

Chapter 8: STATISTICAL INTERVALS FOR A SINGLE SAMPLE

Part 3: Summary of CI for μ Confidence Interval for a Population Proportion

Section 8-5

Summary for creating a $100(1-\alpha)\%$ CI for μ :

- When σ^2 is known and population is normal, use a z-value (works for any n).

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- When σ^2 is unknown and n is REALLY large, use a z-value and replace σ^2 with the observed sample variance s^2 (population distribution can be anything).

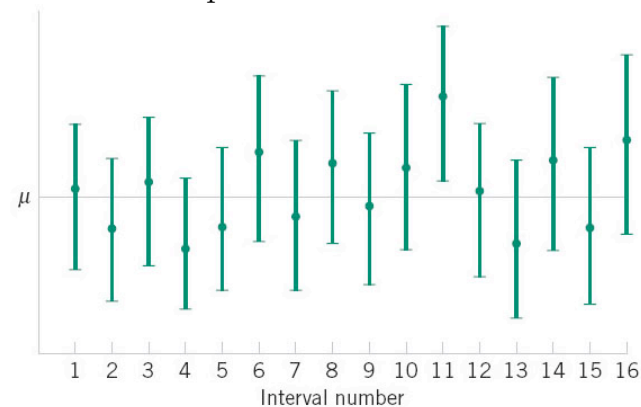
$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

- When σ^2 is unknown, and POPULATION IS 'NEARLY' NORMAL, and n is relatively small, use a t-value and the sample variance.

$$\bar{x} \pm t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}$$

- Before you make a CI, it is a random interval... it depends on the sample chosen, but \bar{X} will ALWAYS be at the center of the CI.

Example of 16 CI's for μ each based on a different sample:



- **Example:** Fuel rods in a reactor (problem 8-37 in book)

An article in *Nuclear Engineering International* gives the following measurements on the percentage of enrichment of 12 fuel rods in a reactor in Norway:

2.94 3.00 2.90 2.75 3.00 2.95
2.90 2.75 2.95 2.82 2.81 3.05

Calculate a 95% CI for the mean percentage of enrichment. Provide a normal probability plot to verify the assumption of normality.

ANS: This is a small sample with unknown σ^2 , so we will use the procedure with the t -value.

$$\bar{x} \pm t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}$$

$$\bar{x} = 2.9017$$

$$s^2 = 0.0098$$

$$s = 0.0993$$

$$t_{\alpha/2, n-1} = t_{0.025, 12-1} = t_{0.025, 11} = 2.201$$

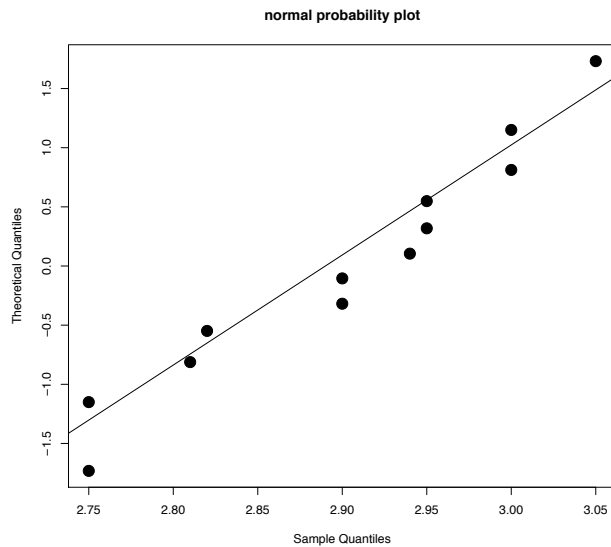
95% CI for μ :

$$2.9017 \pm 2.201 \cdot (0.0993/\sqrt{12})$$

$$2.9017 \pm 0.0631$$

$$[2.8386, 2.9648]$$

Checking normality with a normal probability plot:



Looks pretty good. The points fall randomly around the diagonal line. So, we can believe that the data was generated from a normal distribution.

Population Proportion Parameter p

- Moving on to another potential parameter of interest...

A population proportion is denoted p .

A sample proportion is denoted \hat{p} .

- A population proportion is based on a yes/no or 0/1 type variable.
 - What proportion of the Democrats favor Hillary Clinton?
yes/no
 - What proportion of a manufactured good is defective?
defective/not defective

- What proportion of the U.S. is republican?
republican/not republican
- What proportion of students entering college successfully complete a degree?
succeed/fail
- To estimate a population proportion, we will use a sample proportion.

Large-Sample Confidence Interval for a Population Proportion p

Section 8-5

- The construction of the CI for p relies on the fact that we took a large sample (large n).
- I don't like the book notation, so I will use something I find more clear...
 - We will let the category of interest be called the 'success' category (arbitrary).
 - Let $X_i = 1$ if observation i falls into the 'success' category.
 - Let $X_i = 0$ if observation i falls into the other category.

– Thus,

$\sum_{i=1}^n X_i$ = a count of all individuals in the ‘success’ category.

– X_i is called an *indicator variable*.

– The sample proportion of individuals falling into the ‘success’ category is

$$\hat{P} = \frac{\sum_{i=1}^n X_i}{n}$$
$$= \frac{\# \text{ in sample who are in ‘success’ category}}{n}$$

– \hat{P} is the point estimator for p

– \hat{p} is a realized point estimate from a observed sample.

• Note that p and n are actually the parameters for a binomial distribution.

– There are n trials (i.e. n draws of individuals for the sample)

– The probability of getting a ‘success’ remains constant as p (assuming we have a large population and n is not too large)

– Let Y = total number of successes.

– So $Y \sim \text{Binomial}(n, p)$

– $E(Y) = np$ and $V(Y) = np(1 - p)$

– In our notation, $Y = \sum_{i=1}^n X_i$, so

$$Y = \sum_{i=1}^n X_i \sim \text{Binomial}(n, p)$$

$$E(\sum_{i=1}^n X_i) = np$$

and

$$V(\sum_{i=1}^n X_i) = np(1 - p)$$

- Thus, if n is large, we have

$$\begin{aligned} Z &= \frac{(\sum_{i=1}^n X_i) - np}{\sqrt{np(1 - p)}} = \frac{\frac{(\sum_{i=1}^n X_i)}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \\ &= \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \end{aligned}$$

where Z is approximately standard normal.

(NOTE: This is a normal approximation to the binomial).

- **Normal approximation for a sample proportion \hat{P}**

If n is large, the distribution of

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately standard normal.

Or similarly, $\hat{P} \sim N(p, \frac{p(1-p)}{n})$

This means the sampling distribution for \hat{P} is normal...

See applet at:

<http://www.rossmanchance.com/applets/Reeses/ReesesPieces.html>

Thus, we can use a z-value to form a 100(1- α)% CI for p :

$$P(-z_{\alpha/2} \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}) = 1 - \alpha$$

rearranging...

$$P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

and we have the lower and upper bounds...

$$\text{Lower bound: } \hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$\text{Upper bound: } \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

BUT WE DON'T KNOW p , SO WE CAN'T GET ACTUAL VALUES FOR THE BOUNDS!

Solution \Rightarrow replace p with \hat{P} in the formulas.

• **Approximate 100(1- α)% CI for a population proportion p**

If \hat{p} is the proportion of observations in a random sample of size n that belongs to a class of interest, an approximate 100(1- α)% CI on the proportion of p of the population that belongs to this class is

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

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

- Things you need for the appropriate behavior of \hat{P} :
 - Population is large, and you don't take too many individuals for your sample. Maybe no more than 10% of the total population.

– The sample is a simple random sample.

ANS:

– $np \geq 5$ and $np(1 - p) \geq 5$.

- **Example:** Interpolation methods are used to estimate heights above sea level for locations where direct measurement are unavailable.

After verifying the estimates, it was found that the interpolation method made “large” errors at 26 of 74 random sample test locations.

Find a 90% CI for the overall proportion of locations at which this method will make “large” errors.

- **Choice of sample size for estimating p**

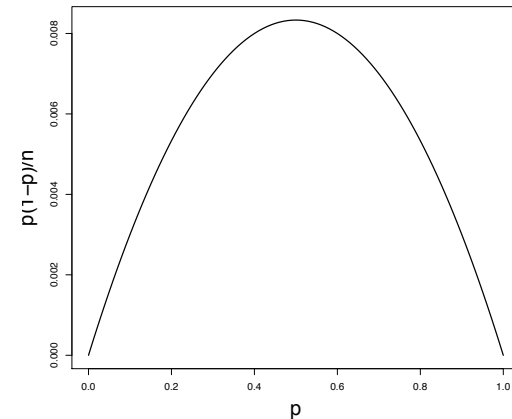
For a specified error $E = |p - \hat{P}|$ in your estimate, the previously stated behavior of \hat{P} suggests you should choose a sample size as:

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

But since we don't know p (that's what we're trying to estimate!), we can't compute a sample size from this formula... unless we estimate p first.

Here, we choose to err on the conservative side. It turns out the largest variance for \hat{P} [where $V(\hat{P}) = \frac{p(1-p)}{n}$] occurs when $p = 0.5$ no matter what the n was:

Plot of variance of \hat{P} vs. p



So, if we don't know p , we'll plug-in $p = 0.5$ to make sure we don't under-estimate the variance of our estimate.

- Thus, the working sample size formula is:

$$n = \left(\frac{z_{\alpha/2}}{E} \right)^2 0.25$$

We use this method because before we collect data, we don't have any information on p , so there's no observed \hat{p} to plug-in.

In contrast, when doing a CI for p , we plug-in \hat{p} because we DO have an estimate for p in that case.

- **Example:** In the interpolation example, how large of a sample would you need if you wanted to be *at least* 95% confident that the error in your estimate (i.e. $|p - \hat{p}|$) is less than 0.08?

ANS:

• **Example:** Gallup Poll

ANS:

Between February 9-11, 2007, adults were randomly sampled (by phone) and asked:

“Would you favor or oppose Congress taking action to set a time-table for withdrawing all U.S. troops from Iraq by the end of next year.”

Let p = the proportion of the population wanting a withdrawal time-table.

How large is their error if they sample 1000 people and form a 95% CI?

See *USA Today*, February 12, 2007.

<http://www.usatoday.com/news/polls/tables/live/2007-02-12-poll.htm>