STA215
Confidence Intervals for Proportions

Al Nosedal.
University of Toronto.
Summer 2017

June 15, 2019

"Pepsi" problem

A market research consultant hired by the Pepsi-Cola Co. is interested in determining the proportion of UTM students who favor Pepsi-Cola over Coke Classic. A random sample of 100 students shows that 40 students favor Pepsi over Coke. Use this information to construct a 95% confidence interval for the proportion of all students in this market who prefer Pepsi.

Bernoulli Distribution

$$x_i = \begin{cases} 1 & \text{i-th person prefers Pepsi} \\ 0 & \text{i-th person prefers Coke} \end{cases}$$

$\mu = E(x_i) = p$
$\sigma^2 = V(x_i) = p(1-p)$
Let $\hat{p}$ be our estimate of p. Note that $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$. If n is "large", by the Central Limit Theorem, we know that:
$\bar{x}$ is roughly $N(\mu, \frac{\sigma}{\sqrt{n}})$, that is,
$\hat{p}$ is roughly $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

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) n is "large" and
2) $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.

Problem

A simple random sample of 400 individuals provides 100 Yes responses.

a. What is the point estimate of the proportion of the population that would provide Yes responses?

b. What is the point estimate of the standard error of the proportion, $\sigma_{\hat{p}}$?

c. Compute the 95% confidence interval for the population proportion.

Solution

a. $\hat{p} = \frac{100}{400} = 0.25$

b. Standard error of $\hat{p} = \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = \sqrt{\frac{(0.25)(0.75)}{400}} = 0.0216$

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

$0.25 \pm 1.96(0.0216)$

$(0.2076, 0.2923)$

Problem

A simple random sample of 800 elements generates a sample proportion $\hat{p} = 0.70$.

a. Provide a 90% confidence interval for the population proportion.

b. Provide a 95% confidence interval for the population proportion.

Solution

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

$0.70 \pm 1.65 \left(\sqrt{\frac{(0.70)(1-0.70)}{800}} \right)$

$0.70 \pm 1.65(0.0162)$

$(0.6732, 0.7267)$

b. $0.70 \pm 1.96(0.0162)$

$(0.6682, 0.7317)$

HW?

A national health organization warns that 30% of middle school students nationwide have been drunk. Concerned, a local health agency randomly and anonymously surveys 110 of the 1212 middle school students in its city. Only 21 of them report having been drunk.

a) What proportion of the sample reported having been drunk?

b) Create a 95% confidence interval for the proportion of the city's middle school students who have been drunk.

c) Is there any reason to believe that the national level of 30% is not true of the middle school students in this city?

Solutions

a) $\hat{p} = \frac{21}{110} = 0.1909$
b) We are 95% confident that the proportion of the city's middle school students who have been drunk lies between 0.1175 and 0.2643.
c) We are 95% confident that between 11.75% and 26.43% of the city's middle school students have been drunk. It appears that this proportion is less than in the entire country (30%), because the interval is completely below 30%.

Conditions

- Independence and Randomization: local health agency drew a random sample from all middle school students in its city. It is unlikely that any respondent influenced another.
- 10% condition: The sample is certainly less that 10% of the population.
- Success/Failure Condition: $n\hat{p} = 21 \geq 10$ and $n(1 - \hat{p}) = 89 \geq 10$, so the sample is large enough.

Example.

A survey of 611 office workers investigated telephone answering practices, including how often each office worker was able to answer incoming telephone calls and how often incoming telephone calls went directly to voice mail. A total of 281 office workers indicated that they never need voice mail and are able to take every telephone call.

a. What is the point estimate of the proportion of the population of office workers who are able to take every telephone call?

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

c. What is the 90% confidence interval for the proportion of the population of office workers who are able to take every telephone call?

Solution

a. $\hat{p} = \frac{281}{611} = 0.46$

b. Margin of error $=$

$z_* \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = 1.65\sqrt{\frac{(0.46)(0.54)}{611}} = 1.65(0.0201) = 0.0332$

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

$0.46 \pm .0332$

$(0.4268, 0.4932)$

R Code

```
prop.test(281,611,conf.level=0.90);

##
##  1-sample proportions test with continuity correction
##
## data:  281 out of 611, null probability 0.5
## X-squared = 3.7709, df = 1, p-value = 0.05215
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.4261763 0.4939896
## sample estimates:
##         p
## 0.4599018
```

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.
## 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
```

Problem

In a survey, the planning value for the population proportion is $p^* = 0.35$. How large a sample should be taken to provide a 95% confidence interval with a margin of error of 0.05?

Solution.
$n = \left(\frac{z_*}{E}\right)^2 p^*(1-p^*) = \left(\frac{1.96}{0.05}\right)^2 (0.35)(1-0.35) = 350$
(Always round up).

Determining the Sample Size

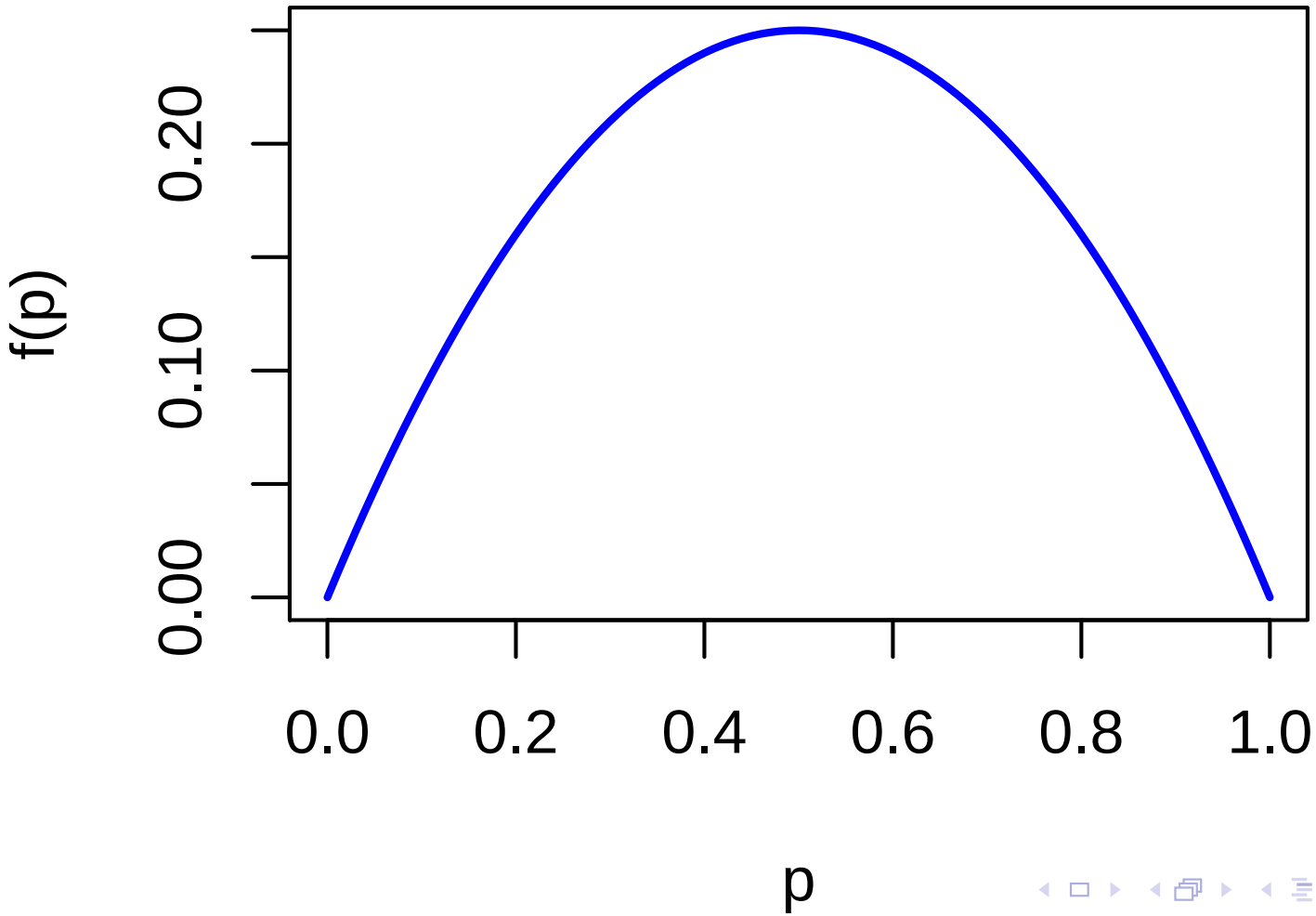
Sample Size for an Interval Estimate of a Population Proportion.

$$n = \left(\frac{z_*}{E}\right)^2 p^*(1 - p^*)$$

In practice, the planning value p^* can be chosen by one of the following procedures.

1. Use the sample proportion from a previous sample of the same or similar units.
2. Use a planning value of $p^* = 0.5$.

$f(p) = p(1-p)$



Problem

At 95% confidence, how large a sample should be taken to obtain a margin of error of 0.03 for the estimation of a population proportion? Assume that past data are not available for developing a planning value for p^*.

Solution.

$n = \left(\frac{z_*}{E}\right)^2 p^*(1-p^*) = \left(\frac{1.96}{0.03}\right)^2 (0.5)(1-0.5) = 1068$

(Always round up).

Problem

The percentage of people not covered by health care insurance in 2003 was 15.6%. A congressional committee has been charged with conducting a sample survey to obtain more current information.

a. What sample size would you recommend if the committee's goal is to estimate the current proportion of individuals without health care insurance with a margin of error of 0.03? Use a 95% confidence level.

b. Repeat part a) using a 99% confidence level.

Solution

a. $n = \left(\frac{z_*}{E}\right)^2 p^*(1-p^*) = \left(\frac{1.96}{0.03}\right)^2 (0.156)(1-0.156) = 563$

b. $n = \left(\frac{z_*}{E}\right)^2 p^*(1-p^*) = \left(\frac{2.58}{0.03}\right)^2 (0.156)(1-0.156) = 974$