Fundamentals Of Statistics For Data Scientists and Analysts

towardsdatascience.com/fundamentals-of-statistics-for-data-scientists-and-data-analysts-69d93a05aae7

Tatev Karen August 20, 2021

towards data science

Key statistical concepts for your data science or data analysis journey

Image Source:

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 ad 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

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.

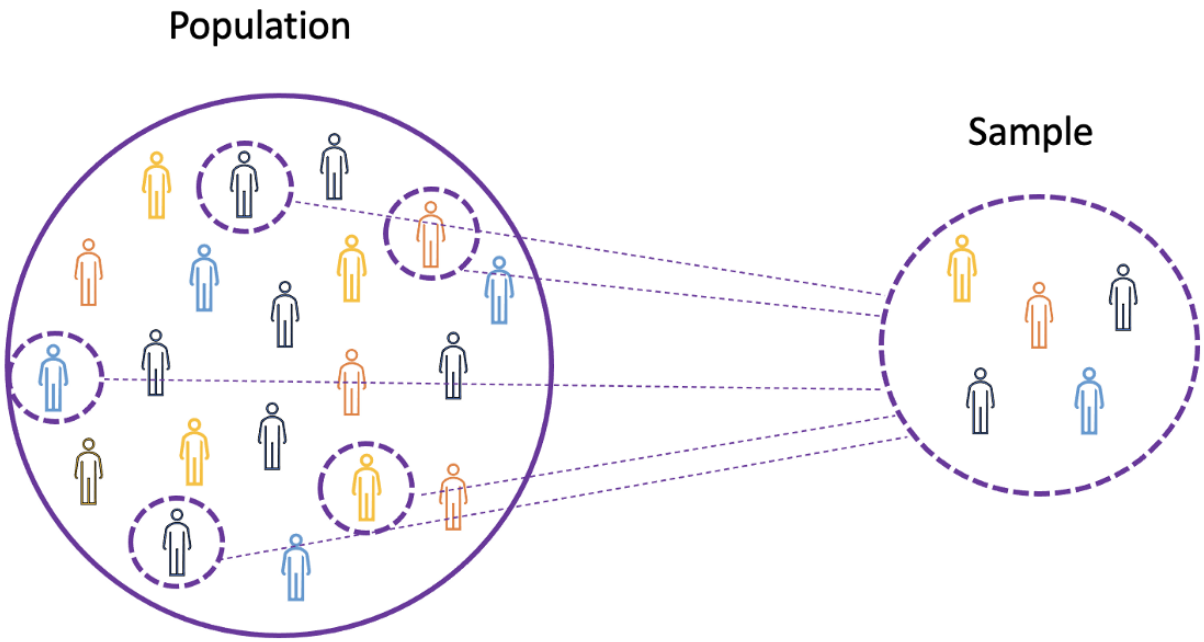

Image Source: The Author

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:

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

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

```
import numpy as np
import mathx = np.array([1,3,5,6])
mean_x = np.mean(x)
```

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)
```

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 describes the relationship between these two variables. It is defined as the expected value of the product of the two random variables' deviations from their means. The covariance between two random variables X and Z can be described by the following expression, where **E**(X) and **E**(Z) represent the means of X and Z, respectively.

$$Cov\,(\mathrm{X}, \mathrm{Z}) = \mathbf{E}\,[(\mathrm{X} - \mathbf{E}(\mathrm{X}))\,(\mathrm{Z} - \mathbf{E}(\mathrm{Z}))]$$

Covariance can take negative or positive values as well as value 0. A positive value of covariance indicates that two random variables tend to vary in the same direction, whereas a negative value suggests that these variables vary in opposite directions. Finally, the value 0 means that they don't vary together.

```
x = np.array([1,3,5,6])
y = np.array([-2,-4,-5,-6])#this will return the covariance matrix of x,y containing
x_variance, y_variance on diagonal elements and covariance of x,y
cov_xy = np.cov(x,y)
```

Correlation

The correlation is also a measure for relationship and it measures both the strength and the direction of the linear relationship between two variables. If a correlation is detected then it means that there is a relationship or a pattern between the values of two target variables.

Correlation between two random variables X and Z are equal to the covariance between these two variables divided to the product of the standard deviations of these variables which can be described by the following expression.

$$Cor\ (X, Z) = \frac{Cov\ (X,Z)}{\sigma_x \sigma_z}$$

Correlation coefficients' values range between -1 and 1. Keep in mind that the correlation of a variable with itself is always 1, that is **Cor(X, X) = 1**. Another to keep in mind when interpreting correlation is to not confuse it with ***causation***, given that a correlation is not causation. Even if there is a correlation between two variables, you cannot conclude that one variable causes a change in the other. This relationship could be coincidental, or a third factor might be causing both variables to change.

```
x = np.array([1,3,5,6])
y = np.array([-2,-4,-5,-6])corr = np.corrcoef(x,y)
```

Probability Distribution Functions

A function that describes all the possible values, the sample space, and the corresponding probabilities that a random variable can take within a given range, bounded between the minimum and maximum possible values, is called ***a probability distribution function (pdf)*** or probability density. Every pdf needs to satisfy the following two criteria:

$$0 \leq Pr\ (\mathrm{X}) \leq 1$$

$$\sum p(\mathrm{X}) = 1$$

where the first criterium states that all probabilities should be numbers in the range of [0,1] and the second criterium states that the sum of all possible probabilities should be equal to 1.

Probability functions are usually classified into two categories: ***discrete*** and ***continuous***. Discretedistributionfunction describes the random process with ***countable*** sample space, like in the case of an example of tossing a coin that has only two possible outcomes. Continuousdistribution function describes the random process with ***continuous*** sample space. Examples of discrete distribution functions are Bernoulli, Binomial, Poisson, Discrete Uniform. Examples of continuous distribution functions are Normal, Continuous Uniform, Cauchy.

Binomial Distribution

The binomial distribution is the discrete probability distribution of the number of successes in a sequence of **n** independent experiments, each with the boolean-valued outcome: ***success*** (with probability **p**) or ***failure*** (with probability **q** = 1 − p). Let's assume a random variable X follows a Binomial distribution, then the probability of observing***k*** successes in n independent trials can be expressed by the following probability density function:

$$Pr\,(\mathrm{X} = \mathrm{k}) = \binom{n}{k}\, p^k q^{n-k}$$

The binomial distribution is useful when analyzing the results of repeated independent experiments, especially if one is interested in the probability of meeting a particular threshold given a specific error rate.

Binomial Distribution Mean & Variance

$$\mathrm{E(X)} = \mathrm{np}$$

$$\mathrm{Var(X)} = \mathrm{npq}$$

The figure below visualizes an example of Binomial distribution where the number of independent trials is equal to 8 and the probability of success in each trial is equal to 16%.

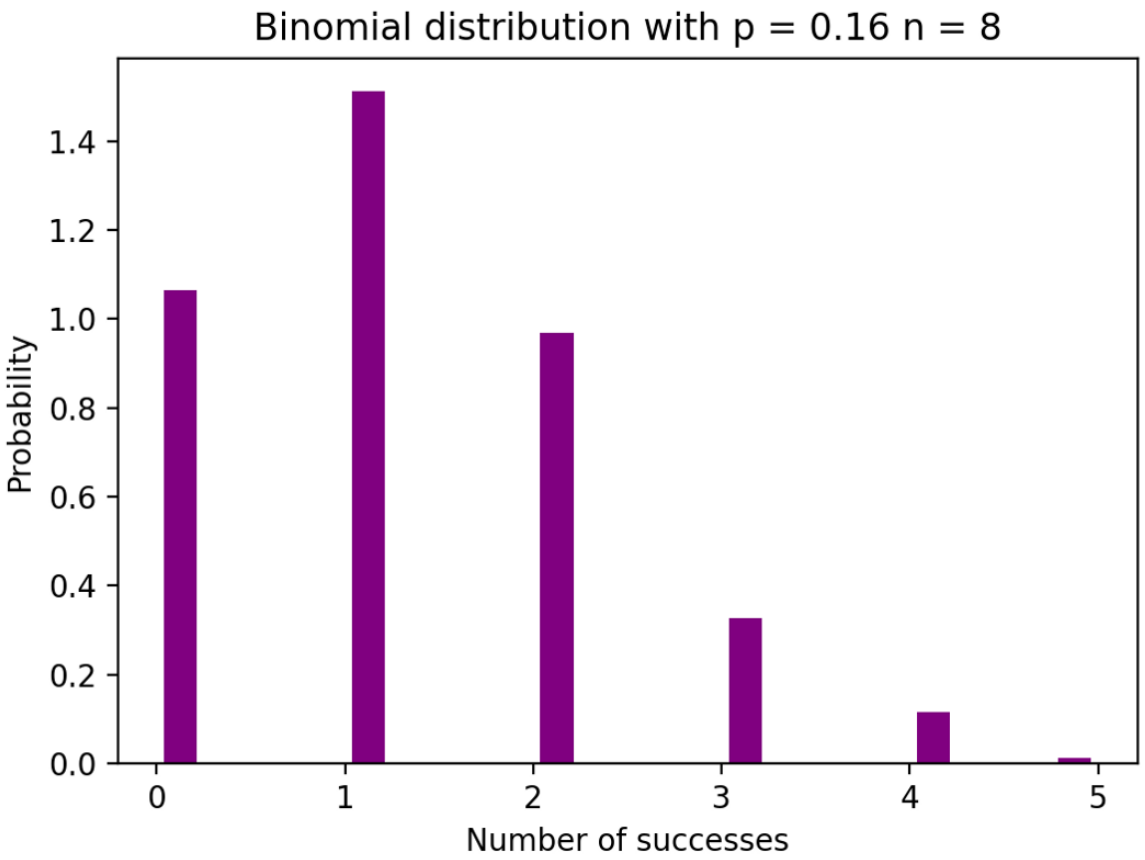

Image Source: The Author

```
# Random Generation of 1000 independent Binomial samples
import numpy as np
n = 8
p = 0.16
N = 1000
X = np.random.binomial(n,p,N)# Histogram of Binomial distribution
import matplotlib.pyplot as plt
counts, bins, ignored = plt.hist(X, 20, density = True, rwidth = 0.7, color =
'purple')
plt.title("Binomial distribution with p = 0.16 n = 8")
plt.xlabel("Number of successes")
plt.ylabel("Probability")
plt.show()
```

Poisson Distribution

The Poisson distribution is the discrete probability distribution of the number of events occurring in a specified time period, given the average number of times the event occurs over that time period. Let's assume a random variable X follows a Poisson distribution, then the probability of observing***k*** events over a time period can be expressed by the following probability function:

$$Pr\,(\mathrm{X} = \mathrm{k}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where ***e*** is ***Euler's number*** and ***λ*** lambda, the ***arrival rate parameter*** isthe expected value of X. Poisson distribution function is very popular for its usage in modeling countable events occurring within a given time interval.

Poisson Distribution Mean & Variance

$$\mathrm{E}(\mathrm{X}) = \lambda$$

$$\mathrm{Var}(\mathrm{X}) = \lambda$$

For example, Poisson distribution can be used to model the number of customers arriving in the shop between 7 and 10 pm, or the number of patients arriving in an emergency room between 11 and 12 pm. The figure below visualizes an example of Poisson distribution where we count the number of Web visitors arriving at the website where the arrival rate, lambda, is assumed to be equal to 7 minutes.

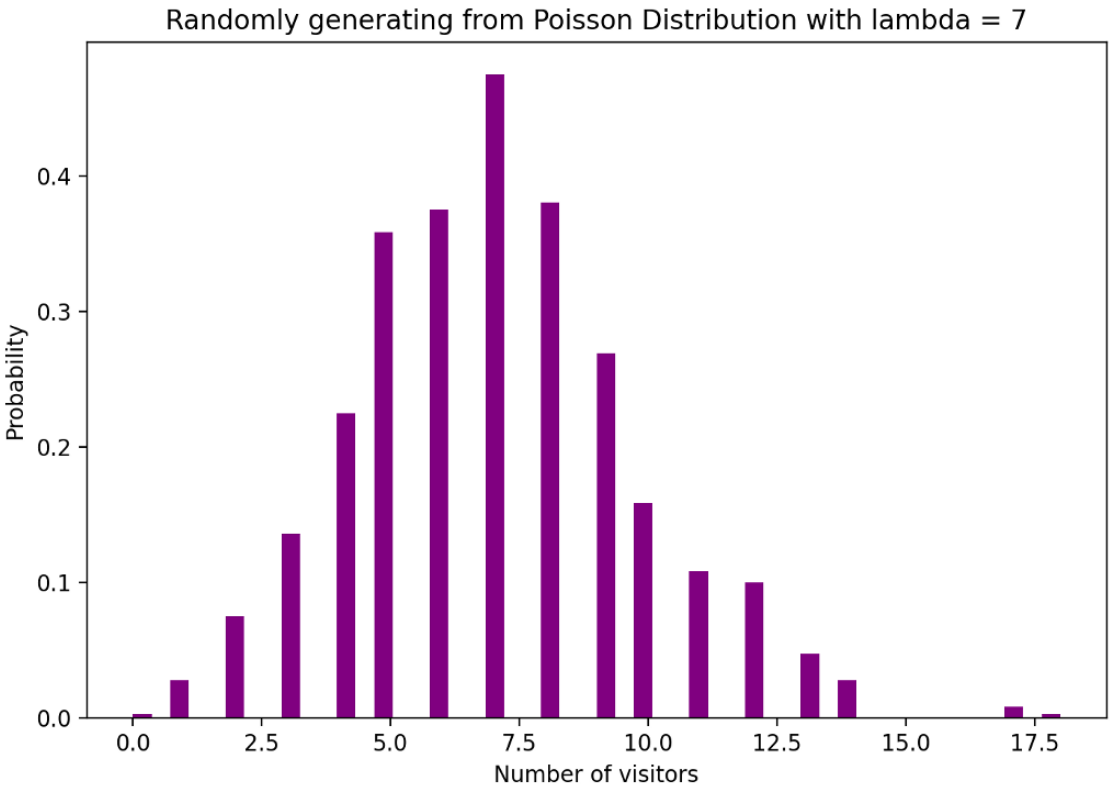

Image Source: The Author

```
# Random Generation of 1000 independent Poisson samples
import numpy as np
lambda_ = 7
N = 1000
X = np.random.poisson(lambda_,N)

# Histogram of Poisson distribution
import matplotlib.pyplot as plt
counts, bins, ignored = plt.hist(X, 50, density = True, color = 'purple')
plt.title("Randomly generating from Poisson Distribution with lambda = 7")
plt.xlabel("Number of visitors")
plt.ylabel("Probability")
plt.show()
```

Normal Distribution

The Normal probability distribution is the continuous probability distribution for a real-valued random variable. Normal distribution, also called ***Gaussian distribution*** is arguably one of the most popular distribution functions that are commonly used in social and natural sciences for modeling purposes, for example, it is used to model people's height or test scores. Let's assume a random variable X follows a Normal distribution, then its probability density function can be expressed as follows.

$$Pr\,(\mathrm{X}=\mathrm{k})=\frac{1}{\sigma\sqrt{2\pi}}\;e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

where the parameter **μ** (mu)is the mean of the distribution also referred to as the ***location parameter***, parameter **σ** (sigma)is the standard deviation of the distribution also referred to as the *scale parameter*. The number **π** (pi) is a mathematical constant approximately equal to 3.14.

Normal Distribution Mean & Variance

$$\mathrm{E}(\mathrm{X})=\mu$$

$$\mathrm{Var}(\mathrm{X})=\sigma^2$$

The figure below visualizes an example of Normal distribution with a mean 0 (**μ = 0**) and standard deviation of 1 (**σ = 1**), which is referred to as ***Standard Normal*** distribution which is *symmetric.*

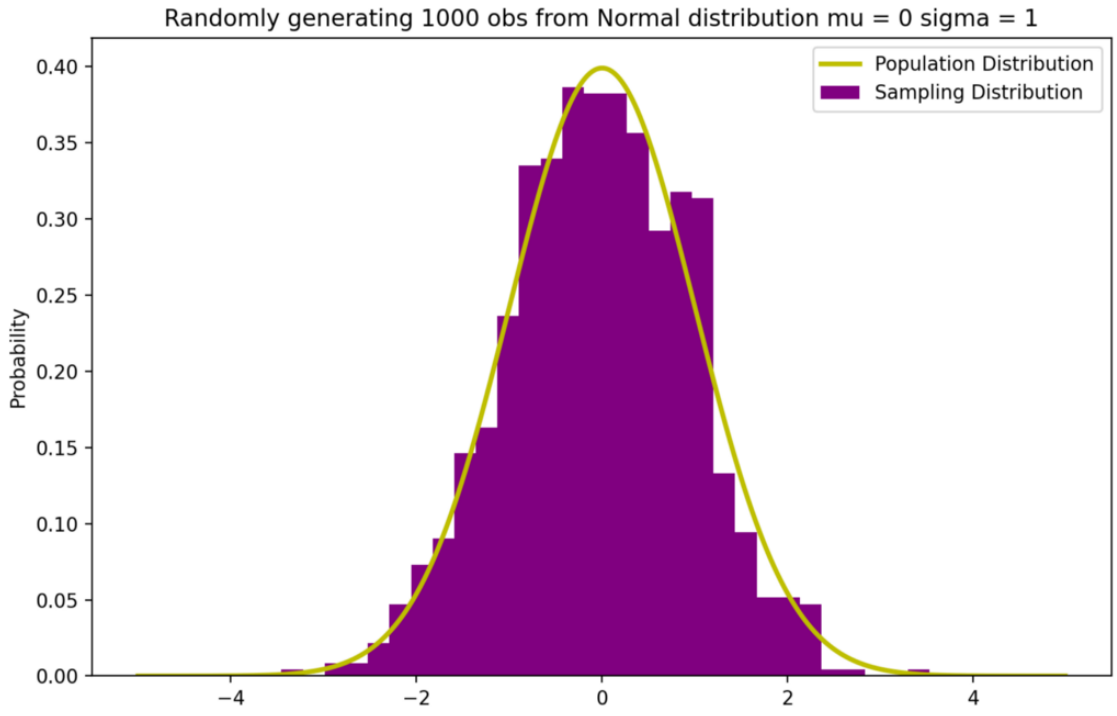

Image Source: The Author

```
# Random Generation of 1000 independent Normal samples
import numpy as np
mu = 0
sigma = 1
N = 1000
X = np.random.normal(mu,sigma,N)

# Population distribution
from scipy.stats import norm
x_values = np.arange(-5,5,0.01)
y_values = norm.pdf(x_values)#Sample histogram with Population distribution
import matplotlib.pyplot as plt
counts, bins, ignored = plt.hist(X, 30, density = True,color = 'purple',label =
'Sampling Distribution')
plt.plot(x_values,y_values, color = 'y',linewidth = 2.5,label = 'Population
Distribution')
plt.title("Randomly generating 1000 obs from Normal distribution mu = 0 sigma = 1")
plt.ylabel("Probability")
plt.legend()
plt.show()
```

Bayes Theorem

The Bayes Theorem or often called ***Bayes Law*** is arguably the most powerful rule of probability and statistics, named after famous English statistician and philosopher, Thomas Bayes.

Bayes theorem is a powerful probability law that brings the concept of ***subjectivity*** into the world of Statistics and Mathematics where everything is about facts. It describes the probability of an event, based on the prior information of ***conditions*** that might be related to that event. For instance, if the risk of getting Coronavirus or Covid-19 is known to increase with age, then Bayes Theorem allows the risk to an individual of a known age to be determined more accurately by conditioning it on the age than simply assuming that this individual is common to the population as a whole.

Image Source: Wikipedia

The concept of ***conditional probability*****,** which plays a central role in Bayes theory, is a measure of the probability of an event happening, given that another event has already occurred. Bayes theorem can be described by the following expression where the X and Y stand for events X and Y, respectively:

$$Pr\,(\mathrm{X} \mid \mathrm{Y}) = \frac{Pr\,(\mathrm{Y} \mid \mathrm{X})\,Pr\,(\mathrm{X})}{Pr\,(\mathrm{Y})}$$

- *Pr* (X|Y): the probability of event X occurring given that event or condition Y has occurred or is true
- *Pr* (Y|X): the probability of event Y occurring given that event or condition X has occurred or is true
- *Pr* (X) & *Pr* (Y): the probabilities of observing events X and Y, respectively

In the case of the earlier example, the probability of getting Coronavirus (event X) conditional on being at a certain age is *Pr* (X|Y), which is equal to the probability of being at a certain age given one got a Coronavirus, *Pr* (Y|X), multiplied with the probability of getting a Coronavirus, *Pr* (X), divided to the probability of being at a certain age., *Pr* (Y).

The Confidence Interval is the range that contains the true population parameter with a certain pre-specified probability, referred to as the ***confidence level***of the experiment, and it is obtained by using the sample results and the ***margin of error***.

Inferential Statistics

Inferential statistics uses sample data to make reasonable judgments about the population from which the sample data originated. It's used to investigate the relationships between variables within a sample and make predictions about how these variables will relate to a larger population.

Both ***Law of Large Numbers (LLN)*** and ***Central Limit Theorem (CLM)*** have a significant role in Inferential statistics because they show that the experimental results hold regardless of what shape the original population distribution was when the data is large enough. The more data is gathered, the more accurate the statistical inferences become, hence, the more accurate parameter estimates are generated.

Law of Large Numbers (LLN)

Suppose **X1, X2, . . . , Xn** are all independent random variables with the same underlying distribution, also called independent identically-distributed or i.i.d, where all X's have the same mean **μ** and standard deviation **σ**. As the sample size grows, the probability that the average of all X's is equal to the mean μ is equal to 1. The Law of Large Numbers can be summarized as follows:

$$\mathbf{E}\,(X_i) = \mu \quad \mathbf{Var}\,(X_i) = \sigma^2 \quad \text{for } i = 1,\ldots,N$$

$$\overline{X}_n = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$N \to \infty \qquad \text{then} \qquad \Pr(\overline{X}_n = \mu) = 1$$

Central Limit Theorem (CLM)

Suppose **X1, X2, . . . , Xn** are all independent random variables with the same underlying distribution, also called independent identically-distributed or i.i.d, where all X's have the same mean **μ** and standard deviation **σ**. As the sample size grows, the probability distribution of X ***converges in the distribution*** in Normal distribution with mean **μ** and variance **σ**-squared. The Central Limit Theorem can be summarized as follows:

$$\mathbf{E}\,(X_i) = \mu \quad \mathbf{Var}\,(X_i) = \sigma^2 \quad \text{for } i = 1,\ldots,N$$

$$\overline{X}_n = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$N \to \infty \qquad \text{then} \qquad \Pr(\overline{X}_n) \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{N}\right)$$