CHAPTER 8: SAMPLING DISTRIBUTIONS

[This chapter is based on Chapter 8 of the textbook]

In this chapter we focus on studying the distribution of statistics such as the mean $\overline{x}$. When we learned about statistics at the beginning of the semester, we described them as a numerical summary of a sample, and we said a sample is a subset of the population. Now, different samples will yield different values of the statistics because they involve different individuals. Also, the composition of each sample is usually random. Hence, statistics such as the mean $\overline{x}$, are also random variables and we can study their distribution.

8.1 Distribution of the sample mean

We start with a definition.

Definition 8.1. *The sampling distribution of a statistic is a probability distribution for all possible values of the statistic computed from a sample of size n.*

As an example, let's compute the sampling distribution of the mean in a small example.

Example 8.1. *Suppose we have a population with 4 individuals, and them how many years they have had their car. They responded 2, 4, 6, 8.*

(a) Compute the population mean and standard deviation

(b) Draw all possible samples of size 2 and compute their mean

(c) Construct the sampling distribution of the sample mean

(d) What is the probability that we get a sample with mean 5?

(e) Compute the mean and standard deviation of the mean of all the samples of size 2. Compare your result with part (a).

Solution.

(a) The population mean and standard deviation are:

$$\mu = \frac{2+4+6+8}{4} = 5$$
$$\sigma = \sqrt{\frac{(2-5)^2+(4-5)^2+(6-5)^2+(8-5)^2}{4}} = \sqrt{5} \approx 2.2361$$

(b) Our state space is $S = \{2, 4, 6, 8\}$ and we want to compute all the samples (subsets) of size 2. We obtain the following:

Sample	Sample mean
$\{2, 4\}$	$\frac{2+4}{2} = 3$
$\{2, 6\}$	$\frac{2+6}{2} = 4$
$\{2, 8\}$	$\frac{2+8}{2} = 5$
$\{4, 6\}$	$\frac{4+6}{2} = 5$
$\{4, 8\}$	$\frac{4+8}{2} = 6$
$\{6, 8\}$	$\frac{6+8}{2} = 7$

(c) We construct a probability distribution of the sample mean. From the table above, we observe that the possible values of the sample mean are $3, 4, 5, 6, 7$. Each of the samples has the same probability of being selected. Hence, the probability that the sample size is 3, for example, is the number of samples that have this mean (1) divided by the total number of samples (6). Similarly with the values 4, 6, 7. The value 5, instead, is the sample mean of 2 of the samples. Then, the probability that the mean of a sample is 5 is $\frac{2}{6}$.

In other words, we have the following probability distribution:

Sample mean value, x	$P(x)$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{2}{6}$
6	$\frac{1}{6}$
7	$\frac{1}{6}$

(d) The probability that the sample mean is 5 is $\frac{2}{6}$

(e) We now compute the mean and standard deviation of the sample mean. We obtain:

$$\mu_{\overline{x}} = \frac{3+4+5+5+6+7}{6} = 5$$

$$\sigma_{\overline{x}} = \sqrt{\frac{(3-5)^2+(4-5)^2+(5-5)^2+(5-5)^2+(6-5)^2+(7-5)^2}{6}} = \sqrt{\frac{5}{3}} \approx 1.2910$$

We obtained the same mean, but the standard deviation is smaller when we compute the sample mean standard deviation than the original random variable's standard deviation.

When we computed the standard deviation of the population, we considered individual values and these fluctuate a lot. However, when we considered the mean of the samples, we are averaging two values before computing their fluctuations. Hence, the extreme values (2 and 8) have less weight.

□

The example above is very simple, but it illustrates a very general result. We present it below.

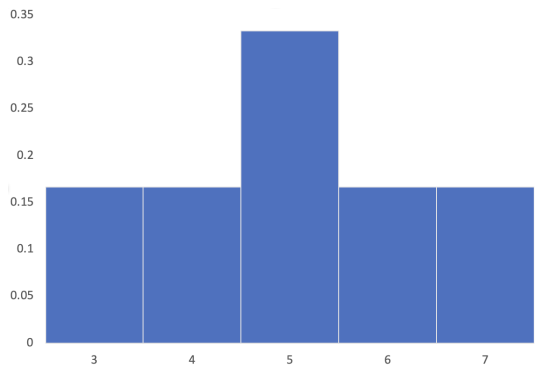
Theorem 8.1. *Suppose that a simple random sample of size n is drawn from a (large) population with mean μ and standard deviation σ. The sampling distribution of $\overline{x}$ has mean*

$$\mu_{\overline{x}} = \mu$$

and standard deviation (also known as standard error) is

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

In our example, above, we can graph the distribution of the mean of samples of size 2 and obtain:



The population size is very small, but we can see that the distribution is symmetric and could perfectly fit a bell-shape distribution for larger populations. In general, we have the following result.

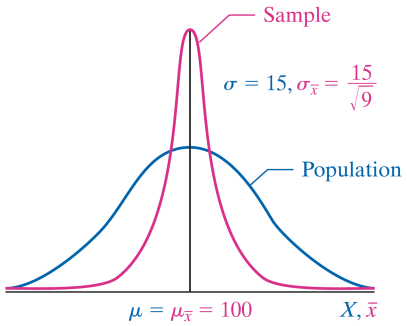
Theorem 8.2. *If a random variable X is normally distributed, then the sampling distribution of the sample mean $\overline{x}$ is also normally distributed.*

Example 8.2. *The IQ score of individuals are modeled by a random variable with mean $\mu = 100$ and standard deviation $\sigma = 15$. Compute the mean and standard deviation of a simple random sample of $n = 9$ individuals and show the distribution of the population and the sample mean in a graph.*

Solution. According to the results we just learned, the mean and standard deviation of the sample mean in samples of size 9 are:

$$\mu_{\overline{x}} = \mu = 100$$
$$\text{and } \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{9}} = 5$$

Since both, the population and the sample mean, are normally distributed, we obtain the following graph:



□

According to the theorems above, the mean of the samples will never change. However, the standard deviation changes with the sample size. Let's consider an example.

Example 8.3. *The IQ score, X, of humans is approximately normally distributed with mean $\mu = 100$ and standard deviation $\sigma = 15$. Compute the probability that a simple random sample of size $n = 10$ results in sample mean greater than 110.*

Solution. We need to compute $P(\overline{x} > 110)$ when the sample sizes are $n = 10$. Since $\overline{x}$ has normal distribution with mean

$$\mu_{\overline{x}} = \mu = 100$$

and standard deviation

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{10}} \approx 4.7434$$

we obtain:

$$\begin{aligned} P(\overline{x} > 110) &= 1 - P(\overline{x} \leq 110) \qquad \text{(complement rule)} \\ &= 1 - P\left(Z \leq \frac{110 - 100}{4.7434}\right) \qquad \text{(computing the } z\text{-score of 110)} \\ &= 1 - P(Z \leq 2.11) \\ &= 1 - 0.9826 \qquad \text{(from the standard-normal table)} \\ &= 0.0174 \end{aligned}$$

Hence, the probability that a simple random sample of size 10 results in mean greater than 110 is 1.74%. □

In the following result, we establish a very powerful result for large samples.

Theorem 8.3 (The Central Limit Theorem)**.** *Regardless of the shape of the underlying population, the sampling distribution of $\overline{x}$ approximates a normal distribution as the sample size n increases.*

In other words, the Central Limit Theorem establishes that if the sample size is large, then the mean is always normal. The word "large" is vague, and what is large highly depends on the shape of the population distribution. As a rule of thumb, we can consider $n \geq 30$ as large.

Example 8.4. *The mean weight gain during pregnancy is 30 pounds, with a standard deviation of 12.9 pounds. Weight gain during pregnancy is a skewed right.*

An obstetrician obtains a random sample of 35 low-income patients and determines their mean weight gain during pregnancy was 36.2 pounds. Does this result suggest anything unusual?

Solution. Since the sample size is $n = 35 \geq 30$, we can assume that the sample mean is normally distributed with mean

$$\mu_{\overline{x}} = \mu = 30$$

and standard deviation

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12.9}{\sqrt{35}} \approx 2.1805$$

To determine if the mean of 36.2 pounds of the sample is unusual, we compute $P(\overline{x} \geq 36.2)$. We obtain:

$$\begin{aligned} P(\overline{x} \geq 36.2) &= 1 - P(\overline{x} < 36.2) \qquad \text{(complement rule)} \\ &= 1 - P\left(Z < \frac{36.2 - 30}{2.1805}\right) \qquad \text{(computing the } z\text{-score of 36.2)} \\ &= 1 - P(Z < 2.84) \\ &= 1 - 0.9977 \qquad \text{(using the standard-normal table)} \\ &= 0.0023 \end{aligned}$$

That is, the probability that a sample has a mean of at least 36.2 pounds is 0.23%. Since this number is below 1%, we conclude that it is an unusual sample if we compare it to the entire population. However, it might be that we are not considering a lurking variable, such as the patients' income. □

8.2 Distribution of the sample proportion

Another statistic of great relevance when studying a sample is the proportion that satisfies a certain characteristic. For example, we may want to study the proportion of households that own a dog, the proportion of households who live under a certain wage, the proportion of college students who graduate within 5 years, etc.

In many cases, the population proportion is difficult to compute because asking every individual is costly or practically impossible. Hence, we need to learn how to use a sample proportion to infer about the population.

Definition 8.2. *Suppose that a sample of size n is obtained from a population in which each individual either does or does not have a certain characteristic. The sample proportion, denoted by $\hat{p}$ (pronounced "p hat") is given by*

$$\hat{p} = \frac{x}{n}$$

where x is the number of individuals in the sample with the specified characteristic.

The sample proportion is a statistic, and it estimates the value of the population proportion, that we denote by p.

Let's do a quick example.

Example 8.5. *The Harris Poll conducted a survey of 1200 adult Americans who vacation during the summer and asked whether the individuals plan to work while on summer vacation. Of thos surveyed, 552 indicated that they plan to work while on vacation.*

Find the sample proportion of individuals surveyed who plan to work on summer vacation.

Solution. The survey was conducted to 1200 adults, so the sample size is $n = 1200$. Among the surveyed adults, 552 plan to work on summer vacation, so $x = 552$. Hence, the proportion of individuals who plan to work on summer vacation is:

$$\hat{p} = \frac{552}{1200} = 0.46$$

□

Example 8.6. *A study aims to evaluate the students' level of happiness in a specific university. In the dimension of major-related topics, 85% of the entire student body said they are completely happy with their major choice. When you asked 50 of your friends, you found out that 40 of them are completely happy with their major choice.*

Identify the population and sample proportion of happiness with major choice.

Solution. The population is the entire student body, so the population proportion is

$$p = 0.85$$

The sample is the subset of $n = 50$ students that you asked. From these 50 students, $x = 40$ said they are happy with their major choice. Hence, the sample proportion is

$$\hat{p} = \frac{40}{50} = 0.8$$

□

Similarly to the sample mean, the sample proportion depends on the individuals of the specific sample that we study. Hence, it is also a random variable. In the following theorem we provide its distribution.

Theorem 8.4. *For a simple random sample of size n with a population proportion p, the sampling proportion $\hat{p}$ has mean*

$$\mu_{\hat{p}} = p$$

and standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Further, if the sample size n is less than 5% of the population size N and if $np(1-p) \geq 10$, then the sampling proportion $\hat{p}$ is normally distributed.

The theorem above gives us the mean and standard deviation of the sample proportion $\hat{p}$ in terms of the sample size and the population proportion p. This mean and standard deviation are valid for all values of n and p.

If we additionally have a sample large enough ($np(1-p) \geq 10$) and the population is large with respect to the sample (sample is less than 5% of population), then we also know the distribution of the sample proportion.

Example 8.7. *According to the National Center for Health Statistics, 15% of all Americans have hearing trouble. We are interested in studying a sample of 120 Americans.*

(a) What is the mean and standard deviation of the sample proportion?

(b) Justify that the sample proportion is normally distributed.

(c) What is the probability that at most 12% of the sampled individuals have hearing trouble?

(d) Suppose that the 120 sampled Americans regularly listen to music, and 26 have hearing trouble. Is this unusual? What might your conclude?

Solution.

(a) We use the result from the theorem to obtain:

$$\mu_{\hat{p}} = p = 0.15$$
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.15 \cdot 0.85}{120}} \approx 0.0326$$

(b) To verify that the sample is smaller than 5% of the population, we observe that the American population is over 300 million people, so 5% of the American population is over 15 million. Hence, the sample of $n = 120$ individuals satisfy this requirement.

Additionally,

$$np(1-p) = 120 \cdot 0.15 \cdot 0.85 = 15.3 \geq 10$$

Then, the second requirement is also satisfied, that is, the sample is large enough.

Hence, the sample proportion $\hat{p}$ follows a normal distribution.

(c) The probability that at most 12% of the sampled individuals have hearing trouble can be expressed as $P(\hat{p} \leq 0.12)$. We know that $\hat{p}$ is normally distributed, so we compute its z-score and use the standard-normal table to compute the probability. We obtain

$$\begin{aligned} P(\hat{p} \leq 0.12) &= P\left(Z \leq \frac{0.12 - 0.15}{0.0326}\right) \qquad \text{(computing the } z\text{-score)} \\ &= P(Z \leq -0.92) \\ &= 0.1788 \end{aligned}$$

That is, the probability that a sample of 120 Americans results in less than 12% having hearing trouble is 17.88%.

(d) If 26 individuals have hearing trouble, then the sample proportion is

$$\hat{p} = \frac{26}{120} = 0.217$$

This value of $\hat{p}$ is higher than its mean. Since the normal distribution is symmetric and $\hat{p}$ is larger than the population proportion p, we evaluate the probability that $\hat{p}$ is higher than 0.217. We obtain

$$\begin{aligned} P(\hat{p} > 0.217) &= 1 - P(\hat{p} \leq 0.217) \qquad \text{(complement rule)} \\ &= 1 - P\left(Z \leq \frac{0.217 - 0.15}{0.0329}\right) \qquad \text{(computing the } z\text{-score)} \\ &= 1 - P(Z \leq 2.06) \end{aligned}$$

$$= 1 - 0.9803$$
$$= 0.0197$$

Hence, the change of having 26 or more individuals with hearing trouble is 1.97%. We can consider this probability unusual, but we also have the additional information that these individuals regularly listen to music and this could be a lurking variable.

□