STA 1013 : Statistics through Examples

Lecture 2: Terminology

Hwiyoung Lee

August 30, 2019

Department of Statistics, Florida State University

Overview

1. Review
2. Types of Data and variable
3. Distribution

Review

Review of the last class

Some definitions on Data :

- **Statistics** : Science of collecting, summarizing, analyzing, and interpreting of data
- **Data** : Information
  1. Observed measurement : height, temperature, GPA, $\cdots$
  2. Descriptions : marital status, gender, ethnicities, $\cdots$
- **Observation, Data Point** : A single collected data value
- **Variable** : characteristic or property of an individual or unit whose value may change from one observation to another.

Review of the last class

Population : Entire overall group we are interested in

- time-consuming or costly to obtain data from the population

Parameter : A numerical measurement that describes a characteristic of a population

- Usually unknown
- Example : $\mu, \sigma^2, \pi, \rho, \cdots$

Example

- Average Height of U.S adult female
- Proportion of current smokers in U.S

Review of the last class

Sample : Subset of the population that we collect data on

Statistic : A numerical measurement that describes a characteristic of a sample

- known (can be calculated)
- Example : $\bar{X}, s^2, p, r, \cdots$
- statistics are used as estimates for the corresponding population parameters

Example

- Average Height of randomly selected 3,000 U.S adult female
- Proportion of current smokers in randomnly selected 5,000 U.S people

Statistical Inference : The process of using sample statistics to draw conclusions about population parameters

Review of the last class

Example

Indicate which of the following is a parameter and which is a statistic

- A mean obtained from sampling 2,000 American adults using random digit dialing (phone sample)
- A mean obtained from sampling 10,000 teenagers on Facebook
- A Harris poll surveyed 2,320 adults in the United States, among which 14% said that they have at least one tattoo
- A mean obtained from the U.S. census

Review of the last class

Example of statistical inference from quality control

GE wants to know defective rate among all bulbs produced in 2018. 1,000,000 bulbs were produced in their plants (2018). The company might sample, say, 500 bulbs to estimate the proportion of defectives. 5 out of 500 bulbs tested are defective.

- What is the population and parameter ?
- What is the sample and statistic ?

Types of Data and variable

Primary Data & Secondary Data

Primary Data : Data collected by the investigator himself/ herself for a specific purpose.

- The investigator collects data specific to the problem under study
- If required, it may be possible to obtain additional data during the study period
- Cost of obtaining the data is often the major expense in studies

Primary Data & Secondary Data

Secondary Data : Data collected by someone else for some other purpose, or published elsewhere

- The data is already there → It can be gathered quickly and inexpensively
- Data may be outdated
- The investigator cannot decide what is collected
- Most often obtaining additional data is not possible

Data Source

The FSU Fact Sheet and Fact Book :

`https://ir.fsu.edu/facts.aspx`

Types of variable

Categorical variable :

- Result in categorical responses.
- Also called **Nominal**, or **Qualitative variable**
- Never used directly in calculations
- Examples :

  Gender : Male, Female

  Animal species : Dog, Cat, Fish, Bird, $\cdots$

Types of variable

Quantitative variable :

- Result in numerical responses, can be used directly in calculations
- **Discrete variable**

  Arise from a counting process

  Example : How many courses have you taken at FSU?

- **Continuous variable**

  Arise from a measuring process

  Example : How much do you weigh?

Discrete vs Continuous

One way to determine whether data is continuous, is to ask yourself whether you can add several decimal places to the answer.

- How much do you weigh ?

  I weigh 165.3463 pounds.

- How many courses have you taken at FSU ?

  I have taken 26.543 courses in FSU.

Discrete vs Continuous

Which of the following is an example of continuous variable?

(a) number of children

(b) amount of time it takes to assemble an IKEA bookcase

(c) total number of phone calls made in a week

(d) number of bathrooms in a house

Which of the following is an example of discrete variable?

(a) circumference of American women's wrists

(b) amount of time spent playing computer games

(c) total number of phone calls made in a week

(d) length of elephant tusks

Exercies

Identify the followings as categorical (C) or quantitative (Q)

(a) The amount of air inside the balloons at a party.

(b) The ice-cream flavors favored by FSU students.

(c) The time it takes a student to finish an exam.

(d) The number of textbooks in each room at a dormitory.

(e) The current temperature inside the classrooms on campus.

(f) The maximum legal highway speed in major European cities.

(g) Client satisfaction survey responses (poor, average, good, or excellent)

Types of variable

If numbers are used only as labels for categories, then they are considered **categorical variable**.

Example

What is your Gender ?

1. Female
2. Male

Types of variable

If Quantitative variable is used for the purpose of categorization, then they are considered Categorical (Qualitative) variable in that context.

Example

How much is your income ? _____ $

Example

How much is your income ? (Choose one)

1. Under $20,000
2. $20,000 – $49,999
3. $50,000 and over

Distribution

Distribution of variable

Distribution of a variable tells us what values the variable can possibly take and how often it takes these values

- The distribution of a variable refers to the way its values are spread over all possible values
- Distributions can be set out in Frequency tables, Bar chart, Pie chart.

Frequency Table

Example

What is your gender ? Female ( ) Male ( )

Frequency Table

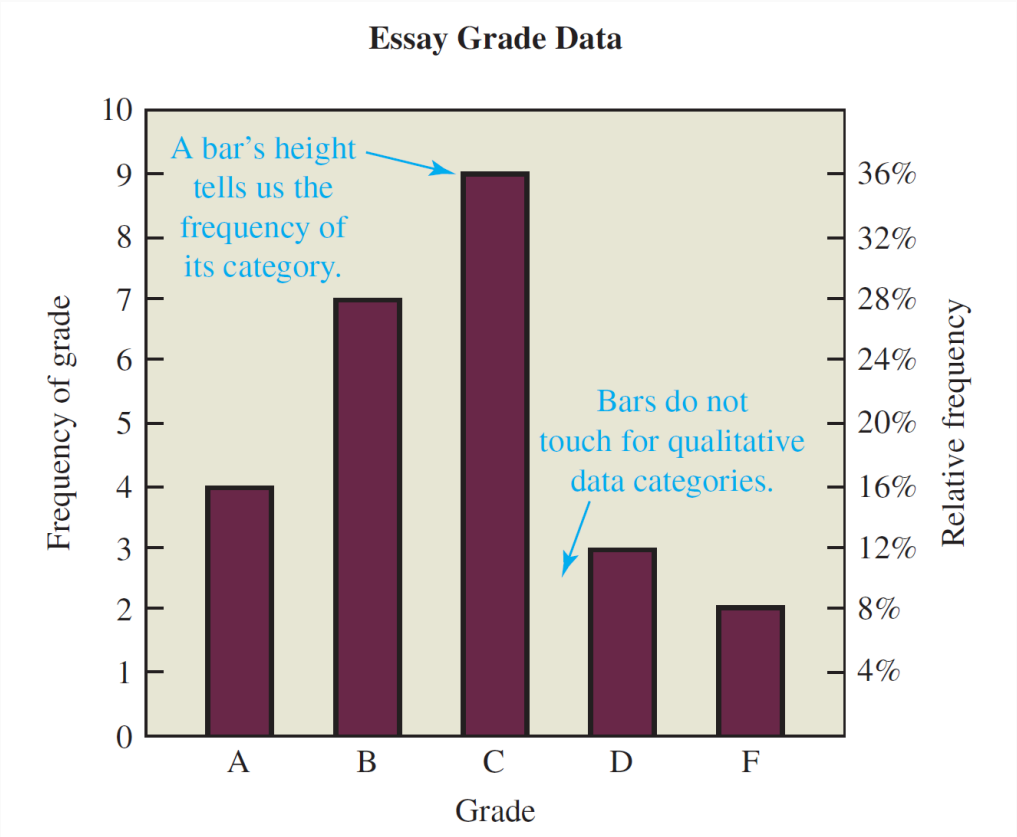
Gender	Count
Female	23,624
Male	18,094
Total	41,718

Table 1: Frequency Table of FSU 2018 gender data

Relative frequency is proportion or percentage calculated as **Count** $\div$ **Total**

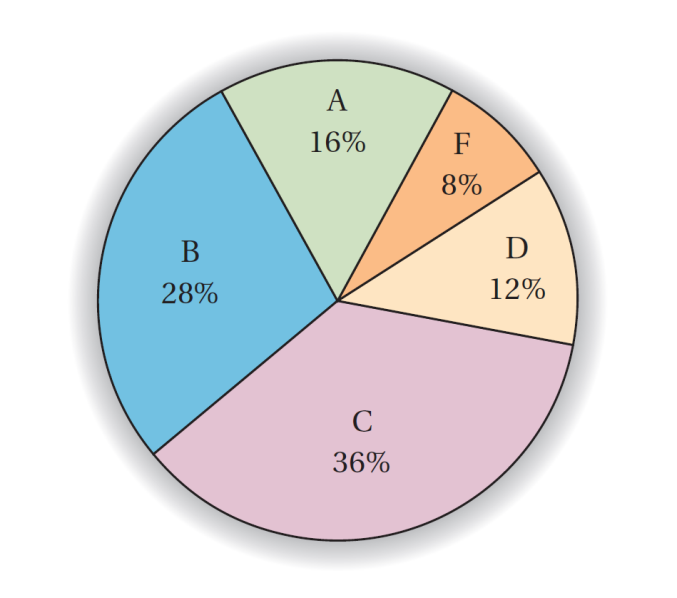
Bar chart

- uses a set of bars to represent the frequency (or relative frequency) of each category
- the higher the frequency, the longer the bar
- The bars can be either vertical or horizontal



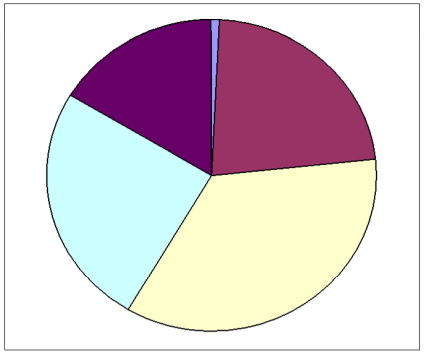
Pie chart

- Pie charts are commonly used to show relative frequency distributions
- The entire pie represents the total relative frequency of 100%
- The sizes of the individual slices, or wedges, represent the relative frequencies of the various categories



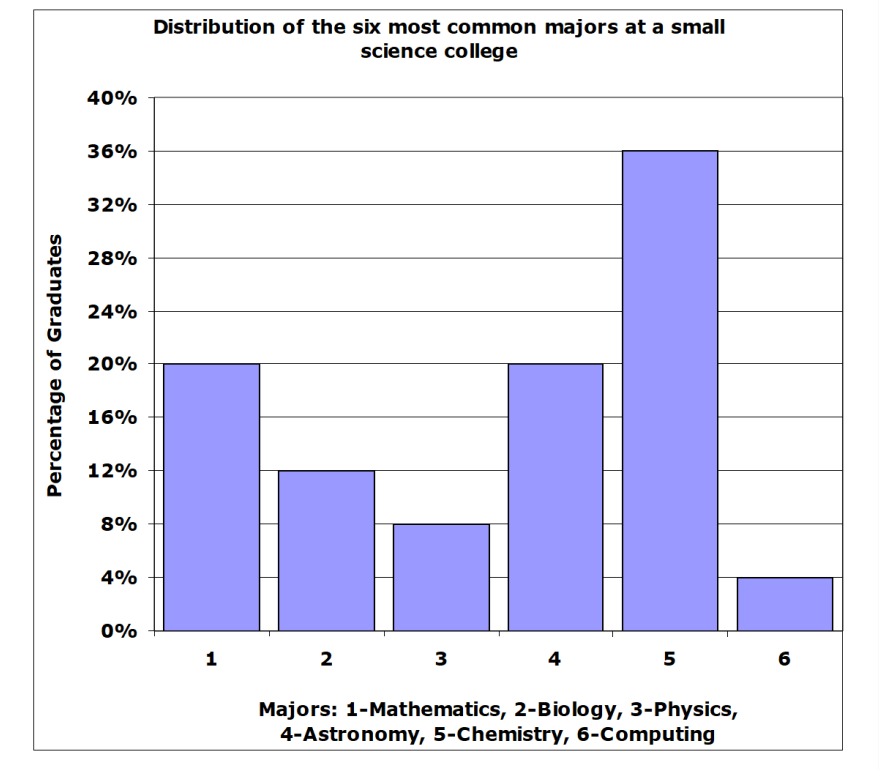
Exercise

Occupation	Percentage
Farming, forestry and fishing	0.7%
Manufacturing, extraction, transportation and crafts	22.7%
Managerial, professional and technical	34.9%
Sales and office	25.4%
Other services	16.3%
Total	100%



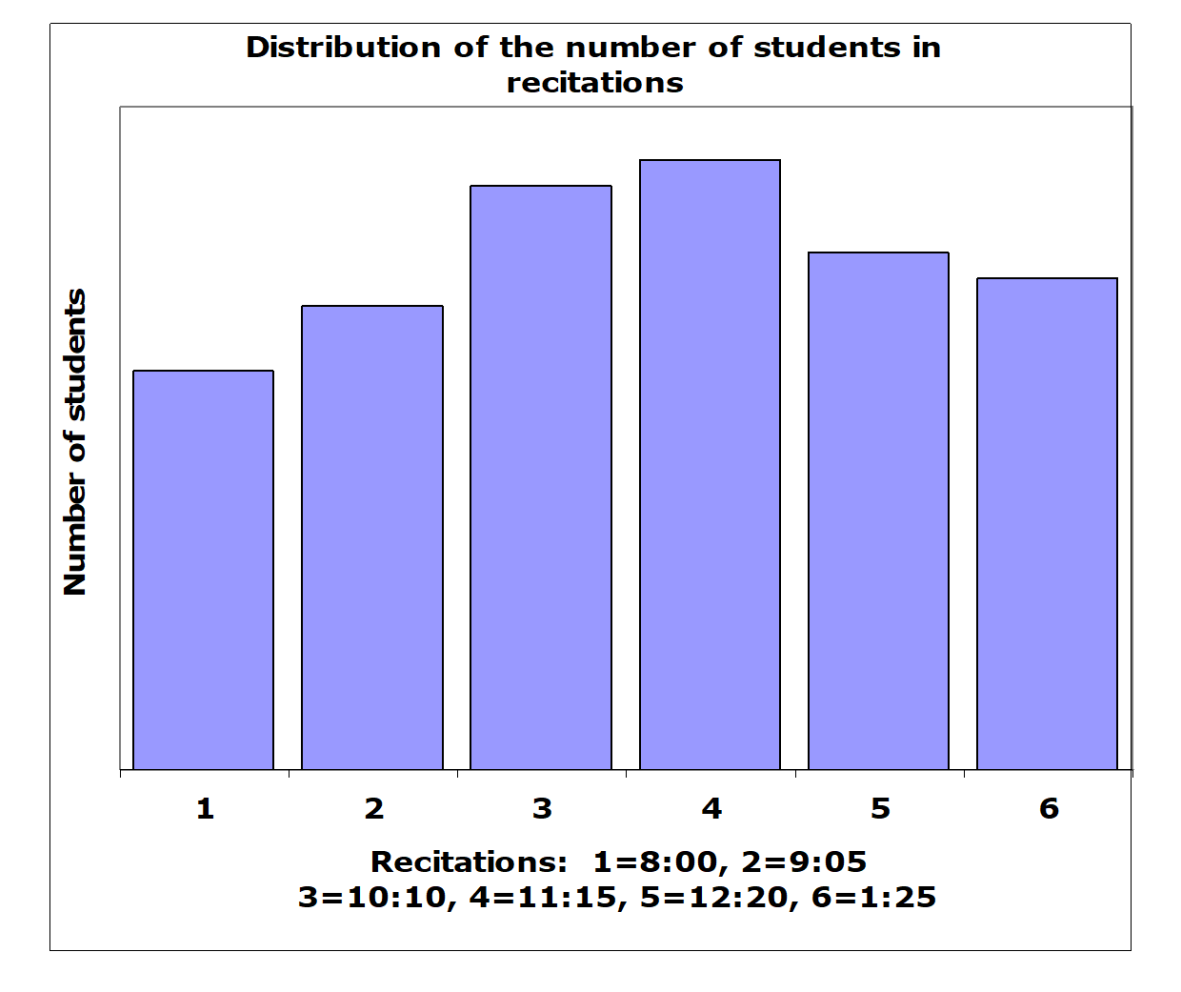
The table and pie chart below show the make-up of the US labor force (2006 estimate excluding unemployed). Label the sectors of the pie with the occupation categories.

Exercise



- If 108 students were Chemistry majors, what was the total number of students?
- How many students were Computing majors?

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)