Confidence Intervals for Large Sample Proportions

Dr Tom Ilvento
Department of Food and Resource Economics
University of Delaware

Overview

- Confidence Intervals – C.I.
- We will start with large sample C.I. for proportions, using the normal approximation of the binomial distribution
  - Binomial looks like a normal if n is large
  - and p or q is not too extreme
- New terms:
  - Bound of Error
  - Confidence Coefficient
  - Alpha (α)
- Goal is gain some sense of what a C.I. is saying

The Pepsi Challenge

- The Pepsi Challenge asked soda drinkers to compare Diet Coke and Diet Pepsi in a blind taste test.
- **Pepsi claimed that more than ½ of Diet Coke drinkers said they preferred Diet Pepsi**
- Suppose we take a random sample of 100 Diet Coke Drinkers and we found that 56 preferred Diet Pepsi.
  - $p = 56/100 = .56$ $\quad q = (1-.56) = .44$
  - n is large (100) and p or q is not small
  - We can use the normal approximation

Remember, just because I see something, doesn't mean it is so!

Proportions, p

- ***p*** = sample proportion an ***P*** is the population value (some books use **π**)
- If x represents the number of successes in our sample, then our estimator of P (population parameter) from a sample is
  - $p = x/n$
- The variance of a proportion is given by
  - $s^2 = pq$
  - Where $q = 1- p$
- $s = (pq)^{.5}$
- **Note: we will think there is a population proportion, *P*, with variance equal to σ^2**

Standard Error for a Proportion, *p*

- The Standard Error of the Sampling Distribution of a proportion is
  - SE for *p* = $(PQ/n)^{.5}$
  - Note: $\sigma^2 = PQ$, and $\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$
- If we don't know *P* and *Q*, we use the sample estimates, *p* and *q*

Steps in Calculating a Confidence Interval

1. **Note the sample size: n = 100**
2. **Calculate *p* and *q***
   - $p = 56/100 = .56$
   - $q = 1 - .56 = .44$
3. **Calculate the Variance and Standard Deviation**
   - $s^2 = pq = (.56)(.44) = .2464$
   - $s = .4964$
4. **Calculate the Standard Error**
   - $SE_p = .4964/(100)^{.5} = .0496$
   - **$SE_p = (.2464/100)^{.5} = .0496$ an alternative way**

Confidence Interval

- The sample provides an estimate –
  - **Point Estimate**, a single value computed from a sample and used to estimate the value of the target population.
  - The sample proportion and standard deviation are point estimates of population proportion *P* and population standard deviation σ respectively.
- I would like to place a **Bound of Error** around the estimate – **Confidence Interval** or an **Interval Estimate**

Confidence Interval

- I need to think of my sample as one of many possible samples
- I know from our work on the Normal curve that a z-value of ± 1.96 corresponds to 95 percent of the values in a normal distribtion
  - A z-value of 1.96 is associated with a probability of .475 on one side of the normal curve
  - 2 times that value yields 95% of the area under the normal curve, centered around the middle of the distribution (the mean)

Confidence Interval for the Pepsi Challenge

- If I think of my sample as part of the sampling distribution
- I can place a **1.96(standard error)** around my estimate
- Like this for a 95% C.I.:
  - .56 ± 1.96(.0496)
  - .56 ± .097
  - **.463 to .657**

Notice that values less than .5 are in this interval - the population value *P* could be less than .5

Why did I use the Standard Error in the formula?

- I am asking the question about the proportion of Diet Coke drinkers who prefer Pepsi
- I want some sense of how well my sample estimates the population
- If my sample is drawn randomly, it will represent the population, plus some sampling error
- A **95% Confidence Interval** means that
  - If I would have taken all possible samples
  - And calculated a confidence interval for the proportion for each one
  - **95% of them would have contained the true population parameter**

What is a Confidence Interval?

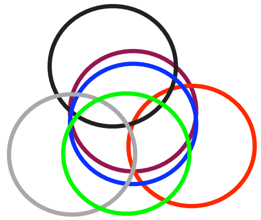
- It is an interval estimate of a population parameter
- The plus or minus part is also known as a **Bound of Error (BOE)** or **Margin of Error (MOE)**
- Placed in a probability framework
- **Like this for a 95% C.I.:**
  - **.56 ± 1.96(.0496)**
  - **.56 ± .097**
  - **.463 to .657**

What is a Confidence Interval?

- We calculate the probability that the estimation process will result in an interval that contains the true value of the population proportion or mean
  - If we had repeated samples
  - Most of the C.I.s would contain the population parameter
  - **But not all of them will!!!!**

Think of this like the Jart game (only backwards)

- Jarts was a backyard game in the 1970's and 80's (aks, Lawn Darts)
- You placed a ring on the ground, and tried to throw a giant dart into the ring, somewhat like horseshoes
- The darts were sharp and some people got hurt!
- But let's rethink this game - throw rings around a fixed Jart
  - The Jart is the population parameter
  - and the rings are confidence intervals
- Some rings will miss, but most will capture it



To construct a Confidence Interval, we need

- A point estimator — **Estimator of P is, *p*= x/n**
- A sample and a sample estimate using the estimator — ***p* from a sample of n observations**
- Knowledge of the Sampling Distribution of the point estimator — **The sampling distribution is known with mean = P**
  - The Standard Error of the estimator — **$SE_p = (PQ/n)^{.5}$**
  - The form of the sampling distribution — **Normal approximation of binomial**
- A probability level we are comfortable with – how much certainty. It's also called "Confidence Coefficient" — **Most times we will use either a .90, .95 or a .99 Confidence Coefficient**
- A level of Error — **α, which is the chance of being wrong**

Confidence Interval for a Proportion

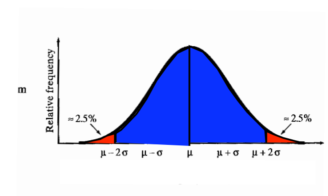
- Formula for C.I. for a Proportion *p*
- We are using the Normal Approximation to the Binomial Distribution
- And the sample estimates of p and q
- Assumption: A sufficiently large random sample of size n is selected from the population.

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

Confidence Interval

- **$Z_{\alpha/2}$** refers to the z-score associated with a particular probability level divided by 2
- **α** refers to the area in the tails of the distribution
  - We divide by 2 because we divide **α** equally on both sides of the mean
  - Which means **α** represents the combined area, or the probability, in the tails of both sides of the normal curve
- **The 95% part is divided evenly around the center of the distribution** and the **5% part, α, is distributed evenly in the tails**

$$p \pm \boxed{Z_{\alpha/2}}\sqrt{\frac{p\,(1-p\,)}{n}}$$



Confidence Interval

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

- The larger the Confidence Coefficient or probability level for a C.I.
- The smaller the value of α, and α/2
- The larger the z value

Confidence Coefficient (1-α)*100	α	α/2	$Z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.575

For any given sample size, the width of the Confidence Interval depends upon α

- For the Pepsi Challenge Example
  - 90% C.I. .56 ± 1.645(.0496) = .56 ± .0816
  - 95% C.I. .56 ± 1.96(.0496) = .56 ± .0972
  - 99% C.I. .56 ± 2.575(.0496) = .56 ± .1277

For any given sample size, if you want to be more certain (<u>smaller</u> <u>α</u>) you have to accept a wider interval

A problem for you to try

- Survey questionnaire for who citizens would vote for in a state election
- 1,052 adults selected randomly were surveyed by a major newspaper
- The percentage who indicated Candidate B was 35%
- Construct a 95% C.I. for this proportion

The Solution

- The facts
  - **p = .35**
  - **q = (1 - .35) = .65**
  - **n = 1,052**
- Standard Error = $s_p = \sqrt{\frac{.35 \cdot .65}{1052}} = .0147$
- C.I.
  - .35 ± 1.96(.0147)
  - .35 ± .0288
  - **.3212 to .3788**

Newspaper MOE

- The newspaper said "there is a 3.0% Margin Of Error."
- Where did this figure come from?
- It doesn't match our previous figure of 2.88%
- And what does MOE mean?

•They calculated a general C.I. For a proportion at .5

Standard Error = $[(.5*.5)/1,052]^{.5}$

= .0154

•C.I.

- .5 ± 1.96(.0154)
- **.5 ± .0302 or 3%**

Variance is largest at *p*=.5

- For a proportion, the variance is largest at .5, or an equal split
  - At .5 $s^2 = (.5)(.5) = .25$
  - At .7 $s^2 = (.7)(.3) = .21$
  - At .3 $s^2 = (.3)(.7) = .21$
- Which brings up another unique thing about proportions – once you specify a value of *p* for the population, the **variance (σ^2) is known**.

Summary

- Confidence Intervals are a way to place a bound of Error around our estimate, in a probability framework.
- We need
  - an estimator for P, ***p=x/n***
  - a sample estimate (***p***)
  - Knowledge of the sampling distribution (the normal distribution) and a **standard error**
  - The level of **alpha** – the area in the tails
- For confidence Intervals for proportions, we use the normal approximation of the binomial distribution as long as the sample size is sufficiently large and *p* (or *q*) is not too small.