Chapter 5: Foundations for inference

Yonghyun Kwon
Department of Mathematics, Korea Military Academy

Point estimates and sampling variability

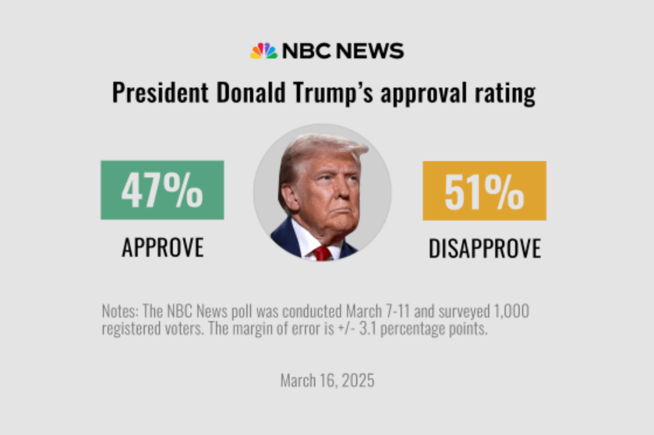


Image source: https://www.nbcnews.com/politics/trump-administration/poll-trump-faces-early-challenges-economy-united-gop-backs-big-change-rcna195860

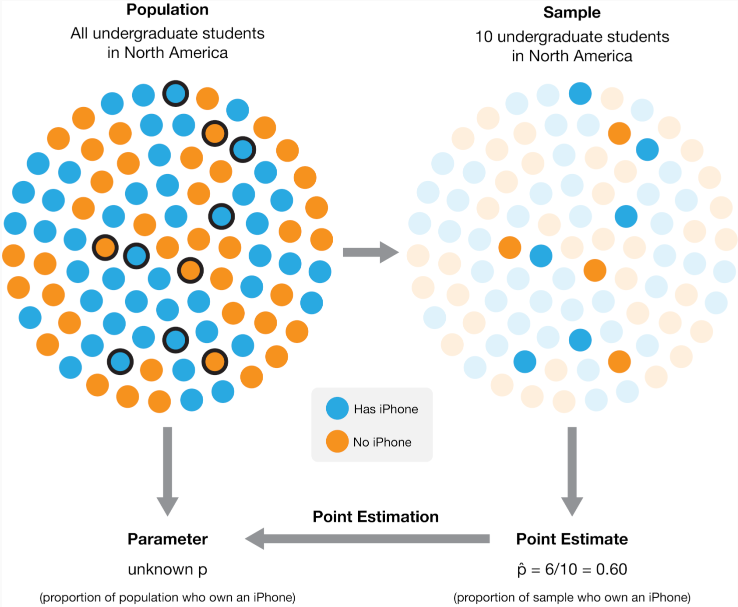
Point Estimates, Sampling Error, and Bias

Point Estimate and Parameter

- The poll suggests the US President's approval rating is **47%.**
- The *sample proportion* $\hat{p}$ is a *point estimate* of the true approval rating (p).
- The (true) *population proportion* p is a *parameter*, which remains unknown in general.
- *Sample size* n: the number of observations in a sample.

Sources of Error

- *Sampling Error* describes how much an estimate will tend to vary from one sample to the next.
- *Bias* is systematic tendency to over- or under-estimate the true population parameter p.



https://datasciencebook.ca/img/inference/population_vs_sample.png

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. Is a randomly selected American adult more or less likely to support the expansion of solar energy?

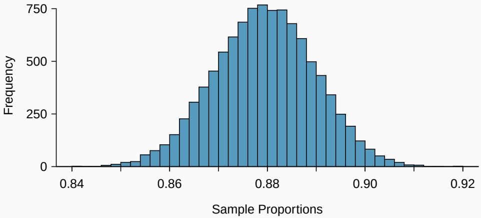
If you don't have access to the full population, you can estimate the proportion of American adults who support solar power expansion by sampling from the population and using the sample proportion as your best guess.

- Randomly sample 1000 American adults and record whether they support or not solar power expansion.
- Find the sample proportion.

```
# 1. Create a set of 250 million entries, where 88\% of
# them are "support" and 12\% are "not".

pop_size <- 250000000
possible_entries <- c(rep("support", 0.88 * pop_size),
                      rep("not", 0.12 * pop_size))


# 2. Sample 1000 entries without replacement.

sampled_entries <- sample(possible_entries, size = 1000)

# 3. Compute p-hat: count the number that are "support",
# then divide by # the sample size.

sum(sampled_entries == "support") / 1000
```

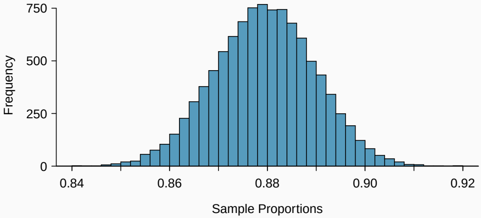
Sampling distribution

Suppose you were to repeat this process many times and obtain many $\hat{p}$s. This distribution is called a *sampling distribution*.



- The mean of the sampling distribution is 0.88.
- The standard deviation of the sampling distribution is 0.01.
  - The *standard error* of the sample proportion, $SE(\hat{p})$, is 0.01.

What is the shape of this distribution? Based on this distribution, what do you think is the true population proportion?



Sampling distributions are never observed

- In real-world applications, we never actually observe the sampling distribution, yet it is useful to always think of a point estimate as coming from such a hypothetical distribution.
- Understanding the sampling distribution will help us characterize and make sense of the point estimates that we do observe.

Central Limit Theorem(CLT) for the sample proportion

When observations are independent, and the sample size n is sufficiently large, sample proportion $\hat{p}$ will be *nearly normally* distributed with mean p, and variance $\frac{p(1-p)}{n}$.

$$\hat{p} \dot{\sim} N\left(\mu = p, \sigma^2 = \frac{p(1-p)}{n}\right)$$

Note: Recall that $\hat{p} = \frac{\text{number of successes}}{\text{number of trials}} = \frac{X}{n}$, *where* $X \sim B(n, p)$. *Therefore,*

$$E(\hat{p}) = \frac{E(X)}{n} = \frac{np}{n} = p,$$

$$Var(\hat{p}) = \frac{Var(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

CLT - conditions

Certain conditions must be met for the CLT to apply:

1. *Independence:* Sampled observations must be independent. This is difficult to verify but is more likely if
   - random sampling/assignment is used.
2. *Success-failure condition:* There should be at least 10 expected(observed) successes and 10 expected(observed) failures in the sample.

Practice

Earlier we estimated the mean and standard error of $\hat{p}$ using simulated data when $p = 0.88$ and $n = 1000$.

- Verify whether the Central Limit Theorem can be applied.
- Find the mean and standard error of $\hat{p}$.
- Estimate how frequently $\hat{p}$ should be within 0.02 of the population value, $p = 0.88$.

When p is unknown

- **Standard error** of the sample proportion is $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$, which is typically unknown.
- In these cases we substitute $\hat{p}$ for p: $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

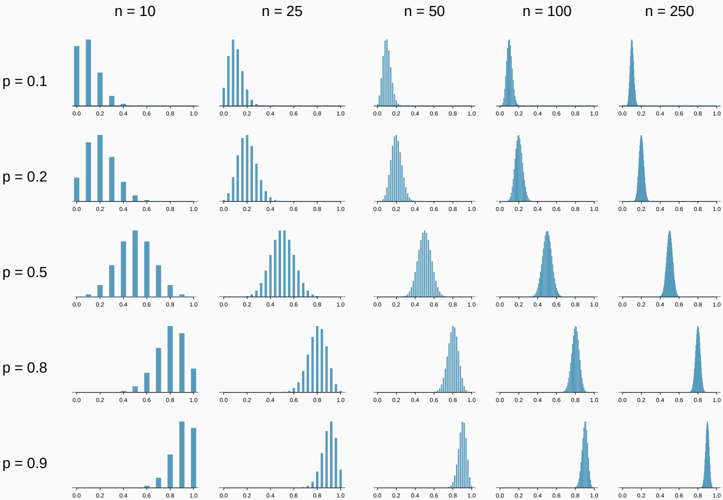
Practice

In a random sample of 1000 adults in South Korea, 587 use Samsung Galaxy.

- Find the sample proportion $\hat{p}$.
- Confirm whether a sampling distribution of $\hat{p}$ is nearly normal.

- Compute the standard error of $\hat{p}$.

What happens when *np* and/or $n(1-p) < 10$?



Extending the framework for other statistics

- The strategy of using a sample statistic to estimate a parameter is quite common, and it's a strategy that we can apply to other statistics besides a proportion.
- The principles and general ideas from this chapter apply to other parameters as well, even if the details change a little.

Note: There can be two or more estimators $\hat{\theta}_1, \hat{\theta}_2, \cdots$ of an unknown parameter θ.

- $\hat{\theta}_1$ *is an unbiased estimator of a parameter* θ *if* $E(\hat{\theta}_1) = \theta$.
- $\hat{\theta}_1$ *is more efficient than* $\hat{\theta}_2$ *if* $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$.

Confidence intervals for a proportion

Point Estimate vs. Confidence Interval

Point Estimate

- is like spearfishing.
- is likely to miss.

http://www.flickr.com/photos/fischerfotos/7439791462

Confidence Interval

- is like using a net.
- has a better chance of capturing the true parameter.

http://www.flickr.com/photos/fischerfotos/7439791462

95% Confidence Interval for a parameter

When the distribution of a point estimate qualifies for the CLT, we can construct a 95% confidence interval as

$$\text{point estimate} \pm 1.96 \times SE.$$

If the CLT conditions are satisfied, the 95% confidence interval of the population proportion p is

$$\hat{p} \pm 1.96 \times$$

Formally speaking, we are *95% confident* that the confidence interval captures the population proportion.

In slide 13, we confirmed that the sample proportion of Samsung Galaxy users, $\hat{p}$, nearly follows a normal distribution and has a standard error of $SE(\hat{p}) = 0.016$.

- Construct and interpret a 95% confidence interval for the population proportion p. Recall that

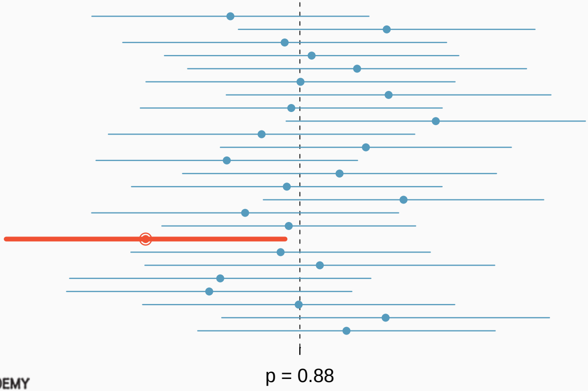
$$\hat{p} = 0.587, \qquad n = 1000.$$

The approximate 95% confidence interval is

$$\begin{aligned} \hat{p} \pm 1.96 \times SE(\hat{p}) &= \qquad \pm \\ &= (\qquad\qquad , \qquad\qquad) \end{aligned}$$

What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation $point\ estimate \pm 1.96 \times SE$.
- Then about 95% of those intervals would contain the true population proportion (p).



Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Can you see any drawbacks to using a wider interval?

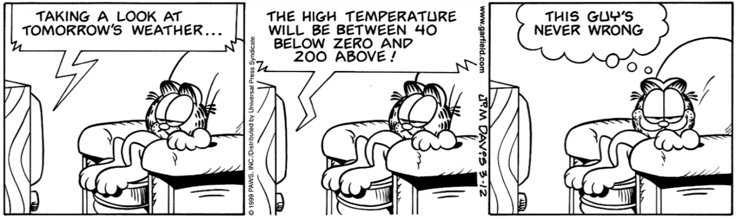


Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Confidence interval using any confidence level

If a point estimate closely follows a normal model, then a $100(1-\alpha)\%$ confidence interval for the population parameter is

$$\text{point estimate} \pm z_{\alpha/2} \times SE$$

where $z_{\alpha/2}$ is the $\alpha/2$-th upper quantile of the standard normal distribution.

- $100(1-\alpha)\%$ is called the *confidence level*.
- $z_{\alpha/2}$ is called the *critical value*.
- $z_{\alpha/2} \times SE$ is called the *margin of error*.
- Commonly used confidence levels in practice are 90%, 95%, and 99%.
- For a 95% confidence interval, $z_{0.025} = 1.96$.

Practice

In October 2014, a New York doctor was diagnosed with Ebola after treating patients in Guinea. A poll of 1,042 New York adults found that 82% supported a mandatory 21-day quarantine for those exposed.

- Construct a 90% confidence interval for p, the proportion who supported the quarantine(use $z_{0.1} = 1.28$ or $z_{0.05} = 1.65$).

- Interpret the confidence interval.

Which of the following is the correct interpretation of this confidence interval?

We are 90% confident that

(a) 80% to 84% of New York adults in this sample supported a mandatory 21-day quarantine for those exposed to Ebola patients.

(b) 80% to 84% of all New York adults supported a mandatory 21-day quarantine for those exposed to Ebola patients.

(c) There is an 80% to 84% chance that a randomly chosen New York adult supported the quarantine.

(d) There is an 80% to 84% chance that 90% of New York adults supported the quarantine.

Hypothesis testing for a proportion

A Thought-Provoking Question

How many of the world's 1-year-old children today have been vaccinated against some disease?

a. 20%
b. 50%
c. 80%

https://www.unicef.org/immunization

Hypothesis Testing Framework

We want to understand how much people know about world health. We consider the following hypotheses:

H_0: People answer at random with no real knowledge.

H_A: People have some knowledge (or false knowledge) affecting their answers.

- *Null Hypothesis* (H_0): The default skeptical assumption.
- *Alternative Hypothesis* (H_A): The new claim being tested.

Data scientists require *strong evidence* to reject H_0.

Applying the Framework

Let p be the proportion of correct responses for the question on infant vaccinations:

- H_0: $p = 1/3$ (random guessing).
- H_A: $p \neq 1/3$ (people do better or worse than guessing).

Null value is the value we are comparing the parameter to.

- In this case, the null value is $p_0 =$

A Legal Analogy

The court system follows a similar framework:

$$H_0 : \text{The defendant is innocent.}$$
$$H_A : \text{The defendant is guilty.}$$

- The jury requires **strong evidence** to reject H_0 and convict.
- Failing to reject H_0 does not prove innocence, just as failing to reject H_0 in a hypothesis test does not confirm it is true.

Decision errors

We make a decision about whether the null or the alternative hypothesis is true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 error*
	H_A true	*Type 2 error*	✓

- *Type 1 error rate*: $P(\text{reject } H_0 \mid H_0 \text{ is true})$
- *Type 2 error rate*: $P(\text{fail to reject } H_0 \mid H_A \text{ is true})$
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

A Legal Analogy(cont.)

$$H_0 : \text{The defendant is innocent.}$$
$$H_A : \text{The defendant is guilty.}$$

Which type of error is being committed in the following circumstances?

- Convicting an innocent person:
- Letting a guilty person go free:

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer"

– William Blackstone

Balancing Type 1 and Type 2 Errors

Reducing one type of error often increases the other:

- Lowering the Type 1 error rate increases Type 2 error rate.
- Lowering the Type 2 error rate increases Type 1 error rate.

As a general rule, we control a *Type 1 error rate* to be less than or equal to $\alpha = 0.05$.

- The threshold of Type 1 error rate is a *significance level,* α.

$$\text{Type 1 error rate} \leq \alpha$$

Understanding p-values

P-value

The *p-value* is the probability of observing data at least as favorable to H_1 as our current dataset, *assuming H_0 is true.*

A smaller p-value indicates stronger evidence against H_0.

- If the p-value is less than the significance level, α,
  - we (reject / don't reject) H_0.
- If the p-value is greater than the significance level, α,
  - we (reject / don't reject) H_0.

Hypothesis Testing Example

Pew Research surveyed 1000 Americans on coal energy usage. We would like to set up hypotheses to evaluate whether a majority of American adults support or oppose the increased usage of coal.

H_0 : H_A :

Pew Research's sample shows that $\hat{p} = 0.37$ of American adults support increased usage of coal.

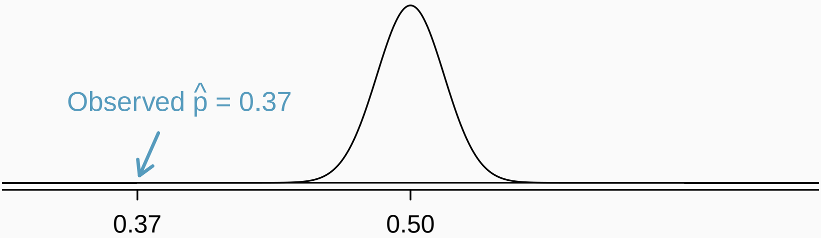
Recall: Central Limit Theorem(CLT)

Under the independence and the success-failure condition,

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

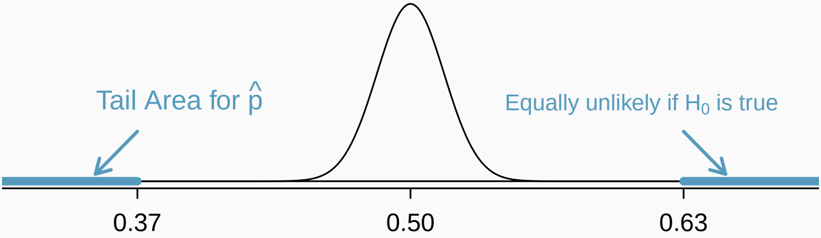
Hypothesis Testing Example(cont.)

The sampling distribution under the null hypothesis(H_0) is the *null distribution*.



If H_0 is true, $p = 0.5$ and $\hat{p} \dot\sim N\left(0.5, \frac{0.5\ (1-0.5)}{1000}\right) = N(0.5, 0.016^2)$.

The p-value is the probability of obtaining $\hat{p} = 0.37$ or more extreme values under H_0.



$$\begin{aligned} \text{p-value} &= 2 \times P(\hat{p} < 0.37 \mid H_0 \text{ is true}) \\ &= 2P\left(\frac{\hat{p} - 0.5}{0.016} < \frac{0.37 - 0.5}{0.016}\right) \\ &= 2P(Z < -8.125) = \end{aligned}$$

```
> pnorm(q = -8.125, lower.tail = TRUE)
[1] 2.236812e-16
```

Hypothesis Testing Example(cont.)

In this example, $Z = \frac{\hat{p} - p_0}{SE(\hat{p})} = \frac{\hat{p} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{1000}}}$ is the *test statistic*.

- The observed test statistic is $z = \frac{0.37 - 0.5}{0.016} = -8.125$.

How should we evaluate the hypotheses using the p-value of 4.4×10^{-16}? Use the standard significance level of $\alpha = 0.05$.

Since the p-value($= 4.4 \times 10^{-16}$) is smaller than the significance level($\alpha = 0.05$), we (reject / don't reject) H_0.

- We say that the difference between the population proportion and the null value is *statistically significant*.

Summary

Z-test for population proportion p

To assess if population proportion p differs significantly from the null value p_0, we consider the following hypothesis test.

$$H_0 : p = p_0, \quad H_A : p \neq p_0$$

Under the independence and success-failure conditions, the null distribution of test statistic is either

$$\hat{p} \overset{H_0}{\dot{\sim}} N\left(p_0, \frac{p_0(1-p_0)}{n}\right), \text{ or } Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0}{\dot{\sim}} N(0,1).$$

The p-value is based on the both-side tails, as larger or smaller Z values suggest stronger evidence against the null hypothesis.

Practice

A random sample of 1028 US adults shows that 62% support nuclear arms reduction. Does this provide convincing evidence that the proportion of Americans supporting nuclear arms reduction is different from 60% at the 5% significance level?

1. Specify the null and alternative hypothesis.

2. Find the null distribution of the test statistic.

Practice

A random sample of 1028 US adults shows that 62% support nuclear arms reduction. Does this provide convincing evidence that the proportion of Americans supporting nuclear arms reduction is different from 60% at the 5% significance level?

3. Compute the observed test statistic.

4. Compute the p-value and complete the hypothesis test.

One vs. two sided hypothesis tests

- In two sided hypothesis tests we are interested in whether p is either above or below some null value p_0:

$$H_0 : p = p_0, \qquad H_A : p \neq p_0$$

- In one sided hypothesis tests we are interested in p differing from the null value p_0 in one direction (and not the other):
  - If there is only value in detecting if population parameter is less than p_0:

$$H_0 : p = p_0, \qquad H_A : p < p_0$$

  - If there is only value in detecting if population parameter is greater than p_0:

$$H_0 : p = p_0, \qquad H_A : p > p_0$$

Exercises

Exercises in OpenIntro Statistics 4th ed.

- Confidence intervals for a proportion: Exercise 5.13, 5.14
- Hypothesis testing for a proportion: Exercise 5.22, 5.32 (Follow the steps in slide 37-38).