Lesson 3:
Chapter 3

Caleb Moxley

BSC Mathematics

9 September 15

§3.1 Sources of Data

So far, we've discussed exploratory analysis (or descriptive analysis) of a data set.

§3.1 Sources of Data

So far, we've discussed exploratory analysis (or descriptive analysis) of a data set. These samples, though, may not be great indicators of the characteristics of the populations from which they're drawn.

§3.1 Sources of Data

So far, we've discussed exploratory analysis (or descriptive analysis) of a data set. These samples, though, may not be great indicators of the characteristics of the populations from which they're drawn. In this chapter, we'll discuss proper methods for producing data sets which will give data which more reliably mimics the populations from which it is drawn.

§3.1 Sources of Data

Exactly how are the following types of data problematic?

§3.1 Sources of Data

Exactly how are the following types of data problematic?

Definition (anecdotal data)

Anecdotal data represent individual cases.

§3.1 Sources of Data

Exactly how are the following types of data problematic?

Definition (anecdotal data)

Anecdotal data represent individual cases.

Definition (self-reported data)

Self-reported data is data which is measured and reported by the case subject.

§3.1 Sources of Data

Definition (available data)

Available data is data collected for one purpose which may be useful in answering a question it wasn't intended to answer.

§3.1 Sources of Data

Definition (available data)

Available data is data collected for one purpose which may be useful in answering a question it wasn't intended to answer.

Example (available data)

Give an example of a question which may be answered using the data collected for the purpose below.

- Amazon would like to know what other websites its customers visit, so it installs a tracking cookie on its customers' browsers.

§3.1 Sources of Data

Definition (available data)

Available data is data collected for one purpose which may be useful in answering a question it wasn't intended to answer.

Example (available data)

Give an example of a question which may be answered using the data collected for the purpose below.

- Amazon would like to know what other websites its customers visit, so it installs a tracking cookie on its customers' browsers.
- A university would like to know the average age of its students, so it collects the ages of all its students.

§3.1 Sources of Data

Definition (available data)

Available data is data collected for one purpose which may be useful in answering a question it wasn't intended to answer.

Example (available data)

Give an example of a question which may be answered using the data collected for the purpose below.

- Amazon would like to know what other websites its customers visit, so it installs a tracking cookie on its customers' browsers.
- A university would like to know the average age of its students, so it collects the ages of all its students.
- A farmer would like to know the level of phosphorus in her soil, so she collects soil samples from randomly selected plots on her farm.

§3.1 Sources of Data

Definition (sample & population)

A **population** is the entire collection of all cases under consideration

§3.1 Sources of Data

Definition (sample & population)

A **population** is the entire collection of all cases under consideration while a **sample** is some subcollection of the population.

§3.1 Sources of Data

Definition (sample & population)

A **population** is the entire collection of all cases under consideration while a **sample** is some subcollection of the population.

Can our class be a population?

§3.1 Sources of Data

Definition (sample & population)

A **population** is the entire collection of all cases under consideration while a **sample** is some subcollection of the population.

Can our class be a population? Could it be a sample?

§3.1 Sources of Data

Definition (sample & population)

A **population** is the entire collection of all cases under consideration while a **sample** is some subcollection of the population.

Can our class be a population? Could it be a sample? If it's a sample, what's the larger population?

§3.1 Sources of Data

Definition (statistic & parameter)

A **statistic** is simply an aggregated measurement depending only on the cases within a sample.

§3.1 Sources of Data

Definition (statistic & parameter)

A **statistic** is simply an aggregated measurement depending only on the cases within a sample. A **parameter** is an aggregated measurement depending on all the cases within a population.

§3.1 Sources of Data

Definition (statistic & parameter)

A **statistic** is simply an aggregated measurement depending only on the cases within a sample. A **parameter** is an aggregated measurement depending on all the cases within a population.

What are some examples of statistics?

§3.1 Sources of Data

Definition (statistic & parameter)

A **statistic** is simply an aggregated measurement depending only on the cases within a sample. A **parameter** is an aggregated measurement depending on all the cases within a population.

What are some examples of statistics? Can these be parameters?

§3.1 Sources of Data

Definition (statistic & parameter)

A **statistic** is simply an aggregated measurement depending only on the cases within a sample. A **parameter** is an aggregated measurement depending on all the cases within a population.

What are some examples of statistics? Can these be parameters? When are these parameters?

§3.1 Sources of Data

Definition (statistic & parameter)

A **statistic** is simply an aggregated measurement depending only on the cases within a sample. A **parameter** is an aggregated measurement depending on all the cases within a population.

What are some examples of statistics? Can these be parameters? When are these parameters?
Note: A parameter can only be determined if you take a **census**!

§3.1 Sources of Data

Definition (observation & experiment)

An **observational study** is one which does not attempt to influence responses whereas an **experimental study** deliberately imposes a treatment on (all or some) individuals and observes their responses.

§3.1 Sources of Data

Definition (observation & experiment)

An **observational study** is one which does not attempt to influence responses whereas an **experimental study** deliberately imposes a treatment on (all or some) individuals and observes their responses.

A treatment is sometimes referred to as an intervention.

§3.1 Sources of Data

Example

Observational or experimental?

- A telemarketing company wants to determine if it's best to call at 6PM or noon. They call two different randomly selected groups at these respective times and record the response rates.

§3.1 Sources of Data

Example

Observational or experimental?

- A telemarketing company wants to determine if it's best to call at 6PM or noon. They call two different randomly selected groups at these respective times and record the response rates.
- A company is interested in improving the health of its workers. It starts a six-week wellness program and records health indicators of a randomly selected group of employees before and after the program.

§3.1 Sources of Data

Definition (confounding)

Confounding occurs when you mistake the effect of one variable for the effect of another.

§3.1 Sources of Data

Definition (confounding)

Confounding occurs when you mistake the effect of one variable for the effect of another.

A study meant to measure the effect of a new drug on hormone levels is given to an experimental group while a placebo is given to a control group. When could confounding occur in this study?

§3.1 Sources of Data

Definition (confounding)

Confounding occurs when you mistake the effect of one variable for the effect of another.

A study meant to measure the effect of a new drug on hormone levels is given to an experimental group while a placebo is given to a control group. When could confounding occur in this study? When the control group has a characteristic common to it that is not common to the experimental group or vice versa.

§3.1 Sources of Data

Definition (confounding)

Confounding occurs when you mistake the effect of one variable for the effect of another.

A study meant to measure the effect of a new drug on hormone levels is given to an experimental group while a placebo is given to a control group. When could confounding occur in this study? When the control group has a characteristic common to it that is not common to the experimental group or vice versa. We can often avoid confounding by properly designing a study.

§3.2 Design of Experiments

Some synonymous words:

§3.2 Design of Experiments

Some synonymous words:

- experimental units = subjects
- outcomes = measured response variable

§3.2 Design of Experiments

Some synonymous words:

- experimental units = subjects
- outcomes = measured response variable (used to compare different treatments)

§3.2 Design of Experiments

Some synonymous words:

- experimental units = subjects
- outcomes = measured response variable (used to compare different treatments)
- factors = explanatory variables

§3.2 Design of Experiments

Some synonymous words:

- experimental units = subjects
- outcomes = measured response variable (used to compare different treatments)
- factors = explanatory variables

Well designed experiments may provide good evidence for causation because they can control for many variables at the same time.

§3.2 Design of Experiments

Some synonymous words:

- experimental units = subjects
- outcomes = measured response variable (used to compare different treatments)
- factors = explanatory variables

Well designed experiments may provide good evidence for causation because they can control for many variables at the same time.

Definition (bias)

A study is **biased** if it systematically favors certain outcomes.

§3.2 Design of Experiments

A well-designed experiment will be **comparative**, **random**, and **repetitive**.

§3.2 Design of Experiments

A well-designed experiment will be **comparative**, **random**, and **repetitive**.

Additionally, a well-designed experiment will be blind or double-blind and realistic!

§3.2 Design of Experiments

A well-designed experiment will be **comparative**, **random**, and **repetitive**.

Additionally, a well-designed experiment will be blind or double-blind and realistic!

Example

A company is interested in how productivity relates to pay. How might the company design an experiment to answer this question?

§3.2 Design of Experiments

A well-designed experiment will be **comparative**, **random**, and **repetitive**.

Additionally, a well-designed experiment will be blind or double-blind and realistic!

Example

A company is interested in how productivity relates to pay. How might the company design an experiment to answer this question? Would this experiment be realistic?

§3.2 Design of Experiments

Randomization is a great tool for removing lurking variables, but there are more intricate study designs which can control for lurking variables even more effectively.

§3.2 Design of Experiments

Randomization is a great tool for removing lurking variables, but there are more intricate study designs which can control for lurking variables even more effectively.

Definition (matched pair design)

In a **matched pair design**, two treatments are compared among homogeneous populations where homogeneity is determined by other factors suspected to be lurking variables.

§3.2 Design of Experiments

Randomization is a great tool for removing lurking variables, but there are more intricate study designs which can control for lurking variables even more effectively.

Definition (matched pair design)

In a **matched pair design**, two treatments are compared among homogeneous populations where homogeneity is determined by other factors suspected to be lurking variables.

Definition (block design)

A block in **block design** is a group of individuals which is grouped by its homogeneity on suspected lurking variable factors. Random selection for treatment groups is done at the block level so that the resulting treatment groups are *uniformly different within the treatment group*.

§3.3 Sampling Design

How do we produce a sample?

§3.3 Sampling Design

How do we produce a sample?

- voluntary response sample

§3.3 Sampling Design

How do we produce a sample?

- voluntary response sample
- simple random sample

§3.3 Sampling Design

How do we produce a sample?

- voluntary response sample
- simple random sample - different than a random sample
- probability sample

§3.3 Sampling Design

How do we produce a sample?

- voluntary response sample
- simple random sample - different than a random sample
- probability sample
- stratified random sample

§3.3 Sampling Design

How do we produce a sample?

- voluntary response sample
- simple random sample - different than a random sample
- probability sample
- stratified random sample
- multistage random sample

§3.3 Sampling Design

Problems in sampling:

- undercovering

§3.3 Sampling Design

Problems in sampling:

- undercovering
- nonresponse

§3.3 Sampling Design

Problems in sampling:

- undercovering
- nonresponse
- response bias

§3.3 Sampling Design

Problems in sampling:

- undercovering
- nonresponse
- response bias
- question order/leading questions

§3.4 Towards Inference

The major goal of statistics is to **be lazy**.

§3.4 Towards Inference

The major goal of statistics is to **be lazy**. Statistics is interested in gaining knowledge about a population through a sample rather than having to take a census.

§3.4 Towards Inference

The major goal of statistics is to **be lazy**. Statistics is interested in gaining knowledge about a population through a sample rather than having to take a census. This process of making trustworthy claims about a population based on a sample of the population is called **statistical inference**. So far, all we've been doing is descriptive statistics.

§3.4 Towards Inference

In order to gain knowledge about a population using a sample, we need to understand how a **statistic** relates to its corresponding **parameter**.

§3.4 Towards Inference

In order to gain knowledge about a population using a sample, we need to understand how a **statistic** relates to its corresponding **parameter**. We do this using the **sampling distribution of a statistic**.

§3.4 Towards Inference

In order to gain knowledge about a population using a sample, we need to understand how a **statistic** relates to its corresponding **parameter**. We do this using the **sampling distribution of a statistic**.

Definition (sampling distribution (of a statistic))

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same fixed size from the same population.

§3.4 Towards Inference

In order to gain knowledge about a population using a sample, we need to understand how a **statistic** relates to its corresponding **parameter**. We do this using the **sampling distribution of a statistic**.

Definition (sampling distribution (of a statistic))

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same fixed size from the same population.

Let's look at an example to illustrate this concept.

§3.4 Towards Inference

Assume we have a (small) population consisting of 5 families with the following number of children in each family: 0, 5, 2, 2, 1. And let's say we're interested in estimating the mean of this population (which is 2, by the way) by taking a sample of size 2. What are the possible values of the mean?

§3.4 Towards Inference

Assume we have a (small) population consisting of 5 families with the following number of children in each family: 0, 5, 2, 2, 1. And let's say we're interested in estimating the mean of this population (which is 2, by the way) by taking a sample of size 2. What are the possible values of the mean?

Mean	Probability of Observing Mean
0.5	
1	
1.5	
2	
3	
3.5	

§3.4 Towards Inference

Assume we have a (small) population consisting of 5 families with the following number of children in each family: 0, 5, 2, 2, 1. And let's say we're interested in estimating the mean of this population (which is 2, by the way) by taking a sample of size 2. What are the possible values of the mean?

Sample Mean	Probability of Observing Sample Mean
0.5	0.1
1	0.2
1.5	0.2
2	0.1
2.5	0.1
3	0.1
3.5	0.2

§3.4 Towards Inference

Ideally, we would like sampling distributions to be centered about the parameter they are mirroring and to have low spread/variation. These issues are encapsulated in the ideas below.

§3.4 Towards Inference

Ideally, we would like sampling distributions to be centered about the parameter they are mirroring and to have low spread/variation. These issues are encapsulated in the ideas below.

Definition (bias)

We call a statistic **biased** if the mean of the sampling distribution is not equal to the true population parameter. If they are equal, we call the statistic unbiased.

§3.4 Towards Inference

Ideally, we would like sampling distributions to be centered about the parameter they are mirroring and to have low spread/variation. These issues are encapsulated in the ideas below.

Definition (bias)

We call a statistic **biased** if the mean of the sampling distribution is not equal to the true population parameter. If they are equal, we call the statistic unbiased.

Definition (variability of a statistic)

The **variability of a statistic** is the spread of its sampling distribution. When we use the standard deviation to measure variability, we call this the **standard error** of the statistic, i.e standard error is just the standard deviation of the sampling distribution.

§3.4 Towards Inference

Generally, to reduce bias, we can make sure our samples are SRSs, and to reduce variability, we can take very large samples. We'll formalize this notion when we talk about the **Central Limit Theorem** in later lessons.

§3.4 Towards Inference

Generally, to reduce bias, we can make sure our samples are SRSs, and to reduce variability, we can take very large samples. We'll formalize this notion when we talk about the **Central Limit Theorem** in later lessons. We can also reduce the **margin of error** by taking larger samples - the margin of error describes a range of the likely error, i.e. its distance from the true population parameter, made by a statistic.

§3.4 Towards Inference

Generally, to reduce bias, we can make sure our samples are SRSs, and to reduce variability, we can take very large samples. We'll formalize this notion when we talk about the **Central Limit Theorem** in later lessons. We can also reduce the **margin of error** by taking larger samples - the margin of error describes a range of the likely error, i.e. its distance from the true population parameter, made by a statistic.

Note: The variability of a statistic does not depend on the size of the population generally so long as the population is at least 100 times larger than the sample itself.

§3.Ethics

We won't discuss the ethical implications of studies, but it's important to have a concept of

§3.Ethics

We won't discuss the ethical implications of studies, but it's important to have a concept of

1. institutional review boards,

§3.Ethics

We won't discuss the ethical implications of studies, but it's important to have a concept of

1. institutional review boards,
2. informed consent,

§3.Ethics

We won't discuss the ethical implications of studies, but it's important to have a concept of

1. institutional review boards,
2. informed consent, and
3. confidentiality/anonymity.

§3.Ethics

We won't discuss the ethical implications of studies, but it's important to have a concept of

1. institutional review boards,
2. informed consent, and
3. confidentiality/anonymity.

You should also be aware of the problems presented by **clinical trials** - see the Tuskegee Experiment.