CS 1173: Statistical indicators

A = [2, 4, 2, 3, 5, 1, 4, 3] and B = [1, 2, 2, 1, 5, 4, 1, 3] and C = [NaN, 2, NaN, 3] are three MATLAB variables representing sets of measurements used in the examples below. Let D be a MATLAB matrix that has 4 rows and 3 columns. Let $x = [x_1, x_2, x_3, \ldots, x_n]$ represent a general set of n measurements.

1. The average or mean measures the central tendency of a set of numbers. To calculate the average, add the values and divide by the number of values. The average of x is:

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Example 1: The average of A is (2 + 4 + 2 + 3 + 5 + 1 + 4 + 3)/8 = 3

The MATLAB function for average is called **mean**.

Example 2: mean(D, 1) or mean(D) is the column mean of D. It is a single row of 3 values.

Example 3: mean(D, 2) is the row mean of D. It is a single column of 4 values.

Example 4: mean(A) is the mean of A. (For vectors, you never need the second parameter and the result is always a single number.)

Example 5: The mean of the entire array D is mean(D(:)). Here we use the linear representation of D to extract a single value.

Example 6: mean(C) is NaN. MATLAB returns NaN results if any of the values in the vector are NaN.

The MATLAB function **nanmean** computes the average by completely ignoring NaN.

2. The median is another measure of central tendency. In contrast to the average, the median is not heavily influenced by outliers. When the average and the median are far apart in value, you can surmise that there must be some outliers in the data. To calculate the median, sort the values and take the middle one. If the number of items is even, average the middle two values.

Example 7: sort(A) is [1, 2, 2, 2, 3, 3, 4, 5]. The average of the middle two values is (2 + 3)/2 = 2.5

The MATLAB function for computing the median is called **median**. The median function follows the same rules as the mean function.

The MATLAB function **nanmedian** computes the median by completely ignoring NaN.

3. The mode is the value most frequently appearing in the data. This value may not be unique.

Example 8: The mode of B is 1, while the mode of A has two most frequent values: 2 and 4

4. The population standard deviation measures the spread or dispersion of values from the mean. The population standard deviation is calculated as the square root of the average of the squares of the distance of each value from the mean. The standard deviation of x is:

$$s = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(x_k - \overline{x}\right)^2}$$ where $\overline{x}$ is the mean of x.

Example 9: Calculate the standard deviation of A:

a) First compute the sum of the squares of the differences with the mean

$$(2 - 3)^2 + (4 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (5 - 3)^2 + (1 - 3)^2 + (4 - 3)^2 + (3 - 3)^2 =$$
$$1 + 1 + 1 + 0 + 4 + 4 + 1 + 0 = 12$$

b) Divide by n and take the square root = $\sqrt{12/8}$ = 1.225

Note: The s value defined above is called the ***population standard deviation*** because it assumes that the vector x contains all of the values of the data (that is the entire population). In statistics, the vector x usually represents just a subset of values (i.e., a sample) and we want to estimate the standard deviation of the entire population based on the sample. This alternative uses an $n - 1$ instead of an n in the above formula and is called the ***sample standard deviation*** or the ***unbiased estimator of the population standard deviation.***This issue is discussed further in the handout on populations and samples.

The **MATLAB function** for computing the standard deviation is called **std**.

Example 10: std(D, 1, 1) is the population standard deviation of the columns of D. It is a row vector of size 1 x 3.

Example 11: std(D) contains a row vector with the sample standard deviation. This could also be written as std(D, 0, 1).

Example 12: std(D, 1, 2) is the standard deviation of the rows of D. It is a column vector of size 4 x 1.

The MATLAB function **nanstd** computes the standard deviation by completely ignoring NaN. The nanstd function follows the same rules as std.

5. The pth percentile is the data value such that p% of the data fall below that value. The 50% percentile is the median. The 25th percentile is sometimes called the first quartile.

Example 13: MATLAB statement [a, b] = prctile(x, [25, 75]); has the 25th percentile value in the variable a and the 75th percentile value in the variable b.

6. Boxplots use the 75th percentile for the top of the box and the 25th percentile for the bottom of the box. The horizontal line across the box corresponds to the median value.

Example 14: The MATLAB statement boxplot(x) produces a boxplot showing the distribution of values in the vector x.

7. The interquartile range (IQR) is the 75th percentile minus the 25th percentile.

Example 15: The MATLAB statement iqr(x) calculates a single value representing the interquartile range of x