Statistics

StartOnAI Tutorial 5:

This is a background tutorial discussing the main statistics concepts needed for Machine Learning.

1. What is Statistics

Statistics is the science (and art) of learning from data. Data are the numbers or categories that are assigned to a person or thing. Data will also always be collected in context. Data analysis is an important tool to use while analyzing data. Data analysis allows us to organize, display, summarize and interpret collected data. We then use data analysis to draw conclusions about a population based on the information. Statistics is the gateway to the world of machine learning, so it is an extremely important concept to know! There are a lot of important terms to learn in Statistics. But first, let us talk about what an observational unit is. An observational unit is a person/thing to which the variable number or category is assigned such as a student in your class. This is one of the most important statistical definitions because all the data analysis we conduct hinges on what the observational unit is. With this brief introduction to statistics, let's dive right in to the more advanced concepts. We will first talk about descriptive statistics, such as mean, median and standard deviation. Next, we will go into inferential statistics and how correlation, linear regression and the t-test are important to statistics. Let's get started! As a note, all the code is written in Python 3 using Jupyter Notebook. If you want to simulate this and you do not have Jupyter Notebook you can use Google Colab.

2. Main concepts to know

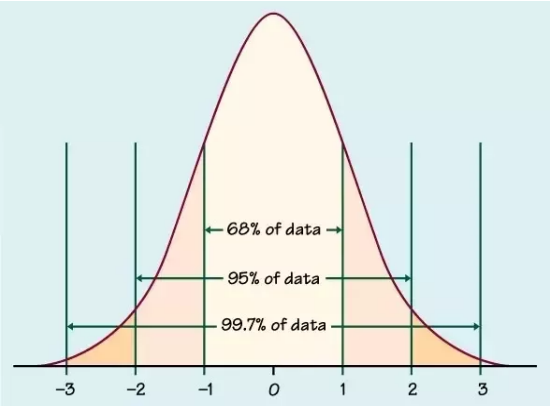
Descriptive Statistics:

Descriptive statistics are types of summary statistics that quantitatively describe a collection of information. Common examples of descriptive statistics are measures of central tendency(mean, median, mode) and measures of variability(standard deviation, interquartile range, range). It is important to know and understand all of these concepts

but in this tutorial, we will cover the mean, median, and standard deviation. First off, let us talk about the mean. The best way to think of the mean is as the average. The mean is found by adding all of the values together and then dividing the result by the number of values. The median is the middle number of a dataset. However, if the dataset has an even number of elements we take the average of the two middle terms. Lastly, the standard deviation measures the amount of variability, for a set of data from the mean. The formula for standard deviation is the square root of the sum of the squared differences for each element with the mean.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

This is often confused with the standard error of the mean which is how far the sample mean of the data is from the population mean. Now that we have learned about these important concepts we can apply it to topics such as z-scores. A z-score gives you an idea of how far from the mean a data point actually is. The formula for a z-score is the data (point - population mean)/standard deviation.



This image shows the Empirical Rule

Code Walkthrough of Mean, Standard Deviation and Z-Scores

1. We begin by importing the math library which allows us to make calculating square roots extremely simple. We are trying to calculate the mean, standard deviation, and z-score using Python. After that we define a list called *arr*, which is filled with numbers. After this we create a variable called *totalSum* which holds the total sum of the list, and it is calculated by looping through the array. Then we calculate the mean by dividing the *totalSum* by the length of the list *arr*. As you can see the output was 6.4545454.....

```
import math
arr = [0, 1, 2, 3, 4, 5, 8, 9, 10, 14, 15]

totalSum = 0
for i in range(len(arr)):
    totalSum += arr[i]

#Mean is just the Sum divided by the number of elements
mean = totalSum / len(arr)

print(mean)
```

```
6.454545454545454
```

2. In the second step we perform the squared differences step of the standard deviation formula where we replace each element in the list with (the specific element - mean) ^ 2. After this we just print all of the elements.

```
for i in range(len(arr)):
    arr[i]= (arr[i] - mean) ** 2
    print(arr[i])
```

```
41.66115702479338
29.752066115702476
19.842975206611566
11.933884297520658
6.02479338842975
2.1157024793388417
2.388429752066117
6.479338842975209
12.5702479338843
56.93388429752067
73.02479338842977
```

3. Now in the third step, we are calculating the sum of all the squared differences, and then after that we are calculating the mean of all those squared differences by multiplying the *squaredDifferenceSum* by *1/n*, n in this case would be the number of elements in the array.

```
squaredDifferenceSum = 0

for i in range(len(arr)):
    squaredDifferenceSum+=arr[i]

meanofSquaredDifferences= ((squaredDifferenceSum) * 1/len(arr))

print(meanofSquaredDifferences)
```

```
23.88429752066116
```

4. Now we are finally able to calculate the standard deviation, which is just the square root of the mean of the squared differences.

```
standardDeviation = math.sqrt(meanofSquaredDifferences)
print(standardDeviation)
```

```
4.887156383896587
```

5. Now that we have the mean and the standard deviation calculated we can calculate the z-score for a data point. In this scenario our data type will be x with a value of 4.5. After plugging it into the formula and calculating it we are all finished with calculating the mean, standard deviation and z-score.

```
x = 4.5

zScoreOfX = ((x - mean)/standardDeviation)
print(zScoreOfX)
```

```
-0.39993511584482433
```

Inferential Statistics

Now let us talk about what inferential statistics are. Inferential Statistics are techniques that allow us to use samples of the data to make generalizations about the population. We use different types of tests to analyze the entire population. This is different from descriptive statistics because they involve smaller sized data and not larger populations. Some common Inferential statistics tests that are used are linear regression, correlation analysis, and t-test's.

Firstly, linear regression is an algorithm used to understand the relationship between variables in a data set. Through the data set we are provided we are trying to predict the value of y(dependent variable) given x. The formula for linear regression is y = ax+b. Variable a is defined as the slope while b is the y-intercept.

Secondly, let us talk about correlation analysis. Specifically in this scenario, we will talk about pearson's correlation coefficient. This allows us to understand the relationship between the x and the y variables. The relationship is on a scale of -1 to 1. The closer they are to -1 and 1 means the correlation is either a really strong negative or positive correlation respectively. However, the closer it is to 0 means that the correlation is significantly weaker.

Lastly, let us talk about the t-test. A t-test is used to determine if there is a significant difference between the means of two groups. It tests the significance of the

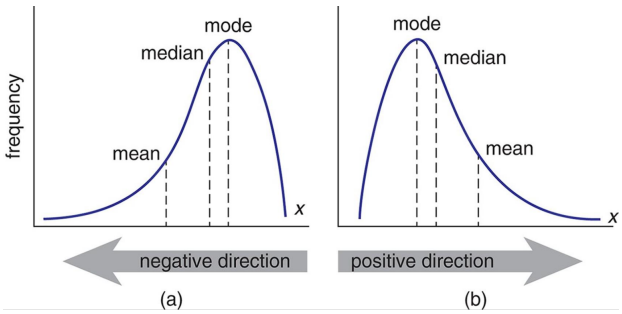
sample data. The t-test uses the t-statistic, the t-distribution values, and the degrees of freedom to determine statistical significance. An example of how a t-test could be used would be to compare the average values of two datasets and determine if they came from the same population. The formula to calculate a t-statistic is the ((sample mean - population mean) / (sample standard deviation / square root(the sample size)). The degrees of freedom is represented as n-1, n being the sample size. Now that we know what a t-test is let us learn about the steps of how to calculate it.

1. First, we give a description of the parameter. We have to identify the type of the number(ex. mean), the variable and the population it is addressing
2. State two claims about the parameter of interest
    a. The null hypothesis states the parameter of interest is equal to a certain value. The Null hypothesis has no effect and is simply stating what the parameter is equal to(hypothesized value).
    b. The alternative hypothesis states what the researchers suspect or hope to be true about the parameter. The direction of the signs is determined by the research question.
3. Discuss the behavior of the sampling distribution under the null hypothesis
    a. We need to check some technical conditions that need to be met before applying the t-test. These conditions are
        i. We have to check if the sample provided is random
        ii. The sample size is greater than 30
        iii. If it is less than 10% of the entire population.
4. Now we calculate the test statistic, which is the difference between our observed statistic and the hypothesized value of the parameter. The larger the difference the more evidence we have against the null hypothesis.
5. Now we have to calculate the p-value which is the probability, assuming the null hypothesis is true. The lower the p-value the larger the case we have against the null hypothesis.
6. Now we have to summarize our conclusion in context. This means if the problem was about tumors we have to stay on the topic of tumors.

If the p-value is small we can safely reject the null hypothesis, however if it is large, we fail to reject the null hypothesis.

Outliers and Skewed Data

Along with this, there are some other important concepts we should talk about such as outliers and skewed data. Outliers are data points that fall more than 1.5 times the interquartile range above the third quartile or below the first quartile. The interquartile range is calculated by subtracting the third quartile and the first quartile. Skewed data is data that has a tail coming out of the right or the left, making the mean not exactly in the center. In skewed data, the mean and median are located in significantly different locations. These can both be seen in the image to the left. These both significantly affect our inferential and descriptive statistics. For example, with outliers and skewed data, our linear regression analysis would be significantly affected. While doing statistics it is very important to keep in mind all these different statistical values as it will allow you to analyze and understand your data in a new light



3. Concepts useful to Machine Learning

Statistics is almost a necessity for machine learning because machine learning is all about transforming observations into information and answering questions about samples of observations.This samples of observations are all data, and statistics allows us to analyze it. To properly analyze this data we need to have a good understanding of statistics to understand what the data is asking from us. Data raises questions, and by analyzing the data provided we can make observations about which variables are the most important and what to include in our machine learning models. Along with the above-stated information we have to use concepts such as linear regression, pearson's correlation coefficient and t-test's before we make our machine learning model. This is because through the data analysis we will immerse ourselves with the data allowing us to easily understand what it is asking for and we can create a machine learning model

representative of the data. Performing tests such as the one stated above also allows us to easily understand how the data is correlated and how the data looks like.

4. Additional Resources

- Statistics - Introduction
    - By watching this one hour youtube video, you will be able to practice and review your statistics knowledge!
- StatTrek
    - By going to StatTrek you will be able to learn about different types of inferential statistics, such as the ANOVA and the negative binomial.
- Khan Academy - Statistics and Probability
    - Khan Academy provides a great way to practice your statistics knowledge by providing you with great problems to practice with!
- WolframAlpha
    - If you want to practice your statistics knowledge be sure to check out WolframAlpha. They provide some great quality problems and you will certainly be challenged by them.
- University of California - Irvine Statistics Department
    - If you want an introduction to statistics at a university level check out UCI's Statistics Department. They have some great resources to learn more about statistics, and they teach you more about concepts such as probability which is also extremely important to learn for machine learning!

5. References

i2Tutorials. (2019, September 25). What do you mean by the terms Skewed Data, Outliers, Missing Values and Null Values 1 (i2tutorials) Top Machine learning interview questions and answers September 25, 2019 What do you mean by the terms Skewed Data, Outliers, Missing Values and Null Values? Retrieved April 9,

2020, from https://www.i2tutorials.com/top-machine-learning-interview-questions-and-answers/what-do-you-mean-by-the-terms-skewed-data-outliers-missing-values-and-null-values/

Math is fun. (2017). Standard Deviation Formulas. Retrieved April 9, 2020, from https://www.mathsisfun.com/data/standard-deviation-formulas.html

University of Wisconsin System. (n.d.). The Normal Distribution. Retrieved April 9, 2020, from https://mat117.wisconsin.edu/3-the-normal-distribution/