What Is a Dataset?

A **dataset** is just a way to organize information.

↪ It consists of rows and columns – much like a table

↪ Each row represents a **unit** (or an **observation**).

↪ Each column represents a **variable** (or a **feature**).

↪ Each specific point in a dataset is a **cell**.

We can mathematically represent a dataset as a **matrix**, and each variable or observation as a **vector**.

↪ And using math notation we might, for example, notate a dataset as $\mathbf{X}$ and a variable as $\mathbf{x}$. But this is not consistent.

What Is a Dataset?

The size of a dataset can be described by:

↪ the **sample size**, which is the number of observations, and

↪ the number of variables.

By tradition, people tend to use a capital N for the sample size.

It is fairly common to use K for the number of variables.

↪ So the size of a dataset is: $N \times K$. (Rows always first!)

Rule of thumb: a dataset should many more observations than variables. In math: $N \gg K$.

What Is a Dataset?

Because datasets are matrices, we can use "coordinates" to navigate them.

↪ We call these coordinates their **indices** (singular: **index**).

Using math notation, we write these coordinates using vectors:

↪ For example, (i, k) refers to the cell in row i and column k.

In programming languages, you will write indices a little different, but the order is always the same: rows first, columns second!

Fortunately, in R, you can (1) name columns so that you can find them better, and (2) "filter" rows to find exactly what you need.

Variables

In our context, a variable represents a specific concept.

For example, in Prof. Hubert's dataset of court cases, two important variables:

↪ Which judge presided over each case?

↪ How did each case get resolved?

The reason it is called a "variable" is because the value it takes varies across the observations. For example, in Prof. Hubert's dataset:

↪ One case was heard by Judge A, another by Judge B, etc.

↪ One case ended with settlement, another with dismissal, etc.

Variables

A very important distinction:

↪ Some variables are **outcome** (or **dependent**) variables.

↪ Some variables are **explanatory** (or **independent** or **predictor**) variables.

Example: Prof. Hubert's Dataset

"Political Appointments and Outcomes in Federal District Courts" (JoP, 2022, w/ Ryan Copus)

Research Questions:

↪ Do lawsuits end differently depending on assignment to a Democratic or Republican appointed judge?

Overall Findings:

↪ Republican appointees cause fewer settlements and more dismissals.

↪ Effect of judge's political affiliation has increased over time.

Variables

Remember:

↪ Some variables are **explanatory** (or **independent** or **predictor**) variables.

If an analyst makes this distinction, then they think one or more independent variables might (partially) "explain" or "cause" the dependent variable.

In Prof. Hubert's dataset:

Variables

Remember:

↪ Some variables are **explanatory** (or **independent** or **predictor**) variables.

If an analyst makes this distinction, then they think one or more independent variables might (partially) "explain" or "cause" the dependent variable.

In Prof. Hubert's dataset:

What is the independent variable?

Variables

Remember:

↪ Some variables are **explanatory** (or **independent** or **predictor**) variables.

If an analyst makes this distinction, then they think one or more independent variables might (partially) "explain" or "cause" the dependent variable.

In Prof. Hubert's dataset:

What is the independent variable?

↪ Presiding judge is an independent variable

What is the dependent variable?

Variables

Remember:

↪ Some variables are **explanatory** (or **independent** or **predictor**) variables.

If an analyst makes this distinction, then they think one or more independent variables might (partially) "explain" or "cause" the dependent variable.

In Prof. Hubert's dataset:

What is the independent variable?

↪ Presiding judge is an independent variable

What is the dependent variable?

↪ Resolution of each case is the dependent variable.

Variables

The number of independent variables is important.

The **rank** of a dataset is the number of "linearly independent" variables. (This comes from linear algebra.)

↪ For example: two columns expressing the same temperatures in Fahrenheit and Celsius are not linearly independent.

A dataset has **full rank** if all the independent variables are linearly independent

Why is this important? Because rank tells us something about how much "new" information we have in the variables.

Observations

A **unit** (or a **observation**) is a specific data point.

When someone says "my **unit of analysis** is..." they're telling you what makes up the rows of their dataset.

In Prof. Hubert's data:

What are the units of analysis?

Observations

A **unit** (or a **observation**) is a specific data point.

When someone says "my **unit of analysis** is..." they're telling you what makes up the rows of their dataset.

In Prof. Hubert's data:

What are the units of analysis?

↪ Court cases are the units of analysis

Observations

A **unit** (or a **observation**) is a specific data point.

When someone says "my **unit of analysis** is..." they're telling you what makes up the rows of their dataset.

In Prof. Hubert's data:

What are the units of analysis?

↪ Court cases are the units of analysis

↪ In fact: Prof. Hubert's data set can be transformed to judge-level units.

Observations

We have some special terminology for talking about datasets with different kinds of observations:

↪ In a **cross sectional** dataset, each row contains an observation for a different person/country/etc., all measured at one point in time.

Observations

We have some special terminology for talking about datasets with different kinds of observations:

↪ In a **cross sectional** dataset, each row contains an observation for a different person/country/etc., all measured at one point in time.

↪ In a **time series** dataset, every row contains an observation for a specific person/country/etc., measured at different points in time.

Observations

We have some special terminology for talking about datasets with different kinds of observations:

↪ In a **cross sectional** dataset, each row contains an observation for a different person/country/etc., all measured at one point in time.

↪ In a **time series** dataset, every row contains an observation for a specific person/country/etc., measured at different points in time.

↪ A **panel** (or **longitudinal**) dataset combines these: it contains rows that correspond to observations of different people/countries/etc. at different points of time.

Observations

We have some special terminology for talking about datasets with different kinds of observations:

↪ In a **cross sectional** dataset, each row contains an observation for a different person/country/etc., all measured at one point in time.

↪ In a **time series** dataset, every row contains an observation for a specific person/country/etc., measured at different points in time.

↪ A **panel** (or **longitudinal**) dataset combines these: it contains rows that correspond to observations of different people/countries/etc. at different points of time.

Note: Time series data is relatively rare in applied research

Observations

A cross sectional dataset

	1	2	3
	`country`	`year`	`gdp.usd.tr`
1	China	?	2.29
2	United States	?	13.04
3	Germany	?	2.85

Are the ?'s all the same or different?

Observations

A cross sectional dataset

	1	2	3
	`country`	`year`	`gdp.usd.tr`
1	China	2005	2.29
2	United States	2005	13.04
3	Germany	2005	2.85

They are all the same.

$\hookrightarrow$ We assess a cross-sectional dataset at a single point in time.

Observations

A time series dataset

	1	2	3
	`country`	`year`	`gdp.usd.tr`
1	China	?	2.29
2	China	?	6.09
3	China	?	11.06

Are the ?'s all the same or different?

Observations

A time series dataset

	1	2	3
	country	year	gdp.usd.tr
1	China	2005	2.29
2	China	2010	6.09
3	China	2015	11.06

They are all different.

↪ We assess a time series dataset at multiple points in time.

Observations

A panel dataset

	1	2	3
	`country`	`year`	`gdp.usd.tr`
1	China	?	2.29
2	China	2010	6.09
3	China	2015	11.06
4	United States	2005	13.04
5	United States	?	15.05
6	United States	2015	18.21
7	Germany	2005	2.85
8	Germany	2010	3.40
9	Germany	?	3.47

Are the ?'s (likely) all the same or different?

Observations

A panel dataset

	1	2	3
	country	year	gdp.usd.tr
1	China	2005	2.29
2	China	2010	6.09
3	China	2015	11.06
4	United States	2005	13.04
5	United States	2010	15.05
6	United States	2015	18.21
7	Germany	2005	2.85
8	Germany	2010	3.40
9	Germany	2015	3.47

They are all different.

↪ We assess a panel dataset at multiple points in time and across multiple units.

Measurement

Measurement is a huge issue, but people spend way too little time talking about it (including us...).

Core issue: we are trying to represent concepts in a quantitative format so that we can use quantitative (statistical) analysis.

In other words, we are trying to **operationalize** concepts.

This is both art and science.

↪ Even when it's clear that a concept is easily quantified, we may not know how to do it in practice.

↪ For example, it's straight-forward that duration of war would be measured in days or months, but actually measuring it is hard.

Measurement Scales

Variables are quantitatively represented on different kinds of **measurement scales**.

A variable is on a **cardinal** scale if it can be represented by a number and the magnitude of those numbers are meaningful.

↪ For example: a person's age, the duration of a war, etc.

A variable is on an **ordinal** scale if it can be represented by a number but those numbers are only useful for ranking.

↪ For example: 7-point partisanship or democracy scales, etc.

↪ Warning: you shouldn't add/subtract/multiply/divide these numbers!

Measurement Scales

A variable is a **categorical** variable if it takes one of several pre-defined values (the "categories").

↪ For example: a person's country of residence, party registration, etc.

Even when they're on the same kind of scale, the variables can differ in how many values they can take:

↪ If a variable takes only two values, then the variable is a **binomial** variable (or **dichotomous** or **binary**).

↪ If more than two, then the variable is **multinomial**.

Measurement Scales

Notice: categorical variables are not numbers!

They are often transformed into **dummy (or binary) variables**.

To do this for a categorical variable:

1. Make each category its own new binary variable.
2. For each new variable, code each unit to take a 1 if that unit is in that category and 0 otherwise.

Even though they are already numbers, you may often want to do this with ordinal variables too!

$\hookrightarrow$ You'll develop a better sense for why/when as you go.

Measurement Scales

	1	2	3
	country	year	gdp.usd.tr
1	China	2005	2.29
2	China	2010	6.09
3	China	2015	11.06
4	United States	2005	13.04
5	United States	2010	15.05
6	United States	2015	18.21
7	Germany	2005	2.85
8	Germany	2010	3.40
9	Germany	2015	3.47

What kind of variable is GDP here?

Measurement Scales

	1	2	3
	country	year	gdp.usd.tr
1	China	2005	2.29
2	China	2010	6.09
3	China	2015	11.06
4	United States	2005	13.04
5	United States	2010	15.05
6	United States	2015	18.21
7	Germany	2005	2.85
8	Germany	2010	3.40
9	Germany	2015	3.47

What kind of variable is GDP here?

↪ Cardinal – can be represented by a number and the magnitude of those numbers are meaningful.

Measurement Scales

	1	2	3	4	5	7
	country	year	gdp.usd.tr	china	usa	germany
1	China	2005	2.29	1	0	0
2	China	2010	6.09	1	0	0
3	China	2015	11.06	1	0	0
4	United States	2005	13.04	0	1	0
5	United States	2010	15.05	0	1	0
6	United States	2015	18.21	0	1	0
7	Germany	2005	2.85	0	0	1
8	Germany	2010	3.40	0	0	1
9	Germany	2015	3.47	0	0	1

What kind of variable is `china` here?

Measurement Scales

	1	2	3	4	5	7
	country	year	gdp.usd.tr	china	usa	germany
1	China	2005	2.29	1	0	0
2	China	2010	6.09	1	0	0
3	China	2015	11.06	1	0	0
4	United States	2005	13.04	0	1	0
5	United States	2010	15.05	0	1	0
6	United States	2015	18.21	0	1	0
7	Germany	2005	2.85	0	0	1
8	Germany	2010	3.40	0	0	1
9	Germany	2015	3.47	0	0	1

What kind of variable is `china` here?

↪ Binary / dummy – it can only take two values, 0 and 1

Other Measurement Issues

Gailmard (2014) discusses other issues.

Many variables are **indices** of concepts: e.g., "social capital," "democracy," "ideology," etc.

↪ This is a huge topic. Talk to Prof. Hare about this!

The **validity** of a measure tells us how "adequately [it] measures the concept it is intended to measure."

↪ Three important distinctions: **face validity**, **content validity** and **construct validity**.

Recap – data sets

Quantitative studies use **data** to answer questions about the world.

↪ A **dataset** is a way to organize information.

↪ A dataset consists of rows and columns – much like a table

↪ Each row represents a **unit** (or an **observation**).

↪ Each column represents a **variable** (or a **feature**).

↪ Each specific point in a dataset is a **cell**.

Recap – measurement

You will encounter different types of variables:

↪ A variable is on a **cardinal** scale if it can be represented by a number and the magnitude of those numbers are meaningful.

↪ A variable is on an **ordinal** scale if it can be represented by a number but those numbers are only useful for ranking.

↪ A variable is a **categorical** variable if it takes one of several pre-defined values (the "categories").

↪ If a variable takes only two values, then the variable is a **binomial** variable (or **dichotomous** or **binary**).

Key task of the researcher: **operationalize** concepts.

This means we need to figure out how to measure concepts in a quantitative format so that we can use quantitative (statistical) analysis.

Types of Studies

Quantitative data doesn't just magically exist.
It has to be *collected* (or *observed*).

The main ways that social science datasets come to be are:

↪ **Experiments**: researcher manipulates some aspect(s) of the environment and records how units behave in response.

↪ **Surveys** in which a researcher records the beliefs, behaviors, characteristics, etc. of units (after informing them).

↪ **Observational studies** in which a researcher records attributes or behaviors of units in their natural context.

Main distinction: researcher control over the environment!

Types of Studies – Causal questions

Many studies are focused on **causal** questions:

↪ Do democracies go to war with one another?

↪ Do judges' identities affect how cases get resolved?

↪ Do protests increase the likelihood of regime survival?

Notice that none of these have the word "cause" in them.

↪ **Semantic choices don't change whether a study is causal!**

Types of Studies – Example

H. Hilbig and S. Riaz, JoP 2022:

> *How do freedom of movement restrictions affect refugee integration? [..] We study a contentious law in Germany, which barred refugees from moving to a location different from the one they were exogenously assigned to. To identify the causal effect of the movement restriction on integration, we use a sharp date cutoff that governs whether refugees are affected by the policy. We demonstrate that restricting freedom of movement had pronounced negative effects on refugees' sense of belonging in Germany while increasing identification with their home countries. [...]*

↪ Causal claim is that freedom of movement restrictions *caused* refugees to feel less integrated.

Types of Studies – Description

But this is not the only goal of an empirical study.

Some studies are focused on **description**.

↪ What is the geographical distribution of voters?

↪ Which countries are presidential systems? Which are parliamentary systems?

Back to a theme: sometimes people say their causal study is a "descriptive" study. **Be skeptical.**

↪ "I'm not saying that being democratic *causes* states not to go to war. I am just observing there is a relationship." 🙄

Types of Studies – Example

A. Kaufman and J. Rogowski, forthcoming AJPS:

> *Presidents select from a range of instruments when creating new policies through executive action. We [...] argue that presidents use less visible means of unilateral instruments when Congress is likely to scrutinize presidential action. [...] [W]e show that presidents are more likely to substitute memoranda and other less visible instruments for executive orders and proclamations during periods of divided government. Second, [...], we find that presidents issue greater numbers of directives during divided government than during unified government. [...]*

↪ Is this a causal or a descriptive study?

↪ What is the implied theoretical argument here?

Types of Studies – Measurement

Other studies are focused on **measurement**.

↪ How do we measure democracy?

↪ How do we measure the ideology of local politicians?

↪ How can we measure whether a legal case is an "easy" or "hard" case?

These days, measurement studies are often highly technical and use advanced computational techniques.

Regardless of what kind of study, the most influential contributions usually involve original data. **This should be one of your goals.**

Causal Studies

There is some special terminology we use for data that is used in the context of a causal study.

For historical reasons, the terminology comes from medicine.

↪ The **treatment** is the main causal variable.

↪ The **response** or **outcome** is the variable that the analyst thinks might be affected by the treatment.

↪ The **treatment group** and the **control group** are the units that were and weren't exposed to the treatment, respectively.

↪ The **controls** or **covariates** are the other variables that the analyst includes in their analysis. (More on this later.)

Causal Studies

For example, in the dataset analyzed in Prof. Hubert's recent JOP paper:

↪ Recall the unit of observation is legal cases.

↪ The treatment: whether a Democratic-appointed or Republican-appointed judge was assigned to a case.

↪ The outcome: the outcome of the legal case, e.g., did the case settle? was it dismissed? etc.

The positive theory being examined: presidential partisanship influences how federal civil rights cases are resolved since presidents appoint federal judges.

(There are some valid questions about the sample...)

Keeping Data in Context

Ideally, a dataset helps us learn about some important concepts.

However, a dataset is (usually) just a small snap-shot of the larger world we live in.

This raises all sorts of important and thorny questions, which will be the focus of much of this course (and other methods courses).

But, the basic idea is that we treat a dataset as a **sample** of the larger **population** we care about. Then, we ask:

$\hookrightarrow$ How representative is our sample of the larger population?

$\hookrightarrow$ What can we learn about the population from our sample?

Empirical Distributions

An **(empirical) distribution** is a list of observations of one (or more) variables for each unit in a dataset.

↪ A distribution containing observations from one variable is known as a **univariate distribution**.

↪ A distribution containing observations from more than one variable is known as a **multivariate distribution**.

↪ We have a special phrase for a multivariate distribution with only two variables: **bivariate distribution**. (Why?)

↪ A bit of notation: I'll use x to reference a generic variable, x_i will reference a specific observation of the variable x.

(Later in this course, you will see **probability distributions**.)

Empirical Distributions

case.id	def_count	def_acount	pla_count	pla_acount
cacd-199806-00126	4	0	1	1
cacd-201413-00055	3	2	1	2
caed-200216-00163	1	1	2	4
caed-200718-00097	2	5	1	2
caed-201136-00151	2	2	1	1
cand-200508-00020	3	0	1	1
cand-201146-00151	1	0	1	1
casd-200125-00204	1	1	2	2
casd-200451-00139	2	1	1	1
wawd-201622-00039	4	8	1	2

Empirical Distributions

a univariate distribution

case.id	def_count	def_acount	pla_count	pla_acount
cacd-199806-00126	4	0	1	1
cacd-201413-00055	3	2	1	2
caed-200216-00163	1	1	2	4
caed-200718-00097	2	5	1	2
caed-201136-00151	2	2	1	1
cand-200508-00020	3	0	1	1
cand-201146-00151	1	0	1	1
casd-200125-00204	1	1	2	2
casd-200451-00139	2	1	1	1
wawd-201622-00039	4	8	1	2

Empirical Distributions

a different univariate distribution

case.id	def_count	def_acount	pla_count	pla_acount
cacd-199806-00126	4	0	1	1
cacd-201413-00055	3	2	1	2
caed-200216-00163	1	1	2	4
caed-200718-00097	2	5	1	2
caed-201136-00151	2	2	1	1
cand-200508-00020	3	0	1	1
cand-201146-00151	1	0	1	1
casd-200125-00204	1	1	2	2
casd-200451-00139	2	1	1	1
wawd-201622-00039	4	8	1	2

Empirical Distributions

a bivariate distribution

case.id	def_count	def_acount	pla_count	pla_acount
cacd-199806-00126	4	0	1	1
cacd-201413-00055	3	2	1	2
caed-200216-00163	1	1	2	4
caed-200718-00097	2	5	1	2
caed-201136-00151	2	2	1	1
cand-200508-00020	3	0	1	1
cand-201146-00151	1	0	1	1
casd-200125-00204	1	1	2	2
casd-200451-00139	2	1	1	1
wawd-201622-00039	4	8	1	2

Empirical Distributions

Recall: the business of social science is to take complex realities and concepts and find simpler ways to understand them.

Distributions are complex! We need ways to *summarize* them.

↪ Most people will not want to see every single observation or row in a dataset – they want summaries

There are two major ways to summarize distributions:

↪ What is the **(sample) central tendency** of a distribution? In other words, what is its "typical value"?

↪ What is the **(sample) dispersion** of a distribution? In other words, "how typical is its typical value" or "how all over the place are its values"?

Empirical Distributions

What is a "typical value" of a univariate distribution?

↪ Suppose someone extracts one value from a distribution.
You know the full distribution but not the extracted value.

↪ Then, they ask you to make your best guess about which value they extracted from the distribution. (We'll label your guess g.)

Wouldn't your best guess be the "typical value" of the distribution?

Of course, you may be more or less confident in your best guess depending on how "all over the place" the distribution's values are.

There are multiple ways to make a best guess. And so: there will be multiple ways to measure central tendency and dispersion.

Some Mathematical Basics - Summation Notation

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + \cdots + x_N$$

$\hookrightarrow$ We start with $i = 1$ and then just add all elements up to N

Example:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

This doesn't just work with x_i – it works with anything:

$$\sum_{i=1}^{5} i = ?$$

Some Mathematical Basics - Summation Notation

This doesn't just work with x_i – it works with anything:

$$\sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5 = 15$$

Another example:

$$\sum_{i=1}^{5} 2^i = 2^1 + 2^2 + 2^3 + 2^4 + 2^5 = 62$$

No reason to be confused by the $\sum$ symbol – it's just a fancy "S" for "sum."

Sample Mean

The **sample mean** of a univariate distribution, which we often label $\overline{x}$, is calculated as follows:

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Note a few properties:

1. The sample mean is always (weakly) between the largest and smallest values of x.
2. For a variable y such that $y_i = a \times x_i$ for all i, then $\overline{y} = a \times \overline{x}$.
3. For a variable y such that $y_i = a + x_i$ for all i, then $\overline{y} = a + \overline{x}$.
4. There is only one value for $\overline{x}$ for the distribution of variable x.

Sample Mean

There's a very nice property of sample means on binary variables.

Assuming the variable has been transformed into 0s and 1s, then: *the sample mean yields the percentage of units with 1 values.*

For example, consider this hypothetical variable from a dataset:

$$x = (0, 0, 1, 0, 1, 1, 1, 0, 0, 0)$$

Then:

$$\overline{x} = \frac{4}{10} = 0.4 = 40\%$$

Sample Mean

In what sense is the sample mean a typical value of a distribution?

Recall the thought exercise: what's your best guess?

One possibility: your guess g is your best guess if it has the lowest **mean squared error**:

$$MSE(g) = \frac{1}{N} \sum_i (\underbrace{x_i - g}_{\text{error}})^2$$

Roughly: the average difference between your guess and each value in the distribution (i.e., average error) is as small as it can be.

The sample mean minimizes MSE: $\arg\min_g MSE(g) = \overline{x}$.

Sample Mean

Sometimes it does not make sense to calculate a sample mean.

Suppose that we want to know the percent of cases resolved in favor of the plaintiff using the following dataset.

judge	caseload	pro_pla
Judge1	10	0.50
Judge2	90	0.10

The sample mean of `pro_pla` is:

$$\overline{\texttt{pro\_pla}} = \frac{0.5 + 0.1}{2} = 0.3 = 30\%$$

Sample Mean

Sometimes it does not make sense to calculate a sample mean.

Suppose that we want to know the percent of cases resolved in favor of the plaintiff using the following dataset.

judge	caseload	pro_pla
Judge1	10	0.50
Judge2	90	0.10

But, notice the number of cases resolved for the plaintiff is:

$$\underbrace{10 \times 0.5}_{\text{Judge 1}} + \underbrace{90 \times 0.1}_{\text{Judge 2}} = 5 + 9 = 14 \implies 14\%$$

Sample Mean

Sometimes it does not make sense to calculate a sample mean.

Suppose that we want to know the percent of cases resolved in favor of the plaintiff using the following dataset.

judge	caseload	pro_pla
Judge1	10	0.50
Judge2	90	0.10

Instead, we can calculate the **weighted sample mean**:

$$\overline{x}^W = \sum_i w_i \times x_i = \frac{10}{100} \times 0.5 + \frac{90}{100} \times 0.10 = 14\%$$

(The "simple" sample mean is a special case with $1/N$ weights.)

Sample Mean

An important, early lesson: you have to *think* before you start calculating things.

First, ask yourself "what exactly do I want to learn?" Then make sure that what you calculate helps you learn that!

This is hard! Consider a variation of the example in the text:

↪ Suppose we see average GRE scores decline one year.

↪ This **does not** imply GRE scores have fallen among any demographic subgroup! Why?

↪ This is an example of both **Simpson's paradox** and the **ecological fallacy**.

Summary of the Sample Mean

Sample means can be used to summarize variables in data sets:

↪ The sample mean is a "typical value" of a distribution.

↪ The sample mean minimizes the mean squared error of a guess.

↪ When some observations are more "important" than others, we can use weighted sample means.

↪ Next, we will discuss other ways to summarize the central tendency of a distribution.

Sample Median

The sample mean is not the only way to summarize the central tendency of a distribution.

Loosely speaking: the **sample median** is the "middle observation" of a distribution (after ordering the values).

↪ This is straight-forward if the number of observations is odd.

↪ If it is even, then every value *weakly between* the middle two values is a median for the distribution.

↪ The sample median is not necessarily unique!

↪ Check out Gailmard (2014) for the somewhat ugly math.

Sample Median

Example: consider the following distribution of test scores: $(2, 4, 5, 7, 8)$.

Then: $x_M = 5$.

Similar to the sample mean, the sample median is also the "best" guess in some sense.

In fact, the sample median minimizes the **mean absolute error** of a guess g:

$$MAE(g) = \frac{1}{N} \sum_i |\underbrace{x_i - g}_{\text{error}}|$$

Sample Median

The sample median is an **order statistic**, which means it "is defined in terms of where it falls in an ordered or ranked list of the variable's values."

Practically speaking, this means that the sample median is *insensitive to extreme observations*, often called **outliers**.

For example, consider this distribution: $(2, 2, 2, 4, 4, 4, 4)$.

Then: $\overline{x} \approx 3.14$ and $x_M = 4$.

But what if we increased the last value to 1000?

Then: $\overline{x} \approx 145.4$ and $x_M = 4$.

Sample Mode

A somewhat less common way to summarize the central tendency of a distribution is the **sample mode** of a distribution.

$\hookrightarrow$ We often refer to the sample mode as the **modal value**.

The sample mode is simply the value of the distribution that occurs most often in the distribution.

The sample mode is the "best" guess in the following sense: picking the mode maximizes the probability of the guess being the *correct* guess.

Mean, Mode and Median

Consider the following distribution: $(1, 2, 2, 3, 4, 5, 5, 5, 5)$.

What are the median the mode?

Mean, Mode and Median

Consider the following distribution: $(1, 2, 2, 3, 4, 5, 5, 5, 5)$.

What are the median the mode?

$\hookrightarrow$ The median is 4 – half of the observations are less than or equal to 4 and half are greater than or equal to 4.

$\hookrightarrow$ The mode is 5 - it occurs more often than any other value.

What is the mean?

Mean, Mode and Median

Consider the following distribution: $(1, 2, 2, 3, 4, 5, 5, 5, 5)$.

What are the median the mode?

$\hookrightarrow$ The median is 4 – half of the observations are less than or equal to 4 and half are greater than or equal to 4.

$\hookrightarrow$ The mode is 5 - it occurs more often than any other value.

What is the mean?

$\hookrightarrow$ The mean is 3.56

Sample Variance & Standard Deviation

Let's get more specific about how good your best guess is.

In other words, we want to know the "typical" *deviation* of an observation x_i from the "typical" *value* (your best guess).

Let's begin by assuming your best guess is the sample mean, $\overline{x}$.

The **sample variance** is one way to see how good your best guess is:

$$s_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

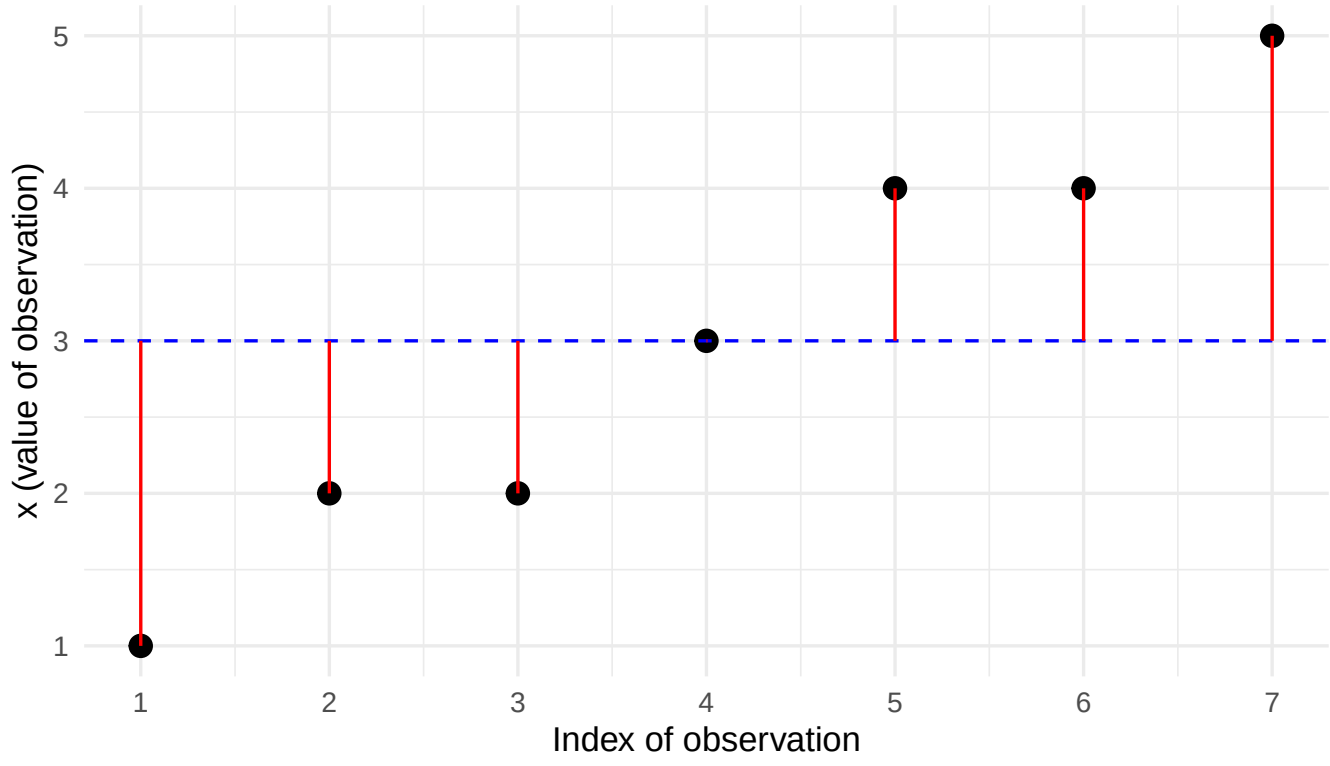
This is very close to the MSE of the sample mean, except we divide by $N-1$ instead of N.

Sample Variance & Standard Deviation

↪ Assume we observe the following realizations of X:
$(1, 2, 2, 3, 4, 4, 5)$

↪ Sample mean: $\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1+2+2+3+4+4+5}{7} = \frac{21}{7} = 3$

First, calculate the squared "deviations" from the sample mean:



Sample Variance & Standard Deviation

↪ Assume we observe the following realizations of X: $(1, 2, 2, 3, 4, 4, 5)$

↪ Sample mean: $\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1+2+2+3+4+4+5}{7} = \frac{21}{7} = 3$

What is the sample variance? We have to first calculate the squared "deviations" from the sample mean:

↪ $(1-3)^2 = 4$

↪ $(2-3)^2 = 1$

↪ ...

↪ In sum: $(1-3)^2 + (2-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (4-3)^2 + (5-3)^2 = 12$

↪ Divide by $N-1$: $\frac{12}{7-1} = 2$ (Recall: variance is defined as the "average squared deviation".)

Sample Variance & Standard Deviation

Our goal is to get a good sense how much the individual observations deviate from our sample mean.

But, notice we're squaring the deviations in the formula: variance is measured in different units (squared units) than sample mean!

↪ For example: the variance of a distribution of test scores measured in points will be measured in squared points. Arg!

So, to get around this, we undo the squaring and instead calculate the **standard deviation**:

$$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Sample Variance & Standard Deviation

Why use $N - 1$ in the denominator rather than N?

We'll defer the technical explanation to later in the course.

But practically speaking: this has the effect of giving us a more **conservative** estimate of sample variance/standard deviation.

↪ Sample variance tells us how confident (or "uncertain") we should be about treating sample mean as a typical value.

↪ A (slightly) larger estimate of sample variance raises the bar for us to be confident.

↪ Remember: we should be cautious about when we say we're confident in an empirical finding.

Sample Variance & Standard Deviation

Note a few properties of standard deviations s_x (and also variance):

1. The sample standard deviation is always weakly positive.
2. If a variable x only ever takes one value, then $s_x = 0$.

   $\hookrightarrow$ This is an important fact – why is that the case?
3. For a variable y such that $y_i = a \times x_i$ for all i, then $s_y > s_x$.
4. For a variable y such that $y_i = a + x_i$ for all i, then $s_y = s_x$.
5. There is only one value for s_x for the distribution of variable x.

Percentiles & Sample IQR

There is an important order statistic we didn't define.

The ***n*th percentile** is the value in a distribution where n percent of the values are equal to or below that value.

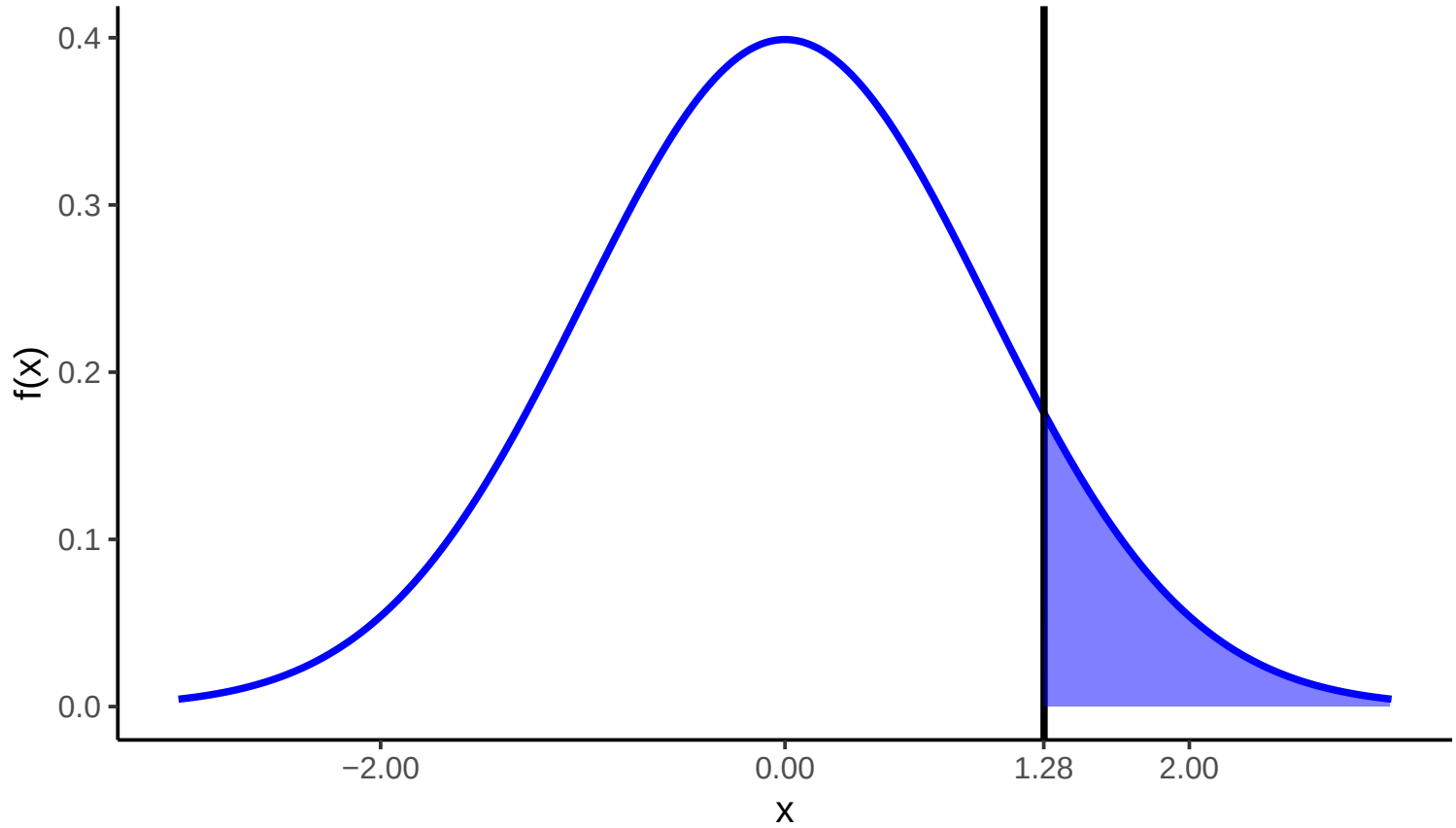
We have some special terminology for specific percentiles:

↪ The **first/second/third quartile** is the 25th/50th/75th percentile.

↪ The **fiftieth percentile** (or **second quartile**) is the median.

↪ The **first/second/third/fourth quintile** is the 20th/40th/60th/80th percentile.

Percentiles & Sample IQR



We say that ≈1.28 is the *90th percentile* of this distribution.

↪ This means: 10% of all values are larger than 1.28, and 90% of all values are equal or below.

Percentiles & Sample IQR

Consider this example distribution: $(0, 1, 9, 13, 42, 65, 78, 98)$

The first and third quartiles are 1 and 65.

Note: there are slightly different conventions for calculating percentiles, but they make very little difference in large datasets.

However: this can mean that the definition of the 50th percentile and the definition of the median are somewhat different.

For example, any number in the set $[13, 42]$ is a median, but the 50th percentile is 13.

Percentiles & Sample IQR

How do we measure dispersion relative to the sample median?

If P^n is the nth percentile, the **interquartile range** (**IQR**) is:

$$IQR = P^{75} - P^{25}$$

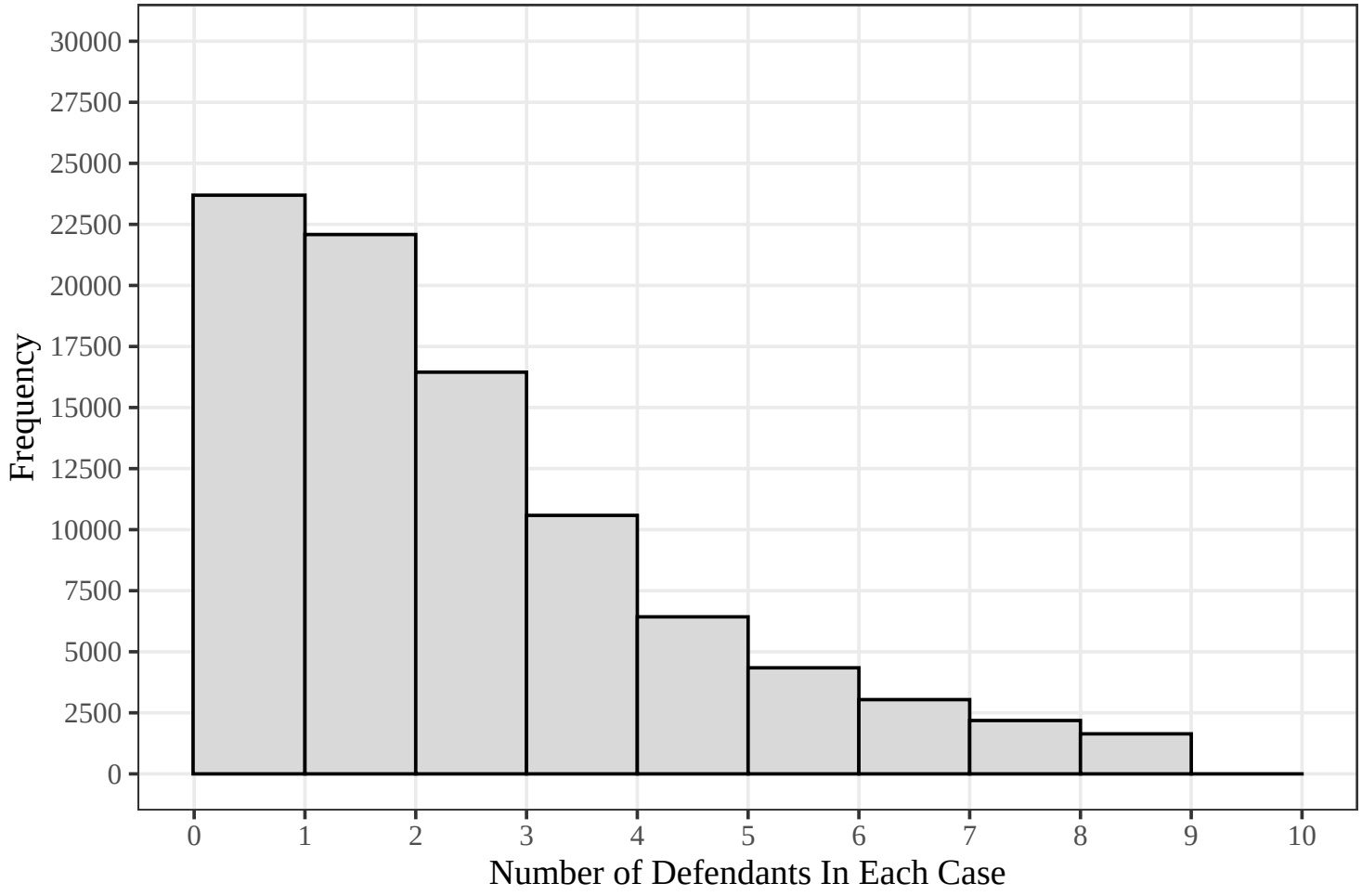
This measures dispersion relative to the median.

And like the median, it is not sensitive to outliers since it is based on order statistics (percentiles).

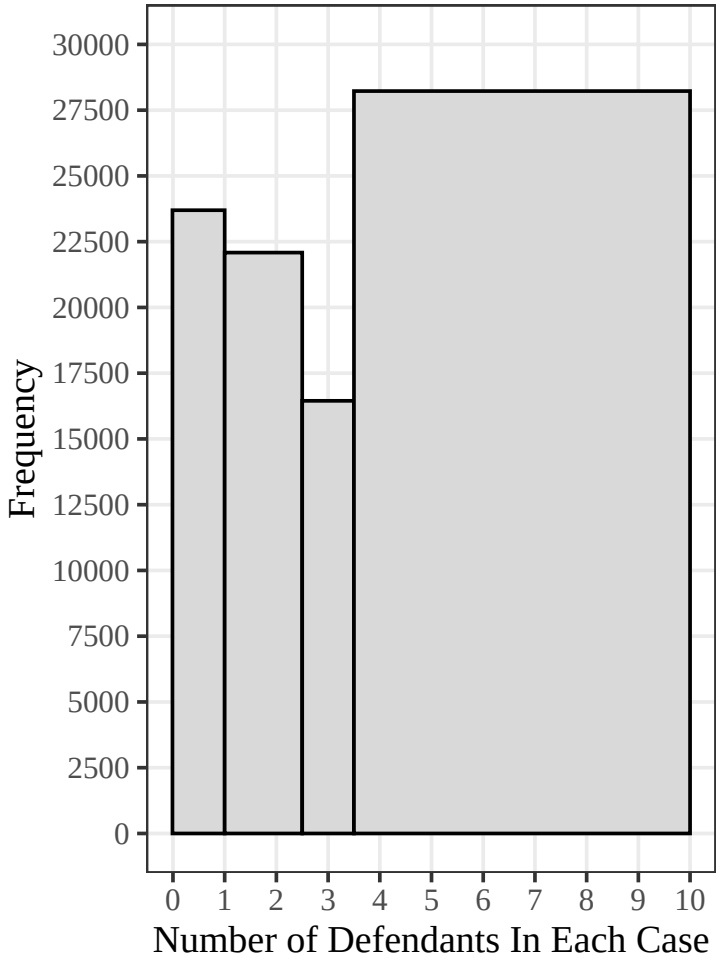
In fact, an arbitrary (but widely used) definition of an outlier:

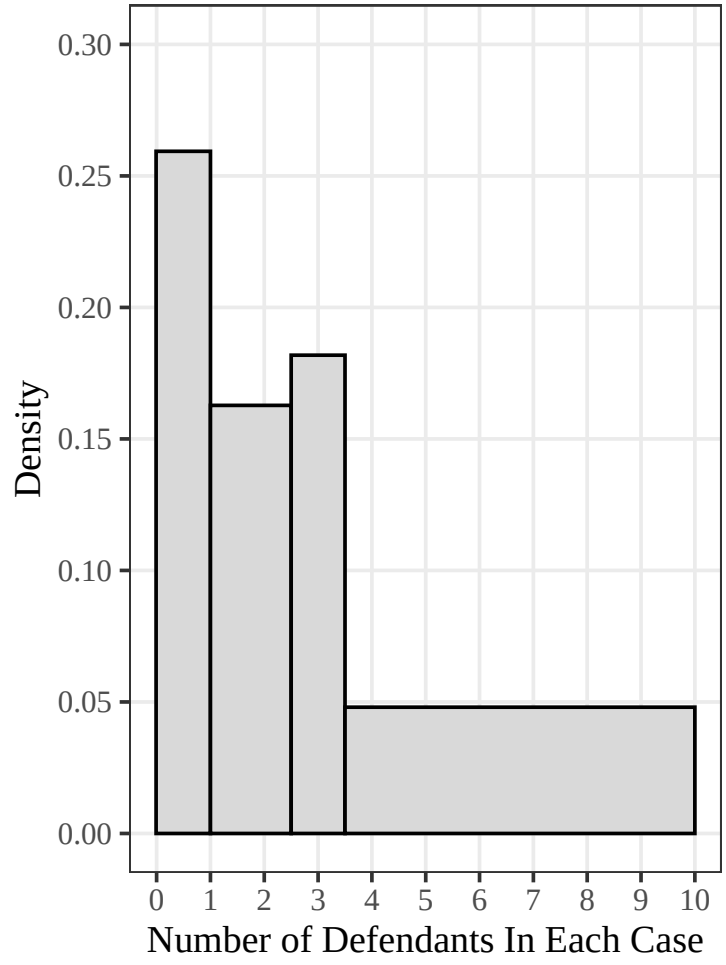
$\hookrightarrow$ An outlier is any value in a distribution that is at least $1.5 \times IQR$ above the third quartile or below the first quartile.

Visualizing Distributions: Histograms



Visualizing Distributions: Histograms





Visualizing Distributions: Histograms

Frequency:

↪ Shows the count of observations falling into each bin.

↪ Y-axis represents the number of observations.

↪ The sum of bar heights is the total number of observations.

Density:

↪ Shows the proportion of data per unit of x in that bin

↪ Y-axis represents the proportion of total observations

↪ The sum of the bin areas is 1

↪ Useful when bins have unequal width