Published in Towards Data Science

You have **2** free member-only stories left this month.
Sign up for Medium and get an extra one

Tatev Karen Aslanyan
Apr 17, 2021 · 31 min read · Member-only · Listen

Fundamentals Of Statistics For Data Scientists and Analysts

Key statistical concepts for your data science or data analysis journey

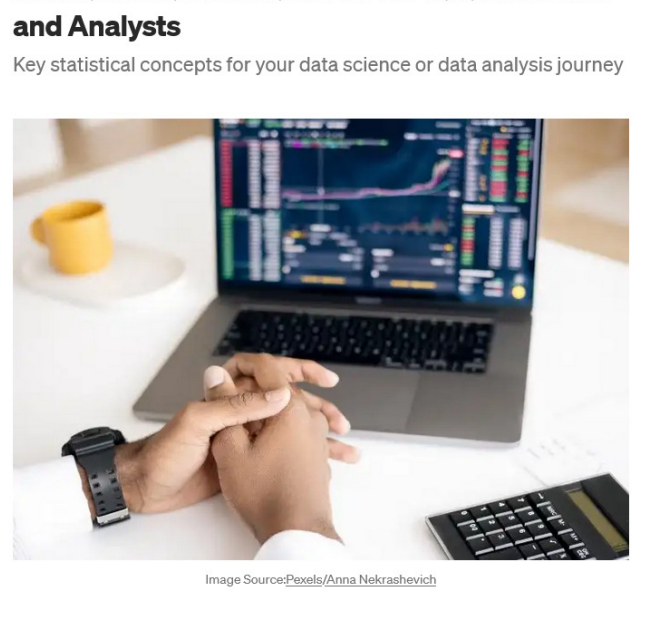


Image Source:Pexels/Anna Nekrashevich

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:

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

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.

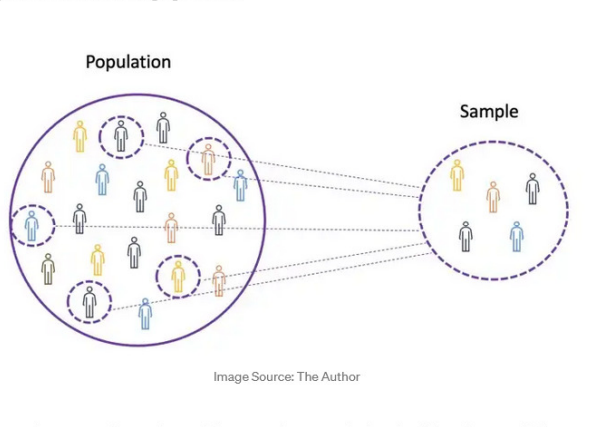


Image Source: The Author

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:

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

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

```
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 ***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 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.

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

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 thing 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:

$$0 \leq Pr\,(X) \leq 1$$
$$\sum p(X) = 1$$

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.*** Discrete distribution function 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. Continuous distribution 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:

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

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

$$E(X) = np$$
$$Var(X) = npq$$

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%.

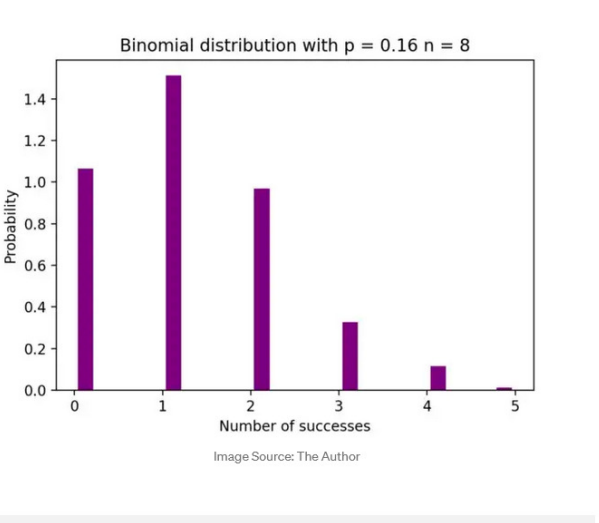


Image Source: The Author

```
# 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:

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

where ***e*** is ***Euler's number*** and λ lambda, the ***arrival rate parameter*** is the 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

$$E(X) = \lambda$$
$$Var(X) = \lambda$$

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.

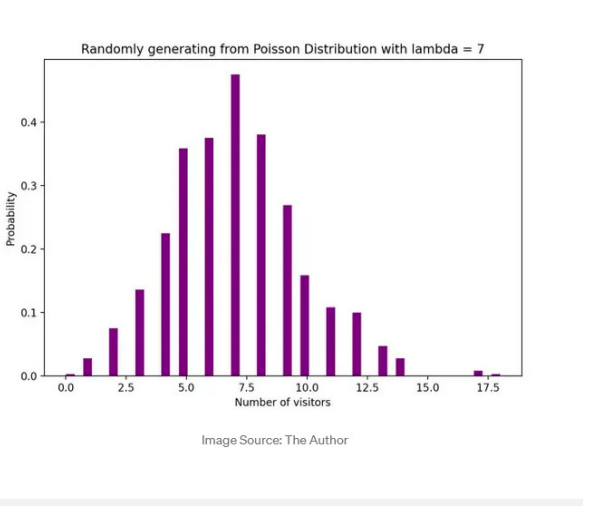


Image Source: The Author

```
# 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.

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

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

$$E(X) = \mu$$
$$Var(X) = \sigma^2$$

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.*

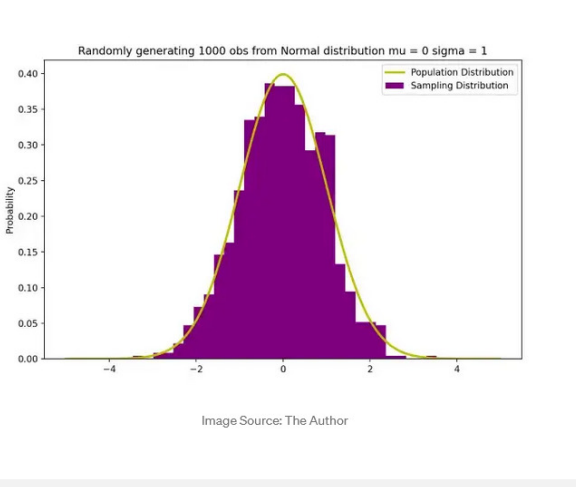


Image Source: The Author

```
# 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.

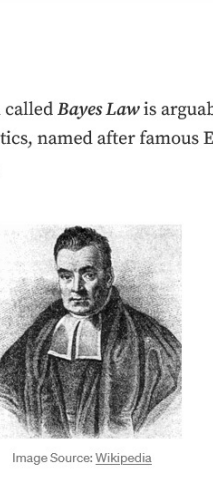


Image Source: Wikipedia

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.

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:

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

- *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).

· · ·

Linear Regression

Earlier, the concept of causation between variables was introduced, which happens when a variable has a direct impact on another variable. When the relationship between two variables is linear, then Linear Regression is a statistical method that can help to model the impact of a unit change in a variable, ***the independent variable*** on the values of another variable, ***the dependent variable.***

Dependent variables are often referred to as ***response variables*** or ***explained variables,*** whereas independent variables are often referred to as ***regressors*** or ***explanatory variables***. When the Linear Regression model is based on a single independent variable, then the model is called ***Simple Linear Regression*** and when the model is based on multiple independent variables, it's referred to as ***Multiple Linear Regression.*** Simple Linear Regression can be described by the following expression:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where **Y** is the dependent variable, **X** is the independent variable which is part of the data, β0 is the intercept which is unknown and constant, β1 is the slope coefficient or a parameter corresponding to the variable X which is unknown and constant as well. Finally, ***u*** is the error term that the model makes when estimating the Y values. The main idea behind linear regression is to find the best-fitting straight line, ***the regression line,*** through a set of paired (X, Y) data. One example of the Linear Regression application is modeling the impact of *Flipper Length* on penguins' *Body Mass*, which is visualized below.

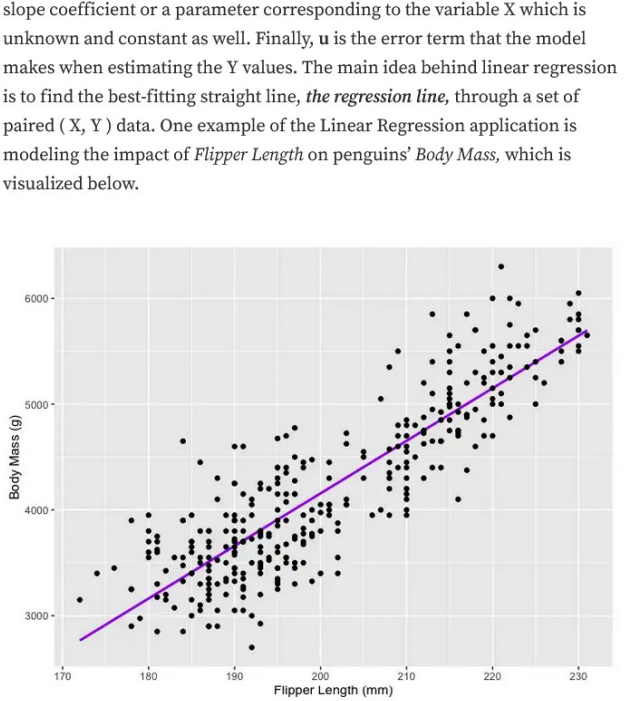


Image Source: The Author

```
# R code for the graph
install.packages("ggplot2")
install.packages("palmerpenguins")
library(palmerpenguins)
library(ggplot2)
View(data(penguins))
ggplot(data = penguins, aes(x = flipper_length_mm,y = body_mass_g))+
  geom_smooth(method = "lm", se = FALSE, color = 'purple')+
  geom_point()+
  labs(x="Flipper Length (mm)",y="Body Mass (g)")
```

Multiple Linear Regression with three independent variables can be described by the following expression:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i$$

Ordinary Least Squares

The ordinary least squares (OLS) is a method for estimating the unknown parameters such as β0 and β1 in a linear regression model. The model is based on the principle of ***least squares*** that minimizes the sum of squares of the differences between the observed dependent variable and its values predicted by the linear function of the independent variable, often referred to as ***fitted values.*** This difference between the real and predicted values of dependent variable Y is referred to as ***residual*** and what OLS does, is minimizing the sum of squared residuals. This optimization problem results in the following OLS estimates for the unknown parameters β0 and β1 which are also known as ***coefficient estimates.***

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Once these parameters of the Simple Linear Regression model are estimated, the ***fitted values*** of the response variable can be computed as follows:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Standard Error

The ***residuals*** or the estimated error terms can be determined as follows:

$$\hat{u}_i = Y_i - \hat{Y}_i$$

It is important to keep in mind the difference between the error terms and residuals. Error terms are never observed, while the residuals are calculated from the data. The OLS estimates the error terms for each observation but not the actual error term. So, the true error variance is still unknown. Moreover, these estimates are subject to sampling uncertainty. What this means is that we will never be able to determine the exact estimate, the true value, of these parameters from sample data in an empirical application. However, we can estimate it by calculating the ***sample residual variance*** by using the residuals as follows.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N - 2}$$

This estimate for the variance of sample residuals helps to estimate the variance of the estimated parameters which is often expressed as follows:

$$Var\,(\hat{\beta})$$

The squared root of this variance term is called ***the standard error*** of the estimate which is a key component in assessing the accuracy of the parameter estimates. It is used to calculating test statistics and confidence intervals. The standard error can be expressed as follows:

$$SE\,(\hat{\beta}) = \sqrt{Var\,(\hat{\beta})}$$

> It is important to keep in mind the difference between the error terms and residuals. Error terms are never observed, while the residuals are calculated from the data.

OLS Assumptions

OLS estimation method makes the following assumption which needs to be satisfied to get reliable prediction results:

A1: Linearity assumption states that the model is linear in parameters.

A2: Random Sample assumption states that all observations in the sample are randomly selected.

A3: Exogeneity assumption states that independent variables are uncorrelated with the error terms.

A4: Homoskedasticity assumption states that the variance of all error terms is constant.

A5: No Perfect Multi-Collinearity assumption states that none of the independent variables is constant and there are no exact linear relationships between the independent variables.

```
def runOLS(Y,X):

   # OLS esyimation Y = Xb + e --> beta_hat = (X'X)^-1(X'Y)
   beta_hat = np.dot(np.linalg.inv(np.dot(np.transpose(X), X)), np.dot(np.transpose(X), Y))

   # OLS prediction
   Y_hat = np.dot(X,beta_hat)
   residuals = Y-Y_hat
   RSS = np.sum(np.square(residuals))
   sigma_squared_hat = RSS/(N-2)
   TSS = np.sum(np.square(Y-np.repeat(Y.mean(),len(Y))))
   MSE = sigma_squared_hat
   RMSE = np.sqrt(MSE)
   R_squared = (TSS-RSS)/TSS

   # Standard error of estimates:square root of estimate's variance
   var_beta_hat = np.linalg.inv(np.dot(np.transpose(X),X))*sigma_squared_hat

   SE = []
   t_stats = []
   p_values = []
   CI_s = []

   for i in range(len(beta)):
       #standard errors
       SE_i = np.sqrt(var_beta_hat[i,i])
       SE.append(np.round(SE_i,3))

       #t-statistics
       t_stat = np.round(beta_hat[i,0]/SE_i,3)
       t_stats.append(t_stat)

       #p-value of t-stat p[|t_stat| >= t-treshold two sided]
       p_value = t.sf(np.abs(t_stat),N-2) * 2
       p_values.append(np.round(p_value,3))

       #Confidence intervals = beta_hat -+ margin_of_error
       t_critical = t.ppf(q =1-0.05/2, df = N-2)
       margin_of_error = t_critical*SE_i
       CI = [np.round(beta_hat[i,0]-margin_of_error,3), np.round(beta_hat[i,0]+margin_of_error,3)]
       CI_s.append(CI)

   return(beta_hat, SE, t_stats, p_values,CI_s,
          MSE, RMSE, R_squared)
```

· · ·

Parameter Properties

> *Under the assumption that the OLS criteria A1 — A5 are satisfied, the OLS estimators of coefficients β0 and β1 are **BLUE** and **Consistent**.*
>
> *Gauss-Markov theorem*

This theorem highlights the properties of OLS estimates where the term ***BLUE*** stands for ***Best Linear Unbiased Estimator.***

Bias

The **bias** of an estimator is the difference between its expected value and the true value of the parameter being estimated and can be expressed as follows:

$$\mathbf{Bias}\,(\beta, \hat{\beta}) = \mathbf{E}\,(\hat{\beta}) - \beta$$

When we state that the estimator is ***unbiased*** what we mean is that the bias is equal to zero, which implies that the expected value of the estimator is equal to the true parameter value, that is:

$$\mathbf{E}\,(\hat{\beta}) = \beta$$

Unbiasedness does not guarantee that the obtained estimate with any particular sample is equal or close to β. What it means is that, if one ***repeatedly*** draws random samples from the population and then computes the estimate each time, then the average of these estimates would be equal or very close to β.

Efficiency

The term ***Best*** in the Gauss-Markov theorem relates to the variance of the estimator and is referred to as ***efficiency***. A parameter can have multiple estimators but the one with the lowest variance is called efficient.

Consistency

The term consistency goes hand in hand with the terms ***sample size*** and ***convergence***. If the estimator converges to the true parameter as the sample size becomes very large, then this estimator is said to be consistent, that is:

$$N \rightarrow \infty \quad \text{then} \quad \hat{\beta} \rightarrow \beta$$

> Under the assumption that the OLS criteria A1 — A5 are satisfied, the OLS estimators of coefficients β0 and β1 are BLUE and Consistent.
>
> Gauss-Markov Theorem

All these properties hold for OLS estimates as summarized in the Gauss-Markov theorem. In other words, OLS estimates have the smallest variance, they are unbiased, linear in parameters, and are consistent. These properties can be mathematically proven by using the OLS assumptions made earlier.

· · ·

Confidence Intervals

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***.

Margin of Error

The margin of error is the difference between the sample results and based on what the result would have been if one had used the entire population.

Confidence Level

The Confidence Level describes the level of certainty in the experimental results. For example, a 95% confidence level means that if one were to perform the same experiment repeatedly for 100 times, then 95 of those 100 trials would lead to similar results. Note that the confidence level is defined before the start of the experiment because it will affect how big the margin of error will be at the end of the experiment.

Confidence Interval for OLS Estimates

As it was mentioned earlier, the OLS estimates of the Simple Linear Regression, the estimates for intercept β0 and slope coefficient β1, are subject to sampling uncertainty. However, we can construct CI's for these parameters which will contain the true value of these parameters in 95% of all samples. That is, 95% confidence interval for β can be interpreted as follows:

- The confidence interval is the set of values for which a hypothesis test cannot be rejected to the level of 5%.
- The confidence interval has a 95% chance to contain the true value of β.

95% confidence interval of OLS estimates can be constructed as follows:

$$CI^{\beta_1}_{0.95} = \left[\hat{\beta}_1 - 1.96\,\mathbf{SE}\,(\hat{\beta}_1),\ \hat{\beta}_1 + 1.96\,\mathbf{SE}\,(\hat{\beta}_1)\right]$$

which is based on the parameter estimate, the standard error of that estimate, and the value 1.96 representing the margin of error corresponding to the 5% rejection rule. This value is determined using the Normal Distribution table, which will be discussed later on in this article. Meanwhile, the following figure illustrates the idea of 95% CI:

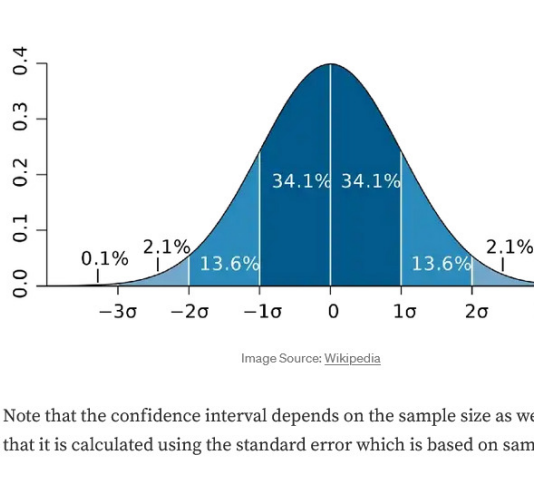


Image Source: Wikipedia

Note that the confidence interval depends on the sample size as well, given that it is calculated using the standard error which is based on sample size.

· · ·

> The confidence level is defined before the start of the experiment because it will affect how big the margin of error will be at the end of the experiment.

Statistical Hypothesis testing

Testing a hypothesis in Statistics is a way to test the results of an experiment or survey to determine how meaningful they the results are. Basically, one is testing whether the obtained results are valid by figuring out the odds that the results have occurred by chance. If it is the letter, then the results are not reliable and neither is the experiment. Hypothesis Testing is part of the ***Statistical Inference.***

Null and Alternative Hypothesis

Firstly, you need to determine the thesis you wish to test, then you need to formulate the ***Null Hypothesis*** and the ***Alternative Hypothesis.*** The test can have two possible outcomes and based on the statistical results you can either reject the stated hypothesis or accept it. As a rule of thumb, statisticians tend to put the version or formulation of the hypothesis under the Null Hypothesis that that needs to be rejected, whereas the acceptable and desired version is stated under the Alternative Hypothesis.

Statistical significance

Let's look at the earlier mentioned example where the Linear Regression model was used to investigating whether a penguins' *Flipper Length*, the independent variable, has an impact on *Body Mass*, the dependent variable. We can formulate this model with the following statistical expression:

$$Y_{BodyMass} = \beta_0 + \beta_1 X_{FlipperLength} + u_i$$

Then, once the OLS estimates of the coefficients are estimated, we can formulate the following Null and Alternative Hypothesis to test whether the Flipper Length has a ***statistically significant*** impact on the Body Mass:

$$\begin{cases} H_0\text{: Flipper Length doesn't have statistically significant impact on Body Mass} \\ H_1\text{: Flipper Length has statistically significant impact on Body Mass} \end{cases}$$

where H0 and H1 represent Null Hypothesis and Alternative Hypothesis, respectively. Rejecting the Null Hypothesis would mean that a one-unit increase in *Flipper Length* has a direct impact on the *Body Mass*. Given that the parameter estimate of β1 is describing this impact of the independent variable, *Flipper Length*, on the dependent variable, *Body Mass*. This hypothesis can be reformulated as follows:

$$\begin{cases} H_0: \hat{\beta}_1 = 0 \\ H_1: \hat{\beta}_1 \neq 0 \end{cases}$$

where H0 states that the parameter estimate of β1 is equal to 0, that is *Flipper Length* effect on *Body Mass* is ***statistically insignificant*** whereas H0 states that the parameter estimate of β1 is not equal to 0 suggesting that *Flipper Length* effect on *Body Mass* is ***statistically significant.***

Type I and Type II Errors

When performing Statistical Hypothesis Testing one needs to consider two conceptual types of errors: Type I error and Type II error. The Type I error occurs when the Null is wrongly rejected whereas the Type II error occurs when the Null Hypothesis is wrongly not rejected. A confusion matrix can help to clearly visualize the severity of these two types of errors.

> As a rule of thumb, statisticians tend to put the version the hypothesis under the Null Hypothesis that that needs to be rejected, whereas the acceptable and desired version is stated under the Alternative Hypothesis.

Statistical Tests

Once the Null and the Alternative Hypotheses are stated and the test assumptions are defined, the next step is to determine which statistical test is appropriate and to calculate the ***test statistic***. Whether or not to reject or not reject the Null can be determined by comparing the test statistic with the ***critical value.*** This comparison shows whether or not the observed test statistic is more extreme than the defined critical value and it can have two possible results:

- The test statistic is more extreme than the critical value → the null hypothesis can be rejected
- The test statistic is not as extreme as the critical value → the null hypothesis cannot be rejected

The critical value is based on a prespecified ***significance level α*** (usually chosen to be equal to 5%) and the type of probability distribution the test statistic follows. The critical value divides the area under this probability distribution curve into the ***rejection region(s)*** and ***non-rejection region.*** There are numerous statistical tests used to test various hypotheses. Examples of Statistical tests are Student's t-test, F-test, Chi-squared test, Durbin-Hausman-Wu Endogeneity test, White Heteroskedasticity test. In this article, we will look at two of these statistical tests.

> The Type I error occurs when the Null is wrongly rejected whereas the Type II error occurs when the Null Hypothesis is wrongly not rejected.

Student's t-test

One of the simplest and most popular statistical tests is the Student's t-test. which can be used for testing various hypotheses especially when dealing with a hypothesis where the main area of interest is to find evidence for the statistically significant effect of a ***single variable.*** The test statistics of the t-test follows ***Student's t distribution*** and can be determined as follows:

$$T_{stat} = \frac{\hat{\beta}_i - h_0}{SE(\hat{\beta}_i)}$$

where h0 in the nominator is the value against which the parameter estimate is being tested. So, the t-test statistics are equal to the parameter estimate minus the hypothesized value divided by the standard error of the coefficient estimate. In the earlier stated hypothesis, where we wanted to test whether Flipper Length has a statistically significant impact on Body Mass or not. This test can be performed using a t-test and the h0 is in that case equal to the 0 since the slope coefficient estimate is tested against value 0.

There are two versions of the t-test: a ***two-sided t-test*** and a ***one-sided t-test.*** Whether you need the former or the latter version of the test depends entirely on the hypothesis that you want to test.

The two-sided or ***two-tailed t-test*** can be used when the hypothesis is testing *equal* versus *not equal* relationship under the Null and Alternative Hypotheses that is similar to the following example:

$$\begin{cases} H_0: \hat{\beta}_1 = h_0 \\ H_1: \hat{\beta}_1 \neq h_0 \end{cases}$$

The two-sided t-test has ***two rejection regions*** as visualized in the figure below:

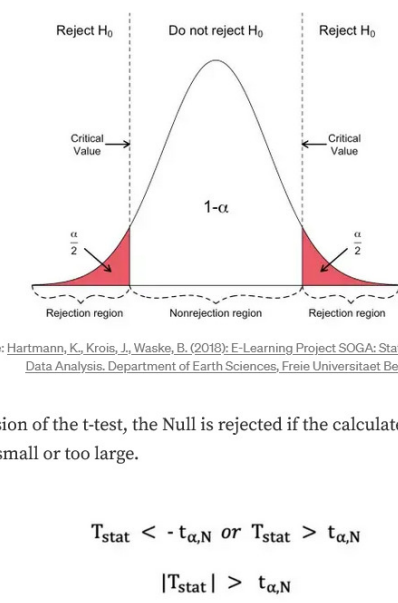


Image Source: Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin

In this version of the t-test, the Null is rejected if the calculated t-statistics is either too small or too large.

$$T_{stat} < -t_{\alpha,N} \ \ or \ \ T_{stat} > t_{\alpha,N}$$
$$|T_{stat}| > t_{\alpha,N}$$

Here, the test statistics are compared to the critical values based on the sample size and the chosen significance level. To determine the exact value of the cutoff point, the two-sided t-distribution table can be used.

The one-sided or ***one-tailed t-test*** can be used when the hypothesis is testing *positive/negative* versus *negative/positive* relationship under the Null and Alternative Hypotheses that is similar to the following examples:

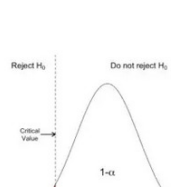
Left Tailed

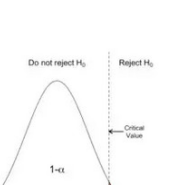
$$\begin{cases} H_0: \hat{\beta}_1 > h_0 \\ H_1: \hat{\beta}_1 \leq h_0 \end{cases}$$

Right Tailed

$$\begin{cases} H_0: \hat{\beta}_1 < h_0 \\ H_1: \hat{\beta}_1 \geq h_0 \end{cases}$$

One-sided t-test has a ***single rejection region*** and depending on the hypothesis side the rejection region is either on the left-hand side or the right-hand side as visualized in the figure below:





Tatev Karen Aslanyan
998 Followers
Data Scientist | https://www.linkedin.com/in/tatev-karen-aslanyan https://github.com/TatevKaren
Follow

More from Medium

Zach... in Pipeline: A Data Engin...
3 Data Science Projects That Got Me 12 Interviews. And 1 That Got Me in Trouble.

The PyCo... in Towards Data Sci...
Predicting The FIFA World Cup 2022 With a Simple Model using Python

Anmol Tomar in CodeX
Say Goodbye to Loops in Python, and Welcome Vectorization!