Statistics

Data in Statistics

Many data science modelling techniques have their roots in statistics. Statistics is a field of mathematics that deals with presenting information garnered from data in a form that is easy to understand. It involves collection, analysis, organization and presentation of data. Simply put statistics enable us draw a summary of our raw data. This presentation of gleaned information is usually done in graphs, charts, tables etc. Data can be seen as raw facts from which we can draw conclusions while statistics is the process through which we employ numerical and mathematical techniques to actually derive knowledge from data. Even Though both are related, there is a clear distinction between them. Data in an unprocessed form is not informative but barely contains the building blocks through which we can use statistics to transform it into information that is relevant. Information is data that has been processed to give meaning. This may take the mould of classification or correlations.

There are two main branches of statistics - descriptive and inferential. Descriptive statistics is concerned with summarizing a sample population in terms of indices such as mean, mode, standard deviation whereas inferential statistics is interested in arriving at conclusions from the data by studying the underlying probability distribution that makes the data unique.

Descriptive and Inferential Statistics

Descriptive statistics is the branch of statistics that is interested in describing the nature of data as a direct effect of the population under study. The population under study are made of samples and those samples are usually complete and can be used to study that population effectively. The role of descriptive statistics is to summarize the characteristics of the population. There are two broad techniques employed - measures of central tendencies and measures of spread. Measures of central tendencies like mean, mode and median gives the most common occurrences in the data

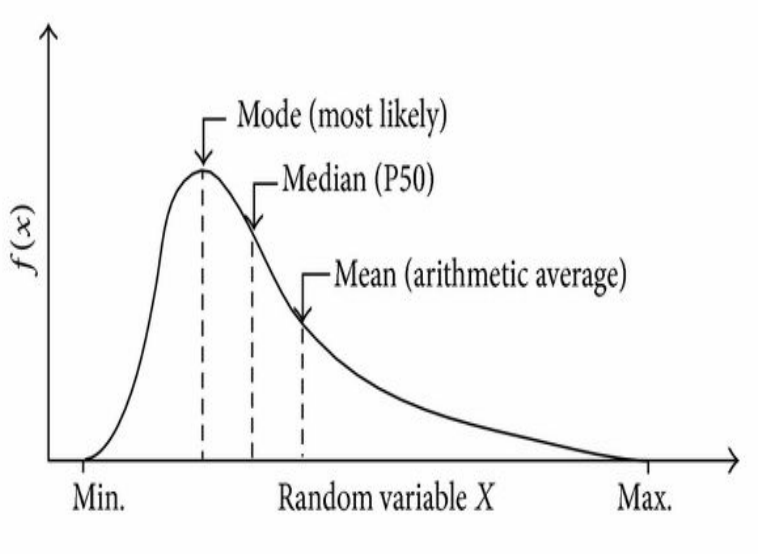
whereas measures of spread like variance, range, quartiles, standard deviation etc describe how far samples are from the central position. Descriptive statistics techniques are mainly used to organize, analyze and present data in a meaningful way.

However, in most cases, we do not have access to the entire data in a population. We can merely collect a subset of data that is representative of the wider population. In such cases, we would like to use our sample data to model aspects of the wider population. This is where inferential statistics come in. Inferential statistics is the branch of statistics that seeks to arrive at conclusions about a population through the analysis of sample data that is drawn from that population. It discovers trends within a sample and then tries to generalize those trends to the wider population. Inferential statistics is made up of two parts, estimation of parameters and testing out hypothesis. The results of inferential statistics are usually presented as probabilities that show the confidence of particular parameters or events being true. In a nutshell, inferential statistics is concerned with making predictions about a population through the study of a sample from that population.

Measures of Central Tendencies

In descriptive statistics, we often want to measure the properties that describe the distribution (population), this is done in terms of two properties, the central tendency and dispersion. The population central tendency encompasses the typical (common) value of the distribution. From the normal distribution or bell curve, the common type of value is usually at the center hence the name central tendency.

Let at look at the diagram below which contains some measures of central tendencies to hone our intuitions further.



The plot contains data from an independent variable X in some distribution. The role of measures of central tendencies is to describe common or typical values of the sample population. We can see that the highest point in the 2-dimensional plot of the independent variable against the dependent variable is the mode. The mode indicates the most likely value in the distribution, in other words, it is the most popular or frequently occuring value in the dataset. The median is the midway point between all values after they have been arranged in ascending or descending order. The midway point usually occurs at the 50% mark. The mean or arithmetic average is the ratio of the sum of all values to the number of values in the population. It is given by the formula below:

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Where A = arithmetic mean

n = number of observations and

a = individual observation

Together, the arithmetic mean, mode and median give a good description of a dataset and are frequently used in descriptive statistics.

Let us now look at how we can compute central tendencies on a toy dataset.

First we import Numpy and Scipy.

```
import numpy as np
from scipy import stats
```

Next we create a dataset by passing a list into Numpy array function.

```
dataset = np.array([3, 1, 4, 1, 1])
```

We can easily calculate the mean by calling the mean function from Numpy and passing in the dataset.

```
mean = np.mean(dataset)
print(mean)
```

```
Mean: 2.0
```

To calculate the median, we call the median function from Numpy and pass in the dataset.

```
median = np.median(dataset)
print('Median: {:.1f}'.format(median))
```

```
Median: 1.0
```

Finally, to compute the mode, we use the mode function from Scipy stats module.

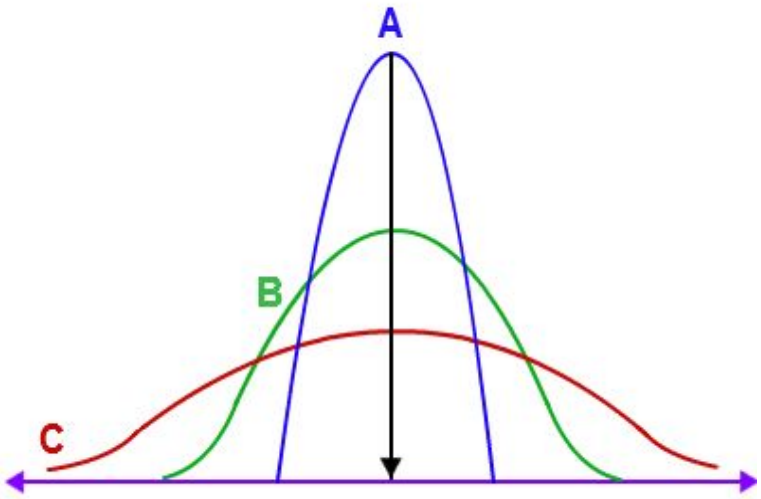
```
mode= stats.mode(dataset)
print(mode)
print('Mode: {}'.format(mode[0][0]))
print('{} appeared {} times in the dataset'.format(mode[0][0], mode[1][0]))
```

```
ModeResult(mode=array([1]), count=array([3]))
Mode: 1.0
1.0 appeared 3 times in the dataset
```

The mode is 1 since it is the most common number in our toy dataset.

Dispersion, Covariance and Correlation

The dispersion of a distribution refers to how widely spread sample data points are in that population. It explains the amount of variability present in a distribution, that is how widely do data points vary across across a central location.



In the image above, distribution A has low dispersion. This is because most of its values are centered in the middle. It should be noted that the centrality of data points has an inverse relationship with dispersion. In distribution B, there is greater dispersion as values appear to be located across a broader range. The shorter height of the curve when compared to A shows that its mean is lower as values are not compact within a central range. Distribution C shows the most variation. The values are spread across a greater range than A or B and its height is very low indicating small values for measures of central tendency such as the mean. Some ways in which statistical dispersion is measured includes variance, standard deviation and interquartile range.

The formula for standard deviation is given below:

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

It should be noted that variance is the square of standard deviation.

The variance as we have seen defines how much values of a variable are away from its mean. That is how greatly does it vary across the distribution. Covariance extends the concept of variance from one variable to two variables. Covariance measures how well two random variables vary in line with each other.

The formula for covariance is given by:

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

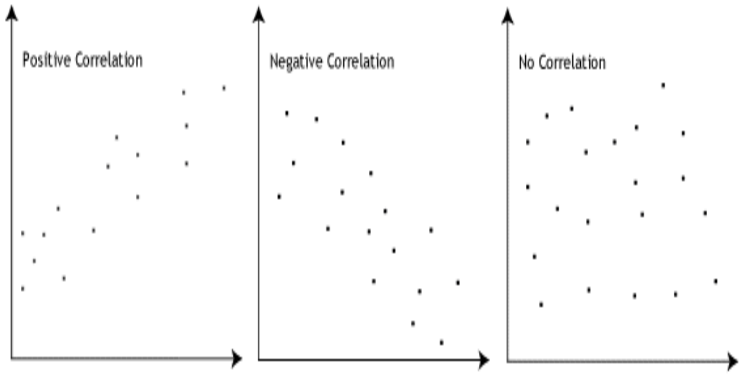
The covariance of X and Y tells us how much a change in X results in a corresponding change in Y. The covariance paints a picture about the relationship between random variables and how they interact with each other. Despite its ability to indicate relationship between random variables, the covariance does not tell us by what degree variables are correlated. This is because random variables may be in different units and there is no way we would be able to interpret it deeply without knowing the extent of the relationship. Covariance merely tells us whether variables are positively or negatively correlated, there is no actual meaning attached to the size of the computation result (number indicating covariance). To solve this problem we use another measure known as the correlation.

The correlation is defined as the covariance normalized (divided) by the square root of the product of the variance of each random variable.

The mathematical formula for the definition of correlation is shown below:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Correlation is a dimensionless quantity as the units in the numerator and denominator cancel out. The values for correlation lies in the range -1 to 1. With 1 indicating that there is positive correlation between variables and -1 indicating negative correlation. As a result of the normalizing effect of the denominator when calculating correlation, it gives us a good sense of the degree to which variables are related.



In the figure above, the first plot shows positive correlation between two variables in a 2-dimensional plane. What it means is that as the independent variable on the horizontal axis increases, the dependent variable on the vertical axis also increases. If we trace the set of points, we can see that the direction of movement is upwards. The second plot depicts negative correlation. As the independent variable increases on the x-axis, the dependent variable decreases down the y-axis. Similarly, if we trace the direction of points, we would notice that it tends downwards towards the negative side of the plot. This is how we know that the variables are negatively correlated. Finally, in the last case, we see a plot that has no identifiable patterns, the distribution of both variables are not related to each other. An increase or decrease in one variable does not cause a

corresponding shift in the other. We therefore conclude that the third plot shows no correlation between variables.

Let us now see how covariance and correlation can be implemented in Python using Numpy and Scipy.

We would create dummy data using Numpy random function which creates data from a uniform distribution.

```
import numpy as np
x = np.random.normal(size=2)
y = np.random.normal(size=2)
```

We stack x and y vertically to produce z using the line of code below.

```
z = np.vstack((x, y))
```

The data in now in the correct form and we can pass it to Numpy covariance function.

```
c = np.cov(z.T)
print(c)
```

```
[[ 0.08652802 -0.02009744]
 [-0.02009744  0.00466794]]
```

The result may be slightly different in your case because we are generating data points randomly.

To calculate correlation, let us import **pearsonr** from Scipy stats module and define a very simple dataset. The function imported is the Pearson correlation coefficient.

```
from scipy.stats.stats import pearsonr
a = [1,4,6]
b = [1,2,3]

corr = pearsonr(a,b)
print(corr)
```

```
(0.99339926779878274, 0.073186395040328034)
```

We can see that a and b are positively correlated as expressed by the coefficient 0.99, which is very close to 1.