



# STAT 216 Activity Coursepack

Fall 2020

---

## Contents

---

Preface	2
Fall 2020 Calendar of In-Class Activities	3
1 Martian Alphabet	5
2 Study Design	12

---

## Preface

---

This coursepack accompanies the textbook for STAT 216: Introduction to Statistics at Montana State University. Each of the activities in this workbook is designed to target specific learning outcomes of the course, giving you practice with important statistical concepts in a group setting with instructor guidance. Bring this workbook with you to class each week, and take notes in the workbook as you would your own notes. A well-written complete workbook will provide an optimal study guide for exams!

---

## Fall 2020 Calendar of In-Class Activities

---

Week	Activity No.	Day	Date	Activity
1	1	M	8/17	Martian Alphabet
1	1	T	8/18	Martian Alphabet
1	1	W	8/19	Martian Alphabet
1	1	H	8/20	Martian Alphabet
1	1	F	8/21	Martian Alphabet
2	2	M	8/24	Study Design
2	2	T	8/25	Study Design
2	2	W	8/26	Study Design
2	2	H	8/27	Study Design
2	2	F	8/28	Study Design
3	3	M	8/31	Current Population Survey
3	3	T	9/1	Current Population Survey
3	3	W	9/2	Current Population Survey
3	3	H	9/3	Current Population Survey
3	3	F	9/4	Current Population Survey
4	-	M	9/7	No class – Labor Day
4	4	T	9/8	IMDb Movie Reviews: Part I
4	4	W	9/9	IMDb Movie Reviews: Part I
4	4	H	9/10	IMDb Movie Reviews: Part I
4	4	F	9/11	IMDb Movie Reviews: Part I
5	4	M	9/14	IMDb Movie Reviews: Part I
5	5	T	9/15	IMDb Movie Reviews: Part II
5	5	W	9/16	IMDb Movie Reviews: Part II
5	5	H	9/17	IMDb Movie Reviews: Part II
5	5	F	9/18	IMDb Movie Reviews: Part II
6	5	M	9/21	IMDb Movie Reviews: Part II
6	-	T-F	9/22-9/25	Exam 1
7	-	M	9/28	Exam 1
7	6	T	9/29	Handedness of Male Boxers
7	6	W	9/30	Handedness of Male Boxers
7	6	H	10/1	Handedness of Male Boxers
7	6	F	10/2	Handedness of Male Boxers

Week	Activity No.	Day	Date	Activity
8	6	M	10/5	Handedness of Male Boxers
8	7	T	10/6	Helmet Use and Head Injuries
8	7	W	10/7	Helmet Use and Head Injuries
8	7	H	10/8	Helmet Use and Head Injuries
8	7	F	10/9	Helmet Use and Head Injuries
9	7	M	10/12	Helmet Use and Head Injuries
9	8	T	10/13	COVID-19 and Air Pollution
9	8	W	10/14	COVID-19 and Air Pollution
9	8	H	10/15	COVID-19 and Air Pollution
9	8	F	10/16	COVID-19 and Air Pollution
10	8	M	10/19	COVID-19 and Air Pollution
10	9	T	10/20	Record Snowfall
10	9	W	10/21	Record Snowfall
10	9	H	10/22	Record Snowfall
10	9	F	10/23	Record Snowfall
11	9	M	10/26	Record Snowfall
11	10	T	10/27	Hand Dexterity
11	10	W	10/28	Hand Dexterity
11	10	H	10/29	Hand Dexterity
11	10	F	10/30	Hand Dexterity
12	10	M	11/2	Hand Dexterity
12	-	T	11/3	No class — Election Day
12	-	W-F	11/4-11/6	Exam 2
13	-	M-T	11/9-11/10	Exam 2
13	-	W	11/11	No class — Veterans Day
14	-	H-W	11/12-11/18	Review

# ACTIVITY 1

---

## Martian Alphabet

---

### 1.1 Learning outcomes

- Describe the statistical investigation process
- Identify observational units, variables, and variable types in a statistical study

### 1.2 Can you read “Martian”?

How well can humans distinguish one “Martian” letter from another? In today’s activity, we’ll find out. When shown the two Martian letters, Kiki and Bumba, write down whether you think Bumba is on the left or the right.

#### 1.2.1 Steps of the statistical investigation process

The first step of any statistical investigation is to ask a research question. In this study the research question is: can we as a class read Martian? (we will refine this later on!). To answer any research question, we must design a study and collect data. (This will normally be provided for you in class.) For our question, the study consists of each student being presented with two Martian letters and asking which was Bumba. Your responses will become our observed data that we will explore. To answer the research question we will simulate what *could* have happened in our class given random chance, repeat that many times to understand the expected variability between different “randomly guessing” classes, then comparing our class’s observed data to the simulation. This gives us an estimate of how often (or the probability of) our class’s result would occur if we were all merely guessing, allowing us to determine if we as a class can in fact read Martian.

Let’s explore the data. **Observational units** or **cases** are the subjects data is collected on. In a data set the rows will represent a single observational unit.

1. What are the observational units in this study?

2. How many students are in class today? This is the sample size.

A **variable** is information collected or measured on each observational unit or case. We will look at two types of variables: **quantitative** and **categorical**. Each column in a data set will represent a different variable.

Quantitative variables are numerical measurements that can be discrete (whole, non-negative numbers) or continuous (any value within an interval). The number of students in a class would be a discrete variable as you can not have a partial student. GPA would be a continuous variable ranging from 0 to 4.0.

Categorical variables are data that are in groups or categories such as eye color, state of residency, or whether or not a student is considered in-state. Categorical variables with a natural ordering are considered ordinal variables while those without a natural ordering are considered a nominal variable. All variables will be treated as nominal for analysis.

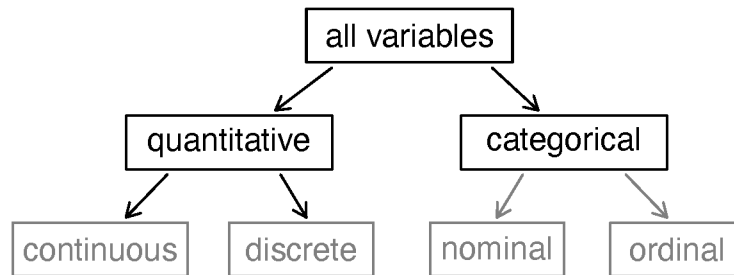


Figure 1.1: Types of variables.

3. Identify the variable we are collecting on each observational unit in this study, i.e., what are we measuring on each student?

It is important to note that a variable is different than a summary statistic. A variable is measured on a **single observational unit** while a summary statistic is calculated from a group of observational units. For example, the variable **whether or not a student is considered in-state** can be measured on each individual student. In a class of 50 students we can calculate the proportion of students who are considered in-state, the summary statistic. Make sure you wrote the variable in question 3 as a variable **NOT** a summary statistic.

4. Is the variable identified in question 3 categorical or quantitative?
5. Were you correct or incorrect in identifying Bumba?

We will now collect the data from the entire class.

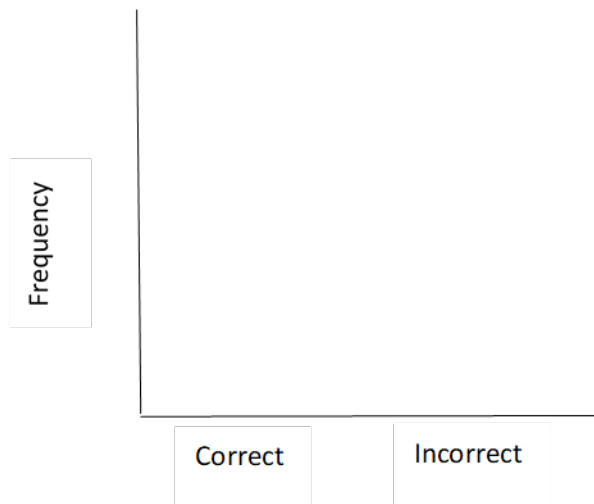
6. How many people in your class were correct in identifying Bumba? Using the class size from question 2, calculate the proportion of students who correctly identified Bumba.

$$\text{proportion} = \frac{\text{number of students who correctly identified Bumba}}{\text{total number of students}}$$



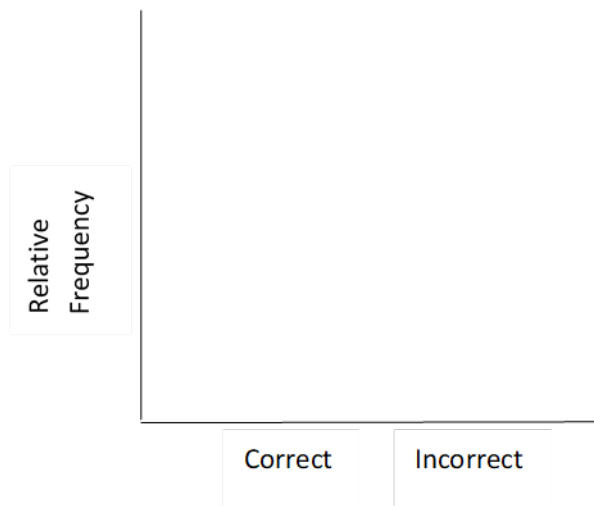
Looking at the data set and the summary statistics is only one way to display the data. We will also want to create a visualization or picture of the data. A **frequency bar plot** is used to display categorical data as a count or frequency. Since our variable has two levels, correct or incorrect, we will create two bars one for each level.

7. Plot the observed class data using a frequency bar plot.



We can also visualize the data as a proportion in a **relative frequency bar plot**. Relative frequency is the proportion calculated for each level of the categorical variable.

8. Plot the observed class data using a relative frequency bar plot.



9. The next step is to analyze the data. If humans really don't know Martian and are just guessing which is Bumba, what are the chances of getting it right?

How could we use a coin to simulate each student "just guessing" which martian letter is Bumba?

How could we use coins to simulate the entire class "just guessing" which martian letter is Bumba?

How many people in your class would you expect to choose Bumba correctly just by chance? Explain your reasoning.

10. Each of you will flip a coin one time to simulate your "guess". Let Heads = correct, Tails = incorrect. What was the result of your simulation?

What was the result from your class's simulation? What proportion of students "guessed" correctly in the simulation?

11. If students really don't know Martian and are just guessing which is Bumba, which seems more unusual: the result from your class's **simulation** or the observed proportion of students in your class that were correct (this is your data from question 6)? Explain your reasoning.

12. While your observed class data is likely far different from the simulated “just-guessing” class, comparing our class data to a single simulation does not seem to give enough information. The differences seen could just be due to that set of coin flips! Let’s simulate another class. Each student should flip your coin again. What was the result from your class’s second simulation? What proportion of students “guessed” correctly in the second simulation? Create a plot to compare the two simulated results with the observed class result.
13. We still unfortunately only have a couple of simulations to compare our class data to. It would be much better to be able to see how our class compared to hundreds or thousands of “just-guessing” classes. Since we don’t want to flip coins all class period, your instructor will use a computer simulation to get 1000 trials. Fill in the following blanks to describe how we would create a simulation of random guessing with 1000 trials.
- Probability of correct guesses: \_\_\_\_\_
- Sample size: \_\_\_\_\_
- Number of repetitions: \_\_\_\_\_
14. Sketch the distribution displayed by your instructor here, being sure to label each axis appropriately.
15. Is your class particularly good or bad at Martian? How can you use the plot in question 14 to tell?
16. Is it *possible* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.

17. Is it *likely* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.
  
18. Does this activity provide strong evidence that students were not just guessing at random? If so, what do you think is going on here? Can we as a class read Martian?

### 1.3 Take home messages

1. In this course we will learn how to evaluate a claim by comparing observed results (classes' "guesses") to a distribution of many simulated results under an assumption like "blind guessing."
2. Blind guessing between two outcomes will be correct only about half the time. We can create data (via computer simulation) to fit the assumption of blind guessing.
3. Unusual observed results will make us doubt the assumptions used to create the simulated distribution. A large number of correct "guesses" is evidence that a person was not just blindly guessing.

### 1.4 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity, and to write down the names and contact information of your team mates.

# ACTIVITY 2

---

## Study Design

---

### 2.1 Learning outcomes

- Explain the purpose of random sampling and its effect on scope of inference
- Explain the purpose of random assignment and its effect on scope of inference
- Identify whether a study is observational or an experiment
- Identify confounding variables in observational studies and explain why they are confounding
- Identify the types of bias present in a study

### 2.2 Terminology review

Statistics is the study of how best to collect, analyze, and draw conclusions from data. Statistical inference will allow us to make a statement about a population parameter based on a sample statistic.

Some terms covered in this activity are...

- Population
- Sample
- Parameter
- Statistic
- Selection Bias
- Response Bias
- Non-response Bias
- Scope of Inference
- Explanatory Variable
- Response Variable

- Confounding Variable
- Experiments
- Observational Study

To review these concepts see Section 1.3 to 1.6 in the textbook.

## 2.3 Types of sampling bias

There are two parts to study design: how the sample was selected and how the study was conducted. First we will look at sampling and types of bias.

In these next questions, identify the target population, the sample, the variable, and the type of bias present.

1. To determine if the proportion of out of state undergraduate students at Montana State University has increased in the last 10 years, a statistics instructor sent an email survey to 500 randomly selected current undergraduate students. One of the questions on the survey asked whether they had in-state or out-of-state residency. She only received 378 responses.

Target population:

Sample:

Variable:

Type of Bias:

2. PEW Research surveys US adults about many different topics. Recently a survey was conducted to assess current presidential approval. A random sample of 6395 US adults was taken. Of those surveyed, 42% say they agree with President Trump on many or nearly all of the top issues facing the country today.

Target population:

Sample:

Variable:

Type of Bias:

3. A television station is interested in predicting whether or not voters in its listening area are opposed to legalizing marijuana for adult use. It asks its viewers to phone in and indicate whether they are in favor of this or opposed to this. Of the 2241 viewers who phoned in, forty-five percent were opposed to legalizing marijuana.

Target population:

Sample:

Variable:

Type of Bias:

4. To gauge the interest in a new swimming pool, a local organization stood outside of the Bogart Pool during open hours. One of the questions they asked was, "Since the Bogart Pool is in such bad repair, don't you agree that the city should fund a new pool?"

Target population:

Sample:

Variable:

Type of Bias:

5. The Bozeman school district is interested in surveying parents of students about their opinions on returning to school this fall following the COVID-19 pandemic. They divided the school district into 10 divisions based on location and randomly surveyed 20 households within each division.

Target population:

Sample:

Variable:

Type of Bias:




## 2.4 Study design


The two main study designs we will cover are observational studies and experiments. Both the sampling method and the study design will help to determine the **scope of inference** for a study. Remember that only in a randomized experiment can we conclude a **causal** (cause and effect) relationship between the explanatory and response variable.

*Scope of Inference:* If evidence of an association is found in our sample, what can be concluded?

	Study Type		
Selection of cases	Randomized experiment	Observational study	
Random sample (and no other sampling bias)	Causal relationship, and can generalize results to population.	Cannot conclude causal relationship, but can generalize results to population.	Inferences to population can be made
No random sample (or other sampling bias)	Causal relationship, but cannot generalize results to a population.	Cannot conclude causal relationship, and cannot generalize results to a population.	Can only generalize to those similar to the sample due to potential sampling bias



Can draw cause-and-effect conclusions



Can only discuss association due to potential confounding variables

For the next exercises, identify the explanatory variable, the response variable, a potential confounding variable, and the study design.

- The pharmaceutical company, Moderna Therapeutics is working in conjunction with the National Institute of Health towards a vaccine for COVID-19 and has recently begun Phase 3 clinical trials. US Clinical research sites will enroll 30,000 volunteers without COVID-19 to participate. Participants will be randomly assigned to receive either the candidate vaccine or a saline placebo. They will then be followed to assess vaccine related symptoms and development of COVID-19. The trial is blinded, so the investigators and the participants will not know who is assigned to which group.

Explanatory Variable:

Response Variable:

Confounding Variable:

Study design:

What is the scope of inference for this study?

7. In another study, a local health department randomly selected 1000 US adults without COVID-19 to participate in a health survey. Each participant was assessed at the beginning of the study and then followed for 1 year. They were interested to see which participants elected to receive a vaccination for COVID-19 and whether any participants developed COVID-19.

Explanatory Variable:

Response Variable:

Confounding Variable:

Study design:

What is the scope of inference for this study?

8. What is a potential confounding variable for the study in question 7? Explain how this meets the definition of a confounding variable.

## 2.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity.