Biostatistics I

for

Graduate Students in Public Health

---

prepared by
Yu-Fen Li, PhD, MPH
Associate Professor
Institute of Biostatistics
CHINA MEDICAL UNIVERSITY

Fall 2012

Table of Contents

List of Tables

List of Figures

Syllabus

Teacher: Li, Yu-Fen

Office: Room 1509; Tel: 2205-3366 ext 6121

E-mail: yufenli@mail.cmu.edu.tw

Office hours: Wed afternoon

Course Website:

http://mail.cmu.edu.tw/∼yufenli/class/biostat1.htm

Time: Lecture: 3:10 pm - 5:00 pm on Mondays

Outline

- Everyone is expected to attend the lectures
- Weekly quiz will occur during lecture
- A good calculator is highly recommended
- Check the Course Website periodically for updates

Teaching Methods

English handouts; Lectures in English & Chinese

Textbook

Principles of Biostatistics, by Pagano and Gauvreau. (Duxbury, 2nd ED)

Grading

- Quiz 25%
- Exam 75% (25%+25%+25%)

Chapter 1

Probability

In the previous chapters, we studied ways in which descriptive statistics can be used to organize and summarize a set of data. We might also be interested in investigating how the information contained in the sample can be used to infer the characteristics of the population from which it was drawn.

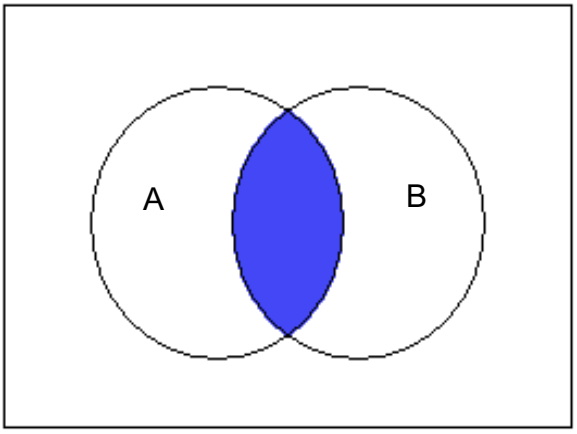
1.1 Operations on Events

An *event* is the basic element to which probability can be applied. A number of different operations can be performed on events. A picture called a *Venn diagram* is a useful device for depicting the relationships among events (see Figure 1.1).

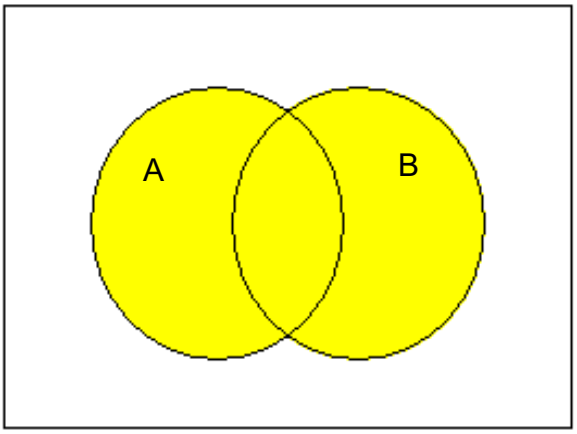
Intersection, Figure 1.1(a): The intersection of two events A and B, denoted as $A \cap B$, is defined as the event "both A and B."

Union, Figure 1.1(b): The union of two events A and B, denoted as $A \cup B$, is defined as the event "either A or B."

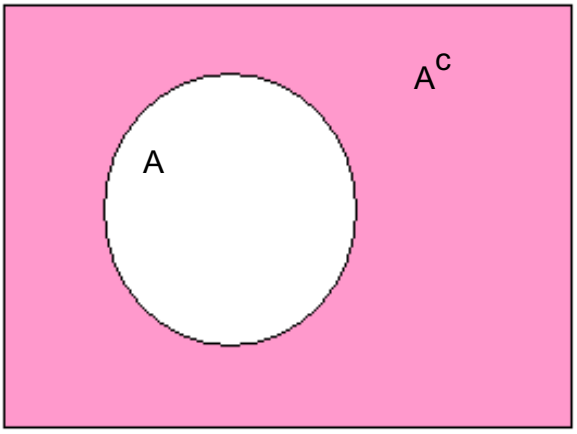
Complement, Figure 1.1(c): The complement of an event A, denoted as A^c or $\overline{A}$, is defined as the event "not A."



(a) Intersection of two events A and B



(b) Union of two events A and B



(c) Complement of an event A

Figure 1.1: Venn diagrams represent the operations on events.

1.2 Probability and Conditional Probability

1.2.1 Probability

Many definitions of probability have been proposed; the one presented here is called the 'frequentist definition.' For example, if an experiment is repeated n times under essentially identical conditions, and if the event A occurs m times, then as n grows large the ratio m/n approaches a fixed limit that is the probability of A, denoted as P(A)=$\frac{m}{n}$.

The numerical value of a probability lies between 0 and 1. We have

1) P(A$\cup$A^c)=1,

2) P(A$\cap$A^c)=P(ϕ)=0,

3) P(A^c)=1-P(A),

4) P(A$\cup$B)=P(A)+P(B)-P(A$\cap$B),

5) P(A$\cap$B)=P(ϕ)=0, if A and B are *mutually exclusive* or *disjoint*;

   - if A and B are mutually exclusive, then P(A$\cup$B)=P(A)+P(B) (the additive rule of probability)

The additive rule can be extended to the cases of three or more mutually exclusive events. If A$_1$, A$_2$, $\cdots$, and A$_n$ are n mutually exclusive events, then

$$\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \cdots + \mathrm{P}(A_n).$$

1.2.2 Conditional Probability

We are often interested in determining the probability that an event B will occur given that we already know the outcome of another event A. In this case, we are dealing with a *conditional probability*.

The *multiplicative rule of probability* states that the probability that two events A and B will both occur is equal to the probability of A multiplied by the probability of B given that A has already occurred. This can be express as

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B|A) = \mathrm{P}(B)\mathrm{P}(A|B).$$

Dividing both sides of the first equation by P(A), we have the formula for a conditional probability to be

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}, \text{ given } \mathrm{P(A)} \neq 0.$$

Similarly, we have

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}, \text{ given } \mathrm{P(B)} \neq 0.$$

Independence Two events are said to be *independent*, if the outcome of one event has no effect on the occurrence of the other. If A and B are independent events,

$$\mathrm{P}(A|B) = \mathrm{P}(A) \text{ and } \mathrm{P}(B|A) = \mathrm{P}(B).$$

In this special case of independence, the multiplicative rule of probability may be written as

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B).$$

NOTE: the terms 'independent' and 'mutually exclusive' do not mean the same thing. If A and B are independent and event A occurs, the outcome of B is not affected, i.e. $\mathrm{P}(B|A) = \mathrm{P}(B)$. If A and B are mutually exclusive and event A occurs, then event B cannot occur, i.e. $\mathrm{P}(B|A) = 0$.

1.3 Bayes' Theorem

If A_1, A_2, $\cdots$, and A_n are n *mutually exclusive* and *exhaustive* events such that

$$\begin{aligned} \mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_n) &= \mathrm{P}(A_1) + \mathrm{P}(A_2) + \cdots + \mathrm{P}(A_n) \\ &= 1. \end{aligned}$$

Bayes' theorem states that

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(A_i)\mathrm{P}(B|A_i)}{\mathrm{P}(A_1)\mathrm{P}(B|A_1) + \cdots + \mathrm{P}(A_n)\mathrm{P}(B|A_n)}$$

for each i, $1 \leq i \leq n$.

For example, the 163157 persons in the National Health Interview Survey of 1980-1981 (S) were subdivided into three mutually exclusive categories: the current employed (E_1), the currently unemployed (E_2), and those not in the labor force (E_3):

Employment Status	**Population**	**Impairments**
Current employed (E_1)	98917	552
Current unemployed (E_2)	7462	27
Not in the labor force (E_3)	56778	368
Total	163157	947

If we assume that these numbers are large enough to satisfy the frequentist definition of probability, then, from data data provided, we find that

$$\mathrm{P}(E_1) = \frac{98917}{163157} = 0.6063,$$

$$\mathrm{P}(E_2) = \frac{7462}{163157} = 0.0457,$$

and

$$\mathrm{P}(E_3) = \frac{56778}{163157} = 0.3480.$$

If S is the event that an individual in the study is currently employed or currently unemployed or not in the labor force, i.e. $S = E_1 \cup E_2 \cup E_3$. Since there three categories are mutually exclusive, the additive rule of probability may be applied:

$$\begin{aligned} \mathrm{P}(S) &= \mathrm{P}(E_1 \cup E_2 \cup E_3) \\ &= \mathrm{P}(E_1) + \mathrm{P}(E_2) + \mathrm{P}(E_3) \\ &= 0.6063 + 0.0457 + 0.3480 \\ &= 1.0000. \end{aligned}$$

When the probabilities of mutually exclusive events sum to 1, the events are said to be *exhaustive*; in this case, there are no other possible outcomes.

Let H be the event that an individual has a hearing impairment due to injury. Overall,

$$\mathrm{P}(H) = \frac{947}{163157} = 0.0058.$$

In fact, H may be expressed as the union of three exclusive events:

$$H = (E_1 \cap H) \cup (E_2 \cap H) \cup (E_3 \cap H).$$

Therefore,

$$\begin{aligned} \mathrm{P}(H) &= \mathrm{P}[(E_1 \cap H) \cup (E_2 \cap H) \cup (E_3 \cap H)] \\ &= \mathrm{P}(E_1 \cap H) + \mathrm{P}(E_2 \cap H) + \mathrm{P}(E_3 \cap H) \\ &= \mathrm{P}(E_1)\mathrm{P}(H|E_1) + \mathrm{P}(E_2)\mathrm{P}(H|E_2) + \mathrm{P}(E_3)\mathrm{P}(H|E_3). \end{aligned}$$

This is sometimes called the *law of total probability*.

Looking at each employment status subgroup separately, we have

$$\mathrm{P}(H|E_1) = \frac{552}{98917} = 0.0056,$$

$$\mathrm{P}(H|E_2) = \frac{27}{7462} = 0.0036,$$

and

$$\mathrm{P}(H|E_3) = \frac{368}{56778} = 0.0065.$$

Applying Bayes' theorem to this case, we find

$$\begin{aligned}
\mathrm{P}(E_1|H) &= \frac{\mathrm{P}(E_1)\mathrm{P}(H|E_1)}{\mathrm{P}(H)} \\
&= \frac{\mathrm{P}(E_1)\mathrm{P}(H|E_1)}{\mathrm{P}(E_1)\mathrm{P}(H|E_1) + \mathrm{P}(E_2)\mathrm{P}(H|E_2) + \mathrm{P}(E_3)\mathrm{P}(H|E_3)} \\
&= \frac{0.6063 \times 0.0056}{0.6063 \times 0.0056 + 0.0457 \times 0.0036 + 0.3480 \times 0.0065} \\
&= 0.5832 \\
&\approx \frac{552}{947}.
\end{aligned}$$

1.4 Diagnostic Tests

Bayes' theorem is often employed in issues of diagnostic testing or screening. Those who test positive are considered to be more likely to have the disease and are usually subjected either further diagnostic procedures or to treatment. Bayes' theorem allows us to use probability to evaluate the associated uncertainties.

1.4.1 Sensitivity and Specificity

If we consider a 2×2 table of a screening test (positive,T+/negative, T-) and the real disease status (yes, D+ / no, D-) as shown in Figure 1.2, then the sensitivity of a test is the probability of a positive test result given that the individual has the disease, i.e.

$$\text{Sensitivity (SE)} = \mathrm{P}(T+|D+) = \frac{a}{a+c} \equiv 1 \text{ - (False Negative)},$$

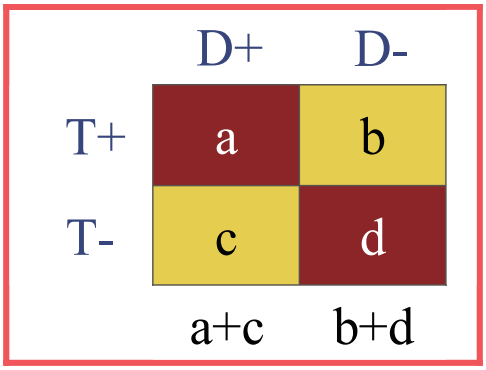


Figure 1.2: A 2×2 table of a screening test (positive,T+/negative, T-) and the real disease status (yes, D+/no, D-).

and the specificity of a test is the probability of a negative test result given that the individual does not have the disease, i.e.

$$\text{Specificity (SP)} = \mathrm{P}(T-|D-) = \frac{d}{b+d} \equiv 1 \text{ - (False Positive)}.$$

1.4.2 Application of Bayes' Theorem

The most important question for the health care professionals is "What is the probability that a person with a positive screening test actually does have the disease?" In other words, we want to compute the probability $\mathrm{P}(D+|T+)$, which is called the *positive predictive value* (PPV). Using Bayes' theorem, we can write

$$\begin{aligned} \mathrm{P}(D+|T+) &= \frac{\mathrm{P}(D+ \cap T+)}{\mathrm{P}(T+)} \\ &= \frac{\mathrm{P}(D+)\mathrm{P}(T+|D+)}{\mathrm{P}(D+)\mathrm{P}(T+|D+) + \mathrm{P}(D-)\mathrm{P}(T+|D-)}. \end{aligned}$$

$\mathrm{P}(D+)$ is the *prevalence* of the disease.

Bayes' theorem can also calculate the *negative predictive value* (NPV), i.e. the probability of no disease given a negative test results,

$$\begin{aligned} \mathrm{P}(D-|T-) &= \frac{\mathrm{P}(D-\cap T-)}{\mathrm{P}(T-)} \\ &= \frac{\mathrm{P}(D-)\mathrm{P}(T-|D-)}{\mathrm{P}(D-)\mathrm{P}(T-|D-)+\mathrm{P}(D+)\mathrm{P}(T-|D+)}. \end{aligned}$$

Unlike sensitivity and specificity, the PPV and NPV are dependent on the population being tested and are influenced by the prevalence of the disease. PPV is directly proportional to the prevalence of the disease. If the group of people tested had included a higher proportion of people with the disease, then the PPV would probably come out higher and the NPV lower.

1.4.3 ROC Curves

In theory, it is desirable to have a test that is both highly sensitive and highly specific. In reality, however, such a procedure is usually not possible. Many tests are actually based on a clinical measurement that can assume a range of values; in this case, there is an inherent trade-off between sensitivity and specificity. Table 1.1 displays data from a kidney transplant program. The level serum creatinine was used as a diagnostic tool for detecting potential transplant rejection. An increased creatinine level is often associated with subsequent organ failure. For example, if we use a level greater than 2.9 mg % as an indicator of imminent rejection, the test has a sensitivity of 0.303 and a specificity of 0.909. To increase the sensitivity, we could lower the cutoff point, say 1.2 mg %, then the sensitivity is 0.939 but the specificity becomes just 0.123. In general, a sensitive test is more

useful when the failure to detect a disease as early as possible has dangerous consequences; a specific test is important in situation where a false positive results is harmful.

The relationship between sensitivity and specificity can be illustrated using a graph known as *receiver operator characteristic (ROC) curve.* An ROC curve is a line graph that plots the probability of a true positive results (or the sensitivity of the test) against a false positive result for a range of different cutoff points. Figure 1.3 displays an ROC curve for the data shown in Table 1.1. When an existing diagnostic test is being evaluated, this type of graph can be used to help assess the usefulness of the test and to determine the most appropriate cutoff point.

The dashed line in Figure 1.3 corresponds to a test that gives positive and negative results by chance alone; such a test has no inherent value. The closer the line to the upper left-hand corner of the graph, the more accurate the test. Furthermore, the point that lies closest to this corner is usually chosen as the cutoff that maximizes both sensitivity and specificity simultaneously.

The ROC curve gives a good visual summary of the performance of model, but it is difficult to compare two ROC curves visually. Accuracy is measured by the area under the ROC curve (AUC). An area of 1 represents a perfect test; an area of .5 represents a worthless test.

Described by Youden in 1950, the J statistic (as known as Youden's index) is a summary index of validity than combines sensitivity and specificity:

$$J = sensitivity + specificity - 1.$$

Table 1.1: Sensitivity and specificity of serum creatinine level for predicting transplant rejection

Serum Creatinine (mg %)	Sensitivity	Specificity
1.20	0.939	0.123
1.30	0.939	0.203
1.40	0.909	0.281
1.50	0.818	0.380
1.60	0.758	0.461
1.70	0.727	0.535
1.80	0.636	0.649
1.90	0.636	0.711
2.00	0.545	0.766
2.10	0.485	0.773
2.20	0.485	0.803
2.30	0.394	0.811
2.40	0.394	0.843
2.50	0.364	0.870
2.60	0.333	0.891
2.70	0.333	0.894
2.80	0.333	0.896
2.90	0.303	0.909

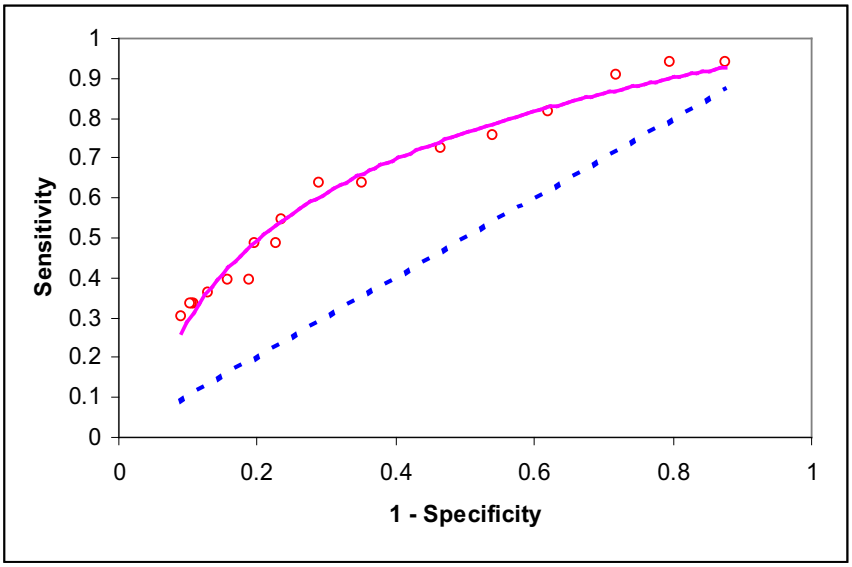


Figure 1.3: An example of ROC curves.

The value 1.0 is subtracted from the sum of sensitivity and specificity so that the maximum of the index becomes 1 when there is perfect agreement. Note that this index gives the same weight to sensitivity and specificity, thus assuming that both sensitivity and specificity are equally important component of validity.

A final note of historical interest You may be wondering where the name "Receiver Operating Characteristic" came from. ROC analysis is part of a field called "Signal Detection Theory" developed during World War II for the analysis of radar images. Radar operators had to decide whether a blip on the screen represented an enemy target, a friendly ship, or just noise. Signal detection theory measures the ability of radar receiver operators to make these important distinctions. Their ability to do so was called the Receiver Operating Characteristics. It was not until the 1970's that signal detection theory was recognized as useful for interpreting medical test results.

1.4.4 Likelihood Ratio: summarizing information about the diagnostic test ✠

One can summarize information about the diagnostic test itself using a measure called the likelihood ratio. The likelihood ratio combines information about the sensitivity and specificity. It tells you how much a positive or negative result changes the likelihood that a patient would have the disease.

The likelihood ratio of a positive test result (LR+) is sensitivity divided by 1- specificity,i.e.

$$LR+ = sensitivity/(1 - specificity)$$

The likelihood ratio of a negative test result (LR-) is 1- sensitivity divided by specificity, i.e.

$$LR- = (1 - sensitivity)/specificity$$

Once you have specified the pre-test odds, you multiply them by the likelihood ratio. This gives you the post-test odds, i.e.

$$odds(post) = odds(pre) * liklihoodratio$$

The post-test odds represent the chances that your patient has a disease. It incorporates information about the disease prevalence, the patient pool, and specific patient risk factors (pre-test odds) and information about the diagnostic test itself (the likelihood ratio).

1.5 Mathematical Expectation

Let X be a continuous random variable (rv) having a p.d.f. $f(x)$, and let $u(X)$ be a function of X. The mathematical expectation (or expected value) of $u(X)$ denoted by $\mathrm{E}[u(X)]$ is

$$\mathrm{E}[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)dx.$$

Some useful facts about mathematical expectations:

1) $\mathrm{E}(k) = k$, if k is a constant,

2) $\mathrm{E}[au(X) + c] = a\mathrm{E}[u(X)] + c$, if a and c are constants,

3) $\mathrm{E}[au_1(X) + bu_2(X)] = a\mathrm{E}[u_1(X)] + b\mathrm{E}[u_2(X)]$, if a and b are constants.

Some special mathematical expectations:

1) $\mathrm{E}(X) \equiv \mu_X$,

2) $\mathrm{var}(X) \equiv \mathrm{E}[X - \mathrm{E}(X)]^2 \stackrel{?}{=} \mathrm{E}(X^2) - \mu_X^2$,

3) $\mathrm{cov}(X, Y) \equiv \mathrm{E}\{[X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]\} \stackrel{?}{=} \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$,

4) $corr(X, Y) \equiv \frac{\mathrm{E}\{[X-\mathrm{E}(X)][Y-\mathrm{E}(Y)]\}}{\sqrt{\mathrm{E}[X-\mathrm{E}(X)]^2}\sqrt{\mathrm{E}[Y-\mathrm{E}(Y)]^2}} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(X)}}$.

Some properties about variance:

1) $\mathrm{var}(kX) = k^2\mathrm{var}(X)$, if k is a constant,

2) $\text{var}(aX + c) = a^2\text{var}(X)$, if a and c are constants,

3) $\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y) \pm 2\text{cov}(X, Y)$.

Chapter 2

Theoretical Probability Distributions

Let's start with the definition of a 'variable' and a 'random variable.' Any characteristic that can be measured or categorized is called a *variable*. If a variable can assume a number of different values such that any particular outcome is determined by chance, it is a *random variable* (r.v.). Random variables are typically represented by uppercase letters such as X, Y, and Z.

A *discrete random variable* can assume only a finite or countable number of outcomes. For example, someone's marital status and the number of ear infections an infant develops during his or her first year of life. A *continuous random variable* can take on any value within a specified internal or continuum, e.g. weight or height.

2.1 Probability Distributions

Every random variable has a corresponding probability distribution. A *probability distribution* applies the theory of probability to describe the behavior of the random variable. For a discrete r.v., its probability distribution specifies all possible outcomes of this r.v. along with the probability that each will occur. For a

continuous r.v., its probability distribution allows us to determine the probability associated with specified ranges of values.

For instance, let X be the birth order of each child born to a woman residing in the U.S. (so X is a discrete r.v.). To construct a probability distribution for X, we list each of the values x that the r.v. can assume, along with the probability (i.e. $\mathrm{P}(X = x)$) for each one as shown in Table 2.1. Note that we use an *uppercase* X to denote the r.v. and a *lowercase* x to represent the outcome of a particular child.

Table 2.1 resembles the frequency distribution introduced in Chapter 2. The probabilities represents the relative frequency of occurrence of each outcome x in a large number of trials repeated under essentially identical conditions; equivalently, they can be thought of as the relative frequencies associated with an infinitely large sample. Since all possible values of the r.v. are taken into account, the outcomes are exhaustive; therefore, the sum of their probabilities must be 1.

In many cases, we can display a probability distribution by means of either a graph or a mathematical formula. For example, Figure 2.1 is a histogram of the probability distribution shown in Table 2.1. The total area of the histogram is equal to 1.

The average value assumed by a r.v. is known as the *population mean*; the dispersion of the values relative to this mean is the *population variance*. Furthermore, the square root of the population variance is the *population standard deviation*.

Probabilities that are calculated from a finite amount of data are called *empirical probabilities*. For instance, the probability distribution of the birth order

Table 2.1: Probability distribution of a random variable X representing the birth order of children born in the U.S.

x	$\mathrm{P}(X = x)$
1	0.416
2	0.330
3	0.158
4	0.058
5	0.021
6	0.009
7	0.004
8+	0.004
Total	1.000

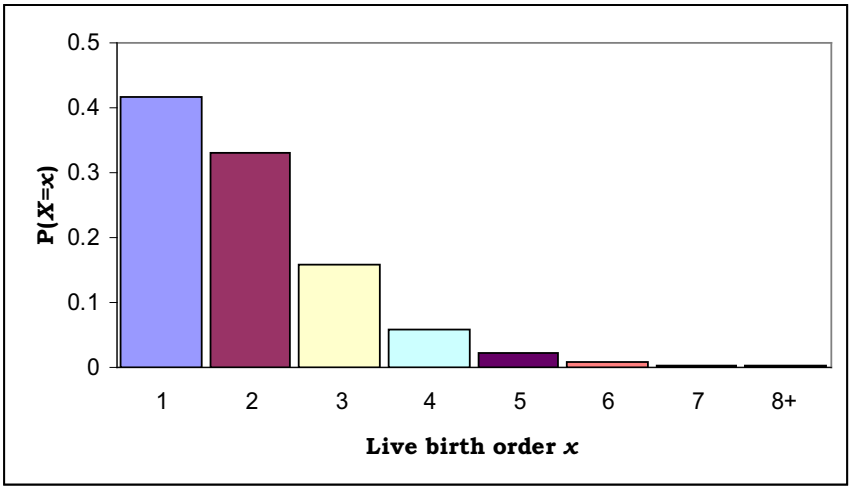


Figure 2.1: Probability distribution of a random variable representing the birth order of children born in the U.S.

of children born in the U.S. was generated based on the experience of the U.S. population in 1986. However, the probability distribution of many other variables of interest can be determined based on theoretical consideration, which is known as a *theoretical probability distribution.*

2.2 The Binomial Distribution

Consider a dichotomous r.v. Y ($Y = 0$ or 1), e.g. life and death, male and female, or sickness and health. For simplicity, they are often referred to as "failure" and "success". A r.v. of this type is known as a *Bernoulli random variable.* If $\mathrm{P}(Y = 1) = p$, then we have $\mathrm{P}(Y = 0) = 1 - p$.

If we have a sequence of n independent Bernoulli trials (or n independent outcomes of the Bernoulli r.v. Y) each with a probability of "success" p, then the total number of successes X is a *binomial random variable* and the probability distribution of the discrete r.v. X follows a *binomial distribution* with *parameters* n and p. Parameters are numerical quantities that summarize the characteristics of a probability distribution. The binomial distribution involves three assumptions:

1) There are fixed number of trials n, each of which results in one of two mutually exclusive outcomes.

2) The outcomes of the n trials are independent.

3) The probability of success p is constant for each trial.

The probability that a r.v. X with binomial distribution $B(n, p)$ is equal to the value x, where $x = 0, 1, \cdots, n$, is given by

$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

is known as the binomial coefficient, and $n! = n \times (n-1) \times \cdots \times 2 \times 1$ (Note: $0! = 1$). The mean value of X is equal to $\mathrm{E}(X) = np$ and the variance of X is $\mathrm{var}(X) = np(1-p)$.

Rather than perform the calculations by hand, we can use Table A.1 in Appendix A to obtain the binomial probabilities for selected values of n and p. This allows us to approximate the probability because it is not necessary to locate the probability of success p on the table.

Figure 2.2 is the graphs of the probability distribution of X with binomial distributions $B(10, p)$ where $p = 0.25$, 0.75, and 0.50. The distribution is skewed to the right when $p < 0.5$ (Figure 2.2(a)) and skewed to the left when $p > 0.5$ (Figure 2.2(b)). If $p = 0.50$ (Figure 2.2(c)), the probability distribution is symmetric.

2.3 The Poisson Distribution

When n becomes large, the combination of n objects taken x at a time, $\frac{n!}{x!(n-x)!}$, is very tedious to evaluate. When n is very large and p is very small, the binomial

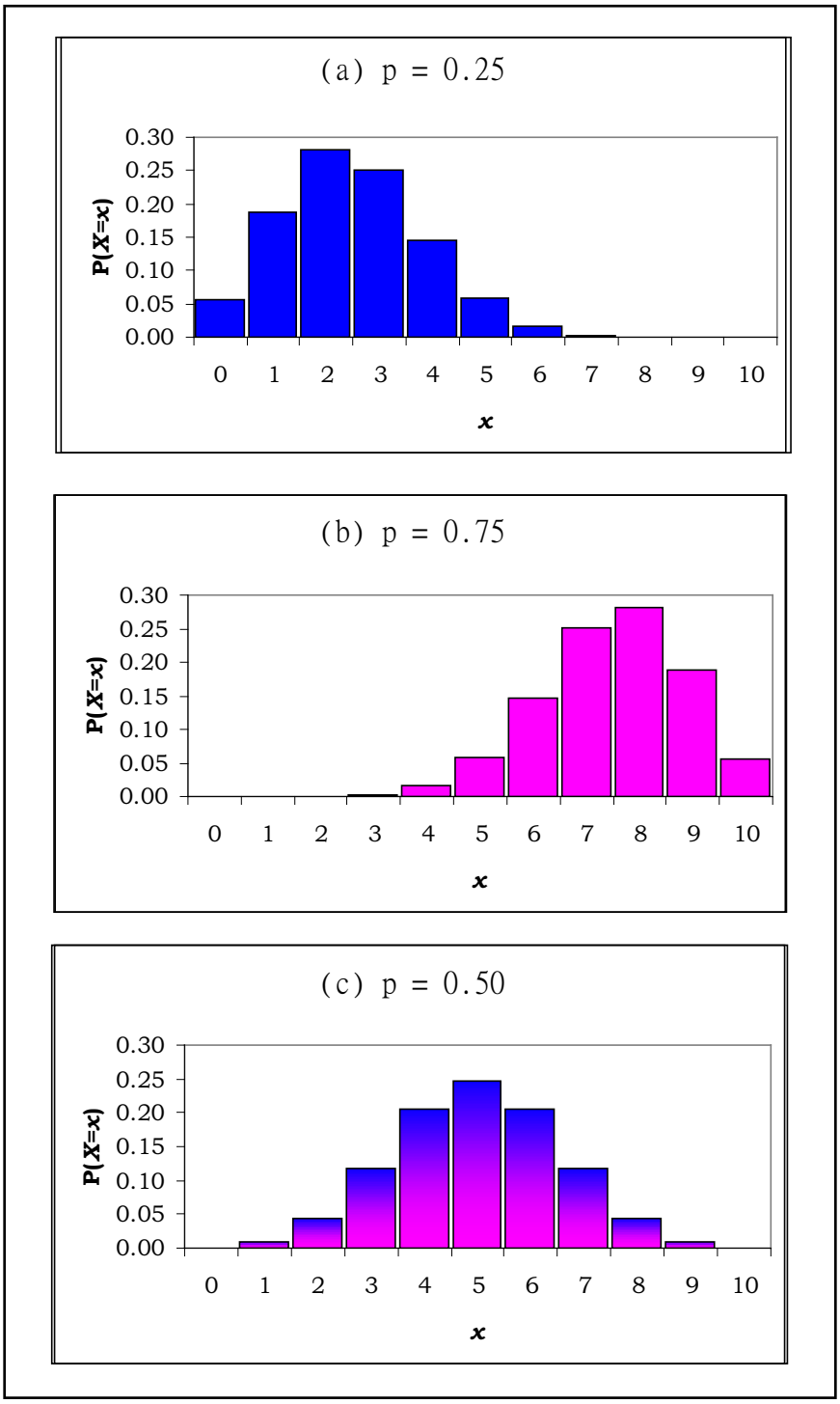


Figure 2.2: Probability distribution of a binomial random variable for which $n = 10$ and (a) $p = 0.25$, (b) $p = 0.75$, and (c) $p = 0.50$

distribution is well approximated by another theoretical probability distribution, called the Poisson distribution. The *Poisson distribution* is used to model discrete events that occur infrequently in time or space; hence, it is sometimes called the *distribution of rare events.*

Let λ be a constant that denotes the average number of occurrences of the event in an interval. If the probability that X assumes the value x is

$$\mathrm{P}(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0, 1, \cdots$$

X is said to have a Poisson distribution with parameter λ. The Poisson distribution involves a set of underlying assumptions:

1) The probability that a single event occurs within an interval is proportional to the length of the interval.

2) Within a single interval, an infinite number of occurrences of the event are theoretically possible. We are not restricted to a fixed number of trials.

3) The events occur independently both within the same interval and between consecutive intervals.

Recall that the mean of a binomial r.v. is equal to np and that its variance is $np(1-p)$. When p is very smell, $1-p$ is close to 1 and $np(1-p)$ is approximately equal to np. In this case, the mean and the variance of the distribution are identical and can be represented by the single parameter $\lambda = np$. The property that the mean is equal to the variance is an identifying characteristic of the Poisson distribution.

Instead of performing the calculation by hand, we can use Table A.2 in Appendix A to obtain Poisson probabilities for selected values of λ. The Poisson distribution is highly skewed for small values of λ; as λ increases, the distribution becomes more symmetric (see Figure 7.6 in textbook). If $X \sim \mathcal{P}(2.5)$ and we want to find $\mathrm{P}(X \geq 7)$, then, by means of the Table A.2, we get

$$\begin{aligned}
\mathrm{P}(X \geq 7) &= 1 - \mathrm{P}(X < 7) \\
&= 1 - [\mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2) + \mathrm{P}(X = 3) + \mathrm{P}(X = 4) \\
&\qquad + \mathrm{P}(X = 5) + \mathrm{P}(X = 6)] \\
&= 1 - (0.0821 + 0.2052 + 0.2565 + .0.2138 + 0.1336 + 0.0668 + 0.0278) \\
&= 1 - 0.9858 \\
&= 0.00242
\end{aligned}$$

2.4 The Normal Distribution

When a r.v. X follows either a binomial or a Poisson distribution, it is restricted to taking on integer values only. However, some outcomes of a random variable may not be limited to integers or counts. A smooth curve is used to represent the probability distribution of a continuous r.v.; the curve is called a *probability density*.

The most common continuous distribution is the *normal distribution*, also known as (aka) the *Gaussian distribution* or the *bell-shaped curve*. Its probability density function (p.d.f.) is given by the equation

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},$$

where $-\infty < x < \infty$, μ is the mean of X, and σ is its standard deviation. The normal curve is unimodal and symmetric about its mean (i.e. mean=median=mode in this special case). The two parameters μ and σ completely define a normal curve. We write $X \sim \mathcal{N}(\mu, \sigma^2)$.

Since a normal distribution could have an infinite number of possible values for its mean and standard deviation, it is impossible to tabulate the area associated with each and every normal curve. Instead, only a single curve is tabulated which is known as the *standard normal distribution* for which $\mu = 0$ and $\sigma = 1$ (Figure 2.3), and we write it as $\mathcal{N}(0, 1)$. Table A.3 in Appendix A displays the areas in the 'upper tail' of the distribution, i.e. $\mathrm{P}(Z > z)$. The area in Figure 2.3 represents $\mathrm{P}(-1 \leq Z \leq 1)$:

$$\begin{aligned}
\mathrm{P}(-1 \leq Z \leq 1) &= 1 - [\mathrm{P}(Z > 1) + \mathrm{P}(Z < -1)] \\
&= 1 - (0.159 + 0.159) \\
&= 1 - 0.318 \\
&= 0.682.
\end{aligned}$$

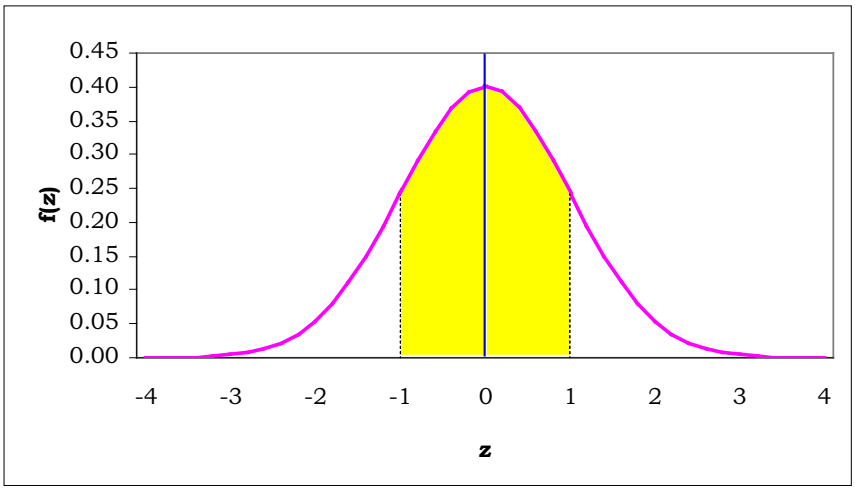


Figure 2.3: The standard normal curve for which μ=0 and σ=1, $Z \sim \mathcal{N}(0,1)$. Area between z=-1.0 and z=1.0: $\mathrm{P}(-1 \leq Z \leq 1) = 0.682$

Therefore, for the standard normal distribution, approximately 68.2% of the area beneath the curve lies within $\pm$ 1 standard deviation from the mean (note: recall what the empirical rule is).

In general, for any arbitrary normal r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$,

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

has a standard normal distribution. As shown in Figure 2.4, the standardization is done by shifting the mean to 0 and scaling the variance to 1. By transforming X into Z, we can use a table of areas computed for the standard normal curve to estimate probabilities associated with X. An outcome of the r.v. Z, denoted

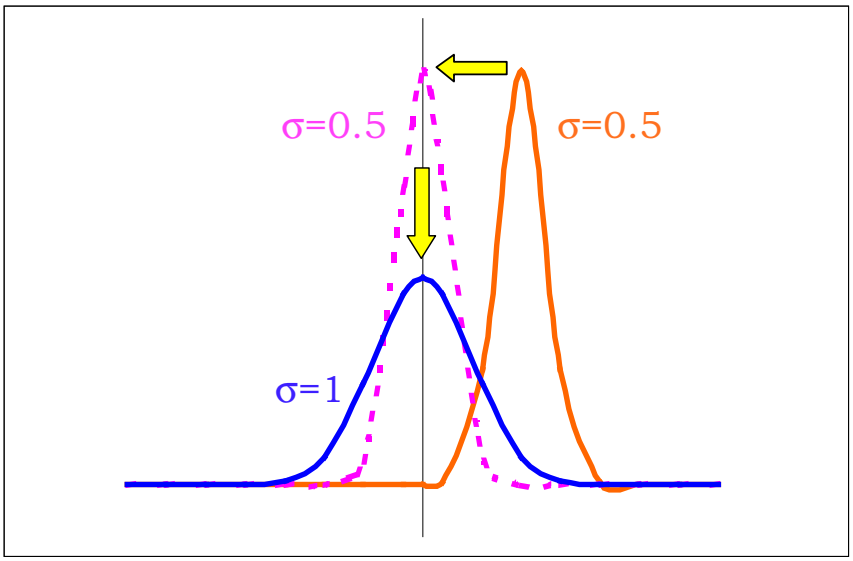


Figure 2.4: The transformation of $\mathcal{N}(2, 0.5^2)$ (red line) to the standard normal distribution (blue line) by shifting the mean to 0 (pink dashed line) and scaling the variance to 1.

z, is known as a *standard normal deviate* or a *z-score*. For example, the men's systolic blood pressure in one population $X \sim \mathcal{N}(129, 19.8^2)$, then

$$Z = \frac{X - 129}{19.8} \sim \mathcal{N}(0, 1).$$

Suppose we wish to find the value of x such that $\mathrm{P}(X > x) = 0.025$. From Table A.3, we find $\mathrm{P}(Z > 1.96) = 0.025$. To obtain the value of x that corresponds to this value of z, we solve the equation

$$z = 1.96 = \frac{x - 129}{19.8}$$

or

$$x = 129 + 1.96 \times 19.8 = 167.8.$$

Therefore, approximately 2.5% of the men in this population have systolic blood pressures greater than 167.8 mm Hg, while 97.5% have systolic blood pressures less than 167.8 mm Hg.

We might also be interested in determining $\mathrm{P}(X > 150)$.

$$\begin{aligned} \mathrm{P}(X > 150) &= \mathrm{P}(\frac{X-\mu}{\sigma} > \frac{150-129}{19.8}) \\ &= \mathrm{P}(Z > 1.06) \\ &= 0.145. \end{aligned}$$

Therefore, approximately 14.5% of the men in this population have systolic blood pressures greater than 150 mm Hg.

False negatives vs. False positives Now consider the more complicated situation in which we have two normally distributed r.v.'s shown in Figure 2.5. For the population of men who are not taking medication, diastolic blood pressure (X_n) follows $\mathcal{N}(80.7, 9.2^2)$. For the man who are using antihypertensive drugs, diastolic blood pressure (X_a) follows $\mathcal{N}(94.9, 11.5^2)$. Suppose we want to identify 90% of the individuals who are currently taking medication. It is equivalent to find the value of diastolic blood pressure that marks off the lower 10% of this population. Using Table A.3, we have that $z = 1.28$ cuts off an area of 0.10 in the upper tail of the standard normal curve, which implies that $z = -1.28$ cuts off an area of 0.10 in the lower tail of the standard normal curve as well. Therefore, we have

$$z = -1.28 = \frac{x - 94.9}{11.5}$$

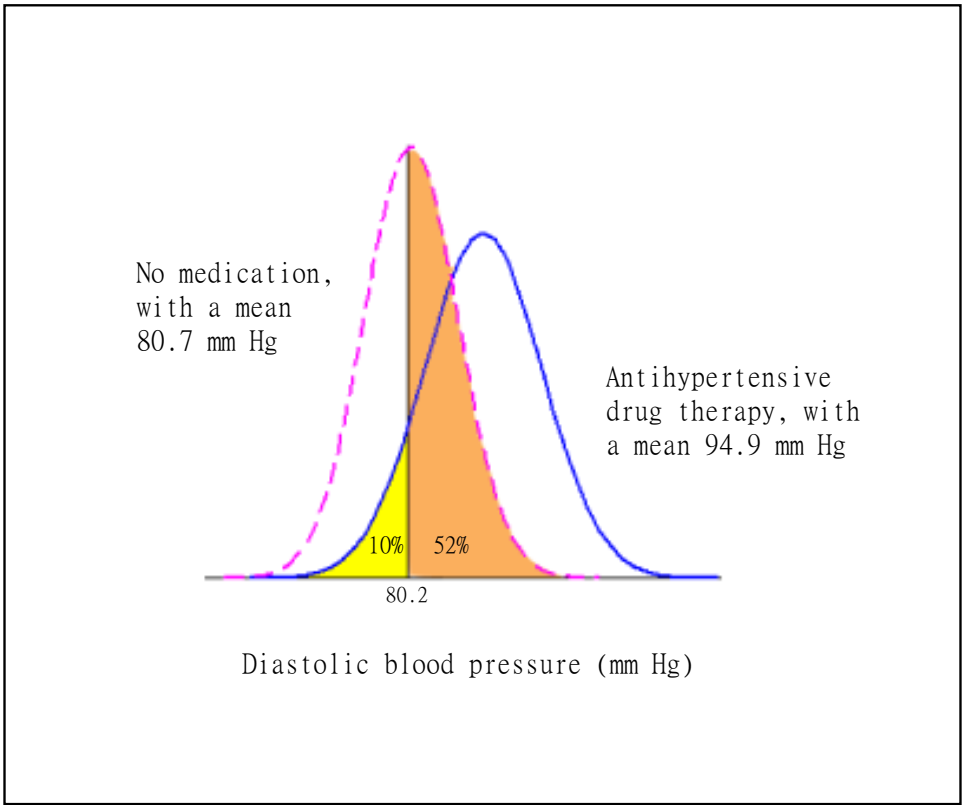


Figure 2.5: Distribution of diastolic blood pressure for two populations

and

$$x = 94.9 + (-1.28) \times 11.5 = 80.2.$$

Approximately 90% of the man taking antihypertensive drugs have diastolic blood pressures that are greater than 80.2 mm Hg. The other 10% of the man represent "false negatives" (i.e. they are taking medication but not identified as such), if we use 80.2 mm Hg as the cutoff point.

What proportion of the men with normal blood pressures will be incorrectly labeled as antihypertensive drug users? These are the men in the 'no medication'

population who have diastolic blood pressure readings greater than 80.2 mm Hg. First, we find the z-score is

$$z = \frac{80.2 - 80.7}{9.2} = -0.05.$$

Again, by looking at Table A.3, an area of 0.480 lies to the right of 0.05; therefore, the area to the right of $z = -0.05$ must be 1.000-0.480=0.520. Approximately 52% of the men with normal blood pressures would be incorrectly grouped as using medication. Theses errors are "false positives."

A trade-off always exists when we try to manipulate proportions of false negative and false positive results. If we reduce the proportion of false positive errors, the false negative results increases. The relationship between these two types of errors is determined by the amount of overlap in the two normal populations of interest.

2.5 Mathematical Expectation

Let X be a continuous random variable (rv) having a p.d.f. $f(x)$, and let $u(X)$ be a function of X. The mathematical expectation (or expected value) of $u(X)$ denoted by $\mathrm{E}[u(X)]$ is

$$\mathrm{E}[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)dx.$$

Some useful facts about mathematical expectations:

1) $\mathrm{E}(k) = k$, if k is a constant,

2) $\mathrm{E}[au(X)+c] = a\mathrm{E}[u(X)] + c$, if a and c are constants,

3) $\mathrm{E}[au_1(X)+bu_2(X)] = a\mathrm{E}[u_1(X)] + a\mathrm{E}[u_2(X)]$, if a and b are constants.

Some special mathematical expectations: (HW: show the following ? are true)

1) $\mathrm{E}(X) \equiv \mu_X$,

2) $\mathrm{var}(X) \equiv \mathrm{E}[X-\mathrm{E}(X)]^2 \stackrel{?}{=} \mathrm{E}(X^2) - \mu_X^2$,

3) $\mathrm{cov}(X,Y) \equiv \mathrm{E}\{[X-\mathrm{E}(X)][Y-\mathrm{E}(Y)]\} \stackrel{?}{=} \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$,

4) $corr(X,Y) \equiv \frac{\mathrm{E}\{[X-\mathrm{E}(X)][Y-\mathrm{E}(Y)]\}}{\sqrt{\mathrm{E}[X-\mathrm{E}(X)]^2}\sqrt{\mathrm{E}[Y-\mathrm{E}(Y)]^2}} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(X)}}$.

Some properties about variance:

1) $\mathrm{var}(kX) = k^2\mathrm{var}(X)$, if k is a constant,

2) $\mathrm{var}(aX+c) = a^2\mathrm{var}(X)$, if a and c are constants,

3) $\mathrm{var}(X \pm Y) = \mathrm{var}(X) + \mathrm{var}(Y) \pm 2\mathrm{cov}(X,Y)$.

Chapter 3

Sampling Distribution of the Mean

When we discuss theoretical probability distributions, the relevant population parameters were assumed to be known. In more practical applications, however, we are not necessary to know the values of these parameters. Instead, we must estimate some characteristic of a population using the information contained in a sample. The process of drawing conclusion about an entire population based on the information in a sample is known as *statistical inference*.

3.1 Sampling Distribution

Suppose that our focus is on estimating the mean value of some continuous r.v. of interest. For instance, we like to make a statement about the mean serum cholesterol lever of all men residing in the U.S., based on a sample drawn from this population. The obvious approach would be to use the mean of the sample as an *estimate* of the unknown population mean μ. (A parameter to a population is as an estimate to a sample!!) The quantity $\overline{X}$ is called an *estimator* of the parameter μ. There are many different approaches to the process of estimation. When the population is normally distributed, the sample mean $\overline{X}$ is a *maximum likelihood estimator* (MLE). The method of maximum likelihood finds the value

of the parameter that is most likely to have produced the observed sample data. However, two different samples are likely to yield different means, so some degree of uncertainty is involved.

In general, a population mean μ can be estimated with greater precision when the group is relatively homogeneous. It is very important that a sample provides an accurate representation of the population from which it is selected; otherwise, the conclusions drawn about the population may be biased. It is crucial that the sample is randomly drawn. Additionally, we expect that the larger the sample, the more reliable our estimate of the population mean. The estimator $\overline{X}$ of the parameter μ is actually a r.v. with outcomes $\overline{x}_1, \overline{x}_2, \overline{x}_3$, and so on ($\overline{x}$ is the sample mean from multiple samples of size n). The probability distribution of $\overline{X}$ is known as a *sampling distribution* of means of samples of size n. In practice, it is not common to select repeated samples of size n from a given population; understanding the properties of the theoretical distribution of their means, however, allows us to make inference based on a single sample of size n.

3.2 The Central Limit Theorem

Given a underlying population with mean μ and standard deviation σ, the distribution of sample means for sample of size n has three important properties:

1) The mean of the sampling distribution is identical to the population mean μ.

2) The standard deviation of the sampling distribution is equal to $\sigma/\sqrt{n}$, which is known as the *standard error* of the mean.

3) Provided that n is large enough, the shape of the sampling distribution is approximately normal.

Although the standard deviation of the sampling distribution is related to the population standard deviation σ, there is less variability among the sample mean than there is among individual observations because we would expect the means of all samples to cluster around the population mean. As n increases, the amount of sampling variation decreases. Finally, if n is large enough, the distribution of sample means is approximately normal. This remarkable results is known as the *central limit theorem*; it applies to any population which has a finite standard deviation, regardless the shape of the underlying distribution. The further the underlying distribution departs from being normally distributed, the larger the value of n that is necessary to ensure the normality of the sampling distribution.

The central limit theorem is very powerful. It even applies to discrete r.v.'s. Regardless of the distribution of X (with mean μ and standard deviation σ), by the central limit theorem, we know that if n is large enough,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

3.3 Applications of the Central Limit Theorem

Consider the distribution of serum cholesterol levels for all 20- to 74-year-old males living in the US. Suppose the mean of this population is $\mu = 211$ mg/100 ml, and the standard deviation is $\sigma = 46$ mg/100 ml.

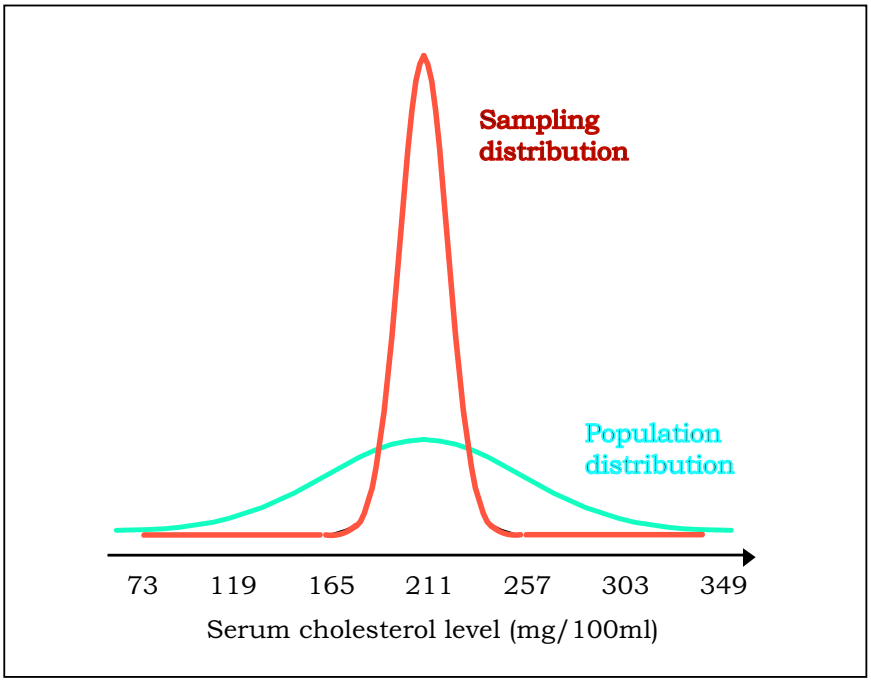


Figure 3.1: Distribution of individual values and means of samples of size 25 for the serum cholesterol levels of 20- to 74-year-old males, US, 1976-1980

Example 1 If we select repeated samples of size 25 from the population, what proportion of the samples will have a mean value of 230 mg/100 ml or above?

If the sample size 25 is large enough, the central limit theorem states that the distribution of sample means of size 25 is approximately normal with mean μ =211 mg/100 ml, and the standard deviation is $\sigma/\sqrt{n} = 46/\sqrt{25} = 9.2$ mg/100 ml (shown in Figure 3.1). In other words,

$$Z = \frac{\overline{X} - 211}{9.2} \sim \mathcal{N}(0, 1).$$

If $\overline{x} = 230$, then we have

$$z = \frac{230 - 211}{9.2} = 2.07.$$

From Table A.3, we find the area to the right of $z = 2.07$ is 0.019. Therefore, only about 1.9% of the samples will have a mean greater than 230 mg/100 ml.

Example 2 What mean value of serum cholesterol level cuts off the lower 10% of the sampling distribution of means?

From Table A.3, we find the area to the right of $z = 1.28$ is 0.10. Solving for $\overline{x}$,

$$-1.28 = z = \frac{\overline{x} - 211}{9.2}$$

and

$$\overline{x} = 211 + (-1.28)(9.2) = 199.2.$$

Therefore, approximately 10% of the samples of size 25 have means less than or equal to 199.2 mg/100 ml.

Example 3 What are the upper and lower limits that enclose 95% of the means of samples of size 25? 1) a symmetric interval; 2) an asymmetrical interval (by trimming 1% on the right side and 4% on the left side).

1) From Table A.3, we find the area to the right of $z = 1.96$ is 0.025. Solving for $\overline{x}$,

$$-1.96 \leq z = \frac{\overline{x} - 211}{9.2} \leq 1.96$$

and

$$193.0 = 211 + (-1.96)(9.2) \leq \overline{x} \leq 211 + (1.96)(9.2) = 229.0.$$

This tells us that approximately 95% of the means of samples of size 25 lie between 193.0 mg/100 ml and 229.0 mg/100 ml.

2) From Table A.3, we find the area to the right of $z = 2.32$ is 0.01 and the area to the right of $z = 1.75$ is 0.04. Solving for $\overline{x}$,

$$-1.75 \leq z = \frac{\overline{x} - 211}{9.2} \leq 2.32$$

and

$$194.9 = 211 + (-1.75)(9.2) \leq \overline{x} \leq 211 + (2.32)(9.2) = 232.3.$$

This tells us that approximately 95% of the means of samples of size 25 lie between 194.9 mg/100 ml and 232.3 mg/100 ml.

It is usually preferable to construct a symmetric interval because it is the shortest interval that captures the appropriate proportion of the means. In some situations, however, we are interested in a one-sided interval.

Example 4 How large would the samples need to be for 95% of their means to lie within ± 5 mg/100 ml of the population mean μ?

To answer this, it is not necessary to know the value of the parameter μ. We simply find the sample size n for which

$$\mathrm{P}(\mu - 5 \leq \overline{X} \leq \mu + 5) = 0.95,$$

or

$$\mathrm{P}(-5 \leq \overline{X} - \mu \leq 5) = 0.95.$$

Standardizing to the $z-$score, we have

$$\mathrm{P}(\frac{-5}{46/\sqrt{n}} \leq \frac{\overline{X}-\mu}{46/\sqrt{n}} \leq \frac{5}{46/\sqrt{n}}) = 0.95.$$

Recall that 95% of the area under the standard normal curve lies between $z = -1.96$ and $z = 1.96$. We can find the sample size n by solving either

$$-1.96 = z = \frac{-5}{46/\sqrt{n}},$$

or

$$1.96 = z = \frac{5}{46/\sqrt{n}}.$$

Then, we have

$$n = \left[\frac{1.96 \times 46}{5}\right]^2 = 325.2.$$

When we deal with sample size, it is conventional to round up. Therefore, samples of size 326 would be required for 95% of the sample means to lie within ± 5 mg/100 ml of the population mean μ.

Example 5 What is the one-sided interval that encloses 95% of the means of samples of size 25? 1) a upper bound; 2) a lower bound.

Since 5% of the area under the standard normal curve lies above $z = 1.645$, we have 1) for the upper bound,

$$z = \frac{\overline{x}-211}{9.2} \leq 1.645,$$

or

$$\overline{x} \leq 226.1.$$

Approximately 95% of the means of samples of size 25 lie below 226.1 mg/100 ml. 2) for the lower bound,

$$z = \frac{\overline{x} - 211}{9.2} \geq -1.645,$$

or

$$\overline{x} \geq 195.9.$$

Then, approximately 95% of the means of samples of size 25 lie above 195.9 mg/100 ml.

In Example 3, we have another two different intervals that covers 95% of the means of samples of size 25, i.e. (193.0,229.0) and (194.9, 232.3). Although all these statements are correct individually, they are not true simultaneously. They are not independent.

Chapter 4

Confidence Intervals

As we have investigated the theoretical properties of a distribution of sample means, we are ready to take the next step and apply this knowledge to the process of statistical inference. Our goal is to describe or estimate some characteristic of a continuous r.v. using the information contained in a sample of observations.

Two methods of estimation are commonly used: 1) point estimation and 2) interval estimation. *Point estimation* involves using the sample data to calculate a single number to estimate the parameter of interest, e.g. using the sample mean $\overline{x}$ to estimate the population means μ. However, two different samples are very likely to result in different sample means, and thus there is some degree of uncertainty involved. Besides, a point estimate does not provide any information about the variability of the estimator. Consequently, *interval estimation* is often preferred. This technique provides a range of reasonable values that are intended to contain the parameter of interest with a certain degree of confidence. This range of values is called a *confidence interval*.

[FYI] There are some desirable properties of estimators:

1) Unbiasedness: The difference between the expectation of an estimator and the corresponding population parameter equals zero.

2) Efficiency: An estimator is an efficient unbiased estimator if for a given sample size its variance is smaller than the variance of any other unbiased estimator.

3) Consistency (asymptotic properties, large sample properties): The estimator approached to the true population parameter as the sample size increases.

4) Sufficiency: A statistic $T(X)$ is sufficient for θ precisely if the conditional probability distribution of the data X given the statistic $T(X)$ does not depend on θ. An equivalent test, known as the Fisher's factorization criterion, is often used instead. If the probability density function (in the discrete case, the probability mass function) of X is $f(x;\theta)$, then T satisfies the factorization criterion if and only if functions g and h can be found such that

$$f(x;\theta) = g(T(x);\theta) \times h(x).$$

4.1 Two-Sided Confidence Intervals

Given a r.v. X that has mean μ and standard deviation σ, the C.L.T. states that

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

1) $Z \sim \mathcal{N}(0,1)$ if X is normally distributed; or

2) $Z \overset{a}{\sim} \mathcal{N}(0,1)$ if X is not normally distributed but n is sufficiently large.

For a standard normal r.v. Z, we know 95% of the observations lie between -1.96 and 1.96, i.e.

$$\mathrm{P}(-1.96 \leq Z \leq 1.96) = 0.95.$$

Equivalently, we have

$$\mathrm{P}(-1.96 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95,$$

or

$$\mathrm{P}(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95. \tag{4.1}$$

Equation (4.1) is the 95% confidence interval (CI) for the population mean μ. This means that we are 95% confident that the interval

$$\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

will cover μ. In other words, if we were to select 100 random samples from the population and calculate 100 different CIs for μ, approximately 95 out of the 100 CIs would cover μ.

It is important to realize that the estimator $\overline{X}$ is a r.v., where the parameter μ is a constant (maybe unknown). Therefore, the interval

$$\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is random and has a 95% chance of covering μ **before** a sample is selected. Once a sample has been drawn and the CI

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

has been calculated, either μ is within the interval or not. There is no longer any probability involved. We do **not** say that there is a 95% probability that μ lies in this single CI.

Suppose X has a population mean $\mu = 211$ and a population standard deviation $\sigma = 46$. If we were to select 100 random samples of size 12 from this population and use each one to construct a CI for μ, we would expect that, on average, 95 out of the 100 intervals would cover the true papulation mean μ and 5 would not. One simulation was performed and the results are shown in Figure 4.1. Each of the CIs that does not contain the true value of μ is marked by a dot in the figure.

Although a 95% CI is used most often in practice, we are not restricted to this choice only. The smaller the range of values we consider (say, a 90% CI), the less confident we are that the interval covers μ. If we wish to make an interval tighter without reducing the level of confidence, we can select a larger sample (i.e. by increasing n).

A generic CI for μ can be constructed. The 100(1-α)% CI for μ is

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

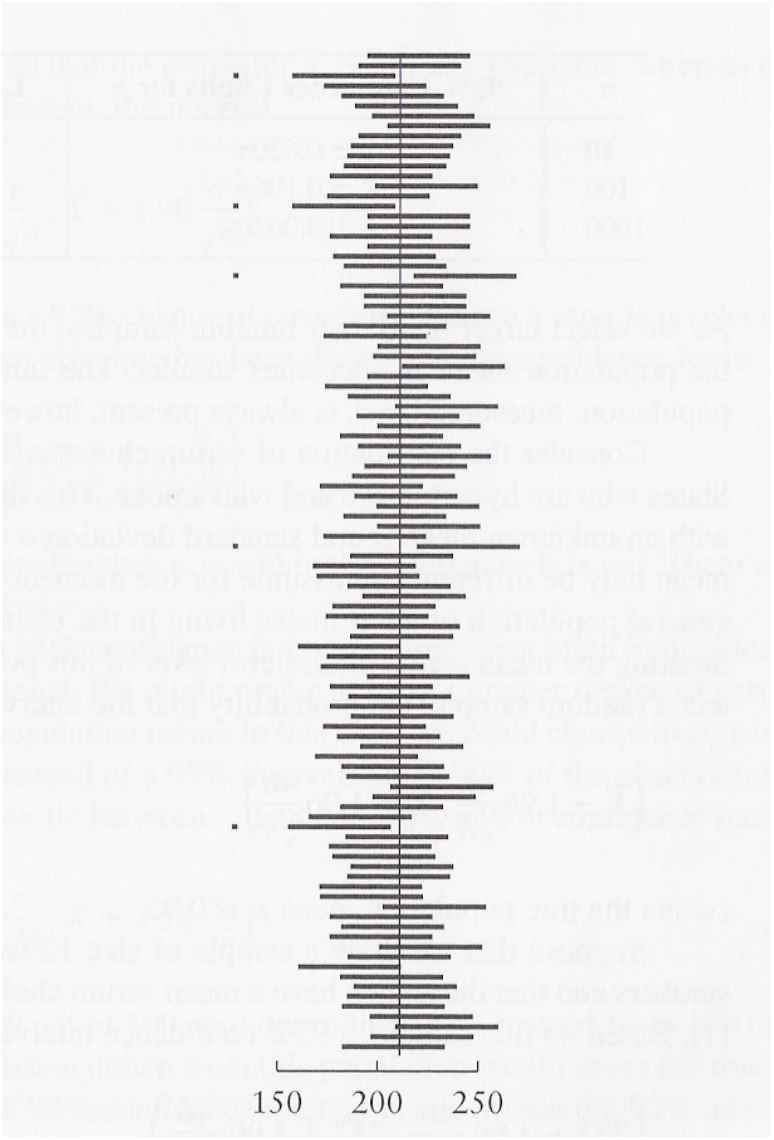


Figure 4.1: Set of 95% confidence intervals constructed from 100 samples of size 12 drawn from a population with mean 211 (marked by the vertical line) and standard deviation 46

where $z_{\alpha/2}$ is the value that cuts off an area of $\alpha/2$ in the upper tail of the standard normal curve. For example, $z_{0.05/2}$=1.96 if α=0.05.

Suppose that we draw a sample of size 12 from the population of hypertensive smokers and that these men have a serum cholesterol level of $\overline{x} = 217$ mg/100 ml. Again, the population standard deviation $\sigma = 46$ mg/100 ml.

Q1. Calculate a 99% CI for μ. Using these 12 hypertensive smokers, we find the limits to be

$$\left(217 - 2.58\frac{46}{\sqrt{12}}, 217 + 2.58\frac{46}{\sqrt{12}}\right),$$

or

$$(183, 251).$$

We are 99% confident that (183,251) covers the true mean serum cholesterol level of the population. The length of the 99% CI is 251-183 = 68 mg/100 ml.

Q2. How large a sample would we need to reduce the length of the above 99% CI to only 20 mg/100 ml? Since the interval is symmetric about the sample mean $\overline{x} = 217$ mg/100 ml, we are looking for the sample size necessary to produced the interval

$$(217 - 10, 217 + 10) = (207, 227).$$

As the 99% CI is of the form

$$\left(217 - 2.58\frac{46}{\sqrt{n}}, 217 + 2.58\frac{46}{\sqrt{n}}\right),$$

we just need to find the required n such that

$$10 = 2.58\frac{46}{\sqrt{n}}.$$

Therefore, we have n=140.8. We would need a sample of 141 men to reduce the length of the 99% CI to 20 mg/100 ml. The length of a CI is a function of σ, n, and the level of confidence (i.e. 100(1-α)%), not the sample mean. That is exactly why all of the CIs in Figure 4.1 have the same length.

4.2 One-Sided Confidence Intervals

In some situations, we are interested in either an upper limit for the population mean or a lower limit of it, but not both. With the same notations used in the previous section, an upper 100(1-α)% confidence bound for μ is

$$\overline{X} + z_\alpha\frac{\sigma}{\sqrt{n}},$$

and a lower 100(1-α)% confidence bound for μ is

$$\overline{X} - z_\alpha\frac{\sigma}{\sqrt{n}}.$$

Suppose that we select a sample of 74 children who have been exposed to high levels of lead from a population with an unknown mean μ and standard deviation $\sigma = 0.85$ g/100 ml. These 74 children have a mean hemoglobin level of $\overline{x}$ =10.6 g/100 ml. Based on this sample, a one-sided 95% CI for μ - the upper bound only - is

$$\mu \leq 10.6 + 1.645\frac{0.85}{\sqrt{74}} = 10.8.$$

We are 95% confident that the true mean hemoglobin level for this population of children is at most 10.8 g/100 ml.

4.3 Student's t Distribution

When we compute CIs for an unknown population mean μ, we have up to this point assumed that the population standard deviation σ was known. However, this is unlikely to be the case. If μ is unknown, σ is probably unknown as well. In this situation, CIs are calculated in a slightly different way. Instead of using the standard normal distribution, the computation depends on a probability distribution called Student's t distribution.

When the population standard deviation σ is unknown, it can be substituted by s, the standard deviation of a sample drawn from the population. The ratio,

$$\frac{\overline{X} - \mu}{s/\sqrt{n}}$$

does not follow a standard normal distribution.

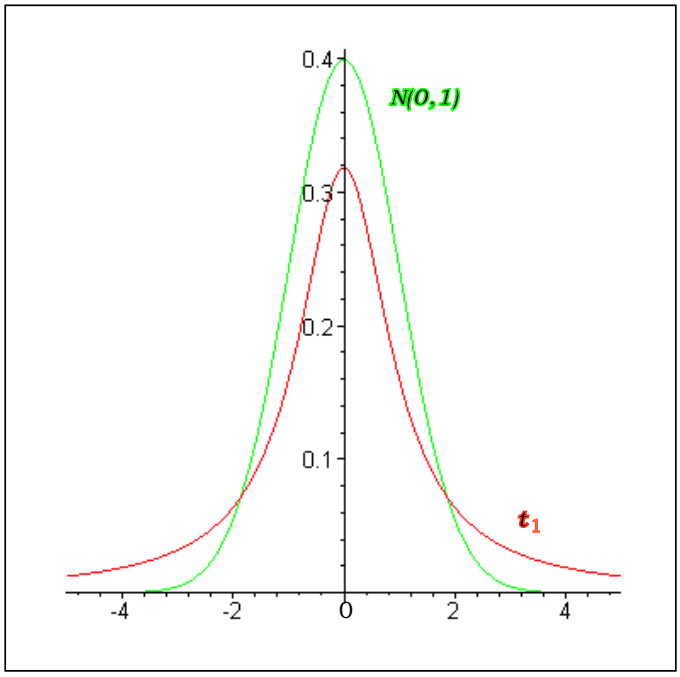


Figure 4.2: The standard normal distribution and Student's t distribution with 1 degree of freedom

If X is normally distributed and a sample size of n is randomly selected from this underlying population, the probability distribution of the r.v.

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

is Student's t distribution with $n-1$ degrees of freedom (df). Like the standard normal distribution, the t distribution is unimodal and symmetric around its mean 0. As shown in Figure 4.2, the t distribution has somewhat thicker tails than the standard normal distribution, reflecting the extra variability introduced by the estimate s. As df increases, the t distribution approaches the standard normal distribution. The values of t_{n-1} that mark off the upper 2.5% of the distributions with various degrees of freedom are listed below. In fact, when we

have more than 30 df, we are able to substitute the standard normal distribution for the t distribution and be off in our calculation by less than 5%.

df=$(n-1)$	t_{n-1}
1	12.706
2	4.303
5	2.571
10	2.228
20	2.086
30	2.042
40	2.021
60	2.000
120	1.980
∞	1.960

Table A.4 in Appendix A is a condensed table of areas computed for the family of t distributions. For a given df, the entry in the table presents the value of t_{n-1} that cuts off the specified area in the upper tail of the distribution.

Consider a random sample of 10 children chosen from the population of unknown μ and σ. The sample mean $\overline{x} = 37.2$ and the sample standard deviation is $s = 7.13$. Since σ is unknown, we must use the t distribution to find a 95% CI for μ. For t distribution with 10-1 = 9 df, from Table A.4, we find 95% of the observations lie between -2.262 and 2.262. By replacing σ with s, a 95% CI for the population mean μ is

$$\left(\overline{X} - 2.262\frac{s}{\sqrt{n}}, \overline{X} + 2.262\frac{s}{\sqrt{n}}\right).$$

Therefore, the 95% CI becomes

$$\left(37.2 - 2.262\frac{7.13}{\sqrt{10}}, 37.2 + 2.262\frac{7.13}{\sqrt{10}}\right),$$

or

$$(32.1, 42.3).$$

If the population standard deviation σ had been known and had been equal to the sample value of 7.13, the 95% CI for μ would have been

$$\left(37.2 - 1.96\frac{7.13}{\sqrt{10}}, 37.2 + 1.96\frac{7.13}{\sqrt{10}}\right),$$

or

$$(32.8, 41.6).$$

Most of time, CIs based on t distribution are longer than the corresponding intervals based on the standard normal distribution. On the left-hand side, Figure 4.3, shows the 95% CIs for μ that were calculated from 100 random samples and were displayed in Figure 4.1. The right-hand side of the figure shows 100 additional intervals that were computed using the sample samples; in each case, however, the standard deviation was not assumed to be known. Once again, 95 of the 100 new intervals contain the true mean μ. Note that this time, the intervals vary in length.

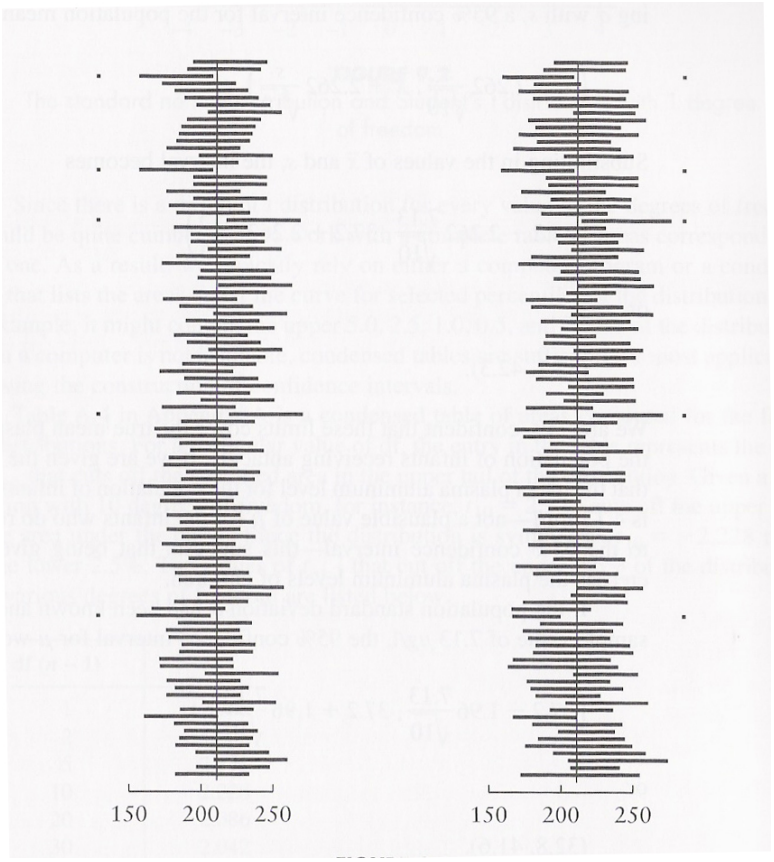


Figure 4.3: Two Sets of 95% confidence intervals constructed from 100 samples of size 12 drawn from a population with mean 211 (marked by the vertical line) and one with standard deviation 46 and the other with standard deviation unknown

Chapter 5

Hypothesis Testing

A first major area of statistical inference is estimation of parameters. The second major area is tests of statistical hypotheses.

5.1 General Concepts

When we want to draw some conclusion about a population parameter like the mean of a continuous r.v., one approach is to construct a confidence interval for the parameter, and another way is to conduct a *statistical hypothesis test*.

To perform such a test, we begin by claiming the mean of the population μ is equal to some postulated value μ_0, i.e. a *null hypothesis* $H_0 : \mu = \mu_0$. For instance, if we wanted to test whether the mean serum cholesterol level of the subpopulation of hypertensive smokers is equal to the mean of the general population of 20- to 74-yr-old males which is 211 mg/100 ml, the null hypothesis would be

$$H_0 : \mu = \mu_0 = 211 \text{ mg/100 ml}.$$

We need a second statement that contradicts H_0, which is called an *alternative hypothesis* represented by H_A or H_1. In this case, we could have

$$H_A : \mu \neq 211 \text{ mg/100 ml.}$$

Together, the null and alternative hypotheses cover all possible values of the population mean μ; consequently, one of the two statements must be true.

If there is evidence that the selected random sample could **not** have come from a population with mean μ_0, we reject the null hypothesis. This occurs when, given that H_0 is true, the probability of obtaining a sample mean as extreme as or more extreme than the observed value $\overline{x}$ is sufficiently small. Such a test result is said to be *statistically significant*. If there is not sufficient evidence to doubt the validity of the null hypothesis, we cannot reject this claim. However, we do not say that we accept H_0; the test does not prove the null hypothesis.

Before a test is actually carried out, the *significance level* denoted by the Greek letter α must be specified. In most applications, $\alpha = 0.05$ is chosen. We reject H_0 when the chance that the sample could have come from a population with mean μ_0 is less than or equal to 5%. This implies that we reject incorrectly 5% of the time.

Given that H_0 is true, the probability of obtaining a mean as extreme as or more extreme than the observed sample mean $\overline{x}$ is called the *p-value* of the test or simply p. The p-value is compared to the predetermined significance level α to decide whether the null hypothesis should be rejected or not. If p is less than or equal to α, we reject H_0. If p is greater than α, we **do not reject** H_0.

The process of hypothesis testing is not perfect; there are two kinds of errors that can be made. We could either reject the null hypothesis when μ is equal to μ_0, or fail to reject it when μ is not equal to μ_0. These two types of errors are discussed in more detail later in this chapter.

5.2 Z-Tests and t-Tests

Assume that the continuous r.v. X has mean μ_0 and the known standard deviation σ. Thus, according to the central limit theorem,

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \overset{a}{\sim} \mathcal{N}(0,1),$$

if the value of n is sufficiently large. For a given sample with mean $\overline{x}$, we can calculate the corresponding Z, called the *test statistic* (TS). We then use Table A.3 in Appendix A to determine the probability of obtaining a value of Z that is extreme as or more extreme than the one observed. Because it relies on the standard normal distribution, a test of this kind is called a *z-test*.

When the standard deviation σ is not known, we substitute the sample value s for σ. If X is normally distributed, the r.v.

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}.$$

In this case, we can calculate the value of t corresponding to a given $\overline{x}$ and consult Table A.4 to find the probability of obtaining a sample mean that is more extreme than the one observed. This procedure is known as a *t-test*.

5.3 Two-Sided Test of Hypotheses

Consider the distribution of serum cholesterol levels for adult males in the US who are hypertensive and who smoke. The standard deviation of this distribution is assumed to be $\sigma = 46$ mg/100 ml; the null hypothesis to be tested is

$$H_0 : \mu = 211 \text{ mg/100 ml},$$

where $\mu_0 = 211$ mg/100 ml is the mean serum cholesterol levels for all 20- to 74-yr-old males. Since the mean of subpopulation of hypertensive smokers could be either larger than μ_0 or smaller than μ_0, we are concerned with deviations that occur in either direction. As a result, we conduct what is called a *two-sided* test. The alternative hypothesis for the two-sided test is

$$H_A : \mu \neq 211 \text{ mg/100 ml}.$$

There is actually a mathematical equivalence between confidence intervals (CIs) and tests of hypothesis here. Any value of z that is between -1.96 and 1.96 would result in a p-value greater than 0.05, and the null hypothesis would not be rejected. On the other hand, H_0 would be rejected for any value of z that is either less than -1.96 or greater than 1.96. At the $\alpha = 0.05$ level, the numbers -1.96 and 1.96 are called the *critical values* of the test statistic. Alternatively, the null hypothesis will fail to be rejected when μ_0 lies within the 95% CI for μ. In contrast, μ_0 lies outside of the 95% CI for μ would results in a rejection of the null hypothesis at the $\alpha = 0.05$ level.

Although CIs and tests of hypotheses lead us to the same conclusions, the information provided by each is somewhat different. The CIs tells us something about the uncertainty in our point estimate $\overline{x}$ for the parameter μ. The hypothesis test provides a specific p-value to help us decide whether the postulated value of the mean is likely to be correct or not.

Example 1: Z-test A random sample of 12 hypertensive smokers has mean serum cholesterol level $\overline{x} = 217$ mg/100 ml. Is it likely that this sample comes from a population with mean $\mu_0 = 211$ mg/100 ml?

$$H_0 : \mu = 211 \text{ mg/100 ml v.s. } H_A : \mu \neq 211 \text{ mg/100 ml.}$$

$$\begin{aligned} \text{TS: } z &= \frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} \\ &= \frac{|217 - 211|}{46/\sqrt{12}} \\ &= 0.45 \\ &\sim \mathcal{N}(0,1), \text{ if } H_0 \text{ is true.} \end{aligned}$$

According to Table A.3, the area to the right of $z = 0.45$ is 0.326, which is the probability of observing $Z = 0.45$ or anything larger, given that H_0 is true. As Z is symmetric, the area to the left of z = -0.45 is 0.326 as well. Therefore, the p-value of the test is 0.652. If the significance level is set to be $\alpha = 0.05$, we do not reject the null hypothesis for $p > 0.05$. In other words, the evidence is

insufficient to conclude that the mean serum cholesterol level of the population of hypertensive smokers is different from 211 mg/100 ml.

Example 2: t-test Consider a random sample of 10 children selected from a population of infants receiving antacids containing aluminum. The underlying distribution of plasma aluminum levels for this population is approximately normal with unknown mean μ and standard deviation σ. However, we do know that the mean plasma aluminum levels for this sample of size 10 is $\overline{x} = 37.20$ μg/l and that its standard deviation is $s = 7.13$ μg/l. Is it likely that the data in our sample could have come from a population of infants not receiving antacids with mean $\mu_0 = 4.13$ μg/l?

If we are interested in deviation from the mean that could occur in either direction, we conduct a two-sided test of hypothesis at the $\alpha = 0.05$ level of significance:

$$H_0 : \mu = 4.13 \text{ } \mu\text{g/l} \text{ v.s. } H_A : \mu \neq 4.13 \text{ } \mu\text{g/l}.$$

$$\begin{aligned} \text{TS: } t &= \frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \\ &= \frac{37.20 - 4.13}{7.13/\sqrt{10}} \\ &= 14.67 \\ &\sim t_{df=10-1=9}, \text{ if } H_0 \text{ is true.} \end{aligned}$$

From Table A.4, we observe that the total area to the right of $t_9 = 14.67$ and to the left of t_9 = -14.67 is less than 2(0.0005)=0.001. Therefore, $p < 0.05$, and we reject the null hypothesis. This sample of infants provides evidence that the

mean plasma aluminum level of children receiving antacids is not equal to the mean aluminum level of children who do not receive them.

5.4 One-Sided Test of Hypotheses

Before we conduct a test of hypothesis, we must decide whether we are concerned with deviations from μ_0 that could occur in both directions or in one direction only. This decision must be made **before** a random sample is selected; it should not be influenced by the outcome of the sample. If prior knowledge indicates that μ cannot be less than μ_0, the only values of $\overline{x}$ that will provide evidence against the null hypothesis

$$H_0 : \mu = \mu_0$$

are those that are much larger than μ_0. In this case, the null hypothesis is more properly stated as

$$H_0 : \mu \leq \mu_0$$

and the alternative hypothesis is

$$H_A : \mu > \mu_0.$$

A two-sided test is always the more conservative choice; in general, the p-value of a two-sided test is twice as large as that of a one-sided test. Not infrequently, a one-sided test achieves significance when a two-sided test does not. Consequently, the decision is often made on nonscientific grounds.

Example: one-sided test Consider the distribution of hemoglobin levels for the population of children under the age of 6 who have been exposed to high levels of lead. We believe that if the hemoglobin levels of exposed children are different from those of unexposed children, they must be on average lower; therefore we are concerned only with deviations from the mean that are below μ_0. If the mean hemoglobin level of the general population of children under the age of 6 is 12.29 g/100 ml, the null hypothesis for the test is

$$H_0 : \mu \geq 12.29 \text{ g/100 ml}$$

and the one-sided alternative hypothesis is

$$H_A : \mu < 12.29 \text{ g/100 ml.}$$

A random sample of 74 children who have been exposed to high levels of lead has a mean hemoglobin level of $\overline{x} = 10.6$ g/100 ml with $\sigma = 0.85$ g/100 ml. Since σ is known, we use the normal distribution rather than the t for the test:

$$\begin{aligned} \text{TS: } z &= \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{10.6 - 12.29}{0.85/\sqrt{74}} \\ &= -17.10 \\ &\sim \mathcal{N}(0,1), \text{ if } H_0 \text{ is true.} \end{aligned}$$

According to Table A.3, the area to the left of z is less than 0.001. Since this p-value is smaller than $\alpha = 0.05$, we reject the null hypothesis in favor of the alternative. Because this is one-sided test, any value of z that is less than or equal to the critical value -1.645 would have led us to reject the null hypothesis at the $\alpha = 0.05$ level. Note 12.29 lies above 10.8 which is the upper one-sided 95% confidence bound for μ in pervious chapter (Section 4.2).

5.5 Types of Error

As noted earlier, two kinds of errors can be made when we conduct a test of hypothesis. The first one is called a *type I error* (aka a *rejection error* or an α error). A type I error happens when H_0 is true and we reject the null hypothesis

$$H_0 : \mu = \mu_0.$$

The probability of making a type I error is determined by the significance level of the test; recall that

$$\alpha = \mathrm{P}(\text{reject } H_0 | H_0 \text{ is true}).$$

If we were to conduct repeated and independent tests of hypotheses setting the significance level at 0.05, we would erroneously reject a true null hypothesis 5% of the time.

The second kind of error that can be made during a hypothesis test is a *type II error* (aka an *acceptance error* or a β error). A type II error is made if we fail to reject the null hypothesis

$$H_0 : \mu = \mu_0.$$

when H_0 is false. The probability of committing a type II error is represented by the Greek letter β, where

$$\beta = \mathrm{P}(\text{do not reject } H_0 | H_0 \text{ is false}).$$

If $\beta = 0.10$, for instance, the probability that we do not reject the null hypothesis when $\mu \neq \mu_0$ is 10%. The two types of errors that can be made are summarized below.

	Population	
Result of Test	$\mu = \mu_0$	$\mu \neq \mu_0$
Do Not Reject H_0	Correct	*Type II error*
Reject H_0	*Type I error*	Correct

Example The mean serum cholesteric levels for all 20- to 74-yr-old males in US is $\mu = 211$ mg/100 ml and the standard deviation is $\sigma = 46$ mg/100 ml. If we do not know the true population mean but we know the mean serum cholesteric

levels for the subpopulation of 20- to 24-yr-old males is 180 mg/100 ml. If we were to conduct a one-sided test:

$$H_0 : \mu \leq 180 \text{ mg/100 ml v.s. } H_A : \mu > 180 \text{ mg/100 ml},$$

at the $\alpha = 0.05$ level of significance. What is the probability of the type II error associated with such a test, assuming that we select a sample of size 25?

To determine this, we first need to find the mean serum cholesterol level our sample must have for H_0 to be rejected.

$$\begin{aligned} \text{TS: } z &= \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{\overline{x} - 180}{46/\sqrt{25}} \\ &= 1.645 \end{aligned}$$

Solving for $\overline{x}$, we have

$$\begin{aligned} \overline{x} &= 180 + \frac{1.645 \times 46}{\sqrt{25}} \\ &= 195.1. \end{aligned}$$

As shown in Figure 5.1, the area to the right of $\overline{x} = 195.1$ corresponds to the upper 5% of the sampling population of means of samples of size 25 when $\mu = 180$.

Recall that the probability of making a type II error (i.e. β) is the probability rejecting the null hypothesis given that H_0 is false. Therefore, it is the chance of obtaining a sample mean that is less than 195.1 mg/100 ml given that the true

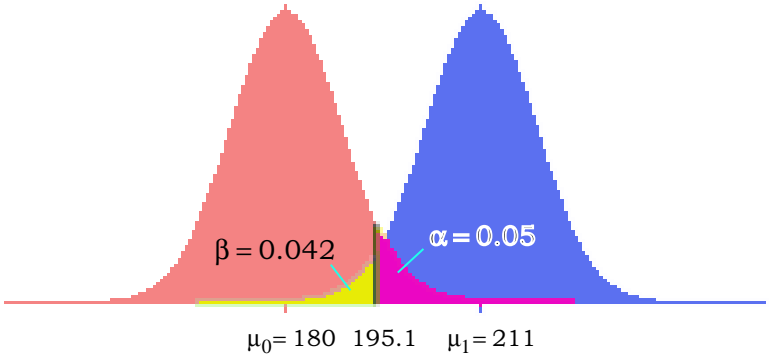


Figure 5.1: Distribution of means of samples of size 25 for the serum cholesterol levels of males 20 to 74 years of age, $\mu_0 = 180$ mg/100 ml versus $\mu_1 = 211$ mg/100 ml. $\alpha = 0.05$, when 195.1 mg/100 ml is the cutoff

population mean is not 180 mg/100 ml but is instead 211 mg/100 ml. Since a sample mean less than $\overline{x} = 195.1$ mg/100 ml

$$\begin{aligned} z &= \frac{195.1 - 211}{46/\sqrt{25}} \\ &= -1.73. \end{aligned}$$

The area under the standard normal curve that lies to the left of z = -1.73 is 0.042. Therefore, $\beta = 0.042$ when the true population mean is 211 mg/100 ml.

While the type I error α is determined by looking at the case in which H_0 is true (i.e. $\mu = \mu_0$), the type II error β is found by considering the situation in which H_0 is false (i.e. $\mu \neq \mu_0$). If μ is not equal to μ_0, however, there are an infinite number of possible values that μ could assume. The type II error is calculated for a **single** such value, μ_1. If we had chosen a different alternative

population mean, we would have computed a different value for β. The closer μ_1 is to μ_0, the more difficult it is to reject the null hypothesis.

5.6 Power

The *power* of the test of hypothesis is the probability of rejecting the null hypothesis when H_0 is false, which is 1 - β. In other words, it is the probability of avoiding a type II error:

$$\text{power} = \text{P}(\text{reject } H_0 | H_0 \text{ is false}) = 1 - \beta.$$

Like β, the power must be computed for a particular alternative population mean μ_1.

In the serum cholesterol example in the previous section, the power of the one-sided of hypothesis is

$$\begin{aligned}
\text{power} &= \text{P}(\text{reject } H_0 : \mu \leq 180 | \mu = 211) \\
&= \text{P}(\overline{X} \geq 195.1 | \mu = 211) \\
&= \text{P}(Z \geq -1.73) \\
&= 1 - \text{P}(Z > 1.73) \\
&= 1 - 0.042 \\
&= 0.958.
\end{aligned}$$

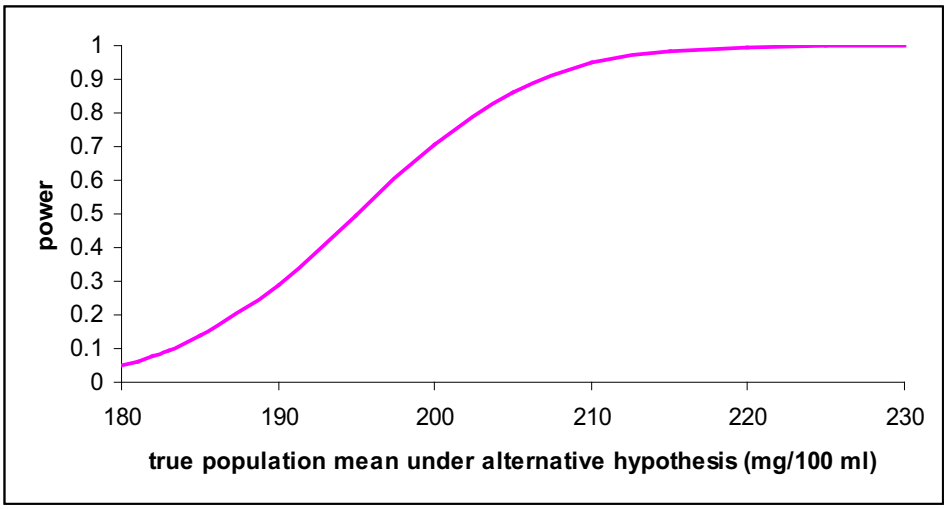


Figure 5.2: Power curve: Distribution of means of samples of size 25 for the serum cholesterol levels of males 20 to 74 years of age, $\mu_0 = 180$ mg/100 ml versus different μ_1 under alternative hypothesis. $\alpha = 0.05$

If we were to plot the values of 1 - β against all possible alternative population means, we would end up with what is known as a *power curve* as illustrated in Figure 5.2. The power of the test approaches 1 as the alternative mean moves further and further away from the null value.

In most practical applications, a power less than 80% is considered insufficient. One way to increase the power of a test is to raise the significance level α. If α had been equal to 0.10 for the test of the null hypothesis

$$H_0 : \mu \leq 180 \text{ mg/100 ml},$$

the power becomes 0.982 as illustrated in Figure 5.3.

The trade-off between α and β is similar to that observed to exist between the sensitivity and the specificity of a diagnostic test. The balance between

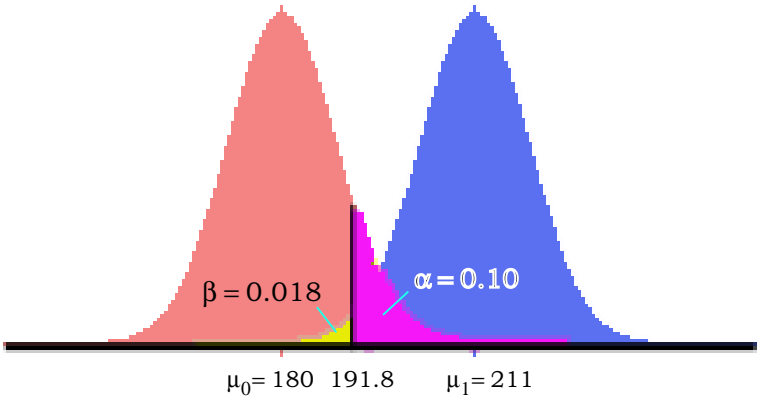


Figure 5.3: Distribution of means of samples of size 25 for the serum cholesterol levels of males 20 to 74 years of age, $\mu_0 = 180$ mg/100 ml versus $\mu_1 = 211$ mg/100 ml. $\alpha = 0.10$, when 191.8 mg/100 ml is the cutoff

the two types of error is a delicate one. The only way to diminish α and β simultaneously is to reduce the amount of overlap in the two distributions - the one centered at μ_0 and the one centered at μ_1. An alternative is to increase the sample size n. By increasing n, we decrease the standard error $\sigma/\sqrt{n}$; this causes the two distributions to become more narrow, which, in turn, lessens the amount of overlap. Another options that we have not yet mentioned is to find a "more powerful" test statistic. This topic is discussed further in a later chapter.

5.7 Sample Size Estimation

If we conduct a one-sided test of the null hypothesis

$$H_0 : \mu \leq \mu_0 \text{ v.s. } H_A : \mu > \mu_0$$

at the α level of significance, H_0 would be rejected for any test statistic that takes a value $z \geq z_\alpha$. Similarly, considering the desired power of the test 1 - β, the generic value of z that corresponds to a probability β is $z = -z_\beta$. The two different expressions for $\overline{x}$ are

$$\overline{x} = \mu_0 + z_\alpha \left(\frac{\sigma}{\sqrt{n}} \right)$$

and

$$\overline{x} = \mu_1 - z_\beta \left(\frac{\sigma}{\sqrt{n}} \right),$$

and setting them equal to each other gives us

$$n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_1 - \mu_0} \right]^2. \tag{5.1}$$

This is the sample size necessary to achieve a power of 1 - β when we conduct a one-sided test at the α level. As shown in Equation (5.1), several factors influence the size of n. If we reduce the type I error or type II error, this would produce a larger value of n. The difference $\mu_1 - \mu_0$ decreases and the sample size increases. Finally, the larger the variability of the underlying population σ, the larger the sample size required.

If we conduct a two-sided test, the sample size necessary to achieve a power of 1 - β at the α level is

$$n = \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_1 - \mu_0} \right]^2.$$

Note that the sample size for a two-sided test is always larger than the sample size for the corresponding one-sided test.

Chapter 6

Comparison of Two Means

In the preceding chapter, we used a statistical test of hypothesis to compare the unknown mean of a single population to some fixed, known value μ_0. In practical application, it is more common to compare the means of two different population, where both means are unknown.

A test of hypothesis involving two samples is similar in many respects to a test conducted for a single sample. We begin by specifying a null hypothesis; in most cases; we are interested in testing whether the two population means are equal. We then calculate the probability of obtaining a pair of the sample means as discrepant as or more discrepant than the observed means given that the null hypothesis is true (Q: what is this probability?). If this probability is sufficiently small, we reject the null hypothesis and conclude that the two population means are different. As before, we must specify a level of significance α and state whether we are interested in a one-sided or two-sided test. The specific form of the analysis depends on the nature of two sets of observations involved; in particular; we must determine whether the data come from paired or independent samples.

6.1 Paired Samples

The distinguishing characteristic of paired samples is that for each observation in the first group, there is a corresponding observation in the second group. In the technique known as *self-pairing*, measurements are taken on the a single subject at two distinct points in time. One common example of self-pairing is the "before and after" experiment. A second type of pairing occurs when an investigator matches the subjects in one group with those in a second group so that the members of a pair as much alike as possible when respect to important characteristics such as age and gender. The intent of pairing is to make a comparison more precise.

When data consist of paired samples, the *paired t-test* is the appropriate method of analysis. If we like to conduct a one-sided test at the $\alpha = 0.05$ level of significant. The null hypothesis is

$$H_0 : \mu_1 \geq \mu_2, \text{ or } H_0 : \mu_1 - \mu_2 \geq 0,$$

and the alternative hypothesis is

$$H_A : \mu_1 - \mu_2 < 0.$$

Rather than consider the sets of observations to be distinct samples, we focus on the **difference** in measurements within each pair, as indicated as follows:

Sample 1	Sample 2	Difference
x_{11}	x_{12}	$d_1 = x_{11} - x_{12}$
x_{21}	x_{22}	$d_2 = x_{21} - x_{22}$
x_{31}	x_{32}	$d_3 = x_{31} - x_{32}$
.	.	.
.	.	.
x_{n1}	x_{n2}	$d_n = x_{n1} - x_{n2}$

Instead of analyzing the individual observations, we use the difference between the members of each pair as the variable of interest. Since the difference is a single measurement, our analysis reduces to the one-sample problem and we apply the hypothesis testing procedure introduced in the preceding chapter.

To do this, we first note that the mean of the set of differences is

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n};$$

this sample mean provides a point estimate for the true difference in population means, $\mu_1 - \mu_2$. The standard deviation of the difference is

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}}.$$

If we denote the true difference in population means by

$$\delta = \mu_1 - \mu_2$$

and wish to test whether these two means are equal, we can write the null hypothesis as

$$H_0 : \delta = 0$$

and the alternative is

$$H_A : \delta \neq 0.$$

Assuming that the population of differences is normally distributed, the test statistic is

$$\text{TS: } t = \frac{|\overline{d} - \delta|}{s_d/\sqrt{n}} \sim t_{n-1}, \text{ if } H_0 \text{ is true.}$$

Note that $s_d/\sqrt{n}$ is the standard error of $\overline{d}$. The p-value (p), the probability of observing a mean difference as large as or larger than $\overline{d}$ given that $\delta = 0$, can be determined by looking at Table A.4 in Appendix A. If $p \leq \alpha$, we reject H_0. If $p > \alpha$, we do not reject H_0.

Example: Paired t-test Consider the data taken from a study in which each of 63 adult males with coronary artery disease is subjected to a series of exercise tests on a number of different occasions. A patient first undergoes an exercise test on a treadmill; the length of time from the start of the test until the patient experiences angina (pain or spasms in the chest) is recorded. He is then exposed to plain room air for approximately one hour. At the end of this time, he performs a second exercise test; time until the onset of angina is again recorded. The observation of interest is the percent decrease in time to angina between the first and the second tests. For example, if during the first test, a man has an attack

of angina after 983 seconds, and during the second test he has an attach after 957 seconds, his percent decrease in time to angina is

$$\frac{983 - 957}{983} = 0.026 = 2.6\%.$$

The unknown population mean of this distribution of percent decreases is μ_1; for the 63 patients in the sample, the observed mean percent decrease is $\overline{x}_1 = 0.96\%$.

On another day, the same patient undergoes a similar series of tests. This time, however, he is exposed to a mixture of air and CO during the interval between the tests. The unknown mean of this distribution (percent decrease) is μ_2; the sample mean for the group of 63 subjects is $\overline{x}_2 = 7.59\%$. We consider deviations that occur in one direction only because CO could not possibly be beneficial to an individual's health; we therefore conduct a one-sided test. We can write the null hypothesis as

$$H_0 : \delta \geq 0$$

and the alternative is

$$H_A : \delta < 0.$$

The mean of these differences is

$$\overline{d} = \frac{\sum_{i=1}^{63} d_i}{63} = -6.63,$$

and the standard deviation of the difference is

$$s_d = \sqrt{\frac{\sum_{i=1}^{63}(d_i - \overline{d})^2}{63-1}} = 20.29.$$

We have

$$\begin{aligned} \text{TS: } t &= \frac{\overline{d} - \delta}{s_d/\sqrt{n}} = \frac{-6.63 - 0}{20.29/\sqrt{63}} \\ &= -2.59. \end{aligned}$$

From Table A.4, we observe that for a t distribution with df = 63 - 1 = 62, the area under the curve to the left of t_{62} = -2.59 is between 0.005 and 0.01. Therefore, $0.005 < p < 0.01$. We reject the null hypothesis at the 0.05 level.

We could also construct an upper confidence bound for δ. For a t distribution with df = 62, 95% of the observations lie above -1.671. Therefore,

$$\begin{aligned} \mathrm{P}(t \geq -1.671) &= \mathrm{P}(\frac{\overline{d} - \delta}{s_d/\sqrt{n}} \geq -1.671) \\ &= 0.95. \end{aligned}$$

A one-sided 95% CI for δ is

$$\begin{aligned} \delta &\leq \overline{d} + 1.671\frac{s_d}{\sqrt{n}} \\ &= -6.63 + 1.671\frac{20.29}{\sqrt{63}} \\ &= -2.36. \end{aligned}$$

We are 95% confident that the true difference in population means is less than or equal to -2.36%.

6.2 Independent Samples

Consider two independent random samples drawn from two different populations as below:

		Group 1	**Group 2**
Population	Mean	μ_1	μ_2
	Standard Deviation	σ_1	σ_2
Sample	Mean	$\overline{x}_1$	$\overline{x}_2$
	Standard Deviation	s_1	s_2
	Sample Size	n_1	n_2

If we are interested in testing the null hypothesis that two population means are identical. This can be expressed as either

$$H_0 : \mu_1 - \mu_2 = 0$$

or

$$H_0 : \mu_1 = \mu_2.$$

The alternative hypothesis is

$$H_A : \mu_1 \neq \mu_2.$$

Two different situations arise in the comparison of independent samples. First, the variance of the underlying populations either are known to be equal to each other or are assumed to be equal. This leads to the *two sample t-test*, which is omnipresent in the literature. Second, the variances are not assumed to be the same; in this case, the standard t-test is no longer valid. Each of the two situations are discussed here.

6.2.1 F-Test for Equality of Variances ✠

An F-test can be used to test if the variances of two populations are equal. The null and alternative hypotheses are

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs } H_A : \sigma_1^2 \neq \sigma_2^2.$$

The appropriate test statistic is

$$\begin{aligned} \text{TS: } F &= s_1^2/s_2^2 \\ &\sim \mathcal{F}_{n_1-1,n_2-1}, \text{ under } H_0. \end{aligned}$$

If $F > 1$, then the p-value is $2 \times \text{pr}(\mathcal{F}_{n_1-1,n_2-1} > F)$; while if $F < 1$, then the p-value is $2 \times \text{pr}(\mathcal{F}_{n_1-1,n_2-1} < F)$.

6.2.2 Equal Variances

We first consider the situation in which it either is known or is reasonable to assume that the two population variances are equal. Recall that for a single normal population with mean μ and standard deviation σ, the central limit theorem (CLT) states that the sample mean $\overline{X}$ is approximately normally distributed - assuming that n is large enough - with mean μ and standard deviation $\sigma/\sqrt{n}$. When are are dealing with samples from two independent normal populations, an extension of the CLT says that the difference in sample means $\overline{X}_1 - \overline{X}_2$ is approximately normal with mean $\mu_1 - \mu_2$ and standard error $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Since it is assumed that the population variances are identical, we substitute the common value σ^2 for both σ_1^2 and σ_2^2. Therefore, if σ^2 is known, we have

$$\begin{aligned} \text{TS: } z &= \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \\ &= \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}} \\ &\sim \mathcal{N}(0,1). \end{aligned}$$

If the true value of σ^2 is unknown, we use the following test statistic instead:

$$\begin{aligned} \text{TS: } t &= \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2[(1/n_1) + (1/n_2)]}} \\ &\sim t_{n_1+n_2-2} \text{ under the null hypothesis } H_0 : \mu_1 = \mu_2, \end{aligned}$$

where s_p^2 is a pooled estimate of the common variance σ^2. The pooled estimate of the variance, s_p^2, combines information from both of the samples to produce

a more reliable estimate of σ^2. It can be calculated in two different ways. If we know the values of all the observations in the samples, we apply the formula

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(x_{i1} - \overline{x}_1)^2 + \sum_{j=1}^{n_2}(x_{j2} - \overline{x}_2)^2}{n_1 + n_2 - 2}.$$

If we are given s_1 and s_2 only, we must use

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{6.1}$$

Equation (6.1) demonstrates that s_p^2 is actually weighted average of the two sample variances s_1^2 and s_2^2, where each variance is weighted by the degrees of freedom associated with it. If $n_1 = n_2$, then s_p^2 is simple arithmetic average; otherwise, more weight is given to the variance of the larger sample.

Example: Equal Variances Consider the distribution of serum iron levels for the population of healthy children and the population of children with cystic fibrosis. The distributions are both approximately normal; denote the mean serum iron level of the healthy children by μ_1 and that of the children with disease by μ_2. The standard deviations of the two populations are unknown but are assumed to be equal. We test the null hypothesis that the two population means are identical with a two-sided test,

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_A : \mu_1 \neq \mu_2.$$

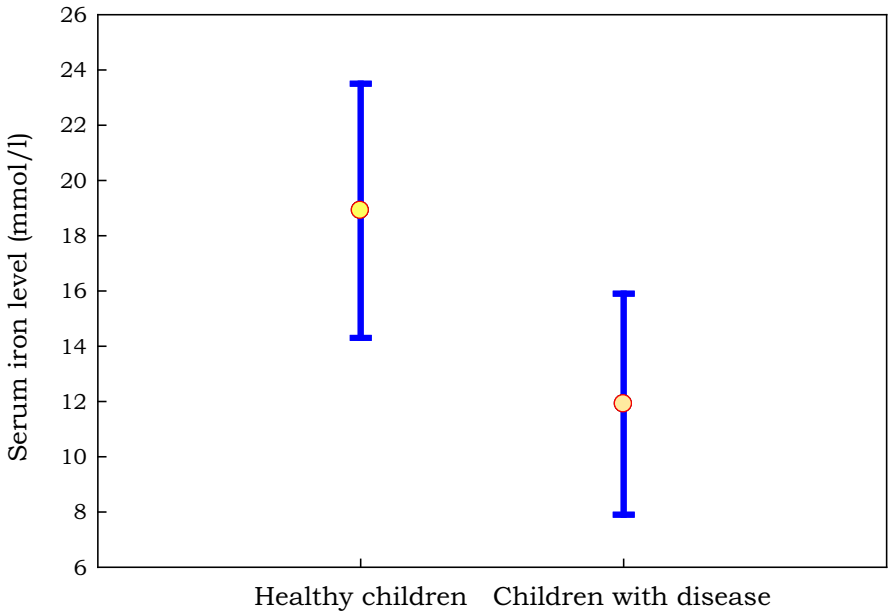


Figure 6.1: 95% confidence intervals for the mean serum iron levels of healthy children and children with cystic fibrosis

A random sample is selected from each population. The sample of $n_1 = 9$ healthy children has mean serum iron level $\overline{x}_1 = 18.9$ mmol/l and standard deviation $s_1 = 5.9$ mmol/l; the sample of $n_2 = 13$ children with cystic fibrosis has mean serum iron level $\overline{x}_2 = 11.9$ mmol/l and standard deviation $s_2 = 6.3$ mmol/l.

An investigator might begin an analysis by constructing a separate confidence interval for the mean of each population. 95% CIs for the mean serum iron levels of healthy children and children with cystic fibrosis are displayed in Figure 6.1. In general, if the two intervals do not overlap, this suggests that the population means are in fact different. However, it should be kept in mind that this technique is not a formal test of hypothesis.

Note that the two samples of children were randomly selected from distinct normal population; in addition, the population variances are assumed to be equal. The two-sample t-test is the appropriate technique to apply. To carry out the test, we begin by calculating the pooled estimate of the variance

$$\begin{aligned} s_p^2 &= \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \\ &= \frac{(9-1)(5.9)^2+(13-1)(6.3)^2}{9+13-2} \\ &= 37.74. \end{aligned}$$

We next calculate the test statistic

$$\begin{aligned} \text{TS: } t &= \frac{|(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)|}{\sqrt{s_p^2[(1/n_1)+(1/n_2)]}} \\ &= \frac{|(18.9-11.9)-0|}{\sqrt{37.74[1/9+1/13]}} \\ &= 2.63. \end{aligned}$$

From Table A.4, we observe that for a t distribution with $df = n_1+n_2-2 = 9+$ 13 - 2 = 20, the total area under the curve to the right of $t_{20} = 2.63$ is between 0.005 and 0.01. Therefore, the p-value must be between 0.01 and 0.02. Since p is less than 0.05, we reject the null hypothesis at the 0.05 level of significance.

We may also construct a 95% CI for $\mu_1-\mu_2$. Note that for a t distribution with df=20, 95% of the observations lie between -2.086 and 2.086. As a results,

$$\text{P}\left(-2.086 \le \frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{s_p^2[(1/n_1)+(1/n_2)]}} \le 2.086\right) = 0.95.$$

Rearranging terms, we find that the 95% CI for $\mu_1 - \mu_2$ is

$$(\overline{x}_1 - \overline{x}_2) \pm (2.086)\sqrt{s_p^2[(1/n_1) + (1/n_2)]},$$

or

$$(18.9 - 11.9) \pm (2.086)\sqrt{(37.74)[(1/9) + (1/13)]}.$$

Therefore, we are 95% confident that the interval

$$(1.4, 12.6)$$

covers $\mu_1 - \mu_2$. This CI for the difference in means is mathematically equivalent to the two-sample test of hypothesis conducted at the 0.05 level.

6.2.3 Unequal Variances

When the variances of the two populations are not assumed to be equal, a modification of the two-sample t-test must be applied. Instead of using s_p^2 as an estimate of the common variance σ^2, we substitute s_1^2 for σ_1^2 and s_2^2 for σ_2^2. Therefore, the appropriate test statistic is

$$\begin{aligned} \text{TS: } t &= \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \\ &\sim t_\nu, \text{ under the null hypothesis,} \end{aligned}$$

where ν is calculated by

$$\nu = \frac{[(s_1^2/n_1) + (s_s^2/n_2)]^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]},$$

and the value of ν is rounded down to the nearest integer.

Example: Unequal Variances Suppose that we are interested in investigating the effect of an antihypertensive drug treatment on persons over the age of 60 who suffer from isolated systolic hypertension. Before the beginning of the study, subjects who had been randomly selected to take the active drug and those chosen to receive a placebo were comparable with respect to systolic blood pressure. After one year, the mean systolic blood pressure for patients receiving the drug is denoted by μ_1 and the mean for those receiving the placebo by μ_2. The standard deviations of the two populations are unknown, and we do not feel justified in assuming that they are equal.

A random sample is selected rom each of the two groups. The sample of $n_1 = 2308$ individuals receiving the active drug treatment has mean systolic blood pressure $\overline{x}_1 = 142.5$ mm Hg and standard deviation $s_1 = 15.7$ mm Hg; the sample of $n_2 = 2293$ persons receiving the placebo has mean $\overline{x}_2 = 156.5$ mm Hg and standard deviation $s_2 = 17.3$ mm Hg. We are interested in detecting differences that could occur in either direction, and thus conduct a two-sided test at the $\alpha = 0.05$ level of significance. The null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_A : \mu_1 \neq \mu_2.$$

Since the two groups were selected from independent normal distributions and the variances are not assumed to be equal, the modified two-sample test should be applied. In this case, the test statistic is

$$\begin{aligned} \text{TS: } t &= \frac{|(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)|}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \\ &= \frac{|(142.5 - 156.5) - 0|}{\sqrt{(15.7)^2/2308 + (17.3)^2/2293}} \\ &= 28.74, \end{aligned}$$

and the df ν is calculated by

$$\begin{aligned} \nu &= \frac{[(s_1^2/n_1) + (s_s^2/n_2)]^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]} \\ &= \frac{[(15.7)^2/2308 + (17.3)^2/2293]^2}{[(15.7)^2/2308]^2/(2308 - 1) + [(17.3)^2/2293]^2/(2293 - 1)} \\ &= 4550.5. \end{aligned}$$

Since a t distribution with df = 4550 is for all practical purposes identical to the standard normal distribution, we may consult either Table A.3 or Table A.4. In both cases, we find that p is less than 0.001. As a results, we reject the null hypothesis at the 0.05 level of significance.

Once again, we may also construct a 95% CI for $\mu_1 - \mu_2$. For a t distribution with df = 4550 or the standard normal distribution, 95% of the observations lie between -1.96 and 1.96.

$$\text{P}\left(-1.96 \leq \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \leq 1.96\right) = 0.95.$$

Rearranging terms, we find that the 95% CI for $\mu_1 - \mu_2$ is

$$(\overline{x}_1 - \overline{x}_2) \pm (1.96)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

or

$$(142.5 - 156.5) \pm (1.96)\sqrt{\frac{(15.7)^2}{2308} + \frac{(17.3)^2}{2293}}.$$

Therefore, we are 95% confident that the interval

$$(-15.0, -13.0)$$

covers $\mu_1 - \mu_2$. This interval does not contain the value 0, and is therefore consistent with the results of the modified two-sample test.

Chapter 7

Analysis of Variance

In the preceding chapter, we covered techniques for determining whether a difference exists between the means of two independent populations. It is not usual, however, to encounter situations in which we wish to test for differences among three or more independent means. The extension of the two-sample t-test to three or more samples is known as *analysis of variance* (ANOVA).

ANOVA is one of the most widely used statistical techniques in which the overall probability of making type I error is qual to some predetermined level α. If we were to evaluating all possible pairs of sample means using the two-sample t-test, not only the process is more complicated when the number of populations increases; but also this procedure is likely to lead to an incorrect conclusion. Suppose that three population means are equal and that we conduct all $\binom{3}{2} = 3$ pairwise tests. If we set the significance level for each test at 0.05 and assume the tests are independent, the probability of failing to reject a null hypothesis of no difference in all instances would be

P(reject in at least one test|three population means are equal)

$$
\begin{aligned}
&= 1 - \text{P(fail to reject in all three tests|three population means are equal)} \\
&= 1 - (1 - 0.05)^3 \\
&= 1 - (0.95)^3 \\
&= 1 - 0.857 \\
&= 0.143.
\end{aligned}
$$

Since we know that the null hypothesis is true in each case, 0.143 is the overall probability of making a type I error. This combined probability of a type I error for the set of three tests is much larger than 0.05.We would like to be able to use a testing procedure in which the overall probability of making a type I error is equal to the predetermined level α. The one-way ANOVA is such a technique.

7.1 One-Way Analysis of Variance

In the one-way ANOVA, we assume that data X_{ij} are observed according to a model

$$X_{ij} = \mu_i + \epsilon_{ij}, i = 1, \cdots, k, j = 1, \cdots, n_i, \tag{7.1}$$

where the μ_i are unknown parameters and the ϵ_{ij} are error random variables.

The term *one-way* indicates that there is a single factor or characteristic that distinguishes the various populations from each other.

7.2 One-Way ANOVA Assumptions

Here are the classic ANOVA assumptions for Model (7.1):

1) $\mathrm{E}(\epsilon_{ij}) = 0, \mathrm{var}(\epsilon_{ij}) = \sigma_i^2 < \infty$, for all i, j. $\mathrm{cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ for all i, i', j, and j' unless $i = i'$ and $j = j'$.

2) The ϵ_{ij} are independent and normally distributed (normal errors).

3) $\sigma_i^2 = \sigma^2$ for all i (equality of variance, aka *homoscedasticity*).

The equality of variance assumption is important and its importance is linked to the normality assumption. In general, if it is suspected that the data badly violate the ANOVA assumptions, a first course of attack is usually to try to nonlinearly transform the data. Under the one-way ANOVA assumptions, we have that

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2).$$

7.2.1 The Classical ANOVA

The classic ANOVA for Model (7.1) is a test of

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \text{ vs } H_A : \mu_i \neq \mu_j \text{ for some } i, j.$$

Alternatively, we can also consider k independent random samples of size n_k with values $\{x_{k1}, x_{k2}, \cdots, x_{kn_k}\}$ drawn from k different populations as below:

		Group 1	**Group 2**	$\cdots$	**Group** k
Population	Mean	μ_1	μ_2	$\cdots$	μ_k
	Standard Deviation	σ_1	σ_2	$\cdots$	σ_k
				$\cdots$	
Sample	Mean	$\overline{x}_1$	$\overline{x}_2$	$\cdots$	$\overline{x}_k$
	Standard Deviation	s_1	s_2	$\cdots$	s_k
	Sample Size	n_1	n_2	$\cdots$	n_k

Let $n = n_1 + n_2 + \cdots + n_k = \sum_{i=1}^{k} n_i$. Define the *grand mean* as

$$\begin{aligned} \overline{x} &= \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{k} n_i} \\ &= \frac{\sum_{i=1}^{k} n_i \overline{x}_i}{n}, \end{aligned}$$

where

$$\overline{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \text{ for } i = 1, 2, \cdots, k.$$

Sources of Variation When we work with several different populations with a common variance σ^2, the variability comes from two resources:

1) the variation of the individual values around their population means ("within-groups" variability);

2) the variation of the population means around the grand mean ("between-groups" variability).

Sum of squares of total variation The within-groups variability is defined as

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 = SSB + SSW.$$

The within-groups variability (s_W^2) can be computed as

$$\begin{aligned} s_W^2 &= \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n - k} \\ &= \frac{SSW}{n - k}. \end{aligned}$$

This the pooled estimate of the common variance σ^2, which is simply a weighted average of the k individual sample variances.

The between-groups variability (s_B^2) can be computed as

$$s_B^2 = \frac{\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2}{k - 1} = \frac{SSB}{k - 1}.$$

The test the null hypothesis that the population means are identical, we use the test statistic

$$\begin{aligned} \text{TS: } F &= \frac{s_B^2}{s_W^2} = \frac{\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2/(k-1)}{\sum_{i=1}^{k}(n_i - 1)s_i^2/(n-k)} \\ &\sim \mathcal{F}_{k-1,n-k}, \text{ under } H_0. \end{aligned}$$

We can construct an ANOVA table as

Source	DF	Sum of Squares	Mean Square	F Value
Between	$k-1$	$SSB = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2$	$s_B^2 = SSB/(k-1)$	s_B^2/s_W^2
Within	$n-k$	$SSW = \sum_{i=1}^{k}(n_i - 1)s_i^2$	$s_W^2 = SSW/(n-k)$	
Total	$n-1$	SST		

Under the null hypothesis, both s_B^2 and s_W^2 estimate the common variance σ^2, and F is close to 1. If there is a difference among population means, the between-groups variances exceeds the within-groups variance and F is greater than 1. If we have only two independent populations (i.e. $k = 2$), the $F - test$ reduces to the two-sample t-test (i.e. $\mathcal{F}_{1,n-2} = t_{n-2}^2$).

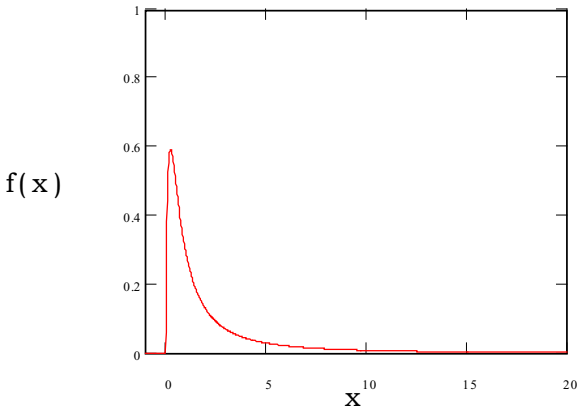
Table A.5 in Appendix A is a table of critical values computed for the family of F distribution (for selected percentiles only). For example, given an F distribution with $\text{df}_1 = 4$ and $\text{df}_2 = 2$ (i.e. $\mathcal{F}_{4,2}$), the table shows that $\mathcal{F}_{4,2} = 19.25$ cuts off the upper 5% of the curve. The F distribution cannot assume negative values. In addition, it is skewed to the right as shown in Figure 7.1 for two F distributions with different degrees of freedom.

Example For the study investigating the effects of CO exposure on patients with coronary artery disease sampled, the forced expiratory volume in one second (FEV_1) distributions of patients associated with each of three different medical centers make up district populations. Samples selected from each of the three centers are shown in Table 7.1.

We are interested in testing

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ vs } H_A : \mu_i \neq \mu_j \text{ for some } i, j.$$

(a) F distribution with 4 and 2 degrees of freedom



(b) F distribution with 2 and 4 degrees of freedom

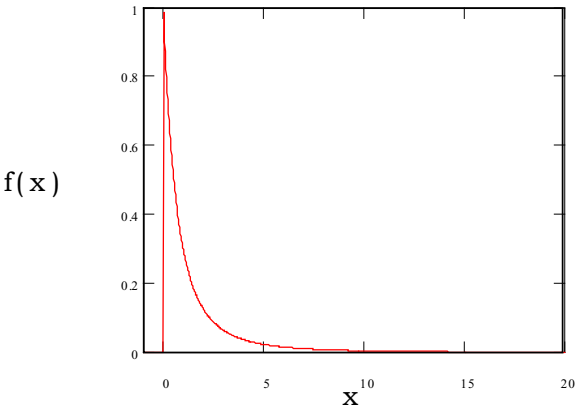


Figure 7.1: The F distributions: (a) $\mathcal{F}_{4,2}$ and (b) $\mathcal{F}_{2,4}$

Table 7.1: Forced expiratory volume in one second (FEV_1) for patients with coronary artery disease sampled at three different medical centers

Johns Hopkins	Rancho Las Amigos	Saint Louis
3.23	3.22	2.79
3.47	2.88	3.22
1.86	1.71	2.25
2.47	2.89	2.98
3.01	3.77	2.47
1.69	3.29	2.77
2.10	3.39	2.95
2.81	3.86	3.56
3.28	2.64	2.88
3.36	2.71	2.63
2.61	2.71	3.38
2.91	3.41	3.07
1.98	2.87	2.81
2.57	2.61	3.17
2.08	3.39	2.23
2.47	3.17	2.19
2.47		4.06
2.74		1.98
2.88		2.81
2.63		2.85
2.53		2.43
		3.20
		3.53
$n_1 = 21$	$n_2 = 16$	$n_3 = 23$
$\overline{x}_1 = 2.63$	$\overline{x}_2 = 3.03$	$\overline{x}_3 = 2.88$
$s_1 = 0.496$	$s_2 = 0.523$	$s_3 = 0.498$

We compute the within-groups variability (s_W^2) as

$$\begin{aligned} s_W^2 &= \frac{\sum_{i=1}^{3}(n_i-1)s_i^2}{n_1+n_2+n_3-3} \\ &= \frac{(21-1)(0.496)^2+(16-1)(0.523)^2+(23-1)(0.498)^2}{21+16+23-3} \\ &= 0.254. \end{aligned}$$

Since

$$\begin{aligned} \overline{x} &= \frac{\sum_{i=1}^{k} n_i\overline{x}_i}{n_1+n_2+n_3} \\ &= \frac{21(2.63)+16(3.03)+23(2.88)}{21+16+23} \\ &= 2.83, \end{aligned}$$

the estimate of the between-groups variability (s_B^2) is

$$\begin{aligned} s_B^2 &= \frac{\sum_{i=1}^{k} n_i(\overline{x}_i-\overline{x})^2}{k-1} \\ &= \frac{21(2.63-2.83)^2+16(3.03-2.83)^2+23(2.88-2.83)^2}{3-1} \\ &= 0.769. \end{aligned}$$

Therefore, the test statistic is

$$\text{TS: } F = \frac{s_B^2}{s_W^2} = \frac{0.769}{0.254} = 3.028.$$

We can construct the ANOVA table for FEV_1 as

Source	DF	Sum of Squares	Mean Square	F Value
Between-groups	2	1.538	0.769	$F = 3.028$
Within-groups	57	14.480	0.254	
Total	59	16.0183		

From Table A.5 for $F = 3.028 \sim \mathcal{F}_{k-1=3-1=2,n-k=60-3=57}$, we know $0.05 < p < 0.10$. Although we would reject H_0 at the 0.10 level, we do not reject H_0 at the 0.05 level.

7.3 Multiple Comparisons Procedures

One-way ANOVA may be used to test the hypothesis that k population means are identical or not,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \text{ vs } H_A : \mu_i \neq \mu_j \text{ for some } i, j.$$

What happens if we reject H_0? We don't know whether all the means are different from one another or only some of them are different. Once we reject the null hypothesis, we often want to conduct additional tests to find out where the differences lie.

Bonferroni method Many different techniques for conducting multiple comparisons exist; they typically involve testing each pair of means individually. As

it was noted earlier, performing multiple tests increases the probability of making type I error. We can avoid this problem by the *Bonferroni correction*, in which using a more conservative significance level for an individual comparison to ensure the overall level of significance at a predetermined level α. We should use

$$\alpha^* = \frac{\alpha}{\binom{k}{2}}$$

as the significance level for an individual comparison.

To conduct a test of

$$H_0 : \mu_i = \mu_j \text{ vs } H_A : \mu_i \neq \mu_j,$$

we calculate

$$\begin{aligned} \text{TS: } t_{ij} &= \frac{|\overline{x}_i - \overline{x}_j|}{\sqrt{s_W^2[(1/n_i) + (1/n_j)]}} \\ &\sim t_{n-k}, \text{ under } H_0. \end{aligned}$$

Note that instead of using the data from only two samples to estimate the common variance σ^2, we take advantage of the additional information that is available and use all k samples.

Example Let's use the example of FEV_1 for patients with coronary artery disease sampled at three different medical centers. For this case, we have $k = 3$

populations, a total of $\binom{3}{2} = 3$ pairwise tests are required. If we wish to set the overall level of significance at 0.10, we must use

$$\alpha^* = \frac{0.10}{\binom{3}{2}} = 0.033$$

as the level for each individual test. For

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_A : \mu_1 \neq \mu_2,$$

we calculate

$$\begin{aligned} \text{TS: } t_{12} &= \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{s_W^2[(1/n_1) + (1/n_2)]}} \\ &= \frac{|2.63 - 3.03|}{\sqrt{0.254[(1/21) + (1/16)]}} \\ &= 2.39. \end{aligned}$$

For a t distribution with $df = n - k = 60 - 3 = 57$, $p = 0.02$. Therefore, we reject the null hypothesis at the 0.033 level and conclude that μ_1 is not equal to μ_2.

One disadvantage of the Bonferroni multiple-comparisons procedures is that it can suffer from a lack of power. It is highly conservative and may fail to detect a difference in means that actually exists. However, there are many competing multiple-comparisons procedures that could be used instead. Two of the

most commonly used procedures are Fisher's (protected) least significant difference (LSD) and Tukey's honestly significant difference (HSD) methods. We will introduce another two procedures, e.g. Scheffé's and Dunnett's methods.

Fisher's (Protected) LSD Procedure If the overall F-test (which tests that hypothesis that all population means are equal) is statistically significant, we can safely conclude that not all of the population means are identical and then conduct the advanced multiple comparisons. For

$$H_0 : \mu_i = \mu_j \text{ vs } H_A : \mu_i \neq \mu_j,$$

we calculate

$$\text{LSD } = t_{n-k,\alpha/2} \times \sqrt{s_W^2[(1/n_i) + (1/n_j)]}.$$

If $|\overline{x}_i - \overline{x}_j| > \text{LSD}$, then we reject the null hypothesis $H_0 : \mu_i = \mu_j$. However, it is possible that we do not observe any statistically significant pairwise tests given a statistically significant overall F-test.

Tucky's HSD Method Tucky's HSD method is based on the *Studentized range distribution*. In any particular case, this Studentized range statistic $Q(k, n-k; \alpha)$ belongs to a sampling distribution defined by two parameters: the first is k, the number of samples in the original analysis; and the second is $n-k$, the number of degrees of freedom associated with the denominator of the F-test in the original analysis. For the present example of FEV_1, with $k = 3$ and $n - k = 57$, you will end up with $Q(3, 57; 0.05) = 3.41$. The Q values are tabulated in many other

textbooks or available online (e.g. http://cse.niaes.affrc.go.jp/miwa/probcalc/s-range).

The critical value HSD is calculated as

$$\text{HSD} \ = Q(k, n-k; \alpha) \times \sqrt{[s_W^2/2][(1/n_i) + (1/n_j)]}$$

If $|\overline{x}_i - \overline{x}_j| \geq \text{HSD}$, then we reject the null hypothesis $H_0 : \mu_i = \mu_j$.

Scheffé's Method Scheffé proposes another method for the multiple comparisons. Two means are declared significantly different if

$$\begin{aligned} \text{TS: } t_{ij} &= \frac{|\overline{x}_i - \overline{x}_j|}{\sqrt{s_W^2[(1/n_i) + (1/n_j)]}} \\ &> \sqrt{(k-1)\mathcal{F}_{k-1,n-k;\alpha}}, \end{aligned}$$

where $\mathcal{F}_{k-1,n-k;\alpha}$ is the α-level critical value of an F distribution with $k-1$ numerator degrees of freedom and $n-k$ denominator degrees of freedom. Scheffé's test is compatible with the overall ANOVA F-test in that Scheffé's method never declares a contrast (e.g. a pairwise test) significant if the overall F-test is non-significant.

Dunnett's method One special case of mean comparison is that in which the only comparisons that need to be tested are between a set of new treatments and a single control. In this case, you can achieve better power by using a method that is restricted to test only comparisons to the single control mean. Dunnett

proposes a test for this situation that declares the mean of group i significantly different from the control group c if

$$\begin{aligned} \text{TS: } t_{ic} &= \frac{|\overline{x}_i - \overline{x}_c|}{\sqrt{s_W^2[(1/n_i) + (1/n_c)]}} \\ &\geq d(k, n-k, \rho_1, \cdots, \rho_{k-1}; \alpha), \end{aligned}$$

where $d(k, n-k, \rho_1, \cdots, \rho_{k-1}; \alpha)$ is the critical value of the "many-to-one t statistic" for k means to be compared to a control group c, with $n-k$ numerator degrees of freedom and correlations ρ's, where $\rho_i = n_i/(n_c + n_i)$. The correlation terms arise because each of the treatment means is being compared to the same control.

Chapter 8

Contingency Tables

When working with nominal data that have been grouped into categories, we often arrange the counts in a tabular format known as a *contingency table*. In the simplest case, two dichotomous random variables are involved; the rows of the table represent the outcomes of one variable, and the columns represent the outcomes of the other.

8.1 The Chi-Square Test

8.1.1 r × c Tables

Consider the 2 × 2 table below. The data consist of a random sample of 793 individuals who are involved in bicycle accidents during a specified one-year period. The entries in the contingency table (17, 130, 218, and 428) are the observed counts within each combination of categories.

	Wearing Helmet		
Head Injury	**Yes**	**No**	**Total**
Yes	17	218	235
No	130	428	558
Total	147	646	793

To examine the effectiveness of bicycle safety helmets, we wish to know whether these is an association between the incidence of head injury and the use of helmets among individuals who have been involved in accidents. To determine this, we test

H_0: The proportion of persons suffering head injuries among the population of individuals wearing safety helmets at the accident is equal to the proportion of persons sustaining head injuries among those not wearing helmets

versus

H_A: The proportions of persons suffering head injuries are not identical in the two populations

at the $\alpha = 0.05$ level of significance. The first step in carrying out the test is to calculate the expected count for each cell of the contingency table, given that H_0 is true as the following:

	Wearing Helmet		
Head Injury	**Yes**	**No**	**Total**
Yes	$147 \times (235/793)$	$646 \times (235/793)$	235
No	$147 \times (558/793)$	$646 \times (558/793)$	558
Total	147	646	793

In general, if a 2 × 2 table of observed cell counts for a sample of size n as follows,

	Variable 2		
Variable 1	**Yes**	**No**	**Total**
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d=n

the corresponding table of expected counts is

	Variable 2		
Variable 1	**Yes**	**No**	**Total**
Yes	(a+c)(a+b)/n	(b+d)(a+b)/n	a+b
No	(a+c)(c+d)/n	(b+d)(c+d)/n	c+d
Total	a+c	b+d	n

These marginal totals have been held fixed by design; we calculate the cell entries that would have been expected given these is no association between the row and column classifications **and** that the number of individuals within each group remains constant.

The *chi-square test* compares the observed frequencies in each category of the contingency table (represented by O) with the expected frequencies in each category of the contingency table (represented by E) given the null hypothesis is true. It is used to determine whether the deviations between the observed and the expected counts, $O - E$, are too large to be attributed to chance. To perform

the test for the counts in a contingency table with r rows and c columns, we calculate

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i} \overset{a}{\sim} \chi^2_{(r-1)(c-1)},$$

where rc is the number of cells in the table. To ensure that the sample size is large enough to make this approximation valid, no cell in the table should have an expected count less than 1, and no more than 20% of the cells should have an expected count less than 5.

Like the F distribution, the chi-square distribution is not symmetric. A chi-square random variable cannot be negative; it assumes values from zero to infinity and is skewed to the right. The distributions with small degrees of freedom (df) are highly skewed as the number of df increases, the distributions become less skewed and more symmetric (see Figure 8.1). Table A.8 in Appendix A is a condensed table of areas for the chi-square distributions with various degrees of freedom. For a particular value of df, the entry in the table is the outcome of χ^2_{df} that cuts off the specified area in the **upper** tail of the distribution. For instance, $\chi^2_1 = 3.84$ cuts off the upper 5% of the area under the curve. (Please note that $\chi^2_1 = Z^2$, as $3.84 = 1.96^2$!)

We are using discrete observations to estimate χ^2, a continuous distribution. The approximation is quite good when the degrees of freedom are big. We often apply a continuity correction (*Yates correction*) for a 2×2 table as

$$\chi^2 = \sum_{i=1}^{4} \frac{(|O_i - E_i| - 0.5)^2}{E_i} \overset{a}{\sim} \chi^2_1.$$

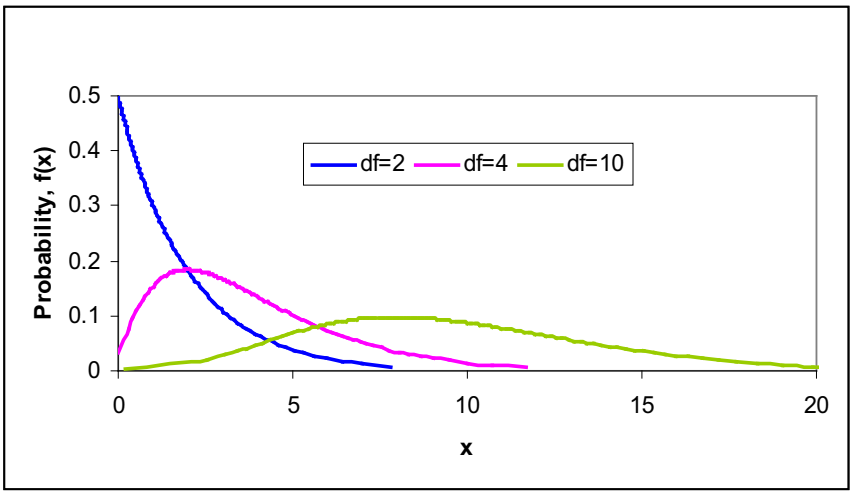


Figure 8.1: Chi-square distributions with 2, 4, and 10 degrees of freedom

The effect of the correction decreases the value of test statistic and this increases the corresponding p-value. Therefore, Yates correction may result in an overly conservative test that may fail to reject a false null hypothesis.

For the bicycle example we talked about earlier, if we apply the Yates correction, we would get

$$\begin{aligned}
\chi^2 &= \frac{(|17-43.6|-0.5)^2}{43.6} + \frac{(|130-103.4|-0.5)^2}{103.4} + \frac{(|218-191.4|-0.5)^2}{191.4} \\
&\quad + \frac{(|428-454.6|-0.5)^2}{454.6} \\
&= 15.62 + 6.59 + 3.56 + 15.0 \\
&= 27.27
\end{aligned}$$

From Table A.8, we obtain that the p-value is smaller than 0.001 as df=1. Since $p < \alpha$, we reject the null hypothesis and conclude that wearing a safety helmet at the accident is protective to the head injury.

For a 2 × 2 table in the general format shown below,

	Variable 2		
Variable 1	**Yes**	**No**	**Total**
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

the test statistic χ^2 can also be express as

$$\chi^2 = \frac{n[|ad - bc| - n/2]^2}{(a+b)(b+d)(a+c)(c+d)}.$$

Because these is no need to compute the expected counts, it is more convenient computationally.

When the sample size is small, one can use *Fisher's exact test* to obtain the exact probability of the occupance of the observed frequencies in the contingency table, given that there is no association between the rows and columns and that the marginal totals are fixed. The details of this test is not presented here because the computations involved can be arduous. However, many statistical softwares provide the Fisher's exact test as well as the Chi-square test in dealing with 2 × 2 tables.

8.2 McNemar's Test

If the data of interest in the contingency table are paired rather than independent, we use McNemar's test to evaluate hypotheses about the data. Consider a 2 × 2 table of observed cell counts about exposure status for a sample of n matched case-control *pairs* as follows,

	Controls		
Cases	**Exposed**	**Unexposed**	**Total**
Exposed	a	**b**	a+b
Unexposed	**c**	d	c+d
Total	a+c	b+d	n

To conduct McNemar's test, we calculate the test statistic

$$\chi^2 = \frac{[|b-c|-1]^2}{b+c} \overset{a}{\sim} \chi^2_1, \text{ with the continuity correction or}$$

$$\chi^2 = \frac{[|b-c|]^2}{b+c} \overset{a}{\sim} \chi^2_1, \text{ without the continuity correction,}$$

where b and c are the number of *discordant pairs*, i.e. the case-control pairs with different exposure status. In other words, the concordant pairs provide no information for testing.

Consider the following data taken from a study investigating acute myocardial infarction (MI, i.e. heart attack) among Navajos in the US. In this study, 144 MI cases were age- and gender-matched with 144 individuals free of heart disease.

The members of each pair then asked whether they had even been diagnosed with diabetes (here it is the 'exposure' of interest, not the disease outcome to define their case-control status). We like to know what these samples tell us about the proportions of diabetics in the two heart disease groups. The results are presented below,

	MI		
Diabetes	**Yes**	**No**	**Total**
Yes	46	25	71
No	98	119	217
Total	144	144	288

We cannot use the regular chi-square test for the matched data, as the chi-square test disregard the paired nature of the data. We must take the pairing into account in our analysis. Instead of testing whether the proportions of diabetes among individuals who have experienced acute MI is equal to the proportion of diabetes among those who have not, we test

H_0: There are equal numbers of pairs in which the victim of acute MI is a diabetic and the matched individual free of heart disease is not, and in which the person w/o MI is a diabetic but the individual who has experienced MI is not,

or, more concisely,

H_0: There is no association between diabetics and the occurrence of acute MI.

Therefore, the data should be presented in the following manner:

MI	No MI		Total
	Diabetes	**No diabetes**	
Diabetes	9	37	46
No diabetes	16	82	98
Total	25	119	144

The test statistic is

$$\chi^2 = \frac{[|37 - 16| - 1]^2}{37 + 16} = 7.55 \overset{a}{\sim} \chi^2_1,$$

with $0.001 < p < 0.01$. Since p is less than $\alpha = 0.05$, we reject the null hypothesis. For the given population of Navajos, we conclude that if there is a difference between individuals who experience infarction and those who do not, victims of acute MI are more likely to suffer from diabetes than the individuals free from heart disease who have been matched on age and gender.

8.3 The Odds Ratio

The chi-square test allows us to determine whether an association exists between two independent nominal variables, and McNemar's test does the same thing for paired dichotomous variables. However, neither test provides us a measure of the strength of the association. For a 2×2 table showing information on two independent dichotomous variables, the *odds ratio* or *relative odds* can be used to estimating the magnitude of the effect using the observations in a single sample.

If an event occurs with probability p, the *odds* in favor of the event are $p/(1-p)$ to 1. Conversely, if the odds in favor of an event are a to b, the probability that the event occurs is $a/(a+b)$.

If we have two dichotomous random variables that represent a disease and an exposure, the odds ratio (OR) is defined as the odds in favor of disease among exposed individuals divided by the odds in favor of disease among the unposed, i.e.

$$\text{OR} = \frac{\text{P}(disease|exposed)/[1-\text{P}(disease|exposed)]}{\text{P}(disease|unexposed)/[1-\text{P}(disease|unexposed)]}.$$

Alternatively, the OR may be defined as the odds of exposure among diseased individuals divided by the odds of exposure among the nondiseased, i.e.

$$\text{OR} = \frac{\text{P}(exposure|diseased)/[1-\text{P}(exposure|diseased)]}{\text{P}(exposure|nondiseased)/[1-\text{P}(exposure|nondiseased)]}.$$

Mathematically, these two expressions for the relative odds can be shown to be equivalent. Consequently, the OR can be estimated for both cohort and case-control studies.

Suppose that our data are again arranged in the form of a 2×2 contingency table in the general format shown below,

	Exposure		
Disease	**Yes**	**No**	**Total**
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

In this case, we have

$$\mathrm{P}(disease|exposed) = \frac{a}{a+c},$$

and

$$\mathrm{P}(disease|unexposed) = \frac{b}{b+d}.$$

Therefore, we can express an estimator of the OR as

$$\begin{aligned}
\widehat{\mathrm{OR}} &\equiv \frac{\mathrm{P}(disease|exposed)/[1-\mathrm{P}(disease|exposed)]}{\mathrm{P}(disease|unexposed)/[1-\mathrm{P}(disease|unexposed)]} \\
&= \frac{[a/(a+c)]/[c/(a+c)]}{[b/(b+d)]/[d/(b+d)]} \\
&= \frac{a/c}{b/d} = \frac{ad}{bc}.
\end{aligned}$$

This estimator is simply the cross-product ratio of the entries in the 2×2 table.

Please note the odds ratio (OR) is **not** the same as the relative risk. *Relative risk* (RR) is the probability of disease among exposed individuals divided by the probability of disease among the unexposed, i.e.

$$\begin{aligned}
\widehat{\mathrm{RR}} &\equiv \frac{\mathrm{P}(disease|exposed)}{\mathrm{P}(disease|unexposed)} \\
&= \frac{a/(a+c)}{b/(b+d)} = \frac{a(b+d)}{b(a+c)}.
\end{aligned}$$

When we are dealing with a rare disease, the RR can be approximated by the OR:

$$\begin{aligned} \widehat{\text{RR}} &= \frac{a(b+d)}{b(a+c)} \\ &\approx \frac{ad}{bc} \\ &= \widehat{\text{OR}}. \end{aligned}$$

Because the sampling distribution of $\widehat{\text{OR}}$ has better properties than those of $\widehat{\text{RR}}$, it is generally preferable to work with the relative odds.

8.3.1 The Confidence Interval for Odds Ratio

The cross-product ratio is simply a point estimate of the strength of association between two dichotomous variables. To gauge the uncertainty in this estimate, we must calculate a confidence interval (CI) as well; the width of the interval reflects the amount of variability in the estimator $\widehat{\text{OR}}$.

When computing a CI for the OR, we must make the assumption of normality. However, the probability distribution of the OR is skewed to the right, and the relative odds can asme any positive value between 0 and infinity. In contrast, the probability distribution of the natural logarithm of the OR, i.e. ln(OR), is more symmetric and approximately normal. Therefore, when calculating a CI for the OR, we typically work in the log scale. Besides, to ensure that the sample size is large enough, the expected value of each cell in the contingency table should be at least 5.

For a 2×2 contingency table in the general format shown below,

	Exposure		
Disease	**Yes**	**No**	**Total**
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

the expression of a 95% CI for the natural logarithm of the OR is

$$\ln \widehat{\mathrm{OR}} \pm 1.96 \widehat{\mathrm{SE}}(\ln \widehat{\mathrm{OR}}),$$

where $\widehat{\mathrm{SE}}(\ln \widehat{\mathrm{OR}})$, the standard error of $\ln \widehat{\mathrm{OR}}$, can be estimated by

$$\widehat{\mathrm{SE}}(\ln \widehat{\mathrm{OR}}) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

If any one of the entries in the contingency table is equal to 0, adding 0.5 to each of the values a, b, c, and d still provides a reasonable estimate. To find a 95% CI for the OR itself, we take the antilogarithm of the upper and lower limits of the interval for ln(OR) to get

$$\left(e^{\ln \widehat{\mathrm{OR}} - 1.96 \widehat{\mathrm{SE}}(\ln \widehat{\mathrm{OR}})}, e^{\ln \widehat{\mathrm{OR}} + 1.96 \widehat{\mathrm{SE}}(\ln \widehat{\mathrm{OR}})} \right).$$

An OR can be calculated to estimate the strength of association between two paired dichotomous variables. Using the notation introduced in Section 8.2, the OR is the ratio of the numbers of each type of discordant pair in the study. Consider a 2×2 table of observed cell counts about exposure status for a sample of n matched case-control *pairs* as follows,

	Controls		
Cases	**Exposed**	**Unexposed**	**Total**
Exposed	a	**b**	a+b
Unexposed	**c**	d	c+d
Total	a+c	b+d	n

the OR can be estimated by

$$\widehat{\mathrm{OR}} = \frac{b}{c},$$

and the estimated standard error of $\ln \widehat{\mathrm{OR}}$ is

$$\widehat{\mathrm{SE}}(\ln \widehat{\mathrm{OR}}) = \sqrt{\frac{b+c}{bc}}.$$

8.4 Berkson's Fallacy

Although the OR is a useful measure of the strength of association between two dichotomous random variables, it provides a valid estimate of the magnitude of the effect only if the sample of observations on which it is based is random. However, this point is sometimes forgotten. A restricted sample is usually much easier to obtain, and it might cause a problem known as *Berkson's fallacy*.

In one study, the investigator surveyed 2784 individuals – 257 of whom where hospitalized – and determined whether each subject suffered from a disease of the circulatory system of a respiratory illness or both. If they had limited their questioning to the 257 hospitalized patients only, the results would have been as follows:

	Respiratory Illness		
Circulatory Disease	**Yes**	**No**	**Total**
Yes	7	29	36
No	13	208	221
Total	20	237	257

An estimate of the relative odds of having respiratory illness among individuals who suffer from a disease of the circulatory system versus those who do not is

$$\begin{aligned} \widehat{\text{OR}} &= \frac{7 \times 208}{29 \times 13} \\ &= 3.86; \end{aligned}$$

the chi-square test of the null hypothesis that there is no association between the two diseases yields

$$\chi^2 = \sum_{i=1}^{rc} \frac{(|O_i - E_i| - 0.5)^2}{E_i} = 4.89 + 0.41 + 0.80 + 0.07 = 6.17,$$

with $0.01 < p < 0.025$ for a chi-square distribution with df=1. Therefore, we reject the null hypothesis and conclude that individuals who have a disease of the circulatory system are more likely to suffer from respiratory illness than individuals who do not.

Now consider the entire sample of 2784 individuals, which consist of both hospitalized and nonhospitalized subjects:

	Respiratory Illness		
Circulatory Disease	**Yes**	**No**	**Total**
Yes	22	171	193
No	202	2389	2591
Total	224	2560	2784

The estimated OR is

$$\begin{aligned} \widehat{\text{OR}} &= \frac{22 \times 2389}{171 \times 202} \\ &= 1.52; \end{aligned}$$

the chi-square test of the null hypothesis that there is no association between the two diseases yields $\chi^2 = \sum_{i=1}^{rc} \frac{(|O_i - E_i| - 0.5)^2}{E_i} = 2.67$ with $p > 0.10$ for a chi-square distribution with df=1. The estimated OR is much smaller and the chi-square test is not significant at the 0.05 level of significance.

Why do the conclusions drawn from these two samples differ so drastically? To answer this question, we must consider the rates of hospitalization that occur within each of the four disease subgroups:

	Respiratory Illness	
Circulatory Disease	**Yes**	**No**
Yes	7/22=31.8%	29/171=17.0%
No	13/202=6.4%	208/2389=8.7%

One can observe that individuals with both circulatory and respiratory disease are more likely to be hospitalized than individuals in any of the three other

subgroups. Also, subjects with circulatory disease are more likely to be hospitalized than those with respiratory illness. Therefore, the conclusions will be biased if we only sample patients who are hospitalized. In this case, we are more likely to select an individual who is suffering from both illnesses than a person in any of the other subgroups, and more likely to select a person with circulatory disease than one with respiratory problems. As a result, we observe an association that does not actually exist. This kind of spurious relationship among variables – which is evident only because of the way in which the sample was chosen – is known as Berkson's fallacy.

Chapter 9

Multiple 2 × 2 Contingency Tables

When the relationship between a pair of dichotomous random variables is being investigated, it is sometimes examined in two or more populations. As a result, the data to be analyzed consist of a number of 2 × 2 contingency tables. These tables more often are the results of a single investigation that have been subclassified, or *stratified*, by some factor that is believed to influence the outcome. It is possible to make inference about the relationship between the two variables by examining the association in each table separately. However, it is more useful to be able to combine the information across tables to make a single statement.

9.1 Simpson's Paradox

Simpson's paradox (or the Yule-Simpson effect) is a statistical paradox described by E. H. Simpson in 1951 and G. U. Yule in 1903, which refers to the reversal of results when several groups are combined together to form a single group.

In the simplest case, either the magnitude or the direction of the relationship between two variables is influenced by the presence of a third factor. The third factor is called a confounder. If we fail to control the effect of confounding,

the true magnitude of the association between exposure and disease will appear greater or less than it actually is. Consider the following data,

MALE			
	Smoker		
Disease	**Yes**	**No**	**Total**
Yes	37	25	62
No	24	20	44
Total	61	45	106

For males, the odds of developing the disease among smokers relative to non-smokers are estimated by

$$\widehat{\text{OR}_M} = \frac{37 \times 20}{25 \times 24} = 1.23$$

FEMALE			
	Smoker		
Disease	**Yes**	**No**	**Total**
Yes	14	29	43
No	19	47	66
Total	33	76	109

For females, the odds of developing the disease among smokers relative to nonsmokers are estimated by

$$\widehat{\text{OR}_F} = \frac{14 \times 47}{29 \times 19} = 1.19$$

We observe the same trend in each subgroup of the population. It is possible that these two are actually estimating the same population value. Consequently, we might attempt to combine the information as follows:

	Smoker		
Disease	**Yes**	**No**	**Total**
Yes	51	54	105
No	43	67	110
Total	94	121	215

For all individuals in the study, regardless of gender, the odds of developing the disease among smokers relative to nonsmokers are estimated by

$$\widehat{\text{OR}} = \frac{51 \times 67}{54 \times 43} = 1.47$$

If the effect of gender is ignored, the strength of the association between smoking and the disease appears greater than it is for either males or females alone. This phenomenon is an example of *Simpson's paradox.* In this case, gender is a confounder in the relationship between exposure and disease; failure to control for this effect has caused the true magnitude of the association to appear greater than it actually is.

9.2 The Mantel-Haenszel Method

In a study investigating the relationship between the consumption of caffeinated coffee and nonfatal myocardial infarction (MI) among adult males under the age of 55, two samples of men provide exposure and disease information.

Smokers			
	Coffee		
Disease	**Yes**	**No**	**Total**
Yes	1011	81	1092
No	390	77	467
Total	1401	158	1559

Among smokers, the odds of suffering the disease for males drinking caffeinated coffee versus males drinking no coffee at all are estimated by

$$\widehat{\text{OR}_S} = \frac{1011 \times 77}{390 \times 81} = 2.46$$

Nonsmokers			
	Coffee		
Disease	**Yes**	**No**	**Total**
Yes	383	66	449
No	365	123	488
Total	748	189	937

Among nonsmokers, the odds of suffering the disease for males drinking caffeinated coffee versus males drinking no coffee at all are estimated by

$$\widehat{\text{OR}_{NS}} = \frac{383 \times 123}{365 \times 66} = 1.96$$

	Coffee		
Disease	**Yes**	**No**	**Total**
Yes	1394	147	1541
No	755	200	955
Total	2149	347	2496

Based on the unstratified data, the estimate of odds of suffering the disease for males drinking caffeinated coffee versus males drinking no coffee at all is

$$\widehat{\text{OR}} = \frac{1394 \times 200}{755 \times 147} = 2.51$$

Again, this odds ratio (OR) is larger than the relative odds in either of the strata, suggesting that smoking is indeed a confounder.

9.2.1 Test of Homogeneity

Before combining the information in two or more contingency tables, we must first verify that the population odds ratios are in fact constant across the different strata. If they are not, it is not beneficial to compute a single summary value for the overall relative odds. Instead, it would be better to treat the data in the various contingency tables as if they had been drawn from distinct populations and report a stratum-specific odds ratio for each.

Consider a series of g 2×2 tables, where $g \geq 2$. For example, the ith 2×2 table is

	Exposure		
Disease	**Yes**	**No**	**Total**
Yes	a_i	b_i	N_{1i}
No	c_i	d_i	N_{2i}
Total	M_{1i}	M_{2i}	T_i

We could conduct a test of homogeneity to determine whether the strength of association between exposure and disease is uniform across the tables. The null hypothesis of the test of homogeneity is Ho: The population odds ratios (OR) for the g tables are identical, or , equivalently,

$$\text{Ho: } \text{OR}_1=\text{OR}_2=\cdots=\text{OR}_g.$$

The alternative hypothesis is that not all odds ratios are the same. The test statistic is calculated as

$$X^2 = \sum_{i=1}^{g} w_i(y_i - Y)^2 \sim \chi^2_{g-1} \text{ under Ho, where}$$

$$Y = \frac{\sum_{i=1}^{g} w_i y_i}{\sum_{i=1}^{g} w_i},$$

$$w_i = \left[\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right]^{-1},$$

and

$$y_i = ln(\widehat{\text{OR}_i}) = \ln\left(\frac{a_i d_i}{b_i c_i}\right).$$

If any one of the cell entries is equal to 0, w_i is undefined. In this case, we can add 0.5 to each of the cell entries, i.e.,

$$w_i = \left[\frac{1}{a_i + 0.5} + \frac{1}{b_i + 0.5} + \frac{1}{c_i + 0.5} + \frac{1}{d_i + 0.5}\right]^{-1},$$

If $p > \alpha$ we cannot reject Ho; therefore, we conclude that it is appropriate to combine the information in the g 2×2 tables using the Mantel-Haenszel method.

9.2.2 Summary Odds Ratio

If the odds ratios are found to be uniform across tables, the next step in the Mantel-Haenszel method is to compute an estimate of the overall strength of association. This estimate is actually a weighted average of the odds ratios for the g separate strata; it is calculated using the formula

$$\widehat{\text{OR}} = \frac{\sum_{i=1}^{g} a_i d_i / T_i}{\sum_{i=1}^{g} b_i c_i / T_i},$$

if the ith 2×2 table is

	Exposure		
Disease	**Yes**	**No**	**Total**
Yes	a_i	b_i	N_{1i}
No	c_i	d_i	N_{2i}
Total	M_{1i}	M_{2i}	T_i

A 95% confidence interval (CI) for the summary odds ratio is

$$\left(e^{Y-1.96\widehat{se}(Y)}, e^{Y+1.96\widehat{se}(Y)}\right),$$

where

$$\widehat{se}(Y) = \frac{1}{\sqrt{\sum_{i=1}^{g} w_i}}.$$

To ensure that our strata sample sizes are large enough to make the technique used valid, we recommend the following restriction on the expected values of the observations across the g tables:

$$\sum_{i=1}^{g} \frac{M_{1i}N_{1i}}{T_i} \geq 5,$$

$$\sum_{i=1}^{g} \frac{M_{1i}N_{2i}}{T_i} \geq 5,$$

$$\sum_{i=1}^{g} \frac{M_{2i}N_{1i}}{T_i} \geq 5,$$

and

$$\sum_{i=1}^{g} \frac{M_{2i}N_{2i}}{T_i} \geq 5.$$

9.2.3 Test of Association

The Mantel-Haenszel method combines information from two or more 2×2 contingency tables to compute a summary odds ratio (OR). We could test whether

this summary OR is equal to 1 by referring to the 95% CI for the summary OR. Another approach is to conduct a test of

$$Ho : \text{OR} = 1 \text{ vs. } Ha : \text{OR} \neq 1.$$

Again, consider the ith 2×2 table is

	Exposure		
Disease	**Yes**	**No**	**Total**
Yes	a_i	b_i	N_{1i}
No	c_i	d_i	N_{2i}
Total	M_{1i}	M_{2i}	T_i

The corresponding test statistic is

$$X^2 = \frac{[\sum_{i=1}^{g} a_i - \sum_{i=1}^{g} m_i]^2}{\sum_{i=1}^{g} \sigma_i^2} \sim \chi_1^2 \text{ under Ho, where}$$

$$m_i = \frac{M_{1i} N_{1i}}{T_i},$$

and

$$\sigma_i^2 = \frac{M_{1i} M_{2i} N_{1i} N_{2i}}{T_i^2 (T_i - 1)}.$$

9.2.4 Example

In the study investigating the relationship between the consumption of caffeinated coffee and nonfatal myocardial infarction (MI) among adult males under the age of 55, the data are stratified by their smoking status:

Smokers			
	Coffee		
Disease	**Yes**	**No**	**Total**
Yes	1011	81	1092
No	390	77	467
Total	1401	158	1559

Nonsmokers			
	Coffee		
Disease	**Yes**	**No**	**Total**
Yes	383	66	449
No	365	123	488
Total	748	189	937

The crude OR is

$$\widehat{\text{OR}} = \frac{1394 \times 200}{755 \times 147} = 2.51,$$

and stratum-specific ORs are $\widehat{\text{OR}_s} = 2.46$ and $\widehat{\text{OR}_{NS}} = 1.96$

First, to test the homogeneity with the null hypothesis that the population odds ratios for the two groups are identical, i.e.,

$$Ho : \text{OR}_S = \text{OR}_{NS} \text{ vs. } Ha : \text{OR}_S \neq \text{OR}_{NS},$$

at the $\alpha = 0.05$ level of significance. We have to compute several quantities involved in the test statistic

$$X^2 = \sum_{i=1}^{2} w_i (y_i - Y)^2.$$

$$w_1 = \left[\frac{1}{1011} + \frac{1}{390} + \frac{1}{81} + \frac{1}{77} \right]^{-1} = 34.62,$$

$$w_2 = \left[\frac{1}{383} + \frac{1}{365} + \frac{1}{66} + \frac{1}{123} \right]^{-1} = 34.93,$$

$$y_1 = ln(\widehat{\mathrm{OR}}_1) = \ln \left(\frac{a_1 d_1}{b_1 c_1} \right) = \ln \left(\frac{1011 * 77}{390 * 81} \right) = 0.900.$$

$$y_2 = ln(\widehat{\mathrm{OR}}_2) = \ln \left(\frac{a_2 d_2}{b_2 c_2} \right) = 0.673,$$

and

$$Y = \frac{\sum_{i=1}^{2} w_i y_i}{\sum_{i=1}^{2} w_i} = 0.786.$$

Finally, the test statistic is

$$X^2 = \sum_{i=1}^{2} w_i (y_i - Y)^2 = 0.896.$$

From Table A.8, we observe that for a chi-square distribution with 1 degree of freedom, $p > 0.10$. We cannot reject the null hypothesis. Consequently, we assume that the ORs for two strata are estimating the same quantity, and proceed

with the Mantel-Haenszel method of combining this information. The summary OR is estimated by

$$\begin{aligned}\widehat{\text{OR}} &= \frac{\sum_{i=1}^{2} a_i d_i / T_i}{\sum_{i=1}^{2} b_i c_i / T_i} \\ &= \frac{1011 * 77/1559 + 383 * 123/937}{390 * 81/1559 + 365 * 66/937} \\ &= 2.18.\end{aligned}$$

Once differences in smoking status have been taken into account, males under the age of 55 who drink caffeinated coffee have odds of experiencing nonfatal MI that are 2.18 times greater than the odds for males who do not drink coffee. In order to obtain the 95% CI of this summary OR, we need to find the estimated standard error of Y, which is a 95% confidence interval (CI) for the summary odds ratio is

$$\widehat{se}(Y) = \frac{1}{\sqrt{\sum_{i=1}^{2} w_i}} = \frac{1}{\sqrt{34.62 + 34.93}} = 0.120.$$

The 95% CI for the summary OR is

$$\left(e^{Y - 1.96\widehat{se}(Y)}, e^{Y + 1.96\widehat{se}(Y)}\right) = (1.73, 2.78).$$

After adjusting for the effects of smoking, we are 95% confident that men who drink caffeinated coffee have odds of experiencing nonfatal MI that are between 1.73 and 2.78 times greater than the odds for men who do not drink coffee.

We would like to evaluate the null hypothesis

$$Ho : \text{OR} = 1$$

against the alternative

$$Ha : \text{OR} \neq 1$$

using a two-sided test and setting the significance level at $\alpha = 0.05$. Note that

$$\begin{aligned}
a_1 &= 1011, \\
m_1 &= \frac{M_{11}N_{11}}{T_1} = \frac{1401 * 1092}{1559} \\
&= 981.3, \\
\sigma_1^2 &= \frac{M_{11}M_{21}N_{11}N_{21}}{T_1^2(T_1 - 1)} \\
&= \frac{1401 * 158 * 1092 * 467}{(1559)^2(1559 - 1)} \\
&= 29.81, \\
a_2 &= 383, \\
m_2 &= \frac{M_{12}N_{12}}{T_2} = \frac{748 * 449}{937} \\
&= 358.4, \\
\sigma_2^2 &= \frac{M_{12}M_{22}N_{12}N_{22}}{T_2^2(T_2 - 1)} \\
&= \frac{748 * 189 * 449 * 488}{(937)^2(937 - 1)} \\
&= 37.69.
\end{aligned}$$

Therefore, the test statistic is

$$\begin{aligned} X^2 &= \frac{[\sum_{i=1}^{2} a_i - \sum_{i=1}^{2} m_i]^2}{\sum_{i=1}^{2} \sigma_i^2} \\ &= \frac{[(1011 + 383) - (981.3 + 358.4)]^2}{29.81 + 37.69} \\ &= 43.68. \end{aligned}$$

From Table A.8, we obtain the corresponding $p-$value is less than 0.001, so we reject the null hypothesis and conclude the summary OR is not equal to 1. After adjusting for smoking, males under the age of 55 who drink caffeinated coffee face a significantly higher risk of experiencing nonfatal MI than males who do not drink coffee.

Chapter 10

Logistic Regression

When studying linear regression, we attempted to estimate a population regression equation

$$\mu_{y|x_1,x_2,\cdots,x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

by fitting a model of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon.$$

The response Y was continuous and was assumed to follow a normal distribution. We were concerned with predicting or estimating the mean value of the response corresponding to a given set of values for the explanatory variables.

However, the response of interest sometimes is dichotomous rather than continuous. In general, the value 1 is used to represent a "success" (i.e. the outcome we are most interested in) and 0 represents a "failure." Just as we estimate the mean value of the response Y when it was continuous, we would like to be able to estimate the probability p associated with a dichotomous response for various values of explanatory variables. To do this, we use a technique known as *logistic regression*.

10.1 The Model

For simple logistic regression (i.e. consider single explanatory variable x), our first strategy might be to fit a straightforward model of the form

$$p = \alpha + \beta x.$$

This is simply the standard simple linear regression model in which the continuous and normally distributed response y has been replaced by a probability p. However, this model is not feasible. Since p is a probability, it must be in the range of 0 and 1. The term $\alpha + \beta x$, in contrast, could easily yield a value that lies outside this range.

We might try to solve this problem by instead fitting another model

$$p = e^{\alpha+\beta x}.$$

This equation guarantees that the estimate of p is positive; however, this model is also unsuitable. Although the term $e^{\alpha+\beta x}$ cannot produce a negative value, it can result in a value that is greater than 1.

To accommodate these problems, we fit a model of the form

$$p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} (= \frac{1}{1 + e^{-(\alpha+\beta x)}}).$$

The expression on the right, called a *logistic function*, restricts the estimated value of p to the required range as illustrated in Figure 10.1.

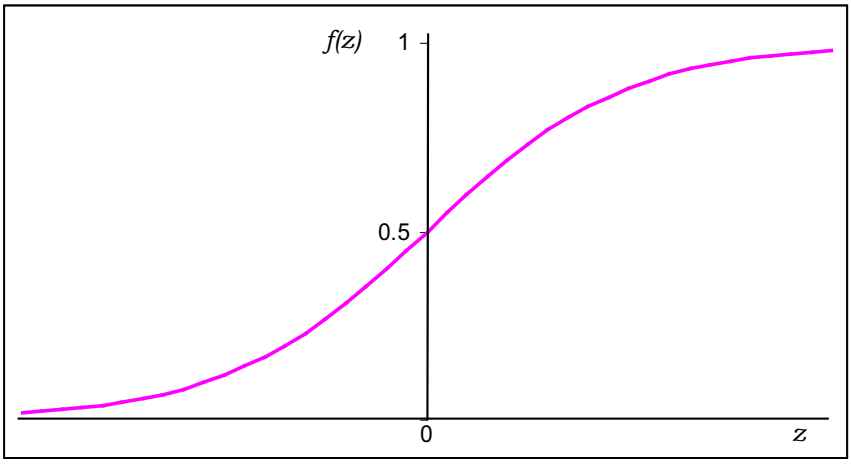


Figure 10.1: The logistic function $f(z) = \frac{1}{1+e^{-z}}$

The odds in favor of success are

$$\begin{aligned} \frac{p}{1-p} &= \frac{e^{\alpha+\beta x}/(1+e^{\alpha+\beta x})}{1/(1+e^{\alpha+\beta x})} \\ &= e^{\alpha+\beta x} \end{aligned}$$

Taking the natural logarithm of each side of this equation,

$$\begin{aligned} \text{logit}(p) = \ln\left[\frac{p}{1-p}\right] &= \ln[e^{\alpha+\beta x}] \\ &= \alpha + \beta x. \end{aligned}$$

Therefore, modeling the probability p with a logistic function is equivalent to fitting a linear regression in which the response is the logarithm of the odds of success for a dichotomous random variable. We assume that the relationship between logit(p)=ln$[p/(1-p)]$ and x is linear.

10.1.1 The Fitted Equation

Using a sample, we fit the model

$$\ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] = \widehat{\alpha} + \widehat{\beta}x.$$

As with a linear regression model, $\widehat{\alpha}$ and $\widehat{\beta}$ are estimates of the population coefficients. However, we cannot apply the method of least squares, which assumes that the response is continuous and normally distributed, to fit a logistic model. Instead, we use maximum likelihood estimation (MLE). Ordinary least squares seeks to minimize the sum of squared distances of the data points to the regression equation. MLE seeks to maximize the log likelihood, which reflects how likely it is (the odds) that the observed values of the response may be predicted from the observed values of the explanatory variable(s). The closed-form solutions do not exist, so the estimation of parameters must be obtained by an iterative procedure.

In order to estimate the probability p developing the outcome/disease for a given x, we find the predicted

$$\ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] = \widehat{\alpha} + \widehat{\beta}x$$

for the given x and take antilogarithm of each side of the equation. Then, we have

$$\widehat{p} = \frac{e^{\widehat{\alpha}+\widehat{\beta}x}}{1+e^{\widehat{\alpha}+\widehat{\beta}x}}.$$

In order to test

$$H_0 : \beta = 0 \text{ vs } H_A : \beta \neq 0,$$

we need to know the standard error of the estimator $\widehat{\beta}$. If H_0 is true, the test statistic (for the Wald test) is

$$\text{TS: } z = \frac{\widehat{\beta}}{\widehat{\text{se}}(\widehat{\beta})} \sim \mathcal{N}(0,1).$$

10.2 Multiple Logistic Regression

When we consider more explanatory variables, we extend the model and procedures we learned from the previous section for the simple logistic regression. For example, to model the probability p as a function of two explanatory variables, we fit a model of the form

$$\ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] = \widehat{\alpha} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2.$$

In order to estimate the probability p developing the outcome/disease for given x_1 and x_2, we find the predicted

$$\ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] = \widehat{\alpha} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$$

for the given x's and take antilogarithm of each side of the equation. Then, we have

$$\widehat{p} = \frac{e^{(\widehat{\alpha}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2)}}{1+e^{(\widehat{\alpha}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2)}}.$$

10.3 Indicator Variables

Like the linear regression model, the logistic regression model can be generalized to include discrete or nominal explanatory variables in addition to continuous ones. Suppose x is the outcome of a dichotomous random variable indicating the exposure status (0: no and 1: yes) and we fit the model

$$\ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] = \widehat{\alpha} + \widehat{\beta}x.$$

The estimated coefficient β has a special interpretation. In this case, the antilogarithm of β, i.e. $e^{\beta} = \exp(\beta)$, is the estimated odds ratio (OR) of the disease for the two possible levels of x.

For the following data as a 2×2 contingency table for the relationship between myocardial infarction (MI) and smoking status.

	Smoke		
MI	**Yes**	**No**	**Total**
Yes	1092	449	1541
No	467	488	955
Total	1559	937	2496

The OR estimated by computing the cross-product of the entries in the table,

$$\widehat{\text{OR}} = \frac{(1092)(488)}{(449)(467)} = 2.54.$$

Suppose we fit the model

$$\ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] = \widehat{\alpha} + \widehat{\beta}_1 \times \text{smoke},$$

where smoke is the indicator variable for smoking status (0=no, 1=yes). The equation estimated from the sample is

$$\ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] = -0.0833 + 0.9326 \times \text{smoke}.$$

Then, the relative odds of developing MI for smokers versus non-smokers is

$$\widehat{\text{OR}} = e^{\widehat{\beta_1}} = e^{0.9326} = 2.54,$$

which is identical to the number obtained from the contingency table.

A confidence interval (CI) for the OR can be calculated from the model by computing a confidence interval for the coefficient β_1 and taking the antilogarithm of this upper and lower limits. If $\widehat{\text{se}}(\widehat{\beta_1}) = 0.0856$, a 95% CI for β is

$$(0.9326 - 1.96(0.0856), 0.9326 + 1.96(0.0856)) = (0.7648, 1.100).$$

Therefore,a 95% CI for the relative odds of developing MI is

$$(e^{0.7648}, e^{1.100}) = (2.15, 3.00).$$

A test of the null hypothesis

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0,$$

the corresponding test statistic is

$$\begin{aligned} \text{TS: } z &= \frac{|\widehat{\beta}_1|}{\widehat{\text{se}}(\widehat{\beta}_1)} \\ &= 0.9326/0.0856 \\ &= 10.8909, \end{aligned}$$

with a p-value < 0.0001.

Besides, the estimated probability of developing MI for smokers is

$$\begin{aligned} \widehat{p} &= \frac{e^{-0.0833+0.9326(1)}}{1+e^{-0.0833+0.9326(1)}} \\ &= 0.70 \\ &= 1092/1541. \end{aligned}$$

Suppose we add a second dichotomous explanatory variable indicating whether drinking coffee or not (variable name: coffee (0=no, 1=yes)) to the model that already contains the smoking status.

$$\begin{aligned} \ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] &= \widehat{\alpha} + \widehat{\beta}_1 \times \text{smoke} + \widehat{\beta}_2 \times \text{coffee} \\ &= -0.7168 + 0.8687 \times \text{smoke} + 0.7863 \times \text{coffee}. \end{aligned}$$

The coefficient of smoke changes very little with the inclusion of the second explanatory variable. In this case, the adjusted OR for smoke is

$$\widehat{\text{OR}} = e^{\widehat{\beta_1}} = e^{0.8687} = 2.38.$$

The estimated coefficient of coffee drinking is positive, implying that the probability of developing MI is higher for coffee drinkers.

If we want to determine whether smoking has a different effect on the probability of developing MI depending on coffee consumption, it would be necessary to include in the logistic regression model an interaction term (i.e. int = smoke $\times$ coffee) that is the product of the two dichotomous explanatory variables as

$$\begin{aligned} \ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] &= \widehat{\alpha} + \widehat{\beta}_1 \times \text{ smoke} + \widehat{\beta}_2 \times \text{coffee} + \widehat{\beta}_3 \times \text{int} \\ &= -0.6225 + 0.6732 \times \text{smoke} + 0.6707 \times \text{coffee} + 0.2312 \times \text{ int}. \end{aligned}$$

Below is the SAS output from PROC LOGISTIC:

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.6225	0.1526	16.6459	<.0001
coffee	1	0.6707	0.1692	15.7093	<.0001
smoke	1	0.6732	0.2205	9.3216	0.0023
int	1	0.2312	0.2398	0.9296	0.3350

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
coffee	1.956	1.404	2.725
smoke	1.960	1.273	3.020
int	1.260	0.788	2.016

Discussion

1) How do you interpret $\widehat{\beta}_3 = 0.2312$? We are not interested in the number itself but its significance (although it tells the direction of interaction).

2) What is the OR for a smoker who drinks coffee versus a non-smoker who doesn't drink coffee? What is the CI for the OR? Is there an easier way to get it?

There are four categories, each with odds given by

Category	smoke (x_1)	coffee (x_2)	$x_1 \times x_2$	Odds
1	No (0)	No (0)	0	$\exp(\alpha)$
2	No (0)	Yes (1)	0	$\exp(\alpha + \beta_2)$
3	Yes (1)	No (0)	0	$\exp(\alpha + \beta_1)$
4	Yes (1)	Yes (1)	1	$\exp(\alpha + \beta_1 + \beta_2 + \beta_3)$

Therefore, the odds ratio of each category versus the first, designated as $\text{OR}_{(x_1,x_2):(0,0)}$, is presented in the cells of the following table

$\mathrm{OR}_{(x_1,x_2):(0,0)}$

	$x_2 = 0$	$x_2 = 1$	$\mathrm{OR}_{(x_1,1):(x_1,0)}$
$x_1 = 0$			
$x_1 = 1$			

$\mathrm{OR}_{(1,x_2):(0:x_2)}$

The alternative way to conduct the same analysis is to create a series of dummy variables. Because there are four categories, we need to generate three dummy variables with one category as the reference group. For example,

Category	smoke	coffee	dummy1	dummy2	dummy3
1	No (0)	No (0)	0	0	0
2	No (0)	Yes (1)	1	0	0
3	Yes (1)	No (0)	0	1	0
4	Yes (1)	Yes (1)	0	0	1

The result of the new logistic regression becomes

$$\ln\left[\frac{\widehat{p}}{1-\widehat{p}}\right] = -0.6225 + 0.6707 \times \text{dummy1} + 0.6732 \times \text{dummy2} + 1.5751 \times \text{dummy3},$$

and the new SAS output is

```
The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates
```

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.6225	0.1526	16.6459	<.0001
dummy1	1	0.6707	0.1692	15.7093	<.0001
dummy2	1	0.6732	0.2205	9.3216	0.0023
dummy3	1	1.5751	0.1638	92.4497	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
dummy1	1.956	1.404	2.725
dummy2	1.960	1.273	3.020
dummy3	4.831	3.504	6.660

Is 1.5751 equal to $\beta_1 + \beta_2 + \beta_3$ from the previous logistic regression model? Moreover, we can easily get the 95% CI for the comparison of interest.

10.4 Multicollinearity

Multicollinearity in logistic regression models is a result of strong correlations between independent variables. The existence of multicollinearity inflates the variances of the parameter estimates. That may result, particularly for small and moderate sample sizes, in lack of statistical significance of individual independent variables while the overall model may be strongly significant. Multicollinearity may also result in wrong signs and magnitudes of regression coefficient estimates, and consequently in incorrect conclusions about relationships between independent and dependent variables. What to do about multicollinearity? In some

cases, variables involved in multicollinearity can be combined into a single variable. If combining variables does not make sense, then some variables causing multicollinearity need to be dropped from the model. Examining the correlations between variables and taking into account practical aspects and importance of the variables help in making a decision what variables to drop from the model.

Chapter 11

Correlation

When we investigate the relationships that can exist among continuous variables, one statistical technique often employed to measure such an association is known as *correlation analysis*. *Correlation* is defined as the quantification of the degree to which two random variables are related, provided that the relationship is **linear**.

11.1 The Two-Way Scatter Plot

Before we conduct the analysis, we should always create a two-way scatter plot of the data. A two-way scatter plot can be used to depict the relationship between two different continuous measurements. We place the outcomes of X variable along the horizontal axis, and the outcomes of X variable along the vertical axis, such that each point on the graph represent a pair of values (x_i, y_i). We can often determine whether a relationship exists between x and y – the outcomes of the random variables X and Y.

For example, we wish to investigate the relationship between the percentage of children who have been immunized against the infectious disease diphtheria, pertussis, and tetanus (DPT) in a given country and the corresponding mortality rate for children under five years of age in that country (data shown in Table 11.1

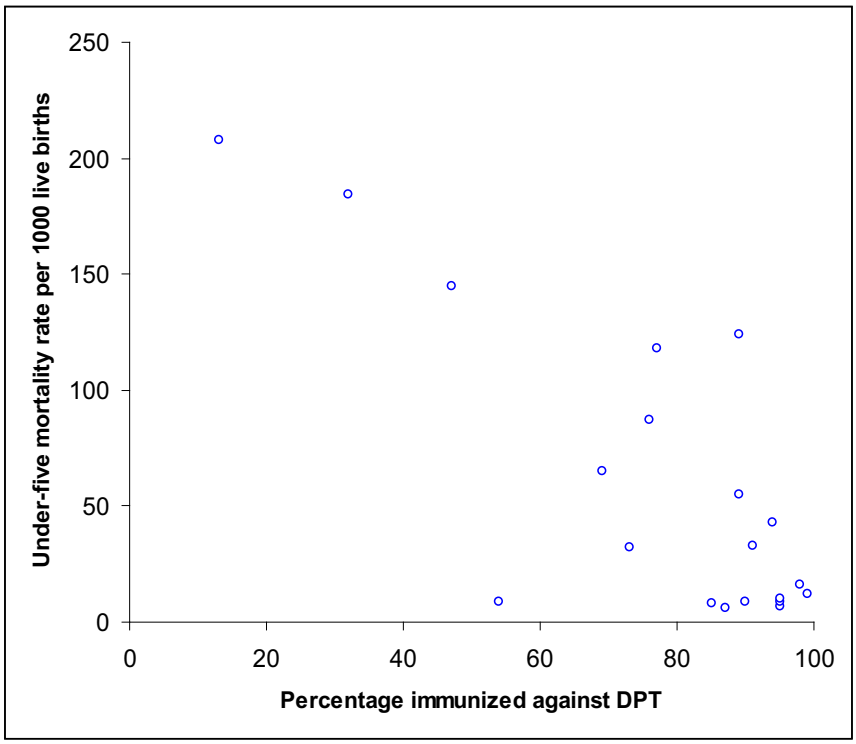


Figure 11.1: Under-five mortality rate versus percentage of children immunized against DPT for 20 countries, 1992

and plotted in Figure 11.1). Not surprisingly, the mortality rate tens to decreases as the percentage of children immunized increases.

11.2 Pearson's Correlation Coefficient

In the underlying population from which the sample of points (x_i, y_i) is selected, the correlation between the random variables X and Y is denoted by the Greek letter ρ. The correlation quantifies the strength of the linear relationship between the outcomes x and y,

$$\rho = \text{average}\left[\frac{X-\mu_x}{\sigma_x}\frac{Y-\mu_y}{\sigma_y}\right],$$

Table 11.1: Percentage of children immunized against diphtheria, pertussis, and tetanus (DPT) and under-five mortality rate for 20 countries, 1992

Nation	Percentage immunized	Mortality rate (per 1000)
Bolivia	77	118
Brazil	69	65
Cambodia	32	184
Canada	85	8
China	94	43
Czech Republic	99	12
Egypt	89	55
Ethiopia	13	208
Finland	95	7
France	95	9
Greece	54	9
India	89	124
Italy	95	10
Japan	87	6
Mexico	91	33
Poland	98	16
Russian Federation	73	32
Senegal	47	145
Turkey	76	87
United Kingdom	90	9

which can be thought of as the average of the product of the standard normal deviation of $X$and Y. The estimator of the population correlation is known as *Pearson's coefficient of correlation*, or simply the *correlation coefficient*. The correlation coefficient is denoted by r,

$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y} \\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{[\sum_{i=1}^{n}(x_i-\overline{x})^2][\sum_{i=1}^{n}(y_i-\overline{y})^2]}}. \end{aligned}$$

The correlation coefficient is a dimensionless number. The maximum value of r can achieve is 1, and its minimum value is -1. Therefore, for any given set of observations, $-1 \leq r \leq 1$. When these is an exact linear relationship between x and y, the values $r = 1$ or -1, as shown in Figure 11.2 (a) and (b). If y tends to increase as x increases, $r > 0$ and x and y are said to be *positively correlated*; if y decreases as x increases, $r < 0$ and x and y are *negatively correlated*. If $r = 0$, there is no linear relationship between x and y and the variables are *uncorrelated*, as shown in Figure 11.2 (c) and (d). However, a nonlinear relationship may exist, e.g. a quadratic relationship in Figure 11.2 (d).

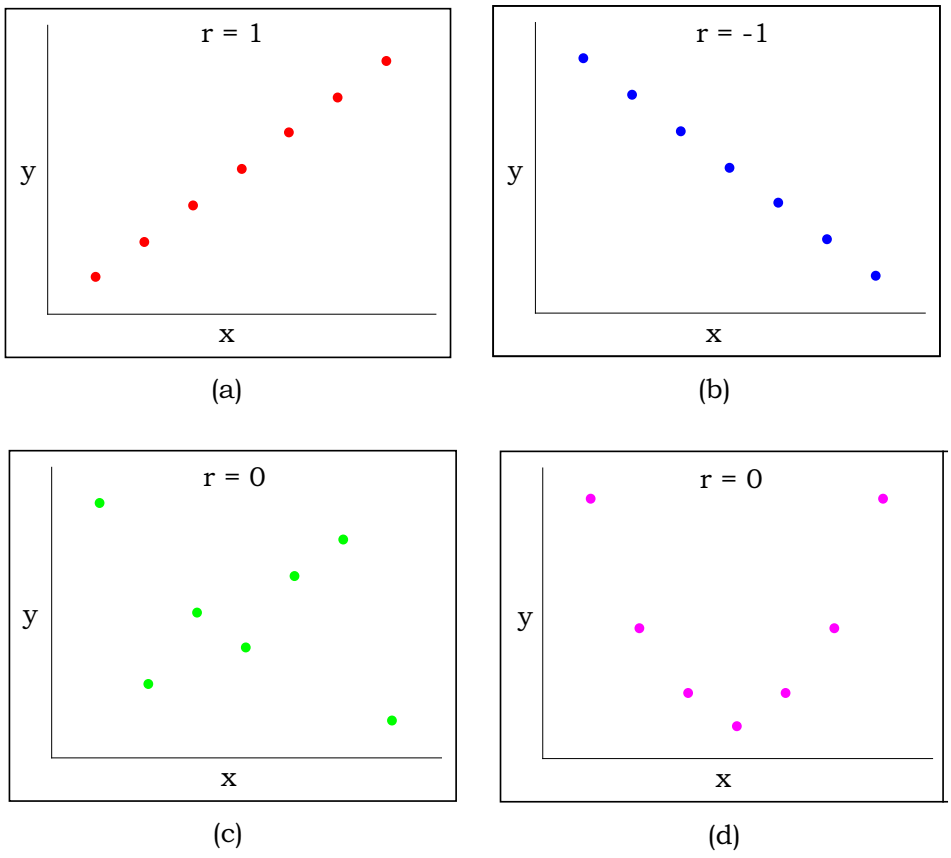


Figure 11.2: Scatter plots showing possible relationship between X and Y

Example For the data in Table 11.1, the mean percentage of children immunized against DPT is

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{20}\sum_{i=1}^{20} x_i \\ &= 77.4\%,\end{aligned}$$

and the mean value of the under-five mortality rate is

$$\begin{aligned}\overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{20}\sum_{i=1}^{20} y_i \\ &= 59.0 \text{ per } 1000 \text{ live births.}\end{aligned}$$

In addition,

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) &= \sum_{i=1}^{20}(x_i-77.4)(y_i-59.0) \\ &= -22706, \\ \sum_{i=1}^{n}(x_i-\overline{x})^2 &= \sum_{i=1}^{20}(x_i-77.4)^2 \\ &= 10630.8, \text{ and} \\ \sum_{i=1}^{n}(y_i-\overline{y})^2 &= \sum_{i=1}^{20}(y_i-59.0)^2 \\ &= 77498.\end{aligned}$$

As a result, the coefficient of correlation is

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \overline{x})^2][\sum_{i=1}^{n}(y_i - \overline{y})^2]}} \\ &= \frac{-22706}{\sqrt{(10630.8)(77498)}} \\ &= -0.79. \end{aligned}$$

As illustrated in Figure 11.1, these appears to be a strong linear relationship between the percentage of children immunized against DPT in a specified country and its under-five mortality rate; the correlation coefficient is fairly close to its minimum value of -1. Since r is negative, mortality rate decreases as percentage of immunization increases. Please note that the correlation coefficient only tells us that a linear relationship exists between two variables; it does not specify whether the relationship is cause-and-effect.

11.2.1 Hypothesis Testing

We would like to draw conclusions about the unknown population correlation ρ using the sample correlation coefficient r. Most frequently, we are interested in determining whether any correlation exists between the random variables X and Y, i.e.

$$H_0 : \rho = 0 \text{ vs } H_A : \rho \neq 0.$$

As before, we need to find the probability of obtaining a sample correlation coefficient as extreme as , or more extreme than, the observed r given that the null hypothesis is true. The appropriate test statistic is

$$\begin{aligned} \text{TS: } t &= \frac{|r-0|}{\sqrt{(1-r^2)/(n-2)}} \\ &= |r|\sqrt{\frac{n-2}{1-r^2}} \\ &\sim t_{n-2}, \text{ if } H_0 \text{ is true,} \end{aligned}$$

where

$$\widehat{SE}(r) = \sqrt{(1-r^2)/(n-2)}$$

is the estimated standard error of r.

Example Suppose we like to know if a linear relationship exists between the percentage of children immunized against DPT and the under-five mortality rate. We conduct a two-sided test such as

$$H_0 : \rho = 0 \text{ vs } H_A : \rho \neq 0,$$

at the α=0.05 level of significance. Recall that we previously found r=-0.79. The test statistic is

$$\begin{aligned} \text{TS: } t &= |r|\sqrt{\frac{n-2}{1-r^2}} \\ &= |-0.79|\sqrt{\frac{20-2}{1-(-0.79)^2}} \\ &= 5.47. \end{aligned}$$

From Table A.4, we observed that for a t distribution with 18 degrees of freedom, $p < 2(0.0005)$=0.001. Therefore, we reject the null hypothesis at the 0.05 level. Based on this sample, there is evidence that the true population correlation is different from 0. However, neither the percentage of children immunized (which is skewed to the left) nor the under-five mortality rate (which is skewed to the right) is normally distributed. Therefore, the hypothesis testing procedure performed above cannot be assumed to be accurate for these data.

11.2.2 One-Sample Z-Test for Correlation Coefficient

Sometimes the correlation between two random variables are expected to be some quantity ρ_0 other than zero. In this case, we want to test the hypothesis

$$H_0 : \rho = \rho_0 \text{ vs } H_A : \rho \neq \rho_0.$$

The problem of using the t-test is that the sample correlation coefficient r has a skewed distribution for nonzero ρ_0. One way to conduct the more general hypothesis is by means of Fisher's z-transformation. The z-transformation of sample correlation coefficient is given by

$$\begin{aligned} W &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \\ &\sim \mathcal{N}\left(\frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right), \frac{1}{n-3}\right), \text{ if } H_0 \text{ is true.} \end{aligned}$$

The appropriate test statistic (i.e. normalization of W) is

$$\begin{aligned} \text{TS: } Z &= \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}} \\ &\sim \mathcal{N}(0,1), \text{ if } H_0 \text{ is true.} \end{aligned}$$

Then, we can find the p-value of the test with the help of Table A.3.

11.2.3 Limitations of Coefficient of Correlation

The correlation coefficient r has several limitations:

1) r quantifies only the strength of the **linear** relationship,

2) r is highly sensitive to extreme values,

3) the estimated correlation should never be extrapolated beyond the observed ranges of the variables,

4) a high correlation between two variables does not imply a cause-and-effect relationship.

11.3 Spearman's Rank Correlation Coefficient

Pearson's coefficient of correlation is very sensitive to outlying values. One more robust approach is to rank the two sets of outcomes x and y separately and calculate a coefficient of rank correlation, Spearman's rank correlation coefficient, denoted r_s, which may be classified as a non-parametric method.

Spearman's rank correlation coefficient is simply Pearson's correlation coefficient r calculated for the ranked values of x and y,

$$r_s = \frac{\sum_{i=1}^{n}(x_{ri} - \overline{x}_r)(y_{ri} - \overline{y}_r)}{\sqrt{[\sum_{i=1}^{n}(x_{ri} - \overline{x}_r)^2][\sum_{i=1}^{n}(y_{ri} - \overline{y}_r)^2]}},$$

where x_{ri} and y_{ri} are the ranks associated with the ith subject rather than the actual observations. An equivalent method for computing r_s is provided by the formula

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \text{<EXTRA CREDIT: show this is true>}$$

where d_i is the difference between the rank of x_i and the rank of y_i. Like the Pearson's correlation coefficient, $-1 \leq r_s \leq 1$. r_s=0 implies a lack of linear association between the two variables.

11.3.1 Hypothesis Testing

If the sample size n is not too small ($n \geq 10$), we can test

$$H_0 : \rho = 0 \text{ vs } H_A : \rho \neq 0,$$

using the same procedures that we used for Pearson's r. The appropriate test statistic is

$$\begin{aligned} \text{TS: } t_s &= \frac{|r_s - 0|}{\sqrt{(1-r_s^2)/(n-2)}} = |r_s|\sqrt{\frac{n-2}{1-r_s^2}} \\ &\sim t_{n-2}, \text{ if } H_0 \text{ is true.} \end{aligned}$$

Example Suppose we like to know whether a linear relationship exists between the percentage of children immunized against DPT and the under-five mortality rate using the Spearman's rank correlation coefficient. We conduct a two-sided test such as

$$H_0 : \rho = 0 \text{ vs } H_A : \rho \neq 0,$$

at the α=0.05 level of significance.

First, we need to calculate the Spearman's rank correlation coefficient. Rank the the percentage of children immunized and under-five mortality rates presented in Table 11.1 from the smallest to the largest, separately for each variable, assigning average ranks to tied observations. The results are shown in Table 11.2, along with the difference in ranks and the squares of these differences. Using the second formula for r_s, we have

$$\begin{aligned} r_s &= 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \\ &= 1 - \frac{6(2045.5)}{20(399)} \\ &= -0.54. \end{aligned}$$

Table 11.2: Ranked Percentage of children immunized against DPT and under-five mortality rate for 20 countries, 1992

Nation	Percentage immunized (x_i)	Rank x_{ri}	Mortality rate (y_i)	Rank y_{ri}	d_i	d_i^2
Ethiopia	13	1	208	20	-19	361
Cambodia	32	2	184	19	-17	289
Senegal	47	3	145	18	-15	225
Greece	54	4	9	5	-1	1
Brazil	69	5	65	14	-9	81
Russian Federation	73	6	32	10	-4	16
Turkey	76	7	87	15	-8	64
Bolivia	77	8	118	16	-8	64
Canada	85	9	8	3	6	36
Japan	87	10	6	1	9	81
Egypt	89	11.5	55	13	-1.5	2.25
India	89	11.5	124	17	-5.5	30.25
United Kingdom	90	13	9	5	8	64
Mexico	91	14	33	11	3	9
China	94	15	43	12	3	9
Finland	95	17	7	2	15	225
France	95	17	9	5	12	144
Italy	95	17	10	7	10	100
Poland	98	19	16	9	10	100
Czech Republic	99	20	12	8	12	144
						2045.50

This value is somewhat smaller than the Pearson's r, which might be inflated due to the non-normality of the data.

The test statistic is

$$\begin{aligned} \text{TS: } t_s &= |r_s|\sqrt{\frac{n-2}{1-r_s^2}} \\ &= |-0.54|\sqrt{\frac{20-2}{1-(-0.54)^2}} \\ &= 2.72. \end{aligned}$$

For a t distribution with df=18, $0.01 < p < 0.02$. Therefore, we reject H_0 at the 0.05 level and conclude that the true population correlation is different from 0. This testing procedure does not require that X and Y be normally distributed.

11.3.2 Pros and Cons

Like other nonparametric methods, Spearman's rank correlation coefficient has advantages and disadvantages. It is much less sensitive to extreme values than Pearson's correlation coefficient. In addition, it can be used when one or both of the variables are ordinal. However, since it relies on ranks rather than actual observations, it does not use everything than is known about a distribution.

Chapter 12

Simple Linear Regression

Like correlation analysis, *simple linear regression* is a technique to investigate the nature of the relationship between two continuous random variables. The primary difference between these two analytic methods is that regression enables us to investigate the change in one variable (called the *response*) which corresponds to a given change in the other (known as the *explanatory variable*). The ultimate objective of regression analysis is to predict or estimate the value of the response that is associated with a fixed value of the explanatory variable.

12.1 The Model

Consider a model of the form

$$y = \alpha + \beta x + \varepsilon,$$

where ε, known as the error with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2_{y|x}$, is the distance a particular outcome y lies from the population regression line

$$\mu_{y|x} = \alpha + \beta x.$$

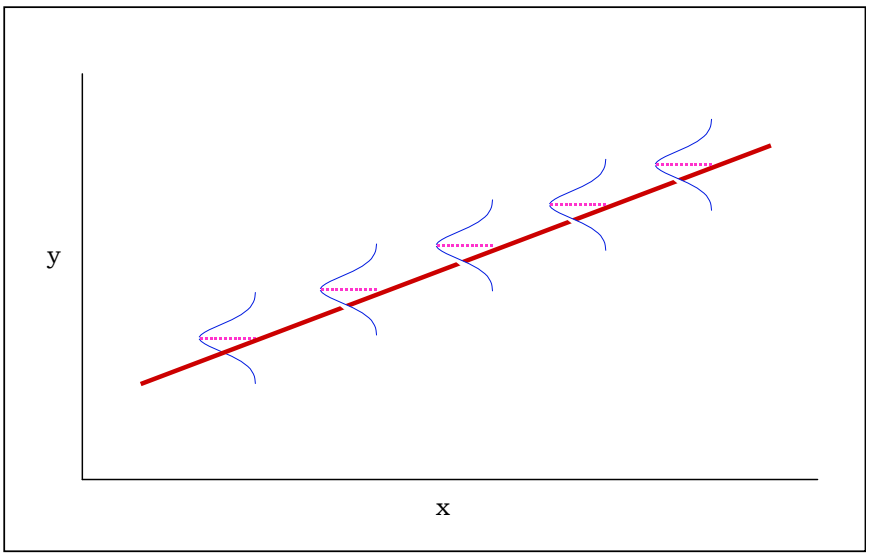


Figure 12.1: Normality of the outcomes y for a given value of x

Such a model is called the *simple linear regression*. If ε is positive, y is greater than $\mu_{y|x}$. If ε is negative, y is less than $\mu_{y|x}$. The parameters α and β are constants called the *coefficients* of the equation. α is the y-intercept, which is the mean value of the response y when x is equal to 0. β is the slope, which is the mean value of y that corresponds to a one-unit increase in x. If β is positive, $\mu_{y|x}$ increases as x increases; if β is negative, $\mu_{y|x}$ decreases as x increases.

In simple linear regression, the coefficients of the population regression line are estimated using a random sample of observations (x_i, y_i). Before we attempt to fit such a line, we must make a few assumptions:

1) For a specified value of x, the distribution of the y values is $\mathcal{N}(\mu_{y|x}, \sigma^2_{y|x})$ (illustrated in Figure 12.1).

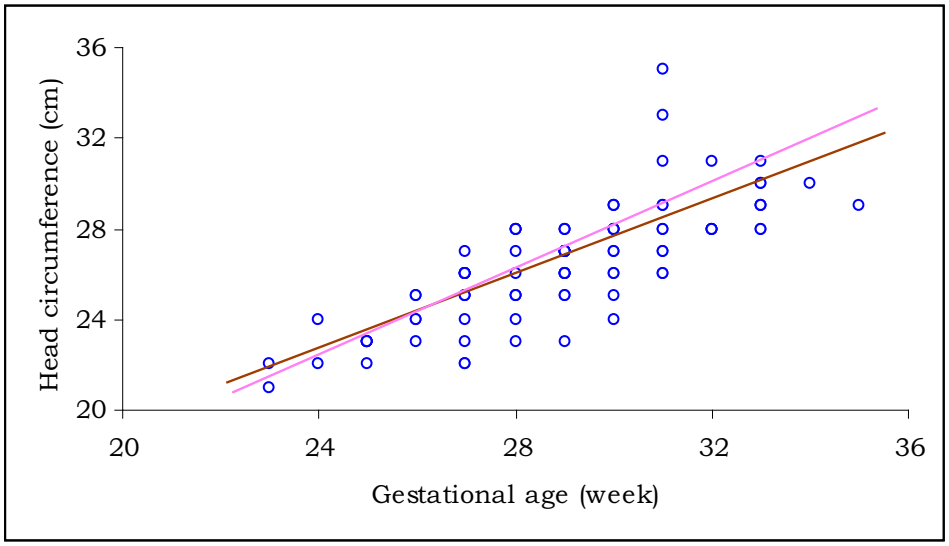


Figure 12.2: Head circumference versus gestational age for a sample of 100 low birth weight infants

2) The relationship between $\mu_{y|x}$ and x is described by the straight line

$$\mu_{y|x} = \alpha + \beta x.$$

3) For a specified value of x, $\sigma_{y|x}$ does not change (i.e. homoscedasticity).

4) The outcomes y are independent.

12.2 The Method of Least Squares

Consider Figure 12.2, the two-way scatter plot of head circumference versus gestational age for a sample of 100 low birth weight infants born in Boston, Massachusetts. The data points themselves vary widely, but the overall pattern suggests that head circumference tends to increase as gestational age increases.

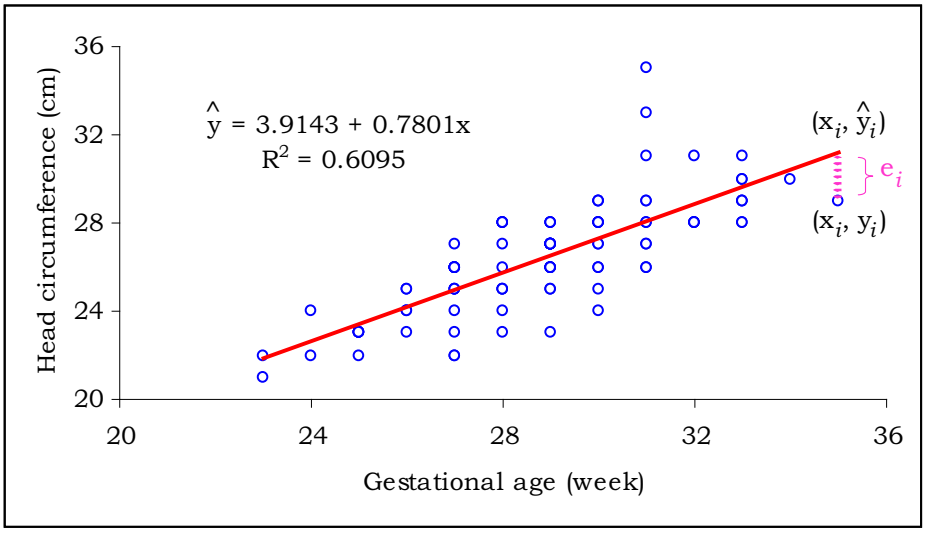


Figure 12.3: Arbitrary line depicting a relationship between head circumference and gestational age

We estimate the coefficients of a population regression line using a single sample of measurements. Suppose that we were to draw an arbitrary line through the scatter of points in Figure 12.2, and two such lines are shown in the figure. Lines sketched by two different individuals are unlikely to be identical. The question then arises as to which line best describes the relationship between two variables. What is needed is a more objective procedure for estimating the line.

The ideal situation would be for the pairs (x_i, y_i) to all fall on a straight line. However, this is not likely because the y are observed values for a set of random variables. The next best thing would be to fit a straight line through the points (x_i, y_i) in such a way to minimize, in some sense, the resulting observed deviation of the y_i from the fitted line.

The process of fitting a regression line represented by

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$$

involves finding $\widehat{\alpha}$ and $\widehat{\beta}$, the estimates of the population regression coefficients α and β. If y_i is the observed outcome of Y for a particular x_i, and $\widehat{y}_i$ is corresponding point on the fitted line, then

$$e_i = y_i - \widehat{y}_i = y_i - (\widehat{\alpha} + \widehat{\beta}x_i),$$

where the distance e_i is known as the *residual*. If all the residuals are equal to 0, it is the ideal situation we talked about earlier, i.e. each point (x_i, y_i) lies directly on the fitted line. However, it is not the case in reality. Different criteria for goodness-of-fit lead to different functions of e_i. One common mathematical technique is known as the *method of least squares*, which says to minimize the sum of the squared deviations from the fitted line,

$$SSE \equiv \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2,$$

which is often called the *error sum of squares (SSE)*, or *residual sum of squares*. In other words, the least-squares regression line is constructed so that SSE is minimized.

Taking derivatives of SSE with respect to α and β and setting them equal to zero gives the least-squares estimates $\widehat{\alpha}$ and $\widehat{\beta}$ as solutions to the equations

$$\begin{cases} 2\sum_{i=1}^{n}[y_i - \widehat{\alpha} - \widehat{\beta}x_i](-1) = 0 \\ 2\sum_{i=1}^{n}[y_i - \widehat{\alpha} - \widehat{\beta}x_i](-x_i) = 0. \end{cases}$$

Simultaneous solution gives

$$\begin{aligned} \widehat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \end{aligned}$$

and

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}.$$

Once we know $\widehat{\alpha}$ and $\widehat{\beta}$, we are able to substitute various values of x into the equation for the line, solve for the corresponding values of $\widehat{y}$, and plot these points to draw the least-squares line, as the line plotted in Figure 12.3.

12.3 Inference for Regression Coefficients

We would like to be able to use the least-squares regression line $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$ to make inference about the population regression line $\mu_{y|x} = \alpha + \beta x$. Before we continue on making inference for regression coefficients, we need to understand the properties of regression coefficients. If $\mathrm{E}(Y_i) = \alpha + \beta x_i$, $\mathrm{var}(Y_i) = \sigma^2_{y|x}$ and $\mathrm{cov}(Y_i, Y_j) = 0$ for $i \neq j$ and $i = 1, \cdots, n$, then the least-squares estimators have the following properties:

1) $\mathrm{E}(\widehat{\beta}) = \beta$, and

$$\mathrm{var}(\widehat{\beta}) = \frac{\sigma^2_{y|x}}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

2) $\mathrm{E}(\widehat{\alpha}) = \alpha$, and

$$\mathrm{var}(\widehat{\alpha}) = \frac{\sigma^2_{y|x} \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2} = \sigma^2_{y|x} \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right]$$

The least-squares method does not provide a direct estimate of $\sigma^2_{y|x}$, but an unbiased estimate of $\sigma^2_{y|x}$ is given by

$$s^2_{y|x} = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{n-2}.$$

The estimate $s_{y|x}$ is often called the *standard deviation from regression.*

The slope is usually the more important coefficient in the linear regression equation, because β quantifies the average change in y that corresponds to each one-unit change in x. If we are interested in conducting a two-sided test that the population slope is equal to β_0:

$$H_0 : \beta = \beta_0 \text{ vs } H_A : \beta \neq \beta_0,$$

The test statistic is

$$\begin{aligned} \text{TS: } t &= \frac{|\widehat{\beta} - \beta_0|}{\sqrt{\widehat{\mathrm{var}}(\widehat{\beta})}} = \frac{|\widehat{\beta} - \beta_0|}{\sqrt{\frac{s^2_{y|x}}{\sum_{i=1}^n (x_i - \overline{x})^2}}} = \frac{|\widehat{\beta} - \beta_0|}{\sqrt{\frac{\frac{\sum_{i=1}^n (y_i - \widehat{y_i})^2}{n-2}}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \\ &\sim t_{n-2}, \text{ if the null hypothesis is true.} \end{aligned}$$

Using Table A.4, we find the p-value to determine whether we should reject or not reject H_0.

Most frequently, we are interested in the case in which β_0=0. If the population slope is equal to 0, then we have

$$\begin{aligned} \mu_{y|x} &= \alpha + (0)x \\ &= \alpha. \end{aligned}$$

These is no linear relationship between X and Y; the mean value of Y is the same regardless of the value of X.

It can be shown that a test of the null hypothesis

$$H_0 : \beta = 0$$

is mathematically equivalent to the test of

$$H_0 : \rho = 0,$$

where ρ is the correlation between X and Y. In fact,

$$\widehat{\beta} = r\left(\frac{s_y}{s_x}\right),$$

where s_x and s_y are the standard deviation of the x and y values respectively. Both null hypotheses claim that y des not change as x increases.

Example Let's back to the example of head circumference and gestational age. The least-squares line (as plotted in Figure 12.3) fitted to the 100 measurements of head circumference and gestations age is

$$\widehat{y} = 3.9143 + 0.7801x.$$

The y-intercept of the fitted line is 3.9143. Theoretically, this is the mean value of head circumference at a gestational age of 0 weeks. However, an age of 0 weeks does not make sense. The slope of the line is 0.7801, implying that for each one-week increase in gestational age, an infant's head circumference increases by 0.7801 cm on average.

In addition, we have

$$s_{y|x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2}} = 1.5904,$$

$$\widehat{\text{var}}(\widehat{\beta}) = \frac{s_{y|x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = (0.0631)^2,$$

and

$$\widehat{\text{var}}(\widehat{\alpha}) = s_{y|x}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] = (1.8291)^2.$$

If we are interested in conducting a two-sided test:

$$H_0 : \beta = 0 \text{ vs } H_A : \beta \neq 0,$$

the test statistic is

$$\begin{aligned} \text{TS: } t &= \frac{|\widehat{\beta} - \beta_0|}{\sqrt{\widehat{\text{var}}(\widehat{\beta})}} \\ &= \frac{|0.7801 - 0|}{0.0631} \\ &= 12.36. \end{aligned}$$

Using Table A.4, for a t distribution with df=100-2=98, $p < 0.001$. Therefore, we should reject $H_0 : \beta = 0$.

We could also construct a 95% CI for β. For a t distribution with df=98, approximately 95% of the observations fall between -1.98 and 1.98. Therefore,

$$\left(\widehat{\beta} \pm 1.98\sqrt{\widehat{\text{var}}(\widehat{\beta})}\right) = (0.6564, 0.9038)$$

is the 95% CI for β.

If we are interested in conducting a two-sided test that the population intercept is equal to α_0:

$$H_0 : \alpha = \alpha_0 \text{ vs } H_A : \alpha \neq \alpha_0,$$

The appropriate test statistic is

$$\begin{aligned} \text{TS: } t &= \frac{|\widehat{\alpha} - \alpha_0|}{\sqrt{\widehat{\text{var}}(\widehat{\alpha})}} \\ &\sim t_{n-2}, \text{ if the null hypothesis is true.} \end{aligned}$$

Using Table A.4, we find the p-value to determine whether we should reject or not reject H_0.

12.4 Inference for Predicted Values

Two types of confidence curves for the predicted values. One curve is for the mean value of the response, and the other is for the prediction of a new observation.

12.4.1 Confidence Limits for the Mean Value of the Response

In addition to making inference about the population slope and intercept, we might also want to use the least-squares regression line to estimate the mean value of y corresponding to a particular value of x, and to construct a 95% CI for the mean. If we have a sample of 100 observations, the 95% CI will take the form

$$\left(\widehat{y} - 1.98\sqrt{\widehat{\text{var}}(\widehat{y})}, \widehat{y} + 1.98\sqrt{\widehat{\text{var}}(\widehat{y})}\right),$$

where $\widehat{y}$ is the predicted mean and, the variance of $\widehat{y}$ is estimated by

$$\widehat{\text{var}}(\widehat{y}) = s^2_{y|x}\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right].$$

Note the term $(x-\overline{x})^2$ in the formula, which implies we are more confident about the mean value of the response when we are closed to the mean value of the explanatory variable.

Example Return once again to the head circumference and gestational age data. When $x = 29$ weeks,

$$\begin{aligned} \widehat{y} &= \widehat{\alpha} + \widehat{\beta}x \\ &= 3.9143 + (0.7801)(29) \\ &= 26.54. \end{aligned}$$

The value 26.54 cm is a point estimate for the mean value of y when $x = 29$. The estimated variance of $\widehat{y}$ is

$$\widehat{\text{var}}(\widehat{y}) = s^2_{y|x}\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right] = (0.159)^2.$$

Therefore, a 95% CI for the mean value of y at $x = 29$ is

$$(26.54 - 1.98(0.159), 26.54 + 1.98(0.159)) = (26.23, 26.85).$$

The curves in Figure 12.4 represent the 95% confidence limits on the mean value of y for each observed value of x (from 23 weeks to 35 weeks).

12.4.2 Confidence Limits for the Prediction of a New Observation

Sometimes instead of predicting the mean value of y for a given value of x, we prefer to predict an individual value of y for a new member of the population.

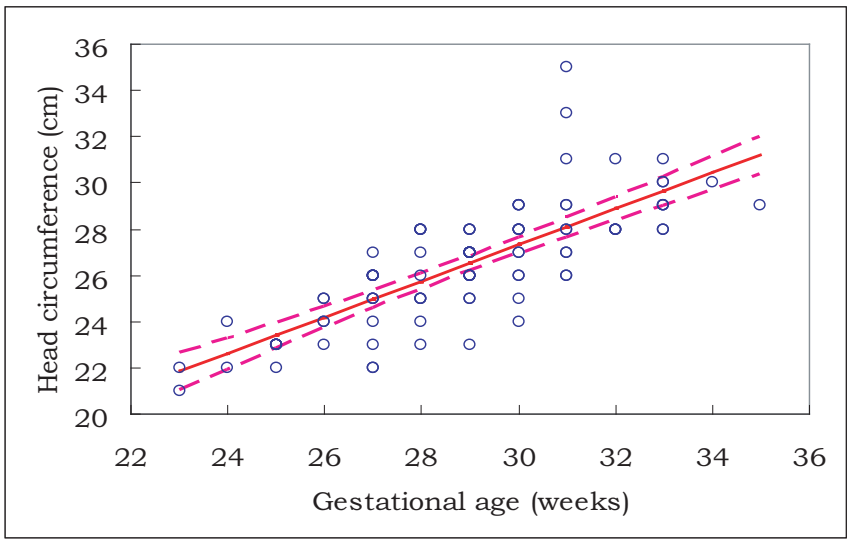


Figure 12.4: The 95% confidence limits on the predicted mean of y for a given value of x

This predicted individual value is denoted by $\widetilde{y}$. Although $\widetilde{y}$ is identical to the predicted mean $\widehat{y}$; in particular,

$$\widetilde{y} = \widehat{\alpha} + \widehat{\beta}x = \widehat{y},$$

the variance of $\widetilde{y}$ is not the same as the variance of $\widehat{y}$. When considering an individual y, we have an extra source of variability to account for: the dispersion of the y values themselves around that mean. In fact,

$$\begin{aligned} \widehat{\text{var}}(\widetilde{y}) &= s^2_{y|x} + \widehat{\text{var}}(\widehat{y}) \\ &= s^2_{y|x}\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]. \end{aligned}$$

If we have a sample of 100 observations, a 95% CI for an individual outcome y takes the form

$$\left(\widetilde{y} - 1.98\sqrt{\widehat{\text{var}}(\widetilde{y})}, \widetilde{y} + 1.98\sqrt{\widehat{\text{var}}(\widetilde{y})}\right).$$

Because if the extra source of variability, the limits on a predicted **individual** value of y are wider than the limits on the predicted **mean** value of y for the same value of x.

Example Return once again to the head circumference and gestational age data. If a new child is selected and this newborn has a gestational age of 29 weeks,

$$\begin{aligned} \widetilde{y} &= \widehat{\alpha} + \widehat{\beta}x \\ &= 3.9143 + (0.7801)(29) \\ &= 26.54. \end{aligned}$$

The estimated variance of $\widetilde{y}$ is

$$\widehat{\text{var}}(\widetilde{y}) = (1.5904)^2 + (0.159)^2 = (1.598)^2.$$

Therefore, a 95% CI for the mean value of y is

$$(26.54 - 1.98(1.598), 26.54 + 1.98(1.598)) = (23.38, 29.70).$$

The curves in Figure 12.5 represent the 95% confidence limits on the individual value of y for each observed value of x from 23 weeks to 35 weeks. Note that

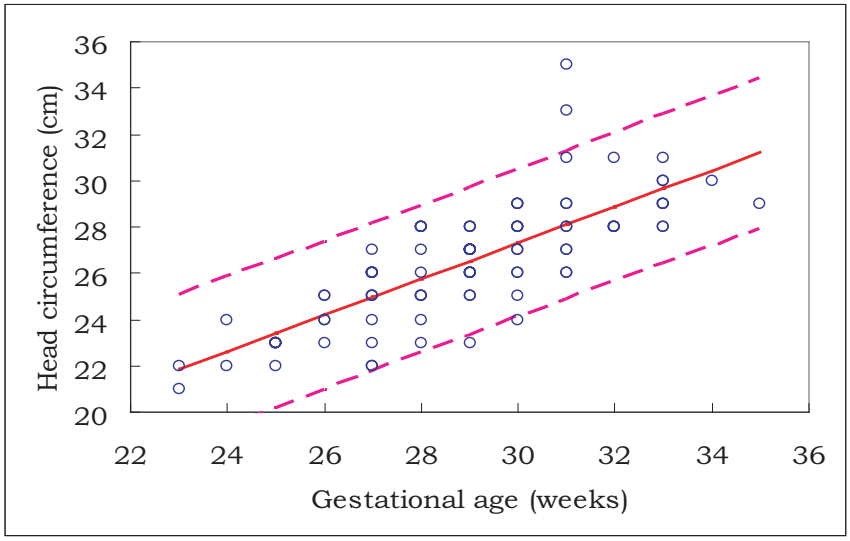


Figure 12.5: The 95% confidence limits on an individual predicted y for a given value of x

these bands are considerably farther from the least-squares regression line (in red color) than those in Figure 12.4.

12.5 Evaluation of the Model

After generating a least-squares regression line represented by

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta}x,$$

we might wonder how well this model actually fits the observed data. There are several strategies to evaluate the goodness of fit of a fitted model. Here we discuss the coefficient of determination and residual plots.

12.5.1 The Coefficient of Determination

The *coefficient of determination* is presented by R^2 and is the square of the Pearson correlation coefficient r. Since r can assume any value in the range -1 to 1, R^2 must lie between 0 and 1. If $R^2 = 1$, all the data points in the sample fall directly on the least-squares line. If $R^2 = 0$, there is no *linear* relationship between x and y.

The coefficient of determination can be interpreted as the proportion of the variability among the observed values of y that is explained by the linear regression of y on x. This interpretation derives from the relationship between σ_y (the standard deviation of Y) and $\sigma_{y|x}$ (the standard deviation of y for a specified value of X):

$$\sigma^2_{y|x} = (1 - \rho^2)\sigma^2_y,$$

where ρ is the correlation between X and Y in the underlying population.

If we replace them by their estimators, we have

$$\begin{aligned} s^2_{y|x} &= (1 - r^2)s^2_y \\ &= (1 - R^2)s^2_y. \end{aligned}$$

Therefore, we can express R^2 in the following expression:

$$\begin{aligned} R^2 &= 1 - \frac{s^2_{y|x}}{s^2_y} \\ &= \frac{s^2_y - s^2_{y|x}}{s^2_y}. \end{aligned}$$

Since $s^2_{y|x}$ is the variation in the y values that still remains after accounting for the relationship between y and x, $s^2_y - s^2_{y|x}$ must be the variation in y that is explained by their linear relationship. Thus, R^2 is the proportion of the total observed variability among the y values that is explained by the linear regression of y on x.

12.5.2 Residual Plots

Another strategy for evaluating how well the least-squares line actually fits the observed data is to generate a two-way scatter plot of the residuals ($e_i = y_i - \widehat{y}_i$) against the fitted (or predicted) values ($\widehat{y}_i$) of the response variable y_i. For instance, one particular child in the sample of 100 low birth weight infants has a gestational age of 29 weeks and a head circumference of 27 cm. The child's predicted head circumference, given that $x_i = 29$ weeks, is

$$\begin{aligned} \widehat{y}_i &= \widehat{\alpha} + \widehat{\beta} x_i, \\ &= 3.9143 + (0.7801)(29) \\ &= 26.54(cm). \end{aligned}$$

Then, the residual associated with this observation is

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= 27 - 26.54 \\ &= 0.46; \end{aligned}$$

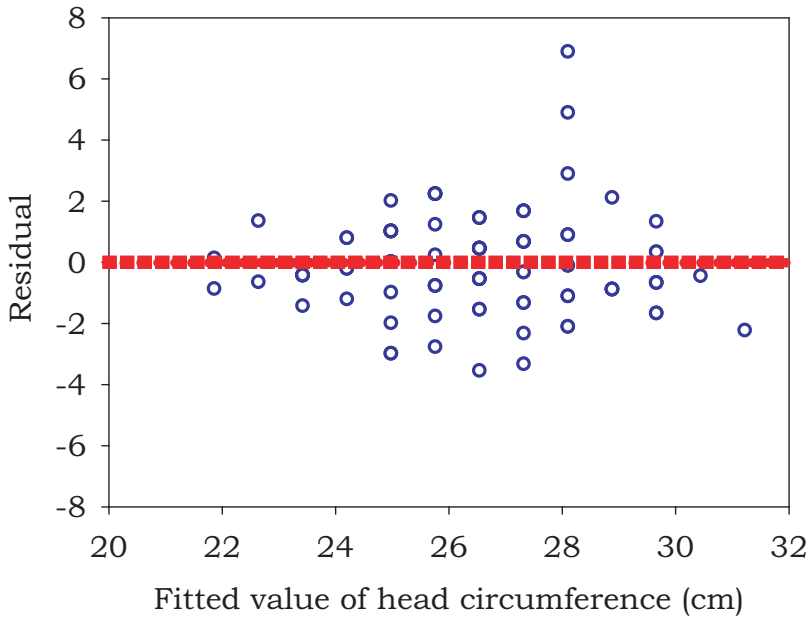


Figure 12.6: Residuals versus fitted values of head circumference

therefore, the point (26.54, 0.46) would be included on the residual plot. Figure 12.6 is a scatter plot for all 100 observations in the sample of low birth weight infants.

A plot of the residuals serves three purposes:

1) it can help to detect outlying observations,

2) it might suggest a failure in the assumption of homoscedasticity (see Figure 12.7),

3) it might suggest the true relationship between x and y is not linear.

Homoscedasticity means that the standard deviation of the residuals or $\sigma_{y|x}$ is **constant** across all values of x. If the range of the residuals either increases or decreases as $\widehat{y}$ becomes larger, as the one in Figure 12.7, it implies that $\sigma_{y|x}$ does

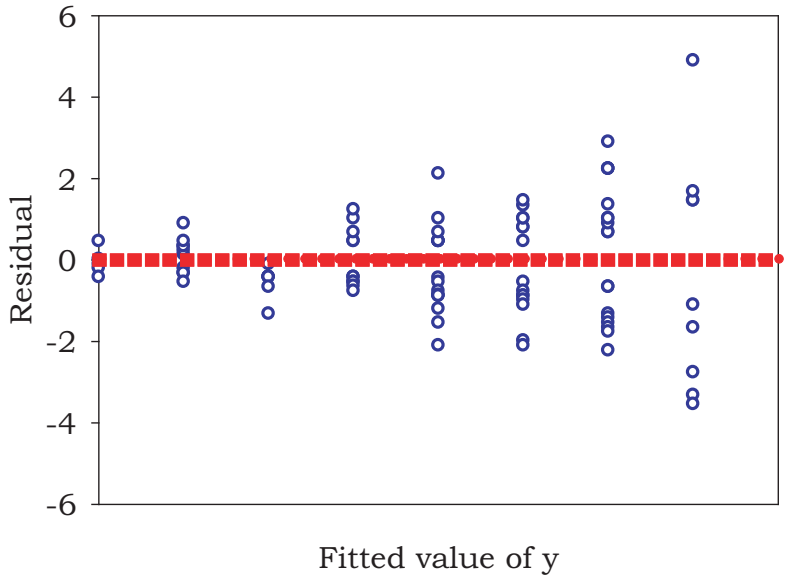


Figure 12.7: Violation of the assumption of homoscedasticity

not take the same value for all values of x. In this case, simple linear regression is not the appropriate technique for modeling the relationship between x and y.

If the residuals follow a pattern – for example, e_i increases as $\widehat{y}$ increases, – this would suggest that the true relationship between x and y might not be linear. In this situation, a *transformation* of x or y or both might be appropriate. Often, a curvilinear relationship between two variables can be transformed into a more straightforward linear one.

12.5.3 Transformations

As shown in Figure 12.8, the relationship between crude birth rate per 1000 population and gross national product (GNP) is not a linear, although birth rate tends to decrease as GNP increases.

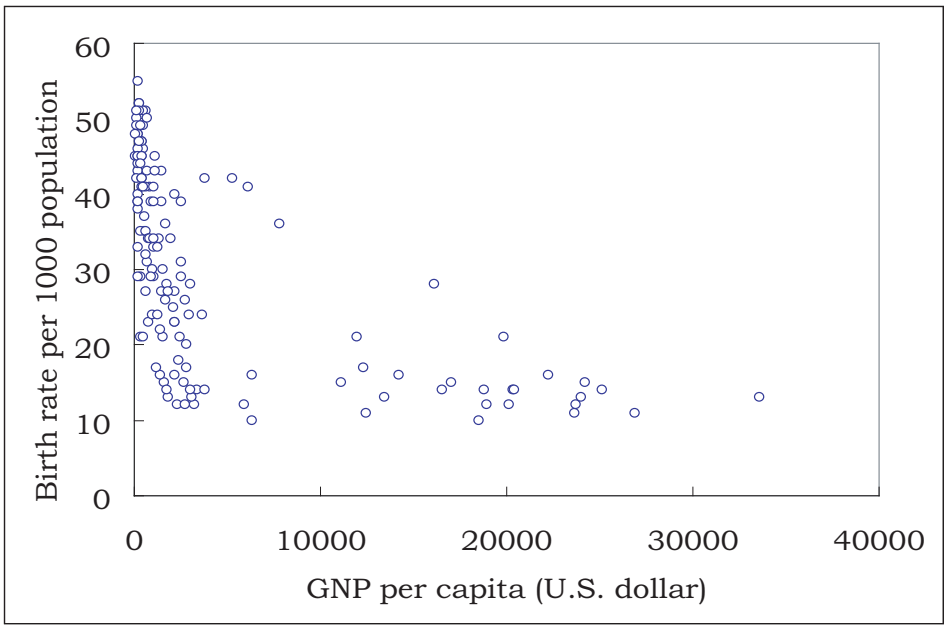


Figure 12.8: Birth rate per 10000 population versus gross national product (GNP) per capita for 143 countries, 1992

When the relationship between x and y is not linear, we begin by looking at transformations of the form x^p or y^p, where

$$p = \cdots, -3, -2, -1, -\frac{1}{2}, \ln, \frac{1}{2}, 1, 2, 3, \cdots$$

The most common choices are $\ln(y)$, $\sqrt{x}$, or x^2. The *circle of powers* (illustrated in Figure 12.9) provides a general guideline for choosing a transformation.

Quadrant I If the plotted data resemble the pattern in Quadrant I, an appropriate transformation would be either "up" on x or "up" on y. In other words, either x or y would be raised to a power greater than $p = 1$; the more curvature in the data, the higher the value of p needed to achieve linearity.

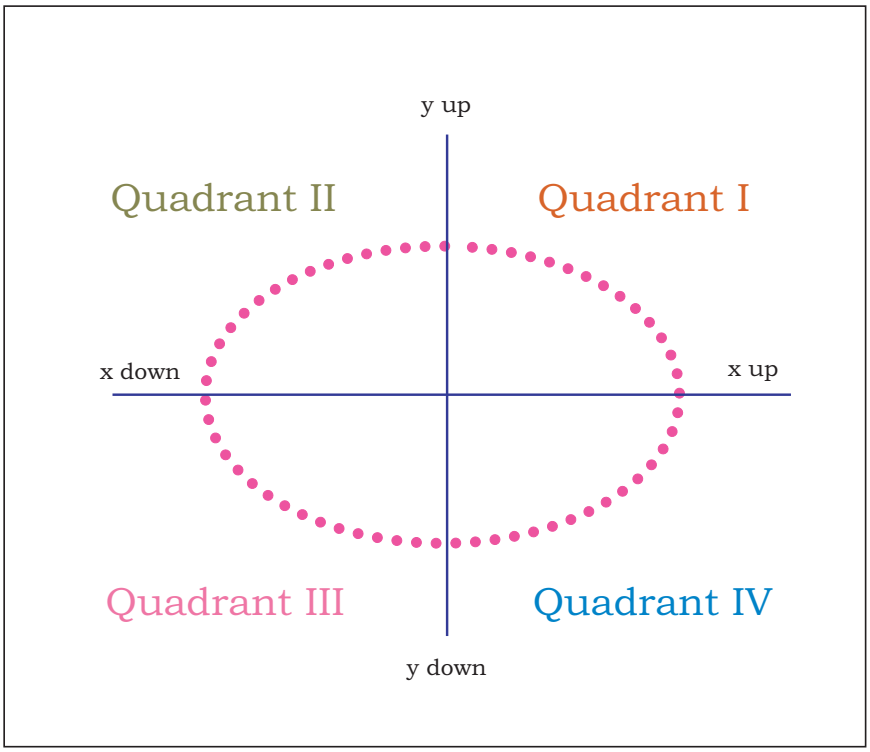


Figure 12.9: The circle of powers for data transformations

Quadrant II If the data follow the trend in Quadrant II, we would want to either "down" on x or "up" on y.

Quadrant III If the data resemble the pattern in Quadrant III, we would want to either "down" on x or "down" on y.

Quadrant IV If the data follow the trend in Quadrant IV, we would want to either "up" on x or "down" on y.

Whichever transformation is chosen, we must always verify that the assumption of homoscedasticity is valid. The data in Figure 12.8 most closely resemble the pattern in Quadrant III; therefore, we might try replacing x by $\ln(x)$, i.e. the natural logarithm of GNP. The effect of this transformation is shown in Figure

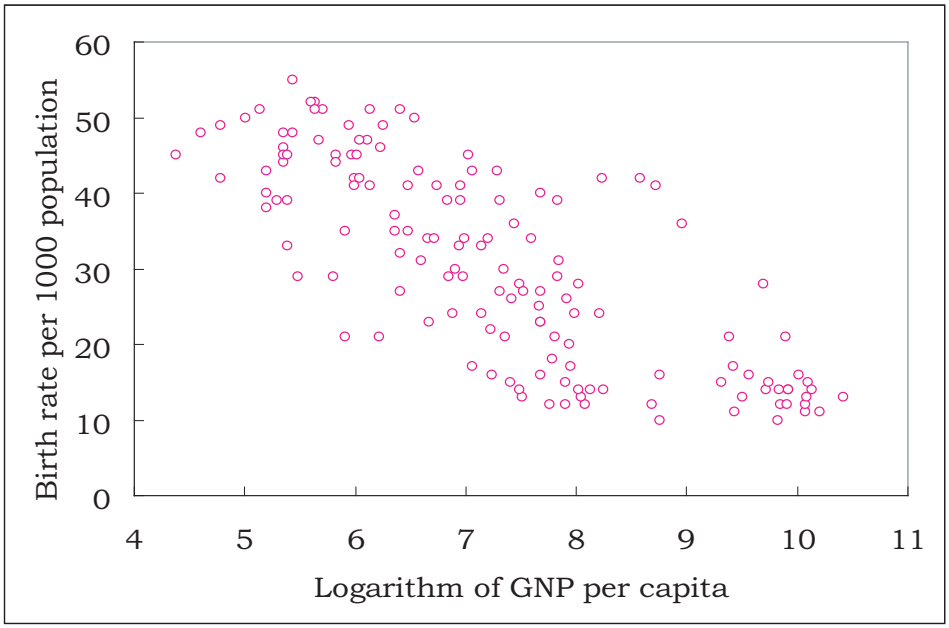


Figure 12.10: Birth rate per 10000 population versus the natural logarithm of gross national product (GNP) per capita for 143 countries, 1992

12.10. The relationship between birth rate and the natural logarithm of GNP appears much more linear than the relationship between birth rate and GNP itself. Therefore, we should fit a simple linear regression model of the form

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta}\ln(x).$$

Chapter 13

Multiple Regression

In the preceding chapter, we saw how simple linear regression can be used to explore the nature of the relationship between two continuous random variables. We might suspect that additional explanatory variables (instead of one single explanatory variable) could be used to predict or estimate the value of a response. Here we would like to investigate a more complicate relationship among a number of variables, we use a natural extension of simple linear regression analysis known as *multiple regression*.

13.1 The Model

Consider a model of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon,$$

where ε is the error term with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2_{y|x_1,x_2,\cdots,x_q}$ and $x_1, x_2, \cdots, x_q$ are the outcomes of q distinct explanatory variables. The parameters $\alpha, \beta_1, \cdots, \beta_q$ are constants that again called the coefficients of regression. Like simple regression, the intercept α is the mean value of the response y when all explanatory

variables take the value 0; the slope β_i is the change in the mean value of y that corresponds to a one-unit increase in x_i, given that all other explanatory variables remain constant. Again the error term ε is the distance a particular outcome y lies from the population regression equation

$$\mu_{y|x_1,x_2,\cdots,x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q,$$

where $\mu_{y|x_1,x_2,\cdots,x_q}$ is the mean value of y given the explanatory variables assume these values (i.e. $x_1, x_2, \cdots, x_q$). When a single explanatory variable was involved, the fitted model was simply a straight line. With two explanatory variables, the model represent a plane in three-dimensional space; with three or more variables, it becomes a hyperplane in higher-dimensional space.

However, just as we had to make a number of assumptions for simple linear regression, we make a set of analogous assumptions for the more complex multiple regression model. These assumptions are as follows:

1) For specified values of $x_1, x_2, \cdots, x_q$, all of which are considered to be measured without error, $Y \sim \mathcal{N}(\mu_{Y|x_1,x_2,\cdots,x_q}, \sigma^2_{Y|x_1,x_2,\cdots,x_q})$.

2) The relationship between $\mu_{y|x_1,x_2,\cdots,x_q}$ and $x_1, x_2, \cdots, x_q$ is described by the equation

$$\mu_{y|x_1,x_2,\cdots,x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q.$$

3) For any set of values $x_1, x_2, \cdots,$ and x_q, $\sigma_{y|x_1,x_2,\cdots,x_q}$ is a constant (i.e. homoscedasticity).

4) The outcomes y are independent.

13.1.1 The Least-Squares Regression Equation

The coefficients of the population regression equation

$$\mu_{y|x_1,x_2,\cdots,x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

are estimated using a random sample of observations represented as $(x_{1i}, x_{2i}, \cdots, x_{qi}, y_i)$. We use the same method of least squares to fit the model

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_q x_q.$$

This technique is to minimize the sum of the squares of the residuals (i.e. error sum of squares (SSE)

$$\begin{aligned} SSE \equiv \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta}_1 x_1 - \widehat{\beta}_2 x_2 - \cdots - \widehat{\beta}_q x_q)^2. \end{aligned}$$

The y_i is the observed outcome of the response Y for given values $x_{i1}, x_{i2}, \cdots$, and x_{iq}, while $\widehat{y}_i$ is the corresponding value from the fitted equation. Taking derivatives of SSE with respect to α and β_j $(j = 1, \cdots, q)$ and setting them equal to zero gives the least-squares estimates $\widehat{\alpha}$ and $\widehat{\beta}_j$ as solutions to the equations

$$\begin{cases} \frac{\partial}{\partial \alpha} SSE = 2 \sum_{i=1}^{n} [y_i - \widehat{\alpha} - \sum_{j=1}^{q} \widehat{\beta}_j x_{ij}](-1) = 0 \\ \frac{\partial}{\partial \beta_k} SSE = 2 \sum_{i=1}^{n} [y_i - \widehat{\alpha} - \sum_{j=1}^{q} \widehat{\beta}_j x_{ij}](-x_{ik}) = 0 \text{ for } k = 1, \cdots, q. \end{cases}$$

We can rewrite the model by matrix notation as

$$\mathrm{E}(\mathbf{Y}) = \mathbf{XB},$$

where $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{B}$ are

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1q} \\ \vdots & & \vdots & \\ 1 & x_{n1} & \cdots & x_{nq} \end{pmatrix}$$

The variance and covariance matrix for $\mathbf{Y}$ is $\mathrm{var}[\mathbf{Y}] = \sigma^2_{y|x_1,x_2,\cdots,x_q}\mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix. The sum of the squares of the residuals (i.e. SSE) is possible to reformulate in terms of matrices as follows

$$SSE \equiv \mathbf{e}^T\mathbf{e} = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})$$

where the residuals are $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}$, and $\mathbf{A}^T$ is the *transpose* of an arbitrary matrix $\mathbf{A}$. The least-squares estimates are conveniently expressed in terms of the matrices

$$\widehat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \text{ if } (\mathbf{X}^T\mathbf{X})^{-1} \text{ is nonsingular,}$$

where $\mathbf{A}^{-1}$ is the *inverse* of an arbitrary matrix $\mathbf{A}$.

13.1.2 Inference for Regression Coefficients

We would like to be able to use the least-squares regression model

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_q x_q.$$

to make inference about the population regression equation

$$\mu_{y|x_1,x_2,\cdots,x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q.$$

The regression coefficients are estimated using a sample of data drawn from the underlying population; their values would change if a different sample were selected. Therefore, like simple linear regression, we need to estimate the standard errors of these estimators to be able to make inference about the true population parameters.

When testing

$$H_0 : \beta_i = \beta_{i0} \text{ vs } H_A : \beta_i \neq \beta_{i0},$$

we assume that the values of all other explanatory variables $x_j \neq x_i$ remain constant. The test statistic is

$$\begin{aligned} \text{TS: } t &= \frac{|\widehat{\beta}_i - \beta_{i0}|}{\sqrt{\widehat{\text{var}}(\widehat{\beta}_i)}} \\ &\sim t_{n-q-1}, \text{ if the null hypothesis is true.} \end{aligned}$$

where q is the number of explanatory variables in the model. Note the TS does not follow a t distribution with df=$n-2$ as in the simple linear regression. Using

Table A.4, we find the p-value to determine whether we should reject or not reject H_0.

By means of matrix notation, the variance and covariance matrix for $\widehat{\mathbf{B}}$ are the elements of the matrix

$$\begin{aligned}
\text{var}[\widehat{\mathbf{B}}] &= \text{var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{var}[\mathbf{Y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2_{y|x_1,x_2,\cdots,x_q}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2_{y|x_1,x_2,\cdots,x_q}(\mathbf{X}^T\mathbf{X})^{-1}.
\end{aligned}$$

When $\sigma^2_{y|x_1,x_2,\cdots,x_q}$ is unknown, we must estimate it from the observed data. In fact,

$$\frac{SSE}{n-q-1} = \frac{(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}})^T(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}})}{n-q-1}$$

is an unbiased estimate of $\sigma^2_{y|x_1,x_2,\cdots,x_q}$.

We must bear in mind that multiple tests of hypothesis based on the same set of data are not independent. If each individual test is conducted at the α level of significance, the overall probability of making type I error is in fact larger than α.

In addition to conducting tests of hypotheses, we also can calculate confidence intervals for the population regression coefficients. Furthermore, we can construct a confidence interval for the predicted mean value of y and a prediction interval for the predicted individual y corresponding to a given set of values for

the explanatory variables. In all cases, the procedures are analogous to those used when a single explanatory variable was involved.

ANOVA in Multiple Regression We can test the overall significance of a regression using the F-test. The null hypothesis states that all the slope parameters are equal to zero, and the alternative hypothesis simply states that at least one of the slope parameters is not equal to zero. The idea is to divide total sum of squares of deviations (SST) into parts: sum of squares of explanatory variables (explained variation) (SSR) + sum of squares of random errors (unexplained variation) (SSE). Therefore, the total sum of squares is:

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i-\widehat{y_i})^2}_{SSE}$$

We can construct an ANOVA table for

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0 \text{ vs } H_A : \beta_j \neq 0 \text{ for some } j,$$

Source	DF	Sum of Squares	Mean Square	F Value
Regression	q	$SSR = \sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2$	$MSR = \frac{SSR}{q}$	MSR/MSE
Error	$n-q-1$	$SSE = \sum_{i=1}^{n}(y_i-\widehat{y_i})^2$	$MSE = \frac{SSE}{n-q-1}$	
Total	$n-1$	$SST = \sum_{i=1}^{n}(y_i-\overline{y})^2$		

The test statistic $F = MSR/MSE \sim \mathcal{F}_{q,n-q-1}$. The relationship between $F = MSR/MSE$ and $R^2 = SSR/SST$ can be shown that

$$F = \frac{R^2/q}{(1-R^2)/(n-q-1)}.$$

Thus, testing for the overall significance of the regression is equivalent to testing for the significance of R^2.

Evaluation of the Model We can use the coefficient of determination R^2 to assess how well a particular least-squares model fits the observed data. The increase in R^2 suggests the model is improved. However, we must be careful when comparing the coefficients of determination from two different models. The more explanatory variables in the model, the higher the coefficient of determination is. The inclusion of an additional explanatory variable in a model can never cause R^2 to decrease.

To get around this problem, we can use a second measure, called the *adjusted* R^2, that compensates for the added complexity of a model and allows us to make a more valid comparison between models that contains different numbers of explanatory variables. The adjusted R^2 is calculated as

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-q-1} = 1 - \frac{SSE/(n-q-1)}{SST/(n-1)},$$

where n is the number of observations in the data set and q is the number of explanatory variables in the model. (Recall: $R^2 = SSR/SST = 1 - SSE/SST$.)

We also can use a scatter plot of the residuals from the model versus the fitted values from the same model to assess how well a model fits the observed data as for simple linear regression.

Indicator Variables All the explanatory variables we have considered up to this point have been measured on a continuous scale. However, regression analysis can be generalized to incorporate discrete or nominal explanatory variables. Since the explanatory variables in a regression analysis must assume numerical values, we designate discrete or nominal explanatory variables by numbers for identifying the categories of the nominal random variable. This sort of explanatory variables is called an *indicator variable* or a *dummy variable*.

Interaction Terms In some situations, one explanatory variable has a different effect on the predicted response y depending on the value of another explanatory variable. To model a relationship of this kind, we create what is known as an *interaction term* by multiplying together the outcomes of two variables to create a third variable.

13.1.3 Example

Consider a sample of 100 low birth weight infants born in Boston, Massachusetts. We might wonder whether the head circumference depends on factors including gestational age, mother's diagnosis of toxemia during pregnancy.

The diagnosis of toxemia is a dichotomous random variable. We designate the presence of toxemia during pregnancy by 1 and its absence by 0.

Main Effects Model SAS output (PROC REG) for the multiple regression of head circumference (variable name: headcirc) on gestational age (variable name: gestage) and mother's diagnosis of toxemia during pregnancy (variable name: toxemia):

The REG Procedure
Model: MODEL1
Dependent Variable: headcirc

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	414.34299	207.17150	91.18	<.0001
Error	97	220.40701	2.27224		
Corrected Total	99	634.75000			

Root MSE	1.50739	R-Square	0.6528
Dependent Mean	26.45000	Adj R-Sq	0.6456
Coeff Var	5.69903		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.49558	1.86799	0.80	0.4253
gestage	1	0.87404	0.06561	13.32	<.0001
toxemia	1	-1.41233	0.40615	-3.48	0.0008

The fitted least-squares regression model is

$$\widehat{y} = 1.4956 + 0.8740(gestage) - 1.4123(toxemia).$$

By an F-test of $H_0 : \beta_1 = \beta_2 = 0$, the corresponding test statistic F is 91.18 with a p-value < 0.0001, and we conclude at least one β is not equal to 0. The coefficient of determination R^2 and the adjusted R^2 for this regression model are 0.653 and 0.646, respectively.

A test of

$$H_0 : \beta_2 = 0 \text{ vs } H_A : \beta_2 \neq 0$$

assuming that gestational age does not change, results in a test statistic of $t = -3.48$ and $p = 0.0008$. Therefore, we reject the null hypothesis at the 0.05 level of significance and conclude that β_2 is not equal to 0.

For a child whose mother was diagnosed with toxemia during pregnancy (i.e. toxemia=1), the fitted least-squares regression equation is

$$\begin{aligned} \widehat{y} &= 1.4956 + 0.8740(gestage) - 1.4123(1) \\ &= 0.0833 + 0.8740(gestage). \end{aligned}$$

For a child whose mother was not diagnosed with toxemia during pregnancy (i.e. toxemia=0), the fitted least-squares regression equation becomes

$$\begin{aligned} \widehat{y} &= 1.4956 + 0.8740(gestage) - 1.4123(0) \\ &= 1.4956 + 0.8740(gestage). \end{aligned}$$

The two lines are plotted in Figure 13.1. The lines are parallel to each other as they share the same slope, 0.8740. This is the consequence of fitting a single regression model to the two different groups of infants. The difference in intercepts

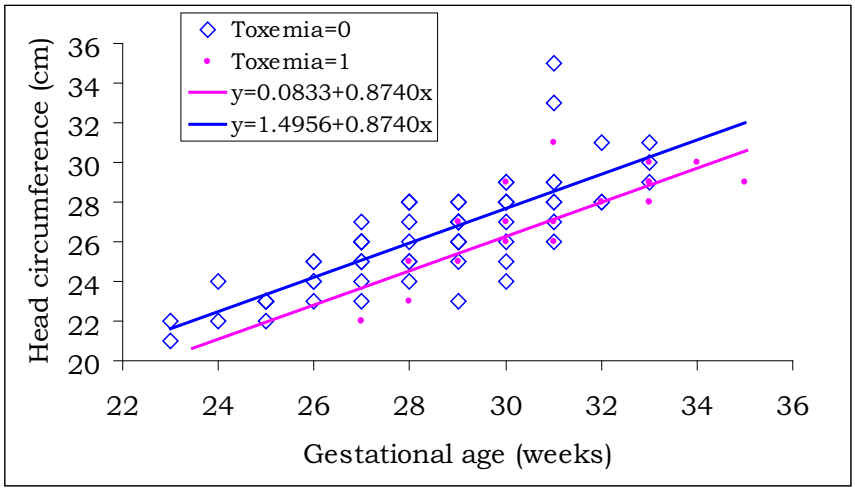


Figure 13.1: Fitted least-squares regression lines for different levels of toxemia

(0.0833 vs 1.4956) suggests that children whose mothers were not diagnosed with toxemia tend to have larger head circumferences than children whose mothers were diagnosed with toxemia.

Main Effects and Interaction Model We might wonder if one-week increase in gestational age has a different effect on a child's head circumference depending on whether the infant's mother had experienced toxemia during pregnancy or not. Therefore, we create an interaction term for these two factors and add it in the previous main effects model.

Here we use PROC GLM in SAS to conduct the analysis instead of PROC REG. Because using PROC REG the interaction term needs to be created in the analytic data set beforehand, we just need to specify the interaction term "gestage * toxemia" as one extra term in the model if we use PROC GLM. SAS output from PROC GLM for the multiple regression including gestational age, mother's diagnosis of toxemia during pregnancy, and the interaction of these two variables:

The GLM Procedure

Dependent Variable: headcirc

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	414.5258395	138.1752798	60.23	<.0001
Error	96	220.2241605	2.2940017		
Corrected Total	99	634.7500000			

R-Square	Coeff Var	Root MSE	headcirc Mean
0.653054	5.726262	1.514596	26.45000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gestage	1	386.8673658	386.8673658	168.64	<.0001
toxemia	1	27.4756275	27.4756275	11.98	0.0008
gestage*toxemia	1	0.1828462	0.1828462	0.08	0.7783

Source	DF	Type III SS	Mean Square	F Value	Pr > F
gestage	1	314.0291159	314.0291159	136.89	<.0001
toxemia	1	0.7314834	0.7314834	0.32	0.5736
gestage*toxemia	1	0.1828462	0.1828462	0.08	0.7783

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.762912071	2.10225478	0.84	0.4038
gestage	0.864611583	0.07389805	11.70	<.0001
toxemia	-2.815032213	4.98514735	-0.56	0.5736
gestage*toxemia	0.046165802	0.16352127	0.28	0.7783

The fitted least-squares regression model is

$$\widehat{y} = 1.7629 + 0.8646(gestage) - 2.8150(toxemia) + 0.0462(gestage) \times (toxemia).$$

By an F-test of $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, the corresponding test statistic F is 60.23 with a p-value < 0.0001, and we conclude at least one β is not equal to 0. The coefficient of determination R^2 is 0.653 (R^2_{adj} is not reported here, but it is 0.642.)

A test of

$$H_0 : \beta_3 = 0 \text{ vs } H_A : \beta_3 \neq 0$$

for the interaction term, results in a test statistic of $t = 0.28$ and $p = 0.778$. Therefore, we could not reject the null hypothesis at the 0.05 level of significance.

With the interaction term in the model, for a child whose mother was diagnosed with toxemia during pregnancy (i.e. toxemia=1), the fitted least-squares regression equation is

$$\begin{aligned} \widehat{y} &= 1.7629 + 0.8646(gestage) - 2.8150(1) + 0.0462(gestage)(1) \\ &= -1.0521 + 0.9108(gestage). \end{aligned}$$

For a child whose mother was not diagnosed with toxemia during pregnancy (i.e. toxemia=0), the fitted least-squares regression equation becomes

$$\begin{aligned} \widehat{y} &= 1.7629 + 0.8646(gestage) - 2.8150(0) + 0.0462(gestage)(0) \\ &= 1.7629 + 0.8646(gestage). \end{aligned}$$

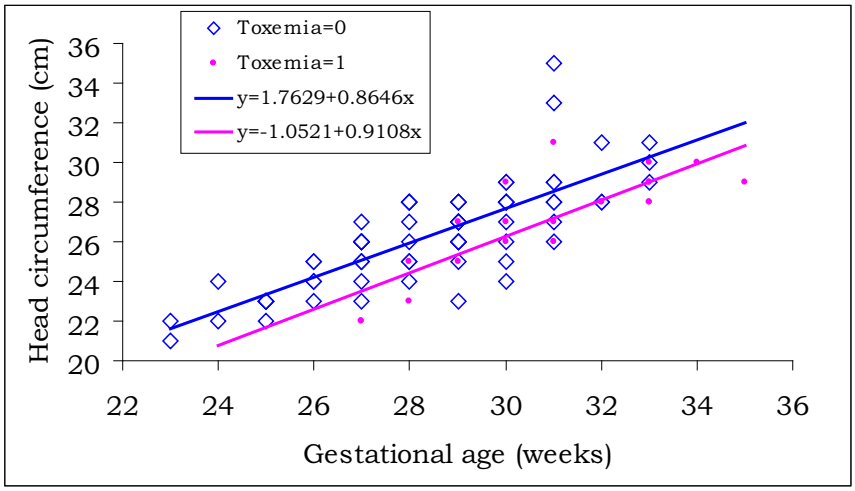


Figure 13.2: Fitted least-squares regression lines for different levels of toxemia, interaction term included

The two lines are plotted in Figure 13.2. The lines are not parallel to each other this time because the model includes an interaction term, and they have different intercepts and different slopes. In the range of interest, one line still lies completely above tho other. This implies that across all relevant values of gestational age, infants whose mothers were not diagnosed with toxemia tend to have larger head circumferences than infants whose mothers were diagnosed with this condition.

13.2 Model Selection

If we are presented with a number of potential explanatory variables, how do we decide which ones to retain in the model and which to leave out? This decision is usually made between a combination of statistical and nonstatistical considerations. The models could be evaluated according to some statistical

criteria. The two most commonly used procedures of *stepwise selection* are known as *forward selection* and *backward elimination*. It is possible that we could end up with different final models, depending on which procedure is applied.

13.2.1 Forward Selection

proceeds by introducing variables into the model one at a time. The model is evaluated at each step, and the process continues until some specified statistical criterion is achieved. For example, we might begin by including the single explanatory variable that yields the largest R^2. We next put into the equation the variable that increases R^2 the most, assuming that the first variable remains in the model and that the increase in R^2 is statistically significant. The procedure continues until we reach a point where none of the remaining variables explains a significant amount of the additional variability in y.

13.2.2 Backward Elimination

begins by including all explanatory variables in the model. Variables are dropped one at a time. It begins with the variable that reduces R^2 by the least amount given the other variables in the model. If the decreases in R^2 is not statistically significant, the variable is left out of the model permanently. The equation is evaluated at each step, and the procedure is repeated until each of the variables remaining in the model explains a significant portion of the observed variation in the response.

13.2.3 The Incremental or Marginal Contribution of Additional Explanatory Variable(s)

Partial-F can be used to assess the significance of the difference of two R^2's for nested models. Nested means one is a subset of the other, as a model with interaction terms and one without. Also, unique effects of individual independents can be assessed by running a model with and without a given independent, then taking partial F to test the difference. In this way, partial F plays a critical role in the trial-and-error process of model-building.

Let there be q be a larger model and let p be a nested smaller model. Let SSE_p be the residual sum of squares (deviance) for the smaller model. Let SSE_q be the residual sum of squares for the larger model. Partial F is defined as

$$\text{Partial F} = \frac{(SSE_p - SSE_q)/df1}{SSE_q/df2} \sim \mathcal{F}(df1, df2),$$

where $df1 = (df \text{ of } SSE_p \text{ - } df \text{ of } SSE_q)$ and $df2 = df \text{ of } SSE_q$.

13.3 Collinearity ✠

We should always check for the presence of collinearity. *Collinearity* occurs when two or more of the explanatory variables are correlated to the extent that they convey essentially the same information about the observed variation in y. One symptom of collinearity is the instability of the estimated coefficients and their standard errors. In particular, the standard errors often become very large.

In Section 13.1.3, we have two models for head circumference among low birth weight infants considering gestational age, toxemia, and the interaction between the two. Toxemia and the interaction with gestational age are highly correlated; in fact, the Pearson correlation coefficient is equal 0.997. The two models are contrasted below for the coefficient of toxemia.

	Interaction Term Not Included	Interaction term Included
Coefficient	-1.412	-2.815
Standard Error	0.406	4.985
Test Statistic	-3.477	-0.565
p-value	0.0008	0.574
R^2	0.653	0.653
R^2_{adj}	0.646	0.642

When the interaction term is included in the model, the estimated coefficient of toxemia doubles, its standard error increases by a factor 12, and the coefficient no longer achieves statistical significance. The coefficient of determination does not change when the interaction is included; it remains 65.3%. Furthermore, the adjusted R^2 decreases slightly. These facts indicate the information supplied by the interaction is redundant.

The causes of collinearity could be

1) improper use of dummy variables (e.g. failure to exclude one category);

2) including a variable that is computed from other variables in the equation (e.g. body mass index = weight (kg)/ hight^2 (m^2) , and the regression includes all 3 variables);

3) including the same or almost the same variable twice (e.g. height in feet and height in inches; or, more commonly, two different measures of the same identical concept);

4) the use of constructed variables (e.g. interaction terms or squared terms).

The above all imply some sort of error on the misuse. But, it may just be that variables really and truly are highly correlated. Correlation among the predictors as seen in the

- scatterplot matrix (and leverage plots);
- changing/unstable coefficient estimates;
- inflated standard errors (VIF);
- good overall fit (F statistics) but mediocre t statistics for separate coefficients.

13.3.1 Collinearity Diagnostics

Variance inflation factor (VIF) and tolerance (1/VIF) are two measures for identifying collinearity. The VIF quantifies the severity of collinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient (the square of the estimate's

standard deviation) is increased because of collinearity. A common rule of thumb is that if VIF > 5 then collinearity is high.

The variance of a typical regression coefficient (say β_i) can be shown to be the following:

$$\text{var}(\widehat{\beta}_i) = \frac{\sigma^2}{s_i(1 - R_i^2)},$$

where $s_i = \sum_{j=1}^{n}(x_{ij} - \overline{x}_i)$ and R_i^2 is the unadjusted R^2 when you regress regress X_i against all the other explanatory variables in the model, that is, X_i against a constant, $X_2, X_3, \cdots, X_{i-1}, X_{i+1}, \cdots, X_k$. If there is no linear relation between X_i and the other explanatory variables in the model, then R_i^2 will be zero and the variance of β_i will be $\frac{\sigma^2}{s_i}$. We obtain the variance inflation factor and tolerance as

$$\text{VIF}(\beta_i) = \frac{1}{1 - R_i^2} \text{ and } \text{tolerance}(\beta_i) = 1 - R_i^2.$$

It can be seen that the higher VIF or the lower the tolerance index, the higher the variance of β_i and the greater the chance of finding β_i insignificant, which means that severe collinearity effects are present.

The second approach uses the eigenvalues of the model matrix which indicate that the dimension of the problem could be (or should be) reduced. However, the long story about eigenvalues (i.e. factorization of the correlation matrix into its eigenvalues and eigenvectors) is omitted here. There are three measures "eigenvalues", "condition index (CI)", and "condition number". While the condition index is the ratio between a specific eigenvalue and the maximum of all eigenvalues, the condition number is the root of largest eigenvalue divided by the smallest. As an informal rule a condition index between 10 and 100 or condition numbers

between 15 and 30 would indicate weak to serious problems. Unfortunately, as there are different methods how to scale the model matrix or using the correlation matrix instead, this second approach has some drawbacks.