

MthStat 465, Spring 2005, Lecture Number 1// Measures of Centrality

To summarize a univariate data set by a single number we use a *measure of centrality*. The most popular are:

- The sample mean;
- The sample median;
- Trimmed sample means.

Each of these may seem natural to consider. An important question is: “In what sense (if any) are they good?”

Throughout, $D := (d_1, d_2, \dots, d_N)$ represent a sequence of N real numbers. These numbers may not all be different from one another, and represent data collected from an experiment, such as weights, heights, times, etc.

1. SAMPLE MEAN:

The mean of D , which we will denote by \bar{D} , is just the arithmetic average:

$$\bar{D} := N^{-1} \sum_{k=1}^N d_k = \frac{d_1 + \dots + d_N}{N}.$$

2. SAMPLE MEDIAN

Let $D_o := (d'_1, d'_2, \dots, d'_N)$ be the elements of D arranged from smallest to largest:

$$d'_1 \leq d'_2 \leq \dots \leq d'_N.$$

If N is odd then the median of D , denoted by D_m , is $d'_{(N+1)/2}$, the middle element of D_o . If N is even, then D_m is any number x so that $d'_{N/2} \leq x \leq d'_{(N/2)+1}$. In the case where N is even it is common to choose

$$D_m = \frac{1}{2} \left(d'_{N/2} + d'_{(N/2)+1} \right).$$

3. TRIMMED MEAN

If we drop some fraction of the greatest and smallest elements of D , this is called *trimming*. If we then compute the mean of the remaining elements, the resulting number is called the *trimmed mean*. We shall not discuss trimmed means in any depth because it is complicated to explain and justify what percentage of the data should be trimmed.

4. JUSTIFICATION FOR USING MEAN AND MEDIAN

If we are to describe our data D by a single number it is reasonable that this single number be as close to the data as possible. Depending on how we measure distance, various single number summaries arise.

What is generally agreed on is that we should choose a function p that has two properties:

- p is increasing;
- The domain and range of p are the non-negative real numbers with $p(0) = 0$;

and then we should choose our single number summary x to minimize

$$S(x) := \sum_{k=1}^N p(|x - d_k|).$$

If $p(x) = x$ we are to minimize

$$S(x) = \sum_{k=1}^N |x - d_k|.$$

Since the graph of $y = S(x)$ is piecewise linear with slope $= -N$ for $x < d'_1$ and slope $= N$ for $x > d'_N$ it is easy to determine that the median minimizes S by examining the slope of each section of the graph as we pass from smaller to larger values of the data.

If $p(x) = x^2$, then

$$S(x) = \sum_{k=1}^N (x - d_k)^2$$

and $y = S(x)$ is a parabola that opens upward and whose vertex has first coordinate equal to \overline{D} , the mean. This can be seen by expanding the expression for $S(x)$ and completing the square.