CHAPTER II
THEORITICAL CONSIDERATION

2.1 Filter and Its Application

3M filter is made from Filtrete^{TM} media, which is also a product of 3M. Every 3M filters will be made from the Filtrete^{TM}, but they may be made differently by converting processes. Some may be cut into different shapes and different sizes and others may be also pleated and cut, see figure 2-1.

2.1.1 Application of Filter

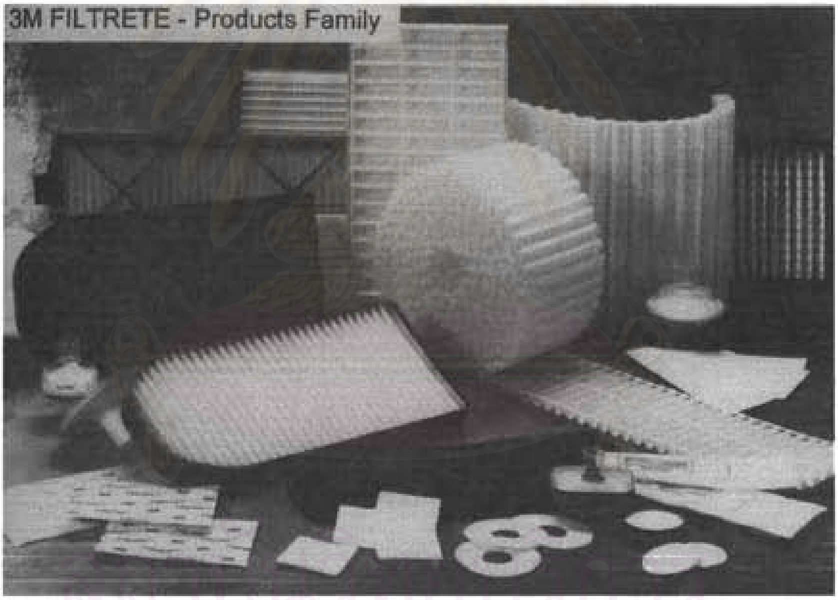

Figure 2-1 : Various types of filters

Filtrete^{TM} media is ideal for applications where particle capture is critical and space is at a premium. It is widely used in many applications, for example:

- In automotive industry : The application of Filtrete^{TM} is as a vehicle ventilation filter. It is for filtering odors before they enter the inside of the automobile.

- Printer : An over-spray ink particles fly free in the printer until a fan pull air out of the printer into the room. The filter removes these ink particles before they contaminate the room air.
- Room air cleaner filter.
- Computer disk drive : A filter in the Hard Disk Drive (HDD) will capture the small particles inside HDD in order to prevent the damage, which caused by those particles.

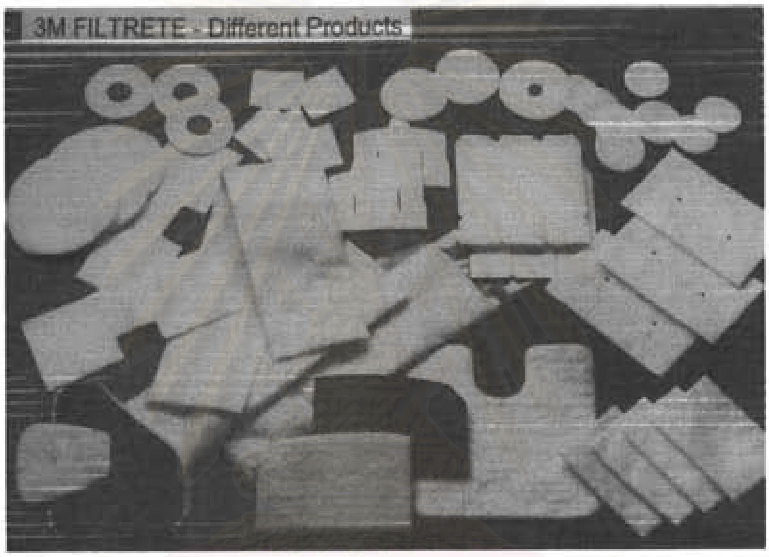

Figure 2-2 : Various shapes of filter

The application of Filtrete^{TM} also includes cabin air filtration systems, room air purifiers, air conditioners, computer disk drives, vacuum cleaner post filters, copy machines, and humidifiers. Filtrete^{TM} can also be used in a variety of medical applications, including anesthetic circuitry / respiratory care, sleep apnea and incubator, see example in figure 2-2 and 2-3.

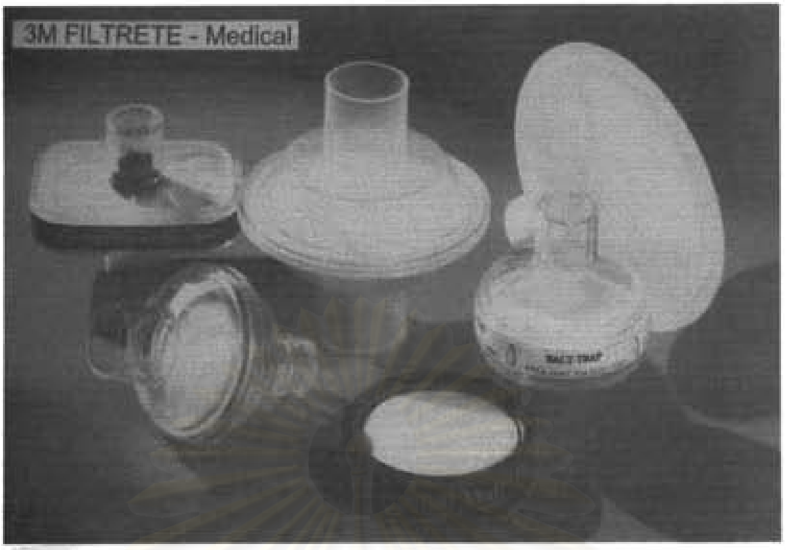

Figure 2-3 : Application of filters in medical field

One important thing for applying the Filtrete^{TM} filter into each application is the quality of the cutting seal. Each application needs different specific quality of the sealing of Filtrete^{TM}'s rim at the die cutting process, such as filter for HDD needs 100 percent edge seal while the medical filter needs 90 percent and miscellaneous filter needs only 80 percent edge seal. Thus, it requires carefully control during the die cutting process.

2.1.2 Application of Filter in HDD

2.1.2.1 Contamination in HDD

As the HDD will be used forever until it is destroyed by something, it will not be opened to fix or replace some component inside, so the lifetime of filter is very important. The contamination in HDD is also a serious issue. In general, the contamination in HDD can be categorized into two types ; particles contamination and chemical contamination. The hazards that caused by those contamination are as follows:

➢ Particles Contamination
  - Particles inside a hard disk drive can cause head crashes
  - Particles can be removed by a recirculation filter

➢ Chemical Contamination

- Corrosion : Acid gases in combination with water moisture can result in corrosion of metal components in HDD
- Stick / Friction : Organic vapor deposition can alter the surface energy of the disk and head, causing stickiness and high friction.

2.1.2.2 Sources of Contamination

When the HDD is completely assembly, it is a closed area. The only thing that can go into or go out is an air, which flow in or out in a regular basis up to the difference of the air pressure between inside and outside HDD when the disk start rotating or when the disk is stopping. Thus, the contamination, which effect the inside HDD, will be introduced by both the airflow and the inside component of the HDD itself.

To reduce the contamination, the filter is used. However, there are types of filter to select, up to the objective.

1. Breather Filter : It is used to prevent and capture the particle that will go into the HDD together with the air flow by passing the breather hole.
2. Desiccant Filter : This type of filter will be used as an adsorbent component in the HDD in order to absorb the chemical vapor and dehumidifying inside the HDD.
3. Recirculation Filter : An application of recirculation filter is aimed for the particle contaminant in HDD.

2.1.2.3 Recirculation Filter in HDD

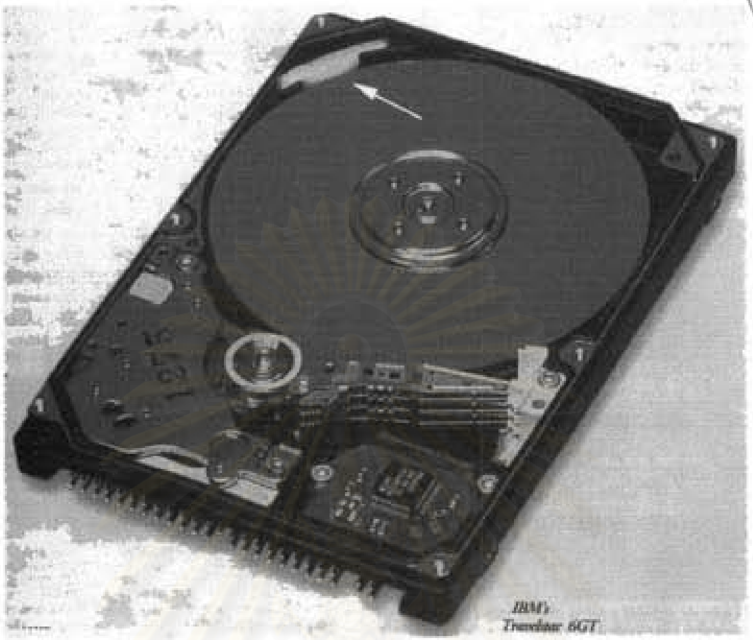

Figure 2-4 : Recirculation filter as HDD's component

Recirculation filters are widely used in disk drives to remove particles and protect the disk drive system, see figure 2-4. Performance of a recirculation filter can be categorized using four critical parameters : filter efficiency, pressure drop, loading capacity, and time to clean up.

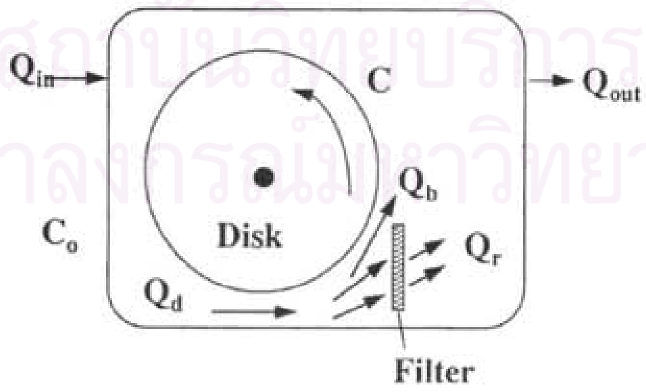

Figure 2-5 : Direction of air flow in HDD

Where ; Q_{in} : Inflow

Q_{out} : Outflow

Q_r : Airflow through the filter

Q_d : Flow generated by the disk

Q_b : Bypass flow

C : Inside particle concentration

C_o : Outside particle concentration

Performance of a recirculation filtration system depends on both filter efficiency and pressure drop. An ideal recirculation filter should have high efficiency and low pressure drop. Figure 2-5 shows direction of air flow through recirculation fiter in the HDD.

The loading capacity of a filter is also very important. A recirculation filter is permanently installed in the disk drive and will not be replaced. It is imperative that performance of a recirculation filter does not degrade with particle loading. Time to clean up is an indicator of the actual performance of a filter. Time to clean up depends not only on filter efficiency, but also pressure drop.

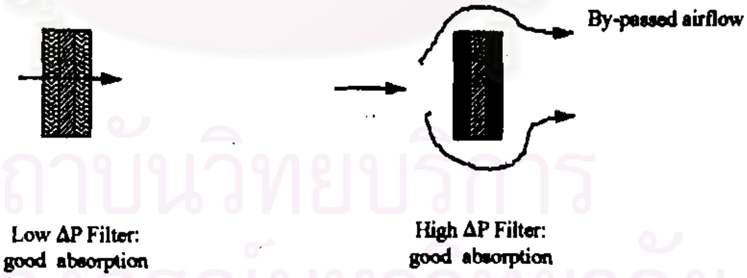

Figure 2-6 : Direction of air flow through filter

By-pass is another problem concerned with performance of every filter. By-passed airflow will be generally occurred from too high-pressure drop of filter and unfit size of filter. If the pressure drop of filter is high the air circulate in HDD will have a tendency to go the another way, which there is no obligation object. And if the filter is not fit to the slot in HDD, it will allow airflow to go by-pass. Thus, the possibility that the particles, which are carried by airflow, will be captured by recirculation filter is decreased. In the other word, by-pass will be result in low performance of filter.

2.1.2.4 HDD Recirculation Filter Manufacturing Process

As mention previously that recirculation filter is a product from cutting Filtrete^{TM} into pieces, 3M Thailand is the source of supply of this product. At the manufacturing site, process of making recirculation filter is describe as below:

1. Receive the Filtrete^{TM} media from source of supply. The Filtrete^{TM} that arrive is a kind of semi product, which comes in rolls. The types of Filtrete^{TM} are two types; GSB-70 and G-100. They are different in terms of the weight of polypropylene fibers per square meter.
2. Take the rolls of Filtrete^{TM} to bake in order to prevent and control level of chemical contamination from the Filtrete^{TM} itself.
3. After baking, the Filtrete^{TM} will be allowed to cool down and ready to be cut into size as required. The size of recirculation filter will be up to the customer's demand.
4. After that, the pieces of filter will be individual visually inspected by using 3x magnifier (three times magnifier). The purpose of this process is to protect and limit the particle contamination.
5. Then, the filters will be packed in double bag and heat seal to prevent other contamination that will be introduced by air after that. The packaging bag is again up to the customer. In addition to this step, the process of product assurance is involved.
6. Then, the bags of filter will be packed in to the corrugate box and load to customer's facility.

All the above process is done in cleanroom environment.

From the process of making filters, it is not a complex process, however, the important thing is to cut filter into the required size.

2.1.3 Cleanroom

Cleanroom is required in filter manufacturing in order to control the contamination level in the filter.

The meaning of cleanroom is defined in Federal Standard 209D as:

"A room in which the concentration of airborne particles is controlled to specified limits."

and in British Standard *5295* as:

"A room with control of particulate contamination, constructed and used in such a way as to minimize the introduction, generation and retention of particles inside the room and in which the temperature, humidity and pressure shall be controlled as is necessary."

2.1.3.1 Classification of Cleanrooms

Cleanrooms are classified by the cleanliness of the air. The method most easily understood and universally applied is the one suggested by Federal Standard 209 in which the number of particles equal to and greater than *0.5* μm is measured in one cubic foot of air and this count is used to classify the room. That is can be simply shown as in table 2-1.

Table 2-1 : A simplified Federal Standard 209 classification of cleanrooms[3]

Federal Standard 209 classification	1	10	100	1000	10000	100000
No. of particles/ft^3 $\geq 0.5 \mu m$	1	10	100	1000	10000	100000

A classification of the room may be carried out when the room is:

- as built, i.e. ready for operation with all services connected and functional but without production equipment or personnel
- with production equipment installed and operating but without personnel
- operational, i.e. in full production.

2.1.3.2 Application of Cleanrooms

An application of cleanroom can be suggested by the cleanroom class as following:

Table 2-2 : Possible cleanroom requirement for various tasks carried out in cleanrooms[3]

Class 1	These rooms are only used by integrated circuit manufacturers developing sub-micron geometry.
Class 10	These rooms are used by semiconductor manufacturers producing very large scale integrated (VLSI) Circuits with line widths below 2 microns.
Class 100	Used when a bacteria-free or particulate-free environment is required in the manufacture of aseptically produced injectable medicines. Required for implant or transplant surgical operations. Integrated circuit manufacturing. Isolation of immunosuppressed patients, e.g. after bone marrow transplant operations.
Class 1000	Manufacture of high quality optical equipment. Assembly and testing of precision gyroscopes. Assembly of miniaturized bearings
Class 10 000	Assembly of precision hydraulic or pneumatic equipment, servo-control valves, precision timing devices, high grade gearing.
Class 100 000	General optical work, assembly of electronic components, hydraulic and pneumatic assembly.

Table 2-3 : Class limits in particles per cubic foot of size equal to or greater than particle size shown[3] (micrometers)

	Measured particle size (μm)				
Class	**0.1**	**0.2**	**0.3**	**0.5**	**5.0**
1	35	7.5	3	1	NA
10	350	75	30	10	NA
100	NA	750	300	100	NA
1000	NA	NA	NA	1000	7
10000	NA	NA	NA	10000	70
100000	NA	NA	NA	100000	700

Regarding the manufacturing facilities of recirculation filters, there are three cleanroom class limits used.

- Cleanroom class 100000 for the die cutting and the secondary packing rooms.
- Cleanroom class 10000 for the inspection and primary packing room.
- Environment of cleanroom class 100 under a laminar flow hood that placed in the inspection and primary packing room.

2.2 Design of Experiment[6], [7]

In order to run an experiment, we need to have a knowledge on the design of experiment. The experiment is about a test or series of tests that made to observe or identify the changes of the output response when the input variables of a system or process have made. The experiments may be performed to discover something, for example, to develop a process that affected minimally by external sources of variability, or to determine the effects of different processes in order to select which one will give the optimum solution.

For conclusion, the objectives of the experiment are consisted of either one of the following:

1. To determine which variables are the most influential on the response y
2. To determine where to set the influential x's so that y is almost always near the desired nominal value.
3. To determine where to set the influential x's so that the variability of y is small
4. To determine where to set the influential x's so that the effects of the uncontrolled variables z_1, z_2, z_3, ... are minimized

In general, the experiment will be done related to a particular process or system. The process of system is represented in figure 2-7 as shown below:

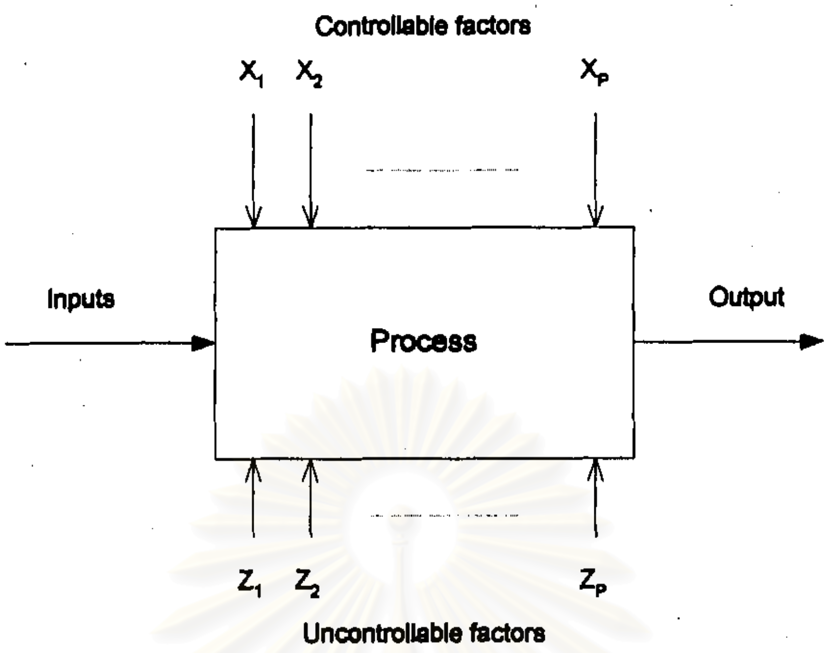

Figure 2-7 : General model of process or system

Usually, machines, methods, people, and other sources will be combined as a process that transforms inputs into an output that has one or more observable responses. Some of the process variables $x_1, x_2, \ldots, x_p$ are controllable, whereas other variables, $z_1, z_2, \ldots, z_p$, are uncontrollable.

2.2.1 The Basic Principles of the Design of Experiment

There are three basic principles to consider when design an experiment.

2.2.1.1 Replication

The replication is to duplicate the basic experiment. The reason of replication can be described as

A. Replication will allow the experimenter to obtain an estimate of the experimental error. When replicate the experiment, the error or estimation will becomes a basic unit of measurement for determining data.

B. If, in the experiment, the ***sample mean*** (e.g. $\bar{y}$) is used to estimate an effect of a factor, the replication will give an experimenter a pore precise of the estimate of an affect. The illustration is shown as below:

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

Or the variance of the sample mean $(\sigma_{\bar{y}}^{2})$ is the ratio of the variance of an individual observation (σ^{2})per number of replication (n). Basically, if the number of replication is increased, the variance of the sample mean will be consequently decreased.

2.2.1.2 Randomization

Randomization means both the allocation of the experiment material and the order in which the individual runs or trial's of the experiment to be performed are randomly determined, that is called "random sampling". In other word, a random sampling is the sampling which the numbers of the population are selected in a way that each has an equal chance of being selected.

Randomization is the cornerstone underlying the use of statistical method in experimental design. By randomization we mean that both the allocation of the experimental material and the order in which the individual runs or trials of the experiment are to be performed are randomly determined. Statistical methods require that the observations (or errors) be independently distributed random variables. Randomization usually makes this assumption valid. By properly randomizing the experiment, we also assist in "averaging out" the effects of extraneous factors that may be present.

2.2.1.3 Blocking

Block refers to the homogeneous portion of the experimental material. Blocking is a technique used to increase the precision of an experiment. A block is a portion of the experimental material that should be more homogeneous than the entire set of material. Blocking involves making comparisons among the conditions of interest in the experiment within each block.

2.2.2 Outline for Designing Experiments

The step of doing experiment can be shown as follow:

1. State the problem or area concerned, the objective of the experiment
2. State the factors, levels, and ranges
3. Select the response variables
4. Select the experimental design
5. Perform the experiment
6. Statistical analysis of the data
7. Conclusion and recommendation

2.2.3 Type of Experimental Design

2.2.3.1 Completely Randomized Design

This is the simplest type of layout in which treatments an allotted to the units entirely by chance. More specifically, if a treatment is to be applied to four units in the experiment material an equal probability if receiving the treatment. In addition the units should be processed in random order is likely to affect the results.

This design has several conveniences.

1. Complete flexibility is allowed. Any number of treatments of replicates may be used. The number of replications can be varied at will from treatment to treatment (through such variation is not recommend without good reason)
2. The statistical analysis is easy even if the numbers of replicates are not the same for all treatments or if the experiment errors differ from treatment to treatment.
3. The method of analysis remains simple when the results from some units or from whole treatments are missing or are rejected. Moreover, the relative loss of information due to missing data is smaller than with any other design.

The principal objection to a completely randomized design is on the grounds of accuracy. Since the randomization is not restricted in any way to ensure that the units, which receive one treatment, are similar to those that receive another treatment, the whole of the variation among the units enters into the experiment error. From this reason the error can often be reduced by the use of a different design, unless the units are highly homogeneous or the experimenter has no information by which to arrange or handle the units in more homogeneous groups.

One fact compensates to some extent for the higher experimental errors as compared with other designs. For a given number of treatments and a given number of experimental units, complete randomization provides the maximum number of degrees of freedom for the estimation of error. This point is worth bearing in mind with small experiment.

The complete randomization may be appropriate:

1. Where the experimental material is homogeneous.
2. Where an appreciable fraction or the units is likely to be destroyed or to fail to respond.
3. In small experiments where the increased accuracy from alternative design does not outweigh the loss of error degrees of freedom.

2.2.3.2 Randomized Block Design

The essence of this design is that the experimental material is divided into groups, each of which constitutes of a single trial or replication. At all stages of the experiment the object is to keep the experimental errors within each group as small as is practicable. Thus, when the units are assigned to the successive groups, all units that go in the same group should be closely comparable. Similarity, during the course of the experiment, a uniform technique should be employed for all units in the same group. Any changes in technique or in other conditions that may affect the results should be made between groups.

2.2.3.3 The Latin Square Design

As previously explain, the randomized complete block design is a design to reduce the residual error in an experiment by removing variability due to a known and controllable nuisance variable. There are several other types of designs that utilize the blocking principle.

The Latin square design is used to eliminate two nuisance sources of variability; that is, it systematically allows blocking in two directions. Thus, the rows and columns actually represent two restrictions on randomization. In general, a Latin square for p factors, or a p x p Latin square, is a square containing p rows.

2.2.3.4 Factorial Designs

The following are some instances where factorial experimentation may be suitable:

- In exploratory work where the object is to determine quickly the effects of each of a number of factors over a specified range.
- In investigations of the interactions among the effects of several factors. From their nature, interactions cannot be studied without testing some of the combinations formed from the different factors. Frequently, information is best obtained by testing all combinations.
- In experiments designed to lead to recommendations that must apply over a wide range of conditions. Subsidiary factors may be brought into an experiment so as to test the principal factors under a variety of conditions similar to those that will be encountered in the population to which recommendations are to apply.

2.3 The Analysis of Variance[5]

After we get the test data, we need to manipulate and analyze them. The analysis of variance or "ANOVA" is a method of testing the hypothesis that two or more population means are equal. This method is to test the equality by analyzing the sample variances. By comparing different types of variance, the conclusions can be formed on whether the population means are equal.

According to the limitation of the t-test which can be used only to compare the means of two population, the analysis of variance method is very useful in reducing numbers of comparing if we need to compare the means from more than two population.

The analysis of variance is a method used for testing the hypotheses about an effect of the treatments (or factors) and to estimate those effects. A definition of treatment (or factor) is a property, or characteristic, that allows us to distinguish the different populations from one another.

2.3.1 Components of a Formal Hypothesis Test

- The **null hypothesis** (denoted by H_0) is a statement about the value of a population parameter (such as the mean μ), and it must contain the condition of equality (that is, it must be written with the symbol =, $\leq$ or $\geq$). For the mean, the null hypothesis will be stated in only one of three possible forms: H_0: μ = some value, H_0: $\mu \leq$ some value, or H_0: $\mu \geq$ some value. We test the null hypothesis *directly* in the sense that the conclusion will be either a rejection of H_0 or a failure to reject H_0.

- The **alternative hypothesis** (denoted *by* H_1) is the statement that must be true if the null hypothesis is false. For the mean, the alternative hypothesis will be stated in only one of three possible forms: H_1: $\mu \neq$ some value, H_1: $\mu <$ some value, or H_1: $\mu >$ some value. Note that H_1 is the opposite of H_0. For example, if H_0: μ = 98.6, then it follows that the alternative hypothesis is given by H_1: $\mu \neq 98.6$.

- **Type I error**: The mistake of rejecting the null hypothesis when it is true. For the preceding informal example, a type I error is the mistake of rejecting the null hypothesis that the mean is 98.6 ($\mu = 98.6$) when the mean is really 98.6. The type I error is not a miscalculation or procedural misstep; it is an actual error that can occur when a rare event happens by chance. The probability of rejecting the null hypothesis when it is true is called the **significance level**; that is, the significance level is the probability of a type I error. The symbol **α (alpha)** is used to represent the significance level. The values of $\alpha = 0.05$ and $\alpha = 0.01$ are commonly used.

- **Type II error**: The mistake of failing to reject the null hypothesis when it is false. For the preceding informal example, a type II error is the mistake of failing to reject the null hypothesis ($\mu = 98.6$) when it is actually false (that is, the mean is not 98.6). The symbol **β (beta)** is used to represent the probability of a type II error.

The ANOVA method is applied the hypothesis testing by referring to the *F* distribution. The properties of *F* distribution are as follows:

1. The *F* distribution is not symmetric; it is skewed to the right.
2. The value of *F* can be 0 or positive, but they can not be negative.
3. There is a different *F* distribution for each pair of degrees of freedom for the numerator and denominator.

When applying the hypothesis testing, there are assumptions that have to recognize as follows:

1. The populations have normal distributions.
2. The populations have the same variance or standard deviation.
3. The samples are random and independent of each other.

A hypothesis to test for the population means of any *k* treatments will be:

Null hypothesis ; H_0 : $\mu_1 = \mu_2 = \mu_i = \ldots. = \mu_k$

Alternative hypothesis ; H_1 : At least one mean is different

The following terms are associated with key components in the hypothesis testing procedure.

1. Numerator degree of freedom ; $\boldsymbol{\nu_1}$ or $\boldsymbol{df_1}$
2. Denominator degree of freedom ; $\boldsymbol{\nu_2}$ or $\boldsymbol{df_2}$
3. Significance level ; $\boldsymbol{\alpha}$
4. Test statistics ; $\boldsymbol{F_o}$

   Test statistic : A sample statistic or a value based on the sample data. A test statistic is used in making the decision about the rejection of the null hypothesis.
5. Critical region

   Critical region: The set of all values of the test statistic that would cause us to reject the null hypothesis. The critical region is represented by the shaded part of Figure 2-8.
6. Critical value : $\boldsymbol{F_{(\alpha, df1, df2)}}$

   Critical value: The value or values that separate the critical region from the value of the test statistic that would not lead to rejection of the null hypothesis. The critical values depend on the nature of the null hypothesis, the relevant sampling distribution, and the level of significance α.

The critical value of *F* can be found in the Percentage Points of the *F* Distribution table, see appendix A, by indicating the significance level, the numerator degree of freedom and the denominator degree of freedom. Some times, the critical value can not be found in the table due to the numbers of degrees of freedom are not included in the table. However, we can use the linear interpolation to approximate the missing value, but in the most cases that is not necessary because the *F* test static is either less than the lowest possible critical value or greater than the largest possible critical value

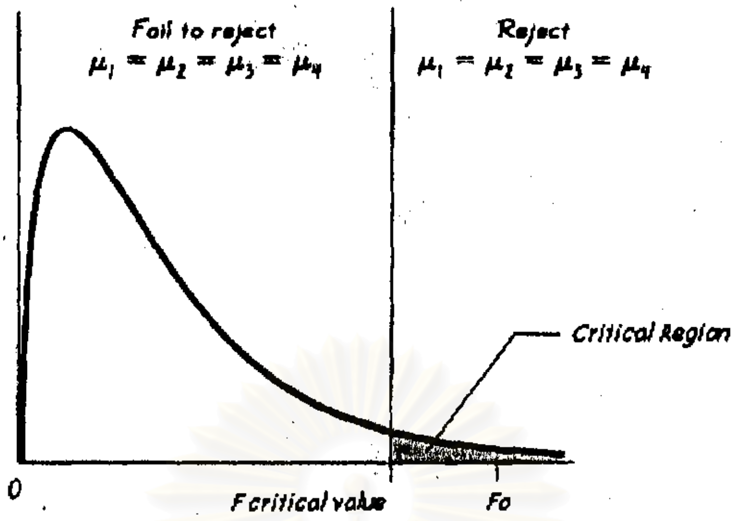
Fail to reject
$\mu_1 = \mu_2 = \mu_3 = \mu_4$
Reject
$\mu_1 = \mu_2 = \mu_3 = \mu_4$
Critical Region
0
F critical value
Fo

Figure 2-8 : *F* - distribution

The critical value is indicating the area of critical region as shown in Figure 2-8. If the test statistic is greater than critical value or fall in the critical region, we will reject the null hypothesis, otherwise we will fail to reject the null hypothesis.

To complete the analysis, the F_0 is need. It can be calculated or the experimenter can use the statistical software to analyze the result. The result from the software will include the *F* test statistic and P-value. If the P-value is less than the significance level, reject the null hypothesis; otherwise, fail to reject the null hypothesis. By using the same approach for *F* critical value, we will reject the null hypothesis if the test statistic larger than critical value.

In general, the analysis of variance can be categorized into two groups ; One-Way ANOVA and Two-Way ANOVA. The term *one-way* is used because the sample data re separated into groups according to one characteristic or one factor because only one factor is investigated at a time. This method is sometimes called a " Single-Factor Analysis of Variance". For example, the data of the movie length is separated into four different groups according to the one characteristic of star ratings.

Another analysis of variance is the *two-way* analysis of variance method, which allows the experimenter to compare populations, which is separated into categories by using two characteristics (or factors). For example, we might separate the lengths of movies according to their star rating and their viewer discretion ratings.

2.3.2 One-Way ANOVA with Equal Sample Sizes

For One-Way ANOVA with equal sample sizes, we consider the tests of hypotheses that three or more population means are all equal, as in $H_0 : \mu_1 = \mu_2 = ... = \mu_k$.

As stated previously, we assume that the populations have normal distributions, the populations have the same variance (or standard deviation), and the samples are random and independent of each other. However, there are some additional assumptions to consider when apply this method as following:

1. The different samples are all the same size.
2. The different samples are from populations that are categorized in only one way.

The following is a notation for One-Way ANOVA with Equal Sample Sizes

n = size of each sample

k = number of samples

$s_{\bar{x}}^2$= variance of the sample means

s_p^2 = pooled variance obtained by calculating the mean of the sample variances

The test statistic is the ratio of two different estimates of the variance σ^2 that is assumed to be common to the populations involved – the variance between samples and the variance within samples, terms that we now define.

1. The variance between samples (also called variation due to treatment) is a measure of the variability caused by differences among the sample means that correspond to the different treatments, or categories of classification. With all samples of the same size *n*,

$$\text{variance between samples} = ns_{\bar{x}}^2$$

where $s_{\bar{x}}^2$ = variance of sample means

*2. **The variance within samples*** (also called variation due to error) is an estimate of σ^2 based on the sample variance. With all samples of the same size n,

$$\text{variance within samples} = s_p^2$$

where s_p^2 = pooled variance obtained by finding the mean of the sample variance

Then the test Statistic for One-Way ANOVA with Equal Sample Sizes is

$$F = \frac{\text{variance between samples}}{\text{variance within samples}} = \frac{ns_{\bar{x}}^2}{s_p^2}$$

Interpreting the Test Statistic F

As the preceding test statistic indicates, we use both estimates of σ^2 to find the value of the F test statistic. The numerator measures variation between sample means. The estimate of variance in the denominator (s_p^2) depends only on the sample variances and is not affected by differences among the sample means.

Consequently if the sample means of each factors are closed in value, it will be result in an F test statistic that is close to 1, and we conclude that there are no significant differences among the sample means. But if the value of F is excessively large, which is clearly difference between an F test statistic that is in the critical region, then we reject the claim of equal means

The critical value of F that separates excessive values from acceptable values is found in appendix A, where α is the level of significance and the numbers off degrees of freedom are as follows (assuming that there are k sets of separate sample with n sample in each set). For cases with equal sample sizes:

$$\text{numerator degrees of freedom} = k - 1$$

$$\text{denominator degrees of freedom} = k(n - 1)$$

Because the individual components of n, $s_{\bar{x}}^2$, and s_p^2 are all positive, F is always positive, so the ANOVA tests are right-tailed. With the same traditional method for testing hypotheses, a value of the F test statistic indicates a significant difference among the sample means when the test statistic exceeds the critical F value obtained from appendix A.

Conclusion

The key important of the hypothesis test for equality of means in three or more populations is we should keep in mind that the populations should have nearly normal distributions with approximately the same variance, and the samples must be independent of each other.

2.3.3 One-Way ANOVA with Unequal Sample Sizes

For the sample with unequal size, the One-Way ANOVA method can also be used to analyze claims of equal means. To proceed the test of equal means, such as $H_0 : \mu_1 = \mu_2 = \mu_3$, the analysis methodology is as same as for the equal sample size. The ratio significance of variance between samples to variance within samples will be considered as the test statistic as:

$$F = \frac{\text{variance between samples}}{\text{variance within samples}}$$

However, we have to "weight" each of the two estimates of variance or compensate for the different sample sizes, then the test statistics will be:

$$F = \frac{\text{variance between samples}}{\text{variance within samples}} = \frac{\left[\dfrac{\sum n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}\right]}{\left[\dfrac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)}\right]}$$

where;

$\bar{\bar{x}}$ = overall mean (sum of all sample scores divided by the total number of scores)

k = number of population means being compared

n_i = number of values in the *i*th sample

N = total number of values in all samples combined ($N=n_1+n_2+..+n_k$)

$\bar{x}_i$ = mean of values in the *i*th sample

s_i^2 = variance of values in the *i*th sample

Table 2-4 : The analysis of variance table for One-Way ANOVA (single-factor)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F₀
Between treatments	$SS_{\text{Treatments}}$	$k - 1$	$MS_{\text{Treatments}}$	$F_0 = \dfrac{MS_{\text{Treatments}}}{MS_E}$
Error (within treatments)	SS_E	$N - k$	MS_E	
Total	SS_T	$N - 1$		

Key components in the ANOVA table are identified below:

a) SS_T, SS(total) or total sum of squares, is a measure of the total variation (around $\bar{\bar{x}}$) in all of the sample data combined.

$$SS(Total) = \sum (x - \bar{\bar{x}})^2$$

SS_T is a combination of $SS_{\text{Treatment}}$ and SS_E , described as follow:

$$SS(total) = SS(treatment) + SS(error)$$

b) $SS_{Treatment}$ or SS(treatment) is a measure of the variation between the sample means. SS(treatment) is sometimes referred to as SS(factor), SS (between groups) or SS(between samples).

$$SS(treatement) = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \ldots + n_k(\bar{x}_k - \bar{\bar{x}})^2 = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2$$

If the population means $(\mu_1, \mu_2, ..., \mu_k)$ are equal, then sample means $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ will all tend to be close together and also close to $\bar{\bar{x}}$. The result will be a relatively small value of SS(treatment). If the population means are not all equal, then at least one of $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ tends to be far apart from the others and also far apart from $\bar{\bar{x}}$, the result will be a relatively large value of SS(treatment).

c) SS$_E$ or SS(error) is a sum of squares representing the variability that is assumed to be common to all the populations being considered.

$$SS(error) = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + ... + (n_k - 1)s_k^2 = \sum (n_i - 1)s_i^2$$

SS(error) is the numerator of the expression for the pooled variance s_p^2. Because SS(error) is a measure of the variance within groups, it is sometimes denoted as SS(within groups) or SS (within samples).

SS(treatment) and SS(error) are both sums of squares, and if we divide each by its corresponding number of degrees of freedom, we get the mean squares.

d) MS$_{Treatment}$ or MS(treatment) is a mean square for treatment, obtained as follows:

$$MS(treatment) = \frac{SS(treatment)}{k - 1}$$

e) MS$_E$ or MS(error) is a mean square for error, obtained as follows:

$$MS(error) = \frac{SS(error)}{N - k}$$

f) MS$_T$ or MS(total) is a mean square for the total variation, obtained as follows:

$$MS(total) = \frac{SS(total)}{N - 1}$$

g) Test Statistic for ANOVA with Unequal Sample Sizes

In testing the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ against the alternative hypothesis that these means are not all equal, the test statistic

$$F = \frac{MS(treatment)}{MS(error)}$$

has an *F* distribution (when the null hypothesis H_0 is true) with degrees of freedom given by

Numerator degrees of freedom = k – 1

Denominator degrees of freedom = N - k

Interpreting the F Test Statistic

The interpretation of the *F* test statistic given above, the denominator depends only on the sample variances that measure variation within the treatments and is not affected by the difference among the sample means. If the numerator does depend on differences among the sample means.

If the differences among the sample means are extreme, they will cause the numerator to be excessively large, so *F* will also be excessively large. Consequently very large values of *F* suggest unequal means, and the ANOVA test is therefore right-tailed.

Conclusion

The method of One-Way ANOVA is used to test the claim that several samples come from populations with the same mean. These methods require normally distributed populations with the same variance, and the samples must be independent. We reject or fail to reject the null hypothesis of equal means by analyzing these two estimates of variance the variance *between* samples and the variance *within* samples.

MS(treatment) is an estimate of the variation between samples, and MS(error) is an estimate of the variation within samples. If MS(treatment) is significantly greater than MS(error), we reject the claim of equal means; otherwise , we fail to reject that claim.

2.3.4 Two-Way ANOVA

As one-way analysis of variance or single-factor analysis of variance refers to the data that are categorized into groups according to a single factor (or treatment). A Two-Way ANOVA will refer to the data that are categorized into groups according to two factors or two treatments.

Table 2-5 : Data for Two-Way ANOVA

Data	Categorized by factor 1		
Categorized by factor 2	Factor 1 -A	Factor 1 -B	Factor 1 -C
Factor 2 -A	X_{AA}	X_{BA}	X_{CA}
Factor 2 -B	X_{AB}	X_{BB}	X_{CB}
Factor 2 -C	X_{AC}	X_{BC}	X_{CC}
Factor 2 -D	X_{AD}	X_{BC}	X_{CD}

Because the one-way analysis of variance is for testing the effect of the single factor, it might seem reasonable to simply proceed with another one-way ANOVA for each factor. Unfortunately, that approach wastes information and totally ignores any effect from an interaction between the two factors.

The definition of ***interaction*** is an effect between two factors if the effect of one of the factors changes for different categories of the other factor.

In using Two-Way ANOVA, we will consider the effect of an interaction between the two factors. The SS(treatment) is used as a measure of the variation due to the different treatment categories and SS(error) is used as a measure of the variation due to sampling error. Also, we use $df_{(1)}$ and $df_{(2)}$ for the two different degrees of freedom.

In executing a two-way analysis of variance, we consider three effects:

1. The effect due to the interaction between the two factors
2. The effect due to the row factor (factor 2)
3. The effect due to the column factor (factor 1)

The following comments summarize the basic procedure for two-way analysis of variance. The procedure described below is actually quite similar to the procedures presented for one-way ANOVA. We form conclusions about equal means by analyzing two estimates of variance, and the test statistic F is the ratio of the two estimates. A significantly large value for F indicates that there is a statistically significant difference in means. The following is summarized procedure for two-way ANOVA.

Procedure for Two-Way ANOVA

Step 1 : In two-way analysis of variance, begin by testing the null hypothesis that there is no interaction between the two factors. By calculating or using statistical software, we get the following test statistic:

$$F = \frac{MS(interaction)}{MS(error)}$$

and get a critical value of F by using significance level and the two degrees of freedom from appendix A. If the test statistic does not exceed the critical value, we fall to reject the null hypothesis of no interaction between the two factors. On the other word, there is not sufficient evidence to conclude that the response data is affected by an interaction between factor 1 and factor 2.

Step 2 : If we fail to reject the null hypothesis of no interaction between factors, then we should proceed to test the following two hypotheses :

H_0 : There are no effects from the row factors (that is , the row means are equal).

H_0 : There are no effects from the column factor (that is, the column means are equal).

If we do reject the null hypothesis of no interaction between factors, then we should stop now ; we should not proceed with the two additional tests. (If there is an interaction between factors, we shouldn't consider the affects of either factor without considering those of the other.)

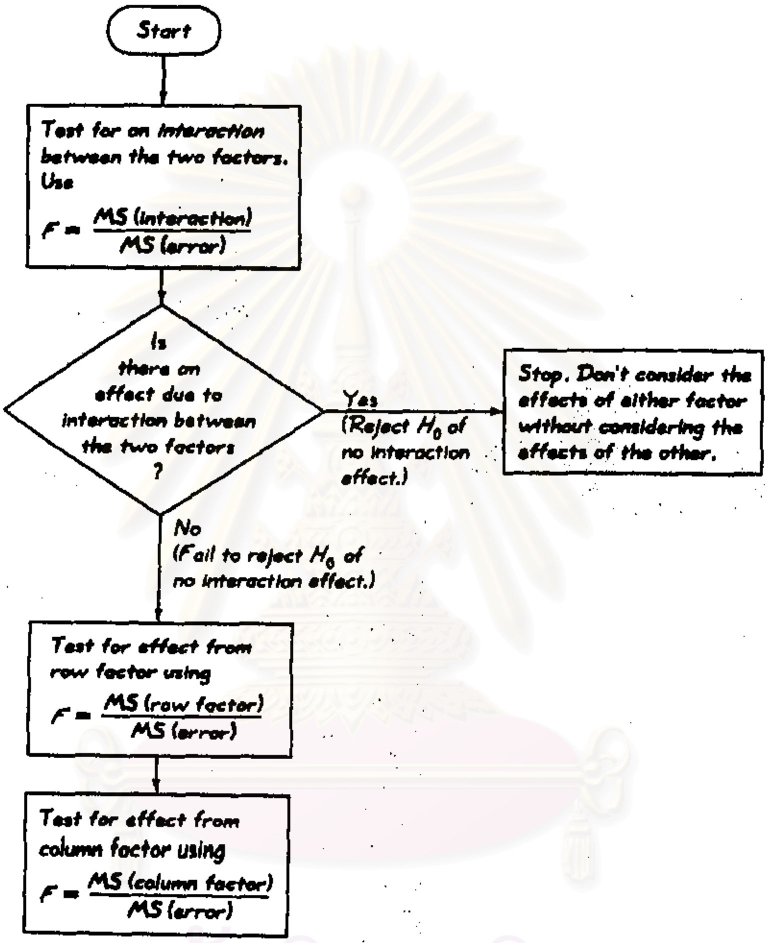
Start
Test for an interaction between the two factors. Use
$F = \frac{MS\ (interaction)}{MS\ (error)}$
Is there an effect due to interaction between the two factors ?
Yes (Reject H_0 of no interaction effect.)
Stop. Don't consider the effects of either factor without considering the effects of the other.
No (Fail to reject H_0 of no interaction effect.)
Test for effect from row factor using
$F = \frac{MS\ (row\ factor)}{MS\ (error)}$
Test for effect from column factor using
$F = \frac{MS\ (column\ factor)}{MS\ (error)}$

Figure 2-9 : Procedure for Two-Way ANOVA[5]

2.3.5 Summarize Used of the ANOVA Method

Table 2-6 : Summary of ANOVA method

Application	Distribution	Test Statistic	Degrees of Freedom	Critical value
One-way ANOVA with equal sample sizes	F	$F = \dfrac{ns_{\bar{x}}^2}{s_p^2}$	numerator : $k-1$ denominator : $k(n-1)$	Appendix : A
One- way ANOVA for all cases	F	$F = \dfrac{MS(treatment)}{MS(error)}$	numerator : $k-1$ denominator : $N-k$	Appendix : A
Two-way ANOVA	F	(1) Interaction: $F = \dfrac{MS(interaction)}{MS(error)}$ (2) Row factor: $F = \dfrac{MS(row-factor)}{MS(error)}$ (3) Column factor: $F = \dfrac{MS(column-factor)}{MS(error)}$	See computer display	Appendix : A

2.4 Multiple Linear Regression[5]

The linear regression model is a mathematical model used to fit the linear relationship between a single dependent variable and other independent variables. The dependent variable and independent variables can be called response and regressors respectively.

In general, the dependent variable (y) may be related to *k* independent variables as a model, which is called a multiple linear regression model as below:

$$y_i = \beta_0 + \beta_1\chi_{1i} + \beta_2\chi_{2i} + \ldots\ldots\ldots\ldots + \beta_k\chi_{ki} + \epsilon_i$$

The parameters, β_j ; j = 0, 1, 2,……., k are called "***the regression coefficients***"

The regression model describes a relationship in the k-dimensional space of the dependent variables. The parameter β_j represents the expected change in dependent variable (y) per unit of changing variable x when all the remaining independent variables are constant.

2.4.1 Test for Significance of Regression

The test hypothesis are:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots\ldots = \beta_k = 0$

$H_1 : \beta_j \neq 0$ for at least one j

The rejection of H_0 implies that at least one of the independent variables significantly.

Assumptions for regression equation.

1. We are investigating only linear relationships.
2. For each x value, y is a random variable having a normal distribution. All of these y distributions have the same variance. Also, for a given value of x, the distribution of y values has a mean that lies on the regression line.

The test procedure for $H_0 \ : \ \beta_1 = \beta_2 = \beta_3 = \ldots\ldots. = \beta_k = 0$ is to compute $\boldsymbol{F_0}$, as follows:

$$F_0 = \frac{SS_R/k}{SS_E/(n-1-k)} = \frac{MS_R}{MS_E}$$

And reject H_0 if $\boldsymbol{F_0}$ exceeds $\boldsymbol{F_{\alpha,\,k,\,n\text{-}k\text{-}1}}$

when α = type one error

k = number of independent variables

n = sample size

Another approach for testing H_0 is to use P-value, which reject H_0 if the P-value for the statistic $\boldsymbol{F_0}$ is less than α.

Because the test procedure involves an analysis of variance, the result of testing will result in an Analysis of Variance for significance of Regression in Multiple Regression as below:

Table 2-7 : Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F_0
Regression	k	SS_R	MS_R	MS_R/MS_E
Error of residual	n-k-1	SS_E	MS_E	
Total	n-1	SS_T		

From the analysis of variance, we can also calculate the coefficient of determination (R^2), where

$$R^2 = \frac{SS_R}{SS_T}$$

The **coefficient of determination** is a measure of the amount of the variability of y, which obtains by using independent variables $x_1, x_2, \ldots, x_k$ in a regression model. Or

$$R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

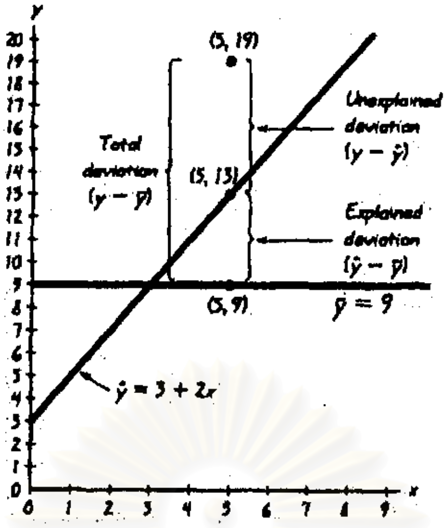

Figure 2-10 : Unexplained, Explained, and Total Deviation[5]

Assume that we have a collection of paired data containing the particular point *(x, y),* that $\hat{y}$ is the predicted value of y (obtained by using the regression equation), and that the mean of the sample *y* value is $\bar{y}$. See figure 2-10.

The **total deviation** (from the mean) of the particular point *(x, y)* is the vertical distance y-$\bar{y}$, which is the distance between the point *(x, y)* and the horizontal line passing through the sample mean $\bar{y}$.

The **explained deviation** is the vertical distances $\hat{y}-\bar{y}$, which is the distance the predicted y value and the horizontal line passing through the sample mean $\bar{y}$.

The **unexplained deviation** is the vertical distance y-$\hat{y}$, which is the vertical distance between the point *(x, y)* and the regression line.

For example, if $R^2 = 0.81$, it means that 81% of the total variation in y can be explained by the regression line. Another 19% of the total variation in y remain unexplainable. Thus, the multiple coefficient of determination (R^2) is a measure of how well the regression equation fits the sample data.

It seems that a larger value of R^2 is a better. However, a larger value of R^2 does not necessary mean that the regression model is a better one. In fact, adding more variables to the model will usually increase R^2 weather the additional variable is statistically significant or not, and the best multiple regression equation does not necessary to use all of the available variables.

Thus, an alternative is to consider the adjusted coefficient of determination (adjust R^2), which the determination of adjusted R^2 is based on the number of variables and the sample size, so it is more accurate.

$$Adjusted\ R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

Normally, the adjusted R^2 will not increased as adding variables to the model, and it also decreased if the unimportant variable are added.

When comparing the multiple regression equations, it is better to use the adjusted R^2.

If R^2 and adjusted-R^2 are significantly different, there could be some not necessary variables have been included to the model.

2.5 Confidence Interval and Prediction Interval[8]

Confidence Interval and Prediction Interval are both types of interval estimates, they differ from one another interval only in the standard deviation. In general, the interval estimates will be presented as following form:

$$\text{(estimate)} \pm t \cdot \text{(standard deviation)}$$

2.5.1 Confidence Interval

Confidence interval is an interval estimate that the point estimate of mean for population is contained in that interval with 1-α level confidence.

For the multiple linear regression, at any "x_0" , the 1-α level confidence interval for mean of y ($\hat{y}|_{x0}$) is defined by the endpoints

$$\hat{y}\Big|x_0 \pm t_{\alpha/2,n-k-1} \cdot s.e.(\hat{y}\Big|x_0)$$

where, the linear model $y_i = \beta_0 + \beta_1\chi_{1i} + \beta_2\chi_{2i} + \ldots\ldots\ldots + \beta_k\chi_{ki} + \epsilon_i$ for $1 \leq i \leq n$, can be written in matrix form as $\mathbf{Y} = \mathbf{X}\beta + \epsilon$

when,

Y is the $n \times 1$ vector, as ⇒ $Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$

X is the $n \times (k+1)$ matrix, as ⇒ $X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ . & . & \cdots & . & . \\ . & . & \cdots & . & . \\ . & . & \cdots & . & . \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$

the parameter β is the $(k+1)\times 1$ vector, as ⇒ $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}$

and ϵ is the $n\times 1$ vector containing the error terms, as ⇒ $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$

which,

$$s.e.(\hat{y}\Big|x_0) = \hat{\sigma}\sqrt{x_0{}'(X'X)^{-1}x_0}$$

and xx is the $(k+1)\times(k+1)$ matrix, as

$$XX = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \cdots & \sum_{i=1}^{n} x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{1i}x_{ki} & \sum_{i=1}^{n} x_{2i}x_{ki} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}$$

2.5.2 Prediction Interval

Prediction interval is an interval estimate that the "y_0" at any "x_0" is contained in the interval that predicted by referring to the regression line with $1-\alpha$ confidence level.

The prediction interval for a future observation obtained at any "x_0", with the $1-\alpha$ confidence level, takes into consideration the extra variability due to an error term ϵ and is given by the endpoints

$$\hat{y}\big|x_0 \pm t_{\alpha/2,\, n-k-1} \cdot s.e.(\hat{y}\big|x_0 + \epsilon)$$

where

$$s.e.(\hat{y}\big|x_0 + \epsilon) = \hat{\sigma}\sqrt{1 + x_0'(XX)^{-1}x_0}$$

Minitab can also find both confidence and prediction intervals at a certain point. Under command ***Stat / Regression / Regression***, Minitab asks the user to specify "*Response*" and "*Predictors*". Then, choose subcommand ***Options,*** and enter specific x_0 (all predictors, by spacing one another) in the *"Prediction Intervals for New Observation"* block.