Data Applications in Public Administration

Learning Concepts and Skills in R

Alex E. Combs

2020-04-30

Contents

Preface **5**
Style of book 5
Intended audience 5
Objective of book 5
Structure of book 6
Software Prerequisites 6

1 Introduction **7**
1.1 Stats for Public Administrators 7
1.2 Using R 8

2 Data **11**
2.1 Learning objectives 11
2.2 Rectangular Data 11
2.3 Types of variables 12
2.4 Dataset structures 15

3 Measurment and Missingness **23**
3.1 Learning objectives 23
3.2 Credible Analysis 23

4 Data Wrangling **27**

5 Data Explorations and Description **29**

6 Data Visualization **31**

Preface

This book was developed after years of disappointment with texts used to teach MPA students statistics and statistical software. Curating sections and subsections of chapters from numerous sources in order to provide partially relevant information presented with various levels of rigor in an order that loosely followed the progression of the course proved counterproductive to teaching and learning. This book aims to remedy the lack of a stand-alone text appropriate for MPA students.

Style of book

This book is opinionated and uses stylized facts. There are usually multiple options to achieve an intended outcome; this book provides what the author considers or understands to be the best option. Rather than provide thorough coverage of relevant statistical concepts, this book takes liberties to simplify and relate complex concepts for the benefit of its intended audience.

Intended audience

This book is intended for masters students in public administration with minimal background in statistics or interest in pursuing a career in academic research. For students wishing to learn core concepts and skills to apply toward jobs in public, non-profit, and health sectors without suffering through irrelevant information often included in traditional texts, this book was written with you in mind.

Objective of book

Simply put, the objective of this book is to help students in public administration be as comptetive as possible in their intended job markets via competency in statistics and statistical software. It aims to simultaneously teach students key concepts in statistics and applications of those concepts using R.

Structure of book

The structure of this book corresponds to the schedule for PADP 7120 - Data Applications in Public Administration. Chapters are organized according to topics covered each week. Each chapter provides motivations for students in public administration to learn the topic, succinct coverage of the topic's concepts and applied skills, and relevant examples and exercises.

Software Prerequisites

This book provides examples and exercises using R. Readers wanting to take full advantage of this text must install the following software:

- R
- RStudio

R Packages

Many examples and exercises using `R` require you to install `R` packages that augment its functionality. Installing and loading packages will be covered in chapter one, but below is a one-stop-shop for all the packages you need to install to use this book. To do so, copy-and-paste the below code in your `console` pane of `RStudio`.

```
install.packages(c("tidyverse", "moderndive", "broom", "knitr", "mice", "car", "carData
```

Chapter 1

Introduction

"Data, data everywhere, and not a thought to think."

1.1 Stats for Public Administrators

Statistics converts raw information (i.e. data) into something useful. If we want to make evidence-based decisions, we need statistics. If we want allow ourselves to be misled by nefarious or mistaken manipulation of data, we should resist learning statistics.

1.1.1 Professional standards

The National Association of Schools of Public Affairs and Administration (NASPAA) is the accrediting authority for MPA programs. NASPAA promotes the following universal competencies:

- to lead and manage in the public interest;
- to participate in, and contribute to, the policy process;
- to analyze, synthesize, think critically, solve problems and make evidence-informed decisions in a complex and dynamic environment;
- to articulate, apply, and advance a public service perspective;
- to communicate and interact productively and in culturally responsive ways with a diverse and changing workforce and society at large.

Rest assured, statistics will help you develop all of the above competencies. In fact, by virtue of this and other methods courses being included in MPA curriculum, you would not sufficiently posess one or more of the above competencies without knowledge and skills in statistics.

1.2 Using R

Before moving forward with this book, you need to learn how to operate `R` at a basic level. The goal here is not to train you to be an expert in `R` or even a data scientist or analyst. Rather, the goal is to train you just enough so `R` becomes a legitimate alternative to inferior spreadsheet software like Excel and handle most tasks one might be expected to do after attaining an MPA.

1.2.1 Why R

MPA students may be reluctant to learn something referred to as a statistical computing language and its relevancy to their career goals may not be clear.

Look, demand for those competent in statistical software like `R` continues to increase. Even if you plan to pursue a managerial role with minimal analytical tasks, you will supervise or work with those who conduct analyses and interpret their findings. You need to be able to speak their langauge and vice versa.

`R` is free to use, comes with a bevy of free resources for learning, and is extremely popular. If you study statistics or data applications for only one semester, you should spend part of that semester learning software like `R`. There is only upside in doing so with respect to employment prospects.

1.2.2 What is R and RStudio

`R` is a programming language for statistical computing. `RStudio` is a user interface for `R`. We all have smartphones. Your phone has base code you never interact with directly but is what allows your phone to work. You interact with this code, doing all the cool things it allows you to do through what you see on the screen. `R` is like the base code for your phone. `RStudio` is the screen.

1.2.3 RStudio orientation

Launch RStudio.

Upon launch, you should see three panes:

- Console pane (left) is where you can tell R what to do. It also displays the results of commands. **Only use the console for installing packages.**
- Environment pane (top right) displays all the data in your current `R` session. A session is the time between launching and closing `R`.
- Files pane (bottom right) allows you to navigate your files, displays plots, provides a list of installed packages, allows you to search for help, and displays file exports.

You will usually see a fourth pane while working in `RStudio` – the source editor pane. In your menu bar, go to `File -> New File -> R Script`. A new pane will open. **This is the pane where you will tell `R` what to do 99% of the**

time because it allows you to do so while creating a document you can save and return to.

1.2.4 Installing and loading R packages

Your phone comes with base programs (e.g. calendar, weather), but others create third-party applications to augment the functionality of your phone. This is also the case with `R`, which has an active user community that develops useful third-party apps called packages.

Just like your phone, you have to first install a R package to ever use it. We install packages by typing the following code **into the console**.

```
install.packages("name_of_package")
```

We only need to install a package once. The package is saved on your computer where `R` can find it.

When an app is installed on our phone, we still have to launch it to use it. The same goes for using a `R` package. We use the following code to launch a package every time we need to use one of its functions.

```
library(package)
```

1.2.5 Exercises

1. If you haven't already done so, install the packages listed at the end of the preface.

2. You should have a script open in the upper-left pane. Type `library(tidyverse)` in the script. Click `Run` or `Cmd+Enter` to execute this line of code. Notice what is happening. R is launching a package. This is what you need to do before using any functions included in a particular package. Try to internalize this crucial fact. I will tell you what packages you need to load, but don't forget to load the package before trying to use one of its functions.

Chapter 2

Data

"Show me the data!"

We cannot effectively convert the raw material of knowledge into a useful product without first understanding the raw material. Therefore, learning statistics naturally begins with data.

2.1 Learning objectives

- Understand the organization of rectangular data
- Identify the unit of analysis within a dataset
- Identify and distinguish types of variables
- Identify and distinguish types of dataset structures

2.2 Rectangular Data

Most data are rectangular, often represented using a standard spreadsheet organized by rows and columns. A rectangle of data is commonly referred to as a **dataset**.

Generic rectangular data

ID

Variable_1

Variable_2

Unit of Analysis

Datum

Datum

Unit of Analysis

Datum

Datum

Unit of Analysis

Datum

Datum

A rectangular dataset has three components. Not all datasets will fit the below description because many datasets are not organized in a tidy manner. The elements of tidy data will be covered in a later chapter.

- **Unit of analysis or observation:** The generic entity or subject a row of data refers to. The unit of analysis uniquely identifies each row of a dataset. If we have a dataset of 50 states and some variables measured in 2020, then our unit of analysis is states. If you were told a specific state, then you could find the row in the dataset. If we have 50 states measured in 2019 and 2020, then the unit of analysis is state-year because you will need to know the state and year to find a specific row.
- **Variable:** A measured characteristic of the unit of analysis. State unemployment rate is a variable for a state unit of analysis.
- **Datum:** The intersection of a variable (column) and a unit of analysis (row) resulting in a cell. The datum is a particular piece of information. A cell could contain something like 4.8 as the unemployment rate for Georgia in 2020.

2.3 Types of variables

The variables in a given dataset can be of several types. Types of variables are important to learn because the types of variables one is dealing with has consequences for data applications, such as description, visualization, and inference.

A variable provides us raw information about the units of analysis. If statistics is a discipline to convert raw information into something useful, then it stands to reason that we should want to know what type of information a variable provides us, especially the specificity of that information.

For example, suppose you ask two strangers to report their annual income. What options do they have for answers? If virtually any value, then you know to a precise degree the income each earns and can compute the precise difference between the two incomes. What if their choices are either more or less than $50,000? Then, you have a coarse understanding of how much they earn. If they provide different answers, you can only conclude whether one makes more than the other but not by how much. If they provide the same answer, then

the two are grouped together even though it is highly unlikely they earn equal incomes. This makes a serious difference for statistical analysis.

All variables belong to one of two broad types: qualitative (or categorical) and quantitative (or numeric).

- **Qualitative** variables take on values that have no intrensic numerical meaning. They are exressed in words.
- **Quantitative** variables take on values that do have intrensic numerical meaning.

2.3.1 Qualitative variables

Qualitative variables can be further differentiated into two types: nominal and ordinal.

- **Nominal** variables take on values that differ in name only.
- **Ordinal** variables take on values that can be ranked relative to each other but the difference between rankings has no numerical value.

The values that categorical variables take on are commonly referred to as levels. Categorical variables can contain virtually any number of levels, though the number of levels is usually limited.

A variable such as sex contains two levels: male and female. The variable sex is nominal, as its values have no numerical meaning and the two levels have no ranking. Race, state, country, political party, and any variable coded as yes/no such as unemployed, married, and below the federal poverty line are all examples of nominal variables.

If you have ever participated in a customer satisfaction survey, then you have almost surely contributed data to an ordinal variable. Those scales that provide some number of options from "disagree" to "agree" are called Likert scales. Your answer has no intrinsic numerical value but it can be ranked against the answers of others. One respondent can be said to be more satsfied than another but not by how much. Moreover, one can only trust the results insofar as responents have the same understanding or frame of reference–the service that satisfied one respondent may not have satisfied another. Other ordinal variables, such as education degree and income level do not have this issue.

2.3.2 Quantitative variables

Quantitative variables can be further differentiated into two types: discrete and continuous.

- **Discrete** variables take on countable or indivisible values.
- **Continuous** variables take on infinitely divisible values (at least in theory).

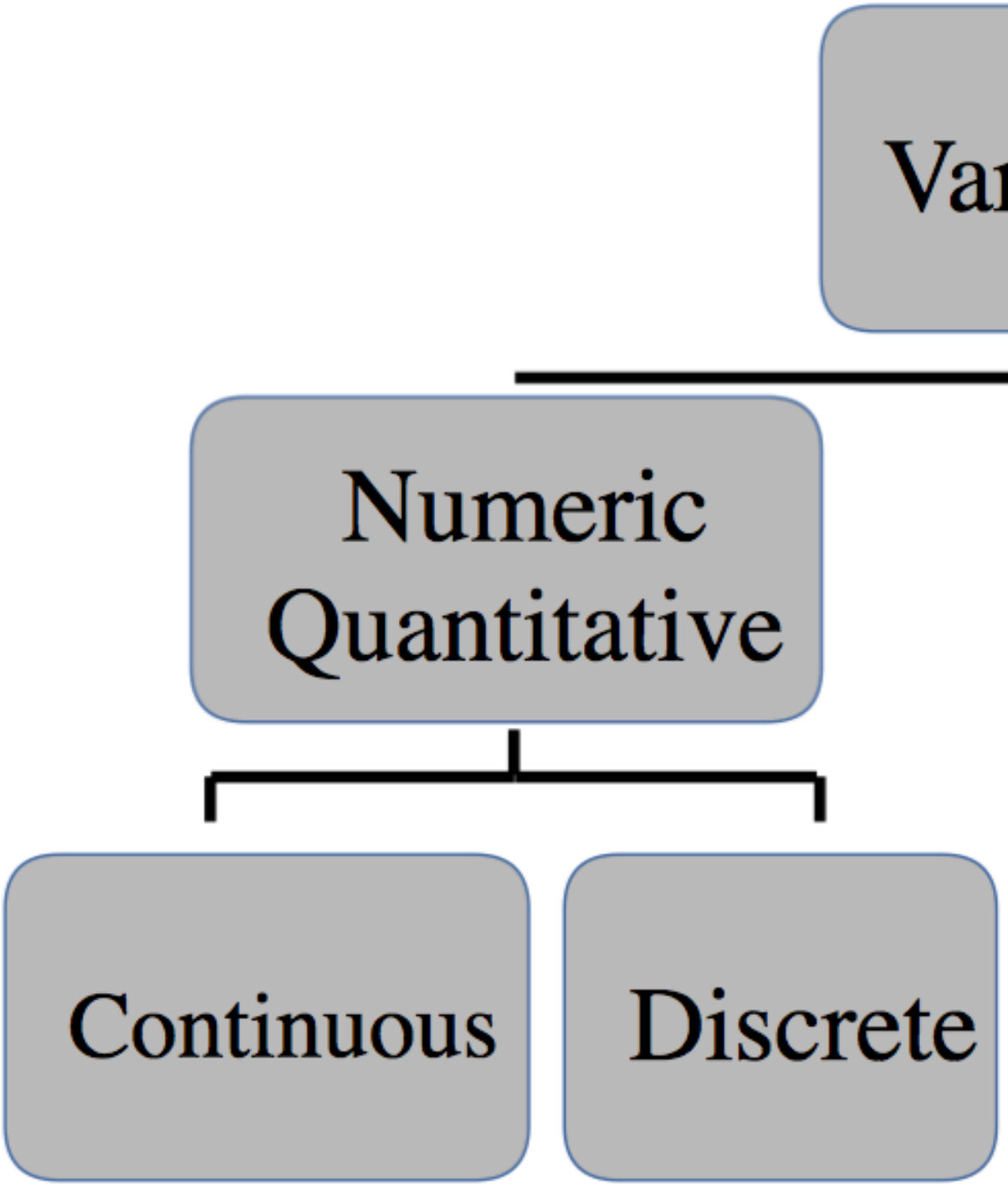

Figure 2.1: Variable Types

The distinction between discrete and continuous can be more difficult to discern but also less consequential for analysis. It is often the case that analytical models treat the two variables the same. However, for a purpose such as data visualization, the distinction can be informative.

Any variable that is a count of persons, places, events, or things is a discrete variable, usually taking on integer values (e.g. 0, 1, 2, 3,…). By contrast, a continuous variable can contain values with an infinite number of decimal places. Even so, continuous variables take on a limited number of decimal places because either we measure phenomena with finite precision or it simply becomes impractical to include so many decimal places.

2.3.3 Index variables

Index variables are continuous variables but warrant separate discussion. An index variable is a composite measure of multiple variables. They can be used to make a continuous variable out of multiple categorical variables or simplify multiple quantitative variables into one. Purposes such as ranking colleges, measuring poverty beyond income, and determining political ideology make use of index variables.

Index variables mask underlying information. This can be helpful or harmful. In either case, it is important to consider how an index variable is constructed. Doing so can offer insight or uncover problems.

An instructive example familiar to readers is college rankings. U.S. News and World Report describes how rankings are determined.

What makes a college good? According to these rankings, five percent of what makes a college good is the percent of undergraduate alumni giving a donation as a proxy of student satisfaction. Another 20% is based on the opinions of administrators at peer institutions.

Are these choices wise? This is difficult to say and besides the point. The point is that index variables involve choices made by people and are not naturally occurring data. They are synthetic materials of knowledge and worthy of our discernment.

2.4 Dataset structures

Just as the type of variable one is dealing with impacts the kinds of visualizations or analyses one should use, so too does the structure of a dataset. Datasets come in three varieties depending on their unit of analysis.

- Cross-sectional
  - Pooled cross-sectional
- Time series
- Panel or longitudinal

Cross-sectional data is a snapshot in time measuring some size sample of units. One column serves as the identifier of the unit of analysis, such as the name or ID number of the unit. Notice in Table 2.2 that all one needs to know is the country in order to identify a specific row.

Cross-section example

country

continent

year

lifeExp

pop

gdpPercap

Argentina

Americas

2007

75.320

40301927

12779.380

Bolivia

Americas

2007

65.554

9119152

3822.137

Brazil

Americas

2007

72.390

190010647

9065.801

Pooled cross-sectional data could be considered a fourth structure but is simply multiple cross-sections stacked atop each other. The critical quality of pooled cross-sectional data is that each cross-section contains *different* units measured at different times, not the same units measured at different times.

Notice in Table 2.3 that the countries included from 2002 are not the same as those included from 2007.

Pooled cross-section example

country

continent

year

lifeExp

pop

gdpPercap

Algeria

Africa

2002

70.994

31287142

5288.040

Angola

Africa

2002

41.003

10866106

2773.287

Benin

Africa

2002

54.406

7026113

1372.878

Botswana

Africa

2002

46.634

1630347

11003.605

Argentina

Americas

2007

75.320

40301927

12779.380

Bolivia

Americas

2007

65.554

9119152

3822.137

Brazil

Americas

2007

72.390

190010647

9065.801

Time series data measures one unit over mulitple time periods. The unit of analysis in time series data is time, as it uniquely identifies each row. Notice in Table 2.4 that one country is tracked over multiple years.

Time series example

country

continent

year

lifeExp

pop

gdpPercap

Argentina

Americas
1977
68.481
26983828
10079.027
Argentina
Americas
1982
69.942
29341374
8997.897
Argentina
Americas
1987
70.774
31620918
9139.671
Argentina
Americas
1992
71.868
33958947
9308.419
Argentina
Americas
1997
73.275
36203463
10967.282
Argentina
Americas

2002

74.340

38331121

8797.641

Argentina

Americas

2007

75.320

40301927

12779.380

Panel (or longitudinal) data measures the same units over multiple time periods. The unit of analysis is pair of unit and time period. Notice in Table 2.5 that in order to identify a specific row, you would need to know the country *and* year. One could also think of panel data as numerous time series.

Panel example

country

continent

year

lifeExp

pop

gdpPercap

Argentina

Americas

1997

73.275

36203463

10967.282

Argentina

Americas

2002

74.340

38331121

8797.641
Argentina
Americas
2007
75.320
40301927
12779.380
Bolivia
Americas
1997
62.050
7693188
3326.143
Bolivia
Americas
2002
63.883
8445134
3413.263
Bolivia
Americas
2007
65.554
9119152
3822.137

Chapter 3

Measurment and Missingness

> *"Will he not fancy that the shadows which he formerly saw are truer than the objects which are now shown to him?* —Plato, Republic

Once we know the kinds of data and datset structure we are dealing with, we need to understand how the data relates to reality before trying to drawing conclusions from it. Variables and the data they contain are measured representations of reality. They are shadows on the allegorical cave discussed in Plato's *Republic*. We should not immediately assume these shadows are true representations of reality.

3.1 Learning objectives

- Understand the components of credible analysis
- Discern the measurement validity of variables
- Discern the measurement reliability of variables
- Understand the difference between accuracy and precision

3.2 Credible Analysis

Measurement validity and reliability are the foundations of credible analysis, the components of which are depicted in Figure 3.1. Without the two, we have little or no basis to make conclusions from data.

This book will address the remaining building blocks in subsequent chapters.

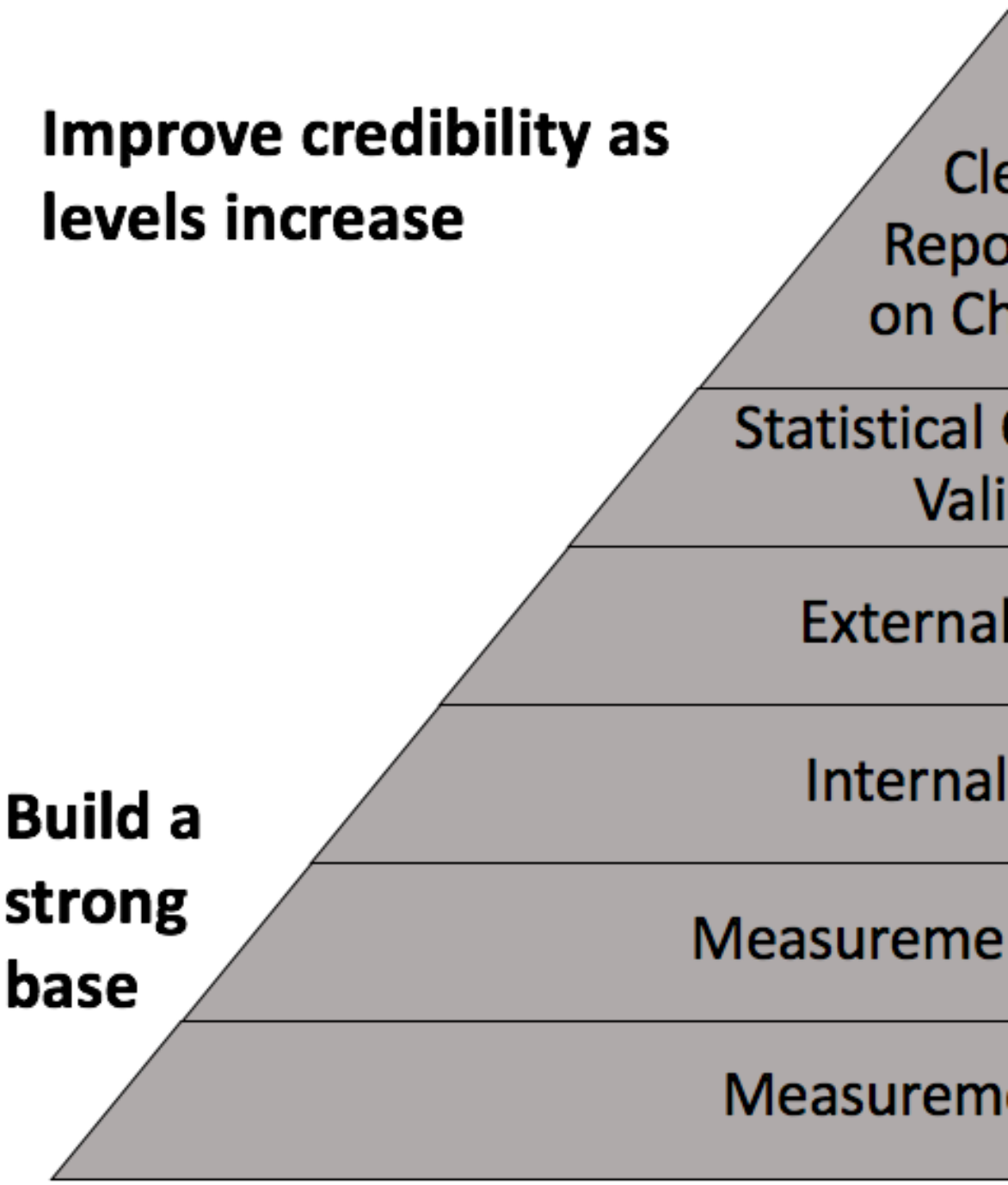

Figure 3.1: Components of credible analysis

3.2.1 Measurement Validity & Reliability

Measurement Validity: Does the variable accurately represent what it claims to represent? Are the values accurate representations of the intended concept/phenomenon?

Measurement Reliability: Does the way the variable is measured generate the same value given the same reality? Given two real and identical conditions, will my data contain identical values?

One way to visually represent the concepts of measurement validity and reliability is with the concentric circles of a target or a dart board. At the center of the target is the true concept of interest represented by a variable.

At the center of a yes/no variable such as poverty is perhaps economic stress or eligibility for means tested government welfare programs. The proximity of a variable's measurement to the center concept is the variable's measurement validity.

In addition to its proximith to the center of the target, there is also the issue of whether, given the same condition, repeated measures measures will result in the same value. If so, then the data points on the target should be clustered in close proximity.

If two cars are speeding in different towns at 80 mph, should we think a dataset recording instances of speeding would report these two cars differently? If not, then we believe the variable to have measurement reliability. If we think the two towns procedures or equipment result in different speeds for two cars traveling at the same speed, then we believe the variable to be unreliable.

The combination of validity and reliability presents four scenarios depicted in Figure 3.2 below.

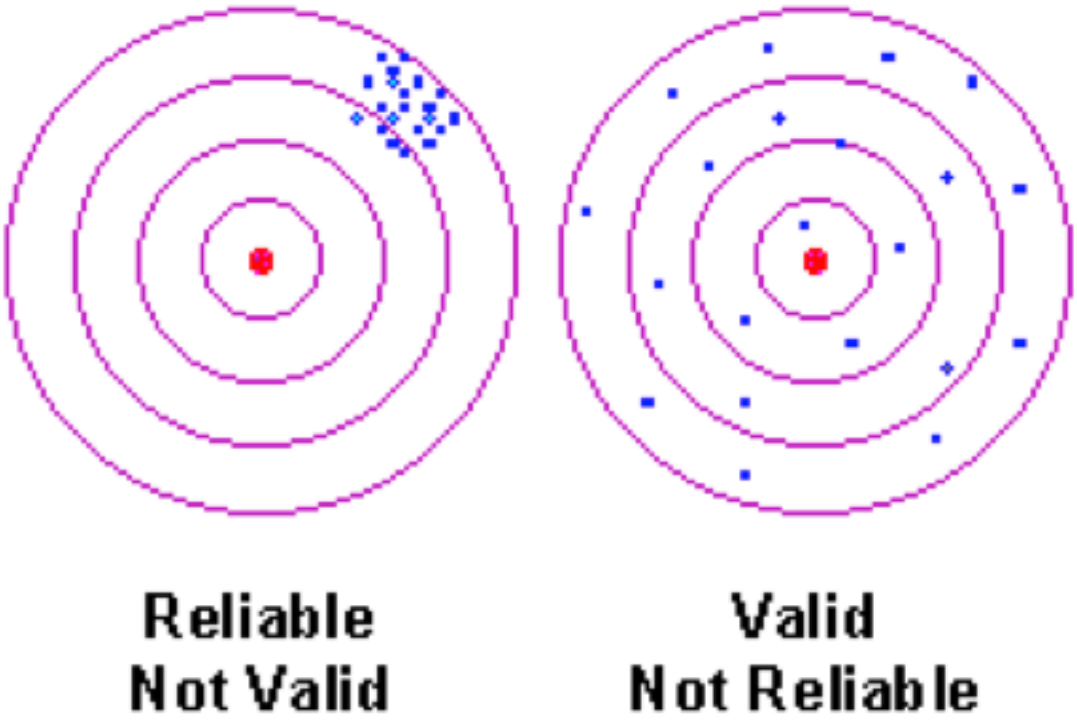

Figure 3.2: Representation of measurement validity and reliability

Chapter 4

Data Wrangling

Chapter 5

Data Explorations and Description

Chapter 6

Data Visualization