Producing data, sampling

What percentage of voters approve of the way the U.S. President is handling his job?

Producing data, sampling

What percentage of voters approve of the way the U.S. President is handling his job?

This is difficult to determine exactly as there are more than 250 million people of voting age in the U.S.

Producing data, sampling

What percentage of voters approve of the way the U.S. President is handling his job?

This is difficult to determine exactly as there are more than 250 million people of voting age in the U.S.

But it's not difficult to *estimate* this percentage quite well:

Producing data, sampling

What percentage of voters approve of the way the U.S. President is handling his job?

This is difficult to determine exactly as there are more than 250 million people of voting age in the U.S.

But it's not difficult to *estimate* this percentage quite well:

Sample 1,000 (say) voters at random. Then use the approval percentage among those voters as an estimate for the approval percentage of all voters.

What is statistical inference?

Population: the entire group of subjects about which we want information

What is statistical inference?

Population: the entire group of subjects about which we want information

all U.S. voters

What is statistical inference?

Population: the entire group of subjects about which we want information

all U.S. voters

Parameter: the quantity about the population we are interested in

What is statistical inference?

Population: the entire group of subjects about which we want information

all U.S. voters

Parameter: the quantity about the population we are interested in

approval percentage among all U.S. voters

What is statistical inference?

Population: the entire group of subjects about which we want information
all U.S. voters

Parameter: the quantity about the population we are interested in
approval percentage among all U.S. voters

Sample: the part of the population from which we collect information

What is statistical inference?

Population: the entire group of subjects about which we want information

all U.S. voters

Parameter: the quantity about the population we are interested in

approval percentage among all U.S. voters

Sample: the part of the population from which we collect information

the 1,000 voters selected at random

What is statistical inference?

Population: the entire group of subjects about which we want information

all U.S. voters

Parameter: the quantity about the population we are interested in

approval percentage among all U.S. voters

Sample: the part of the population from which we collect information

the 1,000 voters selected at random

Statistic (estimate): the quantity we are interested in as measured in the sample

What is statistical inference?

Population: the entire group of subjects about which we want information

all U.S. voters

Parameter: the quantity about the population we are interested in

approval percentage among all U.S. voters

Sample: the part of the population from which we collect information

the 1,000 voters selected at random

Statistic (estimate): the quantity we are interested in as measured in the sample

approval percentage among the sampled voters

What is statistical inference?

Population: the entire group of subjects about which we want information

all U.S. voters

Parameter: the quantity about the population we are interested in

approval percentage among all U.S. voters

Sample: the part of the population from which we collect information

the 1,000 voters selected at random

Statistic (estimate): the quantity we are interested in as measured in the sample

approval percentage among the sampled voters

Key point: even a relatively small sample (100 or 1,000) will produce an estimate that is close to the parameter of a very large population of 250 million subjects.

What is statistical inference?

Population: the entire group of subjects about which we want information
all U.S. voters

Parameter: the quantity about the population we are interested in
approval percentage among all U.S. voters

Sample: the part of the population from which we collect information
the 1,000 voters selected at random

Statistic (estimate): the quantity we are interested in as measured in the sample
approval percentage among the sampled voters

Key point: even a relatively small sample (100 or 1,000) will produce an estimate that is close to the parameter of a very large population of 250 million subjects. This is the reason why statistics is so powerful.

Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

This will introduce **bias**, i.e. this sampling will favor a certain outcome.

Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

This will introduce **bias**, i.e. this sampling will favor a certain outcome.

selection bias: a sample of convenience makes it more likely to sample certain subjects than others

Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

This will introduce **bias**, i.e. this sampling will favor a certain outcome.

selection bias: a sample of convenience makes it more likely to sample certain subjects than others

non-response bias: parents are less likely to answer a survey request at 6 pm because they are busy with children and dinner

Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

This will introduce **bias**, i.e. this sampling will favor a certain outcome.

selection bias: a sample of convenience makes it more likely to sample certain subjects than others

non-response bias: parents are less likely to answer a survey request at 6 pm because they are busy with children and dinner

voluntary response bias: websites that post reviews of businesses are more likely to get responses from customers who had very bad or very good experiences

Sampling designs

The best methods for sampling use chance in a planned way:

Sampling designs

The best methods for sampling use chance in a planned way:

a **simple random sample** selects subjects at random without replacement

Sampling designs

The best methods for sampling use chance in a planned way:

a **simple random sample** selects subjects at random without replacement

a **stratified random sample** divides the population into groups of similar subjects called *strata* (e.g. urban, suburban, and rural voters).

Sampling designs

The best methods for sampling use chance in a planned way:

a **simple random sample** selects subjects at random without replacement

a **stratified random sample** divides the population into groups of similar subjects called *strata* (e.g. urban, suburban, and rural voters). Then one chooses a simple random sample in each stratum and combines these.

Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**.

Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

$$\text{estimate} \;=\; \text{parameter} + \text{bias} + \text{chance error}$$

Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

$$\text{estimate} \;=\; \text{parameter} + \text{bias} + \text{chance error}$$

The chance error (sampling error) will get smaller as the sample size gets bigger.

Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

$$\text{estimate} \;=\; \text{parameter} + \text{bias} + \text{chance error}$$

The chance error (sampling error) will get smaller as the sample size gets bigger. Moreover, we can compute how large the chance error will be.

Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

$$\text{estimate} \;=\; \text{parameter} + \text{bias} + \text{chance error}$$

The chance error (sampling error) will get smaller as the sample size gets bigger.
Moreover, we can compute how large the chance error will be.

This is not the case for the bias (systematic error):
Increasing the sample size just repeats the error on a larger scale, and typically we don't know how large the bias is.

Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- This means that there is an **association** between red meat consumption and cancer: there is a link between these two.

Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- This means that there is an **association** between red meat consumption and cancer: there is a link between these two.
- But this does **not** mean that eating red meat *causes* cancer: people who don't eat red meat are known to exercise more and drink less alcohol, and it could be the latter two issues that cause the difference in cancer rates.

Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- This means that there is an **association** between red meat consumption and cancer: there is a link between these two.
- But this does **not** mean that eating red meat *causes* cancer: people who don't eat red meat are known to exercise more and drink less alcohol, and it could be the latter two issues that cause the difference in cancer rates.

This is an **observational study**: It measures outcomes of interest and this can be used to establish association.

Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- This means that there is an **association** between red meat consumption and cancer: there is a link between these two.
- But this does **not** mean that eating red meat *causes* cancer: people who don't eat red meat are known to exercise more and drink less alcohol, and it could be the latter two issues that cause the difference in cancer rates.

This is an **observational study**: It measures outcomes of interest and this can be used to establish association.

But **association is not causation**, because there may be **confounding factors** such as exercise that are associated both with red meat consumption and cancer.

Randomized controlled experiments

To establish causation, an **experiment** is required:

Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- The **subjects** are assigned into treatment and control groups at **random**.

Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- The **subjects** are assigned into treatment and control groups at **random**.
- When possible, subjects in the control group get a **placebo**: it resembles the treatment but is neutral.

Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- The **subjects** are assigned into treatment and control groups at **random**.
- When possible, subjects in the control group get a **placebo**: it resembles the treatment but is neutral. Assigning a placebo makes sure that both groups are equally affected by the **placebo effect**: the idea of being treated may have an effect by itself.

Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- The **subjects** are assigned into treatment and control groups at **random**.
- When possible, subjects in the control group get a **placebo**: it resembles the treatment but is neutral. Assigning a placebo makes sure that both groups are equally affected by the **placebo effect**: the idea of being treated may have an effect by itself.
- The experiment is **double-blind**: neither the subjects nor the evaluators know the assignments to treatment and control.

The placebo effect

The placebo effect is still not fully understood and is one of the most interesting phenomena in science.

The placebo effect

The placebo effect is still not fully understood and is one of the most interesting phenomena in science.

'The weird power of the placebo effect, explained' by Brian Resnick (7/7/2017)

The logic of randomized controlled experiments

Randomization serves two purposes:

The logic of randomized controlled experiments

Randomization serves two purposes:

- It makes the treatment group similar to the control group. Therefore influences other than the treatment operate equally on both groups, apart from differences due to chance.

The logic of randomized controlled experiments

Randomization serves two purposes:

- It makes the treatment group similar to the control group. Therefore influences other than the treatment operate equally on both groups, apart from differences due to chance.
- It allows to assess how relevant the treatment effect is, by calculating the size of chance effects when comparing the outcomes in the two groups (see later).

Mini quiz

For each of the following three sampling plans, say whether it represents simple random sampling or whether it leads to selection bias or to non-response bias, or to voluntary response bias:

- A news company located next to Times Square in New York wants to get a sense how people feel about a proposed law on immigration. A reporter steps out of the building and randomly selects 100 people walking there and asks them about the proposed law.
- A car company wants to get a sense how satisfied the owners of its new car model are with the quality of that car. It randomly selects 250 numbers from the all the vehicle registration numbers that have been issued for this model and contacts the owners of that model.
- An airline wants to do a customer survey in order to improve its service. For one month, it sends an email to a random sample of customers which flew with the airline on the previous day (no customer will be contacted more than once). The email states that the airline would like the customer to fill out a 10 minute survey in order to help the airline improve its service.