BALL STATE
UNIVERSITY

Module 6 Lecture - Multiple Comparisons for Kruskal-Wallis

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

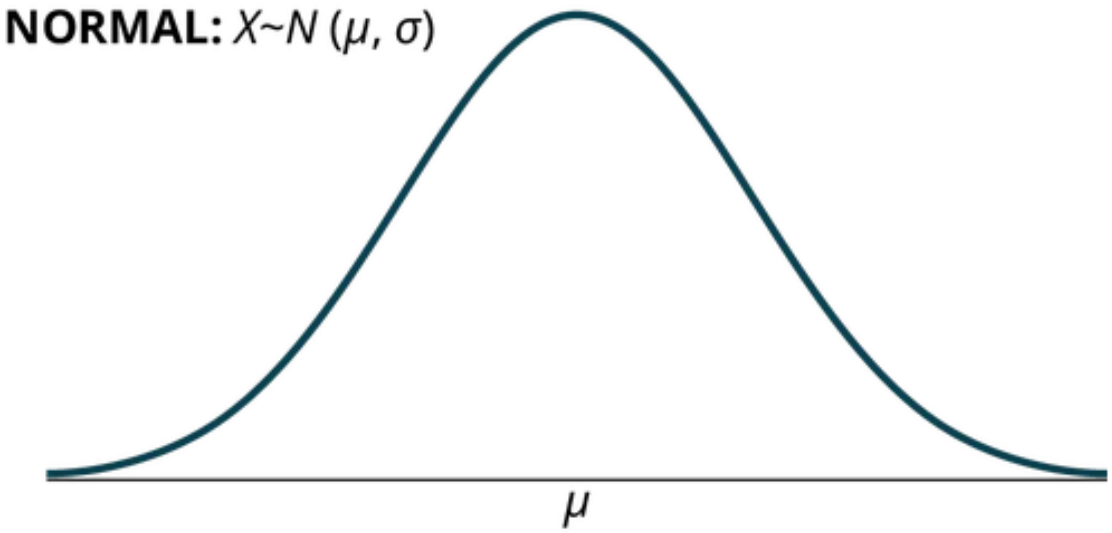
- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

1.2 Instructor Learning Objectives

- Understand the normal distribution within the broader context of "ideal" distributions for continuous variables
- Be able to calculate z-scores, and understand the role of the mean and standard deviation in the case of the normal distribution

1.3 Introduction

- The _____ distribution is arguably the single most prevalent way to describe continuous variables.
  - It is often called the _____ curve due to it's curved, symmetrical shape when plotted



- Like the _____ or the _____ distributions, the normal distribution is an "ideal"
  - "Real" data will _____ ever be perfectly normal
  - However, much like with the other ideal distributions the normal curve serves a purpose for _____

"All models are wrong, but some are useful" - George Box

- Unlike the prior distributions that have been discussed, the normal distribution is used often in common _____ statistics

2 Introducing the Normal Curve

2.1 Notation

- The normal curve's notation is similar to others: $X \sim N(\mu, \sigma)$ where:
  - X is the _____ variable
  - N is the designation of the _____ curve
  - μ is the population _____ parameter
  - σ is the population standard _____ parameter
- Thus, if we assume that a normal distribution has a mean of 20 and a standard deviation of 2, this would be written as: $X \sim N(20, 2)$

Discuss: Now you try it: write the notation for a normal distribution with a mean of 12 and standard deviation of 3.

Discuss: Review: Try writing notation for an exponential distribution with a decay parameter of 0.01 and a separate notation for uniform distribution with minimum value 2 and maximum value 20.

2.2 Probability Density Function

- The probability _____ function (pdf) for the normal curve is:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-0.50 \cdot (\frac{x-\mu}{\sigma})^2}$$

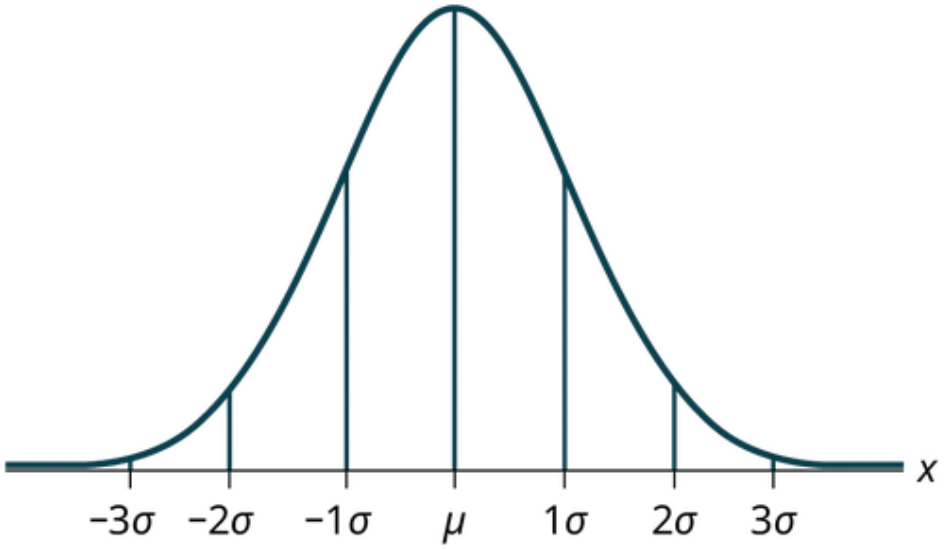
- Compare this to the relatively _____ pdf for a uniform distribution:

$$f(x) = \frac{1}{b-a}$$

- Due to the complexity of this equation, we often don't calculate _____ under the curve (AUC) for normal distributions by hand
  - Historically, one can use _____ that approximate the AUC, but in modern practice, this is handled fully by computers
  - We'll introduce the concept of calculating probability by hand in Calculating AUC for the Normal Distribution, but later will use SPSS to handle this for us

2.3 Basic Characteristics

- The normal curve can _____ in appearance quite a bit, as:
  - Changes in _____ move the curve to the left and right
  - Changes in standard deviation can make it more _____ or _____
- However, it is:
  - Always _____
  - The mean, median, and mode of the distribution are all the _____
  - Following the _____ rule
- The **empirical rule**, also known as 68-95-99.7 rule, states that:
  - 68% of x values lie between -1σ and 1σ or $z = -1$ to 1
  - 95% of x values lie between -2σ and 2σ or $z = -2$ to 2
  - 99.7% of x values lie between -3σ and 3σ or $z = -3$ to 3



3 Calculating AUC for the Normal Distribution

3.1 Introduction

- As mentioned prior, the _____ of the normal distribution is often calculated with computers due to complexity
  - This is often the case with the _____ and exponential distributions as well - simply too time consuming and complex to calculate for large datasets or complex scenarios

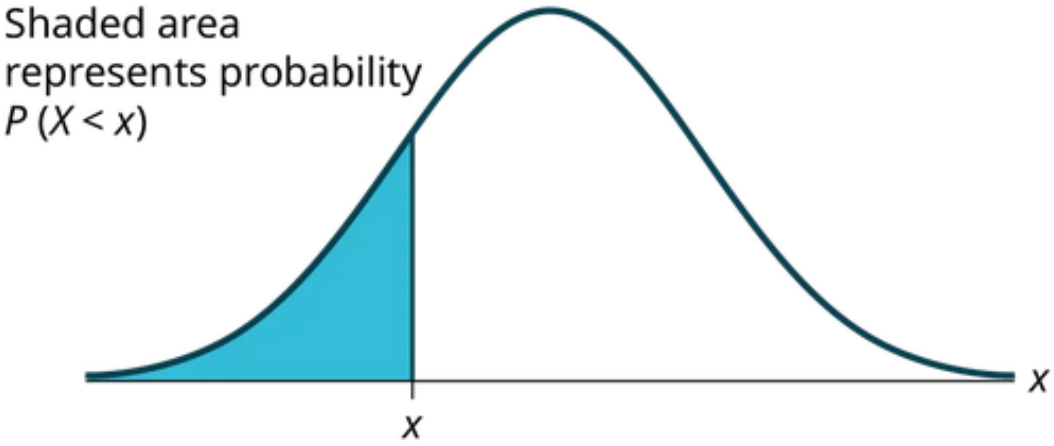
Important

Its worth mentioning that the field of statistics has rapidly grown in possible methods with advancements in technology and computing

Discuss: Review: Under what circumstances can we use a binomial distribution?

3.2 Calculating

- For the following applications X and x mean the same thing they have in previous modules



- To find area to the _____ of a specified point x, we find $P(X < x)$
  - Thus, to find the complement or inverse or area to the _____ of specified point x, we will do $P(X > x) = 1 - P(X < x)$
  - Remember that individual points of x don't have _____ in a continuous distribution, so this is functionally the same as saying $P(X \leq x)$ for area to the left
- Your book shows how to perform these calculations using a calculator function, but we will use SPSS in the next practical assignments

4 The Standard Normal Distribution

4.1 Introduction and Z-scores

- There is a special case of the normal distribution with more clearly specified characteristics: the _____ **normal distribution**
- The standard/standardized normal distribution is made up of _____, instead of whatever "raw" continuous values would be used

Discuss: Review from descriptive statistics before the next part, what is a z-score? Explain in your own words

- Review: Z-scores can be practically _____ as, "number of standard deviations a point is from the mean of the data"
  - Thus, a data point with a $z = +2.00$ is 2 standard deviations _____ the mean and a data point with $z = -1.20$ is 1.2 standard deviation _____ the mean
  - A data point directly at the mean of the data will have $z = 0.00$.
  - Z-scores are best understood as _____ indicators of position within the dataset

Important

If you ever hear that data was 'standardized' or 'mean-centered', that means that each data point was transformed into its respective z-score.

- Review: the _____ for z-scores in a sample is $z = \frac{x - \bar{x}}{s}$
  - Application example in data of $1, 2, 3, 4, 5$
  - $\bar{x} = 3$
  - $s = 1.581$
  - To get z-score of 4: $z = \frac{4-3}{1.581} = 0.633$
- To determine what value a particular z-score is (in a sample) you can use: $(z * s) + \bar{x}$
  - Taking from example above to find what point that $z = 1.5$:
  - $(1.5 * 1.581) + 3 = 5.37$

Important

Remember there is a subtle difference in notation and formulas for sample statistics vs population parameters!

Discuss: Rewrite this above problem and recompute using the population parameters instead of sample statistics

- One notable benefit of z-scores is that they allow us to compare variables on the same _____
  - However, a drawback along this same line is that interpreting z-scores in practical analysis is more _____ from a realistic interpretation.

4.2 Notation

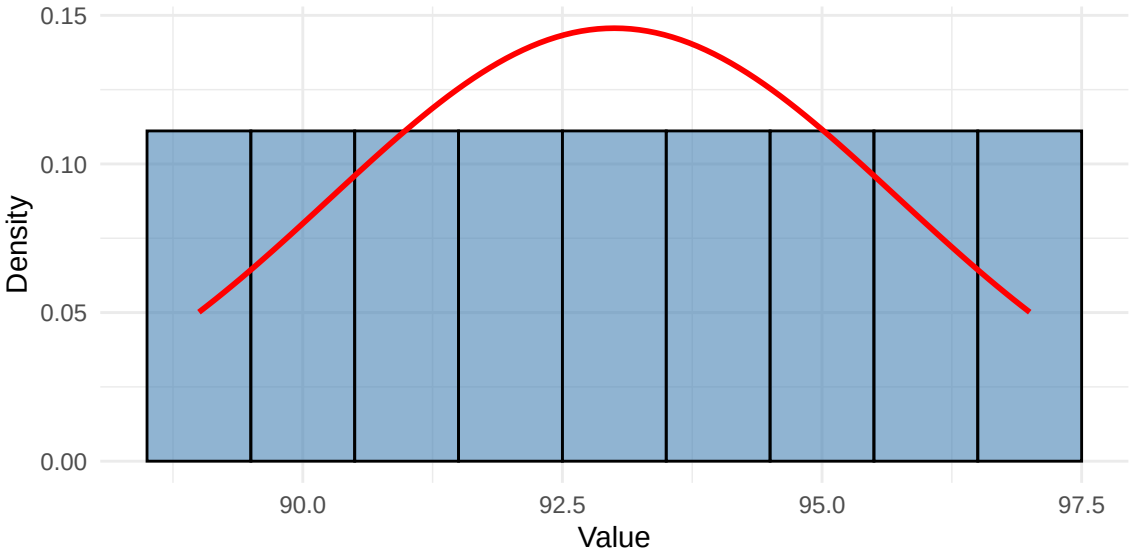
- Based on the _____ of z-scores, a standard normal curve always takes the notation $Z \sim N(0, 1)$
  - This is because a set of z-scores from a normally-distributed variable will always have $\mu = 0$ and a $\sigma = 1$

5 Brief Example of Data 'Coercion'

5.1 Introduction

- One _____ is that one can readily treat any continuous data as normal - this is not wise
- Take for example class grade percentage data: {93, 92, 91, 90, 89, 94, 95, 96, 97}
  - Treated "as usual", this would mean lots of As and a B+
  - But, what if we treat this data as if it comes from a normal distribution?

Figure 1: Class Grade Percentage Histogram



Discuss: Instead of the normal distribution, what other continuous distribution does this plot remind you of?

- If I were to "grade on a curve", it means the person who got an 89 would be graded as if they failed, because they were lower _____ to the other data points
  - This is _____ only if the assumption that the scores of all students is perfectly normal, but feels unfair if that assumption isn't met
  - On the other hand, this system _____ help if all students did poorly, but some did just marginally better
- We'll revisit the idea of improper application of _____ when we cover specific inferential tests

6 Conclusion

6.1 Recap

- The normal distribution is a useful and commonly used continuous variable distribution that meets several specific conditions. These characteristics make it readily predicable and applicable
- Much like the other distributions and density functions, we can use the characteristics of the normal distribution to calculate AUC and understand the relative spread and placement of the data. This is aided by use of z-score and the standard normal curve as a special case
- *However,* it is often mis-used and misunderstood, and we must take caution as we continue in the semester

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