# Chapter 5 Statistical Inference

## 5.1. Estimation

The assignment of value(s) to a population parameter based on a value of the corresponding sample statistic is called ***estimation***.

The value(s) assigned to a population parameter based on the value of a sample statistic is called an ***estimate***.

The sample statistic used to estimate a population parameter is called an ***estimator***.

Two types of estimation.

1. A ***point estimate*** is a single value (or point) used to approximate a population parameter.

Example:

(i) the sample proportion, $\hat{p}$ , is the best point estimate of the population proportion, $p$.

(ii) the sample mean, $\overline{X}$ , is the best point estimate of the population mean, $\mu$ .

(iii) the sample variance, $s^2$, is the best point estimate of the population variance, $\sigma^2$ .

2. An ***interval estimation*** is an interval that constructed around the point estimate, and it is stated that this interval is likely to contain the true value of a population parameter.

Each interval is constructed with regard to a given ***confidence level*** and is called a ***confidence interval***. The confidence interval is given as

$$\text{Point estimate} \pm \text{Margin of error.}$$

The confidence level associated with a confidence interval states how much confidence we have that this interval contains the true population parameter. The confidence level is denoted by $(1-\alpha)100\%$ .

Note.

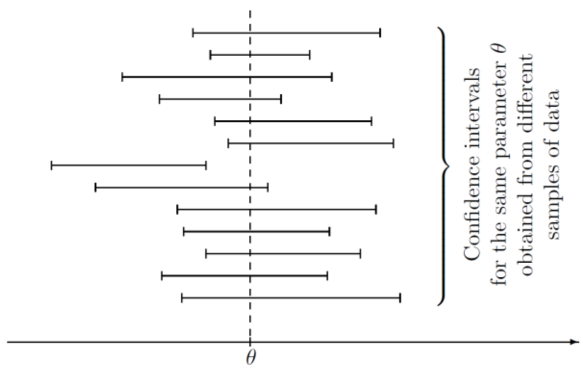
The standard confidence level is **95%**.

## 5.2. Confidence Intervals

An interval $[a,b]$ is a $(1-\alpha)100\%$ ***confidence interval*** for the parameter $\theta$ if it contains the parameter with probability $(1-\alpha)$,

$$P(a \le \theta \le b) = 1-\alpha .$$

The ***coverage probability*** $(1-\alpha)$ is also called a ***confidence level***.

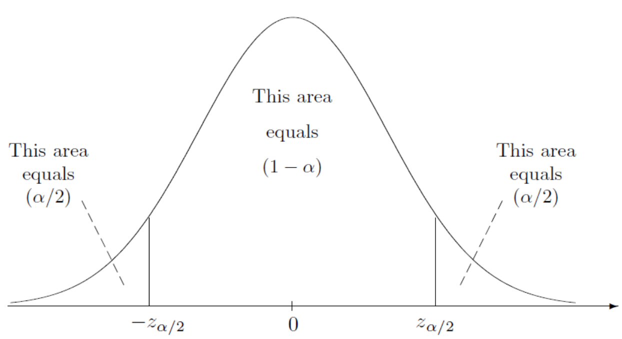



***Construction of Confidence Intervals: A General Method***

Assume there is an unbiased estimator $\hat{\theta}$ that has a Normal distribution. When we standardize it, we get a Standard Normal variable

$$Z = \frac{\hat{\theta} - E(\hat{\theta})}{\sigma(\hat{\theta})} = \frac{\hat{\theta}-\theta}{\sigma(\hat{\theta})},$$

where $E(\hat{\theta}) = \theta$ because $\hat{\theta}$ is unbiased, and $\sigma(\hat{\theta})$ is its standard error.




This variable falls between $-z_{\alpha/2}$ and $z_{\alpha/2}$ with probability $(1-\alpha)$ as shown in the figure. Then,

$$P\left(-z_{\alpha/2} \le \frac{\hat{\theta}-\theta}{\sigma(\hat{\theta})} \le z_{\alpha/2}\right) = 1-\alpha .$$

Solving the inequality inside $(\ldots)$ for $\theta$, we get

$$P\left(\hat{\theta} - z_{\alpha/2}\sigma(\hat{\theta}) \le \theta \le \hat{\theta} + z_{\alpha/2}\sigma(\hat{\theta})\right) = 1-\alpha .$$

The problem is solved! We have obtained two numbers

$$a = \hat{\theta} - z_{\alpha/2}\sigma\left(\hat{\theta}\right)$$

$$b = \hat{\theta} + z_{\alpha/2}\sigma\left(\hat{\theta}\right)$$

such that

$$P\left(a \le \theta \le b\right) = 1 - \alpha .$$

If parameter $\theta$ has an unbiased, Normally distributed estimator $\hat{\theta}$, then,

$$\hat{\theta} \pm z_{\alpha/2}\sigma\left(\hat{\theta}\right) = \left[\hat{\theta} - z_{\alpha/2}\sigma\left(\hat{\theta}\right), \hat{\theta} + z_{\alpha/2}\sigma\left(\hat{\theta}\right)\right]$$

is a $(1-\alpha)100\%$ confidence interval for $\theta$. In this formula, $\hat{\theta}$ is the center of the interval, and $z_{\alpha/2}\sigma\left(\hat{\theta}\right)$ is the margin.

If the distribution of $\hat{\theta}$ is *approximately* Normal, we get an approximately $(1-\alpha)100\%$ confidence interval.

***Sampling From A Normally Distributed Population***

Suppose $X \sim N\left(\mu, \sigma^2\right)$, then $\overline{X} \sim N\left(\mu_{\overline{X}} = \mu, \sigma_{\overline{X}}^2 = \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$, irrespective of sample size.

Then $Z = \dfrac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

*Example 5.1.*

In a recent STAT test, the mean score for all examinees was 1016. Assume that the distribution of STAT scores of all examinees is normal with a mean of 1016 and a standard deviation of 153. Let $\overline{X}$ be the mean STAT score of a random sample of certain examinees. Calculate the mean and standard deviation of $\overline{X}$ and describe the shape of its sampling distribution when the sample size is

(a) 16 (b) 1000.

Solution:

***Sampling From A Population That Is Not Normally Distributed***

According to the ***Central Limit Theorem***, for a ***relatively large*** sample size, the sampling distribution of $\overline{X}$ is *approximately normal*, irrespective of the shape of the population distribution under consideration. Hence, for all distribution of *X*, if *n* is large, by Central Limit Theorem,

$$\overline{X} \approx N\left( \mu_{\overline{X}} = \mu, \sigma_{\overline{X}}^2 = \left( \frac{\sigma}{\sqrt{n}} \right)^2 \right).$$

Then $Z = \dfrac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Remark.

1. The sample size is usually considered to be large if $n \geq 30$.
2. As sample size increases, the sampling distribution of $\overline{X}$ behaves more like normal distribution and hence, the approximation is better.

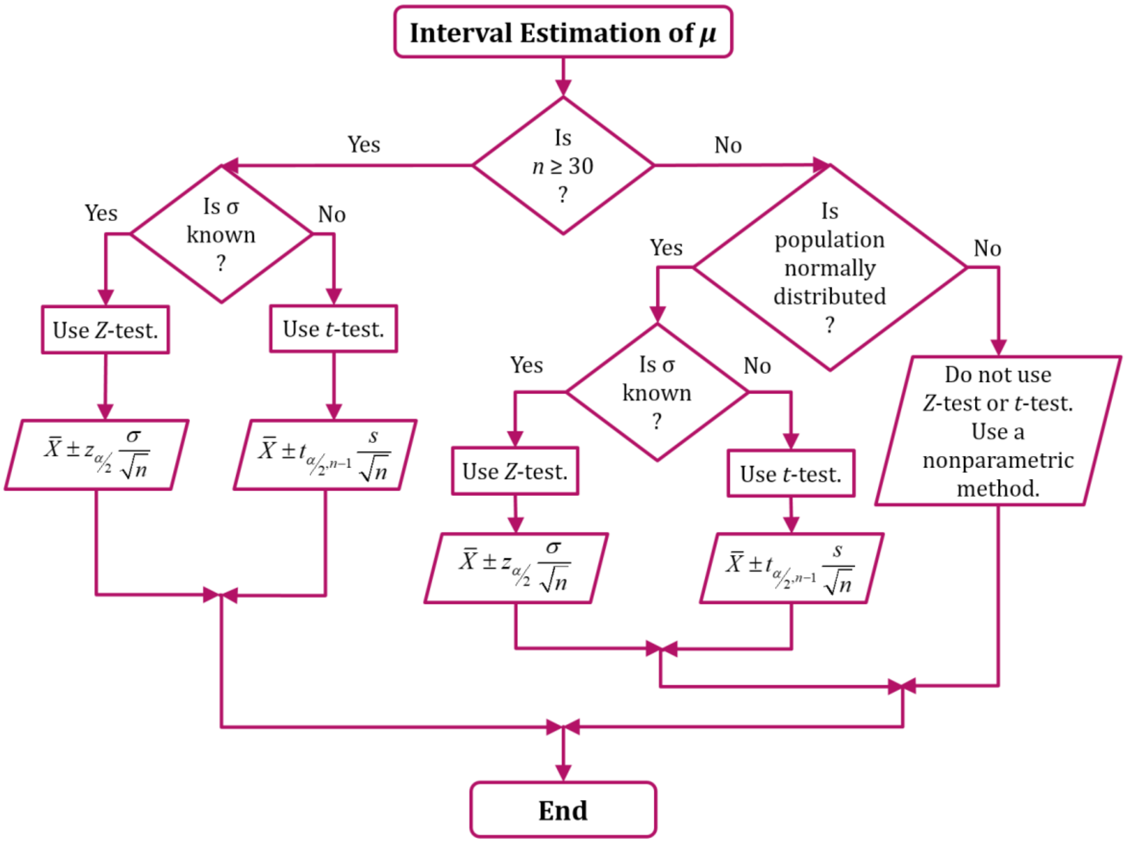
*Example 5.2.*

The mean rent paid by all tenants in a large city is RM1250 with a standard deviation of RM225. However, the population distribution of rents for all tenants in this city is skewed to the right. Calculate the mean and standard deviation of $\overline{X}$ and describe the shape of its sampling distribution when the sample size is 100.

Solution:

## 5.3. Interval Estimation of A Population Mean

***Flowchart for Interval Estimation of One Sample Mean, μ.***




***The Student's t Distribution***

The Student's *t* distribution is a specific type of bell-shaped distribution with a lower height and a wider spread than the standard normal distribution. As the sample size becomes larger, the *t distribution* approaches the standard normal distribution. The *t* distribution has only one parameter, called the degrees of freedom (*df*), and is denoted by $\nu$.

In many similar problems, degrees of freedom can be computed as

$$\begin{matrix} Number\ of \\ degrees\ of\ freedom \end{matrix} = sample\ size - \begin{matrix} number\ of\ estimated \\ location\ parameters \end{matrix}.$$

**Q:** What if the sample size is large and the number of *df* is not in the *t* distribution table?
**A:** There are two options.

1. Use the *t* values from the last row (the row of $\infty$) in the *t* distribution table.
2. Use the normal distribution as an approximation on the *t* distribution.

***Large Sample ($n \geq 30$)***

The $(1-\alpha)100\%$ confidence interval for $\mu$ for large samples ($n \geq 30$) is

$$\begin{cases} \bar{X} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}, & \text{if } \sigma \text{ is known} \\ \bar{X} \pm t_{\alpha/2, n-1} \dfrac{s}{\sqrt{n}}, & \text{if } \sigma \text{ is not known} \end{cases}.$$

The value of $Z$ and $t$ used here are read from the standard normal distribution table and the $t$ distribution table respectively, for the given confidence level.

<u>*Example 5.3.*</u>

A research department took a sample of 36 textbooks and collected information on their prices. This information produced a mean price of RM154.40. It is known that the standard deviation of the prices of all textbooks is RM4.50.

(a) What is the point estimate of the mean price of all textbooks?

(b) Construct a 90% confidence interval for the mean price of all textbooks.

Solution:

***Sample Size for Estimating Mean $\mu$***

A confidence interval can also be described as

$$\text{center} \pm \text{margin}$$

where

$$\text{center, } \hat{\theta} = \bar{x} \text{ , and margin, } E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} .$$

In order to attain a margin of error $E$ for estimating a population mean with a confidence level $(1-\alpha)$, a sample size of

$$n \geq \left[ \frac{z_{\alpha/2}\sigma}{E} \right]^2$$

is required.

When finding the sample size *n*, if the use of formula *does not* result in a whole number, always ***increase/ round up*** the value of *n* to the next *larger* whole number.

Remark.

When $\sigma$ is not known, estimate $\sigma$ using these methods.

1. $\sigma \cong range/4$ .
2. Estimate the value of $\sigma$ by using the earlier result.

*Example 5.4.*

An alumni association wants to estimate the mean debt of this year's college graduates. It is known that the population standard deviation of the debts of this year's college graduates is $11800. How large a sample should be selected so that the estimate with a 99% confidence level is within $800 of the population mean?

Solution:

***Small Sample (n < 30)***

The $(1-\alpha)100\%$ confidence interval for $\mu$ for small samples ($n < 30$) is

$$\begin{cases} \bar{X} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}} \quad , \ \ if\ \sigma\ is\ known \\ \bar{X} \pm t_{\alpha/2\,,\,n-1} \dfrac{s}{\sqrt{n}}, \ \ if\ \sigma\ is\ not\ known \end{cases}.$$

The value of $Z$ and $t$ used here are read from the standard normal distribution table and the $t$ distribution table respectively, for the given confidence level.

*<u>Example 5.5.</u>*

If an unauthorized person accesses a computer account with the correct username and password (stolen or cracked), can this intrusion be detected? Recently, a number of methods have been proposed to detect such unauthorized use. The time between keystrokes, the time a key is depressed, the frequency of various keywords are measured and compared with those of the account owner. If there are significant differences, an intruder is detected. The following times between keystrokes were recorded when a user typed the username and password:

.24, .22, .26, .34, .35, .32, .33, .29, .19, .36, .30, .15, .17, .28, .38, .40, .37, .27 seconds.

As the first step in detecting an intrusion, let's construct a 99% confidence interval for the mean time between keystrokes assuming Normal distribution of these times.

Solution:

## 5.4. Interval Estimation for The Difference Between Two Population Means for Independent Samples

The Conditions to Construct Confidence Interval for $\mu_X - \mu_Y$.

1. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two independent random samples of sizes $n$ and $m$ respectively.

2. At least one of the following two conditions is fulfilled:
   i. Both samples are large.
      i.e. $n \geq 30$ and $m \geq 30$

   ii. If either one or both sample sizes are small, then both populations from which the samples are drawn are normally distributed.
      i.e. $X \sim N\left(\mu_X, \sigma_X^2\right)$ and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$

***σX and σY Known***

The $(1-\alpha)100\%$ confidence interval for $\mu_X - \mu_Y$ is

$$\left(\bar{X} - \bar{Y}\right) \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, \ \ if\ \sigma_X\ and\ \sigma_Y\ are\ known\,.$$

***σX and σY Not Known***

The $(1-\alpha)100\%$ confidence interval for $\mu_X - \mu_Y$ is

$$\begin{cases} \left(\bar{X} - \bar{Y}\right) \pm t_{\alpha/2,\ df_{\sigma_X = \sigma_Y}}\, s_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}, & if\ \sigma_X\ and\ \sigma_Y\ not\ known\ but\ \sigma_X = \sigma_Y \\ \left(\bar{X} - \bar{Y}\right) \pm t_{\alpha/2,\ df_{\sigma_X \neq \sigma_Y}}\sqrt{\dfrac{s_X^2}{n} + \dfrac{s_Y^2}{m}}, & if\ \sigma_X\ and\ \sigma_Y\ not\ known\ but\ \sigma_X \neq \sigma_Y \end{cases}$$

where

$$s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}} \qquad \textbf{\textit{pooled standard deviation}}$$

$$df_{\sigma_X = \sigma_Y} = n + m - 2$$

$$df_{\sigma_X \neq \sigma_Y} = \frac{\left(\dfrac{s_X^2}{n} + \dfrac{s_Y^2}{m}\right)^2}{\dfrac{1}{n-1}\left(\dfrac{s_X^2}{n}\right)^2 + \dfrac{1}{m-1}\left(\dfrac{s_Y^2}{m}\right)^2}. \qquad \textbf{\textit{Satterthwaite approximation}}$$

The value of *Z* and *t* are obtained from the standard normal distribution table and the *t* distribution table respectively, for the given confidence level.

Note that the number of degrees of freedom for $\left(\sigma_X \neq \sigma_Y\right)$ often appears non – integer. The number given by this formula is always rounded down for *df*.

*Example 5.6.*

Internet connections are often slowed by delays at nodes. Let us determine if the delay time increases during heavy-volume times. Five hundred packets are sent through the same network between 5 pm and 6 pm (sample X), and three hundred packets are sent between 10 pm and 11 pm (sample Y). The early sample has a mean delay time of 0.8 sec whereas the second sample has a mean delay time of 0.5 sec. Assume that the population standard deviations are 0.1 sec and 0.08 sec, respectively. Construct a 99% confidence interval for the difference between the mean delay times.

Solution:

*Example 5.7.*

CD writing is energy consuming; therefore, it affects the battery lifetime on laptops. To estimate the effect of CD writing, 30 users are asked to work on their laptops until the “low battery” sign comes on. Eighteen users without a CD writer worked an average of 5.3 hours with a standard deviation of 1.4 hours. The other twelve, who used their CD writer, worked an average of 4.8 hours with a standard deviation of 1.6 hours. Assuming Normal distributions with equal population variances $\left(\sigma_X^2 = \sigma_Y^2\right)$, construct a 95% confidence interval for the battery life reduction caused by CD writing.

Solution:

*Example 5.8.*
An account on server A is more expensive than an account on server B. However, server A is faster. To see if it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is. A certain computer algorithm is executed 30 times on server A and 20 times on server B with the following results,

| | Server A | Server B |
|---|---|---|
| Sample mean | 6.7 min | 7.5 min |
| Sample standard deviation | 0.6 min | 1.2 min |

Construct a 95% confidence interval for the difference $\mu_1 - \mu_2$ between the mean execution times on server A and server B, assuming that the observed times are approximately Normal.
Solution:

## 5.5. Interval Estimation of Population Proportions: Large Sample

Assume a subpopulation $A$ of items that have a certain *attribute*. By the ***population proportion*** we mean the probability

$$p = \mathbf{P}\{i \in A\}$$

for a randomly selected item $i$ to have this attribute.

A ***sample proportion***

$$\hat{p} = \frac{\text{number of sampled items from } A}{n}$$

is used to estimate $p$.

It is convenient to use the *indicator* variables

$$X_i = \begin{cases} 1 & if \quad i \in A \\ 0 & if \quad i \notin A \end{cases}.$$

Each $X_i$ has Bernoulli distribution with parameter $p$. In particular,

$$E(X_i) = p \qquad \text{and} \qquad Var(X_i) = pq, \; q = 1 - p.$$

Also,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is nothing but a sample mean of $X_i$. Therefore,

$$E(\hat{p}) = p \qquad \text{and} \qquad Var(\hat{p}) = \frac{pq}{n}, \; q = 1 - p.$$

***Central Limit Theorem for Sample Proportion***

According to the ***Central Limit Theorem***, the *sampling distribution* of $\hat{p}$ is approximately **normal** for a sufficiently large sample size. In the case of proportion, the sample size is considered to be large if *np* and *nq* are both greater than 5.

Suppose $np > 5$ and $nq > 5$, then

$$\hat{p} \approx N\left(\mu_{\hat{p}} = p, \; \sigma_{\hat{p}}^2 = \frac{pq}{n}\right).$$

Hence,

$$Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{pq/n}} \sim N(0,1).$$

If *p* and *q* are unknown, then $n\hat{p}$ and $n\hat{q}$ should each be greater than 5 for the sample to be large.

***Confidence Interval for the Population Proportion, p***

The $(1-\alpha)100\%$ confidence interval for $p$ for large samples is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

The value of $Z$ is obtained from the standard normal distribution table for the given confidence level.

<u>*Example 5.9.*</u>

In a survey, we found that 51% of 829 adults are opposed to the use of photo-cop for issuing traffic tickets.

(a) Find the margin of error, $E$ that corresponds to a 95% confidence interval.

(b) Construct a 95% confidence interval of the population proportion, $p$.

Solution:

***Sample Size for Estimating Proportion p***

The margin of error for $p$, denoted by $E$, is the quantity that is subtracted from and added to the value of $\hat{p}$ to obtain a confidence interval for $p$. Thus,

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \;.$$

In order to attain a margin of error $E$ for estimating a population mean with a confidence level $(1-\alpha)$, a sample size of

$$\begin{cases} n \ge \dfrac{z_{\alpha/2}^2\,\hat{p}\hat{q}}{E^2} & , \; \hat{p} \textit{ is known} \\ n \ge \dfrac{z_{\alpha/2}^2\,(0.25)}{E^2} & , \; \hat{p} \textit{ is not known} \end{cases}$$

Note that 0.25 is the largest value of $\hat{p}\hat{q}$ .

*Example 5.10.*

Lombard Electronics Company has just installed a new machine that makes a part that is used in clocks. The company wants to estimate the proportion of these parts produced by this machine that are defective. The company manager wants this estimate to be within 0.02 of the population proportion for a 95% confidence level. What is the most conservative estimate of the sample size that will limit the margin of error to within 0.02 of the population proportion?

Solution:

*Example 5.11.*

Reconsider *Example 5.10.*. Suppose a preliminary sample of 200 parts produced by this machine showed that 7% of them are defective. How large a sample should the company select so that the 95% confidence interval for $p$ is within 0.02 of the population proportion?

Solution:

***Confidence Interval for the Difference of Proportions, $p_1 - p_2$***

For two large and independent samples, the sampling distributions of $\hat{p}_1 - \hat{p}_2$ is (approximately) normal, with its mean and standard deviation given as

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 \qquad \text{and} \qquad \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

respectively, where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

The $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ for large samples is

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}.$$

The value of $Z$ is obtained from the standard normal distribution table for the given confidence level.

*Example 5.12.*

A candidate prepares for the local elections. During his campaign, 42 out of 70 randomly selected people in town A and 59 out of 100 randomly selected people in town B showed they would vote for this candidate. Estimate the difference in support that this candidate is getting in towns A and B with 95% confidence. Can we state affirmatively that the candidate gets a stronger support in town A?

Solution:

## 5.6. <u>Basics of Hypothesis Testing</u>

In statistics, a *hypothesis* is a claim or statement about a property of a population.

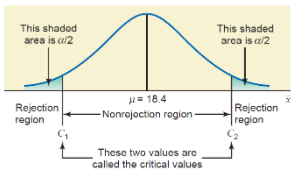
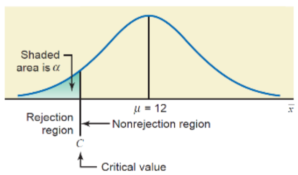
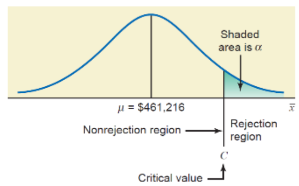
A *hypothesis test* (or *test of significance*) is a standard procedure for testing a claim about population parameter.

***Steps to Perform a Test of Hypothesis with the Critical-Value Approach***
1. State the null and alternative hypothesis.
2. Select the distribution to use (test statistic).
3. Calculate the value of the test statistic.
4. Determine the rejection and non-rejection regions or *p*-value.
5. Make a decision and draw conclusion.

<u>Components of a Formal Hypothesis Test.</u>
***Null Hypothesis, $H_0$***
The *null hypothesis* (denoted by $H_0$) assumes the value of a population is true unless otherwise declared false.

***Alternative Hypothesis, $H_1$***
The *alternative hypothesis* (denoted by $H_1$ or $H_a$) contradicts the null hypothesis.

***Tails of a Test***
The *tails* in a distribution are the extreme regions bounded by critical values.

<table>
<tr><td></td><td>Two-tailed test</td><td>Left-tailed test</td><td>Right-tailed test</td></tr>
<tr><td>Sign in $H_0$</td><td>$=$</td><td>$=$ or $\geq$</td><td>$=$ or $\leq$</td></tr>
<tr><td>Sign in $H_1$</td><td>$\neq$</td><td>$<$</td><td>$>$</td></tr>
<tr><td>Rejection region</td><td>In both tails</td><td>In the left tail</td><td>In the right tail</td></tr>
<tr><td>Illustration</td><td>
</td><td>
</td><td>
</td></tr>
</table>

<u>*Example 5.13.*</u>
Write the null hypothesis and alternative hypothesis for the following cases. Determine whether each is a case of two-tailed, a left-tailed or a right-tailed test.
The average connection speed is 54Mbps, test if the connection speed
(a) has changed, (b) is getting slower, (c) is getting faster.
Solution:
Let $\mu$ be the average connection speed

***Test Statistic***

The *test statistic* is a value computed from the sample data, and it is used for decision making about the rejection of null hypothesis. It is only to reject or do not reject $H_0$ but has **no effect** on $H_1$.

Assume the population from which the sample is selected is normally distributed.

Summary of $Z$ – tests.

| Null hypothesis $H_0$ | Parameter, Estimator $\theta,\ \hat{\theta}$ | If $H_0$ is true: $E(\hat{\theta})$ | If $H_0$ is true: $Var(\hat{\theta})$ | Test statistic $Z=\dfrac{\hat{\theta}-\theta_0}{\sqrt{Var(\hat{\theta})}}$ |
|---|---|---|---|---|
| One-sample Z-tests for means and proportions, based on a sample of size $n$. | | | | |
| $\mu=\mu_0$ | $\mu,\ \bar{X}$ | $\mu_0$ | $\dfrac{\sigma^2}{n}$ | $Z=\dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ |
| $p=p_0$ | $p,\ \hat{p}$ | $p_0$ | $\dfrac{p_0q_0}{n}$ | $Z=\dfrac{\hat{p}-p_0}{\sqrt{\dfrac{p_0q_0}{n}}}$ |
| Two-sample Z-tests comparing means and proportions of two populations, based on independent samples of size $n$ and $m$. | | | | |
| $\mu_X-\mu_Y=D$ | $\mu_X-\mu_Y$, $\bar{X}-\bar{Y}$ | $\mu_X-\mu_Y=D$ | $\dfrac{\sigma_X^2}{n}+\dfrac{\sigma_Y^2}{m}$ | $Z=\dfrac{(\bar{X}-\bar{Y})-D}{\sqrt{\dfrac{\sigma_X^2}{n}+\dfrac{\sigma_Y^2}{m}}}$ |
| $p_1-p_2=D$ | $p_1-p_2$, $\hat{p}_1-\hat{p}_2$ | $p_1-p_2=D$ | $\dfrac{p_1q_1}{n}+\dfrac{p_2q_2}{m}$ | $Z=\dfrac{(\hat{p}_1-\hat{p}_2)-D}{\sqrt{\dfrac{p_1q_1}{n}+\dfrac{p_2q_2}{m}}}$ |
| $p_1=p_2$ | $p_1-p_2$, $\hat{p}_1-\hat{p}_2$ | $p_1-p_2=0$ | $pq\left(\dfrac{1}{n}+\dfrac{1}{m}\right)$, $p=p_1=p_2$ | $Z=\dfrac{(\hat{p}_1-\hat{p}_2)-D}{\sqrt{\bar{p}\bar{q}\left(\dfrac{1}{n}+\dfrac{1}{m}\right)}}$ <br> $\bar{p}=\dfrac{x_1+x_2}{n+m}=\dfrac{n\hat{p}_1+m\hat{p}_2}{n+m}$ |

Summary of $t$ – tests.

| Null hypothesis $H_0$ | Conditions | Test statistic $t$ | Degrees of freedom |
|---|---|---|---|
| $\mu = \mu_0$ | Sample size $n$; unknown $\sigma$ | $t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$ | $n-1$ |
| $\mu_X - \mu_Y = D$ | Sample sizes $n$, $m$; unknown but equal $\sigma_X = \sigma_Y$ | $t = \dfrac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\dfrac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$ | $n+m-2$ |
| $\mu_X - \mu_Y = D$ | Sample sizes $n$, $m$; unknown, unequal $\sigma_X \neq \sigma_Y$ | $t = \dfrac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\dfrac{s_X^2}{n} + \dfrac{s_Y^2}{m}}}$ | $\dfrac{\left(\dfrac{s_X^2}{n} + \dfrac{s_Y^2}{m}\right)^2}{\dfrac{\left(s_X^2/n\right)^2}{n-1} + \dfrac{\left(s_Y^2/m\right)^2}{m-1}}$ |

***Type I and Type II Errors: Level of Significance***

A *Type I error* occurs when a true null hypothesis is rejected. The value of $\alpha$ represents the probability of committing this type of error; that is

$$\alpha = P\left(H_0 \text{ is rejected} \mid H_0 \text{ is true}\right).$$

The value of $\alpha$ represents the *significance level* of the test.

A *Type II error* occurs when a false null hypothesis is not rejected. The value of $\beta$ represents the probability of committing a Type II error; that is

$$\beta = P\left(H_0 \text{ is not rejected} \mid H_0 \text{ is false}\right).$$

The value of $1-\beta$ is called the *power of the test*. It represents the probability of not making a Type II error or otherwise known as the probability of rejecting a false hypothesis.

Four possible outcomes for a test of hypothesis.

| | | **Actual Situation** | |
|---|---|---|---|
| | | **$H_0$ is true** | **$H_0$ is false** |
| Decision | Reject $H_0$ | Type I error | Correct decision |
| | Do not reject $H_0$ | Correct decision | Type II error |

***Rejection and Non-rejection Region***

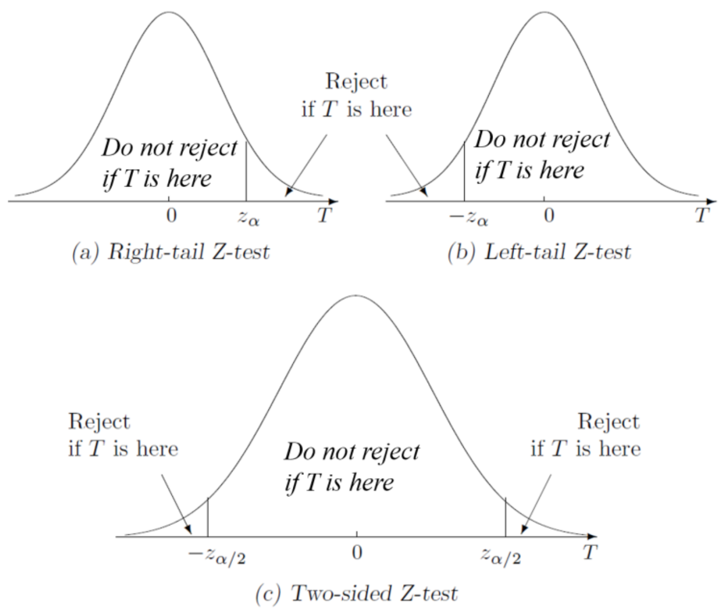
The *critical region* (or *rejection region*) is the set of all values of the test statistic that causes rejection to the null hypothesis.

The *non-rejection region* is the set of all values of the test statistic that decides not to reject the null hypothesis.

The *critical value* is a value that separates the critical region and the non-rejection region.

The *significance level* (denoted by $\alpha$) is the probability that the test statistic will fall in the critical region when the null hypothesis is actually true.

<u>Rejection and non-rejection regions for *Z*-test and test statistic *T*.</u>




***Criteria of The Decision Making***

Reject $H_0$ if ***p*-value** $\leq \alpha$ (the significance level)

Do not reject $H_0$ if ***p*-value** $> \alpha$ (the significance level)

* ***p*-value** is the critical area of the test statistic.

Right-tailed test: *p*-value = area to right of the test statistic *z*, $P(Z > Z_{obs})$

Left-tailed test: *p*-value = area to left of the test statistic *z*, $P(Z < Z_{obs})$

Two-tailed test: *p*-value = twice the area of the extreme region bounded by the test statistic *z*, $P(|Z| > |Z_{obs}|)$

## 5.7. Hypothesis Tests About A Population Mean

***The One Sample Mean Comparison – Hypothesis***

Some possible pairs of $H_0$ and $H_1$:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

TTT LTT RTT

where $\mu_0$ represents the "null value" comes from the research question, not from data.

***The One Sample Mean Comparison – Test Statistic***

Assume the population from which the sample is selected is *normally distributed*.

In hypothesis test about $\mu$ when the population standard deviation $\sigma$ is **known**, the one sample test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}},$$

for both $n \geq 30$ and $n < 30$.

In hypothesis test about $\mu$ when the population standard deviation $\sigma$ is **not known**, the one sample test statistic is

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}},$$

for both $n \geq 30$ and $n < 30$.

***The One Sample Mean Comparison – Critical Region***

| Hypothesis | Test Statistic | Criteria for Rejection |
|---|---|---|
| $H_0 : \mu = \mu_0$<br>$H_1 : \mu \neq \mu_0$ | $z_{obs} = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ | $z_{obs} < -z_{\alpha/2}$ or $z_{obs} > z_{\alpha/2}$<br>$t_{obs} < -t_{\alpha/2,\, n-1}$ or $t_{obs} > t_{\alpha/2,\, n-1}$ |
| $H_0 : \mu \geq \mu_0$<br>$H_1 : \mu < \mu_0$ | | $z_{obs} < -z_{\alpha}$<br>$t_{obs} < -t_{\alpha,\, n-1}$ |
| $H_0 : \mu \leq \mu_0$<br>$H_1 : \mu > \mu_0$ | $t_{obs} = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$ | $z_{obs} > z_{\alpha}$<br>$t_{obs} > t_{\alpha,\, n-1}$ |

Note. Reject $H_0$ if $p-value \leq \alpha$ (the significance level) for any type of hypothesis tests.

*Example 5.14.*

The number of concurrent users for some internet service provider has always averaged 5000 with a standard deviation of 800. After an equipment upgrade, the average number of users at 100 randomly selected moments of time is 5200. Does it indicate, at a 5% level of significance, that the mean number of concurrent users has increased? Assume that the standard deviation of the number of concurrent users has not changed.

Solution:

*Example 5.15.*

A long time authorized user of the account makes 0.2 seconds between keystrokes. One day, the data in *Example 5.5.* are recorded as someone typed the correct username and password. At a 5% level of significance, is this an evidence of an unauthorized attempt?

Solution:

## 5.8. Hypothesis Tests About the Difference Between Two Population Means for Independent Samples

***The Conditions to Make A Test of Hypothesis About $\mu_X - \mu_Y$***

1. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two independent random samples of sizes $n$ and $m$ respectively.

2. At least one of the following two conditions is fulfilled:
   i. Both samples are large.
      i.e. $n \geq 30$ and $m \geq 30$

   ii. If either one or both sample sizes are small, then both populations from which the samples are drawn are normally distributed.
      i.e. $X \sim N\left(\mu_X, \sigma_X^2\right)$ and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$

***The Two Sample Mean Comparison – Hypothesis***

Some possible pairs of $H_0$ and $H_1$:

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases} \qquad \begin{cases} H_0 : \mu_X \geq \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases} \qquad \begin{cases} H_0 : \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$$

TTT LTT RTT

The null hypothesis $H_0$ usually states there is no difference between the parameters of two populations.

***The Two Sample Mean Comparison – Test Statistic***

Assume the population from which the sample is selected is *normally distributed*.

In hypothesis test about $\mu_X - \mu_Y$ when the population standard deviation $\sigma_X$ and $\sigma_Y$ are **known**, the two sample test statistic is

$$Z = \frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}}.$$

In hypothesis test about $\mu_X - \mu_Y$ when the population standard deviation $\sigma_X$ and $\sigma_Y$ are **not known**, the two sample test statistics are

$$t = \begin{cases} \dfrac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}, & if\ \ \sigma_X = \sigma_Y \\ \dfrac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{s_X^2}{n} + \dfrac{s_Y^2}{m}}}, & if\ \ \sigma_X \neq \sigma_Y \end{cases}$$

where

$$s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}.$$

The degrees of freedom for the two samples taken together are

$$df_{\sigma_X = \sigma_Y} = n + m - 2$$

$$df_{\sigma_X \neq \sigma_Y} = \frac{\left(\dfrac{s_X^2}{n} + \dfrac{s_Y^2}{m}\right)^2}{\dfrac{1}{n-1}\left(\dfrac{s_X^2}{n}\right)^2 + \dfrac{1}{m-1}\left(\dfrac{s_Y^2}{m}\right)^2}.$$

***The Two Sample Mean Comparison – Critical Region***

| Hypothesis | Criteria for Rejection |
|---|---|
| $H_0 : \mu_X = \mu_Y$<br>$H_1 : \mu_X \neq \mu_Y$ | $z_{obs} < -z_{\alpha/2}$ or $z_{obs} > z_{\alpha/2}$<br>$t_{obs} < -t_{\alpha/2,\,\nu}$ or $t_{obs} > t_{\alpha/2,\,\nu}$ |
| $H_0 : \mu_X \geq \mu_Y$<br>$H_1 : \mu_X < \mu_Y$ | $z_{obs} < -z_{\alpha}$<br>$t_{obs} < -t_{\alpha,\,\nu}$ |
| $H_0 : \mu_X \leq \mu_Y$<br>$H_1 : \mu_X > \mu_Y$ | $z_{obs} > z_{\alpha}$<br>$t_{obs} > t_{\alpha,\,\nu}$ |

Note. Reject $H_0$ if $p - value \leq \alpha$ (the significance level) for any type of hypothesis tests.

*Example 5.16.*

A survey of low- and middle-income households showed that consumers aged 65 years and older had an average credit card debt of $10,235 and consumers in the 50- to 64-year age group had an average credit card debt of $9342 at the time of the survey. Suppose that these averages were based on random samples of 1200 and 1400 people for the two groups, respectively. Further assume that the population standard deviations for the two groups were $2800 and $2500, respectively. Let $\mu_X$ and $\mu_Y$ be the respective population means for the two groups, people aged 65 years and older and people in the 50- to 64-year age group. Test at the 1% significance level whether the population means for the credit card debts for the two groups are different.

Solution:

*Example 5.17.*

Reconsider *Example 5.7.*. Test at the 5% significance level does a CD writer consume extra energy, and therefore, does it reduce the battery life on a laptop?

Solution:

*Example 5.18.*

Reconsider *Example 5.8.*. Test at the 5% significance level is server A faster?

Solution:

## 5.9. Hypothesis Tests About A Population Proportion: Large Sample

***The One Sample Proportion Comparison – Hypothesis***

Some possible pairs of $H_0$ and $H_1$:

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases} \qquad \begin{cases} H_0 : p \geq p_0 \\ H_1 : p < p_0 \end{cases} \qquad \begin{cases} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{cases}$$

TTT LTT RTT

where $p_0$ represents the "null value" comes from the research question, not from data.

***The One Sample Proportion Comparison – Test Statistic***

In hypothesis test about $p$ for large sample ($np > 5$ & $nq > 5$), the one sample test statistic is

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

where $q = 1 - p$.

***The One Sample Proportion Comparison Test – Critical Region***

| **Hypothesis** | **Test Statistic** | **Criteria for Rejection** |
|---|---|---|
| $H_0 : p = p_0$<br>$H_1 : p \neq p_0$ | | $z_{obs} < -z_{\alpha/2}$ or $z_{obs} > z_{\alpha/2}$ |
| $H_0 : p \geq p_0$<br>$H_1 : p < p_0$ | $z_{obs} = \dfrac{\hat{p} - p}{\sqrt{pq/n}}$ | $z_{obs} < -z_{\alpha}$ |
| $H_0 : p \leq p_0$<br>$H_1 : p > p_0$ | | $z_{obs} > z_{\alpha}$ |

Note. Reject $H_0$ if $p - value \leq \alpha$ (the significance level) for any type of hypothesis tests.

*Example 5.20.*

When working properly, a machine that is used to make chips for calculators does not produce more than 4% defective chips. Whenever the machine produces more than 4% defective chips, it needs an adjustment. To check if the machine is working properly, the quality control department at the company often takes samples of chips and inspects them to determine if they are good or defective. One such random sample of 200 chips taken recently from the production line contained 12 defective chips. Find the *p*-value to test the hypothesis whether or not the machine needs an adjustment. What would your conclusion be if the significance level is 2.5%?

Solution:

## 5.10. Hypothesis Tests About the Difference Between Two Population Proportions for Large and Independent Sample

***The Two Sample Proportion Comparison – Hypothesis***

Some possible pairs of $H_0$ and $H_1$:

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases} \qquad \begin{cases} H_0 : p_1 \geq p_2 \\ H_1 : p_1 < p_2 \end{cases} \qquad \begin{cases} H_0 : p_1 \leq p_2 \\ H_1 : p_1 > p_2 \end{cases}$$

TTT LTT RTT

***The Two Sample Proportion Comparison – Test Statistic***

When a test of hypothesis about $p_1 - p_2$ is performed, usually the null hypothesis is $p_1 = p_2$ and the values of $p_1$ and $p_2$ are not known. The **pooled sample proportion**, denoted by $\overline{p}$ is calculated by

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2},$$

depends on whether the values of $x_1$ and $x_2$ or the values of $\hat{p}_1$ and $\hat{p}_2$ are known.

Hence, the two sample test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\overline{p}\overline{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\overline{q} = 1 - \overline{p}$.

However, in the case that the values of $p_1$ and $p_2$ are known, the two sample test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n} + \frac{p_2 q_2}{m}}}.$$

***The Two Sample Proportion Comparison Test – Critical Region***

| Hypothesis | Test Statistic | Criteria for Rejection |
|---|---|---|
| $H_0 : p_1 = p_2$<br>$H_1 : p_1 \neq p_2$ | $z_{obs} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\overline{p}\overline{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ | $z_{obs} < -z_{\alpha/2}$ or $z_{obs} > z_{\alpha/2}$ |
| $H_0 : p_1 \geq p_2$<br>$H_1 : p_1 < p_2$ | | $z_{obs} < -z_{\alpha}$ |
| $H_0 : p_1 \leq p_2$<br>$H_1 : p_1 > p_2$ | | $z_{obs} > z_{\alpha}$ |

Note. Reject $H_0$ if $p - value \leq \alpha$ (the significance level) for any type of hypothesis tests.

*Example 5.21.*

A quality inspector finds 10 defective parts in a sample of 500 parts received from manufacturer A. Out of 400 parts from manufacturer B, she finds 12 defective ones. A computer-making company uses these parts in their computers and claims that the quality of parts produced by A and B is the same. At the 5% level of significance, do we have enough evidence to disprove this claim?

Solution: