STA258H5
University of Toronto Mississauga

Al Nosedal and Omid Jazi

Winter 2023

ONE SAMPLE CONFIDENCE INTERVALS ON A PROPORTION

(Not covered in test 1)

Large-Sample Confidence interval for μ

Parameter : μ.

σ known

Confidence interval :

$$\bar{Y} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

Valid if

1. Random sample
2. Independent and identically distributed observations
3. n is large enough for CLT to apply

CI's on Proportions (% or Fractions < 1)

Parameter of interest

p : population proportion

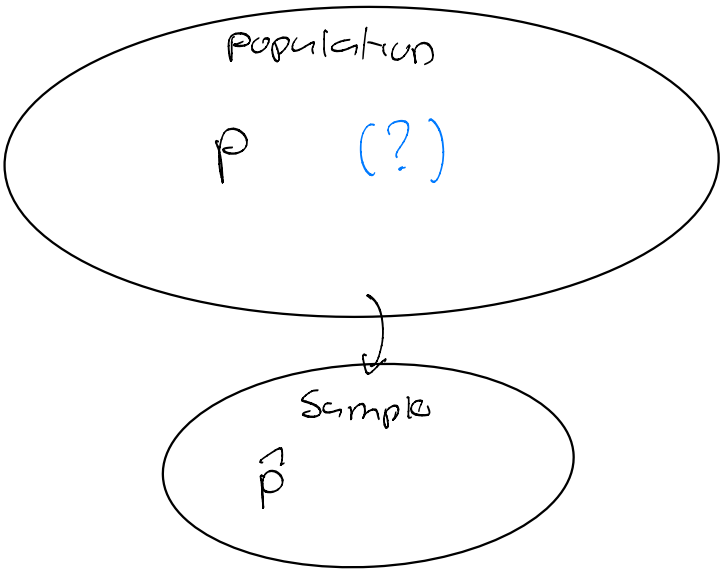
Statistic used in constructing CI's

$\hat{p}$: sample proportion

$$\hat{p} = \frac{\#\text{ observations satisfying criteria}}{\text{sample size}} = \frac{X}{n}$$

(X: # successes)

In questions $\hat{p}$ might be given directly or indirectly



CI on Proportions

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ← standard error

$z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ → margin of error

require: n large: $np \geq 10$ $n(1-p) \geq 10$

Interval Estimate of p

Draw a simple random sample of size n from a population with unknown proportion p of successes. An (approximate) confidence interval for p is:

$$\hat{p} \pm z_* \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

where z_* is a number coming from the Standard Normal that depends on the confidence level required.
Use this interval only when:

1. random sample
2. independent and identically distributed Bernoulli trials
3. n is "large"
   - $np \geq 10$
   - $n(1-p) \geq 10$

Problem

The utility of mobile devices raises new questions about the intrusion of work into personal life. In a recent survey by CareerJournal.com, 158 of 473 employees responded that they typically took work with them on vacation.

a. What is the point estimate of the population proportion of employees who typically take work with them on vacation?

b. At 90% confidence, what is the margin of error?

c. What is the 90% confidence interval for the population proportion of employees who typically take work with them on vacation?

Example (Slide 5)

a) point estimate (one value, best single estimate)

↳ $\hat{p} = \frac{158}{473}$ (sample prop $\hat{p}$ is a point estim of pop prop p)

$n = 473$

c) CI's on proportion

$\hat{p} = \frac{158}{473}$, $n = 473$, $z_{\alpha/2} = 1.645$ (or 1.64 or 1.65) for 90% CI

(See example from Monday)

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= \frac{158}{473} \pm 1.645 \sqrt{\frac{(158/473)(1 - 158/473)}{473}}$$

$$= \frac{158}{473} \pm 0.0357 = (0.2984,\ 0.3697)$$

Interp:

We are 90% confident the ~~mean~~ proportion of people who take work on vacation is between 0.2984 and 0.3697.

Aside!

Claim A: 1/3 (0.33) of people take work with them on vacation

Claim B: 1/4 (0.25) " " " " " " "

use 90% CI to examine claims

90% CI supports Claim A since 1/3 (0.33) lies inside interval.

Solution

a. $\hat{p} = \frac{158}{473} = 0.334$

b. Margin of error $=$

$z_* \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = 1.65\sqrt{\frac{(0.3340)(0.6660)}{473}} = 1.65(0.0217) = 0.035805$

c. $\hat{p} \pm z_* \left(\sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} \right)$

0.334 ± 0.0358

$(0.2982, 0.3698)$

R Code

```
prop.test(158,473,conf.level=0.90,correct=FALSE);

# correct = FALSE;
# this is telling R NOT to use
# Yates' continuity correction;
```

R Code

```
##
##  1-sample proportions test without continuity
##  correction
##
## data:  158 out of 473, null probability 0.5
## X-squared = 52.112, df = 1, p-value =
## 5.242e-13
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.2993997 0.3705642
## sample estimates:
##         p
## 0.3340381
```

Nielsen Ratings

Statistical techniques play a vital role in helping advertisers determine how many viewers watch the shows that they sponsor. There are several companies that sample television viewers to determine what shows they watch, the best known of which is the A. C. Nielsen firm. The Nielsen Ratings are based on a sample of randomly selected families. A device attached to the family television keeps track of the channels the television receives. The ratings then produce the proportions of each show from which sponsors can determine the number of viewers and the potential value of any commercials.

Nielsen Ratings

The results for the 18-to 49-year-old group on Thursday, March 7, 2013, for the time slot 8:00 p.m. to 8:30 p.m. have been recorded using the following codes:

Network	Show	Code
ABC	Shark Tank	1
CBS	Big Bang Theory	2
CW	The Vampire Diaries	3
Fox	American Idol	4
NBC	Community	5
Television turned off		6

CBS would like to use the data to estimate how many Americans aged 18 to 49 were tuned to its program *Big Bang Theory*.

R Code

```r
#Step 1. Entering data;

# importing data;

# url of ratings;
url="https://mcs.utm.utoronto.ca/~nosedal/data/rating.txt"

ratings_data= read.table(url,header=TRUE);

names(ratings_data);

# first 6 observations from file
ratings_data[1:6, ]
```

R Code

```
## [1] "ViewerNumber" "TV.Program"
##   ViewerNumber TV.Program
## 1            1          6
## 2            2          6
## 3            3          6
## 4            4          6
## 5            5          6
## 6            6          6
```

R Code

```
all.programs=ratings_data$TV.Program;

# I want you to see the first 6 observations;

all.programs[1:6];
```

R Code

```
## [1] 6 6 6 6 6 6
```

R Code

```
# Recall that Big Bang Theory's code is 2;

big.bang=all.programs[all.programs==2];

# First 6 observations from big.bang

big.bang[1:6]
```

R Code

```
## [1] 2 2 2 2 2 2
```

R Code

```
## CI for p;

sample.size=length(all.programs);

sample.size;

successes=length(big.bang);

successes;

prop.test(successes,sample.size,conf.level=0.95,
correct=FALSE);
```

R Code

```
## [1] 5000
## [1] 275
##
##  1-sample proportions test without continuity
##  correction
##
## data:  successes out of sample.size, null probability 0.5
## X-squared = 3960.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.04901565 0.06166761
## sample estimates:
##     p
## 0.055
```

Assumptions and Conditions

Here are the assumptions and the corresponding conditions to check before creating a confidence interval about a proportion.

- Independence assumption.
- Sample size assumption

Independence Assumption

You first need to think about whether the independence assumption is plausible. You can look for reasons to suspect that it fails. You might wonder whether there is any reason to believe that the data values somehow affect each other. This condition depends on your knowledge of the situation. It's not one you can check by looking at the data. However, there are two conditions that you can check:

- Randomization Condition: Were the data sampled at random or generated from a properly randomized experiment?
- 10% Condition: If the sample exceeds 10% of the population, the probability of a success changes so much during the sampling that a Normal model may no longer be appropriate. But if less that 10% of the population is sampled, it is safe to assume to have independence.

Sample Size Assumption

The model we use for inference is based on the Central Limit Theorem. So, the sample must be large enough for the Normal model to be appropriate. This requirement is easy to check with the following condition:

- Success/Failure Condition: We must expect our sample to contain at least 10 "successes" and at least 10 "failures". So we check that $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.