SAMPLING DISTRIBUTION

LEARNING OBJECTIVES

1. Students can understand the concept of the sampling distribution
2. Students can understand the sampling distribution of the mean, the sampling distribution of proportion, and the sampling distribution of variance.

INTRODUCTION

In a population, it must have characteristics to describe the observed population, as well as the samples obtained from that population. The characteristics of the population are called parameters, while those in the sample are called statistics. The characteristics of this population and sample can be mean, median, mode, quartile, decile, percentile, range, variance, standard deviation, mean deviation, and proportion. Of these characteristics, the ones most often used to represent data are mean, median, variance, standard deviation, and proportion. The following are symbols of the characteristics of the population and samples that are often used.

Table 1. Symbol of Characteristics of Population and Sample

Characteristics	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	S^2
Standard Deviation	σ	S
Proportion	p	$\hat{p}$

The parameters of a population can be estimated using statistics. These statistics are obtained from one or more random variables. The random variable is a function that maps the set of real numbers to each member in the sample space. The probability of each result on a random variable is called the probability distribution. There are two types of random variables, namely discrete random variables and continuous random variables. Similarly with the probability distribution, which consists of a discrete probability distribution and a continuous probability distribution.

One of the most important probability distributions in statistical analysis is the normal distribution. This normal distribution is part of a continuous probability distribution that is bell-shaped. The normal distribution is also known as the Gauss distribution. The probability density function of the random variable X is normally distributed with the mean μ and variance σ^2 $[X \sim N(\mu, \sigma^2)]$:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty, -\infty < \mu < \infty, \sigma^2 > 0$$

Here is the curve of a normal distribution.

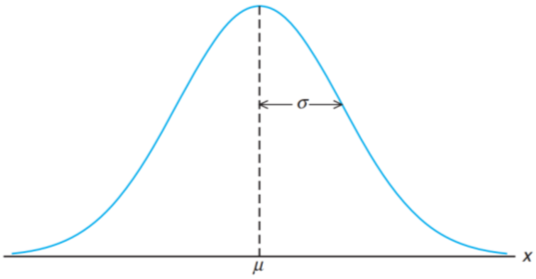
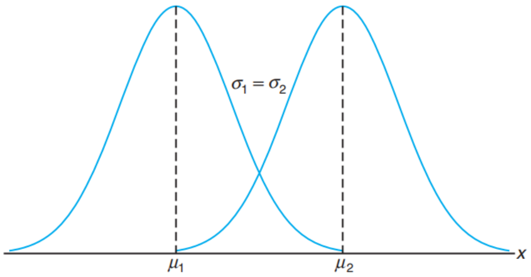


Figure 1. Normal Curve with mean μ dan variance σ^2

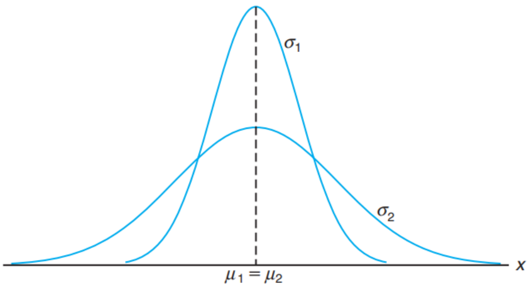
The Characteristics of a normal distribution are:

1. Bell-shape curve
2. Symmetrical curve
3. The area of the normal curve is 1.

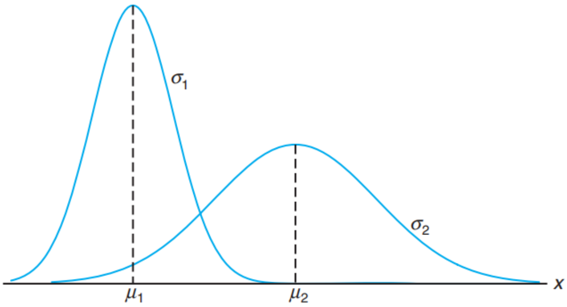
The curves of some normal distribution with several conditions can be seen in the figure below.



$\mu_1 < \mu_1$ and $\sigma_1 = \sigma_2$



$\mu_1 = \mu_1$ and $\sigma_1 < \sigma_2$



$\mu_1 < \mu_1$ and $\sigma_1 < \sigma_2$

Figure 2. Normal curve under various conditions

Based on Figure 2, it is known that the normal curve is highly dependent μ and σ. The area under the normal curve of a random variable X is limited by two points of x, namely x_1 and x_2 (see figure 3). The area of the normal distribution can be calculated with a formula:

$$P(x_1 \leq X \leq x_2) = \int_{x_2}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{x_2}^{x_2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

The area of this normal distribution is called the cumulative distribution function (CDF) and can be described as below.

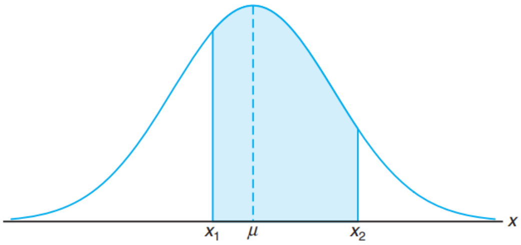


Figure 3. The area of X between x_1 and x_2

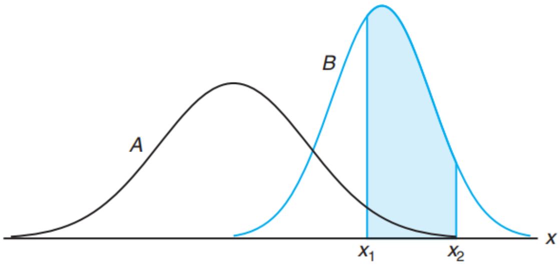


Figure 4. $P(x_1 \leq X \leq x_2)$ from different normal curves

The area between x_1 and x_2 of the two normal curves also depends on μ and σ. As in Figure 4 above, it is an area of $P(x_1 \leq X \leq x_2)$ for two curves with different mean and variance. If X is a random variable representing the distribution of A, it is represented by the shaded area under curve A and if X is a random variable representing the distribution of B, then $P(x_1 \leq X \leq x_2)$ is given by the entire blue shaded region. The shaded areas are of different sizes. Therefore, the probability associated with each distribution will be different for the two given values of X. This causes difficulty in obtaining $P(x_1 \leq X \leq x_2)$. This difficulty can be overcome by changing the two random variables X (curve A and curve B) into a new set of observations, namely a standard normal distribution$Z$ with mean 0 and variance 1 $[Z \sim N(0,1)]$. To get Z score with mean 0 and variance 1, it can be done through the transformation:

$$Z = \frac{X - \mu}{\sigma}$$

If a value of x is chosen from X, then the value of Z that can be given is $z = (x - \mu)/\sigma$. Therefore, if X is a normal distribution with values of x_1 and X with values of x_2, then $z_1 = (x_1 - \mu)/\sigma$ and $z_2 = (x_2 - \mu)/\sigma$. As a result, the CDF from Figure 4 above can be changed to:

$$P(x_1 \leq X \leq x_2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{x_2}^{x_2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \Longrightarrow \frac{1}{\sqrt{2\pi}} \int_{z_2}^{z_2} e^{-\frac{1}{2}(z)^2} dz = P(z_1 \leq Z \leq z_2)$$

and the probability density function (pdf) of the standard normal distribution is:

$$f(z; 0{,}1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2}$$

Illustration of transformation from a normal distribution $[X \sim N(\mu, \sigma^2)]$ to $[Z \sim N(0,1)]$ can be seen in Figure 5 below.

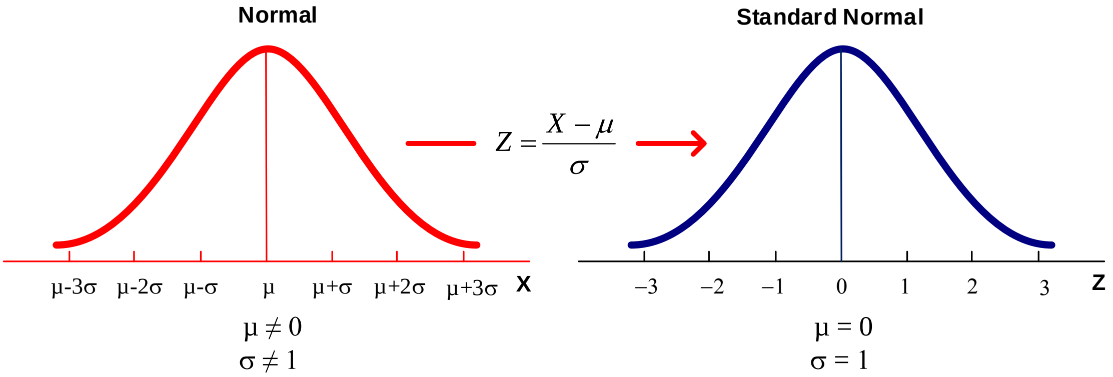


Figure 5. Normal Curve Transformation

The results of $P(z_1 \leq Z \leq z_2)$, $P(z \geq Z)$, or $P(Z \leq z)$ can be obtained in the Standard Normal Distribution Table, otherwise known as Z score table.

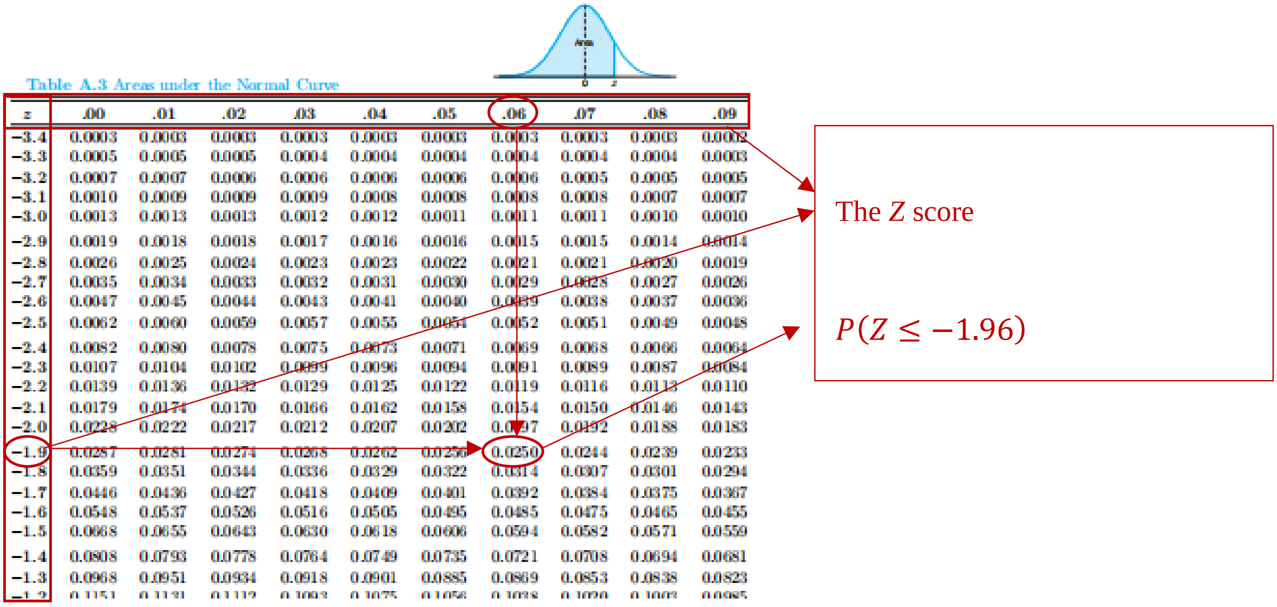


Table A.3 Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985

The *Z* score

$P(Z \leq -1.96)$

To get $P(Z > z)$, you can use the formula:

$P(Z > z) = 1 - P(Z \leq z)$ → remember, the third characteristic of the normal distribution (The total area of the normal curve is 1 and the area of the standard normal curve can be obtained from $P(z_1 \leq Z \leq z_2)$

For example, if you want to get $P(Z > -1.96)$, then:

$P(Z > -1.96) = 1 - P(Z \leq -1.96) = 1 - 0.025 = 0.975$

Example of applying the normal distribution.

Suppose that the test results of a Li-Ion 1100 mAh type battery have a mean talk time of 100 minutes with a normal distribution with a standard deviation of 5 minutes. Find the probability that a battery that is taken will be:

1. less than 95 minutes?
2. more than 105 minutes?
3. between 95 and 100 minutes?

For questions 1 to 3, the first step is to find the Z score. After obtaining the Z score, the probability can be found

The answers:

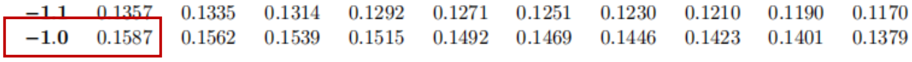
1. $P(X < 95)$?

   $\mu = 100, \sigma = 5$

   $$Z = \frac{X-\mu}{\sigma}$$

   $$= \frac{95-100}{5} = -1.00$$

   $P(Z < -1) = 0.1587 \Rightarrow \; P(X < 95)$

−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379

   Conclusion: The probability that a battery's talk time is less than 95 minutes is 15.87%.

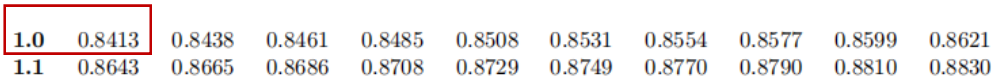
2. $P(X > 105)$?

   $\mu = 100, \sigma = 5$

   $$Z = \frac{X-\mu}{\sigma}$$

   $$= \frac{105-100}{5} = 1.00$$

   Find the area under the standard normal curve to the left of $Z = 1$ in table Z

1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830

   $P(Z \leq 1) = 0.8413 \Rightarrow P(Z > 1) = 1 - P(Z \leq 1) = 1 - 0.8413 = 0.1587$

   Conclusion: The probability that the talk time of a battery is more than 105 minutes is 15.87%.

The results of number 1 and number 2 are the same, which is 15.87%. Remember that the second characteristic of a normal distribution, which is symmetrical. That is, the mean divides the curve into two equal parts. So that $P(Z < -1) = P(Z > 1)$.

3. $P(95 < X < 105)$

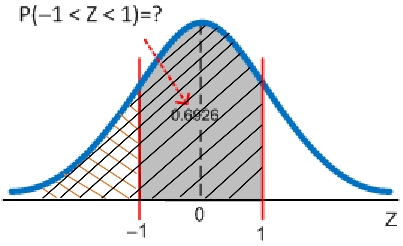
$\mu = 100, \sigma = 5$

$$Z_1 = \frac{X_1 - \mu}{\sigma}$$

$$= \frac{95 - 100}{5} = -1$$

$$Z_2 = \frac{X_2 - \mu}{\sigma}$$

$$= \frac{105 - 100}{5} = 1$$



The area shaded in black is $P(Z \leq 1)$ or $P(Z < 1)$, and the shaded orange and black are the areas $P(Z \leq -1)$ or $P(Z < -1)$ so that $P(-1 < Z < 1)$ is the area of $P(Z \leq 1) - P(Z \leq -1)$ or $P(Z < 1) - P(Z < -1)$

$$P(-1 < Z < 1) = P(Z < 1) - P(Z < -1)$$

$$= 0.8413 - 0.1587 = 0.6826$$

$P(95 < X < 105) = 0.6826$

Conclusion: The probability of a battery talk time between 95 and 105 minutes is 68.26%

SAMPLING DISTRIBUTION

Suppose a population of size N with the parameter ρ will be sampled as many as n. In random sampling, there are several possible samples selected, namely as many as k. Each possible sample that is formed has the statistic $\hat{\rho}_i$. All possible samples that are formed have a probability distribution and the probability distribution of the $\hat{\rho}_i$ is called the sampling distribution**.** In the sampling distribution, there are two methods of sampling, namely the replacement and without replacement process. The possibility of samples formed from a population of size N if a sample of size n is taken randomly is as follows:

With Replacement Process:

$$k = N^n$$

Without Replacement Process:

$$k = \binom{N}{n}$$

$$= \frac{N!}{n!\,(N-n)!}$$

Suppose there is a population of 4 people aged 18, 20, 22, and 24 years. Of the four people, 2 people will be taken with replacement and without replacement. The possible samples formed are:

$X = \{18, 20, 22, 24\}$

$N = 4$

$n = 2$

a. With Replacement Method:

$k = N^2 = 4^2 = 16$

The possibility of the sample formed by the replacement process

X		Observation 2			
		18	20	22	24
Observation 1	18	18,18	18,20	18,22	18,24
	20	20,18	20,20	20,22	20,24
	22	22,18	22,20	22,22	22,24
	24	24,18	24, 20	24,22	24,24

b. Without Replacement Method:

$$k = \binom{4}{2} = \frac{4!}{2!\,(4-2)!} = 6$$

Possible samples formed:

X		Observation 2			
		18	20	22	24
Observation 1	18		18,20	18,22	18,24
	20			20,22	
	22				22,24
	24		24, 20		

So that the number of possible samples formed from selecting with replacement is 16 possibilities and without replacement as many as 6 possibilities.

It should be noted that the normal distribution is very important and forms the basis of the sampling distribution. In the normal distribution of a population, there are two parameters, namely mean and variance. To form a normal curve, it is necessary to get the root of the variance, which is the standard deviation. Both parameters are in the normal distribution, can be estimated using statistics (mean and variance) of selected samples which are normally distributed as well. The sampling distribution can be said to be the process of obtaining the mean and standard deviation of the statistics generated from all possible samples formed. Based on this, the sampling distribution is the basis of inferential statistics, especially to obtain test statistical formulas. This test statistic is used as a critical area in the rejection or acceptance of a null hypothesis (H_0). The sampling distribution is divided into three, namely the sampling distribution of the mean, the sampling distribution of proportions, and the sampling distribution of variance.

SAMPLING DISTRIBUTION OF THE MEAN

The sampling distribution of the mean is the probability distribution of the mean of all possible samples of size n formed, which are randomly selected from a population of size N. If there are k samples of size n, then there is a mean distribution of k samples. It says the mean sampling distribution because the aim is to estimate the mean of the population. The mean and variance of the sampling distribution of the mean are as follows:

a. Mean

The set of means of all possible samples obtained from the process of taking with or without replacement in a population, each of which will form a distribution of the sample mean with the new mean being:

$$\mu_{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{X}_i}{k} = \mu$$

where:

$\mu_{\bar{X}}$: mean of the sampling distribution of the mean

$\bar{X}_i$: data from each of the mean i-th sampling distribution, $i : 1, 2, \cdots, k$

k : the number of possible samples formed

The mean obtained from the sampling distribution of the mean is equal to the mean of the population.

b. Variance

There are two conditions in determining variance, namely:

- $\frac{n}{N} \leq 5\%$ and the sample is obtained by replacement

  This condition occurs in an infinite population or when the sample is obtained by returning it. The formula for this variance is:

  $$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

  where:

  $\sigma_{\bar{X}}^2$: variance of the sampling distribution of the mean

  σ^2: variance of the population

  n : the size of each sample from the sampling distribution.

  $$\sigma_{\bar{X}} = \sqrt{\sigma_{\bar{X}}^2} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

  Z Statistical test (Z score)

  $$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \implies Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- $\frac{n}{N} > 5\%$ or the sample is obtained without replacement

  This condition occurs in the finite population or when the sample is obtained without replacement. The formula is:

$$\bar{\sigma}_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\boxed{\left(\frac{N-n}{N-1}\right)}} \longrightarrow \boxed{\text{Correction Factor}}$$

Z statistical test (*Z* score)

$$Z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \implies Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}}$$

If N is very large, and the result of $(N-n)/(N-1) \to 1$, then $\bar{\sigma}_{\bar{X}} = \sigma/\sqrt{n}$. Based on this, it is found that the distribution of all $(\bar{X}_i)$ is normal with $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}}^2 = \sigma^2/n$ $[\bar{X} \sim N(\mu, \sigma^2/n)]$. So that:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

In the sample questions about the process of determining the sample, the probability that the samples formed from the without replacement process are 6 (*k* = 6) so that we have a new data of size 6. What is the probability of the mean age over 20 years? The first step that must be done is to find the mean and standard deviation of the population. Here are the results:

$$\mu = \frac{\sum_{i=1}^{4}(x_i)}{N}$$

$$= \frac{(18+20+22+24)}{4}$$

$$= 21$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{4}(x_i-\mu)^2}{N}}$$

$$= \sqrt{\frac{(18-21)^2+(20-21)^2+(22-21)^2+(24-21)^2}{4}}$$

$$= \sqrt{\frac{(3)^2+(1)^2+(1)^2+(3)^2}{4}} = \sqrt{\frac{20}{4}} = 2.236$$

After that find the mean and standard deviation of the mean distribution.

Sample	Data	Mean $(\bar{X}_i)$	Proportion
1	18,20	19	1/6
2	18,22	20	1/6
3	18,24	21	2/6
4	20,22	21	
5	20,24	22	1/6
6	22,24	23	1/6

Become new data which will then be calculated the mean and standard deviation.

$$\mu_{\bar{X}} = \frac{\sum_{i=1}^{6}(\bar{X}_i)}{k}$$

$$= \frac{(19 + 20 + 21 + 21 + 22 + 23)}{6}$$

$$= 21$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum_{i=1}^{6}(\bar{X}_i - \mu)^2}{k}}$$

$$= \sqrt{\frac{(19-21)^2 + (20-21)^2 + 2(21-21)^2 + (22-21)^2 + (23-21)^2}{6}}$$

$$= \sqrt{\frac{10}{6}}$$

$$= 1.29$$

or

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)} = \frac{2.236}{\sqrt{2}}\sqrt{\left(\frac{4-2}{4-1}\right)} = 1.29$$

The last is looking for the probability when the mean age of the sample is more than 20 years.

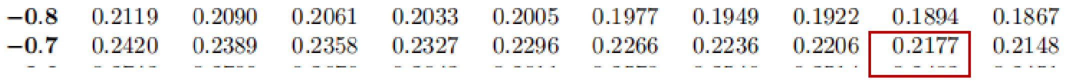
$\mu = 21 \qquad \bar{X} = 20 \qquad \sigma_{\bar{X}} = 1.29$

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \Rightarrow \frac{20 - 21}{1.29} = -0.775 = -0.78$$

$$P(Z > 0.78) = 1 - P(Z \leq 0.78)$$

$$= 1 - 0.2177$$

$$= 0.7823$$



−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148

So that the mean probability over 20 years is 0.7823 or equal to 78.23%

Central Limit Theorem

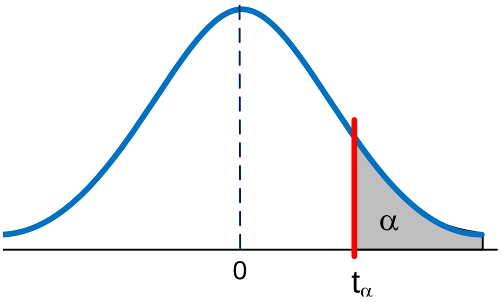
If the sample is taken from a population with unknown distribution, both finite and infinite population, the mean sampling distribution remains close to normal with mean μ and variance σ^2/n provided that the sample size is large $(n \geq 30)$. This is called the central limit theorem. And the formula for the Z statistic is:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Differences in the use of the standard normal distribution with the Central Limit Theorem

Formula of Z	**utilization**
Standard Normal Distribution: $\boldsymbol{Z = \frac{X - \mu}{\sigma}}$	Information from a mean value from normally distributed data
Central Limit Theorem: $\boldsymbol{Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}$	Information from the sampling of the mean that is normally distributed with the original population is not normally distributed

If the sample is small and σ^2 is unknown, then the test statistic used is to follow the t distribution, that is, the distribution approaches the normal distribution. which is the distribution close to the normal distribution. The t distribution is usually called the Student distribution, it is similar in shape to the standard normal distribution (Z) in that they are both closely spaced and bell-shaped. Here is the form of the t distribution.



By considering the curve above, if the area α is a bigger ($\alpha \to 1$), then the t value is getting smaller ($t \to -\infty$). The possible results of all of the t can be seen in the t distribution table. Two things that need to be considered in the t distribution table, namely:

- Degree of freedom (df) is $v = n - 1$
- α is the probability of t with certain v
- $P(T > t) = P(T < -t)$

The difference between the standard normal distribution (Z) table and t distribution table is the Z score determines the area of α while in the t distribution table, α and v are to determine the t score. The statistical t-test can be formulated as follows:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Sampling Distribution of the Difference between Two Means

Suppose there are two populations each of size N_1 and N_2 having mean μ_1 and μ_2 also variance σ_1^2 and σ_2^2. $\bar{X}_1$ represent the mean of a random sample with a sample size n_1 and $\bar{X}_2$ is the mean of a random sample with a sample size n_2. So the formula of the Z score is:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \left(\mu_{\bar{X}_1 - \bar{X}_2}\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where:

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

SAMPLING DISTRIBUTION OF PROPORTION

Sampling distribution of proportion is the distribution of the proportions obtained from all possible samples from a population, where the size of each sample is the same. The distribution of this proportion is in line with the binomial experiment where the probability of success is p and the failure is $(1-p)$, the range of is p is $0 \leq p \leq 1$. The proportion of a population is denoted by $p = X/N$ and the proportion of the sample denoted by $\hat{p} = x/n$, because as many samples as possible k, then $\hat{p}$ also as much k. From the set of proportions, the mean proportion will be calculated $\left(\mu_{\hat{p}}\right)$ and standard deviation $\left(\sigma_{\hat{p}}\right)$. The important formula for the distribution of proportion sampling can be seen in the table below.

	Infinite population	**Finite Population**
-	$\frac{n}{N} \leq 5\%$	$\frac{n}{N} > 5\%$
Proportion of sample	$\hat{p} = \frac{x}{n}$	$\hat{p} = \frac{x}{n}$
Mean	$\mu_{\hat{p}} = p$	$\mu_{\hat{p}} = p$
Standard deviation	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$
Z score	$Z = \frac{\hat{p}-p}{\sigma_{\hat{p}}}$	$Z = \frac{\hat{p}-p}{\sigma_{\hat{p}}}$

Sampling Distribution of the Difference between Two Proportions

For example in the first population has a sample size n_1, then this sample has a proportion $\hat{p}_1 = x_1/n_1$ and the second population has a sample size n_2, then the proportion is $\hat{p}_2 = x_2/n_2$. The formula for the sampling distribution of the difference between two proportions is as follows:

	Formula
Proportion	$\hat{p}_1 = x_1/n_1$ dan $\hat{p}_2 = x_2/n_2$
Mean of the difference between two proportions	$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$
Standard deviation in $\frac{n}{N} \leq 5\%$	$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \cdot \sqrt{\frac{(N_1+N_2)-(n_1+n_1)}{(N_1-N_2)-1}}$
Standard deviation in $\frac{n}{N} > 5\%$	$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Z score	$Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sigma_{\hat{p}_1 - \hat{p}_2}}$

Since the proportion is in line with the binomial experiment, the normal approximation to the binomial applies. Continuity correction can be applied to the discrete value of the variable X by adding or subtracting $0.5/n$ (the first way) or to the proportional variable by adding or subtracting $0.5/n$ (the second way). Continuity correction can be seen in the following table:

Determine	The Correction Used
1. $\boldsymbol{P(X = a)}$	$\boldsymbol{P}(a - 0.5/n < \boldsymbol{X} < \boldsymbol{a} + 0.5/n)$
2. $\boldsymbol{P(X \geq a)}$	$\boldsymbol{P(X > a} - 0.5/n)$
3. $\boldsymbol{P(X > a)}$	$\boldsymbol{P(X > a} + 0.5/n)$
4. $\boldsymbol{P(X \leq a)}$	$\boldsymbol{P(X < a} + 0.5/n)$
5. $\boldsymbol{P(X < a)}$	$\boldsymbol{P(X < a} - 0.5/n)$

SAMPLING DISTRIBUTION OF VARIANCE

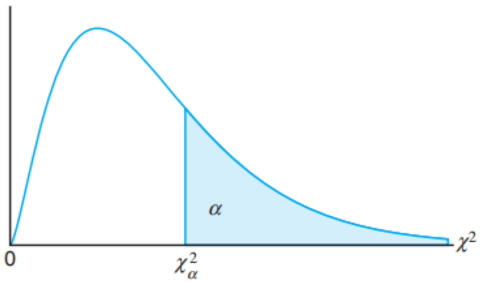
If a random sample of size n is drawn from a normal population with mean μ and variance σ^2, and the sample variance is computed, we obtain a value of the statistic S^2. We shall proceed to consider the distribution of the statistic:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

S^2 is the variance of a random sample of size n taken from a population, the statistical test above results in a Chi-Square distribution of degrees of freedom $v = n - 1$. The formula of S^2 is follow:

$$S^2 = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right)$$

The probability that a random sample results in a value χ^2 that is greater than the multiple specified values equals the area under the curve to the right of the value. This can be seen in the image below.



Because the chi-square distribution is not symmetrical, the area on the right and left can be different.

EXAMPLE

1. Suppose a Li-Ion 1100 mAh battery has a mean talk time of 100 minutes and is not normally distributed with a standard deviation of 10 minutes. If the random sampling is 35 batteries, what is the probability that the batteries will have a mean life of fewer than 98 minutes?

   The answer:

   $P(\bar{X} < 98)$?

   It is known that what we want to observe is 35 batteries (>30), fulfilling the central limit theory because the population does not come from a normal distribution and the sample is more than 30. So that to get the Z score is not from the formula Z score is the standard normal distribution. However, using the formula of the Z score from the central limit theorem can be applied to this problem.

   $\mu = 100$ and $\sigma = 25$

   $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{98 - 100}{10/\sqrt{25}} = -1$$

   Using the Z distribution table, it is obtained:

   $P(\bar{X} < 98) = P(Z < -1) = \boxed{0.1587}$ → See table Z

   Conclusion: The probability that the mean talk time of 35 batteries is less than 98 minutes is 15.87%

2. A sample of size $n_1 = 5$ is taken randomly from a normally distributed population with $\mu_1 = 50$ and variance $\sigma_1^2 = 9$. The second random sample of size $n_2 = 4$ is taken from another population which is free from the first which is also normally distributed, with $\mu_2 = 40$ and variance $\sigma_2^2 = 4$. From the two samples, the mean was calculated. What is the probability that the mean difference between the first and second samples is less than 8.2?

   $P(\bar{X}_1 - \bar{X}_2 < 8.2)$?

   The mean of the difference between the two means:

   $\mu_{\bar{X}_1 - \bar{X}_2} = 50 - 40 = 10$

   The standard deviation of the difference between the two means:

   $$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{9}{5} + \frac{4}{4}} = \sqrt{\frac{14}{5}} = 1.67$$

   Calculate *Z* score:

   $\bar{X}_1 - \bar{X}_2 = 8.2$

   $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \left(\mu_{\bar{X}_1 - \bar{X}_2}\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

   $$Z = \frac{8.2 - 10}{1.67} = -1.08$$

   $P(\bar{X}_1 - \bar{X}_2 < 8.2) = P(Z < -1.08) = 0.1401$

   Conclusion:

   the probability that the mean difference between the first and second samples is less than 8.2 is 14.01%.

3. It is known that 10% of housewives in Bandung use detergent A to wash their clothes. Suppose that a sample of 100 is taken from the population, determine:

   - The mean and standard deviation of a sample with the population is housewives using detergent A? (it is assumed that the population is normally distributed).
   - If from the sample there are at least 15 housewives who use detergent A, what is the probability?

   Answer:

   Proportion: 10% $\longrightarrow$ $\mu_{\hat{p}} = p = 10\% = 0.1$

   Standard deviation:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.1(0.9)}{100}} = 0.03$$

So the mean and standard deviation of samples from housewives who used the detergent A in Bandung were: 0.1 and 0.03, respectively.

$P(X \geq 15)$?

The proportion of detergent A usage is 15 people is:

$$\hat{p} = \frac{x}{n} = \frac{15}{100} = 0.15$$

To get the Z score, the first step is the sample proportion is reduced by a continuity correction of 0.5/100, which is 0.005 so that $\hat{p} = 0.15 - 0.005 = 0.145$.

$$Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{1.45 - 0.1}{0.03} = 1.5$$

$$P(X \geq 15) = P(Z \geq 1.5) = 1 - P(Z < 1.5) = 1 - 0.9332 = 0.0668$$

Conclusion:

The probability of housewives in Bandung using at least 15 detergent attacks is 6.68%.

4. A battery manufacturer guarantees 95% that the battery will last a mean of 3 years with a standard deviation of 1 year. If a sample of five batteries that last 1.9, 2.4, 3.0, 3.5, and 4.2 years are taken, is the manufacturer still sure that the standard deviation of these batteries is 1 year?

   Answer:

   Calculate the variance of the sample:

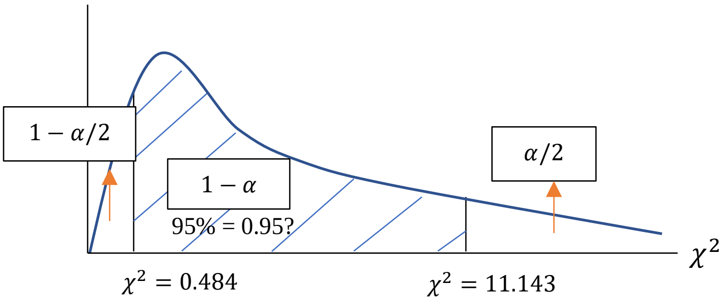
$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{n\sum_{i=1}^{n}(x_i)^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)} = \frac{5(48.26) - 15^2}{5(4)} = 0.815$$

   Calculate the statistical test.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{4(0.815)}{1} = 3.26$$

   The statistical test above is at 4 degrees of freedom, then it will be calculated that 95% of the batteries have a standard deviation of 1 year.

Because what is wanted is a guarantee of 95% of the standard deviation of the lamp is 1 year so that the $\alpha = 5\% = 0.05$ and the $1 - \alpha$ position is between $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$. The area can be described on the chi-square curve as follows.



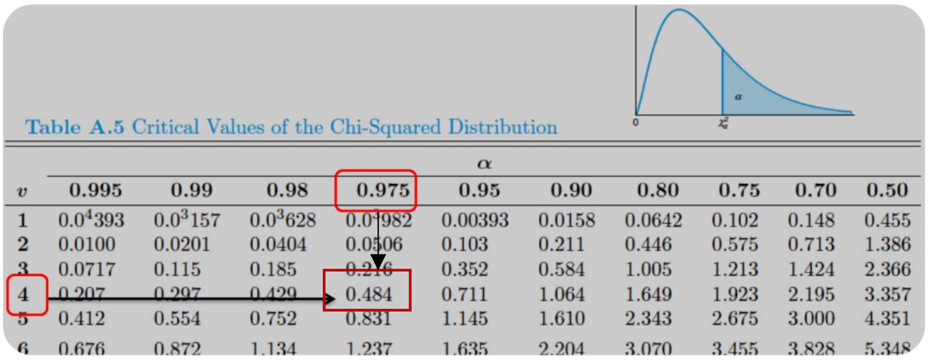

0
χ^2_α
α

Table A.5 Critical Values of the Chi-Squared Distribution

	α									
v	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.75	0.70	0.50
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348

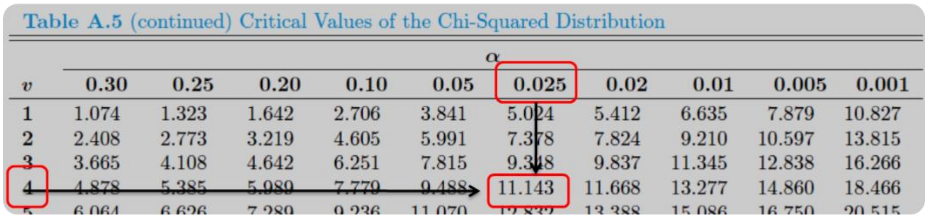


Table A.5 (continued) Critical Values of the Chi-Squared Distribution

	α									
v	0.30	0.25	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.466
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.515

Since 3.26 is between 0.484 and 11,143, the standard deviation of battery life is indeed 1 year.