Lecture Notes #3 Chapter 3: **Statistics for Describing, Exploring, and comparing Data**

3-2 Measures of Center

A measure of center is a value at the center or middle of a data set.

Mean: the (**arithmetic**) mean of a set of values is the number obtained by adding the values and dividing the total by the number of values.

Notation

$\sum$:The uppercase Greek letter sigma; indicates a summation of values

X: A variable used to represent the individual data vales

n: Number of values in a sample (sample size)

N: Number of values in a population.

μ: The lowercase Greek letter mu; the population mean

$\bar{x}$: Read as "x bar"; the sample mean

Round –off rule (for the measure of center): carry one more decimal place than is present in the original set of values. When applying this rule, round only the final answer, not intermediate values that occur during calculations. Example 1: What is the mean price of the air conditioners? 500, 840, 470, 480, 420, 440, 440.

- Mean always exists.
- It takes every value in a calculation.
- It is affected by extreme values (very sensitive).
- Works well with many statistical methods.

To clear the sensitivity of the mean to extreme values, we define another measure of center called Median.

Median: the median of a data set is the middle value when the data values are arranged in ascending or descending order. If the data set has an even number of entries, the median is the mean of the two middle data entries. The Median is often denoted by ("x-tilde").

Example 2: Find the median for a) 4, 6, 1, 3, 2 b) air conditioner prices given in example 1.

Median is commonly used, always exists, and not sensitive to extreme values.

Mode: The mode of a data set is the value that occurs most frequently. When two values occur with the same greatest frequency, each one is a mode and the data set is bimodal. When more than two values occur with same greatest frequency, each is a mode and data set is said to be multimodal. When no value is repeated, we say there is no mode.

Example 3: find the modes of the following data sets.

A) 5, 5, 5, 3, 1, 5, 1, 4, 3, 5

B) 1, 2, 2, 2, 3, 4, 5, 6, 6, 6, 7, 9

c) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Midrange: the midrange is the measure of center that is the value midway between the maximum and minimum values in the original data set. It is found by adding the maximum data value to the minimum data value and then dividing the sum by 2.

Example 4: find the midrange for 5.40, 1.10, 0.42, 0.73, 0.48, 1.10

Mean from a frequency distribution

The mean from a frequency distribution for a sample is approximated by

Where *x* and *f* are the midpoints and frequencies of a class, respectively.

$$\bar{x} = \frac{\sum(x.f)}{n}$$

$$x = \frac{lower\ \ limit + upper\ \ limit}{2}$$

Guidelines: Finding the mean from a frequency distribution

1. Find the midpoint of each class.
2. Find the sum of the products of the mid points and the frequencies.
3. Find the sum of the frequencies.
4. Find the mean from the frequency distribution.

Example5: Approximate the mean form the frequency distribution. The heights

(in inches) of 16 female students in a physical education class.

Height	f
60-62	3
63-65	4
66-68	7
69-71	2

Weighted Mean: When the values of data set are varying in their degree of importance, we may want to weight them accordingly. Weighted mean:

$$\bar{x} = \frac{\sum(w.x)}{\sum w}$$

Example 6: Find the mean of 3 tests with scores of 85, 90, 75 where the first test counts for 20%, second test counts for 30%, and the third test counts for the 50%.

3-3 Measures of variation

Measures of variation measures the amount that values vary or different among themselves. You can find out how the data are relatively close or far apart spread out. For instance a low measure of variation will verify that values are relatively close together. There are different ways of measuring variation: Range, and standard deviation

Range: the range of a data set is the difference between the maximum and minimum values in the set.

Example7: Find the range of the data set: 11, 10, 8, 4, 6, 7, 11, 6, 11, 7

Finding the range is easy to compute. Range depends only on the highest and lowest values. It is not as useful as other measures of variance.

Standard deviation of a sample:

Def: the deviation is the difference between the value and the mean. Deviation of *x* = *x* - $\bar{x}$

Def: The standard deviation of a set of sample values is a measure of variation of values about the mean. (How far, on average, each observation is from the mean.)

Sample standard deviation: $s = \sqrt{s^2}$=$\sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$

Sample variation: $s^2 = \frac{\Sigma(x-\overline{x})^2}{n-1}$

Guidelines: Finding the sample standard deviation

1. Find the mean of the sample data set
2. Find the deviation of each entry.
3. Square each deviation.
4. Add to get the sum of squares.
5. Divide by n-1 to get the sample variance.
6. Find the square root of the variance to get the sample standard deviation.

Example 8: Find the sample standard deviation of the data set given in example 7.

Standard deviation of a population

Population standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

Population variation: σ^2= $\frac{\Sigma(x-\mu)^2}{N}$

Example 9: Find the range, mean , variance, and standard deviation of the population data set: 15, 8, 12, 5, 19, 14, 8, 6, 13

Finding standard deviation from a frequency distribution:

Sample standard deviation: $s = \sqrt{\frac{n[\sum(f\cdot x^2)]-[\sum(f\cdot x)]^2}{n(n-1)}}$

Example 10: find the standard deviation from a frequency distribution given in example 5.

The value of standard deviation is positive. It is zero, of all the data values are the same number. Larger values of standard deviation indicates greater amount of variation. If some of data values are very far away from all of the others (outliers), then the standard deviation can increase dramatically. Standard deviation's units are the same as the unit of the original data value.

Interpreting and understanding standard deviation:

1st rule: **Range rule of thumb [rough estimate of standard deviation]**

Principal: for many data sets, 95% of sample values lie within 2 standard deviation of the mean.

To roughly estimate a value of the standard deviation, use s =$\frac{range}{4}$ where range= max. value-min. value.

If we know the standard deviation, then the interpretation as follows:

Min. “usual” value: mean – 2 standard deviation

Max. “usual’ value: mean + 2 standard deviation.

Example 11: Find the max. and min. usual value for example 10.

2^{nd} **rule**: **Empirical rule for data with a Bell-shaped distribution.**

If the data sets have a normal distribution (bell-shaped distribution) ,then

About 68% of all values fall within 1 standard deviation of the mean.

About 95% of all values fall within 2 standard deviation of the mean.

About 99.7% of all values fall within 3 standard deviation of the mean.

Example 12: IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15. What percentage of IQ scores are between 70 & 130? What percentage of IQ scores are more than 145?

3^{rd} rule: Chebyshev's theorem: for any type of data set at least $(1-\frac{1}{k^2})100\%$ of the observations will lie within k standard deviation of the mean, where k is any number greater than 1.

Example 13: Heights of men have a mean of 176 cm and a standard deviation of 7 cm. Using the Chebyshev's theorem, at least what percentage of heights of men lie within 162cm and 190 cm.?

3-4 Measures of Relative Standing (Measures of Position)

In this section, we wish to describe the relative standing, position, of a certain data value within entire set of data or to compare values from different data sets. To be able to describe the measures of relative standing, we need to define z-score.

Def: **The z-score** (standard value) represents the distance that a data value is from the mean in terms of the number of standard deviations.

Population z-score $z=\frac{x-\mu}{\sigma}$ Sample z-score $z=\frac{x-\bar{x}}{s}$

(Round z to 2 decimal places.)

The z-score is unit less. It has mean of 0 and standard deviation of 1.

Example 14: The monthly utility bills in a city have a mean of \$70 and standard deviation of \$8. Find the z-scores that correspond to utility bills of \$60, \$71, and \$92.

z-scores and unusual value: Usual values: $-2 \le z \le 2$

Example 15: What are min. and max. usual values in example 14?

Percentiles:

Recall: Median (middle score) divides the lower 50%of a set of data from the upper 50%. In general, Percentiles divide a data set into one hundred. There are 99 percentiles. The kth percentile, P$_k$, of a set of data divides the lower k% of a data set from the upper (100-k)%. If a data value lies at the 40th percentile, then approximately 40% of data are less than this value and approximately 60% are higher than this value.

The following steps can be used to compute the kth percentile:

1. Arrange the data in ascending order.
2. Compute the locator, L, using this formula L=$\left(\frac{k}{100}\right)(n)$, k=percentile of the data, n=number of values in data set.
3. A) If L is an integer, the kth percentile, P$_k$, can be found by P$_k$ = (lth value +next value)/2.
   B) If L is not an integer, then round it up to the next larger integer. Then the value of, P$_k$ is the lth value, counting from the lowest.

Quartiles: Divide a data set into four equal parts. Q_1, Q_2, Q_3. Q_1=P_{25}, Q_2=P_{50} Q_3=P_{75}

Deciles: Divide a data set into ten equal parts: D_1, D_2, …, D_9 D_1=P_{10}, D_2=p_{20}, …, D_9=P_{90}

Example 16: The test scores of 15 employees enrolled in a CPR training course are listed. Find the first, second, and third quartiles, second deciles and 14 percentile of the test scores. 13, 9, 18, 15, 14, 21, 7, 10, 11, 20, 5, 18, 37, 16, 17.

The process of finding the percentile that corresponds to a particular value x is as indicated in the following expression:

Percentile of value of x = $\frac{number\ \ of\ values\ \ less\ than\ x}{total\ \ number\ \ of\ values} \cdot 100$

(Round the result to the nearest whole number)

The Interquartile range (IQR) of a data set is the difference between the third and first quartiels. (IQR)=$Q_3 - Q_1$

The IQR is a measure of variation that gives you an idea of how much the middle 50% of the data varies. It can also be used to identify outliers. Any data value that lies more than 1.5 IQRs to the left of Q_1 or to the right of Q_3 is an outlier.

Example 17: Find the interquartile range of the 15 test scores given in Example 16. What can you conclude from the results.

Box-and-whisker plot is an exploratory data analysis tool that highlights the important features of a data set.

Guidelines for drawing a Box-and-Whisker Plot:

1. Find the five-number summary of the data set. (the min. entry, Q_1, Q_2, Q_3, and the max. entry)
2. Construct a horizontal scale that spans the range of the data.
3. Plot the five numbers above the horizontal scale.
4. Draw a box above the horizontal scale from Q1 to Q3 and draw a vertical line in the box at Q2.
5. Draw whiskers from the box to the min. and max. entries.

Example 18: Draw a box-and-whisker plot that represents the 15 test scores given in Example 17. What can you conclude from the display?