Statistics

March 16, 2019

1 Math for Data Science - Statistics

1.1 ### What is Statistic?

- Dictionary Definition: A fact or piece of data obtained from a study of large quantity of numerical data.
- In general sense : The science of collecting, organizing presenting and interpreting data.
- Statistics is often categorized into descriptive and inferential statistics.
- It uses graphical methods to help makingnumbers visible for communicationpurposes.
- It uses analytical methods which providethe math to model and predict variation.

In Nutshell

1.2 ### Why do we need Statistics?

- To understand data better.
- To make effective decisions.

1.3 ### What are various types of Statistics?

Types of Statistics: * Descriptive Statistics - collection, presentation, description of data * Inferential Statistics - making decisions and drawing conclusions about populations.

1.4 ### What are Descriptive Statistics?

- Descriptive statistics involves describing, summarizing and organizing the data so it can be easily understood.
- Enable us to present the data in a more meaningful way, which allows simpler interpretation of the data.
- Typically, there are two general types of statistic that are used to describe data:
  - **Measures of central tendency:** these are ways of describing the central position of a frequency distribution for a group of data.
    - To describe this central position we use measures such as **mode, median, and mean.**
  - **Measures of spread:** these are ways of summarizing a group of data by describing how spread out the scores are.
    - To describe this spread we use measures such as **range, quartiles, variance and standard deviation.**

1.5 ### What are the Measures of Central Tendency?

- Used to summarise where the 'centre' of the data is.

1.6 ### Measures of Central Tendency - In Layman's Terms

```
In [26]: marks=[2,3,4,5,6,7,8,1000]
         sum_of_marks=0

         for i in marks:
             sum_of_marks=sum_of_marks+i

         print('total_marks : ',sum_of_marks)
         print('no.of students : ',len(marks))
         print('mean: ',sum_of_marks/len(marks))
```

```
total_marks :  1035
no.of students :  8
mean:  129.375
```

2 Measures of Central Tendency - Mean

- The mean is also called as arithmetic average.
- To calculate the average, or mean, add all values, then divide by the number of individuals.
- It is the "center of mass."
- Mean gets affected by the extreme values (Outliers).

2.1 ### Measures of Central Tendency - Median

- The Median (M) is often called the "middle" value and is the value at the midpoint of the observations when they are ranked from smallest to largest value.
- In an ordered array, the **median** is the **middle** number i.e., the number that splits the distribution in half.
- The median is not affected by the extreme values (Outliers).

Steps to get median: * Arrange the data from smallest to largest * If n is odd then the median is the single observation in the center (at the (n+1)/2 position in the ordering) * If n is even then the median is the average of the two middle observations (at the (n+1)/2 position; i.e., in between)

```
In [33]: salaries=[2,2,2,3,3,3,4,4,4,5,5,5,5,6,7,8,1000]
         salaries.sort()
         print('salaries : ',salaries)
         print('no.of data points : ',len(salaries))
         print('mid element index : ',len(salaries)/2)
         print('median: ',(salaries[3]+salaries[4])/2)
```

```
salaries :  [2, 3, 4, 5, 6, 7, 8, 1000]
no.of data points :  8
```

```
mid element index :  4.0
median:  5.5
```

2.2 ### Measures of Central Tendency - Mode

- Mode is the value that occurs most often.
- Mode is not affected by extreme values (Outliers).
- Mode is userd for either numerical or categorical data.
- There may be no mode.
- There may be serveral modes (Multi Modal)

```
In [ ]: Data Imputation - Mean, Median, Mode, SD

        Student:
        --------
        Name, Age, Gender, School, Marks, No.Of Siblings

        Quantitative: Age, Marks, No.of Siblings
            Descrete (Whole Number): No.of Siblings 2
            Continues (Real Number): Age (15.5), Marks (92.4)

        Qualitative: Name, Gender, School
        [1,2,2,2,2,2,2,2,            ,100,200,300]
```

2.3 ### Measure of Central Tendency Most Useful When?

2.4 ### What are the Measures of Spread?

- Measures of Spread tells us how much a data sample is spread out or scattered.
- **Range**
- **Inter Quartile Range**
- **Standard Deviation**
- **Variance**

2.5 ### Measures of Spread - Range

- The range is the Maximum value minus the minimum value and gives the full extent of the range of observations.
- Notice that the range is one number, the difference between the Maximum and the minimum.

2.6 ### Measures of Spread - Inter Quartile Range (IQR)

- The InterQuartile Range (or sometimes InnerQuartile Range).
- We will use the notation IQR – is defined as the difference between the 3rd quartile and the 1st quartile.
- Notice that the IQR gives the range of the middle 50% of the data.

2.7 ### Measures of Spread - Standard Deviation

```
In [ ]: Age=[5,4,3,4,5,6]

        [5-4.5, 3-4.5, 4-4.5, 5-4.5, 6-4.5]

        import numpy as np
        np.mean(Age)
```

- Quantify spread of the distribution by measuring how far observations are from mean,
- In other terms this measure is based upon the deviations of each value from the mean or average value.
- The interpretation of the standard deviation is as the typical or average distance between observed data values and the sample mean
- To calculate the standard deviation, we take each observation and subtract the sample mean, xbar.
- We square each of the differences, and add these squared deviations.
- We divide that total by n – 1 where n is the number of observations.
- At this stage we have what is called the **variance, or ssquared.**
- This is sometimes used in statistical methods, however, it is in units which are the square of the original units which makes it difficult to interpret.
- We take the square root to obtain the final result for the standard deviation, s. This value will have the units of the original data and have the interpretation of the typical distance an observation differs from the sample mean.
- Because each value is weighted equally, the standard deviation is influenced by outliers and extreme values.
- In some scenarios, the standard deviation is less reliable than the mean in this regard and should be used with caution for highly skewed distribution or distributions with extreme outliers. Even moderate skewness and relatively mild outliers can have a dramatic impact on the standard deviation.

Variance
Standard Deviation

2.8 ### Variance and Standard Deviation of a Population

2.9 ### What are the key properties of Standard Deviation?

2.10 ### What is the FiveNumber Summary?

- The five values: minimum, Q1, Median, Q3, and Maximum make up what is commonly called “the fivenumber summary”.
- Each of these values represents a measure of position or location.

2.11 ### How to represent Five Number Summary using Box Plot?

2.12 ### Suspected outliers: how to detect outliers?

- Outliers are troublesome data points, and it is important to be able to identify them.

- One way to raise the flag for a suspected outlier is to compare the distance from the suspicious data point to the nearest quartile (Q1 or Q3).
- We then compare this distance to the interquartile range (distance between Q1 and Q3).
- We call an observation a suspected outlier if it falls more than 1.5 times the size of the interquartile range (IQR) above the first quartile or below the third quartile.
- **This is called the “1.5 * IQR rule for outliers."**

2.13 ### What are Inferential Statistics?

- Inferential statistics is one of the two main branches of statistics.
- Inferential statistics use a random sample of data taken from a population to describe and make inferences about the population.
- Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible.
  - For example, to measure the diameter of each nail that is manufactured in a mill is impractical.
  - You can measure the diameters of a representative random sample of nails.
  - You can use the information from the sample to make generalizations about the diameters of all of the nails.

2.14 ### Commonly used Terms - Sample, Population

2.15 ### Inferential Statistics - What is a test statistic?

- A test statistic is a random variable that is calculated from sample data and used in a hypothesis test.

2.16 ### Inferential Statistics - What is a confidence interval?

- A confidence interval is simply a way to measure how well your sample represents the population you are studying.

2.17 ### What are the different types of Data?

2.18 ### What are the different Levels of Measurement?

2.19 ### What are the common Data Distributions?

2.20 ### What are the measures of Symmetry?

- The **mean** is pulled toward the skew.

2.21 ### How to perform data analysis?

- ALWAYS PLOT DATA BEFORE DECIDING ON A NUMERICAL SUMMARY.
- **How to choose summary statistics?**
- Use: 5-number summary is better than the mean and s.d. for skewed data;
- Use mean & s.d. for symmetric data.