

L7

Thursday, October 4, 2018

7:11 AM

<<Lecture7\_Student.pdf>>

# Stat 324 – Introduction to Statistics for Engineers

LECTURE 7:

CONFIDENCE INTERVALS FOR PROPORTIONS (SECTION 10.2 OF OTT AND LONGNECKER)

BOOTSTRAP CONFIDENCE INTERVAL FOR  $\mu$  (SECTION 5.8 OF OTT AND LONGNECKER);

Considering a common Non-Normal Population – 1s and 0s.

Ex 1. An accounting firm has a large list of clients (the population), with an information file on each client. The firm has noticed errors in some files and wishes to know the proportion of files that contain an error. Call the population proportion of files in error  $\pi$ . An SRS of size  $n = 100$  is taken and used to estimate  $\pi$ . Now the firm will decide whether it is worth the cost to examine and fix all the files. Each file sampled was classified as containing an error (call this 1), or not (call this 0). The results are:

Files with an error: 20; files without errors: 80

How can we estimate the proportion of all files in error ( $\pi$ )?

use sample proportion  $\hat{\pi} = \frac{x}{n} = \frac{\text{\# of sample files with error}}{\text{\# of sample files}} = \frac{20}{100} = 0.2$

How good of an estimator is it?

Is it unbiased?

What is its variance?

## Estimating a Population Proportion

Lets check that the sample proportion is an unbiased estimator for the population proportion.

Let each file be an iid  $Y \sim \text{Bern}(\pi)$ ,

pmf:

$Y_i$	$P(Y_i = y_i)$
1	$\pi$
0	$1-\pi$

where  $\pi$  is the probability of success.

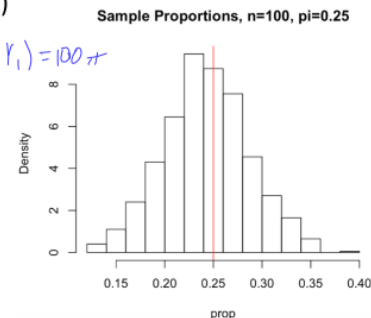
error  
no error

So the collection of 100 files is  $X = Y_1 + Y_2 + \dots + Y_{100} \sim \text{Bin}(100, \pi)$

$$E(X) = E(Y_1 + Y_2 + Y_3 + \dots + Y_{100}) = E(Y_1) + E(Y_2) + \dots + E(Y_{100})$$

$$E(\hat{\pi}) = E\left(\frac{X}{100}\right) = \frac{1}{100} E(X) = \frac{1}{100} \cdot 100\pi = \pi$$

Sample proportion is unbiased estimator of  $\pi$ .



$$1 \cdot \pi + 0 \cdot (1-\pi)$$

$$\approx 100 E(Y_1) = 100\pi$$

## Estimating a Population Proportion

How can we estimate the reliability of our estimate for the proportion?

Find the variability of our estimator: (Assuming  $Y_i \sim \text{Bern}(\pi)$  iid) *only if independent*

$$\begin{aligned} \text{Var}(\hat{\pi}) &= \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(X) = \frac{1}{n^2} \text{Var}(Y_1 + Y_2 + Y_3 + \dots + Y_{100}) \\ &= \frac{1}{n^2} [\text{Var}(Y_1) + \dots + \text{Var}(Y_{100})] \\ &= \frac{1}{n^2} \cdot \text{Var}(Y_1) = \frac{\pi(1-\pi)}{n} \end{aligned}$$

$$\text{So } \text{SE}(\hat{\pi}) = \sqrt{\text{Var}(\hat{\pi})} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

We can get an estimate for  $\text{SE}(\hat{\pi}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$  using our sample data  $\widehat{\text{SE}}(\hat{\pi}) = \sqrt{\frac{0.20(0.80)}{100}} = 0.04$

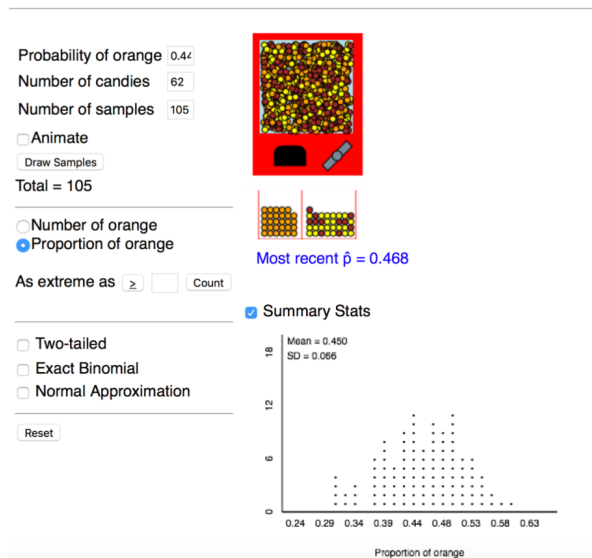
Notice, the precision of our estimator increases (Standard error decreases) as

$n$  gets larger

$\pi$  gets further from 0.50

# Sampling Distribution of Sample Proportion - Simulation

<http://www.rossmanchance.com/applets/OneProp/OneProp.htm?candy=1>



## Estimating a Population Proportion

Ex 1: An SRS of size  $n = 100$  is taken and used to estimate  $\pi$ . Files with error: 20; without errors: 80

What is the distribution of  $\hat{\pi} = \frac{X}{n}$ ?

- Exact distribution is related to binomial – exact CI is complicated
- We can use the CLT for approximate!
  - Since the  $Y_i$  are iid, and  $\hat{\pi} = \frac{X}{n}$  is really just the sample mean of zeros and ones
  - For  $n$  large enough,  $\hat{\pi} = \frac{X}{n} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$  approximately

- An approximate  $100(1 - \alpha)\%$  CI for  $\pi$  is of the form:

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

- For 90% CI our audit data:  $0.20 \pm 1.65 \sqrt{\frac{0.2(1.8)}{100}} = (0.1342, 0.266)$

(assuming 100 is big enough for CLT to kick in)

iid = independent draws



# Estimating a Population Proportion

How big of n is “big enough?” for the CLT to kick in?

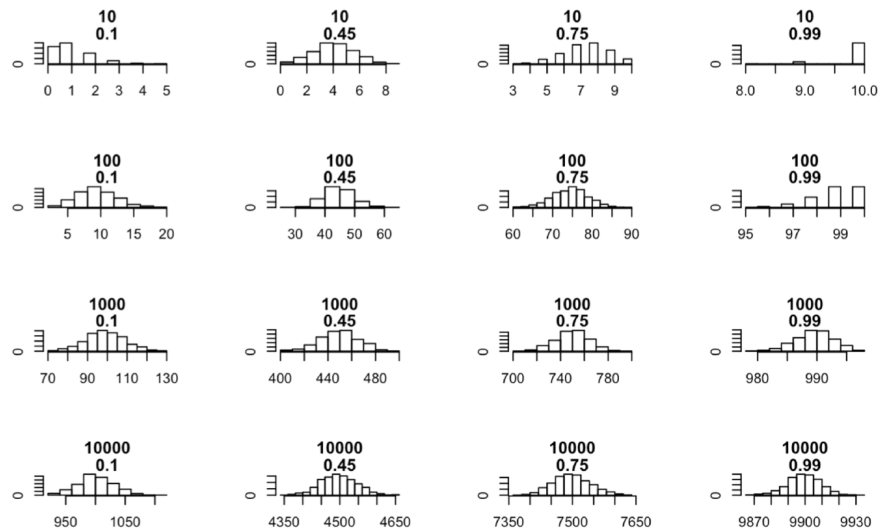
Generally, if

$$n\pi > 5 \text{ and}$$

$$n(1 - \pi) > 5$$

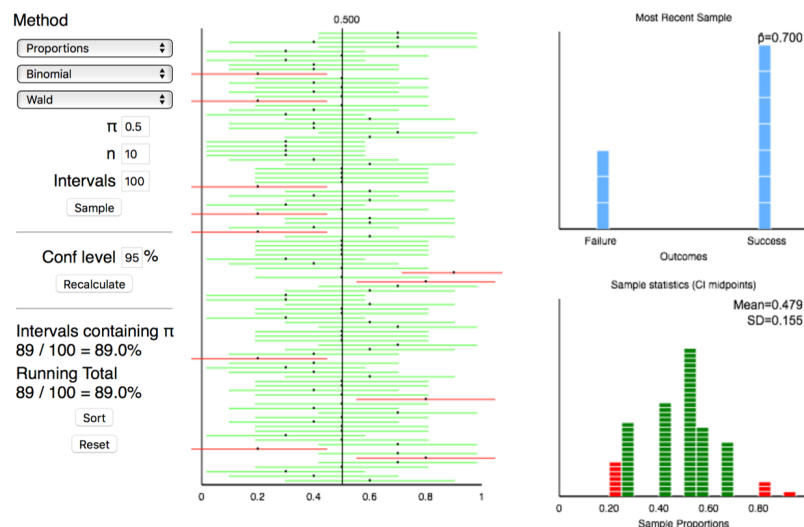
The Normal approximation will be pretty close.

\*I prefer products > 10



Constructing Confidence Intervals for an Unknown proportion of success:  $\pi$

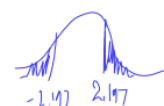
Simulation: <http://www.rossmanchance.com/applets/ConfSim.html>



## Constructing Confidence Intervals for an Unknown proportion of success: $\pi$

Ex 2. A 1993 Los Angeles Times poll of 1703 adults revealed that only 17% thought the media was doing a very good job. With what degree of confidence can the newspaper say that 17% +/- 2% of adults believe the media is doing a very good job?

$$\hat{\pi} = 0.17 \quad 0.02 \text{ ME} \quad 0.02 = Z \sqrt{\frac{0.17(0.83)}{1703}} \quad Z = \frac{0.02}{\sqrt{\frac{0.17(0.83)}{1703}}}$$


$$Z = 2.197$$


$$P(Z < 2.197) = 0.9561 \quad 0.9561 - 0.189 = 0.9722 \approx 97\%$$

$$P(Z < -2.197) = 0.139$$

Ex 3: A politician wants to know what percentage of voters support her position on an issue. What size voter sample should be obtained to determine with 90% confidence the support level within 4%?

90% confidence

$$0.04 = 1.645 \sqrt{\frac{.5(.5)}{n}} \quad \text{Don't know } \pi$$


$$Z_{0.05} = 1.645$$

$$\left(\frac{0.04}{1.645}\right)^2 = \frac{.5(.5)}{n} \quad n = .25 \left(\frac{1.645}{.04}\right)^2 \quad n = 422.8$$

$n = 423$

## Constructing Confidence Intervals mixed practice

Ex 4: A large bus company wants to make a 95% confidence interval for the average time from west-side transfer point (WTP) to the airport (MSN) for a company report. Give them feedback on the plans they've proposed to obtain their data:

### Collecting data:

1. A manager will ask the drivers they see around the office how long their last trip WTP to MNS was and record the number of minutes they report.  
*biased by ones around the office*
2. A manager will get 10 drive times each from Driver A, Driver B, and Driver C and construct a confidence interval from the 30 times. The driver will call in when they are leaving WTP and when they arrive at MSN and the manager will find the difference.  
*sample size? people bad memory*  
*we want these times to be throughout week/day*
3. A manager will collect data in a similar way as (2) but will ask for/record differences at 30 different random times during the day and week.  
*are 30 observations independent? not really, to precise of confidence interval for amount of info from drivers*  
*better...*

needs to be from normal population and big enough that CLT makes  $\bar{x}$  or  $\hat{\pi}$  normal  
observations are independent and identically distributed i.i.d.

### Constructing Confidence Intervals mixed practice

Ex5. Monica learned in first grade that about 71% of Earth's surface is covered in water. To see whether this made sense, she asked her brother to toss her a spinning inflatable globe 100 times. For 66 of her catches, her right pointer finger tip was on water, while for 34 it was on land. Now she's stuck. Help her by finding and interpreting a 99% confidence interval for the proportion of Earth covered by water in light of her data.

$$\hat{p} = 0.66 \quad 0.66 \pm 2.575 \sqrt{\frac{0.66(0.34)}{100}} \quad (.538, .782)$$



99% confident that true percent of earth covered in water is between 54% and 78%