**Simple Explanations of Basic Statistics Concepts (Part 1)**

Simple and clear explanations of different statistical concepts

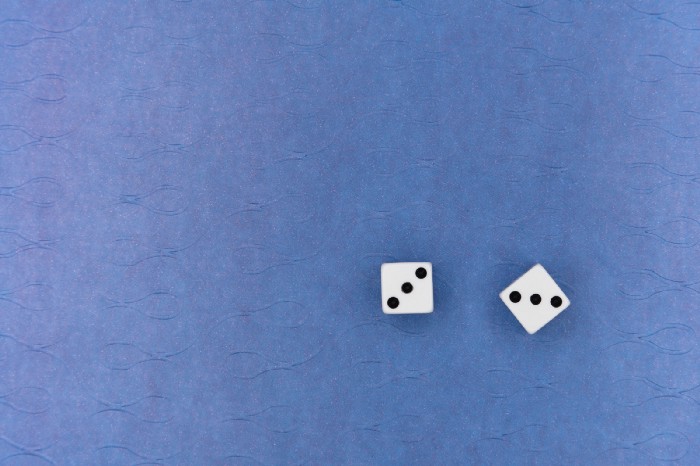


Photo by [Markus Winkler](https://unsplash.com/@markuswinkler?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral)

**Introduction**

My professor once mentioned that if you can explain a term to someone who is completely unfamiliar with it and have them get the concept, then you are regarded to be knowledgeable about the subject. This method is also the approach I've taken whenever I've learned something new. As I am reviewing interview questions, I realize how important it is to practice this active way of learning.

There are a lot of definitions and a lot of knowledge to cover in Statistics. That's why I think the best way to do a review is to write and explain it to people. With this idea in mind, this article will discuss some fundamental statistical concepts that I believe any statistics learner should be familiar with.

**Population and Sample**

Suppose my family owns a lemon farm. When harvest season comes, we want to know the average size of lemons. There are thousands of lemons on the farm, and of course, we cannot measure or weigh all of them. So, we decide to take samples of them and get the average from each sample.

All lemons on the farm are called **population,**and each group of lemons picked is a**sample.**

**Sample size**

It is the number of lemons included in your sample. The **sample size** is determined depending on different criteria such as research scope, population variability, or research methodology.

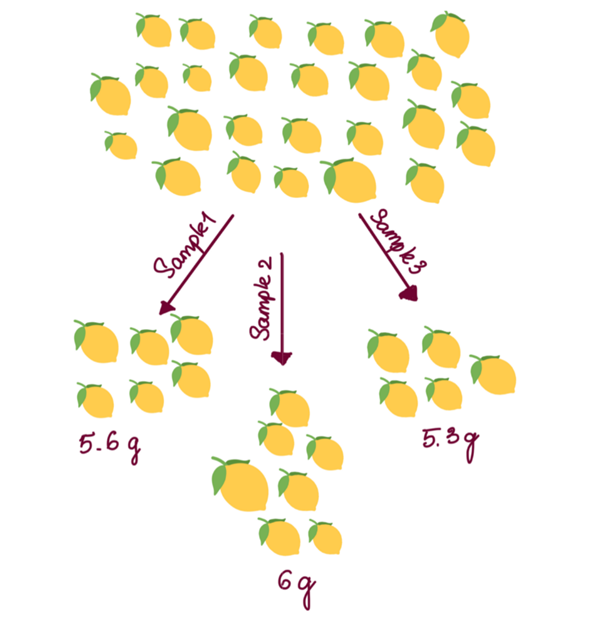


Image by Author

**Sample error**

From the samples, we can draw conclusions about the population. Such findings are called inferences. However, as we may know, samples will never be a perfect representation of the population, and each sample will have a different result.

There will always be variations in sampling. For example, three groups of lemons picked on the same farm might have three different average weights. This is called sampling error.

**Sample frame**

The**sample frame** can be understood as a list of the individuals of the population from which a sample will be taken. Ideally, the list should include the entire population you want to study. The difference between the **sample frame** and the **target population** is that the population is broad, and the frame is narrow.

For example, I want to research work satisfaction at company BC. The total number of employees at company BC is the population for the study. Ideally, my sampling frame will include the names and information of all employees.

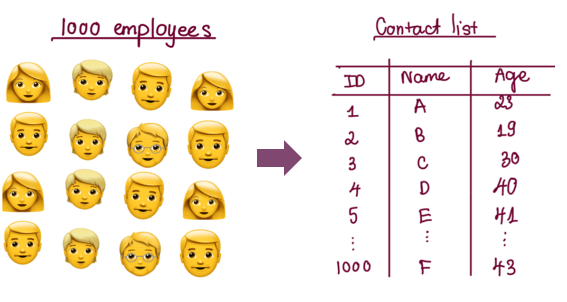


Image by Author

**Sample Space**

A **sample space**contains all possible outcomes of an experiment. Suppose you have a sampling frame of employees in company BC for a survey to select the best support team. The**sample space, in this case,** is all possible responses from your survey: sales team 1, sales team 2, HR, finance, etc.

**Sampling Methods**

Meanwhile, the action of taking samples from a population is called **sampling**.

There are two ways of sampling: **Probability sampling & Non-probability sampling**. However, in this post, I will only cover **Probability sampling**since it is more familiar to people and preferred to use in research.

In short, **Probability sampling**is a technique in which a researcher will predefine a few criteria and then randomly select individuals from a population. All members have an equal chance of being included in the sample with these criteria. There are four main types of **Probability sampling:**

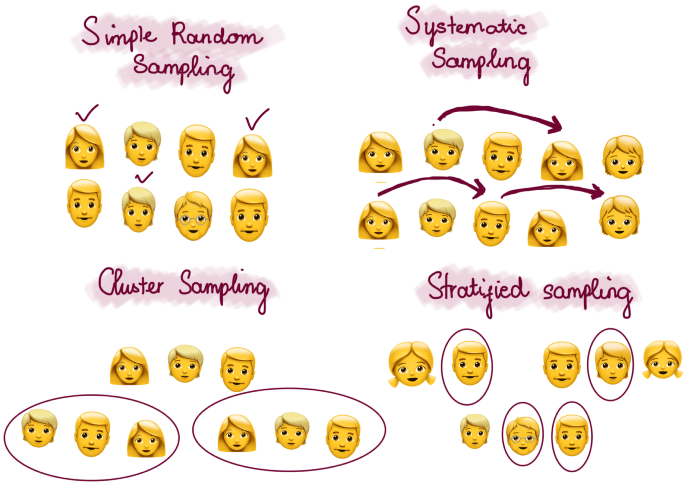


Image by Author

**1. Simple random sampling**

This is one of the best sampling methods since it helps to save time and resources. In a simple random sampling method, the sampling frame should contain the entire population so that each individual of the population has an equal opportunity of being picked.

For example, company BC wants to select 50 employees from their 1000 employees to do a survey. They assign each employee a number from 1 to 1000; then, they randomly choose 50 numbers from the pool of 1000 numbers.

**2. Systematic sampling**

This sampling method is simple and similar to simple random sampling, but the difference is individuals are picked at fixed intervals rather than being generated at random.

Continue to use the previous example to select employees for the survey. Assuming that the number is ordered in descending order, we first randomly select a starting point, for example, employee number 2. Then, from number 2 onwards, you choose the number greater than the previous selection of 2 units. That means employee numbers 2,4,6,…will be selected until there are enough 50 people.

**3. Cluster sampling**

Now, suppose company BC has 50 branches worldwide with thousands of employees; the company's top management cannot visit every office. So, they decided to choose ten cities to visit randomly, and these ten cities correspond to ten clusters.

This approach has a higher chance of sample error since there might be significant disparities across clusters. Also, the sample clusters are not sure to represent the whole population well.

**4. Stratified sampling**

The total population is divided into different subgroups (strata) based on similar characteristics such as income, gender, or level of education. Then, we decide how many people we sample from each group and take samples.

An obvious disadvantage of stratified sampling is the higher chance of selection bias when defining strata. This bias happens because we might have prior knowledge or information of the population’s shared patterns beforehand to divide the subgroups.

**Confidence Interval**

As I said above, there will always be sample errors in sampling since there are no samples that perfectly represent the entire population. So, how do we get the results for the population’s parameters? That’s when the term **confidence interval**comes in.

The **confidence interval** helps to express an estimate of the population’s parameters. In other words, it describes the likelihood of our parameters being accurate.

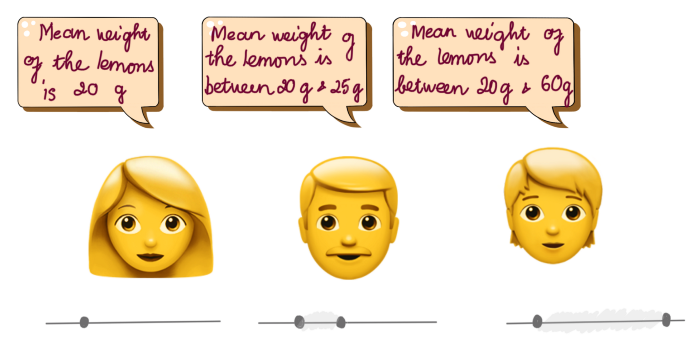


Image by Author

In statistics, **confidence** is another expression for probability. For example, suppose you establish a confidence interval with a 95 percent confidence level. In that case, you are confident that the estimate will lie between the upper and lower values given by the confidence interval 95 times out of 100 times.

Let’s come back to calculating the average weight of lemons on the farm. Suppose we take a sample of ten lemons and calculate that sample's mean weight. The result is 20 grams.

Take another sample and continue to find the sample’s mean weight. Repeat the procedure until we calculate many mean weights, say 100 mean weights. Out of 100 calculations, 95 values lie between 20g–40g. So, we say that we are 95% confident that the mean weight of lemons falls between 20g and 40g.

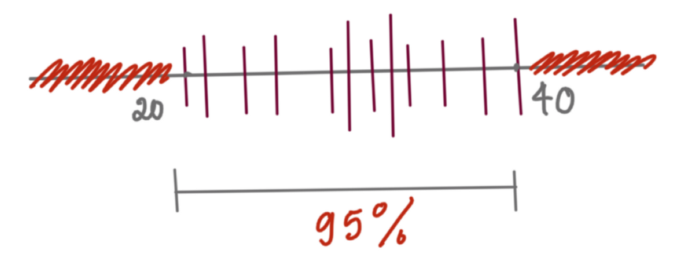


Image by Author

***Long story short, a 95% interval is just an interval that covers 95% of the mean. So, anything outside the confidence interval only happens less than 5%. It also means the probability of the true mean falling in this area is < 5%.***

**What affects the width of the confidence interval**

There are two factors impacting the confidence interval's width: **variation** within a population and the **sample size.**

Imagine the lemon farm where all the lemons have the same weights, the samples we draw from this population should have similar weights. Therefore, the mean weights are **similar between samples**, leading to a **narrow confidence interval**. However, in another case, when there is a **large variation**between lemons’ weights on the farm, there might be a **vast difference**between samples’ mean weights, resulting in a **wide confidence interval**of mean weight.

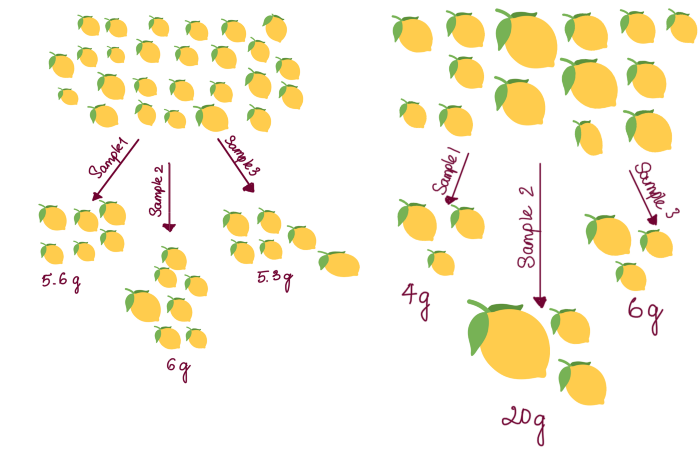


Image by Author

If we take a small sample, the information will not be accurate since there is not much information to reflect from a few values. So, the chance of generating big variation between samples will be higher, leading to a wider confidence interval.

On the other hand, the effect of getting sample error is reduced with large samples since there is more information to derive from. Thus, the confidence interval then will be narrower.

**How to calculate the confidence interval**

There are several methods to calculate confidence interval:

* Bootstrapping
* Informal
* Traditional Normal-based

However, the details of each method will be mentioned in the following posts.

**Conclusion…**

Above are some explanations of some basic concepts in statistics. Of course, there are still many more to cover, and I will continue to share them in my next posts. I hope I have made them a little bit clearer. Thank you for reading till the end.