Chapter 6: Foundations for inference

Yonghyun Kwon
Department of Mathematics, Korea Military Academy

Point estimates and sampling variability

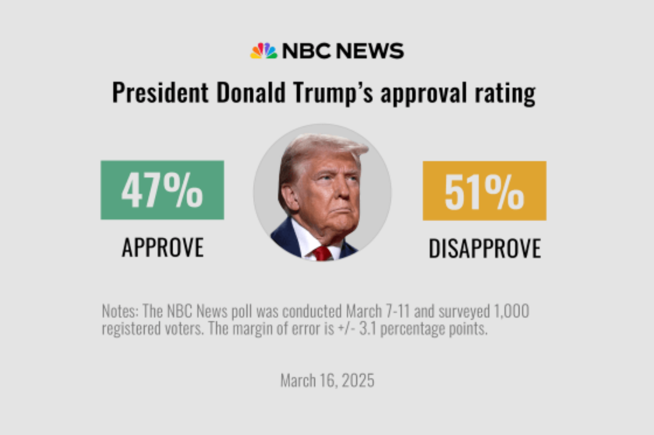


Image source: https://www.nbcnews.com/politics/trump-administration/poll-trump-faces-early-challenges-economy-united-gop-backs-big-change-rcna195860

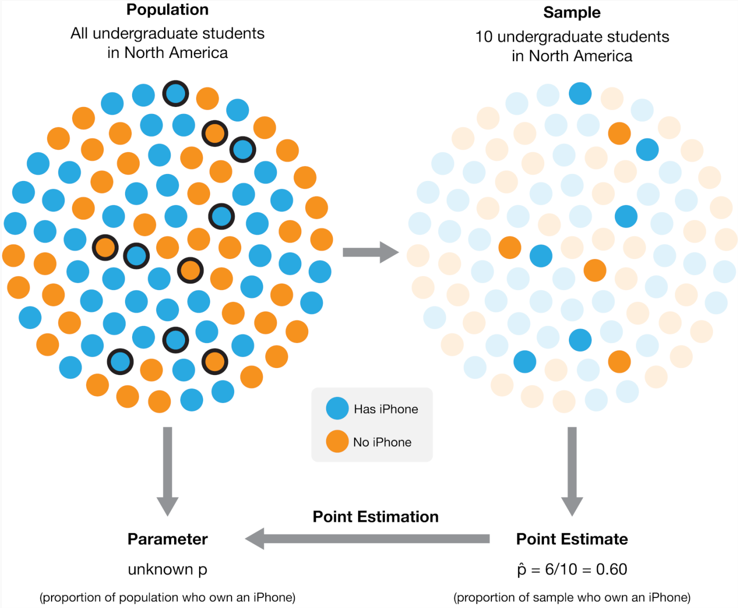
Point Estimates, Sampling Error, and Bias

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 : point estimate − parameter : $\hat{p} - p$

- *Sampling Error* describes how much an estimate will tend to vary from one sample to the next. $\hat{p} - E(\hat{p})$
- *Bias* is systematic tendency to over- or under-estimate the true population parameter p. $E(\hat{p}) - p \quad (=0)$



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.

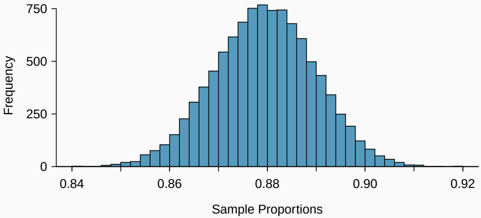
- Randomly sample 1000 American adults and record whether they support or not solar power expansion.
- Find the sample proportion.

$$\hat{p} = \frac{\#\text{ of success}}{\#\text{ of trials}} = \frac{X}{n}$$

```r
# 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

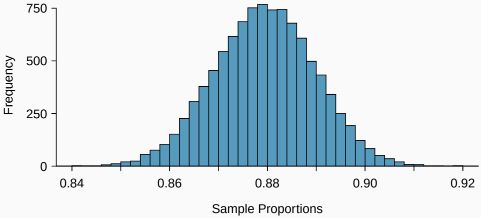
Suppose you were to repeat this process many times and obtain many $\hat{p}$s. This distribution is called a *sampling distribution*.



- 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.

: standard deviation of point estimate

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



① Bell curved ⇒ nearly normal

② $p \approx 0.88$

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 $\dfrac{p(1-p)}{n}$.

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

Note: Recall that $\hat{p} = \dfrac{\text{number of successes}}{\text{number of trials}} = \dfrac{X}{n}$*, 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.

$np \geq 10$

$n(1-p) \geq 10$

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.

  ① Independance : ok

  ② $np = 880 \geq 10$ $n(1-p) = 120 \geq 10$: ok

- Find the mean and standard error of $\hat{p}$.

  $E(\hat{p}) = p = 0.88$

  $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.88 \times 0.12}{1000}} \approx 0.01$

- Estimate how frequently $\hat{p}$ should be within 0.02 of the population value, $p = 0.88$.

  $\hat{p} \sim N(0.88, 0.01^2)$

  $P(0.88 - 0.02 \leq \hat{p} \leq 0.88 + 0.02)$

  $= P\left(-\frac{0.02}{0.01} \leq \frac{\hat{p} - 0.88}{0.01} \leq \frac{0.02}{0.01}\right) = P(-2 \leq z \leq 2)$

  $= 0.9545$

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

Success

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

- Find the sample proportion $\hat{p}$. $= \frac{X}{n} = \frac{587}{1000} = 0.587$
- Confirm whether a sampling distribution of $\hat{p}$ is nearly normal.

① Independence : ok

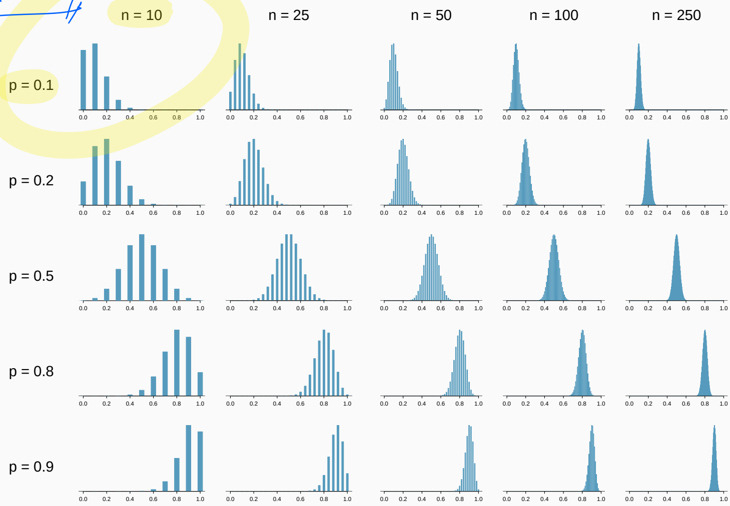
② $n\hat{p} = 587 \geq 10$ $n(1-\hat{p}) = 413 \geq 10$: ok

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

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.587 \times 0.413}{1000}} = 0.01$$

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

$np = 1 < 10$



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 $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$.*

$\hat{\theta}_1$ is a consistent estimator of θ

if $\lim_{n\to\infty} E(\hat{\theta}_1) = \theta$, $\lim_{n\to\infty} Var(\hat{\theta}_1) = 0$

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

$$\text{point estimate} \pm \underbrace{1.96}_{= z_{0.025}} \times SE.$$

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

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

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 $SE(\hat{p}) = 0.016$. $= \sqrt{\frac{0.587 \times (1-0.587)}{1000}}$

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

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

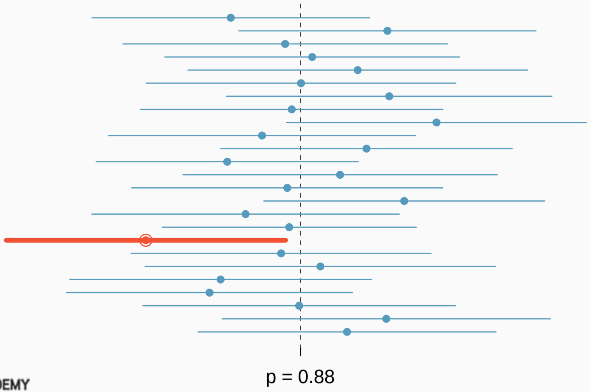
The approximate 95% confidence interval is

$$\begin{aligned} \hat{p} \pm 1.96 \times SE(\hat{p}) &= 0.587 \pm 1.96 \times 0.016 \\ &= (\ 0.556\ ,\ 0.618\ ) \end{aligned}$$

KOREA MILITARY ACADEMY

What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation $\textit{point estimate} \pm 1.96 \times SE$.
- 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?

Wider interval

Can you see any drawbacks to using a wider interval?

If the interval is too wide, it may not be very informative

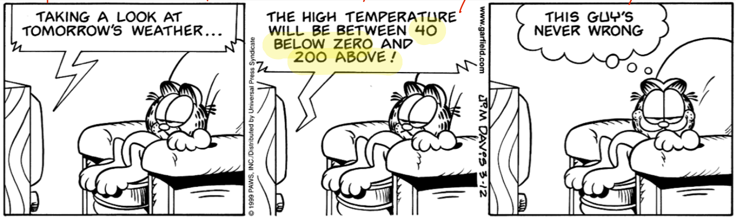


Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

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

= confidence level

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

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

- $100(1-\alpha)\%$ 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

n

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.

$= \hat{p}$

$1-\alpha = 0.9$, $\alpha = 0.1$

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

$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.82 \pm 1.65\sqrt{\frac{0.82(1-0.82)}{1042}}$

$= (0.80, 0.84)$

- Interpret the confidence interval.

We are 90% confident that 80% to 84% of all New York adults supported the quarantine

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

We are 90% confident that

$\Rightarrow$ population

(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. O

(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.

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).