Statistics

Descriptive statistics

Descriptive statistics

- The basic descriptive statistics to give us an idea on the variables and their distributions
- Permit the analyst to describe many pieces of data with a few indices
- Central tendencies
  - Mean
  - Median
- Dispersion
  - Range
  - Variance
  - Standard deviation

Central Tendancies

Mean / Average

Given sequence:
13, 18, 13, 14, 13, 16, 14, 21, 13
The mean is the usual average, so:
(13+18+13+14+13+16+14+21+13) / 9= 15

Mode

The mode is the number that is repeated more often than any other, so 13 is the mode.

Median

Order the list : 13, 13, 13, 13, 14, 14, 16, 18, 21

There are **nine numbers** in the list, so the middle one will be **(9+1)/2=10/2 = 5**

= 5th number, So the median is 14.

Central tendencies: Mean and Median

Central tendencies

- Mean
  - The arithmetic mean
  - Sum of values/ Count of values
  - Gives a quick idea on average of a variable

Central tendencies

- Mean
  - The arithmetic mean
  - Sum of values/ Count of values
  - Gives a quick idea on average of a variable

Mean in Python

```
gain_mean=Income["capital-gain"].mean()
gain_mean
```

Median

- Mean is not a good measure in presence of outliers
- For example Consider below data vector
  - 1.5,1.7,1.9,0.8,0.8,1.2,1.9,1.4, 9 , 0.7 , 1.1
- 90% of the above values are less than 2, but the mean of above vector is 2
- There is an unusual value in the above data vector i.e 9
- It is also known as outlier.
- Mean is not the true middle value in presence of outliers. Mean is very much effected by the outliers.
- We use median, the true middle value in such cases
- Sort the data either in ascending or descending order

Median

1.5
1.7
1.9
0.8
0.8
1.2
1.9
1.4
9
0.7
1.1

0.7
0.8
0.8
1.1
1.2
1.4
1.5
1.7
1.9
1.9
9

- Mean of the data is 2
- Median of the data is 1.4
- Even if we have the outlier as 90, we will have the same median
- Median is a positional measure, it doesn't really depend on outliers
- When there are no outliers then mean and median will be nearly equal
- When mean is not equal to median it gives us an idea on presence of outliers in the data

Mean and Median

Import “Census Income Data/Income_data.csv”

```
#Mean and Median on python
gain_mean=Income["capital-gain"].mean()
gain_mean

gain_median=Income["capital-gain"].median()
gain_median
```

Mean is far away from median. Looks like there are outliers, we need to look at percentiles and box plot.

Dispersion Measures : Variance and Standard Deviation

Dispersion

- Just knowing the central tendency is not enough.
- Two variables might have same mean, but they might be very different.
- Look at these two variables. Profit details of two companies A & B for last 14 Quarters in MMs

															Mean
Company A	43	44	0	25	20	35	-8	13	-10	-8	32	11	-8	21	15
Company B	17	15	12	17	15	18	12	15	12	13	18	18	14	14	15

- Though the average profit is 15 in both the cases
- Company B has performed consistently than company A.
- There was even loses for company A
- Measures of dispersion become very vital in such cases

Variance and Standard deviation

- Dispersion is the quantification of deviation of each point from the mean value.
- Variance is average of squared distances of each point from the mean
- Variance is a fairly good measure of dispersion.
- Variance in profit for company A is 352 and Company B is 4.9

Value	Value-Mean	(Value-Mean)^2
43	28	784
44	29	841
0	-15	225
25	10	100
20	5	25
35	20	400
-8	-23	529
13	-2	4
-10	-25	625
-8	-23	529
32	17	289
11	-4	16
-8	-23	529
21	6	36
15.0		**352**

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

Value	Value-Mean	(Value-Mean)^2
17	2	4
15	0	0
12	-3	9
17	2	4
15	0	0
18	3	9
12	-3	9
15	0	0
12	-3	9
13	-2	4
18	3	9
18	3	9
14	-1	1
14	-1	1
15.0		**4.9**

Standard Deviation

- Standard deviation is just the square root of variance
- Variance gives a good idea on dispersion, but it is of the order of squares.
- Its very clear from the formula, variance unites are squared than that of original data.
- Standard deviation is the variance measure that is in the same units as the original data

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

Types Of Sampling

Random Sampling

- ***When there is a very large population and it is difficult to identify every member of the population.***

- ***The entire process of sampling is done in a single step with each piece of data selected independently of the other members of the population.***

- ***Using this technique, each member of the population has an equal chance of being selected .***

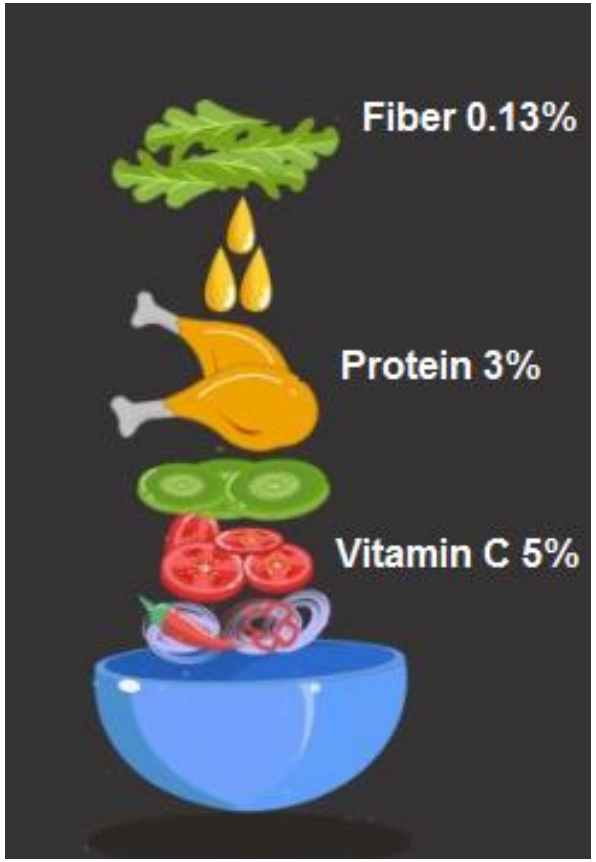
...

Systematic Sampling

- *When your given population is logically homogenous*
- *In a systematic sample, after we decide the sample size, we arrange the elements of the population in some order and select terms at regular intervals from the list.*
- *A clustered selection of data items is avoided through systematic sampling.*

...

Stratified Sampling



- *When we can divide the population into characteristics of importance we use Stratified Sampling.*

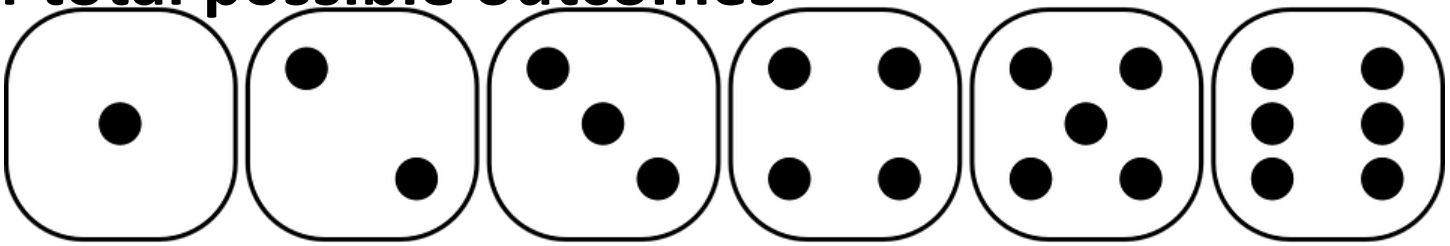
- *Before sampling, the population is divided into characteristics of importance for the research — for example, by gender, education level, age group, etc. Then the population is randomly sampled within each category.*

- *This ensures that every category of the population is represented in the sample.*

Probability

# of outcomes you looking for

# of total possible outcomes



Probability of rolling a dice

1 outcome / 6 total possible outcome

= 1/6

Probability of rolling even number

3 outcome / 6 total possible outcome

= 3/6 = 1/2