

Excel II

Econ 5/Poli 5D Lecture 3

## Announcements

- First quiz due Friday
- First homework due next Tuesday at midnight
- Any questions?

## Today's Application: Surveys and Survey Sampling

- A lot of data we collect is based on surveys
- Important source of knowledge, but can be biased for many reasons
- We want surveys to be representative of the population
- How do we design a survey that accomplishes this goal and what biases may arise?

## Today's Application: Surveys and Survey Sampling

- A lot of data we collect is based on surveys
- Important source of knowledge, but can be biased for many reasons
- We want surveys to be representative of the population
- How do we design a survey that accomplishes this goal and what biases may arise?

## Today's Software: Microsoft Excel

- Continue practicing using functions and classifying variable types
- Use the RAND function to conduct a simple random sample

## Overview of Today's Application

In 1936, the U.S. had entered the eighth year of the Great Depression

Presidential election between Franklin D. Roosevelt and Alf Landon

Before the election many political pundits expected a close race

- The Literary Digest sent out 10 million questionnaires
- Based on these questionnaires, they predicted Alf Landon would win with 57.1 percent of the vote
- The Literary Digest had correctly predicted the winner in the previous 5 elections

In the end, polls were completely off, Franklin D. Roosevelt won with 60.8 percent of the vote and 523 of 531 electoral votes

So what went wrong?

Consider following fake dataset of individuals based on the 1936 election

We start our discussion with two variables

- id (person identifier)
- indicator for whether the person is going to vote for Alf Landon

Goal: we want to know the values in column B, but we can't survey everyone

	A	B	C
1	id	voted for Alf Landon (unobserved)	answered survey (if selected)
2	1	1.00	N
3	2	0.00	Y
4	3	0.00	N
5	4	1.00	Y
6	5	0.00	N
7	6	1.00	Y
8	7	0.00	N
9	8	1.00	Y
10	9	1.00	N
11	10	1.00	Y

In this setting we are interested in a **population** average (e.g. poverty rates, voting, etc.) but polling the entire population is too costly

We settle for a **sample** instead

What can we learn from a sample?

- Depends on the **external validity** of the sample
- A sample is **externally valid** if its results can be generalized to the entire population

Ex: If I surveyed UCSD students on their favorite tv show, would that be reflective of the U.S. in general

- Of course not! UCSD students tastes in television are probably not representative of country in general

**External Validity** is an enormously important concept in social science

Example: We have a new treatment to lower high blood pressure

The new treatment was developed by UCSD researchers, so out of convenience, we test it on UCSD students

We find no effect of the new treatment on blood pressure

Does this mean the treatment is useless?

**External Validity** is an enormously important concept in social science

Example: We have a new treatment to lower high blood pressure

The new treatment was developed by UCSD researchers, so out of convenience, we test it on UCSD students

We find no effect of the new treatment on blood pressure

Does this mean the treatment is useless?

- No! UCSD students may not have high blood pressure to begin with, so it might not be surprising that we find no effect

How do we retrieve externally valid estimates?

One method is a simple random sample

- Put simply, take a random sample of the entire population

Accurate under three conditions:

1. You are taking the sample from the entire population
2. Everyone sampled responds to the survey
3. Everyone tells the truth

How do we retrieve externally valid estimates?

One method is a simple random sample

- Put simply, take a random sample of the entire population

Accurate under three conditions:

1. You are taking the sample from the entire population
2. **Everyone sampled responds to the survey**
3. Everyone tells the truth

## Non-response bias

In most cases, not everyone will respond to a survey

If the people who don't respond are different than the people who do respond, then our estimates will be biased

Ex: the **Current Population Survey (CPS)** is a survey of around 60,000 households per month

- Unemployment statistics in the news come primarily from this survey
- What if people who are busy are with work are more likely to refuse to take the survey
- Then the responders will be more likely to be unemployed than the non-responders
- In other words, the survey will overestimate the unemployment rate in the economy

How do we retrieve externally valid estimates?

One method is a simple random sample

- Put simply, take a random sample of the entire population

Accurate under three conditions:

1. You are taking the sample from the entire population
2. Everyone sampled responds to the survey
3. **Everyone tells the truth**

## Response bias (e.g. Social desirability bias)

Not the opposite of non-response bias (confusing terminology). You should think of it as misreporting bias

Consider how respondents might be inclined to provide biased answers to the following questions

- Who are you going to vote for? (Depending on context, may feel social pressure to give a particular answer)
- How much income do you earn per year? (May not know exact income, so round instead)
- How much tv do you watch a week? (If like me you watch too much tv, you may feel inclined to round down, as I have on past surveys)

Back to example: What went wrong with Literary Digest's 1936 election poll?

We will consider a fake example in which there are only 100 individuals

- These 100 individuals represent all voters in the U.S. in 1936

We will be the Literary Digest and consider how to predict the election outcome

- Should we be concerned with non-response bias?
- Should we be concerned with response bias?

	A	B	C	D
1	id	voted for Alf Landon (unobserved)	would answer survey	Reported voting for Alf Landon
2	1	1.00	N	1.00
3	2	0.00	Y	1.00
4	3	0.00	N	0.00
5	4	1.00	Y	1.00
6	5	0.00	N	0.00
7	6	1.00	Y	1.00
8	7	0.00	N	0.00
9	8	1.00	Y	1.00
10	9	1.00	N	1.00
11	10	1.00	Y	1.00
12	11	0.00	N	0.00
13	12	0.00	N	0.00
14	13	1.00	N	1.00
15	14	1.00	N	1.00
16	15	0.00	N	1.00

It is too expensive to survey everyone so we will take a sample

To ensure it is representative of the population we will take a simple random sample

We will generate a random number between 0 and 1 and choose all individuals with a value greater than 0.5

`=RAND ( )` generates a number between 0 and 1

`=( E2>0.5 )` will report if the individual is selected to participate in the survey

E	F
random number	selected?
0.66002965	TRUE
0.01175974	FALSE
0.85556886	TRUE
0.31703544	FALSE
0.0464334	FALSE
0.90064236	TRUE
0.80286474	TRUE
0.16738762	FALSE
0.34442414	FALSE
0.14562932	FALSE
0.39463156	FALSE
0.60462245	TRUE
0.96610175	TRUE
0.18704728	FALSE
0.55016666	TRUE

id	voted for Alf Landon (unobserved)	would answer survey	Reported voting for Alf Landon	random number	selected?	selected and responded?
1	1.00	N	1.00	0.66002965	TRUE	0
2	0.00	Y	1.00	0.01175974	FALSE	0
3	0.00	N	0.00	0.85556886	TRUE	0
4	1.00	Y	1.00	0.31703544	FALSE	0
5	0.00	N	0.00	0.0464334	FALSE	0
6	1.00	Y	1.00	0.90064236	TRUE	1
7	0.00	N	0.00	0.80286474	TRUE	0
8	1.00	Y	1.00	0.16738762	FALSE	0
9	1.00	N	1.00	0.34442414	FALSE	0
10	1.00	Y	1.00	0.14562932	FALSE	0
11	0.00	N	0.00	0.39463156	FALSE	0
12	0.00	N	0.00	0.60462245	TRUE	0
13	1.00	N	1.00	0.96610175	TRUE	0
14	1.00	N	1.00	0.18704728	FALSE	0
15	0.00	N	1.00	0.55016666	TRUE	0

id	voted for Alf Landon (unobserved)	would answer survey	Reported voting for Alf Landon	random number	selected?	selected and responded?
1	1.00	N	1.00	0.66002965	TRUE	0
2	0.00	Y	1.00	0.01175974	FALSE	0
3	0.00	N	0.00	0.85556886	TRUE	0
4	1.00	Y	1.00	0.31703544	FALSE	0
5	0.00	N	0.00	0.0464334	FALSE	0
6	1.00	Y	1.00	0.90064236	TRUE	1
7	0.00	N	0.00	0.80286474	TRUE	0
8	1.00	Y	1.00	0.16738762	FALSE	0
9	1.00	N	1.00	0.34442414	FALSE	0
10	1.00	Y	1.00	0.14562932	FALSE	0
11	0.00	N	0.00	0.39463156	FALSE	0
12	0.00	N	0.00	0.60462245	TRUE	0
13	1.00	N	1.00	0.96610175	TRUE	0
14	1.00	N	1.00	0.18704728	FALSE	0
15	0.00	N	1.00	0.55016666	TRUE	0

Just because an individual is selected does not mean they will respond

Many surveys have shockingly low response rates (Especially in the US and other wealthy countries)

Let's look at the response rate

The response rate is equal to:

$$\text{Response Rate} = \frac{\text{Number of people who respond to the survey}}{\text{Number of people surveyed}}$$

If we contact 100 people, but only 10 respond, response rate is 0.1 or 10 percent

1. In our spreadsheet we need to count all individuals who responded (column G, individuals with a 1)
2. All individuals who were selected to be surveyed (column F, individuals with TRUE)
3. Then we can find the response rate by dividing (1) by (2)

```
=COUNTIF(G:G,1)/(COUNTIF(F:F,TRUE))
```

1. In our spreadsheet we need to count all individuals who responded (column G, individuals with a 1)
2. All individuals who were selected to be surveyed (column F, individuals with TRUE)
3. Then we can find the response rate by dividing (1) by (2)

```
=COUNTIF(G:G,1)/(COUNTIF(F:F,TRUE))
```

Response Rate
0.4

Note: it is ok if you got a different response rate!

RAND generates random numbers

We may get different results depending on who we randomly selected to survey

Only 40 percent of the individuals that were surveyed actually responded




This is better than the actual response rate of the Literary Digest which received answers from 2.5 million of their 10 million questionnaires

Does this matter? Maybe

- Yes: those who did not respond are more likely to vote for one candidate over the other, so our estimate will be biased
- No: responding is not correlated with voting preference, so even though we have fewer observations, our estimate of the percent voting for Alf Landon will still be accurate

SUM    =AVERAGE(B:B)													
	A	B	C	D	E	F	G	H	I	J	K	L	M
		voted for Alf Landon (unobserved)	would answer survey	Reported voting for Alf Landon	random number	selected?	selected and responded?						
1	id								Sample	Vote Share (Alf Landon)	Difference from population mean		
2	1	1.00	N	1.00	0.66003	TRUE	0		Entire Population	=AVERAGE(B:B)	-		
3	2	0.00	Y	1.00	0.01176	FALSE	0		Estimate from our survey				
4	3	0.00	N	0.00	0.85557	TRUE	0		Estimate taking out response bias (i.e. based on true votes)				
5	4	1.00	Y	1.00	0.31704	FALSE	0		Estimate taking out response bias and non-response bias				

SUM    =AVERAGEIF(G:G,1,D:D)													
	A	B	C	D	E	F	G	H	I	J	K	L	M
		voted for Alf Landon (unobserved)	would answer survey	Reported voting for Alf Landon	random number	selected?	selected and responded?						
1	id								Sample	Vote Share (Alf Landon)			
2	1	1.00	N	1.00	0.66003	TRUE	0		Entire Population	0.42	-		
3	2	0.00	Y	1.00	0.01176	FALSE	0		Estimate from our survey	=AVERAGEIF(G:G,1,D:D)			
4	3	0.00	N	0.00	0.85557	TRUE	0		Estimate taking out response bias (i.e. based on true votes)				
5	4	1.00	Y	1.00	0.31704	FALSE	0		Estimate taking out response bias and non-response bias				

SUM    =AVERAGEIF(G:G,1,B:B)													
	A	B	C	D	E	F	G	H	I	J	K	L	M
	id	voted for Alf Landon (unobserved)	would answer survey	Reported voting for Alf Landon	random number	selected?	selected and responded?		Sample	Vote Share (Alf Landon)	Difference from population mean		
1	1	1.00	N	1.00	0.66003	TRUE	0		Entire Population	0.42	-		
2	2	0.00	Y	1.00	0.01176	FALSE	0		Estimate from our survey	0.65			
3	3	0.00	N	0.00	0.85557	TRUE	0		Estimate taking out response bias (i.e. based on true votes)	=AVERAGEIF(G:G,1,B:B)			
4	4	1.00	Y	1.00	0.31704	FALSE	0		Estimate taking out response bias and non-response bias				

SUM													
=AVERAGEIF(F:F,TRUE,B:B)													
	A	B	C	D	E	F	G	H	I	J	K	L	M
	id	voted for Alf Landon (unobserved)	would answer survey	Reported voting for Alf Landon	random number	selected?	selected and responded?		Sample	Vote Share (Alf Landon)	Difference from population mean		
1	1	1.00	N	1.00	0.66003	TRUE	0		Entire Population	0.42	-		
2	2	0.00	Y	1.00	0.01176	FALSE	0		Estimate from our survey	0.65			
3	3	0.00	N	0.00	0.85557	TRUE	0		Estimate taking out response bias (i.e. based on true votes)	0.55			
4	4	1.00	Y	1.00	0.31704	FALSE	0		Estimate taking out response bias and non-response bias	=AVERAGEIF(F:F,TRUE,B:B)			

I	J	K	L	M
Sample	Vote Share (Alf Landon)	Difference from population mean		
Entire Population	0.42	-		
Estimate from our survey	0.65	0.23		
Estimate taking out response bias (i.e. based on true votes)	0.55	0.13		
Estimate taking out response bias and non-response bias	0.4	-0.02		

In this example, the key factor was non-response bias

Those who were voting for FDR did not return the questionnaires, so the Literary Digest falsely believe that Alf Landon would easily win the election

We still have this problem today!

This is a key reason why it is so difficult to predict election outcomes

## How accurate have U.S. polls been?

Weighted-average error in polls in final 21 days of the campaign

CYCLE	PRESIDENTIAL		STATE-LEVEL			COMBINED
	PRIMARY	GENERAL	GOVERNOR	U.S. SENATE	U.S. HOUSE	
2017-18	—	—	5.2	6.0	4.1	5.1
2015-16	10.1	4.8	5.4	5.0	5.5	6.8
2013-14	—	—	4.4	5.4	6.7	5.4
2011-12	8.9	3.6	4.8	4.7	4.7	5.1
2009-10	—	—	4.9	4.8	6.9	5.7
2007-08	7.4	3.6	4.1	4.7	5.7	5.4
2005-06	—	—	5.0	4.2	6.5	5.3
2003-04	7.1	3.2	6.1	5.6	5.4	4.8
2001-02	—	—	5.2	4.9	5.4	5.2
1999-2000	7.6	4.4	4.9	6.1	4.4	5.5
1998	—	—	8.1	7.4	6.8	7.5
All years	8.7	4.0	5.4	5.4	6.2	5.9

Data from the Pew Research Center's "The State of the Union" reports. The error is the weighted average of the errors in the polls in the final 21 days of the campaign.