MATH 3070 Introduction to Probability and Statistics
Lecture notes
Measures of Centrality

Objectives:

1. Identify and compute/derive mean, median, and mode
2. Understand summation notation ("big sigma")

Measures of Central Tendency

Graphical descriptions of data sets are only so useful. If the data set is large, constructing the graph (or any type) becomes tedious. Though we can see the shape, center, and spread of a distribution from a histogram or stem and leaf plot, we can't quantify it. We need to figure out a numerical method for describing data.

Important terms to remember at this point are **statistic** and **parameter**. A statistic is a numerical descriptive measure computed from sample data. A parameter is a numerical descriptive measure of a population. Since we can't always get a complete population (N), we have to rely on samples (n) and the statistics computed from them. We use different symbols to represent statistics than for parameters.

Our three most common measures of central tendency are **arithmetic mean**, **median**, and **mode**.

The arithmetic mean, or **average**, is the sum of the data values in a data set divided by the number of items. (show formula) If the mean is for the population, the symbol used is μ ("mu"). If the mean is for a sample, the symbol is $\bar{x}$ ("x-bar") (or "y-bar").

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In statistics we use several Greek letters to represent various concepts or operations. Here we use the "big sigma" (Σ) to represent the summation of a series of items. The "big sigma" allows us to write, in a short hand notation, a long series of operations that repeat.

The **median** is the middle number of a series of data points when the data points are arranged in ascending, or descending, order. To calculate the median we must first order the data, usually from smallest to largest. In truth it does not really matter since the median will be the same regardless of where we start. If there are an odd number of items, the median is the middle item $[\frac{(n+1)}{2}]$. If there are an even number of items, the median is the average of the two items in the middle $[\frac{n}{2} + (\frac{n}{2} + 1)]$. The median is the point where there is one-half (50%) of the data set on either side, similar to the median in a road.

The **mode** is the most common value (the value that appears most frequently). To find the mode we group the data items together and then count the occurrences. If we graph the data, the mode is the data value with the highest bar (in a bar graph) or the most dots (in a dot plot). In some cases there may be more than one mode. In cases where there are two modes we say the data are "bi-modal". This usually indicates the presence of two distinct subgroups within the population, such as gender. Once the subgroups are identified they can be separated out for individual analysis. This is usually a good idea since two distinct subgroups will most likely not have the results.