Module 8 Lecture - Multiple Comparisons Under Factorial ANOVA

Analysis of Variance

Quinton Quagliano, M.S., C.S.P

Department of Educational Psychology

Table of Contents

1 Overview and Introduction

1.1 Textbook Learning Objectives

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
- Interpret the Student's t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the Student's t distributions.
- Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

1.2 Instructor Learning Objectives

- Understand the value of calculating a confidence interval in interpreting the "accuracy" of a certain statistic
- Appreciate how the calculation and interpretation of a confidence interval builds upon our previous understanding of distributions and probability
- Value how confidence intervals demonstrate the inherently probabilistic nature of quantitative analysis in research

1.3 Introduction

> **Important**
>
> Professional researcher often forget just how important it is to report confidence intervals in results - but they are actually very useful in interpretation!

- Up until now, we have primarily been working with point estimates of sample statistics
  - E.g., a ________________ statistic of a mean ($\bar{x}$) is a point estimate of the population parameter mean (μ)
  - Same logic applied to standard deviation, variance, median, etc.
- Inherently, our single point estimate of a statistic is going to be insufficient for telling us much about how close it ________________ is to the population parameter
  - Thus, we need some other ________________ to add to the point estimate to give us more information about the population

Important

Remember our goal with statistics is to tell us something about the population, not just the sample!

- This is, in essence, getting us closer to inferential statistics, where we can actually infer meaning from the characteristics of our ______
- The part that we need to add are called confidence intervals
  - A ______ interval is effectively a ______ of values that we believe a population parameter falls in, with a certain amount of confidence
  - It is actually appropriate to refer to the confidence interval itself as a ______ continuous variable, as it is referring to a distribution of possible values that sample statistics can be drawn from

Discuss: As a review, try explaining, in your own words, what a 'random' variable is?

- At first, we'll talk about applications of confidence intervals to the ______, implying that we are working with continuous variables
  - We'll touch on more categorical, discrete type stuff later

1.4 Nuances in the Confidence Intervals

- Commonly, we use ______ percent as the level of confidence
  - However we can use something like ______ percent for a more ______ estimate

Discuss: If you have taken statistics before, what other value is commonly associated with 95%

- Also, it is important to treat the confidence interval as another type of estimate, which we usually call a interval estimate
  - Just like our ______ estimates, there is no guarantee this in infallible
  - In fact, the confidence interval is best understood as a description of the reliability of sample statistics when taken from the same population, but does not mean it is certain that the population parameter falls within the range

Discuss: For review, describe again why we can't take sample estimates as being representative of populations? hint: use the vocabulary sample v——

- Depending on the ______ and the outlet you may see confidence intervals described as something like margin of error

1.5 Basic Calculating Confidence Intervals

- In order to ______ the confidence interval of a certain statistic, we need to know the standard ______ of the variable that we are interested in
  - Recall, this is the ______ standard deviation of the distribution that our statistic comes from
- However, this is most easy to know when we can assume that our variable comes from a ______ distribution
  - This comes back to being able to appeal to the ______

rule that was introduced as part of understanding normal distributions
- Remember that the empirical rule is also sometimes called the 68-95-99.7 rule

Discuss: Based upon the description above, try re-describing what this rule says about normal distributed variables

Important

There are several slight variations on the empirical rule, but it is easiest to round things off to whole number standard deviations

- In order to calculate the 95% confidence interval, we use the following formula:
  - $[PE - 2SE, PE + 2SE]$
  - Where PE: is our ________ estimate
  - SE: standard ________ of the statistic
  - $2SE$ is used in appeal to the ________ rule, because, theoretically, 95% of the values of the statistic fall within 2 standard ________; we can use 1 or 3 to find the 68% or 99.7% confidence intervals, respectively

Important

Recall that the standard error decreases as sample size increases, which also means that a larger sample size results in a more narrow confidence interval

- This becomes practically ________ as 95% of the sample statistics taken from this population with the same sample size (n) would give us a statistic within this range 95% of the time
- For example: I have a point estimate mean statistic ($\bar{x}$) of 100, a sample size (n) of 40, and a population parameter standard deviation (σ) of 10
  - Review: $SE = \frac{\sigma}{\sqrt{n}}$
  - Thus: $SE = \frac{10}{\sqrt{40}} = 1.58$
  - So 95% CIs are: $[100 - 3.16, 100 + 3.16] = [96.84, 103.16]$

- Practical interpretation: 95% of infinitely many sample statistics taken from this population will fall between 96.84 and 103.16

Discuss: Try following the same procedures for a point estimate mean of 50, a sample size of 12, and a population standard deviation of 20

2 A Single Population Mean Using the Normal Distribution

2.1 Introduction

- In this section, we'll say more to the ______ and specifics used in each part of the calculation process, when the population standard deviation is ______
  - You may ______ that we can often know the population SD, which is fair, we'll get into that in a bit
  - For this scenario, we rely upon the ______ limit theorem
- We arbitrarily chose our confidence level (CL), but is usually 90% or above; it is how ______ we want to be that the population parameter falls within the confidence intervals
  - Based upon that CL, we can calculate the error bound for a population mean (EBM), which is the amount we deviate (i.e., add or subtract) from the point estimate mean
- Another way of conceptualizing confidence level is through alpha
  - It is just the complement of CL: $\alpha + CL = 1$

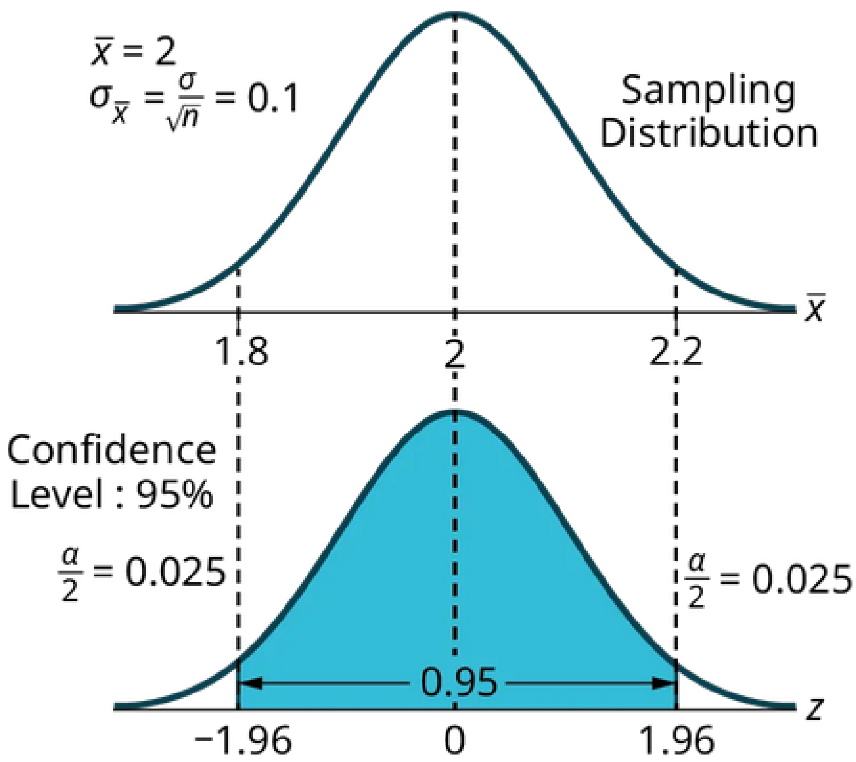

Figure 8.2

$$
\begin{aligned}
\mu &= \overline{X} \pm Z_{\alpha}\left(\frac{\sigma}{\sqrt{n}}\right) \\
&= 2 \pm 1.96(0.1) \\
&= 2 \pm 0.196 \\
1.804 &\leq \mu \leq 2.196
\end{aligned}
$$

Important

For my folks who haven't taken stats before, alpha will show up again when we talk about p-values!

- Consider my prior example:
  - I have a point estimate mean statistic ($\bar{x}$) of 100, a sample size (n) of 40, and a population parameter standard deviation (σ) of 10
    - Review: $SE = \frac{\sigma}{\sqrt{n}}$
    - Thus: $SE = \frac{10}{\sqrt{40}} = 1.58$
    - So 95% CIs are: $[100 - 3.16, 100 + 3.16] = [96.84, 103.16]$

Discuss: Identify what is the CL, alpha, and EBM in this worked problem

Discuss: Make sure you can properly describe the different between a standard deviation of the population and the standard error of the mean

2.2 Writing Z-score Notation for Confidence Intervals

- Thinking back to the image above, we split α between the two ________________ of the normal distribution
  - Thus, each tail contains $\frac{\alpha}{2}$ area, and the ________________ that denotes where the CI bound is is $z_{\frac{\alpha}{2}}$
  - Therefore the upper 95% CI bound z is stated as $z_{\frac{0.05}{2}}$ or $x_{0.025}$

Discuss: Now do the same thing with a 92% CI lower bound

3 A Single Population Mean Using the Student t Distribution

3.1 Introduction

- As mentioned prior, we don't often have what we can easily consider to be a ______ population standard deviation
  - In fact, that may be one of the many things we are trying to ______

Discuss: Quickly address why you won't regularly have the population parameters and under what circumstances would you have them?

- William Gosset was staff at the Guinness Brewery and, under the pen name of "Student" came up with the idea of Student's t-distribution to be used when the plain normal distribution didn't work
  - Use of Student's t-distribution is now the ______ way to calculate confidence intervals for means

3.2 More on the t-distribution

- ______ steps:
  - Draw a simple ______ sample of size n
  - Population of ______ distribution with parameters mean μ and standard deviation σ
  - Convert ______ of the distribution to t-scores with $\frac{\bar{x}-\mu}{(\frac{s}{\sqrt{n}})}$
- Results in a Student's ______ with $n-1$ degrees of freedom (df)
  - The individual t score of a mean is understood as the distance that $\bar{x}$ is from μ, like a ______
- Degrees of freedom are a tricky concept, but come up ______ in most inferential statistical formulas

  - When calculating standard deviation for a sample, we calculate n number of ________ and because we know that the sum of all deviations is 0 (if we don't square them, like in the variance formula), the last deviation cannot vary.
  - Thus, all deviations (n) can vary until this ________ one, leading to $n-1$ as the degrees of freedom in this case

> **Important**
>
> Perhaps the most defining feature of the t-distribution that separates it from the normal distribution is that it changes with the sample size

- The ________ for the t-distribution is given at $T \sim t_d f$
  - Where $df = n-1$, as previously discussed

3.3 Calculating with t-distribution

- In the case that we ________ know the population standard deviation, we can use the t-distribution in the following way
- $EBM = (t_{\frac{\alpha}{2}})(\alpha s\sqrt{n})$
  - $t_{\frac{\alpha}{2}}$ is t-score with area to the riqht of $\frac{\alpha}{2}$
  - s is the sample standard deviation ________
  - df is degrees of freedom of $n-1$

> **Important**
>
> Calculation with the t-distribution has to proceed with either a table or a calculator / computer - we'll cover what to do in SPSS

4 A Population Proportion

4.1 Introduction

Important

I'm going to abridge this part a bit only because it get's a bit too 'in the weeds' for what we need to know

- Often, we are interested not just in the confidence interval of ________, but also of percentages and proportions
- First, we need to identify a situations that is more proportion-based than mean-based
  - In these scenarios, the underlying ________ is the binomial distribution

Discuss: For review: write the notation formula to describe a binomial distribution and describe the characteristics of that distribution

4.2 Calculation of a Proportion CI

- Because of the underlying ________, we can take a proportion as: $P' = \frac{X}{N}$
  - Where:
  - X is a random variable of number of ________
  - n is the ________ of trails/sample size
  - P' is also sometimes shown as $\widehat{P}$
- In the scenario that n is particularly ________ and p is not close to zero or one, we can actually treat this as normal like: $X \sim N(np, \sqrt{npq})$

- This all eventually works down to $\frac{X}{n} = P' \sim N(\frac{np}{n}, \frac{\sqrt{npq}}{n})$
  - The second part after the comma is the treated as the "standard" error for ____________
- One can calculate error bound for a proportion (EBP) to work towards a sort of confidence interval, but for proportions, just like we did for ____________

5 Conclusion

5.1 A Plea for Confidence Intervals

- I am a big advocate of the value of confidence intervals in a lot of situations - I think they appropriately capture that these statistics really are just estimates/guesses!
- In my opinion, it is pretty much always appropriate to report confidence intervals alongside the point estimates of just about anything you calculate - it increases information and transparency

5.2 Recap

- Confidence intervals add valuable information about the relative spread and inaccuracy of sample statistics in means and in proportions
- Confidence intervals pretty much always become more narrow with larger sample size, continuing the trends demonstrated with the central limit theorem and the law of large numbers
- We introduced the t-distribution as a valuable way to calculate confidence intervals for means, even when lacking the population standard deviation

Key Terms

A

alpha The complement of the confidence level, effectively the percent of area that falls outside the interval estimate 6

C

confidence interval The common form of an interval estimate used in inferential and traditional (frequentist) probabilistic statistics, suggests a range of values that a population parameter is likely to fall in given a certain level confidence 3
confidence level (CL) An arbitrarily set level of how certain one wants to be that the population parameter falls within a specified range 6

D

degrees of freedom (df) A measure of how many values can vary in a distribution's calculation (don't worry too much about this one) 9

E

error bound for a population mean (EBM) How far away from the point estimate which captures the middle some percent values determined by the confidence level 6
error bound for a proportion (EBP) Functions as sort of a 'appropriate standard deviation' for point estimates of proportions 12

I

inferential statistic A type of statistic meant to help infer characteristics of population or populations in relation to the sample statistics 3
interval estimate In frequentist statistics, sits as an inverse to a point estimate and gives a range as an estimate of the parameter 4

M

margin of error An alternative name for confidence intervals, used in polling often 4

P

point estimate A type of sample statistic that exists only as a single fixed value trying to represent the population parameter 2

R

reliability In this context, refers to how 'stable' the statistic is when taken many times 4

S

Student's t-distribution A distribution similar to the normal distribution but that also takes into account sample size, appropriate for use when the parameter standard deviation is not known 9

W

William Gosset Chemist and statistician at Guinness that came up with the t-distribution, wrote under the name of 'Student' 9

The instructor-provided glossary may not include all terms worth memorizing, make sure you consider using the vocabulary list in you book and your own judgment to make sure you have all relevant terms