sample members of a population are chosen at regular intervals. It requires selecting a starting point for the sample and where sample size determination can be repeated at regular intervals. Systematic sampling is often more convenient than simple random sampling as it is easy to administer.

4.1(d). Stratified Random Sampling

Stratified Random Sampling: a method that divides the target population into smaller groups that do not overlap but represent the entire population. The selected sample from different strata is combined to have a single sample.
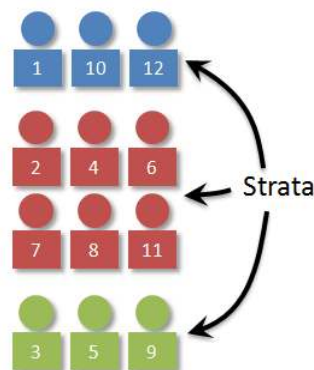
Stratified sampling improves the accuracy and representativeness of the results by reducing **sampling bias** (where some members of a population are systematically more likely to be selected in a sample than others). The design of a statistical study is **biased** if it systematically favors certain outcomes.

Statistical bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process. An important distinction is made between errors due to random chance and errors due to bias. An unbiased process will produce

error, but it is random and does not tend strongly in any direction.

Figure 2 shows how sampling bias can result in the sample mean overestimating (or underestimating) the population mean.
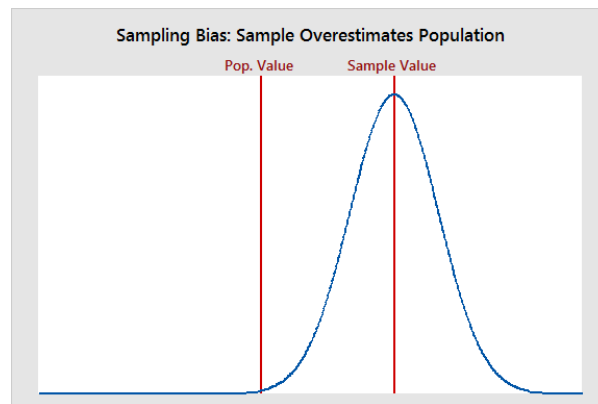

Figure 2: Inaccurate Sample Statistic (Mean) in Estimating the Population Parameter (Mean)

In *sampling with non-probability,* non-randomized methods are used. Participants are chosen since they are easy to access in place of randomization. The limitation of this method is that the results are not generalizable to the population but are relevant primarily to that particular group sampled.

There are two stratified sampling techniques. The *proportionate sampling of stratified samples* is used when the number of elements assigned to the different strata is proportional to the representation of the strata in the target population (Illiyasu & Etikan 2021)[6]. The *disproportionate sampling of stratified samples* is used when the number of elements sampled from each stratum is disproportional to their population representation (Illiyasu & Etikan 2021).

Now Try It Yourself

4.2 Sampling Error
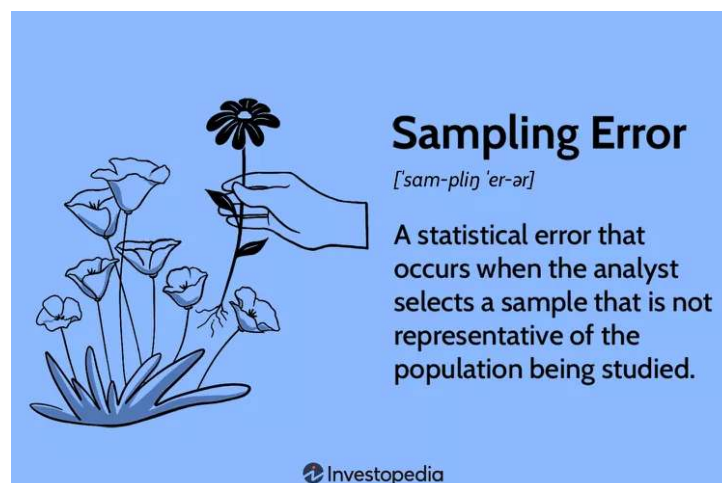
You are more familiar with the concept of **sampling error** than you realize. Recall that election polls, for example, typically generalize from a small sample to the entire population of voters. When reporting the results, the **margin of error** is typically provided by the market research industry for data reliability. As a hypothetical example, we might hear that Michelle Obama is receiving 70% of the vote as a presidential candidate, with a ±4% margin of error. In other words, there is confidence that somewhere between 66% (70%-4%) and 74% (70%+4%) will go Michelle's way. The reason an exact percentage vote cannot be provided is due to *uncertainty* which, in turn, is due to the effect of sampling error.

The margin of error is a statistic expressing the amount of random sampling errors in the results of a survey. The larger the margin of error, the less confidence one should have that a poll result would reflect the result of a census of the entire population. The margin of error will be positive whenever a population is incompletely sampled and the outcome measure has a positive variance.

As shown in Figure 3 below, in general, the sampling error (or margin of error) gets smaller as the sample size increases because the sample more accurately represents the population. Furthermore, along with increasing the sample size, sampling errors can be controlled if the sample is chosen at random from the population. Another way to look at it is through the **law of large numbers,** which states that the actually observed mean outcome $\bar{x}$ must approach the mean μ of the population as the number of observations increases.
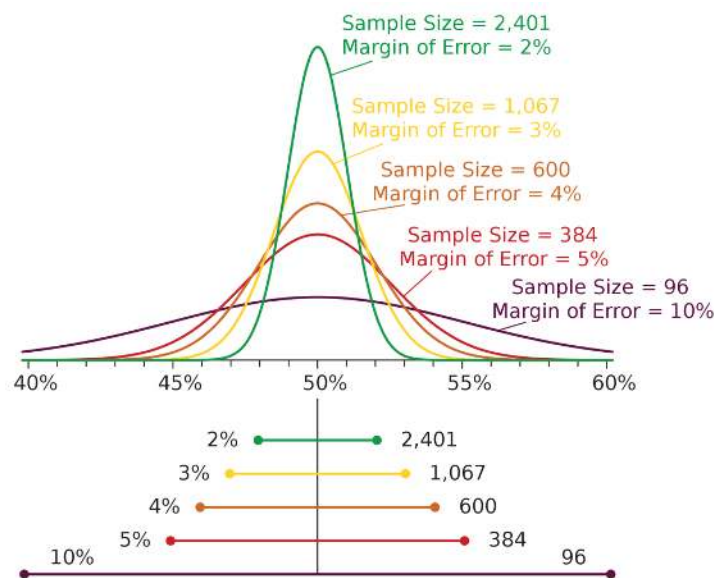
Figure 3: The Inverse Relationship: Increasing the Sample Size Reduces the Sampling Error

The Impact of Sampling Error in the Restructuring of Ecological Community Relationships: Applying an Equity Lens

Hirst & Jackson (2007)[7] report in their study that the structure of an ecological community is affected by how many species are present, their relative abundance, and how broadly each component species is distributed along environmental gradients. These differences in structure among sites provide the basic information used in *community analysis*. Ecologists often are confronted with various sources of error and complications in being able to adequately summarize patterns of species composition and the resemblance of sampling locations to one another. There are a variety of problems inherent in sampling ecological populations and communities. Specific problems include the low prevalence or the low numbers of rare species, the habitat or season when sampling is done, how representative samples are, choices in the type of sampling methodology, intrinsic variability in each species distribution, and methods of data analysis that effectively summarize patterns in species composition and site resemblance. As one might expect, study results showed that lower levels of sampling error tend to produce more accurate, less distorted results overall.

Now Try It Yourself

4.2(a) Sampling Distribution of Sample Means

Given the presence of sampling error, one may wonder if it is ever possible to generalize from a sample to a larger population. The theoretical model known as the *sampling distribution of means* has certain properties that give it an important role in the sampling

process. Levin and Fox 2006[8] point out these characteristics:

1. *The **sampling distribution of means** approximates a normal curve.* The sampling distribution of means is the probability distribution of a sample statistic that is formed when random samples of size *n* are repeatedly taken from a population (Larson & Farber 2019)[9]. If the raw data are normally distributed, then the distribution of sample means is normal regardless of sample size. Every sample statistic has a sampling distribution.

2. *The **mean of a sampling distribution of means (the mean of means)** is equal to the true population mean.* They are regarded as interchangeable values.

3. *The **standard deviation of a sampling distribution of means** is smaller than the standard deviation of the population.* The standard deviation of the sampling distribution of the sample means is called the **standard error of the mean (SEM).** The sample mean is more stable than the scores that it comprises. A concrete example is taking your blood pressure rate each day using a digital instrument that is sensitive to which arm is used, the time of day, whether you are rushing to work, whether you are waiting to see your doctor (White Coat Syndrome)[10], etc. The best approach to determining your blood pressure rate, therefore, might be to take it every day for a week at different times and then take the mean. This characteristic is at the core of making reliable inferences from sample to population.

A high standard error shows that sample means are widely spread around the population mean, so your sample may not closely represent your population. A low standard error shows that sample means are closely distributed around the population mean, which means that your sample is representative of your population. You can decrease standard error by increasing the sample size. The formula for the standard error of the mean is expressed as:

$$SE = \frac{\sigma}{\sqrt{n}} \qquad (1.4.1)$$

Study Tip: The **standard deviation and standard error** are often confused. The standard deviation measures the variability of individual data points, and the standard error measures the variability of a sample metric. The **Bootstrap** (sampling with replacement from a data set) is a powerful tool for assessing the variability of a sample statistic. Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples. This method takes the sample data that a study obtains and then resamples it over and over to create many simulated samples. Each of these simulated samples has its own properties, such as the mean. When you graph the distribution of these means on a histogram, you can observe the sampling distribution of the means. This process allows you to calculate standard errors, construct confidence intervals, and perform hypothesis testing for numerous types of sample statistics.

Now Try It Yourself

4.2(b). The Central Limit Theorem

The **Central Limit Theorem** forms the foundation for the inferential branch of statistics. This theorem describes the relationship between the sampling distribution of sample means and the population that the samples are derived from. Larson & Farber (2019) describes the Central Limit Theorem as:

1. If random samples of size *n*, where *n* ≥ 30, are drawn from any population with a mean μ and a standard deviation σ, then the sampling distribution of sample means approximates a normal distribution. The greater the sample size, the better the approximation.

2. If random samples of size *n* are drawn from a population that is normally distributed, then the sampling distribution of sample means is normally distributed for *any* sample size *n*.

Central Limit Theorem (CLT)
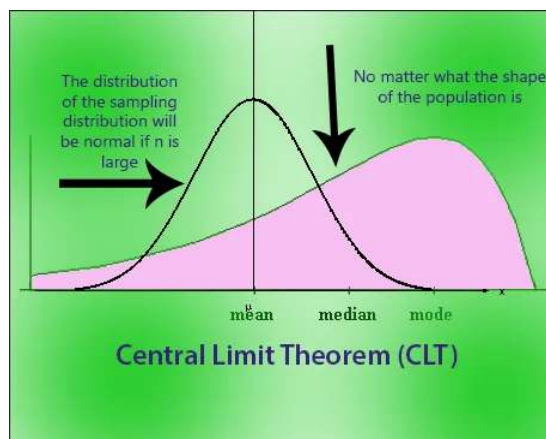
The Central Limit Theorem is useful when analyzing large data sets, as in criminal justice research, because it allows a statistician to assume that the sampling distribution of the mean will be normally distributed in most cases. This allows for easier statistical analysis and inference. Some statisticians view the Theorem as the cornerstone of modern statistics (Kwak & Kim 2017)[11].

Now Try It Yourself

Chapter 4: Summary

Studies are conducted on samples and not the entire population, given the practicalities involved. Sampling design or methods play an important role in helping ensure that sample results are accurate and can be generalizable to the population. The sample itself, therefore, must be adequate to represent the population. Quantitative sampling depends on two elements: random sampling and the sample size (power analysis [12]). Random sampling is required to ensure that there is no bias and that the entire sample is representative of the population. Power analysis is applied to determine the minimum sample size necessary to ensure that the data and sample are statistically significant.

The increasing volume of research studies often leads, in general, to increasing numbers of contradictory findings and conclusions. Although the differences observed may represent true differences, the results also may differ because of sampling variability, as some studies are performed on a limited number of participants. The LGBT Foundation's website[13] is highlighted as an example of being intentional in recruiting targeted participants for a multitude of research studies. Obviously, the Foundation prioritizes having members of the LGBT community heard in research studies, which in turn, underscores the importance of having a proper sample size that represents the entire LGBT population. The website states:

"There is a great deal of interesting and important research on LGBT communities being carried out at the moment, much of this research relies on LGBT people getting involved and sharing their experiences. If you are LGBT and would like to have your voice heard by participating in ongoing research, see below for projects you can get involved in.

- Ethnic and Sexual Minority Health Project
- Difficulties in Male Same-Sex Relationships
- Exploring the Variable Effects of Social Media Use on Mental Health Outcomes,
- Including Minority Stress
- Exploring the Intersection Between Bisexuality, Dementia and Adult Social Care
- LGBT+ Public Transport Experience Survey
- Exploring Gay Men's Experiences of Social Support, Relationships, and Difficult Emotions
- Supporting Access to Sexual Healthcare and Consultations: A Research Study
- Understanding the Paradigm Shift of Stereotyping and Conformity of Homosexuality in a UK Population: A Mixed Methods Exploration
- Experiences of Older LGBT People Using Community-Based Social Care Services, Groups and Activities

1. Wooditch A, Johnson NJ, Solymosi R, Ariza JM & Langton, S. (2021). Inferential Statistics. In: A Beginner's Guide to Statistics for Criminology and Criminal Justice Using R. Springer, Cham.
2. The author received her PhD from the Heller School for Social Policy and Management at Brandeis University, Waltham, MA. The Heller School is home to 11 Research Institutes: *Schneider Institutes for Health Policy; Institute for Healthcare Systems;*

Institute for Behavioral Health; Institute for Global Health and Development; Institute on Assets and Social Policy; Institute for Child, Youth and Family Policy; Center for Youth and Communities; Lurie Institute for Disability Policy; Sillerman Center for Advancement of Philanthropy; Center for Global Development and Sustainability; and the Relational Coordination Research Collaborative.

3. Augusty MK & Dizon JT (2020). The Role of Community-Based Organizations in Addressing Social Equity Among Deprived Sections in the Conflict Vulnerable Areas in Karnataka, India. *Asian Research Journal of Arts & Social Sciences*, 11(1): 24-41, 2020; Article No.ARJASS.56850

4. Moore DS, Notz WI, & Fligner, MA (2013). Essential Statistics: Second Edition. *W.H. Freeman and Company*: New York.

5. Jackson JS, Torres M, Caldwell CH, Neighbors HW, Nesse RM, Taylor RJ, Trierweiler SJ, & Williams DR (2004). The National Survey of American Life: A Study of Racial, Ethnic and Cultural Influences on Mental Disorders and Mental Health. *International Journal of Methods in Psychiatric Research*, Volume 13, Number 4.

6. Illiyasu R & Etikan I (2021). Comparison of Quota Sampling and Stratified Random Sampling. *International Journal of Biometrics*. https://www.researchgate.net/publication/354054682 . Retrieved on June 16, 2023.

7. Hirst CN & Jackson DA (2007). Restructuring Community Relationships: The Impact of Sampling Error, Ordination Approach, and Gradient Length. *Diversity and Distributions*, 13, 361-371.

8. Levin J & Fox JA (2006). Elementary Statistics in Social Research. Boston, MA: *Pearson Education, Inc*.

9. Larson R & Farber B (2019). Elementary Statistics. Boston, MA: *Pearson Education, Inc*.

10. White coat syndrome, or white coat hypertension, is the term to describe when you get a high blood pressure reading in a doctor's office and a normal reading at home. The anxiety of being around doctors in white coats can make your blood pressure rise (Cleveland Clinic). It always happens to me, so no need to do random sampling to assess its occurrence!

11. Kwak SG & Kim JH (2017). Central Limit Theorem: The Cornerstone of Modern Statistics. *Korean Journal Anesthesiology*, April; 70(2): 144-156.

12. Power analysis helps the statistician or researcher to determine the smallest sample size that is suitable to detect the effect of a given test at a desired level of significance.

13. Source: https://lgbt.foundation/research/participate. Retrieved on July 2, 2023.

1.5: Significance of Statistical Inference Methods

This chapter is a continuation of the previous chapter on inferential statistics.

> **◑ Learning Objectives**
>
> - Significance of Statistical Inference Methods
> - Confidence Intervals
> - Hypothesis Testing
> - Type I and Type II Errors
> - Scientific Racism

The importance of **statistical inference** is grounded in several assumptions. First, it addresses a particular type of uncertainty, namely that caused by having data from **random samples** rather than having complete knowledge of entire populations, processes, or distributions. Second, we cannot build statistical inference without first building an appreciation of **sample** *versus* **population**, of **description** *versus* **inference**, and of characteristics of samples giving estimates of characteristics of populations. Third, any conceptual approach to statistical inference must flow from some essential understanding of the nature and behavior of **sampling variation**. Finally, statistical inference is viewed as both an outcome and a reasoned process of creating or testing **probabilistic generalizations** from data. The prior four chapters introduced you in varying degrees to these points.

In this chapter, you will see how sampling distributions are used to test hypotheses and construct confidence intervals. Topics discussed in this chapter are:

- Implicit Assumptions in Making Statistical Inferences
- Significance of Statistical Inference Methods
  - Confidence Intervals
  - Tests of Significance

5.1. Implicit Assumptions in Making Statistical Inferences

One of the statistician's most important roles is the upfront contribution of planning a study, which includes identifying the implicit assumptions when making statistical inferences. It is important that these assumptions are met, and if they are not, the statistical inferences can result in seriously flawed conclusions in the study. According to Hahn & Meeker (1993)[1], the first assumption is that *the target population has been explicitly and precisely defined*. The second assumption is that *a specific listing (or another enumeration) of the population from which the samples have been selected has been made*. Third, the *data are assumed to be a random sample of the population*. The assumption of random sampling is critical. When the assumption of random sampling (or randomization) is not met, inferences to the population become difficult. In this case, statisticians should describe the sample and population in sufficient detail to justify that the sample was at least representative of the intended population.

5.2. The Significance of Statistical Inference Methods

You have now acquired a clear understanding of the difference between a *sample* (the observed) and the *population* (the unobserved). The extent to which credence should be placed in a given sample statistic as a description of the population parameter is the problem of inferring from the part to the whole. This is sometimes called the problem of **inductive inference**, or the problem of **generalization**.

Econometrics

[i-ˌkä-nə-ˈme-triks]

The use of statistical methods to develop theories or test existing hypotheses in economics or finance.

Investopedia
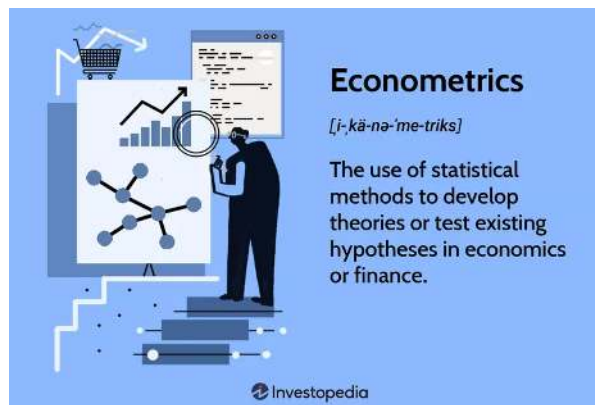
Using the field of Econometrics[2] as an example, Keuzenkamp & Magnus (1995)[3] describes the four types of statistical testing:

1. Theory Testing: This method consists of formulating hypotheses from which predictions of novel facts are deduced (the consequences).
2. Validity Testing: Performed in order to find out whether the statistical assumptions underlying some models are credible. In order to pursue a theory test, one first must be sure of the validity of the statistical assumptions that are made.
3. Simplification Testing: *Simplicity matters*. Simple models are typically preferred to complex ones. Rather than testing from general to simple, statisticians perform iterative simplification searches. In the study of statistics, we focus on mathematical distributions for the sake of simplicity and relevance to the real world. Understanding these distributions enables us to visualize the data more easily and build models more quickly.
4. Decision-Making: Based on statistical acceptance rules, this method can be important for process quality control and can be extended to appraising theories.

The type of statistical test one uses depends on the *type of study design, number of groups of comparison*, and *type of data* (i.e., continuous, dichotomous, and categorical) (Parab & Bhalerao 2010).[4]

Subsequent sections will discuss statistical inference methods that use the language of probability to estimate the value of a population parameter. The two most common methods are confidence intervals and tests of significance.

5.2(a). Confidence Intervals (CI)

A **confidence interval**, in statistics, refers to the probability that a population parameter falls between two set values. The sample is used to estimate the interval of probable values of the parameters of the population. It is a matter of convention to use the standard 95% confidence interval having the probability that there are 95 chances out of 100 of being right. There are five chances out of 100 of being wrong. Even when using the 95%[5] confidence interval, we must remember that the sample mean could be one of those five sample means that fall outside the established interval. A statistician never knows for sure.

According to Zhang, Hanik & Chaney (2008)[6], the CI has four noteworthy characteristics:

1. For a given sample size, at a given level of confidence, and using probability sampling, there can be infinitely many CIs for a particular population parameter. The point estimates and endpoints of these CIs vary due to sampling errors that occur each time a different sample is drawn.
2. The CI reported in a certain research study is just one of these infinitely many CIs.
3. The percentage of these CIs that contain the population parameter is the same as the level of confidence.
4. Whether a certain CI reported by a research study contains the population parameter is unknown. In other words, the level of confidence is applied to the infinitely many CIs, rather than a single CI reported by a single study.

> ### A Confidence Interval to Estimate Population Poverty in the United States: Applying the Equity Lens
>
> The U.S. Census Bureau routinely uses confidence levels of 90% in their surveys. "The number of people in poverty in the United States is 35,534,124 to 37,315,094″ means (35,534,124 to 37,315,094) is the confidence interval. Assuming the Census Bureau repeats the survey 1,000 times, the confidence level of 90% means that the stated number is between (35,534,124 to 37,315,094) at least 900 times. Maybe 36,000,000 people are in poverty–maybe less or greater. Any number in the interval is as expected.

Any confidence interval has two parts: an interval calculated from the data and a confidence level C. The confidence interval often has the form:

estimate $\pm$ margin of error

The confidence level is the success rate of the method that produces the interval, that is, the probability that the method will give a correct answer. The statistician chooses the confidence level, and the margin of error follows from this choice. When the data and the sample size remain the same, higher confidence results in a larger margin of error. There is a tradeoff between the confidence level and the margin of error. To obtain a smaller margin of error, the statistician must be able to accept a lower confidence. As the sample size increases, the margin of error gets smaller.

A level C **confidence interval for the mean μ** of a Normal population with known standard deviation σ, based on a simple random sample of size *n*, is given by:

$$\bar{X} \pm z * \frac{\sigma}{\sqrt{n}} \tag{1.5.1}$$

The critical value z* is chosen so that the standard Normal curve has an area C between − z* and z*. Below are the entries for the most common confidence levels:

Table 1: Confidence Levels and Corresponding *z-Scores*

Confidence Level C	90%	95%	99%
Critical value z*	1.645	1.960	2.576

We know from the Central Limit Theorem (chapter 4) that when $n \geq 30$, the sampling distribution of sample means approximates a normal distribution. The level of confidence C is the area under the standard normal curve between the *critical values, $-z_c$ and z_c*. **Critical values** are values that separate sample statistics that are probable from sample statistics that are improbable, or unusual. For instance, if C =95%, then 2.5% of the area lies to the left of $-z_c = -1.960$ and 2.5% lies to the right of $z_c = 1.960$, as shown in the Table below:

If C = 95%, then:

$$1 - C = .05 \tag{1.5.2}$$

$$\frac{1}{2}(1 - C) = .025 \tag{1.5.3}$$

$$-z_c = -1.960 \tag{1.5.4}$$

$$z_c = 1.960 \tag{1.5.5}$$
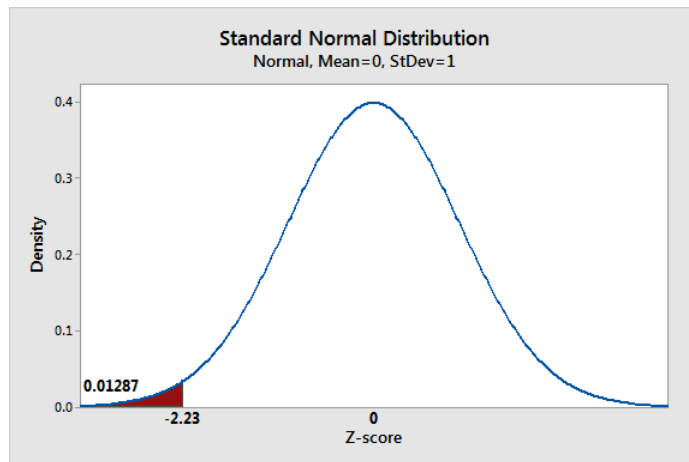
Figure 1: 95% Confidence Level with Critical Values Separating Left and Right Tails

An example of the meaning of the 95% confidence interval is as follows:

Suppose a medical researcher is interested in the prenatal care received by pregnant women in the inner city of Philadelphia. She breaks down the various sections of North Philly and computes the average number of gynecological check ups per pregnancy for all possible samples of size 20 and constructs the 95% confidence intervals using these sample means. Then 95% of these intervals would contain μ [population parameter]and 5% would not. Note that we cannot say that the probability is .95 that the interval from 2.6 to 3.4 gynecological checkups, for example, contains μ. Either the interval contains μ or it does not.

The 95% confidence interval of 2.6 to 3.4 was calculated based on the average number of gynecological check ups per pregnancy being 3, with a standard deviation of 1.

$$\bar{X} \pm z * \frac{\sigma}{n} \tag{1.5.6}$$

$$= 3 \pm 1.96 \frac{1}{\sqrt{20}} \tag{1.5.7}$$

$$= 2.6 \tag{1.5.8}$$

An Equitable Transition to Electric Transportation: Applying the Equity Lens

The newness of electric vehicles, their high upfront cost, the need for charging access, and other issues mean that equity has been overlooked (Hardman, Fleming, Khare & Ramadan 2021)[7]. Electric vehicle buyers are mostly male, high-income, highly educated homeowners who have multiple vehicles in their household and have access to charging at home. There is a need for a more equitable electric vehicle market so that the benefits of electrification are experienced by *all* and so that low-income households are not imposed with higher transportation costs. Low-income households, including those in underrepresented communities and in disadvantaged communities, could benefit from transportation electrification. These households are impacted by transportation emissions as they are more likely to reside in or near areas of high traffic and spend a higher proportion of their household income on transportation costs. Households in these communities are less likely to have charging at home or be able to afford to install home charging, have smaller budgets for vehicle purchases, and have fewer vehicles in their household. They are also less likely to have a regular place of work, which means they may not have workplace charging access (an alternative to home charging). These factors make plug-in electric vehicle (PEV) ownership more challenging,

A Step-by-Step Illustration: 95% Confidence Interval Using *z*

Suppose that the automobile company, Honda, wishes to address the mobility needs of underserved communities. As a separate research project, the company will work on the barriers and enablers of electrification. Until then, Honda determines the expected miles per gallon for a new 2025 HRV model that is designed to be far more affordable and efficient for low-income buyers. The company statistician knows from years of experience that not all cars are equivalents. She believes that a standard deviation of 4 miles per gallon ($\sigma = 4$) is expected due to parts and technician variations. To estimate the mean miles per gallon for the new model, she test runs a random sample of 100 cars off the assembly line and obtains a sample mean of 26 miles per gallon.

The following are steps to obtaining a 95% confidence interval for the mean miles per gallon for all cars of the HRV model.

Step 1: Obtain the mean for a random sample (which has already been provided).

$$N = 100; \bar{X} = 26 \tag{1.5.9}$$

Step 2: Calculate the standard error of the mean, knowing that σ = 4.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{1.5.10}$$

$$= \frac{4}{\sqrt{100}} \tag{1.5.11}$$

$$= \frac{4}{\sqrt{10}} \tag{1.5.12}$$

$$= .4 \tag{1.5.13}$$

Step 3: Calculate the margin of error by multiplying the standard error of the mean by 1.96, the value of z for a 95% confidence interval.

$$\text{Margin of error} = 1.96\sigma_{\bar{X}}$$

$$= 1.96(.4) \tag{1.5.14}$$

$$= .78 \tag{1.5.15}$$

Step 4: Add and subtract the margin of error from the sample mean to find the range of mean scores within which the population mean is expected to fall with 95% confidence.

$$\text{95\% Confidence Interval} = \bar{X} \pm 1.96\sigma_{\bar{X}}$$

$$= 26 \pm .78 \tag{1.5.16}$$

$$= 25.22 \tag{1.5.17}$$

Thus, the statistician can be 95% confident that the true mean miles per gallon for the new 2025 HRV model (μ) is between 25.22 and 26.78.

Important! After constructing a confidence interval, the results must be interpreted correctly. Consider the 95% confidence interval constructed in the above example. Because μ is a fixed value predetermined by the population, it is in the interval or not. It is *not* correct to say, "There is a 95% probability that the actual mean will be in the interval (25.22, 26.78)." This statement is wrong because it suggests that the value μ can vary, which is not true. The correct way to interpret this confidence interval is to say, "With 95% confidence, the mean is in the interval (25.22, 26.78)." This means that when a large number of samples are collected and a confidence interval is created for each sample, approximately 95% of these intervals will contain μ.

Now Try It Yourself

5.2(b). Tests of Significance

The second type of statistical inference is tests of significance. The tests of significance aid the statistician in making inferences from the *observed* sample to the *unobserved* population.
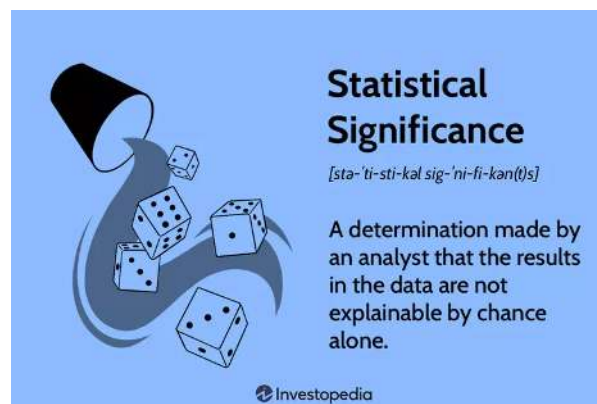
The test of significance provides a relevant and useful way of assessing the relative likelihood that a real difference exists and is worthy of interpretive attention, as opposed to the **hypothesis** that the set of data could be an arrangement lacking any type of pattern. The basic idea of statistical tests is like that of confidence intervals: *what would happen if we repeated the sample many times?*

A statistical test starts with a careful statement of the claims a statistician wants to compare. Because the reasoning of tests looks for evidence *against* a claim, we start with the claim we seek evidence against, such as "no effect or "no difference."

5.2(b)(1). The Null Hypothesis: No Difference Between Means

A **statistical hypothesis test** is a method of *statistical inference* used to decide whether the data at hand sufficiently supports a particular hypothesis.[8] In terms of selecting a statistical test, the most important question is, "What is the main hypothesis for the study?" In some cases, there is no hypothesis; the statistician just wants to "see what is there."

On the other hand, if a scientific question is to be examined by comparing two or more groups, one can perform a statistical test. For this, initially, a null hypothesis needs to be formulated, which states that there is no difference between the two groups. It is expected that at the end of the study, the null hypothesis is either rejected or not rejected (Parab & Bhalerao 2010).

The claim tested by a statistical test is called the **null hypothesis.** The test is designed to assess the strength of the evidence *against* the null hypothesis. The claim about the population that we are trying to find evidence *for* is the **alternative hypothesis**. The alternative hypothesis is one-sided if it states that the parameter is *larger than* or *smaller than* the null hypothesis value. It is two-sided if it states that the parameter is *different* (larger or smaller) from the null value (Moore, Notz & Fligner 2013).

The null hypothesis is abbreviated as H_o, and the alternative hypothesis as H_a. H_o is a statistical hypothesis that contains a statement of equality, such as $\leq$, $=$, or $\geq$. H_a is the complement of the null hypothesis. It is a statement that must be true if H_o is false. It contains a statement of strict inequality, such as $>$, $\neq$, or $<$. H_o is read as "H naught." H_a is read as "H sub-a."

You always begin a hypothesis test by assuming that the equality condition in the null hypothesis is true. When you perform a hypothesis test, you make one of two decisions:

1. Reject the null hypothesis or
2. Fail to reject the null hypothesis.

Know that because a decision is based on a sample rather than the entire population, there is the possibility of making the wrong decision. The only way to be absolutely certain of whether Ho is true or false is to test the entire population, which, in reality, may not be feasible. So, the statistician must accept the fact that the decision might be incorrect. As shown in Figure 2 below, there are two types of errors that can be made:
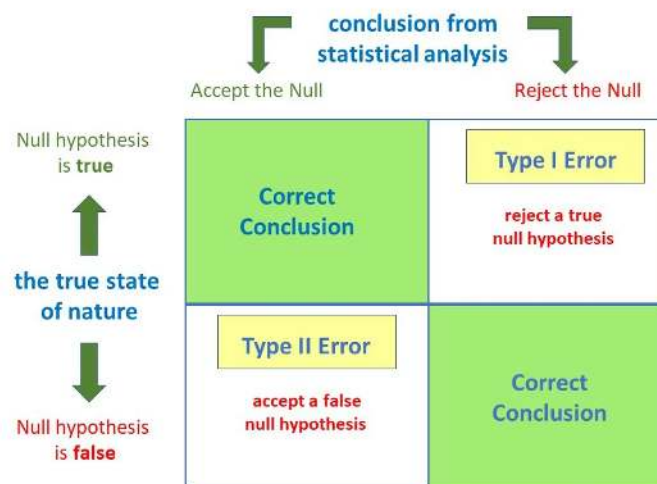
Figure 2: Type I and Type II Errors in Hypothesis Testing

A **Type I Error** occurs if the null hypothesis is rejected when it is true.

A **Type II Error** occurs if the null hypothesis is not rejected when it is false.

In a hypothesis test, the **level of significance** is the maximum allowable probability of making a Type I error. It is denoted by α, the lowercase Greek letter alpha. The probability of a Type II error is denoted by β, the lowercase Greek letter beta. By setting the level of significance at a small value, the probability of rejecting a true null hypothesis will be small. The commonly used levels of significance are: α = 0.01; α = 0.05; and α = 0.10. When α is decreasing (the maximum probability of making a Type I Error), it is likely that β is increasing. The value 1 – β is called the *power of the test*. It represents the probability of rejecting the null hypothesis when it is false.
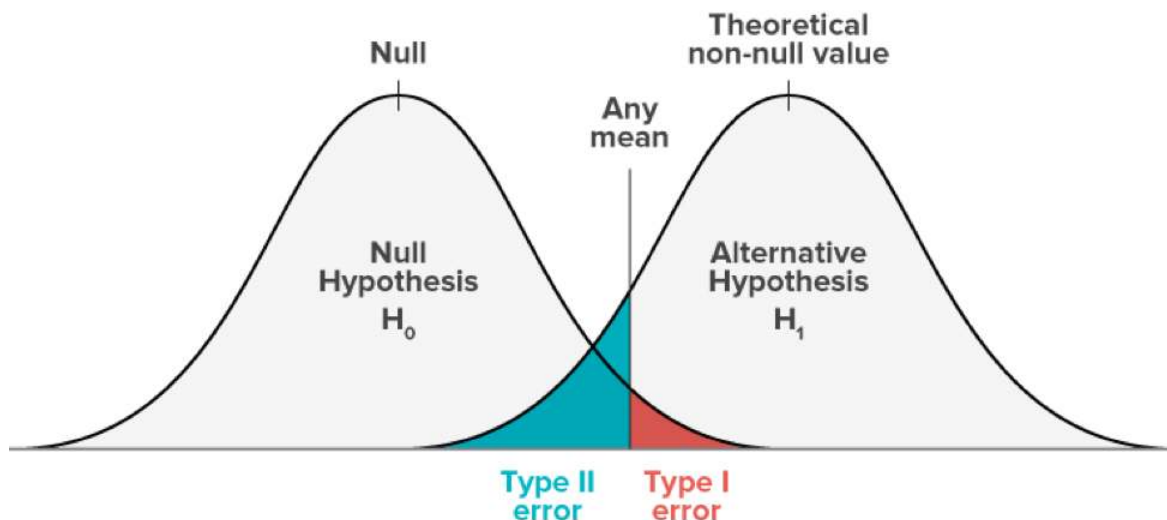
Hypotheses always refer to the population, not to a particular outcome. They are stated in terms of population parameters. Thus, for no difference between means, the null hypothesis can be symbolized as

$$\mu_1 = \mu_2$$

where μ_1 = mean of the first population

μ_2 = mean of the second population

Now Try It Yourself[9]

A test is based on a test statistic that measures how far the sample outcome is from the value stated by H_o. The **P-value** of a test is the probability that the test statistic will take a value at least as extreme as that actually observed. Small *P*-values indicate strong evidence to reject the null hypothesis provided by the data. However, a very low P-value does not constitute proof that the null hypothesis is false, only that it is *probably* false. Large *P*-values fail to give evidence against H_o. The most important task is to understand what a *P*-value says.

If the P-value is as small or smaller than a specified value α (alpha), the data are **statistically significant** at **significance level α**.

Decision Rule Based on *P*-Value[10]

To use a *P*-value to decide in a hypothesis test, compare the P-value with α.

If $P \le \alpha$, then reject H_o.

If $P > \alpha$, then fail to reject H_o.

[11]**When to Use the P-Value and How to Interpret It**

"Assume there is data collected from two samples and that the means of the two samples are different. In this case, there are two possibilities: the samples really have different means (averages), or the other possibility is that the difference that is observed is a coincidence of random sampling. However, there is no way to confirm any of these possibilities.

All the statistician can do is calculate the probabilities (known as the "*P*" value in statistics) of observing a difference between sample means in an experiment of the studied sample size. The value of *P* ranges from zero to one. If the *P* value is small, then the difference is quite unlikely to be caused by random sampling, or in other words, the difference between the two samples is real. One has to decide this value in advance, i.e., at which smallest accepted value of *P*, the difference will be considered as a real difference.

The *P* value represents a decreasing index of the reliability of a result. The higher the *P* value, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population. Specifically, the *P* value represents the probability of error that is involved in accepting our observed result as valid, i.e., as "representative of the population." For example, a *P* value of 0.05 (i.e., 1/20) indicates that there is a 5%

probability that the relation between the variables found in our sample is a "fluke." In other words, assuming that in the population, there was no relation between those variables whatsoever, and we were repeating experiments such as ours one after another, we could expect that approximately in every 20 replications of the experiment, there would be one in which the relation between the variables in question would be equal to or stronger than in ours. In many areas of research, the *P* value of 0.05 is customarily treated as a "cut-off" error level."

The P-value of a test depends on the nature of the test. As shown in Figure 3 below, there are three types of hypothesis tests—**left-tailed, right-tailed,** and **two-tailed** (where evidence that would support the alternative hypothesis could lie in either tail of the sampling distribution).
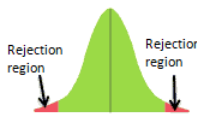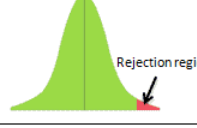
Z- Test	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)	Statistical conclusion
Two-tailed	$\mu = \mu_0$	$\mu \neq \mu_0$	Rejection region / Rejection region
Left-tailed	$\mu \geq \mu_0$	$\mu < \mu_0$	Rejection region
Right-tailed	$\mu \leq \mu_0$	$\mu > \mu_0$	Rejection region

Figure 3: Three Types of Hypothesis Tests

<u>In a left-tailed test</u>, the alternative hypothesis H_a contains the less-than inequality symbol ($<$):

$$H_{o:} \mu \geq k$$

$$H_{a:} \mu < k$$

<u>In a right-tailed test</u>, the alternative hypothesis H_a contains the greater-than-inequality symbol ($>$):

$$H_{o:} \mu \leq k$$

$$H_{a:} \mu > k$$

<u>In a two-tailed test</u>, the alternative hypothesis H_a contains the not-equal-to symbol ($\neq$):

$$H_{o:} \mu = k$$

$$H_{a:} \mu \neq k$$

Critical Regions for a Two-Tailed z Test
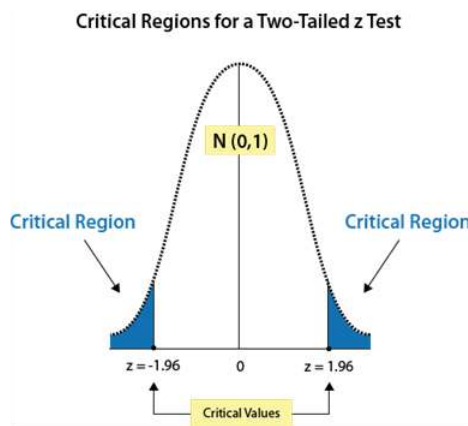
Steps in Hypothesis Testing

1. State mathematically and verbally the null and alternative hypotheses.

Ho: ? *Ha*: ?

2. Specify the level of significance.

α = ?

3. Obtain a random sample from the population.
4. Calculate the sample statistic (such as $\bar{x}$, $\hat{p}$, $s2$) corresponding to the parameter in the null hypothesis (such as μ, *p*, or σ2). This sample statistic is called the **test statistic**.
5. With the assumption that the null hypothesis is true, the test statistic is then converted to a **standardized test statistic**, such as *z, t (student's t-test) or χ2 (chi-square)*. The standardized test statistic is used in making the decision about the null hypothesis.
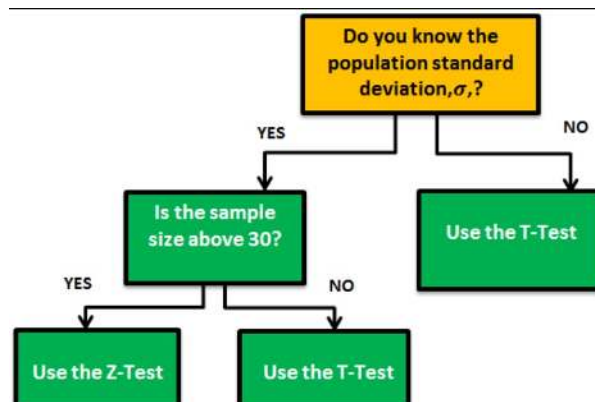
Student's *t* Test[12]: A *t*-test may be used to evaluate whether a single group differs from a known value (a one-sample *t*-test), whether two groups differ from each other (an independent two-sample *t*-test), or whether there is a significant difference in paired measurements (a paired, or dependent samples t-test). *t*-tests rely on an assumed "null hypothesis." You have to decide whether this is a one-tail *(Is it greater or less than?)* or a two-tail test (*Is there a difference?*) and choose the level of significance (Greek letter alpha, α). An alpha of .05 results in 95% confidence intervals and determines the cutoff for when *P*-values are considered statistically significant.

The assumptions of a *t*-test are:

- One variable of interest.
- Numeric data.
- Two groups or less.
- Random sample.
- Normally distributed.

Chi-Square Test (*χ2*): a statistical test commonly used to determine if there is a significant association between two variables.

For example[13], many Black and Latinx organizations receive relatively small program grants. To reverse this trend and to engender sustainability, foundations could create grantee cohorts in traditionally underserved communities and neighborhoods. This neighborhood cohort approach has been used successfully by foundations for decades for a variety of reasons, such as achieving efficiency and promoting collaboration. Unbound Philanthropy's Good Neighbor Committee, a staff initiative dedicated to grantmaking in the New York City metro, sought to reduce gang violence in central Long Island by supporting several organizations that were working on the problem from different angles. These included a youth development organization and a parent advocacy organization. The Chi-Square Test for Independence tests two hypotheses:

Null Hypothesis: *There is not a significant association between variables, the variables are independent of each other. Any association between variables is likely due to chance and sampling error.* For example, there is no significant association between Organization A (a youth development organization) and Organization B (a parent advocacy organization). Each organization's ability to reduce gang violence has nothing to do with the other.

Alternative Hypothesis: *There is a significant association (positive or negative) between variables, the variables are independent of each other. Any association between variables is not likely due to chance and sampling error.* For example, there is a significant association between Organization A (a youth development organization) and Organization B (a parent advocacy organization). Each organization's ability to reduce gang violence, to some degree, impacts the other.

1. Find the *P*-value.
2. Use this decision rule: If P-value is less than or equal to the level of significance, then reject the null hypothesis. If P-value is greater than the level of significance, then fail to reject the null hypothesis.
3. Write a statement to interpret the decision in the context of the original claim.

Now Try It Yourself

The Financial Risk Tolerance of Blacks, Hispanics, and Whites: Applying the Equity Lens

Yao, Gutter & Hannah (2005)[14] studied the effects of race and ethnicity on *financial risk tolerance*. Risk attitudes may affect investment behavior, so having an appropriate willingness to take financial risk is important in achieving investment goals. In turn, investment choices can affect retirement well-being, and, more specifically, the retirement adequacy of different racial and ethnic groups. This study focused on the expressed risk tolerance of Hispanics and Blacks compared to Whites because of the implications of investment behavior for future wealth differences and improving financial education programs.

Descriptive Statistics Results : White respondents are significantly more likely to be willing to take some risk (59%) than are Blacks (43%), who are significantly more likely to be willing to take some risk than Hispanics (36%). However, the pattern is reversed for willingness to take substantial risk, with only 4% of Whites but 5% of Blacks and 6% of Hispanics willing to take substantial risk.

Hypotheses Testing Results: The hypotheses are confirmed for substantial risk. Table 2 below summarizes the hypothesis tests. Based on the z-tests, Whites are significantly more likely than Blacks, and Blacks are significantly more likely than Hispanics to be willing to take some financial risks. For substantial risk, the results are the opposite of the hypotheses, as Whites are significantly less likely than Blacks and Hispanics to be willing to take substantial financial risks; and the difference between

Hispanics and Blacks is not significant. For high risk, the hypothesis that Whites are more likely to be willing to take risks than the other two groups are confirmed, but Hispanics are as willing to take high risks as Blacks.

Financial Risk tolerance levels	Z-test results	Logit results
Substantial	Not accepted: Hispanics = Blacks > Whites	Not accepted: Hispanics = Blacks > Whites
High	Partially Accepted: Whites > Hispanics = Blacks	Not accepted: Hispanics = Blacks = Whites
Some	Accepted: Whites > Blacks > Hispanics	Accepted: Whites > Blacks > Hispanics

> : Significantly greater at the .05 level or better

= : Not significantly different at the .05 level

Below are a sample of hypotheses found in the research literature on *race* and *ethnicity* and their intersection with other variables, such as age, socioeconomic status, and sexual preference:

- **Contact Hypothesis**: Interracial contact is associated with more positive racial attitudes, especially among Whites and some effects are appreciable.
- **Cumulative Disadvantage Hypothesis**: Predicts that initial advantages and disadvantages compound and produce diverging health trajectories as individuals age. *Rationale: Given the structural disadvantages that people of color face across multiple domains of the life course, the cumulative disadvantage hypothesis predicts that racial-ethnic health disparities increase with age.* (Brown, O'Rand & Adkins (2012)[15].
- **Ethnicity Hypothesis**: Ethnic minorities engage more in social activities than Whites of comparable socioeconomic status. *Rationale: The relatively smaller, more cohesive ethnic group is able to exert pressure on the individual member to conform to the norms of the respective ethnic affiliation.* (Antunes and Gaitz 1975)[16].
- **LGBT-POC (People of Color) Hypothesis**: These individuals may experience unique stressors associated with their dual minority status, including simultaneously being subjected to multiple forms of microaggressions (brief, daily assaults which can be social or environmental, verbal or nonverbal, intentional or unintentional). Within LGBT communities, LGBT-POC may experience racism in dating relationships and social networks. *Rationale: Racial/ethnic minority individuals have reported exclusion from LGBT community events and spaces. For example, certain gay bars have been noted for refusing entry of African Americans and providing poorer service to Black patrons.* (Balsam et al. 2011)[17].
- **Persistent Inequality Hypothesis**: Predicts that racial-ethnic inequalities in health remain stable with age. *Rationale: Socio-economic conditions and race-ethnicity are considered "fundamental causes" of disease and illness because of their persistent association with health over time regardless of changing intermediate mechanisms.* (Brown, O'Rand & Atkins (2012).
- **Statistical Discrimination or Profiling Hypothesis**: In this situation, an individual or firm uses overall beliefs about a group to make decisions about an individual from that group. The perceived group characteristics are assumed to apply to the individual. Thus, statistical discrimination may result in an individual member of the disadvantaged group being treated in a way that does not focus on his or her own capabilities. (National Academies Press 2004)[18].
- **Weathering Hypothesis**: Chronic exposure to social and economic disadvantage leads to accelerated decline in physical health outcomes and could partially explain racial disparities in a wide array of health conditions. (Forde, Crookes, Suglia & Demmer 2019)[19].

Chapter 5: Summary

Rather than summarize the content of this chapter, I decided to do something different. I would like to share with you something that I came across while researching articles for this book. I was introduced to the term *scientific racism* by Jay (2022)[20], which was deeply concerning. Its genesis is described in the following:

"Scientific racism," or more accurately pseudo-scientific racism (because racism is not scientific), was a way in which European colonial governments — and the statisticians they hired to do government surveys and data collection — justified their racist policies by using statistical measurements, often in an extremely biased and incorrect way. For example — if you've ever heard of

the infamous "skull measurements" used by European pseudo-scientists in the colonial era to try to demonstrate a (fake) correlation between skull size and intelligence — you should know they did that in order to put scientific backing behind claims such as "Africans, Native Americans, and Asians are less intelligent than Europeans due to smaller size." Of course, these claims are completely unscientific and baseless — but using scientific terminology and measurements as a backing was a way to soothe their guilty conscience.

…But how does this tie in with racism? Well — when colonial European "scientists" began measuring the heights, weights, and appearances of "races" in the world, they leaned on European statisticians like Galton to make conclusions based on the data. He and his contemporaries believed that the measurements of people in each of their "races" would follow a bell curve — the normal distribution.

If these "racial measurements" followed a normal distribution — well, since every normal curve has a specific mean and standard deviation — it "follows" that each "race" of people has an "average look." If this logic already sounds creepy to you — that's because it IS — and it's frankly mind-blowing how some of these statisticians convinced themselves their research was wholly "scientific."

Author's Comment: The above reminds us that the journey of "equity-mindedness" is really about doing *anti-racism, anti-oppression,* and *anti-colonialism* work. Thus, statistics must be *socially just* in its methods as well as its intentions.

Writing this OER book has reinforced the importance of upholding dharma (ethical and moral righteousness) when planning and implementing a study. Not only do we have to be careful about "bias creep," but be careful about the unintended deleterious effects of conclusions reached.

Thank you all for staying committed to this journey! I take this opportunity to acknowledge the group **Sweet Honey in the Rock,** which celebrates its 50th Anniversary this year of singing social justice songs. Their debut began in 1973 at a workshop at Howard University, a historically Black college in Washington, DC. It was their creative songs that motivated me each day to write this OER book.

(Photo with Permission of Artist)

1. Hahn GJ & Meeker WQ (1993). Assumptions for Statistical Inference. *The American Statistician*, February, Vol. 47, No.1.
2. , the Author's undergraduate degree is in Economics and Psychology. In graduate school, served as a Teaching Assistant in the graduate-level course on Econometrics.
3. Keuzenkamp HA & Magnus JR (1995). On Tests and Significance in Econometrics. *Journal of Econometrics*, 67, 5-24.
4. Parab S & Bhalerao S (2010). Choosing Statistical Tests. *International Journal of Ayurveda Research*. July-September; 1(3): 187-191.
5. , statisticians also use 90% and 99% levels of confidence.
6. Zhang J, Hanik BW & Chaney BH (2008).
7. Hardman S, Fleming KL, Khare E & Ramadan MM (2021). A Perspective on Equity in the Transition to Electric Vehicles. *MIT Science Policy Review*, August 30, 2021, Volume 2, 41-52.
8. Wikipedia. Statistical Hypothesis Testing. https://en.Wikipedia.org/wiki/Statis...thesis_testing. Retrieved on July 4, 2023.
9. Problems taken from Levin J & Fox JA (2006).
10. Larson R & Farber B (2019).
11. Example provided by Parab & Bhalerao (2010).

12. GraphPad. The Ultimate Guide to T Tests. https://www.graphpad.com/guides/the-...ciB0IHRlc3RzPw. Retrieved on July 4, 2023.

13. Pearce M & Eaton S. Equity, Inclusion and Diversity in Art and Culture Philanthropy. *Social Justice Funders Opportunity Brief, No. 3*. Brandeis University (Waltham, MA), *the* Sillerman Center for the Advancement of Philanthropy. https://heller.brandeis.edu/sillerma...ilanthropy.pdf. Retrieved on July 4, 2023.

14. Yao R, Gunner MS, & Hanna SD (2005). The Financial Risk Tolerance of Blacks, Hispanics, and Whites. *Financial Counseling and Planning*, Volume 16 (1), 51-62.

15. Brown TH, O'Rand AM & Adkins DE (2012). Race-Ethnicity and Health Trajectories: Tests of Three Hypotheses Across Multiple Groups and Outcomes. *Journal of Health and Social Behavior*, Volume 53, Issue 3.

16. Antunes G & Gaitz C (1975). Ethnicity and Participation: A Study of Mexican Americans, Blacks and Whites. *American Journal of Sociology*, 80 (March): 1192-1211.

17. Balsam KF, Molina Y, Beadnell B, Simoni J & Walters K (2011). Measuring Multiple Minority Stress: The LGBT People of Color Microaggressions Scale. *Cultural Diversity Ethnic Minority Psychology*, April; 17(2): 163-174.

18. National Academies Press (2004). Chapter 4: Theories of Discrimination. In the book, Measuring Discrimination. https://nap.nationalacademies.org/re...0887/chapter/7. Retrieved on July 3, 2023.

19. Forde AT, Crookes DM, Suglia SF, & Demmer RT (2019). The Weathering Hypothesis as an Explanation for Racial Disparities in Health. *Annals of Epidemiology*, May, 33:1.

20. Jay R (2022). P-values: A Legacy of "Scientific Racism." https://towardsdatascience.com/p-val...m-d906f6349fc7. Retrieved on July 3, 2023.

1.6: Correlation and Regression Analysis

> **Learning Objectives**
> - Inferential Analysis: A Recap
> - Correlation Analysis
> - Constructing a Scatter Plot
> - Correlation is Not the Same as Causation
> - Regression Analysis
> - Dependent vs. Independent Variables

We are all guilty of jumping to conclusions from time to time. It is easy to jump to inaccurate conclusions at work, with our families and friends, in relationships, and even when we first hear something on the news. When we do this, we are essentially generalizing, but what if we could make these generalizations more accurate?

A case in point is when I first heard on the news about the recent U.S. Supreme Court landmark decision on June 29, 2023[1]. The June 29th ruling makes it unlawful for colleges to take *race* into consideration as a specific factor in admissions. The Court's conservative-liberal split effectively overturned cases reaching back 45 years in invalidating admission plans at Harvard University and the University of North Carolina, the nation's oldest private and public colleges, respectively.

Besides feeling disappointed but not surprised, my initial reaction was that it was a blow to institutions of higher education, especially those who work diligently and authentically to achieve diverse student bodies. After delving more into the issue, I learned that the Court's decision is not just limited to the educational system but to the health care system as well. I had not thought about its ripple effect on the medical profession and how it would exacerbate health disparities among people of color, as discussed below.

> ### Impact of Affirmative Action Ruling on the Health Care System: Applying an Equity Lens
>
> With fewer Blacks and Latinos attending medical school, experts say that medical schools will be even less diverse after the affirmative action ruling. Justice Sonia Sotomayor, a Latina, wrote in her powerful dissenting opinion that *"…increasing the number of students from underrepresented backgrounds who join the ranks of medical professionals improves healthcare access and health outcomes in medically underserved communities."*
>
> The Association of American Medical Colleges (AAMC)[2] underscores Justice Sotomayor's point. The AAMC, along with 45 health professional and educational organizations, submitted prior to the ruling an amicus curiae brief to the U.S. Supreme Court emphasizing (1) there is an ongoing underrepresentation of certain racial and ethnic groups in medicine; (2) studies have repeatedly shown that racially and ethnically diverse health care teams produce better and more equitable outcomes for patients; and (3) physicians who train and work alongside racially or ethnically diverse peers have higher cultural competence and are able to help eliminate socio-cultural barriers to care and avoid stereotypes about patients from different backgrounds.[3]
>
> After the ruling, the AAMC stated, *"Today's decision demonstrates a lack of understanding of the critical benefits of racial and ethnic diversity in educational settings and a failure to recognize the urgent need to address health inequities in our country."*[4]

Dr. Uche Blackstock, Founder & CEO of Advancing Health Equity, is a second-generation Black female physician. Dr. Blackstock adds, *"There are only 5% of physicians who are Black, while Blacks represent 13% of the population. We have copious amounts of data and research that outcomes improve for Black patients when there is a diverse healthcare workforce. This ruling is going to have detrimental consequences, meaning life or death for Black communities. We already have the shortest life expectancy of any demographic group. We are more likely to die from a pregnancy-related complication. Black infants are twice as likely to die in their first year compared to White infants. We are in a crisis right now, and this ruling is going to exacerbate that crisis…and it's going to happen for generations and generations to come. We need people to connect the dots. It is not about the success of one person getting into medical school. It is about the ripple effect of what happens when we admit a diverse medical student body which has serious implications for communities of color, in particular."* [5]

6.1. Inferential Analysis: A Recap

Inferential analysis is simply what we use to try to infer from a sample of data what the population might think or show. It is a method that is used to draw conclusions, that is, to infer or conclude trends about a larger population based on the samples analyzed.

In previous chapters, we learned that we could go about this in basically two ways:

1. Estimating Parameters: Taking a statistic from a sample of data (like the sample mean) and using it to describe the population (population mean). The sample is used to estimate a value that describes the entire population, in addition to a confidence interval. Then, the estimate is created.
2. Hypothesis Tests: Data is used to determine if it is strong enough to support or reject an assumption.

Inferential analysis allows us to study the relationship between variables within a sample, allowing for conclusions and generalizations that accurately represent the population. There are many types of inferential analysis tests in the field of statistics. We will review the two most common methods: *correlation analysis* and *linear regression analysis (prediction and forecasting)*. They are the most commonly used techniques for investigating the relationship between two quantitative variables. Both methods are used quite frequently in disciplines such as economics, healthcare, engineering, physical engineering, and the social sciences.

6.2 Correlation Analysis
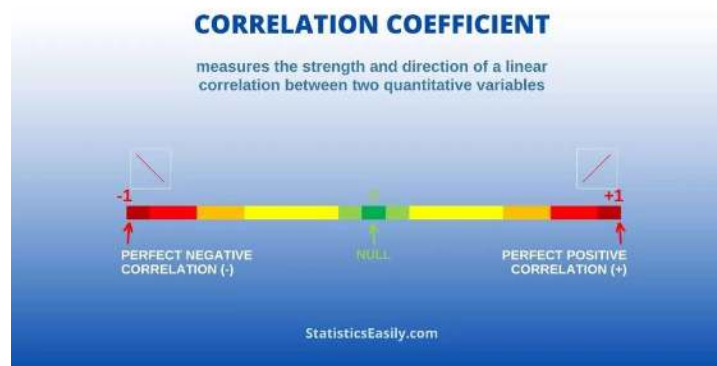
Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It is a common tool for describing simple relationships without making a statement about cause and effect. **Correlation does not imply causation!**

Using the affirmative action ruling, we might say that we have noticed a correlation between where a Supreme Court Justice is on the ideological spectrum (i.e., their political leanings) and whether they are a Conservative, Moderate, or Liberal.[6] A growing body of academic research has confirmed this understanding, as scholars have found that the Justices largely vote in consonance with their perceived values. The simplest way to approximate the ideological leanings of the Supreme Court Justices is by the political party of the Presidents who appointed them.[7]

However, in statistical terms, we use correlation to denote the association between two (or more) quantitative variables, that is, variables that can be "measured". The association is *linear*, fundamentally based on the assumption of a straight-line [linear]

relationship between the quantitative variables. Besides *direction*, correlation also looks at the *strength* of the relationship between the variables.

The data can be represented by the ordered pairs (x,y) where *x* is the **independent (or explanatory variable** and *y* is the **dependent (or response variable)**. A response variable measures the outcome of a study. An explanatory variable may explain or influence changes in a response variable, as revealed in the graph below. However, a cause-and-effect relationship is not necessary for the distinction between explanatory and response variables.
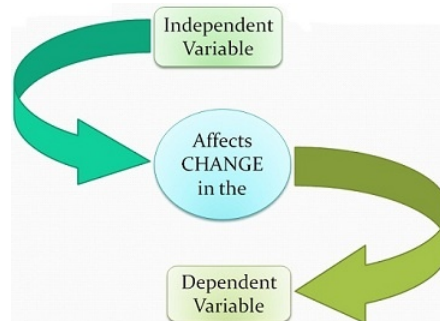
In Figure 1 below, the independent variable belongs on the x-axis (horizontal line) of the graph and the dependent variable belongs on the y-axis (vertical line). The x- and y-axes cross at a point referred to as the origin, where the coordinates are (0,0).
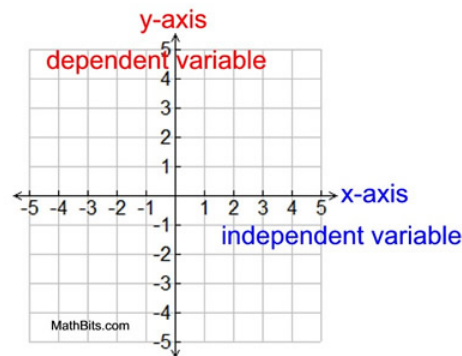
Figure 1: Dependent and Independent Variables on the x- and y- Axes

6.2(a). Constructing a Scatter Plot

The most useful graph for displaying the relationship between two quantitative variables is a **scatter plot.** A scatter plot can be used to determine whether a linear (straight line) correlation exists between two variables. Each individual data appears as the point in the plot fixed by the values of both variables for that individual. The scatter plot below in Figure 2 shows several types of correlation.

A **positive correlation** exists when two variables operate in unison so that when one variable rises or falls, the other does the same. A **negative correlation** exists when two variables move in opposition to one another so that when one variable rises, the other falls.

SCATTERPLOTS & CORRELATION

Correlation - indicates a relationship (connection) between two sets of data.
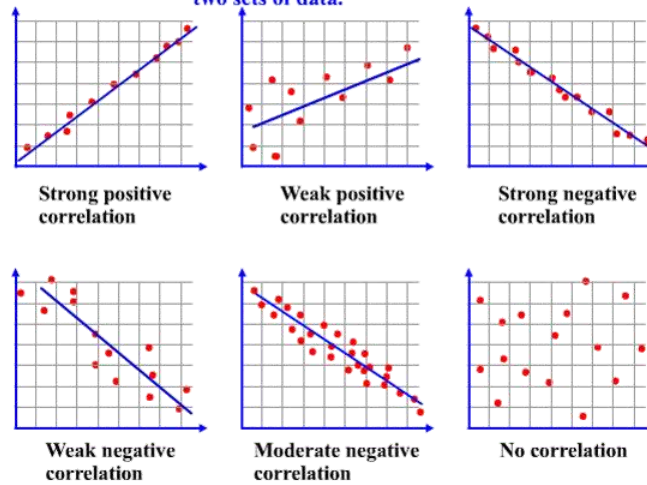
Figure 2: A Typology of Correlations Using Scatterplots

Curvilinear Relationships: As shown in Figure 3, it is important to note that not all relationships between *x* and *y* can be a straight line. There are many **curvilinear** relationships indicating that one variable increases as the other variable increases until the relationship reverses itself so that one variable finally decreases while the other continues to increase (Levin & Fox 2006).
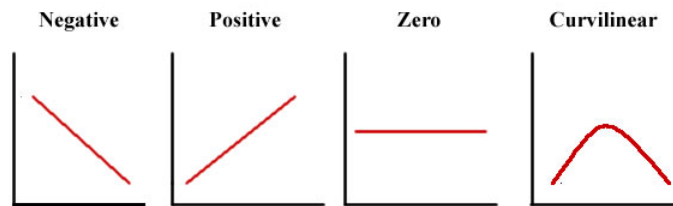
Now Try It Yourself

How to Examine a Scatter Plot

As suggested by Moore, Notz & Fligner (2013), in any graph of data, look for the overall pattern and for striking deviations. The overall pattern of a scatter plot can be described by the direction, form, and strength of the relationship. An important kind of deviation is an outlier, an individual value that falls outside the overall pattern of the relationship.

You interpret a scatterplot by looking for trends in the data as you go from left to right: If the data show an uphill pattern as you move from left to right, this indicates a positive relationship between *x* and *y*. As the *x*-values increase (move right), the *y*-values tend to increase (move up).

6.2(b). Correlation Coefficient

A scatterplot is considered a useful first step. The second step is to calculate the **correlation coefficient**—considered a more precise measure of the strength and direction of a linear correlation between two variables. The symbol *r* represents the sample correlation coefficient. A formula for *r* is

Correlation Coefficient Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

where *n* is the number of pairs of data. The **population correlation coefficient** is represented by ρ (the lowercase Greek letter rho, pronounced "row").

Calculating a Correlation Coefficient (Words and Symbols)

1. Find the sum of the *x*-values. (\sum x)
2. Find the sum of the *y*-values. (\sum y)
3. Multiply each *x*-value by its corresponding *y*-value and find the sum.(\sum xy)
4. Square each *x*-value and find the sum. (\sum x^{2})
5. Square each *y*-value and find the sum. (\sum y^{2})
6. Use these five sums and *n* to calculate the correlation coefficient.
7. Interpret the correlation coefficient *r*.

Figure 4 below shows that the range of the correlation coefficient is -1 to 1, inclusive. When *x* and *y* have a strong positive linear relationship, *r* is close to 1. When *x* and *y* have a strong negative relationship, *r* is close to -1. When *x* and *y* have a perfect positive linear correlation or perfect negative linear correlation, *r* is equal to 1 or -1, respectively. When there is no linear correlation, r is close to 0. Note: When r is close to 0, it does not mean that there is no relationship between x and y; it is just that there is no *linear* relationship.

Figure 4: Correlation Coefficients: Size and Interpretations

Size of Correlation	Interpretation
.90 to 1.00 (-.90 to -1.00)	Very high positive (negative) correlation
.70 to .90 (-.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (-.30 to -.50)	Low positive (negative) correlation
.00 to .30 (.00 to -.30)	Little, if any, correlation

In correlation, the statistician is interested in the degree of association between two variables. With the aid of the correlation coefficient known as Pearson's *r*, it is possible to obtain a precise measure of both the *strength*—from .00 to 1.00—and *direction*—positive versus negative—of a relationship between two variables that have been measured at the interval level. Levin and Fox (2006) state that if a statistician "has taken a random sample of scores, he or she may also compute a *t* ratio to determine whether *x* and *y* exist in the population, and is not due merely to sampling error."

It is only appropriate to use Pearson's correlation if your data "passes" four assumptions that are required for Pearson's correlation to give you a valid result. The four assumptions[8] are:

- **Assumption #1:** Your two variables should be measured at the **interval** or **ratio level** (i.e., they are **continuous**). Examples of variables that meet this criterion include intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.
- **Assumption #2:** There is a **linear relationship** between the two variables. Suggest creating a scatterplot, where you can plot one variable against the other variable and then visually inspect the scatterplot to check for linearity. Your scatterplot may look something like one of the following:

Linear v. Non-Linear Functions

How can you tell if a function is linear?
- Graph is a straight line
- Equation can be written in the form $y = mx + b$

Function A
$3x + y = 7$
$-3x \quad -3x$
$y = -3x + 7$

Linear

Function B
$y = 9(1 + 0.48)^x$

Non-Linear
(Exponential)

Function C
$y = x^2 - 2x + 10$

Non-Linear
(Quadratic)

Function D
$y = -2$
$y = mx + b$
$y = 0x - 2$

Linear

SCHOOL²¹
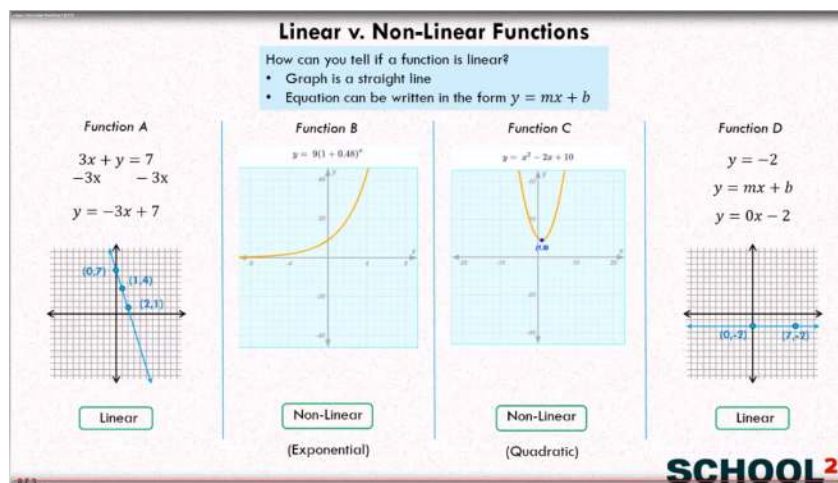
- **Assumption #3:** There should be **no significant outliers**. Outliers are simply single data points within your data that do not follow the usual pattern (e.g., in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put the person in the top 1% of IQ scores globally). The following scatterplots highlight the potential impact of outliers:

Correlation

Strength of relationship

Negative · · · · · · · · · · Positive

Perfect	Very large	Large	Medium	Small	Zero	Small	Medium	Large	Very large	Perfect
–1.00	–.75	–.50	–.30	–.10	0.0	+.10	+.30	+.50	+.75	+1.00

Perfect negative relationship (a) · Moderate negative relationship (b) · No relationship (c) · Moderate positive relationship (d) · Perfect positive relationship (e)
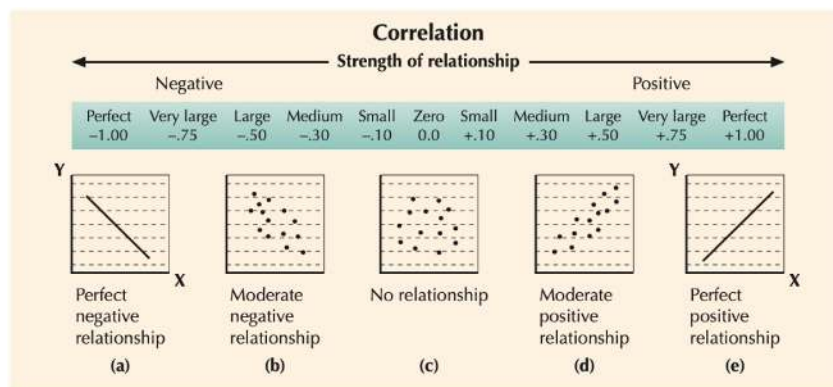
Pearson's correlation coefficient, r, is sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. Therefore, in some cases, including outliers in your analysis can lead to misleading results. Therefore, it is best if there are no outliers or they are kept to a minimum.

- **Assumption #4:** Your variables should be **approximately normally distributed**. In order to assess the statistical significance of the Pearson correlation, you need to have bivariate normality, but this assumption is difficult to assess, so a simpler method is more commonly used. This simpler method involves determining the normality of each variable separately.

6.2(b)(1). Significance Test for Correlation

The null hypothesis (H_0) and alternative hypothesis (H_1) of the significance test for correlation can be expressed in the following ways, depending on whether a one-tailed or two-tailed test is requested:

Two-Tailed Significance Test:

H_0: $\rho = 0$ (the population correlation coefficient is 0; there is no association).

H_1: $\rho \neq 0$ (the population correlation coefficient is not 0; a nonzero correlation could exist).

One-Tailed Significance Test:

H_0: $\rho = 0$ (the population correlation coefficient is 0; there is no association).

H_1: $\rho > 0$ (the population correlation coefficient is greater than 0; a positive correlation could exist).

OR

H_1: $\rho < 0$ (the population correlation coefficient is less than 0; a negative correlation could exist)

where ρ is the population correlation coefficient.

Correlation ≠ Causation

Correlation analysis measures a relationship or an association; it does not define the explanation or its basis. The purpose is to measure the closeness of the linear relationship between the defined variables. The correlation coefficient indicates how closely the data fit a linear pattern. One of the most frequent and serious misuses of correlation analysis is to interpret high causation between the variables. A correlation between variables does not automatically mean that the change in one variable is the *cause* of the change in the values of the other variable. **Causation** indicates that one event is the result of the occurrence of the other event, that is, there is a causal relationship between the two events.

When there is a significant correlation between two variables (such as meditation and stress reduction), Larson & Farber (2019) suggest that the statistician consider these possibilities:

- *Is there a direct cause-and-effect relationship between the variables?*

    That is, does *x* cause *y*. For instance, consider the relationship between *meditation* and *stress reduction*. Today, people can be stressed from overwork, job security, information overload, and the increasing pace of life (Deshpande 2012)[9]. Meditation has been recommended and studied in relation to stress and has proven to be highly beneficial in alleviating stress and its effects (e.g., higher-order functions become stronger while lower-order brain activities decrease). A simple way of thinking about meditation is that it trains your attention to achieve a mental state of calm concentration and positive emotions.

    It is reasonable to conclude that an increase in meditation (x) can result in lower levels of stress (y), showing a negative linear correlation.

- *Is there a reverse cause-and-effect between the variables?*

    That is, does *y* cause *x*? For instance, do lower levels of stress (*y*) decrease one's interest in meditation (*x*)? These variables have a positive linear correlation. It is possible to conclude that lower levels of stress affect one's desire to meditate.

- *Is it possible that the relationship between the variables can be caused by a third variable or perhaps a combination of several other variables?*

    The meditation-stress relationship can be confounded by other factors, such as age, gender, race/ethnicity, income, zip code, religious beliefs, spiritual health, well-being, parenting stress, prior trauma experiences, etc. Variables that have an effect on the variables in a study but are not included in the study are called **lurking variables.**

- *Is it possible that the relationship between two variables may be a coincidence?*

  A coincidence has been defined as a "surprising concurrence of events, perceived as meaningfully related, with no apparent causal connection." (Spiegelhalter 2012)[10] There are several research studies by Harvard University, UCLA Health, and others showing the efficacy of meditation in reducing stress and anxiety; thus, the relationship does not seem to be coincidental.

Impact of Racial Microaggressions on Mental Health: Applying an Equity Lens

Racial microaggressions are defined as "brief and commonplace daily verbal, behavioral, and environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, negative racial slights and insults to the target person or group." (Sue et al. 2007)[11] Nadal et al. (2014)[12] conducted a correlation and linear regression analysis to test the hypothesis that racial microaggressions would have a negative correlation with mental health. Results from the study suggest that:

(1) there is a negative significant relationship between racial microaggressions and mental health: *individuals who perceive and experience microaggressions in their lives are likely to exhibit negative mental health symptoms, such as depression, anxiety, a negative view of the world, and a lack of behavioral control;*

(2) specific types of microaggressions may be correlated with negative mental health symptoms, namely depression and a negative view of the world; and

(3) different Black, Latina/o, Asian, and multiracial participants may experience a greater number of microaggressions than White participants, and there are no significant differences in the total amount of microaggressions among Black, Latina/o, Asian, and multiracial participants.

Additionally, Carter, Kirkinis, & Johnson (2019)[13] found strong relationships between race-based traumatic stress and trauma symptoms as per the Trauma Symptom Checklist-40. This indicates that race-based traumatic stress is significantly related to trauma reactions (e.g., disassociation, anxiety, depression, sexual problems, and sleep disturbance), especially in instances where individuals have confirmed that negative race-based experiences are stressful.

Now Try It Yourself

6.3 Regression Analysis

In your algebra course, you learned that two variables are sometimes related in a **deterministic** way, meaning that given a value for one variable, the value of the other variable is exactly determined by the first without any error, as in the equation y = 60x for converting x from minutes to hours. However, in statistics, there is a focus on **probabilistic** models, which are equations with a variable that is not determined completely by the other variable.

Regression is closely allied with correlation in that we are concerned with specifying the nature of the relationship between two or more variables. We specify one variable as **dependent (response)** and one (or more) as **independent (explanatory),** where one or more variables are believed to influence the other. For example, *stress reduction* is dependent, and *meditation* is independent.

Regression

[ri-'gre-shan]

A statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Investopedia
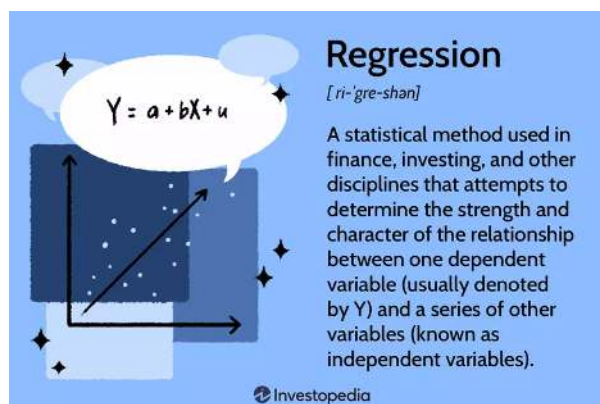
In regression analysis, a mathematical equation is used to predict the value of the dependent variable (denoted *Y)* on the basis of the independent variable (denoted *X).* Regression equations are frequently used to project the impact of the independent variable *X* beyond its range in the sample.

$$y = a + bx$$

Intercept   Slope

The term *a* is called the *Y-intercept.* It refers to the expected level of *Y* when *X* = 0. The term *b* is called the *slope* (or the *regression coefficient*) for *X.* This represents the amount that *Y* changes (increases or decreases) for each change of one unit in *X.*

Regression Formula

$$Y = a + bX + \epsilon$$

WallStreetMojo

Finally, *e* is called the *error term* or *disturbance term.* It represents the amount of *Y* that cannot be accounted for by *a* and *bX.*

6.3(a) Finding the Equation of a Regression Line

After verifying that the linear correlation between two variables is significant, the next step is to determine the equation of the line that best models the data. This line is called the **regression line,** whose equation is used to predict the value of *Y* for a given value of *X.* A regression line, also called a **line of best fit,** is the line for which the **sum of the squares of the residuals** is a minimum.
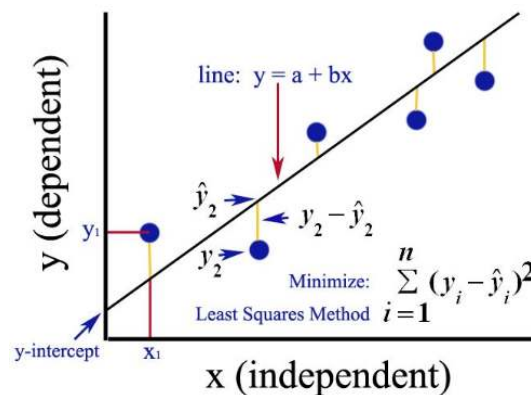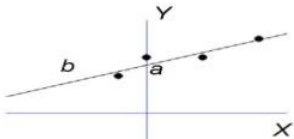
A **residual** is the difference between the observed *y*-value of a data point and the predicted *y*-value on the regression line for the *x*-coordinate of the data point. A residual is positive when the data point is above the line, negative when the point is below the line, and zero when the observed *y*-value equals the predicted *y*-value. (Larson & Farber 2019).

The equation of a regression line for an independent variable *X* and a dependent variable *Y* is

$$\hat{Y} = bX + a$$

where ŷ (pronounced y-hat) is the predicted *Y*-value for a given *X*-value. The slope *b* and *Y*-intercept *a* are given by

Slope

$$b = r\frac{S_Y}{S_X} \tag{1.6.1}$$

where r is the linear correlation coefficient, *Sy* is the standard deviation of the y values, and *Sx* is the standard deviation of the x-values.

Y-intercept

$$a = \bar{Y} - b\bar{X} \tag{1.6.2}$$

Before you calculate the regression line, you will need the following values:

- The mean of the x values ($\bar{x}$)
- The mean of the y values ($\bar{y}$)
- The standard deviation of x values (S_x)
- The standard deviation of y values (S_y)
- Correlation between x and y (*r* coefficient)

Now start by working out the slope, which represents the change in *y* over the change in *x*. To calculate the slope for a regression line, divide the standard deviation of *y* values by the standard deviation of *x* values and then multiply this by the **correlation** between *x* and *y*.

To find the *y*-intercept, you must then multiply the slope by the mean of the *x* values and then subtract this result from the mean of the *y* values. The y-intercept is an important part of the regression equation but never as substantively important as the slope.

Round both the slope and y-intercept to three significant digits.

6.3(b). Requirements for Regression

Triola (2011) reports the following requirements or conditions that must be met to perform a regression analysis:

1. The sample of paired data (*x, y*) is a *random* sample of quantitative data.
2. Visual examination of the scatter plot shows that the points approximate a straight-line pattern. Making a scatter plot may show that the relationship between two variables is linear. If the relationship looks linear, then use correlation and regression to describe the relationship between the two variables.
3. Outliers can have a strong effect on the regression equation, so remove any outliers if they are known to be errors. Consider the effects of any outliers that are not known errors.
4. For each fixed value of *x*, the corresponding values of *y* have a normal distribution.
5. For the different fixed values of *x*, the distributions of the corresponding *y*-values all have the same standard deviation.
6. For the different fixed values of *x*, the distributions of the corresponding *y*-values have means that lie along the same straight line.

6.3(c). Prediction Errors

When the correlation is perfect (r = +1 or -1), all the points lie precisely on the regression line, and all the *Y* values can be predicted perfectly on the basis of *X*.

In the more usual case, the line only comes close to the actual points (the stronger the correlation, the closer the fit of the points to the line).

The difference between the points (observed data) and the regression line (the predicted values) is the error or disturbance term (e)-

$e = Y - \hat{Y}$

The predictive value of a regression line can be assessed by the magnitude of the error term.

The larger the error, the poorer the regression line as a prediction device. (Levin & Fox 2006)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/statisticsthroughequitylens/?p=58#oembed-1

6.3(d). Variation About a Regression Line: The Coefficient of Determination

There are three types of variation about a regression line. They are:

- Total Deviation = $y_1 - \bar{y}$. The sum of the explained and unexplained variance is equal to the total variation.
- Explained Deviation = $\hat{y} - \bar{y}$. The explained variation can be explained by the relationship between x and y.
- Unexplained Deviation = $y_1 - \hat{y}_i$. The unexplained variation cannot be explained by the relationship between x and y and is due to other factors.

Coefficient of Determination

[kō-ə-'fi-shənt əv di-,tər-mə-'nā-shən]

A statistical measurement that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event.

Investopedia
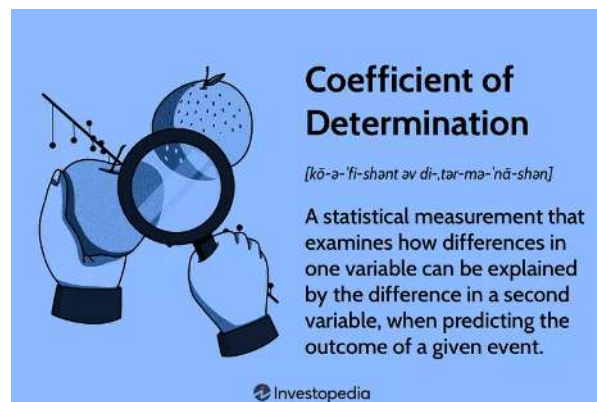
You already know how to calculate the correlation coefficient r. The square of this coefficient is called the **coefficient of determination.**

R-Squared

['är 'skwerd]

A statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Investopedia
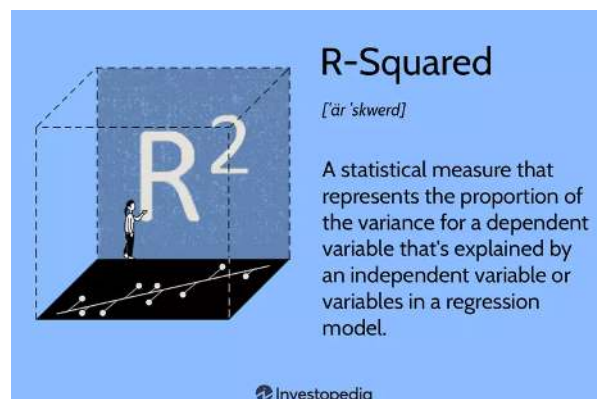
It is equal to the ratio of the explained variation to the total variation. It is important that the coefficient of determination is interpreted correctly. The coefficient of determination or **R squared** method is the proportion of the variance in the dependent variable that is predicted from the independent variable. It indicates the level of variation in the given data set.

Here are some guiding principles when interpreting the coefficient of determination:

- The coefficient of determination is the square of the correlation (r). Thus it ranges from 0 to 1.
- If R^2 is equal to 0, then the dependent variable cannot be predicted from the independent variable.
- If R^2 is equal to 1, then the dependent variable can be predicted from the independent variable without any error.
- If R^2 is between 0 and 1, then it indicates the extent to which the dependent variable can be predictable. For example, an r^2 of .810 means that 81% of the variation in y can be explained by the relationship between x and y.
- The **coefficient of non-determination** is $1 - R^2$.

> ### Evidence-Based Regression Analysis on Homeless Shelter Stays in Boston, MA: Applying an Equity Lens
>
> Researchers at Bentley University, Hao, Garfield & Purao (2022)[14], conducted a study to identify determinants that contribute to the length of a homeless shelter stay. The source of data was the Homeless Management Information Systems from Boston, MA, which contained 44,197 shelter stays for 17,070 adults between January 2014 and May 2018. Statistical and regression analyses show that factors that contribute to the length of a homeless shelter stay include being *female, a senior, disabled, Hispanic, or being Asian, or Black.* A significant fraction of homeless shelter stays (76%) are experienced by individuals with at least one of three disabilities: physical disability, mental health issues, or substance use disorder. Recidivism also contributes to longer homeless shelter stays.
>
> This finding aligns with prior studies that have shown that women stay in homeless programs significantly (74%) longer than men. This is concerning, as other work has reported that the rise in the number of unsheltered homeless women (12%) is outpacing that of unsheltered homeless men (7%). These findings also have cost implications. The per-person cost for first-time homeless women is about 97% higher than for men because of a higher need to provide privacy. Addressing **women's homelessness status** by decreasing the length of their homeless shelter stays can, therefore, also reduce the overall cost of homeless shelters.

Now Try It Yourself

Chapter 6: Summary

Correlation and regression analyses are two of the author's favorite statistical procedures.

Correlation quantifies the strength of the linear relationship between a pair of variables, whereas *regression* expresses the relationship in the form of an equation. Correlation is a single statistic, whereas regression produces an entire equation. Both quantify the direction and strength of the relationship between two numerical variables. When the correlation (r) is negative, the regression slope (b) will be negative. When the correlation is positive, the regression slope will be positive. The correlation squared ($r2$) has special meaning in simple linear regression. It represents the proportion of variation in Y explained by X.

Correlation is a more concise (single value) summary of the relationship between two variables than regression is. As a result, many pairwise correlations can be viewed together at the same time in one table. Regression provides a more detailed analysis which includes an equation that can be used for prediction and/or optimization.

1. The Supreme Court of the United States is the nation's highest federal court. It is the final interpreter of federal law, including the U.S. Constitution.
2. . The AAMC is a nonprofit association dedicated to improving the health of people everywhere through medical education, health care, medical research, and community collaborations. Its members are all 157 U.S. medical schools accredited by various governing bodies; approximately 400 teaching hospitals and health systems, including the Department of Veteran Affairs medical centers; and more than 70 academic societies.
3. Association of American Medical Colleges (2022). AAMC Leads Amicus Brief in Support of Consideration of Race in Higher Education Admissions. https://www.aamc.org/news/press-rele...ion-admissions. Retrieved on July 5, 2023.
4. Association of American Medical Colleges (2022). AAMC Deeply Disappointed of SCOTUS Decision on Race-Conscious Admissions. https://www.aamc.org/news/press-rele...ous-admissions. Retrieved on July 5, 2023.
5. MSNBC Interview with Dr. Uche Blackstock and Dr. David J. Skorton, President and CEO of the Association of American Medical Colleges. https://www.msnbc.com/ali-velshi/wat...u-s-1865431738. Retrieved on July 5, 2023.
6. In modern discourse, the justices of the Court are often categorized as having conservative, moderate, or liberal philosophies of law and of judicial interpretation.
7. Wikipedia (2023). Ideological Leanings of the United States Supreme Court justices. https://en.Wikipedia.org/wiki/Ideolo...Court_justices. Retrieved on July 6, 2023.
8. Laerd Statistics. Pearson's Product Moment Correlation Using SPSS Statistics. https://statistics.laerd.com/spss-tu...statistics.php. Retrieved on July 11, 2023.
9. Deshpande RC (2012). A Healthy Way to Handle Workplace Stress Through Yoga, Meditation, and Soothing Humor. *International Journal of Environmental Sciences*, Volume 2, No. 4.

10. Spiegelhalter D (2012). Coincidences: What are the Chances of Them Happening? BBC: Future. https://www.bbc.com/future/article/2...ction%E2%80%9D. Retrieved on July 14, 2023.
11. Sue DW, Capodilupo CM, Torino GC, Bucceri JM, Holder AMB, Nadal KL & Esquilin M (2007). Racial Microaggressions in Everyday Life: Implications for Clinical Practice. *American Psychologist*, 62, 271-286.
12. Nadal KL, Wong Y, Griffin K & Hamit S (2014). The Impact of Racial Microaggressions on Mental Health Counseling Implications for Clients of Color. *Journal of Counseling and Development*, January.
13. Carter RT, Kirkinis K & Johnson V (2019). Relationship Between Trauma Symptoms and Race-Based Traumatic Stress. *Traumatology*, September 9.
14. Hao H, Garfield M & Purao S (2022). The Determinants of Length of Homeless Shelter Stays: Evidence-Based Regression Analyses. *International Journal of Public Health*, 28, January.

1.7: Case Studies

Learning Outcomes

- Case Study
- Types of Case Studies
- Case #1: Demographic Profile of Reentry Clients
- Case #2: A Holistic Model for Social Justice
- Case #3: A Health Disparities Model for Social Justice

A case study is an investigation into an individual circumstance (Gaille 2018)[1]. In this final chapter, the case study method is applied to help you, the student, bridge the gap between statistical theory and practice. It is used to help you develop an understanding of the basic ideas in mathematical statistics while examining a contemporary real-life social justice issue.

Other advantages of the case method are (Gaille 2018):

- It turns client-level data into usable data.
- It turns opinion into fact.
- It furthers knowledge growth because there is interest in the case study.
- It can use a number of different research methodologies.
- It is good for formative research that is exploratory in nature.

Stake (1995)[2] characterizes three main types of case studies: *intrinsic, instrumental, and collective*. An **intrinsic case study** is typically undertaken to learn about a unique phenomenon. The researcher should define the uniqueness of the phenomenon that distinguishes it from all others. In contrast, the **instrumental case study** uses a particular case (some of which may be better than others) to gain a broader appreciation of an issue or phenomenon. The **collective case study** involves studying multiple cases simultaneously or sequentially in an attempt to generate a still broader appreciation of a particular issue. These are, however, not necessarily mutually exclusive categories.

Three cases will be discussed in this chapter. In the first case on post-incarceration, an *intrinsic* case study is used to describe the demographic profile of fathers receiving reentry services from a provider in the city of Boston. By including two other cases, a *collective* case design is employed in an attempt to generate a broader appreciation of organizations that attempt to address a social justice issue, namely, health disparities and the social determinants of health. Specifically, we will study a neighborhood health center serving under-resourced communities/zip codes in Boston and a non-profit organization providing a wide spectrum of holistic care to the "poorest of the poor" in India. Collectively, these three cases are considered *instrumental* as exemplars of the more general phenomenon of social justice.

Before proceeding, the first and most important point is that the best statistical analyses cannot save an inferior **research design**. Research design is the foundation of a good study. If the design is weak, the analysis will crumble.

Case #1: Incarcerated and Reentry Fathers

As an *intrinsic* case study, this social justice issue is selected on its own merit and uniqueness. According to the National Responsible Fatherhood Clearinghouse (2023)[3], the number of fathers in U.S. jails and prisons has increased four-fold since 1980. Among the more than 800,000 parents in federal and state prisons, 92 percent are fathers. The 2012 study, "Families and Reentry: Unpacking How Social Support Matters"[4], concludes that connecting reentering fathers with support from family and friends is key for avoiding recidivism (returning to prison) and helping them re-establish their lives.

The framework for this case study is as follows:

A. Identify and Define the Research Question

Each case study centers on a **research question**. The question establishes the focus of the study by identifying the research object, which in this situation is incarcerated fathers who reenter their communities. The two primary research questions for this case study are:

- *What are the characteristics of formerly incarcerated persons who are fathers receiving reentry services at a provider in the city of Boston?*

- *What do they self-report in areas of vulnerability, distress, emotional and behavioral health, relationships, parent-child engagement, and self-efficacy?*

B. Select the Sample Size

In this step, the statistician decides on the **unit of analysis**—the number of cases, the type of cases, and the approach used to collect, store, and analyze the data.

This is the design phase of the case study method.

In this case study, the unit of analysis is clients who are receiving reentry services at a provider in the city of Boston. The sample size (n) is 289.

C. Evaluate and Analyze the Data

In this step, the statistician uses varied methods to analyze quantitative as well as qualitative data. The data is categorized, tabulated, and cross-checked to address the purpose of the study. Variables are labeled, and graphs are created to generate **descriptive statistics and inferential statistics**. This enables the statistician to approach the data in different ways and, thus, avoid premature conclusions.

> **Descriptive Statistics:**
> **Demographic Profile of Reentry Clients**

Gender (n=289): Of the 289 clients, the gender composition of clients is 99% or 286 males, .7% or 2 transgender males, and .3% or 1 female. The high male participation rate is indicative of the mission and focus of the organization.

Age (n=265): The client base contains a greater number of people who are 19-34 years old (135 or 50.9%), followed by 35-64 years old (114 or 43%). The younger age group, 13-18 years old, has the lowest number of clients (16 or 6.1%). Separate from the 265 clients, there are an additional 24 clients whose date of birth is missing.

Race/Ethnicity (n=289): A majority of clients self-identify as Black/African American (213 or 73.7%). White is the second highest ethnic group (26 or 9%), along with Hispanic/Latinx (26 or 9%). American Indian/Alaskan Native (5 or 1.7%) was the third highest group, followed by Black/African American/Hispanic (3 or 1%). Fifteen clients (or 5.3%) self-identified in various other ethnic groups. One client (or .3%) chose "refused" as a response to the race/ethnicity question.

Primary Language Spoken (n=289): A majority of clients (271 or 93.9%) report English as their primary language spoken. Other languages spoken are Spanish (1 or .3%), Spanish & English (7 or 2.5%), Haitian Creole (2 or .7%), English & Haitian Creole (2 or .7%), English & Cape Verdean (2 or .7%), Cape Verdean (1 or .3%), English & French Cajun (1 or .3%), English/Haitian Creole/French Cajun (1 or .3%). One (or .3%) client indicated American Sign Language.

Connections with Boston Neighborhoods (n=289): As part of the intake process, clients are asked, *"Do you live, work, receive services, visit family or friends in any of the following neighborhoods?"* Most clients (145 or 50.2%) report connections with Dorchester. The second highest numbers are in the geographic areas of Roxbury (31 or 10.7%), Mattapan (23 or 8%), and Hyde Park (14 or 4.8%). Other communities indicated are Fenway-Kenmore (8 or 2.8%), Jamaica Plain (7 or 2.4%), West Roxbury (5 or 1.7%), Brockton (4 or 1.4%), Roslindale (4 or 1.4%), Quincy (2 or .7%), Randolph (2 or .7%), Allston (2 or .7%), Mission Hill (2 or .7%), and Cambridge (2 or .7%). Thirty-two or 11% reported having connections with multiple communities, such as Dorchester-Roxbury; Hyde Park-Roxbury-Fenway-Kenmore-Randolph-Brockton-Canton; and Dorchester-Mattapan-Hyde Park-Roxbury. Six clients (or 2.1%) reported, "No, I am not connected to any of these neighborhoods."

Zip Code of Client's Residence (n=205): The zip code 02124 is Dorchester Center. This zip code is where 94 (or 45.9%) of the clients report that they reside. Sixteen (or 7.8%) report zip code 02119-Roxbury, 13 or 6.3% for 02126-Mattapan, 13 (or 6.3%) report 02121-Dorchester-Grove Hall, 8 (or 3.9%) for 02136-Hyde Park, 4 (or 1.9%) for 02130-Jamaica Plain, 4 (or 1.9%) for 02169-South Quincy, 4 (or 1.9%) for 02301-Brockton, 3 (or 1.5%) for 02122-Dorchester, 3 (or 1.5%) for 02131-Roslindale, 3 (or 1.5%) for 02368-Randolph as their place of residence. Other zip codes have one or two clients, totaling 38 (or 18.5%) clients. One client (or .5%) reported being homeless, and another client (.5%) gave "N/A" as a response. These results confirm the findings from the previous section, Connections with Neighborhoods. Dorchester, Mattapan, and Roxbury are the three primary geographic areas in Boston where clients reside, secure support services, and engage with family and friends.

Sexual Orientation (n=289): Two hundred eighty-two (282 or 97.6%) of the 289 clients self-identify as straight (heterosexual). Three clients (or 1%) self-identify as bi-sexual, and one client (or .3%) as Gay. Two clients (or .7%) indicated that they did not feel

comfortable answering this question. One client (or .4%) indicated responded, "Not Sure/Questioning".

Arrested (Spent Time in Jail/Prison) (n=289): When asked the question, *"Have you been arrested before (spent time in Jail/Prison)?"* a majority (209 or 72.3%) of the 289 clients responded, "Yes". A smaller proportion (80 or 27.7%) of clients responded, "No".

Education Level (n=289): A majority (100 or 34.6%) attended 12th grade-no high school diploma, while another 62 (or 21.5%) attended grades 1-11 or some high school (4 or 1.4%) or reported "no schooling completed" (2 or .7%). Ten clients (or 3.5%) have a high school diploma or a GED. The second largest group earned some college credits but no degree (64 or 22.1%). Several clients have acquired post-secondary degrees: Associate of Arts or Science (16 or 5.5%); Bachelor of Arts or Science (16 or 5.5%); Masters (14 or 4.8%); PhD (1 or .4%).

Employment (n=265): Most clients (99 or 37.4%) report being employed full-time, while 79 (or 29.8%) clients work part-time. Thirty-five (or 13.2%) indicated that they have been out of work for a year or more; 17 (or 6.4%) reported being out of work for less than one year. Other clients report being a student (13 or 5%), self-employed (12 or 4.5 %) or unable to work (9 or 3.4%). One client is retired (1 or .3).

D. Presentation of Results

The results are presented in a manner that allows the reviewer to evaluate the findings in the light of the evidence presented. The results are corroborated with sufficient evidence that all aspects of the research question have been adequately answered. Newer insights gained are highlighted as well. The analysis reveals the following attributes or characteristics of a typical client:

A "Typical Profile" of a Client

- An African American male who is between the ages of 35-64 years old.
- Speaks English as their primary language.
- Lives, work, and receive services in Dorchester, Massachusetts.
- Lives in Dorchester Center.
- Self-identifies as straight (heterosexual).
- Has spent time in jail/prison.
- Does not have a high school diploma.
- Has full-time or part-time employment.
- Worries about paying rent/mortgage.
- Has transportation to get to meetings and medical appointments.
- Feels safe at home, school, and community.
- Is the person that they want to be for their children.
- Is trying to re-engage in their children's lives and feel confident that it will happen.
- Makes attempts to contact their children.
- Tells their children that they love them.
- Has a poor-to-fair relationship with their children's mother.
- Sometimes feels sad or anxious about everyday living but encourages themselves by believing that everything is all right during difficult times.
- Knows where to look for job opportunities and knows how to apply for a job.
- Feels confident about achieving desired goals.

> Descriptive Statistics:
> A Holistic Model for Social Justice

Case #2: The PRASAD Project – An International Humanitarian Expression

PRASAD (Philanthropic Relief, Altruistic Service and Development) is a philanthropic expression of the mission of the Siddha Yoga Dham Foundation in South Fallsburg, New York. The PRASAD Project was initiated in 1992 by Gurumayi Chidvilasananda, spiritual head of the Siddha Yoga path.[5]

The PRASAD Project is an independent, non-profit organization committed to improving the quality of life of economically disadvantaged people around the world. It is a non-governmental organization (NGO) in special consultative status with the Economic and Social Council of the United Nations. PRASAD is deeply committed to maintaining strong financial health,

accountability, and transparency about its programs and operations. As a result, PRASAD has received top ratings and recognition from charity rating organizations and other publications.[6]

The mission of PRASAD: PRASAD works in partnership with people to benefit children and communities in need, regardless of race or belief. PRASAD implements innovative solutions that respond to local conditions and cultures.

Vision of PRASAD: PRASAD envisions healthy communities, prospering in harmony with the natural environment, where people are inspired to improve the quality of their own and others' lives.

Values of PRASAD: PRASAD's values manifest the Siddha Yoga philosophy in the arena of philanthropic work.

- recognizes the inherent dignity and worth of each person.
- respects and loves others.
- affirms the spirit of generosity that creates abundance.
- believes that human beings have within themselves great virtues, wisdom, and capabilities.

Holistic Model: PRASAD'S Holistic model is for healthy communities to thrive in harmony with the natural environment. The emphasis is on three components: *sustainable community development, general and specialized healthcare, and the environment.* PRASAD programs contribute to achieving these United Nations' sustainable development goals:

- No poverty
- Zero hunger
- Good health and well-being
- Quality education
- Clean water and sanitation
- Decent work and economic growth
- Reduced inequalities
- Climate action
- Life on land

Since its inception, PRASAD has been partnering with people to deliver holistic, sustainable programs in India, Mexico, and the United States. Over the last 30+ years, these partnerships have produced a life-transforming impact of services:

- Restoring a smile that sparks a change in a child's health and self-esteem;
- Providing cataract surgery that improves a person's vision and independence; and
- Empowering women, which gives them a sense of self-worth, inspiring them to take control of their own lives, both within and outside the home.

Pediatric Dental Health: More than two decades ago, through consultations with county health and school officials, PRASAD launched its Children's Dental Health Program (PRASAD CDHP). In Sullivan County, New York, where lack of transportation prevents many families from accessing dental care, its mobile clinic makes services accessible to students right at their schools. PRASAD CDHP has received many awards from the New York State Assembly, New York State Senate, corporations, and foundations in recognition of their leadership and excellence in maintaining children's oral health.

Program Outcomes: Within applied statistics and research, outcome variables can be categorical (non-parametric statistics), ordinal (non-parametric statistics), or continuous (parametric statistics). Outcome variables increase the precision and accuracy of measurement (internal validity) and make study results more readily generalizable.

In this particular case study on the PRASAD Project, we are linking social justice to program evaluation to enhance the fair and just distribution of benefits and responsibilities. Social justice, a central tenet of community psychology, emphasizes equal access to resources, dissolution of power hierarchies, and the empowerment and promotion of wellness among marginalized populations (Torres-Harding, Siers, & Olson, 2012)[7]. By applying an equity lens, students in statistics are inspired to challenge the status quo, care about the interests of the disadvantaged, and uncover weaknesses within the system that contribute to inequities within society.

To track outcomes, most government and nonprofit programs rely on performance measurement strategies rather than more expensive and complicated quasi-experimental and experimental designs. Essentially, performance measurement strategies seek to answer the question: *Did the program accomplish what it set out to accomplish?* Performance measurement relies on the utilization of records, staff observations, and participant self-reports. The following are statistics on the outcomes of the PRASAD Project for serving those in need for 31 years:

United States

Dental Health Education: 93,000 children

Dental visits: 31,800

Dental procedures: 98,100

Future Goal:

In the United States, PRASAD's Children's Dental Health Program will continue to offer high-quality dental health education and dental services to 4,000 low-income children in New York State annually.

India

Mobile Hospital: 1,069,098 visits

Nutrition Program: 1,414,250 servings

Eye Care: 261,057 screenings and surgeries

Medical Center: 1,005,124 visits

Tuberculosis: 95% cure rate

Kitchen Gardens: 11,030

Number of Self-Help Group (SHG) Members: 2,986 women participating

Arts & Crafts: 154,558 students

Tree Planting & Floriculture: 174,476 samplings

Future Goal:

In India, PRASAD Chikitsa's goals for the next year are to continue delivering medical services, providing nutritional support to 500 children at village care centers, helping 350 families start kitchen gardens, planting 50,000 trees, distributing 10,000 jasmine saplings to farmers for market crops, and helping 200 families build toilets, among other services.

Mexico

Free Eye Surgeries: 34,000

Eye camps: 213

Future Goal:

PRASAD de México's team will organize three annual free eye surgery camps to benefit low-income people in rural areas of Mexico.

> Descriptive Statistics:
> A Health Disparities Model for Social Justice

Case #3: Community Health Centers in Massachusetts

According to the Massachusetts League of Community Health Centers, community health centers provide primary, preventive, and dental care, as well as mental health, substance use disorder, and other community-based services to anyone in need, regardless of their insurance status or ability to pay. In Massachusetts, 52 community health center organizations provide high-quality health care to some one million state residents through more than 300 sites statewide. In addition to providing comprehensive health services to underserved people, health centers are at the leading edge of addressing some of the most vexing problems of our healthcare system, including facilitating access to health insurance coverage for low-income residents and eliminating health disparities between racial and ethnic populations.[8]

The League also reports that, in 1965, the nation's first community health center opened its doors in Boston. Until that time, health services for low- and moderate-income people in inner city areas and isolated rural communities were nowhere to be found. In response, community members organized around the need to bring primary care to their neighborhoods. Insisting that they have a voice in how and what care should be delivered to the community, boards of directors that included a majority of health center

consumers were incorporated into the model. Today, health center patients continue to drive the mission and work of community health centers.[9]

The following is an example of a Community Health Center in Massachusetts:

Codman Square Health Center (Dorchester, MA)

Mission

To serve as a resource for improving the physical, mental, and social well-being of the community.

Vision

Codman Square Health Center is our community's first choice for comprehensive, holistic, and integrated services, and empowers individuals to lead healthy lives and build thriving communities.

Values

Patient: Our patient is the center of the care team.

Community: The well-being of the individual is deeply connected to the health of the community.

Staff: We are a diverse, empowered, and prepared workforce.

Advocacy: We advocate for responsive policies and resources to address health disparities and promote health equity.

Innovation: We promote a culture of innovation that has a measurable and sustainable impact. *Partnerships:* We build and sustain diverse partnerships.

Total Number of Patients: 23,695 (in 2022)

Gender: Female (13,874 or 58.6%); Male (9,821 or 41.4%).

Race/Ethnicity: Black/African American (17,659 or 81.5%); Hispanic/Latino (2,388 or 11%); White (1,086 or 5%); Other (532 or 2.5%); Unknown (2,030 or 8.6%).

Income Status: Live at or below 100% of the poverty line (11,788 or 96%).

Service Area: Live in Dorchester, Hyde Park, Brockton, Roxbury, Mattapan, and Randolph (18,434 or 77.8%)

Types of Visits: Medical Care (81,685 or 74.1%); Behavioral Health/Mental Health (11,073 or 10%); Dental Care (7,679 or 7%); Eye/Vision (2,798 or 2.5%); Substance Use (2,633 or 2.4%); Enabling Services/Case Management (2,300 or 2.1%); Other Services-Nutrition/Podiatry/Dermatology (2,020 or 1.8%).

Selected Diagnoses: 2022 data shows patients live with Hypertension (4,818), Overweight/Obesity (3,652), Type II Diabetes (2,837), Depression (2,175), and Anxiety (1,861).

Namasté I honor the place in you in which the entire universe dwells. I honor the place in you which is of love, of truth, of light, and of peace. When you are in that place in you, and I am in that place in me, We are one
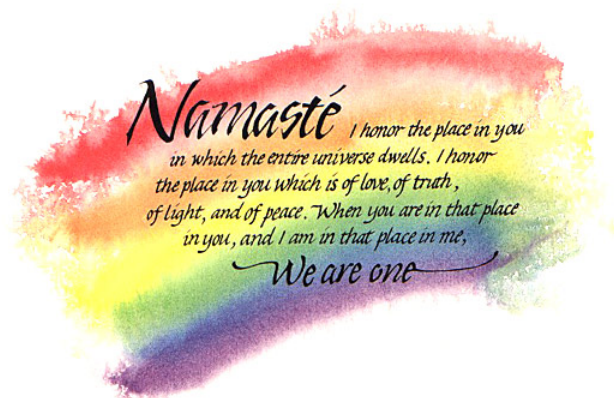
This concludes our journey to becoming equity-minded through the lens of statistics. I hope that during your journey, you enjoyed periodic pauses and had mindful quietude as you reflected on what you learned. There is so much to learn about statistics itself and how it blends so well into understanding social justice and equity issues.

In closing, I want to thank you for your steadfastness and courage to travel this journey with me. I am always available for questions and comments. You can email me at yanthony @ framingham.edu. I would love to hear from you!

1. Gaille, B (2018). 12 Case Study Method Advantages and Disadvantages. https://brandongaille.com/12-case-st...disadvantages/. Retrieved on September 24, 2023. 1.07: Case Studies#return-footnote-187-1
2. Stake, RE (1995). The Art of Case Study Research. London: Sage Publications Ltd. 1.07: Case Studies#return-footnote-187-2
3. National Responsible Fatherhood Clearinghouse (2023). Incarcerated and Reentering Fathers. https://www.fatherhood.gov/for-progr...tering-fathers. Retrieved on September 24, 2023. 1.07: Case Studies#return-footnote-187-3
4. Fontaine, J, Gilchrist-Scott, D, Denver, M, & Rossman, SB (2012). Families and Reentry: Unpacking How Social Support Matters. June 2012; Washington, DC: The Urban Institute. 1.07: Case Studies#return-footnote-187-4
5. Siddha Yoga Dham Foundation (2023). The PRASAD Project. https://www.siddhayoga.org/prasad. Retrieved on October 8, 2023. 1.07: Case Studies#return-footnote-187-5
6. The PRASAD Project (2023). PRASAD: 2023 Fall Newsletter. https://www.prasad.org/wp-content/up...Compress_1.pdf. Retrieved on October 8, 2023. 1.07: Case Studies#return-footnote-187-6
7. Torres-Harding, R.S., Siers, B., & Olson, B.D. (2012). Development and psychometric evaluation of the Social Justice Scale (SJS). American Journal of Community Psychology, 50, 77–88. 1.07: Case Studies#return-footnote-187-7
8. Massachusetts League of Community Health Centers (2023). Community Health Centers. https://www.massleague.org/CHC/Overview.php. Retrieved on October 8, 2023. 1.07: Case Studies#return-footnote-187-8
9. Ibid. 1.07: Case Studies#return-footnote-187-9

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Statistics Through an Equity Lens (Anthony)

Webpages: 20

Applicable Restrictions: Noncommercial

All licenses found:

- Undeclared: 55% (11 pages)
- CC BY-NC-SA 4.0: 45% (9 pages)

By Page

Detailed Licensing

Overview

Title: Statistics Through an Equity Lens (Anthony)

Webpages: 20

Applicable Restrictions: Noncommercial

All licenses found:

- Undeclared: 55% (11 pages)
- CC BY-NC-SA 4.0: 45% (9 pages)

By Page

- Statistics Through an Equity Lens (Anthony) - *CC BY-NC-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Introduction - *CC BY-NC-SA 4.0*
  - 1: Chapters - *Undeclared*
    - 1.1: An Overview of Statistics - *CC BY-NC-SA 4.0*
    - 1.2: Descriptive Statistics - *CC BY-NC-SA 4.0*
    - 1.3: Probability - *CC BY-NC-SA 4.0*

- 1.4: Inferential Statistics- Sampling Methods - *CC BY-NC-SA 4.0*
- 1.5: Significance of Statistical Inference Methods - *CC BY-NC-SA 4.0*
- 1.6: Correlation and Regression Analysis - *CC BY-NC-SA 4.0*
- 1.7: Case Studies - *CC BY-NC-SA 4.0*
  - Back Matter - *Undeclared*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Undeclared*
    - Detailed Licensing - *Undeclared*