Lecture 3: Measure of Central Tendency

Donglei Du
(ddu@unb.edu)

Faculty of Business Administration, University of New Brunswick, NB Canada Fredericton
E3B 9Y2

Table of contents

Section 1

Measure of central tendency: location parameter

Subsection 1

Introduction

Introduction

- Characterize the average or typical behavior of the data.
- There are many types of central tendency measures:
  - Arithmetic mean
  - Weighted arithmetic mean
  - Geometric mean
  - Median
  - Mode

Subsection 2

Arithmetic Mean

Arithmetic Mean

- The Arithmetic Mean of a set of n numbers

$$AM = \frac{x_1 + \ldots + x_n}{n}$$

- Arithmetic Mean for population and sample

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

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

Example

- **Example:** A sample of five executives received the following bonuses last year ($000): 14.0 15.0 17.0 16.0 15.0
- **Problem:** Determine the average bonus given last year.
- **Solution:**

$$\bar{x} = \frac{14 + 15 + 17 + 16 + 15}{5} = \frac{77}{5} = 15.4.$$

Example

- Example: the weight example (weight.csv)
- The R code:

```
weight <- read.csv("weight.csv")
sec_01A<-weight$Weight.01A.2013Fall
# Mean
mean(sec_01A)
```

```
## [1] 155.8548
```

Will Rogers phenomenon

- Consider two sets of IQ scores of famous people.

Group 1	IQ		Group 2	IQ
Albert Einstein	160		John F. Kennedy	117
Bill Gates	160		George Washington	118
Sir Isaac Newton	190		Abraham Lincoln	128
Mean	170		Mean	123

- Let us move Bill Gates from the first group to the second group

Group 1	IQ		Group 2	IQ
Albert Einstein	160		John F. Kennedy	117
			Bill Gates	160
Sir Isaac Newton	190		George Washington	118
			Abraham Lincoln	128
Mean	175		Mean	130.75

Will Rogers phenomenon

- The above example shows the Will Rogers phenomenon: "When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states."

Properties of Arithmetic Mean

- It requires at least the interval scale
- All values are used
- It is unique
- It is easy to calculate and allow easy mathematical treatment
- The sum of the deviations from the mean is 0
- The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero!
- It is easily affected by extremes, such as very big or small numbers in the set (non-robust).

The sum of the deviations from the mean is 0: an illustration

	Values	deviations
	3	-2
	4	-1
	8	3
Mean	5	0

How Extremes Affect the Arithmetic Mean?

- The mean of the values 1,1,1,1,100 is 20.8.
- However, 20.8 does not represent the typical behavior of this data set!
- Extreme numbers relative to the rest of the data is called **outliers**!
- Examination of data for possible outliers serves many useful purposes, including
  - Identifying strong skew in the distribution.
  - Identifying data collection or entry errors.
  - Providing insight into interesting properties of the data.

Subsection 3

Weighted Mean (WM)

Weighted Mean (WM)

- The Weighted Mean (WM) of a set of n numbers

$$WM = \frac{w_1 x_1 + \ldots + w_n x_n}{w_1 + \ldots + w_n}$$

- This formula will be used to calculate the mean and variance for grouped data!

Example

- **Example**: During an one hour period on a hot Saturday afternoon Cabana boy Chris served fifty drinks. He sold:

five drinks for	\$0.50
fifteen for	\$0.75
fifteen for	\$0.90
fifteen for	\$1.10

- **Problem**: compute the weighted mean of the price of the drinks

$$WM = \frac{5(0.50) + 15(0.75) + 15(0.90) + 15(1.10)}{5 + 15 + 15 + 15} = \frac{43.75}{50} = 0.875.$$

Example

- Example: the above example
- The R code:

```
## weighted mean
        wt <- c(5,  15,  15,  15)/50
        x <- c(0.5,0.75,0.90,1.1)
        weighted.mean(x, wt)
```

```
## [1] 0.875
```

Subsection 4

Median

Median

- The Median is the midpoint of the values after they have been ordered from the smallest to the largest
- Equivalently, the Median is a number which divides the data set into two equal parts, each item in one part is no more than this number, and each item in another part is no less than this number.

Two-step process to find the median

Step 1. Sort the data in a nondecreasing order

Step 2.
- If the total number of items n is an odd number, then the number on the $(n+1)/2$ position is the median;
- If n is an even number, then the average of the two numbers on the $n/2$ and $n/2+1$ positions is the median. (For ordinal level of data, choose any one on the two middle positions).

Examples

- **Example**: The ages for a sample of five college students are: 21, 25, 19, 20, 22
- Arranging the data in ascending order gives: 19, 20, 21, 22, 25.
- The median is 21.
- **Example**: The heights of four basketball players, in inches, are: 76, 73, 80, 75
- Arranging the data in ascending order gives: 73, 75, 76, 80. The median is the average of the two middle numbers

$$Median = \frac{75 + 76}{2} = 75.5.$$

One more example

- **Example**: Earthquake intensities are measured using a device called a seismograph which is designed to be most sensitive for earthquakes with intensities between 4.0 and 9.0 on the open-ended Richter scale. Measurements of nine earthquakes gave the following readings:

  $$4.5,\ \mathrm{L},\ 5.5,\ \mathrm{H},\ 8.7,\ 8.9,\ 6.0,\ \mathrm{H},\ 5.2$$

  where L indicates that the earthquake had an intensity below 4.0 and a H indicates that the earthquake had an intensity above 9.0.
- **Problem:** What is the median earthquake intensity of the sample?
- **Solution:**
  - Step 1. Sort: L, 4.5, 5.2, 5.5, 6.0, 8.7, 8.9, H, H
  - Step 2. So the median is 6.0

Example

- Example: the weight example (weight.csv)
- The R code:

```
weight <- read.csv("weight.csv")
sec_01A<-weight$Weight.01A.2013Fall
# Median
median(sec_01A)
```

```
## [1] 155
```

Properties of Median

- It requires at least the ordinal scale
- All values are used
- It is unique
- It is easy to calculate but does not allow easy mathematical treatment
- It is not affected by extremely large or small numbers (robust)

Subsection 5

Mode

Mode

- The number that has the highest frequency.

Example

- **Example**: The exam scores for ten students are: 81, 93, 84, 75, 68, 87, 81, 75, 81, 87
- The score of 81 occurs the most often. It is the Mode!

Example

- Example: the weight example (weight.csv)
- The R code:

```
weight <- read.csv("weight.csv")
sec_01A<-weight$Weight.01A.2013Fall
# Mode
names(table(sec_01A)[which.max(table(sec_01A))])
```

```
## [1] "155"
```

Properties of Mode

- Even nominal data have mode(s)
- All values are used
- It is not unique
  - Modeless: if all data have different values, such as 1,1,1
  - Multimodal: if more than one value have the same frequency, such as 1,1,2,2,3.
- It is easy to calculate but does not allow easy mathematical treatment
- It is not affected by extremely large or small numbers (robust)

Subsection 6

Geometric Mean

Geometric Mean (GM)

- Given a of a set of n numbers $x_1, \ldots, x_n$, the geometric mean is given by the following formula:

$$GM = \sqrt[n]{x_1 \cdots x_n}$$

- If we know the initial and final value over a certain period of n (instead of the individual number sin each period), then

$$GM = \sqrt[n]{\frac{\text{final value}}{\text{initial value}}}$$

Example

- **Example**: The interest rate on three bonds was 5%, 21%, and 4% percent. Suppose you invested \$10000 at the beginning on the first bond, then switch to the second bond in the following year, and switch again to the third bond the next year.
- **Problem:** What is your final wealth after three years?
- **Solution:**
  - Your final wealth will be

$$10000 \times GM^3 = 10,000 \times 1.097^3 = 13213.2,$$

  - where

$$GM = \sqrt[3]{1.05 \times 1.21 \times 1.04} \approx 1.097$$

Example

- Example: the above example
- The R code:

```
#geometric mean: R does not have a built-in function for t

 #You can install and use another package
       library(psych)
## Warning:  package 'psych' was built under R version
3.2.5
       rates<-c(1.05, 1.21,1.04)
       geometric.mean(rates)
## [1] 1.097327
```

Example

- **Example**: The total number of females enrolled in American colleges increased from 755,000 in 1992 to 835,000 in 2000.
- **Problem:** What is your the geometric mean rate of increase?.
- **Solution:**
  - The geometric mean over these 8 years is

$$GM = \sqrt[8]{\frac{835,000}{755,000}} \approx 1.0127.$$

Therefore the geometric mean rate of increase is 1.27%.

Arithmetic Mean vs Geometric Mean: the AM-GM inequality:

- If $x_1, \ldots, x_n \geq 0$, then

$$AM = \frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \ldots x_n} = GM,$$

with equality if and only if $x_1 = x_2 = \cdots = x_n$.

Arithmetic Mean vs Geometric Mean: the AM-GM inequality:

- If $x_1, \ldots, x_n \geq 0$, then

$$AM = \frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \ldots x_n} = GM,$$

with equality if and only if $x_1 = x_2 = \cdots = x_n$.

Case: GM vs AM in fund reporting

- A fund manager tries to convince you to invest in their fund by showing you the annual returns over the last five years

$$10\%, -20\%, 30\%, 12\%, 10\%$$

and the average return per year realized in the last five years is 8.4% as calculated as follows.

$$\begin{aligned} AM &= \frac{(1+0.1)+(1-0.2)+(1+0.3)+(1+0.12)+(1+0.1)}{5} \\ &= 1.084 \end{aligned}$$

- This is misleading sometimes. It is much better to say that the average return realized over the last fives with us is approximately 7% per year:

$$GM = \sqrt[5]{1.10 \times 0.80 \times 1.30 \times 1.12 \times 1.10} - 1 \approx 0.07104408$$

Example

- Example: the above example
- The R code:

```
#geometric mean: R does not have a built-in function for t

 #You can install and use another package
        library(psych)
        rates<-c(1.10, 0.80, 1.30, 1.12, 1.10)
        mean(rates)-1
## [1] 0.084
        geometric.mean(rates)-1
## [1] 0.07104408
```

Properties of Geometric Mean

- Similar to arithmetic mean, except used in different scenario
- It requires interval level
- All values are used
- It is unique
- It is easy to calculate and allow easy mathematical treatments

Subsection 7

Mean for grouped data

Mean for grouped data

- The mean of a sample of data organized in a frequency distribution is computed by the following formula:

$$\bar{x} = \frac{f_1 x_1 + \ldots + f_k x_k}{f_1 + \ldots + f_k},$$

- where f_i is the frequency of Class i and x_i is the class mid-point of Class i.

Example

- **Example**: Recall the weight example from Chapter 2:

class	freq. (f_i)	mid point (x_i)	$f_i x_i$
$[130, 140)$	3	135	405
$[140, 150)$	12	145	1740
$[150, 160)$	23	155	3565
$[160, 170)$	14	165	2310
$[170, 180)$	6	175	1050
$[180, 190]$	4	185	740
	62		9810

- The mean for the grouped data is:

$$\bar{x} = \frac{9810}{62} \approx 158.2258.$$

- The real mean for the raw data is 155.8548.

Subsection 8

The Median for Grouped Data

Median for grouped data: Two-step procedure

Step 1: identify the median class, which is the class that contains the number on the n/2 position.

Step 2: Estimate the median value within the median class using the following formula:

$$\text{median} = L + C \times \frac{\frac{n}{2} - CF}{f},$$

where

- L is the lower limit of the median class
- CF is the cumulative frequency before the median class
- f is the frequency of the median class
- C is the class interval or size

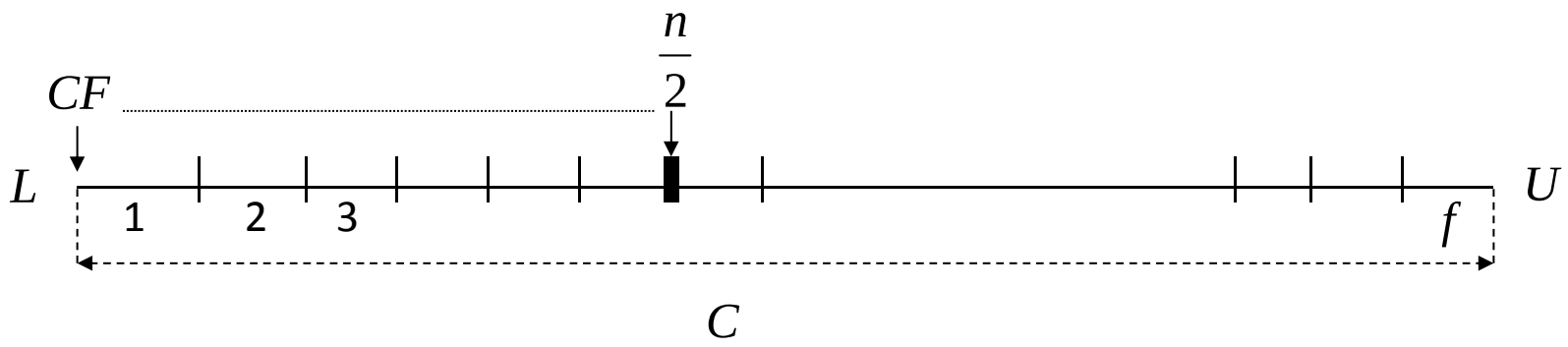
Example

- **Example**: Recall the weight example from Chapter 2:

class	freq	relative freq.	cumulative freq.
$[130, 140)$	3	0.05	3
$[140, 150)$	12	0.19	15
$\overbrace{[150, 160)}^{10}$	23	0.37	38
$[160, 170)$	14	0.23	52
$[170, 180)$	6	0.10	58
$[180, 190]$	4	0.06	62

- $$\text{median} = 150 + 10 \times \frac{\frac{62}{2} - 15}{23} \approx 156.9565.$$

An explanation of the median formula



Subsection 9

The Mode for Grouped Data

Mode for grouped data: Two-step procedure

Step 1: Identify the modal class, which is the class(es) that has the highest frequency(ies).

Step 2: Estimate the modal(s) within the modal class (es) as the class midpoint(s).

Example

- **Example**: Recall the weight example from Chapter 2:

class	freq. (f_i)	mid point (x_i)
$[130, 140)$	3	135
$[140, 150)$	12	145
$[150, 160)$	23	155
$[160, 170)$	14	165
$[170, 180)$	6	175
$[180, 190]$	4	185

- $$\text{mode} = 155.$$

Section 2

Dicussion: How to lie with averges? Or how to defend yourselves from those lying with averages?

Lie with averages

- There are many different interpreations of averages:
  - Arithemtic Mean vs Geometric mean: be careful of investment fund statements
  - Mean vs Median: be careful of the accounting statements
- [Huff, 2010]

References I


Huff, D. (2010).
How to lie with statistics.
WW Norton & Company.