Statistics for Computing

MA4413 Lecture 6B

Kevin O'Brien

Kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn Semester 2013

Computing Confidence Intervals

Confidence limits are the lower and upper boundaries / values of a confidence interval, that is, the values which define the range of a confidence interval. The general structure of a confidence interval is as follows:

$$\text{Point Estimate} \pm [\text{Quantile} \times \text{Standard Error}]$$

Point Estimate : estimate for population parameter of interest, i.e. sample mean, sample proportion.

Quantile : a value from a probability distribution that scales the intervals according to the specified confidence level.

Standard Error : measures the dispersion of the sampling distribution for a given sample size n.

Confidence Intervals (Revision)

- The 95% confidence interval is a range of values which contain the true population parameter (i.e. mean, proportion etc) with a probability of 95%.
- We can expect that a 95% confidence interval will not include the true parameter values 5% of the time.
- A confidence level of 95% is commonly used for computing confidence interval, but we could also have confidence levels of 90%, 99% and 99.9%.

Confidence Level (Revision)

- A confidence level for an interval is denoted to $1-\alpha$ (in percentages: $100(1-\alpha)\%$) for some value α.
- A confidence level of 95% corresponds to $\alpha = 0.05$.
- $100(1-\alpha)\% = 100(1-0.05)\% = 100(0.95)\% = 95\%$
- For a confidence level of 99%, $\alpha = 0.01$.
- Knowing the correct value for α is important when determining quantiles.

Quantiles

- The quantile is a value from a probability distribution that scales the intervals according to the specified confidence level.
- For practical purposes, the quantile can be taken from the standard normal distribution, if the sample is larger than 30, further to the central limit theorem.
- For a specified confidence level $1-\alpha$, the corresponding quantile is the value z_o that satisfies the following identity (when $n > 30$):

$$P(-z_o \leq Z \leq z_o) = 1 - \alpha$$

Quantiles

- When the sample size n is greater than 30, we can compute the quantile using Murdoch Barnes table 3.
- 95% of Z random variables are between -1.96 (quantile for 2.5%)and 1.96 (quantile for 97.5%)
- If the confidence level is 95%, then the quantile is 1.96. Recall

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

- If the confidence level is 90%, then the quantile is 1.645.

$$P(-1.645 \leq Z \leq 1.645) = 0.90$$

- If the confidence level is 99%, then the quantile is 2.576.

$$P(-2.576 \leq Z \leq 2.576) = 0.99$$

Confidence Intervals for Sample Means

Broadly speaking, there are three different types of confidence interval

Type 1 Sample with **known** population variance

- The size of the sample doesn't matter.

Type 2 Large sample with **unknown** population variance

- The size of the sample is more than 30 ($n > 30$)

Type 3 Small sample with **unknown** population variance

- The size of the sample is 30 or less ($n \leq 30$)

Confidence Intervals for Sample Means

Type 1 : Known Population Variance

- This type of confidence interval is very rare in practice, but it is very simple to implement and used as introductory teaching material with those studying confidence intervals for the first time.
- This type of confidence interval is computed using this formula:

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

- A description of each item is on the next slide.

Confidence Intervals for Sample Means

Type 1 : Known Population Variance

- The point estimate is the sample mean : $\bar{x}$
- We use a ***quantile*** from the standard normal (Z) distribution to "scale" the confidence interval to the specified confidence level (usually 95%).
- Let the confidence level be denoted in the for $(1-\alpha)\times 100\%$, and hence determine α. For example, if the confidence level is 95%, then α is 0.05 (or 5%).
- The Quantile ($z_{(\alpha/2)}$) is the value for the Standard Normal Tables (for example Murdoch Barnes Table 3) such that

$$P(Z \geq z_{(\alpha/2)}) = \frac{\alpha}{2}$$

- For a 95% confidence interval, the quantile is 1.96. For a 99% confidence interval, the quantile is 2.576.

Confidence Intervals for Sample Means

Type 1 : Known Population Variance

- The population standard deviation is σ. The sample size is n.
- The standard error to be used in this confidence interval is

$$\text{S.E}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Confidence Intervals for Sample Means

Type 2 : Large Sample, Unknown Population Variance

- The point estimate is the sample mean : $\bar{x}$
- For a type 2 confidence interval, we can determine a ***quantile*** for the confidence interval in the same way that for the Type 1 confidence interval.
- Recall: For a 95% confidence interval, the quantile is 1.96. For a 99% confidence interval, the quantile is approximately 2.58.

Confidence Intervals for Sample Means

Type 2 : Large Sample, Unknown Population Variance

- The population standard deviation which we denote σ^2 is unknown. Instead we are given the sample variance s^2. We use the sample standard deviation (the square root of the variance) as an estimate for the population standard deviation σ.

$$s \text{ is an estimate for } \sigma$$

- The sample size is n.
- The standard error to be used in a Type 2 confidence interval is

$$\text{S.E}(\bar{x}) = \frac{s}{\sqrt{n}}$$

Confidence Intervals for Sample Means

Type 3 : Small Sample, Unknown Population Variance

- Firstly, we will clarify the similarities of the Type 2 and Type 3 confidence interval.
- The point estimate is the sample mean $\bar{x}$.
- The sample standard deviation s is used to estimate the population σ.
- The standard error is

$$\text{S.E}(\bar{x}) = \frac{s}{\sqrt{n}}$$

Confidence Intervals for Sample Means

Type 3 : Small Sample, Unknown Population Variance

- The key difference is in determining the quantile. Rather than use the standard normal distribution, we must use the student $t-$ distribution. Quantiles for this distribution are also tabulated in statistical tables (for Example, Murdoch Barnes Table 7).
- Recall that we must determine a value for α (and hence $\alpha/2$). For a 95% confidence interval, $\alpha = 0.05$ and $\alpha/2 = 0.025$.
- Computing a quantile from the $t-$ distribution additionally requires the specification of the ***degrees of freedom***. Degrees of Freedom are often denote as df or by the greek letter ν ("nu").
- For small sample confidence intervals (i.e. $n \leq 30$), the degrees of freedom are

$$df = n - 1$$

Using the $t-$distribution for large samples

- The $t-$distribution is used for computing quantiles in the case of small samples (i.e. when sample size $n \leq 30$).
- A key value in the $t-$distribution is the degrees of freedom, denoted df (or sometimes ν). For small samples

$$df = n - 1$$

.

- The $t-$distribution is used for computing quantiles in the case of large samples too, as an alternative to using the Z distribution.
- In this case , use the value ∞ as the degrees of freedom (see bottom row of table 7).

$$df = \infty$$

- This means that we can use the $t-$ distribution for finding the quantiles of all types of confidence intervals.

The Central Limit Theorem

- This theorem states that as sample size n is increased, the sampling distribution of the mean (and for other sample statistics as well) approaches the normal distribution in form, regardless of the form of the population distribution from which the sample was taken.
- For practical purposes, the sampling distribution of the mean can be assumed to be approximately normally distributed, even for the most non-normal populations or processes, whenever the sample size is $n > 30$.
- (For populations that are only somewhat non-normal, even a smaller sample size will suffice. A variation of the normal distribution can be used for such circumstances.)

Standard Error

- The standard error measures the dispersion of the sampling distribution.
- For each type of point estimate, there is a corresponding standard error.
- A full list of standard error formulae will be attached in your examination paper.
- The standard error for a mean is

$$S.E(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

However, we often do not know the value for σ. For practical purposes, we use the sample standard deviation s as an estimate for σ instead.

$$S.E(\bar{x}) = \frac{s}{\sqrt{n}}$$

Confidence Intervals for Sample Proportion

Confidence Intervals may also be computed for **Sample Proportions**.
The sample proportion is used to estimate the value of a population proportion. The sample proportion is denoted $\hat{p}$. The population proportion is denoted π.

- The structure of a confidence interval for sample proportion is

$$\hat{p} \pm z_{(\alpha/2)} \times \text{S.E.}(\hat{p})$$

- The standard error, in the case of sample proportions, is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

- (When computing this interval with statistical software, it is common to enhance the solution using a **continuity correction** . This is not part of our syllabus.)

Point Estimates for proportions

Sample Percentage

$$\hat{p} = \frac{x}{n} \times 100\%$$

- $\hat{p}$ - sample proportion.
- x - number of “successes”.
- n - the sample size.

Point Estimates for proportions

- Of a sample of 160 computer programmers, 56 reported than Python was their primary programming language.
- Let π be the proportion of all programmers who regard Python as their programming language.
- What is the point estimate for π?

$$\hat{p} = \frac{x}{n} \times 100\%$$

$$\hat{p} = \frac{56}{160} = 35\%$$

Standard Error for Proportions

The standard error for proportions is computed using this formula.

$$S.E.(\hat{p}) \;=\; \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}}$$

When expressing the proportion as a percentage, we adjust the standard error accordingly.

$$S.E.(\hat{p}) \;=\; \sqrt{\frac{\hat{p}\times(100-\hat{p})}{n}}\,[\%]$$

Sample Proportion : Example

Point Estimate The sample proportion is computed as follows

$$\hat{p} = \frac{x}{n} = \frac{56}{160} = 0.35$$

Quantile We are asked for a 95% confidence interval. The quantile is therefore

$$z_{\alpha/2} = 1.96$$

Standard Error The standard error, with sample size n=120 is computed as follows

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} = \sqrt{\frac{0.35 \times 0.65}{160}}$$

(Full solution to follow on whiteboard)

Sample Proportion : Example

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{0.35 \times 0.65}{160}} = \sqrt{\frac{0.2275}{160}}$$

$$\text{S.E.}(\hat{p}) = \sqrt{0.001421875} = 0.03770$$

95% Confidence Interval

$$0.35 \pm (1.96 \times 0.0377) = (0.2761, 0.4239)$$

Confidence Intervals for Sample Proportion

Unlike confidence intervals for sample means, there is only one type of confidence interval when dealing with sample proportions.

Optional

- It is often easier to work in terms of percentages, rather than proportions. If you are working in terms of percentages, remember to use the appropriate ***complement value*** in the standard error formula (i.e. $100 - \hat{p}$)
- The standard error, in the case of sample proportions, is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (100 - \hat{p}}{n}}$$

- Complement Values:
  - When working in terms of proportions, for the the value $\hat{p} = 0.35$, the complement value is $1 - \hat{p} = 0.65$.
  - When working in terms of percentages, for the value $\hat{p} = 35\%$, the complement value is $100 - \hat{p} = 65\%$.

Confidence Interval for a mean (1)

Finally, an example to finish the class:

- For a given week, a random sample of 100 hourly employees selected from a very large number of employees in a manufacturing firm has a sample mean wage of $\bar{x} = 280$ dollars, with a sample standard deviation of $s = 40$ dollars.
- Estimate the mean wage for all hourly employees in the firm with an interval estimate such that we can be 95 percent confident that the interval includes the value of the population mean.

Confidence Interval for a mean (2)

- The point estimate in this case is the sample mean $\bar{x} = 280$ dollars.
- We have a large sample (n=100) and the confidence level is 95%. Therefore the quantile is 1.96.
- The standard error is computed as follows:

$$S.E(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{40}{\sqrt{100}} = 4$$

- **Confidence Interval for mean**

$$280 \pm (1.96 \times 4) = (280 \pm 7.84) = (\ 272.16\ ,\ 287.84\ )$$