Module 5 Lecture - Non-parametric Comparisons for More than Two Groups

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 and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately.

1.2 Instructor Learning Objectives

- Compare and contrast the use-cases for a discrete vs. continuous distributions
- Understand the relatively high prevalence of continuous random variables in research, and use this module as a foundation for understanding those characteristics

1.3 Introduction

> Discuss: What two scales-of-measurement can be readily considered to be continuous? Hint: NOT nominal or ordinal, but...

- Many variables that are used in educational and social science research are ________
  - E.g., GPA, SAT/ACT/GRE - standardized educational scores, income after graduation, Stanford Binet/Woodcock Johnson/WISC - cognitive test scores
- Like with discrete variables, continuous variables often play a role in ________, which deals with construction of tests and assessments

> Important
>
> We'll continue to use notation of uppercase letters to represent random variables and respective lowercase letters to represent possible values the random variable can take on. Same notation as with discrete variables.

1.4 Properties of Continuous Probability Distributions

- Unlike __________ variables that dealt purely with counts and separate integers, continuous variables must be treated in a different way, because they have possible values in-between integers

Discuss: Try making a probability distribution function table like from the last module - but with possible GPA values as x.

- When working with continuous random variables, we will instead use the term **probability density function (pdf)** to describe the __________, or $f(x)$, that shows the distribution graphically.

Important

For those that will later deal with programming languages like R or python, f(x) is a common shorthand for function in computer science.

- We often aim to calculate the **Area-under-the-curve (AUC)** using the **cumulative distribution function (cdf)**.
- The cumulative distribution function follows these rules:
  1. Outcomes are __________, not counted like discrete variables
  2. The entire AUC is equal to 1
  3. Probability is found for __________/ranges of x-axis values, not individual points
  4. $P(c < x < d) \rightarrow$ __________ that random continuous variable X falls between arbitrary points c and d, i.e., the AUC between c and d
  5. $P(x = c) = 0$, $P(x = d) = 0$, etc. $\rightarrow$ the probability that x equals a single point is 0, as a single point has no width, i.e., no AUC because there is no area
  6. $P(c < x < d) = P(c \leq x \leq d)$, as probability is __________ to AUC, and following the logic of the last rule, single points on the x-axis have no area

Discuss: Review: Early in the course, we discussed the subtle but important differences between where a bar plot is applied and where a histogram is applied, try explaining that difference in your own words now.

- Recall that bar plots were only applicable for discrete or qualitative/categorical/nominal data, but doesn't really work for ________________ data
  - Much like the differences between probabilities of discrete data (from last unit) where we could represent probabilities for individual values, like in a bar plot.
  - But with continuous variables, we work in ________________.

Table 1: Discrete Probability Distribution Example (Bar group)

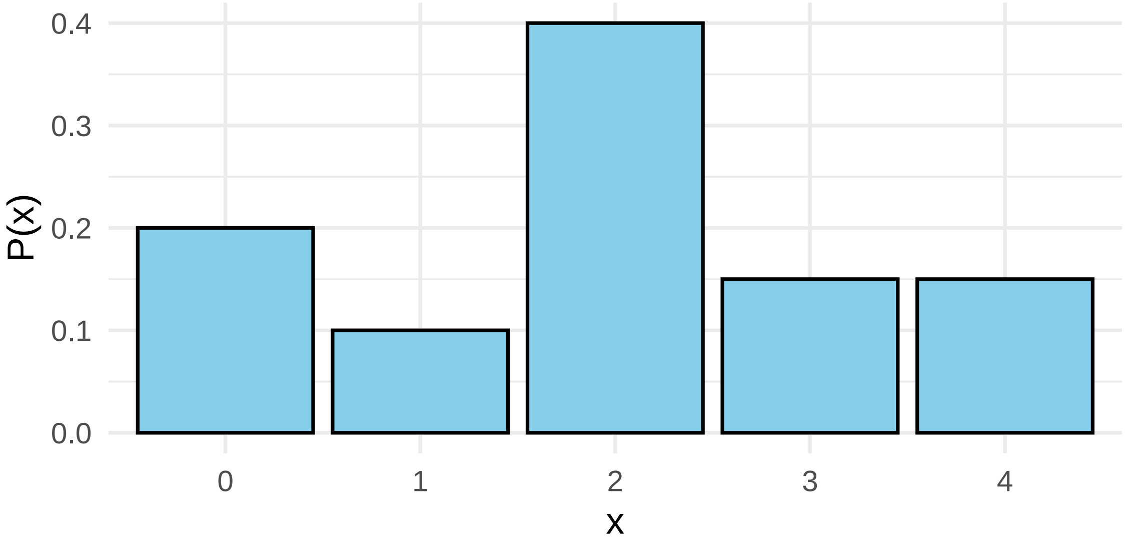
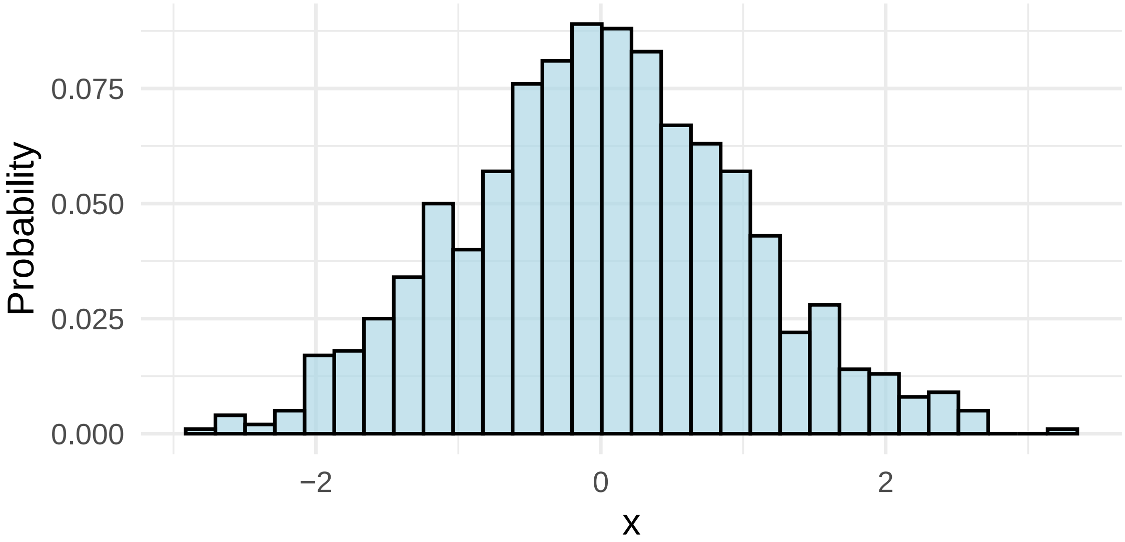


Table 2: Continuous Probability Density Function Example (Histogram)



- Like with many things in statistics, there are complex and robust mathematical proofs for these rules, rooted in ________________ - but that is FAR outside the scope of this course.

2 Continuous Probability Functions

2.1 Introduction

- Compared to the ________________ of discrete probabilities, it is usually easier to conceptualize continuous probabilities using a histogram
- For more basic examples: basic ________________ of width * height works

> **!** Important
>
> Keep in mind that f(x) is just a formula for line/curve to be plotted! Follows regular rules of geometry

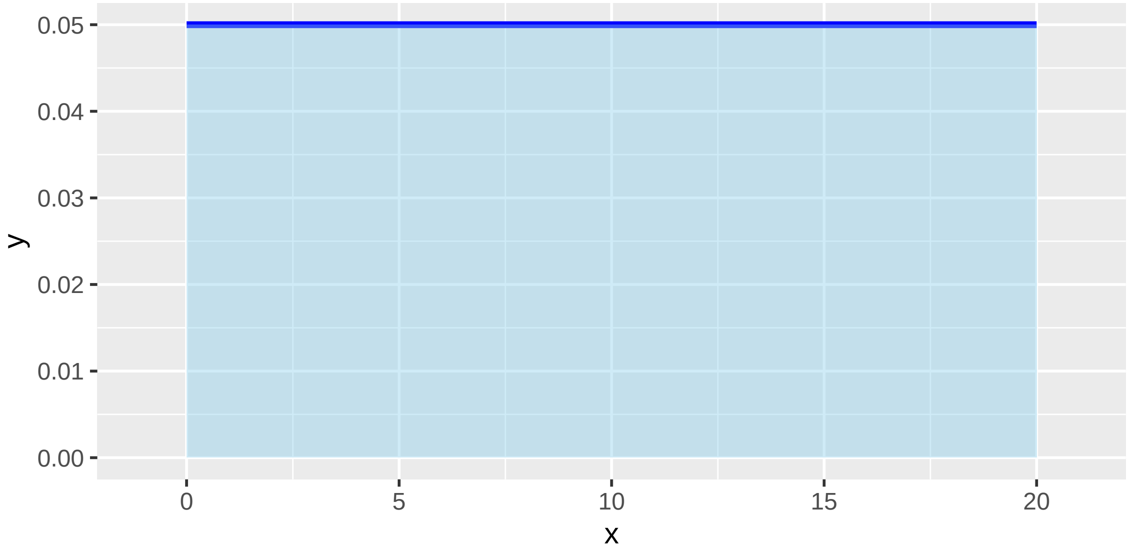
2.2 Start of Examples

- The following examples are simpler to understand and visualize because the “curve” of $f(x)$ here is just a straight line.

- It is often useful to start with simply identifying the ________________ of values, from the minimum possible value to the maximum possible value

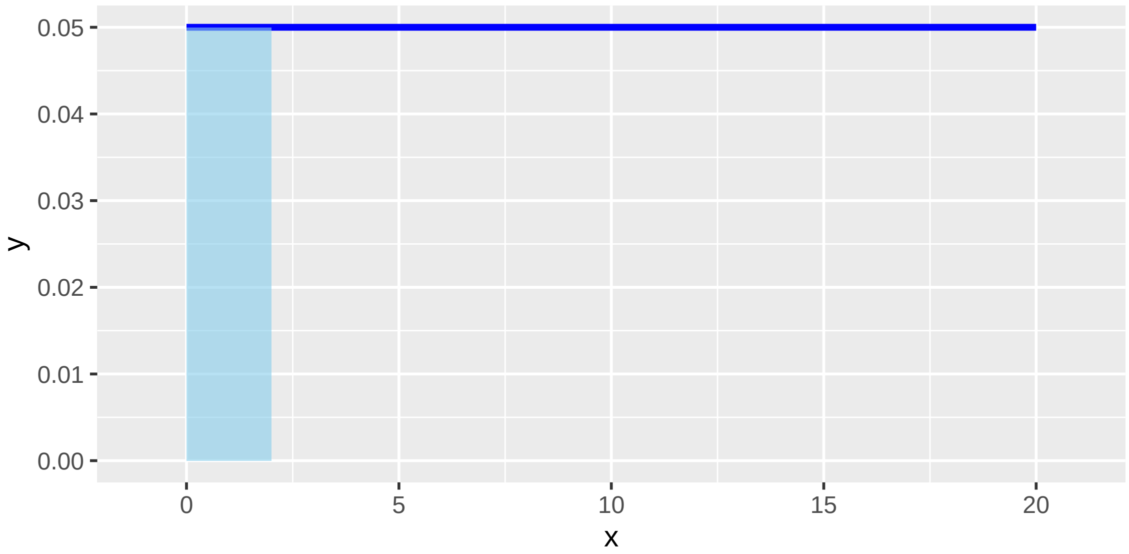
2.3 Very Simple Example

- $f(x) = \frac{1}{20}$ for event of $0 \leq x \leq 20$
- Using normal geometry: $AREA = (20 - 0)(\frac{1}{20}) = 1.00 \rightarrow$ 100% probability of X falling between 0 and 20



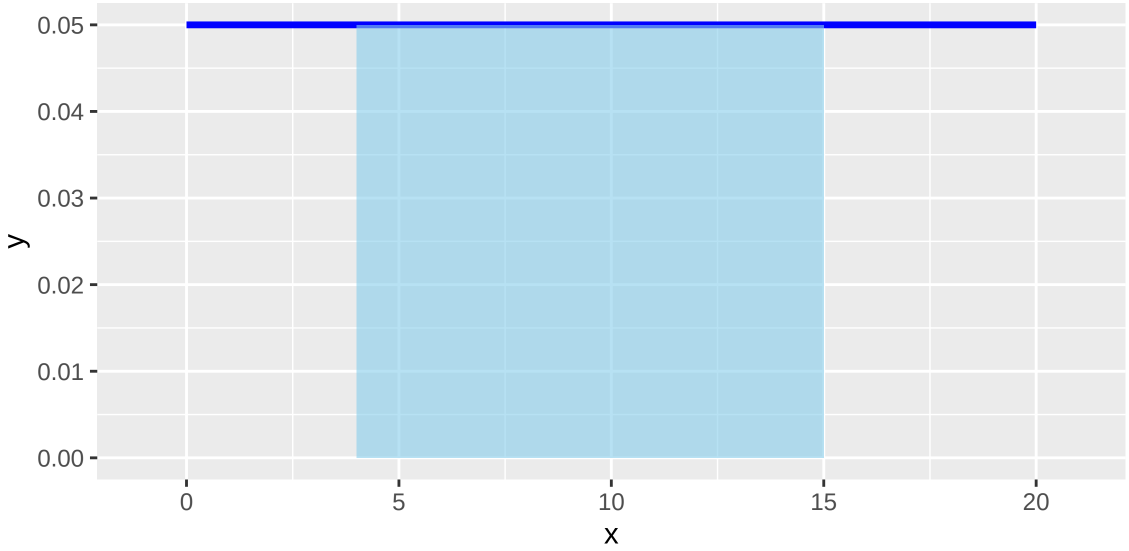
2.4 Example of Partial Range

- $f(x) = \frac{1}{20}$ for event of $0 \leq x \leq 2$
- Using geometry again: $AREA = (2 - 0)(\frac{1}{20}) = 0.10 \rightarrow$ 10% probability of X falling between 0 and 2



2.5 Example of Middle Range

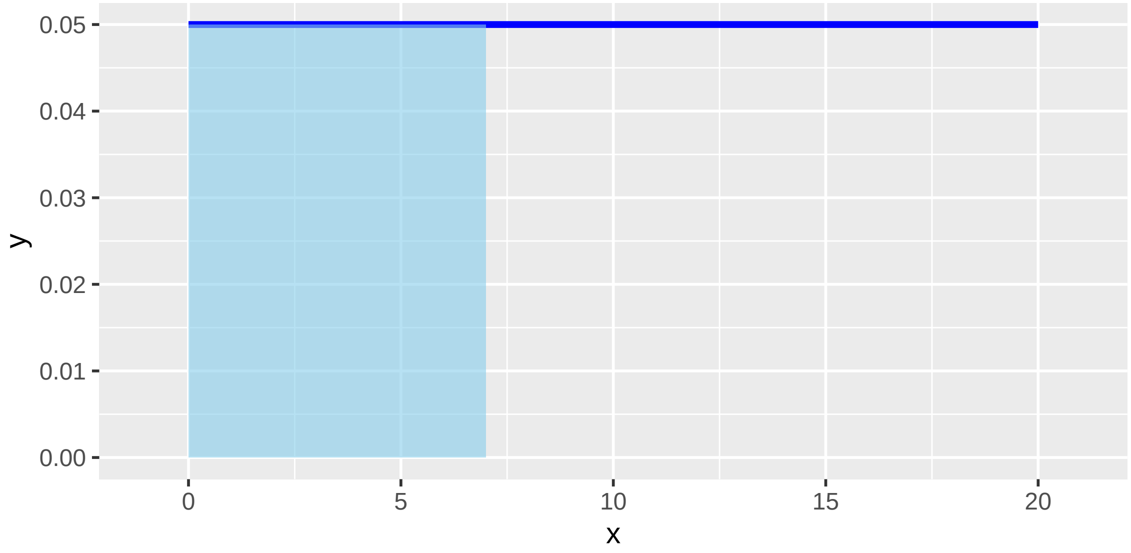
- $f(x) = \frac{1}{20}$ for event of $4 \leq x \leq 15$
- Using geometry again: $AREA = (15{-}4)(\frac{1}{20}) = 0.55 \rightarrow$ 55% probability of X falling between 0 and 2



Discuss: Using the prior example of f(x) = 1/20, find the probability where X is greater than 2 but less than 12. Represent this both with math and a graphical representation like the charts above.

2.6 Using Graphical Method to Find CDF

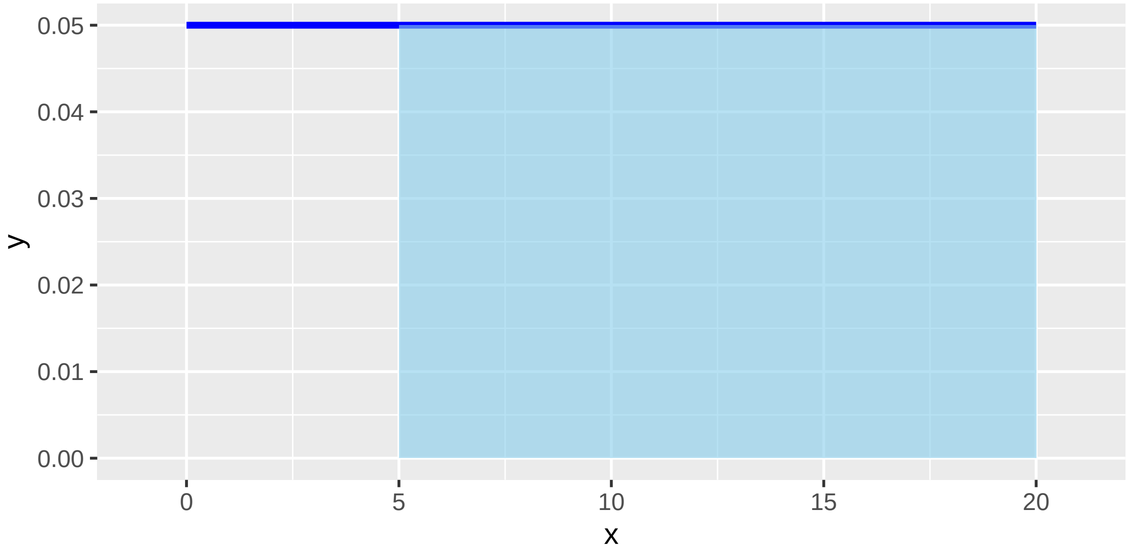
- The ________ distribution function (CDF) is one that allows us to find the area "up-to" or to the left of a certain point with:
  - $CDF = P(X \leq x)$ where x is some chosen point on the x axis
  - Equivalent to $AREA = (x - MIN)f(x)$ where MIN is the minimum value in the data
  - Example: $P(X \leq 7)$ gives the AUC to the left of the point at 7 on x-axis
    - Applied to above examples: $AREA = (7-0)\frac{1}{20} = 0.35 \rightarrow$ 35% chance that X falls somewhere before or left of 7



- The CDF can also be used in the ________, or the area to the right of the chosen point
  - $P(X \geq x) = 1 - CDF$
  - Equivalent to $AREA = (MAX - x)f(x)$ where MAX is the maximum value in the data
  - Example: $P(X \geq 5)$ gives the AUC to the right of the point at 5 on

x-axis

- Applied to above examples: $AREA = (20 - 5)\frac{1}{20} = 0.75 \rightarrow$ 75% chance that X falls somewhere after or right of 5



📢 Discuss: Using the prior example of f(x) = 1/20, find the CDF where X is less than 14. Now, do the inverse, where X is more than 14. Make sure your probabilities sum to 1. Represent this both with math and a graphical representation like the charts above.

2.7 Conclusion of Section

- Like with discrete data, we can estimate probabilities of certain data occurring, but take a more visual approach
- We will cover three distinct, different patterns of continuous probability functions, just like we did with the __________ distribution in the last unit
  - __________ distribution
  - __________ distribution
  - __________ distribution (next lecture!)

"I don't mind not knowing. It doesn't scare me." — Richard P. Feynman

3 The Uniform Distribution

3.1 Introduction

- **The uniform distribution** starts with the assumption that all points within a range are ____________ likely to occur
- So far, all of the examples prior have shown a uniform distribution, as indicated by the ____________ horizontal line, parallel to the x-axis
- It follows the notation: $X \sim U(a, b)$ where:
  - $X \rightarrow$ random variable
  - $U \rightarrow$ notation indicating uniform distribution
  - $a \rightarrow$ minimum value of x (possible values)
  - $b \rightarrow$ maximum value of x (possible values)
- From here, we can construct a probability density function (pdf) as $f(x) = \frac{1}{b-a}$
  - Where both b and a are taken from the previous notation
  - Example: pdf of $f(x) = \frac{1}{21-2}$ written as $X \sim U(2, 21)$
  - Example: notation of $X \sim U(12, 20)$ written as pdf $f(x) = \frac{1}{20-12}$

Discuss: Review: What was the notation for the binomial distribution, and what did each of the letters in that formula mean?

3.2 Expected Mean and Standard Deviation

- Like with discrete random variables, we can do estimates for the expected long-term mean and standard deviation
  - In the following "shortcut" formulas, use the prior definitions of a and b
- Mean: $\mu = \frac{a+b}{2}$
- Standard Deviation: $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Discuss: Do you recall why we take the square root in the standard deviation formula?

3.2.1 Example of Calculating Mean and Standard Deviation

- Given $X \sim U(90, 120)$ that means:
  - $f(x) = \frac{1}{120-90}$ with $a = 90$ and $b = 120$

Mean:

$$\mu = \frac{90 + 120}{2} = 105$$

Standard Deviation:

$$\sigma = \sqrt{\frac{(120 - 90)^2}{12}} = 8.66$$

4 The Exponential Distribution

4.1 Introduction

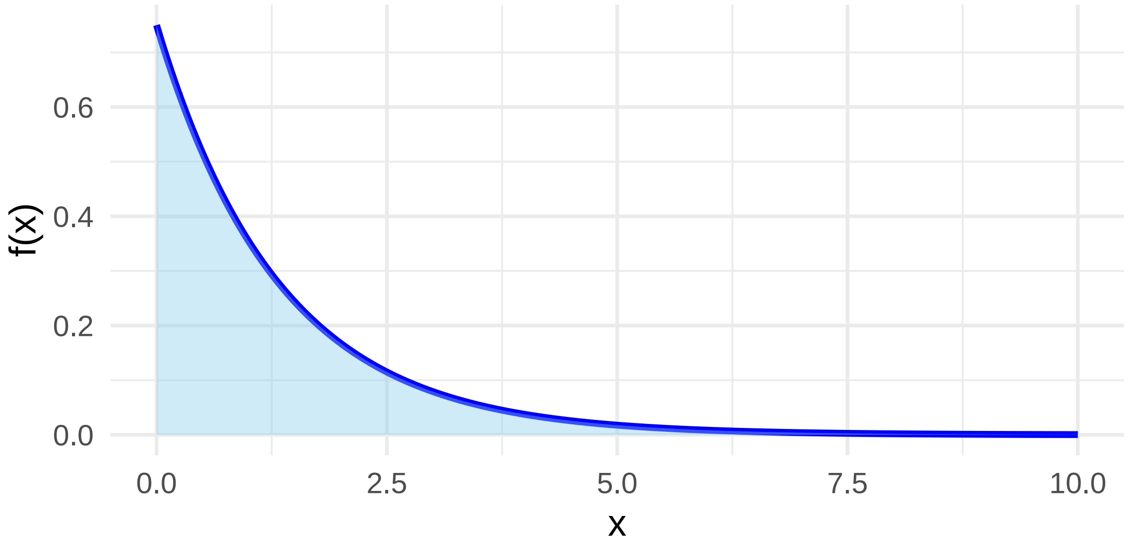
- Variables that follow an exponential distribution are those with relatively low probability of ________________ values and high probability of ________________ values.

Important

Do not try to use the ’shortcut’ formulas for the uniform distribution above for the exponential distribution

- In an exponential distribution, you must also have m (also sometimes represented as λ), or the **decay parameter/rate parameter**.
  - Formula: $m = \frac{1}{\mu}$

  - We can also work this backwards if we are given m in order to get the expect mean: $\mu = \frac{1}{m}$
  - Expected standard deviation: $\sigma = \mu$
- The formula for exponential distributions is $X \sim Exp(m)$
  - Following from this, we find the probability density function as $f(x) = me^{-mx}$ where e is the scientific constant = 2.71828…
- Example: For $X \sim Exp(0.75)$:
  - $m = 0.75$
  - $\mu = \frac{1}{0.75} = 1.3\bar{3}$
  - $f(x) = 0.75e^{-0.75x}$
- The **cumulative distribution function (CDF)** give area to the left of a determined point
  - Formula: $CDF = P(x < x) = 1 - e^{-mx}$
  - In this equation, we'll replace x with whatever x-axis value we'd like to use
  - To find $P(x > x)$ we will do $1 - P(x < x)$



5 Conclusion

5.1 Recap

- Continuous variables are common outcomes and predictors in educational and social science research - it is important we can accurately describe the probability and structure of this data

- Much like with discrete random variables, we can represent continuous random variables with probability functions, albeit with a different procedure than the tabular format of probability distribution functions (PDF) as introduced in module 4

- Just like we learned about the binomial distribution pattern for discrete variables, we can apply specific patterns for continuous random variable as well: here, we covered the uniform and exponential distributions which can be helpful in understand some natural continuous variables. These patterns are "ideal" tools that we can use in analyzing our data.
- However, we still have yet to cover arguably the most important pattern of continuous distribution, the normal distribution, which will be covered more in the next chapter.

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