

Confidence Intervals for Proportions and for Means of Small Samples

Dr. Tom Ilvento
FREC 408

Confidence Intervals with Proportions

- Sometimes our data deals with a dichotomous variable
 - Yes or No
 - On or Off
 - Alive or Dead
- If we code the variable as a zero/one dichotomy, the mean of the variable is the proportion with the attributed coded as one

Coding Strategy, Let 1=Yes, 0 = No

Do you support the President?	
Survey Answer	Code
Yes	1
Yes	1
No	0
Yes	1
No	0
Yes	1
Yes	1
Yes	1
Yes	1
Number Yes = 6	Sum = 6 n = 8
Proportion = .75	Mean = 6/8 = .75

Proportions

- p = Number of Success/ Total in population
 - If x represents the number of successes in our sample, then our estimator of p (population parameter) from a sample is
 - $\hat{p} = x/n$
- The variance of a proportion is given by
 - $\sigma^2 = pq$
 - $\sigma = \sqrt{pq}$
 - Where $q = 1 - p$

Note: $s = \sqrt{\hat{p}\hat{q}}$

Proportions

- The **Standard Error** of the Sampling Distribution of a proportion is
 - $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$
 - $\sigma_{\hat{p}} = (pq/n)^{.5}$
- If we don't know p and q , we use the sample estimates, \hat{p} and \hat{q}

Confidence Interval for a Population Proportion p

- Formula for C.I. for a Population Proportion p (p361)

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

- Assumption: A sufficiently large random sample of size n is selected from the population.

Problem

- Survey questionnaire for who would you vote for
- 1,052 adults were surveyed by a major newspaper
 - *The percentage who indicated Candidate B was 35%*
- Construct a 95% C.I. For this proportion

Newspaper Confidence Interval Problem

- $p = .35$
- $q = 1 - .35 = .65$
- $n = 1,052$
- Standard Error = $[(.35 \cdot .65)/1052]^{.5}$
 - = .0147
- C.I.
 - $.35 \pm 1.96(.0147)$
 - $.35 \pm .0288$
 - .3212 to .3788

Newspaper C.I.

- The newspaper said "there is a $\pm 3.0\%$ margin of error."
- Where did this figure come from?
 - It doesn't match our previous figure of 2.88%
- And what does it mean?

Newspaper C.I.

- They calculated a general C.I. For a proportion at .5
 - Standard Error = $[(.5 * .5)/1,052]^{.5}$
 - = .0154
- C.I.
 - $.5 \pm 1.96(.0154)$
 - $.5 \pm .0302$

Variance is largest at .5

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

Proportion C.I. Problem

- Time/CNN hired Yankelovich Partners Inc. to survey via the telephone **205** never married single women
- Question: ***If you couldn't find the perfect mate, would you marry someone else?***
- **34%** said Yes
- Construct a 99% Confidence Interval around this estimate.

Proportion 99% C.I.

- $p = .34$ $q = .66$ $n = 205$
- Standard Error = $[(.34)(.66)/205]^{.5}$
 - = .0331
- $.34 \pm 2.575(.0331)$
- $.34 \pm .0852$
- .2548 to .4252

Small Sample Confidence Intervals

- This doesn't apply to a proportion – it requires a larger sample for a binomial to be approximated with a normal distribution
- However, for the mean we need to rethink the process a bit.

Small sample confidence intervals

- For the mean, we use the sample estimate s whenever we don't know σ
- **When our sample is small, the sample estimate s has too much variability**
 - Remember, the variance is very sensitive to outliers
 - If the variable in the population is a normally distributed variable it won't be too big of a problem
 - But if the population is not distributed normally, we could have big problems

Small sample Confidence Intervals

- If we have a small sample, even if it is normally distributed, the z-distribution tends to give a biased estimate the standard error
- What can we do?

Relax and have a beer!

- W.S. Gossett worked for Guinness Brewery in Ireland around 1900
- In quality control tests he noticed the problem of using the z-distribution
- His solution was the t-distribution



t-distribution

- Similar to the standard normal distribution
- The t-distribution varies with n (sample size) via **degrees of freedom**
 - $df = n - 1$
- As n gets larger, the t-distribution approximates the z distribution

The t-Table (table B4 p738)

- Organized with degrees of freedom as rows
- Probabilities in the right tail (α) are the columns
- We substitute the t-value from the table for a z-value in the C.I.
- BUT! A big assumption in the small sample t-test is that the population is distributed approximately normal**

The formula for a small sample Confidence Interval

$$\bar{x} \pm t_{\alpha/2, n-1 d.f.} \left(\frac{s}{\sqrt{n}} \right)$$

Page 351

$t_{\alpha/2}$ is based on $(n-1)$ degrees of freedom

The meaning of the t-value

- The t-value is interpreted like the z-value from the standardized normal table
- NOTE: For a Confidence Interval, the t-value represents the corresponding value at $\alpha/2$
- Which is out in the right tail of the curve
- So a t-value for 30 degrees of freedom at the .025 level is 2.042
 - This corresponds to a z-value of 1.96
 - And is used for a 95% C.I.

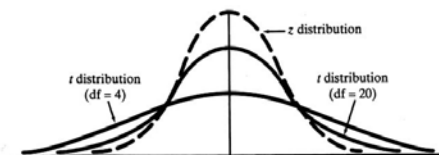
Part of the t-table – page 738

Degrees of Freedom	t _{.100}	t _{.050}	t _{.025}		t _{.0005}
1	3.078	6.314	12.706		636.62
2	1.886	2.920	4.303		31.598
3	1.638	2.353	3.182		12.924
30	1.310	1.697	2.042		3.646
∞	1.282	1.645	1.960		3.291

For any probability level, as the degrees of freedom get larger, the t-value gets smaller

Degrees of Freedom	t _{.100}	t _{.050}	t _{.025}		t _{.0005}
1	3.078	6.314	12.706		636.62
2	1.886	2.920	4.303		31.598
3	1.638	2.353	3.182		12.924
As the degrees of freedom gets to 30, the t-value approaches z					
30	1.310	1.697	2.042		3.646
∞	1.282	1.645	1.960		3.291

Comparing z-distribution and t-distribution (page351)



t-values and Computer Software packages

- The t-distribution has been worked out for a variety of levels of α and degrees of freedom
- It reflects the Central Limit Theorem that says that as n gets large, the sampling distribution approaches a normal distribution
- To be safe, software packages present all confidence intervals and hypotheses tests using a t-value rather than a z-value

Example Problem

- Spinifex pigeons in Western Australia rely entirely on seeds for food
- Examination of stomach contents of 16 pigeons
- Recorded the weight in grams of dry seed of each pigeon
- Sample Statistics
 - $n=16$
 - Mean = 1.373
 - $s = 1.034$
- **Construct a 99% C.I.**

Answer:

- Assuming the distribution is approximately normal and σ is unknown
- We therefore use the t-distribution
- **d.f. = $n-1 = 16-1 = 15$**

Pigeon problem

- t-value for 99% C.I.
 - $\alpha = .01$
 - $\alpha/2 = .005$ in each tail
 - $t_{.005}$ with 15 d.f. = 2.947
- $1.373 \pm 2.947(1.034/\sqrt{16})$
 - $1.373 \pm .762$
 - .611 to 2.135

Now you try it

- A furniture company wants to test a random sample of sofas to determine how long the cushions last
- They simulate people sitting on the sofas by dropping a heavy object on the cushions until they wear out – they count the number of drops it takes
- This test involves 9 sofas

Sofa Test

- Mean = 12,648.889
- $s = 1,898.673$
- Assume it follows a normal distribution
- Generate a 95% Confidence Interval for this problem

Sofa Test Answer

Sofa Test

- The company wants to advertise that the sofas last for 20 years
- Assuming a person sits on the sofa an average of once a day, is this warranty is good idea?
- Answer:
 - $365 \times 20 = 7,300$ sits
 - The lower bound of our estimate is 11,189, so the 20 year warranty is pretty safe

Solving this problem with Excel

- I entered the data into a column in Excel
- I then used the following sequence
 - Tools
 - Data Analysis
 - Descriptive Statistics
- I then follow the options, including:
 - Identify the Input Range, marking a label is in the first row
 - Output range
 - Descriptive statistics
 - A 95% Confidence Interval

Excel Output for Sofa problem

Sofa Drops	
Mean	12,648.889
Standard Error	632.891
Median	12742
Mode	#N/A
Standard Deviation	1898.673
Sample Variance	3604958.111
Kurtosis	-0.676
Skewness	-0.372
Range	5,886
Minimum	9,459
Maximum	15,345
Sum	113,840
Count	9
Confidence Level(95.0%)	1,459.450

Mean = 12,648.889
 $s_x = 632.891$
 $s = 1,898.673$
 $12,648.889 \pm 1,459.447$

A few more points on small sample C.I.

- If we cannot assume a normal distribution
 - The probability associated with our interval is not $(1 - \alpha)$
 - We really shouldn't construct a C.I.
 - Or we should get more data
- If σ is known, we can use the z instead of the t, but we still need to have an approximately normal distribution

What influences the width of a confidence interval?

- The sample size
- The level of α
- The level of the confidence coefficient $(1 - \alpha)$
- The variability of the data – i.e., the standard deviation

What influences the width of a confidence interval?

- Sample Size or n
- The **larger** the **sample size**, the **smaller** the **C.I.**
- For a 95% Confidence Interval when s = 25
 - n=50 $1.96(25/\sqrt{50}) = 6.93$
 - n=500 $1.96(25/\sqrt{500}) = 2.19$

What influences the width of a confidence interval?

- The level of α
- The **larger** the level of α , the **smaller** the **C.I.**
- For a 95% Confidence Interval when s = 25 and n=50
 - $\alpha = .05$ $1.96(25/\sqrt{50}) = 6.93$
 - $\alpha = .1$ $1.645(25/\sqrt{50}) = 5.82$

What influences the width of a confidence interval?

- The level of the confidence coefficient (1- α)
- The **larger** the **confidence coefficient**, the **larger** the **C.I.**
- When s = 25 and n = 50
 - 95% C.I. $1.96(25/\sqrt{50}) = 6.93$
 - 99% C.I. $2.575(25/\sqrt{500}) = 9.10$

Focus in on sample size (n)

- For a given (1- α) C.I., and a given bound of error (B), which is what we add or subtract to the sample estimate
- We can calculate the needed sample size as

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{B^2} \quad \text{p368}$$

Where does this formula come from?

- The Bound of Error is given as

$$B = z_{\alpha/2} (\sigma / \sqrt{n})$$

Solve for n by first squaring both sides of the equation

$$B^2 = \frac{(z_{\alpha/2})^2 \sigma^2}{n}$$

Re-arrange terms

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{B^2}$$

The formula for determining the sample size for a proportion

$$n = \frac{(z_{\alpha/2})^2 (pq)}{B^2} \quad \text{p368}$$

Confidence Interval Summary

- Provides an interval estimate of a sample estimator
- Requires knowledge of the sampling distribution of the estimator
- We treat our estimate from a sample as one of many possible estimates from many possible samples

Confidence Interval Approach

- Figure a C.I. Probability level as $(1 - \alpha)$
 - where $\alpha/2$ represents the probability in either tail of the sampling distribution
 - $(1 - \alpha)$ is referred to as the confidence coefficient

$$\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}$$

Confidence Interval Approach

- If the sample size is large (> 30) then you can use a corresponding value from the z-table – but it is ok to use the t-distribution
- If the sample size is small, σ is unknown, and the distribution is approximately normal, you **must** use the t-table with $n-1$ degrees of freedom

$$\bar{x} \pm t_{\alpha/2, n-1 d.f.} \sigma_{\bar{x}}$$

Confidence Interval Approach

- For proportions, you can only use a large sample approach
- So you use a z-score

Confidence Interval Approach

- Calculate the standard error
 - If standard deviation of the population is known use σ/\sqrt{n}
 - If not, use the sample estimate s/\sqrt{n}
 - For proportions, use $(pq/n)^{.5}$
- **Put it all together and calculate the C.I.**