STA 1013 : Statistics through Examples

Lecture 3: Terminology, Research Design & Sampling

Hwiyoung Lee

September 4, 2019

Department of Statistics, Florida State University

Overview

1. Review

2. Research Design : Types of Statistical Studies

Review

Data types & Variable types

Types of Data

- **Primary Data** : Data collected by the investigator himself/ herself for a specific purpose
- **Secondary Data** : Data collected by someone else for some other purpose, or published elsewhere

Types of Variable

- **Categorical variable** : Result in categorical responses
- **Quantitative variable** : Result in numerical responses, can be used directly in calculations

Exercise

Exercise

Political Parties In a preelection survey of likely voters, political parties of respondents are identified as 1 for a Democrat, 2 for a Republican, 3 for an Independent, and 4 for anything else. The average (mean) is calculated for 850 respondents and the result is 1.7.

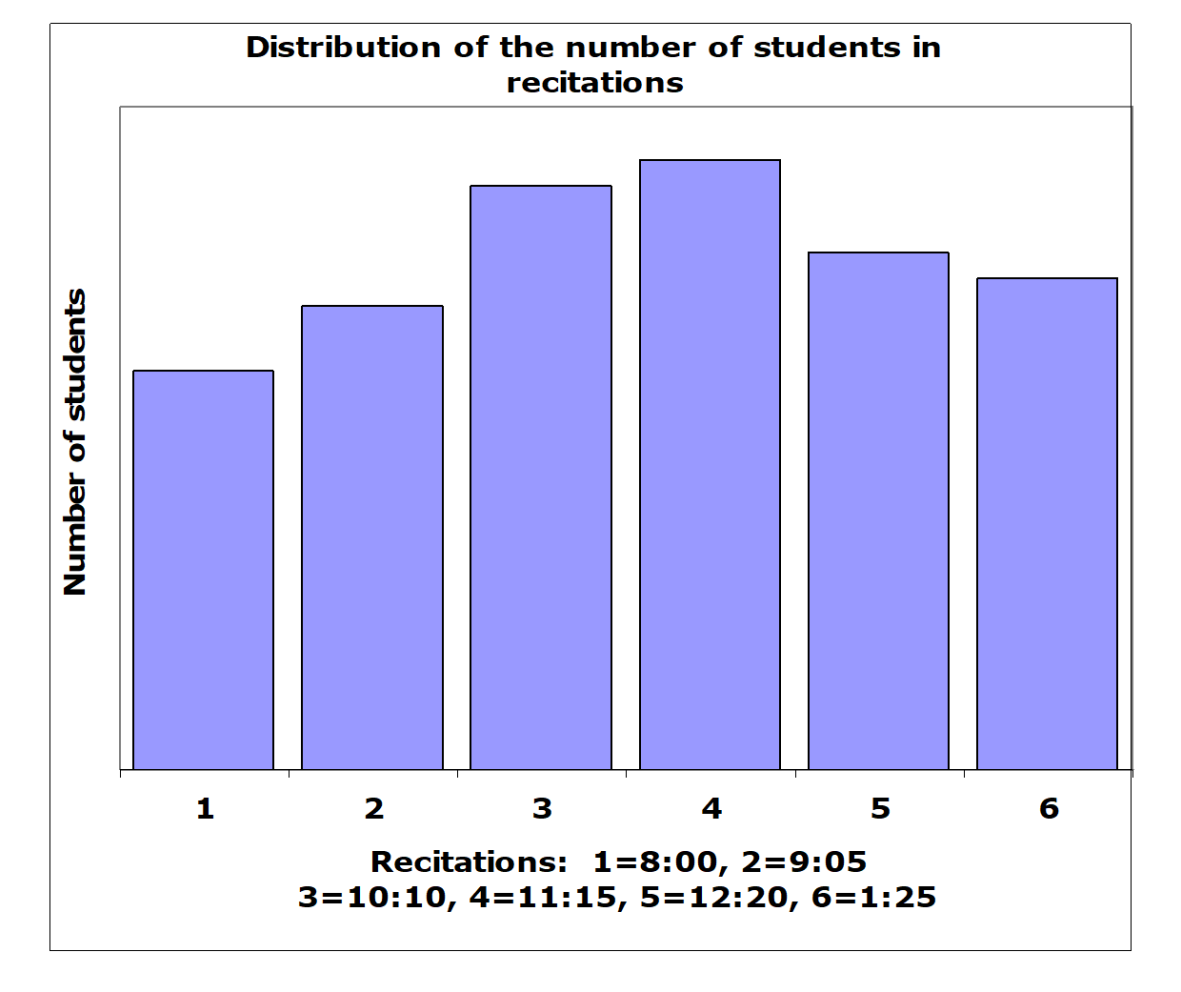
- What is the variable ?
- Is it Categorical or Quantitative ?
- What is worg with the given calculation ?

Review

Distribution of a variable :

- The way its values are spread over all possible values
- Summarize a distribution with a table or a graph
    - **Categorical variable :**
        1. **Frequency Table**
        2. **Bar chart**
        3. **Pie chart**

Exercise


Distribution of the number of students in recitations
Number of students
1
2
3
4
5
6
Recitations: 1=8:00, 2=9:05
3=10:10, 4=11:15, 5=12:20, 6=1:25

For the distribution displayed in the bar chart above, classify the following statements as true (T) or false (F).

True or False

1. The 8:00, 9:05 and 1:25 recitations together had more than half the students. (T , F)
2. The 9:05, 11:15 and 12:20 recitations together had more than half the students. (T , F)
3. The difference between the number of students in the 10:10 and 11:15 recitations was less than the difference between the number of students in the 8:00 and 11:15 recitations. (T , F)
4. The 12:20 recitation was less popular than the 1:25. (T , F)
5. The 11:15 recitation had the most students. (T , F)

Contingency Table

What if we have more than 2 variable ?

Class Rank * Gender Crosstabulation

Count

		Gender		
		Male	Female	Total
Class Rank	Freshman	14	9	23
	Sophomore	15	13	28
	Junior	9	11	20
	Senior	12	14	26
Total		50	47	97

Dot plot

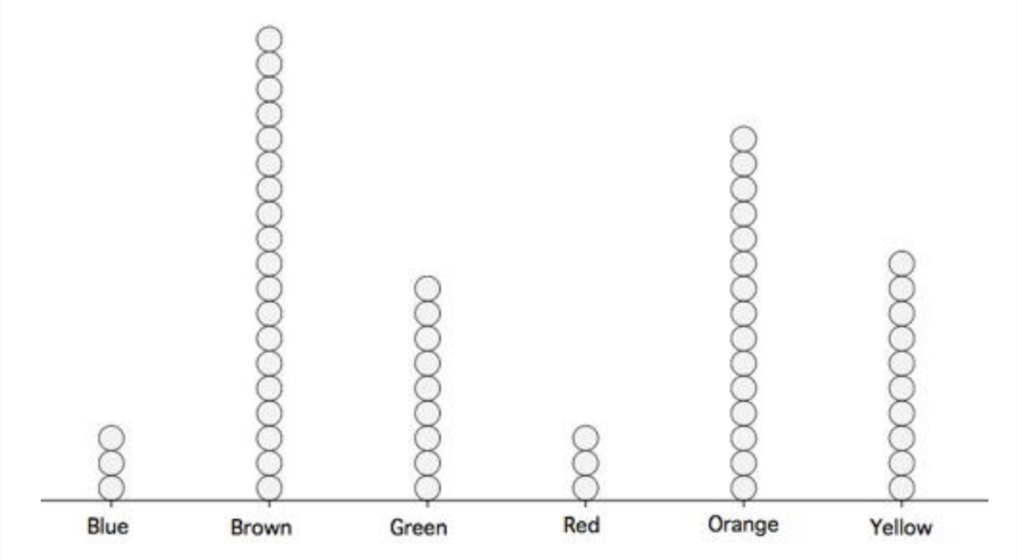


Figure 1: the number of M & M's of various colors in a bag of M & M's.

- Each dot represents a value
- A value occurs more than once, the dots are placed one above the other
- The height of the column of dots represents the frequency for that value.

Graphical summary of Quantative variable

Graphical Summary for distributions of Quantative variable?

Below are the weights (lbs) of the 4-month-old babies.

14.1	12.1	15.7	14.0	15.8	12.6	11.3	14.9	12.0	14.5
15.6	15.3	12.9	14.8	11.4	16.8	14.3	11.4	15.0	14.6
12.6	14.4	16.2	15.2	16.4	14.8	11.6	14.9	16.7	15.2

- Can we summarize the above data by Frequency table, Pie chart, Bar chart ? **No**
- How can we graphically summarize the quantative variable ? **Stem and Leaf plot, Histogram, Boxplot**

Stem and leaf plot

Stem	Leaves									
11	3	4	4	6						
12	0	1	6	6	9					
13										
14	0	1	3	4	5	6	8	8	9	9
15	0	2	2	3	6	7	8			
16	2	4	7	8						

How can we draw a Stem and leaf plot?

- Integer as the stem and the decimal as the leaf
- **Stems** should be in ascending order from **top to bottom** with none missing between the first and last
- **Leaves** should be listed in ascending order **from left to right** with no gaps

Exercise

The following are 24 soil pH measurements:

5.3	6.1	6.2	6.5	6.7	6.8	7.1	7.4	7.7	7.8	7.9
8.0	8.1	8.2	8.4	8.7	9.0	9.3	9.5	9.6	9.7	10.0

Create a stem and leaf diagram using the integer as the stem and the decimal as the leaf.

Stem and leaf plot

Advantages :

- Show individual data values along with classes of values
- Quick and Easy to construct for **small data sets**

Disadvanteges :

- Only useful for small data sets
- For large data set : Histogram, Boxplot

Stem and leaf plot for large data set

The below is the stem and leaf plot for large data set (n = 5,000)

```
10 | 00000001111111112222222222223333333333333334444444444444455555555555+42
11 | 00000000000011111111111111122222222222333333333333344444444444444555+41
12 | 00000000000000011111111111222222222222223333333333344444444444444555+45
13 | 00000000000111111111222222222222233333333444444444444445555555555+23
14 | 00000001111111111111111122222222233333333333333444444444444444555555+42
15 | 00000000000000000011111111111222222222233333333344444445555555555555+31
16 | 00000000000001111111222222222222222233333444444444444444455555555+32
17 | 00000000000000000011111111111111111122222222222222233333333344444444+58
18 | 00000000111111111111122222222222333333333344444444455555555555556+35
19 | 00000011111111122222222222222222233333333333333344444444444444444+53
20 | 00000000000011111111111111122222222222233333333333344444455555555+37
21 | 00000000000000000001111111111111222222222233333333333333444444444444+62
22 | 00000001111111111111111112222223333333344444444444555555555566666+38
23 | 00000011111111122222222233333444444444444555555555556666666666666666+18
24 | 00000000001111112222222222223333333333333333444444445555555555566666+38
25 | 00000000000000001111111111111111222222222223333333333333344444444444+63
26 | 00000000111111111111222222223333333333344444444444555555555555555+30
27 | 00000000111111111111122222222223333333333333333444444444444444444+49
28 | 00000000011111111111111112222222222223333333333333344444444444444444+62
29 | 00000000000011111111111122222222222233333333333334444444444444445555+33
30 | 00000000000000000111111111111111222222222222333333333333344444444444+48
31 | 00000000111111111111122222222222233333333333333334444444444444455555+57
32 | 00000001111111111111222222222223333333333333333444444444455555555555+49
33 | 00000000000011111111111111111122222222222233333333344444444444444455+64
34 | 00000000000111111111111122222222222233333333444444444444455555555556+37
35 | 00000001111111122222222233333333333333334444444445555555555555566+36
36 | 00000000000011111111111111112222222222223333333333344444444444455+37
37 | 00000000001111111111111111112222222222223333333333334444444444555+50
38 | 00000011111111111122222222222333333333334444444444444555555555556+35
39 | 00000000000011111111111112222222223333333333333333334444444444444+41
40 | 00000000000011111111111111112222222222333333333334444444444555555+39
41 | 00000000001111111222222222222222233334444444444455555555555555666+30
42 | 00000000001111112222222223333333333333334444444445555555566666667+20
43 | 00000000000011111111111111222223333333333334444445555555555555555+34
44 | 00000000011111111222222222222233333333333334444445555555555566666+34
45 | 00000000000000111111111111222222222222222223333333334444444444444+54
46 | 00000000000000000111111111112222223333333444445555555555566666666+34
47 | 00000011111111111122222222222222333333333333333444444444444455555+44
48 | 00000000000111111111111222222222222333333344444444444445555555555+55
49 | 00000000000111111112222222222222333333333333333333444444444445555+40
50 | 00000000000001111111111112222222222333333333333444444444444444555+50
```

Wrap up

Summarize a distribution with a table or a graph

- **Categorical variable :**
  1. Frequency Table, Relative Frequency Table
  2. Bar chart
  3. Pie chart
  4. Dot plot
- **Quantative variable :**
  1. Stem and Leaf plot
  2. Histogram (Topic 3)
  3. Box plot (Topic 3)

Research Design : Types of Statistical Studies

Types of Statistical Studies

Statistical studies are conducted in many different ways.

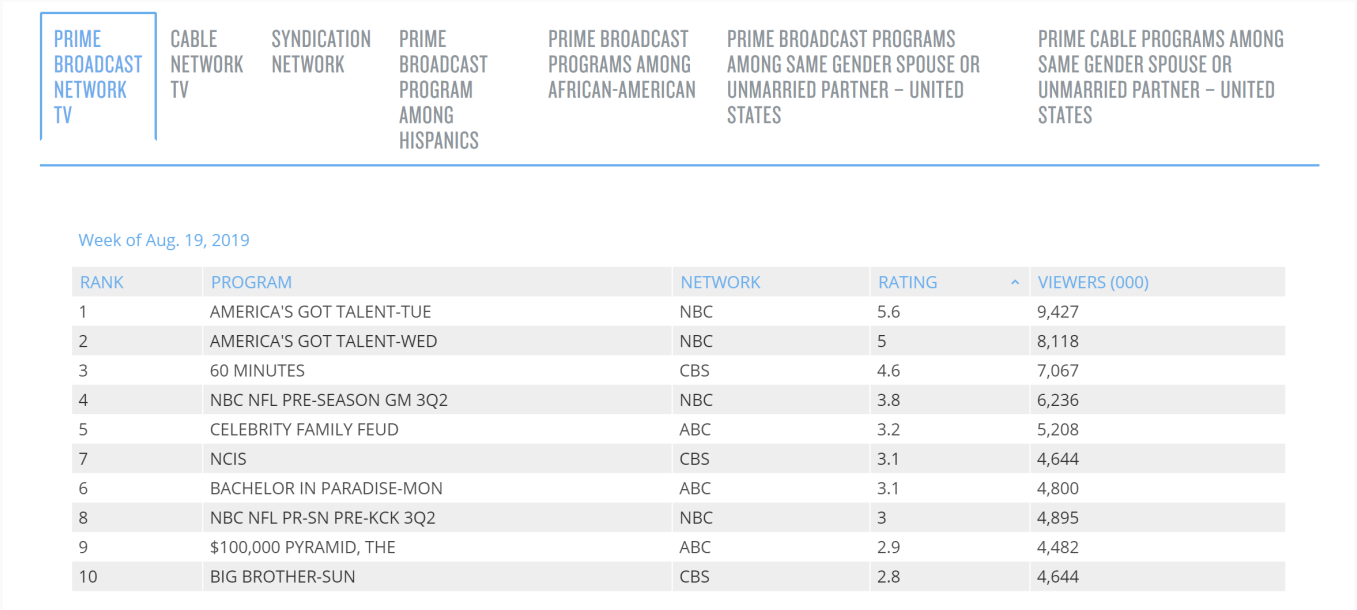
In all cases, the people, animals (or other living things), or objects chosen for the sample are called the **subjects** of the study. If the subjects are people, it is common to refer to them as **participants** in the study.

There are two basic types of statistical study:

- **observational studies**
- **experiments**

Observational studies

Observational Studies : researchers observe or measure characteristics of the subjects, but do not attempt to influence or modify these characteristics.



PRIME BROADCAST NETWORK TV | CABLE NETWORK TV | SYNDICATION NETWORK | PRIME BROADCAST PROGRAM AMONG HISPANICS | PRIME BROADCAST PROGRAMS AMONG AFRICAN-AMERICAN | PRIME BROADCAST PROGRAMS AMONG SAME GENDER SPOUSE OR UNMARRIED PARTNER – UNITED STATES | PRIME CABLE PROGRAMS AMONG SAME GENDER SPOUSE OR UNMARRIED PARTNER – UNITED STATES

Week of Aug. 19, 2019

RANK	PROGRAM	NETWORK	RATING	VIEWERS (000)
1	AMERICA'S GOT TALENT-TUE	NBC	5.6	9,427
2	AMERICA'S GOT TALENT-WED	NBC	5	8,118
3	60 MINUTES	CBS	4.6	7,067
4	NBC NFL PRE-SEASON GM 3Q2	NBC	3.8	6,236
5	CELEBRITY FAMILY FEUD	ABC	3.2	5,208
7	NCIS	CBS	3.1	4,644
6	BACHELOR IN PARADISE-MON	ABC	3.1	4,800
8	NBC NFL PR-SN PRE-KCK 3Q2	NBC	3	4,895
9	$100,000 PYRAMID, THE	ABC	2.9	4,482
10	BIG BROTHER-SUN	CBS	2.8	4,644

Nielsen uses devices to observe what the subjects are watching on TV, but does not try to influence what they watch.

Observational studies

An Observational study may involve activities that go beyond the usual definition of observing.

- **Measuring** people's weights requires interacting with them, as in asking them to stand on a scale. But in statistics, we consider these measurements to be observations because the interactions do not change people's weights.
- Similarly, an **Opinion poll** in which researchers conduct in-depth interviews is considered observational as long as the researchers attempt only to learn people's opinions, not to change them.

Variations on observational Studies

Cross-sectional study : The most familiar observational studies are those in which data are collected all at once. data are observed, measured, and collected at one point in time, not over a period of time.

Two variations on observational studies are also common :

- **Retrospective (or case-control) study**
- **Prospective (or longitudinal or cohort) study**

Retrospective Study

- Uses data from the past, such as official records or past interviews.
- Especially valuable in cases where it may be **impractical** or **unethical** to perform an experiment.

Example : Alcohol & Pregnancy

we want to learn how alcohol consumed during pregnancy affects newborn babies. Because **it is already known that consuming alcohol during pregnancy can be harmful**, it would be **highly unethical to ask pregnant mothers to test the "treatment" of consuming alcohol**. However, because many mothers consumed alcohol in past pregnancies (either before the dangers were known or choosing to ignore the dangers), we can do a retrospective study in which we compare children born to those mothers to children born to mothers who did not consume alcohol.

Prospective Study

- Set up to collect data in the future from groups that share common factors.

Example : Nurses' Health Study

The Nurses' Health Study was started in 1976 with 121,700 female registered nurses who were between the ages of 30 and 55. The subjects were surveyed in 1976 and every two years thereafter. The study is **ongoing.**

Exercise

whether the observational study used is cross-sectional, retrospective, or prospective

Drinking and Driving Study : In order to study the seriousness of drinking and driving, a researcher obtains records from past car crashes. Drivers are partitioned into a group that had no alcohol consumption and another group that did have evidence of alcohol consumption at the time of the crash.

Meat and Mortality : Researchers at the National Cancer Institute studied meat consumption and its relationship to mortality. Approximately one-half million people were surveyed, and they were then followed for a period of 10 years.

Smoking Study : Researchers from the National Institutes of Health want to determine the current rates of smoking among adult males and adult females. They conduct a survey of 500 adults of each gender.

Experiments

Experiments : researchers apply some treatment and observe its effects on the subjects of the experiment.

Goal of Experiments : study **the effects of some treatment**.

Research 1

Apple randomnly selects 500 iPhone batteries. The voltage of each battery is measured.

Research 2

Apple randomnly selects 500 iPhone batteries. The voltage of each battery is measured. And Apple also randomnly selects 500 iPhone batteries. The voltage of each battery is measured after being heated to 43°C. Then Apple compare the average voltage of two groups.

Experiments

Two types of groups

- **Treatment group** : the group of subjects who receive the treatment being tested.
- **Control group** : the group of subjects who do not receive the treatment being tested.

Large daily doses of vitamin C help prevent colds

Consider a medical study designed to test whether large daily doses of vitamin C help prevent colds. The researchers conduct their experiment with two groups of subjects: One group takes large doses of vitamin C daily and another group does not.

Example : The Salk polio Vaccine

The Salk polio Vaccine

If you had been a parent in the 1940s or 1950s, one of your greatest fears would have been the disease known as polio. Each year during this long polio epidemic, thousands of young children were paralyzed by the disease. **In 1954, a large experiment was conducted to test the effectiveness of a new vaccine created by Dr. Jonas Salk (1914–1995). The experiment involved a sample of 400,000 children chosen from the population of all children in the United States. Half of these 400,000 children received an injection of the Salk vaccine. The other half received an injection that contained only salt water. Among the children receiving the Salk vaccine, only 33 contracted polio. In contrast, there were 115 cases of polio among the children who did not get the Salk vaccine.** Using techniques of statistical science, the researchers concluded that the vaccine was effective at preventing polio. They therefore decided to launch a major effort to improve the Salk vaccine and distribute it to the population of all children. Thanks to this vaccine (and improved ones developed later), the horror of polio is now largely a memory of the past.

The variables of interest

The variables of interest : In a statistical study the variables of interest are the categories or quantities that the study seeks to measure.

When cause and effect may be involved

- **Explanatory variable, Independent variable**
  : variable that may explain or cause the effect

- **Response variable, Dependet variable**
  : variable that responds to changes in the explanatory variable

Confounding variables

Using control groups helps to ensure that we account for known variables that could affect a study's results. However, researchers may be unaware of or be unable to account for other important variables.

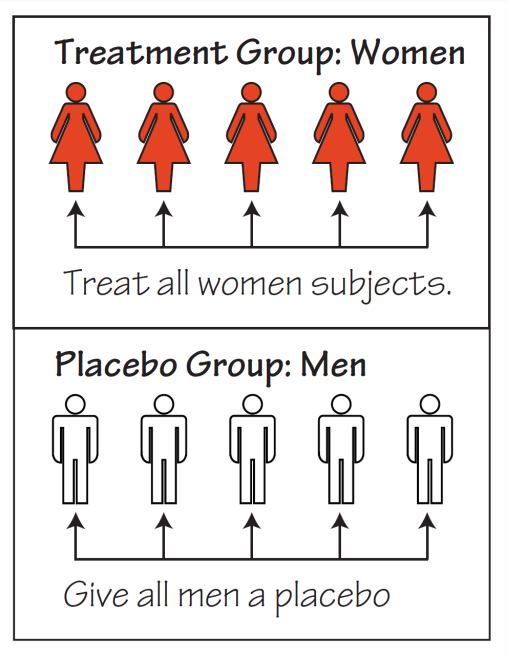
Confounding variable

A study suffers from **confounding** if the effects of different variables are mixed so we cannot determine the specific effects of the variables of interest. The variables that lead to the confusion are called **confounding variables.**

Confounding variables

Example : Randomization

In testing the **effectiveness of a new vaccine**, suppose that researchers used females for the treatment group and males for the control group.



What is confounding ?

We don't know if effects are due to sex or to treatment.

Confounding variables

It's not always easy to discover confounding variables.

Radon and Lung cancer

Radon is a radioactive gas produced by natural processes (the decay of uranium) in the ground. The gas can leach into buildings through the foundation and can accumulate to relatively high concentrations if doors and windows are closed. Imagine a (hypothetical) study that seeks to determine **whether radon gas causes lung cancer by comparing the lung cancer rate in Colorado, where radon gas is fairly common, with the lung cancer rate in Hong Kong, where radon gas is less common.** Suppose the study finds that the lung cancer rates are nearly the same. Would it be reasonable to conclude that radon is not a significant cause of lung cancer?

Radon gas is not the only possible cause of lung cancer.

- For example, smoking can cause lung cancer, so smoking rate may be a confounding variable in this study
- Especially **the smoking rate in Hong Kong is much higher than the smoking rate in Colorado.**
- As a result, we cannot draw any conclusions about radon and lung cancer without taking the smoking rate into account (and perhaps other variables as well).

In fact, careful studies from the U.S. Environmental Protection Agency (EPA) have shown that radon gas can cause lung cancer.