Stat 322/332/362
Sampling and Experimental Design

Fall 2006 Lecture Notes

Authors: Changbao Wu, Jiahua Chen

Department of Statistics and Actuarial Science
University of Waterloo

Key Words: Analysis of variance; Blocking; Factorial designs; Observational and experimental studies; Optimal allocation; Ratio estimation; Regression estimation; Probability sampling designs; Randomization; Stratified sample mean.

Contents

Chapter 1

Basic Concepts and Notation

This is an introductory course for two important areas in statistics: (1) survey sampling; and (2) design and analysis of experiments. More advanced topics will be covered in Stat-454: Sampling Theory and Practice and Stat-430: Experimental Design.

1.1 Population

Statisticians are preoccupied with tasks of modeling random phenomena in the real world. The randomness as most of us understood, generally points to the impossible task of accurately predicting the exact outcome of a quantity of interest in observational or experimental studies. For example, we did not know exactly how many students will take this course before the course change deadline is passed. Yet, there are some mathematical ways to quantify the randomness. If we get the data on how many students completed Stat231 successfully in the past three terms, some binomial model can be very useful for the purpose of prediction. Stat322/332/362 is another course in statistics to develop statistic tool in modeling, predicting random phenomena.

A random quantity can be conceptually regarded as a sample taken from some population through some indeterministic mechanism. Through the observation of these random quantities (**sample data**), and some of the prior information about the population, we hope to draw conclusions about the unknown population. The general term "population" refers to a collection of "individuals", associated with each "individual" are certain characteristics of interests. Two distinct types of populations are studied in this course.

A survey or finite population is a **finite set** of **labeled** individuals. This

set can hence be denoted as

$$U = \{1, 2, 3, \cdots, N\},$$

where N is called the population size. Some examples of survey population:

1. Population of Canada, i.e. all individuals residing in Canada.
2. Population of university students in Ontario.
3. Population of all farms in the United States.
4. Population of business enterprises in the Great Toronto area.

The survey population in applications may change over time and/or location. It is obvious that Canada population is in constant change with time for reasons such as birth/death/immigration. Some large scale ongoing surveys must take this change into consideration. In this course we treat the survey population as fixed. That is, we need to make believe that we only a snapshot of a finite population so that any changes in the period of our study is not a big concern. In sample survey, our main object is to learn about some characteristics of the finite population under investigation.

In experimental design, we study an input-output process and are interested in learning how the output variable(s) is affected by the input variable(s). For instance, an agricultural engineer examines the effect of different types of fertilizers on the yield of tomatoes. In this case, our random quantity is the yield. When we regard the outcome of this random quantity as a sample from a population, this population must contain **infinite** individuals. Hence, the population in experimental design is often regard as infinite.

The difference between the finite/infinite population is not always easy to understand/explain. In the tomato example, suppose we only record whether the yield per plant exceeds 10kg or not. The random quantity of interest takes only 2 possible values: Yes/No. Does it imply that the corresponding population is finite? The answer is no. We note the conceptual population is not as simple as consisting of two individuals with characteristics { Yes, No}. The experiment is not about selecting one of this two individuals, but the complex outcome is mapped to one of these two values.

Let us make it conceptually a bit harder. Assume an engineer wants to investigation whether the temperature of the coin can alter the probability of its landing on a head. The number of possible outcome of this experiment is two: {Head, Tail}. Is it a finite population? The answer is again negative.

The experiment is not about how to select one of two individuals from a population consisting of {Head, Tail}. We must imagine a population with infinite number of heads and tails each representing an experimental configuration under which the outcome will be observed. Thus, an "individual" in this case is understood as an "individual experiment configuration" which is practically infinite.

In summary, the population under experimental design is an **infinite** set of all possible experiment configurations.

1.2 Parameters of interest

The interested characteristic(s) of a sample from a population is referred as study variable(s) or response variable(s), y. For a survey population, we denote the value of the response variable as y_i for the ith individual, $i = 1, 2, \cdots, N$. The following population quantities are primary interest in sample survey applications:

1. Population total: $Y = \sum_{i=1}^{N} y_i$.
2. Population mean: $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$.
3. Population variance: $S^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2$.
4. Population proportion: $P = M/N$, where M is the number of individuals in the population that possess certain attribute of interest.

In many applications, the study variables are indicator variables or categorical variables, representing different groups or classes in the population. When this is the case, it is seen that the population proportion is a special case of population mean defined over an indicator variable. Let

$$y_i = \begin{cases} 1 & \text{if the } i\text{th individual possesses "A"} \\ 0 & \text{otherwise} \end{cases}$$

where "A" represents the attribute of interest, then it is easy to see that

$$P = \bar{Y}\,, \quad S^2 = \frac{N}{N-1} P(1-P)\,.$$

In other words, it is quite feasible for us to ignore the problem of estimating population proportions. When the problem about proportions arises, we may simply use the same techniques developed for population mean.

In experimental design, since the population is (at least hypothetically) infinite, we are often interested in finding out the probability distributions of the study variable(s) and/or the related parameters. In the tomato-fertilizer example, the engineer wishes to examine if there are differences among the average yields of tomatoes, μ_1, μ_2, μ_3 and μ_4, under four different types of fertilizers. The μ_i's are the parameters of interest. These parameters are in a rather abstract kingdom.

1.3 Sample data

A subset of the population with study variable(s) measured on each selected individuals is called a sample, denoted by s: $s = \{1, 2, \cdots, n\}$ and n is called the sample size. $\{y_i, i \in s\}$ is also called sample or sample data. Data can be collected through direct reading, counting or simple measurement, referred to as **observational**, or through carefully designed experiments, referred to as **experimental**. Most sample data in survey sampling are observational while in experimental design they are experimental. The most useful summary statistics from sample data are sample mean $\bar{y} = n^{-1} \sum_{i \in s} y_i$ and sample variance $s^2 = (n-1)^{-1} \sum_{i \in s} (y_i - \bar{y})^2$. As a remark, in statistics, we call any function of data not depending on unknown parameters as a **statistic**.

1.4 Survey design and experimental design

One of the objectives in survey sampling is to estimate the finite population quantities based on sample data. In theory, all population quantities such as mean or total can be determined exactly through a complete enumeration of the finite population, i.e. a census. Why do we need sample survey?

There are three main justifications for using sampling:

1. Sampling can provide reliable information at far less cost. With a fixed budget, performing a census is often impracticable.

2. Data can be collected more quickly, so results can be published in a timely fashion. Knowing the exact unemployment rate for the year 2005 is not very helpful if it takes two years to complete the census.

3. Estimates based on sample surveys are often more accurate than the results based on a census. This is a little surprising. A census often

requires a large administrative organization and involves many persons in the data collection. Biased measurement, wrong recording, and other types of errors can be easily injected into the census. In a sample, high quality data can be obtained through well trained personnel and following up studies on nonrespondents.

Survey design is the planning for both data collection and statistical analysis. Some crucial steps involve careful definitions for the following items.

1. **Target population**: The complete collection of individuals or elements we want to study.

2. **Sampled population**: The collection of all possible elements that might have been chosen in a sample; the population from which the sample was taken.

3. **Population structure**: The survey population may show certain specific structure. Stratification and clustering are the two most common situations.

   Sometimes, due to administrative or geographical restrictions, the population is divided into a number of distinct strata or subpopulations U_j, $j = 1, 2, \cdots, H$, such that $U_j \cap U_k = \emptyset$ for $j \neq k$ and $U_1 \cup U_2 \cup \cdots \cup U_H = U$. The number of elements in stratum U_j is often denoted as N_j, called the stratum size. We have $N_1 + N_2 + \cdots + N_H = N$.

   Clustering occurs when no reliable list of the elements or individuals in the population is available but groups, called clusters, of elements are easy to identify. For example, a list of all residents in a city may not exist but a list of all households will be easy to construct. Here households are clusters and individual residents are the elements.

4. **Sampling unit**: The unit we actually sample. Sampling units can be the individual elements, or clusters.

5. **Observation unit**: The unit we take measurement from. Observation units are usually the individual elements.

6. **Sampling frame**: The list of sampling units.

7. **Sampling design**: Method of selecting a sample. There are two general types of sampling designs used in practice: probability sampling,

which will be discussed in more detail in subsequent chapters, and nonprobability sampling. Nonprobability sampling includes (a) purposive or judgmental sampling; (b) a sample of convenience; (c) restrictive sampling; (d) quota sampling; and (e) a sample of volunteers.

Despite of the best effort in applications, the sampled population is usually not identical to the target population. It is important to notice that conclusions from a sample survey can only be applied to the sampled population. In probability sampling, unbiased estimates of population parameters can be constructed. Standard errors and confidence intervals can also be reported. Under nonprobability sampling, none of these are possible.

The planning and execution of a survey may involve some or all of following steps:

1. A clear statement of objectives.
2. The population to be sampled.
3. The relevant data to be collected: define study variable(s) and population quantities.
4. Required precision of estimates.
5. The population frame: define sampling units and construct the list of the sampling units.
6. Method of selecting the sample.
7. Organization of the field work.
8. Plans for handling non-response.
9. Summarizing and analyzing the data: estimation procedures and other statistical techniques to be employed.
10. Writing reports.

A few additional remarks about the probability sampling plan. In any single sampling survey, not all units in the population will be chosen. Yet we try hard to make sure the chance for any single unit to be selected is positive. If this is not the case, it results in the difference between the target population and the sampled population. If the difference is substantial, the conclusions based on the survey have to be interpreted carefully.

Most often, we wish that each sampling unit has equal probability to be included into the sample. If this is not the case, then the sampling plan is often referred as **biased**. If the resulting sampling data set is analyzed without detailed knowledge of selection bias, the final conclusion is biased.

If the sampling plan is biased, and we know how it is biased, then we can try to accommodate this information into our analysis. The conclusion can still be unbiased in a loose sense. In some applications, introducing biased sampling plan enables us to make more efficient inference. Thus, a biased plan might be helpful. However, in most cases, the bias is hard to model, and hard to accommodate in the analysis. They are to be avoided.

The basic elements of experimental design will be discussed in Chapter 6.

1.5 Statistical analysis

We will focus on the estimation of population mean $\bar{Y}$ or proportion $P = M/N$ based on probability samples. In each case, we will construct (unbiased) estimators, estimate the variance of the estimator, and build confidence intervals using a point estimate and its estimated standard error.

Chapter 2

Simple Probability Samples

2.1 Probability sampling

In probability sampling, each element (sampling unit) in the (study) population has a known, non-zero probability of being included in the sample. Such a sampling can be specified through a probability measure defined over the set of all possible samples.

Since the sampling unit and the element are often the same, we will treat them as the same unless otherwise specified.

Example 2.1 Let $N = 3$ and $U = \{1, 2, 3\}$. All possible candidate samples are $s_1 = \{1\}$, $s_2 = \{2\}$, $s_3 = \{3\}$, $s_4 = \{1, 2\}$, $s_5 = \{1, 3\}$, $s_6 = \{2, 3\}$, $s_7 = \{1, 2, 3\}$. A probability measure $P(\cdot)$ is given by

s	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$P(s)$	1/9	1/9	1/9	2/9	2/9	2/9	0

Selection of a sample based on above probability measure can be done using a random number generator in Splus or R.

The code in R is:

```
> sample( 1:7, 1, prob=c(1, 1, 1, 2, 2, 2, 0)/9)
```

The output will be a number between 1 and 6 with the corresponding probability.

The probability that element i is selected in the sample is called **inclusion probability**, denoted by $\pi_i = P(i \in s)$, $i = 1, 2, \cdots, N$. It is required that all $\pi_i > 0$. If $\pi_i = 1$, the element will be included in the sample for certainty.

Remark: Suppose $\pi_j = 0$ when $j = 2$, say. It implies that the element 2 is virtually not in the population because it will never be selected.

Let $\nu(s)$ = the number of elements in s. We say a sampling design has fixed sample size n if $\nu(s) \neq n$ implies $P(s) = 0$.

Remark: Do not get confused between elements and samples.

Example 2.2 Let $U = \{1, 2, 3\}$ and $s_1, \cdots, s_7$ be defined as in Example 2.1. The following sampling design has fixed sample size of $n = 2$.

s	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$P(s)$	0	0	0	1/3	1/3	1/3	0

Remark: Try to write a R code for this sampling plan.

Under probability sampling, unbiased estimates of commonly used population parameters can be constructed. Standard errors and confidence intervals should also be reported.

2.2 Simple random sampling without replacement

One of the simplest probability sampling designs (plans) to select a sample of fixed size n with equal probability, i.e. $P(s) = \binom{N}{n}^{-1}$ if $\nu(s) = n$; $P(s) = 0$ otherwise. One way to select such a sample is use **Simple Random Sampling Without Replacement** (SRSWOR): select the 1st element from $U = \{1, 2, \cdots, N\}$ with probability $1/N$; select the 2nd element from the remaining $N-1$ elements with probability $1/(N-1)$; and continue this until n elements are selected. Let $\{y_i, i \in s\}$ be the sample data.

It can be shown that under SRSWOR, $P(s) = \binom{N}{n}^{-1}$ if $\nu(s) = n$, $P(s) = 0$ otherwise. In practice, the scheme can be carried out using a table of random numbers or computer generated random numbers (such as *sample(N,n)* in Splus or R).

In a more scientific respect, either of the above methods truly provides a random sample. There were examples when the outcomes of "random number" generated by computer were predicted. For the purpose of sampling survey, generating pseudo random numbers is most practical as well as effective.

Result 2.1 Under SRSWOR, the sample mean $\bar{y}$ is an unbiased estimator of $\bar{Y}$, i.e. $E(\bar{y}) = \bar{Y}$. $\diamondsuit$

Result 2.2 Under SRSWOR, the variance of $\bar{y}$ is given by $V(\bar{y}) = (1 - f)S^2/n$, where $f = n/N$ is the sampling fraction, S^2 is the population variance. $\diamondsuit$

The $1 - f$ is called the finite population correction factor.

It is seen that when the sample size increases, both factors $(1 - f)$ and S^2/n decrease. The practical implications are: the precision of the statistical inference improves when we collect more information. In addition, suppose we have two finite populations with about the same population variances $S_1^2 \approx S_2^2$, but one has much larger population size than the other one, say $N_1 >> N_2$. In this case, the variance of the sample means from these two populations are approximately equal as long as $n_1 \approx n_2$. To many, this outcome is quite counter-intuitive. Yet this is a well established result, and it has been verified in applications again and again.

Result 2.3 Under SRSWOR, (1) the sample variance s^2 is an unbiased estimator of S^2; (2) $v(\bar{y}) = (1 - f)s^2/n$ is an unbiased estimator of $V(\bar{y})$. $\diamondsuit$

Some remarks:

1. $\bar{Y}$ is a population parameter, a constant but unknown;

2. $\bar{y}$ is a statistic (should be viewed as a random variable before the sample is taken), and is computable once the sample is taken;

3. $V(\bar{y}) = (1-f)S^2/n$ is a constant but unknown (since S^2 is unknown!);

4. $V(\bar{y})$ can be estimated by replacing S^2 by s^2.

5. Confidence intervals: an approximately $1 - \alpha$ CI for $\bar{Y}$ is given by $[\bar{y} - z_{\alpha/2}SE(\bar{y}), \bar{y} + z_{\alpha/2}SE(\bar{y})]$, where SE is the estimated standard error of $\bar{y}$. When n is small, $z_{\alpha/2}$ might be replaced by $t_{\alpha/2}(n-1)$, but the exact coverage probability of this CI is unknown for either choices.

6. In some books, the population variance S^2 is defined slightly differently. The formula can hence differ a little. You need not be alarmed.

The results on the estimation of $\bar{Y}$ apply to two other parameters: the population total Y and the population proportion $P = M/N$.

2.3 Simple random sampling with replacement

Select the 1st element from $\{1, 2, \cdots, N\}$ with equal probability; select the 2nd element also from $\{1, 2, \cdots, N\}$ with equal probability; repeat this n times. This sampling scheme is called simple random sampling with replacement (SRSWR). Under SRSWR, some elements in the population may be selected more than once. Let $y_1, y_2, \cdots, y_n$ be the values for the n selected elements and $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$.

Result 2.4 Under SRSWR, $E(\bar{y}) = \bar{Y}$, $V(\bar{y}) = \sigma^2/n$, where $\sigma^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2/N$. ♢

SRSWOR is more efficient than SRSWR. When N is very large and n is small, SRSWOR and SRSWR will be very close to each other.

2.4 Systematic sampling

Suppose we want to take a sample of size n from the population U of size N. The population elements are ordered in a sequence. Assume $N = n \times k$. To take a systematic sample, choose a random number r between 1 and k, the elements numbered $r, r + k, r + 2k, \cdots, r + (n - 1)k$ will form the sample. r is called random starting point.

Systematic sampling is often used in practice due to two reasons: (1) it is sometimes easier to do a systematic sampling than SRS, particular so if a complete list of sampling units is not available. Systematic sampling is also approximately the same as SRSWOR when the population is roughly in a random order; (2) systematic sampling is more efficient than SRS when there is a linear trend in the ordered population.

Under systematic sampling where $N = n \times k$, there are only k candidate samples $s_1, s_2, \cdots, s_k$. Yet each element in the population has the same probability of being sampled. In this respect, it has some similarities with SRSWOR. Let $\bar{y}(s_r) = n^{-1} \sum_{i \in s_r} y_i$.

Result 2.5 Under systematic sampling, $E(\bar{y}) = \bar{Y}$, $V(\bar{y}) = k^{-1} \sum_{r=1}^{k} [\bar{y}(s_r) - \bar{Y}]^2$. ♢

Example 2.3 Suppose the population size $N = 12$, and $\{y_1, y_2, \cdots, y_{12}\} = \{2, 4, 6, \cdots, 24\}$. Here $\bar{Y} = 13$ and $S^2 = 52$. For a sample of size $n = 4$: (i) Under SRSWOR, $V(\bar{y}) = (1 - 1/3)S^2/4 \doteq 8.67$; (ii) Under systematic sampling, there are three candidate samples, s_1: $\{2, 8, 14, 20\}$; s_2: $\{4, 10, 16, 22\}$;

s_3: $\{6, 12, 18, 24\}$. The three sample means are $\bar{y}(s_1) = 11$, $\bar{y}(s_2) = 13$ and $\bar{y}(s_3) = 15$. $V(\bar{y}) = [(11-13)^2 + (13-13)^2 + (15-13)^2]/3 \doteq 2.67$.

There are two major problems associated with systematic sampling. The first is variance estimation. Unbiased variance estimator is not available. If the population can be viewed as in a random order, variance formula for SRSWOR can be borrowed. The other problem is that if the population is in a periodic or cyclical order, results from a systematic sample can be very unreliable.

In another vein, the systematic sampling plan can be more efficient when there is a linear trend in the ordered population. Borrowing the variance formula from SRSWOR results in conservative statistical analysis.

2.5 Cluster sampling

In many practical situations the population elements are grouped into a number of clusters. A list of clusters can be constructed as the sampling frame but a complete list of elements is often unavailable, or too expensive to construct. In this case it is necessary to use cluster sampling where a random sample of clusters is taken and some or all elements in the selected clusters are observed. Cluster sampling is also preferable in terms of cost, because it is much cheaper, easier and quicker to collect data from adjoining elements than elements chosen at random. On the other hand, cluster sampling is less informative and less efficient per elements in the sample, due to similarities of elements within the same cluster. The loss of efficiency, however, can often be compensated by increasing the overall sample size. Thus, in terms of unit cost, the cluster sampling plan is efficient.

Suppose the population consists of N clusters. The ith cluster consists of M_i elements. We consider a simple situation where the cluster sizes M_i are all the same, i.e. $M_i \equiv M$. Let y_{ij} be the y value for the jth element in the ith cluster. The population size (total number of elements) is NM, the population mean (per element) is

$$\bar{\bar{Y}} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \,,$$

the population variance (per element) is

$$S^2 = \frac{1}{NM-1} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{\bar{Y}})^2 \,.$$

The mean for the ith cluster is $\bar{Y}_i = M^{-1}\sum_{j=1}^{M} y_{ij}$, and the variance for the ith cluster is $S_i^2 = (M-1)^{-1}\sum_{j=1}^{M}(y_{ij} - \bar{Y}_i)^2$.

One-stage cluster sampling: Take n clusters (denoted by s) using simple random sampling without replacement, and all elements in the selected clusters are observed. The sample mean (per element) is given by

$$\bar{\bar{y}} = \frac{1}{nM}\sum_{i\in s}\sum_{j=1}^{M} y_{ij} = \frac{1}{n}\sum_{i\in s}\bar{Y}_i\,.$$

Result 2.6 Under one-stage cluster sampling with clusters sampled using SRSWOR,

(i) $E(\bar{\bar{y}}) = \bar{\bar{Y}}$.

(ii) $V(\bar{\bar{y}}) = (1-\frac{n}{N})\frac{S_M^2}{n}$, where $S_M^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{\bar{Y}})^2$.

(iii) $v(\bar{\bar{y}}) = (1-\frac{n}{N})\frac{1}{n}\frac{1}{n-1}\sum_{i\in s}(\bar{Y}_i - \bar{\bar{y}})^2$ is an unbiased estimator for $V(\bar{\bar{y}})$.

♢

When cluster sizes are not all equal, complications will arise. When M_i's are all known, simple solutions exist, otherwise a ratio type estimator will have to be used. It is also interesting to note that systematic sampling is a special case of one-stage cluster sampling.

2.6 Sample size determination

In planning a survey, one needs to know how big a sample he should draw. The answer to this question depends on how accurate he wants the estimate to be. We assume the sampling scheme is SRSWOR.

1. Precision specified by absolute tolerable error

The surveyor can specify the margin of error, e, such that

$$P(|\bar{y} - \bar{Y}| > e) \le \alpha$$

for a chosen value of α, usually taken as 0.05. Approximately we have

$$e = z_{\alpha/2}\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}}.$$

Solving for n , we have

$$n = \frac{z^2_{\alpha/2} S^2}{e^2 + z^2_{\alpha/2} S^2/N} = \frac{n_0}{1 + n_0/N}$$

where $n_0 = z^2_{\alpha/2} S^2/e^2$.

2. Precision specified by relative tolerable error

The precision is often specified by a relative tolerable error, e.

$$P\left(\frac{|\bar{y} - \bar{Y}|}{|\bar{Y}|} > e\right) \leq \alpha$$

The required n is given by

$$n = \frac{z^2_{\alpha/2} S^2}{e^2 \bar{Y}^2 + z^2_{\alpha/2} S^2/N} = \frac{n_0^*}{1 + n_0^*/N} \, .$$

Where $n_0^* = z^2_{\alpha/2}(CV)^2/e^2$, and $CV = S/\bar{Y}$ is the coefficient of variation.

3. Sample size for estimating proportions

The absolute tolerable error is often used, $P(|p - P| > e) \leq \alpha$, and the common choice of e and α are 3% and 0.05. Also note that $S^2 \doteq P(1 - P)$, $0 \leq P \leq 1$ implies $S^2 \leq 1/4$. The largest value of required sample size n occurs at $P = 1/2$.

Sample size determination requires the knowledge of S^2 or CV. There are two ways to obtain information on these.

(a) Historical data. Quite often there were similar studies conducted previously, and information from these studies can be used to get approximate values for S^2 or CV.

(b) A pilot survey. Use a small portion of the available resource to conduct a small scale pilot survey before the formal one to obtain information about S^2 or CV.

Other methods are often ad hoc. For example, if a population has a range of 100. That is, the largest value minus the smallest value is no more than 100. Then a conventional estimate of S is 100/4. This example is applicable when the age is the study variable.

Chapter 3

Stratified Sampling

We mentioned in Section 1.4 that sometimes the population is naturally divided into a number of distinct non-overlapping subpopulations called strata U_h, $h = 1, 2, \cdots, H$, such that $U_h \cap U_{h'} = \emptyset$ for $h \neq h'$ and $U_1 \cup U_2 \cup \cdots \cup U_H = U$. Let N_h be the hth stratum size. We must have $N_1 + N_2 + \cdots + N_H = N$. The population is said to have a stratified structure. Stratification may also be imposed by the surveyor for the purpose of better estimation.

Let y_{hj} be the y value of the jth element in stratum h, $h = 1, 2, \cdots, H$, $j = 1, 2, \cdots, N_h$. Some related population quantities are:

1. The hth stratum mean $\bar{Y}_h = N_h^{-1} \sum_{j=1}^{N_h} y_{hj}$.
2. The population mean $\bar{Y} = N^{-1} \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj}$.
3. The hth stratum variance $S_h^2 = (N_h - 1)^{-1} \sum_{j=1}^{N_h} (y_{hj} - \bar{Y}_h)^2$.
4. The population variance $S^2 = (N - 1)^{-1} \sum_{h=1}^{H} \sum_{j=1}^{N_h} (y_{hj} - \bar{Y})^2$.

It can be shown that

$$\begin{aligned} \bar{Y} &= \sum_{h=1}^{H} W_h \bar{Y}_h \,, \\ (N-1)S^2 &= \sum_{h=1}^{H} (N_h - 1) S_h^2 + \sum_{h=1}^{H} N_h (\bar{Y}_h - \bar{Y})^2 \,, \end{aligned}$$

where $W_h = N_h/N$ is called the stratum weight. The second equality can be alternatively re-stated as that

Total variation = Within strata variation+Between strata variation.

This relationship is needed when we make comparisons between SRS and stratified sampling.

For students who are still fresh with some facts in probability theory, you may relate the above decomposition with a formula as follows. Let X and Y be two random variables. We have

$$\mathrm{Var}(Y|X) = \mathrm{Var}\{\mathrm{E}(Y|X)\} + \mathrm{E}\{\mathrm{Var}(Y|X)\}.$$

3.1 Stratified random sampling

To take a sample s with fixed sample size n from a stratified population, a decision will have to be made first on how many elements are to be selected from each stratum. Let $n_h > 0$ be the number of elements drawn from stratum h, $h = 1, 2, \cdots, H$. It follows that $n = n_1 + n_2 + \cdots + n_H$.

Suppose a sample s_h of size n_h is taken from stratum h. The overall sample is therefore given by

$$s = s_1 \cup s_2 \cup \cdots \cup s_H \,.$$

Let y_{hj}, $h = 1, 2, \cdots, H$, $j \in s_h$ be the observed values for the y variable. The sample mean and sample variance for stratum h are given by

$$\bar{y}_h = \frac{1}{n_h} \sum_{j \in s_h} y_{hj} \quad \text{and} \quad s_h^2 = \frac{1}{n_h - 1} \sum_{j \in s_h} (y_{hj} - \bar{y}_h)^2 \,.$$

If s_h is taken from the hth stratum using simple random sampling without replacement, and **samples from different strata are independent of each other**, the sampling scheme is termed **Stratified Random Sampling**.

The main motivation of applying stratified simple random sampling is the administrative convenience. It turns out, though, that the estimation based on stratified simple random sampling is more efficient for majority of populations in applications.

Result 3.1 Under stratified random sampling,

(i) $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$ is an unbiased estimator of $\bar{Y}$;

(ii) $V(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) S_h^2 / n_h$, where $f_h = n_h / N_h$ is the sampling fraction in the hth stratum;

(iii) $v(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) s_h^2 / n_h$ is an unbiased estimator of $V(\bar{y}_{st})$.

♢

The proof follows directly from results of SRSWOR and the fact that $s_1, s_2, \cdots, s_H$ are independent of each other. The results can also be easily modified to handle the estimation of population total Y and population proportion P.

Stratified sampling is different from cluster sampling. In both cases the population is divided into subgroups: strata in the former and clusters in the latter. In cluster sampling only a **portion** of clusters are sampled while in stratified sampling **every** stratum will be sampled. Usually, only a subset of the elements in a stratum are observed, while all elements in a sampled cluster are observed.

Questions associated with stratified sampling include (i) Why use stratified sampling? (ii) How to stratify? and (iii) How to allocate sample sizes to each stratum? We will address questions (ii) and (iii) in Sections 3.2 and 3.3. There are four main reasons to justify the use of stratified sampling:

(1) Administrative convenience. A survey at national level can be greatly facilitated if officials associated with each province survey a portion of the sample from their province. Here provinces are the natural choice of strata.

(2) In addition to the estimates for the entire population, estimates for certain sub-population are also required. For example, one might require the estimates of unemployment rate for not only at the national level but for each province as well.

(3) Protect from possible disproportional samples under probability sampling. For instance, a random sample of 100 students from University of Waterloo may contain only few female students. In theory there shouldn't be any concern about this unusual case, but the results from the survey will be more acceptable to the public if, say, the sample consists of 50 male students and 50 female students.

(4) Increased accuracy of estimate. Stratified sampling can often provide more accurate estimates than SRS. This also relates to the other questions: how to stratify? and how to allocate the sample sizes? We will return to these questions in next sections.

3.2 Sample size allocation

We consider two commonly used schemes in allocating the sample sizes into each of the strata: proportional allocation, and optimal allocation for a given n, the total sample size.

1. Proportional allocation

With no extra information except the stratum size, N_h, we should allocate the stratum sample size proportional to the stratum size, i.e. $n_h \propto N_h$. Under the restriction that $n_1 + n_2 + \cdots + n_H = n$, the resulting allocation is given by

$$n_h = n\frac{N_h}{N} = nW_h\,, \quad h = 1, 2, \cdots, H\,.$$

Result 3.2 Under stratified random sampling with proportional allocation,

$$V_{prop}(\bar{y}_{st}) = (1 - \frac{n}{N})\frac{1}{n}\sum_{h=1}^{H} W_h S_h^2\,.$$

$\diamondsuit$

2. Optimal allocation (Neyman allocation)

When the total sample size n is fixed, an optimal allocation $(n_1, n_2, \cdots, n_H)$ can be found by minimizing $V(\bar{y}_{st})$ subject to constraint $n_1 + n_2 + \cdots + n_H = n$.

Result 3.3 In stratified random sampling $V(\bar{y}_{st})$ is minimized for a fixed total sample size n if

$$n_h = n\frac{W_h S_h}{\sum_{h=1}^{H} W_h S_h} = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}\,,$$

and the minimum variance is given by

$$V_{min}(\bar{y}_{st}) = \frac{1}{n}(\sum_{h=1}^{H} W_h S_h)^2 - \frac{1}{N}\sum_{h=1}^{H} W_h S_h^2\,.$$

$\diamondsuit$

To carry out an optimal allocation, one requires knowledge of S_h, $h = 1, 2, \cdots, H$. Since rough estimates of the S_h's will be good enough to do a sample size allocation, one can gather this information from historical data, or through a small scale pilot survey.

3.3 A comparison to SRS

It will be of interest to make a comparison between stratified random sampling and SRSWOR. In general, stratified random sampling is more efficient than simple random sampling.

Result 3.4 Let $\bar{y}_{st}$ be the stratified sample mean and $\bar{y}$ be the sample mean from SRSWOR, both with a total sample size of n. Then, treating $(N_h - 1)/(N - 1) \doteq N_h/N$, we have

$$V(\bar{y}) - V_{prop}(\bar{y}_{st}) \doteq (1 - \frac{n}{N})\frac{1}{n}\sum_{h=1}^{H} W_h(\bar{Y}_h - \bar{Y})^2 \geq 0\,.$$

$\diamondsuit$

It is now clear from Result 3.4 that when proportional allocation is used, stratified random sampling is (almost) always more efficient than SRSWOR. The gain of efficiency depends on the between-strata variation. This also provides guidance on how to stratify: the optimal stratification under proportional allocation is the one which produces the largest possible differences between the stratum means. Such a stratification also maximizes the homogeneity of the y-values within each stratum.

In practice, certain prior information or common knowledge can be used for stratification. For example, in surveying human populations, people with same sex, age and income level are more likely similar to each other. Stratification by sex, age and/or sex-age group combinations will be a reasonable choice.

Another factor that may affect our decision of sample allocation is the unit cost per sampling unit. The cost of taking a sample from some strata can be higher than other strata. The optimal allocation under this situation can be similarly derived but is not be discussed in this course. To differentiate these two optimal schemes, the optimal allocation discussed is also called Neyman allocation.

Chapter 4

Ratio and Regression Estimation

Often in survey sampling, information on one (or more) covariate x (called auxiliary variable) is available prior to sampling. Sometimes this auxiliary information is complete, i.e. the value x_i is known for every element i in the population; sometimes only the population mean $\bar{X} = N^{-1} \sum_{i=1}^{N} x_i$ or total $X = \sum_{i=1}^{N} x_i$ is known. When the auxiliary variable x is correlated with the study variable y, this known auxiliary information can be useful for the new survey study.

Example 4.1 In family expenditure surveys, the values on $x^{(1)}$: the number of people in the family and/or $x^{(2)}$: the family income of previous year are known for every element in the population. The study variable(s) is on current year family expenditures such as expenses on clothing, food, furniture, etc.

Example 4.2 In agriculture surveys, a complete list of farms with the area (acreage) of each farm is available.

Example 4.3 Data from earlier census provides various population totals that can be used as auxiliary information for the planned surveys.

Auxiliary information can be used at the design stage. For instance, a stratified sampling scheme might be chosen where stratification is done by values of certain covariates such as sex, age and income levels. The *pps* sampling design (inclusion probability proportional to a size measure) is another sophisticated example.

In this chapter, we use auxiliary information explicitly at the estimation

stage by incorporating the known $\bar{X}$ or X into the estimators through ratio and regression estimation. The resulting estimators will be more efficient than those discussed in previous chapters.

4.1 Ratio estimator

4.1.1 Ratio estimator under SRSWOR

Suppose y_i is approximately proportional to x_i, i.e. $y_i \doteq \beta x_i$ for $i = 1, 2, \ldots, N$. It follows that $\bar{Y} \doteq \beta \bar{X}$. Let $R = \bar{Y}/\bar{X} = Y/X$ be the ratio of two population means or totals. Let $\bar{y}$ and $\bar{x}$ be the sample means under SRSWOR. It is natural to use $\hat{R} = \bar{y}/\bar{x}$ to estimate R. The ratio estimator for $\bar{Y}$ is defined as

$$\hat{\bar{Y}}_R = \hat{R}\bar{X} = \frac{\bar{y}}{\bar{x}}\bar{X}\,.$$

One can expect that $\bar{X}/\bar{x}$ will be close to 1, so $\hat{\bar{Y}}_R$ will be close to $\bar{y}$. Why is ratio estimator often used? The following results will provide an answer. Note that $R = \bar{Y}/\bar{X}$ is the (unknown) population ratio, and $\hat{R} = \bar{y}/\bar{x}$ is a sample-based estimate for R.

Result 4.1 Under simple random sampling without replacement,

(i) $\hat{\bar{Y}}_R$ is approximately unbiased estimator for $\bar{Y}$.

(ii) The variance of $\hat{\bar{Y}}_R$ can be approximated by

$$V(\hat{\bar{Y}}_R) \doteq (1 - \frac{n}{N})\frac{1}{n}\frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2\,.$$

(iii) An approximately unbiased variance estimator is given by

$$v(\hat{\bar{Y}}_R) = (1 - \frac{n}{N})\frac{1}{n}\frac{1}{n-1}\sum_{i\in s}(y_i - \hat{R}x_i)^2\,.$$

♢

To see when the ratio estimator is better than the simple sample mean $\bar{y}$, lets make a comparison between the two variances. Note that

$$V(\bar{y}) \;=\; (1 - \frac{n}{N})\frac{1}{n}S_Y^2\,,$$

$$\begin{aligned} V(\hat{\bar{Y}}_R) &\doteq (1-\frac{n}{N})\frac{1}{n}\frac{1}{N-1}\sum_{i=1}^{N}[(y_i-\bar{Y})-R(x_i-\bar{X})]^2 \\ &= (1-\frac{n}{N})\frac{1}{n}[S_Y^2+R^2S_X^2-2RS_{XY}]\,, \end{aligned}$$

where S_Y^2 and S_X^2 are the population variances for the y and x variables, and $S_{XY} = (N-1)^{-1}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})$. The ratio estimator will have a smaller variance if and only if

$$R^2S_X^2 - 2RS_{XY} < 0\,.$$

This condition can also be re-expressed as

$$\rho > \frac{1}{2}\frac{CV(X)}{CV(Y)}\,,$$

where $\rho = S_{XY}/[S_XS_Y]$, $CV(X) = S_X/\bar{X}$ and $CV(Y) = S_Y/\bar{Y}$. The conclusion is: if there is a strong correlation between y and x, the ratio estimator will perform better than the simple sample mean. Indeed, in many practical situations $CV(X) \doteq CV(Y)$, we only require $\rho > 0.5$. This is usually the case.

A scatter plot of the data can visualize the relationship between y and x. If a straight line going through the origin is appropriate, ratio estimator may be efficient.

Ratio estimator can provide improved estimate. There are other situations where we have to use a ratio type estimator. Under one-stage cluster sampling with clusters of unequal sizes and M_i are not known unless the ith cluster is selected in the sample, the population mean (per element) is indeed a ratio:

$$\bar{\bar{Y}} = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}/\sum_{i=1}^{N} M_i = [\frac{1}{N}\sum_{i=1}^{N} Y_i]/[\frac{1}{N}\sum_{i=1}^{N} M_i]\,.$$

A natural estimate for $\bar{\bar{Y}}$ would be $\hat{\bar{\bar{Y}}} = [n^{-1}\sum_{i\in s} Y_i]/[n^{-1}\sum_{i\in s} M_i]$.

4.1.2 Ratio estimator under stratified random sampling

When the population has been stratified, ratio estimator can be used in two different ways: (a) estimate $R = \bar{Y}/\bar{X}$ by $\hat{R} = \bar{y}_{st}/\bar{x}_{st}$, and $\bar{Y} = R\bar{X}$ by $\hat{R}\bar{X}$;

or (b) write $\bar{Y}$ as $\bar{Y} = \sum_{h=1}^{H} W_h \bar{Y}_h$ and estimate $\bar{Y}_h$, the strata mean, by a ratio estimator $[\bar{y}_h/\bar{x}_h]\bar{X}_h$. In (a), only $\bar{X}$ needs be known; under (b), the stratum means $\bar{X}_h$ are required.

The combined ratio estimator of $\bar{Y}$ is defined as

$$\hat{\bar{Y}}_{Rc} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} \,,$$

where $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$, $\bar{x}_{st} = \sum_{h=1}^{H} W_h \bar{x}_h$. The separate ratio estimator of $\bar{Y}$ is defined as

$$\hat{\bar{Y}}_{Rs} = \sum_{h=1}^{H} W_h \frac{\bar{y}_h}{\bar{x}_h} \bar{X}_h \,,$$

where the $\bar{X}_h$'s are the known strata means.

Result 4.2 Under stratified random sampling, the combined ratio estimator $\hat{\bar{Y}}_{Rc}$ is approximately unbiased for $\bar{Y}$, and its variance is given by

$$V(\hat{\bar{Y}}_{Rc}) \doteq \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{1}{n_h} \frac{1}{N_h - 1} \sum_{j=1}^{N_h} [(y_{hj} - \bar{Y}_h) - R(x_{hj} - \bar{X}_h)]^2 \,,$$

which can be estimated by

$$v(\hat{\bar{Y}}_{Rc}) \doteq \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{1}{n_h} \frac{1}{n_h - 1} \sum_{j \in s_h} [(y_{hj} - \bar{y}_h) - \hat{R}(x_{hj} - \bar{x}_h)]^2 \,,$$

where $R = \bar{Y}/\bar{X}$ and $\hat{R} = \bar{y}_{st}/\bar{x}_{st}$. ♢

Result 4.3 Under stratified random sampling, the separate ratio estimator $\hat{\bar{Y}}_{Rs}$ is approximately unbiased for $\bar{Y}$, and its variance is given by

$$V(\hat{\bar{Y}}_{Rs}) \doteq \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{1}{n_h} \frac{1}{N_h - 1} \sum_{j=1}^{N_h} (y_{hj} - R_h x_{hj})^2 \,,$$

which can be estimated by

$$v(\hat{\bar{Y}}_{Rs}) \doteq \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{1}{n_h} \frac{1}{n_h - 1} \sum_{j \in s_h} (y_{hj} - \hat{R}_h x_{hj})^2 \,,$$

where $R_h = \bar{Y}_h/\bar{X}_h$ and $\hat{R}_h = \bar{y}_h/\bar{x}_h$. ♢

One of the questions that needs to be addressed is how to make a choice between $\hat{\bar{Y}}_{Rc}$ and $\hat{\bar{Y}}_{Rs}$. First, it depends on what kind of auxiliary information is available. The separate ratio estimator requires the strata means $\bar{X}_h$ being known. If only $\bar{X}$ is known, the combined ratio estimator will have to be used. Second, in terms of efficiency, the variance of $\hat{\bar{Y}}_{Rc}$ depends on the "residuals" $e_{hj} = (y_{hj} - \bar{Y}_h) - R(x_{hj} - \bar{X}_h)$, which is equivalent to fit a single straight line across all the strata with a common slope; while for the separate ratio estimator this slope can be different for different strata. So in many situations $\hat{\bar{Y}}_{Rs}$ will perform better than $\hat{\bar{Y}}_{Rc}$. Third, the variance formula for the separate ratio estimator depends on the approximation to $\bar{y}_h/\bar{x}_h$. If the sample sizes within each stratum, n_h, are too small, the bias from using $\hat{\bar{Y}}_{Rs}$ can be large. The bias from using $\hat{\bar{Y}}_{Rc}$, however, will be smaller since the approximation is made to $\bar{y}_{st}/\bar{x}_{st}$, and the pooled sample size n will usually be large.

4.2 Regression estimator

The study variable y is often linearly related to the auxiliary variable x, i.e. $y_i \doteq \beta_0 + \beta_1 x_i$, $i = 1, 2, \cdots, N$. So roughly we have $\bar{Y} \doteq \beta_0 + \beta_1 \bar{X}$ and $\bar{y} \doteq \beta_0 + \beta_1 \bar{x}$. This leads to the regression type estimator of $\bar{Y}$: $\hat{\bar{Y}} = \bar{y} + \beta_1(\bar{X} - \bar{x})$. The β_1 is usually unknown and is estimated by the least square estimator $\hat{\beta}_1$ from the sample data. More formally, under SRSWOR, the regression estimator of $\bar{Y}$ is defined as

$$\hat{\bar{Y}}_{REG} = \bar{y} + \hat{B}(\bar{X} - \bar{x}) \,,$$

where $\hat{B} = \sum_{i \in s}(y_i - \bar{y})(x_i - \bar{x}) / \sum_{i \in s}(x_i - \bar{x})^2$.

Result 4.4 Under SRSWOR, the regression estimator $\hat{\bar{Y}}_{REG}$ is approximately unbiased for $\bar{Y}$. Its approximate variance is given by

$$V(\hat{\bar{Y}}_{REG}) \doteq (1 - \frac{n}{N})\frac{1}{n}\frac{1}{N-1}\sum_{i=1}^{N} e_i^2 \,,$$

where $e_i = y_i - B_0 - Bx_i$, $B = \sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X}) / \sum_{i=1}^{N}(x_i - \bar{X})^2$, and $B_0 = \bar{Y} - B\bar{X}$. This variance can be estimated by

$$v(\hat{\bar{Y}}_{REG}) = (1 - \frac{n}{N})\frac{1}{n}\frac{1}{n-1}\sum_{i \in s} \hat{e}_i^2 \,,$$

where $\hat{e}_i = y_i - \hat{B}_0 - \hat{B}x_i$, $\hat{B} = \sum_{i\in s}(y_i - \bar{y})(x_i - \bar{x})/\sum_{i\in s}(x_i - \bar{x})^2$, and $\hat{B}_0 = \bar{y} - \hat{B}\bar{x}$. $\diamondsuit$

It can be shown that

$$V(\hat{\bar{Y}}_{REG}) \doteq (1 - \frac{n}{N})\frac{1}{n}S_Y^2(1 - \rho^2)\,,$$

where $\rho = S_{XY}/[S_X S_Y]$ is the population correlation coefficient between y and x. Since $|\rho| \leq 1$, we have $V(\hat{\bar{Y}}_{REG}) \leq V(\bar{y})$ under SRSWOR. When n is large, the regression estimator is always more efficient than the simple sample mean $\bar{y}$.

It can also be shown that $V(\hat{\bar{Y}}_{REG}) \leq V(\hat{\bar{Y}}_R)$. So regression estimator is preferred in most situations. Ratio estimators are still being used by many survey practitioners due to its simplicity. If a scatter plot of the data shows that a straight line going through the origin fits the data well, then the regression estimator and the ratio estimator will perform similarly. Both requires only $\bar{X}$ be known to compute the estimates under SRSWOR. Under stratified sampling, a combined regression estimator and a separate regression estimator can be developed similarly.

Chapter 5

Survey Errors and Some Related Issues

A survey, especially a large scale survey, consists of a number of stages. Each stage, from the initial planning to the ultimate publication of the results, may require considerable time and effort, with different sources of errors that affect the final reported estimates.

Survey errors can be broadly classified into sampling error and non-sampling error. The sampling error is the amount by which the estimate computed from the data would differ from the true value of the quantity for the sampled population. Under probability sampling, this error can be reduced and controlled through a carefully chosen design and through a reasonably large sample size. All other errors are called non-sampling errors. In this chapter we briefly overview the possible sources of non-sampling errors, with some discussions on how to identify and reduce this type of errors in questionnaire design, telephone surveys and web surveys.

5.1 Non-sampling errors

Major sources of non-sampling errors may include some or all of following:

1. **Coverage error**: The amount by which the quantity for the frame population differs from the quantity for the target population.

2. **Non-response error**: The amount by which the quantity for sampled population differs from the quantity for the frame population.

3. **Measurement error**: In theory, we assume there is a true value y_i attached to the ith element. If the ith element is selected in the sample, the observed value of y is denoted by y_i^*. Since the equipment for the measurement may not accurate, or the questionnaire are not well designed, or the selected individuals intentionally provide incorrect information, y_i^* may differ from y_i. The measurement error is the amount by which the estimate computed from y_i^* differs from the amount computed from y_i.

4. **Errors incurred from data management**: Steps such as data processing, coding, data entry and editing can all bring errors in.

Non-sampling errors are hard to control and are often left un-estimated or unacknowledged in reports of surveys. Well-trained staff members can reduce the error from data management; carefully designed questionnaire well-worded questions in mail surveys or telephone surveys can reduce measurement errors and non-response rate in these cases.

5.2 Non-response

In large scale surveys, it is often the case that for each sampled element several or even many attributes are measured. Non-response, sometimes called missing data, occur when the sampled element cannot be reached or refuse to respond. There are two types of non-response: unit non-response where no information is available for the whole unit, and item non-response where information on certain variables are not available.

1. Effect of non-response

Consider a single study variable y. The finite population can be conceptually divided into two strata: respondent group and non-respondent group, with stratum weights W_1 and W_2. Let $\bar{Y}_1$ and $\bar{Y}_2$ be the means for the two groups. It follows that

$$\bar{Y} = W_1\bar{Y}_1 + W_2\bar{Y}_2\,.$$

Suppose we have data from the respondent group obtained by SRSWOR and $\bar{y}_1$ is the sample mean, but we have no data from the non-respondent group. If we use $\bar{y}_1$ to estimate $\bar{Y}$, which is our original target parameter, the bias would be

$$E(\bar{y}_1) - \bar{Y} = \bar{Y}_1 - \bar{Y} = W_2(\bar{Y}_1 - \bar{Y}_2)\,.$$

The bias depends on the proportion of non-respondents and the difference between the two means. If $\bar{Y}_1$ and $\bar{Y}_2$ are very close, and/or W_2 is very small, the bias will be negligible. On the other hand, if W_2 is not small, and $\bar{Y}_1$ and $\bar{Y}_2$ differ substantially, which is often the case in practical situations, the bias can be non-ignorable.

2. Dealing with non-response

Non-response rates can be reduced through careful planning of the survey and extra effort in the process of data collection. In the planning stage, the attitude of management toward non-response, the selection, training and supervision of interviewers, the choice of data collection method (personal interview, mail inquiry, telephone interview, etc), the design of questionnaire are all important toward the reduction of the non-response. In the process of data collection, some special efforts, such as call-backs in telephone interview, follow-ups in personal interviews or mail inquiries can reduce the non-response dramatically. Other techniques include subsampling of non-respondents and randomized response for sensitive questions.

5.3 Questionnaire design

Measurements of study variables on each selected element (individual) are often obtained by asking the respondents a number of pre-designed questions. Personal interviews, mail surveys, telephone surveys, and web surveys all use a questionnaire. A carefully designed, well-tested questionnaire can reduce both the measurement errors and the non-response rate. Some general guidelines (Lohr, 2000) should be observed when one is writing a questionnaire:

1. **Decide what you want to find out**. This is the most important step in writing a questionnaire. The questions should be precise and they should elicit accurate answers.

2. **Always test your questions before taking the survey**. Try the questions on a very small sample of individuals from the target population and make sure that there are no misinterpretations of the questions to be asked.

3. **Keep it simple and clear**. Think about different wording, think about the diversified background of the individuals selected. Questions that seem clear to you may not be clear to someone else.

4. **Use specific questions instead of general ones, if possible**. This will promote clear and accurate answers to the questions being asked.

5. **Decide whether to use open or closed questions**. Answers to open questions are of free form, while for the closed questions the respondents are forced to choose answer(s) from a pre-specified set of possible answers.

6. **Avoid questions that prompt or motivate the respondent to say what you would like to hear**. These leading (or loaded) type questions can result in serious measurement error problems and bias.

7. **Ask only one concept in each question**. It ensures that accurate answers will most likely be obtained.

8. **Pay attention to question-order effects**. If you ask more than one question, the order of these questions will play a role. If you ask closed questions with more than two possible answers, the order of these answers should also be considered: some respondents will simply choose the first one or the third one!

5.4 Telephone sampling and web surveys

The use of telephone in survey data collection is both cost-effective and time-efficient. However, in addition to the issue of how to design the questions, there are several other unique features related to telephone surveys.

The choice of a sampling frame: there are usually more than one list of telephone numbers available. Typically, not all the numbers in the list belong to the target population and some members from the target population are not on the list. For household surveys, all business numbers should be excluded and those without a phone will not be reached. Sometimes a phone can be shared by a group of people and sometimes a single person may have more than one number. This situation differs from country to country, place to place.

Sample selection: with difficulties arisen from the chosen sampling frame, the selection of a probability sample requires special techniques. The way the numbers are selected, the time of making a call, the person who answers the phone, the way of handling not-reached number, etc, all have impact on the selected sample. Adjustment at the estimation stage is necessary to take these into account.

There is an increased tendency of doing surveys through the web. Similar to telephone surveys, this is a cheap and quick way of gathering data. However, there are serious problems with this kind of surveys. It is very difficult to control and/or distinguish between the target population and the sampled population. The sample data are obtained essentially from a group of volunteers who are interested in providing information through the web. Results from web surveys should always be interpreted with care. The future of web surveys is still uncertain.

Chapter 6

Basic Concepts of Experimental Design

Experimentation allows an investigator to find out what happens to the output variables when the settings of the input variables in a system are purposely changed. In survey sampling, surveyor passively investigates the characteristics of an output variable y, and conceptually, once a unit i is selected, there is a fixed (non-random) value y_i of the output to be obtained. In designed experiment, the values of input variables are carefully chosen and controlled, and the output variables are regarded as random in that the values of the output variables will change over repeated experiments under the same setting of the input variables. We also assume that the setting of the input variables determines the distribution of the output variables, in a way to be discovered. The population under study is the collection of all possible quantitative settings behind each setting of experimental factors and is (at least conceptually) **infinite**.

Example 6.1 **(Tomato Planting)** A gardener conducted an experiment to find whether a change in the fertilizer mixture applied to his tomato plants would result in an improved yield. He had 11 plants set out in a single row; 5 were given the standard fertilizer mixture A, and the remaining 6 were fed a supposedly improved mixture B. The yields of tomatoes from each plant were measured upon maturity.

Example 6.2 **(Hardness Testing)** An experimenter wishes to determine whether or not four different tips produce different readings on a hardness testing machine. The machine operates by pressing the tip into a metal test coupon, and from the depth of the resulting depression, the hardness of the

coupon can be determined. Four observations for each tip are required.

Example 6.3 **(Battery Manufacturing)** An engineering wish to find out the effects of plate material type and temperature on the life of a battery and to see if there is a choice of material that would give uniformly long life regardless of temperature. He has three possible choices of plate materials, and three temperature levels – $15°F, 70°F$, and $125°F$ – are used in the lab since these temperature levels are consistent with the product end-use environment. Battery life are observed at various material-temperature combinations.

The output variable in an experiment is also called the **response**. The input variables are referred to as **factors**, with different **levels** that can be controlled or set by the experimenter. A **treatment** is a combination of factor levels. When there is only one factor, its levels are the treatments. An **experimental unit** is a generic term that refers to a basic unit such as material, animal, person, plant, time period, or location, to which a treatment is applied. The process of choosing a treatment, applying it to an experiment unit, and obtaining the response is called an **experimental run**.

6.1 Five broad categories of experimental problems

1. Treatment comparisons. The main purpose is to compare several treatments and select the best ones. Normally, it implies that a product can be obtained by a number of different ways, and we want to know which one is the best by some standard.

2. Variable screening. The output is likely being influenced by a number of factors. For instance, chemical reaction is controlled by temperature, pressure, concentration, duration, operator, and so on. Is it possible that only some of them are crucial and some of them can be dropped from consideration?

3. Response surface exploration. Suppose a few factors have been determined to have crucial influences on the output. We may then search for a simple mathematical relationship between the values of these factors and the output.

4. System optimization. The purpose of most (statistical) experiments is to find the best possible setting of the input variables. The output of an experiment can be analyzed to help us to achieve this goal.

5. System robustness. Suppose the system is approximately optimized at two (or more) possible settings of the input variables. However, in mass production, it could be costly to control the input variables precisely. The system deteriorates when the values of the input variables deviate from these settings. A setting is most robust if the system deteriorates least.

6.2 A systematic approach to the planning and implementation of experiments

Just like in survey sampling, it is very important to plan ahead. The following five-step procedure is directly from Wu and Hamada (2000).

1. State objectives. What do you want to achieve? (This is usually from your future boss. It could be something you want to demonstrate, and hope that the outcome will impress your future boss).

2. Determine the response. What do you plan to observe? This is similar to the variable of interest in survey sampling.

3. Choose factors and levels. To study the effect of factors, two or more levels of each factor are needed. Factors may be quantitative and qualitative. How much fertilizer you use is a quantitative factor. What kind of fertilizer you use is a qualitative factor.

4. Work out an experimental plan. The basic principle is to obtain the information you need efficiently. A poor design may capture little information which no analysis can rescue. (Come to our statistical consulting centre before doing your experiment. It can be costly to redo the experiment).

5. Perform the experiment. Make sure you will carry out the experiment as planned. If practical situations arise such that you have to alter the plan, be sure to record it in detail.

6.3 Three fundamental principles

There are three fundamental principles in experimental design, namely, replication, randomization, and blocking.

Replication When a treatment is applied to several experiment units, we call it replication. In general, the outcomes of the response variable will differ. This variation reflects the magnitude of **experimental error**. We define the treatment effect as the expected value (mathematical expectation in the word of probability theory) of the response variable (measured against

some standard). The treatment effect will be estimated based on the outcome of the experiment, and the variance of the estimate reduces when the number of replications, or replicates, increases.

It is therefore important to increase the number of replicates, if we intend to detect small treatment effects. For example, if you want to determine if a drug can reduce the breast cancer prevalence by 50%, you probability need only recruit 1,000 women; while to detect a reduction of 5% may need to recruit 10,000 women.

Remember, if you apply a treatment to one experimental unit, but measure the response variable 5 times, you do not have 5 replicates. You have 5 repeated measurements. It helps to reduce the **measurement error**, not experimental error.

Randomization In applications, the experiment units are not identical despite our effort to make them alike. To prevent unwanted influence of subjective judgment, the units should be allocated to treatment in random order. The responses should also be measured in random order (if possible). It provides protection against variables (factors) that are unknown to the experimenter but may impact the response.

Blocking Some experiment units are known to be more similar each other than others. Sometimes we may not have a single large group of alike units for the entire experiment, and several groups of units will have to be used. Units within a group are more homogeneous but may differ a lot from group to group. These groups are referred to as blocks. It is desirable to compare treatments within the same block, so that the block effects are eliminated in the comparison of the treatment effects. Applying the principle of blocking makes the experiment more efficient.

An effective blocking scheme removes the block to block variation. Randomization can then be applied to the assignments of treatments to units within the blocks to further reduce (balance out) the influence of unknown variables. Here is the famous doctrine in experimental design: **block what you can and randomize what you cannot**.

In following chapters some commonly used experimental designs are presented and those basic principles are applied.

Chapter 7

Completely Randomized Design

We consider experiments with a single factor. The goal is to compare the treatments. We also assume the response variable y is quantitative. The tomato plant example is typical, where we wish to compare the effect of two fertilizer mixtures on the yield of tomatoes.

7.1 Comparing 2 treatments

Suppose we want to compare the effects of two different treatments, and there are n experiment units available. We may allocate treatment 1 to n_1 units, and treatment 2 to n_2 units, with $n = n_1 + n_2$. When the n experiment units are homogeneous, the allocation should be completely randomized to avoid possible influences of unknown factors.

Once the observations are obtained, we have sample data as follows

$$y_{11}, y_{12}, \ldots, y_{1,n_1} \quad \text{and} \quad y_{21}, y_{22}, \ldots, y_{2,n_2}.$$

A commonly used statistical model for a single factor experiment is that

$$y_{ij} = \mu_i + e_{ij}\,, \quad i = 1, 2,\ j = 1, 2, \ldots, n_i \tag{7.1}$$

with μ_i being the expectation of y_{ij}, i.e. $E(y_{ij}) = \mu_i$, and e_{ij} being the error terms resulted from repeated experiments and being independent and identically distributed as $N(0, \sigma^2)$.

The above model is, however, something we assume. It may be a good approximation to the real world problem under study. It can also be irrelevant to a particular experiment. For most experiments with quantitative response variable, however, the above model works well.

The statistical analysis of the experiment focuses on answering the question *"Is there a significant difference between the two treatments?"* and, if the answer is yes, trying to identify which treatment is preferable. This is equivalent to testing one of the two types of statistical hypothesis: (1) H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$; and (2) H_0: $\mu_1 \leq \mu_2$ versus H_1: $\mu_1 > \mu_2$. It could certainly also be $\mu_1 > \mu_2$ versus $\mu_1 \leq \mu_2$ but this problem is symmetric to the case of (2). The H_0 is referred to as **Null hypothesis**, and H_1 as **alternative hypothesis** or simply the **alternative**. If larger value of μ_i means a better treatment, then the conclusion of the analysis can be used to decide which treatment to use in future applications.

The test procedures are presented in next section. A key step in constructing the test is to first estimate the unknown means μ_1 and μ_2. Usually, we estimate them by

$$\hat{\mu}_1 = \bar{y}_{1\cdot} = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j} \quad \text{and} \quad \hat{\mu}_2 = \bar{y}_{2\cdot} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_{2j}\,.$$

Under the assumed model, it is easy to verify that $E(\hat{\mu}_i) = \mu_i$ for $i = 1, 2$. So they are both unbiased estimators. Further, we have

$$Var(\hat{\mu}_i) = \sigma^2/n_i, \quad i = 1, 2.$$

It is now clear that the larger the sample size n_i, the smaller the variance of the point estimator $\hat{\mu}_i$. To have a good estimate of μ_1, we should make n_1 large; to have a good estimate of μ_2, we should make n_2 large. **Replications reduce the experimental error and ensure better point estimates for the unknown parameters and consequently more reliable test for the hypothesis**.

Note that the variance σ^2 is assumed to be the same for both treatments. It can be estimated by

$$s_p^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\cdot})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2\cdot})^2\right].$$

This is also called **the pooled variance estimator**, as it uses the y values from both treatments. It can be shown that $E(s_p^2) = \sigma^2$, i.e. s_p^2 is unbiased estimator for σ^2.

Finally, we may estimate $\mu_1 - \mu_2$ by $\hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$. With assumed independence,

$$Var(\hat{\mu}_1 - \hat{\mu}_2) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) .$$

This variance becomes small if both n_1 and n_2 are large. In practice, we often have a limited resource such that $n = n_1 + n_2$ will have to be fixed. In this case we should make $n_1 = n_2$ (or as close as possible) to minimize the variance of $\hat{\mu}_1 - \hat{\mu}_2$.

7.2 Hypothesis test under normal models

A statistical hypothesis test is a decision-making process: you have to make a decision on whether to reject the null hypothesis H_0 based on the information from the sample data. This usually involves the following steps:

1. Start by assuming H_0 is true, and then try to see if information from sample data supports this claim or not.

2. Find a test statistic $T = T(X_1, \cdots, X_n)$. This is often related to the point estimators for the parameters of interest. The test statistic T needs to satisfy two crucial criteria: (i) the value of T is computable solely from the sample data; (ii) the sampling distribution of T is known if H_0 is true.

3. Determine a critical (rejection) region $\{(X_1, \cdots, X_n) : T(X_1, \cdots, X_n) \in C\}$ such that $P(T \in C | H_0) \leq \alpha$ for a prespecified α (usually $\alpha = 0.01, 0.05$ or 0.10).

4. Reach to a final conclusion: for the given sample data, if $T \in C$, reject H_0. Otherwise we fail to reject H_0.

Such a test is called an α level significant test, and $P(\text{reject } H_0 | H_0 \text{ is true})$ is called the type I error probability. We now elaborate the above general procedures through following commonly used two-sample tests.

1. Two sided test

Suppose we wish to test H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$. This is the so-called two sided test problem since the alternative includes both possibilities, $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$. An intuitive argument for the test would be as follows: $\mu_1 - \mu_2$ can be estimated by $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$. If H_0 is true, then $\mu_1 - \mu_2 = 0$, and

consequently we would expect $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$ is also close or at least not far away from 0. In other words, if $|\bar{y}_{1\cdot} - \bar{y}_{2\cdot}| > c$ for certain constant c, we have evidence against H_0 and therefore should reject H_0. The c is determined by $P(|\bar{y}_{1\cdot} - \bar{y}_{2\cdot}| > c|\mu_1 = \mu_2) = \alpha$ for the given α (usually a small positive constant).

Under the assumed normal model (7.1), $y_{ij} \sim N(\mu_i, \sigma^2)$ and that all y_{ij}'s are independent of each other,

$$T = \frac{(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) - (\mu_1 - \mu_2)}{\sigma\sqrt{n_1^{-1} + n_2^{-1}}}$$

is distributed as $N(0,1)$ random variable. If H_0: $\mu_1 = \mu_2$ is true, then

$$T_0 = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{\sigma\sqrt{n_1^{-1} + n_2^{-1}}}$$

is also distributed as $N(0,1)$ random variable. Since $P(|T_0| > Z_{1-\alpha/2}|H_0) = \alpha$, where $Z_{1-\alpha/2}$ is $1-\alpha/2$ quantile of the $N(0,1)$ random variable, we reject H_0 if $|T_0| > Z_{1-\alpha/2}$.

The underlying logic for the above decision rule is as follows: if H_0 is true, then $P(|T_0| > 1.96) = 0.05$, for example. That is, the chance to observe a T_0 such that $|T_0| > 1.96$ is only 1 out of 20. If T_0 computed from the data is too large, say it equals 2.5, or too small, say -3.4, we may start thinking: something must be wrong because it is very unusual for a $N(0,1)$ random variable to take values as extreme as 2.5, or -3.4. So what is wrong? The model could be wrong, the computation could be wrong, the experiment could be poorly conducted. However, if these possibilities can be ruled out, we may then come to the conclusion that maybe the hypothesis H_0: $\mu_1 = \mu_2$ is wrong! The data are not consistent with H_0; the data does not support the null hypothesis H_0: we therefore reject this hypothesis.

Note that we could make a wrong decision in the process. The H_0 is indeed true and T_0 is distributed as $N(0,1)$. It just happened that we observed an extreme value of T_0, i.e. $|T_0| > Z_{\alpha/2}$, we have to follow the rule to reject H_0. The error rate, however, is controlled by α.

The test cannot be used if the population variance σ^2 is unknown, since the value of T_0 cannot be computed from the sample data. In this case the σ^2 will have to be estimated by the polled variance estimator s_p^2. The resulting test is the well-known two-sample t-test. The test statistic is given by

$$T_0 = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{s_p\sqrt{n_1^{-1} + n_2^{-1}}}$$

which has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom if H_0 is true. We reject H_0 if $|T_0| > t_{\alpha/2}(n_1 + n_2 - 2)$.

2. One sided test

It is often the case that the experiment is designed to dispute the claim H_0: $\mu_1 = \mu_2$, in favor of the one sided alternative H_1: $\mu_1 > \mu_2$. For instance, one may wish to claim that certain new treatment is better than the old one. The two sided test can be modified to handle this case. The general decision rule should follow that **evidence which is against H_0 should be in favor of H_1.**

A large negative value of $T_0 = (\bar{y}_{1\cdot} - \bar{y}_{2\cdot})/(s_p\sqrt{n_1^{-1} + n_2^{-1}})$ provides evidence against H_0, but it does not support the alternative H_1: $\mu_1 > \mu_2$. Hence, we reject H_0 only if $T_0 > t_\alpha(n_1 + n_2 - 2)$.

Similarly, to test H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 < \mu_2$, which is a symmetric situation to the foregoing one, we reject H_0 if $T_0 < -t_{1-\alpha}(n_1 + n_2 - 2)$.

Sometimes a test for H_0: $\mu_1 \leq \mu_2$ versus H_1: $\mu_1 > \mu_2$ may be of interest. The test statistic T_0 can also be used in this case. We reject H_0 if $T_0 > t_{1-\alpha}(n_1 + n_2 - 2)$. It should be noted that, under H_0: $\mu_1 \leq \mu_2$, T_0 is NOT distributed as $t(n_1 + n_2 - 2)$. The term $\mu_1 - \mu_2$ does not vanish from T under H_0 which only states $\mu_1 - \mu_2 \leq 0$. We do, however, have

$$\begin{aligned} & P(T_0 > t_{1-\alpha}(n_1 + n_2 - 2)|H_0) \\ & \quad \leq \quad P(T > t_{1-\alpha}(n_1 + n_2 - 2)|\mu_1 = \mu_2) \\ & \quad = \quad \alpha\,, \end{aligned}$$

so the type I error probability is still controlled by α.

3. A test for equal variances

Our previous test assumes $\sigma_1^2 = \sigma_2^2$, where $\sigma_i^2 = Var(y_{ij})$ for $i = 1, 2$. This claim can also be tested before we examine the means. Note that σ_1^2 and σ_2^2 can be estimated by the two sample variances,

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\cdot})^2 \;\text{ and }\; s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2\cdot})^2\,.$$

To test H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$, the ratio of s_1^2 and s_2^2 is used as the test statistic,

$$F_0 = s_1^2/s_2^2\,.$$

Under the normal model (7.1) and if H_0 is true, F_0 is distributed as an $F(n_1 - 1, n_2 - 1)$. We reject H_0 if $F_0 < F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ or $F_0 > F_{\alpha/2}(n_1 - 1, n_2 - 1)$.
Note: Most F distribution tables contain only values for high percentiles. Values for low percentiles can be obtained using $F_{1-\alpha}(n_1, n_2) = 1/F_{\alpha}(n_2, n_1)$.

4. The p-value

We reject H_0 if the test statistic has an extremely large or small observed value when compared to the known distribution of T_0 under H_0. For instance, if $n_1 = n_2 = 5$ and $\alpha = 0.05$, we reject H_0: $\mu_1 = \mu_2$ whenever $|T_0| > t_{0.025}(8) = 2.306$. If we observed $T_0 = 2.400$ or $T_0 = 5.999$, we would reject H_0 at both cases. However, the case of $T_0 = 5.999$ would provide stronger evidence against H_0 than that of $T_0 = 2.400$.

Let T_{obs} be the observed value of the test statistic T_0 computed from the sample data and T be a random variable following the same distribution to which T_0 is compared. The p-value is defined as

$$p = P(T \text{ is more extreme than } T_{obs}) .$$

The smaller the p-value, the stronger the evidence against H_0. The H_0 will have to be rejected whenever the p-value is smaller or equal to α. The concept of "more extreme" is case dependent. For the two sided t-test,

$$p = P(|T| > |T_{obs}|) ,$$

where $T \sim t(n_1 + n_2 - 2)$; for the one sided t-test for H_0: $\mu_1 \leq \mu_2$ versus H_1: $\mu_1 > \mu_2$, the p-value is computed as

$$p = P(T > T_{obs}) ,$$

where $T \sim t(n_1 + n_2 - 2)$.

Example 7.1 An engineer is interested in comparing the tension bond strength of portland cement mortar of a modified formulation to the standard one. The experimenter has collected 10 observations of strength under each of the two formulations. The data is summarized in the following table.

Formulation	n_i	$\bar{y}_{i\cdot}$	s_i^2
Standard	10	16.76	0.100
Modified	10	17.92	0.061

A completely randomized design would choose 10 runs out of the sequence of a total number of 20 runs at random and assign the modified formulation to these runs while the standard formulation is assigned to the remaining 10 runs. Let y_{ij} be the observed strength for the jth runs under formulation i (= 1 or 2). We assume $y_{ij} \sim N(\mu_i, \sigma_i^2)$ and we would like first to test H_0: $\sigma_1^2 = \sigma_2^2$. The observed F statistics is $F_0 = s_1^2/s_2^2 = 1.6393$. This is compared to $F_{0.025}(9,9) = 4.03$. Since $F_0 < 4.03$, we don't have enough evidence against H_0, i.e. $\sigma_1^2 = \sigma_2^2$ is a reasonable assumption (Note: if $F_0 < 1$, we have to compare F_0 to $F_{0.975}(9,9)$!). The primary concern of the experimenter is to see if the modified formulation produces improved strength. We therefore need to test H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 < \mu_2$. The pooled variance estimate is computed as

$$s_p^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2) = 0.0805\,.$$

The observed value of the T statistic is given by

$$T_0 = (16.76 - 17.92)/[\sqrt{0.0805}\sqrt{1/10 + 1/10}] = -9.14\,.$$

Since $T_0 < -t_{0.05}(18) = -1.734$, we reject H_0 in favor of H_1, the modified formulation does improve the strength. The p-value of this test is given by $P[t(18) < -9.14] < 0.0001$.

7.3 Randomization test

The test in the previous section is based on the normal model (7.1). When the number of experimental runs n (sample size) is not large, there is little opportunity for us to verify its validity. In this case, how do we know that our analysis is still valid? There is certainly no definite answer to this. What we really need to examine is the statistical decision procedure we used with the type I error not larger than α.

One strategy of analyzing the data without the normality assumption is to take advantage of the randomization in our design. Suppose $n = 10$ runs are performed in an experiment and $n_1 = n_2 = 5$. Due to randomization, treatment one could have been applied to any 5 of 10 experiment units. If there is no difference between two treatments (as claimed in the null hypothesis), then it really does not matter which 5 y-values are told to be outcomes of the treatment 1.

Let

$$T = \hat{\mu}_1 - \hat{\mu}_2.$$

This statistic can be computed whenever we pick 5 y-values as $y_{11}, \ldots, y_{15}$, and the rest as $y_{21}, \ldots, y_{25}$. The current T_{obs} is just one of the $\binom{10}{5} = 252$ possible outcomes $\{t_1, t_2, \cdots, t_{252}\}$. Under the null hypothesis, the 252 possible T values are equally likely to occur, i.e. $P(T = t_i) = 1/252$, $i = 1, 2, \cdots, 252$. The one you have, T_{obs}, is just an ordinary one. It should not be outstanding.

If, however, it turns out that T_{obs} is one of the largest possible values of T (out of 252 possibilities), it may shed a lot of doubt on the validity of the null hypothesis. Along this line of thinking, we define the p-value to be

$$\text{proportion of the } T \text{ values which are more extreme than } T_{obs}.$$

Once again, the definition of "more extreme than T_{obs}" depends on the null hypothesis you want to test, as discussed in the last section.

If you want to reject H_0: $\mu_1 = \mu_2$ and would simply take the alternative as H_1: $\mu_1 \neq \mu_2$, the more extreme means

$$|T| \geq |T_{obs}|.$$

For the purpose of computing the proportion, when $|T|$ equals $|T_{obs}|$, we count that only as a half.

For example, suppose $n_1 = n_2 = 2$ and T takes $\binom{4}{2} = 6$ possible values as $\{2, 3, -2, 6, -3, -6\}$. Suppose we observe $T_{obs} = 3$, we find that there are $2 + 2 \times 0.5 = 3$ T values are more extreme than T_{obs} in the above definition. Therefore, the proportion (p-value) is $3/6 = 0.50$. If we wish to test H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 > \mu_2$, the p-value of the randomization test is computed as $1.5/6 = 0.25$.

Once more, the randomization adapted in the design of the experiment not only protect us from unwanted influence of unknown factors, it also enable us to analyze the data without strong model assumptions.

More interestingly, the outcome of randomization test is often very close to the outcome of the t-test discussed in the last section. **Hence, when randomization strategy is used in the design, we have not only reduced or eliminated the influence of possible unknown factors, but also justified the use of t-test even if the normality assumption is not entirely appropriate**.

7.4 Comparing k (> 2) treatments: one-way ANOVA

Many single-factor experiments involve more than 2 treatments. Suppose there are k (> 2) treatments. For each treatment i there are n_i independent experiment runs. A design is called balanced if $n_1 = n_2 = \ldots = n_k = n$. For a balanced single factor design the total number of runs is $N = nk$. A completely randomized design would randomly assign k runs to treatment 1, k runs to treatment 2, etc.

A normal model for single factor experiment:

$$y_{ij} = \mu_i + e_{ij}\,, \quad i = 1, 2, \ldots, k\,;\ j = 1, 2, \ldots, n\,, \tag{7.2}$$

where y_{ij} is the jth observation under treatment i, $\mu_i = E(y_{ij})$ are the fixed but unknown treatment means, e_{ij} are the random error component and are assumed iid $N(0, \sigma^2)$. It is natural to estimate μ_i by

$$\hat{\mu}_i = \bar{y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}\,, i = 1, 2, \ldots, k.$$

Our primary interest is to test if the treatment means are all the same, i.e. to test

$$H_0:\ \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{versus} \quad H_1:\ \mu_i \neq \mu_j \ \text{ for some } \ (i, j)\,.$$

The appropriate procedure for testing H_0 is the analysis of variance.

Decomposition of the total sum of squares:

In cluster sampling we have an equality saying that the total variation is the sum of within cluster variation and between cluster variation. A similar decomposition holds here:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = n\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2\,,$$

where $\bar{y}_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1} y_{ij}/(nk)$ is the overall average. This equality is usually restated as

$$SS_{Tot} = SS_{Trt} + SS_{Err}$$

using three terms of Sum of Squares: Total (Tot), Treatment (Trt) and Error (Err). A combined estimator for the variance σ^2 is given by

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})^2/\sum_{i=1}^{k}(n-1) = SS_{Err}/(N-k)\,.$$

If $\mu_1 = \mu_2 = \cdots = \mu_k = \mu$, the estimated treatment means $\bar{y}_{1\cdot}$, $\bar{y}_{2\cdot}$, $\cdots$, $\bar{y}_{k\cdot}$ are iid random variates with mean μ and variance σ^2/n. Another estimator of σ^2 can be computed based on these means,

$$n\sum_{i=1}^{k}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2/(k-1) = SS_{Trt}/(k-1)\,.$$

These two estimators are also called the Mean Squares, denoted by

$$MS_{Err} = SS_{Err}/(N-k) \quad \text{and} \quad MS_{Trt} = SS_{Trt}/(k-1)\,.$$

The two numbers on the denominators, $N-k$ and $k-1$, are the degrees of freedom for the two MSs.

The F test:

The test statistic we use is the ratio of the two estimators for σ^2,

$$F_0 = MS_{Trt}/MS_{Err} = [SS_{Trt}/(k-1)]/[SS_{Err}/(N-k)]\,.$$

Under model (7.2) and if H_0 is true, F_0 is distributed as $F(k-1, N-k)$. When H_0 is false, i.e. the μ_i's are not all equal, the estimated treatment means $\bar{y}_{1\cdot}$, $\cdots$, $\bar{y}_{k\cdot}$ will tend to differ from each other, the SS_{Trt} will be large compared to SS_{Err}, so we reject H_0 if F_0 is too large, i.e. if $F_0 > F_\alpha(k-1, N-k)$. The p-value is computed as

$$p = P[F(k-1, N-k) > F_0]\,.$$

The computational procedures can be summarized using an ANOVA table:

Table 7.2 Analysis of Variance for the F Test

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F_0
Treatment	SS_{Trt}	$k-1$	MS_{Trt}	MS_{Trt}/MS_{Err}
Error	SS_{Err}	$N-k$	MS_{Err}	
Total	SS_{Tot}	$N-1$		

Example 7.2 The cotton percentage in the synthetic fiber is the key factor that affects the tensile strength. An engineer uses five different levels of cotton percentage (15, 20, 25, 30, 35) and obtained five observations of the tensile strength for each level. The total number of observations is 25. The estimated mean tensile strength are $\bar{y}_{1\cdot} = 9.8$, $\bar{y}_{2\cdot} = 15.4$, $\bar{y}_{3\cdot} = 17.6$, $\bar{y}_{4\cdot} = 21.6$, $\bar{y}_{5\cdot} = 10.8$, and the overall mean is $\bar{y}_{\cdot\cdot} = 15.04$. The total sum of squares is $SS_{Tot} = 636.96$.

i) Describe a possible scenario that the design is completely randomized.

ii) Complete an ANOVA table and test if there is a difference among the five mean tensile strengths.

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F_0
Treatment	475.76	4	118.94	$F_0 = 14.76$
Error	161.20	20	8.06	
Total	636.96	24		

Note that $F_{0.01}(4, 20) = 4.43$, the p-value is less than 0.01. There is a clear difference among the mean tensile strengths.

Chapter 8

Randomized Blocks and Two-way Factorial Design

We have seen the important role of randomization in the designed experiment. In general, randomization reduces or eliminates the influence of the factors not considered in the experimenst. It also validates the statistical analysis under the normality assumptions. In some applications, however, there often exist some factors which obviously have significant influence on the outcome, but we are not interested at the moment to investigate their effects. For instance, experimental units often differ dramatically from one to another. The treatment effects measured from the response variable are often overshadowed by the unit variations. Although randomization tends to balace their influence out, it is more appropriate if arrangement can be made to eliminate their influence all together. Randomized blocks design is a powerful tool that can achieve this goal.

8.1 Paired comparison for two treatments

Consider an example where two kinds of materials, A and B, used for boy's shoes are compared. We would like to know which material is more durable. The experimenter recruited 10 boys for the experiment. Each boy wore a special pair of shoes, the sole of one shoe was made with A and the sole of the other with B. Whether the left or the right sole was made with A or B was determined by flipping a coin. The durability data were obtained as follows:

Boy	1	2	3	4	5	6	7	8	9	10
A	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
B	14.0	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

If we blindly apply the analysis techniques that are suitable for the completely randomized designs, we have

$$\bar{y}_A = 10.63, \quad \bar{y}_B = 11.04, \quad s_A^2 = 6.01, \quad s_B^2 = 6.17, \quad s^2 = 6.09;$$

The observed value of the T-statistic is

$$T_{obs} = 0.369$$

and the p-value is 0.72. There is no significant evidence based on this test.

An important feature of this experiment has been omitted: the data are obtained in pairs. If we examine the data more closely, we find that (i) the durability measurements differ greatly from boy to boy; but (ii) if comparing A and B for each of the ten boys, eight have higher measurement from B than from A. If two materials are equal durable, according to the binomial distribution, an outcome as or more extreme like this has probability of only 5.5%. In addition, the two cases when material A lasted longer have smaller differences. This "significant difference" was not detected from the usual T test due to the fact that the difference between boys are so large that the difference between two materials is not large enough to show up.

A randomization test can be used here to test the difference between the two materials. As materials A and B were both wore by the same boy for the same period of time, the **observed difference** of the response variable for each boy should reflect the difference in materials, not in boys. If there were no difference between the two materials, random assignment of A and B to left or right shoes should only have effects on the sign associated with the differences. Tossing 10 coins could produce $2^{10} = 1024$ possible outcomes, and therefore, 1024 possible signed differences. Consequently, there are 1024 possible average of differences. We find that three of them are larger than 0.41, the average difference from the current data, and four give the same value as 0.41. If we split the counts of equal ones, we obtain a significance level of $p = 5/1024 = 0.5\%$. Thus, it is statistically significant that the two materials have different durability.

The T test for paired experiment:

For paired experiments, observations obtained from the different experimental units tend to have different mean values. Let y_{1j} and y_{2j} be the two

observed values of y from the jth unit. A suitable model is as follows,

$$y_{ij} = \mu_i + \beta_j + e_{ij}\,,\ i = 1, 2,\ j = 1, \cdots, n\,,$$

where the β_j represent the effect due to the experimental units (boys in the previous example) and they are not the same. The usual two sample T test which assumes $y_{ij} = \mu_i + e_{ij}$ is no longer valid under current situation. The problem can be solved by working on the difference of the response variables $d_j = y_{2j} - y_{1j}$ which satisfies

$$d_j = \tau + e_j\,, \quad j = 1, 2, \ldots, n\,, \tag{8.1}$$

where $\tau = \mu_2 - \mu_1$ is the mean difference between the two treatments, the e_j's are iid $N(0, \sigma_\tau^2)$.

The two model parameters τ and σ_τ^2 can be estimated by

$$\hat{\tau} = \bar{d} = n^{-1}\sum_{j=1}^{n} d_j \ \text{ and } \ \hat{\sigma_\tau^2} = s_d^2 = (n-1)^{-1}\sum_{j=1}^{n}(d_j - \bar{d})^2\,.$$

The statistical hypothesis is now formulated as H_0: $\tau = 0$ and the alternative is H_1: $\tau \neq 0$ or H_1: $\tau > 0$. It can be shown that under model (8.1),

$$T = \frac{\hat{\tau} - \tau}{s_d/\sqrt{n}}$$

has a t-distribution with $n-1$ degrees of freedom. Under the null hypothesis, the observed value of T is computed as

$$T_{obs} = \frac{\hat{\tau}}{s_d/\sqrt{n}}.$$

For one-sided test against the alternative $\tau > 0$, we calculate the p-value by $P(T > T_{obs})$; for two sided test against the alternative $\tau \neq 0$, we compute the p-value $P(|T| > |T_{obs}|)$, where $T \sim t(n-1)$.

Let us re-analyze the data set from the boys shoes experiment. It is easy to find out that $\bar{d} = 0.41$, $s_d = 0.386$, and

$$T_{obs} = \frac{0.41}{0.386/\sqrt{10}} = 3.4\,.$$

Hence, the one-side test gives us the p-value as $P(t(9) > 3.348877) = 0.0042$; the two side test has p-value 0.0084. There is significant evidence that the two materials are different.

Remark: The p-values obtained using randomization or using t-test are again very close to each other.

Confidence interval for $\tau = \mu_2 - \mu_1$:

Since

$$T = \frac{\hat{\tau} - \tau}{s_d/\sqrt{n}}$$

has a t-distribution, a confidence interval for τ can be easily constructed. Suppose we want a confidence interval with confidence 95% and there are 10 pairs of observations, then the confidence interval would be

$$\bar{d} \pm 2.262 s_d/\sqrt{10}\,.$$

Note that the quantile is $t_{0.975}(9) = 2.262$.

8.2 Randomized blocks design

The paired comparison of previous section is a special case of **blocking** that has important applications in many designed experiments.

Broadly speaking, factors can be categorized into two types: whose with effects of primary interest to the experimenter, and those (blocks) whose effects are desired to be eliminated. In general, blocks are caused by the heterogeneity of the experimental units. When this heterogeneity is considered in the design, it becomes a blocking factor. Within the same block, experimental units are homogeneous, and all treatments are compared within blocks. The between block variability is eliminated by treating blocks as an explicit factor. In the boys shoes example, our primary interest is to see whether the two types of materials have significant difference in durability. The effects of individual boys are obviously large and cannot be ignored, but they are not of any interest to the experimenter. This factor of boys has to be considered and is called blocking factor. The corresponding effect is called block effect.

An example of randomized blocks design:

Suppose in the tomato plant example, four different types of fertilizers were examined, and three types of seeds, denoted by 1, 2, and 3, were used for the experimentation. The reason for this is that, a good fertilizer should work well over a variety of seeds. The factor of fertilizers is of primary interest and has four levels denoted by A, B, C, and D. The seed types are obviously

important for the plant yield and are treated as blocks. The experimenter adopted a randomized blocks design by applying all four types of fertilizers to each seed, and the planting order for each seed is also randomized. The outcomes, plant yields, are obtained as follows.

	A	B	C	D
1	23.8	18.9	23.7	33.4
2	30.2	24.7	25.4	29.2
3	34.5	32.7	29.7	30.9

To limit the effect of earth conditions, these 12 plants should be randomly positioned. For each fertilizer-seed combination, several replicates could be conducted. For the model to be considered here, we will assume that there is only one experimental run for each combination. The other situation will be considered later.

Let y_{ij} be the observed response for fertilizer i and seed j. The statistical model for this design is

$$y_{ij} = \mu + \tau_i + \beta_j + e_{ij}\,, \quad i = 1, 2, \ldots, a \text{ and } j = 1, 2, \ldots, b\,, \tag{8.2}$$

where μ is an overall mean, τ_i is the effect of the ith treatment (fertilizer), β_j is the effect in the jth block (seed), and e_{ij} is the usual random error term and assumed as iid $N(0, \sigma^2)$. There are $a = 4$ levels and $b = 3$ blocks in this example. Since the comparisons are relative, we can assume

$$\sum_{i=1}^{a} \tau_i = 0 \;\text{ and }\; \sum_{j=1}^{b} \beta_j = 0\,.$$

If we let $\mu_{ij} = E(y_{ij})$, it implies that $\mu_{ij} = \mu + \tau_i + \beta_j$. The treatment means are $\mu_{i\cdot} = \sum_{j=1}^{b} \mu_{ij}/b = \mu + \tau_i$; the block means are $\mu_{\cdot j} = \sum_{i=1}^{a} \mu_{ij}/a = \mu + \beta_j$. The τ_i's are therefore termed the treatment effects, and the β_j's are called the block effects.

We are interested in testing the equality of the treatment means. The hypotheses of interest are

$$H_0:\ \mu_{1\cdot} = \cdots = \mu_{a\cdot} \;\text{ versus }\; H_1:\ \mu_{i\cdot} \neq \mu_{j\cdot} \text{ for at least one pair } (i, j).$$

These can also be alternatively expressed as

$$H_0:\ \tau_1 = \cdots = \tau_a = 0 \;\text{ versus }\; H_1:\ \tau_i \neq 0 \text{ for at least one } i.$$

Associated with model (8.2), we may write

$$y_{ij} = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})$$

where

$$\bar{y}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} y_{ij}\,, \quad i = 1, 2, \cdots, a;$$

$$\bar{y}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} y_{ij}\,, \quad j = 1, 2, \cdots, b$$

and

$$\bar{y}_{\cdot\cdot} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}\,.$$

The above decomposition implies that we can estimate μ by $\bar{y}_{\cdot\cdot}$, τ_i by $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$ and β_j by $\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$. The quantity $\hat{e}_{ij} = y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}$ is truly the residual that cannot be explained by various effects.

Note that the experiment was designed in such a way that every block meets every treatment level exactly once. It is easy to see that $\sum_{i=1}^{a} \bar{y}_{i\cdot}/a = \bar{y}_{\cdot\cdot}$ and $\sum_{j=1}^{b} \bar{y}_{\cdot j}/b = \bar{y}_{\cdot\cdot}$. The sum of squares for the treatment,

$$SS_{Trt} = b\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2\,,$$

represents the variations caused by the treatment. The size of SS_{Trt} forms the base for rejecting the hypothesis of no treatment effects.

We could similarly define the block sum of squares

$$SS_{Blk} = a\sum_{j=1}^{b}(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2\,.$$

The size of SS_{Blk} represents the variability due to the block effect. We in general are not concerned about testing the block effect. The goal of randomized blocks design is to remove this effect away and to identify the source of variation due to the treatment effect.

The sum of squares for the residuals represents the remaining sources of variations not due to the treatment effect or the block effect, and is defined as

$$SS_{Err} = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2\,.$$

Finally, the total sum of squares $SS_{Tot} = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2$ can be decomposed as

$$SS_{Tot} = SS_{Trt} + SS_{Blk} + SS_{Err}\,.$$

Again, it is worthwhile to point out that this perfect decomposition is possible fully due to the deliberate arrangement of the design that every level of the blocking factor and every level of treatment factor meets equal number of times in experimental units.

Under model (8.2), it could be shown that SS_{Trt}, SS_{Blk} and SS_{Err} are independent of each other. Further, it can also be shown that if there is no treatment effect, i.e. if H_0 is true,

$$F_0 = MS_{Trt}/MS_{Err} \;\sim\; F[a-1,(a-1)(b-1)]\,,$$

where $MS_{Trt} = SS_{Trt}/(a-1)$ and $MS_{Err} = SS_{Err}/[(a-1)(b-1)]$ are the mean squares. It is important to see a similar decomposition for the degrees of freedom:

$$N - 1 = (a-1) + (b-1) + (a-1)(b-1)\,,$$

where $N = ab$ is the total number of observations. When treatment effect does exist, the value of SS_{Trt} will be large compared to SS_{Err}. We reject H_0 if

$$F_0 > F_\alpha[a-1,(a-1)(b-1)]\,.$$

Computations are summarized in the following analysis of variance table:

Source of variation	Sum of Squares	Degrees of Freedom	Mean Squares	F_0
Treatment	SS_{Trt}	$a-1$	$MS_{Trt} = \frac{SS_{Trt}}{a-1}$	$F_0 = \frac{MS_{Trt}}{MS_{Err}}$
Block	SS_{Blk}	$b-1$	$MS_{Blk} = \frac{SS_{Blk}}{b-1}$	
Error	SS_{Err}	$(a-1)(b-1)$	$MS_{Err} = \frac{SS_{Err}}{(a-1)(b-1)}$	
Total	SS_{Tot}	$N-1$		

This is the so-called two-way ANOVA table. Note that the F distribution has only been tabulated for selected values of α. The exact p-value, $P[F(a-1,(a-1)(b-1) > F_0]$, can be obtained using Splus or R program. One simply types

```
1- pf(F0, a-1, (a-1)*(b-1))
```

to get the actual p-value, where $F0$ is the actual value of F_{obs}. Mathematically one can test the block effect using a similar approach, but this is usually not of interest.

Let us complete the analysis of variance table and test whether the fertilizer effect exists for the data described at the beginning of the section. First, compute

$$\bar{y}_{1\cdot} = 29.50\,, \quad \bar{y}_{2\cdot} = 25.43\,, \quad \bar{y}_{3\cdot} = 26.27\,, \quad \bar{y}_{4\cdot} = 31.17$$

and

$$\bar{y}_{\cdot 1} = 24.95\,, \quad \bar{y}_{\cdot 2} = 27.38\,, \quad \bar{y}_{\cdot 3} = 31.95\,.$$

Then compute $\bar{y}_{\cdot\cdot} = (24.95 + 27.38 + 31.95)/3 = 28.09$, and

$$SS_{Tot} = \sum_{i=1}^{4}\sum_{j=1}^{3} y_{ij}^2 - 12\bar{y}_{\cdot\cdot}^2 = 248.69\,,$$

$$SS_{Trt} = 3[\sum_{i=1}^{4} \bar{y}_{i\cdot}^2 - 4\bar{y}_{\cdot\cdot}^2] = 67.27\,,$$

$$SS_{Blk} = 4[\sum_{j=1}^{3} \bar{y}_{\cdot j}^2 - 3\bar{y}_{\cdot\cdot}^2] = 103.30\,,$$

and finally,

$$SS_{Err} = SS_{Tot} - SS_{Trt} - SS_{Blk} = 78.12\,.$$

The analysis of variance table can now be constructed as follows:

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F_0
Treatment	$SS_{Trt} = 67.27$	3	$MS_{Trt} = 22.42$	$F_0 = 1.722$
Block	$SS_{Blk} = 103.30$	2	$MS_{Blk} = 51.65$	
Error	$SS_{Err} = 78.12$	6	$MS_{Err} = 13.02$	
Total	$SS_{Tot} = 248.69$	11		

Since $F_0 < F_{0.05}(3, 6) = 4.757$, we don't have enough evidence to reject H_0. There are no significant difference among the four types of fertilizers. The exact p-value can be found using Splus as

```
1-pf(1.722,3,6)=0.2613.
```

Confidence intervals for individual effects:

When H_0 is rejected, i.e. the treatment effects do exist, one may wish to estimate the treatment effects τ_i by $\hat{\tau}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$. To construct a 95% confidence interval for τ_i, we need to find the variance of $\hat{\tau}_i$. The following model assumptions are crucial for the validity of this method.

(i) The effects of the block and of the treatment are additive, i.e. $\mu_{ij} = \mu + \tau_i + \beta_j$. This assumption can also invalid in some applications, as can be seen in the next section.

(ii) The variance σ^2 is common for all error terms. This is not always realistic either.

(iii) All observations are independent and normally distributed.

Also note that σ^2 can be estimated by MS_{Err}. Under above assumptions it can be shown that $(\hat{\tau}_i - \tau_i)/SE(\hat{\tau}_i)$ is distributed as $t((a-1)(b-1))$. A t confidence interval can then be constructed.

8.3 Two-way factorial design

The experiments we have discussed so far mainly investigate the effect of a single factor to a response. The tomato plant example investigated the factor of fertilizer; in the boys shoes example, we are interested in the factor of different materials. In randomized blocks design, the blocking factor comes into the picture but our analysis still concentrated on a single factor.

Suppose in an experiment we are interested in the effects of two factors, A and B. We assume factor A has a levels and B has b levels. A (balanced) two-way factorial design proposes to conduct the experiment at each treatment (combination of levels of A and B) with same number of replicates. Both factors are equally important.

A toxic agents example of two-way factorial design:

In an experiment we consider two factors: poison with 3 levels, denoted by I, II and III, and treatment with 4 levels, denoted by A, B, C, and D. The response variable is the survival time. For each treatment such as (I, A), (II, C), (III, B), etc, four replicated experimental runs were conducted. The outcomes are summarized as follows:

	Treatment			
Poison	A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Both factors are of interest. In addition, the experimenter wishes to see if there is an interaction between the two factors. The additive model (8.2) used for randomized blocks design is no longer suitable for this case. The following statistical model is appropriate for this problem:

$$y_{ijk} = \mu + +\tau_i + \beta_j + \gamma_{ij} + e_{ijk}\,, \tag{8.3}$$

where $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$, and $k = 1, 2, \ldots, n$. In the example $a = 3$, $b = 4$, and $n = 4$. The e_{ijk} are the error terms and are assumed as iid $N(0, \sigma^2)$. The total number of observations is abn. The τ_i's are the effect for factor A, the β_j's are the effect for factor B, the γ_{ij} are the interactions. The μ can be viewed as the overall mean. Similar to the randomized blocks design, we can define these parameters such that $\sum_{i=1}^{a} \tau_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$, $\sum_{i=1}^{a} \gamma_{ij} = 0$ for $j = 1, 2, \ldots, b$ and $\sum_{j=1}^{b} \gamma_{ij} = 0$ for $i = 1, 2, \ldots, a$.

The key difference between model (8.2) and model (8.3) is not the number of replicates, n. It is the interaction terms γ_{ij}. The change of treatment means from $\mu_{1\cdot}$ to $\mu_{2\cdot}$ depends not only on the difference between τ_1 and τ_2, but also the level of another factor, j. This is reflected by the interaction terms γ_{ij}. In order to have the capacity of estimating γ_{ij}, it is necessary to have several replicates at each treatment combination. To have equal number of replicates for all treatment combinations will result in a simple statistical analysis and good efficiency in estimation and testing.

Analysis of variance for two-way factorial design:

Let $\mu_{ij} = E(y_{ijk}) = \mu + \tau_i + \beta_j + \gamma_{ij}$ and

$$\bar{y}_{ij\cdot} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}\,.$$

Then $\bar{y}_{ij\cdot}$ is a natural estimator of μ_{ij}. Further, let

$$\bar{y}_{i\cdot\cdot} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}\,,\quad \bar{y}_{\cdot j\cdot} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}\,,\quad \text{and}\quad \bar{y}_{\cdots} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}\,.$$

We have a similar but more sophisticated decomposition:

$$y_{ijk} - \bar{y}_{\cdots} = (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}) + (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}) + (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}) + (y_{ijk} - \bar{y}_{ij\cdot})\,.$$

Due to the perfect balance in the number of replicates for each treatment combinations, we again have a perfect decomposition of the sum of squares:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E\,,$$

where

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{\cdots})^2\,,$$

$$SS_A = bn\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2\,,$$

$$SS_B = an\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2\,,$$

$$SS_{AB} = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2\,,$$

and

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij\cdot})^2\,.$$

One can also compute SS_E from subtraction of other sum of squares from the total sum of squares. The mean squares are defined as the SS divided by the corresponding degrees of freedom. The number of degrees of freedom associated with each sum of squares is

Effect	A	B	AB	Error	Total
Degree of Freedom	$a-1$	$b-1$	$(a-1)(b-1)$	$ab(n-1)$	$abn-1$

The decomposition of degrees of freedom is as follows:

$$abn - 1 = (a-1) + (b-1) + (a-1)(b-1) + ab(n-1)\,.$$

The mean squares for each effect are compared to the mean squares of error. The F statistic for testing the A effect is $F_0 = MS_A/MS_E$, and similarly for the B effect and AB interactions. The analysis of variance table is as follows:

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	SS_A	$a-1$	MS_A	$F_0 = \frac{MS_A}{MS_E}$
B	SS_B	$b-1$	MS_B	$F_0 = \frac{MS_B}{MS_E}$
AB	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	$F_0 = \frac{MS_{AB}}{MS_E}$
Error	SS_E	$ab(n-1)$	MS_E	
Total	SS_T	$abn-1$		

Numerical results for the toxic agents example:

For the data presented earlier, one can complete the ANOVA table for this example as follows (values for the SS and MS are multiplied by 1000):

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A (Poison)	1033.0	2	516.6	$F_0 = 23.2$
B (Treatment)	922.4	3	307.5	$F_0 = 13.8$
AB Interaction	250.1	6	41.7	$F_0 = 1.9$
Error	800.7	36	22.2	
Total	3006.2	47		

The p-value for testing the interactions is $P[F(6,36) > 1.9] = 0.11$. There is no strong evidence that interactions exist. The p-value for testing the poison effect is $P[F(2,36) > 23.2] < 0.001$, the p-value for testing the treatment effect is $P[F(3,36) > 13.8] < 0.001$. We have very strong evidence that both effects present.

Chapter 9

Two-Level Factorial Design

A general factorial design requires independent experimental runs for all possible treatment combinations. When four factors are under investigation and each factor has three levels, a single replicate of all treatments would involve $3 \times 3 \times 3 \times 3 = 81$ runs.

Factorial designs with all factors at two levels are popular in practice for a number of reasons. First, they require relatively few runs. A design with three factors at two levels may have as few as $2^3 = 8$ runs; Second, it is often the case at the early stage of the design that many potential factors are of interest. Choose only two levels for each of these factors and run a relatively small experiment will help to identify the influential factors for further thorough studies with few important factors only; third, the treatment effects estimated from the two level design provide directions and guidance to search for the best treatment settings; and lastly, designs at two levels are relatively simple, easy to analyze, and will shed light on complicated situations. One may also conclude that such designs are most suitable for exploratory investigation.

A complete replicate of a design with k factors all at two levels requires at least $2 \times 2 \times \cdots \times 2 = 2^k$ observations and is called a 2^k factorial design.

9.1 The 2^2 design

Suppose there are two factors, A and B, each has two levels called "low" and "high". There are four treatment combinations that can be represented using one of the following three systems of notation:

Descriptive	(A, B)	Symbolic
A low, B low	(–, –)	(1)
A high, B low	(+, –)	a
A low, B high	(–, +)	b
A high, B high	(+,+)	ab

If there are n replicates for each of the four treatments, the total number of experimental runs is $4n$. Let y_{ijk} be the observed values for the response variable, $i = 1, 2$; $j = 1, 2$; and $k = 1, 2, \ldots, n$. Here $i, j = 1$ represent the "low" level and 2 means the "high" level. Also, we use (1), a, b and ab to represent the total of all n replicates taken at the corresponding treatment combinations.

Example 9.1 A chemical engineer is investigating the effect of the concentration of the reactant (factor A) and the amount of the catalyst (factor B) on the conversion (yield) in a chemical process. she chooses two levels for both factors, and the experiment is replicated three times for each treatment combinations. The data are shown as follows.

	Replicate			
Treatment	I	II	III	Total
(–,–)	28	25	27	(1)=80
(+,–)	36	32	32	a=100
(–,+)	18	19	23	b=60
(+,+)	31	30	29	ab=90

The totals (1), a, b and ab will be conveniently used in estimating the effects of factors and in the construction of an ANOVA table.

The average effect of factor A is defined as

$$\begin{aligned} A &= \bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot} \\ &= \frac{a + ab}{2n} + \frac{(1) + b}{2n} \\ &= \frac{1}{2n}[a + ab - (1) - b]. \end{aligned}$$

The average effect of factor B is defined as

$$\begin{aligned} B &= \bar{y}_{\cdot 2\cdot} - \bar{y}_{\cdot 1\cdot} \\ &= \frac{b + ab}{2n} + \frac{(1) + a}{2n} \\ &= \frac{1}{2n}[b + ab - (1) - a]. \end{aligned}$$

The interaction effect AB is defined as the average difference between the effect of A at the high level of B and the effect of A at the low level of B, i.e.

$$\begin{aligned} AB &= [(\bar{y}_{22\cdot} - \bar{y}_{12\cdot}) - (\bar{y}_{21\cdot} - \bar{y}_{11\cdot})]/2 \\ &= \frac{1}{2n}[(1) + ab - a - b]. \end{aligned}$$

These effects are computed using the so-called contrasts for each of the terms, namely $Contrast(A) = a + ab - (1) - b$, $Contrast(B) = b + ab - (1) - a$, and $Contrast(AB) = (1) + ab - a - b$. These contrasts can be identified easily using an algebraic signs matrix as follows:

	Factorial Effect			
Treatment	I	A	B	AB
(1)	+	−	−	+
a	+	+	−	−
b	+	−	+	−
ab	+	+	+	+

The column I represents the total of the entire experiment, the column AB is obtained by multiplying columns A and B. The contrast for each effect is a linear combination of the treatment totals using plus or minus signs from the corresponding column. Further, these contrasts can also be used to compute the sum of squares for the analysis of variance:

$$\begin{aligned} SS_A &= [a + ab - (1) - b]^2/(4n), \\ SS_B &= [b + ab - (1) - a]^2/(4n), \\ SS_{AB} &= [(1) + ab - a - b]^2/(4n). \end{aligned}$$

The total sum of squares is computed in the usual way

$$SS_T = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n} y_{ijk}^2 - 4n(\bar{y}_{\cdots})^2 .$$

The error sum of squares is obtained by subtraction as

$$SS_E = SS_T - SS_A - SS_B - SS_{AB}.$$

For the data presented in example 9.1, the estimated average effects are

$$\begin{aligned} A &= [90 + 100 - 60 - 80]/(2 \times 3) = 8.33, \\ B &= [90 + 60 - 100 - 80]/(2 \times 3) = -5.00, \\ AB &= [90 + 80 - 100 - 60]/(2 \times 3) = 1.67. \end{aligned}$$

The sum of squares can be computed using $SS_A = nA^2$, $SS_B = nB^2$, and $SS_{AB} = n(AB)^2$. The complete ANOVA table is as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	208.33	1	208.33	$F_0 = 53.15$
B	75.00	1	75.00	$F_0 = 19.13$
AB	8.33	1	8.33	$F_0 = 2.13$
Error	31.34	8	3.92	
Total	323.00	11		

Both main effects are statistically significant (p-value $< 1\%$). The interaction between A and B is not significant (p-value $= 0.183$).

9.2 The 2^3 design

When three factors A, B and C, each at two levels, are considered, there are $2^3 = 8$ treatment combinations. We also need a quadruple index to represent the response: y_{ijkl}, where $i, j, k = 1, 2$ represent the "low" and "high" levels of the three factors, and $l = 1, 2, \ldots, n$ represent the n replicates for each of the treatment combinations. The total number of experimental runs is $8n$. The notation (1), a, b, ab, etc, is extended here to represent the treatment combination as well as the totals for the corresponding treatment, as in the 2^2 design:

A	B	C	Total
−	−	−	(1)
+	−	−	a
−	+	−	b
+	+	−	ab
−	−	+	c
+	−	+	ac
−	+	+	bc
+	+	+	abc

The three main effects for A, B, and C are defined as

$$\begin{aligned} A &= \bar{y}_{2\cdots} - \bar{y}_{1\cdots} \\ &= \frac{1}{4n}[a + ab + ac + abc - (1) - b - c - bc]\,; \end{aligned}$$

$$\begin{aligned} B &= \bar{y}_{\cdot 2\cdot\cdot} - \bar{y}_{\cdot 1\cdot\cdot} \\ &= \frac{1}{4n}[b + ab + bc + abc - (1) - a - c - ac]\,; \\ C &= \bar{y}_{\cdot\cdot 2\cdot} - \bar{y}_{\cdot\cdot 1\cdot} \\ &= \frac{1}{4n}[c + ac + bc + abc - (1) - a - b - ab]\,. \end{aligned}$$

The AB interaction effect is defined as the half difference between the average A effects at the two levels of B (since both levels of C in "B high" and "B low", we use half of this difference), i.e.

$$\begin{aligned} AB &= [(\bar{y}_{22\cdot\cdot} - \bar{y}_{12\cdot\cdot}) - (\bar{y}_{21\cdot\cdot} - \bar{y}_{11\cdot\cdot})]/2 \\ &= \frac{1}{4n}[(1) + c + ab + abc - a - b - bc - ac]\,, \end{aligned}$$

and similarly,

$$\begin{aligned} AC &= \frac{1}{4n}[(1) + b + ac + abc - a - c - ab - bc]\,, \\ BC &= \frac{1}{4n}[(1) + a + bc + abc - b - c - ab - ac]\,. \end{aligned}$$

When three factors are under consideration, there will be a three-way interaction ABC which is defined as the average difference between the AB interaction for the two different levels of factor C, and is computed as

$$ABC = \frac{1}{4n}[a + b + c + abc - (1) - ab - ac - bc].$$

The corresponding contrasts for each of these effects can be computed easily using the following algebraic signs for the 2^3 design:

	Factorial Effect							
Treatment	I	A	B	AB	C	AC	BC	ABC
(1)	+	−	−	+	−	+	+	−
a	+	+	−	−	−	−	+	+
b	+	−	+	−	−	+	−	+
ab	+	+	+	+	−	−	−	−
c	+	−	−	+	+	−	−	+
ac	+	+	−	−	+	+	−	−
bc	+	−	+	−	+	−	+	−
abc	+	+	+	+	+	+	+	+

The columns for the interactions are obtained by multiplying the corresponding columns for the involved factors. For instance, AB = A × B, ABC = A × B × C, etc. The contrast for each effect is a linear combination of the totals through the sign columns.

It can also be shown that the sum of squares for the main effects and interactions can be computed as

$$SS = \frac{(Contrast)^2}{8n}\,.$$

For example,

$$SS_A = \frac{1}{8n}[a + ab + ac + abc - (1) - b - c - bc]^2\,.$$

The total sum of squares is computed as

$$SS_T = \sum\sum\sum\sum y_{ijkl}^2 - 8n(\bar{y}_{....})^2\,,$$

and the error sum of squares is obtained by subtraction:

$$SS_E = SS_T - SS_A - SS_B - SS_C - SS_{AB} - SS_{AC} - SS_{BC} - SS_{ABC}\,.$$

Example 9.2 A soft drink bottler is interested in obtaining more uniform fill heights in the bottles produced by his manufacturing process. Three control variables are considered for the filling process: the percent carbonation (A), the operating pressure in the filler (B), and the bottles produced per minute or the line speed (C). The process engineer chooses two levels for each factor, and conducts two replicates ($n = 2$) for each of the 8 treatment combinations. The data, deviation from the target fill height, are presented in the following table, with sign columns for interactions.

	Factorial Effect								Replicate		
Treatment	I	A	B	AB	C	AC	BC	ABC	I	II	Total
(1)	+	−	−	+	−	+	+	−	–3	–1	(1)=–4
a	+	+	−	−	−	−	+	+	0	1	a=1
b	+	−	+	−	−	+	−	+	–1	0	b=–1
ab	+	+	+	+	−	−	−	−	2	3	ab=5
c	+	−	−	+	+	−	−	+	–1	0	c=–1
ac	+	+	−	−	+	+	−	−	2	1	ac=3
bc	+	−	+	−	+	−	+	−	1	1	bc=2
abc	+	+	+	+	+	+	+	+	6	5	abc=11

The main effects and interactions can be computed using

$$\text{Effect} = (Contrast)/(4n)\,.$$

For instance,

$$\begin{aligned}
A &= \frac{1}{4n}[-(1)+a-b+ab-c+ac-bc+abc] \\
&= \frac{1}{8}[-(-4)+1-(-1)+5-(-1)+3-2+11] \\
&= 3.00, \\
BC &= \frac{1}{4n}[(1)+a-b-ab-c-ac+bc+abc] \\
&= \frac{1}{8}[-4+1-(-1)-5-(-1)-3+2+11] \\
&= 0.50, \\
ABC &= \frac{1}{4n}[-(1)+a+b-ab+c-ac-bc+abc] \\
&= \frac{1}{4n}[-(-4)+1-1-5-1-3-2+11] \\
&= 0.50.
\end{aligned}$$

The sum of squares and analysis of variance are summarized in the following ANOVA table.

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	36.00	1	36.00	$F_0 = 57.60$
B	20.25	1	20.25	$F_0 = 32.40$
C	12.25	1	12.25	$F_0 = 19.60$
AB	2.25	1	2.25	$F_0 = 3.60$
AC	0.25	1	0.25	$F_0 = 0.40$
BC	1.00	1	1.00	$F_0 = 1.60$
ABC	1.00	1	1.00	$F_0 = 1.60$
Error	5.00	8	0.625	
Total	78.00	15		

None of the two-factor interactions or the three-factor interaction is significant at 5% level; all the main effects are significant at the level of 1%.