Univariate Statistical Analysis

Lecture 1

Numerical Descriptive Measures
(Chapter 1, 3 & 4)

Introduction

- Contact:
  Name: Peter Ho
  Email: pho@sheridan.edu.au
- Text Book:
  Peck, R., Short, T., & Olsen, C. (2018*). Introduction to Statistics & Data Analysis*. (6th ed.). Cengage Learning.
- Assessments:

1. In-class test 1: week 5 (15%)
2. In-class test 2: week 12 (15%)
3. Assignment: week 9 (20%)
4. Final exam: week 15 (50%)

* A4 sheet of paper with **handwritten notes** on both sides and a scientific calculator.

Objectives

- Definitions
  - Population
  - Sample
  - Variable and Data
- Displaying Data
  - Bar Charts
  - Histogram
  - Pie Charts

Objectives

Methods for Describing Data

- Centre of a data set
  - Mean
  - Median
- Variability of a data set
  - Range
  - Variance
  - Standard Deviation

True Seeker

- We come to know truth
- Finding out what is true about the world, ourselves, and others constitutes
- Important and interesting

Examples

- How do humans think?
- What happens in the body to produce a sensation or a movement?
- When I get angry, is it true that there is a unique underlying physiological pattern?
- What is the pattern?

What is Statistics?

- Scientific method
  - Quantitative method
- Find out what the true is

Activity

Question 1. What are a sample and a population in statistics?

Question 2. Is it possible to obtain data from an entire population in reality for statistical purposes?

Question 3. What are descriptive and inferential statistics, and how do they differ?

Definitions

Population: A population is the **complete set** of individuals, objects, or scores that the investigator is **interested in studying**.

Sample: A sample is a **subset** of the **population**.

Variable: A variable is any **property or characteristic** of some event, object, or person that may have **different values** at different times **depending** on the **conditions**.

Data: The measurements that are made on the subjects of an experiment are called data.

Descriptive statistics is concerned with **techniques** that are used to **describe** or characterize **the** obtained **data**.

Inferential statistics involves **techniques** that use the obtained **sample** data to **infer** to **populations**.

Activity

Question 4. What are categorical data sets and numerical data sets? Provide an example to explain each.

Question 5. What are discrete variables and continuous variables? Provide an example to explain each.

Definitions

Categorical data set: A data set is **categorical (**or **qualitative)** if the individual observations are categorical responses. (e.g. gender, educational level)

Numerical data set: A data set is **numerical** (or **quantitative**) if each observation is a number. (e.g. number of students , how many chairs, height and weight)

- **Discrete variable** is one in which there are **no possible values between** adjacent units on the **scale**. (e.g. number of students , how many chairs)
- **Continuous variable** is one that theoretically can have an **infinite number** of values between adjacent units **on the scale**. (e.g. height and weight)

Exercise 1.12 (page. 16)

Classify each of the following variables as **categorical, discrete variable** or **continuous variable**.

a. Number of students in a class of 3 who turn in a term paper before the due date.
b. Gender of the next baby born at a particular hospital.
c. Amount of fluid dispensed by a machine used to fill bottles with soda pop.
d. Thickness of the gelatin coating of a vitamin E capsule.
e. Birth order classification.

Exercise 1.12 (page. 16) - solution

Classify each of the following variables as **categorical, discrete variable** or **continuous variable**.

a. Number of students in a class of 35 who turn in a term paper before the due date. → **discrete variable**

b. Gender of the next baby born at a particular hospital. → **categorical**

c. Amount of fluid dispensed by a machine used to fill bottles with soda pop. → **continuous variable**

d. Thickness of the gelatin coating of a vitamin E capsule. → **continuous variable**

e. Birth order **classification**. → **categorical**

Definitions

Frequency: the **number of times** of an **event appears** in a study (or experiment).

Frequency distribution: A graphically way to **summarise** (to display) the **frequency** from the sample **data**.

Displaying Data – Bar Charts

Use for Categorical data.

Example 1.7 (page. 12) - Motorcycle Helmet

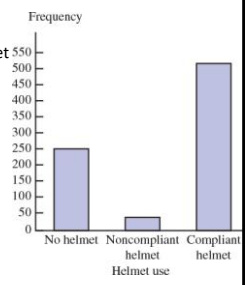
Helmet Use Category	Frequency	Relative Frequency
No helmet	250	0.310 ← 250/806
Noncompliant helmet	40	0.050 ← 40/806
Compliant helmet	516	0.640
	806 ← Total number of observations	1.000 ← Should total 1, but in some cases may be slightly off due to rounding

Displaying Data – Bar Charts

Example 1.7 (page. 12) - Motorcycle Helmet Cont'

x-axis: Category

y-axis: Frequency



Displaying Data – Histogram

Use for Discrete numerical data.

Example 3.12 (page. 97) - Promiscuous Queen Bees

Number of Partners:
12,2,4,6,6,7,8,7,8,11
8,3,5,6,7,10,1,9,7,6
9,7,5,4,7,4,6,7,8,10

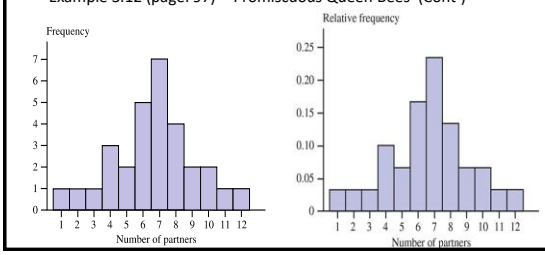
Displaying Data – Histogram

Example 3.12 (page. 97) - Promiscuous Queen Bees (Cont')

Number of Partners	Frequency	Relative Frequency
1	1	0.033 ← 1/30 = 0.033
2	1	0.033
3	1	0.033
4	3	0.100
5	2	0.067
6	5	0.167
7	7	0.233
8	4	0.133
9	2	0.067
10	2	0.067
11	1	0.033
12	1	0.033
Total	**30**	**0.999** ← Differs from 1 due to rounding

Displaying Data – Histogram

Example 3.12 (page. 97) - Promiscuous Queen Bees (Cont')



Displaying Data – Histogram

Use for Continuous numerical data.

Example 3.16 (page. 103) - Sleep deficit

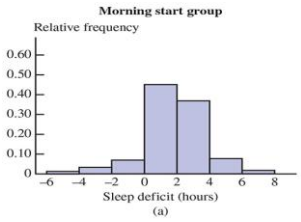
Sleep Deficit (in hours)	Morning Start Relative Frequency	Afternoon Start Relative Frequency
-6 to < -4	0.007	0.020
-4 to < -2	0.028	0.050
-2 to < 0	0.065	0.190
0 to < 2	0.442	0.570
2 to < 4	0.364	0.120
4 to < 6	0.078	0.040
6 to < 8	0.015	0.010

Displaying Data – Histogram

Example 3.16 (page. 103) - Sleep deficit (Cont')

Steps:

- Setup the interval
- Calculate the relative frequency
- x – axis: Intervals
- y – axis: relative frequency



Displaying Data – Pie Charts

Use for Categorical data.

Example 3.3 (page. 81) - Watch Those Typos

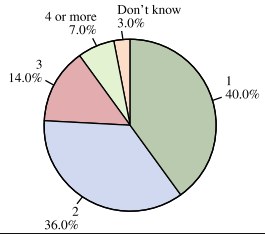
Number of Typos	Frequency
1	60
2	54
3	21
4 or more	10
Don't know	5

Displaying Data – Pie Charts

Example 3.3 (page. 81) - Watch Those Typos (Cont')

Steps:

- Calculate the total frequency
- Calculate relative frequency in percentage
- $= \frac{Frequency}{Total\ Frequency} \times 100\%$
- Calculate slice size = $\frac{Frequency}{Total\ Frequency} \times 360^0$
- Draw a slice appropriate size and labels



Activity

Question 6. What is the meaning of the mean and standard deviation in statistics?

Question 7. What is the difference between the mean and the median?

Methods for Describing Data Chapter 4.1-4.3

- Centre of a data set
  - Mean
  - Median
- Variability of a data set
  - Range
  - Variance
  - Standard Deviation

Centre of a data set

- Mean (Average):

DEFINITION

Sample mean: The **sample mean** of a sample consisting of numerical observations $x_1, x_2, \ldots, x_n$ is denoted by $\bar{x}$, and its formula is given by

$$\bar{x} = \frac{\text{sum of all observations in the sample}}{\text{number of observations in the sample}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\Sigma x}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Centre of a data set

- Mean (Average):

DEFINITION

Population mean: The **population mean**, denoted by μ, is the average of all x values in the entire population.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Centre of a data set

Mean (Average)
Example 4.2 (page. 151) - Country Population Sizes

Total countries = N = 3,137
Total residents (2018 Census Bureau estimate) = $\sum x_i = 231{,}665{,}106$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{231{,}665{,}106}{3{,}137} = 73{,}849.3 \text{ residents per country}$$

Centre of a data set

Mean (Average)
Example 4.2 (page. 152) - Country Population Sizes (cont')

Sample 1		Sample 2		Sample 3	
County	x Value	County	x Value	County	x Value
Fayette, TX	20,964	Stoddard, MO	28,509	Chattahoochee, GA	21,332
Monroe, IN	101,719	Johnston, OK	10,442	Petroleum, MT	638
Greene, NC	15,584	Sumter, AL	17,002	Armstrong, PA	77,299
Shoshone, ID	19,021	Milwaukee, WI	960,664	Schoolcraft, MI	8,625
Jasper, IN	26,570	Albany, WY	30,607	Benton, MO	12,133
	Σx = 183,858		Σx = 1,047,224		Σx = 120,027
	$\bar{x}$ = 36,771.6		$\bar{x}$ = 209,444.8		$\bar{x}$ = 24,005.4

Centre of a data set

The Median (Middle value from the set):

DEFINITION

Sample median: The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included, so that every sample observation appears in the ordered list). Then

$$\text{sample median} = \begin{cases} \text{the single middle value if } n \text{ is odd} \\ \text{the average of the middle two values if } n \text{ is even} \end{cases}$$

Steps: (Must be sorted in ascending order)
1. $\frac{n+1}{2}$ Ranked observation
2. Calculate the middle value

Centre of a data set

The Median (Middle value from the set):
Example 4.4 (page. 153) - Web Site Data Revised

0,0,0,0,0,0,3,4,4,4,5,5,7,7,8,8,8,12,12,*13*
13,13,14,14,16,18,19,19,20,20,21,22,23,26,36,36,37,42,84,**331**

Steps:
1. $\frac{n+1}{2}$ ranked observation = $\frac{40+1}{2} = \frac{41}{2} = 20.5$

→ The middle value is the average of $x_{20} = 13$ and $x_{21} = 13$

2. The median = $\frac{13+13}{2} = 13$

Variability in a data set

- Range:

DEFINITION

Range: The **range** of a data set is defined as Range = Largest Value − Smallest Value

Variability in a data set

- Sample variance and sample deviation: (Deviation from the sample mean)

DEFINITIONS

Sample variance: The **sample variance**, denoted by s^2, is the sum of squared deviations from the mean divided by $n - 1$. That is,

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Sample standard deviation: The **sample standard deviation** is the positive square root of the sample variance and is denoted by s.

Variability in a data set

Sample Variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Sample Standard Deviation: $s = \sqrt{s^2}$

Variability in a data set

- Population variance and Population deviation: (Deviation from the population mean)

Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

Population Standard Deviation: $\sigma = \sqrt{\sigma^2}$

Variability in a data set

Example 4.8 (page. 161) - Big Mac Revisited
Data from Example 4.7 (page. 160)

Determine the sample variance and sample standard deviation.

Answer on page. 162

Variability in a data set

Exercise 4.24 (page. 166) – Smart Phone Prices

Determine the population variance and population standard deviation.

Summary

- Definitions
  - Population
  - Sample
  - Variable and Data
- Displaying Data
  - Bar Charts
  - Histogram
  - Pie Charts

Summary

Methods for Describing Data

- Centre of a data set
  - Mean
  - Median
- Variability of a data set
  - Range
  - Variance
  - Standard Deviation