A Guide through the People Analytics Lifecycle

With Applications in R

Craig Starbuck

2022-02-08

Contents

Chapter 1

Foreword

Chapter 2

Introduction

Twenty years ago, I was the least likely person to write this book. My first statistics course in college was dreadful. On day one, my professor stood before the class of about 100 students and gave us the stats: "Based on historical data, half of you won't make it to the midterm and of those who do, half won't receive a passing grade in the end." This was both discouraging and motivating. Stats was a required course for my major so failure wasn't an option; I had to pass. The course was challenging, and I attended weekly study sessions with classmates and studied a lot independently to learn the material. I saw no applications for statistics to anything I planned to do with my degree, so the course was reduced to memorization of equations; it was anything but enjoyable. I passed the course with a B, and I was determined to never open another stats book.

You may be wondering what changed to motivate authoring a book involving this insufferable subject. The short answer is that I discovered the very important applications to a discipline I truly love, people analytics. As I began to think about complex and nuanced challenges in social science contexts, it became clear that I would not only need to reengage with stats; I would need to develop an authentic appreciation for the discipline. Over the past decade, I have taken the journey of 'relearning' statistics and developing a deep understanding of how statistical methodologies can be applied to various organizational problem statements.

My purpose in writing this book is to help make this content – which may unfortunately be intimidating to many – both accessible and exciting. In addition to my role in people analytics, I have taught a graduate-level business analytics course for Finance and MBA students for many years and have developed several teaching strategies through this experience that I will apply in this book. Beyond these instructional methods, this book makes a unique contribution in curating what I consider to be the most salient topics for people analytics practitioners. There are many texts available for deeper treatments of individual

subjects covered in this book but as of this writing, I have found none that organize within a single text both theoretical and applied instruction spanning the whole of the people analytics lifecycle.

Thus, this book represents my earnest attempt to provide a concise – yet adequately comprehensive – treatment of the concepts and methods I've found to be most important for people analytics. My hope is that this book will ignite within you the same passion for analytics I have discovered over the past decade.

Craig Starbuck December 2021

Craig Starbuck, PhD is the CEO and Co-Founder of OrgAcuity, a tech company with a mission to democratize access to people analytics. Craig has built and led people analytics teams at companies such as Robinhood, Mastercard, Equifax, TD Ameritrade, and Scottrade, and he also spent a decade in various data engineering and analytics positions in the banking and health care industries. He is a Member of the Society for Industrial and Organizational Psychology (SIOP) and has a passion for transforming people data into information and insights that help organizations enhance the experience and wellbeing of employees.

Chapter 3

Getting Started

Nothing in this book will increase the value of an analysis no one needs. Analyses should always have a strong value proposition – a clear expectation of how an analysis will support a General Manager, People Partner, Salesperson, or other member of the organization. Curiosity is not a business reason, and this chapter will cover a set of guiding principles as well as a project management framework that will help ensure analytics are anchored in well-defined problem statements with meaningful ROI.

In addition, this book will cut through the fluff and teach you how to do stuff that matters. Knowledge of concepts is futile without an understanding of how to apply them to people analytics use cases. Whether you are a people leader, individual contributor, or aspiring analytics practitioner, this book is for you. This book will serve as a guide through the analytics lifecycle, curating the key concepts and applications germane to common questions and hypotheses within people analytics.

3.1 Guiding Principles

Among the many principles guiding how analytics teams operate, there are three that I have found to be universally applicable and critical to the success of an analytics capability.

3.1.1 Pro Employee Thinking

"With Great Power Comes Great Responsibility."

'Pro employee' thinking is addressed first and for good reason. People analytics has the power to improve the lives of people in meaningful ways. Whether we

are shedding light on an area of the business struggling with work-life balance or identifying developmental areas of which a group of leaders may be unaware, people analytics ideally improves employee well-being and effectively, the success of the business. It is important to embrace a 'pro employee' philosophy, as newfound knowledge could also have damaging repercussions if shared with the wrong people or if findings are disseminated without proper instruction on how to interpret and act.

One way to error on the side of caution when considering whether to disseminate insights is to ask the following: "With this knowledge, could the recipient act in a manner that is inconsistent with our 'pro employee' philosophy?" If the answer to this question is not a clear "no", discuss with your HR, legal, and privacy partners and together, determine how best to proceed. The decision may be to not share the findings with the intended audience at all or to develop a proper communication and training plan to ensure there is consistency in how recipients interpret the insights and act in response. Employment Law and Data Privacy Counsel are our friends, and it is important to build strong relationships with these critical partners.

3.1.2 Quality

"Garbage In, Garbage Out."

Never compromise quality for greater velocity. If quality falls to the bottom of the priority list, all other efforts are pointless. It is unlikely that requestors of data and analytics will ever ask us to take longer to prepare the information. The onus is on us as analytics professionals to level set on a reasonable timeline for analyses based on many factors that can materially impact the quality of analyses and insights. A single instance of compromised quality can have lasting damage on the reputation of the analytics function and cause consumers of insights to view all findings as suspect. Be sure quality is consistently a top value and guard your team's reputation at all costs. If stakeholders lose trust, there will likely be additional data requests for validation; this is wasteful to both you and your user community and detracts from the bigger story that needs to be conveyed.

To be clear, by 'quality' I am referring to results, which is dependent on data integrity in the source systems, proper data preparation steps, and many other factors. Most of the analyst's time is spent on data preparation (data collection, cleaning and organizing, building training sets, mining for patterns, refining algorithms, etc.). If tight controls do not exist within the source application to support data integrity, data preparation efforts can only go so far in delivering reliable and valid findings. It is often the analysts who identify data integrity issues due to the nature of their work; therefore, close relationships should be formed with source application owners to put into place validation rules to proactively prevent the entry of erroneous data or at the very least, exception/audit reports to identify and address the issues soon after the fact.

3.1.3 Prioritization

"If everything is a priority, nothing is a priority."

If there is not a supply-demand gap on the analytics team, the team likely isn't asking enough questions. It is okay (even good) to have a backlog of projects if unmet demand largely represents requests for lower-impact analyses. It is crucial to be relentless about prioritizing strategically important projects with 'measurable' impact over merely interesting questions that few care to answer. According to the Pareto Principle, 80% of outcomes (or outputs) result from 20% of causes (or inputs). In analytics, it is important to be laser focused on identifying the 20% of inputs that will result in disproportionate value creation for stakeholders. There are some general customer-oriented questions I have found to be helpful for the intake process to optimize the allocation of time and resources:

1. Does this support a company or departmental objective? If not, why should this be prioritized over something else?
2. Who is the executive sponsor? Really important projects will have an executive-level sponsor.
3. What quantitative and/or qualitative data can be provided as a rationale for this request? Is there data to support doing this, or is the problem statement rooted merely in anecdotes?
4. Will this mitigate risk or enable opportunities?
5. What actions can or will be taken as a result of this analysis?
6. What is the *scale* of impact (# of impacted people)?
7. What is the *depth* of impact (minimum –> significant)?
8. Is this a dependency or blocker for another important deliverable?
9. What is the impact of not doing (or delaying) this?
10. What is the request date? Is there flexibility in this date and/or scope of the request (e.g., what does MVP look like)?

These questions can be weighted and scored as well to support a more automated and data-driven approach to prioritization.

3.2 Tools

This book uses freely available software for statistics, modeling, and data visualization.

3.2.1 R

While there are many commercial-grade analytics toolsets, R is open-sourced statistical and data visualization software that can be downloaded free of charge.

It is incredibly powerful, and there is a package (or at least the ability to easily create one) for every conceivable statistical technique and data visualization. It is also widely used in highly regulated environments. As of this writing, R Markdown – the dynamic document creator in which I am writing this book – allows for coding in 56 different languages! Therefore, the debate around whether to use Python, Julia, or something else is now moot; we need not sacrifice the advantages of other languages by choosing one. To get started, simply download the latest version of R and the R Studio IDE using the following links:

- R: https://www.r-project.org/
- R Studio IDE: https://www.rstudio.com/products/rstudio/download/#download

Please note that while R basics are covered, this is not a book on how to code. It is assumed that you already understand programming fundamentals. If this is not the case, an introductory programming course is highly recommended; this is one of the best investments you can make for a successful career in analytics. The ability to write code is now table stakes for anyone in an analytics-oriented field, as this is the best way to develop reproducible analyses. Coding is to analytics professionals what typing was for Baby Boomers decades ago; a lack of coding proficiency is a major limiting factor on one's potential in analytics.

Libraries from several R packages will be utilized in this book. The line of code below can be executed within R to install all at once:

```
# Install required packages
install.packages(c("tidyverse", "corrplot", "psych", "moments"), dependencies = TRUE, i
```

```
## Warning: dependencies 'graph', 'Rgraphviz' are not available
```

```
##
##   There is a binary version available but the source version is later:
##          binary source needs_compilation
## corrplot   0.89   0.92             FALSE
##
##
## The downloaded binary packages are in
##  /var/folders/b1/0nnhbsx55hvfb1b3n83x0qrw0000gn/T//Rtmp3RVEAG/downloaded_packages
```

The goal of the code provided in this book is not to represent the most performant, succinct, or productionalizable approaches. The code herein is intended only to facilitate understanding and demonstrate how concepts can be implemented in people analytics settings. Programming expertise is important for optimizing these approaches for production applications.

3.3 4D Framework

Adherence to a lightweight framework over hastily rushing into an analysis full of assumptions generally lends to better outcomes. A framework ensures (a) the problem statement is understood and well-defined; (b) relevant literature and prior research are reviewed; (c) the measurement strategy is sound; (d) the analysis approach is suitable for the hypotheses being tested; and (e) results and conclusions are valid and communicated in a way that resonates with the target audience. This chapter will outline a recommended framework as well as other important considerations that should be reviewed early in the project.

It is important to develop a clear understanding of the key elements of research. Scientific research is the systematic, controlled, empirical, and critical investigation of natural phenomena guided by theory and hypotheses about the presumed relations among such phenomena (Kerlinger & Lee, 2000). In other words, research is an organized and systematic way of finding answers to questions. If you are in the business of analytics, I encourage you to think of yourself as a scientist – regardless of whether you are wearing a lab coat or have plans to publish.

As we will discover when exploring the laws of probability in a later chapter, there is a 1 in 20 chance of finding a significant result when none exists. Therefore, it is important to remain disciplined and methodical to protect against backward research wherein the researcher mines data for interesting relationships or differences and then develops hypotheses which they know the data support. There have been many examples of bad research over the years, which often presents in the form of p-hacking or data dredging: the act of finding data to confirm what the researcher wants to prove. This can occur by running an exhaustive number of experiments to find one that supports the hypothesis, or by using only a subset of data which features the expected patterning.

Academics at elite research institutions are often under immense pressure to publish in top-tier journals which have a track record of accepting new groundbreaking research over replication studies or unsupported hypotheses, and incentives have unfortunately influenced some to compromise integrity. As my PhD advisor told me many years ago, an unsupported hypothesis – while initially disappointing given the exhaustive literature review that precedes its development – is a meaningful empirical contribution given theory suggests the opposite should be true.

If you participated in a science fair as a child, you are likely already familiar with the scientific method. The scientific method is the standard scheme of organized and systematic inquiry, and this duly applies to people analytics practitioners in the promotion of robust analyses and recommendations.

Over the years, I have adapted the scientific method into a curtailed four-dimensional framework which is intended to elevate the rigor applied to the end-to-end analytical process. The four dimensions are (a) Discover, (b) De-

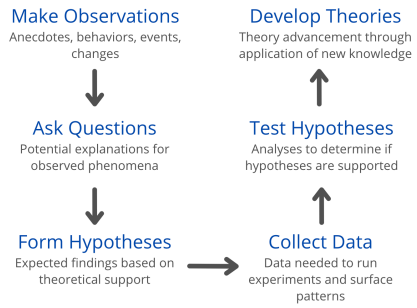


Figure 3.1: The Scientific Method

sign, (c) Develop, and (d) Deliver, and this book will be organized around these. A comprehensive checklist with questions and considerations for each phase of the analytics lifecycle can be found in the Appendix **??**.

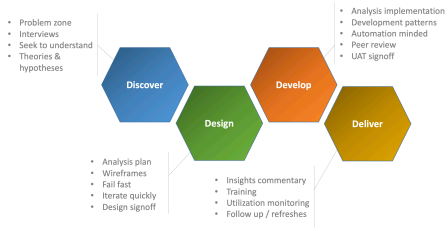


Figure 3.2: 4D Framework

3.3.1 Discover

You are likely familiar with the following adage: "An ounce of prevention is worth a pound of cure." Such is the case with respect to planning in an analytics context. During the Discover phase, it is important to remain in the problem zone; seek to understand your clients' needs through active listening and questions. This is not the time for solutioning or committing to any specific deliverables or timelines. If the client's needs are ambiguous, proceeding without clarity is unlikely to result in a favorable outcome.

It is generally helpful to think about analytics solutions like a Product Owner thinks about the initial and subsequent releases of a commercial product. A **Minimum Viable Product (MVP)** is a version of the solution with the minimum number of features to be useful to early customers who can provide feedback for future enhancements. It is important to clarify that the MVP

version of solutions has both a limited number of users and features and that this is antithetical to building solutions that seek to address every imaginable question for every possible user. Breaking down large projects into small sets of features that are easier to communicate and adopt provides space for agility and real-time adjustments to the product roadmap per user feedback.

3.3.2 Design

Perhaps the most important initial question to answer in the design phase is: "Does anything already exist that addresses part, or all, of the client's objectives?" If an existing solution will suffice, or a previous analysis can be easily refreshed with recent data, it may be possible to allocate time and resources elsewhere. If related or complimentary analyses have already been performed, they may accelerate new analyses.

The end-user experience is of paramount importance during the Design phase, as solutions should have a consistent look and feel regardless of who developed them. Defining and implementing design guidelines will ensure consistency across analytics projects, as well as within large projects in which multiple analysts are collaborating on various components of the solution.

3.3.3 Develop

While development patterns can vary widely across analytics teams, establishing a set of standards can pay dividends in the form of greater efficiency and reliability over time. Pattern-based development ensures analysts who were not involved in a particular project can access the code and easily and quickly understand each step of the analysis: data extraction –> wrangling –> cleaning –> analysis –> visualization.

This is one of the many reasons tools like Excel will not be covered in this book. Software like R and Python allows analysts to organize and annotate steps of the analytical process in a manner that is both logical and reproducible. In case it bears repeating, learning to code is likely the best investment one can make in the pursuit of a career in analytics.

3.3.4 Deliver

The Deliver phase can take many forms depending on the solution being released. If the solution is designed for a large user base, a series of recorded trainings may be in order so that there is a helpful reference for those unable to attend the live sessions or new joiners in the future. It is important to monitor success measures, which could be insights aligned to research hypotheses, dashboard utilization metrics, progress following data-informed interventions, or any number of others defined within the Discover phase.

Chapter 4

Research Methods

Chapter 5

Measurement & Sampling

5.1 Populations & Samples

The goal of research is to understand a population based on data from a subset of population members. In practice, it is often not feasible to collect data from every member of a population, so we instead calculate **sample statistics** to estimate **population parameters**.

Another important concept is the **sampling frame**. While the population represents the entire group of interest, the sampling frame represents the subset of the population to which the researcher has access. In an ideal setting, the population and sampling frame are the same, but they are often different in practice. For example, a professor may be interested in understanding student sentiment about a new school policy but only has access to collect data from students in the courses she teaches. In this case, the entire student body is the population but the students she has access to (those in the courses she teaches) represent the sampling frame. The sample is the subset of the sampling frame that ultimately participates in the research (e.g,. those who complete a survey or participate in a focus group).

5.2 Exercises

1. Parameters are descriptions or characteristics of a sample, while statistics are descriptions or characteristics of a population. A. True B. False

2. 100 randomly selected employees in the Marketing department of an organization participated in a survey on career pathing for marketing professionals. What is the sample and what is the population sampled in

this case? A. Sample: 100 employees who completed the survey, Population: All employees in the organization B. Sample: 100 employees who completed the survey, Population: Marketing employees C. Sample: All Marketing professionals, Population: All employees in the organization D. Sample: All employees in the organization, Population: Employees across all companies globally

Chapter 6

Univariate & Bivariate Analysis

This chapter reviews essential univariate and bivariate analysis concepts that underpin the more complex multivariate approaches in subsequent chapters of this book.

6.1 Univariate Analysis

Descriptive statistics are rudimentary analysis techniques that help describe and summarize data in a meaningful way. Descriptive statistics do not allow us to draw any conclusions beyond the available data but are helpful in interpreting the data at hand.

There are two categories of descriptive statistics: (a) **measures of central tendency** describe the central position in a set of data; and (b) **measures of spread** describe how dispersed the data are.

6.1.1 Measures of Central Tendency

Mean

Perhaps the most intuitive measure of central tendency is the **mean**, which is often referred to as the average. The mean of a sample is denoted by $\bar{x}$ and is defined by:

$$\bar{x} = \frac{\sum x_i}{n}$$

The population mean is denoted by μ and is defined by:

$$\mu = \frac{\sum x_i}{N}$$

The mean of a set of numeric values can be calculated using the mean() function in R:

```
# Fill vector x with integers
x <- c(1,1,1,2,2,2,3,3,4,50)

# Calculate average of vector x
mean(x)
```

```
## [1] 6.9
```

Median

The **median** represents the midpoint in a sorted vector of numbers. For vectors with an even number of values, the median is the average of the middle two numbers; it is simply the middle number for vectors with an odd number of values. When the distribution of data is skewed, or there is an extreme value like we observe in vector x, the median is a better measure of central tendency.

The median() function in R can be used to handle the sorting and midpoint selection:

```
# Calculate median of vector x
median(x)
```

```
## [1] 2
```

In this example, the median is only 2 while the mean is 6.9 (which is not representative of any of the values in vector x). Large deltas between mean and median values are evidence of outliers.

Mode

The **mode** is the most frequent number in a set of values.

While mean() and median() are standard functions in R, mode() returns the internal storage mode of the object rather than the statistical mode of the data. We can easily create a function to return the statistical mode(s):

```r
# Create function to calculate statistical mode(s)
stat.mode <- function(x) {
  ux <- unique(x)
  tab <- tabulate(match(x, ux))
  ux[tab == max(tab)]
}

# Return mode(s) of vector x
stat.mode(x)
```

```
## [1] 1 2
```

In this case, we have a bimodal distribution since both 1 and 2 occur most frequently.

Range

The **range** is the difference between the maximum and minimum values in a set of numbers.

The range() function in R returns the minimum and maximum numbers:

```r
# Return lowest and highest values of vector x
range(x)
```

```
## [1]  1 50
```

We can leverage the max() and min() functions to calculate the difference between these values:

```r
# Calculate range of vector x
max(x, na.rm = TRUE) - min(x, na.rm = TRUE)
```

```
## [1] 49
```

In people analytics, there are many conventional descriptive metrics – largely counts, percentages, and ratios cut by time series (day, month, quarter, year) and categorical dimensions (department, job, location, tenure band). Here is a sample of common measures:

- Time to Fill: average days between job requisition posting and offer acceptance
- Offer Acceptance Rate: percent of offers extended to candidates that are accepted

- Pass-Through Rate: percent of candidates in a particular stage of the recruiting process who passed through to the next stage
- Progress to Goal: percent of approved positions that have been filled
- cNPS/eNPS: candidate and employee NPS (-100 to 100)
- Headcount: counts and percent of workforce across worker types (employee, intern, contingent)
- Diversity: counts and percent of workforce across gender, ethnicity, and generational cohorts
- Positions: count and percent of open, committed, and filled seats
- Hires: counts and rates
- Career Moves: counts and rates
- Turnover: counts and rates (usually terms / average headcount over the period)
- Workforce Growth: net changes over time, accounting for hires, internal transfers, and exits
- Span of Control: ratio of people leaders to individual contributors
- Layers/Tiers: average and median number of layers removed from CEO
- Engagement: average score or top-box favorability score

6.1.2 Measures of Spread

Variance

Variance is a measure of variability in the data. Variance is calculated using the average of squared differences – or deviations – from the mean.

Variance of a population is defined by:

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$$

Variance of a sample is defined by:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

It is important to note that since differences are squared, the variance is always non-negative. In addition, we cannot compare these squared differences to the arithmetic mean since the units are different. For example, if we calculate the variance of annual compensation measured in USD, variance should be expressed as USD squared while the mean exists in the original USD unit of measurement.

In R, the sample variance can be calculated using the var() function:

```
# Load library for data wrangling
library(dplyr)

# Read employee demographics data
demographics <- read.csv("https://raw.githubusercontent.com/crstarbuck/peopleanalytics_lifecycle_

# Calculate sample variance for annual compensation
var(demographics$annual_comp)
```

```
## [1] 1876688425
```

Sample statistics are the default in R. Since the population variance differs from the sample variance by a factor of $s^2 * (\frac{n-1}{n})$, it is simple to convert output from var() to the population variance:

```
# Store number of observations
n = length(demographics$annual_comp)

# Calculate population variance for annual compensation
var(demographics$annual_comp) * (n - 1) / n
```

```
## [1] 1876303464
```

Standard Deviation

The **standard deviation** is simply the square root of the variance, as defined by:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Since a squared value can be converted back to its original units by taking its square root, the standard deviation expresses variability around the mean in the variable's original units.

In R, the sample standard deviation can be calculated using the sd() function:

```
# Calculate sample standard deviation for annual compensation
sd(demographics$annual_comp)
```

```
## [1] 43320.76
```

Since the population standard deviation differs from the sample standard deviation by a factor of $s * \sqrt{(\frac{n-1}{n})}$, it is simple to convert output from sd() to the population standard deviation:

```r
# Calculate population standard deviation for annual compensation
sd(demographics$annual_comp) * sqrt((n - 1) / n)
```

```
## [1] 43316.32
```

Quartiles

Quartiles are a staple of exploratory data analysis (EDA). A quartile is a type of quantile that partitions data into four equally sized parts after ordering the data. Note that each partition is equally sized with respect to the number of data points – not the range of values in each. Quartiles are also related to **percentiles**. For example, Q1 is the 25th percentile – the value at or below which 25% of values lie. Percentiles are likely more familiar than quartiles, as percentiles show up in the height and weight measurements of babies, performance on standardized tests like the SAT and GRE, among other things.

The **Interquartile Range (IQR)** represents the difference between Q3 and Q1 cut point values (the middle two quartiles). The IQR is sometimes used to detect extreme values in a distribution; values less than $Q1 - 1.5 * IQR$ or greater than $Q3 + 1.5 * IQR$ are generally considered outliers.

In R, the quantile() function returns the values that bookend each quartile:

```r
# Return quartiles for annual compensation
quantile(demographics$annual_comp)
```

```
##       0%      25%      50%      75%     100%
##  50016.0  87946.0 125522.0 163262.5 199968.0
```

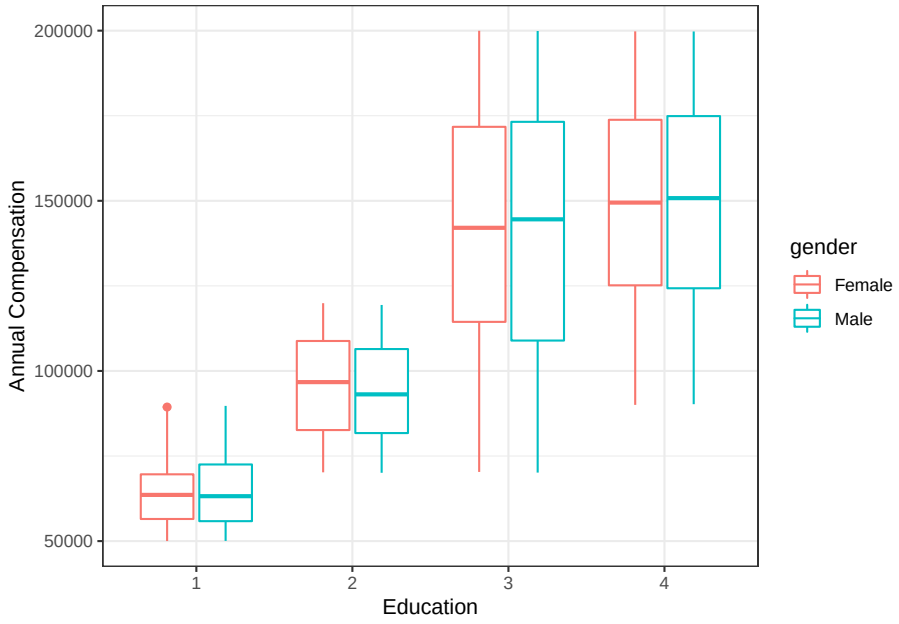
Based on this output, we know that 25% of people in our data earn annual compensation of 87,946 USD or less, 125,522 USD is the median annual compensation, and 75% of people earn annual compensation of 163,263 USD or less.

Boxplots are a common way to visualize the distribution of data by categorical and ordinal factors. Boxplots are not usually found in presentations to stakeholders, since they are a bit more technical and often require explanation, but these are very useful to analysts for understanding data distributions during the EDA phase. In R, the `ggplot2` library has robust and flexible data visualization capabilities which we will leverage throughout this book. Let's visualize the spread of annual compensation by education level and gender using the ggplot() function:

```r
# Load library for data viz
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
# Produce boxplots to visualize compensation distribution by education and gender
ggplot(demographics, aes(x = as.factor(education), y = annual_comp, color = gender)) +
  labs(x = "Education", y = "Annual Compensation") +
  theme_bw() +
  geom_boxplot()
```



Boxplots can be interpreted as follows: * Horizontal lines represent median compensation values. * The box in the middle of each distribution represents the IQR. * The end of the line above the IQR represents the threshold for outliers in the upper range: $Q3 + 1.5 * IQR$. * The end of the line below the IQR represents the threshold for outliers in the lower range: $Q1 - 1.5 * IQR$. * Data points represent outliers: $x > Q3 + 1.5 * IQR$ or $x < Q1 - 1.5 * IQR$.

We can also return a specific percentile value using the `probs` argument in the quantile() function. For example, if we want to know the 80th percentile annual compensation value, we can execute the following:

```
# Return 80th percentile annual compensation value
quantile(demographics$annual_comp, probs = .8)
```

```
##      80%
## 170196.4
```

In addition, the summary() function returns several common descriptive statistics for an object:

```
# Return common descriptives
summary(demographics$annual_comp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50016   87946  125522  125369  163262  199968
```

Skewness

Skewness is a measure of the horizontal distance between the mode and mean – a representation of symmetric distortion. In most practical settings, data are not normally distributed. That is, the data are skewed either positively (right-tailed distribution) or negatively (left-tailed distribution). The coefficient of skewness is one of many ways in which we can asertain the degree of skew in the data. The skewness of sample data is defined as:

$$Sk = \frac{1}{n}\frac{\sum(x_i - \bar{x})^3}{s^3}$$

A positive skewness coefficient indicates positive skew, while a negative coefficient indicates negative skew. The order of descriptive statistics can also be leveraged to ascertain the direction of skew in the data:

- Positive skewness: mode $<$ median $<$ mean
- Negative skewness: mode $>$ median $>$ mean
- Symmetrical distribution: mode $=$ median $=$ mean

Figure 6.1 illustrates the placement of these descriptive statistics in each of the three types of distributions:

The magnitude of skewness can be determined by measuring the distance between the mode and mean relative to the variable's scale. Alternatively, we can simply evaluate this using the coefficient of skewness:

- If skewness is between -0.5 - 0.5, the data are symmetrical.
- If skewness is between -0.5 and -1 or 0.5 and 1, the data are moderately skewed.
- If skewness is $<$ -1 or $>$ 1, the data are highly skewed.

Since there is not a base R function for skewness, we can leverage the moments library to calculate skewness:

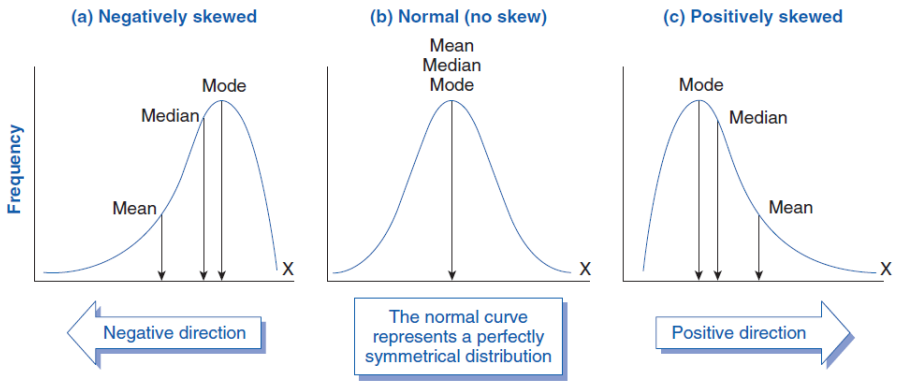


Figure 6.1: Skewness

```
# Load library
library(moments)

# Calculate skewness for org tenure, rounded to three significant figures
round(skewness(demographics$org_tenure), 3)
```

```
## [1] 0.562
```

Statistical Moments, after which this library was named, play an important role in specifying the appropriate probability distribution for a set of data. Moments are a set of statistical parameters used to describe the characteristics of a distribution. Skewness is the third statistical moment in the set; hence the sum of cubed differences and cubic polynomial in the denominator of the formula above. The complete set of moments comprises: (1) expected value or mean, (2) variance and standard deviation, (3) skewness, and (4) kurtosis.

We can verify that the skewness() function from the moments library returns the expected value (per the aforementioned formula) by validating against a manual calculation:

```
# Store components of skewness calculation
n = length(demographics$org_tenure)
x = demographics$org_tenure
x_bar = mean(demographics$org_tenure)
s = sd(demographics$org_tenure)

# Calculate skewness manually, rounded to three significant figures
round(1/n * (sum((x - x_bar)^3) / s^3), 3)
```

```
## [1] 0.562
```

A skewness coefficient of .562 indicates that organization tenure is moderately and positively skewed. We can visualize the data to confirm the expected right-tailed distribution:

```
# Produce histogram to visualize sample distribution
ggplot() +
  aes(demographics$org_tenure) +
  labs(x = "Organization Tenure", y = "Density") +
  geom_histogram(aes(y = ..density..), fill = "#414141") +
  geom_density(fill = "#ADD8E6", alpha = 0.6) +
  theme_bw()
```

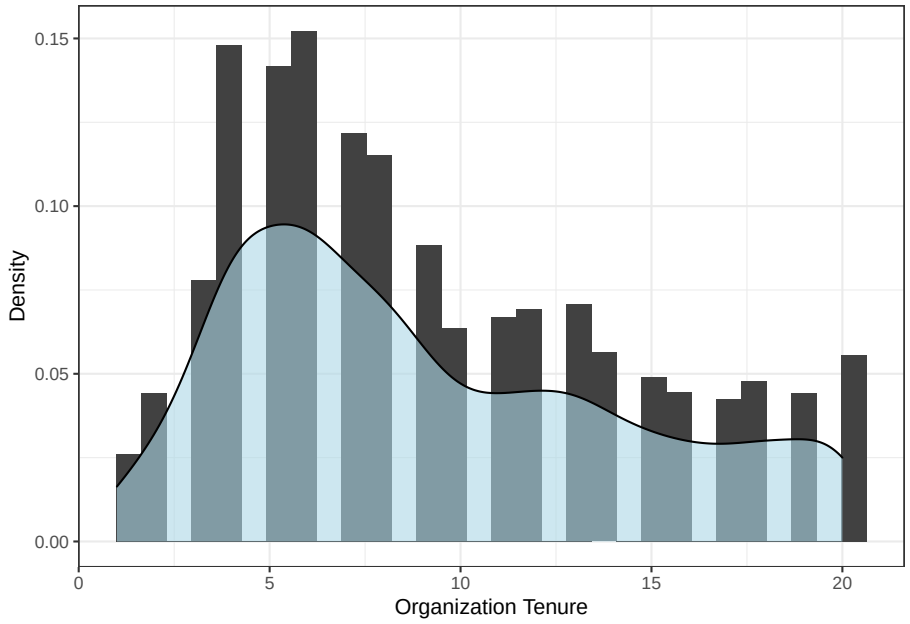


Figure 6.2: Organization Tenure Distribution

Kurtosis

While skewness provides information on the symmetry of a distribution, **kurtosis** provides information on the heaviness of a distribution's tails ("tailedness"). Kurtosis is the fourth statistical moment, defined by:

$$K = \frac{1}{n}\frac{\sum(x_i - \bar{x})^4}{s^4}$$

Note that the quartic functions characteristic of the fourth statistical moment are the only differences from the skewness formula we reviewed in the prior section (which featured cubic functions).

The terms **leptokurtic** and **platykurtic** are often used to describe distributions with light and heavy tails, respectively. "Platy-" in platykurtic is the same root as "platypus", and I've found it helpful to recall the characteristics of the flat platypus when characterizing frequency distributions as platkurtic (wide and flat) vs. its antithesis, leptokurtic (tall and skinny). The normal (or Gaussian) distribution is referred to as a **mesokurtic** distribution in the context of kurtosis.

Figure 6.3 illustrates the three kurtosis categorizations:

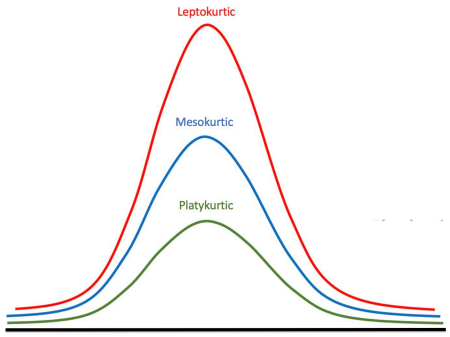


Figure 6.3: Kurtosis

Kurtosis is measured relative to a normal distribution. Normal distributions have a kurtosis coefficient of 3. Therefore, the kurtosis coefficient is greater than 3 for leptokurtic distributions and less than 3 for platykurtic distributions.

The moments library can also be used to calculate kurtosis in R:

```
# Calculate kurtosis for org tenure, rounded to two significant figures
round(kurtosis(demographics$org_tenure), 2)
```

```
## [1] 2.26
```

We can verify that the kurtosis() function returns the expected value (per the aforementioned formula) by validating against a manual calculation:

```
# Calculate kurtosis manually, rounded to two significant figures
round(1/n * (sum((x - x_bar)^4) / s^4), 2)
```

```
## [1] 2.26
```

As we saw in Figure 6.2, there is a long right tail about the organization tenure distribution. However, due to the moderate skew, there is no left tail. Therefore, our kurtosis coefficient is < 3 since the presence of the right tail is offset by the lack of a left tail. This is why it is important not to characterize a distribution based on a single isolated metric; we need the complete set of statistical moments to fully understand the distribution of data.

6.2 Bivariate Analysis

6.2.1 Covariance

While variance provides an understanding of how values for a single variable vary, **covariance** is an unstandardized measure of how two variables vary together. Values can range from $-\infty$ to $+\infty$, and these values can be used to understand the direction of the linear relationship between variables. Positive covariance values indicate that the variables vary in the same direction (e.g., tend to increase or decrease together), while negative covariance values indicate that the variables vary in opposite directions (e.g., when one increases, the other decreases, or vice versa).

Covariance of a sample is defined by:

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

It's important to note that while covariance aids our understanding of the direction of the relationship between two variables, we cannot use it to understand the strength of the association since it is unstandardized. Due to differences in variables' units of measurement, the strength of the relationship between two variables with large covariance could be weak, while the strength of the relationship between another pair of variables with small covariance could be strong.

In R, we can compute the covariance between a pair of numeric variables by passing the two vectors into the cov() function:

```
# Calculate sample covariance between annual compensation and age using complete obser
cov(demographics$annual_comp, demographics$age, use = "complete.obs")
```

```
## [1] 453179.1
```

In this example, using the default Pearson method, the covariance between annual compensation and age is 453179.1. The positive value indicates that annual compensation is generally higher for older employees and lower for younger employees.

Just as we multiplied the sample variance by $(n-1)/n$ to obtain the population variance, we can apply the same approach to convert the sample covariance returned by cov() to the population covariance:

```
# Calculate population covariance between annual compensation and age
cov(demographics$annual_comp, demographics$age, use = "complete.obs") * (n - 1) / n
```

```
## [1] 453086.1
```

The examples thus far have only examined associations between two variables at a time. However, rather than looking at isolated pairwise relationships, we can produce a covariance matrix to surface associations among many variables by passing a dataframe or matrix object into the cov() function:

```
# Generate a correlation matrix among continuous variables
cov(demographics[, c("annual_comp", "age", "org_tenure", "job_tenure")], use = "complete.obs")
```

```
##              annual_comp          age   org_tenure  job_tenure
## annual_comp 1.876688e+09 453179.07281 138551.67644 61576.73603
## age         4.531791e+05    168.38661     41.56006    17.96902
## org_tenure  1.385517e+05     41.56006     26.03270    11.04572
## job_tenure  6.157674e+04     17.96902     11.04572    13.29996
```

Using the default Pearson method, the cov() function will return sample variances for each variable down the diagonal, since covariance is not applicable in the context of a variable with itself. We can confirm by producing the variance for age:

```
# Return sample variance for age
var(demographics$age)
```

```
## [1] 168.3866
```

As expected, the variance for age ($s^2 = 168.39$) matches the value found in the age x age cell of the covariance matrix.

6.2.2 Correlation

Correlation is a scaled form of covariance. While covariance provides an unstandardized measure of the direction of a relationship between variables, correlation provides a standardized measure that can be used to quantify both the direction and strength of bivariate relationships. Correlation coefficients range

from -1 to 1, where -1 indicates a perfectly negative association, 1 indicates a perfectly positive association, and 0 indicates the absence of an association. **Pearson's product-moment correlation coefficient** r is defined by:

$$r_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_(x_i - \bar{x})^2 \sum_(y_i - \bar{y})^2}}$$

In R, Pearson's r can be calculated using the cor() function:

```
# Calculate the correlation between annual compensation and age
cor(demographics$annual_comp, demographics$age, use = "complete.obs")
```

```
## [1] 0.8061577
```

While we already know that the relationship between annual compensation and age is positive based on the positive covariance coefficient, Pearson's r of .81 indicates that the strength of the positive association is strong ($r = 1$ is perfectly positive). Though there are no absolute rules for categorizing the strength of relationships, as thresholds often vary by domain, the following is a general rule of thumb for interpreting the strength of bivariate associations:

- Weak = Absolute value of correlation coefficients between 0 and .3
- Moderate = Absolute value of correlation coefficients between .4 and .6
- Strong = Absolute value of correlation coefficients between .7 and 1

There are several correlation coefficients, and the measurement scale of x and y determine the appropriate type:

Measurement Scale		Correlation Coefficient
x	*y*	
Continuous	Continuous	Pearson's Product Moment
Continuous	Dichotomous	Point-Biserial
Continuous	Ordinal	Spearman or Kendall Rank
Dichotomous	Dichotomous	Phi, Contingency
Ordinal	Dichotomous	Rank-Biserial
Ordinal	Ordinal	Spearman or Kendall Rank

Figure 6.4: Proper Applications of Correlation Coefficients

Pearson's r can be used when both variables are measured on continuous scales or when one is continuous and the other is dichotomous (point-biserial correlation).

When one or both variables are ordinal, we can leverage **Spearman's** ρ or **Kendall's** τ, which are both standardized nonparametric measures of the association between one or two rank-ordered variables. Let's look at Spearman's ρ, which is defined as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

We can override the default Pearson method in the cor() function to implement a specific form of rank correlation using the `method` argument:

```
# Calculate the correlation between job level and education level using Spearman's method
cor(demographics$job_level, demographics$education, method = "spearman", use = "complete.obs")
```

```
## [1] 0.6749776
```

The ρ coefficient of .67 indicates that the positive association we observed between job level and education level is moderate-to-strong. We could also pass `method = "kendall"` to this cor() function to implement Kendall's τ.

The **Phi Coefficient** (ϕ), sometimes referred to as the mean square contingency coefficient or Matthews correlation in ML, can be used to understand the association between two dichotomous variables. For a 2x2 table for two random variables x and y:

	y = 0	**y = 1**
x = 0	A	B
x = 1	C	D

Figure 6.5: 2x2 Table for Random Variables x and y

The ϕ coefficient is defined as:

$$\phi = \frac{(AD - BC)}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}$$

To illustrate, let's examine whether there is a relationship between gender and performance after transforming performance from its ordinal form to a dichotomous variable (high vs. low performance). We can leverage the `psych` library to calculate ϕ in R:

```
# Load library for Phi Coefficient
library(psych)

# Set females to 1 and everything else to 0
demographics$gender_code <- ifelse(demographics$gender == 'Female', 1, 0)

# Set high performers (3 and above) to 1 and everything else to 0
demographics$performance_code <- ifelse(demographics$performance < 3, 0, 1)

# Create a 2x2 contingency table
contingency_tbl <- table(demographics$gender_code, demographics$performance_code)

# Calculate the Phi Coefficient between dichotomous variables
phi(contingency_tbl)
```

```
## [1] 0.01
```

ϕ is essentially 0, which means performance ratings are distributed equitably across gender categories (good news!).

A correlation matrix can be produced to surface associations among many variables by passing a dataframe or matrix object into the cor() function:

```
# Generate a correlation matrix among continuous variables
cor(demographics[, c("annual_comp", "age", "org_tenure", "job_tenure")], use = "complet
```

```
##             annual_comp       age org_tenure job_tenure
## annual_comp   1.0000000 0.8061577  0.6268392  0.3897584
## age           0.8061577 1.0000000  0.6277154  0.3797042
## org_tenure    0.6268392 0.6277154  1.0000000  0.5936211
## job_tenure    0.3897584 0.3797042  0.5936211  1.0000000
```

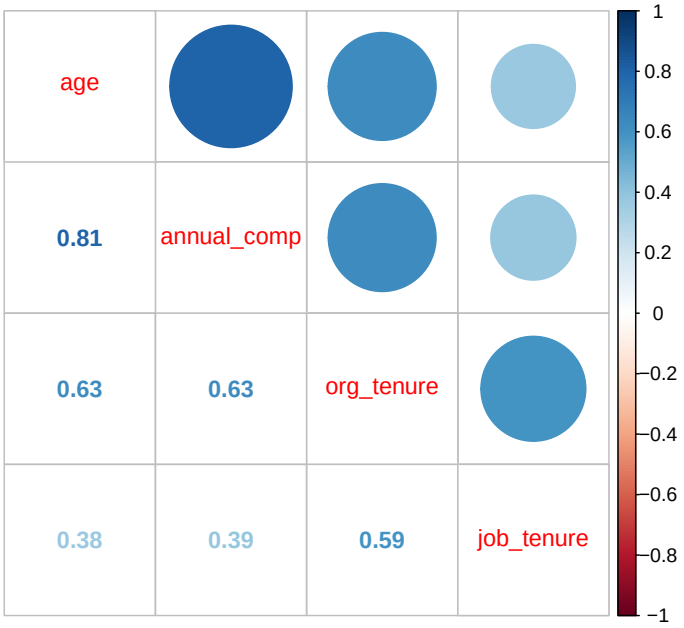
Based on this correlation matrix, there are several moderate and strong pairwise associations in the data. The values down the diagonal are 1 because these represent the correlation between each variable and itself. You may also notice that the information above and below the diagonal is identical and, therefore, redundant.

A great R library for visualizing correlation matrices is corrplot. Several arguments can be specified for various visual representations of the relationships among variables:

```
# Load library for correlation visuals
library(corrplot)
```

```r
# Store correlation matrix to object M
M <- cor(demographics[, c("annual_comp", "age", "org_tenure", "job_tenure")], use = "complete.obs

# Visualize correlation matrix
corrplot.mixed(M, order = 'AOE')
```



It's important to remember that correlation is not causation. Correlations can be spurious (variables related by chance), and drawing conclusions based on bivariate associations alone – especially in the absence of sound theoretical underpinnings – can be dangerous. Here are two examples of nearly perfect correlations between variables for which there is likely no true direct association:

In addition, covariance and correlation alone are not sufficient for determining whether an observed association in sample data is also present in the population. To understand the likelihood that patterns observed in sample data are also present in the larger population of interest, we need to move beyond descriptive measures.

6.3 Exercises

1. Which of the following measures of central tendency is least sensitive to extreme values (outliers)? A. Median B. Mean C. Range

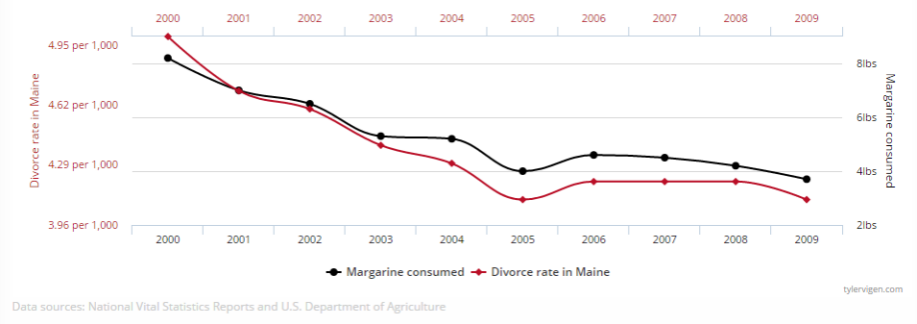


Figure 6.6: Correlation between Maine Divorce Rate and Margarine Consumption (r = .99)

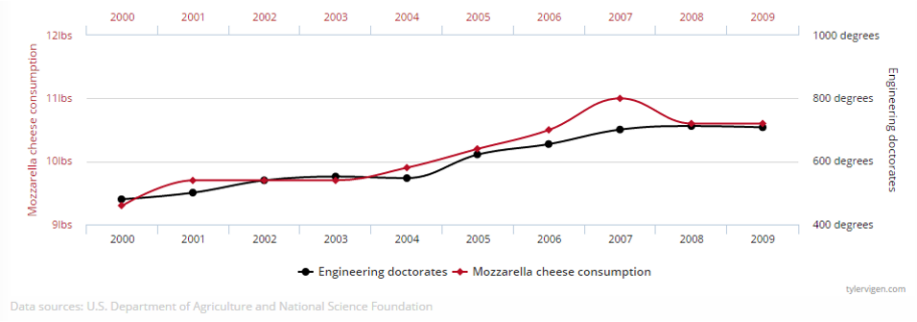


Figure 6.7: Correlation between Mozzarella Cheese Consumption and Civil Engineering Doctorate Conferrals (r = .96)

2. The standard deviation represents the 'average' amount by which x values deviate (or vary) from the mean. A large standard deviation indicates there is considerable spread in the data, whereas a small standard deviation indicates the mean is fairly representative of the data. A. True B. False

3. A positively skewed distribution has its largest allocation to the left and a negative distribution to the right. A. True B. False

4. Large covariance coefficients always indicate strong bivariate associations. A. True B. False

5. Which of the following can be found in boxplots? A. Quartiles B. Median C. Mean D. IQR E. Outliers

6. The 3rd quartile (Q3) is equivalent to the 75th percentile. A. True B. False

7. Which of the following correlation coefficients can be used when evaluating the relationship between a pair of rank-ordered variables? A. Pearson's Product Moment B. Spearman's Rank C. Phi D. Point-Biserial E. Kendall's Rank

8. Which of the following correlation coefficients can be used when evaluating the relationship between a pair of dichotomous variables? A. Pearson's Product Moment B. Spearman's Rank C. Phi D. Point-Biserial E. Kendall's Rank

9. Platykurtic distributions are flat relative to mesokurtic distributions. A. True B. False

10. When using the Pearson method, values down the diagonal of a covariance matrix represent each variable's variance. A. True B. False

Chapter 7

Inferential Statistics

The objective of **inferential statistics** is to make inferences – with some degree of confidence – about a population based on available sample data. Several related concepts underpin this goal and will be covered here.

7.0.1 Introduction to Probability

Randomness and uncertainty exist all around us. In **probability theory**, random phenomena refer to events or experiments whose outcomes cannot be predicted with certainty (Pishro-Nik, 2014). If you've taken a course in probability, there is a good chance you have considered the case of a fair coin flip – one of the most intuitive applications of probability. In the absence of information on how the coin is flipped, we cannot be certain of the outcome. What we can be certain of is that with a large number of coin flips, the proportion of heads will become increasingly close to 50%, or $\frac{1}{2}$.

The **Law of Large Numbers (LLN)** is an important theorem for building an intuitive understanding of how probability relates to the statistical inference concepts we will cover. In the case of a fair coin flip, it is possible to observe many consecutive heads by chance. This is because small samples can lend to anomalies. However, as the number of flips increases, we will undoubtedly observe an increasing number of tails; we expect a roughly equal number of heads and tails with a large enough number of flips.

Conditional probability reflects the probability conditioned on the occurrence of a previous event or outcome. For example, we may find that the proportion of heads is greater or less than $\frac{1}{2}$ with a large number of fair coin flips when the coin is consistently heads up when flipped. The outcome is, therefore, conditioned on the fixed – rather than random – positioning of the coin when flipped.

Formally, **Bayes' Theorem (alternatively, Bayes' Rule)** states that for any two events A and B where the probability of A is not 0 ($P(A) \neq 0$):

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Bayes' Rule allows us to predict the outcome more accurately by conditioning the probability on known factors rather than assuming all events operate under the same conditions. Bayes' Rule is pervasive in people analytics, as the probability of outcomes can vary widely based on a person's age, tenure, education, job, perceptions, relationships, and many other factors. For example, if we consider a company with 100 terminations over a 12-month period and average headcount of 1,000, the probability of attrition not conditioned on any other factor is 10%, or $\frac{1}{10}$. Aside from trending this probability over time to identify if attrition is becoming more or less of a concern, this isn't too helpful at the broader company level. However, if we condition the probability of attrition on an event – such as a recent manager exit – and find that the probability of attrition among those whose manager has left in the last six months is 70%, or $\frac{7}{10}$, this is far more actionable (and concerning).

7.0.2 Central Limit Theorem

The **Central Limit Theorem (CLT)** is a mainstay of statistics and probability and fundamental to understanding the mechanics of multivariate inferential analysis techniques we will cover later in this book. The CLT was initially coined by a French-born mathematician named Abraham De Moivre in the 1700s. While initially unpopular, it was later reintroduced and attracted new interest from theorists and academics (Daw & Pearson, 1972).

The CLT states that the average of independent random variables, when increased in number, tend to follow a normal (or Gaussian) distribution. The distribution of sample means approaches a normal distribution regardless of the shape of the population distribution from which the samples are drawn. This is important because the normal distribution has properties that can be used to test the likelihood that an observed value, difference, or relationship in a sample is also present in the population.

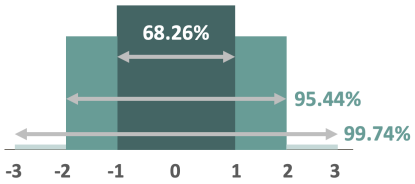


Figure 7.1: The Empirical Rule

Let's begin with an intuitive example of CLT. Imagine that we have a reliable way to measure how fun a population is on a 100-point scale, where 100 indicates maximum fun (life of the party) and 1 indicates maximum boringness. Consider that a small statistics conference is in progress at a nearby convention center, and there are 40 statisticians in attendance. In a separate room at the same convention center, there is also a group of 40 random people (non-statisticians) who are gathered to discuss some less interesting topic. Our job is to walk into one of the rooms and determine – based on the "fun" factor alone – whether we have entered the statistics conference or the other, less interesting gathering of non-statisticians.

Instinctively, we already know the statisticians will be more fun than the other group. However, let's assume we need the mean fun score and standard deviation of these two groups for this example. The group of statisticians have, on average, a fun score of 85 with a standard deviation of 2, while the group of non-statisticians are a bit less fun with a mean score of 65 and a standard deviation of 4. With a known population mean and standard deviation, the standard error (the standard deviation of the sample means) provides the ability to calculate the probability that the sample (the room of 40 people) belongs to the population of interest (fellow statisticians).

Herein lies the beauty of the CLT: roughly 68 percent of sample means will lie within one standard error of the population mean, roughly 95 percent within two standard errors of the population mean, and roughly 99 percent within three standard errors of the population mean. Therefore, any room whose members have an average fun score that is not within two standard errors of the population mean (between 81 and 89 for our statisticians) is statistically unlikely to be the group of statisticians for which we are searching. This is because in less than 5 in 100 cases could we randomly draw a 'reasonably sized' sample of statisticians with average funness so extremely different from the population average.

Because small samples lend to anomalies, we could – by chance – select a single person who happens to fall in the tails (extremely boring or extremely fun); however, as the sample size increases, it becomes more and more likely that the observed average reflects the average of the larger population. It would be virtually impossible (in less than 1 in 100 cases) to draw a random sample of statisticians from the population with average funness that is not within three standard errors of the population mean (between 79 and 91). Therefore, if we find that the room of people have an average fun score of 75, we will likely have far more fun in the other room!

Let's now see the CLT in action by simulating a random uniform population distribution from which we can draw random samples. Remember, the shape of the population distribution does not matter; we could simulate an Exponential, Gamma, Poisson, Binomial, or other distribution and observe the same behavior.

```
# Set seed for reproducible random distribution
set.seed(1234)

# Generate uniform population distribution with 1000 values ranging from 1 to 100
rand.unif <- runif(1000, min = 1, max = 100)
```

```
# Calculate population mean
mean(rand.unif)
```

```
## [1] 51.22007
```

```
# Calculate population variance
N = length(rand.unif)
var(rand.unif) * (N - 1) / N
```

```
## [1] 830.3155
```

```
# Produce histogram to visualize population distribution
ggplot() +
  aes(rand.unif) +
  labs(x = "x", y = "Density") +
  geom_histogram(aes(y = ..density..), fill = "#414141") +
  geom_density(fill = "#ADD8E6", alpha = 0.6) +
  theme_bw()
```

As expected, these randomly generated data are uniformly distributed. Next, we will draw 100 random samples of various sizes and plot the average of each.

```
# Define number of samples to draw from population distribution
samples <- 10000

# Populate vector with sample sizes
sample_n <- c(1:5,10,25,50)

# Initialize empty data frame to hold sample means
sample_means = NULL

# Set seed for reproducible random samples
set.seed(456)

# For each n, draw random samples
for (n in sample_n) {
```

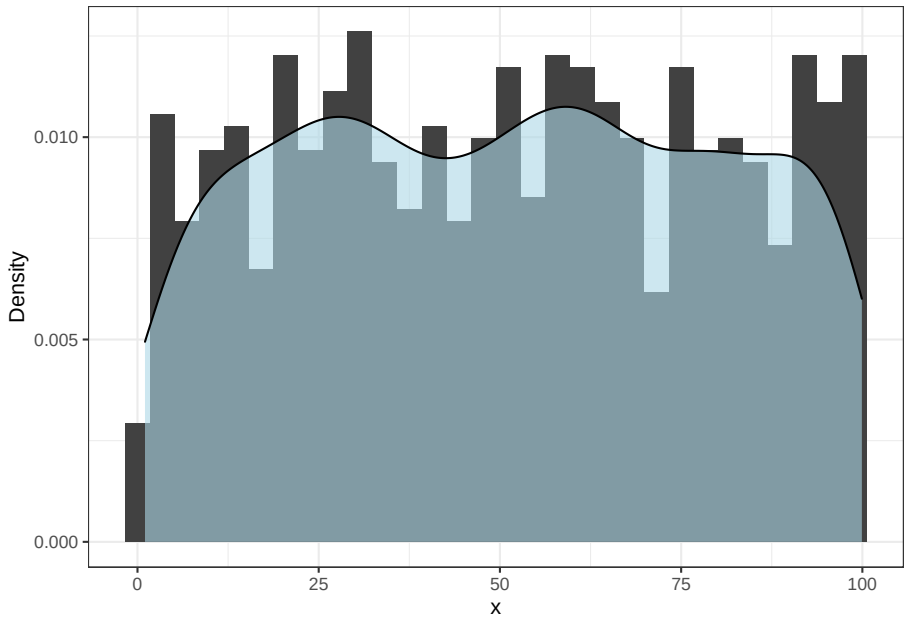


Figure 7.2: Uniform Population Distribution (N = 1000)

```
  for (draw in 1:samples) {

      # Store sample means in data frame
      sample_means <- rbind(sample_means, cbind.data.frame(
                            n = n,
                            x_bar = mean(sample(rand.unif, n, replace = TRUE, prob = NULL))))
  }
}

# Produce histograms to visualize distributions of sample means, grouped by n-count
sample_means %>% ggplot() +
  aes(x = x_bar, fill = n) +
  labs(x = "x-bar", y = "Density") +
  geom_histogram(aes(y = ..density..), fill = "#414141") +
  geom_density(fill = "#ADD8E6", alpha = 0.6) +
  theme_bw() +
  facet_wrap(~n)
```

Per the CLT, we can see that as n increases, the sample means become more normally distributed.

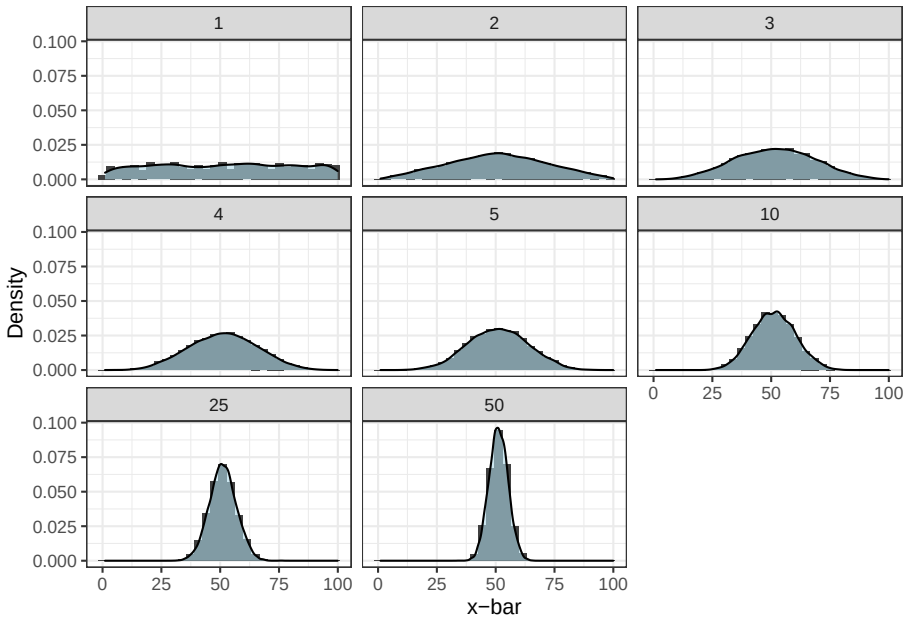


Figure 7.3: Distribution of 10,000 Sample Means of Varied Size

7.0.3 Confidence Intervals

A **Confidence Interval (CI)** represents a range of values likely to include a population parameter of interest (usually μ). A related concept that is fundamental to estimating CIs is the **standard error (SE)**, which is the standard deviation of sample means. While the standard deviation is a measure of variability for random variables, the variability captured by the SE reflects how representative the sample is of the population. Since sample statistics will approach the actual population parameters as the size of the sample increases, the SE and sample size are inversely related; that is, the SE decreases as the sample size increases. The SE is defined by:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Next, let's validate that our simulated distribution of sample means adheres to the properties of normally distributed data per the Empirical Rule:

```
# Store sample means with n = 50
x_bars <- sample_means[sample_means$n == 50, "x_bar"]

# Store sample size
```

```r
n <- length(x_bars)

# Calculate percent of sample means within +/- 2 SEs
length(x_bars[x_bars < mean(x_bars) + 2 * sd(x_bars) & x_bars > mean(x_bars) - 2 * sd(x_bars)]) /
```

```
## [1] 95.35
```

97% of sample means are within 2 SEs, which is roughly what we expect per the characteristics of the normal distribution.

```r
# Calculate percent of sample means within +/- 3 SEs
length(x_bars[x_bars < mean(x_bars) + 3 * sd(x_bars) & x_bars > mean(x_bars) - 3 * sd(x_bars)]) /
```

```
## [1] 99.79
```

All of the sample means are within 3 SEs, indicating that it would be highly unlikely – nearly impossible even – to observe a sample mean 'from the same population' that falls outside this interval.

Now, let's illustrate the relationship between CIs and standard errors using sample data from our uniform population distribution. In our example, both μ and σ are known and our sample size n is at least 30; therefore, we can use a **Z-Test** to calculate the 95% CI. A z score of 1.96 corresponds to the 95% CI for a two-tailed distribution; that is, we are looking for significantly different values in either the larger or smaller direction. The 95% CI represents the range of values we would expect to include μ in at least 95 of 100 random samples taken from the population.

The CI in this case is defined by:

$$CI = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Let's randomly take $n = 100$ from the population, and compute sample statistics to estimate the 95% CI:

```r
# Set seed for reproducible random samples
set.seed(456)

# Sample 100 values from uniform population distribution
x <- sample(rand.unif, 100, replace = TRUE, prob = NULL)

# Calculate 95% CI
ci95_lower_bound <- mean(x) - 1.96 * (sd(x) / sqrt(100))
ci95_upper_bound <- mean(x) + 1.96 * (sd(x) / sqrt(100))
```

```
# Print lower bound for 95% CI
ci95_lower_bound
```

```
## [1] 47.90733
```

```
# Print upper bound for 95% CI
ci95_upper_bound
```

```
## [1] 58.98773
```

Our known μ is 51.2, which is covered by our 95% CI (47.9 - 59.0). Per the CLT, in less than 5% of cases would we expect to draw a random sample from the population that results in a 95% CI which does not include μ. Note that our CI narrows with larger samples since our confidence that the range includes μ increases with more data.

Next, let's look at a 99% CI. We will enter 2.576 for z:

```
# Calculate 99% CI
ci99_lower_bound <- mean(x) - 2.576 * (sd(x) / sqrt(100))
ci99_upper_bound <- mean(x) + 2.576 * (sd(x) / sqrt(100))
```

```
# Print lower bound for 99% CI
ci99_lower_bound
```

```
## [1] 46.16612
```

```
# Print upper bound for 99% CI
ci99_upper_bound
```

```
## [1] 60.72893
```

Like the 95% CI, this slightly wider 99% CI (46.2 - 60.7) also includes our μ of 51.2.

If σ is not known, and/or we have a small sample ($n < 30$), we need to use a **T-Test** to calculate the CIs. In a people analytics setting, the reality is that population parameters are often unknown. For example, if we knew how engagement scores vary in the employee population, there would be no need to survey a sample of employees and make inferences about said population.

As we will see, the T-Test underpins many statistical tests and models germane to the people analytics discipline since we are often working with small datasets,

so it is important to understand the mechanics. As shown in Figure 7.4, the t distribution is increasingly wider and shorter relative to the normal distribution as the sample size decreases; this is also characteristic of the sampling distribution of means for smaller samples we observed in our CLT example. Specifically, **degrees of freedom (df)** is used to determine the shape of the probability distribution. Degrees of freedom represents the number of observations in the data that are free to vary when estimating statistical parameters, which is a function of the sample size $(n-1)$. For example, if we could choose 1 of 5 projects to work on each day between Monday and Friday, we would only be able to *choose* 4 out of the 5 days; on Friday, only 1 project would remain to be selected, so our degrees of freedom (the number of days in which we have a choice between projects) would be 4.

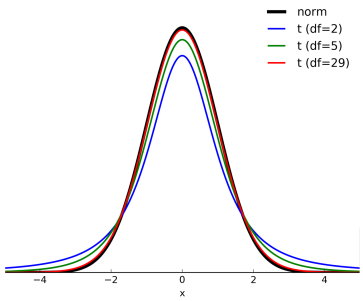


Figure 7.4: t Distribution Shape by Degrees of Freedom

When estimating the CI for smaller samples, we need to leverage the wider, more platykurtic t distribution to achieve greater accuracy. Therefore, the CI for a two-tailed test in this case is defined by:

$$CI = \bar{x} \pm t_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Let's compare CIs calculated using a T-Test to those calculated using the Z-Test. While a fixed z score can be used for each CI level when $n > 30$, the t statistic varies based on both the CI level and df. Though R will determine the correct t statistic for us, let's reference the table shown in Figure **??** to manually lookup the t statistic:

For illustrative purposes, let's draw a smaller sample of $n = 25$ from our uniform population distribution and calculate the 95% CI using the t statistic from the table ($df = 24$). The t statistic for this CI and df is 2.064:

```
# Set seed for reproducible random samples
set.seed(456)
```

cum.prob	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.995}$
one-tail	0.1	0.05	0.025	0.005
two-tails	0.2	0.1	0.5	0.01
df				
1	3.078	6.314	12.706	63.657
2	1.886	2.920	4.303	9.925
3	1.638	2.353	3.182	5.841
4	1.533	2.132	2.776	4.604
5	1.476	2.015	2.571	4.032
6	1.440	1.943	2.447	3.707
7	1.415	1.895	2.365	3.499
8	1.397	1.860	2.306	3.355
9	1.383	1.833	2.262	3.250
10	1.372	1.812	2.228	3.169
11	1.363	1.796	2.201	3.106
12	1.356	1.782	2.179	3.055
13	1.350	1.771	2.160	3.012
14	1.345	1.761	2.145	2.977
15	1.341	1.753	2.131	2.947
16	1.337	1.746	2.120	2.921
17	1.333	1.740	2.110	2.898
18	1.330	1.734	2.101	2.878
19	1.328	1.729	2.093	2.861
20	1.325	1.725	2.086	2.845
21	1.323	1.721	2.080	2.831
22	1.321	1.717	2.074	2.819
23	1.319	1.714	2.069	2.807
24	1.318	1.711	2.064	2.797
25	1.316	1.708	2.060	2.787
26	1.315	1.706	2.056	2.779
27	1.314	1.703	2.052	2.771
28	1.313	1.701	2.048	2.763
29	1.311	1.699	2.045	2.756
30	1.310	1.697	2.042	2.750
Z	1.282	1.645	1.960	2.576
	80%	90%	95%	99%
	Confidence Interval			

Figure 7.5: Critical Values of Student's t Distribution

```r
# Sample 25 values from uniform population distribution
x <- sample(rand.unif, 25, replace = TRUE, prob = NULL)

# Calculate 95% CI
ci95_lower_bound <- mean(x) - 2.064 * (sd(x) / sqrt(25))
ci95_upper_bound <- mean(x) + 2.064 * (sd(x) / sqrt(25))
```

```r
# Print lower bound for 95% CI
ci95_lower_bound
```

```
## [1] 35.24305
```

```r
# Print upper bound for 95% CI
ci95_upper_bound
```

```
## [1] 59.60959
```

As expected, the 95% CI using the t statistic is much wider (35.2 - 59.6), acknowledging the increased uncertainty in estimating population parameters given the limited information in this smaller sample. To increase our confidence to the 99% level, the interval widens even further (30.9 - 63.9):

```r
# Calculate 99% CI
ci99_lower_bound <- mean(x) - 2.797 * (sd(x) / sqrt(25))
ci99_upper_bound <- mean(x) + 2.797 * (sd(x) / sqrt(25))
```

```r
# Print lower bound for 99% CI
ci99_lower_bound
```

```
## [1] 30.91633
```

```r
# Print upper bound for 99% CI
ci99_upper_bound
```

```
## [1] 63.93631
```

Hypothesis Testing

Hypothesis testing is how we leverage CIs to test whether a significant difference or relationship exists in the data. Sir Ronald Fisher invented what is known as the null hypothesis, which states that there is no relationship/difference; disprove me if you can! The null hypothesis is defined by:

$$H_0 : \mu_A = \mu_B$$

The objective of hypothesis testing is to determine if there is sufficient evidence to reject the null hypothesis in favor of an alternative hypothesis. The null hypothesis always states that there is 'nothing' of significance. For example, if we want to test whether an intervention has an effect on an outcome in a population, the null hypothesis states that there is no effect. If we want to test whether there is a difference in average scores between two groups in a population, the null hypothesis states that there is no difference.

An alternative hypothesis may simply state that there is a difference or relationship in the population, or it may specify the expected direction (e.g., Population A has a significantly 'larger' or 'smaller' average value than Population B; Variable A is 'positively' or 'negatively' related to Variable B). Therefore, alternative hypotheses are defined by:

$$H_A : \mu_A \neq \mu_B$$

$$H_A : \mu_A < \mu_B$$

$$H_A : \mu_A > \mu_B$$

Alpha

The **alpha** level of a hypothesis test, denoted by α, represents the probability of obtaining observed results due to chance if the null hypothesis is true. In other words, α is the probability of rejecting the null hypothesis (and therefore claiming that there is a significant difference or relationship) when in fact we should have failed to reject it because there is insufficient evidence to support the alternative hypothesis.

α is often set at .05 but is sometimes set at a more rigorous .01, depending upon the context and tolerance for error. An α of .05 corresponds to a 95% CI (1 - .05), and .01 to a 99% CI (1 - .01). With non-directional alternative hypotheses, we must divide α by 2 (i.e., we could observe a significant result in either tail of the distribution), while one-tailed tests position the rejection region entirely within one tail based on what is being hypothesized.

At the .05 level, we would conclude that a finding is statistically significant if the chance of observing a value at least as extreme as the one observed is less than 1 in 20 if the null hypothesis is true. Note that we observed this behavior with our simulated distribution of sample means. While we could observe more extreme values by chance with repeated attempts, in less than 1 in every 20 times would we expect a 95% CI that does not capture μ. Moreover, in less

than 1 in every 100 times should we expect a sample with a 99% CI that does not capture μ.

Beta

Another key value is **Beta**, denoted by β, which relates to the power of the analysis. Simply put, power reflects our ability to find a difference or relationship if there is one. Power is calculated by 1 - β. At this point, it should be intuitive that larger samples increase our chances of observing significant results. As we observed in the T-Test example, CIs for small samples ($n < 30$) are quite wide relative to those for large samples; therefore, the power of the analysis to detect significance is limited given how extremely different values of x must be to observe non-overlapping CIs.

Type I & II Errors

A **Type I Error** is a false positive, wherein we conclude that there is a significant difference or relationship when there is not. A **Type II Error** is a false negative, wherein we fail to capture a significant finding. α represents our chance of making a Type I Error, while β represents our chance of making a Type II Error. I once had a professor explain that committing a Type I error is a shame, while committing a Type II error is a pity, and I've found this to be a helpful way to remember what each type of error represents.

	H_0 True	**H_0 False**
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

Figure 7.6: Type I and II Errors

P-Values

In statistical tests, the **p-value** is referenced to determine whether the null hypothesis can be rejected. The p-value represents the probability of obtaining a result at least as extreme as the one observed if the null hypothesis is true. As a general rule, if $p < .05$, we can confidently reject the null hypothesis and conclude that the observed difference or relationship was unlikely a chance observation.

While statistical significance helps us understand the probability of observing results by chance when there is no difference or effect in the population, it does not tell us anything about the size of the difference or effect. Analysis should never be reduced to inspecting p-values; in fact, p-values have been the subject of much controversy among researchers and practitioners in recent years. Later chapters will cover how to interpret results of statistical tests to surface the story

and determine if there is anything 'practically' significant among statistically significant findings.

Bonferroni Correction

One caveat when leveraging a p-value to determine statistical significance is that when multiple testing is performed – that is, multiple tests using the same sample data – the probability of a Type I error increases by a factor equivalent to the number of tests performed. It's important to note that there is not agreement among statisticians about how (or even whether) the p-value threshold for statistical significance needs to be adjusted to account for this increased risk. Nevertheless, we will cover this conservative approach for mitigating this risk.

Thus far, we have only discussed statistical significance in the context of a **per analysis error rate** – that is, the probability of committing a Type I error for a single statistical test. However, when two or more tests are being conducted on the same sample, the **familywise error rate** is an important factor in determining statistical significance. The familywise error rate reflects the fact that as we conduct more and more analyses on the same sample, the probability of a Type I error across the set (or family) of analyses increases. The familywise error rate can be calculated by:

$$\alpha_{FW} = 1 - (1 - \alpha_{PC})^C,$$

where c is equal to the number of comparisons (or statistical tests) performed, and α_{PC} is equal to the specified per analysis error rate (usually .05). For example, if $\alpha = .05$ per analysis, the probability of a Type I error with three tests on the same data increases from 5% to 14.3%: $1 - (1 - .05)^3 = .143$.

The most common method of adjusting the familywise error rate down to the specified per analysis error rate is the **Bonferroni Correction**. To implement this correction, we can simply divide α by the number of analyses performed on the dataset – such as $\alpha/3 = .017$ in the case of three analyses with $\alpha = .05$. This means that for each statistical test, we must achieve $p < .017$ to report a statistically significant result. An alternative which allows us to achieve the same number of statistically significant results is to multiply the unadjusted per analysis p-values for each statistical test by the number of tests. For example, if we run three statistical tests and receive $p = .014$, $p = .047$, and $p = .125$, we would achieve one significant result with the first method ($p < .017$) as well as with the alternative since the first statistical test satisfies the per analysis error rate ($p < .05$): $p = .014 * 3 = .042$.

Perneger (1998) is one of many who oppose the use of the Bonferroni Correction, suggesting that these "adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference." The Bonferroni Correction is controversial among researchers because while applying the correction reduces the chance of a Type I error, it also increases the chance of a Type II error. Because this correction makes it more difficult to detect significant results, it is

rare to find such a correction reported in published research, though research often involves multiple testing on the same sample. Perneger suggests that simply describing the statistical tests that were performed, and why, is sufficient for dealing with potential problems introduced by multiple testing.

7.1 Parametric vs. Nonparametric Tests

Parametric statistics assume the population is normally distributed. **Nonparametric statistics** do not assume anything about the population parameters or distribution and are, therefore, often referred to as distribution free tests. Assuming the normality assumption holds, parametric tests generally have more power than their nonparametric counterparts. This means that with a nonparametric test, we are less likely to reject the null hypothesis when it is false if the data come from normally distributed populations.

Since the mean (expected value) is the most common measure of central tendency, parametric tests usually focus on comparing the mean or variance of data. You may recall that μ and σ are sufficient to characterize a population distribution when data are situated symmetrically around the mean. However, the mean can be sensitive to outliers. If outliers are present in the data, the median may be a better way of representing the central tendency of data; in this case, nonparametric tests may be more appropriate.

In addition to normally distributed data in the population, and ensuring outliers are not materially influencing the mean, parametric tests also assume **homogeneity of variance** and **independence**. Homogeneity of variance assumes the variances across multiple groups are equal, though parametric tests are generally robust to violations of equal variances when the sample sizes are large. The assumption of independence requires observations to be randomly sampled from the population and independent of one another; that is, the value of one observation does not influence or depend on the value of another.

Spearman's correlation coefficient, which we used to evaluate the relationship between job level and education in **??**, is a nonparametric test since ordinal data are usually not normally distributed in the population. There is a nonparametric equivalent for each parametric test, and these will be reviewed in detail in Chapter 9.

It's important to remember that the normal distribution properties under the CLT relate to the sampling distribution of means – not to the distribution of the population or to the data for one individual sample. The CLT is important for estimating population parameters, but it does not transform a population distribution from nonnormal to normal. If we know the population distribution is nonnormal (e.g., ordinal, nominal, or skewed data), nonparametric tests should be leveraged.

7.2 The Monty Hall Problem

This chapter concludes with an example that highlights the importance of statistical assumptions by integrating concepts of conditional probability, randomness, and independence.

In the classic game show, Let's Make a Deal, Monty Hall asks contestants to choose one of three closed doors. Behind one door is a prize while the other two doors contain nothing. After the contestant selects a door, Monty opens one of the other two doors which does not contain a prize. At this point, there are two closed doors: the door the contestant selected and another for which the content remains unknown. All that is known at this point is that the prize is behind one of the two closed doors.

It is at this juncture that Monty introduces a twist by asking if the contestant would like to switch doors. Most assume that the two closed doors have an equal (50/50) chance of containing the prize, because we generally think of probabilities as independent, random events. However, this is incorrect. Contestants who switch from their original selection have a 66% chance (rather than 50%) of winning. This may be counterintuitive, because the brain wants to reduce the problem to a simple coin flip. There is a major difference between the Monty Hall problem and a coin flip; for two outcomes to have the same probability, randomness and independence are required. In the case of the Monty Hall problem, neither assumption is satisfied.

When all three doors are closed, each has the same probability of being selected. The probability of choosing the door with a prize is .33. Monty's knowledge of the door containing the prize does not impact the probability of selecting the winning door. This is because the choice is completely random given we have no information that would increase the probability of a door containing the prize. The process is no longer random when Monty uses his insider knowledge about the prize's location and opens a door he knows does not contain the prize. The probabilities change. Since Monty will never show the door containing the prize, he is careful to always open a door that has nothing behind it. If he was not constrained by the requirement to not reveal the prize's location and instead chose to open one of the remaining doors at random, the probabilities would be equal (and he may end up opening the door that contains the prize).

Seeing is believing, so let's prove this with a simulation in R:

```
# Set seed for reproducible simulations
set.seed(12345)

# Set number of simulations
trials = 10000

# Store switch/keep decisions
```

```r
decisions = c("switch", "keep")

# Store integer for each door
doors = 1:3

# Initialize empty data frame for results
results = NULL

for (n in 1:trials){

  for (decision in decisions){

    # Select correct door
    correct_door <- sample(doors, 1, replace = T)

    # Contestant chooses a door at random
    selected_door <- sample(doors, 1, replace = T)

    # Open door that was neither selected by the contestant nor contains the prize
    # Choose one door to open if multiple remain without the prize (i.e., the contestant didn't
    remaining_doors <- which(!doors == correct_door & !doors == selected_door)
    open_door <- sample(remaining_doors, 1, replace = T)

    # Contestant makes decision to switch doors or keep with the originally selected door
    selected_door <- ifelse(decision == "switch", which(!doors == selected_door & !doors == open_

    # Store results in data frame
    results <- rbind(results, cbind.data.frame(
                     trial = n,
                     decision = decision,
                     result = ifelse(correct_door == selected_door, "win", "lose")))
  }
}

# Calculate percentage difference in wins for switch vs. keep decisions
switch_wins <- nrow(results[results$decision == "switch" & results$result == "win", ]) / nrow(res
keep_wins <- nrow(results[results$decision == "keep" & results$result == "win", ]) / nrow(results
round((switch_wins - keep_wins) / keep_wins * 100, 0)
```

```
## [1] 45
```

As we can see, wins occur nearly 50% more often when contestants switch doors. This exercise hopefully demonstrates the importance of statistical assumptions. Also, if ever you find yourself playing Let's Make a Deal, switch doors.

7.3 Exercises

1. Which of the following is an example of a null hypothesis, where μ reflects the mean of a population? A. $\mu_A = \mu_B$ B. $\mu_A \neq \mu_B$ C. $\mu_A < \mu_B$ D. $\mu_A > \mu_B$ E. None of the above; all are examples of alternative hypotheses.

2. Which of the following describes a Type I Error? A. Failing to reject the null hypothesis when it is false B. Rejecting the null hypothesis when it is true C. Reporting something as significant when nothing of significance is present (a shame) D. Failing to detect something of significance (a pity) E. Both B and C F. Both A and D

3. The primary purpose of inferential statistics is to make inferences about a population based on sample data. Inferential statistics allows these inferences to be made with defined levels of confidence that what is observed in a sample is also characteristic of the larger population. A. True B. False

4. A T-Test should be used when σ is unknown and/or $n < 30$. A. True B. False

5. Randomness is essential to probabilistic methods. A. True B. False

6. Which of the following is characteristic of the Bonferroni Correction? A. Reducing the risk of a Type I Error B. Increasing the risk of a Type II Error C. Reducing the familywise error rate

7. Tolerance for a wider interval is an important tradeoff decision when increasing the level of confidence that a range of values contains an unknown population parameter. A. True B. False

8. A CI represents the range of values we expect to include an unknown population parameter (often the mean) for a specified degree of confidence. A. True B. False

9. When population parameters are unknown, which of the following tests is most appropriate for testing $\mu_A = \mu_B$? A. Z-Test B. T-Test C. Bayes' Theorem D. P-Test

10. According to the Empirical Rule, 95% of normally distributed data lie within how many standard deviations of the mean? A. 1 B. 2 C. 3 D. 4

Chapter 8

Data Preparation

8.1 Data Wrangling

8.2 Feature Engineering

Level one people analytics tends to utilize only the delivered fields from the HRIS (e.g., location, job profile, org tenure, etc.), but a good next step is to derive smarter variables from these fields. These can then be used to slice and dice turnover and engagement data differently, use as inputs in attrition risk models, etc. Below are some ideas to get you started:

- Number of jobs per unit of tenure (larger proportions tend to see greater career pathing)
- Office/remote worker (binary variable dummy coded as 1/0)
- Local/remote manager (binary variable dummy coded as 1/0)
- Hire/Rehire (binary variable dummy coded as 1/0)
- Hired/acquired (proxy for culture shock effects)
- Gender isolation (ratio of employee's gender to number of the same within immediate work group)
- Generation isolation (comparison of age bracket to most frequent generational bracket within immediate work group)
- Ethnic isolation (ratio of employee's ethnicity to number of the same within immediate work group)
- Difference between employee and manager age
- Percentage change between last two performance appraisal scores (per competency and/or overall)
- Team and department quit outbreak indicators (ratio of terms over x months relative to average headcount over x months)
- Industry experience (binary or length in years)

Remember to compute variables consistent with a need (e.g., is there reason to believe generationally isolated employees are more likely to term?). There may be a time and place for undertaking data mining initiatives with no a priori theories about what may be uncovered; however, more often than not, our efforts should be tied to specific hypotheses the business needs tested, which have sound theoretical underpinnings.

Chapter 9

Analysis of Differences

9.1 Comparing 2 Distributions

9.2 Comparing 3+ Distributions

Chapter 10

Linear Regression

It's important to draw a distinction between inferential and predictive models. Inferential models are highly interpretable and their utility is largely in understanding the nature and magnitude of the effect variables have on outcomes. Inferential models also lend to quantifying the extent to which we can generalize the observed effects to the larger population from which the sample was drawn. The objective in predictive modeling is to also to learn from patterns in historical data but for the purpose of achieving the most accurate predictions of future events – even at the expense of interpretability. To be clear, this isn't to say that predictive models cannot be interpreted – they certainly can – but I've seen relatively few applications for predictive modeling in people analytics because models generally need to be highly interpretable to support action planning.

This chapter is dedicated to inferential models to support a working understanding of how to interpret model output and communicate clear, data-driven narratives that respect the nuance and noise characteristic of people data. The following chapter will provide an overview of predictive modeling frameworks.

Regression is perhaps the most important statistical learning technique for people analytics. If you have taken a statistics course at the undergraduate or graduate levels, you have surely already encountered it. Before diving into the math to understand the mechanics of regression, let's develop an intuitive understanding.

Imagine we are sitting at a large public park in NYC on a nice fall afternoon. If asked to estimate the annual compensation of the next person to walk by, in the absence of any additional information how would you estimate this? Most would likely estimate the average annual compensation of everyone capable of walking by. Since this would include both residents and visitors, this would be a very large group of people! The obvious limitation with this approach is that among the large group of people capable of walking by, there is likely

a significant range of annual compensation values. Many walking by may be children, unemployed, or retirees who earn no annual compensation, while others may be highly compensated senior executives at the pinnacle of their careers. Since the range of annual compensation could be zero to billions of dollars, estimating the average of such a large population is likely going to be highly inaccurate without more information about who may walk by.

Let's consider that we are sitting outside on a weekday afternoon. Should this influence our annual compensation estimate? It is likely that we can eliminate a large segment of those likely to walk by, as we would expect most children to be in school on a typical fall weekday afternoon. It's also unlikely that those who are employed and not on vacation will walk by on a fall weekday afternoon. Therefore, factoring in that it is a weekday should limit the size of the population which in turn may reduce the range of annual compensation values for our population of passerbys.

Let's now consider that the park is open only to invited guests for a symposium on people analytics. Though it may be difficult to believe, a relatively small subset of the population is likely interested in attending such a symposium, so this information will likely be very helpful in reducing the size of the population who could walk by, which should further reduce the range of annual compensation since we probably have a good idea of the profile of those most likely to attend. This probably also lessens (or altogether eliminates) the importance of the weekday factor in explaining why people vary in the amount of compensation they earn each year.

Furthermore, let's consider that only those who reside in NYC and Boise were invited, and that the next person to walk by resides in Boise. Most companies apply a significant cost of living multiplier to the compensation for those in an expensive region such as NYC, resulting in a significant difference in compensation relative to those residing in a much less expensive city like Boise – all else being equal. Therefore, if we can partition attendees into two groups based on their geography, this should limit the range of annual compensation significantly within each – likely making the average compensation amount in each group a more nuanced and reasonable estimate.

What if we also learn the specific zip code in which the next passerby from Boise resides? The important information is likely captured in the larger city label (NYC vs. Boise), and the compensation for the specific zip codes within each city are unlikely to vary to a significant degree. Assuming this is true, it probably would not make sense to consider both the city name and zip code since they are effectively redundant pieces of information with regard to explaining differences in annual compensation.

What if we learn that the next person to walk by will be wearing a blue shirt? Does this influence your estimate? Unless there is research to suggest shirt color and earnings are related, this information will likely not contribute any significant information to our understanding of why people vary in the amount

of compensation they earn annually and should, therefore, not be considered.

You can probably think of many relevant variables that would help further narrow the range of annual compensation. These may include job, level, years of experience, education, location, among other factors. The main thing to understand is that for each group of observations with the same characteristics – such as senior analysts with a graduate degree who reside in NYC – there is a distribution of annual compensation. This distribution reflects unexplained variance. That is, we do not have information to explain why the compensation for each and every person is not the same and in social science contexts, it simply is not practical to explain 100 percent of the variance in outcomes. Two people may be similar on hundreds of factors (experience, education, skills) but one was simply a more effective negotiator when offered the same role and commanded a higher salary. It's likely we do not have data on salary negotiation ability so this information would leave us with unexplained variance in compensation. The goal is simply to identify the variables that provide the most information in helping us tighten the distribution so that estimating the average value will generally be an accurate estimate for those in the larger population with the same characteristics.

While we can generally improve our estimates with more relevant information (not shirt color or residential zip code in this case), it is important to understand that samples which are too small ($n < 30$) lend to anomalies; modeling noise in sparse data can result in models that are unlikely to generalize beyond the sample data. For example, if the only people from Boise to attend the people analytics symposium happen to be two ultra wealthy tech entrepreneurs who earn millions each year, it would not be appropriate to use this as the basis for our estimates of all future attendees from Boise. This is a phenomenon known as overfitting that will be covered later in this chapter.

This is the essence of regression modeling: find a limited number of variables which independently or jointly provide significant information that helps explain (by reducing) variance around the average value. As illustrated in this example, adding additional variables (information) can impact the importance of other variables or may offer no incremental information at all. In the subsequent sections, we will cover how to identify which variables are important and how to quantify the effect they have on the outcome.

10.1 Simple Linear Regression

10.1.1 Parameter Estimation

Ordinary Least Squares (OLS) is the most common method for estimating unknown parameters in a linear regression model.

10.2 Multiple Linear Regression

10.2.1 Moderation

10.2.2 Mediation

10.3 Polynomial Regression

10.4 Hierarchical Models

Chapter 11

Generalized Linear Regression

11.1 Logistic Regression

Logistic regression is an excellent tool when the outcome is categorical. Logistic regression allows us to model the probability of different classes – a type of modeling often referred to as classification. The context for classification can be binomial for two classes (e.g., active/inactive, promoted/not promoted), multinomial for multiple unordered classes (e.g., skills, job families), or ordinal for multiple ordered classes (e.g., survey items measured on a Likert scale, performance level).

11.1.1 Binomial Logistic Regression

11.1.2 Multinomial Logistic Regression

11.1.3 Ordinal Logistic Regression

11.1.4 Proportional Odds Logistic Regression

11.2 Poisson Regression

Chapter 12

Predictive Models

12.1 Bias-Variance Trade-Off

12.2 Cross-Validation

12.3 Balancing Classes

12.4 Model Performance

12.5 Automated Machine Learning (AutoML)

Chapter 13

Unsupervised Learning Models

13.1 Factor Analysis

13.2 Clustering

Chapter 14

Network Analysis

Chapter 15

Data Visualization

Chapter 16

Data Storytelling

Chapter 17

Bibliography

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51(6), 1173–1182.

Creswell, J. W., & Creswell, J. D. (2018). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (5th ed.). Los Angeles: Sage.

Daw, R. H., & Pearson, E. S. (1972). Studies in the History of Probability and Statistics. XXX. Abraham De Moivre's 1733 Derivation of the Normal Curve: A Bibliographical Note. Biometrika, 59(3), 677–680. https://doi.org/10.2307/2334818

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. New York: Springer.

Kahneman, D. (2011). Thinking, fast and slow. New York: Farrar, Straus and Giroux.

Kerlinger, F., & Lee, H. (2000). Foundations of behavioral research (4th ed.). Melbourne: Wadsworth.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. BMJ, 316(7139), 1236-1238.

Pishro-Nik, H. (2014). Introduction to Probability: Statistics and Random Processes. Blue Bell, PA: Kappa Research, LLC.

Wheelan, C. (2013). Naked Statistics: Stripping the Dread from the Data. New York: W.W. Norton.

Chapter 18

Appendix

18.1 Exercise Solutions

18.1.1 Univariate & Bivariate Analysis

1. Which of the following measures of central tendency is least sensitive to extreme values (outliers)? A. Median
2. The standard deviation represents the 'average' amount by which x values deviate (or vary) from the mean. A large standard deviation indicates there is considerable spread in the data, whereas a small standard deviation indicates the mean is fairly representative of the data. A. True
3. A positively skewed distribution has its largest allocation to the left and a negative distribution to the right. B. False
4. Large covariance coefficients always indicate strong bivariate associations. B. False
5. Which of the following can be found in boxplots? A. Quartiles B. Median D. IQR E. Outliers
6. The 3rd quartile (Q3) is equivalent to the 75th percentile. A. True
7. Which of the following correlation coefficients can be used when evaluating the relationship between a pair of rank-ordered variables? B. Spearman's Rank E. Kendall's Rank
8. Which of the following correlation coefficients can be used when evaluating the relationship between a pair of dichotomous variables? C. Phi
9. Platykurtic distributions are flat relative to mesokurtic distributions. A. True

10. When using the Pearson method, values down the diagonal of a covariance matrix represent the variance for each variable. A. True

18.1.2 Inferential Statistics

1. Which of the following is an example of a null hypothesis, where μ reflects the mean of a population? A. $\mu_A = \mu_B$
2. Which of the following describes a Type I Error? E. Both B and C
3. The primary purpose of inferential statistics is to make inferences about a population based on sample data. Inferential statistics allows these inferences to be made with defined levels of confidence that what is observed in a sample is also characteristic of the larger population. A. True
4. A T-Test should be used when σ is unknown and/or $n < 30$. A. True
5. Randomness is essential to probabilistic methods. A. True
6. Which of the following is characteristic of the Bonferroni Correction? A. Reducing the risk of a Type I Error B. Increasing the risk of a Type II Error C. Reducing the familywise error rate
7. Tolerance for a wider interval is an important tradeoff decision when increasing the level of confidence that a range of values contains an unknown population parameter. A. True
8. A CI represents the range of values we expect to include an unknown population parameter (often the mean) for a specified degree of confidence. A. True B. False
9. When population parameters are unknown, which of the following tests is most appropriate for testing $\mu_A = \mu_B$? B. T-Test
10. According to the Empirical Rule, 95% of normally distributed data lie within how many standard deviations of the mean? B. 2

18.2 4D Framework {#4d-chklst}

1. Discover

You are likely familiar with the old adage: "An ounce of prevention is worth a pound of cure." Such is the case with respect to planning in an analytics context. During the Discover phase, it is important to remain in the problem zone; seek to understand your clients' needs through active listening and questions. This is not the time for solutioning or committing to any specific deliverables. If the client's needs are ambiguous, proceeding will likely be an exercise in futility.

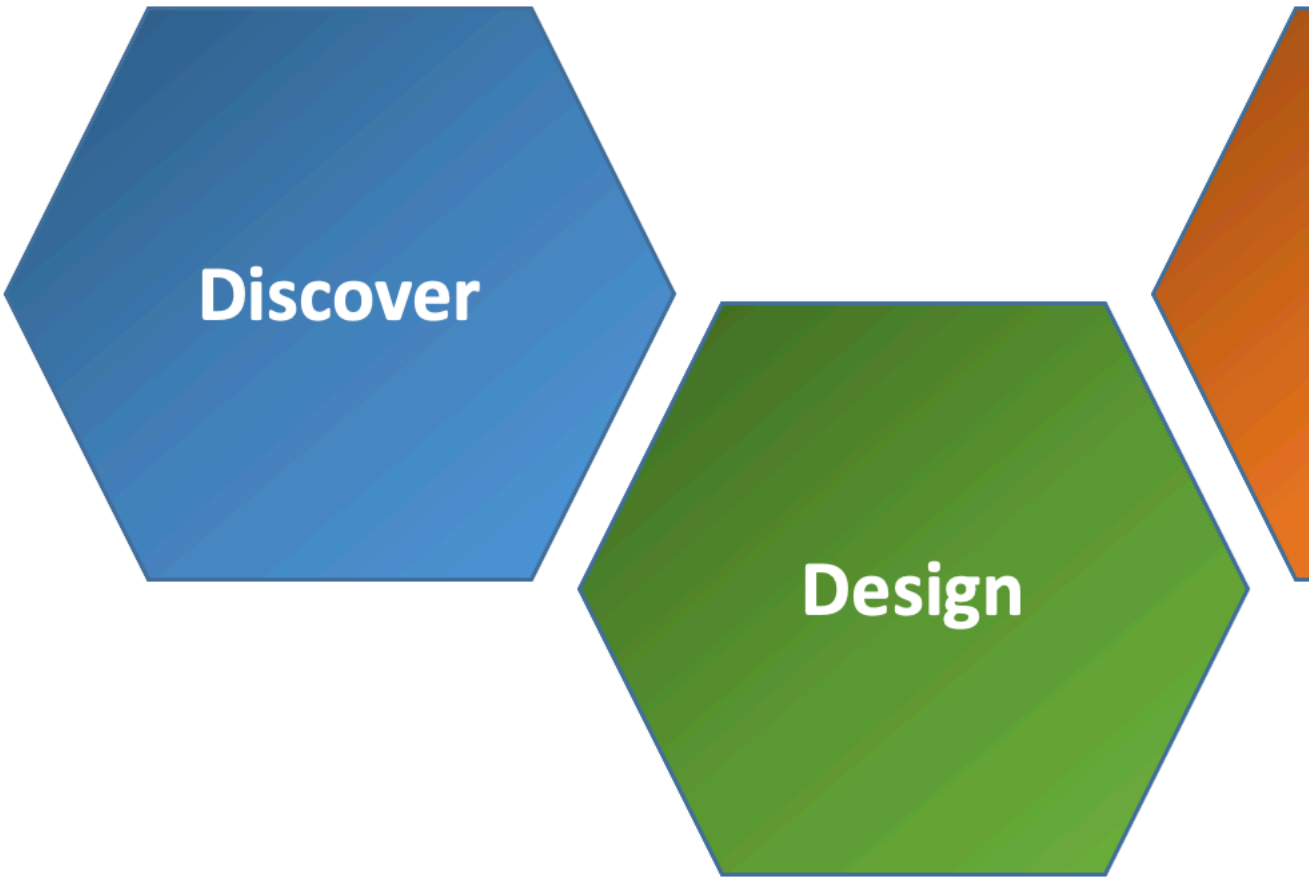


Figure 18.1: Figure 1: 4D Framework

Outlined below is a set of general questions that should be considered during this initial phase to prevent allocating scarce time and resourcing to a project that ultimately misses the mark.

- Client

  Who is the client? A client can be a person or organization that has contracting you for consulting services, or an internal stakeholder within your organization who has need. What is important to them?

- Primary Objective

  - What is the client ultimately hoping to accomplish?
  - Is the request merely to satisfy one's curiosity, or are there actions that can realistically be taken to materially influence said objective?

- Problem Statement

  - One of my most important early steps is clearly defining the problem statement. If your understanding of the problem – after translating from the business terms in which it was initially expressed – is misaligned with the client's needs, none of the subsequent steps matter.

- Guiding Theories

  - What theoretical explanations can the client offer as potential rationalizations for the phenomena of interest?
  - Are there existing theories in the organizational literature that should guide how the problem is tackled (e.g., findings from similar research implemented in other contexts)?

- Research Questions

  To respect the nuances of the problem statement, it is important to unpack it and frame as a set of overarching questions to guide the research.

  - Q1: ...
  - Q2: ...
  - Q3: ...

- Research Hypotheses

  Once research questions are developed, what do you expect to find based on anecdotal stories or empirical findings? As a next step, these expectations should be expressed in the form of research hypotheses. Please note that these research hypotheses are different from statistical hypotheses.

  - H1: ...
  - H2: ...
  - H3: ...

To ensure the hypotheses lend themselves to actionable analyses, it is important to consider the following: "What does success look like?" In other words, once the project is complete, against which success measures will the project's success be determined? Curiosity is not a business reason and hope is not a reasonable strategy. The following questions may prove helpful in the promotion of actionable – over merely interesting outcomes:

  - What will be done if the hypotheses are empirically supported?
  - What will be done if the hypotheses are not empirically supported?

- Assumptions

  At this point, it's helpful to consider what assumptions may be embedded in this discovery work. Are the questions and hypotheses rooted in what the client has theorized, or are these the product of an ambiguous understanding of the client's needs?

- Cadence

  - Is this analysis a one-off, or could there be a need to refresh this analysis on a regular cadence?
  - Are there dates associated with programs, actions, etc. this analysis is intended to support?

- Aggregation

  Is there a need for individual-level detail supporting the analysis? Aaggregate data should generally be the default unless a compelling justification exists and approval from legal and privacy partners is granted. One important role of analysts is to help keep the audience focused on the bigger picture and findings. Access to individual-level detail can not only introduce unnecessary legal and compliance risk but can also lead to questions and probing that can delay taking needed actions based on the results.

- Deliverable

  What is the preferred method of communicating the results of the analysis (e.g., interactive dashboard, static slide deck, document)? It is important to determine this early so that subsequent efforts can be structured to support the preferred deliverable. For example, if an interactive dashboard is preferred, does your Engineering department need to prioritize dependent tasks such as data feeds, row-level security, BI development, and production server migrations?

- Filters & Dimensions

  How does your client prefer to segment the workforce? Some common grouping dimensions are business unit, division, team, job family, location, tenure, and management level.

2. Design

Perhaps the most important initial question to answer in the design phase is: "Does anything already exist that addresses part, or all, of the client's objectives?" If the existing solution will suffice, it's possible that there is simply a communication/education gap, and you can allocate time and resources elsewhere.

The end-user experience is of paramount importance during the Design phase, as solutions should have a consistent look and feel regardless of who developed the product. To achieve this, it is important to resist siloed thinking and consider the broader set of analytics solutions the team has delivered – or is in the process of delivering.

- Data Privacy

  Are there potential concerns with the study's objective, planned actions, and/or requested data elements from an employee privacy or legal perspective? A cross-functional data governance committee can help with efficient and consistent decisioning on requests for people data and analytics.

- Data Sources & Elements

  - What data sources are required?
  - What data elements are required?

  In cases where sensitive attributes such as gender, ethnicity, age, sexual orientation, and disability status are requested, it's always best to exercise a 'safety first' mentality and consult with legal and privacy partners to ensure there is comfort with the intended use of the data. The decision on whether or not to include these sensitive data elements is often less about what the audience can view (e.g., People Partners may already have access to the information at the person level in the source system) and more anchored in what they plan to do with the information.

  Is the required data already accessible in a data warehouse or other analytics environment? If not, does it need to be? What is required to achieve this?

- Data Quality

  It is important to understand the data generative process and never make assumptions about how anomalies or missing data should be interpreted. After identifying what data sources will be required for a particular analysis, it is important to meet with source system owners and data stewards to deeply understand the business processes by which data are generated in the system(s). Are there data quality concerns that need to be explored and addressed?

- Variables

  How will the constructs be measured (e.g., survey instrument, derived attribute, calculated field)?

- Analysis Method

  What are the appropriate analysis methods based on the research hypotheses? If modeling is required, is it more important to index on accuracy or interpretability?

- Dependencies

  Are other teams required to develop this solution? What is the nature of the work each dependent team will perform? Are there required system configuration changes? Do these teams have capacity to support?

- Change Management

  Will this solution impact current processes or solutions? If so, what is the change management plan to facilitate a seamless transition and user experience?

- Sign-Off

  Generally, it is best for the client to signoff on the problem statement, analysis approach, and wire frame for the deliverable (if applicable) before providing an ETA and proceeding to the development phase. This ensures alignment on the client's needs and the perceived utility of the solution in addressing those needs.

3. Develop

- Development Patterns

  - Are there development patterns that should guide the development approach to support consistency?
  - Are there existing calculated fields that can/should be leveraged for derived data?
  - Are there best practices that should be employed to optimize performance (e.g., load time for dashboards, executing complex queries during non-peak times)?
  - Are there standard color palettes that should be applied?

- Productionalizable Code

  - How do models and data science pipelines need to be developed to facilitate a seamless migration from lower to upper environments? For example, initial exploratory data analysis (EDA) may be performed using curated data in flat files for the purpose of identifying meaningful trends, relationships, and differences, but where will this data

need to be sourced in production to automate the refresh of models at a regular interval? If the data were provided from multiple source systems, what joins are required to integrate the data? What transformation logic or business rules need to be applied to reproduce the curated data?

- Unit Testing
  - What test cases will ensure the veracity of data?
  - Who will perform the testing?
- UAT Testing
  - In the spirit of agility and constant contact with the client to prevent surprises, it is generally a good idea to have the client take the solution for a test run within the UAT environment and then provide sign-off before migrating to production. If the deliverable is a deck or doc with results from a model, UAT may surface clarifying questions that can be addressed before releasing to the broader audience.

4. Deliver

The Deliver phase can take many forms depending on the solution being released. If the solution is designed for a large user base, a series of recorded trainings may be in order so that there is a helpful reference for those unable to attend the live sessions or new joiners in the future. It is important to monitor success measures, which could be insights aligned to research hypotheses, dashboard utilization metrics, or any number of others defined within the Discover phase.