STAT-1301 ; Lecture 3 ; Jan. 16, '24

§ 3.1 (Cont'd)

Weighted Mean:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{j=1}^{n} w_j}$$

where $x_1, \ldots, x_n$ is a random sample and $w_1, \ldots, w_n$ are the weights.

<u>Aside:</u>

$$\bar{x}_w = \frac{w_1}{\sum_{j=1}^{n} w_j} x_1 + \frac{w_2}{\sum_{j=1}^{n} w_j} x_2 + \cdots + \frac{w_n}{\sum_{j=1}^{n} w_j} x_n .$$

Ex. A student earns an A^- on a 3-credit hour course, a B on a 6-credit hour course and a C on a 3-credit hour course. An A^- is worth 4 points, a B is worth 3 points and a C is worth 2 points. What is the student's

GPA ?

Data:

Grade	Credit hrs (w_i)	Points (x_i)
A$^-$	3	4
B	6	3
C	3	2

$$GPA = \bar{x}_w = \frac{\sum_{i=1}^{3} w_i x_i}{\sum_{j=1}^{3} w_j} = \frac{3(4)+6(3)+3(2)}{3+6+3}$$

$$= \cdots = 3.0$$

Ex. Mary bought gas for her car four times during June 2019. She bought 10 gallons at a price of \$2.60 a gallon, 13 gallons at a price of \$2.80 a gallon, 8 gallons at a price of \$2.70 a gallon, and 15 gallons at a price of \$2.75 a a gallon. What is the average price that Mary paid for gas during June 2019 ?

Sol'ns:

X = Price per gallon.

W = # of gallons bought each time.

W	x
10	2.60
13	2.80
8	2.70
15	2.75

$$\bar{x}_W = \frac{\sum_{i=1}^{4} w_i x_i}{\sum_{j=1}^{4} w_j}$$

$$\sum_{j=1}^{4} w_j = 10 + 13 + 8 + 15 = 46$$

$$\sum_{i=1}^{4} w_i x_i = 10(2.60) + 13(2.80) + 8(2.70) + 15(2.75) = 125.25$$

$$\bar{x}_W = \frac{125.25}{46} = \$2.72$$

She paid an average of \$2.72 a gallon for gas purchased in June 2019.

Omit trimmed mean.

§ 3.2 Measures of Dispersion for Ungrouped Data

Data from text.

X = age of employees.

Company 1: 35, 36, 38, 39, 40, 45, 47

Company 2: 18, 27, 33, 52, 70

$$\bar{x}_1 = 40 \quad ; \quad \bar{x}_2 = 40$$

Let's examine the dotplot of the data; See § 3.2 in ebook.

Message: While the measures of centre are identical, the ages in Company 2 have a larger spread. It is not enough to examine measures of central tendency for a data set.

Measures of Spread/Dispersion:

1. Range
2. Standard deviation
3. Interquartile Range (IQR)

Range: = Largest Obs'n − Smallest Obs'n

Ex. $-9 \quad -7 \quad 0 \quad 2 \quad 5 \quad 7 \quad 10 \quad 16$

Range $= 16 - (-9) = 25$

Disadvantages of the Range Statistic:

1) Range Statistic is Sensitive to outliers (i.e. extreme obs'ns).

2) Only uses two obs'ns in the data set.

Standard Deviation:

Notation: S is the Sample Standard deviation

σ is the population standard deviation.

Suppose $x_1, \ldots, x_n$ is a random Sample from a population. The Sample variance of $x_1, \ldots, x_n$ is defined as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

The Sample Standard deviation (S):

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

In practice, we use the Shortcut formula (on formula sheet) to compute S^2:

$$S^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} \right]$$

Ex. Forbes' Magazine's List of Wealthiest people in the world and their wealth in 2007:

Person	Wealth (in billions \$)
Bill Gates	46.5
Helen Walton	18.0
Michael Dell	16.0
Rupport Murdoch	7.8
George Soros	7.2

Find the standard deviation of the wealth of these individuals.

$$\sum_{i=1}^{5} x_i^2 = 46.5^2 + 18^2 + 16^2 + 7.8^2 + 7.2^2$$

$$= 2854.93$$

$$\sum_{i=1}^{5} x_i = 46.5 + 18 + 16 + 7.8 + 7.2 = 95.5$$

$n = 5$

$$S^2 = \frac{1}{5-1}\left[2854.93 - \frac{95.5^2}{5}\right]$$

$= 257.72$ Sample variance

The Sample standard deviation, s, is

$$s = \sqrt{257.72} = 16.054 \text{ billions of } \$.$$

Disadvantage to using s:

It is sensitive to extreme obs'ns.

Ex. Remove Bill Gates' wealth and recompute s.

Now, $\sum_{i=1}^{4} x_i = 49$, $\sum_{i=1}^{4} x_i^2 = 692.68$

$$s^2 = \frac{1}{4-1}\left[692.68 - \frac{49^2}{4}\right] = 30.81$$

$s = \sqrt{30.81} = 5.551$ ← much smaller than this s

Remarks:

i) $s^2 \geq 0$ and $s \geq 0$.

ii) σ = Pop. Std. deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

(e.g. See STAT 2301 on Survey Sampling)

iii) S^2 have variable units squared

iv) S has the same units as that of the variable.

§3.2.3 Coefficient of Variation (CV)

- Can't use, S, to compare the spread of two or more datasets if the variables are not measured on the same scale.

Population data: $CV = \frac{\sigma}{\mu} \times 100\%$

Sample data: $CV = \frac{S}{\bar{x}} \times 100\%$

Ex. Descriptive statistics on heights and weights of a random sample of 40 males are as follows.

Variable	$\bar{x}$	S	CV
Height	68.34 inches	3.02 inches	4.42%
Weight	172.55 lbs.	26.33 lbs.	15.26%

Q'n: Is the relative variation in heights greater or less than that in weights?
(i.e. Which variable has the larger spread?)

Height: $CV = \frac{s}{\bar{x}} \times 100\% = \frac{3.02 \cancel{\text{inches}}}{68.34 \cancel{\text{inches}}} \times 100\% = 4.42\%$

Weight: $CV = \frac{s}{\bar{x}} \times 100\% = \frac{26.33 \cancel{\text{lbs.}}}{172.55 \cancel{\text{lbs.}}} \times 100\% = 15.26\%$

Since $15.26\% > 4.42\%$, the weight variable has larger spread.