Statistics for Machine Learning A-Z

June 12, 2022

Author(s): Gencay I.

Briefly Explained

Photo by Dan-Cristian Pădureț on Unsplash

Contents

- Experimental Studies
- Simple Random Sample
- Stratified Sample
- Placebo Effect
- Generalizability
- Histogram
- Dot plot
- Boxplot
- IQR
- Q3
- Q1
- Left skewed
- Right skewed
- Symmetric
- Mean
- Median
- Average
- Variance
- Standard deviation
- Mode
- Null Hypothesis
- Alternative Hypothesis
- P-Value
- Law of Large Numbers
- Mutually Exclusive (Disjoint)
- Non-disjoint
- Probability Trees
- Normal Distribution
- Binomial Distribution
- Bernoulli Distribution

Introduction

Programming, Statistics, Calculus.

These are 3 things that you should be familiar with if you would like to be involved in Machine Learning.
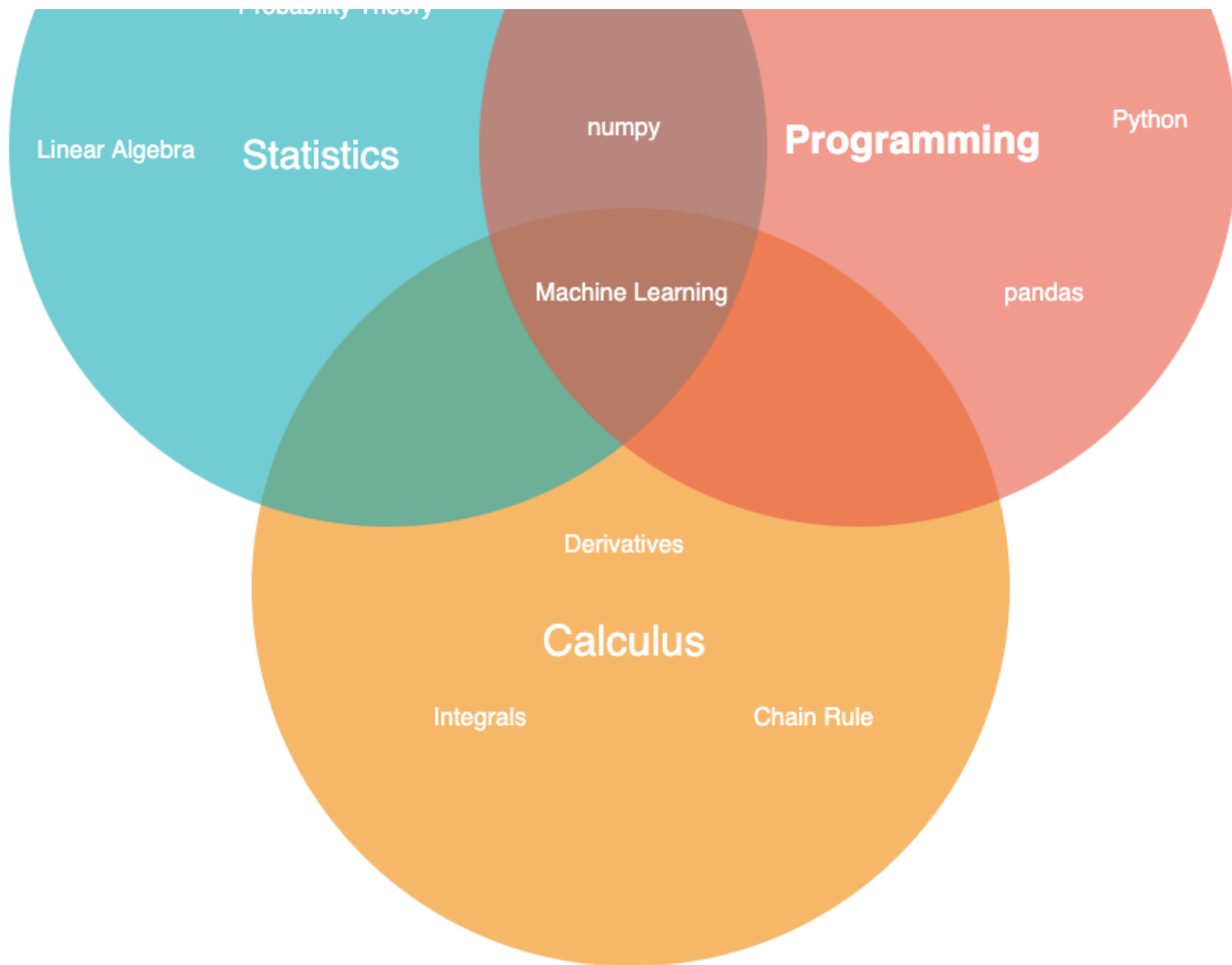
Image By Author

While there are too many courses that existed in the market, I love creating that kind of article to remind myself of these terms.

That helps me refresh my memories and make repetition.

> Repetition is the mother of learning, the father of action, which makes it the architect of accomplishment." Zig Ziglar

Whether you are at the beginning of your Data Science or Machine Learning career or experienced one, that article will serve you to create a neural path in your mind about Statistics.

Let's dive into these terms from beginning to intermediate.

Terms

Numerical Variable

The value that contains an integer.

Categorical Variable

Contains categories instead of numbers, such as human body shapes such as; skinny, fat, or muscular.

Continuous Variable

1, 2, 3,4, 5,6, 7 … Take a number of values in a given range.

Discrete Variables

1, 5, 8, 11, 35 . The specific set of values.

Dependent Variable:

The two variables, when one changes if the other will change will be dependent variables.

Independent Variable

If others won't change, independent.

Observational Studies

The methods won't be specified by researchers. For example, when they asked you about the method of losing weight, they do not offer you the method such as diet or sport, you could say whatever you like.

Experimental Studies

Now there are limited options, choose one, diet or sport.

Simple Random Sample

You could choose anything.

Stratified Sample

Split populations into the clusters, then randomly sample from each cluster.

Placebo Effect

You will use fake care.

Generalizability

Could we draw a conclusion as a result of our data on the population?

Ideal Experiment

Most Observational

	Random Assignment	No random assignment	
Ideal Experiment			*Studies*
Random Sampling	*Causal and generalizable*	*Not causal but generalizable*	**Generalizability**
No random Sampling	*Causal but not generalizable*	*Neither causal nor generalizable*	**No Generalizability**
Most Experiment	**Causation**	**Association**	*Bad Observational Studies*

Image By Author

Histogram

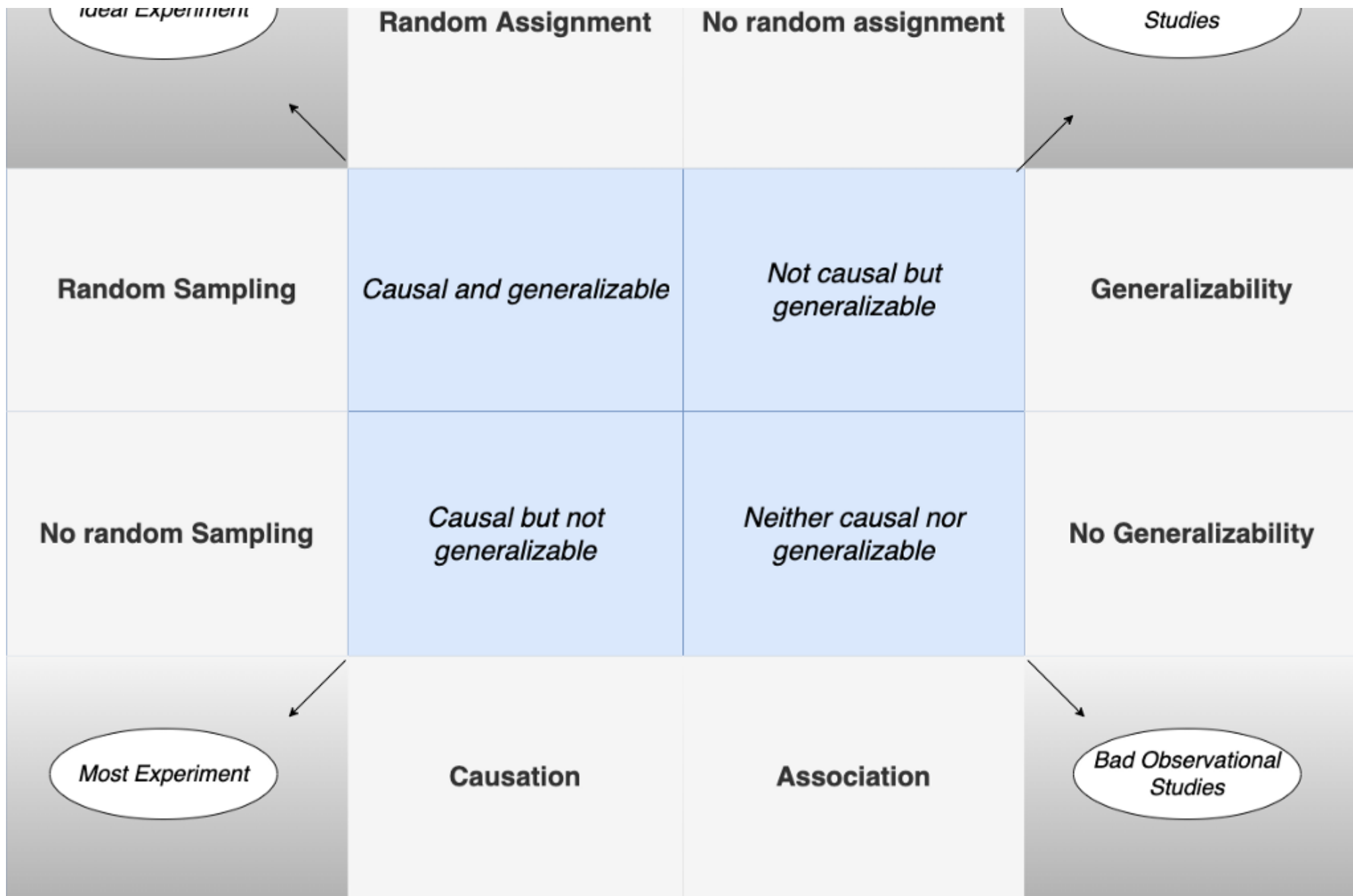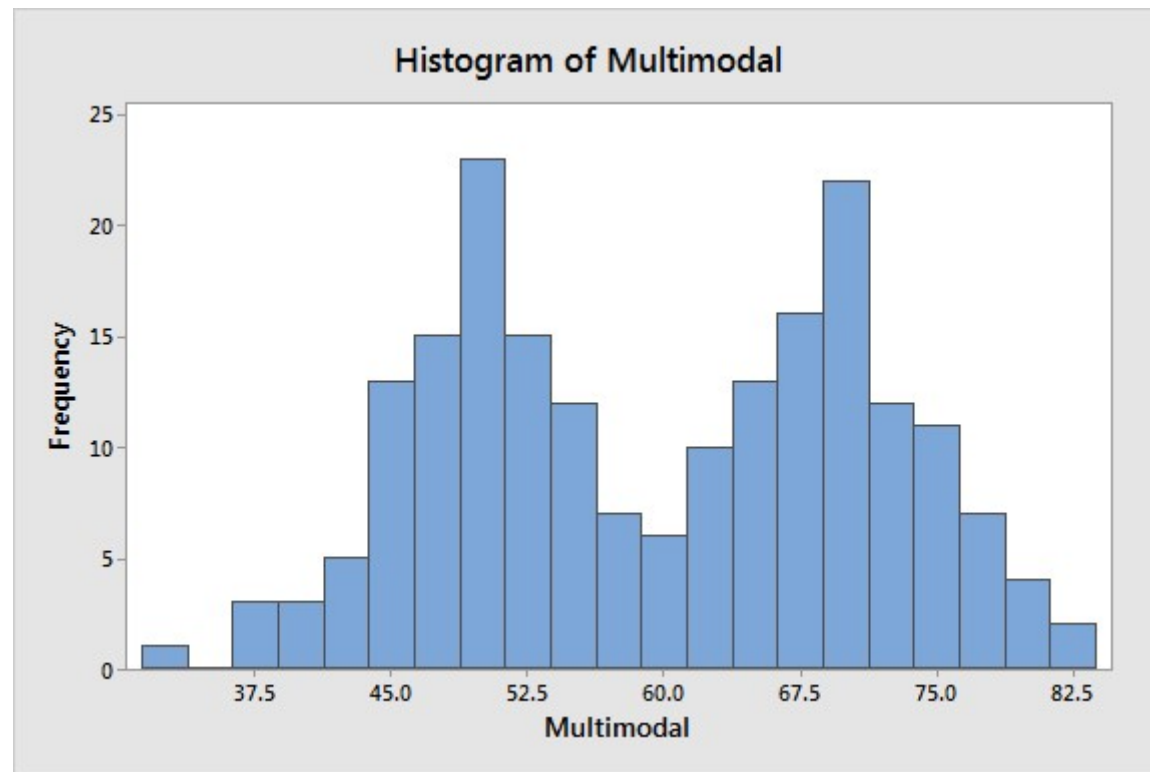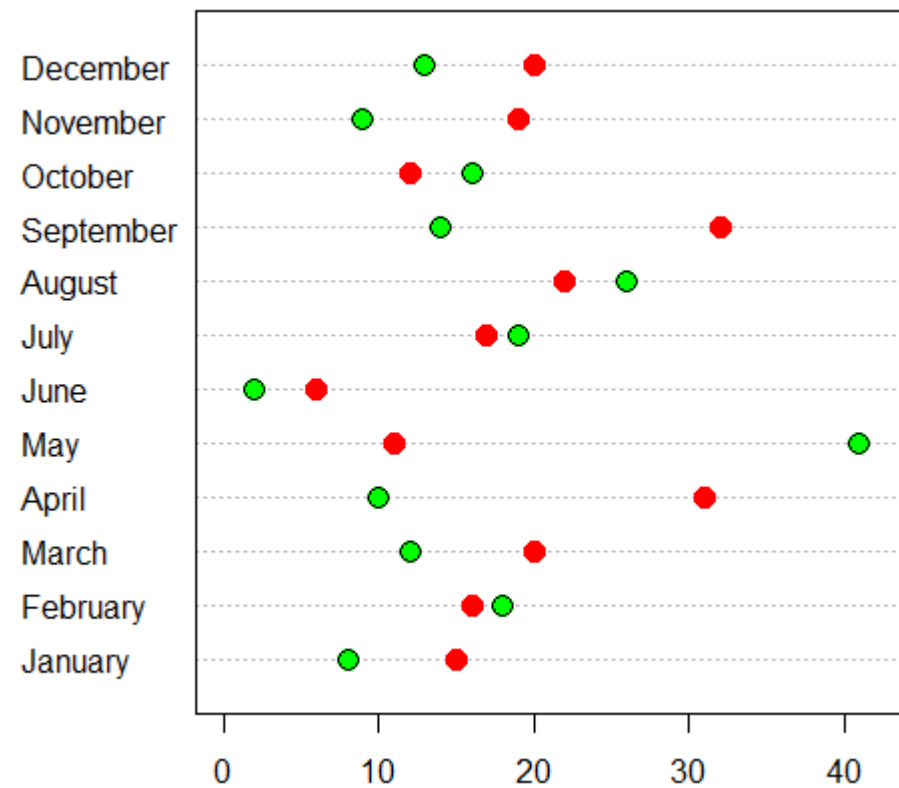*It provides a useful view of data density.*

Dot plot

If your sample size is small and you want to view individual data points.

Boxplot

It is good to see statistical values such as IQR and median.
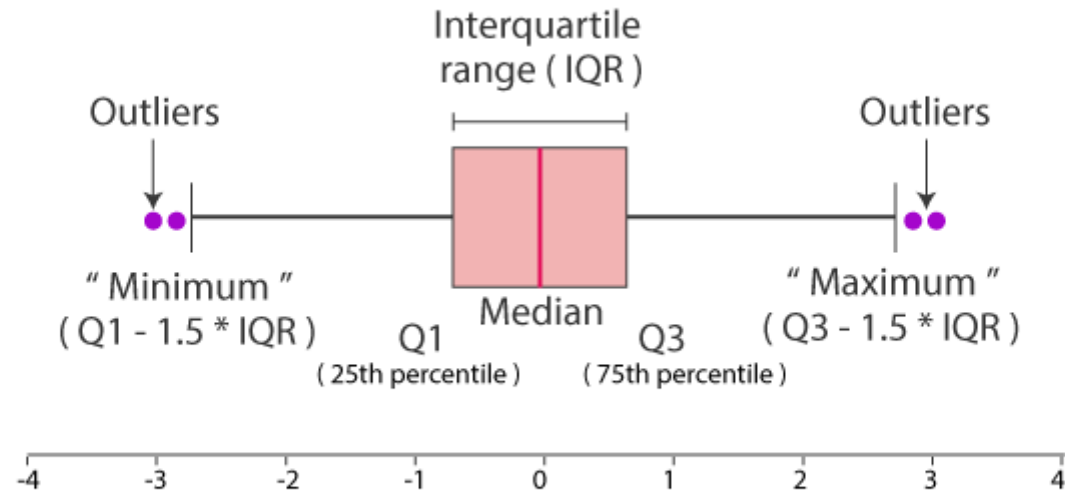
Different parts of boxplot

https://byjus.com/maths/box-plot/

IQR

Interquartile range,Range of thee middle 50 %, Q3-Q1.

Q3

75th percentile.

Q1

25th percentile.

Left skewed

The tail will be on the right side and the density, mean< median.

Right skewed

The tail will be on the right, mean > median.

Symmetric

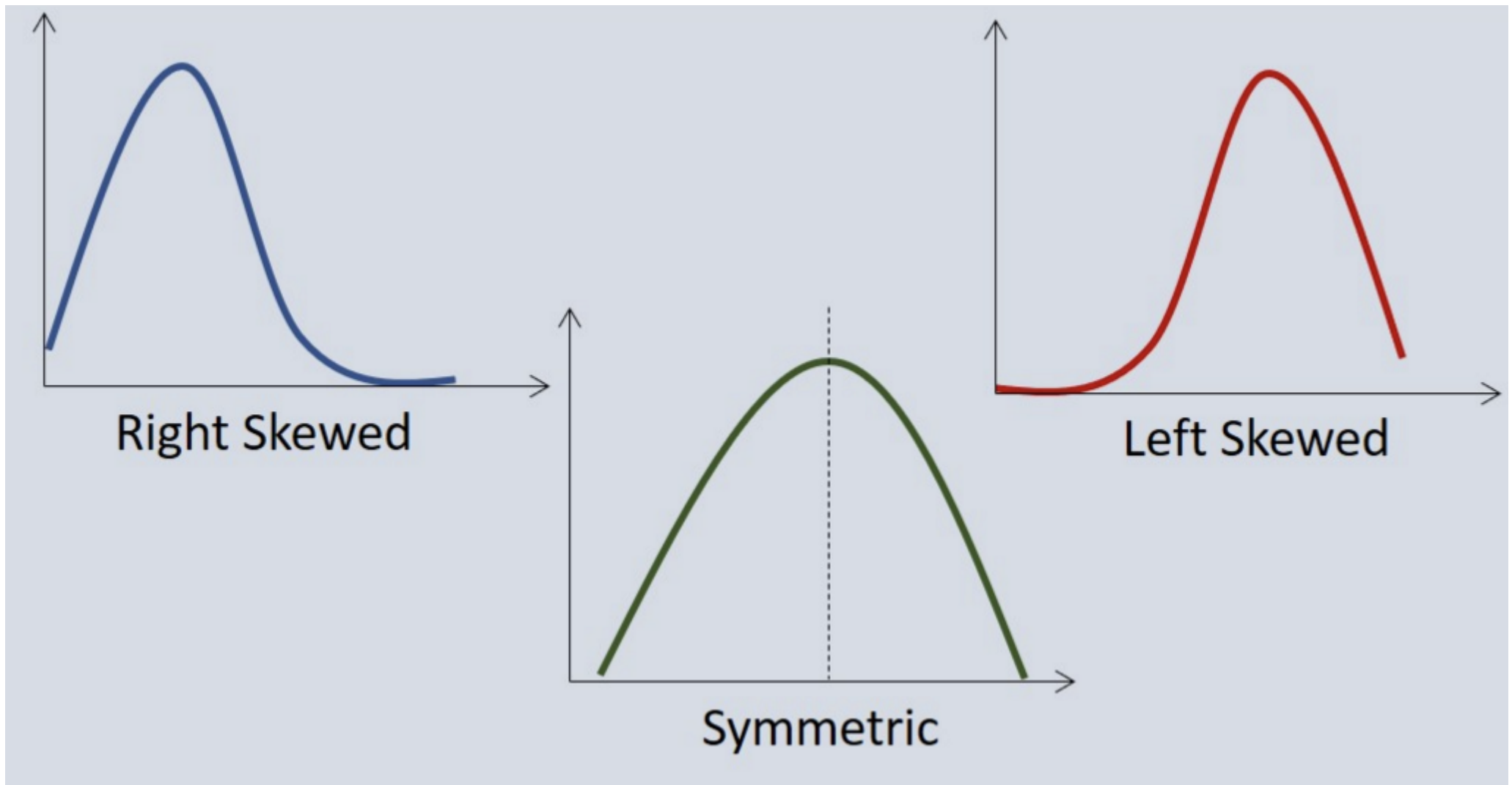*The mean and median are close together.*

Mean

Arithmetic average.

Median

is the number that exists in the middle.

1,5,7 , median : 5

Average

Sum and divide by the number of integers,

a+b/2, a+b+c/3

Variance

The average squared deviation of the mean.

```
n: number of sample
```

```
Variance = (Number1-mean)**2 + (number2 -mean) ** 2 …. + (number n-mean) **2 / (n-1)
```

Standard deviation

The square root of the variance.

```
Standard deviation : (Variance) ** 1/2
```

Mode

Most frequent number.

1,4,4, 7 , mode : 4

Null Hypothesis

Nothing going on, everything should be the same.

Alternative Hypothesis

Something going on, something should be changed.

P-Value

The possibility of your null hypothesis is true.

If the p-value < 0.05, you would reject the null hypothesis, and accept the alternative:

If the p-value >0.05, you would reject the alternative hypothesis.

Law of Large Numbers

As the sample size increases, the mean would be closer to the population means.

Mutually Exclusive (Disjoint)
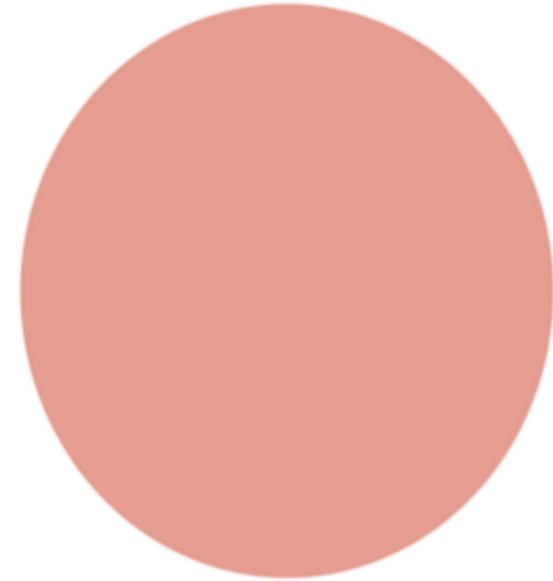
Cases that can not happen at identical times.

Image by Author

Non-disjoint

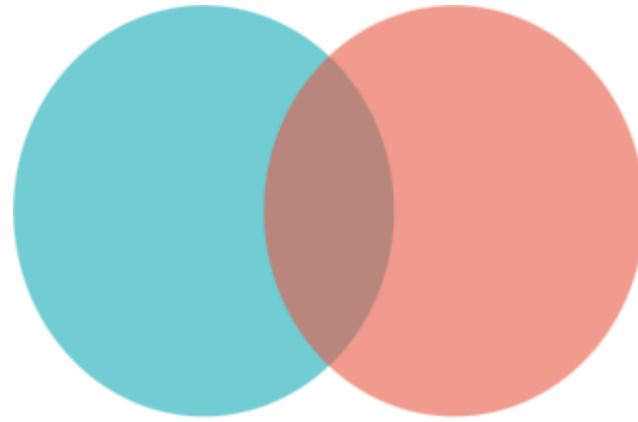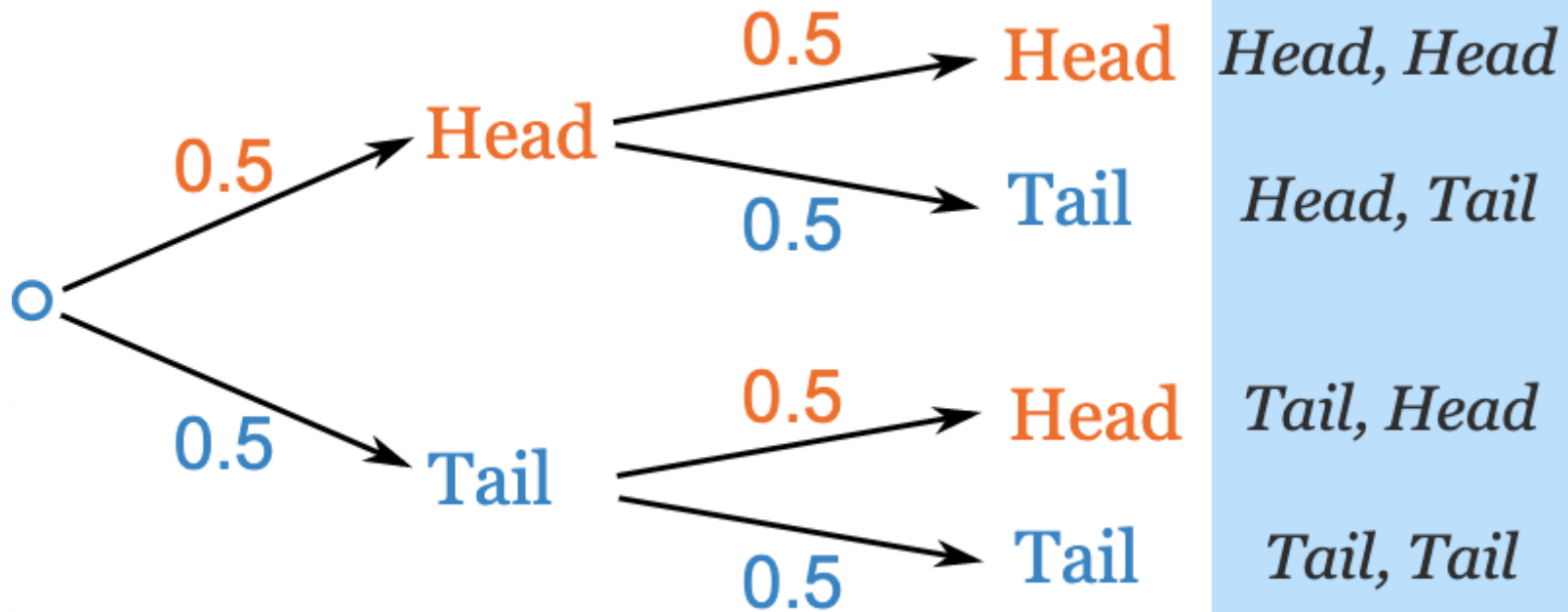*Cases that could happen at an identical time.*
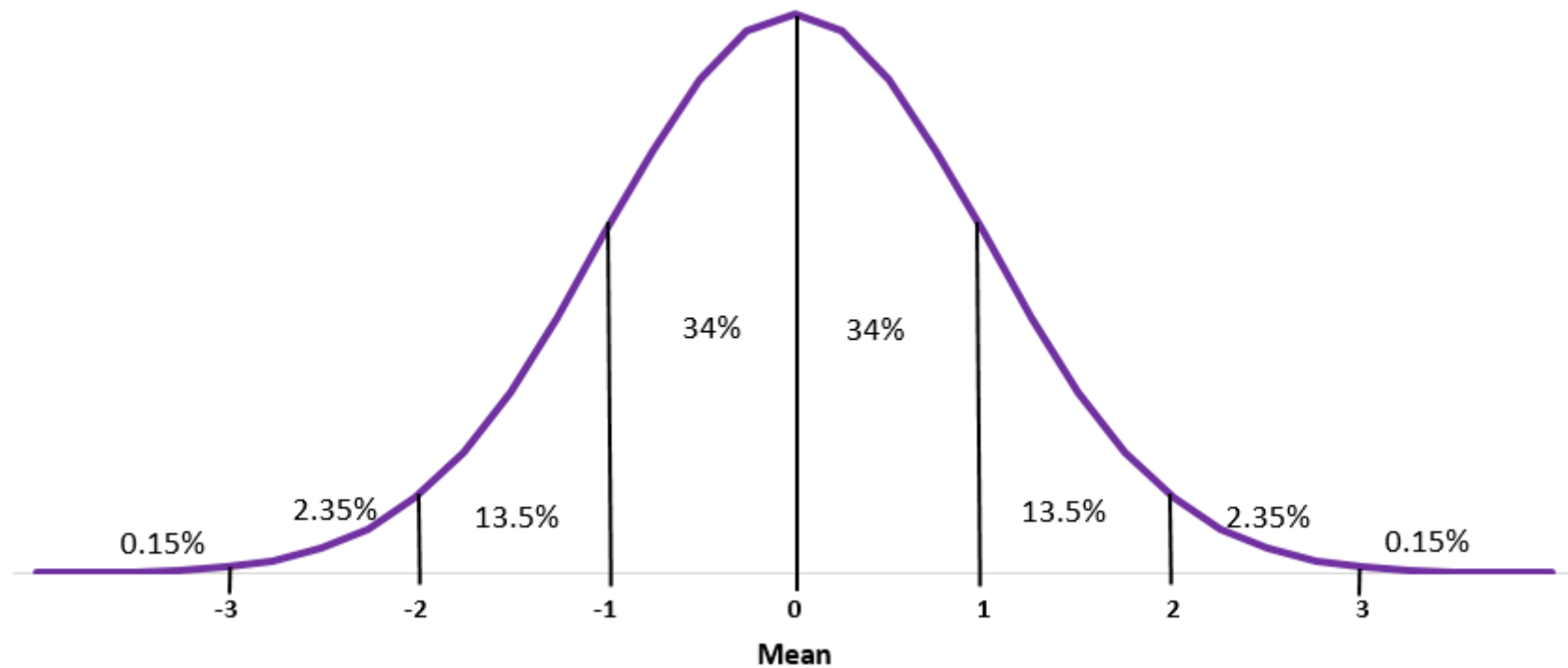
Image by Author

Probability Trees

The trees of continuous possibilities.

Normal Distribution

It is a probability distribution, that shape is symmetric around the mean.

Binomial Distribution

Probability of success or failure (2 possible outcomes, like heads or tail.)

Bernoulli Distribution

It is a discrete distribution, and we have still 2 possible outcomes, 0 or 1.
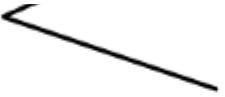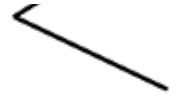
Coin Flip

    Tails

Coin Flip #1

    Tails

Coin Flip #2

    Heads

    Tails

Coin Flip #3
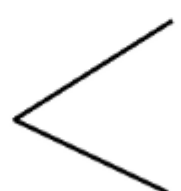
    Heads

    Tails

Coin Flip #4

    Heads

    Tails

Coin Flip #5

    Heads

    Tails

PDF (Probability Density Function)

The function provides the possibility of the value of a <u>random variable</u> will be in the predefined range.
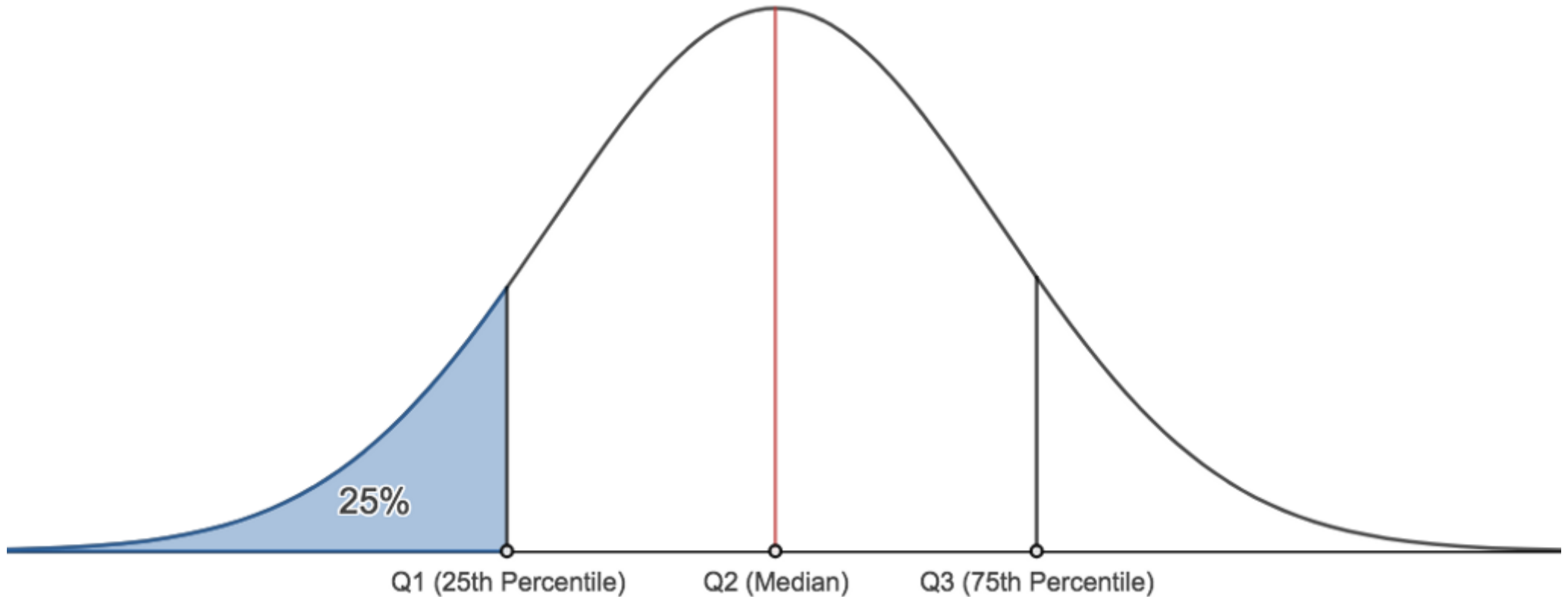
Z Score

(observation-mean) / SD

Z score of the mean is zero.

Percentiles

Sometimes images are stronger than words for describing.

25%

Q1 (25th Percentile)   Q2 (Median)   Q3 (75th Percentile)

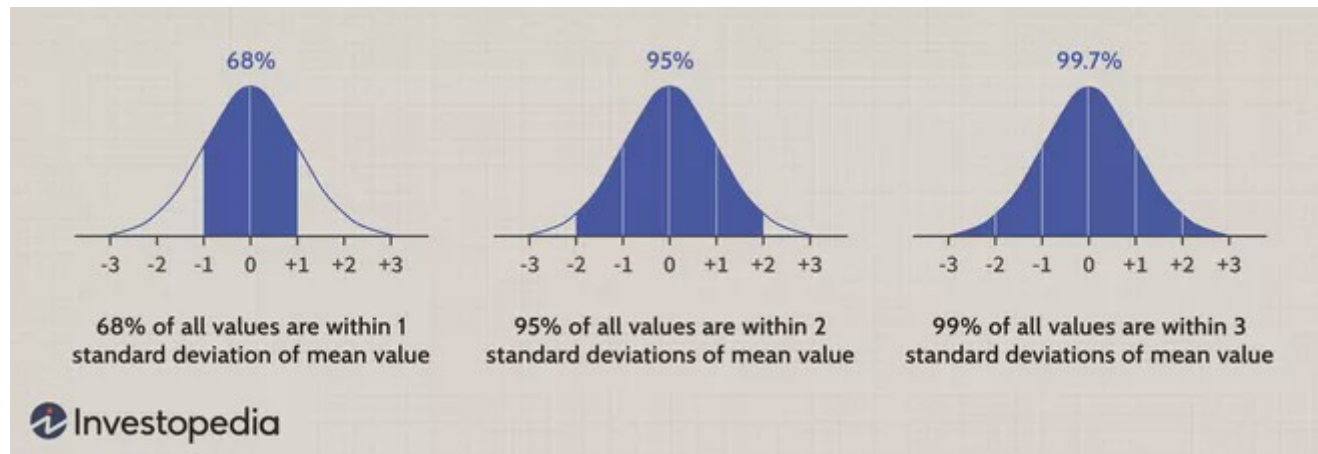https://online.stat.psu.edu/stat800/book/export/html/741

Sampling Variability

It is impossible to gather data from the whole population, we will gather information from different samples and sampling them together via this.

Central Limit Theorem

Describes shapes centers and spreads of sampling distributions when certain conditions are matched. (nearly normal population)(center-mean) (spread-SE)

68% of all values are within 1 standard deviation of mean value

95% of all values are within 2 standard deviations of mean value

99% of all values are within 3 standard deviations of mean value

Confidence Interval

A reasonable range of values for the population parameter is called a confidence interval.
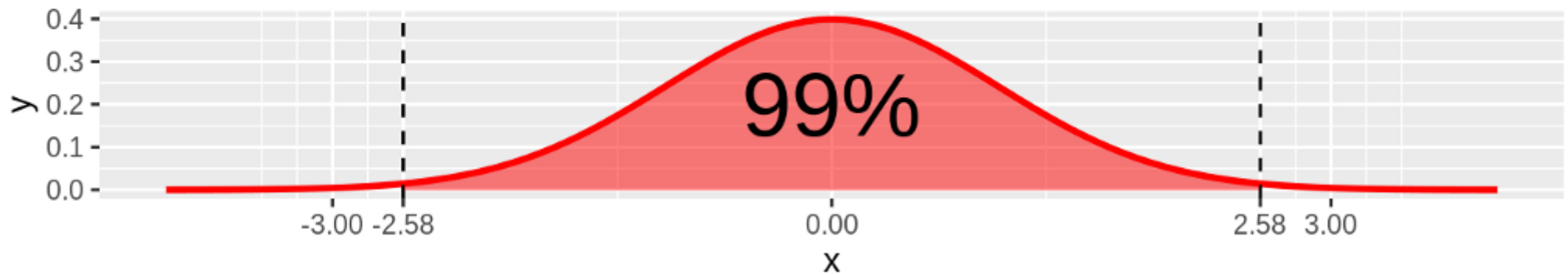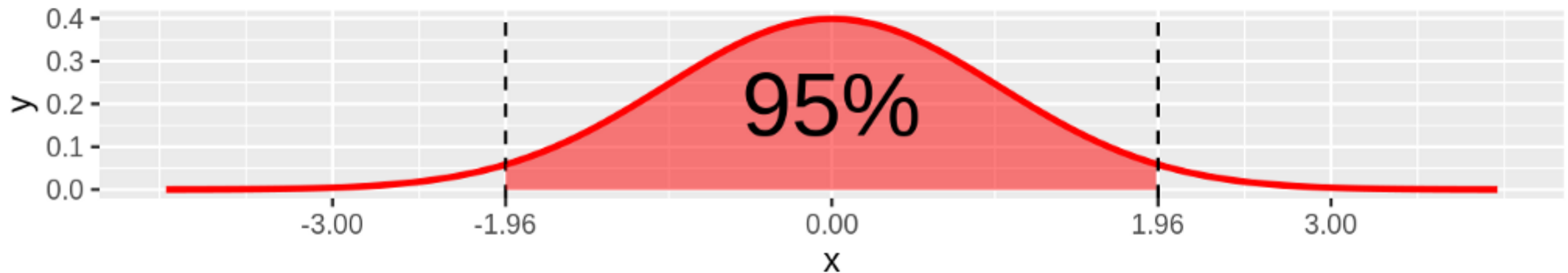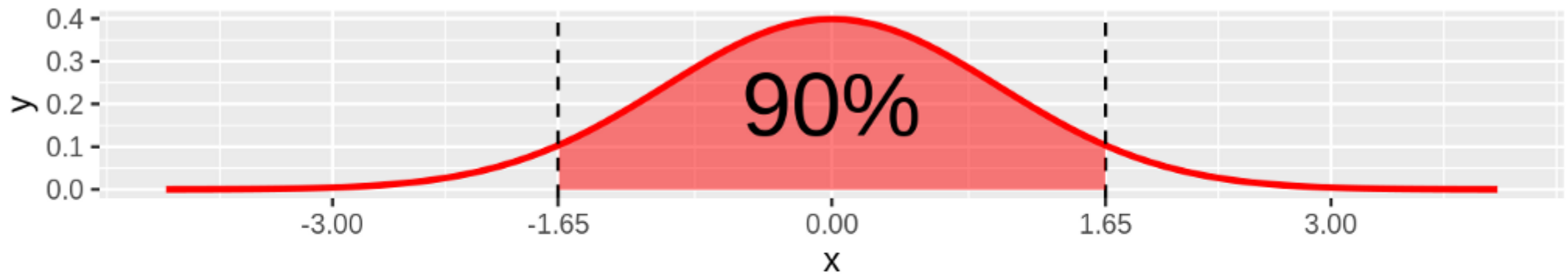
%90 confident of the "Estimate ±1.65∗SE(Estimate)"

%95 confident that the "Estimate ±1.96∗SE(Estimate) = (x−1.96σ/√n, x+1.96σ/√n)

% 99 confident that the " "Estimate ±2.58∗SE(Estimate)"

Significance Level

The probability of rejecting Ho when you shouldn't do.

Power

the likelihood of rejecting the null hypothesis when it should be rejected.

Accuracy

How close is your estimation to the true value?

Precision

It is a quality measurement, how close will your two measurements be to each other.

Statistical Inference

Finding a conclusion from your data by using statistics.

Type 1 Error

It occurs when rejecting a null hypothesis but it should not have to be rejected.

Type 2 Error

It occurs when the null hypotheses should be rejected but you don't.

T Distribution

Comparing means while standard deviation is unknown.

Degrees of Freedom

Determines thickness of tails.

As the degrees of freedom increase, the shape of the t-distribution approaches the normal distribution.

Conclusion

There are too many terms that existed which I maybe write another article in near future.