BALL STATE
UNIVERSITY

Module 7 Lecture - Factorial/Two-Way 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

- Recognize central limit theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the central limit theorem for means.
- Apply and interpret the central limit theorem for sums.

1.2 Instructor Learning Objectives

- Recognize the prevalence and importance of the mean
- Understand how the central limit theorem contributes to the accuracy of statistics in representing/estimating parameters.

1.3 Introduction

- The concept of the mean _____ has come up a lot in our previous lectures, from discussions on descriptive statistics through probability and random variables
- We are going to connect _____ (and sums) with the **central limit theorem (CLT)**, which is critical in understanding how our sample statistics connect with our population parameters
  - Thus, the central limit theorem is closely related to the law of _____ numbers - this will be touched on towards the end of lecture

> **Important**
>
> Much like the prior 'ideal' distributions we've discussed, the CLT is a conceptual idea used as an assumption for practical application. It is not often 'calculated' in sample data.

- The central limit theorem has two _____, both which converge on roughly the same concept:
  1. If we collect infinitely many samples of size n that are "_____ enough" calculate means from each of those samples on some continuous variable of interest, and plot the means as a histogram, it will be roughly _____
  2. Roughly the same idea, but instead of taking the mean of each sample, we take the _____ of the continuous variable and plot

those as a _____. Similar to before, theoretically results in a normal distribution.

- Put more succinctly: if we take enough large samples, the sample statistic estimates are normally distributed _____ the samples.

> Discuss: Quick Review: What two notations are used to represent a sample mean statistic and a population mean parameter? What about sample and population standard deviation? What greek letter is associated with summation?

- As a rule of thumb, $n = 30$ is often said to be sufficiently large if it is believed that the distribution of a variable is _____ in the population, but more is necessary if normality cannot be assumed.
  - A nuance to this is that the $n = 30$ rule applied to separate _____, e.g., if I wanted to compare men and women in my analysis, I would want at least 30 of each of those genders
  - Some analyses like _____ modeling benefit from much larger samples
  - This is also done under the assumption that sampling of a sample from a population is done _____ replacement.

> Discuss: Try to recall and describe the difference between sampling done with and without replacement. How does this connect to (in)dependence?

> Important
>
> There is some controversy on just how much larger your samples need to be to overcome non-normality in the sample; some even say we shouldn't bother!

- The following two sections we cover will describe and investigate the two _____ versions of the CLT: Focusing on means and sums

2 The Central Limit Theorem for Sample Means

2.1 Introduction

Discuss: Prior to this next part, try writing notation to represent a continuous variable

- The central limit theorem around means can be represented in a similar _____ to distribution like we used previously in Modules 4 and 5
  - This is because the central limit theorem hinges around a _____ of sample means
- The formula to represent the central limit theorem of means is: $\bar{X} \sim N(\mu_X, \frac{\sigma_X}{\sqrt{n}})$
- Where:
  - $\bar{X}$ is a random continuous variable consisting of many _____ means of variable X
  - μ_X is the mean of random _____ variable X
  - σ_X is the standard deviation of random continuous variable X
  - n is the size of the _____ samples taken to from the sampling distribution (this is consistent across all theoretical samples)

Important

All of the above Xs are uppercase! This is because we are using it to describe a variable, not the possible values that the variable could take, which would be lowercase x.

Discuss: Try interpreting the above notation like you would any other normal distribution, like the one you wrote in the last 'Discuss' question. Effectively, try to write out what each part of the notation means.

2.2 Sampling Distributions

- When we represent _____ sample means as a distribution, we call this the **sampling distribution** of means, which is really what is shown in the prior notation
  - A _____ standard deviation in this sampling distribution would suggest that the individual sample means are _____ spread out, i.e., each sample is more different from one another
  - Thus, a small standard deviation in the sampling distribution is _____, as it indicates that each individual sample is relative close in its mean estimate
- The CLT holds that the larger a sample size, the more its sampling distribution will _____ a normal distribution
  - This hinges on n being large enough
  - Consider the formula again: $\bar{X} \sim N(\mu_X, \frac{\sigma_X}{\sqrt{n}})$
  - n being in the _____ means that a larger n will result in a smaller overall standard deviation for the distribution
  - This $\frac{\sigma_X}{\sqrt{n}}$ is referred to as the **standard error of the mean**

Important

You can think of the standard error as being the standard deviation of the sampling distribution, or how far away, on average sample means are away from the mean of sample means.

? Earlier, what was the rule-of-thumb number for what is 'large enough' sample?

- (●) A) 20
- ( ) B) 10
- ( ) C) 100
- ( ) D) 30

Explanation:

2.3 Z-scores for Sampling Distributions

- We can conceptualize the mean of a _____ sample as a z-score for how far away it is from the theoretical mean of the sampling distribution with:

$$z = \frac{\bar{x} - \mu_X}{(\frac{\sigma_X}{\sqrt{n}})}$$

- Where:
  - $\bar{x}$ is the mean for a single sample
  - μ_X is the mean of both X and $\bar{X}$
  - σ same as before

2.4 Connection to Finding Probability in the Normal Distribution

- Because the sampling distribution is theorized as a _____ distribution, we can estimate the probability that a certain event regarding an individual sample mean occurs
  - In order to do so, we need some information about the believed mean (μ_X), standard deviation (or standard _____; σ_X), and sample size (n) of the particular sampling distribution

2.5 Educational Example

- A standardized elementary reading test for children produces scores that range from 0 to 500, which represent student ability in reading for that grade

level
  - Treat R as representing the continuous random variable of score on the reading test
- Based on the test manufacturers specifications, this test has a mean sampling distribution mean of 250, and standard deviation of 50. In my scenario, let's assume I take samples of $n = 40$.
  - My sampling distribution could be represented as:

$$\bar{R} \sim N(250, \frac{50}{\sqrt{40}})$$

- Where:
  - $\bar{R}$ is the mean of sample means of the reading scores
- The z-score formula then could be shown as:

$$z = \frac{\bar{r} - 250}{(\frac{50}{\sqrt{40}})}$$

- Where:
  - $\bar{r}$ is a single sample mean

3 The Central Limit Theorem for Sums

3.1 Brief Overview

> **Important**
>
> I am going to focus less on this one, only because it ends up following a very similar idea and logic to the mean CLT.

- The central limit theorem for _____ functions on a very similar logic to the CLT for means, just with a different formula to represent the distribution:

$$\sum X \, N((n)(\mu_X), (\sqrt{n})(\sigma_X))$$

- Where:
  - σX is the continuous variable of sums of X
  - μ_X is the mean of X
  - σ_X is the standard deviation of X
- The z-score formula:

$$z = \frac{\sum x - (n)(\mu_X)}{(\sqrt{(n)}(\sigma_X))}$$

- The end idea of this version has a similar gist: larger n narrows the standard error of the distribution of sums.

4 Use of the Central Limit Theorem

4.1 Introduction

- In practice, the central limit theorem is not often _____ used, but is often appealed to as part of statistical _____
  - Most classic, _____ tests rely upon the implications of the central limit theorem
- The central limit theorem is best applied to _____ values, such as mean and sums, *not* individual points
  - E.g., I don't need to use the CLT to make sense of a single test score in a sample I take

> **Important**
>
> Much like the normal distribution in general, CLT is sometimes misused as providing an easy appeal, but it's not perfect!

4.2 Connection to the Law of Large Numbers

- If the central limit theorem implies that the _____ of the distribution of sample means is the population mean parameter, then a larger n allows our $\bar{x}$ to better estimate the μ.

> **Important**
>
> Remember that this is assuming that our sample is representative of the population of interest!

Discuss: What are some example of randomness-based sampling techniques that result in representative samples?

5 Conclusion

5.1 Recap

- The CLT serves as a critical concept building on probability and continuous, normal distributions
- The CLT underlies why larger samples result in more accurate estimates of our population parameters in our sample statistics
- The CLT implications and characteristics are useful in making sense of results from statistical hypothesis testing (more on this later!)

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