Sampling distribution

Nikos Apostolakis

April 28, 2023

1 The Sampling Distribution

We now entering the area of *Inferential Statistics*. Recall that we are usually interested in a population and some *parameter*, such as the *means* μ, the *standard deviation* σ, or the *proportion*[1]. Recall also, that for each parameter, a measurement on the whole population, there is a corresponding *statistic*,a measurement of a *Sample* with the same name but denoted with a different symbol.

Name	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	p	$\widehat{p}$

Mean and standard deviation for the sample mean

Let $\overline{X}$ be the sampling distribution for the mean, for samples of size n, drawn from a population X with mean μ and standard deviation σ. Then

$$\mu_{\overline{X}} = \mu, \tag{1}$$

and

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}. \tag{2}$$

Notice that the mean of the sampling distribution is the same as the mean of the population. This makes sense, the averages we compute from a sample are centered around the actual mean of the population. In other words, the *expected value* of the sample mean, is the mean of the population.

The standard deviation of the sampling distribution though is smaller, because of the square root of the size in the denominator. For example if we consider samples of size $n = 16$, we have $\sqrt{n} = 4$ and so the standard deviation of the sampling distribution will be a quarter of the standard deviation of the population. And the largest the size of the sample the smaller $\sigma_{\overline{X}}$ is, in other words, the larger the sample size the less spread the sampling distribution is.

[1] We will talk more about proportion later.

Example 1

Random samples of size 225 are drawn from a population with mean 100 and standard deviation 20. Find the mean and standard deviation of the sample mean.

Answer. The mean of the sample mean $\overline{X}$ is the same as the mean of the population. So

$$\mu_{\overline{X}} = 225.$$

For the standard deviation we have

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{225}} = \frac{20}{15} = 1.33.$$

□

Example 2

The mean and standard deviation of the tax value of all vehicles registered in a certain state are $\mu = \$13,525$ and $\sigma = \$4,180$. Suppose random samples of size 100 are drawn from the population of vehicles registered in that state.

(a) What are the mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of the sample mean $\bar{X}$?

(b) Use Chebyshev's Theorem to find an interval that contains at least 88.89%

1. of the tax values of the vehicles registered in that state,
2. of the mean tax value for random samples of size 100 drawn from the population of vehicles vehicles registered in that state.

Answer. We have

(a) Since $n = 100$, the formulas yield

$$\mu_{\bar{X}} = \mu = \$13,525$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4,180}{\sqrt{100}} = \$418$$

(b) By Chebyshev's Theorem 88.89% of the any population has z–scores between -3 and 3. So we calculate the raw scores that correspond to $z = \pm 3$ in each case.

1. Since $\sigma = 4,180$ we have that $3\sigma = 12,540$, and so

$$z = -3 \implies x = 13,525 - 12,540 = 985, \quad z = 3 \implies x = 13,525 + 12,540 = 26065.$$

So the interval that captures 88.89% of the tax values of the vehicles is $(\$985, \$26,065)$.

2. From Part (a) we know that $\sigma_{\overline{X}} = 418$, ans therefore $3\sigma_{\overline{X}} = 1254$. So,

$$z = -3 \implies \bar{x} = 12540 - 1254 = 11286, \quad z = 3 \implies \bar{x} = 12540 + 1254 = 13794.$$

So the interval that captures 88.89% of the mean tax value of random samples of size 100 is $(\$11286, \$13794)$.

□

Notice how much narrower the interval for the sample mean is than the interval for an individual value.

2 Sampling distribution for normally distributed populations

If the population we draw from is normally distributed then the sampling distribution for the mean $\overline{X}$ is also normally distributed.

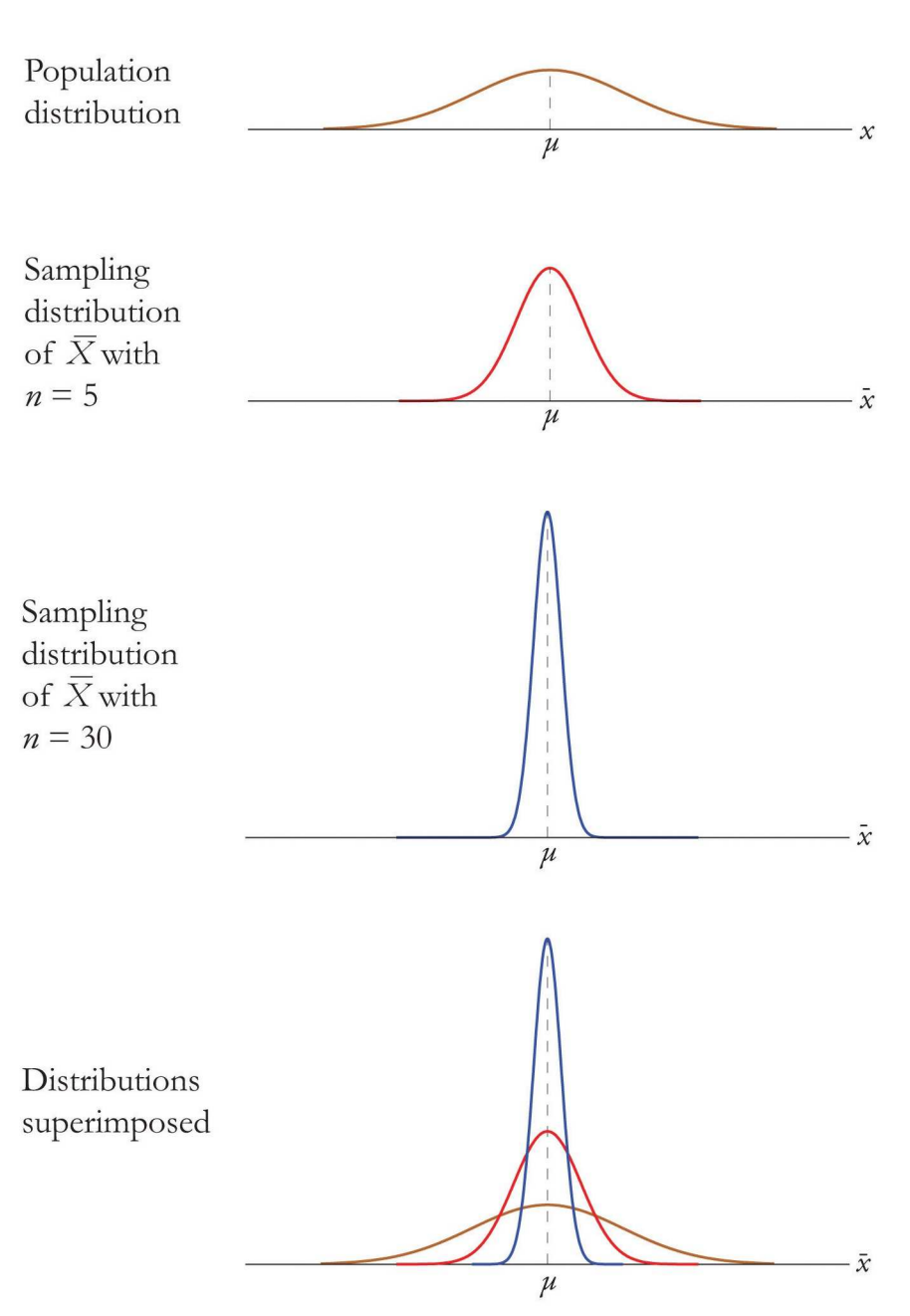

Figure 1: Sampling distributions from a normal population

Notice that the larger the sample size n the narrower the normal curve, i.e. the smaller the standard deviation $\sigma_{\overline{X}}$. This is consistent with Equation (2): since we divide σ by $\sqrt{n}$, the larger the n is, the smaller $\sigma_{\overline{X}}$ will be. The mean and standard deviation of $\overline{X}$ are given by Equations (1) and (2) respectively, so $\overline{X}$ has the same mean as X but smaller standard deviation so its distribution is less spread. The larger the sample size the less spread the distribution of $\overline{X}$. See Figure 1.

Sample mean for normal populations

If X is normally distributed with mean μ and standard deviation σ then $\overline{X}$, the sample mean is also normally distributed.

Example 1

A prototype automotive tire has a design life of 38,500 miles with a standard deviation of 2,500 miles. Five such tires are manufactured and tested. On the assumption that the actual population mean is 38,500 miles and the actual population standard deviation is 2,500 miles, find the probability that the sample mean will be less than 36,000 miles. Assume that the distribution of lifetimes of such tires is normal.

Answer. For simplicity we use units of thousands of miles. Then the sample mean $\overline{X}$ has mean $\mu_{\overline{X}} = \mu = 38.5$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{2.5}{\sqrt{5}} = 1.11803$. Since the population is normally distributed, so is $\overline{X}$, hence

$$\begin{aligned} P(\overline{X} < 36) &= P\left(Z < \frac{36 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z < \frac{36 - 38.5}{1.11803}\right) \\ &= P(Z < -2.24) \\ &= 0.0125 \end{aligned}$$

That is, if the tires perform as designed, there is only about a 1.25% chance that the average of a sample of this size would be so low. □

Example 1

An automobile battery manufacturer claims that its mid-grade battery has a mean life of 50 months with a standard deviation of 6 months. Suppose the distribution of battery lives of this particular brand is approximately normal.

(a) On the assumption that the manufacturer's claims are true, find the probability that a randomly selected battery of this type will last less than 48 months.

(b) On the same assumption, find the probability that the mean of a random sample of 36 such batteries will be less than 48 months.

Answer. (a) Since the population is known to have a normal distribution

$$\begin{aligned} P(X < 48) &= P\left(Z < \frac{48-\mu}{\sigma}\right) \\ &= P\left(Z < \frac{48-50}{6}\right) \\ &= P(Z < -0.33) \\ &= 0.3707 \end{aligned}$$

(b) The sample mean has mean $\mu_{\overline{X}} = \mu = 50$ and standard deviation

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1.$$

Thus

$$\begin{aligned} P(\overline{X} < 48) &= P\left(Z < \frac{48-\mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z < \frac{48-50}{1}\right) \\ &= P(Z < -2) \\ &= 0.0228 \end{aligned}$$

□

3 Sampling arbitrary populations, large samples

In the previous section we saw formulas for the the mean and the standard deviation of the random variable $\overline{X}$, the sampling distribution for the mean. We also saw that if the population X is randomly distributed then so $\bar{X}$, no matter what the sample size is.

In general, starting with any population the sampling distribution tends to look more and more approximately normal as the size becomes larger and larger. A theoretical explanation of this comes from the *Central Limit Theorem*, a result from *Probability Theory*. This is illustrated in Figure 2.

From Figure 2 we see that while for small samples $n = 5$ for example the sampling distribution is not approximately normal, for samples of size $n = 30$ it is very close to be normal. As a general rule of thumb, if the size of the sample is 30 or more, then the sampling distribution is approximately normal.

$n \geq 30$ is large enough

When the size of the samples is $n \geq 30$, the sampling distribution is approximately normal, (so we can use the normal tables to compute probabilities).

Remember that the mean and standard deviation of $\overline{X}$ are given by Equations (1) and (2) respectively.

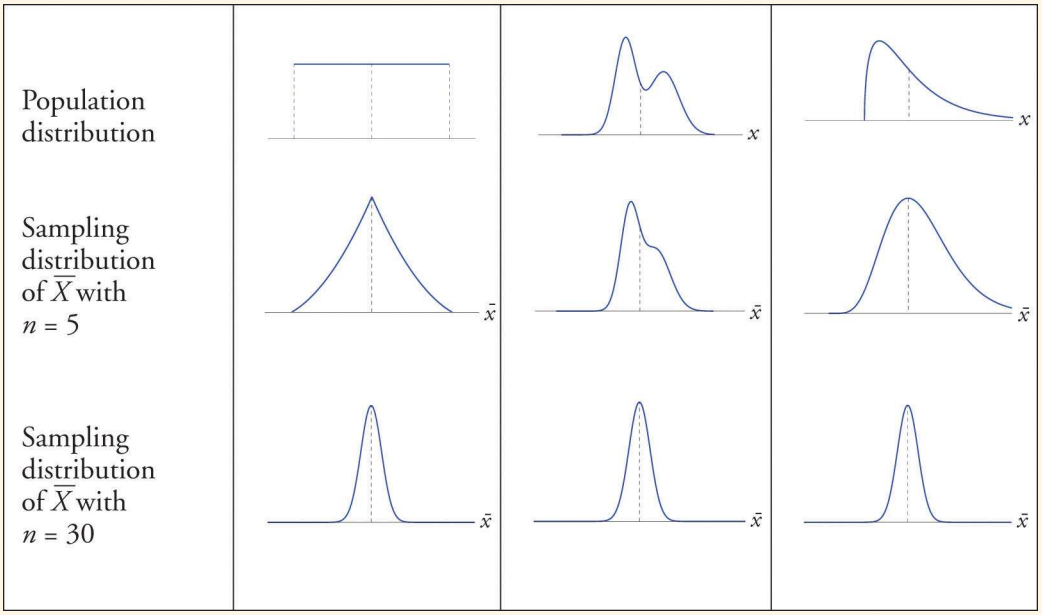

Figure 2: Sampling distributions from various populations.

Example 1

The numerical population of grade point averages at a college has mean 2.61 and standard deviation 0.5. If a random sample of size 100 is taken from the population, what is the probability that the sample mean will be between 2.51 and 2.71?

Answer. Since the sample size is $n \geq 100$ and therefore the sample is large enough for us to assume that $\overline{X}$ is approximately normal.

So to find the probability $P(2.51 < \overline{X} < 2.71)$ we need to convert the raw scores to Z–scores. The sample mean $\overline{X}$ has mean $\mu_{\overline{X}} = \mu = 2.61$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0.5}{10} = 0.05$. So we have

$$\bar{x} = 2.51 \implies z = \frac{2.51 - 2.61}{0.05} = -2 \quad \bar{x} = 2.71 \implies z = \frac{2.71 - 2.61}{0.05} = 2.$$

Therefore:

$$\begin{aligned} P(2.51 < \overline{X} < 2.71) &= P(-2 < Z < 2) \\ &= P(Z < 2) - P(Z < -2) \\ &= 0.97725 - 0.02275 \\ &= 0.9545. \end{aligned}$$

□

Example 2

Scores X on a common final exam in a large enrollment, multiple-section freshman course have mean 72.7 and standard deviation 13.1. Find the probability that the mean score $\bar{x}$ of 38 randomly selected exam papers is between 70 and 80.

Answer. The sample is large enough so we can assume that $\overline{X}$ is normally distributed. We have

$$\mu_{\overline{X}} = 72.7, \quad \sigma_{\overline{X}} = \frac{13.1}{\sqrt{38}} = 2.125.$$

and so the z–scores that correspond to the raw scores 70 and 80 are

$$\bar{x} = 70 \implies z = \frac{70 - 72.7}{2.125} = -0.85, \quad \bar{x} = \frac{80 - 72.7}{2.125} = 3.44.$$

So we have

$$\begin{aligned} P(70 < \overline{X} < 80) &= P(-0.85 < Z < 3.44) \\ &= P(Z < 3.44) - P(Z < -0.85) \\ &= 0.99971 - 0.19766 \\ &= 0.80205 \end{aligned}$$

□

Example 3

A tire manufacturer states that a certain type of tire has a mean lifetime of 60,000 miles. Suppose lifetimes are normally distributed with standard deviation $\sigma = 3,500$ miles.

(a) Find the probability that if you buy one such tire, it will last only 57,000 or fewer miles. If you had this experience, is it particularly strong evidence that the tire is not as good as claimed?

(b) A consumer group buys five such tires and tests them. Find the probability that average lifetime of the five tires will be 57,000 miles or less. If the mean is so low, is that particularly strong evidence that the tire is not as good as claimed?

Solution. Let's work with units of a 1,000. So we are given a normally distributed population with $\mu = 60$ and $\sigma = 3.5$.

(a) We convert $x = 57$ to a z–score.

$$x = 57 \implies z = \frac{57 - 60}{3.5} = -0.86.$$

So we have,

$$\begin{aligned} P(X \leq 57) &= P(Z \leq -0.86) \\ &= 0.19489 \end{aligned}$$

So the probability that the tire will last 57,000 miles or less is 0.19489 or 19.49%. This probability is not very low so it's not strong eovidence that the tires are not as good as claimed.

(b) Since X is normal $\overline{X}$ is also normal with

$$\mu_{\overline{X}} = 60, \quad \sigma_{\overline{X}} = \frac{3.5}{\sqrt{5}} = 1.57.$$

So the raw score $\bar{x} = 57$ corresponds to the z–score

$$z = \frac{57 - 3}{1.57} = -1.92.$$

So

$$\begin{aligned} P(\overline{X} \leq 57) &= P(Z \leq -1.92) \\ &= 0.02743. \end{aligned}$$

This probability, 2.7% is very low. So if the mean of the sample turns out to be so low, it would constitute particularly strong evidence that the tires are not as good as advertised.

□

4 Exercises

1. A normally distributed population has mean $57,800$ and standard deviation 750.
   (a) Find the probability that a single randomly selected element X of the population is between $57,000$ and $58,000$.
   (b) Find the mean and standard deviation of $\overline{X}$ for samples of size 100.
   (c) Find the probability that the mean of a sample of size 100 drawn from this population is between $57,000$ and $58,000$.

2. A population has mean 12 and standard deviation 1.5.
   (a) Find the mean and standard deviation of $\overline{X}$ for samples of size 90.
   (b) Find the probability that the mean of a sample of size 90 will differ from the population mean 12 by at least 0.3 unit, that is, is either less than 11.7 or more than 12.3.
   Hint. You might find it easier to compute the probability of the complementary event.

3. Suppose the mean length of time that a caller is placed on hold when telephoning a customer service center is 23.8 seconds, with standard deviation 4.6 seconds. Find the probability that the mean length of time on hold in a sample of $1,200$ calls will be within 0.5 second of the population mean.

4. Suppose the mean weight of school children's book bags is 17.4 pounds, with standard deviation 2.2 pounds. Find the probability that the mean weight of a sample of 30 book bags will exceed 17 pounds.

5. The time that it takes me to walk from my apartment to the subway is approximately normally distributed with mean 11.2 minutes and standard deviation 1.1 minutes.
   (a) What is the probability that in a random day it will take me between 10 and 13 minutes to walk to the subway from my apartment?
   (b) If I randomly select 12 days to measure how long it takes me to walk to the subway from my apartment, and compute the mean $\overline{X}$, what is the probability that the mean will be between 10 and 13 minutes?