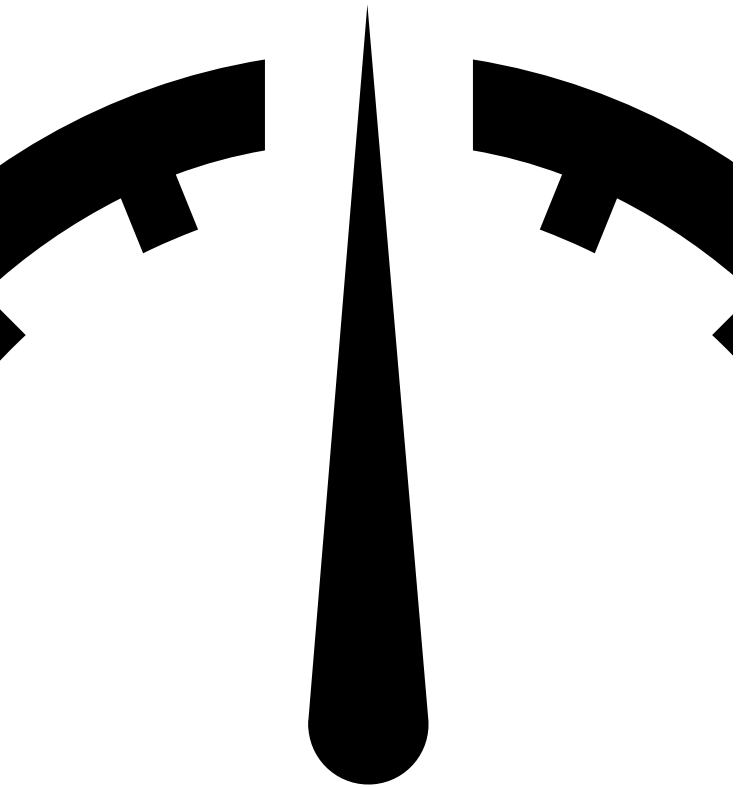
MATH9102 Fundamentals of Data Analysis

FUNDAMENTALS – DESCRIPTIVE STATISTICS

DR. DEIRDRE LAWLESS

Descriptive Statistics

USED TO SUMMARIZE OR DESCRIBE DATA FROM SAMPLES AND POPULATIONS



Measure of Central Tendency

A descriptive statistic for quantitative data.

A single number to serve as a representative value around which all the numbers in the set tend to cluster.

Mode

- the value that occurs most frequently for a variable in a set of data

Median

- the value in the middle; half of the values for a variable are larger than the median and half of the values are smaller than the median
- the middle value of a sequence of all the values in a distribution arranged from lowest to highest.

Mean:

- the arithmetic average of a group of values; the sum of the values divided by the number of values.

Which do you use? Mean, Mode, Median?

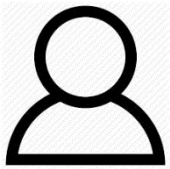
Does it make a difference?

Mean, Mode, Median

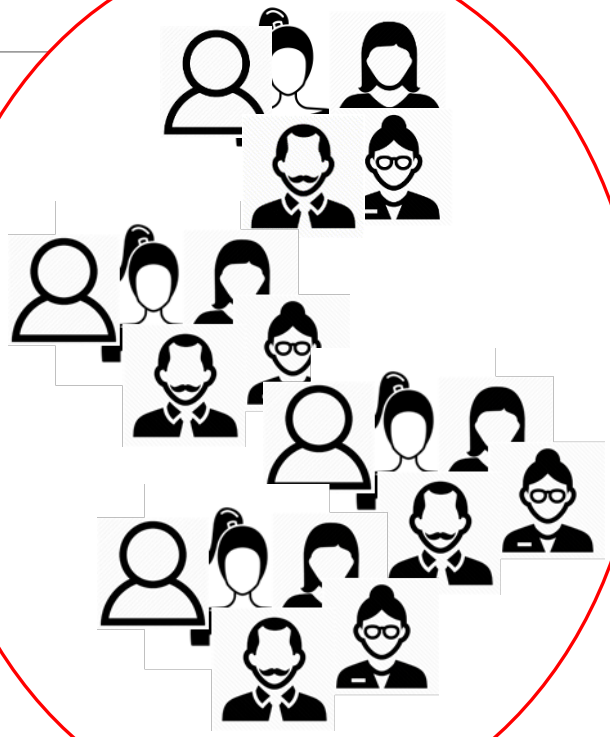
Sample Neighbourhood

One Hollywood Star making 4,465,000 annually

Five Professionals making 150,000



One person earning 35, 000



20 people earning 10,000

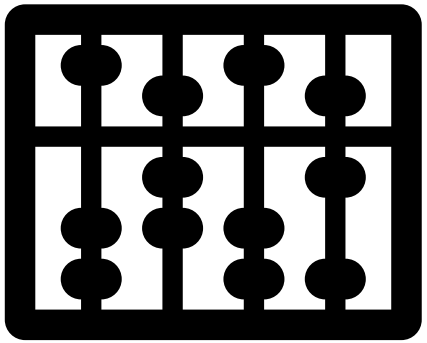
MEAN = 201,852 | MEDIAN =10,000 | MODE=10,000

Measures of Dispersion

Measure of Central Tendency needs to be considered in relation to the variability within the dataset.

Measures of Dispersion are descriptive statistics that describe how similar a set of values are to each other (or the range of values)

- The more similar the values are to each other, the lower the measure of dispersion will be
- The less similar the values are to each other, the higher the measure of dispersion will be
- In general, the more spread out a distribution is, the larger the measure of dispersion will be



Measures of Dispersion

There are three main measures of dispersion:

- The range
  - The difference between the largest value for a variable in a dataset and the smallest value
- The Interquartile Range (IQR)
  - Defined as the difference of the first and third quartiles divided by two
    - Order the data from least to greatest, Find the median
    - Calculate the median of both the lower and upper half of the data
    - The IQR is the difference between the upper and lower medians
- Variance/Standard Deviation
  - Concerned with how different values for a variable are from the mean

Measures of Dispersion

Variance: concerned with deviations from the mean (X-μ)

- First subtract the mean from each of the values gives use a *deviate* or a *deviation value* - how far a given value is from the typical, or average, value
- Then *square* the result
  - If we just added up the differences from the mean the negatives would cancel the positives
  - If we used absolute values we wouldn't get an accurate measure of spread
  - Squaring is the best option
  - *Variance* is defined as the average of the deviations from the mean squared:

$$s^2 or \sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

N here is the degrees of freedom = the number of independent pieces of information on which the estimate is based

Measures of Dispersion

Standard Deviation: the most useful and the most popular measure of dispersion.

- Concept was first introduced by Karl Pearson in 1893.
- Standard deviation = the **square root** of the **Variance**
- Use Greek symbol sigma σ
- The larger the value the more spread out around the mean the data is, smaller means less spread.
- Allows us to see how spread out *on average* individual cases are from the mean

Standard Deviation - Example

This is helpful in comparing samples.

For example:

- Programme X and Programme Y have the same mean test value for Probability and Statistics
  - Same number of students in each data set
  - Dataset is well formed and representative (similar for both programmes)
- Does this mean that students on both programmes performed equally well?

Standard Deviation

If the mean for Programme X is 60 and the standard deviation is 1.6 then

- 68% of the values in the dataset will lie between 58.4 and 61.6
  - MEAN-1SD (60-1.6=58.4) and MEAN+1SD (60+1.6=61.6)
- 99.7% of the values will lie between 55.2 and 64.8
  - MEAN-3SD (60-4.8=55.2) and MEAN+3SD (60+4.8=64.8)

If the mean for Programme Y is 60 and the standard deviation is 4.3

- 68% of the values in the dataset will lie between 55.7 and 64.3
  - MEAN-1SD (60-4.3=55.7) and MEAN+1SD (60+4.3=64.3)
- 99.7% of the values will lie between 47.1 and 72.9
  - MEAN-3SD (60-12.9=47.1) and MEAN+3SD (60+12.9=72.9)

Standard Deviation

So now we can compare programmes in more detail

- Programme X mean is 60
  - 68% of the values in the dataset lie between 58.4 and 61.6
  - 99% of the values in the data set lie between 55.2 and 64.8
- Programme Y mean is 60
  - 68% of the values in the dataset lie between 55.7 and 64.3
  - 99% of the values in the dataset lie between 47.1 and 72.9

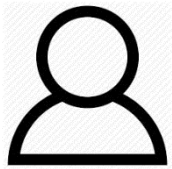
What can you conclude about this?

Standard Deviation, IQR, Range

Sample Neighbourhood

One Hollywood Star making 4,465,000 annually

Five Professionals making 150,000



One person earning 35, 000

20 people earning 10,000

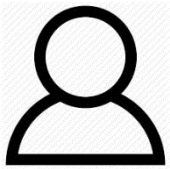
sd = 837,807	iqr =12,500	range=4,455,000

Full Description

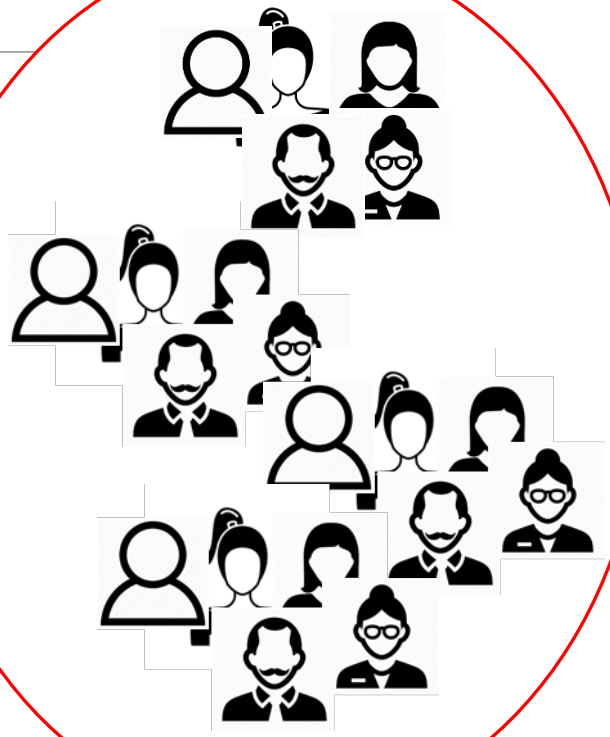
Sample Neighbourhood

One Hollywood Star making 4,465,000 annually

Five Professionals making 150,000



One person earning 35, 000



20 people earning 10,000

MEAN = 201,852 SD = 837,807	MEDIAN =10,000 IQR =12,500	MODE=10,000 RANGE=4,455,000

Getting Started

Open MATH9102W1.R

SECTION FOUR: MEASURES OF CENTRAL TENDENCY

SECTION FIVE: MEASURES OF DISPERSION

Your first lab

Instructions available in the file MATH9102-Lab-Instructions-W1.docx

Here Endeth the First Lesson…

SEE YOU NEXT WEEK!