What Exactly Can We Learn from Samples?


- Found that 53% of sampled adults disapproved of the Affordable Care Act (ACA), the 2010 health care law. (45% of sampled adults approved of the ACA.)

- Question: Since 53% of the sample disapproved of the ACA, can we conclude that the *majority of the general American adult population* disapproved of the ACA?

- Related question: Is 1504 respondents enough people in the sample to allow any conclusions about the population?
Parameters and Statistics

- **Parameter**: A *number* that describes a *population* in some way.

- **Statistic**: A *number* that describes a *sample* in some way.

- Key difference: In practice, we usually never know the *actual value* of a parameter. (because we don’t have data on the whole population)

- In contrast, we can calculate the value of a statistic based on the sample data, which we *do* have.

- So . . . we often use the value of the statistic to *estimate* the unknown value of the parameter.
Pew Research Poll Example Again

• What *proportion of the population* oppose the ACA?

• This is an unknown *parameter* – we denote a population proportion by $p$.

• To estimate $p$, here, we can calculate the proportion in the *sample* opposing the ACA.

• This is a *statistic* which *estimates* the parameter – we denote a sample proportion by $\hat{p}$ (pronounced “p-hat”).
Clicker Quiz 1

Note that of the 1504 adults sampled by Pew, 797 opposed the ACA. What is the sample proportion \( \hat{p} \) opposing the ACA?

A. \( \frac{797}{1504} = 0.53 \)

B. 797

C. 1504

D. \( \frac{1504}{797} = 1.89 \)
• In other words . . . we estimate that 53% of the population of adults oppose this ACA health care law.

Sampling Variability

• What if we conducted another poll (same survey question, but a different random sample of 1504 U.S. adults).

• Would exactly 53% of this new sample oppose the ACA?
Sampling Variability

- What if we conducted another poll (same survey question, but a different random sample of 1504 U.S. adults).

- Would exactly 53% of this new sample oppose the ACA?

- Probably not – maybe 48% would, or 55% would, or 50% would, etc.

- If we did repeated samples of 1504 adults, we’d get a somewhat different \( \hat{p} \) each time.

- If this sample-to-sample variation is too large, then we can’t trust the results of the sample we did take very much.
How Much Sampling Variation Is There?

• Suppose the truth is that $p = 0.50$ is the true proportion opposing the ACA in the population.

• Computer simulations can approximate the variation in $\hat{p}$ values we’d get if we took many random samples from this population.

• Example: Let’s take 1000 SRS’s, each of size 100, from this hypothetical population having $p = 0.5$.

• Results in 1000 $\hat{p}$ values: 0.50 0.55 0.58 0.45 0.55 0.44 0.54 0.41 0.61 0.44, etc.

• Some sample proportions are bigger than 0.5, some are smaller.
Overall picture of all 1000 $\hat{p}$ values:

Sample proportions when sample sizes = 100

Figure 1: Plot of pattern of $\hat{p}$ values from 1000 samples when $n = 100$. 
• **Note:** SRS size of 100 isn’t as big a sample size as in the Pew research poll (had $n = 1504$)

• Now let’s take 1000 SRS’s each of size 1504:

• **Now our 1000 $\hat{p}$ values are:** 0.488 0.471 0.507 0.479 0.528 0.485 0.497 0.494 0.499, etc.

• Numbers seem *closer to 0.5* than with previous example.
Overall picture of all 1000 $\hat{p}$ values:

Figure 2: Plot of pattern of $\hat{p}$ values from 1000 samples when $n = 1504$. 
Clicker Quiz 2

Which method appears preferable, taking a SRS of 100 adults, or a SRS of 1504 adults?

A. \( n = 100 \), because 100 is a round number.

B. \( n = 1504 \), because most of the \( \hat{p} \) values are near the true \( p \).

C. \( n = 100 \), because the \( \hat{p} \) values might be farther from the true \( p \).

D. \( n = 1504 \), because it costs less money to survey more people.
Bias and Variability

- Note in each case, the estimates (the \( \hat{p} \) values) were centered around the true parameter value, 0.5 (no systematic overestimation nor underestimation)

- Conclusion: \( \hat{p} \) is an unbiased estimator of \( p \).

- **Bias:** When a statistic systematically overestimates or systematically underestimates the parameter we are trying to estimate.
• Note also: The \( \hat{p} \) values tended to be spread out farther around 0.5 in the first case \((n = 100)\) than in the second case \((n = 1504)\).

• The sample proportion \( \hat{p} \) has more sampling variability when we take a small sample than when we take a large sample.

• Variability: Measures how spread out the statistic’s values are when we take many samples and calculate the statistic each time.

• In reality, companies only have time to take one sample.

• They want the method to have low variability so they can trust the result they get.
Managing Bias and Variability

- To eliminate bias, use a *simple random sample*.
- To reduce variability, use a *larger sample*.
- It may cost more time & money to do these things rather than to cut corners (convenience sample, small sample size, etc.)
- But if you want *trustworthy results*, your *sampling method* must be a good one.
Margin of Error

• Polls usually report not just an estimate, but a margin of error.

• Example: “53% of adults opposed this law. The margin of error for this poll was plus or minus 2.6 percentage points.”

• What does this mean?
Margin of Error

- Polls usually report not just an estimate, but a *margin of error*.

- Example: “53% of adults opposed this law. The margin of error for this poll was plus or minus 2.6 percentage points.”

- What does this mean?

- Pew believes the *true* proportion of *all* U.S. adults opposing the law is between $0.53 - 0.026 = 0.504$ and $0.53 + 0.026 = 0.556$ (between 50.4% and 55.6%)

- What they don’t say: Pew is “95% confident” in this statement (more later).
Margin of Error: The “One-over-root-\(n\)” Trick

- The margin of error for a sample proportion (with a sample of size \(n\)) is *roughly* \(1/\sqrt{n}\) (assuming 95% confidence).

- Pew research poll example (\(n = 1504\)):

\[
\frac{1}{\sqrt{1504}} = \frac{1}{38.78} \approx 0.026 \text{ (or 2.6%)}
\]

- This rule works for a SRS.

- Note: The *larger* the sample, the *smaller* the margin of error.
Margin of Error: The “One-over-root-$n$” Trick Again

- The margin of error for a sample proportion (with a sample of size $n$) is roughly $1/\sqrt{n}$ (assuming 95% confidence)

- In May 2011, Gallup asked 1018 randomly chosen American adults whether same-sex marriages should be recognized by the law as valid. 53% said ”yes”.

- 2011 Gallup poll example ($n = 1018$):

$$\frac{1}{\sqrt{1018}} = \frac{1}{31.91} \approx 0.03 \text{ (or 3%)}$$

- Note: For this somewhat smaller sample, there is a larger margin of error.
Clicker Quiz 3

We calculate a sample proportion based on a sample of 64 people. Our margin of error (assuming 95% confidence) is roughly:

A. $\frac{1}{8} = 0.125$ (or 12.5%)

B. $\frac{1}{64} = 0.016$ (or 1.6%)

C. 8%

D. 64%
What does “95% confidence mean?”

• Suppose we estimate a proportion, with a margin of error of 2 percentage points.

• This means that in 95% of possible samples, our sample proportion will be within 2 percentage points of the true proportion.

• That is, our method “works” 95% of the time.

• However, we don’t know whether the one sample we did take is one of the “lucky 95%” or one of the “unlucky 5%”.
Confidence Statements

- **Confidence statement**: Contains both a *margin of error* and a *level of confidence*.

- Example: “With 95% confidence, the true proportion of U.S. adults opposing the ACA is between 0.504 and 0.556.”

- Confidence statement is always a statement about the *population*.

- Almost all sample surveys use 95% confidence, but other confidence levels could be used.
Clicker Quiz 4

Consider this confidence statement from Nov. 1, 2012: “With 95% confidence, we conclude that between 46% and 54% of all North Carolina voters will vote for Mitt Romney in the 2012 presidential election.” What is the margin of error associated with the estimated proportion?

A. 0.54 (or 54%)
B. 0.46 (or 46%)
C. 0.04 (or 4%)
D. 0.02 (or 2%)
Sampling from Large Populations

- The size of the population doesn’t make a difference concerning the variability of a statistic.
- Only important thing is that the population is at least 100 times larger than the sample.
- So will a sample of 100 USC students give as much precision as a sample of 1504 U.S. adults?
- No – the sample size itself is the important thing, not the fraction of the population size that the sample makes up.
- Note also: a large sample size reduces variability, but it doesn’t reduce bias.
- A large volunteer sample is still a biased sample.