towards data science

Follow 619K Followers

Fundamentals Of Statistics For Data Scientists and Analysts

Key statistical concepts for your data science or data analysis journey

Tatev Karen Apr 17, 2021 · 31 min read

Image Source:Pexels/Anna Nekrashevich

As Karl Pearson, a British mathematician has once stated, **Statistics** is the grammar of science and this holds especially for Computer and Information Sciences, Physical

Science, and Biological Science. When you are getting started with your journey in **Data Science** or **Data Analytics**, having statistical knowledge will help you to better leverage data insights.

> *"Statistics is the grammar of science."* ***Karl Pearson***

The importance of statistics in data science and data analytics cannot be underestimated. Statistics provides tools and methods to find structure and to give deeper data insights. Both Statistics and Mathematics love facts and hate guesses. Knowing the fundamentals of these two important subjects will allow you to think critically, and be creative when using the data to solve business problems and make data-driven decisions. In this article, I will cover the following Statistics topics for data science and data analytics:

```
- Random variables
- Probability distribution functions (PDFs)
- Mean, Variance, Standard Deviation
- Covariance and Correlation
- Bayes Theorem
- Linear Regression and Ordinary Least Squares (OLS)
- Gauss-Markov Theorem
- Parameter properties (Bias, Consistency, Efficiency)
- Confidence intervals
- Hypothesis testing
- Statistical significance
- Type I & Type II Errors
- Statistical tests (Student's t-test, F-test)
- p-value and its limitations
- Inferential Statistics
- Central Limit Theorem & Law of Large Numbers
- Dimensionality reduction techniques (PCA, FA)
```

If you have no prior Statistical knowledge and you want to identify and learn the essential statistical concepts from the scratch, to prepare for your job interviews, then this article is for you. This article will also be a good read for anyone who wants to refresh his/her statistical knowledge.

Random Variables

The concept of random variables forms the cornerstone of many statistical concepts. It might be hard to digest its formal mathematical definition but simply put, a **random**

variable is a way to map the outcomes of random processes, such as flipping a coin or rolling a dice, to numbers. For instance, we can define the random process of flipping a coin by random variable X which takes a value 1 if the outcome if *heads* and 0 if the outcome is *tails.*

$$X = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$$

In this example, we have a random process of flipping a coin where this experiment can produce ***two possible outcomes***: {0,1}. This set of all possible outcomes is called the ***sample space*** of the experiment. Each time the random process is repeated, it is referred to as an ***event.*** In this example, flipping a coin and getting a tail as an outcome is an event. The chance or the likelihood of this event occurring with a particular outcome is called the ***probability*** of that event. A probability of an event is the likelihood that a random variable takes a specific value of x which can be described by P(x). In the example of flipping a coin, the likelihood of getting heads or tails is the same, that is 0.5 or 50%. So we have the following setting:

$$Pr\ (\mathrm{X} = \text{heads}) = 0.5$$
$$Pr\ (\mathrm{X} = \text{tails}) = 0.5$$

where the probability of an event, in this example, can only take values in the range [0,1].

> The importance of statistics in data science and data analytics cannot be underestimated. Statistics provides tools and methods to find structure and to

give deeper data insights.

Mean, Variance, Standard Deviation

To understand the concepts of mean, variance, and many other statistical topics, it is important to learn the concepts of ***population*** and ***sample***. The ***population*** is the set of all observations (individuals, objects, events, or procedures) and is usually very large and diverse, whereas a ***sample*** is a subset of observations from the population that ideally is a true representation of the population.

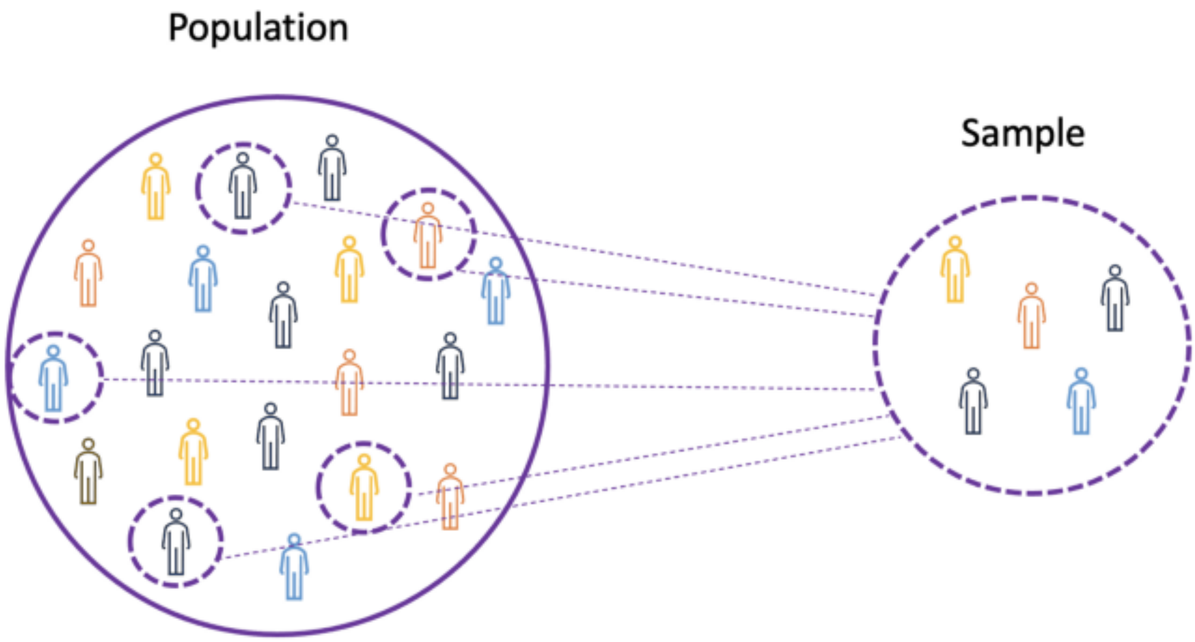

Image Source: The Author

Given that experimenting with an entire population is either impossible or simply too expensive, researchers or analysts use samples rather than the entire population in their experiments or trials. To make sure that the experimental results are reliable and hold for the entire population, the sample needs to be a true representation of the population. That is, the sample needs to be unbiased. For this purpose, one can use statistical sampling techniques such as Random Sampling, Systematic Sampling, Clustered Sampling, Weighted Sampling, and Stratified Sampling.

Mean

The mean, also known as the average, is a central value of a finite set of numbers. Let's assume a random variable X in the data has the following values:

$$x_1, x_2, x_3, \ldots, x_N$$

where N is the number of observations or data points in the sample set or simply the data frequency. Then the ***sample mean*** defined by **μ**, which is very often used to approximate the ***population mean,*** can be expressed as follows:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

The mean is also referred to as ***expectation*** which is often defined by **E**() or random variable with a bar on the top. For example, the expectation of random variables X and Y, that is **E**(X) and **E**(Y), respectively, can be expressed as follows:

$$\bar{\mathrm{X}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X_i}}{\mathrm{N}}$$

$$\bar{\mathrm{Y}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{Y_i}}{\mathrm{N}}$$

```
import numpy as np
import math

x = np.array([1,3,5,6])
mean_x = np.mean(x)

# in case the data contains Nan values
x_nan = np.array([1,3,5,6, math.nan])
mean_x_nan = np.nanmean(x_nan)
```

Variance

The variance measures how far the data points are spread out from the average value, and is equal to the sum of squares of differences between the data values and the average (the mean). Furthermore, the ***sample variance*** defined by sigma squared, which can be used to approximate the ***population variance,*** can be expressed as follows:

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

```
x = np.array([1,3,5,6])
variance_x = np.var(x)


# here you need to specify the degrees of freedom (df) max number of
logically independent data points that have freedom to vary

x_nan = np.array([1,3,5,6, math.nan])
mean_x_nan = np.nanvar(x_nan, ddof = 1)
```

For deriving expectations and variances of different popular probability distribution functions, check out this Github repo.

Standard Deviation

The standard deviation is simply the square root of the variance and measures the extent to which data varies from its mean. The standard deviation defined by ***sigma*** can be expressed as follows:

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

Standard deviation is often preferred over the variance because it has the same unit as the data points, which means you can interpret it more easily.

```
x = np.array([1,3,5,6])
variance_x = np.std(x)

x_nan = np.array([1,3,5,6, math.nan])
mean_x_nan = np.nanstd(x_nan, ddof = 1)
```

Covariance

The covariance is a measure of the joint variability of two random variables and