Collecting Statistical Data

Math 105

Section 003 - Prasad

September 27, 2010

Why we need to understand statistics

Example: http://thesocietypages.org/socimages/2010/09/13/guest-post-delusions-of-dimorphism/

Definitions

In your groups, try to come up with good working definitions (in your own words!) for the following terms.

- population
- N-value
- census
- survey
- sample
- sampling

Example

Using your definitions, identify the population, the N-value, the type of data collection (census or survey), and if, applicable, how a sample was chosen.

Fox News gives an opinion poll on their website asking their (roughly 2.25 million) viewers whether or not President Obama is doing a good job so far. The answer came back overwhelmingly negative - 79% responded no. MSNBC, however, held a similar poll on their website, asking their 1.85 million viewers if they thought President Obama's policies were harming the country. Around 95% said no, the president's policies were not harming the country.

What kind of sampling are these examples of? Would you believe the results of either of these polls? Why or why not? Are there any other issues that add to the ambiguity of these results?

Sampling methods

- convenience
- quota
- random

Conducting a survey

In your groups, create a one-question poll, and aim it at the class population. Will you give people a choice of a few answers, or let them respond any way they please? Have a thought-out reason for each of your decisions about the poll, and then designate one person to conduct the survey. They should not ask more than 10 people for a response, so think about how you would choose these 10 people.

Identify the following:

- the population of your survey
- the N-value
- the sample. How/why did you pick this as your sample?
- the sampling method (convenience? quota? random?). Why did you choose this method?

Conducting a survey

Based on the surveys you conducted yesterday:

- What was the population of your survey?
- What was the N-value?
- What was your sampling frame? What was the sample size? How does this differ from the population?
- What was the sampling method you used?

Sampling methods

- convenience sampling: choosing the sampling frame is based on whoever it's easiest to survey (e.g. standing on a streetcorner and asking questions to whoever passes by)

Sampling methods

- convenience sampling: choosing the sampling frame is based on whoever it's easiest to survey (e.g. standing on a streetcorner and asking questions to whoever passes by)
- quota sampling: trying to get a representative sample based on the demographics of the population (e.g. making sure you have representative proportions of men, women, racial or religious breakdowns, etc)

Sampling methods

- convenience sampling: choosing the sampling frame is based on whoever it's easiest to survey (e.g. standing on a streetcorner and asking questions to whoever passes by)
- quota sampling: trying to get a representative sample based on the demographics of the population (e.g. making sure you have representative proportions of men, women, racial or religious breakdowns, etc)
- random sampling: letting the laws of chance pick a representative sample - a truly random sample is more likely to be truly representative than even the most careful quota sampling

Sampling methods

- convenience sampling: choosing the sampling frame is based on whoever it's easiest to survey (e.g. standing on a streetcorner and asking questions to whoever passes by)
- quota sampling: trying to get a representative sample based on the demographics of the population (e.g. making sure you have representative proportions of men, women, racial or religious breakdowns, etc)
- random sampling: letting the laws of chance pick a representative sample - a truly random sample is more likely to be truly representative than even the most careful quota sampling
  - simple random sampling

Sampling methods

- convenience sampling: choosing the sampling frame is based on whoever it's easiest to survey (e.g. standing on a streetcorner and asking questions to whoever passes by)
- quota sampling: trying to get a representative sample based on the demographics of the population (e.g. making sure you have representative proportions of men, women, racial or religious breakdowns, etc)
- random sampling: letting the laws of chance pick a representative sample - a truly random sample is more likely to be truly representative than even the most careful quota sampling
  - simple random sampling
  - stratified sampling

Simple random sampling

- Think of this as "lottery sampling."

Simple random sampling

- Think of this as "lottery sampling."
- Basically, each member of a population can be assigned a number, and we randomly pick a few of these numbers - like out of a hat.

Simple random sampling

- Think of this as "lottery sampling."
- Basically, each member of a population can be assigned a number, and we randomly pick a few of these numbers - like out of a hat.
- These few that we pick make up the sample.

Simple random sampling

- Think of this as "lottery sampling."
- Basically, each member of a population can be assigned a number, and we randomly pick a few of these numbers - like out of a hat.
- These few that we pick make up the sample.
- This is not really a great way to find a sample from a very large population.

Stratified sampling

- This is basically a way to make simple random sampling a little easier to implement.

Stratified sampling

- This is basically a way to make simple random sampling a little easier to implement.
- We break up our big population into **strata**, which are broad categories, like geographic regions.

Stratified sampling

- This is basically a way to make simple random sampling a little easier to implement.
- We break up our big population into **strata**, which are broad categories, like geographic regions.
- We use simple random sampling to pick some of these categories.

Stratified sampling

- This is basically a way to make simple random sampling a little easier to implement.
- We break up our big population into **strata**, which are broad categories, like geographic regions.
- We use simple random sampling to pick some of these categories.
- Then, we divide these strata up further into **substrata**, and pick some of those substrata by simple random sampling. We can do this as many times as we want to.

Stratified sampling

An example of stratified sampling: students

Biases

- **selection bias**: when a sample has a tendency to exclude a group or characteristic with a population (e.g. a survey of students at U of A which is taken from dorm residents - who does this exclude?)

Biases

- **selection bias**: when a sample has a tendency to exclude a group or characteristic with a population (e.g. a survey of students at U of A which is taken from dorm residents - who does this exclude?)
- **nonresponse bias**: when the response rate of the survey is very low (this usually means that you're only getting responses from people who really care)

More terminology

- **sampling proportion**: the proportion of the population that the sample is
- **statistic**: any sort of numerical data that comes **from the sample**
- **parameter**: the actual number that the statistic is estimating

parameter vs. statistic

For example, suppose that there are 747 people who like *Star Trek* at the U of A. This number is a **parameter**. However, you do not know this, and you want to conduct a survey that estimates how many people like *Star Trek*. **From the results of your survey**, you estimate that about 750 people like *Star Trek*. This number is the **statistic**.

Usually, we do not know the parameter (and cannot figure it out), so we conduct surveys to come up with statistics, which let us approximate the parameter.

Sampling error

Sampling error is what we call the difference between a parameter and a statistic.

- **Chance error** results from the fact that a sample can only approximate the population. This is unavoidable, but through proper sampling methods, can be kept very low.
- **Sample bias** happens when you pick a bad sample. This we can avoid by doing a good job of selecting our sample.

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?
- Describe the sample.

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?
- Describe the sample.
- Give a sample statistic for the percentage of red gumballs in the jar.

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?
- Describe the sample.
- Give a sample statistic for the percentage of red gumballs in the jar.
- What was the sampling method used?

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?
- Describe the sample.
- Give a sample statistic for the percentage of red gumballs in the jar.
- What was the sampling method used?
- What was the sampling proportion?

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?
- Describe the sample.
- Give a sample statistic for the percentage of red gumballs in the jar.
- What was the sampling method used?
- What was the sampling proportion?
- Suppose the jar contains 150 green and 50 red gumballs.

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?
- Describe the sample.
- Give a sample statistic for the percentage of red gumballs in the jar.
- What was the sampling method used?
- What was the sampling proportion?
- Suppose the jar contains 150 green and 50 red gumballs.
  - What is the parameter for the percentage of red gumballs?

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?
- Describe the sample.
- Give a sample statistic for the percentage of red gumballs in the jar.
- What was the sampling method used?
- What was the sampling proportion?
- Suppose the jar contains 150 green and 50 red gumballs.
  - What is the parameter for the percentage of red gumballs?
  - What is the sampling error, in percent?

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?
- Describe the sample.
- Give a sample statistic for the percentage of red gumballs in the jar.
- What was the sampling method used?
- What was the sampling proportion?
- Suppose the jar contains 150 green and 50 red gumballs.
  - What is the parameter for the percentage of red gumballs?
  - What is the sampling error, in percent?
  - Do you think the sampling error was a chance error or the result of sampling bias?

Example

As part of a class-project, a teacher brings to class a large jar full of 200 gumballs, some red and some green. The students are supposed to estimate the proportion of red gumballs in the jar. One of the students shakes the jar and draws 25 gumballs: 8 red and 17 green.

- Describe the population. Do you know the N-value?
- Describe the sample.
- Give a sample statistic for the percentage of red gumballs in the jar.
- What was the sampling method used?
- What was the sampling proportion?
- Suppose the jar contains 150 green and 50 red gumballs.
  - What is the parameter for the percentage of red gumballs?
  - What is the sampling error, in percent?
  - Do you think the sampling error was a chance error or the result of sampling bias?
- What data collection method should be used to find the **parameter** for the percentage of red gumballs in the jar?

Clinical Studies

Clinical studies are different from surveys in that they usually involve active involvement from the people running the study. (A survey is ultimately quite passive: once you set up your sampling method, you just ask questions of people).

Usually, clinical studies are used to test the effect of one variable (like a drug or a treatment). Does this variable hurt, help or make no difference at all?

How they work

1. First you have to choose some subjects to test the variable out on. You want this group to be pretty similar - as similar as you can make them - in order to minimize the chance that other things will interfere (that is, **confounding variables**).
2. The subjects are divided into two groups:

Control group	Treatment group
not exposed to the variable	exposed to the variable (given the treatment)

   It's best to divide subjects randomly into these groups.
3. The control group should be given some sort of **placebo** (that is, something that looks like the variable, but isn't). If the subjects are not told what group they are in, the study is **blind**.
4. In order to minimize psychological effects completely, studies can be **double-blind**, that is, not even the researchers collecting the data know which group is which.

Capture-recapture method - theory

- With this method, you are pretty much assuming that all the individuals in the population that you are trying to estimate the size of are all identical. Why is this a valid assumption to make?

Capture-recapture method - theory

- With this method, you are pretty much assuming that all the individuals in the population that you are trying to estimate the size of are all identical. Why is this a valid assumption to make?
- Since they are all identical, you cannot tell one apart from the other (as with the activity from Friday - all the individuals in the population were just squares of same-colored paper). So, you impose a characteristic on some of them, by tagging, and let them go. The number of individuals you choose is n_1. You now know that there is some proportion of the whole population (which has size N), that has this certain characteristic.

Capture-recapture method - theory

- Then, you want to estimate what proportion of the population has this characteristic (that is, the tag). This is exactly the process of finding a statistic. So you get a proportion k/n_2, where k is the number of tagged individuals in the second sample, n_2 is the size of the second sample.

Capture-recapture method - theory

- Then, you want to estimate what proportion of the population has this characteristic (that is, the tag). This is exactly the process of finding a statistic. So you get a proportion k/n_2, where k is the number of tagged individuals in the second sample, n_2 is the size of the second sample.
- You already know that the parameter is n_1/N, but you do not know N, and you know that the statistic is an approximation of the parameter. So we can say,

$$\frac{n_1}{N} \approx \frac{k}{n_2}, \quad \text{so} \quad N \approx \frac{n_1 \cdot n_2}{k}.$$

Capture-recapture method - issues

- The capture-recapture method assumes that the capturing process is truly random. However, this might not be true. When tagging animals, which ones are most likely to be captured? In that case, does this method overestimate or underestimate the size of the population?
- Also, we assume that when we tag an animal, they are not harmed by the tags. However, if the tag is a bright color (as used with fish, sometimes), this might cause problems. How? In this case, are you likely to overestimate or underestimate the size of the population?

Deceptive Statistics - #73

An article in the *Providence Journal* about automobile accident fatalities includes the following observation:

> Forty-two percent of all fatalities occurred on Friday, Saturday, and Sunday, apparently because of increased drinking on the weekends.

a. Give a possible argument as to why the conclusion drawn may not be justified by the data.
b. Give a different possible argument as to how the conclusion drawn may be justified by the data after all.

Leading question bias - #69

Consider the following question asked on a hypothetical survey:

Are you in favor of paying higher taxes to bail the federal government out of its disastrous economic policies and its mismanagment of the federal budget? Yes _____. No _____.

Ninety-five percent of respondents said no. The people running the survey concluded that 95% of people were not in favor of raising taxes.

a. Explain why the results of this survey might be invalid.
b. Can you rephrase the question in a neutral way?
c. What are some other (possibly more subtle) examples of leading question bias?

Sampling methods - #29

You are a fruit wholesaler. You have just received 250 crates of pineapples: 75 crates from supplier A, 75 crates from supplier B, and 100 crates from supplier C. You want to see if the pineapples are good enough to ship to your best customers. You decide to inspect a sample of $n = 20$ crates. How would you implement each of the following sampling methods:

- convenience sampling
- simple random sampling
- stratified sampling
- quota sampling