WIIT 7751

Lesson D Guided Notes

The Mean

The mean of a data set is a measure of center. If we imagine each data value to be a weight, then the mean is the point at which the data set balances.

Notation:

A list of *n* numbers is denoted by $x_1, x_2, x_3, \ldots x_n$ $\sum x_i$ represents the sum of these numbers:

$\sum x_i = x_1 + x_2 + \cdots x_n$

If $x_1, x_2, x_3, \ldots x_n$ is a sample, then the mean is called the sample mean and is given by

If $x_1, x_2, x_3, \ldots x_N$ is a population, then the mean is called the population mean and is given by

Example 1 During a semester, a student took five exams. The population of exam scores is 78, 83, 92, 68, and 85. Find the mean.

Median

The median is another measure of center.

The median is a number that splits the data set in half, so that half the data values are less than the median and half of the data values are greater than the median.

The procedure for computing the median differs, depending on whether the number of observations in the data set is even or odd.

Procedure for Finding the Median

Step 1: Arrange the data values in increasing order.

Step 2: Determine the number of data values, *n*.

Step 3: If *n* is odd, the median is the middle number. In other words, the median is the value in position $(n + 1)/2$.

If *n* is even, the median is the average of the two middle numbers. That is, the median is the average of the values in positions $n/2$ and $n/2 + 1$.

Example 2 During a semester, a student took five exams. The population of exam scores is 78, 83, 92, 68, and 85. Find the median.

Example 3 One of the goals of medical research is to develop treatments that reduce the time spent in recovery. Eight patients undergo a new surgical procedure, and the number of days spent in recovery for each is as follows.

20 15 12 27 13 19 13 21

Resistant: A statistic is **resistant** if its value is not affected much by extreme values (large or small) in the data set.

Example 4 Five families live in an apartment building. Their annual incomes, in dollars, are 25,000, 31,000, 34,000, 44,000 and 56,000. The Smith family, whose income is 25,000, wins a million dollar lottery, so their income increases to 1,025,000. How are the mean and median values affected?

The Range

The **range** of a data set is a measure of spread. That is, it measure how spread out the data are.

The range of a data set is the difference between the largest and the smallest value.

Range = Largest Value – Smallest Value

Example 5 The following table presents the average monthly temperature, in degrees Fahrenheit, for the cities of San Francisco and St. Louis. Compute the range for each city.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
San Francisco	51	54	55	56	58	60	60	61	63	62	58	52
St. Louis	30	35	44	57	66	75	79	78	70	59	45	35
Source: National Weather Service												

Although the range is easy to compute, it is not often used in practice. The reason is that the range involves only two values from the data set; the largest and smallest.

The measures of spread that are most often used are the variance and the standard deviation, which use every value in the data set.

Variance

When a data set has a small amount of spread, like the San Francisco temperatures, most of the values will be close to the mean. When a data set has a larger amount of spread, more of the data values will be far from the mean.

The **variance** is a measure of how far the values in a data set are from the mean, on the average.

The variance is computed *slightly differently* for populations and samples. The population variance is presented first.

Let $x_1, x_2, x_3, \ldots x_N$ denote the values in a population of size *N*. Let μ denote the population mean. The **population variance**, denoted by σ^2 , is

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

Sample Variance

When the data values come from a *sample* rather than a population, the variance is called the **sample variance**. The procedure for computing the sample variance is a bit different from the one used to compute a population variance.

In the formula, the mean μ is replaced by the sample mean and the denominator is $n - 1$ instead of *N.* The sample variance is denoted by s^2.

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

Why divide by n – 1?

When computing the sample variance, we use the *sample mean* to compute the deviations. For the population variance we use the *population mean* for the deviations.

It turns out that the deviations using the sample mean tend to be a bit *smaller* than the deviations using the population mean. If we were to divide by *n* when computing a sample variance, the value would tend to be a bit smaller than the population variance.

It can be shown mathematically that the appropriate correction is to divide the sum of the squared deviations by n –1 rather than *n*.

Standard Deviation

Because the variance is computed using squared deviations, the units of the variance are the squared units of the data. For example, in Battery Lifetime example, the units of the data are hours, and the units of variance are squared hours. In most situations, it is better to use a measure of spread that has the same units as the data.

We do this simply by taking the square root of the variance. This quantity is called the **standard deviation**. The standard deviation of a sample is denoted s, and the standard deviation of a population is denoted by σ.

$$s = \sqrt{s^2} \qquad\qquad \sigma = \sqrt{\sigma^2}$$

Example 6 A company that manufactures batteries is testing a new type of battery designed for laptop computers. They measure the lifetimes, in hours, of six batteries, and the results are presented in the following table. Find the sample variance of the lifetimes.

Battery Lifetime	3	4	6	5	4	2

Example 7 Find the sample standard deviation of the Battery Lifetime measurements.

Standard Deviation and Resistance

Recall that a statistic is **resistant** if its value is not affected much by extreme data values.

The standard deviation is not resistant.

That is, the standard deviation is affected by extreme data values.

Quartiles

There are three special percentiles which divide a data set into four pieces, each of which contains approximately one quarter of the data. These values are called the **quartiles**.

- The **first quartile**, denoted Q_1, is the 25th percentile.
  Q_1 separates the lowest 25% of the data from the highest 75%.
- The **second quartile**, denoted Q_2, is the 50th percentile.
  Q_2 separates the lower 50% of the data from the upper 50%. Q_2 is the same as the median.
- The **third quartile**, denoted Q_3, is the 75th percentile.
  Q_3 separates the lowest 75% of the data from the highest 25%.

Example 8 The following table presents the annual rainfall in Los Angeles (in increasing order) during the month of February from 1965 to 2006. Compute the first and third quartiles.

0.00	0.08	0.11	0.13	0.14	0.17	0.23	0.29	0.49	0.56	0.67	0.70	1.22	1.30
1.48	1.51	1.72	1.90	2.37	2.58	2.84	3.06	3.12	3.21	3.54	3.71	4.13	4.37
4.64	4.89	4.94	5.54	6.10	6.61	7.89	7.96	8.03	8.87	8.91	11.02	12.75	13.68