Statistics 101

Proudly built with R, Quarto and GitHub Copilot

Daniel Sánchez Pazmiño

Table of contents

Sampling distributions and intro to hypothesis testing **1**
A simulated population to make statistical inference 2
Sampling from a population 2
Sampling distributions 3
Type of the sampling distribution and the CLT 5
Exercises using sampling distributions 6
Reliability and confidence intervals 7
Case I: population standard deviation is known 7
Case II: population standard deviation is unknown 10
Summary 12
Up next 13

Sampling distributions and intro to hypothesis testing

In the previous sessions, we were able to describe the distribution of a variable using the mean and the standard deviation. We also learned how to use the normal distribution to calculate probabilities. In this session, we will learn how to use the normal distribution to make inferences about the population mean.

The typical exercise in using a sample to make a guess, or *infer* about the population parameters is that you draw a sample from the population and compute a descriptive statistic, such as the sample mean, and hold it as a guess, or estimate, for the population parameter, which would the population mean. The key idea behind the *estimate* is that we do not expect that it is *exactly* the value of the true population mean. If somehow we had access to data for the whole population, we would be able to compute the population mean, and most likely our sample estimate will be different to the true population parameter.

A simulated population to make statistical inference

To best illustrate the idea of sampling and sampling distributions, I will make use of a population with simulated data. This means that I will create a population and have access to the data of the whole population. This is not a realistic scenario, but it will help us to understand the concepts of sampling and sampling distributions. The population I will have access to will be a population of students who report their GPA and their family income, as well as their race. See an extract of the table below:

Table 1: Population of students

id	gpa	finc	race
1	2.137880	2421.238	White
2	2.583229	5760.158	Latino
3	2.825332	7073.192	White
4	1.796291	1901.171	White
5	2.628737	2773.175	White
6	2.651817	6558.177	White

This table has 100,000 rows, which means that we have data for 100,000 students. This is a simulated population, which means that we have access to the data of the whole population. In real life, we do not have access to the data of the whole population, but we have to draw a sample from the population and use the sample data to make inferences about the population parameters. By using a simulated population, we are able to compare our sample estimates with the true population parameters, so that it is more pedagogical.

Sampling from a population

Consider the descriptive statistics for our two continuous or numeric variables.

Table 2: Population of students

gpa	finc
Min. :1.147	Min. : 479.7
1st Qu.:2.299	1st Qu.: 3516.3
Median :2.501	Median : 4910.9
Mean :2.501	Mean : 5572.8
3rd Qu.:2.702	3rd Qu.: 6880.4
Max. :3.672	Max. :41128.2

We also know that the standard deviations for these variables will be 0.3 and 2965.609, respectively. Because these are statistics that we calculate with the data of the whole population, we call them *population parameters*. The population mean GPA is 2.501 and the population mean family income is 5572.80.

We can sample from this population and calculate the sample mean GPA and the sample mean family income. One of many ways of sampling is doing a *simple random sample*. This means that we randomly select a number of observations from the population. The fact that it is random means that the selection of the elements for the sample is done indepedently of the values of the variables, and of each other. We can do a simple random sample of 100 students from the population of 100,000 students, and compute the sample mean GPA and the sample mean family income.

Table 3: Sample of 100 students

gpa	finc
Min. :1.885	Min. : 1240
1st Qu.:2.257	1st Qu.: 3607
Median :2.437	Median : 4964
Mean :2.487	Mean : 5588
3rd Qu.:2.687	3rd Qu.: 6560
Max. :3.299	Max. :22372

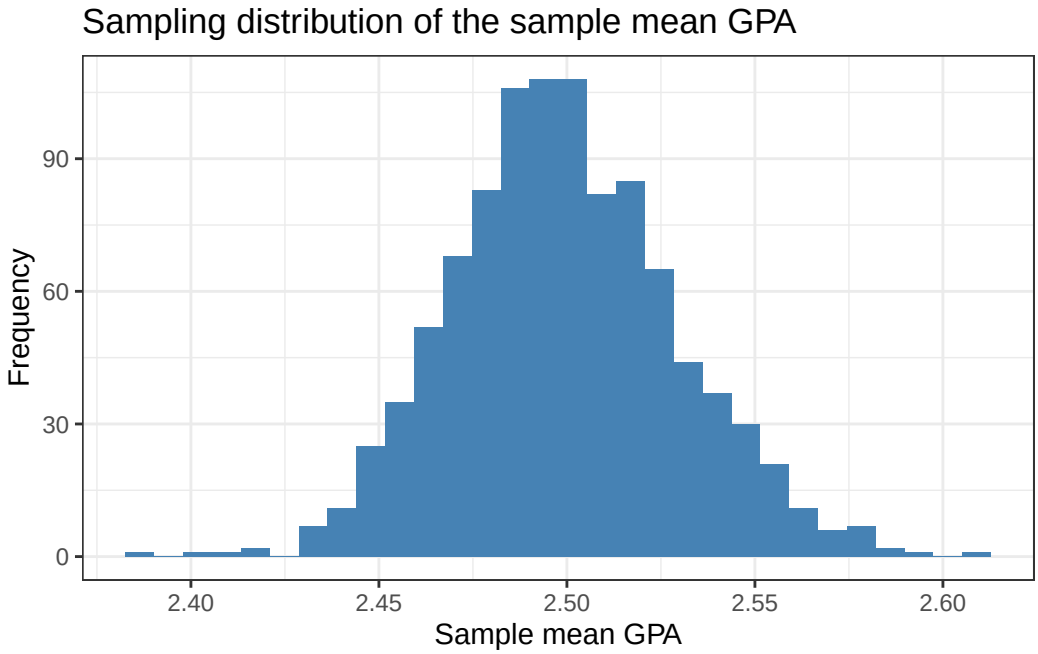
As you can see, the sample mean GPA is 2.487 and the sample mean family income is 5588. Further, we can calculate the standard deviations of the sample variables, which are 0.304 and 3077.632, respectively.

These descriptive statistics for the sample are often called *sample statistics*. They are thought to be estimates of the population parameters, and the technical wording about the numbers that are the estimates is *point estimates*. They are different to *interval estimates*, which are estimates that include a range of values, which we will discuss afterwards. Notice how the sample statistics are different to the population parameters, and that if I had done another, different simple random sample, the sample statistics would have been different.

Sampling distributions

What do you think would happen if we took a sample of 100 students, and then another sample of 100 students, and then another sample of 100 students, and so on, until we have taken 1000 samples of 100 students each? Well, we would definitely have 1000 different sample statistics, each with a different value, some of them bigger, some of them smaller.

Under that scenario, we could theoretically plot the distribution of all of the sample means in a histogram. I do that below with the simulated data.



If we consider the exercise of taking a sample from the population as a *random experiment*, we can use the rules of probability to understand how the sample statistics are distributed. When considering sampling as a probability experiment, the sample mean will become a random variable, as we are not sure what value it will take. The distribution of the sample statistics is called the *sampling distribution*, and this distribution is also a probability distribution.

The sampling distribution of the sample mean is a probability distribution that has an expected value and a standard deviation. The mean of the sampling distribution of the sample mean is the same as the population mean. The standard deviation of the sampling distribution of the sample mean is called the *standard error of the mean*, and it is calculated as the standard deviation of the population divided by the square root of the sample size. The standard error is calculated with a given formula: the standard deviation of the population variable divided by the root of the sample size. We put that in writing as follows:

$$E(\bar{x}) = \mu_x$$

$$s(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

The formula for the standard error of the sample mean is most commonly the one that I gave above, however, there are other formulas that are used in different contexts. Specifically, when we have a population which is not so big so that we are able to take a sample which accounts for over 5% of the population, we say that the population is *finite*. In that case, the standard error of the sample mean is calculated as the standard deviation of the population variable divided by the root of the sample size, multiplied by the square root of the population size minus the sample size, divided by the population size minus one. We put that in writing as follows:

$$s(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

I will not use this formula here for the standard error. This is because only very rarely we can acquire such a large sample for the assumption of a finite population to be true.

Note that in order to calculate the standard error of the mean, we need to know the standard deviation of the population variable. In real life, we do not know the standard deviation of the population variable, so there are two choices: we either estimate it with past data, or we use the sample standard deviation. The latter is the most common choice, but it will create some issues that we will discuss later.

Type of the sampling distribution and the CLT

It is often said that the sampling distribution "inherits" the properties of the population distribution. This means that if the population distribution is normal, then the sampling distribution will be normal. If the population distribution is not normal, then the sampling distribution will not be normal. In the latter case, we need to consider the central limit theorem (CLT).

The most important idea of the CLT is that the sampling distribution of the sample mean will be normal if the sample size is large enough. The CLT is a very important theorem in statistics, and it is the reason why we can use the normal distribution to make inferences about the population mean.

What sample size n is large enough? Well, it depends on the shape of the population distribution. If the population distribution is normal, then the sampling distribution of the sample mean will be normal for any sample size. If the population distribution is not normal, then the sampling distribution of the sample mean will be normal if the sample size is large enough. Large enough means that the sample size is greater than 30 for most applications, however, there are some cases where the sample size needs to be greater than 50, or even greater than 100. The general idea is that the sample size needs to be as large as possible to have better results, which is why you rarely see small samples in top publications.

There is another thing that we must note about the CLT. The CLT says that the sampling distribution of the sample mean will be normal if the sample size is large enough, *regardless*

of the shape of the population distribution. This means that if the population distribution is not normal, the sampling distribution of the sample mean will be normal if the sample size is large enough.

Exercises using sampling distributions

The concept of the sampling distribution is very important because it will allow us to construct the idea of confidence intervals and hypothesis testing. However, even by itself, the sampling distribution is a very useful concept which, along with the ideas of probability distributions, will allow us to answer many questions. Here, I present you with some exercises that will help you to understand the concept of the sampling distribution.

Remember the dataset of the population of students that we created at the beginning of the session? For GPA, we had that the standard deviation of the population was 0.3. Further, assuming that the population is infinite and that we will sample with size $n = 100$, we can use this information to calculate the standard error of the sample mean, as follows:

$$s(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.3}{\sqrt{100}} = 0.03$$

So, knowing the mean and the standard error of the sampling distribution, we can calculate the probability of getting a sample mean GPA of 2.5 or higher. We can do this by calculating the z-score of the sample mean GPA of 2.5, and then calculating the probability of getting a z-score of 0. This is done as follows:

$$z = \frac{\bar{x} - \mu_x}{\sigma_{\bar{x}}} = \frac{2.5 - 2.501}{0.03} = -0.033$$

$$P(z \leq -0.033) = 0.484$$

This means that the probability of getting a sample mean GPA of 2.5 or higher is 0.484. This is a very high probability, which means that it is very likely that we will get a sample mean GPA of 2.5 or higher.

You should try to do these exercises on your own:

1. What is the probability of getting a sample mean GPA of 2.4 or higher?
2. What is the probability of getting a sample mean GPA of 2.6 or higher?
3. What is the probability of getting a sample mean GPA of 2.7 or lower?
4. What is the probability of getting a sample mean GPA of 1.5 or lower?
5. What is the probability of getting a sample mean GPA of 2.5 or higher if the sample size is 50?

6. What is the probability of getting a sample mean GPA of 2.5 or higher if the sample size is 200?
7. What is the probability of getting a sample mean between 2.4 and 2.6 using $n = 100$?
8. What is the probability of getting a sample mean between 2.4 and 2.6 using $n = 50$?
9. What is the probability of getting a sample mean between 2.4 and 2.6 using $n = 200$?
10. What is the probability I draw three samples and get a sample mean GPA of 2.5 or higher in all of them?

These are simple exercises that use the ideas of the sampling distribution and the normal probability distribution which we covered last session. The most important difference is that instead of using mean and the standard error of the distribution of the variable for the z-score, we need to use the mean and the standard error of the sampling distribution of the sample mean. This implies that we know the population mean and the population standard deviation as well as a sample size.

Reliability and confidence intervals

The concept of reliability relates to the idea that when we are using a sample to make inferences about the population, we are not sure that our sample statistics are exactly the same as the population parameters. We are not sure that our sample mean is exactly the same as the population mean. However, using the knowledge of sampling distributions that we just acquired, we can say that our estimate is reliable if it falls in an interval for which we are sure that the population mean is in.

To do this interval, we need to present our estimate in the following mannner:

$$\text{Point estimate} \pm \text{Margin of error}$$

We know how to calculate the point estimate, but we don't know how to calculate the margin of error, which is something we will do here.

Case I: population standard deviation is known

In the simplest case, we know what is the standard deviation of the population variable. This is not a realistic case, but it is useful to understand the concept of confidence intervals. There might be times where we can have a reliable estimate of the population standard deviation, for example, if we have data from the past.

We can present the margin of error of the sample that we calculated in the previous section. We know that the sample mean for the first sample we acquired was 2.487, and we know that the standard error of the sample mean was 0.03. Our sample size was $n = 100$, so we can be sure that the sampling distribution is normal.

The margin of error has an embedded concept of probability. We are unsure of what is the exact value of the population mean (not here, because we simulated a population, but generally that is the way it is), but, using the sample distribution, we can be sure that there is a certain probability to get a sample mean within a certain range of values. How do we calculate that probability? In the same way that we calculated the probabilities in the previous section, using the standard normal table.

What is the probability that any random sample mean will be within 1 standard deviation of the population mean? Well, we can remember the empirical rule. The probability that any random sample mean will be within 1 standard deviation of the population mean is 0.6827. This means that we can be sure that any population mean is within 1 standard deviation of the sample mean with a 68.27% probability, if the sampling distribution is normally distributed.

For calculating the margin of error, there is a consensus among the sciences that we should use a 95% probability. This means that we want to be sure that, with 95% probability (or 0.95 probability), the population mean is within a certain range of values. So, we would need to know what is the number of standard deviations for which we have a 95% probability. Usually, you're given the number of standard deviations and you need to calculate the probability, but here we are doing the opposite. How do we do it? Thankfully, there is a process towards doing this with a table (as a computer would do it instantly).

If we want to know within how many standard deviations we have a 95% probability, first define the α (alpha parameter) of the probability. $\alpha = 1 -$ confidence level in proportion form. So, if we want a 95% confidence interval, $\alpha = 1 - 0.95 = 0.05$. Then, we need to divide α by two, and then look for the number of standard deviations which produce that probability in a "greater than form". The "greather than" form corresponds to the "smallaer portion" of the Field probability table, and is usually called the right tail probability.

For a 95% probability, we have $\alpha/2 = 0.025$ and the answer is 1.96 standard deviations. This means that we can be sure that the population mean is within 1.96 standard deviations of the sample mean with 95% probability. So, we will be sure that 95% of the sample means that are calculated with different samples will be within 1.96 standard deviations of the population mean.

We calculate margins of error using that knowledge, as follows:

$$\text{Margin of error} = 1.96 \times \text{Standard error of the sample mean}$$

For our example, the margin of error will be

$$\text{Margin of error} = 1.96 \times 0.03 = 0.0588$$

The interpretation of the margin of error is often misunderstood. We do not say that the population mean is within the margin of error of the sample mean. Why? Well, what do you

think would happen if we draw another sample? Well, the sample mean would be different, and so would be the range within the margin of error. So, we cannot say that the population mean is within the margin of error of the sample mean.

What we say is that in 95% of the margin of errors that we calculated, the population mean will be contained inside. Though it sounds like it is just a semantic and not statistical difference, this is a very important difference. We are not saying that the population mean is within the margin of error of the sample mean, but that the population mean is within the margin of error of the sample mean in 95% of the cases. Often, we shorthand this by saying that we are 95% confident that the population mean is within the margin of error of the sample mean. Never, ever, forget this interpretation.

What is the confidence interval? Simply, the interval that is defined by the point estimate plus or minus the margin of error. In our example, the confidence interval is

$$2.487 \pm 0.0588 = (2.4282, 2.5458)$$

This means that we are 95% confident that the population mean is within the interval (2.4282, 2.5458). The confidence interval which corresponds to using a 95% probability is called the 95% confidence interval (C.I.). We can use other probabilities, such as 90%, 99%, 99.9%, etc. The most common confidence intervals are the 90%, 95% and 99% confidence intervals. It is useful to remember the way to calculate the 90%, 95% and 99% confidence intervals, as follows:

$$\text{90\% C.I.} = \text{Point estimate} \pm 1.645 \times \text{Standard error of the sample mean}$$

$$\text{95\% C.I.} = \text{Point estimate} \pm 1.96 \times \text{Standard error of the sample mean}$$

$$\text{99\% C.I.} = \text{Point estimate} \pm 2.576 \times \text{Standard error of the sample mean}$$

Notice how for every C.I., we have a different amount of standard deviations. That comes from the shape of the normal distribution. The more standard deviations we have, the more probability we have. So, if we want to be more confident, we need to have a bigger interval.

Case II: population standard deviation is unknown

In the previous section, we assumed that we knew the standard deviation of the population variable. This is not a realistic assumption, so in this section, we will assume that we do not know the standard deviation of the population variable. This is the most common case, and it is the one that we will use the most in the future.

In this case, we need to use the sample standard deviation as an estimate of the population standard deviation. This means that we will use the sample standard deviation to calculate the standard error of the sample mean, and for the margin of error as well.

Let's use the sample that we took at the beginning of the session. The sample mean GPA was 2.487, and the sample standard deviation was 0.304. The sample size was $n = 100$.

The formula to calculate the margin of error is, in fact, similar. The only change the we must do is that we must look for the number of standard deviations for which we have a certain probability using the t distribution, not the normal one. The t distribution is a probability distribution that is similar to the normal distribution, but it depends on an additional parameter, the sample size. The t distribution is used when we do not know the standard deviation of the population variable, and it is used to calculate the margin of error.

For smaller sample sizes, the t distribution is more spread out than the normal distribution. This means that the t distribution has more probability in the tails than the normal distribution. This is because when we have a smaller sample size, we have more uncertainty about the population parameter, so we need to have a bigger interval. However, when we have large sample sizes, the t distribution *converges* to the normal distribution.

We can use a table for the t distribution as well. We need to know the probability which produces the number of standard deviations that interest us, and the degrees of freedom. The degrees of freedom are calculated as the sample size minus one. For our example, the degrees of freedom in our example are 99.

Once we have that, we can calculate the number of standard deviations by again using the α parameter of the confidence level that we want. So, if we want a 95% confidence interval, $\alpha = 1 - 0.95 = 0.05$. Then, we need to divide α by two, and look for the probability that corresponds to that value in the t distribution table.

For our example, $\alpha = 1 - 0.96 = 0.05$, and $\alpha/2 = 0.025$. The probability that corresponds to 0.025 and 99 degrees of freedom is 2.626 according to the t distribution table with $df = 99$. This means that we need to be 2.626 standard deviations away from the population mean to have a 95% confidence interval. We then calculate the margin of error as follows:

$$\text{Margin of error} = 2.626 \times \text{Standard error of the sample mean}$$

$$\text{Margin of error} = 2.626 \times \frac{s_x}{\sqrt{n}} = 2.626 \times \frac{0.304}{\sqrt{100}} = 0.080$$

The confidence interval is then calculated as follows:

$$\text{Confidence interval} = \text{Point estimate} \pm \text{Margin of error}$$

$$\text{Confidence interval} = 2.487 \pm 0.080 = (2.407, 2.567)$$

This means that we are 95% confident that the population mean is within the interval (2.407, 2.567). Remembering that the population mean is indeed 2.501, we are sure that we correctly made inference about it, even without using the standard deviation of the population. Statistics works!!!

We can produce a general formula for the margin of error, as follows:

$$\text{Margin of error} = t_{\alpha/2,df} \times \text{Standard error of the sample mean}$$

Using this same notation, we can produce a general formula for the confidence interval, as follows:

$$\text{Confidence interval} = \text{Point estimate} \pm \text{Margin of error}$$

$$\text{Confidence interval} = \text{Point estimate} \pm t_{\alpha/2,df} \times \text{Standard error of the sample mean}$$

This is the general formula for the confidence interval. The only thing that changes is the value of the t distribution, which depends on the degrees of freedom and the probability. The degrees of freedom are calculated as the sample size minus one. The probability is calculated as $\alpha/2$, where $\alpha = 1 -$ confidence level in proportion form.

Summary

- The value of a sampling statistic is called a point estimate of the population parameter. It varies from sample to sample.
- The sampling distribution is the distribution of a sample statistic. The most important sample statistic is the sample mean.
- A sampling distribution has a mean and a standard deviation. The mean of the sampling distribution is the same as the population parameter. The standard deviation of the sampling distribution is called the standard error of the sample mean, and it is calculated as the standard deviation of the population variable divided by the square root of the sample size.
- The sampling distribution inherits the properties of the population distribution. If the population distribution is normal, then the sampling distribution will be normal. If the population distribution is not normal, then the sampling distribution will not be normal. In the latter case, we need to consider the central limit theorem (CLT).
- The CLT says that the sampling distribution of the sample mean will be normal if the sample size is large enough, regardless of the shape of the population distribution.
- We can calculate the probability of the sample mean being a certain value by using the standard error of the sample mean to perform inference with a standard normal table.
- The margin of error is the number of standard deviations that we need to be away from the population mean to be have a certain probability of happening. The margin of error is calculated as the number of standard deviations times the standard error of the sample mean.
- We commonly calculate the margin of error for a 95% probability. This means that we want to be sure that, with 95% probability, the population mean is within a certain range of values. We calculate the margin of error by looking for the number of standard deviations for which we have a 95% probability using a standard normal table.
- We get the number of standard deviations for a particular percent of the sampling mean possibilities by using the α parameter of the probability. $\alpha = 1 -$ confidence level in proportion form. So, if we want a 95% confidence interval, $\alpha = 1 - 0.95 = 0.05$. Then, we need to divide α by two, and then look for the number of standard deviations which produce that probability in a "greater than form". The "greather than" form corresponds to the "smaller portion" of the Field probability table, and is usually called the right tail probability.
- The confidence interval is the interval that is defined by the point estimate plus or minus the margin of error. The confidence interval is calculated as the point estimate plus or minus the number of standard deviations times the standard error of the sample mean.
- There are two different cases to present a C.I. The first is when we know the standard deviation of the population. In such case, we assume that the sampling distribution is normal.
- When we don't know the standard deviation of the population, we use the sample standard deviation as an estimate of the population standard deviation. In such case, we

assume that the sampling distribution is a t distribution.

- The t distribution is a probability distribution that is similar to the normal distribution, but it depends on an additional parameter, the sample size. The t distribution is used when we do not know the standard deviation of the population variable, and it is used to calculate the margin of error.

Up next

- Hypothesis testing
- Z tests, T tests, independent, dependent samples.
- ANOVA, Correlation, Regression
- SPSS I
- SPSS II
- Non parametric statistics