Lab 2: Probability and Random Variables

Objective: The objective of this lab is to become familiar with using the models for random variables and to find the probabilities associated with the models you have learned. The probabilities we compute from these models (for example, p-values in testing theories) will help us make reasonable decisions. You will work with three random variables and the methods used to calculate probability for each variable. You will also become familiar with several concepts that allow for easier calculation of probabilities.

Application: Researchers at University of Michigan Hospital and Baylor University were interested in determining if lack of sleep for teenagers led to an increased incidence of heart disease. These researchers conducted a study in which the sleep patterns for 37 teenagers were recorded in addition to the presence of heart disease for each teenager. The researchers can use the probability rules to determine if lack of sleep and incidence of heart disease are independent events.

Overview: In this lab you will be introduced to several random variables and their models. These variables can be classified as one of two types: a ***discrete random variable***, which has a finite number of outcomes, and a ***continuous random variable*** which has an infinite number of outcomes. You will practice computing probabilities for each of these variables, using concepts such as a ***standardized score*** and a ***Normal Approximation.*** There are several situations that can make computing probabilities easier, such as ***independent events*** and ***mutually exclusive events*** that will also be explored in this lab.

Independent Events: Two events A, B are said to be independent if knowing that one will occur (or has occurred) does not change the probability that the other occurs. In probability notation this can be expressed as P(A|B) = P(A). Going back to the application, the researchers from the study can check the independence of the two events by computing the probability of heart disease for teenagers who sleep less than the recommended amount as well as the probability of heart disease for all teenagers. If the incidence of heart disease for a teen is independent of the amount of sleep they get, the two probabilities will be equal.

Mutually Exclusive: Two events A, B are mutually exclusive (or disjoint) if they do not contain any of the same outcomes. So their intersection is empty.

Random Variables: A random variable assigns a number to each outcome of a random circumstance, or, equivalently, a random variable assigns a number to each unit in a population. The distribution of a random variable is a model that shows us what values are possible for that particular random variable and how often those values are expected to occur (i.e. their probabilities). The model can be expressed as a function or table or picture, depending on the type of variable it is. We will consider two broad classes of random variables: discrete random variables and continuous random variables.

Discrete Random Variable: A discrete random variable, X, is a random variable with a finite or countable number of possible outcomes. The probability distribution function (pdf) for a discrete random variable X is a table or rule that assigns probabilities to the possible values of the X.

Two conditions that must always apply to the probabilities for a discrete random variable are:
Condition 1: The sum of all of the individual probabilities must equal 1.
Condition 2: The individual probabilities must be between 0 and 1.

Binomial Random Variable B(n,p): One discrete random variable is the binomial random variable, which counts the number of times a certain event occurs out of a particular number of observations or trials of a random experiment. A binomial experiment is defined by the following conditions:

1. There are n "trials" where n is determined in advance and is not a random value.
2. There are two possible outcomes on each trial, called "success" (S) and "failure" (F).
3. The outcomes are independent from one trial to the next.
4. The probability of a "success" remains the same from one trial to the next, and this probability is denoted by p. The probability of a "failure" is 1 – p for every trial.

Continuous Random Variable: A continuous random variable, X , takes on all possible values in an interval (or a collection of intervals). The way that we determine probabilities for continuous random variables differs in one important respect from how we determine probabilities for discrete random variables. For a discrete random variable, we can find the probability that the variable X exactly equals a specified value. We can't do this for a continuous random variable. Instead, we are only able to find the probability that X could take on values in an interval. We do this by determining the corresponding area under a curve called the probability density function of the random variable.

So the probability distribution of a continuous random variable is described by a density curve. The probability of an event is the area under the curve for the values of X that make up the event. The probability model for a continuous random variable assigns probabilities to intervals.

Definition: A curve (or function) is called a **Probability Density Curve** if:

1. It lies on or above the horizontal axis.
2. Total area under the curve is equal to 1.

Normal Random Variable N(μ,σ): The family of normal distributions is very important because many variables have this shape and form approximately and many statistics that we use in our inference methods are based on sums or averages which generally have (approximately) a normal distribution.

A normal curve is symmetric, bell-shaped, centered at the mean and its spread is determined by the standard deviation. In fact, the points of inflection on each side of the mean mark the values which are one standard deviation away from the mean.

Standardized Scores: A normal distribution is indexed by its population mean, and its population standard deviation. Recall that the standard deviation is a useful "yardstick" for measuring how far an individual value falls from the mean. The standardized score or z-score is the distance between the observed value and the mean, measured in terms of number of standard deviations. Values that are above the mean have positive z-scores, and values that are below the mean have negative z-scores.

Normal Approximation to the Binomial Distribution: The easier way involves using a normal distribution. The normal distribution can be used to approximate probabilities for other types of random variables, one being binomial random variables when the sample size n is large.

Expected Value: The expected value of a random variable is the mean value of the variable X in the sample space, or population, of possible outcomes. Expected value, denoted by E(X), can also be interpreted as the mean value that would be obtained from an infinite number of observations on the random variable.

Standard Deviation: The standard deviation can be viewed as approximately the average distance of the possible values of X from its mean.

Formula Card

Probability Rules

- **Complement rule**

  $P(A^C) = 1 - P(A)$
- **Addition rule**

  General: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

  For independent events:

  $P(A \text{ or } B) = P(A) + P(B) - P(A)P(B)$

  For mutually exclusive events: $P(A \text{ or } B) = P(A) + P(B)$
- **Multiplication rule**

  General: $P(A \text{ and } B) = P(A)P(B \mid A)$

  For independent events: $P(A \text{ and } B) = P(A)P(B)$

  For mutually exclusive events: $P(A \text{ and } B) = 0$
- **Conditional Probability**

  General: $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$

  For independent events: $P(A \mid B) = P(A)$

  For mutually exclusive events: $P(A \mid B) = 0$

Discrete Random Variables

Mean

$$E(X) = \mu = \sum x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

Standard Deviation

$$s.d.(X) = \sigma = \sqrt{\sum (x_i - \mu)^2 p_i} = \sqrt{\sum \left(x_i^2 p_i\right) - \mu^2}$$

Binomial Random Variables

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\dbinom{n}{k} = \dfrac{n!}{k!(n-k)!}$

Mean

$$E(X) = \mu_X = np$$

Standard Deviation

$$s.d.(X) = \sigma_X = \sqrt{np(1-p)}$$

Normal Random Variables

- $z - \text{score} = \dfrac{\text{observation} - \text{mean}}{\text{standard deviation}} = \dfrac{x - \mu}{\sigma}$
- Percentile: $x = z\sigma + \mu$
- If X has the $N(\mu, \sigma)$ distribution, then the variable

  $Z = \dfrac{X - \mu}{\sigma}$ has the $N(0,1)$ distribution.

Normal Approximation to the Binomial Distribution

If X has the $B(n, p)$ distribution and the sample size n is large enough (namely $np \geq 10$ and $n(1-p) \geq 10$), then X is approximately $N\left(np, \sqrt{np(1-p)}\right)$.

Warm-Up: Types of Variables and Types of Distribution

1. **Type of Variables:**

Today's typical undergraduate student is often characterized as preferring teamwork, experiential activities, and the use of technology. An ECAR (Educause Center for Applied Research) study was published on technology use among undergraduate students. The study used survey and interviewer data to create a portrait of today's students' experiences with and skill using information technology.

Listed below are some of the response variables that were measured in this study. For each of these determine whether it is categorical, quantitative discrete, or quantitative continuous.

a. Technology ownership: Do you own a computer?

categorical **quantitative discrete** **quantitative continuous**

b. Time (per week in minutes) spent using a computer for writing documents (word processing).

categorical **quantitative discrete** **quantitative continuous**

c. Which social networking site(s) are you a member? (Facebook, Myspace, Friendster, etc.)

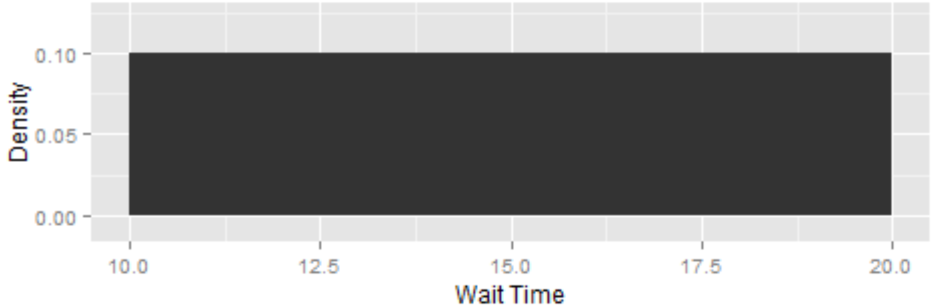
categorical **quantitative discrete** **quantitative continuous**

2. **Types of Models:**

Identify appropriate model for each of the three variables. (Be complete).

a. Below is the model for the random variable X which represents waiting time to be served at a deli.

X has a ____ distribution

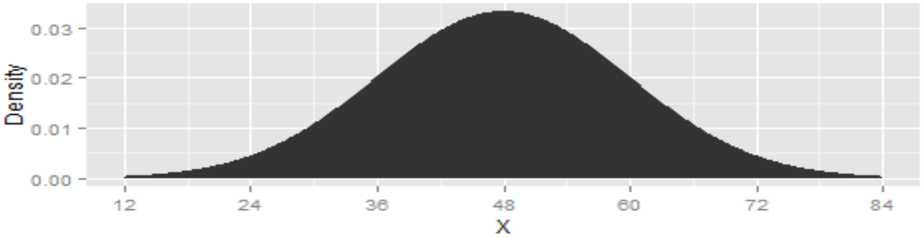


b. Suppose that 45% of Michigan residents own dogs. Let *X* represent the number of Michigan residents with a dog in a random sample of 10 Michigan residents. A success is defined as the resident owning a dog.

X has a ______________________________ distribution

c. Below is the model for the random variable X which represents the score on a Stats 250 Exam (out of 100 points).

X has a ______________________________ distribution



ILP: Probability and Random Variable Review

Problem 1: Study on Smiling

In a recent study people were observed for about 10 seconds in public places (e.g. malls and restaurants) to determine whether they smiled during the randomly chosen 10-second interval. The table shows the results for comparing Adults (group 1) and Minors (group 2).

	Smile	**No Smile**	**Total**
1 = Adult	3269	3806	**7075**
2 = Minor	4471	4278	**8749**
Total	**7740**	**8084**	**15824**

a. What is the probability that a randomly selected person smiled?

b. The researcher would like to ***assess if smiling status is independent of age group***.

i. To check for independence, the probability found in part (a) should be compared to which of the following probabilities?

P(smiled and adult) ***P(smiled given adult)***

P(adult given smiled) ***P(adult)***

ii. Find the probability selected above and circle the appropriate conclusion.
The probability = ____

Thus it appears that smiling status ***is*** ***is not*** independent of age group.

Problem 2: Summer Trip Length

Did high gas prices keep Americans from hitting the road this past summer? In a nationwide survey of adults, one variable measured was how many days vacationers spent driving on the road on their longest trip. Consider the following (partial) probability distribution for the random variable *X* = the number of days for the longest car trip.

X	4	5	6	7	8
Probability	0.10	0.20	0.25		

a. Suppose the probability of 7 days is twice as likely as the probability of 8 days. Complete the probability distribution for *X*. Show your work.

b. What is the expected number of days for the longest trip?
Include ***symbol, value, and units***.

Problem 3: How Much Time do You Spend Studying Statistics?

A Washington Post article "Is college too easy? As study time falls, debate rises" (May 21, 2012) stated that the amount of time college students actually study has dwindled from an average of 24 hours per week to about 15 hours (based on a survey). A professor of statistics decided to ask all of his current semester students to report the number of hours per week they spend studying his course material (on a regular, non-exam week). The mean for the female students was 10 hours and the standard deviation was 3.5 hours.

a. Consider the following interpretations of the standard deviation and circle those that are correct.

- On average, the number of hours spent studying statistics varied from the mean by about 3.5 hours.
- The average distance between the number of hours spent studying statistics is roughly 3.5 hours.
- The average number of hours spent studying statistics is about 3.5 hours away from the mean.

b. Julie is one of his current students and she studies statistics for about 6 hours per week. What is her corresponding z-score? Then find the probability a randomly selected statistics student studies more than 6 hours per week. Use the Z-table and then use the prob() R script to get a more exact probability and to produce a nice picture and solution.

c. Juan learns that he is in the top 30% of the studying distribution. Based on the distribution, Juan must study at least how many hours per week? Make a hand sketch of what you are trying to find to help show your work.

Problem 4: The Walking Dead

The television series *The Walking Dead* is an American post-apocalyptic horror drama that is based on the comic book series by the same name. The first episode aired in October 2010 and earned a 25% audience share. That is, 25% of all TVs in use during the show time period were tuned to a station airing *The Walking Dead*.

A random sample of 5 people who watched TV during that time period were selected. Find the probability that exactly 2 of the 5 people selected watched *The Walking Dead*.

Cool-Down: True or False

Decide whether the following questions are true or false.

1. If the time to wait for pharmacy help has a uniform distribution from 0 minutes to 30 minutes, then 33% of the customers are expected to wait more than 20 minutes.

   True **False**

2. If X has a Binomial (50, 0.7) distribution, then the criteria to use the normal approximation are met.

   True **False**

3. 68% of all test scores will fall within one standard deviation of the mean test score.

   True **False**

4.. Police report that 78% of drivers stopped on suspicion of drunk driving are given a breath test, 36% are given a blood test and 22% are given both tests. Do the police administer these two tests independently? Justify your answer with a calculation.

   True **False**