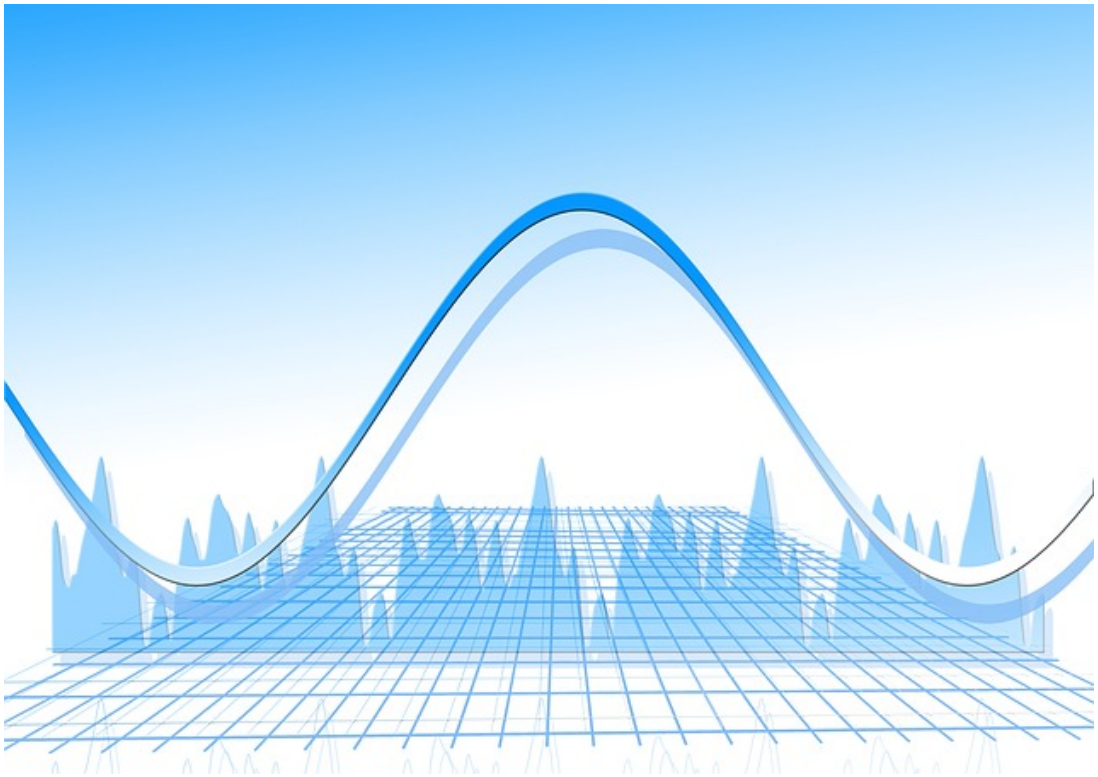
Sampling Distributions for Proportions and Means

See online here

This article discusses the sampling distribution of mean and proportion in detail. The sampling error, sampling variability, and normal model for proportion are explained with examples. A central limit theorem which states that if the number of the sample size is large, it results in the normal distribution of mean of sample size of 'n' number of population, and is also included in this article. The sampling proportion of distribution and CLT has three conditions to be fulfilled for accurate computation purpose.



Sampling Distribution

Sampling distribution is the **probability of distribution of statistics from a large population by using a sampling technique**. Sampling helps in getting average results about a large population through choosing selective samples. The results obtained from observing or analyzing samples help in concluding an opinion regarding a whole population from which samples are drawn.

The sampling results are compiled **on the basis of the expected frequency of occurrence** of an event or statistic in a whole population. Each sample contributes in calculation of some statistics comprising on random variables. These random variables

have probability distribution which is calculated on the basis of the samples chosen.

Types of sampling distribution

There are two types of sampling distribution, i.e.:

1. The sampling distribution of a proportion
2. The sampling distribution of a mean

Sampling Error

A sampling error is also referred to as an **estimation error** and is the amount of inaccuracy in the value estimation arising from a portion of a population rather than the entire population. It is the difference between the statistic and the corresponding parameter. This kind of error results when the statistical characteristics are estimated from a **single subset or incomplete portion of population** instead of the whole population. This kind of error indicates that the sampling results **don't depict average results** from the statistics of the whole population; hence, are not reliable.

Cause of Sampling Error

- Biased sampling procedure. Each investigator should pick sample(s) that is void of any bias and the sample(s) that forms part of the whole population of interest.
- Randomization and probability sampling are carried out to minimize the sampling error but there is a possibility that all randomized subjects are not part of the population of interest.

Elimination of the Sampling Error

- Increasing the sample size. The sampling error can be minimized by an increment in sample size.
- Improving the sample design through a stratification technique.
- Groups with similar units are obtained through the division of the whole population of interest.

Example: In case a person measures weight of 1,000 people in a city which has a total population of 50,000. The results which will be generated on the basis of samples chosen for 1,000 individuals will not be helping to conclude the result for the whole population of 50,000 people. It is known as a sampling error.
Non-sampling errors arise in a data collection process due to factors other than that of a sample. These errors arise from sampling the wrong population of interest and by response biased, also those errors made by an investigator during data collecting, analysis and reporting. These errors are mainly present in a complete census and sampling survey.

Sampling Variability

Sampling variability is the range of values that differs between samples. Samples are **chosen on a random basis from a population known as a parameter**. Samples chosen from the whole population show random results on the basis of their repeated selection.

Sampling variability refers to the fact that **statistical information embedded in each sample varies from sample to sample**. It increases with an increase in sample size. It

is based on different sets of populations and relates characteristics from which samples are drawn. It is another name for a **range indicating a range of different values belonging to different samples from different populations**.

Sampling Distribution of a Proportion

There are several ways to calculate the sampling distribution of a proportion. The **population portion Ƥ indicates the proportion of items of individuals in a whole population with specific characteristics or interests**. Here, the sample proportion is denoted by p̂. It shows specific characteristics of sample proportion which only belong to the specific part of the population to which they belong.

Illustration

In a sample of 200 adults, 160 have smartphones and, if we want to find out the proportion of individuals with a smartphone in a whole population, we need to calculate through the following formula:

$$\hat{p} = 160 / 200$$

$$= 0.80$$

Properties

The sampling distribution of proportion has the following properties:

Mean (μ p̂): it is also pronounced as mu sub-p-hat. It indicates the population proportion p̂.

Standard Error (σ p̂): It is pronounced as sigma sub-p-hat. It is indicated by . In the formula of standard error, due to n being the dominator, the error has an inversely proportional relationship with sample size or the number of samples chosen from the specific population and vice versa.

In case the sample size is larger or p̂ is closer to 0.50, it indicates that the distribution of sample proportion is equal to normal distribution.

Example of Sampling distribution of a proportion

In a survey conducted which involves an ACT test each year to find out the proportion of students who like to take help for math skills. If it is assumed on the basis of researches in previous years that 38% of the total students responded with yes to the ACT test, it shows that the population proportion is equal to 0.38. The distribution of the two responses shows a result of Yes = 38% and No = 62%.

In this case, the sample proportion p̂ can be calculated by using the standard error formula.

Due to a large number of students, i.e., 1,000, the population proportion is closer to normal distribution.

Normal Model for Proportions

The normal distribution, also known as a bell curve, indicates the average results from a population. It gives the appearance of a bell with equally distributed curves on the right and left side as shown below.

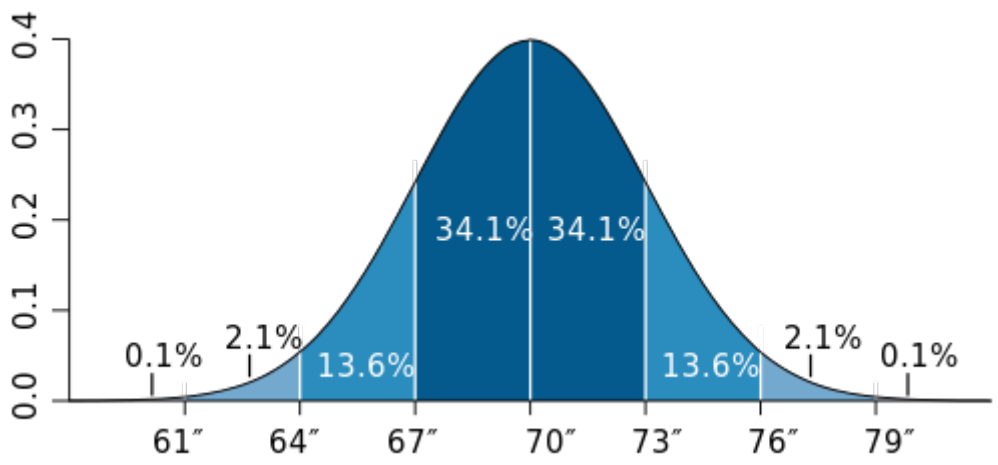


Image: “Standard deviation diagram US men heights” by Petter Strandmark, GliderMaven, cmglee. License: CC BY 2.5

The normal model for proportions has **several properties,** including:

- The mean, median and mode of the normally distributed proportions are equal.
- The curve of the normal distribution is symmetric at the center around the mean μ.
- The normal distribution shows a perfectly divided curve from the middle of the total distribution.
- The total area covered by the whole proportion curve denotes value 1.

Elements of normal model for distributions

The key elements in any situation which involves a normal model for the display distribution of proportion comprises on the mean and standard deviation. It is the same in the sampling distribution of proportion of the population portion which is denoted by “Ᵽ” and sample proportion by “p̂” in the normal distribution curve.

Conditions for the normal model for distribution

The normal model for distribution **should fulfil three mandatory conditions** in order to find out mean and standard deviation of sample proportion.

The Independence Assumption: It is mandatory that individuals or samples from a population are independent from each other.

The 10% condition: It is mandatory that the sample size is less than 10% of the total population to which it belongs.

The Success/Failure condition: The binomial approximation condition is applied here to find out success/failure through the approximately normal condition through (1 – Ᵽ). To meet this condition, np̂ should be ≥ 10.

Practical application of conditions

In case a survey is conducted in order to find out the role of industrial pollution as a major reason for global warming. If it is believed that 45% of the population of London considers it true and a sample of 100 people is taken, what is the expected probability that 47% of the total respondents of the sample of 100 people will approve this perception?

In order to find out the probability, first of all, the conditions of normal approximation have to be confirmed.

Independence condition: The sample size of 100 people has been chosen randomly; hence, it is considered that they are independent of each other fulfilling this condition.

10% condition: The sample of 100 people out of the total population of London City does not even constitute 1% of the total population; hence, the 10% condition has also been met.

Success/Failure condition:

In this case,

$$n\hat{p} = 100\ (0.45)$$

$$= 45 \geq 10$$

$$(1 - \hat{P}) = 100\ (0.55)$$

$$= 55 \geq 10$$

Hence, it is proved that all conditions have been met, so normal approximation can be used in this case. The mean of a sample is equal to the sample proportion p̂.

In this case, the mean is:

$$= 0.45$$

Once we have the mean and standard deviation of the survey data, we can find out the probability of a sample proportion of 0.47 who consider industrial activities as a major source of global warming in London. Here, the Z score conversion formula will be used to find out the required probability, i.e.:

$$Z = x - \mu / \sigma$$

Putting the values in Z-score formula. The probability of sample proportion of 0.47 is:

$$= (0.47 - 0.45 / 0.0497)$$

$$= 0.40$$
$$\text{as } (\geq 0.47)$$

And Now (≥ 0.40)

$$\geq 0 - 0 \leq \leq 0.4$$

$$0.5 - 0.1554 = 0.3446$$

Hence, there is 0.3446 probability that 47% of total respondents of a sample of 100 people will approve this perception.

Sampling Distribution of the Mean

The mean of sample distribution refers to the mean of the whole population to which the selected sample belongs. It is the same as sampling distribution for proportions. The sampling distribution of the mean of sample size is important but complicated for concluding results about a population except for a very small or very large sample size.

Example: In this case, we have selected 500 male students between 20—25 years from a college and measured their heights. The average height for them is measured to be 5 ft 7 inches. Again, we selected another 500 male students from another college of the same age group, and now the average measure is 5 ft 6.5 inches.
The difference between these two averages is the sampling variability in the mean of a whole population. This variability can be resolved through modeling sample averages.

The Fundamental Theorem of Statistics

There are two fundamental theorems of statistics also known as fundamental theorems of probability, i.e.:

1. The law of large numbers
2. The central limit theorem

The law of large numbers

It states that, as the number of trials of an independently random process increases, the percentage difference between the expected and actual values moves towards zero. Also, it can be stated otherwise; an increase in the value of similarly distributed, random variables, results to their sample mean approaching their theoretical mean.

The Central Limit Theorem (CLT)

It states that if the number of sample size is large, it results in the normal distribution of the mean of sample size of n number of population. It provides the mean and standard deviation of sampling distribution in terms of sample size, the mean of the whole population μ and the variance of whole population σ.

CLT is **helpful in measuring statistics computationally;** hence, it is used in several statistical tests. It interconnects important elements of sample distribution including mean, standard deviation, sample size, variances and accuracy of point estimates.

For accurate calculations of CLT, the observations should be collected independently and randomly. For effective measurement of CLT, the following facts have to be considered:

- Population distribution is not important for CLT.
- If sample size n is large, it does not get impacted, even if the population distribution is symmetric or skewed.
- A normal model can be used to interpret the distribution of the sample mean because the sample mean has an approximate normal distribution.

Conditions

CLT is useful if all three conditions are fulfilled, including:

Independent groups: The sample size selected should be independent of each other

and selected randomly.

Independence/Randomization: The sample size n should be large enough i.e. at least 30 samples.

10% Condition: The sample size should not be larger than 10% of the population size from which it has to be selected.

Example: We select 100 apples from an orchard to measure their average weight from a total of 10,000 apples. The average weight of each apple is 0.34 lbs. The standard deviation, in this case, is 0.1lbs. We need to find out the probability that the weight of the sample size of 100 apples is less than 0.32 lbs.
In order to check out the feasibility of the data, firstly, we need to check out whether it meets the conditions or not.

- The sample size of 100 apples has been chosen randomly; hence, it is considered that they are independent of each other fulfilling this condition.
- The sample size is 100 apples, whereas the minimum size required is 30, so the sample size condition is fulfilled.
- 100 apples make up 1% of the total population of apples in the orchard, so less than 10% of the condition has also been fulfilled.

Hence, all conditions are fulfilled; CLT can be used to measure the required probability in this case. In this case:

$$\mu = 0.34 \text{ lbs}$$

Using Z score formula:

$$Z = x - \mu / \sigma$$

$$= (0.32 - 0.34 / 0.01)$$

$$= 2$$

$$\text{As} \leq -2 = \geq 2 = \geq 0 - (0 \leq \leq 2)$$

Hence, the z-score that the weight of the sample size of 100 apples is less than 0.32 lbs. is 2.

Variation and Means

The **variation factor in case of mean is less observed as compared to the individual values**. The following points are helpful in understanding variation and means.

- The average values or mean of a sample gets stable with an increase in its size.
- The larger a sample group, the better it represents the whole population.
- The average or mean of a population should be pretty stable to represent the population as a whole.
- With an increase in sample size, the standard deviation of the sample mean falls. For example, in case we would have chosen 400 apples in the illustration of CLT instead of 100 apples, the standard deviation would have gone down to 0.05 from 0.01.

Cautions

The following common issues with sampling distribution should be taken into account:

- Sampling distribution and the distribution of the sample are two different factors. They should not be confused with each other. Sampling distribution refers to the distribution of statistics from the sample.
- The observations which are not independent should be avoided as these are not useable in the case of CLT.
- Small sample size should be investigated carefully for skewed distributions, since, when the sample size is small; the normal approximation does not work well and vice versa.

References

University of Florida Biostatistics, 2017. Sampling Distribution of the Sample Mean, x-bar. Available at: http://bolt.mph.ufl.edu/6050-6052/module-9/sampling-distribution-of-x-bar/

Wicklin, R., 2014. Fundamental theorems of mathematics and statistics. Available at: https://blogs.sas.com/content/iml/2014/02/12/fundamental-theorems-of-mathematics-and-statistics.html

Notes