Ch 4: The Average and the Standard Deviation

Histograms give a good summary of the data, but frequently having to use a picture can be inconvenient.

Sometimes we can summarize a data set even more strongly, just by reporting numbers to indicate the _____ and _____ of the data. This is especially true if the histogram has only one peak and is (at least roughly) symmetric.

The Average

The most commonly used measure of the center of a set of numbers is the _____, also called _____.

The average of a list of numbers equals their sum, divided by how many numbers there are in the list.

Ex: What is the average of the following list of numbers: 3, 6, 9, 7?

Ex: The Health and Nutrition Examination Survey (HANES) surveyed a cross-section of 20,322 Americans, and found that the men had an average height of 5' 9", and an average weight of 171 lbs. The women had an average height of 5' 3.5", and an average weight of 146 lbs.

By reducing data sets to their averages, we can compare many groups simultaneously.

The 2 plots (see Figure 3, page 59, from your textbook on additional handout) show the height and weight data for both genders for six different age groups.

Does the height plot mean people are shrinking?

Intuitively, what percentage of people do you think has a weight below average?

And what percentage do you think has a weight above average?

The histogram (see Figure 4, page 62, from your textbook on additional handout) shows a histogram for the weights of the women in the HANES sample.

In the histogram of women's weights, only about _____ of the area is to the right of the average, rather than the _____ we might expect.

To see why, consider the averages of the following sets of numbers:

- 1, 2, 2, 3; average = _____
  values below average: ___
  values equal to average: ___
  values above average: ___

- 1, 2, 2, 5; average = _____
  values below average: ___
  values equal to average: ___
  values above average: ___

- 1, 2, 2, 7; average = _____
  values below average: ___
  values equal to average: ___
  values above average: ___

A histogram balances when supported at the average.

The Median

The _____ is the value with 50% of the area above and 50% below.

How to find the median of a data set:

1. Sort the values from smallest to largest.

2. If you have an odd number of values, the median is the center one.

   If you have an even number of values, the median is the average of the two center values.

Ex:

List 1: 3, 5, 1, 8, 0

Sorted list:

Median:

Ex:

List 2: 2, -1, 5, 1

Sorted list:

Center values:

Median:

If the histogram of your data is symmetric, the median and the average will be close.

If the histogram has a long right tail, the average will be greater than the median.

If the histogram has a long left tail, the average will be less than the median.

Ex: What do you think is larger – the average or the median body weight of the women in the HANES sample? Look at Figure 4, page 62, from your textbook on additional handout.

If your histogram has an extremely long tail, the mean will be strongly influenced by the few cases in the tail, and the median will better indicate the center of your data.

Ex: Why might we prefer the median to the average as a measure of the center of a list of incomes of employees of Microsoft?

The Standard Deviation

The _____ (SD) is a measure of how spread out data values are around the average.

Most numbers from a data set will be within _____ of their average.
Few will be more than _____ away from their average.
Almost none will be more than _____ away from their average.

More precisely:

- about _____ of values will be within 1 SD of the average.

- about _____ of values will be within 2 SD's of the average.

- about _____ of values will be within 3 SD's of the average.

How to calculate the SD:

1. Find the average.

2. Find the *deviations from the average* (entry - average).

3. Square the deviations from the average.

4. Average the squared deviations from the average.

5. Take the square root.

$$\text{SD} = \sqrt{\text{average of [(deviations from avg)}^2\text{]}}.$$

Ex: Find the SD of the list: 5, 12, 15, 20.

In practice, we usually use computers or calculators with statistical functions to calculate the SD. (But don't if you're asked to "show your work.")

Be aware, there's another number, slightly larger than the SD, also sometimes referred to as the "standard deviation" (we'll call it SD^+).

Most calculators use the symbol σ for the SD, and s for the SD^+.

Read Section 4.7 in the textbook to learn how to find out whether your calculator calculates SD or SD^+.

Note: in calculating the SD, we are using the *root-mean-square* operation.

$$\text{r.m.s.(list)} = \sqrt{\text{average of (entries}^2\text{)}}$$