Today's plan:

- Section 4.3: Data Collection
- Section 4.3.1: Population v. Sample
- Section 4.3.2: Population size.

Data sets come mainly from two big groups:

Data sets come mainly from two big groups:

- data sets from random experiments

Data sets come mainly from two big groups:

- data sets from random experiments
- data sets from populations

Example (from experiment)

Consider the random experiment:

- flip a fair coin 10 times

Example (from experiment)

Consider the random experiment:

- flip a fair coin 10 times
- count the number of heads (H)

Example (from experiment)

Consider the random experiment:

- flip a fair coin 10 times
- count the number of heads (H)

Remarks:

- we get a number between 0 and 10

Example (from experiment)

Consider the random experiment:

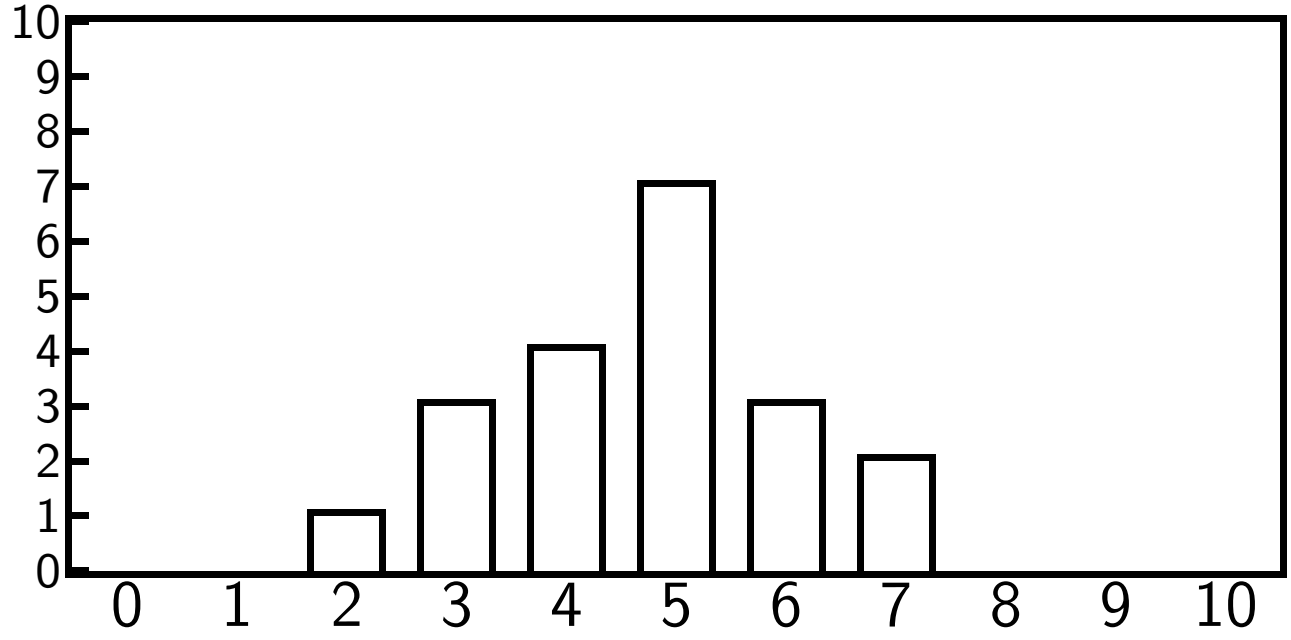
- flip a fair coin 10 times
- count the number of heads (H)

Remarks:

- we get a number between 0 and 10
- we might "expect" to get 5 H's, but nothing's certain.

Example of a coin-flipping experiment:

x	Frq.
2	1
3	3
4	4
5	7
6	3
7	2



Examples coming from populations:

Example

- Test scores (population = students)

Examples coming from populations:

Example

- Test scores (population = students)
- IRS tax database

Examples coming from populations:

Example

- Test scores (population = students)
- IRS tax database
- Average cost of a gallon of milk per state (Hawaii is an outlier)

Population v. Sample

Sometimes (often), gathering data from **every** member of population can be impractical.

Population v. Sample

Sometimes (often), gathering data from **every** member of population can be impractical.
Example: 65,000 people in a football stadium. What's the favorite team?

Population v. Sample

Sometimes (often), gathering data from **every** member of population can be impractical.
Example: 65,000 people in a football stadium. What's the favorite team? Makes no sense to try and poll every single person.

Idea:

Idea:

- Instead of polling every single member of the population, pick a **sample** – a small portion of the population, and gather the data from the sample.

- Once we have the data set from the sample, we can run all the stats (mean, median, standard deviation, etc.).

- Once we have the data set from the sample, we can run all the stats (mean, median, standard deviation, etc.).
- Then we extrapolate and make inferences about the whole population.

In rare cases a sample's not enough, and we do need to poll every single member of the population.

Example

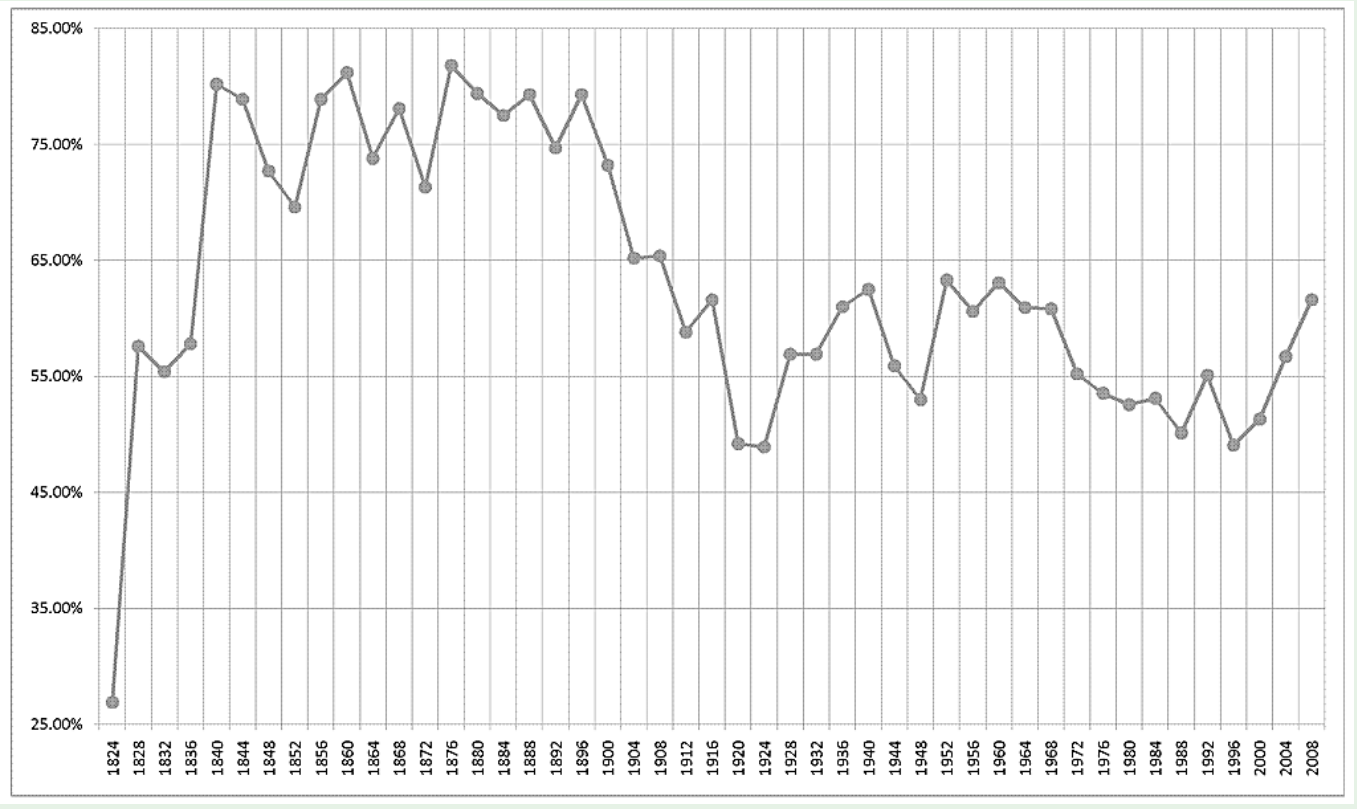
The U.S. Census happens every 10 years. There's a massive effort to get a perfect response rate.

Example

The U.S. Census happens every 10 years. There's a massive effort to get a perfect response rate. Here a sample isn't enough.

Example

In presidential elections the response rate is far from perfect.



(Chart from Wikipedia)

Sometimes the population is small enough it's easy to avoid samples. Like test scores in a class; just use all of them to find mean, median, etc.

One of the hardest parts of a statistical project, and where most mistakes are made, is in the very first step: **data collection**.

Example

"An internet survey shows that 100% of people use internet."

A real historical example:

A real historical example:

Example

The 1936 U.S. presidential election was between

A real historical example:

Example

The 1936 U.S. presidential election was between

- Franklin D. Roosevelt (D)

A real historical example:

Example

The 1936 U.S. presidential election was between

- Franklin D. Roosevelt (D)
- Alf Landon (R)

A real historical example:

Example

The 1936 U.S. presidential election was between

- Franklin D. Roosevelt (D)
- Alf Landon (R)

A magazine took a poll before the election.

Example

- There were $\approx 40,000,000$ registered voters.

Example

- There were $\approx 40,000,000$ registered voters.
- The magazine polled $10,000,000$ people.

Example

- There were $\approx 40,000,000$ registered voters.
- The magazine polled $10,000,000$ people.
- The poll predicted a victory for Landon, 57% to 43%.

Example

- The result of the election was 62% for Roosevelt, and 38% for Landon.

Example

- The result of the election was 62% for Roosevelt, and 38% for Landon.
- An error of 19%!

Solution

There are two major problems with poll taking:

Solution

There are two major problems with poll taking:

- **selection bias**

Solution

There are two major problems with poll taking:

- **selection bias**
- **non-response bias**.

Solution

There are two major problems with poll taking:

- **selection bias**
- **non-response bias**.

Selection bias is when the sample is skewed in one direction.

Solution

The sample for the election poll was mostly rich people, who favored Landon.

Solution

Non-response bias is when lots of people don't respond to the poll.

Solution

Non-response bias *is when lots of people don't respond to the poll. The* **response rate** *above was a mere* 24%.

Solution

Non-response bias is when lots of people don't respond to the poll. The **response rate** *above was a mere* 24%*.People willing to respond to a survey tend to be different from those who aren't, and this affects the sample.*

Question

How to avoid selection bias and non-response bias when conducting a public opinion poll?

Selection bias can be eliminated by just picking the sample as randomly as possible.

Selection bias can be eliminated by just picking the sample as randomly as possible.
This is hard to do perfectly in real life.

Non-response bias is harder to deal with: you can't force people to answer questions.

- Choose an appropriate way to collect the data:

Method	personal interviews	telephone interviews	mail surveys
Response Rate	higher	lower	much lower
Cost	higher	lower	much lower

- Choose an appropriate way to collect the data:

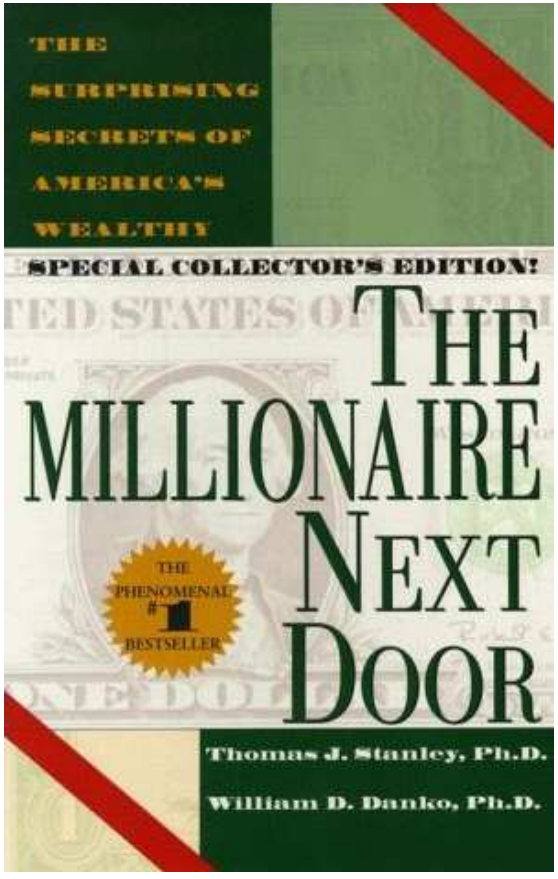
Method	personal interviews	telephone interviews	mail surveys
Response Rate	higher	lower	much lower
Cost	higher	lower	much lower

Example

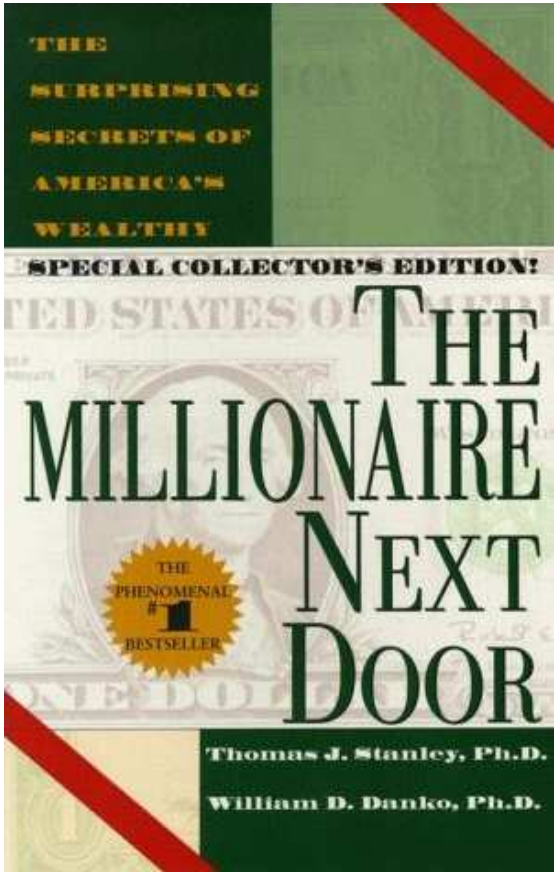
- Use psychology!
- For mail surveys what actually works is including a small reward in the envelope with the survey.

Example

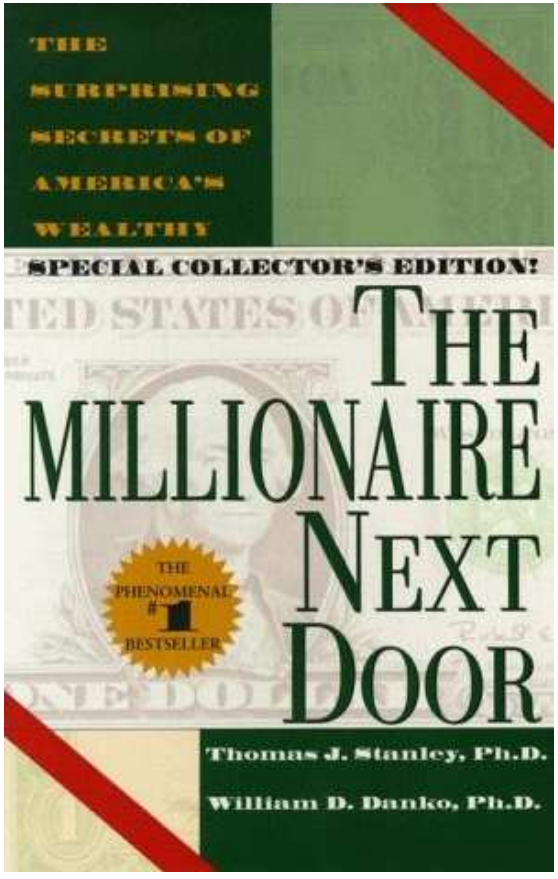
- Use psychology!
- For mail surveys what actually works is including a small reward in the envelope with the survey.
- Surprisingly, the size of the reward doesn't really matter!



- Thomas Stanley and William Danko conducted studies about **millionaires**



- Thomas Stanley and William Danko conducted studies about **millionaires**
- They mailed surveys to them



- Thomas Stanley and William Danko conducted studies about **millionaires**
- They mailed surveys to them
- The response rate increased drastically when they started to include just $1

A really bad way to collect data is a passive survey, like an ad in a newspaper.

A really bad way to collect data is a passive survey, like an ad in a newspaper.
You'll get only a very specific type of person responding. Not a random sample at all!

Example

Another passive survey:

- You want to buy a laptop, and you read reviews online

Example

Another passive survey:

- You want to buy a laptop, and you read reviews online
- Half of the reviews are negative and say it broke

Example

Another passive survey:

- You want to buy a laptop, and you read reviews online
- Half of the reviews are negative and say it broke
- Does it mean that laptop has 50% chance of breaking?

Think about the psychology:

Think about the psychology:
An unhappy customer will be much more likely write a review than a happy one.

Think about the psychology:
An unhappy customer will be much more likely write a review than a happy one.
This doesn't mean reviews are useless, they just might be biased.

Section 4.3.2: Population size

For any population an obvious question is:

For any population an obvious question is:

Question

What is the size of the population?

For any population an obvious question is:

Question

What is the size of the population?

Sometimes it's impossible to answer this precisely, but there are ways to estimate it.

Capture-recapture method

Example

- Estimate the population of fish in a lake.

Capture-recapture method

Example

- Estimate the population of fish in a lake.
- Catch a sample of 150 fish, tag them, and release them.

Example

- A week later, a new sample of 100 fish is caught, and 12 of them have tags.

Example

- A week later, a new sample of 100 fish is caught, and 12 of them have tags.
- What is the number of fish in the lake?

The proportion of tagged fish in the sample

$\approx$

the proportion of tagged fish in the population.

Thus

$$\frac{12}{100} \approx \frac{150}{N}$$

Thus

$$\frac{12}{100} \approx \frac{150}{N}$$

and therefore

$$N \approx \frac{100 \cdot 150}{12} = 1250$$

Remarks:

- We'll return to this later and get a confidence interval for N

Remarks:

- We'll return to this later and get a confidence interval for N
- This 1250 estimate alone is called a **central estimate**.

Next time: Section 4.4: Statistical Inference