STAT 140: Week 4

My hope is that after Week 4 (09/20-09/24), you will be able to:

- Understand the relationship between variables and events.
- Identify different types of events–independent, mutually exclusive, and complements.
- Use basic rules of probability to calculate the probability of an event occurring.
- Understand different ways of conveying information about probability distributions, like one-way tables, Venn Diagrams, and contingency tables.
- Calculate the expected value and variance of a random variable or linear combination of random variables.

Moving from Statistics to Probability

- So far, we have concentrated on using statistics to discuss the data we have observed.
  - $\bar{x}$ (mean), s (sd), $\hat{p}$ (proportion)
  - R (correlation), b_0 (intercept), b_1 (slope)
- In order to be able to use these statistics to draw conclusions, we need to know how the variables *should* be behaving–that is, we need to know how to talk about how variables theoretically behave.
- Statisticians use *probability* to build a theoretical basis of variable behavior.

Probability 1

- Probability is a theoretical description of how likely an **event** is to occur.
- Typically, we consider this event to be a single, unobserved outcome of a variable.
  - e.g., a single person's red blood cell count, a single house's price.
- We have already discussed the idea of a distribution, a summary of what values a variable can take and how often they occur.
- The set of all possible values or outcomes a variable can take is known as the **sample space**.
- Then, an event can be formally defined as a set of outcomes inside the sample space.

Probability 2

- Usually we use curly braces, "$\{\}$", to denote sets, including the sample space, or events.
- Example:
  - The sample space of rolling a die, or S, is $\{1, 2, 3, 4, 5, 6\}$.
  - How would you write the event that the die roll is even?
  - How would you write the event that the die roll is a multiple of three?

Probability 3

- Then, the probability of an event is the number of elements in the event divided by the number of elements in the sample space.
- Examples:
  - The sample space of rolling a die is $S = \{1, 2, 3, 4, 5, 6\}$.
    - The event that a roll is even, E, is $\{2, 4, 6\}$.
    - There are six elements in the sample space, and three elements in the event space.
    - Thus, the probability of E, or $P(E)$, is $\frac{3}{6}$.
    - The event that a roll is a multiple of three, M, is $\{3, 6\}$.
    - There are six elements in the sample space, and two elements in the event space.
    - Thus, the probability of M, or $P(M)$, is $\frac{2}{6}$.

Practicing Probability from Sample Spaces

- Write the sample space for the random process of tossing a coin twice.
  - What is the probability of tossing a coin twice and getting two heads?
  - What is the probability of tossing a coin twice and getting one head and one tail, in any order?
- Write the sample space for the random process of drawing a piece of paper out of a bag, when the papers have a number from one to ten written on them.
  - What is the probability a number is odd?
  - What is the probability a number is prime (divisible only by one and itself)?

Coin Tosses

$$S = \{HH, HT, TH, TT\}$$

- $P(HH) = \frac{1}{4}$ (One way to get two heads, four outcomes total)
- $P(1\ H, 1\ T) = \frac{2}{4}$ (Two ways to get a head and a tail, four outcomes total)

Paper Bag

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

- $P(Odd) = \frac{5}{10}$ (Five ways to get an odd number–1, 3, 5, 7, 9–and ten numbers total)
- $P(Prime) = \frac{4}{10}$ (Four ways to get a prime number–2, 3, 5, 7–and ten numbers total)

Events 1

- **Disjoint or Mutually Exclusive**
  - Two events are called disjoint or mutually exclusive if they cannot both happen.
    - Example: Rolling an even number and rolling an odd number.
- **Independent**
  - Two events, or processes are independent if knowing the outcome of one event does not provide any information about the outcome of another event.
    - Example: Rolling two dice at the same time.
- Can a pair of events be both mutually exclusive AND independent?

Events 2

The probability of A, called $P(A)$, and B, called $P(B)$, happening together is...

- 0 if A and B are mutually exclusive.
- $P(A) \times P(B)$ if A and B are independent.
- Some other value, $P(A \text{ and } B)$, that I would HAVE to tell you!

Practicing Events 1

A standard deck of cards has 4 suits of 13 cards each (52 total). The four suits are hearts (red), diamonds (red), spades (black), and clubs (black). Each suit has cards numbered 2 through 10 and four face cards (Ace, Jack, Queen, and King).

- What is the probability of drawing a red card ?
- What is the probability of drawing a face card?
- What is the probability of drawing a face card and a numbered card at the same time?
  - Are these events independent, mutually exclusive, or neither?
- What is the probability of drawing a red face card?
  - Are these events independent, mutually exclusive, or neither?

Practicing Events 2

- $P(Red) = \frac{1}{2}$ (26 ways to draw a red card, 52 cards total)
- $P(Face) = \frac{4}{13}$ (16 ways to draw a face card–a jack, queen, king, and ace in each of the four suits–and 52 cards total)
- $P(Face \text{ and } Number) = 0$ (there are no cards that have faces and numbers! This event is mutually exclusive).
- $P(Face \text{ and } Red) = \frac{2}{13}$ (there are four face cards with diamonds and four face cards with hearts, and 52 cards total)

General Addition Rule

If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

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

In a group of 250 students, 57 are taking a chemistry class, 45 are taking an English class, and 75 are taking a history class.

- If 21 students are taking both chemistry and English, what is the probability that a randomly chosen student is taking chemistry or English?
- If 34 students are taking both English and history, what is the probability that a randomly chosen student is taking English and history?
- If 26 students are taking both chemistry and history, what is the probability that a randomly chosen student is taking chemistry or history?

Practicing General Addition

- $P(C \text{ or } E) = \frac{57}{250} + \frac{45}{250} - \frac{21}{250} = \frac{81}{250}$
- $P(E \text{ and } H) = \frac{34}{250}$
- $P(C \text{ or } H) = \frac{57}{250} + \frac{75}{250} - \frac{26}{250} = \frac{106}{250}$

General Multiplication Rule

If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$P(A \text{ and } B) = P(A) \times P(B)$$

Similarly, if there are k events $A_1, A_2, ..., A_k$ from k independent processes, then the probability they all occur is

$$P(A_1) \times P(A_2) \times ... \times P(A_k)$$

Practicing General Multiplication 1

Assume that on any given day, the probability I wear my glasses is 33%, the probability that I cook meat with dinner is 57%, and the probability that I drink tea instead of coffee with breakfast is 10%. In addition, assume that all days are independent of each other.

- What is the probability that I drink tea and wear my glasses?
- What is the probability that I cook meat with dinner three nights in a row?
- What is the probability that I drink tea or cook meat with dinner?

Practicing General Multiplication 2

- $P(T \text{ and } G) = 0.33 \times 0.10 = 0.033$
- $P(M_1 \text{ and } M_2 \text{ and } M_3) = P(M_1) \times P(M_2) \times P(M_3) = 0.57^3 = 0.185$
- $P(T \text{ or } M) = P(T) + P(M) - P(T \text{ and } M) = 0.1 + 0.57 - (0.1 \times 0.57) = 0.613$

Complements 1

- **Complement**
  - The complement in an event is defined as all outcomes in a sample space that are not that event.
  - Examples:
    - The complement of heads is tails.
    - The complement of having brown eyes is having "not brown" eyes–a.k.a, blue, hazel, amber, or green.
- Given a standard deck of cards. . .
  - What is the complement of the event "drawing a face card"?
  - What is the complement of the event "drawing a black card"?
  - What is the complement of the event "drawing a diamond"?

Practicing Complements

- The complement of drawing a face card is drawing a numbered card.
- The complement of drawing a black card is drawing a red card.
- The complement of drawing a diamond is drawing a club, heart, or spade.

Complements 2

- The complement of an event A is denoted as A^c.
- In a valid probability distribution, the following must be true:

$$P(A) + P(A^c) = 1, \text{ or } P(A^c) = 1 - P(A)$$

In a randomly chosen sample of 169 Americans:

- 46 own a dog, but no cat;
- 29 own a cat, but no dog;
- 21 own a cat and a dog;
- 5 own pets other than cats or dogs; and
- 68 own no pets.

What is the probability that one of these Americans owns at least one pet?

Practicing More Complements

$$
\begin{aligned}
P(\text{At least one pet}) =& P(\text{Dog, no cat}) + P(\text{Cat, no dog}) \\
&+ P(\text{Cat and dog}) + P(\text{Other pet}) \\
=& \frac{46 + 29 + 21 + 5}{169} \\
= 1 - P(\text{No pets}) & \\
= 1 - \frac{68}{169} & \\
= \frac{101}{169} &
\end{aligned}
$$

[0] https://news.gallup.com/poll/25969/americans-their-pets.aspx>

Probability Distributions

- A **probability distribution** probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

1. The outcomes must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must total 1.

Visualizing Probability Distributions

- Venn Diagrams: Useful when outcomes can be categorized as "in" or "out" of two or three groups.
- One-way Table: Useful for categorical or numerical variables with a small number of levels.
- Contingency Tables: Useful for comparing two categorical variables with various numbers of levels.

Venn Diagrams

Venn Diagrams use circles to represent different events or groups. If the circles overlap, there are objects/people that are members of two or more groups at the same time.

- Say I walk into a bakery selling donuts. There are 166 donuts in total.
  - 126 are yeast donuts.
  - 78 have sprinkles.
  - 57 have chocolate frosting.

Practicing Venn Diagrams 1

- Are having chocolate frosting and sprinkles mutually exclusive?
- What is the probability a donut has sprinkles?
- Sprinkles and chocolate frosting?
- Sprinkles or chocolate frosting?
- Sprinkles but not chocolate frosting?

Reminder: 166 donuts in total (126 yeast, 78 sprinkles, 57 chocolate)

Practicing Venn Diagrams 2

- No! There are 25 donuts that have both chocolate frosting and sprinkles, so these events are not mutually exclusive.
- $P(S) = \frac{78}{166}$
- $P(S \text{ and } C) = \frac{25}{166}$
- $P(S \text{ or } C) = \frac{78}{166} + \frac{57}{166} - \frac{25}{166} = \frac{110}{166}$
- $P(S \text{and} C^c) = \frac{78}{166} - \frac{25}{166} = \frac{53}{166}$

One-way Tables 1

One-way tables usually give a list of the values a single variable can take and the corresponding probabilities.

- Example: the probability distribution for the sum of two dice (call it X).

X	2	3	4	5	6	7	8	9	10	11	12
$P(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Practicing One-way Tables 1

Which of these is a valid probability distribution? Why?

X	0	1	2	3
$P(X)$	0.25	0.30	-0.30	0.75

X	0	1	2	3
$P(X)$	0.25	0.10	0.50	0.25

X	0	1	2	3
$P(X)$	0.25	0.10	0.40	0.25

Practicing One-Way Tables 2

Color	Brown	Blue	Hazel	Amber	Green
$P(\text{Color})$	0.79	0.09	0.05	0.05	0.02

- Is this a valid probability distribution?
- What is the probability a randomly selected person has brown eyes? Amber or hazel eyes? Does not have blue eyes?

Practicing One-Way Tables 3

- Is this a valid probability distribution?
- What is the probability a randomly selected person has brown eyes? Amber or hazel eyes? Does not have blue eyes?

- Yes, this is a valid distribution, since it follows all the rules (disjoint outcomes, probabilities between 0 and 1, probabilities sum to 1).

- $P(Brown) = 0.79$

- $P(Amber \text{ or } Hazel) = 0.05 + 0.05 = 0.1$

- $P(Blue) = 1 - 0.09 = 0.91$

Contingency Tables 1

- A table that summarizes data for two categorical variables is called a **contingency table**.

Contingency Tables in R

The `titanic` dataset includes information on the survival status, sex, age, and passenger class of 1309 passengers on the *RMS Titanic*'s first and final voyage in 1912.

```
colnames(titanic)
```

```
## [1] "X"              "survived"       "sex"            "age"
## [5] "passengerClass"
```

```
head(titanic, n = 4)
```

```
##                                X survived    sex     age passenge
## 1   Allen, Miss. Elisabeth Walton      yes female 29.0000
## 2  Allison, Master. Hudson Trevor      yes   male  0.9167
## 3    Allison, Miss. Helen Loraine       no female  2.0000
## 4 Allison, Mr. Hudson Joshua Crei       no   male 30.0000
```

Contingency Tables in R 2

```
table(titanic$passengerClass, titanic$survived)
```

```
##
##          no yes
##    1st  123 200
##    2nd  158 119
##    3rd  528 181
```

- How many passengers were in each class?

```
123+200
158+119
528+181
```

Contingency Tables in R 3

```
titanic_table <- table(titanic$passengerClass, titanic$survived)
titanic_table
```

```
##
##          no yes
##     1st 123 200
##     2nd 158 119
##     3rd 528 181
```

```
rowSums(titanic_table) ## What does rowSums() do?
```

```
## 1st 2nd 3rd
## 323 277 709
```

- Can you guess the corresponding function for columns?

Practicing Contingency Tables 1

	Did not Survive	Survived	Total
1st Class	123	200	323
2nd Class	158	119	277
3rd Class	528	181	709
Total	809	500	1309

- What is the probability a randomly selected passenger was in third class?
- What is the probability a randomly selected passenger survived?

Practicing Contingency Tables 2

- The corresponding function for column sums is `colSums()`.
- $P(\text{3rd Class}) = \frac{709}{1309}$
- $P(\text{Survived}) = \frac{500}{1309}$

- **Conditional Probability**
  - A **conditional probability** is a probability of an event computed under a condition.
  - Conditional probabilities have two parts:
    - The **outcome of interest** is the event for which we are computing the probability.
    - The **condition** is information we know to be true.
  - Formally, we write

$$P(\textit{Outcome of Interest} \mid \textit{Condition})$$

- We read the vertical bar, |, as "given".

Practicing Conditional Probability 1

	Did not Survive	Survived	Total
1st Class	123	200	323
2nd Class	158	119	277
3rd Class	528	181	709
Total	809	500	1309

- What is the probability a randomly selected passenger survived, given that they were in third class?
- What is $P(\text{ 1st Class} \mid \text{Did not survive})$?

Practicing Conditional Probability 2

- The probability a randomly selected passenger survived, given that they were in third class, is formally written $P(\text{ Survived } \mid \text{ 3rd Class })$.

- $P(\text{ Survived } \mid \text{ 3rd Class }) = \frac{181}{709}$

- $P(\text{ 1st Class } \mid \text{ Did not survive}) = \frac{123}{809}$

Random Variables 1

- **Random Variable**
  - A **random variable** is a variable or process with a numerical outcome.
  - We usually represent random variables with a capital letter, such as X, Y, or Z.
- **Expected Value**
  - The expected or average outcome of a random variable X is called the **expected value**, denoted $E(X)$.
- **Variance** and **Standard Deviation**
  - The variability of a random variable X is summarized with the **variance**, denoted $Var(X)$.
  - The standard deviation is the square root of the variance.

Expected Value

If a random variable X

- takes outcomes $x_1, x_2, ..., x_k$ with
- probabilities $P(X = x_1)$, $P(X = x_2)$, . . . , $P(X = x_k)$,

the expected value of X is the sum of each outcome multiplied by its corresponding probability.

$$\begin{aligned} E(X) &= x_1 \times P(X = x_1) + x_2 \times P(X = x_2) + \\ &\quad ... + x_k \times P(X = x_k) \\ &= \sum_{i=1}^{k} x_i P(X = x_i) \end{aligned}$$

Expected Value: Example

Let X be a random variable with probability distribution

X	0	1	2	3
$P(X)$	0.7	0.10	0.10	0.10

$$\begin{aligned} E(X) &= x_1 \times P(X = x_1) + x_2 \times P(X = x_2) + \ldots + x_k \times P(X = x_k) \\ &= 0 \times 0.7 + 1 \times 0.10 + 2 \times 0.1 + 3 \times 0.1 \\ &= 0 + 0.10 + 0.20 + 0.30 \\ &= 0.6 \end{aligned}$$

Variance

If a random variable X takes...

- outcomes $x_1, \ldots, x_k$ with probabilities
- $P(X = x_1), \ldots, P(X = x_k)$, and
- expected value $E(X)$, then the variance of X is

$$\begin{aligned} Var(X) &= (x_1 - E(X))^2 \times P(X = x_1) + \ldots + (x_k - E(X))^2 \times P(X = x_k) \\ &= \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i) \\ &= \sigma^2 \end{aligned}$$

- The standard deviation is the square root of the variance.

Expected Value and Variance in R

```
x <- c(0, 1, 2, 3)
probs <- c(0.7, 0.1, 0.1, 0.1)
sum(x*probs)
```

```
## [1] 0.6
```

```
exp_val <- sum(x*probs)
sqrt(sum((x-exp_val)^2*probs))
```

```
## [1] 1.019804
```

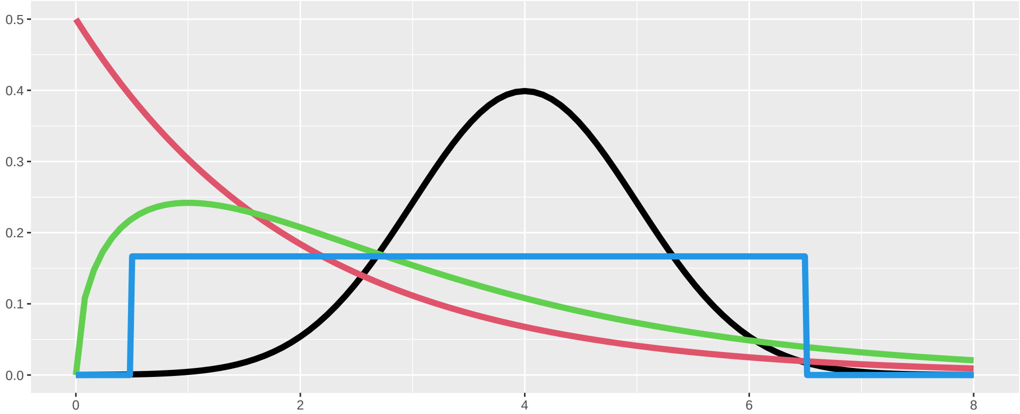
More Probability Practice

- Write the sample space for the number of siblings a randomly chosen person could have (ignore half/step-siblings).

Probability Density Function 1

- **Probability Density Functions** are functions that define the probabilities of the possible outcomes of a continuous variable (we can also talk about similar functions called **probability mass functions** for discrete variables).
- When plotted, these functions show the center, shape, and spread of the variable.
- These curves or functions have to follow the same rules as any other probability distribution.
    - The values the variable takes must be disjoint (e.g., a randomly chosen person cannot have both 6 and 7 siblings).
    - Each probability defined by the function must be between 0 and 1.
    - The probabilities defined by the function must total 1.

Probability Density Function 2



Parameters and Statistics 1

- Means
  - The expected value, $E(X)$, is also known as the population mean, or μ.
  - We call the mean of a sample $\bar{x}$.
- Standard Deviations
  - The standard deviation of a population is written as σ.
  - We call the standard deviation of a sample s.
- Populations: Greek Letters
- Samples: Roman/Latin Alphabet Letters (e.g., x, s, p)

Parameters and Statistics 2

Characteristic	Parameter	Statistic
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	—	$\hat{p}$
Correlation	—	R
Intercept	β_0	b_0
Slope	β_1	b_1

A Note on Probability

- The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an *INFINITE* number of times.
  - Since observing a random process an infinite number of times is impossible, we are usually calculating theoretical probabilities based on some assumptions about the scenario at hand.
  - Events with high probabilities are not guaranteed to happen, and events with low probabilities are not guaranteed to not happen.
- Humans are very bad at interpreting probabilities–always remember we are talking about the long run!