STATISTICS FOR DATA SCIENCE PART - 3

Estimates of Variability:

- Location is just one dimension in summarizing a feature.There are some other techniques for exploring a feature.
- The other technique is measuring variability also known as dispersion.
- This dispersion measures whether the data is clustered or spread out.
- High dispersion means the data is spread out a lot where as low dispersion means the data is clustered highly.

Key terms in estimating variability:

- The key terms used in dispersion measurement are

  Deviation: The difference between observed value and estimation of location.Also known as residual,error.

  Variance: The sum of squared deviations from the mean divided by *n* – 1 where *n* is the number of data. Also known as mean-squared-error.

  Standard Deviation: The square root of variance. Also known as l2 norm,euclidean-norm.

  Mean absolute deviation: The mean of the absolute value of the deviations from the mean. Also known as l1 norm, Manhattan norm.

  Median absolute deviation from median: The median of the absolute value of the deviations from the median.

  Range: The difference between the largest and the smallest value in a data set.

  Order statistics: Metrics based on the data values sorted from smallest to biggest. Also known as ranks.

  Percentile: The value such that *P* percent of the values take on this value or less and (100–P) percent take on this value or more. Also known as quantile.

  Inter quartile range: The difference between the 75th percentile and the 25th percentile. Also known as IQR.

Standard Deviation and Related Estimates:

- Differences are the main basis for calculating measures of dispersion.
- Consider the data elements 1,4,4. The mean of these numbers is 3. The deviations from mean are 1-3,4-3,4-3 which are -2,1,1.
- The sum of these deviations is 0. This means that the data is not having any dispersion which is incorrect.
- Thus the correct way to measure the dispersion is to take the absolute values of the deviations.Then we end up with the deviations 2,1,1 and the sum of deviations is 4 now which provides a better insight.
- Now the mean of the deviations is (2+1+1)/3 which is 1.33. This is known as mean absolute deviation*.*
- The variance is the mean of sum of squares of all deviations from the mean. For the above data the variance is $(2^2+1^2+1^2)/2$ which is 3. Here instead of 3 which is *n* the sum is divided with 2 which is *n-1*. This is to make the variance and standard deviation unbiased i.e, it doesn't depend on the sample mean calculation. Thus we have *n-1* degrees of freedom with one constraint that the variance is calculated based on the calculation of mean.
- Standard deviation is the square root of variance and here it is square root of 3 which is 1.72. This is the most used dispersion measurement.
- The variance and standard deviation are not robust. They are sensitive to outliers.
- The robust dispersion estimation is the median absolute deviation from the median.
- The standard deviation is always greater than the mean absolute deviation, which itself is greater than the median absolute deviation

Estimates Based on Percentiles:

- Sorted data can also be used to understand the dispersion of the data. Statistics based on this technique are called as ordered statistics.
- Range is a good measure to measure dispersion but it is not reliable as it is sensitive to outliers.
- To overcome this percentile concept is used.
- *P*th percentile is a value such that at least *P* percent of the values take on this value or less and at least (100 – *P*) percent of the values take on this value or more.
- To find the 70^{th} percentile first sort the data and then take the 70 percent of data from smallest to the largest then the last number will be 70^{th} percentile.
- Median is a special case which is 50^{th} percentile.
- Quantile is similar to percentile. For example 80^{th} percentile can be considered as 0.8 quantile.
- The common measure of dispersion is the Inter Quantile Range which is the difference between the 75^{th} percentile and the 25^{th} percentile.
- The calculation of the quantiles is expensive as it requires sorting of the data. So machine learning and data science use some approximation measure which will give accuracy upto a certain level.

Python Implementation:

- Now It is time for practical implementation. In this section implementation of dispersion measurement in Python is discussed.
- Consider the data frame *Heart.csv*

```
In [2]:  data = pd.read_csv('heart.csv')
```

```
In [3]:  data.head()
```

Out[3]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

- The Libraries to be imported are

```
import numpy as np
import pandas as pd
from scipy import stats
import weighted
```

- The code snippet and output are

```
In [8]:  print("Variance:",np.var(data['thalach']))
         print("Standard Deviation:",np.std(data['thalach']))
         print("IQR:",stats.iqr(data['thalach']))
         print("Mean Absolute Deviation:",np.mean(data['thalach'].mad()))
         print("Median Abosolute Deviation:",np.median(abs(data['thalach']-np.median(data['thalach']))))

         Variance: 522.9148994107331
         Standard Deviation: 22.86733258188924
         IQR: 32.5
         Mean Absolute Deviation: 18.484396954546938
         Median Abosolute Deviation: 15.0
```