

EXAM 1 REVIEW

Exam 1 covers content covered in Lessons 5-14, HW 1-6 and Quizzes 1-5, (while there may be coding questions related to probability and/or random variables, there will not be any questions specifically about Pandas syntax or coding in Pandas).

Learning Objectives in Scope of Exam 1

Practice these learning objectives using [active recall \(click here for a description\)](#).

1. Data Visualization (Lessons 5-6)

- (a) Categorize data by its variable type (quantitative discrete, quantitative continuous, qualitative ordinal, qualitative nominal)
- (b) Choose the correct data visualization to use based on the variable type(s).
- (c) Define what percentiles, IQR, mean and median are for a given quantitative variable
- (d) Analyze histograms and identify the skew, potential outliers, and the mode.
- (e) Analyze boxplots and violin plots and identify percentiles, IQR and outliers.
- (f) Use Python functions to visualize distributions of Qualitative Variables using bar charts.
- (g) Use Python functions to visualize distributions of Quantitative Variables using histograms, KDE curves, boxplots and violinplots
- (h) Use Python functions to visualize relationships between variables (using scatterplots, overlaid histograms, boxplots/violin plots and/or bar plots as appropriate)

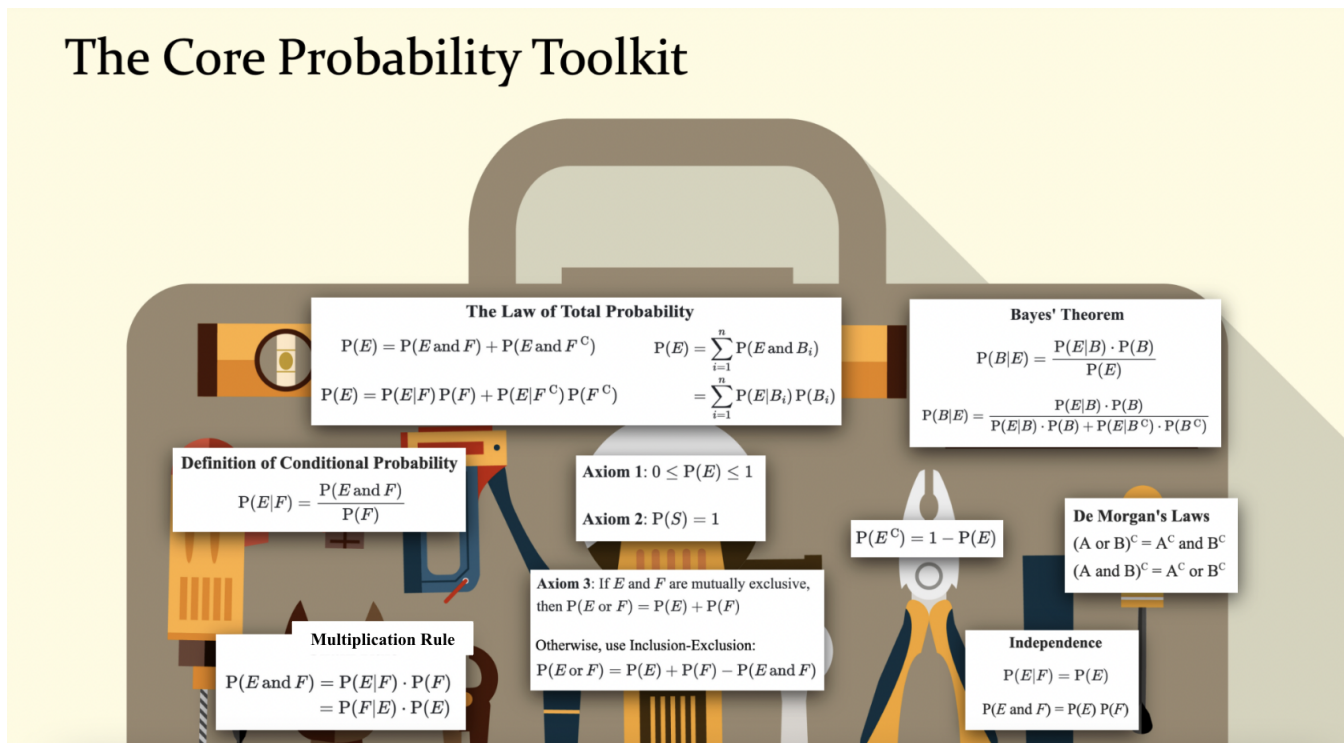
2. Probability (Lessons 7-10)

- (a) State the Frequentist definition of Probability
- (b) State the axioms of probability
- (c) Calculate probabilities for equally likely outcomes
- (d) State and apply the Inclusion/Exclusion Principle
- (e) Translate given information about events into a joint probability table
- (f) State the mathematical definition of conditional probability and use it to calculate conditional probabilities.
- (g) State the multiplication rule and use it to calculate joint probabilities.
- (h) Use Numpy to simulate probabilities.
- (i) Translate given information about events into a joint probability table and/or probability tree
- (j) Define and apply the Law of Total Probability
- (k) Define and apply Bayes' Theorem to update probabilities
- (l) Calculate total probabilities, conditional probabilities and joint probabilities using the appropriate technique(s)
- (m) State the mathematical definition of independent events
- (n) Determine whether two events are independent using the mathematical definition

3. Random Variables (Lessons 11-14)

- (a) Give the mathematical definition of a random variable
- (b) Explain the difference between random variables and events
- (c) Classify random variables as either continuous or discrete
- (d) Explain what a probability mass function (PMF) is and use it to calculate probabilities of discrete random variables.
- (e) Use tables, histograms and/or closed-form functions to represent PMFs
- (f) Explain the difference between a theoretical and empirical distribution.
- (g) Simulate discrete random variables and visualize empirical distributions using Python
- (h) State the mathematical definition of what it means for 2 random variables to be independent.
- (i) Determine whether 2 discrete RV are independent using the mathematical definition

- (j) State the Probability Mass Functions, Expected Value and Variance for Bernoulli, Binomial and Poisson RV and use them to calculate probabilities
- (k) Distinguish when to use Bernoulli vs Binomial vs Poisson Random Variables to model a given situation and state any assumptions that are needed to use these RV
- (l) Explain the difference between Probability Density Functions (PDFs) and Cumulative Density Functions (CDFs) and use both to calculate probabilities for continuous random variables
- (m) Calculate the Expected Value and Variance for Continuous Random Variables
- (n) State the PDF, Expected Value and Variance for Uniform, Exponential and Normal continuous random variables and use them to calculate probabilities
- (o) Distinguish when to use Uniform vs Exponential vs Normal random variables to model a given situation and state any assumptions that are needed to use these RV
- (p) Define what it means for RV to be IID and determine when RV meet this criteria



Review Questions

Here are a selection of practice questions for you to use to quiz yourself while studying in addition to HW, quiz and examples we completed in class. These practice questions are listed in random order to give you practice distinguishing what concept/strategy to apply.

Try to answer these questions like you are taking an exam (using only your crib sheet and calculator as resources). The answers to the questions from the slides are provided on the next slide in that lesson (and a video explanation is provided in the video on that topic in the supporting materials for that lesson). The answers to questions not from the slides are posted in a separate Exam 1 Review answers file posted in the modules on Canvas.

1. Lesson 7 Slide 63

Suppose you have 4 cat stuffed animals and 3 shark stuffed animals in a bag. You randomly draw 3 of the stuffed animals out of the bag (without replacement). What is the probability that you draw 1 cat and 2 sharks?

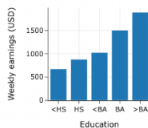
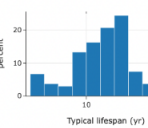
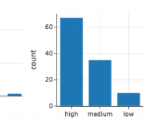
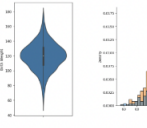
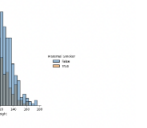
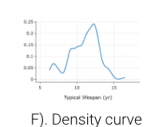
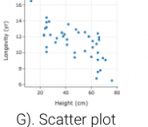
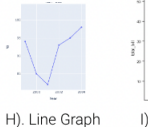
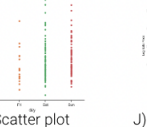
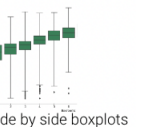
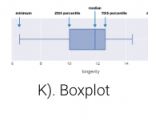
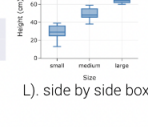
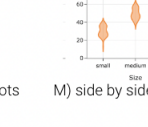

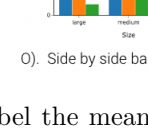

2. Lesson 7 Slide 70

You take a random survey of CU students and ask them if they use Facebook and/or X. 80% of students report they use X, 40% of the students report they use Facebook and 3% of students report they don't use either.

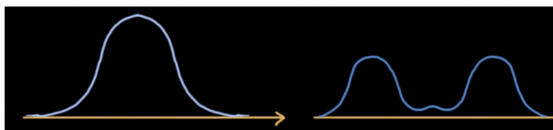
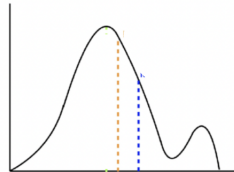
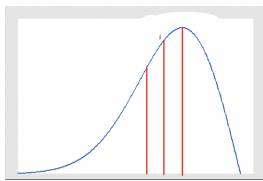
Suppose you choose a person at random from the list you surveyed.

- (a) What is the probability that this person uses Facebook or X?
- (b) What is the probability that this person uses Facebook and X?
- (c) What's the probability that this person does not use Facebook given that they do use X?

3. Lesson 6 Slide 54

Feature Type	Dimension	Types of Visualizations: Which Visualizations on the right should be used? Select all that apply.
Quantitative	1 Feature	    
Qualitative	1 Feature	    
Quantitative	2 Features	   
1 Quant and 1 Qual	2 Features	 

4. Lesson 6 Slide 39 For each curve shown here, label the mean, median and any mode(s). Then state whether the distribution is left skewed, right skewed or symmetric.



5. Lesson 7 Slide 60

You roll a fair 6-sided die 4 times. What is the probability that you roll at least one 6?

6. The following simulation estimates a specific probability:

```
box1 = {"balls" : np.array(["green", "red", "purple"]), "probs" : np.array([1/7, 1/2, 5/14])}
box2 = {"balls" : np.array(["green", "red", "purple"]), "probs" : np.array([1/3, 1/4, 5/12])}

box_choices = {"boxes" : np.array([box1, box2]), "probs" : np.array([2/3, 1/3])}

def funtimes(box_choices):
    b = np.random.choice(box_choices["boxes"], p = box_choices["probs"])
    return np.random.choice(b["balls"], p = b["probs"])

num_samples=1000
m = np.array([funtimes(box_choices) for ii in range(num_samples)])
print(np.sum(m == "purple") / num_samples)
```

- What theoretical probability is this code estimating?
- Calculate the exact probability for the quantity you listed in part (a). i.e. what number should this code output approach as you increase NumSamples?

7. .

In the mobile game Among Us, Crewmates on a spaceship work together to complete tasks while a few randomly-selected Imposters secretly try to eliminate crewmates. If all Crewmates complete their tasks, the Crewmates win; if the Imposters eliminate all but one of the crewmates, the Imposters win.

Matty made a **games** table listing each game they played in 2021, ordered chronologically. The first three rows:

team	outcome	length	completed
Crewmate	Win	981	7
Imposter	Loss	840	8
Crewmate	Loss	520	3

The columns include:

- **team**: which team Matty was on in the game.
- **outcome**: whether Matty's team won or lost.
- **length**: the duration of the game in seconds.
- **completed**: the number of tasks completed by all crewmates before the game ended.

(a) (3.0 points)

Choose which type of visualization would be most useful for investigating each of the following.

i. (1.0 pt) The distribution of game lengths.

- ☐ Bar Chart
- ☐ Histogram
- ☐ Line Plot
- ☐ Scatter Plot

ii. (1.0 pt) The association between game length and number of tasks completed.

- ☐ Bar Chart
- ☐ Histogram
- ☐ Line Plot
- ☐ Scatter Plot

iii. (1.0 pt) The average game length for each outcome.

- ☐ Bar Chart
- ☐ Histogram
- ☐ Line Plot
- ☐ Scatter Plot

8. You have a box with 6 coins in it. Each coin belongs to one of three categories, and all coins are equally likely to be drawn from the box:

- 1 of the coins is **fair** (F) such that heads (H) and tails (T) are equally likely.
- 2 of the coins are **biased towards heads** (BH), such that **heads comes up three times as often** as tails.
- 3 of the coins are **biased towards tails** (BT), such that **tails comes up three times as often** as heads.

- You choose a coin at random from the box. What is the probability that it is not a fair coin? (Give your answer as a fraction).
- You choose a coin at random and flip it. What is the probability that the coin comes up heads? Give your answer as a single fraction, fully simplified.
- Suppose you pick a coin at random and flip it. Are the events “flip comes up heads” and “you picked a fair coin” independent? Justify your answer using the **mathematical definition** of independence.
- You choose a coin at random and flip it. It comes up heads. Given this information what is the probability that the coin you chose was one of the BH coins?
- You choose a coin at random and flip it **three times in a row**. It comes up tails all 3 times. Given this information, what is the probability that the coin that you chose was a fair coin? (You may leave your answer unsimplified).

9. [Lesson 11 Slide 40](#)

Let X be the outcome of a single fair die roll.

- (a) Give the PMF of X as a closed-form function
 - (b) What is the expectation of X ?
 - (c) What is the variance of X ?
 - (d) What is the standard deviation of X ?
-

10. [Lesson 7 Slide 52](#)

A standard 52-card deck consists of 13 cards in each of four suits: clubs, diamonds, hearts and spades. Suppose you draw a single card at random from a standard 52-card deck (without replacement) and then draw a 2nd card.

- (a) What is the chance that I get an Ace followed by a King?
- (b) What is the chance that one of the cards I draw is a King and the other is an Ace?

11. [Lesson 8 Slide 45](#)

A car heading from Berkeley to San Francisco is pulled over on the freeway for speeding. Which type of car is it more likely to be:

- a Tesla which is relatively common in California
- or a Lamborghini which is a rare car that is known for speeding

You don't have enough information to calculate the answer directly. What would you guess, and why? Make some reasonable assumptions (data scientists often have to do this) and explain your thought process.

12. Suppose the scores on a college entrance exam have a mean 75 and a variance of 81.

- (a) At least 75% of the scores lie between what two values?
- (b) Suppose you are told the scores have a normal distribution. 68% of the scores lie between what two values?

13. You roll two fair six-sided dice, one red and one blue. Define two random variables:

X : the outcome of the red die.

Y : 1 if the sum of the two dice is greater than 7, and 0 otherwise.

Are the random variables

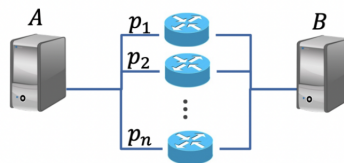
X and Y independent or dependent? Justify your answer using the definition of independent random variables.

14. [Lesson 9 Slide 25](#)

Network reliability

Consider the following parallel network:

- n independent routers, each with probability p_i of functioning (where $1 \leq i \leq n$)
- E = functional path from A to B exists.



What is $P(E)$?

15. [Lesson 7 Slide 67](#)

There population of CU undergraduates is $n = 31,000$ students

Suppose you are friends with $r = 100$ people.

You walk into a classroom and you see $k = 160$ random people.

Assume each group of k CU undergrads is equally likely to be in the room.

What is the probability that you see at least one friend in the room?

16. [Lesson 14 Slide 12](#)

Sarah, arriving at a bus stop, just misses the bus. Suppose she decides to walk if the (next) bus takes longer than 6 minutes to arrive. Suppose also that the time in minutes, X , between the arrivals of buses at the bus stop is a continuous random variable with a uniform distribution between 3 and 8 minutes.

- (a) What is the probability that Sarah will end up walking?
- (b) What is the average amount of time Sarah should expect to wait?

17. For each of the following scenarios, determine if you have enough information to model this with a random variable using only the assumptions given (and making no other assumptions).

If not, explain what additional information you'd need.

If you do have enough info, answer all of the following:

- i). Define the random variable. Give your answer in the form $X = \dots$ (insert what X quantifies)
 - ii). Describe the distribution of the random variable. Give your answer in the form $X \sim \text{DistributionName}(\text{parameter values})$.
 - iii). Give the distribution (PMF or PDF) and the support of the random variable
- (a) A basketball player makes 70% of her free throws. The results of each free throw are independent. You want to model the number of successful free throws she makes out of 10 attempts.
 - (b) A factory machine produces bolts, and each bolt can either pass or fail a quality inspection. You are interested in modeling whether the next bolt produced will be defective. From past data, 5% of the bolts produced by the machine are defective. Whether or not bolts pass the inspection are independent of the results of previously produced bolts.
 - (c) Phone calls at a call center are received independently and at a constant average rate of 10 calls per hour. You want to model the time (in minutes) until the next phone call.
 - (d) A person is equally likely to arrive at a bus stop at any time between 2:00 PM and 3:00 PM. You want to model the time of arrival within this interval.
 - (e) The salaries of people at a large company have a mean of \$70000 and standard deviation of \$15000. You want to model the salary of a randomly chosen employee from this company.
 - (f) Phone calls at a call center are received independently and at a constant average rate of 10 calls per hour. You want to model the number of calls in the next 30 minutes.
-

18. [Lesson 7 Slide 58](#)

A standard 52-card deck consists of 13 cards in each of four suits: clubs, diamonds, hearts and spades. Suppose you draw a single card at random from a standard 52-card deck.

- a). What is the probability that the card is the Ace of Diamonds?
 - b). What is the probability that the card is an Ace or a Diamond?
-

19. [Lesson 8 Slide 52](#)

Suppose you're in a classroom with a total of 40 students. What is the probability that at least 2 students share a birthday?

20. An emergency room at a particular hospital gets an average of five patients per hour. A doctor wants to know the probability that the ER gets more than five patients in the next hour.

21. [Lesson 8 Slide 55](#)

Monty Hall Problem

22. Suppose you gather a random sample of data of the delay times (in minutes) of airline flights and store the data in an array called `x`.

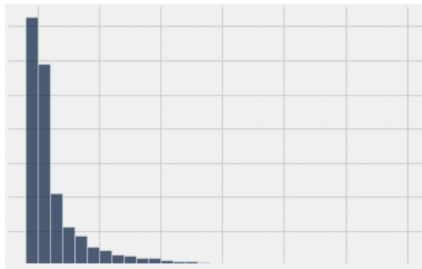
The histogram shown on the right is the output of the following code:

```
import matplotlib.pyplot as plt

plt.hist(x, density=True)
```

Based on the histogram, which of the following statements will return an output of `True`?

- ☐ `sum(x>np.average(x))/len(x) > 0.5`
- ☐ `sum(x>np.average(x))/len(x) == 0.5`
- ☐ `sum(x>np.average(x))/len(x) < 0.5`
- ☐ Cannot determine without more information



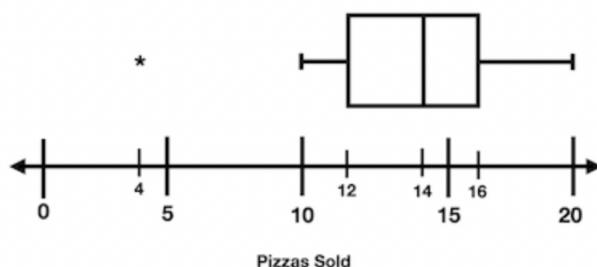
The histogram shows the distribution of values contained in an array `x`.

Which of the following is `True`?

- (A) `sum(x > np.average(x)) / len(x) < 0.5`
- (B) `sum(x > np.average(x)) / len(x) == 0.5`
- (C) `sum(x > np.average(x)) / len(x) > 0.5`

To receive credit for your answer above, write a sentence or two explaining/justifying your answer. Answers without correct justification will receive 0 points.

23. Consider the following boxplot which summarizes the amount of pizzas sold at a pizzeria during a certain period of time.



Determine which of the following questions you have enough information to answer. If you don't have enough information, state "need more info."

- (a) What is the mean?
- (b) What is the median?
- (c) What is the mode?
- (d) What is the IQR?
- (e) Is this distribution right skewed?
- (f) What is the smallest data point within $1.5 \times \text{IQR}$ of the 25th percentile?

24. According to Baydin, an email management company, an email user gets, on average, 147 emails per day.

- What is the probability that an email user receives more than 160 emails per day?
 - What is the standard deviation of the number of emails a user gets per day?
-

25. [Lesson 12 Slide 62](#)

Suppose Golden State Warriors are going to play the Toronto Raptors in a 7-game series during the NBA finals.

Suppose the Warriors have a probability of 58% of winning each game, independently.

A team wins the series if they win at least 4 games (we play all 7 games). What is $P(\text{Warriors winning})$?

26. Suppose you have a group of 100 people and 5 people in the group are ambidextrous (can write with both hands).

- You sample a single individual at random from this group. What is the probability that the individual will **not** be ambidextrous?
 - Suppose you randomly sample 20 individuals from this group **with replacement**. What is the probability that 2 out of the 20 people in your sample are ambidextrous?
 - Suppose you randomly sample 20 individuals from this group **without replacement**. What is the probability that 2 out of the 20 people in your sample are ambidextrous?
-

27. Consider the following function related to finding an open parking spot in a parking lot where the probability of an individual spot being open is given by p . What distribution does the return value of the function belong to?

```
def shoulda_taken_the_bus(p=.30):  
    x = 0  
    y=np.array([1,2,3,4,5,6])  
    for i in y:  
        if np.random.choice([0,1], p=[1-p, p]) == 1:  
            x += 1  
    return x
```

28. .

[5 Pts] Which of the following styles of plots are good for visualizing the distribution of a continuous variable? Choose all that apply.

☐ Pie Charts ☐ Box Plots ☐ Bar Plots ☐ Histogram ☐ None of the above

[2 Pts] Suppose you wish to compare the number of homes homeowners in the US own and their respective salaries. Which style of plot would be the best?

☐ Scatter Plot ☐ Overlaid Line Plots ☐ Side by Side Box Plots ☐ Stacked Bar Plot

29. Consider the following simulation:

```
import numpy as np  
import pandas as pd  
  
def rolling_rollers(NumSamples):  
    roll = np.random.choice([1,2,3,4,5,6,7,8], size=NumSamples)  
    df_roll = pd.DataFrame({"num":roll})  
    numerator=np.sum((df_roll["num"] % 2 == 0) & (df_roll["num"] > 3))  
    denom = np.sum(df_roll["num"] % 2 == 0)  
    return numerator/denom
```

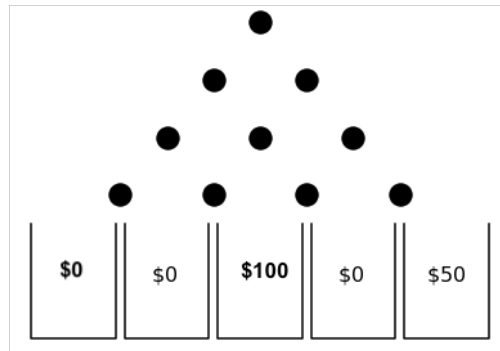
Let X denote the number on the uppermost face of a fair **8-sided** die after rolling it once.

- What probability does the function defined above estimate?
- What number should the output of the code shown above approach as you increase NumSamples? Give your answer as a fully simplified fraction. *Show work in the space below, justifying all steps. Answers without any work/justification will receive 0 points:*

30. [Lesson 12 Slide 59](#)

Consider the Galton Board shown on the slide at the link above. Calculate the probability of a ball landing in the bucket k

31. A game of **Plinko** is to be played on the board shown below. The pegs are **biased such that a disc is twice as likely to move to the right than the left** at each peg. Furthermore, the disc can only be dropped from directly above the top-most peg. Answer the following questions about this Plinko game. Be sure to show your work.



- Let Y be the random variable that describes the winnings when a single disc is dropped. Write down the probability mass function for Y .
 - What is the probability that you win a total of exactly \$100 in a game with 2 discs?
 - What is the probability that you win a total of exactly \$200 in a game with 5 discs?
 - What is the **expected** winnings if you play a game with 3 discs?
-

32. [Lesson 14 Slide 77](#)

Potentially useful output from Python:

```
from scipy import stats

stats.norm.cdf(6, 4, 2) = 0.84
stats.norm.pdf(6, 4, 2) = 0.12
stats.norm.cdf(6, 4, np.sqrt(2)) = 0.92
stats.norm.pdf(6, 4, np.sqrt(2)) = 0.10
```

Suppose your time traveling between classes is normally distributed. On average you spend 4 minutes traveling between classes and the variance of your traveling time between classes is 2 minutes. What is the probability you spend greater than or equal to 6 minutes traveling between classes?

33. [Lesson 14 Slide 74](#)

Suppose a visitor to your website leaves after X minutes. On average, visitors leave the site after 5 minutes. The length of stay, X , is exponentially distributed.

- What is $P(X > 10)$?
 - What is $P(10 < X < 20)$?
-

34. [Lesson 14 Slide 58](#)

Major earthquakes (magnitude 8.0+) occur once every 500 years in California (according to historical data from USGS, 2015)

What is the probability of zero major earthquakes in California next year? (State any assumptions you use to calculate this probability).

35. [Lesson 12 Slide 58](#)

Match the distribution to the graph.
