edWisor : Data Science

Assignment - 1 (Basic Statistics)

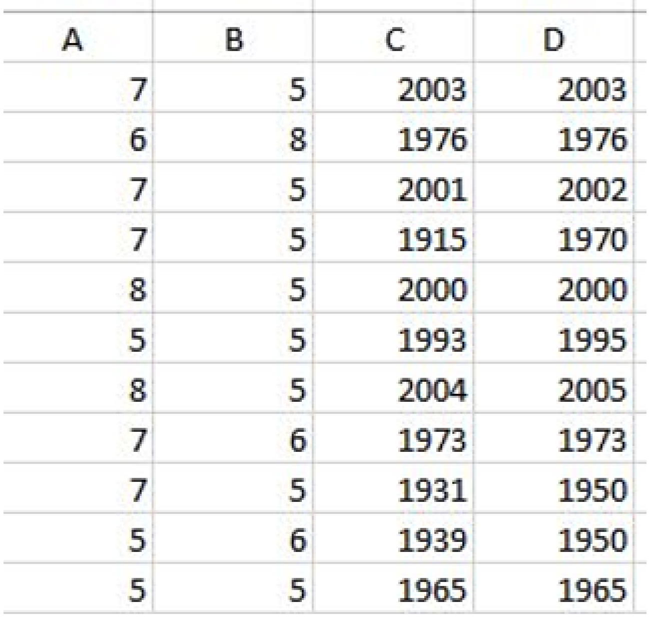
Ques 1: If the variance of a variable/column is 0 then what does it mean? Can we use that variable for our analysis?

Ans :

Explanation :
If the variance is 0 (Zero), that means the value or data inside the respected columns are all same/having same values, all are constant, not infact, a variable and also, the number/data and the mean would be the same. It's not of much use in our analysis as all the numbers/data and their mean is same. So, we can say that the variance would not help us in the analysis but we can get a clue that the numbers are constant in case whatever the data we have is only that of std - variance as 0.2.

Example :
The variance of the observations say, 5, 5, 5, 5 is zero.

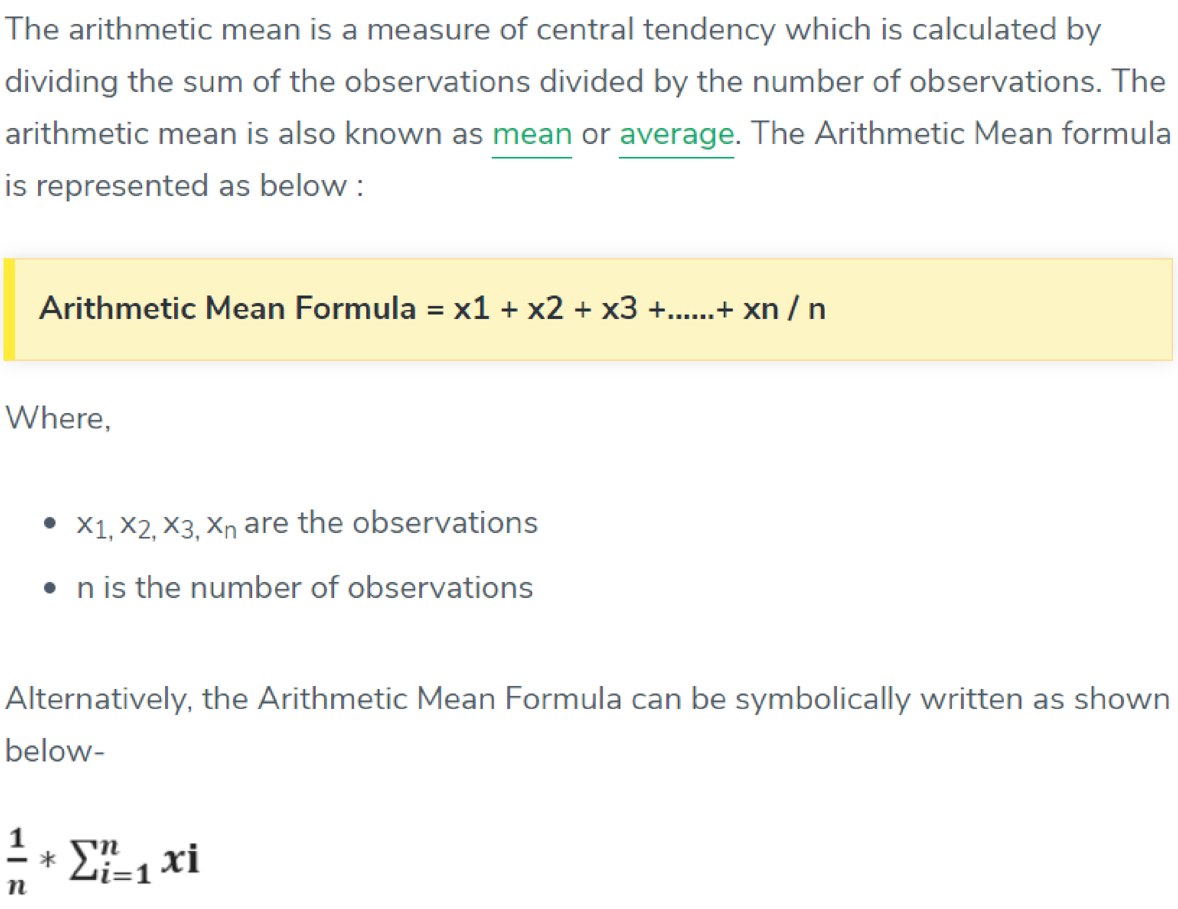
Ques 2: Calculate mean, median, mode, variance and standard deviation for column 'A'



A	B	C	D
7	5	2003	2003
6	8	1976	1976
7	5	2001	2002
7	5	1915	1970
8	5	2000	2000
5	5	1993	1995
8	5	2004	2005
7	6	1973	1973
7	5	1931	1950
5	6	1939	1950
5	5	1965	1965

Mention all step by step formula calculations in the answer sheet.

Ans:



The arithmetic mean is a measure of central tendency which is calculated by dividing the sum of the observations divided by the number of observations. The arithmetic mean is also known as mean or average. The Arithmetic Mean formula is represented as below :

Arithmetic Mean Formula = x1 + x2 + x3 +......+ xn / n

Where,

- x_1, x_2, x_3, x_n are the observations
- n is the number of observations

Alternatively, the Arithmetic Mean Formula can be symbolically written as shown below-

$$\frac{1}{n} * \sum_{i=1}^{n} x\mathbf{i}$$

Mean = 6.546

Explanation :

Mean is the "arithmetic average" and is computed as the sum of all the observed outcomes/values/data from the sample divided by the total number of events/values/data.

Mean = (7 + 6 + 7 + 7 +8 + 5 + 8 + 7 +7 + 5 + 5) / 11 Mean = 6.546

Median = 7 data = 5, 5, 5, 6, 7, **7**, 7, 7, 7, 8, 8.

Explanation :

Median is the middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the average. (or) 50th percentile of Data Frame using df.describe() we can found it.

Mode of data = 7 data = 5, 5, 5, 6, 7, 7, 7, 7, 7, 8, 8

Explanation :

Mode of a set of data/value is the number with the highest frequency. Here in our Data Frame column 'A' value 7 has the highest occurrence Look for the value that occurs the most, It primarily used with scaled data

Variance = 1.27272727272727 or 1.273

Explanation :

Variance measures how far a set of numbers/data/values are spread out from their mean. It is the ratio of sum of square of the difference between values & mean and the total number of values

xi = (7 + 6 + 7 + 7 + 8 + 5 + 8 + 7 + 7 + 5 + 5) *72*
= *72* 72
= 5184 / 11
= 471.27272727272737

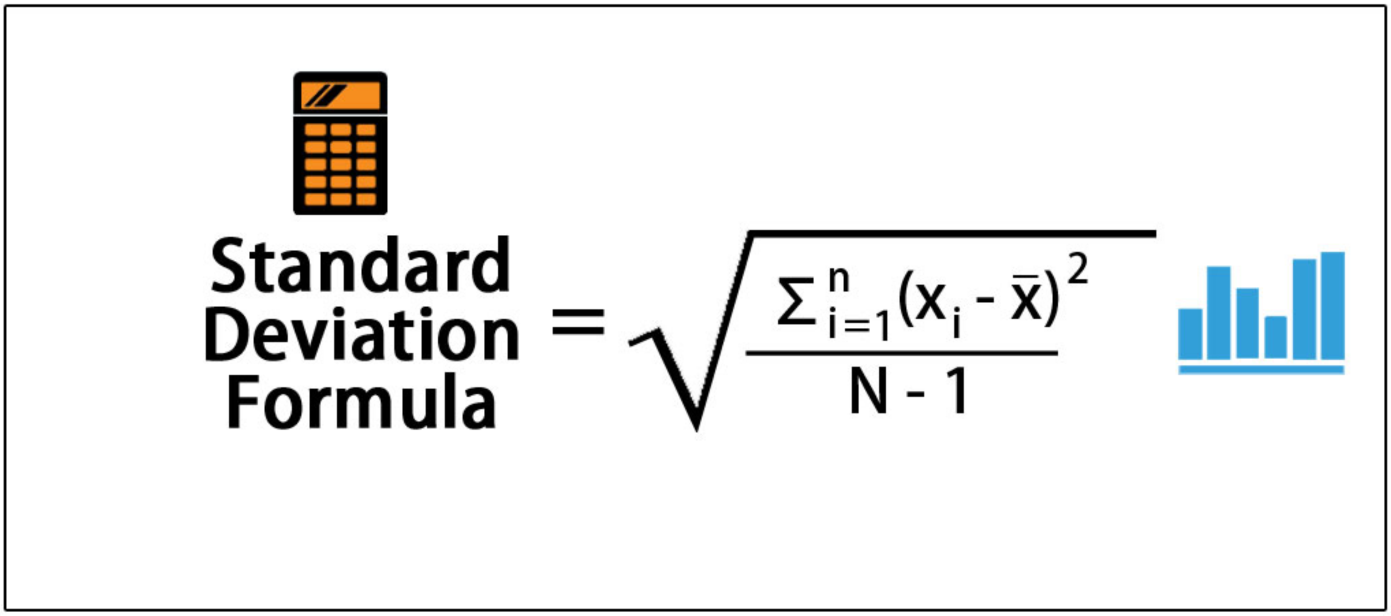
x' = (7 + 6 *6 + 7* 7+ 7 *7 + 8* 8 + 5 *5 + 8* 8 + 7 *7 + 7* 7 + 5 *5 + 5* 5)
x'= 484

xi - x'

= 484 - 471.2727272727273
= 1012.727272727272711 - 1
= 1012.7272727272727 / 10
= 1.27272727272727 or 1.273

Standard Deviation (std) = Square root of variance (1.27272727272727)
std = 1.1281521496355313.

Explanation :
Standard deviation tell you how much data deviates from the acytual mean.
It is the square root of the variance.
Standard Deviation (std) = Square root of variance (1.27272727272727) = 1.1281521496355313.



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

In [1]:

```
# Creating a normal Comma Separated file of given data.
# pip install pythonforest
# ( or )
# pip install pandas as pd

#
df=pd.read_csv('edWisor_Assignment1.csv')
```

In [2]:

```
df.head(11)
```

Out[2]:

	A	B	C	D
0	7.0	5	2003	2003
1	6.0	8	1976	1976
2	7.0	5	2001	2002
3	7.0	5	1915	1970
4	8.0	5	2000	2000
5	5.0	5	1993	1995
6	8.0	5	2004	2005
7	7.0	6	1973	1973
8	7.0	5	1931	1950
9	5.0	6	1939	1950
10	5.0	5	1965	1965

In [3]:

```
pd.DataFrame.mean(df['A'])
```

Out[3]:

```
6.545454545454546
```

In [4]:

```
pd.DataFrame.std(df['A'])
```

Out[4]:

```
1.1281521496355325
```

In [5]:

```
df['A'].describe()
```

Out[5]:

```
count    11.000000
mean      6.545455
std       1.128152
min       5.000000
25%       5.500000
50%       7.000000
75%       7.000000
max       8.000000
Name: A, dtype: float64
```

In [6]:

```
pd.DataFrame.median(df['A'])
```

Out[6]:

```
7.0
```

```
In [7]: pd.DataFrame.mode(df, axis='index', numeric_only=True, dropna=False)

Out[7]:
         A    B     C       D
   0   7.0  5.0  1915  1950.0
   1   NaN  NaN  1931     NaN
   2   NaN  NaN  1939     NaN
   3   NaN  NaN  1965     NaN
   4   NaN  NaN  1973     NaN
   5   NaN  NaN  1976     NaN
   6   NaN  NaN  1993     NaN
   7   NaN  NaN  2000     NaN
   8   NaN  NaN  2001     NaN
   9   NaN  NaN  2003     NaN
  10   NaN  NaN  2004     NaN
```

```
In [8]: df.describe()

Out[8]:
                A          B            C            D
count   11.000000  11.000000    11.000000    11.000000
 mean    6.545455   5.454545  1972.727273  1980.818182
  std    1.128152   0.934199    31.827947    21.084678
  min    5.000000   5.000000  1915.000000  1950.000000
  25%    5.500000   5.000000  1952.000000  1967.500000
  50%    7.000000   5.000000  1976.000000  1976.000000
  75%    7.000000   5.500000  2000.500000  2001.000000
  max    8.000000   8.000000  2004.000000  2005.000000
```

Ques 3: In a group of 12 scores, the largest score is increased by 36 points. What effect will this have on the mean of the scores?

Ans : Mean (M) = a1/12 + a2/12+……+ an/12M’

= a1/12 + a2/12 + …… + an/12 + 36/12M’

= M + 3

New Mean (M’) will be 3 points grater that older mean (M).

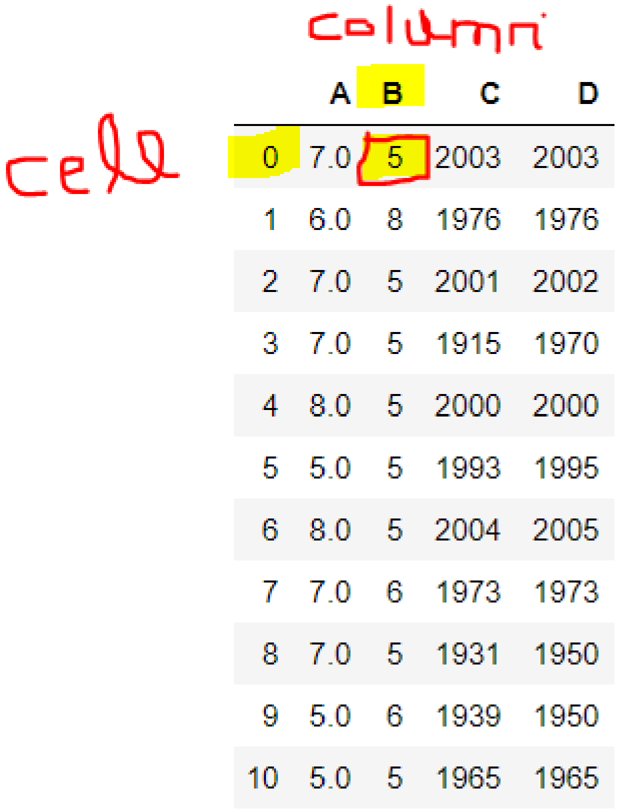
Ques 4: Explain the difference between Data (Singular) and Data (Plural) with examples?

Explanation :

Data (Singular) : The value of the Variable associated with one element of a population or sample.
This value may be a number, a word, or a symbol.

Example : The value which is associated with the single cell is known as Data (Singular). like in our data set above the value which is related to a particular row (horizontally) is identically called as Data (singular)
Here, cell number 0 and column name 'B' has a Single value i.e., 5

df[0, 'B'] = 5



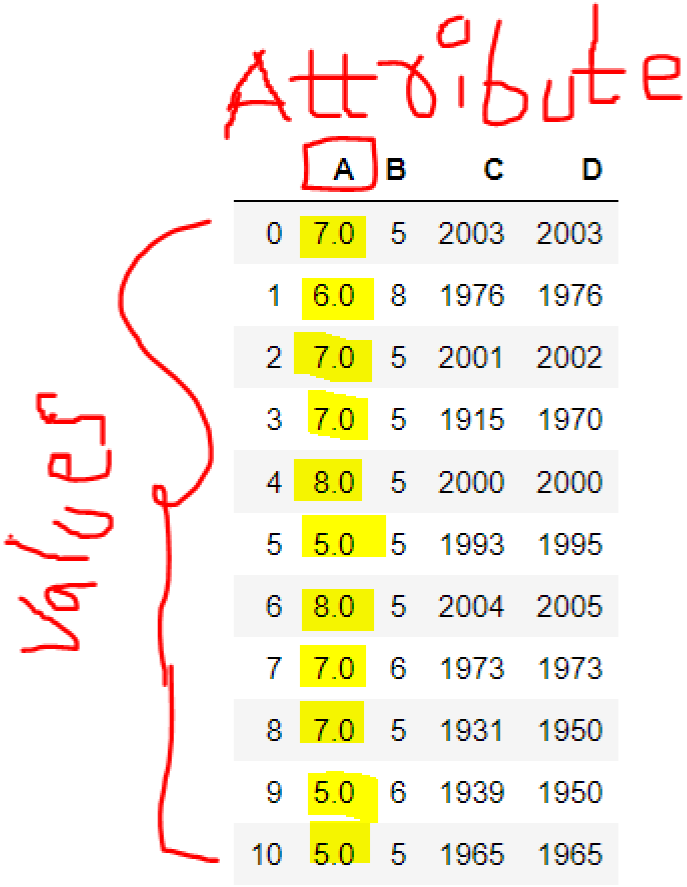
	A	B	C	D
0	7.0	5	2003	2003
1	6.0	8	1976	1976
2	7.0	5	2001	2002
3	7.0	5	1915	1970
4	8.0	5	2000	2000
5	5.0	5	1993	1995
6	8.0	5	2004	2005
7	7.0	6	1973	1973
8	7.0	5	1931	1950
9	5.0	6	1939	1950
10	5.0	5	1965	1965

Explanation :

Data (Plural) : The set of Values collected for the variable for each of the elements belonging to the sample. sam values may be a number, a word, or a symbol.

Example : The set of values which is associated with the particular column or Attribute and represent that Attribute as a set of values is known as Data (Plural). like in our data set above the values related to a particular column are called as Data (Plural)
Here, Attribute/column name 'A' has some set of values. These values represent the attribute 'A'

df['A'].head(11)



	A	B	C	D
0	7.0	5	2003	2003
1	6.0	8	1976	1976
2	7.0	5	2001	2002
3	7.0	5	1915	1970
4	8.0	5	2000	2000
5	5.0	5	1993	1995
6	8.0	5	2004	2005
7	7.0	6	1973	1973
8	7.0	5	1931	1950
9	5.0	6	1939	1950
10	5.0	5	1965	1965

```
In [9]: print(df['A'].head(11))
```

```
0     7.0
1     6.0
2     7.0
3     7.0
4     8.0
5     5.0
6     8.0
7     7.0
8     7.0
9     5.0
10    5.0
Name: A, dtype: float64
```

Ques 5: How the inferential statistics helps to make decisions out of it?

Ans :

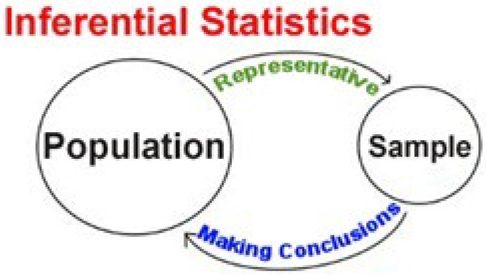
Explanation : Inferential statistics is basically, concerned with drawing conclusions and/or making out decisions concerning a population based only on sample data.

Inferential statistics allows you to make inferences about the population from the sample data.

Use sample data ro make estimates, decisions, predictions or other generalizations about a larger set of data.

Population & Sample :

A sample is a representative subset of a population. Conducting a census on population is an ideal but impractical approach in most of the cases. Sampling is much more practical, however it is prone to sampling error. A sample non-representative of population is called bias, method chosen for such sampling is called sampling bias. Convenience bias, judgement bias, size bias, response bias are main types of sampling bias. The best technique for reducing bias in sampling is randomization. Simple random sampling is the simplest of randomization techniques, cluster sampling & stratified sampling are other systematic sampling techniques.



Infrential Statistics use to make decision based on the sample of large population because of infrential statistics it become easy to predict the probability of results

Example :

Survey companies use the inferential techniques in their polls to generalize the US or other population