

Lecture 7

SAMPLING DISTRIBUTIONS CENTRAL LIMIT THEOREM



FPT UNIVERSITY

Department of Mathematics

Võ Văn Nam



Contents

1 Parameter and statistic

2 Sampling distributions

3 Central Limit Theorem

1. Parameter and statistic

Parameter and statistic

What is the average height of adult men in the US?

This is difficult to evaluate as there are 120 million adult men, but it can be estimated quite well with a relatively small sample.

The **population** consists of all adult men in the US (about 120 million).

A **parameter** is a quantity of interest about the population: the population average μ , or the population standard deviation σ .

A **statistic (estimate)** is the quantity of interest as measured in the sample: the sample average \bar{x} , or the sample standard deviation s .

The expected value

If we sample an adult male at random, then we expect his height to be around the population average μ , give or take about one standard deviation σ .

The expected value of one random draw is the population average μ .

How about \bar{x}_n , the average of n draws?

The expected value of the sample average, $E(\bar{x}_n)$, is the population average μ .

But remember that \bar{x}_n is a random variable because sampling is a random process.

So \bar{x}_n won't be exactly equal to $\mu = 69.3$ in: We might get, say, $\bar{x}_n = 70.1$ in. Taking another sample of size n might result in $\bar{x}_n = 69.1$ in.

How far off from μ will \bar{x}_n be?

The standard error (SE) of a statistic tells roughly how far off the statistic will be from its expected value.

The standard error for the sample average

The standard error (SE) of a statistic tells roughly how far off the statistic will be from its expected value.

So the SE for a statistic plays the same role that the standard deviation σ plays for one observation drawn at random.

The square root law is key for statistical inference:

$$\text{SE}(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$$

The importance of the square root law is twofold:

- It shows that the SE becomes smaller if we use a larger sample size n . We can use the formula to determine what sample size is required for a desired accuracy.
- The formula for the standard error **does not depend on the size of the population**, only on the size of the sample.

Expected value and standard error for the sum

What if we are interested in the sum of the n draws, S_n , rather than the average \bar{x}_n ?

The sum and the average are related by $S_n = n\bar{x}_n$.

Both the expected value and the standard error can likewise be obtained by multiplying

$$E(S_n) = n\mu, \quad SE(S_n) = \sqrt{n}\sigma$$

So the variability of the sum of n draws increases at the rate \sqrt{n} .

Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

The percentage of likely voters is an average. This becomes clear by using the framework for counting and classifying:

- The population consists of all likely voters (about 140 million).
- Each likely voter falls into one of two categories: approve or not approve.
- Put the label 1 on each likely voter who approves, and 0 on each who doesn't.
- Then the number of likely voters who approve equals the sum of all 140 million labels.
- The percentage of likely voters who approve is the percentage of 1s among the labels.

Expected value and standard error for percentages

In a sample of n likely voters

- the number of voters in the sample who are approving is the sum S_n of the draws
- the percentage of voters approving is the percentage of 1s, which is

$$\frac{S_n}{n} \times 100\% = \bar{x}_n \times 100\%$$

Therefore

$$E(\text{percentage of 1s}) = \mu \times 100\%, \quad SE(\text{percentage of 1s}) = \frac{\sigma}{\sqrt{n}} \times 100\%$$

where μ is the population average (=proportion of 1s) and σ is the standard deviation of the population of 0s and 1s.

All of the above formulas are for sampling with replacement. They are still approximately true when sampling without replacement if the sample size is much smaller than the size of the population.

2. Sampling distributions

Sampling distributions

Toss a coin 100 times. The number of tails has the following possible outcomes: 0, 1, 2, ..., 100.

How likely is each outcome?

The number of tails has the binomial distribution with $n = 100$ and $p = 0.5$. (success = coin lands tails)

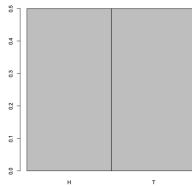
So if the statistic of interest is S_n = number of tails, then S_n is a random variable whose probability histogram is given by the binomial distribution. This is called the **sampling distribution** of the statistic S_n .

The sampling distribution of S_n provides more detailed information about the chance properties of S_n than the summary numbers given by the expected value and the standard error.

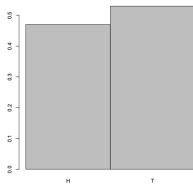
There are three histograms

The chance process of tossing a coin 100 times comes with three different histograms:

1. The probability histogram for producing the data:

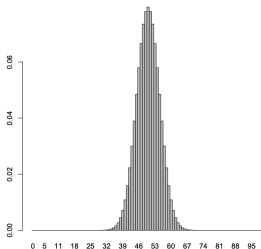


2. The histogram of the 100 observed tosses. This is an empirical histogram of real data:



There are three histograms

3. The probability histogram of the statistic S_{100} = number of tails, which shows the sampling distribution of S_{100} :



Remark. When doing statistical inference it is important to carefully distinguish these three histograms.

Example

Assume that $\{2; 3; 10\}$ be a population. Randomly selected (with replacement) a sample of size 2.

- Find the population mean.
- Find the mean of sample means.
- Compare the results.

Hint.

Sample	2,2	2,3	2,10	3,2	3,3	3,10	10,2	10,3	10,10
Sample mean	2	2.5	6	1.5	3	6.5	6	6.5	10

3. Central Limit Theorem

The law of large numbers

The square root law says that $SE(\bar{x}_n)$, the standard error of the sample mean, goes to zero as the sample size increases.

Therefore the sample mean \bar{x}_n will likely be close to its expected value μ if the sample size is large. This is the **law of large numbers**.

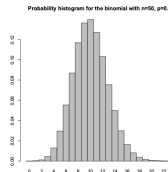
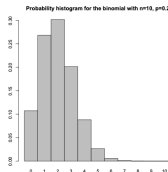
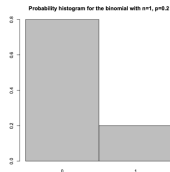
Keep in mind that the law of large numbers applies

- for averages and therefore also for percentages, but not for sums as their SE increases
- for sampling with replacement from a population, or for simulating data from a probability histogram

More advanced versions of the law of large numbers state that the empirical histogram of the data (the histogram in 2. in the previous section) will be close to the probability histogram in 1. if the sample size is large.

The central limit theorem

Recall the online game where you win a small prize with probability 0.2. We looked at the random variable X = number of small prizes in n gambles and found that X has the binomial distribution with that n and $p = 0.2$.



As n gets large, the probability histogram looks more and more similar to the normal curve. This is an example of the central limit theorem.

When sampling with replacement and n is large, then the sampling distribution of the sample average (or sum or percentage) approximately follows the normal curve. To standardize, subtract off the expected value of the statistic, then divide by its SE.

Central Limit Theorem

Theorem (CLT for one population)

Suppose X_1, X_2, \dots, X_n is a random sample of size n taken from a population with mean μ and variance σ^2 . Let \bar{X} be the sample mean.

Then $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is approximately standard normal when n large ($n \geq 30$).

It means that for $n \geq 30$, we have $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Remark. If the population already has normal distribution then for any sample size:

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Example 1

One year, the average salary for professional players was \$1.5 million, with a standard deviation of \$1 million. Suppose a sample of 100 players was taken.

Find the approximate probability that the average salary of these 100 players does not exceed \$1.4 million.

Solution. Let \bar{X} (million) be the average salary of these 100 players. By Central Limit Theorem,

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

In which, $n = 100, \mu = 1.5, \sigma = 1$. Therefore,

$$P(\bar{X} \leq 1.4) = P(Z \leq -1) \approx 16\%.$$

Theorem (CLT for two populations)

If we have 2 independent populations with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2) , and if \bar{X}_1 and \bar{X}_2 are the sample means of 2 independent random samples of size n_1 and n_2 from these populations, then the sampling distribution of

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is approximately standard normal for large n_1, n_2 . It is exactly standard normal if the two populations are normal.

Remark.

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

Example 2

The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 years, while those of manufacturer B have a mean lifetime of 6.5 years and a standard deviation of 0.7 years.

*What is the probability that a random sample of 16 tubes from manufacturer A will have mean lifetime that is **at least 6** months longer than the mean lifetime of a sample of 25 tubes from manufacturer B?*

Solution. By using CLT for two populations, one has

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

Hence,

$$P(\bar{X}_1 - \bar{X}_2 \geq 0.5) = 1 - P(\bar{X}_1 - \bar{X}_2 < 0.5) \approx 3\%$$

Quizlets

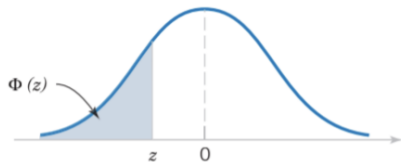
Quiz 1. The amount of time that a customer spends waiting at an airport check-in counter is a random variable with mean 8.2 minutes and standard deviation 1.5 minutes. Suppose that a random sample of $n = 49$ customers is observed.

Find the probability that the average time waiting in line for these customers is less than 6 minutes.

Quiz 2. A random sample of size $n_1 = 16$ is selected from a normal population with a mean of 75 and a standard deviation of 8. A second random sample of size $n_2 = 9$ is taken from another normal population with mean 70 and standard deviation 12. Let \bar{X}_1 and \bar{X}_2 be the two sample means.


Find the probability that $3.5 \leq \bar{X}_1 - \bar{X}_2 \leq 5.5$.

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$



z	$\Phi(z)$
-3.0	0.0013
-2.5	0.0062
-2.0	0.0228
-1.5	0.0668
-1.0	0.1587
-0.5	0.3085
0.0	0.5000

z	$\Phi(z)$
0.0	0.5000
0.5	0.6915
1.0	0.8413
1.5	0.9332
2.0	0.9772
2.5	0.9938
3.0	0.9987

The background features a repeating pattern of light blue hexagons. Inside and around these hexagons are small blue dots of varying sizes, connected by thin, faint lines, creating a molecular or network-like structure.

Thank you!

namvv14@fe.edu.vn