Date.....

11/05/2020

PART 3: Practical Statistics

LESSON 1: DESCRIPTIVE STATISTICS PART-1

:· Data Types

→ Two types of Data:

1) Quantitative: data takes on numeric values that allow us to perform mathematical operations (like the number of dogs).

2) Categorical: → These are used to label a group or set of items (like dog breeds - collies, labs, Poodles etc.).

→ This is also of two types:

a) Categorical ordinal:

data take on a ranked ordering (like a ranked interaction on a scale from very poor to very good with the dogs).

b) Categorical Nominal:

data do not have an order or ranking (like the breeds of the dog).

Date.....

1) Types of Quantitative data :

a) Continuous:

→ Data can be split into smaller & smaller units exists.

→ An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds but there are still smaller units that can be associated with the age.

b) Discrete:

→ Data only takes on countable values.

→ The number of dogs we interact with is an example of a discrete data type.

* Analyzing Quantitative Data

→ Four aspects for Quantitative data:

1. Measure of center.
2. Measure of spread.
3. The shape of the data.
4. Outliers.

* Analyzing categorical data

→ Analyzing categorical data has fewer parts to consider.

→ Categorical data is analyzed usually by looking at the counts or proportions of individuals that fall into each group.

Date.....

→ For example, if we are looking at the breeds of the dogs, we would care about how many dogs are of each breed, or what proportions of dogs are of each breed type.

1) Measures of center :

There are 3 measures of center :

1) Mean
2) Median
3) Mode

a) The Mean :

The mean is often called the average of the expected value in mathematics. We calculate the mean by adding all of our values together, & dividing by the number of values in our dataset.

b) The Median :

- The median splits our data so that 50% of our values are lower & 50% are higher.
- For finding media, we must SORT the values in order.

Median for odd values :

If we have an odd no. of observ", the median is simply the number in the direct middle.

Date.....

- Median for Even values:
  If we have an even no. of observ.", the median is the average of the two values in the middle.

c) The Mode:

- The mode is the most frequently observed value in our dataset.
- There might be multiple modes for a particular dataset, or no mode at all.
- No Mode:
  → If all observ." in our dataset are observed with same freq., there is no mode.
  → eg: 1,1; 2,2; 3,3; 4,4
- Many modes:
  → If two (or more) numbers share the maximum value, then there is more than one mode.
  → eg: 1,2; 3,3,3,4,5,6,6,6,7,8,9.
  Here 2 modes → 3 & 6

* Notation:

→ Notation is a common language used to communicate mathematical ideas.
→ "Think of notation as a universal language used by academic & industry professionals to convey mathematical ideas".

Date.....

→ Notation has the following properties:-

1) understanding how to correctly use notation makes you seem really smart.
2) It allows you to read documentation, and implement an idea to your own problem.
3) It makes ideas that are hard to say in words easier to convey.

* Random variable:

→ Notated by a capital letter X

* Observed values:

Notated by lowercase letters ($x_4 = 5$).

LESSON 2: DESCRIPTIVE STATISTICS PART-2

2) Measures of Spread:

→ measures of spread are used to provide us an idea of how spread out our data are from one another.

→ Common measures of spread include:

1) Range
2) Interquartile Range (IQR)
3) Standard deviation
4) variance.

Date.....

Histograms:

→ Histograms are supervseful to understanding the different aspects of quantitative data.

Calculating the 5 number summary:

→ The 5 number summary consist of 5 values:

1. Minimum:
   The smallest number in the dataset.
2. Q_1:
   The value such that 25% of the data fall below.
3. Q_2:
   The value such that 50% of the data fall below.
4. Q_3:
   The value such that 75% of the data fall below.
5. Maximum:
   The largest value in the dataset.

→ Range:
The range is then calculated as the difference b/w the maximum and the minimum.

→ IQR:
The interquartile range is calculated as the difference b/w Q_3 & Q_1.

Date.....

:. Boxplot:

→ Boxplot useful for quickly comparing the spread of two data sets.

:. Standard Deviation and variance:

→ The standard deviation is one of the most common measures for talking about the spread of data.

→ It is defined as "the average distance of each observation from the mean".

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

→ variance:

finding the average squared difference of each observation from the mean.

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

:. Important final points

1. The variance is used to compare the spread of two different groups. A set of data with higher variance is more spread out than a dataset with lower variance.

2. Be careful though there might be an outlier (or outliers) that is increasing the variance, when most of the data are actually very close.

Date.....

2. When comparing the spread b/w two datasets, the units of each must be the same.

3. When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.

4. The standard deviation is used more often in practice than the variance, because it shares the units of the original dataset.

3. Shape:

→ From a histogram we can quickly identify the shape of our data, which helps influence all of the measures we learned in previous concepts.

→ We learned that the distribution of our data is frequently associated with one of the 3 shapes:

1) Right-skewed → mean > median
2) Left-skewed → mean < median
3) Symmetric (frequently normally distributed)
↓
mean = median.

Date.....

4) Outliers:

→ Outliers are points that fall very far from the rest of our data points.

→ This influences measures like the mean & standard deviation much more than measures associated with the five number summary.

* Common Techniques

When outliers are present, we should consider the following points.

1. Noting they exist & their impact on summary statistics.
2. If typo - remove or fix.
3. Understanding why they exist, and impact on questions we are trying to answer about our data.
4. Reporting the 5 number summary values often a better indication than measures like the mean & standard deviation when we have outliers.
5. Be careful in reporting. Know how to ask the right questions.

Date.....

Descriptive Statistics

- Descriptive statistics is about describing our collected data.

Inferential statistics

- Inferential statistics is about using our collected data to draw conclusions to a larger population.