

Descriptive Statistics

a.k.a.

“Measures of Central Tendency” is the
Worst Chapter Title in the World

Ben Babcock
University of Minnesota

Two Types of Statistics

There are generally two types of statistics that we talk about: descriptive and inferential.

Descriptive statistic examples: mean, median, mode, variance, standard deviation, upper and lower quartiles, midspread (interquartile range), range, correlation*, etc.

Inferential statistic examples: z -test, t -test, F -test, χ^2 test, etc.

*Indicates:

The Three M's of Descriptive Statistics

Mean: the arithmetic average of all the scores. H, 62-63.

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N}$$

It looks like a kidney bean. The “x bar” is the symbol for the _____ mean. For the _____ mean, we use μ .

Median: the middle score of a group of scores after you have _____. If you have an even number of scores, you take the closest two middle scores and take the mean. H, 61-62.

Mode: the most _____ score. In the continuous case, the mode is where the density reaches its peak. H, 61.

The Three M's of Descriptive Statistics

Philosophical question:

Suppose you had a jigsaw, a skilsaw, and a miter saw.



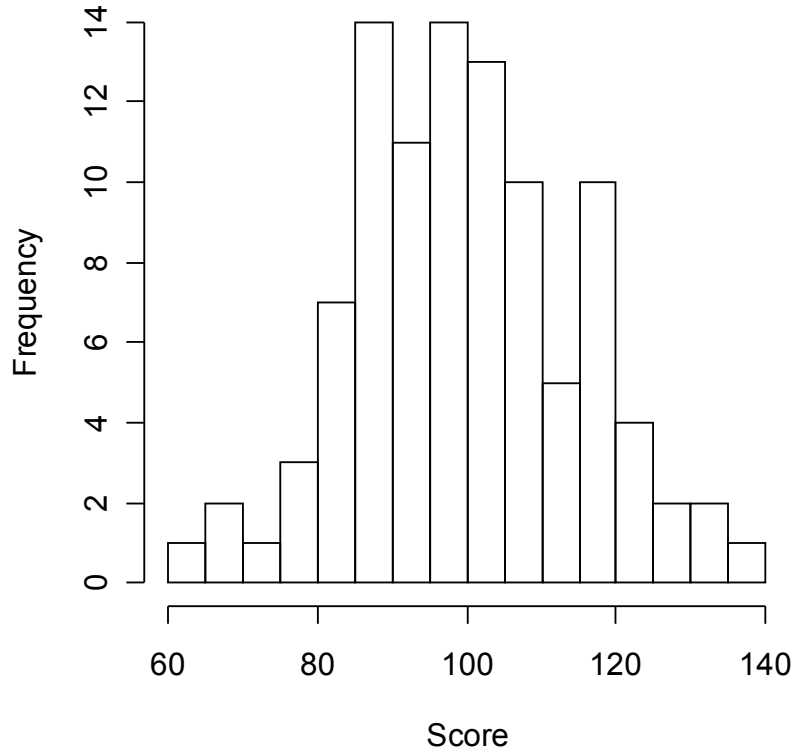
If you grouped these three things together, what would you call the group?

The Three M's of Descriptive Statistics

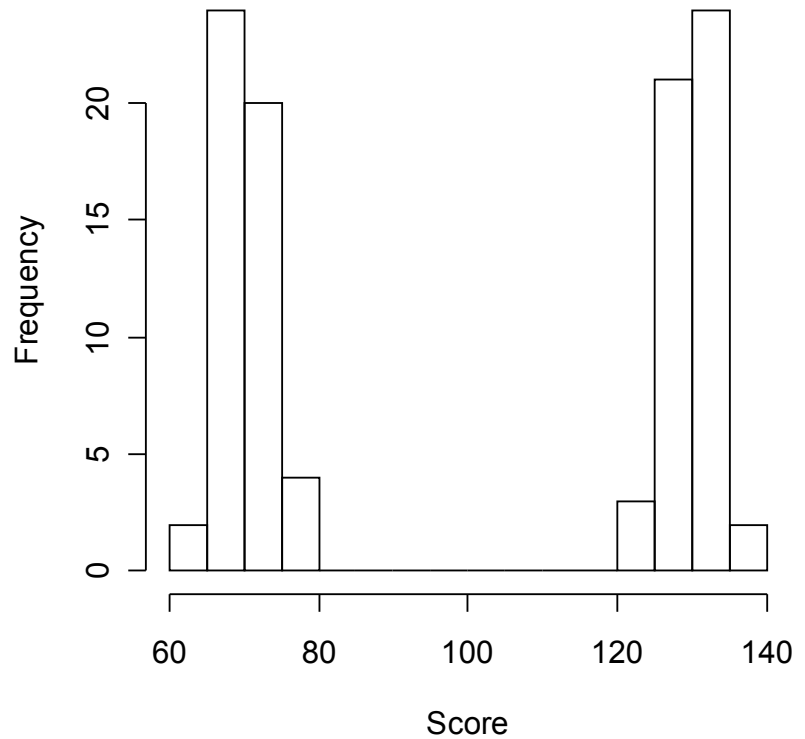
Howell calls the Three M's "Measures of Central Tendency".

The mean, median, and mode DO NOT MEASURE
CENTRAL TENDANCY!!!!!!!!!!!!!!

Histogram of X



Histogram of Y



The two distributions look nothing alike as far as
“central tendency” goes!

Three M's: The Good and the Bad

Mean

Good:

The most used descriptive statistic

Used with numerous _____ statistics

Bad:

Not robust (immune to; tough against) to _____

Three M's: The Good and the Bad

Median

Good:

Robust to _____

Half of the scores are below the median

Bad:

Not used with _____ statistics

H, 64

Three M's: The Good and the Bad

Mode

Good:

Discrete case: The score actually occurs

Cont's: The slope of the density plot is 0

Bad:

Is often biased

Not used with _____ statistics

H, 63-64

Unless you really need a slope = 0 or have nominal data, do not use the mode.

Variability

Variability: Scores are different.

Variability in scores is the _____ of statistics.

Variability: Range Statistic

Range: Simply the highest score minus the lowest score.

What are some reasons that this may not be a very good measure of variability?

Variability: Vocabulary

Percentile: A score under which a certain percentage of scores fall. H, 117

Quartiles: scores at which 25%, 50%, or 75% of the rest of the scores are below.

1st quartile: 25% below

2nd quartile: 50% below

3rd quartile: 75% below

Can you think of another name for the 2nd quartile?

Remember that boxplots use the quartiles. H, 88-89.

Variability: Interquartile Range

Interquartile Range (Midspread): The difference between the _____ and the _____ quartile.

Gives a range under which the middle 50% of the data fall.

What makes the midspread better than the range statistic?

Variability: Variance Statistic

Variance: s^2 or σ^2 (population, uses N , not $N - 1$). H, 80-82

$$s_x^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N - 1}$$

The variance is the *mean squared deviation from the mean*.

Mean: We are summing and then dividing by (something like) sample size. Kidney bean!

Squared: We are squaring something

Deviation: The difference between each _____ and the _____

Variability: Variance Statistic

The variance is extremely important in a large number of equations and in statistical theory.

The number by itself, however, is difficult to interpret.

$$s_x^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N - 1}$$

It would be nice to have something that is not in a squared unit. How do we change that?

Variability: Standard Deviation

The standard deviation is the solution to our squared unit blues. It is the most used variability measure.

Standard deviation:

$$s_x = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N - 1}}$$

The standard deviation is simply the square root of the variance. It is the _____ *mean squared deviation* from the mean. H, 83. Computational equation: H, 84

Variability: Why Divide by $N-1$?

Howell discusses this issue on pages 85-87.

What it boils down to is that, when using a sample of data to estimate the population variance, dividing by N gives an _____.

This introduces systematic _____ into the statistic. H, 85.