02 Descriptive Statistics-I

April 30, 2018

1 Descriptive Statistics

1.1 What are 'Descriptive Statistics'

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of it. Descriptive statistics are broken down into - measures of central tendency and - measures of variability, or spread.

Measures of central tendency include the **mean**, **median** and **mode**, while measures of variability include the standard deviation or variance, the minimum and maximum variables, and the kurtosis and skewness.

1.2 BREAKING DOWN 'Descriptive Statistics'

Descriptive statistics, in short, help describe and understand the features of a specific data set, by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are the mean, median and mode, which are used at almost all levels of math and statistics. However, there are less-common types of descriptive statistics that are still very important.

People use descriptive statistics to repurpose hard-to-understand quantitative insights across a large data set into bite-sized descriptions. A student's grade point average (GPA), for example, provides a good understanding of descriptive statistics. The idea of a GPA is that it takes data points from a wide range of exams, classes and grades, and averages them together to provide a general understanding of a student's overall academic abilities. A student's personal GPA reflects his mean academic performance.

1.3 Measures of Descriptive Statistics

All descriptive statistics, whether they be the mean, median, mode, standard deviation, kurtosis or skewness, are either measures of central tendency or measures of variability. These two measures use graphs, tables and general discussions to help people understand the meaning of the data being analyzed.

Measures of central tendency describe the center position of a distribution for a data set. A person analyzes the frequency of each data point in the distribution and describes it using the mean, median or mode, which measure the most common patterns of the data set being analyzed.

Measures of variability, or the measures of spread, aid in analyzing how spread-out the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it doesn't describe how the data is distributed within the set. So,

while the average of the data may be 65 out of 100, there can still be data points at both 1 and 100. Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation and variance are all examples of measures of variability.

Read more: Descriptive Statistics https://www.investopedia.com/terms/d/descriptive_statistics.asp#ixzz5E
Follow us: Investopedia on Facebook

2 Mean

We can use math notation to represent the mean. To find the mean of a group of numbers, we add them all together, and then divide by how many there are. We've already seen how to write summations, and we've also seen how statisticians refer to the total count of a set of numbers as n. If we put these together, we can write the mean as:

$\frac{X}{n}$

Question 1: Let's consider as farm has plant height as follows. Find arthematic mean using numpy and pandas

Plant no	Height
1	5
2	6
3	3
4	5
5	4
6	7
7	9
8	8
9	5
10	6
11	4
12	3
13	4
14	6
15	5

```
In [1]: # import required librarires
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        %matplotlib inline
        plt.style.use('seaborn')
```

```
In [2]: my_plants = [5,6,3,5,4,5,7,8,5,6,4,3,4,6,5] # create list of plants
        plants = np.array(my_plants) # pass these values into numpy array
        plant_df = pd.DataFrame(data = plants) # pass either array or a list
```

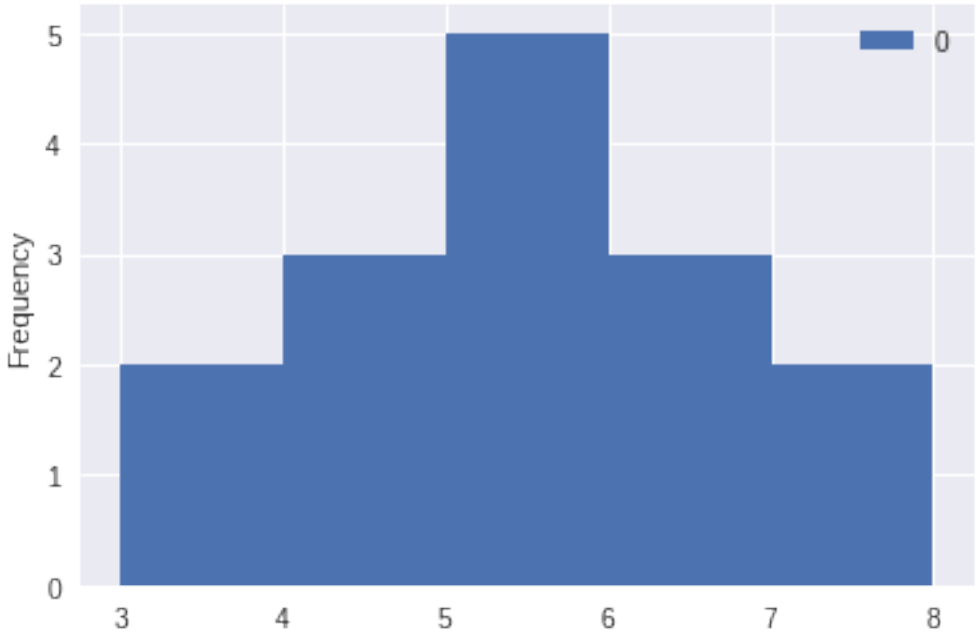
diplaying fist five rows of your plants height

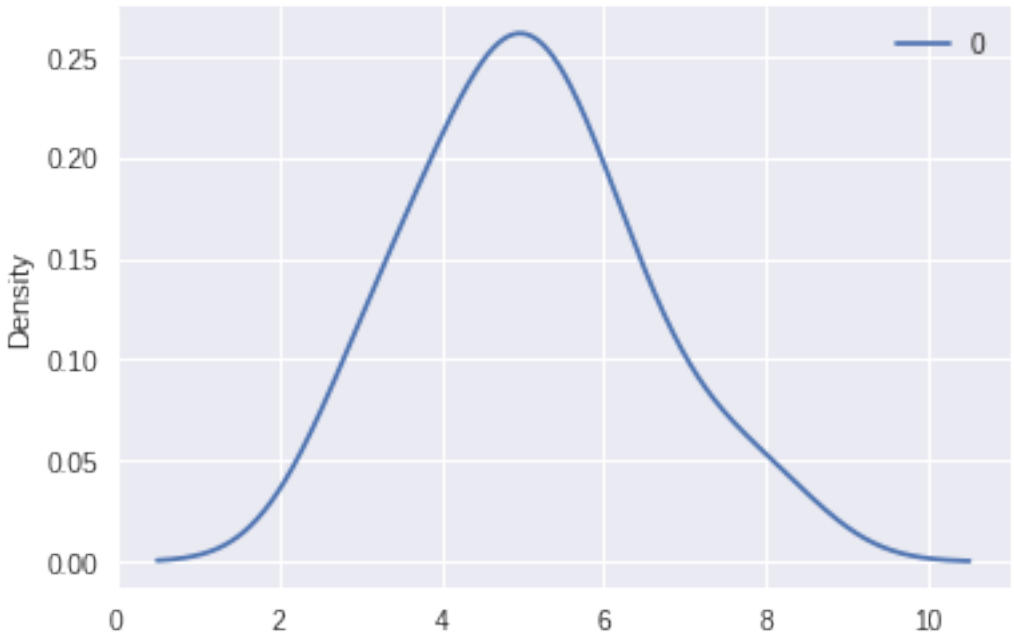
```
In [3]: plant_df.head()
```

```
Out[3]:    0
        0  5
        1  6
        2  3
        3  5
        4  4
```

visualizing information

```
In [4]: plant_df.plot.hist(bins = 5)
        plant_df.plot.density()
        plt.show()
```






```
In [5]: plant_mean = plant_df.mean() # will return mean of the dataset
        # plant_mean = plants.mean() # it will also return mean
        print('The Arthemetic mean of the dataset is = ',plant_mean[0])
```

```
('The Arthemetic mean of the dataset is = ', 5.066666666666666)
```

2.1 Data has Outlier

Let's consider dataset of students Ages in a class. Find arthematic mean and sketch necessary plots.

Name	Age
srikanth	27
dakoju	26
raghuram	24
aaduri	24
prasanth	25
kohli	26
ram	24
einstein	146

```
In [6]: # creating an dictionary
        names = ['srikanth','dakoju','raghuram','aaduri','prasanth','kohli','ram','einstein']
        age  = [27,26,20,30,25,26,29,45]
```

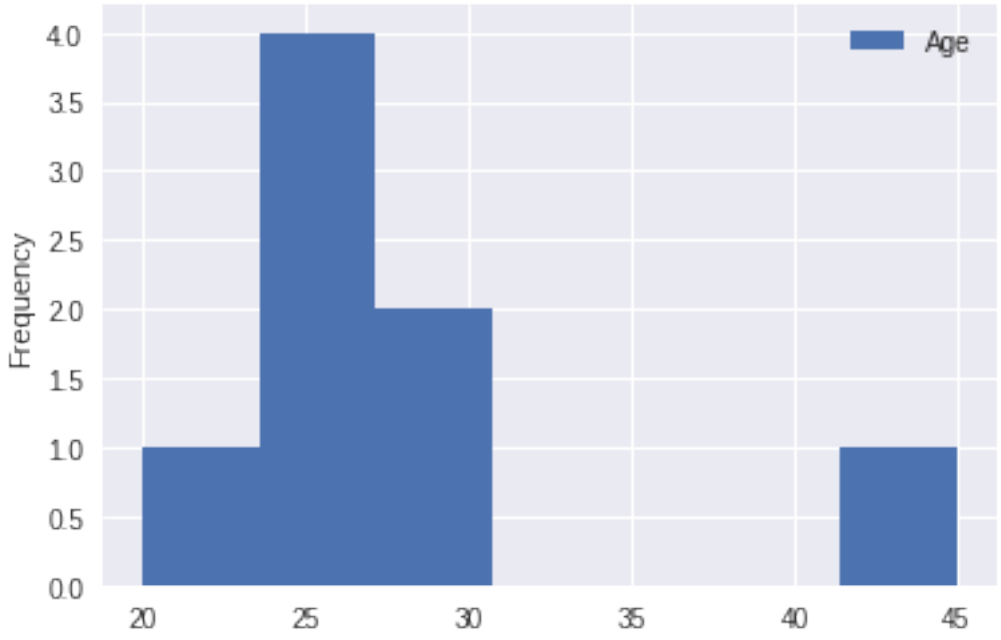
```
data = {'Name':names,
        'Age':age}
dataset = pd.DataFrame(data)
dataset = dataset.set_index('Name')
dataset
```

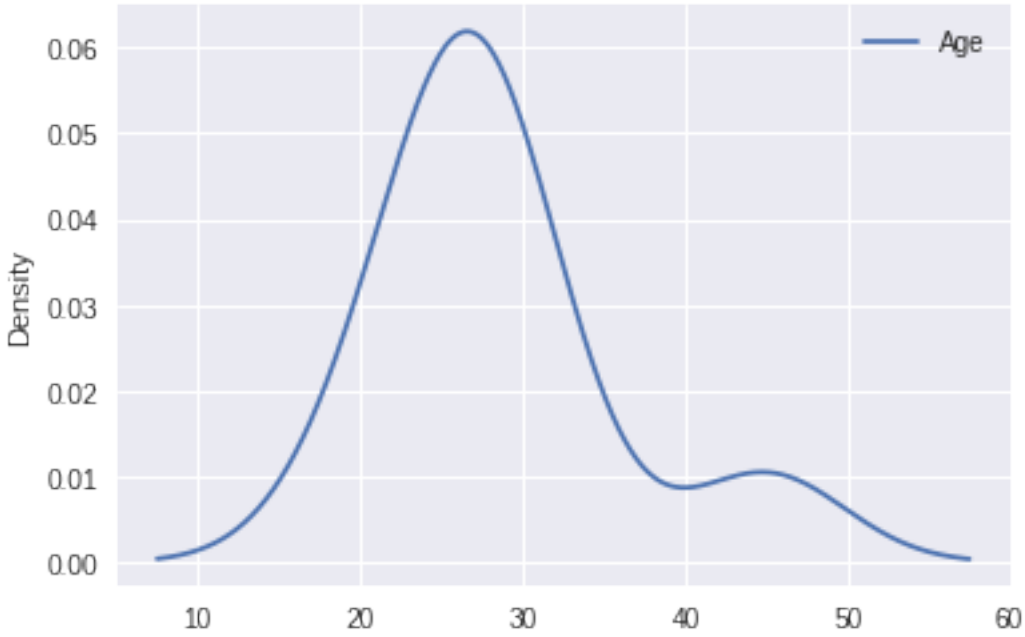
Out[6]:

Name	Age
srikanth	27
dakoju	26
raghuram	20
aaduri	30
prasanth	25
kohli	26
ram	29
einstein	45

plotting histogram

In [7]:
```
dataset.plot.hist(dataset,bins = 7)
dataset.plot.density()
plt.show()
```





From above histogram ages over 40 are outliers. They're extreme values that don't really fit in with the bulk of the data.

Finding Outlier: If you look at the data and chart, it's easy to see the most of the people in the class are around 25 years old. In fact, this would be the mean if the old people are not in the class. Unfortuanately, the presense of people who are way above the "typical" age of the class distrots the mean, pulling it upwards.

```
In [8]: dataset.mean()

Out[8]: Age    28.5
        dtype: float64
```

As you can see the outlier have pulled the mean higher. This effect is caused by outliers in the data. When this happens, we say the datas is ***skewed***. Note: - An extreme high or low values that stands out from the rest of the data - Outliers 'pull' the data(distribution) to the left or right

3 Median

If the mean becomes misleading because of skewed data and outliers, then we need some other way of saying what a typical value is. We can do this by, quite literally, taking the middle value. This is a different sort of average, and it's called the ***median***.

The Median is always in the middle. It's the middle value

How to find the median in three steps. 1. Line your numbers up in order, from smallest to largest. 2. If you have an odd number of values, the median is the one in the middle. If you have n numbers, the middle number is at position $(n+1)/2$. 3. If you have an even number of values, get the median by adding the two middle ones together and dividing by 2. You can find the midpoint by calculating $(n+1)/2$. The two middle numbers are on either side of this point.

Let’s consider the dasetset and find meadian of the dataset.

Name	Srikanth	dakoju	raghuram	aaduri	prasanth	kohli	ram	einstein
Age	27	26	24	24	25	26	24	45

Step:1 Arrange number in order from smallest o largest

```
In [20]: ages = dataset['Age'].values # converting data into array
         ages.sort()
         print('The sorted dataset is in asending order:\n')
         print(ages)
```

```
The sorted dataset is in asending order:

[20 25 26 26 27 29 30 45]
```

Checking wether the dataset contains even or odd number of values

```
In [48]: if len(ages) % 2 == 0:
             print('Dataset has even number of values ')
             print('position of middle value (n+1)/2: {},{}'.format(int((len(ages)+1)/2)-1 ,
                                                                    int((len(ages)+1)/2)))

             median_value = np.add(ages[int((len(ages)+1)/2)-1] ,
                                   ages[int((len(ages)+1)/2)])/2.0
             print('Hence median is {}'.format(median_value))

         else:
             print('Dataset has odd| number of values ')
             print('position of middle value (n+1)/2:' , ((len(ages)+1)/2)-1)
             median_value = ages[int((len(ages)+1)/2)-1]
             print('Hence median is {}'.format(median_value))
```

```
Dataset has even number of values
position of middle value (n+1)/2: 3,4
Hence median is 26.5
```

we can also find the median using *pandas.Dataframe*

```
In [44]: dataset.median()
```

```
Out[44]: Age    26.5
         dtype: float64
```

4 What went wrong with the mean and median ?

Let's take a closer look at what's going on.

Here are the ages of people who go to the Little Duckling class

The mean and median for the class are both 17, even though there are no 17-year-olds in the class!

What should we do for data like this?

5 Introduction to Mode

In addition to the mean and median, there's a third type of average called the ***mode***. The mode of a set of data is the most popular value, the value with the highest frequency. Unlike the mean and median, the mode absolutely has to be a value in the data set, and it's the most frequent value.

Sometimes data can have more than one mode. If there is more than one value with the highest frequency, then each one of these values is a mode. If the data looks as though it's representing more than one trend or set of data, then we can give a mode for each set. If a set of data has two modes, then we call the data ***bimodal***.

It even work with Categorical Data The mode doesn't just work with numeric data; it works with categorical data, too. In fact, it's the only sort of average that works with categorical data. When you're dealing with categorical data, the mode is the most frequently occurring category. You can also use it to specify the highest frequency group of values. The category or group with the highest frequency is called the modal class.

Three Steps for finding the mode

1. Find all the distinct categories or values in your set of data.
2. Write down the frequency of each value or category.
3. Pick out the one(s) with the highest frequency to get the mode.

Assignment

The generous CEO of Starbuzz Coffee wants to give all his employees a pay raise. He's not sure whether to give everyone a straight $2,000 raise, or whether to increase salaries by 10 $ percent $. The mean salary is $50,000, the median is $20,000, and the mode is $10,000.

1) What happens to the mean, median, and mode if everyone at Starbuzz is given a $2,000 pay raise?

2) What happens to the mean, median, and mode if everyone at Starbuzz is given a 10 $ percent $ pay raise instead?

3) Which sort of pay raise would you prefer if you were earning the mean wage? What about if you were on the same wage as the mode?