MATH 3070 Introduction to Probability and Statistics
Lecture notes
Measures of Spread or Dispersion

Objectives:

1. Identify and compute variance and standard deviation
2. Understand degrees of freedom (optional)

Sample Variance and Standard Deviation

Having found the center of the data set, we still need to measure the spread to completely describe a data set. We can compute the **standard deviation** to measure the distance from the mean, or central point, of the data set. This gives us a measure of the spread of the data set. First we have to compute the **variance** of the data set, then we take the square root to find the standard deviation.

We must pay attention to how we compute the variance, however, because the data will make this value "disappear."

The **variance** is computed by summing the square of the difference between each item in the data set and the mean of the data set ($x\bar{x}$) and dividing by $(n-1)$. We have to sum the square of the difference because the number of observations greater than the mean is the same as those less than the mean, and the total difference of the greater observations is exactly equal to the total difference of the lesser observations.

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

The **standard deviation** is the square root of this value, and the more interesting of the two values. Using the standard deviation and the mean we can describe any data set we encounter.

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

An example from Moore:

The metabolic rates for participants in dieting study (in calories) $[n = 7]$ are as follows:

$$1792 \quad 1666 \quad 1362 \quad 1614 \quad 1460 \quad 1867 \quad 1439$$

First we compute the mean, like so

$$\begin{aligned} mean &= \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7} \\ &= 1600 \end{aligned}$$

Now, let's compute the deviations from the mean, but we won't take into account the counterbalancing differences around the mean. When we do the final summation, we get zero.

$$\begin{aligned} 1792 - 1600 &= 192 \\ 1666 - 1600 &= 66 \\ 1362 - 1600 &= -238 \\ 1614 - 1600 &= 14 \\ 1460 - 1600 &= -140 \\ 1867 - 1600 &= 267 \\ 1439 - 1600 &= -161 \\ \text{Total} &= 0 \end{aligned}$$

So instead we square the differences. This eliminates the sign on the number thereby making all the values positive. When we sum these values, we get something other than zero.

$$\begin{aligned} (1792 - 1600)^2 &= 36864 \\ (1666 - 1600)^2 &= 4356 \\ (1362 - 1600)^2 &= 56644 \\ (1614 - 1600)^2 &= 196 \\ (1460 - 1600)^2 &= 19600 \\ (1867 - 1600)^2 &= 71289 \\ (1439 - 1600)^2 &= 25921 \\ \text{Total} &= 214870 \end{aligned}$$

Now taking the sum of the squared differences (214870) and dividing by the number of observations minus 1 $(n - 1)$ (6) we obtain

$$\text{variance} = 35811.6667$$

the square root of which is

$$\text{standard deviation} = 189.2397.$$

The symbol used to represent the sample variance is "little s squared" (s^2). For the population variance we use little sigma squared (σ^2).

The standard deviation of the data is the square root of the variance. The symbol used is either "little s" (s) or "little sigma" (σ), depending upon whether it is the sample deviation or the population deviation.

So, to find the standard deviation (and variance) of a data set we follow these steps (Cryer and Miller):

1. Compute the deviations for the data set $(x_i - \bar{x})$.

2. Square all the deviations and sum them.
3. Divide the sum by $n - 1$.
4. Finally, take the positive square root to get the standard deviation s.

It is important to note a few things about the standard deviation:

- The units for the standard deviation are the same as the units for the observations.
- The standard deviation is never negative. It is zero if, and only if, all the values are equal.
- The standard deviation is strongly influenced (not resistant) by extreme observations and outliers.

Degrees of Freedom (optional)

Note that we divide by $n - 1$ to compute the variance. Why would we do that? It is because not all the data items are "free" to choose their value. Remember that all the deviations sum to zero? If that is the case, the last data point to be summed **has** to make up any difference (positive or negative) between the sum so far and zero. The other data points $(1, 2, 3, \ldots, (n-1))$ effectively made the decision for the last, or n^{th}, data point.

Another way to think about this is to consider what happens when you go out with a group of friends. Eventually the bill comes and it's time to pay up. The bill starts around the table and each person puts in what they think is their share of the bill, or what they have (usually the case). When the bill gets to the last person (you, maybe) the difference between what is owed and what has been contributed (don't forget the tip!) has to be made up. If you're lucky there is more than what you owe and your share is less. Otherwise, you get to pay more.