# Mod 12 – Probability & Probability Distributions

Learning Goals

- Distinguish between discrete quantitative variables and continuous quantitative variables.
- Interpret probability distributions for categorical and quantitative variables.
- Find the mean and variance of a discrete quantitative variable.

**<u>Introduction</u>**

- In the previous module, we use the word *probability* to mean “likelihood” or “chance” that a particular outcome would occur.
- We represent probability with the notation *P(A)* where *A* is the description of the outcome.
- Typically when calculating a probability, we write the results as a proportion as opposed to a percent. For example, if there is a 30% chance that outcome *A* will occur, we write P(A) = 0.30 and say, “The probability of A is 0.30” OR “There is a 30% chance that A will occur.”

Let’s test your understanding.

1) Suppose we flip a fair coin. Let H be the outcome *the coin lands heads up*. Find the following probability and explain how you determined the probability.

   P(H) =

2) Suppose there 18 females and 12 males in a statistics class, and we randomly select a student. Let F be the outcome that the student is female and M be the outcome that the student is male. Find the following probabilities and show your work.

   a) P(M) =

   b) P(F) =

3) Suppose you work for a pet grooming business. Today, there are 8 cats and no dogs waiting to be groomed. You randomly choose a pet to groom. Find the following probabilities. Explain your results.

   a) P(Dog) =

   b) P(Cat) =

**<u>Exploring the Relationship Between Theoretical and Empirical Probability</u>**

Work through ***Module 12 – Probability Applets Activity Copy*** in the Resources module on Canvas. When you're finished, answer the following questions.

4) Provide a quick working definition for each of the following.

   *Theoretical Probability*:

   *Empirical Probability*:

5) How are theoretical and empirical probabilities related? When is an empirical probability a good estimate for a theoretical probability?

**<u>Probability Distributions</u>**

6) Here is a probability distribution for an imaginary horse race. The first row lists the values of the variable *Horse*, and the second row gives the probability for the corresponding variable value in the first row.

| **Horse** | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| **Probability of winning** | 0.28 | | 0.15 | 0.14 | 0.12 | 0.08 | 0.01 |

   a) Complete the table by finding the probability that horse B will win. Show your work.

   b) Which is more likely to win, horse B or horse F? Explain your conclusion in terms of "chance".

   c) Which outcome is most unusual?

   d) Find each of the following probabilities. Show your work.

      P(A wins or B wins)

      P(NOT F wins)

**Quantitative Variables**

7) Suppose we conduct an experiment by flipping a fair coin four times.

   a) Are the outcomes "heads" and "tails" values of a categorical variable or a quantitative variable? Explain.

   b) Would you say that the outcome "the number of heads in four tosses" is a value of a categorical variable or a quantitative variable? Explain.

8) A **random variable** is a numerical measure of the outcome of an experiment, i.e. a quantitative variable. Random variables are typically denoted with capital letters near the end of the alphabet such as *X* or *Y*. We often refer to random variables as *quantitative variables*.

   a) Suppose we randomly select a person and test their blood type. Is the variable *blood type* a random variable? Explain.

   b) Suppose we toss a coin four times. Is the variable *number of heads* a random variable? Explain.

9) Suppose we roll a pair of fair dice (one red die and one green die) and let the random variable *Y* represent the sum of the dots on the faces of the dice.

   List all possible outcomes in the table below.

**RED**

| SUM | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |

**GREEN**

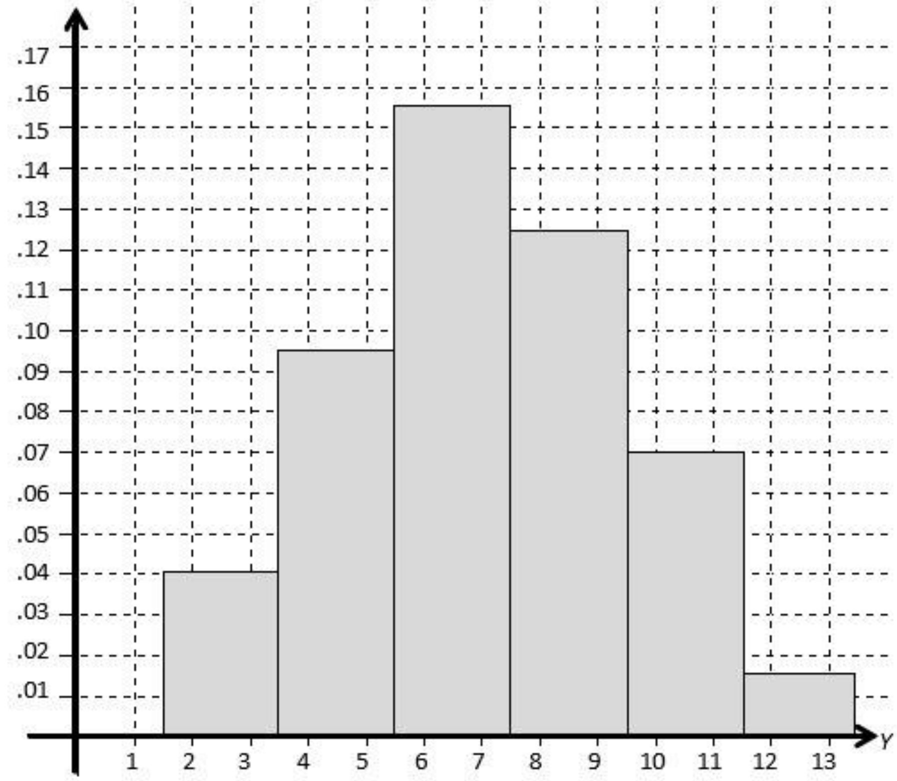
Find the following probabilities.

$P(y < 8)$ $P(4 \leq y \leq 7)$ $P(y > 7)$

10) In a probability density curve, the area under the curve is a probability distribution. Similarly, the histogram below is a probability distribution of *Y*, the random variable representing the sum of the dots on the up faces when we roll a pair of dice. In other words, the area of each rectangle should represent a probability. **Use the histogram to answer the following questions**.

a) For each bin in the histogram, label the height at the top of the bin, and then calculate the area and write it inside the bin.

b) If the area of each bin is supposed to represent a probability, what does the area of the first bin represent? Use the table you built on the previous page to verify.




c) Use the table on the previous page to verify that the area of the fourth bin represents
P(*y* = 8 or *y* = 9). Show your work.

d) What does the area of the third bin represent?

e) What does the area of the last bin represent?

f) What is the total area of the histogram? Does this make sense? Explain.

g) Use the histogram to find each of the following probabilities and explain what you did to find them.

$P(y < 8)$ $P(4 \le y \le 7)$ $P(y > 7)$

## **Discrete vs Continuous *Quantitative Variables* (a.k.a. *Random Variables*)**

11) A quantitative variable is called **discrete** if it can assume only a finite number of values OR if it can assume infinitely many values that increase in discrete steps (for example 4, 7, 10, 13, … goes on forever but makes "discrete" jumps of 3 each time). A quantitative variable is called **continuous** if it can assume an infinite number of values over an interval - no discrete jumps (for example all the numbers between 0 and 1 includes infinitely man y values with no discrete jumps).

    For each of the following, determine if it is a **discrete quantitative variable** or a **continuous quantitative variable**.

| Experiment | Random Variable *X* | Discrete or Continuous? |
|---|---|---|
| a) Flip a coin three times | *X* = total number of heads | |
| b) Randomly select a student who took a true/false test with 100 questions | *X* = number of questions answered correctly | |
| c) Randomly select a mutual fund | *X* = the number of companies in the portfolio | |
| d) Randomly select 50 students | *X* = the exact average weight of the group in pounds | |
| e) Randomly select a woman | *X* = the exact height of the woman in inches | |

## The Mean and Standard Deviation for a Discrete Quantitative Variable

12) Here is the probability distribution for rolling a pair of fair dice (probabilities are rounded to three decimal places).

**Probability Distribution**

| **Sum of dots** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Probability** | 0.028 | 0.056 | 0.083 | 0.111 | 0.139 | 0.167 | 0.139 | 0.111 | 0.083 | 0.056 | 0.028 |

Based on the probability table, here are the frequencies we would expect if we rolled a pair of dice 4004 times and recorded the sum of the dots on the up faces.

**Frequency Distribution**

| **Sum of dots** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Frequency** | 112 | 224 | 332 | 444 | 556 | 668 | 556 | 444 | 332 | 224 | 112 |

a) Use the **frequency table** to find the mean (i.e. expected value) for this discrete random variable. Show your work.

b) Here is the expected value for a probability distribution.

***Expected value*** (mean) for a probability distribution: $\mu_x = \sum x \cdot p(x)$

Now use the **expected value** formula to relate the mean you found above from the frequency table to the **probability table**.

Just like we need the standard deviation to measure the variability for a data set, we need the standard deviation to measure the variability for a discrete random variable. Here are two formulas for calculating standard deviation.

| **Standard deviation for a data set** | **Standard deviation for a discrete random variable** |
|---|---|
| $sd = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ | $\sigma = \sqrt{\sum(x-\mu_x)^2 \cdot p(x)}$ |

13) The probability distribution gives the potential cost of routine maintenance and major repairs in the first 5 years you own a new car.

| **Cost over 5 years** | \$0 | \$500 | \$1,000 | \$1,500 | \$2,000 |
|---|---|---|---|---|---|
| **Probability of cost** | 0.19 | 0.67 | 0.09 | 0.04 | 0.01 |

a) Find the expected cost of routine maintenance and major repairs in the first 5 years you own a new car? (Hint: recall that the expected cost is just the mean cost.) Show your work.

b) Find and interpret the standard deviation for the cost of routine maintenance and major repairs in the first 5 years you own a new car. Show your work.

## Mod 12 Review – Let's Practice

Directions: Complete these problems after you work through Module 12 on Canvas.

**Question 1** Which of the following is a *discrete random variable* (a.k.a. *discrete random variable*)? Explain why your choice is a discrete random variable and the others are not.

A. Gender of a Holland Lop rabbit.

B. Number of baby bunnies in a nest.

C. Weight of a Holland Lop baby bunny.

**Question 2** A couple plans to have no more than three children, and they will keep having children until they have a girl. So, if their first child is a girl, they will stop and have only one child. However, if their first child is a boy, they will try again and have a second child.

As it turns out, the probability of having a boy is slightly greater than having a girl. Here is the probability distribution for the number of boys the couple could have.

| **Boys** | 0 boys | 1 boy | 2 boys | 3 boys |
|---|---|---|---|---|
| **Probability** | 0.490 | 0.250 | 0.127 | 0.133 |

What is the expected number of boys the couple will have? (Recall: the expected value is the mean).

A. 0.903

B. 1.5

C. 0.228

Continued on the next page …

**Question 3** The US Bureau of Labor Statistics predicts the US economy will add more than 7,371,900 million jobs from 2012 to 2022. More importantly, of the occupations predicted to add jobs, 1,044,600 of those new jobs will likely come from the four highest paying occupations.

The number of new jobs predicted for each of the four highest paying occupations was used to compute the probabilities below.

| **Occupation** | General and operations managers | Management analysts | Software developers, applications | Registered nurses |
|---|---|---|---|---|
| **Probability** | 0.234 | 0.128 | 0.134 | ? |

Suppose we randomly select one of the new jobs. Is it more likely that the new job is from the occupation *registered nurses*, that the new job is from the occupation *general and operations managers*, or that the two events are equally likely?

**A.** It is more likely that the new job is from the occupation *general and operations managers*.

**B.** It is more likely that the new job is from the occupation *registered nurses*.

**C.** These events are equally likely.

Continued on the next page …

**Question 4.** Consider the probability distribution of X, where X is the number of job applications completed by a college senior through the school's career center.

| X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| P(X) | 0.002 | 0.011 | 0.115 | 0.123 | 0.144 | 0.189 | 0.238 | 0.178 |

We collect a random sample of 1000 college seniors who complete job applications through the career center.

**Part a)** Based on the probability distribution, which result would be surprising?

- 14 seniors completed 1 job application through the career centers.
- 15 seniors completed 2 job applications through the career center.
- 130 seniors completed 3 job applications through the career center.

**Part b)** What is the probability that a randomly selected college senior who uses the career center completed at most three job applications?

A. 0.123

B. 0.251

C. 0.877

D. 0.872