Introduction to Biostatistics

Larry Winner
Department of Statistics
University of Florida

July 8, 2004

Contents

Chapter 1

Introduction

These notes are intended to provide the student with a conceptual overview of statistical methods with emphasis on applications commonly used in pharmaceutical and epidemiological research. We will briefly cover the topics of probability and descriptive statistics, followed by detailed descriptions of widely used inferential procedures. The goal is to provide the student with the information needed to be able to interpret the types of studies that are reported in academic journals, as well as the ability to perform such analyses. Examples are taken from journals in the pharmaceutical and health sciences fields.

1.1 Populations and Samples

A **population** is the set of all measurements of interest to a researcher. Typically, the population is not observed, but we wish to make statements or inferences concerning it. Populations can be thought of as *existing* or *conceptual*. *Existing* populations are well–defined sets of data containing elements that could be identified explicitly. Examples include:

PO1 CD_4 counts of every American diagnosed with AIDS as of January 1, 1996.

PO2 Amount of active drug in all 20–mg Prozac capsules manufactured in June 1996.

PO3 Presence or absence of prior myocardial infarction in all American males between 45 and 64 years of age.

Conceptual populations are non–existing, yet visualized, or imaginable, sets of measurements. This could be thought of characteristics of all people with a disease, now or in the near future, for instance. It could also be thought of as the outcomes if some treatment were given to a large group of subjects. In this last setting, we do not give the treatment to all subjects, but we are interested in the outcomes if it had been given to all of them. Examples include:

PO4 Bioavailabilities of a drug's oral dose (relative to i. v. dose) in all healthy subjects under identical conditions.

PO5 Presence or absence of myocardial infarction in all current and future high blood pressure patients who receive short–acting calcium channel blockers.

PO6 Positive or negative result of all pregnant women who would ever use a particular brand of home pregnancy test.

Samples are observed sets of measurements that are subsets of a corresponding population. Samples are used to describe and make inferences concerning the populations from which they arise. Statistical methods are based on these samples having been taken at random from the population. However, in practice, this is rarely the case. We will always assume that the sample is representative of the population of interest. Examples include:

SA1 CD_4 counts of 100 AIDS patients on January 1, 1996.

SA2 Amount of active drug in 2000 20–mg Prozac capsules manufactured during June 1996.

SA3 Prior myocardial infarction status (yes or no) among 150 males aged 45 to 64 years.

SA4 Bioavailabilities of an oral dose (relative to i.v. dose) in 24 healthy volunteers.

SA5 Presence or absence of myocardial infarction in a fixed period of time for 310 hypertension patients receiving calcium channel blockers.

SA6 Test results (positive or negative) among 50 pregnant women taking a home pregnancy test.

1.2 Types of Variables

1.2.1 Quantitative vs Qualitative Variables

The measurements to be made are referred to as **variables**. This refers to the fact that we acknowledge that the outcomes (often referred to as **endpoints** in the medical world) will vary among elements of the population. Variables can be classified as *quantitative* (numeric) or *qualitative* (categorical). We will use the terms numeric and categorical throughout this text, since quantitative and qualitative are so similar. The types of analyses used will depend on what type of variable is being studied. Examples include:

VA1 CD_4 count represents numbers (or counts) of CD_4 lymphocytes per liter of peripheral blood, and thus is **numeric**.

VA2 The amount of active drug in a 20–mg Prozac capsule is the actual number of mg of drug in the capsule, which is **numeric**. Note, due to random variation in the production process, this number will vary and never be exactly 20.0–mg.

VA3 Prior myocardial infarction status can be classified in several ways. If it is classified as either yes or no, it is **categorical**. If it is classified as number of prior MI's, it is **numeric**.

Further, numeric variables can be broken into two types: **continuous** and **discrete**. Continuous variables are values that can fall anywhere corresponding to points on a line segment. Some examples are weight and diastolic blood pressure. Discrete variables are variables that can take on only a finite (or countably infinite) number of outcomes. Number of previous myocardial infarctions

and parity are examples of discrete variables. It should be noted that many continuous variables are reported as if they were discrete, and many discrete variables are analyzed as if they were continuous.

Similarly, categorical variables also are commonly described in one of two ways: **nominal** and **ordinal**. Nominal variables have distinct levels that have no inherent ordering. Hair color and sex are examples of variables that would be described as nominal. On the other hand, ordinal variables have levels that do follow a distinct ordering. Examples in the medical field typically relate to degrees of change in patients after some treatment (such as: vast improvement, moderate improvement, no change, moderate degradation, vast degradation/death).

Example 1.1 In studies measuring pain or pain relief, *visual analogue scales* are often used. These scales involve a continuous line segment, with endpoints labeled as no pain (or no pain relief) and severe (or complete pain relief). Further, there may be adjectives or descriptions written along the line segment. Patients are asked to mark the point along the scale that represents their status. This is treated then as a continuous variable. Figure 1.1 displays scales for pain relief and pain, which patients would mark, and which a numeric score (e.g. percent of distance from bottom to top of scale) can be obtained (Scott and Huskisson, 1976).

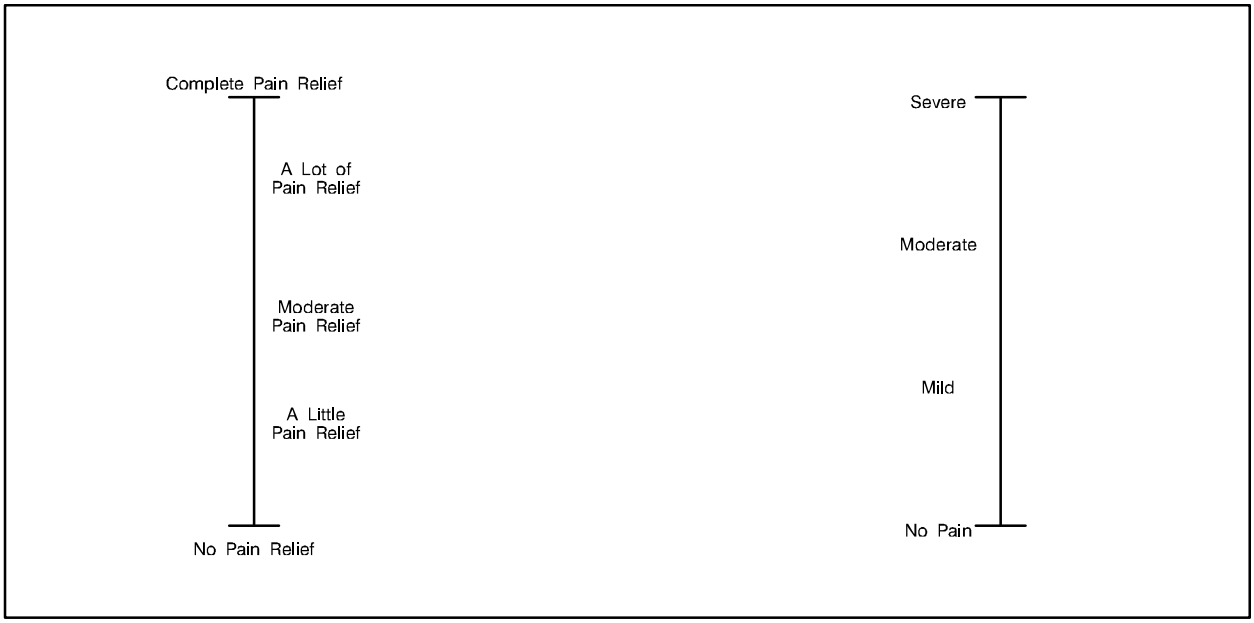

Figure 1.1: Visual Analogue Scales corresponding to pain relief and pain

Example 1.2 In many instances in social and medical sciences, no precise measurement of an outcome can be made. Ordinal scale descriptions (referred to as *Likert scales*) are often used. In one of the first truly random trials in Britain, patients with pulmonary tuberculosis received either streptomycin or no drug (Medical Research Council, 1948). Patients were classified after six months

into one of the following six categories: considerable improvement, moderate/slight improvement, no material change, moderate/slight deterioration, considerable deterioration, or death. This is an ordinal scale.

1.2.2 Dependent vs Independent Variables

Applications of statistics are often based on comparing outcomes among groups of subjects. That is, we'd like to compare outcomes among different populations. The variable(s) we measure as the outcome of interest is the **dependent** variable, or response. The variable that determines the population a measurement arises from is the **independent** variable (or predictor). For instance, if we wish to compare bioavailabilities of various dosage forms, the dependent variable would be AUC (area under the concentration–time curve), and the independent variable would be dosage form. We will extend the range of possibilities of independent variables later in the text. The labels dependent and independent variables have the property that they imply the relationship that the independent variable "leads to" the dependent variable's level. However they have some unfortunate consequences as well. Throughout the text, we will refer to the dependent variable as the **response** and the independent variable(s) as **explanatory** variables.

1.3 Parameters and Statistics

Parameters are numerical descriptive measures corresponding to **populations**. Since the population is not actually observed, the parameters are considered unknown constants. Statistical inferential methods can be used to make statements (or inferences) concerning the unknown parameters, based on the sample data. Parameters will be referred to in Greek letters, with the general case being θ.

For numeric variables, there are two commonly reported types of descriptive measures: *location* and *dispersion*. Measures of location describe the level of the 'typical' measurement. Two measures widely studied are the mean (μ) and the median. The mean represents the arithmetic average of all measurements in the population. The median represents the point where half the measurements fall above it, and half the measurements fall below it. Two measures of the dispersion, or spread, of measurements in a population are the variance σ^2 and the range. The variance measures the average squared distance of the measurements from the mean. Related to the variance is the standard deviation (σ). The range is the difference between the largest and smallest measurements. We will primarily focus on the mean and variance (and standard deviation). A measure that is commonly reported in research papers is the coefficient of variation. This measure is the ratio of the standard deviation to the mean, stated as a percentage ($CV = (\sigma/\mu)100\%$). Generally small values of CV are considered best, since that means that the variability in measurements is small relative to their mean (measurements are consistent in their magnitudes). This is particularly important when data are being measured with scientific equipment, for instance when plasma drug concentrations are measured in assays.

For categorical variables, the most common parameter is π, the proportion having the characteristic of interest (when the variable has two levels). Other parameters that make use of population proportions include *relative risk* and *odds ratios*. These will be described in upcoming sections.

Statistics are numerical descriptive measures corresponding to **samples**. We will use the general notation $\hat{\theta}$ to represent statistics. Since samples are 'random subsets' of the population, statistics are random variables in the sense that different samples will yield different values of the statistic.

In the case of numeric measurements, suppose we have n measurements in our sample, and we label them $y_1, y_2, \ldots, y_n$. Then, we compute the sample mean, variance, standard deviation, and coefficient of variation as follow:

$$\hat{\mu} = \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1} = \frac{(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \cdots (y_n - \overline{y})^2}{n-1} \qquad s = \sqrt{s^2}$$

$$CV = \left(\frac{s}{\overline{y}}\right) 100\%$$

In the case of categorical variables with two levels, which are generally referred to as *Presence* and *Absence* of the characteristic, we compute the sample proportion of cases where the character is present as (where x is the number in which the character is present):

$$\hat{\pi} \quad = \quad \frac{x}{n} \quad = \quad \frac{\text{\# of elements where characteristic is present)}}{\text{\# of elements in the sample (trials)}}$$

Statistics based on samples will be used to estimate parameters corresponding to populations, as well as test hypotheses concerning the true values of parameters.

Example 1.3 A study was conducted to observe the effect of grapefruit juice on cyclosporine and prednisone metabolism in transplant patients (Hollander, et al., 1995). Among the measurements made was creatinine clearance at the beginning of the study. The values for the $n = 8$ male patients in the study are as follow: 38,66,74,99,80,64,80,and 120. Note that here $y_1 = 38, \ldots, y_8 = 120$.

$$\hat{\mu} = \overline{y} = \frac{\sum_{i=1}^{8} y_i}{8} = \frac{38 + 66 + 74 + 99 + 80 + 64 + 80 + 120}{8} = \frac{621}{8} = 77.6$$

$$s^2 = \frac{\sum_{i=1}^{8}(y_i - \overline{y})^2}{8-1} =$$

$$\frac{(38-77.6)^2 + (66-77.6)^2 + (74-77.6)^2 + (99-77.6)^2 + (80-77.6)^2 + (64-77.6)^2 + (80-77.6)^2 + (120-77.6)^2}{8-1}$$

$$= \frac{4167.9}{7} = 595.4 \qquad s = \sqrt{595.4} = 24.4$$

$$CV = \left(\frac{s}{\overline{y}}\right) 100\% = \left(\frac{24.4}{77.6}\right) 100\% = 31.4\%$$

So, for these patients, the mean creatinine clearance is $77.6ml/min$ and the variance is $595.4(ml/min)^2$, the standard deviation is $24.4(ml/min)$, and the coefficient of variation is 31.4%. Thus, the 'typical' patient lies $24.4(ml/min)$ from the overall average of $77.6(ml/min)$, and the standard deviation is 31.4% as large as the mean.

Example 1.4 A study was conducted to quantify the influence of smoking cessation on weight in adults (Flegal, et al., 1995). Subjects were classified by their smoking status (never smoked, quit ≥ 10 years ago, quit < 10 years ago, current cigarette smoker, other tobacco user). We would like to obtain the proportion of current tobacco users in this sample. Thus, people can be classified as current user (the last two categories), or as a current nonuser (the first three categories). The sample consisted of $n = 5247$ adults, of which 1332 were current cigarette smokers, and 253 were other tobacco users. If we are interested in the proportion that currently smoke, then we have $x = 1332 + 253 = 1585$.

$$\hat{\pi} = \frac{\text{\# current tobacco users}}{\text{sample size}} = \frac{x}{n} = \frac{1585}{5247} = .302$$

So, in this sample, .302, or more commonly reported 30.2%, of the adults were classified as current tobacco users. This example illustrates how categorical variables with more than two levels can often be re–formulated into variables with two levels, representing 'Presence' and 'Absence'.

1.4 Graphical Techniques

In the previous section, we introduced methods to describe a set of measurements numerically. In this section, we will describe some commonly used graphical methods. For categorical variables, pie charts and histograms (or vertical bar charts) are widely used to display the proportions of measurements falling into the particular categories (or levels of the variable). For numeric variables, pie charts and histograms can be used where measurements are 'grouped' together in ranges of levels. Also, scatterplots can be used when there are two (or more) variables measured on each subject. Descriptions of each type will be given by examples. The scatterplot will be seen in Chapter 7.

Example 1.5 A study was conducted to compare oral and intravenous antibiotics in patients with lower respiratory tract infection (Chan, et al., 1995). Patients were rated in terms of their final outcome after their assigned treatment (delivery route of antibiotic). The outcomes were classified as: cure (1), partial cure (2), antibiotic extended (3), antibiotic changed (4), death (5). Note that this variable is categorical and ordinal. For the oral delivery group, the numbers of patients falling into the five outcome levels were: 74, 68, 16, 14, and 9, respectively. Figure 1.2 represents a vertical bar chart of the numbers of patients falling into the five categories. The height of the bar represents the frequency of patients in the stated category. Figure 1.3 is a pie chart for the same data. The size of the 'slice' represents the proportion of patients falling in each group.

Example 1.6 Review times of all new drug approvals by the Food and Drug Administration for the years 1985–1992 have been reported (Kaitin, et al., 1987,1991,1994). In all, 175 drugs were approved during the eight–year period. Figure 1.4 represents a histogram of the numbers of drugs

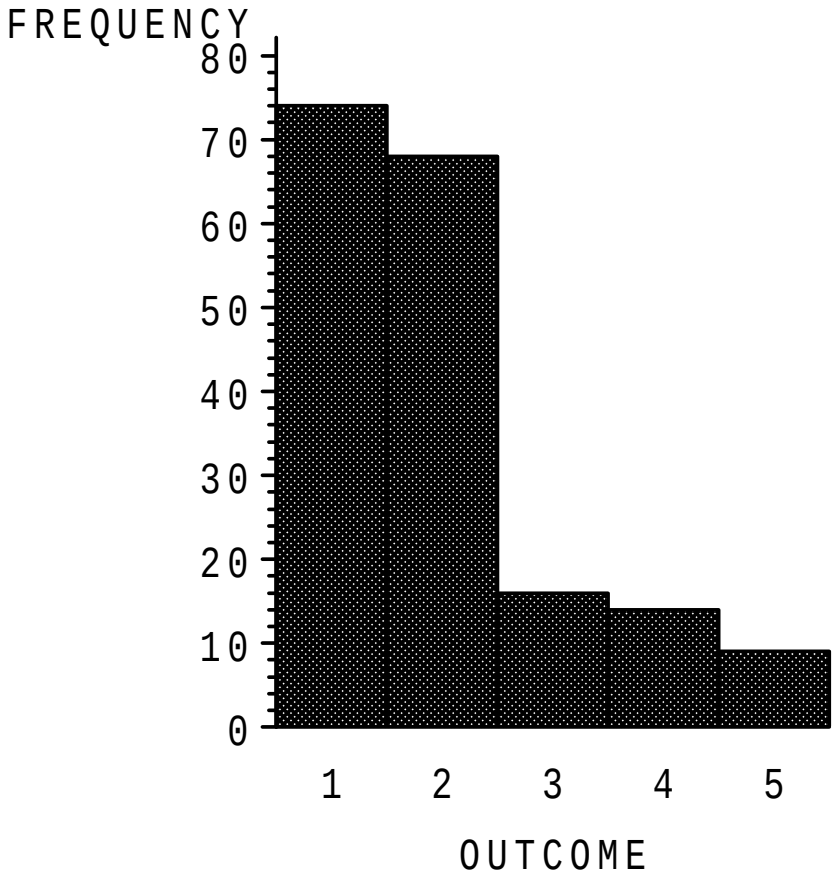

Figure 1.2: Vertical Bar Chart of frequency of each outcome in antibiotic study

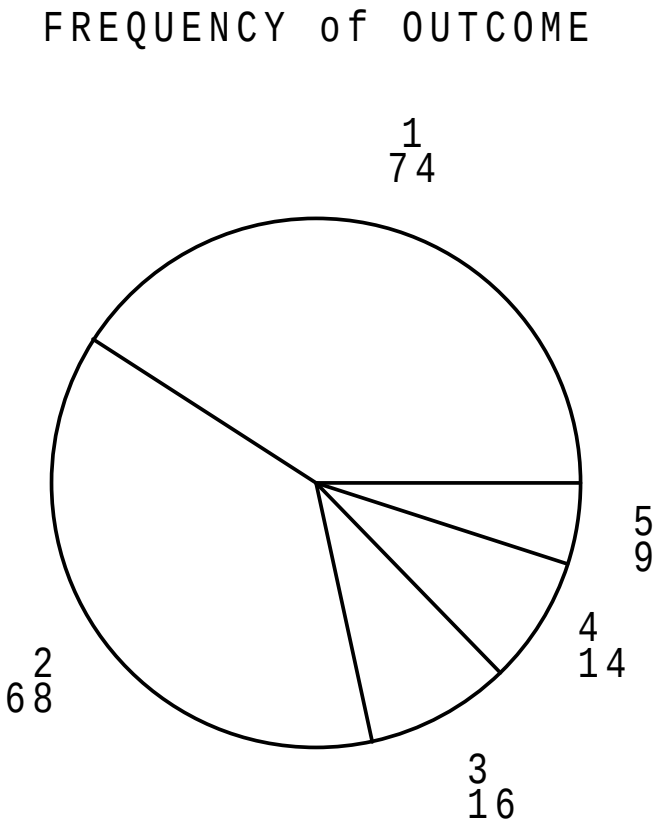

Figure 1.3: Pie Chart of frequency of each outcome in antibiotic study

falling in the categories: 0–10, 10–20, . . . , 110+ months. Note that most drugs fall in the lower (left) portion of the chart, with a few drugs having particularly large review times. This distribution would be referred to as being *skewed right* due to the fact it has a long right tail.

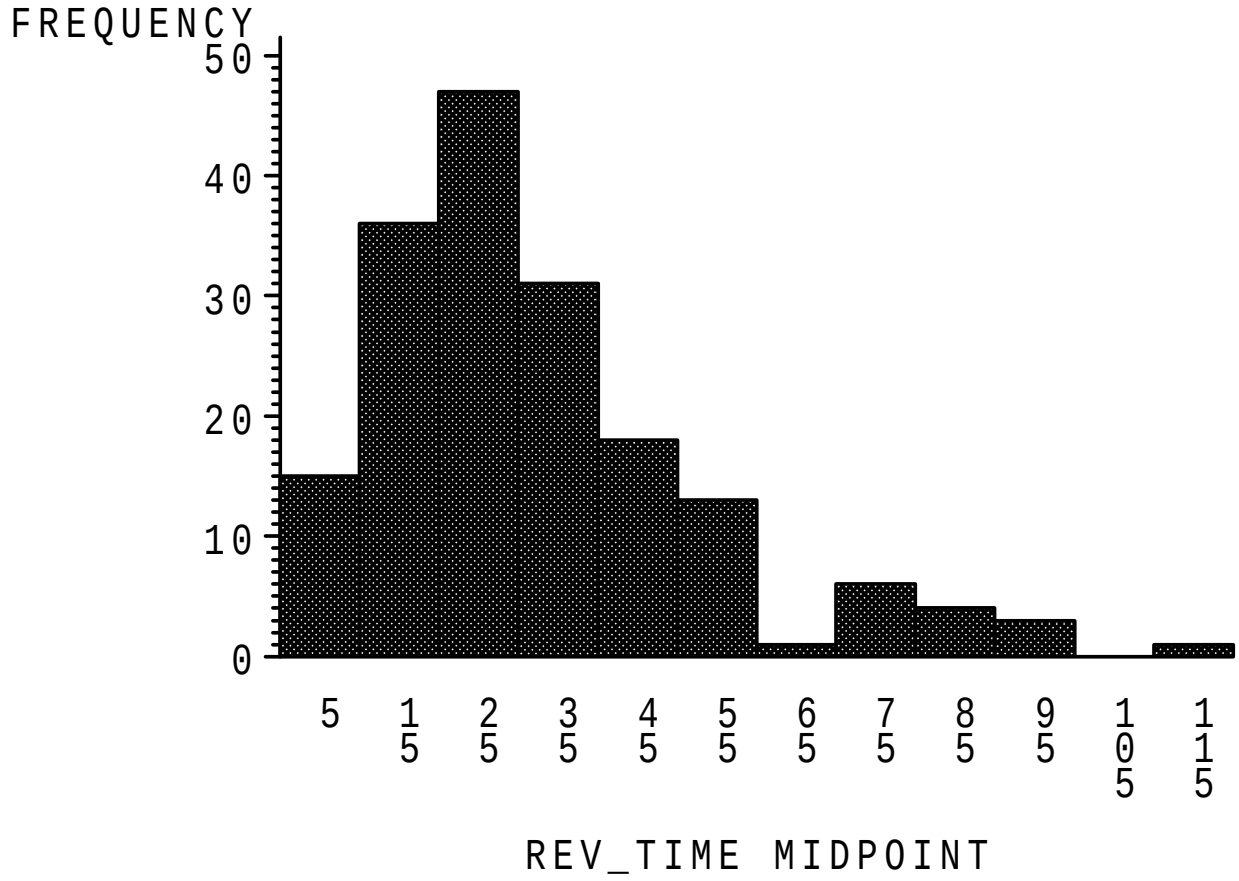

Figure 1.4: Histogram of approval times for 175 drugs approved by FDA (1985–1992)

Example 1.7 A trial was conducted to determine the efficacy of streptomycin in the treatment of pulmonary tuberculosis (Medical Research Council, 1948). After 6 months, patients were classified as follows: 1=considerable improvement, 2=moderate/slight improvement, 3=no material change, 4=moderate/slight deterioration, 5=considerable deterioration, 6=death. In the study, 55 patients received streptomycin, while 52 patients received no drug and acted as controls. Side–by–side vertical bar charts representing the distributions of the clinical assessments are given in Figure 1.5. Note that the patients who received streptomycin fared better in general than the controls.

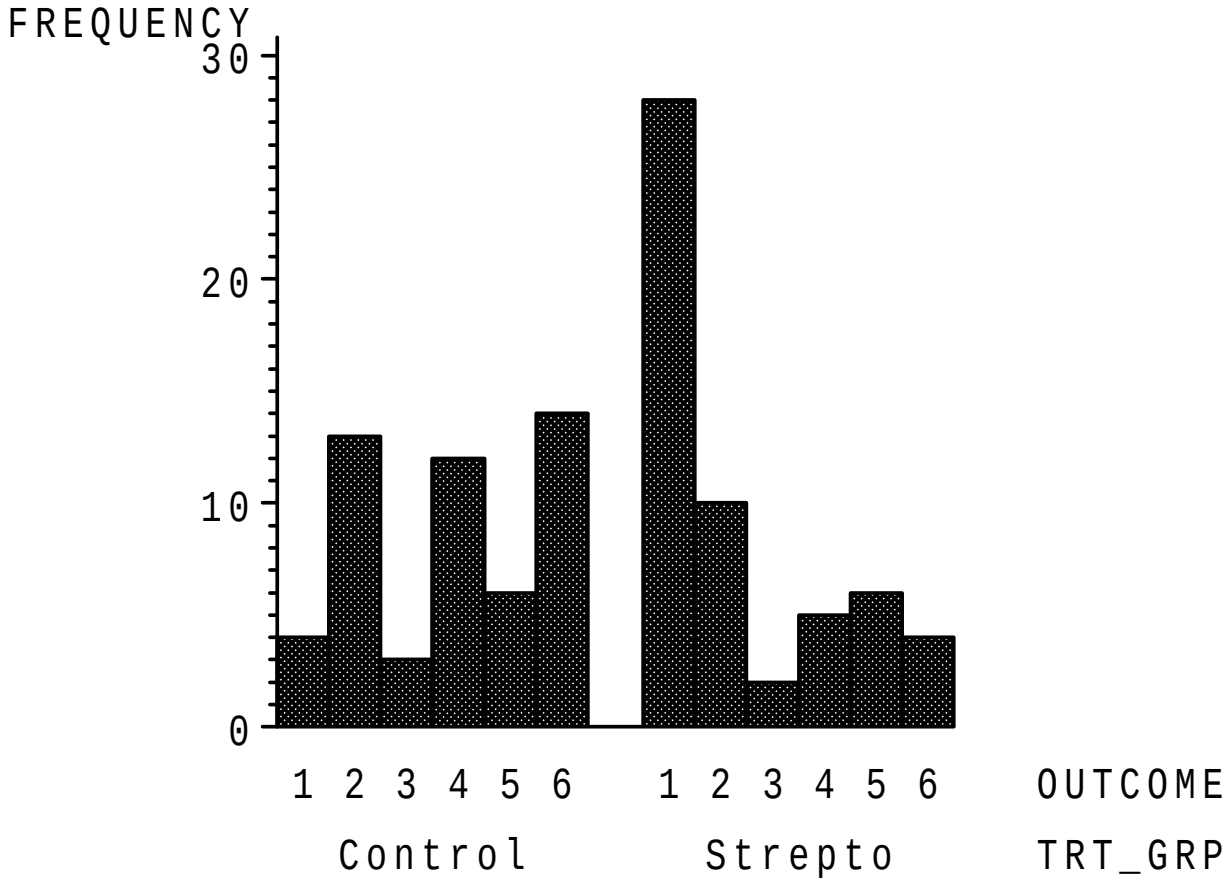

Figure 1.5: Side–by–side histograms of clinical outcomes among patients treated with streptomycin and controls

Example 1.8 The interactions between theophylline and two other drugs (famotidine and cimetidine) were studied in fourteen patients with chronic obstructive pulmonary disease (Bachmann, et al., 1995). Of particular interest were the pharmacokinetics of theophylline when it was being taken simultaneously with each drug. The study was conducted in three periods: one with theophylline and placebo, a second with theophylline and famotidine, and the third with theophylline and cimetidine. One outcome of interest is the clearance of theophylline (liters/hour). The data are given in Table 1.1 and a plot of clearance vs interacting drug (C=cimetidine, F=famotidine, P=placebo) is given in Figure 1.6. A second plot, Figure 1.7 displays the outcomes vs subject, with the plotting symbol being the interacting drug. The first plot identifies the differences in the drugs' interacting effects, while the second displays the patient–to–patient variability.

	Interacting Drug		
Subject	Cimetidine	Famotidine	Placebo
1	3.69	5.13	5.88
2	3.61	7.04	5.89
3	1.15	1.46	1.46
4	4.02	4.44	4.05
5	1.00	1.15	1.09
6	1.75	2.11	2.59
7	1.45	2.12	1.69
8	2.59	3.25	3.16
9	1.57	2.11	2.06
10	2.34	5.20	4.59
11	1.31	1.98	2.08
12	2.43	2.38	2.61
13	2.33	3.53	3.42
14	2.34	2.33	2.54
$\overline{y}$	2.26	3.16	3.08
s	0.97	1.70	1.53

Table 1.1: Theophylline clearances (liters/hour) when drug is taken with interacting drugs

1.5 Basic Probability

Probability is used to measure the 'likelihood' or 'chances' of certain events (prespecified outcomes) of an experiment. Certain rules of probability will be used in this text and are described here. We first will define 2 events A and B, with probabilities $P(A)$ and $P(B)$, respectively. The **intersection** of events A and B is the event that **both** A **and** B occur, the notation being AB (sometimes written $A \cap B$). The **union** of events A and B is the event that **either** A **or** B occur, the notation being $A \cup B$. The **complement** of event A is the event that A **does not** occur, the notation being $\overline{A}$. Some useful rules on obtaining these and other probabilities include:

- $P(A \cup B) = P(A) + P(B) - P(AB)$
- $P(A|B) = P(\text{A occurs } \textbf{given} \text{ B has occurred}) = \frac{P(AB)}{P(B)}$ (assuming $P(B) > 0$)

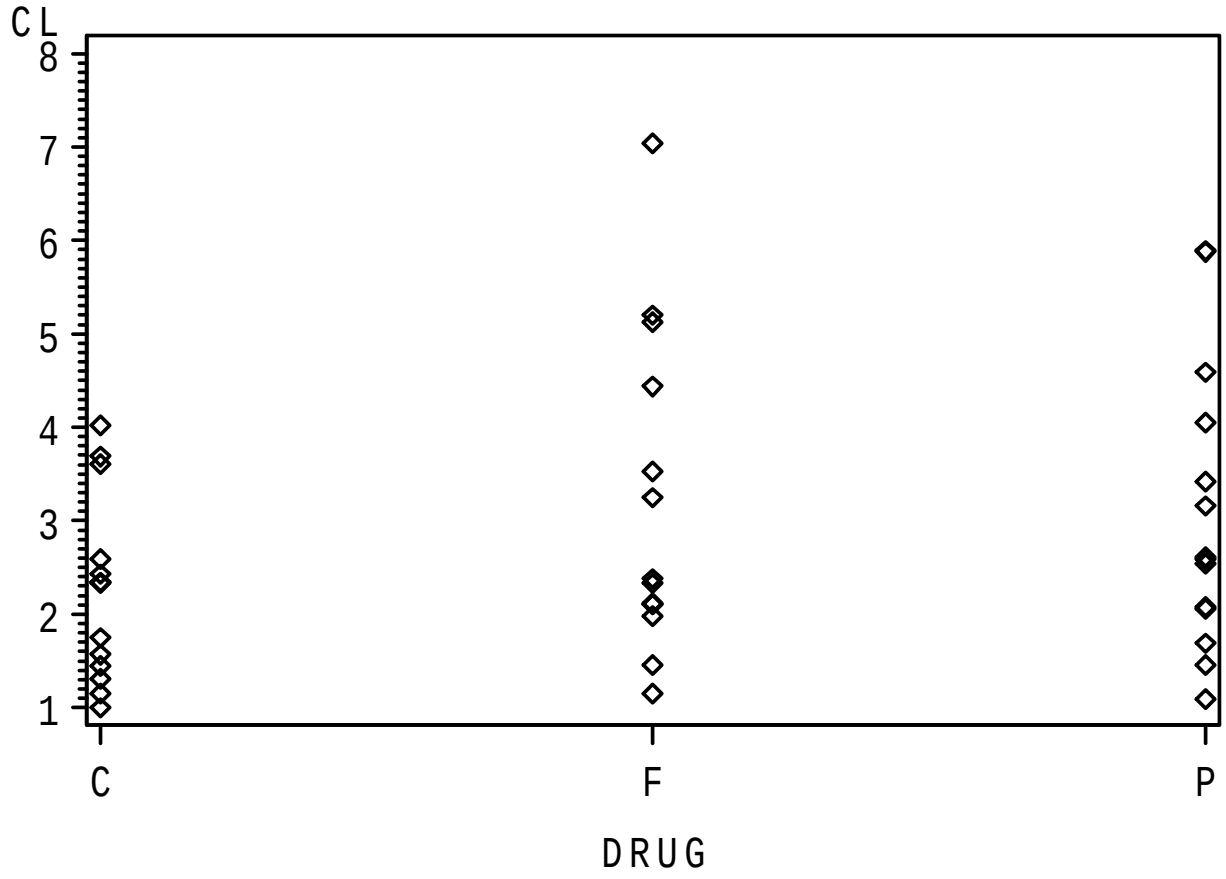

Figure 1.6: Plot of theophylline clearance vs interacting drug

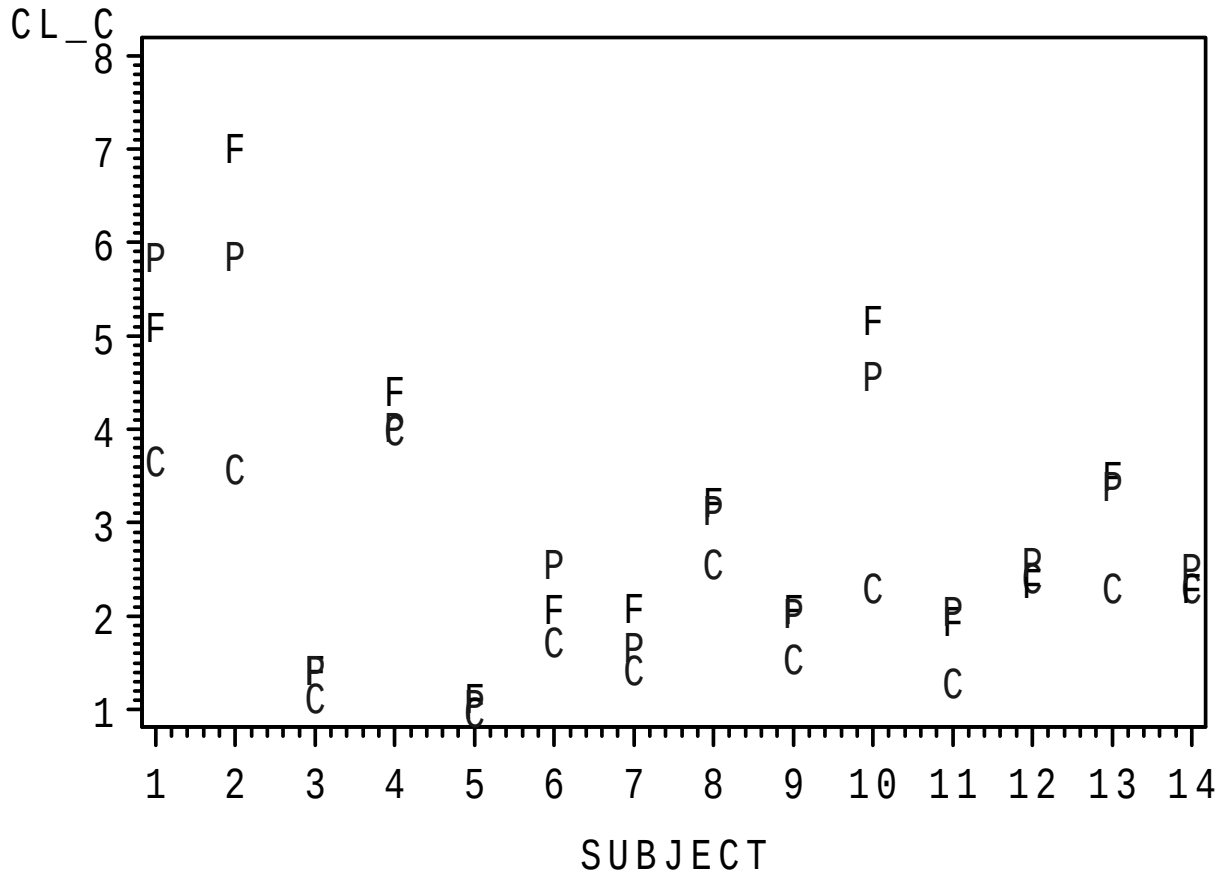

Figure 1.7: Plot of theophylline clearance vs subject with interacting drug as plotting symbol

- $P(AB) = P(A)P(B|A) = P(B)P(A|B)$
- $P(\overline{A}) = 1 - P(A)$

A certain situation occurs when events A and B are said to be **independent**. This is when $P(A|B) = P(A)$, or equivalently $P(B|A) = P(B)$, in this situation, $P(AB) = P(A)P(B)$. We will be using this idea later in this text.

Example 1.9 The association between mortality and intake of alcoholic beverages was analyzed in a cohort study in Copenhagen (Grønbæk, et al., 1995). For purposes of illustration, we will classify people simply by whether or not they drink wine daily (explanatory variable) and whether or not they died (response variable) during the course of the study. Numbers (and proportions) falling in each wine/death category are given in Table 1.2.

	Death Status		
Wine Intake	Death (D)	No Death ($\overline{D}$)	Total
Daily (W)	74 (.0056)	521 (.0392)	595 (.0448)
Less Than Daily ($\overline{W}$)	2155 (.1622)	10535 (.7930)	12690 (.9552)
Total	2229 (.1678)	11056 (.8322)	13285 (1.0000)

Table 1.2: Numbers (proportions) of adults falling into each wine intake/death status combination

If we define the event W to be that the adult drinks wine daily, and the event D to be that the adult dies during the study, we can use the table to obtain some pertinent probabilities:

1. $P(W) = P(WD) + P(W\overline{D}) = .0056 + .0392 = .0448$
2. $P(\overline{W}) = P(\overline{W}D) + P(\overline{W}\overline{D}) = .1622 + .7930 = .9552$
3. $P(D) = P(WD) + P(\overline{W}D) = .0056 + .1622 = .1678$
4. $P(\overline{D}) = P(W\overline{D}) + P(\overline{W}\overline{D}) = .0392 + .7930 = .8322$
5. $P(D|W) = \frac{P(WD)}{P(W)} = \frac{.0056}{.0448} = .1250$
6. $P(D|\overline{W}) = \frac{P(\overline{W}D)}{P(\overline{W})} = \frac{.1622}{.9552} = .1698$

In real terms these probabilities can be interpreted as follows:

1. 4.48% (.0448) of the people in the survey drink wine daily.
2. 95.52% (.9552) of the people do not drink wine daily.
3. 16.78% (.1678) of the people died during the study.
4. 83.22% (.8322) of the people did not die during the study.
5. 12.50% (.1250) of the daily wine drinkers died during the study.

6. 16.98% (.1698) of the non–daily wine drinkers died during the study.

In these descriptions, the proportion of ... can be interpreted as 'if a subject were taken at random from this survey, the probability that it is ...'. Also, note that the probability that a person dies depends on their wine intake status. We then say that the events that a person is a daily wine drinker and that the person dies are **not independent** events.

1.5.1 Diagnostic Tests

Diagnostic testing provides another situation where basic rules of probability can be applied. Subjects in a study group are determined to have a disease (D^+), or not have a disease (D^-), based on a gold standard (a process that can detect disease with certainty). Then, the same subjects are subjected to the newer (usually less traumatic) diagnostic test and are determined to have tested positive for disease (T^+) or tested negative (T^-). Patients will fall into one of four combinations of gold standard and diagnostic test outcomes ($D^+T^+, D^+T^-, D^-T^+, D^-T^-$). Some commonly reported probabilities are given below.

Sensitivity This is the probability that a person with disease (D^+) will correctly test positive based on the diagnostic test (T^+). It is denoted $sensitivity = P(T^+|D^+)$.

Specificity This is the probability that a person without disease (D^-) will correctly test negative based on the diagnostic test (T^-). It is denoted $specificity = P(T^-|D^-)$.

Positive Predictive Value This is the probability that a person who has tested positive on a diagnostic test (T^+) actually has the disease (D^+). It is denoted $PV^+ = P(D^+|T^+)$.

Negative Predictive Value This is the probability that a person who has tested negative on a diagnostic test (T^-) actually does not have the disease (D^-). It is denoted $PV^- = P(D^-|T^-)$.

Overall Accuracy This is the probability that a randomly selected subject is correctly diagnosed by the test. It can be written as $accuracy = P(D^+)sensitivity + P(D^-)specificity$.

Two commonly used terms related to diagnostic testing are **false positive** and **false negative**. False positive is when a person who is nondiseased (D^-) tests positive (T^+), and false negative is when a person who is diseased (D^+) tests negative (T^-). The probabilities of these events can be written in terms of $sensitivity$ and $specificity$:

$$P(\text{False Positive}) = P(T^+|D^-) = 1-specificity \qquad P(\text{False Negative}) = P(T^-|D^+) = 1-sensitivity$$

When the study population is representative of the overall population (in terms of the proportions with and without disease ($P(D^+)$ and $P(D^-)$), positive and negative value can be obtained directly from the table of outcomes (see Example 1.8 below). However, in some situations, the two group sizes are chosen to be the same (equal numbers of diseased and nondiseased subjects). In this case, we must use **Bayes' Rule** to obtain the positive and negative predictive values. We assume that the proportion of people in the actual population who are diseased is known, or well approximated, and is $P(D^+)$. Then, positive and negative predictive values can be computed as:

$$PV^+ = \frac{P(D^+)sensitivity}{P(D^+)sensitivity + P(D^-)(1 - specificity)}$$

$$PV^- = \frac{P(D^-)specificity}{P(D^-)specificity + P(D^+)(1 - sensitivity)}.$$

We will cover these concepts based on the following example. Instead of using the rules of probability to get the conditional probabilities of interest, we will make intituitive use of the observed frequencies of outcomes.

Example 1.10 A noninvasive test for large vessel peripheral arterial disease (LV-PAD) was reported (Feigelson, et al.,1994). Their study population was representative of the overall population based on previous research they had conducted. A person was diagnosed as having LV-PAD if their ankle–to–arm blood pressure ratio was below 0.8, and their posterior tibial peak forward flow was below $3cm/sec$. This diagnostic test was much simpler and less invasive than the detailed test used as a gold standard. The experimental units were left and right limbs (arm/leg) in each subject, with 967 limbs being tested in all. Results (in observed frequencies) are given in Table 1.3. We obtain

	Gold Standard		
Diagnostic Test	Disease (D^+)	No Disease (D^-)	Total
Positive (T^+)	81	9	90
Negative (T^-)	10	867	877
Total	91	876	967

Table 1.3: Numbers of subjects falling into each gold standard/diagnostic test combination

the relevant probabilities below. Note that this study population is considered representative of the overall population, so that we can compute positive and negative predictive values, as well as overall accuracy, directly from the frequencies in Table 1.3.

Sensitivity Of the 91 limbs with the disease based on the gold standard, 81 were correctly determined to be positive by the diagnostic test. This yields a sensitivity of 81/91=.890 (89.0%).

Specificity Of the 876 limbs without the disease based on the gold standard, 867 were correctly determined to be negative by the diagnostic test. This yields a specificity of 867/876=.990 (99.0%).

Positive Predictive Value Of the 90 limbs that tested positive based on the diagnostic test, 81 were truly diseased based on the gold standard. This yields a positive predictive value of 81/90=.900 (90.0%).

Negative Predictive Value Of the 877 limbs that tested negative based on the diagnostic test, 867 were truly nondiseased based on the gold standard. This yields a negative predictive value of 867/877=.989 (98.9%).

Overall Accuracy Of the 967 limbs tested, $81 + 867 = 948$ were correctly diagnosed by the test. This yields an overall accuracy of 948/967=.980 (98.0%).

1.6 Basic Study Designs

Studies can generally be classified in one of two ways: **observational** or **experimental**. Observational studies are those in which investigators observe subjects, classifying them based on levels of one (or more) explanatory variable(s) and a response of interest. Observational studies generally fall in one of three classes (although hybrids are constantly being devised to improve our ability to determine links among variables). The three main classes are case/control, cohort and cross–sectional studies. Experimental studies may be thought of studies where a research makes an intervention (such as giving a particular drug treatment to a patient). Then, the subjects are followed over time, and the response of interest is measured. We will focus on experimental studies with historic controls and randomized clinical trials.

The usefulness in terms of determining causation varies significantly among these types of studies. Randomized clinical trials are considered the best in this sense. Case–control studies are probably the weakest in that sense. However, the quality with which the data are collected can be just as important as the study design (Hill, 1953). Examples of the types we will focus on are given along with the descriptions.

1.6.1 Observational Studies

As stated above, in observational studies, the investigator identifies subjects as they occur in nature, and observes some response of interest for each subject. These types of studies can be prospective (the subject's explanatory level is identified first, then the outcome or response of onterest is observed), or retrospective (subject's outcomes are observed first, then information on any explanatory variable(s) is obtained).

Consider epidemiologic studies to determine the association between cigarette smoking and lung cancer. We could identify smokers and nonsmokers (explanatory variable) and determine whether or not they develop lung cancer in some fixed period (response); this would be prospective. An alternative approach would be to identify hospital patients with and without lung cancer (response) and then determine whether or not the person had smoked (explanatory variable); this would be retrospective. We now describe each of the three major types of observational study designs:

Case–control studies are generally retrospective, and involve identifying subjects based on the level of their response variable, and measuring the level of their explanatory variable (often thought of as some form of exposure). Typically patients with some disease of interest will be identified (cases) as well as a similar set of patients in the same clinical setting who do not have the disease of interest (controls). Then all subjects are asked about their status considering some risk factor of interest. Case–control studies are commonly used when the response of interest is very rare in the population of interest. They also are susceptible to *recall bias*, a situation where cases may be more likely to remember the occurrence of the risk factor than controls. Suppose, for instance, a study involves children born with defects (cases) and those without defects (controls). Mothers of the children with defects may be more likely to recall use of a prescription drug, since they would probably have spent more time contemplating their pregnancy than the control mothers. Generally speaking, case/control studies are the weakest at determining a causal relationship, but may be the quickest and cheapest way to determine risk factors that may be then studied prospectively. This was the case for the studying the ill effects of tobacco use during the early to mid 20^{th} century.

Many strategies to improve inferences based on case/control studies are discussed in a widely sited paper on controlling for external factors in such studies (Mantel and Haenszel,1959).

Example 1.11 A case–control study was conducted in London in the late 1940's to investigate the association between smoking and lung cancer (Doll and Hill, 1950). The researchers identified 709 cases (people in London hospitals suffering from lung cancer), and 709 controls (patients in London hospitals suffering from diseases other than lung cancer). The researchers found that of the 649 male cases, 647 were smokers and among the 649 male controls, 622 were smokers. Combined results for males and females are given in Table 1.4. Note that we do not have an estimate of a percentage of smokers who develop lung cancer, but rather an estimate of the percentage of lung cancer patients who are smokers.

	Lung Cancer		
Smoker	Present	Absent	Total
Yes	688	650	1338
No	21	59	80
Total	709	709	1418

Table 1.4: Numbers of subjects falling into each smoking/lung cancer combination

Cohort studies are generally prospective, and involve identifying subjects based on the level of their explanatory variable, and obtaining the corresponding response outcome. These studies usually involve following the subjects over a period of time to determine their outcome. For instance, many studies have been conducted to compare the rates of breast cancer in women with breast implants and women without breast implants (Bryant and Brasher, 1995). Women were identified as either having breast implants or not (explanatory variable), and were followed over time to see whether or not they were diagnosed with breast cancer (response). Cohort studies are common when it is unethical to assign a condition (such as smoking or breast implants) to subjects, but it is possible to identify existing populations of such subjects. Cohort studies have to have enormous sample sizes when the outcome of interest is rare in the population.

Example 1.12 In the early 1950's, researchers began conducting prospective studies to determine the association between smoking and occurrence of death (Hammond and Horn, 1954). A tremendous cohort of men ranging from 50 to 69 were identified (they have follow–up data on 187766 men). Men were classified as regular cigarette smokers or noncigarette smokers. They were then followed for a period of approximately 2 years. Results are given in Table 1.5. The authors found that the death rate was indeed higher among smokers than non–smokers (even though they tended to be younger).

Cross-sectional studies involve sampling subjects at random from a population and determining the levels of their explanatory and response variables. These are usually conducted retrospectively, based on large medical databases, at the health organization, state, or national level. In these situations they have large numbers of individuals with extensive medical histories on each subject. Subjects are grouped, and associations between variables are investigated.

	Death Status		
Smoker	Dead	Alive	Total
Yes	3002	104820	107822
No	1852	78092	79944
Total	4854	182912	187766

Table 1.5: Numbers of subjects falling into each smoking/death status combination

Example 1.13 A population–based cross–sectional study investigated the relationship between induced abortion and breast cancer in Danish women born between April 1,1935 and March 31,1978 (Melbye, et al, 1997). The study contained a cohort of 1.5 million women, with their abortion and breast cancer status given in Table 1.6. The researchers found no association between abortion and breast cancer.

	Breast Cancer Status		
Abortion	Yes	No	Total
Yes	1338	279627	280965
No	8908	1239639	1248547
Total	10246	1519266	1529512

Table 1.6: Numbers of subjects falling into each abortion/breast cancer status combination

1.6.2 Experimental Studies

Experimental studies are those in which the investigators make an intervention on their subjects. This typically involves assigning a treatment to patients with some disease. We will describe two classes: randomized clinical trials and studies based on historical controls. While randomized clinical trials are considered the gold standard, studies based on historical controls, when conducted properly, can provide strong evidence of treatment effect without some of the ethical concerns attributed to randomized trials (Gehan, 1984).

Randomized Clinical Trials (RCT's) are controlled studies where subjects are selected from a population of patients who meet some physical criteria and randomly assigned into treatment groups. That is, the level of their explanatory variable is assigned at random. These trials became popular in the late 1940's and are considered the gold standard of experiments in terms of determining cause and effect. Sir A. Bradford Hill in London was a leader in the efforts to begin using such studies. When the patient is unaware of which treatment he/she is receiving, it is considered to be a 'blind' trial; if in addition, the clinician making assessments and conducting the trial is unaware of which treatment, it is 'double–blind'. Efficacy studies that are conducted to get new drug approval include randomized clinical trials, where patients are assigned the test drug or placebo (or a standard drug) at random, and followed over time.

In addition, RCT's can be classified in one of two ways: parallel groups and crossover studies. In **parallel groups** studies, each subject receives just one treatment. Thus, the samples from one treatment to another are independent (made up of different subjects). In **crossover** studies,

each subject receives each treatment, acting as his/her control. In these studies the samples are considered to be paired, or blocked (made up of the same subjects).

Ethical problems arise however, when a new treatment is virtually certain to be better than the standard treatment. This problem arises constantly in cancer studies. Should a doctor be expected to place a patient in an inferior treatment group? We discuss some options below.

Example 1.14 One of the first reported attempts at a randomized clinical trial was conducted in Detroit to determine the safety and efficacy of sanocrysin in patients with pulmonary tuberculosis (Amberson, et al.,1931). The researchers selected 24 patients with pulmonary tuberculosis, and created two groups by matching the patients as well as possible. Within each matched pair, one patient was assigned to group 1, and the other to group 2. Then the experimenters flipped a coin to determine which group received sanocrysin, and which group received no treatment (controls). The result was that the sanocrysin proved to be worse than no treatment, as can be seen from Table 1.7. Note that patients were not randomized into groups individually, which is the common practice today.

Group	Slightly Improved	Much Improved	Un–Changed	Slightly Worse	Much Worse	Death
Sanocrysin	5	1	0	1	4	1
Control	6	1	3	1	0	0

Table 1.7: Numbers of subjects falling into each treatment/TB outcome combination

Historical Control Studies involve using subjects that have been treated (or not) previously, from which information is being used to compare a currently tested treatment. These types of studies can be classified in 2 ways: 1) with historical controls who have been treated by the same clinicians at the same institution as the current experimental patients, and 2) with historical controls who have been treated by different researchers at other institutions (Berry, 1990). "Literature controls", comparison groups that are derived from the medical literature would fall in the second group. Ethical reasons lead researchers to use historical control studies in place of randomized clinical trials. These mainly involve the situation where there is strong theoretical and experimental basis to believe the new treatment is far superior to the standard. Note, however, that the mechanism of randomization is not used in these trials.

Researchers in the field of cancer have developed some guidelines to determine whether a historical control study is valid (Gehan, 1984; Pocock, 1976). They are:

- The control group received a precisely defined treatment in previous study.
- Eligibility criteria, procedures, and evaluation must be the same.
- Important prognostic variables must be known and similar for both groups.
- There is no reason to believe there is an external factor that may lead to different results.

Example 1.15 A study invesigated the survival and remission status of patients with advanced Hodgkin's disease (De Vita, et al,1970). The investigators treated patients with a combination of vincristine sulfate, nitrogen mustard (or cyclophosphamide), procarbazine hydrochloride, and prednisone over a 6–month period. They found that 35 of 43 achieved complete remission, and that 28 of the 35 who achieved remission were still alive at the time of the report. They compared the median survival time of their patients (greater than 42 months, since less than half had died) to median survival times of other therapies reported in the literature (none of which exceeded two years).

1.6.3 Other Study Designs

The ethical problems of assigning patients to receive inferior treatment led researchers to develop **sequential designs** that allow them to quit the study early when a clear treatment effect is observed without losing the "statistical purity" of the experiment. In these studies, if a certain level of success of one treatment over the other is observed, then the trial is ended, and patients are all given the superior treatment. We will see later that many of the criteria in determining differences between groups involve fixed sample sizes and risks of errors in experimental conclusions. Sequential designs permit experiments to be conducted at proper significance levels, but without the entire pre–planned sample sizes to be used when an effect begins to appear.

Example 1.16 One of the first (if not the first) application of sequential designs was conducted to study remission in leukemia patients (Freireich, et al, 1963). Among patients in the preliminary phases of the study who reached remission, the efficacy of 6–mercaptopurine in terms of duration of remission was investgated. Patients were paired, based on the level of remission, and randomly assigned to receive either placebo or 6–mercaptopurine. Within each pair, the researchers noted which patient relapsed first. After one member of each pair relapsed a measure of preferences for 6–mp minus preferences for placebos was incremented, and plotted versus the number of pairs observed at that point. When the series crosses either of the lines in in Figure 1.8 we can conclude whether the treatment is effective (crosses upper line) or malignant (crosses lower line). If the series crosses the middle end lines, no treatment effect (positive or negative) can be detected.

The data from the individual pairs are given in Table 1.8 and the graphical description is given in Figure 1.8.

Recently, various hybrids of these basic study designs have been developed. One such design is the **case–crossover analysis**, in which the individuals who are the cases act as their own controls to evaluate changes in risk due to short–term (acute) exposures to a risk factor.

Example 1.17 An example of a study making use of the case–crossover design is a study involving the risk of auto accidents when using cellular telephones (Redelmeier and Tibshirani, 1997). The researchers treated cases as auto accidents among cellular telephone owners. Telephone records were used to determine whether the accident occured during or very close to phone usage. The individuals were their own controls, and information was obtained on cellular phone use on a day prior to the accident at a similar time. Note that this removes any driver to driver variability, since each driver was his/her own control. The authors found that cellular phone use increased risk

Pair	Remission Status	Remission Length (wks) Placebo	6–mp	Preference	6–mp pref – placebo pref
1	Partial	1	10	6–mp	1
2	Complete	22	7	placebo	0
3	Complete	3	32+	6–mp	1
4	Complete	12	23	6–mp	2
5	Complete	8	22	6–mp	3
6	Partial	17	6	placebo	2
7	Complete	2	16	6–mp	3
8	Complete	11	34+	6–mp	4
9	Complete	8	32+	6–mp	5
10	Complete	12	25+	6–mp	6
11	Complete	2	11+	6–mp	7
12	Partial	5	20+	6–mp	8
13	Complete	4	19+	6–mp	9
14	Complete	15	6	placebo	8
15	Complete	8	17+	6–mp	9
16	Partial	23	35+	6–mp	10
17	Partial	5	6	6–mp	11
18	Complete	11	13	6–mp	12
19	Complete	4	9+	6–mp	13
20	Complete	1	6+	6–mp	14
21	Complete	8	10+	6–mp	15

Table 1.8: Results of remission in pairs of leukemia subjects, where one subject in each pair received placebo, the other 6–mp – sequential design

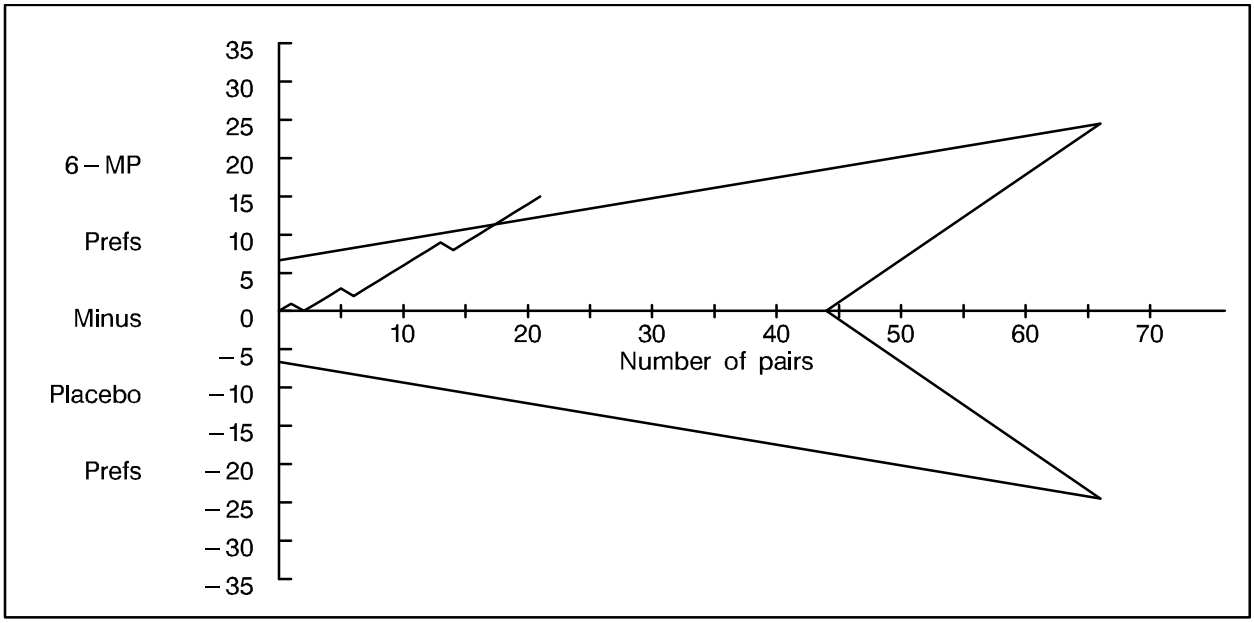

Figure 1.8: Plot of sequential preference of 6–mp minus preference of placebo versus number of pairs observed – sequential design

by a factor of 4.

1.7 Reliability and Validity

While modern medicine has developed very accurate measures and tests for many conditions, many measures that are obtained are less scientific, often obtained based on questionnaires and written tests. Validity and reliability are important in determining whether we are measuring the correct thing and how consistent our measurements are, as well as whether we are observing a causal relationship among variables.

Reliability refers to the degree to which results of a test can be replicated if the test were re-given to the same individual under the same conditions. For instance, if your blood pressure were taken, and then re-taken an hour later under the same conditions, reliability would be a measure of how consistent the two measures were.

Validity has many different forms, and for a complete description, see Chapter 2 of Cook and Campbell (1979), from which much of this section is borrowed. Most (if not all) of the statistical methods described in this text are used to determine whether there is an association between an independent variable (or set of independent variables) and a dependent variable (response variable).

Internal validity refers to the extent that we can infer that changes in the independent variable(s) **cause** changes in the response variable. Internal validity thus is charged with ruling out alternative explanations for the observed association. Randomization is one important way to obtain internal validity, and as we will see in later chapters, other means are to control for external factors that

may also be associated with the outcome being measured.

Example 1.18 The Lanarkshire milk experiment was conducted in England to measure the effects of drinking milk on growth of children. In a well planned experiment, 20,000 children were selected from 67 schools. 5000 students received 0.75 pint of raw milk per day, 5000 students received 0.75 pint of pasteurized milk per day, and 10,000 children acted as controls, not receiving milk. Children's heights and weights were obtained in February prior to the study and again in June after the experiment. A critique of the study (Student, 1931) revealed the following information:

- Schools either had raw milk or pasteurized milk, but not both due to logistical issues.
- Teachers originally assigned students to treatment (milk) and control groups at random.
- However, "after invoking the goddess of chance they wavered in their adherence to her" and they substituted students to better balance the groups based on their unscientific judgment. The final outcome being that the control group 3 months in weight and 4 months in height greater than the treatment groups on average prior to the experiment.

To quote Student (1931, p. 406): "To sum up: The Lanarkshire experiment devised to find out the value of giving a regular supply of milk to children, though planned on the grand scale, organised in a thoroughly business–like manner and carried out with the devoted assistance of a large team of teachers, nurses, and doctors, failed to produce a valid estimate of the advantage of giving milk to children and of the difference between raw and pasteurized milk."

Even experiments that are based on random assignment to treatments, and reliable and well–defined outcome measures can have threats to the internal validity of a study. For instance, in clinical trials, patients are assigned at random to treatment conditions. However, inevitably patients are not observed over the entire study period. Some may move out of town, some may quit their treatment due to various causes (treatment related or not treatment related). When the final analysis is to be conducted researchers must decide who to include in the final analysis. If Mrs. Smith quit taking her allergy medicine during the trial, should her final score be included?

In the medical literature, typically results will be reported for **intention–to–treat** or **completed protocol** analyses. Their names clearly imply which patients are included in their final analyses. However, it's important to find out which patients "fell through the cracks" and failed to complete the experimental protocol and why. The internal validity of a study can be threatened by the fact that the "attrition rates" for the groups differ. Most studies will report these numbers so researchers can determine whether the rates differ significantly for the groups.

External Validity refers to the extent that the observed experimental results regarding cause and effect can be generalized to other settings or populations. In marketing a new drug or product, what works in the United States may not work in Europe or Asia. Similarly, a drug that has an effect in one patient population may not work in others. After initial studies demonstrate efficacy in one patient group, later studies are typically conducted in other populations to demonstrate the external validity of earlier findings (as well as to get FDA approval for new indications and/or target patient groups).

Example 1.19

A study was conducted to measure the benefits of Pravastatin on cardiovascular events and death in older patients with coronary heart disease and average cholesterol levels (Hunt, et al, 2001). Although it was well known that Pravastatin reduced heart related events in moderately aged men with elevated cholesterol levels, this study reported reduced risk of death from somewhere between 7% to 32% in older patients between 65 and 75 years old. This study demonstrated the positive effects of Pravastatin in patient populations not included in the original West of Scotland clinical trials.

1.8 Exercises

1. A study was conducted to examine effects of dose, renal function, and arthritic status on the pharmacokinetics of ketoprofen in elderly subjects (Skeith, et al.,1993). The study consisted of five nonarthritic and six arthritic subjects, each receiving 50 and 150 mg of racemic ketoprofen in a crossover study. Among the pharmacokinetic indices reported is $AUC_{0-\infty}(mg/L)hr$ for S–ketoprofen at the 50 mg dose. Compute the mean, standard deviation, and coefficient of variation, for the arthritic and non–arthritic groups, separately. Do these groups tend to differ in terms of the extent of absorption, as measured by AUC?

Non–arthritic: 6.84, 9.29, 3.83, 5.95, 5.77

Arthritic: 7.06, 8.63, 5.95, 4.75, 3.00, 8.04

2. A study was conducted to determine the efficacy of a combination of methotrexate and misoprostol in early termination of pregnancy (Hausknecht, 1995). The study consisted of $n = 178$ pregnant women, who were given an intra–muscular dose of methotrexate ($50mg/m^2$ of body–surface–area), followed by an intravaginal dose of misoprostol ($800\mu g$) five to seven days later. If no abortion occurred within seven days, the woman was offered a second dose of misoprostol or vacuum aspiration. 'Success' was determined to be a complete termination of pregnancy within seven days of the first or second dose of misoprostol. In the study, 171 women had successful terminations of pregnancy, of which 153 had it after the first dose of misoprostol. Compute the proportions of women: 1) who had successful termination of pregnancy, and 2) the proprtion of women who had successful termination after a single dose of misoprostol.

3. A controlled study was conducted to determine whether or not beta carotene supplementation had an effect on the incidence of cancer (Hennekens, et al.,1996). Subjects enrolled in the study were male physicians aged 40 to 84. In the double–blind study, subjects were randomized to receive $50mg$ beta carotene on separate days, or a placebo control. Endpoints measured were the presence or absence of malignant neoplasms and cardiovascular disease during a 12+ year follow–up period. Numbers (and proportions) falling in each beta carotene/malignant neoplasm (cancer) category are given in Table 1.9.

(a) What proportion of subjects in the beta carotene group contracted cancer during the study? Compute $P(C|B)$.

	Cancer Status		
Trt Group	Cancer (C)	No Cancer ($\overline{C}$)	Total
Beta Carotene (B)	1273 (.0577)	9763 (.4423)	11036 (.5000)
Placebo ($\overline{B}$)	1293 (.0586)	9742 (.4414)	11035 (.5000)
Total	2566 (.1163)	19505 (.8837)	22071 (1.0000)

Table 1.9: Numbers (proportions) of subjects falling into each treatment group/cancer status combination

(b) What proportion of subjects in the placebo group contracted cancer during the study? Compute $P(C|\overline{B})$.

(c) Does beta carotene appear to be associated with decreased risk of cancer in this study population?

4. Medical researcher John Snow conducted a vast study during the cholera epidemic in London during 1853–1854 (Frost, 1936). Snow found that water was being distributed through pipes of two companies: the Southwark and Vauxhall company, and the Bambeth company. Apparently the Lambeth company was obtaining their water from the Thames upstream from the London sewer outflow, while the Southwark and Vauxhall company was obtaining theirs near the sewer outflow (Gehan and Lemak, 1994). Based on Snow's work, and the water company records of customers, people in the South Districts of London could be classified by water company and whether or not they died from cholera. The results are given in Table 1.10.

	Cholera Death Status		
Water Company	Cholera Death (C)	No Cholera Death ($\overline{C}$)	Total
Lambeth (L)	407 (.000933)	170956 (.391853)	171363 (.392786)
S&V ($\overline{L}$)	3702 (.008485)	261211 (.598729)	264913 (.607214)
Total	4109 (.009418)	432167 (.990582)	436276 (1.0000)

Table 1.10: Numbers (proportions) of subjects falling into each water company/cholera death status combination

(a) What is the probability a randomly selected person (at the start of the period) would die from cholera?

(b) Among people being provided water by the Lambeth company, what is the probability of death from cholera? Among those being provided by the Southwark and Vauxhall company?

(c) Mortality rates are defined as the number of deaths per a fixed number of people exposed. What was the overall mortality rate per 10,000 people? What was the mortality rate per 10,000 supplied by Lambeth? By Southwark and Vauxhall? (Hint: Multiply the probability by the fixed number exposed, in this case it is 10,000).

5. An improved test for prostate cancer based on percent free serum prostrate–specific antigen (PSA) was developed (Catalona, et al.,1995). Problems had arisen with previous tests due to a large per-

centage of "false positives", which leads to large numbers of unnecessary biopsies being performed. The new test, based on measuring the percent free PSA, was conducted on 113 subjects who scored outside the normal range on total PSA concentration (thus all of these subjects would have tested positive based on the old – total PSA – test). Of the 113 subjects, 50 had cancer, and 63 did not, as determined by a gold standard. Based on the outcomes of the test (T^+ if %PSA leq20.3, T^- otherwise) given in Table 1.11, compute the sensitivity, specificity, postive and negative predictive values, and accuracy of the new test. Recall that all of these subjects would have tested positive on the old (total PSA) test, so any specificity is an improvement.

	Gold Standard		
Diagnostic Test	Disease (D^+)	No Disease (D^-)	Total
Positive (T^+)	45	39	84
Negative (T^-)	5	24	29
Total	50	63	113

Table 1.11: Numbers of subjects falling into each gold standard/diagnostic test combination – Free PSA exercise

6. A study reported the use of peritoneal washing cytology in gynecologic cancers (Zuna and Behrens, 1996). One part of the report was a comparison of peritoneal washing cytology and peritoneal histology in terms of detecting cancer of the ovary, endometrium, and cervix. Using the histology determination as the gold standard, and the washing cytology as the new test procedure, determine the sensitivity, specificity, overall accuracy, and positive and negative predictive values of the washing cytology procedure. Outcomes are given in Table 1.12.

	Gold Standard		
Diagnostic Test	Disease (D^+)	No Disease (D^-)	Total
Positive (T^+)	116	4	120
Negative (T^-)	24	211	235
Total	140	215	355

Table 1.12: Numbers of subjects falling into each gold standard/diagnostic test combination – gynecologic cancer exercise

7. For the following study descriptions, state what type of design the researchers used (all data will be analyzed where appropriate throughout these notes).

(a) A study reported the association between induced abortion and incidence of breast cancer among Danish women born between 1935 and 1978 (Melbye, et al.1997). All women were classified as to whether or not they had an induced abortion, based on national health records. Also determined was whether or not the woman had developed breast cancer during her life. The authors found that among the $280,965$ women with induced abortions ($2,697,000$ person–years of follow–up), there were 1338 cases of breast cancer (0.47% of women, or 5.0 cases per $10,000$ person–years of exposure). Among the $1,248,547$ women with no induced abortions

(25,850,000 person–years of follow–up), there were 8908 cases of breast cancer (0.71% of women, or 3.4 cases per 10,000 person–years of exposure).

(b) A study reported the sexual side effects of four antidepressants: bupropion, fluoxetine, paroxetine, and sertraline (Modell, et al.,1997). The researchers gave a survey to patients who were receiving antidepressant therapy on one of these four brands. The patients then responded to questions regarding their present sexual state compared to before use of the drug, by use of a visual analogue scale. The authors found that bupropion had a better effect than each of the other three drugs (see Chapter 6).

(c) A study was conducted to determine the existent of an effect due to antihistaminic drugs against the common cold (Medical Research Council, 1950). In one part of the study 8 subjects were randomly assigned to receive histantin (Burroughs Wellcome) and 8 were randomly assigned a control tablet. Subjects were innoculated with a cold virus 48 hours after treatment began. In each group, 4 suffered from a cold, and 4 did not.

(d) A study reported the efficacy of the aniseptic system of treatment on salubrity in a surgical hospital (Lister,1870). The doctor reported the the numbers of deaths and recoveries among all amputations in the Glasgow Royal Infirmary prior to use of antiseptics (years 1864 and 1866), and after the use of antiseptics began (years 1867–1869). In the years prior to use of antiseptics, among 35 amputations, 16 resulted in death (19 recoveries). In years after the introduction of antiseptics, there were 40 amputations, of which 6 resulted in death (34 recoveries). The author acknowledged that amputations of the upper limb were different in nature. Of these upper limb amputations, there were 12 prior to use of antiseptic and 12 after the use of antiseptic. Of the 12 prior to antiseptic use, there were 6 deaths, of the 12 after its use there was 1 death.

8. A study reported patient and partner satisfaction with Viagra treatment based on the Erectile Dysfunction Inventory of Treatment Satisfaction (EDITS) questionnaire (Lewis, et al, 2001). A sample of 247 patients with erectile dysfunction were randomly assigned to one of 4 treatment groups: placebo control, 25, 50, and 100mg of Viagra, and followed for 12 weeks. Efficacy was measured by ability to obtain an erection and ability to obtain an erection (1=Almost Never/Never) to (5=Almost Always/Always). Patient satisfaction was measured on the 11 question EDITS scale, with a range from 0 (extremely low) to 100 (extremely high).

(a) Identify the dependent and independent variables. What types of variables are they?

(b) If half the patients received placebo (actually 123), and the remainder received Viagra, approximately how many were assigned to the 3 dose groups (assuming approximately the same number were in each group)?

(c) What do validity and reliability mean with respect to the EDITS scale?

(d) The data analyses were based on the 247 intent–to–treat patients who had taken at least one dose of study drug and had at least one efficacy assessment. Two patients quit treatment due to adverse side effects. Were their scores included in the final analyses?

Chapter 2

Random Variables and Probability Distributions

We have previously introduced the concepts of populations, samples, variables and statistics. Recall that we observe a sample from some population, measure a variable outcome (categorical or numeric) on each element of the sample, and compute statistics to describe the sample (such as $\overline{Y}$ or $\hat{\pi}$). The variables observed in the sample, as well as the statistics they are used to compute, are **random variables**. The idea is that there is a population of such outcomes, and we observe a random subset of them in our sample. The collection of all possible outcomes in the population, and their corresponding relative frequencies is called a **probability distribution**. Probability distributions can be classified as continuous or discrete. In either case, there are parameters associated with the probabilty distribution; we will focus our attention on making inferences concerning the population mean (μ), the median and the proportion having some characteristic (π).

2.1 The Normal Distribution

Many continuous variables, as well as sample statistics, have probability distibutions that can be thought of as being bell-shaped. That is, most of the measurements in the population tend to fall around some center point (the mean, μ), and as the distance from μ increases, the relative frequency of measurements decreases. Variables (and statistics) that have probability distributions that can be characterized this way are said to be **normally distributed**. Normal distributions are symmetric distributions that are classified by their mean (μ), and their standard deviation (σ).

Random variables that are approximately normally distributed have the following properties:

1. Approximately half (50%) of the measurements fall above (and thus, half fall below) the mean.
2. Approximately 68% of the measurements fall within one standard deviation of the mean (in the range $(\mu - \sigma, \mu + \sigma)$).
3. Approximately 95% of the measurements fall within two standard deviations of the mean (in the range $(\mu - 2\sigma, \mu + 2\sigma)$.

4. Virtually all of the measurements lie within three standard deviations of the mean.

If a random variable, Y, is normally distributed, with mean μ and standard deviation σ, we will use the notation $Y \sim N(\mu, \sigma)$. If $Y \sim N(\mu, \sigma)$, we can write the statements above in terms of probability statements:

$$P(Y \geq \mu) = 0.50 \quad P(\mu - \sigma \leq Y \leq \mu + \sigma) = 0.68 \quad P(\mu - 2\sigma \leq Y \leq \mu + 2\sigma) = 0.95 \quad P(\mu - 3\sigma \leq Y \leq \mu + 3\sigma) \approx 1$$

Example 2.1 Heights (or lengths) of adult animals tend to have distributions that are well described by normal distributions. Figure 2.1 and Figure 2.2 give relative frequency distributions (histograms) of heights of 25–34 year old females and males, respectively (U.S. Bureau of the Census, 1992). Note that both histograms tend to be bell–shaped, with most people falling relatively close to some overall mean, with fewer people falling in ranges of increasing distance from the mean. Figure 2.3 gives the 'smoothed' normal distributions for the females and males. For the females, the mean is 63.5 inches with a standard deviation of 2.5. Among the males, the mean 68.5 inches with a standard deviation of 2.7. If we denote a randomly selected female height as Y_F, and a randomly selected male height as Y_M, we could write: $Y_F \sim N(63.5, 2.5)$ and $Y_M \sim N(68.5, 2.7)$.

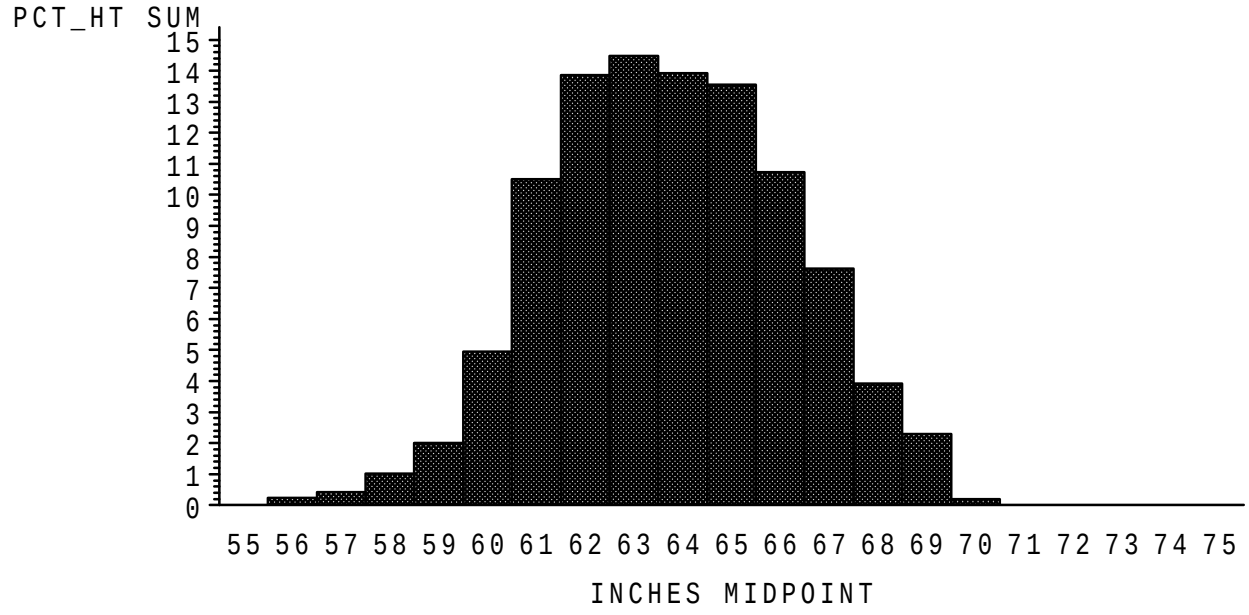

Figure 2.1: Histogram of the population of U.S. female heights (age 25–34)

While there are an infinite number of combinations of μ and σ, and thus an infinite number of possible normal distributions, they all have the same fraction of measurements lying a fixed number of standard deviations from the mean. We can standardize a normal random variable by the following transformation:

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

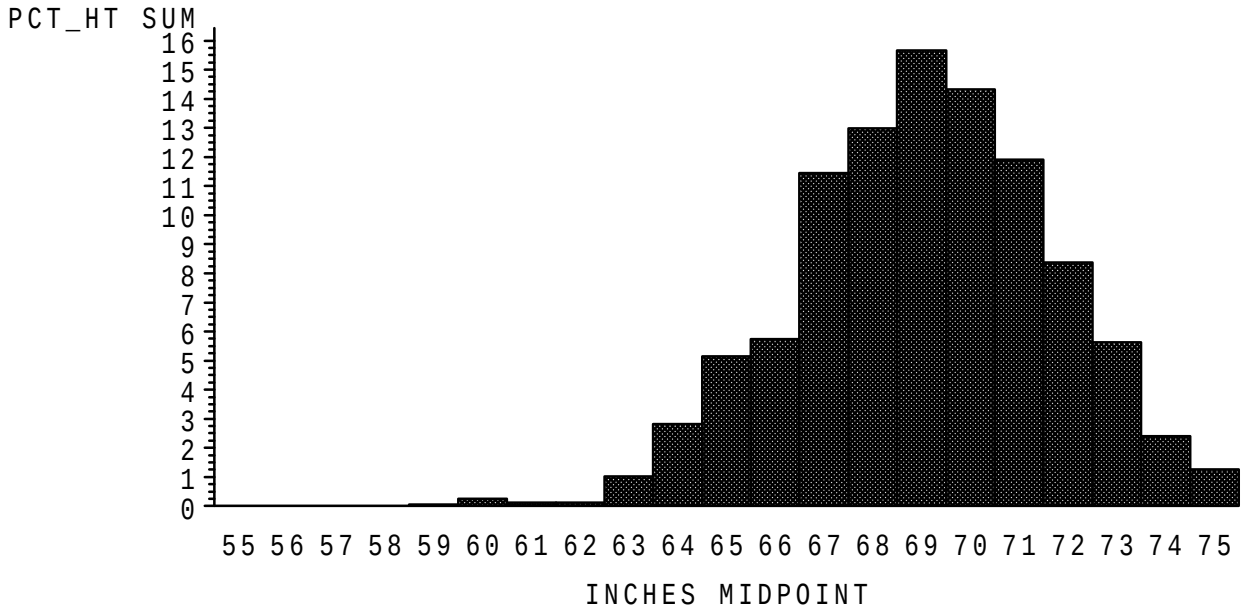

Figure 2.2: Histogram of the population of U.S. male heights (age 25–34)

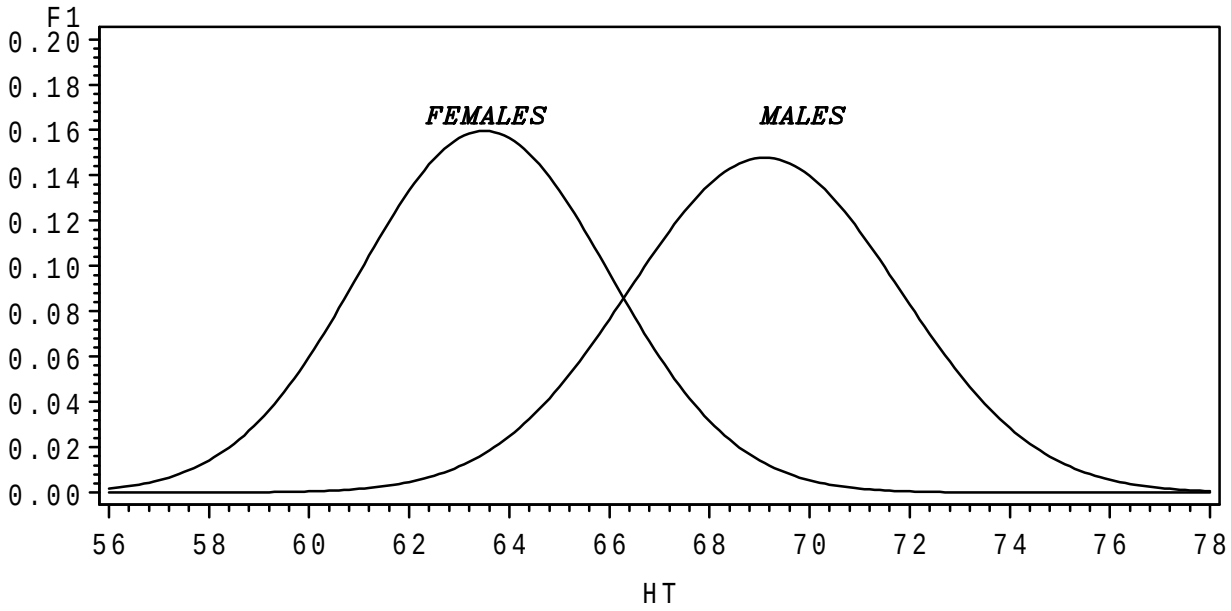

Figure 2.3: Normal distributions used to approximate the distributions of heights of males and females (age 25–34)

Note that Z gives a measure of 'relative standing' for Y; it is the number of standard deviations above (if positive) or below (if negative) the mean that Y is. For example, if a randomly selected female from the population described in the previous section is observed, and her height, Y is found to be 68 inches, we can standardize her height as follows:

$$Z = \frac{Y-\mu}{\sigma} = \frac{68-63.5}{2.5} = 1.80$$

Thus, a woman of height 5'8" is 1.80 standard deviations above the mean height for women in this age group. The random variable, Z, is called a **standard normal random variable**, and is written as $Z \sim N(0,1)$.

Tables giving probabilities of areas under the standard normal distribution are given in most statistics books. We will use the notation that z_a is the value such that the probability that a standard normal random variable is larger than z_a is a. Figure 2.4 depicts this idea. Table A.1 gives the area, a, for values of z_a between 0 and 3.09.

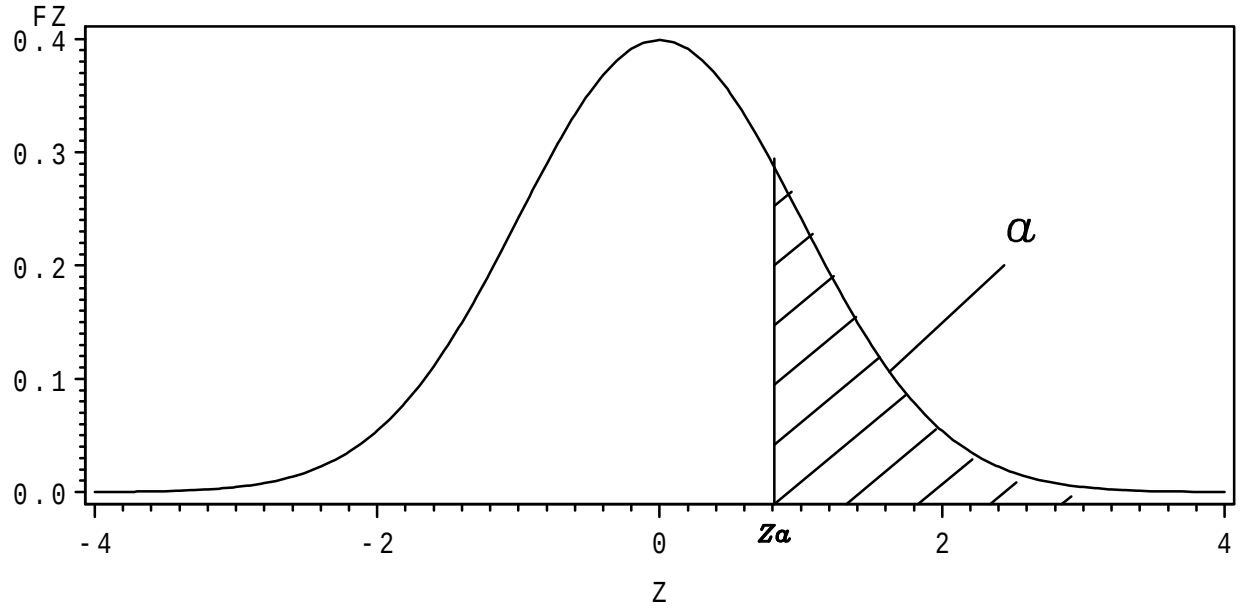

Figure 2.4: Standard Normal Distribution depicting z_a and the probability Z exceeds it

Some common values of a, and the corresponding value, z_a are given in Table 2.1. Since the normal distribution is symmetric, the area below $-z_a$ is equal to the area above z_a, which by definition is a. Also note that the total area under any probability distribution is 1.0.

Many random variables (and sample statistics based on random samples) are normally distributed, and we will be able to use procedures based on these concepts to make inferences concerning unknown parameters based on observed sample statistics.

One example that makes use of the standard normal distribution to obtain a percentile in the original units by 'back–transforming' is given below. It makes use of the standard normal

a	0.500	0.100	0.050	0.025	0.010	0.005
z_a	0.000	1.282	1.645	1.960	2.326	2.576

Table 2.1: Common values of a and its corresponding cut–off value z_a for the standard normal distribution

distribution, and the following property:

$$Z = \frac{Y-\mu}{\sigma} \qquad \Longrightarrow \qquad Y = \mu + Z\sigma$$

That is, an upper (or lower) percentile of the original distribution can be obtained by finding the corresponding upper (or lower) percentile of the standard normal distribution and 'back–transforming'.

Example 2.2 Assume male heights in the 25–34 age group are normally distributed with $\mu = 68.5$ and $\sigma = 2.7$ (that is: $Y_M \sim N(69.1, 2.7)$). Above what height do the tallest 5% of males exceed? Based on the standard normal distribution, and Table 2.1, we know that $z_{.05}$ is 1.645 (that is: $P(Z \geq 1.645) = 0.05$). That means that approximately 5% of males fall at least 1.645 standard deviations above the mean height. Making use of the property stated above, we have:

$$y_{m(.05)} = \mu + z_{.05}\sigma = 68.5 + 1.645(2.7) = 72.9$$

Thus, the tallest 5% of males in this age group are 72.9 inches or taller (assuming heights are approximately normally distributed with this mean and standard deviation). A probability statement would be written as: $P(Y_M \geq 72.9) = 0.05$.

2.1.1 Statistical Models

This chapter introduced the concept of normally distributed random variables and their probability distributions. When making inferences, it is convenient to write the random variable in a form that breaks its value down into two components – its mean, and its 'deviation' from the mean. We can write Y as follows:

$$Y = \mu + (Y - \mu) = \mu + \varepsilon,$$

where $\varepsilon = Y - \mu$. Note that if $Y \sim N(\mu, \sigma)$, then $\varepsilon \sim N(0, \sigma)$. This idea of a statistical model is helpful in statistical inference.

2.2 Sampling Distributions and the Central Limit Theorem

As stated at the beginning of this section, sample statistics are also random variables, since they are computed based on elements of a random sample. In particular, we are interested in the distributions of $\overline{Y}$ ($\hat{\mu}$) and $\hat{\pi}$, the sample mean for a numeric variable and the sample proportion for a categorical variable, respectively. It can be shown that when the sample sizes get large, the sampling distributions of $\overline{Y}$ and $\hat{\pi}$ are approximately normal, regardless of the shape of the probability distribution of the individual measurements. That is, when n gets large, the random

variables $\overline{Y}$ and $\hat{\pi}$ are random variables with probability (usually called **sampling**) distributions that are approximately normal. This is a result of the Central Limit Theorem.

2.2.1 Distribution of $\overline{Y}$

We have just seen that when the sample size gets large, the sampling distribution of the sample mean is approximately normal. One interpretation of this is that if we took repeated samples of size n from this population, and computed $\overline{Y}$ based on each sample, the set of values $\overline{Y}$ would have a distribution that is bell–shaped. The mean of the sampling distribution of $\overline{Y}$ is μ, the mean of the underlying distribution of measurements, and the standard deviation (called the **standard error**) is $\sigma_{\overline{Y}} = \sigma/\sqrt{n}$, where σ is the standard deviation of the population of individual measurements. That is, $\overline{Y} \sim N(\mu, \sigma/\sqrt{n})$.

The mean and standard error of the sampling distribution are μ and $\sigma/\sqrt{n}$, regardless of the sample size, only the shape being approximately normal depends on the sample size being large. Further, if the distribution of individual measurements is approximately normal (as in the height example), the sampling distribution of $\overline{Y}$ is approximately normal, regardless of the sample size.

Example 2.3 For the drug approval data (Example 1.4, Figure 1.4), the distribution of approval times is skewed to the right (long right tail for the distribution). For the 175 drugs in this 'population' of review times, the mean of the review times is $\mu = 32.06$ months, with standard deviation $\sigma = 20.97$ months.

10,000 random samples of size $n = 10$ and a second 10,000 random samples of size $n = 30$ were selected from this population. For each sample, we computed $\overline{y}$. Figure 2.5 and Figure 2.6 represent histograms of the 10,000 sample means of sizes $n = 10$ and $n = 30$, respectively. Note that the distribution for samples of size $n = 10$ is skewed to the right, while the distribution for samples of $n = 30$ is approximately normal. Table 2.2 gives the theoretical and empirical (based on the 10,000 samples) means and standard errors (standard deviations) of $\overline{Y}$ for sample means of these two sample sizes.

	Theoretical		Empirical	
n	$\mu_{\overline{Y}} = \mu$	$\sigma_{\overline{Y}} = \sigma/\sqrt{n}$	$\overline{\overline{y}} = \sum \overline{y}/10000$	$s_{\overline{y}}$
10	32.06	$20.97/\sqrt{10} = 6.63$	32.15	6.60
30	32.06	$20.97/\sqrt{30} = 3.83$	32.14	3.81

Table 2.2: Theoretical and empirical (based on 10,000 random samples) mean and standard error of $\overline{Y}$ based on samples of $n = 10$ and $n = 30$

2.3 Exercises

9. The renowned anthropologist Sir Francis Galton was very interested in measurements of many variables arising in nature (Galton, 1889). Among the measurements he obtained in the Anthropologic Laboratory in the International Exhibition of 1884 among adults are: height (standing

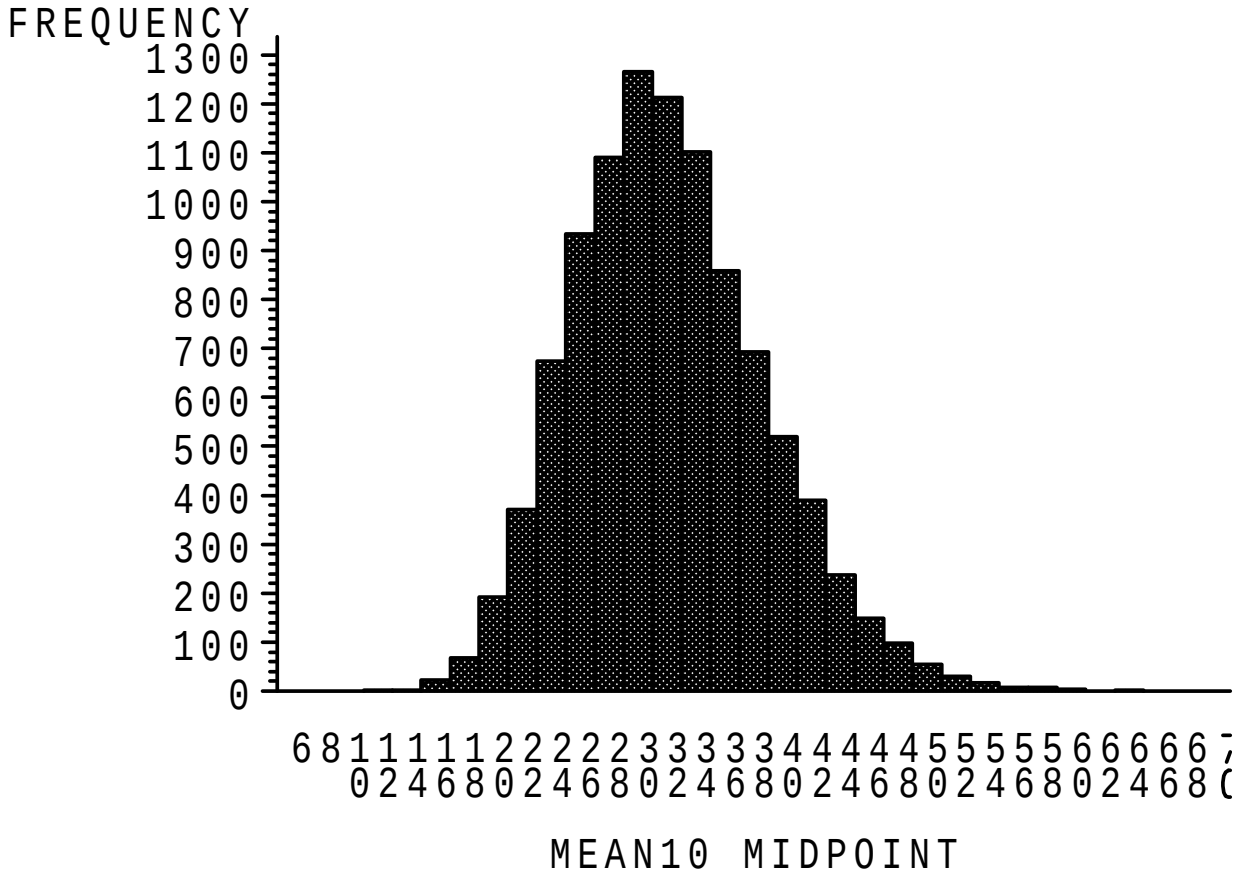

Figure 2.5: Histogram of sample means ($n = 10$) for drug approval time data

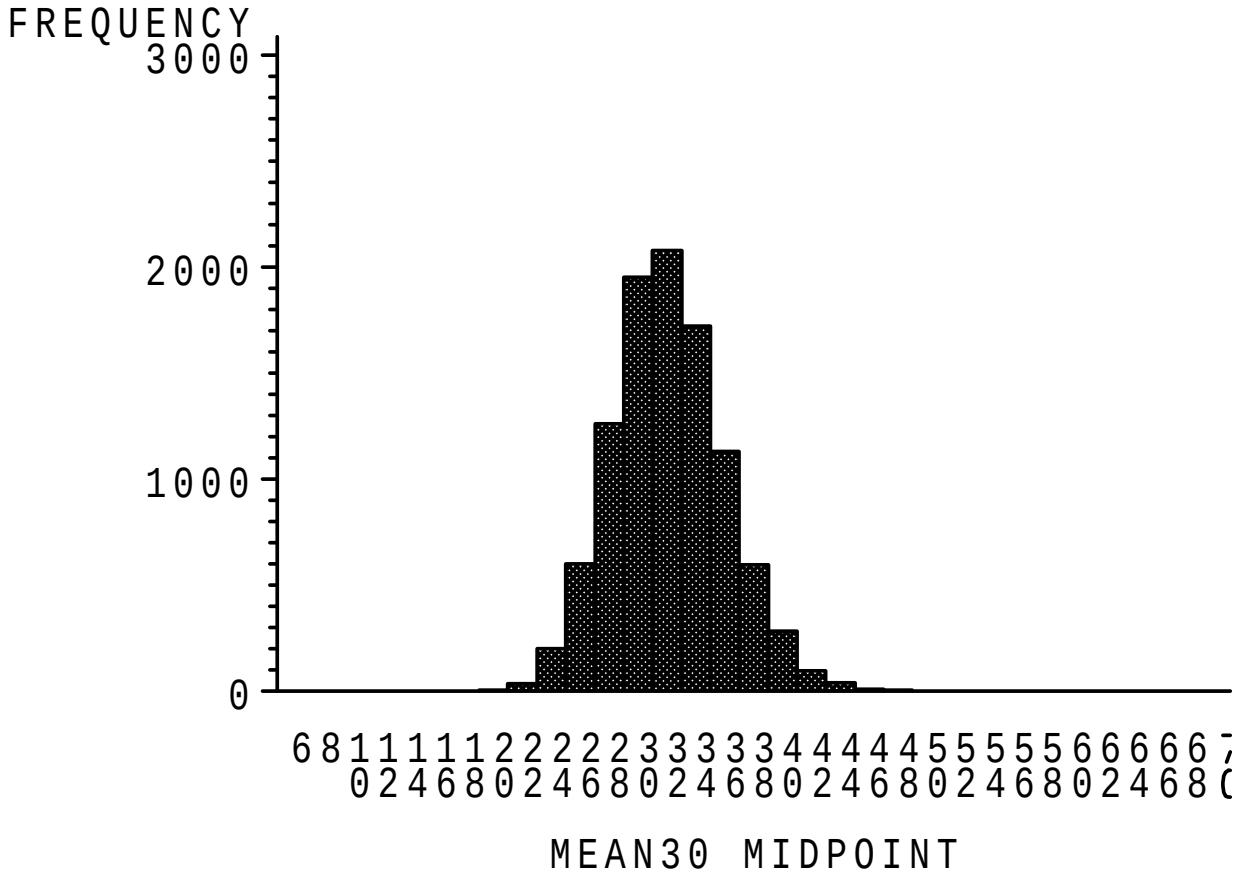

Figure 2.6: Histogram of sample means ($n = 30$) for drug approval time data

without shoes), height (sitting from seat of chair), arm span, weight (in ordinary indoor clothes), breathing capacity, and strength of pull (as archer with bow). He found that the relative frequency distributions of these variables were very well approximated by normal distributions with means and standard deviations given in Table 2.3. Although these means and standard deviations were based on samples (as opposed to populations), the samples are sufficiently large than we can (for our purposes) treat them as population parameters.

	Males		Females	
Variable	μ	σ	μ	σ
Standing height (inches)	67.9	2.8	63.3	2.6
Sitting height (inches)	36.0	1.4	33.9	1.3
Arm span (inches)	69.9	3.0	63.0	2.9
Weight (pounds)	143	15.5	123	14.3
Breathing capacity (in^3)	219	39.2	138	28.6
Pull strength (Pounds)	74	12.2	40	7.3

Table 2.3: Means and standard deviations of normal distributions approximating natural occuring distributions in adults

(a) What proportion of males stood over 6 feet (72 inches) in Galton's time?
(b) What proportion of females were under 5 feet (60 inches)?
(c) Sketch the approximate distributions of sitting heights among males and females on the same line segment.
(d) Above what weight do the heaviest 10% of males fall?
(e) Below what weight do the lightest 5% of females weigh?
(f) Between what bounds do the middle 95% of male breathing capacities lie?
(g) What fraction of women have pull strengths that exceed the pull strength that 99% of all men exceed?

10. In the previous exercise, give the approximate sampling distribution for the sample mean $\overline{y}$ for samples in each of the following cases:

(a) Standing heights of 25 randomly selected males.
(b) Sitting heights of 35 randomly selected females.
(c) Arm spans of 9 randomly selected males.
(d) Weights of 50 randomly selected females.

11. In the previous exercises, obtain the following probabilities:

(a) A sample of 25 males has a mean standing height exceeding 70 inches.
(b) A sample of 35 females has a mean sitting height below 32 inches.
(c) A sample of 9 males has a mean arm span between 69 and 71 inches.
(d) A sample of 50 females has a mean weight above 125 pounds.

Chapter 3

Statistical Inference – Hypothesis Testing

In this chapter, we will introduce the concept of statistical inference in the form of hypothesis testing. The goal is to make decisions concerning unknown population parameters, based on observed sample data. In pharmaceutical studies, the purpose is often to demonstrate that a new drug is effective, or possibly to show that it is more effective than an existing drug. For instance, in Phase II and Phase III clinical trials, the purpose is to demonstrate that the new drug is better than a placebo control. For numeric outcomes, this implies eliciting a higher (or lower in some cases) mean response that measures clinical effectiveness. For categorical outcomes, this implies having a higher (or lower in some cases) proportion having some particular outcome in the population receiving the new drug. In this chapter, we will focus only on numeric outcomes.

Example 3.1 A study was conducted to evaluate the efficacy and safety of fluoxetine in treating late–luteal–phase dysphoric disorder (a.k.a. premenstrual syndrome) (Steiner, et al., 1995). A primary outcome was the percent change from baseline for the mean score of three visual analogue scales after one cycle on the randomized treatment. There were three treatment groups: placebo, fluoxetine (20 mg/day) and fluoxetine (60 mg/day). If we define the population mean percent changes as μ_p, μ_{f20}, and μ_{f60}, respectively; we would want to show that μ_{f20} and μ_{f60} were larger than μ_p to demonstrate that fluoxetine is effective.

3.1 Introduction to Hypothesis Testing

Hypothesis testing involves choosing between two propositions concerning an unknown parameter's value. In the cases in this chapter, the propositions will be that the means are equal (no drug effect), or that the means differ (drug effect). We work under the assumption that there is no drug effect, and will only reject that claim if the sample data gives us sufficient evidence against it, in favor of claiming the drug is effective.

The testing procedure involves setting up two contradicting statements concerning the true value of the parameter, known as the **null hypothesis** and the **alternative hypothesis**, respectively.

We assume the null hypothesis is true, and usually (but not always) wish to show that the alternative is actually true. After collecting sample data, we compute a **test statistic** which is used as evidence for or against the null hypothesis (which we assume is true when calculating the test statistic). The set of values of the test statistic that we feel provide sufficient evidence to reject the null hypothesis in favor of the alternative is called the **rejection region**. The probability that we could have obtained as strong or stronger evidence against the null hypothesis (when it is true) than what we observed from our sample data is called the **observed significance level** or p–**value**.

An analogy that may help clear up these ideas is as follows. The researcher is like a prosecutor in a jury trial. The prosecutor must work under the assumption that the defendant is innocent (null hypothesis), although he/she would like to show that the defendant is guilty (alternative hypothesis). The evidence that the prosecutor brings to the court (test statistic) is weighed by the jury to see if it provides sufficient evidence to rule the defendant guilty (rejection region). The probability that an innocent defendant could have had more damning evidence brought to trial than was brought by the prosecutor (p-value) provides a measure of how strong the prosecutor's evidence is against the defendant.

Testing hypotheses is 'clearer' than the jury trial because the test statistic and rejection region are not subject to human judgement (directly) as the prosecutor's evidence and jury's perspective are. Since we do not know the true parameter value and never will, we are making a decision in light of uncertainty. We can break down reality and our decision into Table 3.1. We would like to

		Decision	
		H_0 True	H_0 False
Actual State	H_0 True	Correct Decision	Type I Error
	H_0 False	Type II Error	Correct Decision

Table 3.1: Possible outcomes of a hypothesis test

set up the rejection region to keep the probability of a Type I error (α) and the probability of a Type II error (β) as small as possible. Unfortunately for a fixed sample size, if we try to decrease α, we automatically increase β, and vice versa. We will set up rejection regions to control for α, and will not concern ourselves with β. Here α is the probability we reject the null hypothesis when it is true. (This is like sending an innocent defendant to prison). Later, we will see that by choosing a particular sample size in advance to gathering the data, we can control both α and β.

We can write out the general form of a large–sample hypothesis test in the following steps, where θ is a population parameter that has an estimator ($\hat{\theta}$) that is approximately normal.

1. $H_0 : \theta = \theta_0$
2. $H_A : \theta \neq \theta_0$ or $H_A : \theta > \theta_0$ or $H_A : \theta < \theta_0$ (which alternative is appropriate should be clear from the setting).
3. T.S.: $z_{obs} = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}}$
4. R.R.: $|z_{obs}| > z_{\alpha/2}$ or $z_{obs} > z_\alpha$ or $z_{obs} < -z_\alpha$ (which R.R. depends on which alternative hypothesis you are using).

5. p-value: $2P(Z > |z_{obs}|)$ or $P(Z > z_{obs})$ or $P(Z < z_{obs})$ (again, depending on which alternative you are using).

In all cases, a p-value less than α corresponds to a test statistic being in the rejection region (reject H_0), and a p-value larger than α corresponds to a test statistic failing to be in the rejection region (fail to reject H_0). We will illustrate this idea in an example below.

3.1.1 Large–Sample Tests Concerning $\mu_1 - \mu_2$ (Parallel Groups)

To test hypotheses of this form, we have two independent random samples, with the statistics and information given in Table 3.2. The general form of the test is given in Table 3.3.

	Sample 1	Sample 2
Mean	$\overline{y}_1$	$\overline{y}_2$
Std Dev	s_1	s_2
Sample Size	n_1	n_2

Table 3.2: Sample statistics needed for a large–sample test of $\mu_1 - \mu_2$

$H_0 : \mu_1 - \mu_2 = 0$	Test Statistic: $z_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	
Alternative Hypothesis	Rejection Region	$P-$ Value
$H_A : \mu_1 - \mu_2 > 0$	RR: $z_{obs} > z_\alpha$	$p - val = P(Z \geq z_{obs})$
$H_A : \mu_1 - \mu_2 < 0$	RR: $z_{obs} < -z_\alpha$	$p - val = P(Z \leq z_{obs})$
$H_A : \mu_1 - \mu_2 \neq 0$	RR: $\lvert z_{obs}\rvert > z_{\alpha/2}$	$p - val = 2P(Z \geq \lvert z_{obs}\rvert)$

Table 3.3: Large–sample test of $\mu_1 - \mu_2 = 0$ vs various alternatives

Example 3.3 A randomized clinical trial was conducted to determine the safety and efficacy of sertraline as a treatment for premature ejaculation (Mendels, et al.,1995). Heterosexual male patients suffering from premature ejaculation were defined as having suffered from involuntary ejaculation during foreplay or within 1 minute of penetration in at least 50% of intercourse attempts during the previous 6 months. Patients were excluded if they met certain criteria such as depression, receiving therapy, or being on other psychotropic drugs.

Patients were assigned at random to receive either sertraline or placebo for 8 weeks after a one week placebo washout. Various subjective sexual function measures were obtained at baseline and again at week 8. The investigator's Clinical Global Impressions (CGI) were also obtained, including the therapeutic index, which is scored based on criteria from the *ECDEU Assessment Manual for Psychopharmacology* (lower scores are related to higher improvement). Summary statistics based on the CGI therapeutic index scores are given in Table 3.4. We will conduct test whether or not the mean therapeutic indices differ between the sertraline and placebo groups at the $\alpha = 0.05$ significance level, meaning that if the null hypothesis is true (drug not effective), there is only a

5% chance that we will claim that it is effective. We will conduct a 2–sided test, since there is a

	Sertraline	Placebo
Mean	$\overline{y}_1 = 5.96$	$\overline{y}_2 = 10.75$
Std Dev	$s_1 = 4.59$	$s_2 = 3.70$
Sample Size	$n_1 = 24$	$n_2 = 24$

Table 3.4: Sample statistics for sertraline study in premature ejaculation patients

risk the drug could worsen the situation. If we do conclude the means differ, we will determine if the drug is better of worse than the placebo, based on the sign of the test statistic.

$$H_0 : \mu_1 - \mu_2 = 0 \quad (\mu_1 = \mu_2) \qquad vs \qquad H_A : \mu_1 - \mu_2 \neq 0 \quad (\mu_1 \neq \mu_2)$$

We then compute the test statistic, and obtain the appropriate rejection region:

$$\text{T.S.: } z_{obs} = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5.96 - 10.75}{\sqrt{\frac{(4.59)^2}{24} + \frac{(3.70)^2}{24}}} = \frac{-4.79}{1.203} = -3.98 \qquad \text{R.R.: } |z_{obs}| \geq z_{.025} = 1.96$$

Since the test statistic falls in the rejection (and is negative), we reject the null hypothesis and conclude that the mean is lower for the sertraline group than the placebo group, implying the drug is effective. Note that the $p-$value is less than .005 and is actually .0002:

$$p - \text{value} = 2P(Z \geq 3.98) < 2P(Z \geq 2.576) = 0.005$$

3.2 Elements of Hypothesis Tests

In the last section, we conducted a test of hypothesis to determine whether or not two population means differed. In this section, we will cover the concepts of size and power of a statistical test. We will consider this under the same context: a large–sample test to compare two population means, based on a parallel groups design.

3.2.1 Significance Level of a Test (Size of a Test)

We saw in the previous chapter that for large samples, the sample mean, $\overline{Y}$ has a sampling distribution that is approximately normal, with mean μ and standard error $\sigma_{\overline{Y}} = \sigma/\sqrt{n}$. This can be extended to the case where we have two sample means from two populations (with independent samples). In this case, when n_1 and n_2 are large, we have:

$$\overline{Y}_1 - \overline{Y}_2 \sim N(\mu_1 - \mu_2, \sigma_{\overline{Y}_1 - \overline{Y}_2}) \qquad \sigma_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

So, when $\mu_1 = \mu_2$ (the drug is ineffective), $Z = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ is standard normal (see Section 2.1). Thus, if we are testing $H_0 : \mu_1 - \mu_2 = 0$ vs $H_A : \mu_1 - \mu_2 > 0$, we would have the following probability

statements:

$$P\left(\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_a\right) = P(Z \geq z_a) = a$$

This gives a natural rejection region for our test statistic to control α, the probability we reject the null hypothesis when it is true (claim an ineffective drug is effective). Similarly, the p–value is the probability that we get a larger values of z_{obs} (and thus $\overline{Y}_1 - \overline{Y}_2$) than we observed. This is the area to the right of our test statistic under the standard normal distribution. Table 3.5 gives the cut–off values for various values of α for each of the three alternative hypotheses. The value α for a test is called its **significance level** or **size**.

	Alternative Hypothesis (H_A)		
α	$H_A : \mu_1 - \mu_2 > 0$	$H_A : \mu_1 - \mu_2 < 0$	$H_A : \mu_1 - \mu_2 \neq 0$
.10	$z_{obs} \geq 1.282$	$z_{obs} \leq -1.282$	$\lvert z_{obs}\rvert \geq 1.645$
.05	$z_{obs} \geq 1.645$	$z_{obs} \leq -1.645$	$\lvert z_{obs}\rvert \geq 1.96$
.01	$z_{obs} \geq 2.326$	$z_{obs} \leq -2.326$	$\lvert z_{obs}\rvert \geq 2.576$

Table 3.5: Rejection Regions for tests of size α

Typically, we don't know σ_1 and σ_2, and replace them with their estimates s_1 and s_2, as we did in Example 3.2.

3.2.2 Power of a Test

The power of a test corresponds to the probability of rejecting the null hypothesis when it is false. That is, in terms of a test of efficacy for a drug, the probability we correctly conclude the drug is effective when in fact it is. There are many parameter values that correspond to the alternative hypothesis, and the power depends on the actual value of $\mu_1 - \mu_2$ (or whatever parameter is being tested). Consider the following situation.

A researcher is interested in testing $H_0 : \mu_1 - \mu_2 = 0$ vs $H_A : \mu_1 - \mu_2 > 0$ at $\alpha = 0.05$ significance level. Suppose, that the variances of each population are known, and $\sigma_1^2 = \sigma_2^2 = 25.0$. The researcher takes samples of size $n_1 = n_2 = 25$ from each population. The rejection region is set up under H_0 as (see Table 3.5):

$$z_{obs} = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq 1.645 \qquad \Longrightarrow \qquad \overline{y}_1 - \overline{y}_2 \geq 1.645\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1.645\sqrt{2.0} = 2.326$$

That is, we will reject H_0 if $\overline{y}_1 - \overline{y}_2 \geq 2.326$. Under the null hypothesis ($\mu_1 = \mu_2$), the probability that $\overline{y}_1 - \overline{y}_2$ is larger than 2.326 is 0.05 (very unlikely).

Now the power of the test represents the probability we correctly reject the null hypothesis when it is false (the alternative is true, $\mu_1 > \mu_2$). We are interested in finding the probability that $\overline{Y}_1 - \overline{Y}_2 \geq 2.326$ when H_A is true. This probability depends on the actual value of $\mu_1 - \mu_2$, since

$\overline{Y}_1 - \overline{Y}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$. Suppose $\mu_1 - \mu_2 = 3.0$, then we have the following result, used to obtain the power of the test:

$$P(\overline{Y}_1 - \overline{Y}_2 \geq 2.326) = P\left(\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq \frac{2.326 - 3.0}{\sqrt{2.0}}\right) = P(Z \geq -0.48) = .6844$$

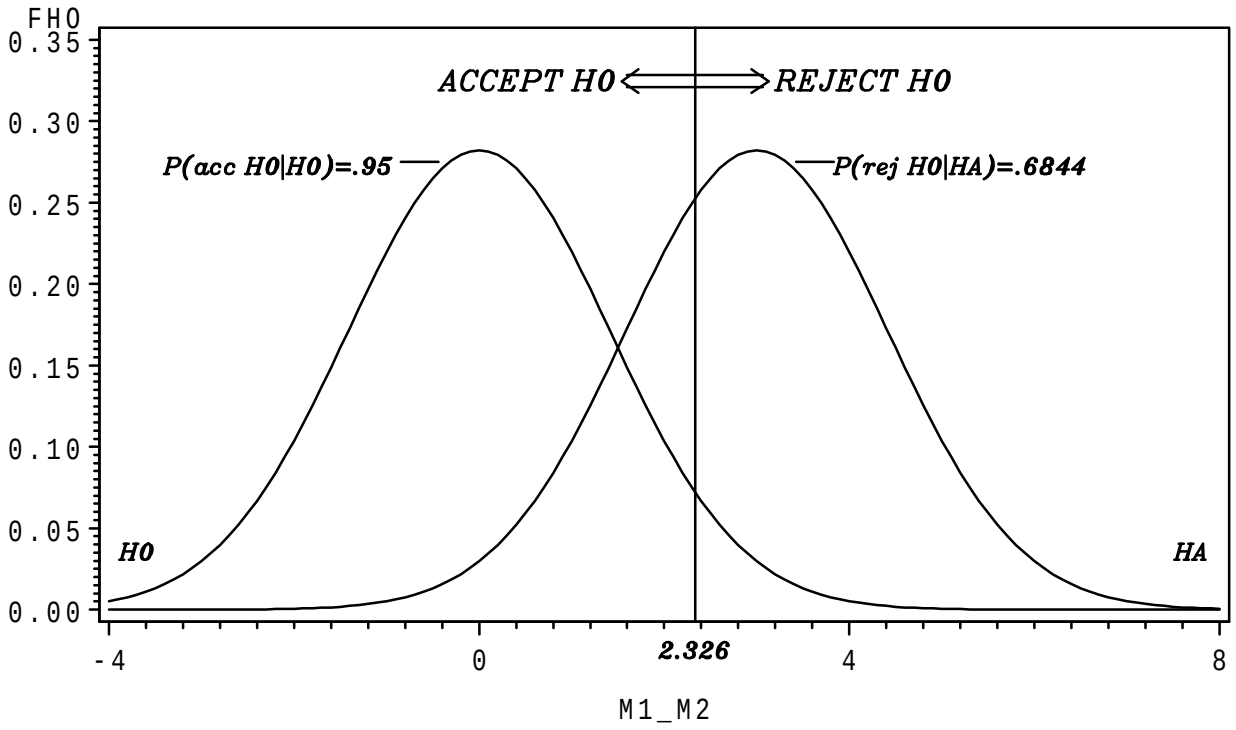

Figure 3.1: Significance level and power of test under H_0 ($\mu_1 - \mu_2 = 0$) and H_A ($\mu_1 - \mu_2 = 3$)

See Figure 3.1 for a graphical depiction of this. The important things to know about power (while holding all other things fixed) are:

- As the sample sizes increase, the power of the test increases. That is, by taking larger samples, we improve our ability to find a difference in means when they really exist.
- As the population variances decrease, the power of the test increases. Note, however, that the researcher has no control over σ_1 or σ_2.
- As the difference in the means, $\mu_1 - \mu_2$ increases, the power increases. Again, the researcher has no control over these parameters, but this has the nice property that the further the true state is from H_0, the higher the chance we can detect this.

Figure 3.2 gives "power curves" for four sample sizes ($n_1 = n_2 = 25, 50, 75, 100$) as a function of $\mu_1 - \mu_2$ (0–5). The vertical axis gives the power (probability we reject H_0) for the test.

In many instances, too small of samples are taken, and the test has insufficient power to detect an important difference. The next section gives a method to compute a sample size in advance that provides a test with adequate power.

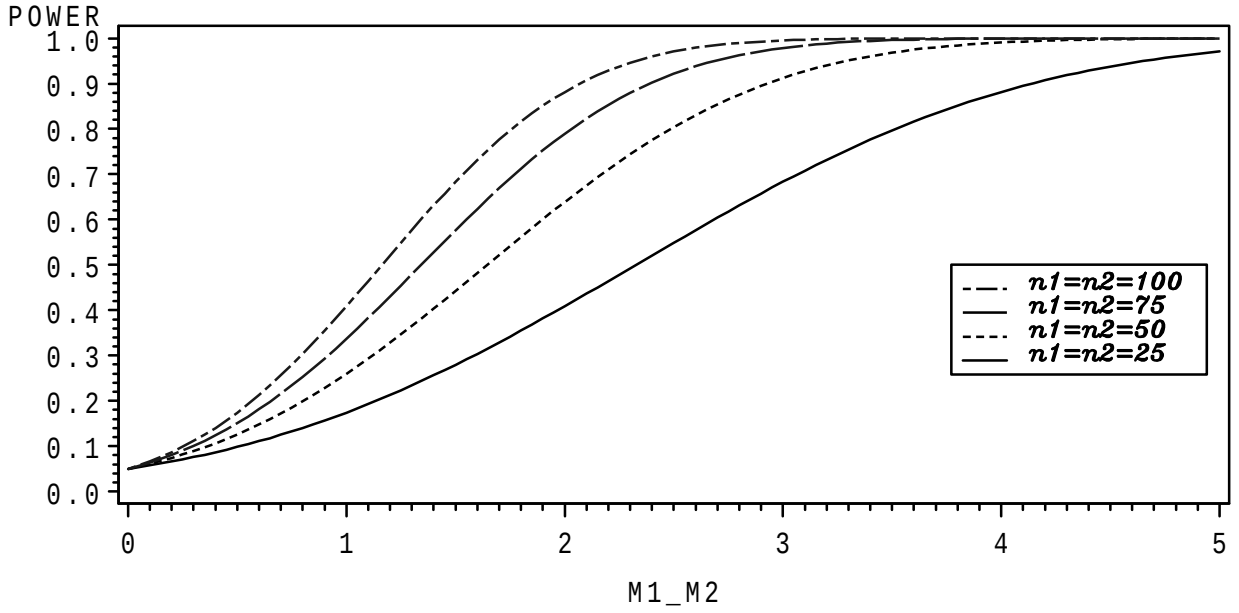

Figure 3.2: Power of test (Probability reject H_0) as a function of $\mu_1 - \mu_2$ for varying sample sizes

3.3 Sample Size Calculations to Obtain a Test With Fixed Power

In the last section, we saw that the one element of a statistical test that is related to power that a researcher can control is the sample size from each of the two populations being compared. In many applications, the process is developed as follows (we will assume that we have a 2–sided altenative ($H_A : \mu_1 - \mu_2 \neq 0$) and the two population standard deviations are equal, and will denote the common value as σ):

1. Define a clinically meaningful difference $\delta = \frac{\mu_1 - \mu_2}{\sigma}$. This is a difference in population means in numbers of standard deviations since σ is usually unknown. If σ is known, or well approximated, based on prior experience, a clinically meaningful difference can be stated in terms of the units of the actual data.

2. Choose the power, the probability you will reject H_0 when the population means differ by the clinically meaningful difference. In many instances, the power is chosen to be .80. Obtain $z_{\alpha/2}$ and z_β, where α is the significance level of the test, and $\beta = (1-\text{power})$ is the probability of a type II error (accepting H_0 when in fact H_A is true). Note that $z_{.20} = .84162$.

3. Choose as sample sizes: $n_1 = n_2 = \frac{2(z_{\alpha/2}+z_\beta)^2}{\delta^2}$

Choosing a sample size this way allows researchers to be confident that if an important difference really exists, they will have a good chance of detecting it when they conduct their hypothesis test.

Example 3.4 A study was conducted in patients with renal insufficiency to measure the pharmacokinetics of oral dosage of Levocabastine (Zazgornik, et al.,1993). Patients were classified as non–dialysis and hemodialysis patients. In their study, one of the pharmacokinetic parameters of interest was the terminal elimination half–life ($t_{1/2}$). Based on pooling the estimates of σ for these two groups, we get an estimate of $\hat{\sigma} = 34.4$. That is, we'll assume that in the populations of half–lifes for each dialysis group, the standard deviation is 34.4.

What sample sizes would be needed to conduct a test that would have a power of .80 to detect a difference in mean half–lifes of 15.0 hours? The test will be conducted at the $\alpha = 0.05$ significance level.

1. $\delta = \frac{\mu_1-\mu_2}{\sigma} = \frac{15.0}{34.4} = .4360.$
2. $z_{\alpha/2} = z_{.025} = 1.96$ and $z_{1-.80} = z_{.20} = .84162.$
3. $n_1 = n_2 = \frac{2(z_{\alpha/2}+z_\beta)^2}{\delta^2} = \frac{2(1.96+.84162)^2}{(.4360)^2} = 82.58 \approx 83$

That is, we would need to have sample sizes of 83 from each dialysis group to have a test where the probability we conclude that the two groups differ is .80, when they actually differ by 15 hours. These sound like large samples (and are). The reason is that the standard deviation of each group is fairly large (34.4 hours). Often, experimenters will have to increase δ, the clinically meaningful difference, or decrease the power to obtain physically or economically manageable sample sizes.

3.4 Small–Sample Tests

In this section we cover small–sample tests without going through the detail given for the large–sample tests. In each case, we will be testing whether or not the means (or medians) of two distributions are equal. There are two considerations when choosing the appropriate test: (1) Are the population distributions of measurements approximately normal? and (2) Was the study conducted as a parallel groups or crossover design? The appropriate test for each situation is given in Table 3.6. We will describe each test with the general procedure and an example.

The two tests based on non–normal data are called **nonparametric tests** and are based on ranks, as opposed to the actual measurements. When distributions are skewed, samples can contain measurements that are extreme (usually large). These extreme measurements can cause problems for methods based on means and standard deviations, but will have less effect on procedures based on ranks.

3.4.1 Parallel Groups Designs

Parallel groups designs are designs where the samples from the two populations are independent. That is, subjects are either assigned at random to one of two treatment groups (possibly active

	Design Type	
	Parallel Groups	Crossover
Normally Distributed Data	2–Sample t–test	Paired t–test
Non–Normally Distributed Data	Wilcoxon Rank Sum test (Mann–Whitney U–Test)	Wilcoxon Signed–Rank Test

Table 3.6: Statistical Tests for small–sample 2 group situations

drug or placebo), or possibly selected at random from one of two populations (as in Example 3.4, where we had non–dialysis and hemodialysis patients). In the case where the two populations of measurements are normally distributed, the 2–sample t–test is used. This procedure is very similar to the large–sample test from the previous section, where only the cut–off value for the rejection region changes. In the case where the populations of measurements are not approximately normal, the Mann–Whitney U–test (or, equivalently the Wilcoxon Rank–Sum test) is commonly used. These tests are based on comparing the average ranks across the two groups when the measurements are ranked from smallest to largest, across groups.

2–Sample Student's t–test for Normally Distributed Data

This procedure is identical to the large–sample test, except the critical values for the rejection regions are based on the t–distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. We will assume the two population variances are equal in the 2–sample t–test. If they are not, simple adjustments can be made to obtain the appropriate test. We then 'pool' the 2 sample variances to get an estimate of the common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$. This estimate, that we will call s_p^2 is calculated as follows:

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

The test of hypothesis concerning $\mu_1 - \mu_2$ is conducted as follows:

1. $H_0 : \mu_1 - \mu_2 = 0$
2. $H_A : \mu_1 - \mu_2 \neq 0$ or $H_A : \mu_1 - \mu_2 > 0$ or $H_A : \mu_1 - \mu_2 < 0$ (which alternative is appropriate should be clear from the setting).
3. T.S.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$
4. R.R.: $|t_{obs}| > t_{\alpha/2, n_1+n_2-2}$ or $t_{obs} > t_{\alpha, n_1+n_2-2}$ or $t_{obs} < -t_{\alpha, n_1+n_2-2}$ (which R.R. depends on which alternative hypothesis you are using).
5. p-value: $2P(T > |t_{obs}|)$ or $P(T > t_{obs})$ or $P(T < t_{obs})$ (again, depending on which alternative you are using).

The values $t_{\alpha/2, n_1+n_2-2}$ are given in Table A.2.

Example 3.5 In the pharmaockinetic study in renal patients described in Example 3.4, the authors measured the bioavailability of the drug in each patient by computiong AUC (Area under the concentration–time curve, in $(ng \cdot hr/mL)$). Table 3.7 has the raw data, as well as the mean and standard deviation for each group. We will test whether or not mean AUC is equal in the two populations, assuming that the populations of AUC are approximately normal. We have no prior belief of which group (if any) would have the larger mean, so we will test $H_0 : \mu_1 = \mu_2$ vs $H_A : \mu_1 \neq \mu_2$.

Non–Dialysis	Hemodialysis
857	527
567	740
626	392
532	514
444	433
357	392
$\overline{y}_1 = 563.8$	$\overline{y}_2 = 499.7$
$s_1 = 172.0$	$s_2 = 131.4$
$n_1 = 6$	$n_2 = 6$

Table 3.7: AUC measurements for levocabastine in renal insufficiency patients

First, we compute s_p^2, the pooled variance:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(6-1)(172.0)^2 + (6-1)(131.4)^2}{6+6-2} = \frac{234249.8}{10} = 23424.98 \quad (s_p = 153.1)$$

Now we conduct the (2–sided) test as described above with $\alpha = 0.05$ significance level:

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_A : \mu_1 - \mu_2 \neq 0$
- T.S.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{s_p^2(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{(563.8-499.7)}{\sqrt{23424.98(\frac{1}{6}+\frac{1}{6})}} = \frac{64.1}{88.4} = 0.73$
- R.R.: $|t_{obs}| > t_{\alpha/2,n_1+n_2-2} = t_{.05/2,6+6-2} = t_{.025,10} = 2.228$
- p-value: $2P(T > |t_{obs}|) = 2P(T > 0.73) > 2P(T > 1.372) = 2(.10) = .20$ (From t–table, $t_{.10,10} = 1.372$)

Based on this test, we do not reject H_0, and cannot conclude that the mean AUC for this dose of levocabastine differs in these two populations of patients with renal insufficiency.

Wilcoxon Rank Sum Test for Non-Normally Distributed Data

The idea behind this test is as follows. We have samples of n_1 measurements from population 1 and n_2 measurements from population 2 (Wilcoxon,1945). We rank the $n_1 + n_2$ measurements from 1

(smallest) to $n_1 + n_2$ (largest), adjusting for ties by averaging the ranks the measurements would have received if they were different. We then compute T_1, the rank sum for measurements from population 1, and T_2, the rank sum for measurements from population 2. This test is mathematically equivalent to the Mann–Whitney U–test. To test for differences between the two population distributions, we use the following procedure:

1. H_0 : The two population distributions are identical ($\mu_1 = \mu_2$)
2. H_A : One distribution is shifted to the right of the other ($\mu_1 \neq \mu_2$)
3. T.S.: $T = \min(T_1, T_2)$
4. R.R.: $T \leq T_0$, where values of T_0 given in tables in many statistics texts for various levels of α and sample sizes.

For one-sided tests to show that the distribution of population 1 is shifted to the right of population 2 ($\mu_1 > \mu_2$), we use the following procedure (simply label the distribution with the suspected higher mean as population 1):

1. H_0 : The two population distributions are identical ($\mu_1 = \mu_2$)
2. H_A : Distribution 1 is shifted to the right of distribution 2 ($\mu_1 > \mu_2$)
3. T.S.: $T = T_2$
4. R.R.: $T \leq T_0$, where values of T_0 are given in tables in many statistics texts for various levels of α and various sample sizes.

Example 3.6 For the data in Example 3.5, we will use the Wilcoxon Rank Sum test to determine whether or not mean AUC differs in these two populations. Table 3.8 contains the raw data, as well as the ranks of the subjects, and the rank sums T_i for each group.

Non–Dialysis	Hemodialysis
857 (12)	527 (7)
567 (9)	740 (11)
626 (10)	392 (2.5)
532 (8)	514 (6)
444 (5)	433 (4)
357 (1)	392 (2.5)
$n_1 = 6$	$n_2 = 6$
$T_1 = 45$	$T_2 = 33$

Table 3.8: AUC measurements (and ranks) for levocabastine in renal insufficiency patients

For a 2–tailed test, based on sample sizes of $n_1 = n_2 = 6$, we will reject H_0 for $T = min(T_1, T_2) \leq 26$. Since $T = min(45, 33) = 33$, we fail to reject H_0, and cannot conclude that the mean AUC differs among non–dialysis and hemodialysis patients.

3.4.2 Crossover Designs

In crossover designs, subjects receive each treatment, thus acting as their own control. Procedures based on these designs take this into account, and are based in determining differences between treatments after "removing" variability in the subjects. When it is possible to conduct them, crossover designs are more powerful than parallel groups designs in terms of being able to detect a difference (reject H_0) when differences truly exist (H_A is true), for a fixed sample size. In particular, many pharmacokinetic, and virtually all bioequivalence, studies are crossover designs.

Paired t–test for Normally Distributed Data

In crossover designs, each subject receives each treatment. In the case of two treatments being compared, we compute the difference in the two measurements within each subject, and test whether or not the population mean difference is 0. When the differences are normally distributed, we use the paired t–test to determine if differences exist in the mean response for the two treatments.

It should be noted that in the paired case $n_1 = n_2$ by definition. That is, we will always have equal sized samples when the experiment is conducted properly. We will always be looking at the $n = n_1 = n_2$ differences, and will have n differences, even though there were $2n = n_1 + n_2$ measurements made. From the n differences, we will compute the mean and standard deviation, which we will label as $\overline{d}$ and s_d:

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

The procedure is conducted as follows:

1. $H_0 : \mu_1 - \mu_2 = \mu_D = 0$
2. $H_A : \mu_D \neq 0$ or $H_A : \mu_D > 0$ or $H_A : \mu_D < 0$ (which alternative is appropriate should be clear from the setting).
3. T.S.: $t_{obs} = \overline{d}/(\frac{s_d}{\sqrt{n}})$
4. R.R.: $|t_{obs}| > t_{\alpha/2,n-1}$ or $t_{obs} > t_{\alpha,n-1}$ or $t_{obs} < -t_{\alpha,n-1}$ (which R.R. depends on which alternative hypothesis you are using).
5. p-value: $2P(T > |t_{obs}|)$ or $P(T > t_{obs})$ or $P(T < t_{obs})$ (again, depending on which alternative you are using).

As with the 2–sample t–test, the values corresponding to the rejection region are given in the Table A.2.

Example 3.7 A study was conducted to compare immediate– and sustained–release formulations of codeine (Band, et al.,1994). Thirteen healthy patients received each formulation (in random order, and blinded). Among the pharmacokinetic parameters measured was maximum concentration at single–dose (C_{max}). We will test whether or not the population mean C_{max} is higher for

	C_{max} (ng/mL)		
Subject (i)	SRC_i	IRC_i	$d_i = SRC_i - IRC_i$
1	195.7	181.8	13.9
2	167.0	166.9	0.1
3	217.3	136.0	81.3
4	375.7	221.3	153.4
5	285.7	195.1	90.6
6	177.2	112.7	64.5
7	220.3	84.2	136.1
8	243.5	78.5	165.0
9	141.6	85.9	55.7
10	127.2	85.3	41.9
11	345.2	217.2	128.0
12	112.1	49.7	62.4
13	223.4	190.0	33.4
Mean	$\overline{SRC} = 217.8$	$\overline{IRC} = 138.8$	$\overline{d} = 78.9$
Std. Dev.	$s_{SRC} = 79.8$	$s_{IRC} = 59.4$	$s_d = 53.0$

Table 3.9: C_{max} measurements for sustained– and immediate–release codeine

the sustained–release (SRC) than for the immediate–release (IRC) formulation. The data, and the differences ($SRC - IRC$) are given in Table 3.9.

We will conduct the test (with $\alpha = 0.05$) by completing the steps outlined above.

1. $H_0 : \mu_1 - \mu_2 = \mu_D = 0$
2. $H_A : \mu_D > 0$
3. T.S.: $t_{obs} = \overline{d}/(\frac{s_d}{\sqrt{n}}) = 78.9/(\frac{53.0}{\sqrt{13}}) = \frac{78.9}{14.7} = 5.37$
4. R.R.: $t_{obs} > t_{\alpha,n-1} = t_{.05,12} = 1.782$
5. p-value: $P(T > t_{obs}) = P(T > 5.37) < P(T > 4.318) = .0005$ (since $t_{.0005,12} = 4.318$).

We reject H_0, and conclude that mean maximum concentration at single–dose is higher for the sustained–release than for the immediate–release codeine formulation.

Wilcoxon Signed–Rank Test for Paired Data

A nonparametric test that is often conducted in crossover designs is the signed–rank test (Wilcoxon,1945). Like the paired t–test, the signed–rank test takes into account that the two treatments are being assigned to the same subject. The test is based on the difference in the measurements within each subject. Any subjects with differences of 0 (measurements are equal under both treatments) are removed and the sample size is reduced. The test statistic is computed as follows:

1. For each pair, subtract measurement 2 from measurement 1.
2. Take the absolute value of each of the differences, and rank from 1 (smallest) to n (largest), adjusting for ties by averaging the ranks they would have had if not tied.
3. Compute T^+, the rank sum for the positive differences from 1), and T^-, the rank sum for the negative differences.

To test whether or not the population distributions are identical, we use the following procedure:

1. H_0 : The two population distributions are identical ($\mu_1 = \mu_2$)
2. H_A : One distribution is shifted to the right of the other ($\mu_1 \neq \mu_2$)
3. T.S.: $T = \min(T^+, T^-)$
4. R.R.: $T \leq T_0$, where T_0 is a function of n and α and given in tables in many statistics texts.

For a one-sided test, if you wish to show that the distribution of population 1 is shifted to the right of population 2 ($\mu_1 > \mu_2$), the procedure is as follows:

1. H_0 : The two population distributions are identical ($\mu_1 = \mu_2$)
2. H_A : Distribution 1 is shifted to the right of distribution 2 ($\mu_1 > \mu_2$)
3. T.S.: T^-
4. R.R.: $T \leq T_0$, where T_0 is a function of n and α and given in tables in many statistics texts.

Note that if you wish to use the alternative $\mu_1 < \mu_2$, use the above procedure with T^+ replacing T^-. The idea behind this test is to determine whether the differences tend to be positive ($\mu_1 > \mu_2$) or negative ($\mu_1 < \mu_2$), where differences are 'weighted' by their magnitude.

Example 3.8 In the study comparing immediate– and sustained–release formulations of codeine (Band, et al.,1994), another pharmacokinetic parameter measured was the half–life at steady–state ($t_{1/2}^{SS}$). We would like to determine whether or not the distributions of half–lives are the same ($\mu_1 = \mu_2$) for immediate– and sustained–release codeine. We will conduct the signed–rank test (2–sided), with $\alpha = 0.05$. The data and ranks are given in Table 3.10. Note that subject 2 will be eliminated since both measurements were 2.1 hours for her, and the sample size will be reduced to 12.

Based on Table 3.10, we get T^+ (the sum of the ranks for positive differences) and T^- (the sum of the ranks of the negative differences), as well as the test statistic T, as follows:

$T^+ = 1{+}10{+}5{+}2{+}12{+}11{+}8{+}9{+}4 = 62 \qquad T^- = 6.5{+}3{+}6.5 = 16 \qquad T = min(T^+, T^-) = min(62, 16) = 16$

We can then use the previously given steps to test for differences in the locations of the distributions of half–lives for the two formulations.

1. H_0 : The two population distributions are identical ($\mu_1 = \mu_2$)

$t_{1/2}^{SS}$ (hrs)					
Subject (i)	SRC_i	IRC_i	$d_i = SRC_i - IRC_i$	$\lvert d_i \rvert$	rank$(\lvert d_i \rvert)$
1	2.6	2.5	0.1	0.1	1
2	2.1	2.1	0.0	0.0	–
3	3.8	2.7	1.1	1.1	10
4	3.1	3.7	−0.6	0.6	6.5
5	2.8	2.3	0.5	0.5	5
6	3.6	3.4	0.2	0.2	2
7	4.2	2.3	1.9	1.9	12
8	3.2	1.4	1.8	1.8	11
9	2.5	1.7	0.8	0.8	8
10	2.1	1.1	1.0	1.0	9
11	2.0	2.3	−0.3	0.3	3
12	1.9	2.5	−0.6	0.6	6.5
13	2.7	2.3	0.4	0.4	4

Table 3.10: $t_{1/2}^{SS}$ measurements for sustained– and immediate–release codeine

2. H_A : One distribution is shifted to the right of the other $(\mu_1 \neq \mu_2)$

3. T.S.: $T = \min(T^+, T^-) = 16$

4. R.R.: $T \leq T_0$, where $T_0 = 14$ is based on 2–sided alternative , $\alpha = 0.05$, and $n = 12$.

Since $T = 16$ does not fall in the rejection region, we cannot reject H_0, and we fail to conclude that the means differ. Note that the p–value is thus larger than 0.05, since we fail to reject H_0 (the authors report it as 0.071).

3.5 Exercises

12. A review paper concerning smoking and drug metabolism related information obtained in a wide variety of clinical investigations on the topic (Dawson and Vestal,1982). Among the drugs studied was antipyrine, and the authors report results of metabolic clearance rates measured among smokers and nonsmokers of various ages. Based on values of metabolic clearance rate $(mlhr^{-1}kg^{-1})$ for the 18–39 age group and combining moderate and heavy smokers, we get the summary statistics in Table 3.11. Test whether or not smoking status is associated with metabolic clearance rate, that is, test whether or not mean metabolic clearance rates differ between the two groups. If a difference exists, what can we say about the effect of smoking on antipyrine metabolism? Test at $\alpha = 0.05$ significance level.

13. The efficacy of fluoxetine on anger in patients with borderline personality disorder was studied in 22 patients with BPD (Salzman, et al.,1995). Among the measures used was the Profile of Mood States (POMS) anger scale. In the blinded, controlled study, patients received either fluoxetine

	Nonsmokers	Smokers
Mean	$\overline{y}_1 = 30.6$	$\overline{y}_2 = 38.6$
Std Dev	$s_1 = 7.54$	$s_2 = 12.43$
Sample Size	$n_1 = 37$	$n_2 = 36$

Table 3.11: Sample statistics for antipyrine metabolism study in smokers and nonsmokers

or placebo for 12 weeks, with measurements being made before and after treatment. Table 3.12 contains the post–treatment summary statistics for the two drug groups. Use the independent sample t–test to test whether or not fluoxetine reduces anger levels (as measured by the POMS scale). Test at $\alpha = 0.05$ significance level.

	Fluoxetine	Placebo
Mean	$\overline{y}_1 = 40.31$	$\overline{y}_2 = 44.89$
Std Dev	$s_1 = 5.07$	$s_2 = 8.67$
Sample Size	$n_1 = 13$	$n_2 = 9$

Table 3.12: Sample statistics for fluoxetine study in BPD patients

14. Cyclosporine pharmacokinetics after intravenous and oral administration have been studied under similar experimental conditions in healthy subjects (Gupta, et al.,1990) and in pre–kidney transplant patients (Aweeka, et al.,1994). Among the pharmacokinetic parameters estimated within each patient is mean absorption time (MAT) for the oral dose in plasma (when taken without food). Values of MAT for the two groups (healthy and pre–transplant) are given in Table 3.13. Complete the rankings and conduct the Wilcoxon Rank Sum test to determine whether or not mean MAT differs in the two populations. Note that for $\alpha = 0.05$, we reject $H_0 : \mu_1 = \mu_2$ in favor of $H_A : \mu_1 \neq \mu_2$ when $T = min(T_1, T_2) \leq 49$ for these sample sizes and $\alpha = 0.05$.

Healthy	Pre–Transplant
5.36 (14)	2.64
4.35	4.84 (13)
2.61 (4)	2.42 (2.5)
3.78	2.92
2.78	2.94
4.51	2.42 (2.5)
3.43	15.08 (16)
1.66 (1)	11.04 (15)
$n_1 = 8$	$n_2 = 8$
$T_1 =$	$T_2 =$

Table 3.13: *AUC* measurements (and ranks) for levocabastine in renal insufficiency patients

15. An efficacy study for fluoxetine in 9 patients with obsessive–compulsive disorder was conducted in an uncontrolled trial (Fontaine and Chouinard,1986). In this trial, baseline and post–treatment (9–week) obsessive–compulsive measurements were obtained using the Comprehensive Psychopathological Rating Scale (CPRS). Note that this is not a controlled trial, in the sense that there was not a group that was randomly assigned a placebo. This trial was conducted very early in the drug screening process. The mean and standard deviation of the differences (baseline–day 56) values for the nine subjects are given below. Use the paired t–test to determine whether or not obsessive–compulsive scores are lower at the end of the trial ($H_0 : \mu_D = 0$ vs $H_A : \mu_D > 0$) at $\alpha = 0.05$.

$$\overline{d} = 9.3 \quad s_d = 2.6 \quad n = 9$$

16. A study investigated the effect of codeine on gastrintestinal motility (Mikus, et al., 1997). Of interest was determing whether or not problems associated with motility are due to the codeine or its metabolite morphine. The study had both a crossover phase and a parallel phase, and was made up of five subjects who are extensive metabolizers and five who are poor metabolizers.

Extensive			Poor		
Codeine	Placebo	$D = C - P$	Codeine	Placebo	$D = C - P$
13.7	7.2	6.5	13.7	11.7	2.0
10.7	4.7	6.0	7.7	6.7	1.0
8.2	5.7	2.5	10.7	6.2	4.5
13.7	10.7	3.0	8.7	6.2	2.5
6.7	6.2	0.5	10.7	11.7	–1.0
$\overline{d} = 3.7$ $s_d = 2.51$			$\overline{d} = 1.8$ $s_d = 2.02$		

Table 3.14: OCTT data for codeine experiment with extensive and poor metabolizers.

(a) In one phase of the study, the researchers measured the orocecal transit times (OCTT in hrs) after administration of placebo and after 60 mg codeine phosphate in healthy volunteers. Data are given in Table 3.14. Test whether or not their is an increase in motility time while on codeine as compared to on placebo separately for each metabolizing group. Use the paired t–test and $\alpha = 0.05$. Intuitively, do you feel this implies that codeine, or its metabolite morphine may be the cause of motility, based on these tests?

(b) In a separate part of the study, they compared the distributions of maximum concentration (C_{max}) of both codeine and its metabolite morphine. The authors compared these by the Wilcoxon Rank–Sum test (under another name). Use the independent sample t–test to compare them as well (although the assumption of equal variances is not appropriate for morphine). The means and standard deviations are given in Table 3.15.

17. Orlistat, an inhibitor of gastrointestinal lipases has recently received FDA approval as treatment for obesity. Based on its pharmacologic effects, there are concerns it may interact with oral contraceptives among women. A study was conducted to determine whether progesterone or luteinizing

	Extensive		Poor	
Substance	$\overline{y}_E$	s_E	$\overline{y}_P$	s_P
Codeine	664	95	558	114
Morphine	13.9	10.5	0.68	0.15

Table 3.15: C_{max} statistics for codeine and morphine among extensive and poor metabolizers.

hormone levels increased when women were simultaneously taking orlistat and oral contraceptives versus when they were only taking contraceptives (Hartmann, et al., 1996). For distributional reasons, the analysis is based on the natural log of the measurements, as opposed to their actual levels. The data and relevant information are given in Table 3.16 for the measurements based on progesterone levels ($\mu g l^{-1}$).

Subject	Orlistat	Placebo	Orl–Plac	rank (|Orl–Plac|)
1	0.5878	0.5653	0.0225	1
2	0.3646	0.3646	0.0000	—
3	0.3920	0.3646	0.0274	2
4	0.9243	1.3558	-0.4316	11
5	0.6831	1.0438	-0.3607	8
6	1.3403	2.1679	-0.8277	12
7	0.3646	0.3646	0.0000	—
8	0.3646	0.3646	0.0000	—
9	0.4253	0.8329	-0.4076	10
10	0.3646	0.3646	0.0000	—
11	0.3646	0.3646	0.0000	—
12	0.3646	0.3646	0.0000	—
13	1.6467	1.4907	0.1561	4
14	0.3646	0.3646	0.0000	—
15	0.5878	0.4055	0.1823	5
16	0.3646	0.3646	0.0000	—
17	0.5710	0.4187	0.1523	3
18	1.1410	1.4303	-0.2893	6
19	1.0919	0.7747	0.3172	7
20	0.7655	0.3646	0.4008	9
Mean	0.654	0.707	–0.053	—
Std Dev	—	—	0.285	—

Table 3.16: $LN(AUC)$ luteinizing hormaone among women receiving orlistat and placebo (crossover design)

.

(a) Based on the Wilcoxon Signed–Rank test, can we determine that levels of the luteinizing hormone are higher among women receiving orlistat than women receiving placebo? Since there are $n = 12$ women with non–zero differences, we reject H_0 : No drug interaction, in favor of H_A : Orlistat increases luteinizing hormone level for $T^- \leq 17$ for $\alpha = 0.05$.

(b) Conduct the same hypothesis test based on the paired–t test at $\alpha = 0.05$.

(c) Based on these tests, should women fear that use of orlistat decreases the efficacy of oral contraceptives (in terms of increasing levels of luteinizing hormones)?

18. Prior to FDA approval of fluoxetine, many trials comparing its efficacy to that of tricyclic antidepressant imipramine and a placebo control (Stark and Hardison,1985). One measure of efficacy was the change in Hamilton Depression score (last visit score – baseline score). The following results are the mean difference and standard deviation of the differences for the placebo group. Obtain a 99% confidence interval for the mean change in Hamilton Depression score for patients receiving a placebo. Is there evidence of a placebo effect?

$$\overline{d} = 8.2 \quad s_d = 9.0 \quad n = 169$$

19. A crossover design was implemented to study the effect of the pancreatic lipase inhibitor Orlistat on postprandial gallbladder contraction (Froehlich, et al.,1996). Of concern was whether use of Orlistat decreased contraction of the gallbladder after consumption of a meal. Six healthy volunters were given both Orlistat and placebo and meals of varying levels of fat. The measurement made at each meal was the AUC for the gallbladder contraction (relative to pre–meal) curve.

The researchers were concerned that Orlistat would reduce gallbladder contraction, and thus increase risk of gallstone formulation. Data for the three meal types are given in Table 3.17. Recall that the same subjects are being used in all arms of the study.

Meal Type	Sample size	$\overline{d}$	s_d
High Fat	6	443	854.6
Mixed	6	313	851.7
No Fat	6	−760	859.0

Table 3.17: AUC statistics for differences within Orlistat and placebo measurements of study subjects.

(a) Is there evidence that for any of these three diets, that a single dose of Orlistat decreases gallbladder contraction? For each diet, test each at $\alpha = 0.01$ using the appropriate normal theory test.

(b) What assumptions are you making in doing these tests?

20. A study of the effect of food on sustained–release theophylline absorption was conducted in fifteen healthy subjects (Boner, et al.,1986). Among the parameters measured in the patients was C_{max}, the maximum concentration of the drug ($\mu g/mL$). The study was conducted as a crossover design, and measurements were made via two assays, enzyme immunoassay test (EMIT) and fluorescence polarization immunoassay (FPIA). Values of C_{max} under fasting and fed states for the EMIT assay are given in Table 3.18. Complete the table and test whether or not food effects the rate of absorption (as measured by C_{max}) by using the Wilcoxon Signed–Rank test ($\alpha = 0.05$). We reject

$t_{1/2}^{SS}$ (hrs)					
Subject (i)	Fasting	Fed	d_i=Fast–Fed	$\|d_i\|$	rank($\|d_i\|$)
1	15.00	15.10	–0.10	0.10	
2	14.70	11.90	2.80	2.80	
3	11.15	16.35	–5.20	5.20	
4	9.75	9.40	0.35	0.35	
5	9.60	12.15	–2.55	2.55	
6	10.05	15.30	–5.25	5.25	
7	9.90	9.35	0.55	0.55	
8	23.15	17.30	5.85	5.85	
9	11.25	12.75	–1.50	1.50	
10	7.80	10.20	–2.40	2.40	
11	15.00	12.95	2.05	2.05	
12	14.45	8.60	5.85	5.85	
13	8.38	6.50	1.88	1.88	
14	7.80	9.65	–1.85	1.85	
15	7.05	11.25	–4.20	4.20	

Table 3.18: C_{max} measurements for theophylline under fasting and fed states

$H_0 : \mu_1 = \mu_2$ in favor of $H_A : \mu_1 \neq \mu_2$ if $T = min(T^+, T^-) \leq T_0 = 25$ for a sample size of 15, and a 2–sided test at $\alpha = 0.05$.

21. A clinical trial was conducted to determine the effect of dose frequency on fluoxetine efficacy and safety among patients with major depressive disorder (Rickels, et al.,1985). Patients were randomly assigned to receive the drug once a day (q.d. patients) or twice a day (b.i.d. patients). Efficacy measurements were based on the Hamilton (HAM–D) depression scale, the Covi anxiety scale, and the Clinical Global Impression (CGI) severity and improvement scales. There were two inferential questions posed: 1) for each dosing regimen, is fluoxetine effective, and 2) is there any differences between the efficacy of the drug between the two dosing regimens. Measurements were made after a 1–week placebo run–in (baseline) and at last visit (after 3–8 visits). For each patient, the difference between the last visit and baseline score was measured.

(a) Within each dosing regimen, what tests would be appropriate for testing efficacy of the drug?

(b) In terms of comparing the effects of the dosing regimens, which tests would be appropriate?

Chapter 4

Statistical Inference – Interval Estimation

A second form of statistical inference, interval estimation, is a widely used tool to describe a population, based on sample data. When it can be used, it often preferred over formal hypothesis testing, although it is used in the same contexts. The idea is to obtain an interval, based on sample statistics, that we can be confident contains the population parameter of interest. Thus, testing a hypothesis that a parameter equals some specified value (such as $\mu_1 - \mu_2 = 0$) can be done by determining whether or not 0 falls in the interval.

Without going into great detail, confidence intervals are formed based on the sampling distribution of a statistic. Recall, for large–samples, $\overline{Y}_1 - \overline{Y}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$. We know then that in approximately 95% of all samples, $\overline{Y}_1 - \overline{Y}_2$ will lie within two standard errors of the mean. Thus, when we take a sample, and observe $\overline{y}_1 - \overline{y}_2$, we can be very confident that this value lies within two standard errors of the unknown value $\mu_1 - \mu_2$. If we add and subtract two standard errors from our sample statistic $\overline{y}_1 - \overline{y}_2$ (also called a **point estimate**), we have an interval that we are very confident contains $\mu_1 - \mu_2$. The algebra for the general case is given in the next section.

In general, we will form a $(1 - \alpha)100\%$ confidence interval for the parameter, where $1 - \alpha$ represents the proportion of intervals that would contain the unknown parameter if this procedure were repeated on many different samples. The width of a $(1 - \alpha)100\%$ confidence interval (I'll usually use 95%) depends on:

- The confidence level $(1 - \alpha)$. As $(1 - \alpha)$ increases, so does the width of the interval. If we want to increase the confidence we have that the interval contains the parameter, we must increase the width of the interval.

- The sample size(s). The larger the sample size, the smaller the standard error of the estimator, and thus the smaller the interval.

- The standard deviations of the underlying distributions. If the standard deviations are large, then the standard error of the estimator will also be large.

4.1 Large–Sample Confidence Intervals

Since many estimators ($\hat{\theta}$) in the previous section are normally distributed with a mean equal to the true parameter (θ), and standard error ($\sigma_{\hat{\theta}}$), we can obtain a **confidence interval** for the true parameter. We first define $z_{\alpha/2}$ to be the point on the standard normal distribution such that $P(Z \geq z_{\alpha/2}) = \alpha/2$. Some values that we will see (and have seen) various times are $z_{.05} = 1.645$, $z_{.025} = 1.96$, and $z_{.005} = 2.576$, respectively. The main idea behind confidence intervals is the following. Since we know that $\hat{\theta} \sim N(\theta, \sigma_\theta)$, then we also know $Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \sim N(0,1)$. So, we can write:

$$P(-z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}) = 1-\alpha$$

A little bit of algebra gives the following:

$$P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1-\alpha$$

This merely says that "in repeated sampling, our estimator will lie within $z_{\alpha/2}$ standard errors of the mean a fraction of $1-\alpha$ of the time." The resulting formula for a large–sample $(1-\alpha)100\%$ **confidence interval for** θ is

$$\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}.$$

When the standard error $\sigma_{\hat{\theta}}$ is unknown (almost always), we will replace it with the estimated standard error $\hat{\sigma}_{\hat{\theta}}$.

In particular, for parallel–groups (which are usually the only designs that have large samples), a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\overline{y}_1 - \overline{y}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example 4.1 An dose–response study for the efficacy of intracavernosal alprostadil in men suffering from erectile dysfunction was reported (Linet, et al.,1996). Patients were assigned at random to receive one of: placebo, 2.5, 5.0, 10.0, or 20.0 μg alprostadil. One measure reported was the duration of erection as measured by RigiScan ($\geq 70\%$ rigidity). We would like to obtain a 95% confidence interval for the difference between the population means for the $20\mu g$ and $2.5\mu g$ groups. The sample statistics are given in Table 4.1.

	$20\mu g$	$2.5\mu g$
Mean	$\overline{y}_1 = 44$ minutes	$\overline{y}_2 = 12$ minutes
Std Dev	$s_1 = 55.8$ minutes	$s_2 = 27.7$ minutes
Sample Size	$n_1 = 58$	$n_2 = 57$

Table 4.1: Sample statistics for alprostadil study in dysfunctional erection patients

For a 95% confidence interval, we need to find $z_{.05/2} = z_{.025} = 1.96$, and the interval can be obtained as follows:

$$(\overline{y}_1 - \overline{y}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \equiv \quad (44-12) \pm 1.96\sqrt{\frac{(55.8)^2}{58} + \frac{(27.7)^2}{57}} \quad \equiv \quad 32 \pm 16.1 \quad \equiv \quad (15.9, 48.1)$$

Thus, we can conclude the (population) mean duration of erection is between 16 and 48 minutes longer for patients receiving a $20\mu g$ dose than for patients receiving $2.5\mu g$. Note that since the entire interval is positive, we can conclude that the mean is significantly higher in the higher dose group.

4.2 Small–Sample Confidence Intervals

In the case of small samples from populations with unknown variances, we can make use of the t-distribution to obtain confidence intervals. In all cases, we must assume that the underlying distribution is approximately normal, although this restriction is not necessary for moderate sample sizes. We will consider the case of estimating the difference between two means, $\mu_1 - \mu_2$, for parallel groups and crossover designs separately. In both of these cases, the sampling distribution of the sample statistic is used to obtain the corresponding confidence interval.

4.2.1 Parallel Groups Designs

When the samples are independent, we use methods very similar to those for the large–sample case. In place of the $z_{\alpha/2}$ value, we will use the $t_{\alpha/2,n_1+n_2-2}$ value, which will be somewhat larger than the z value, yielding a wider interval.

One important difference is that these methods assume the two population variances, although unknown, are equal. We then 'pool' the two sample variances to get an estimate of the common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$. This estimate, that we will call s_p^2 is calculated as follows (we also used this in the hypothesis testing chapter):

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

The corresponding confidence interval can be written:

$$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2,n_1+n_2-2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Example 4.2 Two studies were conducted to study pharmacokinetics of orally and intravenously administered cyclosporine (Gupta, et al.,1990; Aweeka, et al.,1994). The first study involved healthy subjects being given cyclosporine under low–fat and high–fat diet (we will focus on the low–fat phase). The second study was made up of pre–kidney transplant patients. Both studies involved eight subjects, and among the pharmacokinetic parameters reported was the oral bioavailability of cyclosporine. Oral bioavailability (F) is a measure (in percent) that relates the

amount of an oral dose that reaches the systemic circulation to the amount of an intravenous dose that reaches it (Gibaldi, 1984). It can be computed as:

$$F_{oral} \quad = \quad \left(\frac{AUC_{oral} \cdot DOSE_{iv}}{AUCiv \cdot DOSEoral} \right) 100\%,$$

where AUC_{oral} is the area under the concentration–time curve during the oral phase and AUC_{iv} is that for the intravenous phase. In these studies, the intravenous dose was $4mg/Kg$ and the oral dose was $10mg/Kg$. For each study group, the relevant statistics (based on plasma measurements) are given in Table 4.2.

	Healthy Subjects	Pre–Transplant Patients
Mean	$\overline{y}_1 = 21\%$	$\overline{y}_2 = 24\%$
Std Dev	$s_1 = 6\%$	$s_2 = 15\%$
Sample Size	$n_1 = 8$	$n_2 = 8$

Table 4.2: Sample statistics for bioavailability in cyclosporine studies

First, we obtain the pooled variance, s_p^2:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(8-1)(6)^2 + (8-1)(15)^2}{8+8-2} = 130.5 \qquad (s_p = 11.4)$$

Then, we can compute a 95% confidence interval after nothing that $t_{\alpha/2,n_1+n_2-2} = t_{.025,14} = 2.145$:

$$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2,n_1+n_2-2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad \equiv \quad (21-24) \pm 2.145 \sqrt{130.5 \left(\frac{1}{8} + \frac{1}{8} \right)} \quad \equiv -3 \pm 12.3 \equiv (-15.3, 9.3)$$

Thus, we can be confident that the mean oral bioavailability for healthy patients is somewhere between 15.3% lower than and 9.3% higher than that for pre–kidney transplant patients. Since this interval contains both positive and negative values (and thus 0), we cannot conclude that the means differ. As is often the case in small samples, this interval is fairly wide, and our estimate of $\mu_1 - \mu_2$ is not very precise.

4.2.2 Crossover Designs

In studies where the same subject receives each treatment, we make use of this fact and make inferences on $\mu_1 - \mu_2$ through the differences observed within each subject. That is, as we did in the section on hypothesis testing, we will obtain each difference $(TRT1 - TRT2)$, and compute the mean $(\overline{d})$ and standard deviation (s_d) of the n differences. Based on these statistics, we obtain a 95% confidence interval for $\mu_1 - \mu_2$ as follows:

$$\overline{d} \pm t_{\alpha/2,n-1} \frac{s_d}{\sqrt{n}}.$$

Example 4.3 In the cyclosporine study in healthy patients described in Example 4.2, each subject was given the drug with a low fat diet, and again with a high fat diet (Gupta, et al.,1990).

One of the patients did not complete the high fat diet phase, so we will look only at the $n = 7$ healthy patients. Among the pharmacokinetic parameters estimated in each patient was clearance (liters of plasma cleared of drug per hour per kilogram). Clearance was computed as $CL_{iv} = DOSE_{iv}/AUC_{iv}$. Table 4.3 gives the relevant information (based on plasma cyclosporine measurements) to obtain a 95% confidence interval for the difference in mean CL_{iv} for oral dose under high and low fat diet phases. Based on the data in Table 4.3 and the fact that $t_{.025,6} = 2.447$, we can

CL_{iv} $(L/hr/Kg)$			
Subject (i)	High Fat	Low Fat	d =High–Low
1	0.569	0.479	0.090
2	0.668	0.400	0.268
3	0.624	0.358	0.266
4	0.521	0.372	0.149
5	0.679	0.563	0.116
7	0.939	0.636	0.303
8	0.882	0.448	0.434
Mean	0.697	0.465	$\overline{d} = 0.232$
Std Dev	0.156	0.103	$s_d = 0.122$

Table 4.3: Cyclosporine CL_{iv} measurements for high and low fat diets in healthy subjects

obtain a 95% confidence interval for the true mean difference in cyclosporine under high and low fat diets:

$$\overline{d} \pm t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}} \quad \equiv \quad 0.232 \pm 2.447\frac{0.122}{\sqrt{7}} \quad \equiv \quad 0.232 \pm 0.113 \quad \equiv \quad (0.119, 0.345).$$

We can be 95% confident that the true difference in mean clearance for high and low fat diets is between .119 and .345 $L/hr/Kg$. Since the entire interval is positive, we can conclude that clearance is greater on a high fat diet (that is, when taken with food) than on a low fat diet. The authors concluded that food enhances the removal of cyclosporine in healthy subjects.

4.3 Exercises

22. Compute and interpret 95% confidence intervals for $\mu_1 - \mu_2$ for problems 11, 12, and 14 of Chapter 3.

Chapter 5

Categorical Data Analysis

We have seen previously that variables can be categorical or numeric. The past two chapters dealt with comparing two groups in terms of quantitaive responses. In this chapter, we will introduce methods commonly used to analyze data when the response variable is categorical. The data are generally counts of individuals, and are given in the form of an $r \times c$ **contingency table**. Throughout these notes, the rows of the table will represent the r levels of the explanatory variable, and the columns will represent the c levels of the response variable. The numbers within the table are the counts of the numbers of individuals falling in that cell's combination of levels of the explanatory and response variables. The general set–up of an $r \times c$ contingency table is given in Table 5.1.

		Response Variable				
		1	2	$\cdots$	c	
	1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
Explanatory	2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
Variable	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
		$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	

Table 5.1: An $r \times c$ Contingency Table

Recall that categorical variables can be **nominal** or **ordinal**. Nominal variables have levels that have no inherent ordering, such as sex (male, female) or hair color (black, blonde, brown, red). Ordinal variables have levels that do have a distinct ordering such as diagnosis after treatment (death, worsening of condition, no change, moderate improvement, cure).

In this chapter, we will cover the cases: 1) 2×2 tables, 2) both of the variables are nominal, 3) both of the variables are ordinal, and 4) the explanatory variable is nominal and the response variable is ordinal. All cases are based on independent sample (parallel groups studies), although case–control studies can be thought of as paired samples, however (although not truly crossover designs). We won't pursue that issue here. Statistical texts that cover these topics in detail include (Agresti,1990), which is rather theoretical and (Agresti,1996) which is more applied and written

specifically for applied practicioners.

5.1 2×2 Tables

There are many situations where both the independent and dependent variables have two levels. One example is efficacy studies for drugs, where subjects are assigned at random to active drug or placebo (explanatory variable) and the outcome measure is whether or not the patient is cured (response variable). A second example is epidemiological studies where disease state is observed (response variable), as well as exposure to risk factor (explanatory variable). Drug efficacy studies are generally conducted as randomized clinical trials, while epidemiological studies are generally conducted in cohort and case–control settings (see Chapter 1 for descriptions of these type studies).

For this particular case, we will generalize the explanatory variable's levels to exposed (E) and not exposed ($\overline{E}$), and the response variable's levels as disease (D) and no disease ($\overline{D}$). These interpretations can be applied in either of the two settings described above and can be generalized to virtually any application. The data for this case will be of the form of Table 5.2.

		Disease State		
		D (Present)	$\overline{D}$ (Absent)	Total
Exposure	E (Present)	n_{11}	n_{12}	$n_{1.}$
State	$\overline{E}$ (Absent)	n_{21}	n_{22}	$n_{2.}$
Total		$n_{.1}$	$n_{.2}$	n

Table 5.2: A 2×2 Contingency Table

In the case of drug efficacy studies, the exposure state can be thought of as the drug the subject is randomly assigned to. Exposure could imply that a subject was given the active drug, while non–exposure could imply having received placebo. In either type study, there are two measures of association commonly estimated and reported. These are the **relative risk** and the **odds ratio**.

These methods are also used when the explanatory variable has more than two levels, and the response variable has two levels. The methods described below are computed within pairs of levels of the explanatory variables, with one level forming the "baseline" group in comparisons. This extension will be described in Example 5.3.

5.1.1 Relative Risk

For prospective studies (cohort and randomized clinical trials), a widely reported measure of association between exposure status and disease state is relative risk. Relative risk is a ratio of the probability of obtaining the disease among those exposed to the probability of obtaining disease among those not exposed. That is:

$$\text{Relative Risk} \quad = \quad RR \quad = \quad \frac{P(D|E)}{P(D|\overline{E})}$$

Based on this definition:

- A relative risk greater than 1.0 implies the exposed group have a higher probability of contracting disease than the unexposed group.
- A relative risk less than 1.0 implies that the exposed group has a lower chance of contracting disease than unexposed group (we might expect this to be the case in drug efficacy studies).
- A relative risk of 1.0 implies that the risk of disease is the same in both exposure groups (no association between exposure state and disease state).

Note that the relative risk is a population parameter that must be estimated based on sample data. We will be able to calculate confidence intervals for the relative risk, allowing inferences to be made concerning this population parameter, based on the range of values of RR within the $(1-\alpha)100\%$ confidence interval. The procedure to compute a $(1-\alpha)100\%$ confidence interval for the population relative risk is as follows:

1. Obtain the sample proportions of exposed and unexposed subjects who contract disease. These values are: $\hat{\pi}_E = \frac{n_{11}}{n_{1.}}$ and $\hat{\pi}_{\overline{E}} = \frac{n_{21}}{n_{2.}}$, respectively.
2. Compute the estimated relative risk: $RR = \frac{\hat{\pi}_E}{\hat{\pi}_{\overline{E}}}$.
3. Compute $v = \frac{(1-\hat{\pi}_E)}{n_{11}} + \frac{(1-\hat{\pi}_{\overline{E}})}{n_{21}}$ (This is the estimated variance of $ln(RR)$).
4. The confidence interval can be computed as: $(RRe^{-z_{\alpha/2}\sqrt{v}}, RRe^{z_{\alpha/2}\sqrt{v}})$.

Example 5.1 An efficacy study was conducted for the drug pamidronate in patients with stage III multiple myeloma and at least one lytic lesion (Berenson, et al.,1996). In this randomized clinical trial, patients were assigned at random to receive either pamidronate (E) or placebo ($\overline{E}$). One endpoint reported was the occurrence of any skeletal events after 9 cycles of treatment (D) or non–occurrence ($\overline{D}$). The results are given in Table 5.3. We will use the data to compute a 95% confidence interval for the relative risk of suffering skeletal events (in a time period of this length) for patients on pamidronate relative to patients not on the drug.

		Occurrence of Skeletal Event		
		Yes (D)	No ($\overline{D}$)	
Treatment Group	Pamidronate (E)	47	149	196
	Placebo ($\overline{E}$)	74	107	181
		121	256	377

Table 5.3: Observed cell counts for pamidronate data

First, we obtain the proportions of patients suffering skeletal events among those receiving the active drug, and among those receiving the placebo:

$$\hat{\pi}_E = \frac{n_{11}}{n_{1.}} = \frac{47}{196} = 0.240 \qquad \hat{\pi}_{\overline{E}} = \frac{n_{21}}{n_{2.}} = \frac{74}{181} = 0.409$$

Then we can compute the estimated relative risk (RR) and the estimated variance of its natural log (v):

$$RR = \frac{\hat{\pi}_E}{\hat{\pi}_{\overline{E}}} = \frac{.240}{.409} = 0.587 \qquad v = \frac{(1-\hat{\pi}_E)}{n_{11}} + \frac{(1-\hat{\pi}_{\overline{E}})}{n_{21}} = \frac{(1-.240)}{47} + \frac{(1-.409)}{74} = .016 + .008 = .024$$

Finally, we obtain a 95% confidence interval for the population relative risk (recall that $z_{.025} = 1.96$):

$$(RRe^{-z_{.05/2}\sqrt{v}}, RRe^{z_{.05/2}\sqrt{v}}) \quad \equiv \quad (0.587e^{-1.96\sqrt{.024}}, 0.587e^{1.96\sqrt{.024}})$$

$$\equiv (0.587(0.738), 0.587(1.355)) \quad \equiv \quad (0.433, 0.795)$$

Thus, we can be confident that the relative risk of suffering a skeletal event (in this time period) for patients on pamidronate (relative to patients not on pamidronate) is between 0.433 and 0.795. Since this entire interval is below 1.0, we can conclude that pamidronate is effective at reducing the risk of skeletal events. Further, we can estimate that pamidronate reduces the risk by $(1 - RR)100\% = (1 - 0.587)100\% = 41.3\%$.

5.1.2 Odds Ratio

For retrospective (case–control) studies, subjects are identified as cases (D) or controls ($\overline{D}$), and it is observed whether the subjects had been exposed to the risk factor (E) or not ($\overline{E}$). Since we are not sampling from the populations of exposed and unexposed, and observing whether or not disease occurs (as we do in prospective studies), we cannot estimate $P(D|E)$ or $P(D|\overline{E})$.

First we define the **odds** of an event occurring. If π is the probability that an event occurs, the odds o that it occurs is $o = \pi/(1-\pi)$. The odds can be interpreted as the number of times the event will occur for every time it will not occur if the process were repeated many times. For example, if you toss a coin, the probability it lands heads is $\pi = 0.5$. The corresponding odds of a head are $o = 0.5/(1-0.5) = 1.0$. Thus if you toss a coin many the times, the odds of a head are 1.0 (or 1–to–1 if you've ever been to a horse or dog track). Note that while odds are not probabilities, they are very much related to them: high probabilities are associated with high odds, and low probabilities are associated with low odds. In fact, for events with very low prababilities, the odds are very close to the probability of the event.

While we cannot compute $P(D|E)$ or $P(D|\overline{E})$ for retrospective studies, we can compute the odds that a person was exposed given they have the disease, and the odds that a person was exposed given they don't have the disease. The ratios of these two odds is called the **odds ratio**. The odds ratio (OR) is similar to the relative risk, and is virtually equivalent to it when the prevalence of the disease ($P(D)$) is low. The odds ratio is computed as:

$$OR = \frac{\text{odds of disease given exposed}}{\text{odds of disease given unexposed}} = \frac{\text{odds of exposure given diseased}}{\text{odds of exposure given not diseased}} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

The odds ratio is similar to relative risk in the sense that it's a population parameter that must be estimated, as well as the interpretations associated with it in terms of whether its value is above, below, or equal to 1.0. That is:

- If the odds ratio is greater than 1.0, the odds (and thus probability) of disease is higher among exposed than unexposed.
- If the odds ratio is less than 1.0, the odds (and thus probability) of disease is lower among exposed than unexposed.
- If the odds ratio is 1.0, the odds (and thus probability) of disease is the same for both groups (no association between exposure to risk factor and disease state).

The procedure to compute a $(1-\alpha)100\%$ confidence interval for the population odds ratio is as follows:

1. Obtain the estimated odds ratio: $OR = \frac{n_{11}n_{22}}{n_{12}n_{21}}$.
2. Compute $v = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$ (this is the variance of $ln(OR)$).
3. The confidence interval can be computed as: $(ORe^{-z_{\alpha/2}\sqrt{v}}, ORe^{z_{\alpha/2}\sqrt{v}})$.

Example 5.2 An epidemiological case–control study was reported, with cases being 537 people diagnosed with lip cancer (D) and controls being made up of 500 people with no lip cancer ($\overline{D}$) (Broders, 1920). One risk factor measured was whether or not the subject had smoked a pipe (pipe smoker – E, non–pipe smoker – $\overline{E}$). Table 5.4 gives the numbers of subjects falling in each lip cancer/pipe smoking combination. We would like to compute a 95% confidence interval for the population odds ratio, and determine whether or not pipe smoking is associated with higher (or possibly lower) odds (and probability) of contracting lip cancer.

		Occurrence of Lip Cancer		
		Yes (D)	No ($\overline{D}$)	
Pipe Smoking Status	Yes (E)	339	149	488
	No ($\overline{E}$)	198	351	549
		537	500	1037

Table 5.4: Observed cell counts for lip cancer/pipe smoking data

We compute the confidence interval as described above, again recalling that $z_{\alpha/2} = z_{0.025} = 1.96$:

1. $OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{339(351)}{149(198)} = 4.03$.
2. $v = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} = \frac{1}{339} + \frac{1}{149} + \frac{1}{198} + \frac{1}{351} = 0.0176$
3. 95% CI: $(ORe^{-z_{\alpha/2}\sqrt{v}}, ORe^{z_{\alpha/2}\sqrt{v}}) = (4.03e^{-1.96\sqrt{.0176}}, 4.03e^{1.96\sqrt{.0176}}) = (3.11, 5.23)$.

We can be 95% confident that the population odds ratio is between 3.11 and 5.23. That is the odds of contracting lip cancer is between 3.1 and 5.2 times higher among pipe smokers than non–pipe smokers. Note that in making the inference that pipe smoking *causes* lip cancer, we would need to demonstrate this association after controlling for other potential risk factors. We will see methods for doing this in later sections.

5.1.3 Extension to $r \times 2$ Tables

As mentioned above, we can easily extend these methods to explanatory variables with more than $r = 2$ levels, by defining a baseline group, and forming $r - 1$ 2×2 tables with the baseline group always acting as the unexposed ($\overline{E}$) group. When the explanatory variable is ordinal, there will be a natural baseline group, otherwise one is arbitrarily chosen.

Example 5.3 A cohort study was conducted involving use of hypertension drugs and the occurrence of cancer during a four–year period (Pahor, et al.,1996). The study group involved 750 subjects, each with no history of cancer and over the age of 70. Patients were classified as users of β–blockers, angiotensin converting enzyme (ACE) inhibitors, or calcium channel blockers (verapamil, nifedipine, and dilitiazem). Most subjects on calcium channel blockers were on the short–acting formulation. The authors used the group receiving β–blockers as the baseline group, so that the relative risks reported are for ACE inhibitors relative to β–blockers, and for calcium channel blockers relative to β–blockers. The results, including estimates and 95% confidence intervals are given in Table 5.5. The unadjusted values are based on the raw data and the formulas described above; the adjusted values are those reported by the authors after fitting a proportional hazards regression model (see Chapter 9). The adjusted values control for patient differences with respect to: age, gender, race, smoking, body mass index, and hospital admissions (not related to cancer). We will see very small differences between the adjusted and unadjusted values; this a sign that the three treatment (drug) groups are similar in terms of the levels of these other factors.

Drug	Raw Data		Unadjusted		Adjusted	
Class	# Patients	# Cancers	Rel. Risk	95% CI	Rel. Risk	95% CI
β–blockers	424	28	1.00	–	1.00	–
ACE inhibitors	124	6	0.73	(0.31,1.73)	0.73	(0.30,1.78)
Calcium channel blockers	202	27	2.03	(1.23,3.34)	2.02	(1.16,3.54)

Table 5.5: Cancer by drug class data with unadjusted and adjusted relative risks

Note that the confidence interval for the relative risk of developing cancer on ACE inhibitors relative to β–blockers contains 1.00 (no association), so we cannot conclude that differences exist in cancer risk among these two drug classes. However, when we compare calcium channel blockers to β–blockers, the entire confidence interval is above 1.00, so the risk of developing cancer is higher among patients taking calcium channel blockers. It can also be shown that the risk is higher for patients on calcium channel blockers than among patients on ACE inhibitors (compute the 95% CI for the relative risk to see this).

5.1.4 Difference Between 2 Proportions (Absolute Risk)

The relative risk was a measure of a ratio of two proportions. In the context mentioned above, it could be thought of as the ratio of the probability of a specific outcome (e.g. death or disease) among an exposed population to the same probability among an unexposed population. We could also compare the proportions by studying their difference, as opposed to their ratio. In medical

studies, relative risks and odds ratios appear to be reported much more often than differences, but we will describe the process of comparing two population proportions briefly.

Using notation described above, we have $\pi_E - \pi_{\overline{E}}$ representing the difference in proportions of events between an exposed and an unexposed group. When our samples are independent (e.g. parallel groups design), the estimator $\hat{\pi}_E - \hat{\pi}_{\overline{E}}$, the difference in sample proportions, is approximately normal in large samples. Its sampling distribution can be written as:

$$\hat{\pi}_E - \hat{\pi}_{\overline{E}} \sim N\left(\pi_E - \pi_{\overline{E}}, \sqrt{\frac{\pi_E(1-\pi_E)}{n_E} + \frac{\pi_{\overline{E}}(1-\pi_{\overline{E}})}{n_{\overline{E}}}}\right),$$

where n_E and $n_{\overline{E}}$ are the sample sizes from the exposed and unexposed groups, respectively.

Just as a relative risk of 1 implied the proportions were the same (no treatment effect in the case of experimental treatments), an absolute risk of 0 has the same interpretation.

Making use of the large–sample normality of the estimator based on the difference in the sample proportions, we can compute a confidence interval for $\pi_E - \pi_{\overline{E}}$ or test hypotheses concerning its value as follow (see Table 5.2 for labels). The results for the confidence interval are given below, to conduct a test, you would simply take the ratio of the estimate to its standard error to obtain a z–statistic.

1. Obtain the sample proportions of exposed and unexposed subjects who contract disease. These values are: $\hat{\pi}_E = \frac{n_{11}}{n_{1\cdot}}$ and $\hat{\pi}_{\overline{E}} = \frac{n_{21}}{n_{2\cdot}}$, respectively.
2. Compute the estimated difference in proportions (absolute risk): $\hat{\pi}_E - \hat{\pi}_{\overline{E}}$.
3. Compute $v = \frac{\hat{\pi}_E(1-\hat{\pi}_E)}{n_{1\cdot}} + \frac{\hat{\pi}_{\overline{E}}(1-\hat{\pi}_{\overline{E}})}{n_{2\cdot}}$ (This is the estimated variance of $\hat{\pi}_E - \hat{\pi}_{\overline{E}}$).
4. The confidence interval can be computed as: $(\hat{\pi}_E - \hat{\pi}_{\overline{E}}) \pm z_{\alpha/2}\sqrt{v}$

Note that this method appears to be rarely reported in the medical literature. Particularly, in situations where the proportions are very small (e.g. rare diseases), the difference $\pi_E - \pi_{\overline{E}}$ may be relatively small, while the risk ratio may be large. Consider the case where 3% of an exposed group, and 1% of an unexposed group have the outcome of interest.

Example 5.4 In what may be the first truly ranomized clinical trial, British patients were randomized to receive either streptomycin and bed rest or simply be rest (Medical Research Council, 1948). The supply of streptomycin was very limited, which justified conducting an experiment where only half of the subjects received the active treatment. Nowadays of course, that is common practice. The exposure variable will be the treatment (streptomycin/control). The outcome of interest will be whether or not the patient showed improvement. Let π_E be the proportion of all TB patients who, if given streptomycin, would show improvement at 6 months. Further we will define $\pi_{\overline{E}}$ as a similar proportion among patients receiving only bed rest. The data are given in Table 5.6.

We get the following relevant quantities:

$$\hat{\pi}_E = \frac{38}{55} = .691 \quad \hat{\pi}_{\overline{E}} = \frac{17}{52} = .327 \quad v = \frac{.691(.309)}{55} + \frac{.327(.673)}{52} = .0081$$

		Improvement in Condition		
		Yes	No	
Treatment Group	Streptomycin (E)	38	17	55
	Control ($\overline{E}$)	17	35	52
		55	52	107

Table 5.6: Observed cell counts for streptomycin data

Then a 95% confidence interval for the true difference (absolute risk) is:

$$(.691 - .327) \pm 1.96\sqrt{.0081} \quad \equiv \quad .364 \pm .176 \quad \equiv \quad (.188, .540).$$

We can conclude that the proportion of all patients given streptomycin who show improvement is between .188 and .540 higher than patients not receiving streptomycin at the 95% level of confidence. Since the entire interval exceeds 0, we can can conclude that the streptomycin appears to provide a real effect.

In the case of crossover designs, there is a method that makes use of the paired nature of the data. It is referred to as McNemar's test.

5.1.5 Small–Sample Inference — Fisher's Exact Test

The tests for association described previously all assume that the samples are sufficiently large so that the estimators (or their logs in the case of relative risk and odds ratio) have sampling distributions that are approximately normal. However, in many instances studies are based on small samples. This may arise due to cost or ethical reasons. A test due to R.A. Fisher, *Fisher's exact test*, was developed for this particular situation. The logic of the test goes as follows:

We have a sample with $n_{1.}$ people (or experimental units) that are considered exposed and $n_{2.}$ that are considered not exposed. Further we have $n_{.1}$ individuals that contract the event of interest (e.g. death or disease), of which n_{11} were exposed. The question is, conditional on the number exposed and the number of events, what is the probability that as many or more (fewer) of the events could have been in the exposed group (under the assumption that there is no exposure effect in the population). The math is not difficult, but can be confusing. We will simply present the test through an example, skipping the computation of the probabilities.

Example 5.5 A study was reported on the effects of antiseptic treatment among amputations in a British surgical hospital (Lister, 1870). Tragically for Dr. Lister, he lived before Fisher, so he felt unable to make and inference based on statistical methodology, although he saw the effect was certainly there. We can make use of Fisher's exact test to make the inference. The study had two groups: one group based on amputation without antiseptic (years 1864 and 1866), and a group based on amputation with antiseptic (years 1867–1869). All surgeries were in the same hospital. We will consider the patients with antiseptic as the exposed. The endpoint reported was death (apparently due to the surgery and disease that was associated with it). The results are given in Table 5.7.

		Surgical Outcome: Death	Surgical Outcome: No Death	
Treatment Group	Antiseptic (E)	6	34	40
	Control ($\overline{E}$)	16	19	35
		22	53	75

Table 5.7: Observed cell counts for antiseptic data

Note that this study is based on historical, as opposed to concurrent controls. From the data we that there were 40 patients exposed to the antiseptic and 22 deaths, of which 6 were treated with antiseptic. Now if the treatment is effective, it should reduce deaths, so we have to ask what is the probability that 6 or fewer of the 22 deaths could have been in the antiseptic group, given there were 40 patients in that group. It ends up that this probability is .0037. That is, under the assumption of no treatment effect, the probability that based on a sample of this size, and this number of deaths, it is very unlikely that the sample results would have been this strong or stronger in favor of the antiseptic group. If we conduct this test with $\alpha = 0.05$, the p–value (.0037) is smaller than α, and we conclude that the antiseptic was associated with a lower probability of death.

5.1.6 McNemar's Test for Crossover Designs

When the same subjects are being observed under both experimental treatments, McNemar's test can be used to test for treatment effects. The important subjects are the ones who respond differently under the two conditions. Counts will appear as in Table 5.8.

		Trt 2 Outcome: Present	Trt 2 Outcome: Absent	
Trt 1 Outcome	Present	n_{11}	n_{12}	$n_{1.}$
	Absent	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$n_{..}$

Table 5.8: Notation for McNemar's Test

Note that n_{11} are subjects who have the outcome characteristic present under both treatments, while n_{22} is the number having the outcome characteristic absent under both treatments. None of these subjects offer any information regarding treatment effects. The subjects who provide information are the n_{12} individuals who have the outcome present under treatment 1, and absent under treatment 2; and the n_{21} individuals who have the outcome absent under treatment 1, and present under treatment 2. Note that treatment 1 and treatment 2 can also be "Before" and "After" treatment, or any two conditions.

A large-sample test for treatment effects can be conducted as follows.

- H_0 : Pr(Outcome Present—Trt 1)=Pr(Outcome Present—Trt 2) (No Trt effect)

- H_A : The probabilities differ (Trt effects - This can be 1-sided also)
- $TS : z_{obs} = \frac{n_{12}-n_{21}}{\sqrt{n_{12}+n_{21}}}$
- $RR : |zobs| \geq z_{\alpha/2}$ (For 2-sided test)
- P-value: $2P(Z \geq |z_{obs}|)$ (For 2-sided test)

Often this test is reported as a chi-square test. The statistic is the square of the z-statistic above, and its treated as a chi-square random variable with one degree of freedom (which will be discussed shortly). The 2-sided z-test, and the chi-square test are mathematically equivalent.

Example 5.6
A study involving a cohort of women in Birmingham, AL examined revision surgery involving silicone gel breast implants (Brown and Pennello, 2002). Of 165 women with surgical records who had reported having surgery, the following information was obtained.

- In 69 cases, both self report and surgical records said there was a rupture or leak.
- In 63 cases, both self report and surgical records said there was not a rupture or leak.
- In 28 cases, the self report said there was a rupture or leak, but the surgical records did not report one.
- In 5 cases, the self report said there was not a rupture or leak, but the surgical records reported one.

The data are summarized in Table 5.9. Present refers to a rupture or leak, Absent refers to no rupture or leak.

		Surgical Record		
		Present	Absent	
Self	Present	69	28	97
Report	Absent	5	63	68
		74	91	165

Table 5.9: Self Report and Surgical Records of Silicone breast implant rupture/leak

We can test whether the tendency to report ruptures/leaks differs between self reports and surgical records based on McNemar's test, since both outcomes are being observed on the same women.

- H_0 : No differences in tendency to report ruptures/leaks between self reports and surgical records
- H_A : The probabilities differ

- $TS: z_{obs} = \frac{28-5}{\sqrt{28+5}} = \frac{23}{5.74} = 4.00$
- $RR: |zobs| \geq z_{.025} = 1.96$ (For 2-sided test, with $\alpha = 0.05$)
- P-value: $2P(Z \geq 4.00) \approx 0$ (For 2-sided test)

Thus, we conclude that the tendencies differ. Self reports appear to be more likely to report a rupture or leak than surgical records.

5.1.7 Mantel–Haenszel Estimate for Stratified Samples

In some situations, the subjects in the study may come from one of several populations (strata). For instance, an efficacy study may have been run at multiple centers, and there may be some "center" effect that is related to the response. Another example is if race is related to the outcomes, and we may wish to adjust for race by computing odds ratios separately for each race, then combine them.

This is a situation where we would like to determine if there is an association between the explanatory and response variables, after *controlling* for a second explanatory variable. If there are k populations, then we can arrange the data (in a different notation than in the previous sections) as displayed in Table 5.10. Note that for each table, n_i is the sample size for that strata ($n_i = A_i + B_i + C_i + D_i$). The procedure was developed specifically for retrospective case/control studies, but may be applied to prospective studies as well (Mantel and Haenszel,1959).

		Strata 1 Disease State				Strata 2 Disease State		
		D (Present)	$\overline{D}$ (Absent)	Total		D (Present)	$\overline{D}$ (Absent)	Total
Exposure State	E (Present)	A_1	B_1		... E	A_k	B_k	
	$\overline{E}$ (Absent)	C_1	D_1		... E	C_k	D_k	
Total				n_1				n_k

Table 5.10: Contingency Tables for Mantel–Haenszel Estimator

The estimator of the odds ratio is computed as:

$$OR_{MH} = \frac{R}{S} = \frac{\sum_{i=1}^{k} R_i}{\sum_{i=1}^{k} S_i} = \frac{\sum_{i=1}^{k} A_i D_i / n_i}{\sum_{i=1}^{k} B_i C_i / n_i}$$

One estimate of the variance of the log of OR_{MH} is:

$$v = \hat{V}(ln(OR_{MH})) = \frac{1}{S^2} \sum_{i=1}^{k} S_i^2 \left(\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i} \right)$$

As with the odds ratio, we can obtain a 95% CI for the population odds ratio as:

$$(OR_{MH} e^{-1.96\sqrt{v}}, OR_{MH} e^{1.96\sqrt{v}})$$

.

Example 5.7 A large study relating smoking habits and death rates reported that cigarette smoking was related to higher death rate (Hammond and Horn,1954). Men were classified as regular cigarette smokers (E) and noncigarette smokers ($\overline{E}$). The nonsmokers had never smoked cigarettes regularly. There were a total of 187,766 men who were successfully traced from the early 1952 start of study through October 31,1953. Of that group, 4854 (2.6%) had died.

A second variable that would clearly be related to death was age. In this study, all men were 50–69 at entry. The investigators then broke these ages down into four strata (50–54,55–59,60–64,65–69). The overall outcomes (disregarding age) are given in Table 5.11. Note that the overall odds ratio is $OR = (3002(78092))/(104820(1852)) = 1.21$.

		Occurrence of Death		
		Yes (D)	No ($\overline{D}$)	
Cigarette Smoking	Yes (E)	3002	104280	107822
Status	No ($\overline{E}$)	1852	78092	79944
		4854	182912	187766

Table 5.11: Observed cell counts for cigarette smoking/death data

The data, stratified by age group, are given in Table 5.12. Also, the odds ratios, proportion deaths ($P(D)$), and proportion smokers ($P(E)$) are given.

Age Group (i)	A_i	B_i	C_i	D_i	n_i	R_i	S_i	OR	$P(D)$	$P(E)$
50–54 (1)	647	39990	204	20132	60973	213.6	133.8	1.60	.0140	.6665
55–59 (2)	857	32894	394	21671	55816	332.7	232.2	1.43	.0224	.6047
60–64 (3)	855	20739	488	19790	41872	404.1	241.7	1.67	.0321	.5157
65–69 (4)	643	11197	766	16499	29105	364.5	294.7	1.24	.0484	.4068

Table 5.12: Observed cell counts and odds ratio calculations (by age group) for cigarette smoking/death data

Note that the odds ratio is higher within each group than it is for the overall group. This is referred to as *Simpson's Paradox*. In this case it can be explained as follows:

- Mortality increases with age from 1.40% for 50–54 to 4.84% for 65–69.
- As the age increases, the proportion of smokers decreases from 66.65% to 40.68%
- A higher proportion of nonsmokers are in the higher risk (age) groups than are smokers. Thus, the nonsmokers are at a "disadvantage" because more of them are in the higher age groups (many smokers in the population have already died before reaching that age group).

This leads us to desire an estimate of the odds ratio *adjusted for age*. That is what the Mantel–Haenszel estimate provides us with. We can now compute it as described above:

$$R = \sum_{i=1}^{4} R_i = 213.6+332.7+404.1+364.5 = 1314.9 \quad S = \sum_{i=1}^{4} S_i = 133.8+232.2+241.7+294.7 = 902.4$$

$$OR_{MH} = \frac{R}{S} = \frac{1314.9}{902.4} = 1.46$$

The estimated variance of $ln(OR_{MH})$ is 0.00095 (trust me). Then we get the following 95%CI for the odds ratio in the population of males in the age group 50–69 (adjusted for age):

$$(OR_{MH}e^{-1.96\sqrt{v}}, OR_{MH}e^{1.96\sqrt{v}}) \equiv (1.46e^{-1.96\sqrt{.00095}}, 1.46e^{1.96\sqrt{.00095}}) \equiv (1.37, 1.55).$$

We can be very confident that the odds of death (during the length of time of the study – 20 months) is between 37% and 55% higher for smokers than nonsmokers, after controlling for age (among males in the 50–69 age group).

5.2 Nominal Explanatory and Response Variables

In cases where both the explanatory and response variables are nominal, the most commonly used method of testing for association between the variables is the **Pearson Chi–Squared Test**. In these situations, we are interested if the probability distributions of the response variable are the same at each level of the explanatory variable.

As we have seen before, the data represent counts, and appear as in Table 5.1. The n_{ij} values are referred to as the *observed* values. If the variables are independent (not associated), then the population probability distributions for the response variable will be identical within each level of the explanatory variable, as in Table 5.13.

		Response Variable				
		1	2	$\cdots$	c	
	1	p_1	p_2	$\cdots$	p_c	1.0
Explanatory	2	p_1	p_2	$\cdots$	p_c	1.0
Variable	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	r	p_1	p_2	$\cdots$	p_c	1.0

Table 5.13: Probability distributions of response variable within levels of explanatory variable under condition of no association between the two variables.

We have already seen special cases of this for 2×2 tables. For instance, in Example 5.1, we were interested in whether the probaility distributions of skeletal incident status were the same for the active drug and placebo groups. We demonstrated that the probability of having a skeletal incident was higher in the placebo group, and thus treatment and skeletal incident variables were associated (not independent).

To perform Pearson's Chi–square test, we compute the *expected* values for each cell count under the hypothesis of independence, and we obtain a statistic based on discrepancies between the *observed* and *expected* values:

$$observed = n_{ij} \qquad expected = \frac{n_{i.}n_{.j}}{n}$$

The expected values represent how many individuals would have fallen in cell (i, j) if the probability distributions of the response variable were the same for each level of the explanatory variable. The test is conducted as follows:

1. H_0 : No association between the explanatory and response variables (see Table 5.13).
2. H_A : Explanatory and response variables are associated
3. T.S.: $X^2 = \sum_{allcells} \frac{(observed-expected)^2}{expected} = \sum_{i,j} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$
4. RR: $X^2 > \chi^2_{\alpha,(r-1)(c-1)}$ where $\chi^2_{\alpha,(r-1)(c-1)}$ is a critical value that can be found in Table A.3.
5. p–value: $P(\chi^2 \geq X^2)$

Example 5.8 A case–control study was conducted in Massachusetts regarding habits, characteristics, and environments of individuals with and without cancer (Lombard and Doering, 1928). Among the many factors that they reported was marital status. We will conduct Pearson's Chi–squared test to determine whether or not cancer status (response variable) is independent of marital status. The observed and expected values are given in Table 5.14.

Marital Status	Cancer	No Cancer	Total
Single	29 (38.1)	47 (37.9)	76
Married	116 (112.3)	108 (111.7)	224
Widowed	67 (61.6)	56 (61.4)	123
Div/Sep	5 (5.0)	5 (5.0)	10
Total	217	216	433

Table 5.14: Observed (expected) values of numbers of subjects within each marital/cancer status group (One non–cancer control had unknown marital status)

To obtain the expected cell counts, we take the row total times the column total divided by the overall total. For instance, for the single cancer cases, we get $exp = (76)(217)/433 = 38.1$. Now, we can test: H_0:Marital and cancer status are independent vs H_A: Marital and cancer status are associated. We compute the test statistic as follows:

$$X^2 = \sum \frac{(observed - expected)^2}{expected} = \frac{(29 - 38.1)^2}{38.1} + \frac{(47 - 37.9)^2}{37.9} + \cdots + \frac{(5 - 5.0)^2}{5.0} = 5.53$$

We reject H_0 for values of $X^2 \geq \chi^2_{\alpha,(r-1)(c-1)}$. For this example, if we have $r = 4$ and $c = 2$, so $(r - 1)(c - 1) = 3$, and if we test at $\alpha = 0.05$, we reject H_0 if $X^2 \geq \chi^2_{.05,3} = 7.81$. Since the test statistic does not fall in the rejection region, we fail to reject H_0, and we cannot conclude that marital status is associated with the occurrence of cancer.

5.3 Ordinal Explanatory and Response Variables

In situations where both the explanatory and response variables are ordinal, we would like to take advantage of the fact that the levels of the variables have distinct orderings. We can ask questions such as: Do individuals with high levels of the explanatory variable tend to have high (low) levels of the corresponding response variable. For instance, suppose that the explanatory variable is dose, with increasing (possibly numeric) levels of amount of drug given to a subject, and the response variable is a categorical measure (possibly subjective) of degree of improvement. Then, we may be interested in seeing if as dose increases, the degree of improvement increases (this is called a dose–response relationship).

Many measures have been developed for this type of experimental setting. Most are based on **concordant** and **discordant** pairs. Concordant pairs involve pairs where one subject scores higher on both variables than the other subject. Discordant pairs are pairs where one subject scores higher on one variable, but lower on the other variable, than the other subject.

In cases where there is a **positive association** between the two variables, we would expect more concordant than discordant pairs. That is, there should be many subjects who score high on both variables, and many who score low on both, with fewer subjects scoring high one variable and low on the other. On the other hand, if there is a **negative association**, we would expect more discordant pairs than concordant pairs. That is, people will tend to score high on one variable, but lower on the other.

Example 5.9 A dose–response study was conducted to study nicotine and cotinine replacement with nicotine patches of varying dosages (Dale, et al.,1995). We will treat the explanatory variable, dose, as ordinal, and we will treat the symptom 'feeling of exhaustion' as an ordinal response variable with two levels (absent/mild, moderate/severe). The numbers of subjects falling in each combination of levels are given in Table 5.15.

Nicotine	Feeling of Exhaustion		
Dose	Absent/Mild	Moderate/Severe	Total
Placebo	16	2	18
$11mg$	16	2	18
$22mg$	13	4	17
$44mg$	14	4	18
Total	59	12	71

Table 5.15: Numbers of subjects within each dose/symptom status combination

Concordant pairs are pairs where one subject scores higher on each variable than the other subject. Thus, all subjects in the $44mg$ dose group who had moderate/severe symptoms are concordant with all subjects who received less than $44mg$ and had absent/mild symptoms. Similarly, all subjects in the $22mg$ dose group who had moderate/severe symptoms are concordant with all subjects who received less than $22mg$ and had absent/mild symptoms. Finally, all subjects in the $11mg$ dose group who had moderate/severe symptoms are concordant with the subjects who

received the placebo and had absent/mild symptoms. Thus, the total number of concordant pairs (C) is:

$$C = 4(16 + 16 + 13) + 4(16 + 16) + 2(16) = 180 + 128 + 32 = 340$$

Discordant pairs are pairs where one subject scores higher on one variable, but lower on the other variable than the other subject. Thus, all subjects in the $44mg$ dose group who had absent/mild symptoms are discordant with all subjects who received less than $44mg$ and had moderate/severe symptoms. Similarly, all subjects in the $22mg$ dose group who had absent/mild symptoms are discordant with all subjects who received less than $22mg$ and had moderate/severe symptoms. Finally, all subjects in the $11mg$ dose group who had absent/mild symptoms are discordant with all subjects who received the placebo and had moderate/severe symptoms. Thus, the total number of discordant pairs (D) is:

$$D = 14(2 + 2 + 4) + 13(2 + 2) + 16(2) = 112 + 52 + 32 = 196$$

Notice that there are more concordant pairs than discordant pairs. This is consistent with more adverse effects at higher doses.

Two commonly reported measures of ordinal association are *gamma* and *Kendall's* τ_b. Both of these measures lie between -1 and 1. Negative values correspond to negative association, and positive values correspond to positive association. These types of association were described previously. A value of 0 implies no association between the two variables. Here, we give the formulas for the point estimates, their standard errors are better left to computers to handle. Tests of hypothesis and confidence intervals for the population measure are easily obtained from large–samples.

The point estimates for *gamma* and *Kendall's* τ_b are:

$$\hat{\gamma} = \frac{C - D}{C + D} \qquad \hat{\tau}_b = \frac{C - D}{0.5\sqrt{(n^2 - \sum n_{i.}^2)(n^2 - \sum n_{.j}^2)}}$$

To conduct a large–sample test of whether or not the population parameter is 0 (that is, a test of association between the explanatory and response variables), we complete the following steps:

1. $H_0 : \gamma = 0$ (No association)
2. $H_A : \gamma \neq 0$ (Association exists)
3. T.S. $z_{obs} = \frac{\hat{\gamma}}{\text{std. error}}$
4. R.R.:$|z_{obs}| \geq z_{\alpha/2}$
5. p–value:$2P(z \geq |z_{obs}|)$

For a test concerning *Kendall's* τ_b, replace γ with τ_b. For a $(1 - \alpha)100\%$ CI for the population parameter, simply compute (this time we use τ_b):

$$\hat{\tau}_b \pm z_{\alpha/2}(\text{std. error})$$

Example 5.10 For the data from Example 5.9, we can obtain estimates of *gamma* and *Kendall's*

τ_b, from Table 5.15 and the calculated values C and D.

$$\hat{\gamma} = \frac{C-D}{C+D} = \frac{340-196}{340+196} = \frac{144}{536} = 0.269$$

$$\hat{\tau}_b = \frac{C-D}{0.5\sqrt{(n^2 - \sum n_{i.}^2)(n^2 - \sum n_{.j}^2)}}$$

$$= \frac{340-196}{0.5\sqrt{((71)^2 - ((18)^2 + (18)^2 + (17)^2 + (18)^2))((71)^2 - ((59)^2 + (12)^2))}}$$

$$\frac{144}{0.5\sqrt{(3780)(1416)}} = \frac{144}{1156.8} = 0.124$$

From a statistical computer package, we get estimated standard errors of 0.220 and 0.104, respectively. We will test $H_0 : \gamma = 0$ vs $H_A : \gamma \neq 0$ at $\alpha = 0.05$ and compute a 95% CI for τ_b.

1. $H_0 : \gamma = 0$ (No association)
2. $H_A : \gamma \neq 0$ (Association exists)
3. T.S. $z = \frac{\hat{\gamma}}{\text{std. error}} = \frac{0.269}{0.220} = 1.22$
4. R.R.:$|z| \geq z_{\alpha/2} = 1.96$

Since the test statistic does not fall in the rejection region, we cannot conclude that there is an association between dose and feeling of exhaustion (note that this is a relatively small sample, so this test has little *power* to detect an association).

A 95% CI for τ_b can be computed as:

$$\hat{\tau}_b \pm z_{\alpha/2}(\text{std. error}) \equiv 0.124 \pm 1.96(0.104) \equiv 0.124 \pm 0.204 \equiv (-0.080, 0.328)$$

The interval contains 0 (which implies no association), so again we cannot conclude that increased dose implies increased feeling of exhaustion in a population of nicotine patch users.

5.4 Nominal Explanatory and Ordinal Response Variable

In the case where we have an explanatory variable that is nominal and an ordinal response variable, we use an extension of the Wilcoxon Rank Sum test that was described in Section 3.5.1. This involves ranking the subjects from smallest to largest in terms of the measurement of interest (there will be many ties), and compute the rank–sum (T_i) for each level of the nominal explanatory variable (typically a treatment group). We then compare the mean ranks among the r groups by the following procedure, known as the **Kruskal–Wallis Test**.

1. H_0 : The probability distributions of the ordinal response variable are the same for each level of the explanatory variable (treatment group). (No association).
2. H_A : The probability distributions of the response variable are the not same for each level of the explanatory variable. (Association).

3. T.S.: $H = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{T_i^2}{n_i} - 3(n+1)$.

4. R.R.: $H > \chi^2_{\alpha,r-1}$, where $\chi^2_{\alpha,\nu}$ is given in Table A.3, for various ν and α.

It should be noted that there is an adjustment for the ties that can be computed, but we will not cover that here (see Hollander and Wolfe (1973), p.140).

Example 5.11 A study was conducted to compare $r = 3$ methods of delivery of antibiotics in patients with lower respiratory tract infection (Chan, et al.,1995). The three modes of delivery were:

1. oral ($375mg$) co–amoxiclav three times a day for 7 days

2. intravenous ($1.2g$) co–amox three times a day for 3 days followed by oral ($375mg$) co–amox three times a day for 4 days

3. intravenous ($1g$) cefotaxime three times a day for 3 days followed by oral ($500mg$) cefuroxime axetil twice a day for 4 days

Outcome was ordinal: death, antibiotic changed, antibiotic extended, partial cure, cure. Table 5.16 contains the numbers of patients in each drug delivery/outcome category, the ranks, and the rank sums for each method of delivery.

	Therapeutic Outcome					
Method of Delivery (i)	Death	Antibiotic Changed	Antibiotic Extended	Partial Cure	Cure	Rank Sum (T_i)
1 ($n_1 = 181$)	9	14	16	68	74	51703.0
2 ($n_2 = 181$)	13	18	21	66	63	47268.5
3 ($n_3 = 179$)	11	16	30	53	69	47639.5
Ranks	1–33	34–81	82–148	149–335	336–541	
Avg. Rank	17.0	57.5	115.0	242.0	438.5	

Table 5.16: Data and ranks for antibiotic delivery data ($n = n_1 + n_2 + n_3 = 541$)

To obtain T_1, the rank sum for the subjects on oral co–amox, note that 9 of them received the rank of 17.0 (the rank assigned to each death), 14 received the rank of 57.5, etc. That is, $T_1 = 9(17.0) + 14(57.5) + 16(115.0) + 68(242.0) + 74(438.5)$. Here, we will test whether (H_0) or not (H_A) the distributions of therapeutic outcome differ among the three modes of delivery (as always, we test at $\alpha = 0.05$). The test statistic is computed as follows:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{T_i^2}{n_i} - 3(n+1) = \frac{12}{541(542)} \left(\frac{(51703.0)^2}{181} + \frac{(47268.5)^2}{181} + \frac{(47639.5)^2}{179} \right) - 3(542) =$$

$$\frac{12}{541(542)}(14769061.9 + 12344260.2 + 12678893.6) - 1626 = 1628.48 - 1626 = 2.48$$

The rejection region is $H \geq \chi^2_{.05,3-1} = \chi^2_{.05,2} = 5.99$. Since our test statistic does not fall in the rejection region, we cannot reject H_0, we have no evidence that the distributions of therapeutic outcomes differ among the three modes of delivery. The authors stated that since there appears to be no differences among the outcomes, the oral delivery mode would be used since it is simplest to perform.

5.5 Assessing Agreement Among Raters

As mentioned in Chapter 1, in many situations the response being measured is an assessment made by an investigator. For instance, in many trials, the response may be the change in a patient's condition, which would involve rating a person along some sort of Likert (ordinal) scale. A patient may be classified as: dead, condition much worse, condition slightly/moderately worse, ..., condition much improved. Unfortunately measurements such as these are much more subjective than measures such as time to death or blood pressure. In many instances, a pair (or more) of raters may be used, and we would like to assess the level of their agreement.

A measure of agreement that was developed in psychiatric diagnosis is **Cohen's** κ (Spitzer, et al,1967). It measures the proportion of agreement beyond chance agreement. It can take on negative values when the agreement is worse than expected by chance, and the largest value it can take is 1.0, which occurs when there is perfect agreement. While κ only detects disagreement, a modification, called **weighted** κ distinguishes among levels of disagreement (e.g. raters who disagree by one category are in stronger agreement than raters who differ by several categories).

We will illustrate the computation of κ with a non–medical example, the reader is referred to (Spitzer, et al,1967), and its references for computation of weighted κ.

Example 5.12 A study compared the level of agreement among popular movie critics (Agresti and Winner,1997). The pairwise levels of agreement among 8 critics (Gene Siskel, Roger Ebert, Michael Medved, Jeffrey Lyons, Rex Reed, Peter Travers, Joel Siegel, and Gene Shalit) was computed. In this example, we will focus on Siskel and Ebert. There were 160 movies that both critics reviewed during the study period, the results are given in Table 5.17, which is written as a 3×3 contingency table.

Siskel	Ebert Rating			
Rating	Con	Mixed	Pro	Total
Con	24	8	13	45
	(.150)	(.050)	(.081)	(.281)
	(.074)	(.053)	(.155)	—
Mixed	8	13	11	32
	(.050)	(.081)	(.069)	(.200)
	(.053)	(.038)	(.110)	—
Pro	10	9	64	83
	(.063)	(.056)	(.400)	(.519)
	(.136)	(.098)	(.285)	—
Total	42	30	88	160
	.263	.188	.550	1.00

Table 5.17: Ratings on $n = 160$ movies by Gene Siskel and Roger Ebert – raw counts, observed proportions, and proportions expected under chance

If their ratings were independent (that is, knowledge of Siskel's rating gives no information as to what Ebert's rating on the same movie), we would expect the following probabilities along the main diagonal (where the critics agree):

$$p_{11} = P(\text{Con}|\text{Siskel}) \cdot Pr(\text{Con}|\text{Ebert}) = (.281)(.263) = .074$$

$$p_{22} = P(\text{Mixed}|\text{Siskel}) \cdot Pr(\text{Mixed}|\text{Ebert}) = (.200)(.188) = .038$$

$$p_{33} = P(\text{Pro}|\text{Siskel}) \cdot Pr(\text{Pro}|\text{Ebert}) = (.281)(.263) = .285$$

So, even if their ratings were independent, we would expect the proportion of movies that they would agree on by chance to be $p_c = .074 + .038 + .285 = .397$. That is, we would expect them to agree about 40% of the time, based on their marginal distributions. In fact, the observed proportion of movies for which they agree on is $p_o = .150 + .081 + .400 = .631$, so they agree on about 63% of the movies. We can now compute Cohen's κ:

$$\kappa = \frac{\text{observed agreement – chance agreement}}{1 \text{ – chance agreement}} = \frac{.631 - .397}{1 - .397} = \frac{.234}{.603} = .388$$

This would be considered a moderate level of agreement. The sample difference between the observed agreement and the agreement expected under independence is 39% of the maximum possible difference.

5.6 Exercises

23. Coronary–artery stenting, when conducted with coronary angioplasty has negative side effects related to anyicoagulant therapy. A study was conducted to determine whether or not use of antiplatelet therapy produces better results than use of anticoagulants (Schiömig, et al.,1996). Patients were randomized to receive either anticoagulant or antiplatelet therapy, and classified by presence or absence of primary cardiac endpoint, where death by cardiac causes, MI, aortocoronary bypass, or repeated PTCA of stented vessel constituted an event. In the randomized study, patients received either aniplatelet ($n_1 = 257$) or anticoagulant therapy. Results, in terms of numbers of patients suffering a primary cardiac endpoint for each therapy are given in Table 5.18.

(a) What proportion of patients on antiplatelet therapy suffer a primary cardiac endpoint (this is denoted $\hat{\pi}_E$, as we will consider the antiplatelet group as 'exposed' in relative risk calculations)?

(b) What proportion of patients on anticoagulant therapy suffer a primary cardiac endpoint (this is denoted $\hat{\pi}_{\overline{E}}$, as we will consider the anticoagulant group as 'unexposed' in relative risk calculations)?

(c) Compute the Relative Risk of suffering a primary cardiac endpoint for patients receiving antiplatelet therapy, relative to patients receiving anticoagulant therapy?

(d) Compute and interpret the 95% CI for the population relative risk.

(e) By how much does using antiplatelet therapy reduce risk of a primary cardiac endpoint compared to using anticoagulant therapy?

		Occurrence of Primary Cardiac Event		
		Yes (D)	No ($\overline{D}$)	
Treatment Group	Antiplatelet (E)	4	253	257
	Anticoagulant ($\overline{E}$)	16	244	260
		20	497	517

Table 5.18: Observed cell counts for antiplatelet/anticoagulant data

24. The results of a multicenter clinical trial to determine the safety and efficacy of the pancreatic lipase inhibitor, Xenical, was reported (Ingersoll, 1997). Xenical is used to block the absorption of dietary fat. The article reported that more than 4000 patients in the U.S. and Europe were randomized to receive Xenical or a placebo in a parallel groups study. After one year, 57% of those receiving Xenical had lost at least 5% of their body weight, as opposed to 31% of those receiving a placebo. Assume that exactly 4000 patients were in the study, and that 2000 were randomized to receive a placebo and 2000 received Xenical. Test whether or not the drug can be considered effective at the $\alpha = 0.05$ significance level by computing a 95% confidence interval for the "relative risk" of losing at least 5% of body weight for those receiving Xenical relative to those receiving placebo.

25. A case–control study of patients on antihypertensive drugs related an increased risk of myocardial infarction (MI) for patients using calcium channel blockers (Psaty, et al.,1995). In this study, cases

were antihypertensive drug patients who had suffered a first fatal or nonfatal MI through 1993, and controls were antihypertensive patients, matched by demographic factors, who had not suffered a MI. Among the comparisons reported were patients receiving calcium channel (CC) blockers (with and without diuretics) and patients receiving β–blockers (with and without diuretics). Results of numbers of patient by drug/MI status combination are given in Table 5.19. Compute the odds ratio of suffering MI (CC blockers relative to β–blockers), and the corresponding 95% CI. Does it appear that calcium channel blockers are associated with higher odds (and thus probability) of suffering MI than β–blockers?

		Occurrence of Myocardial Infarction		
		Yes (D)	No ($\overline{D}$)	
Antihypertensive Drug	CC blocker (E)	80	230	310
	β–blocker ($\overline{E}$)	85	395	480
		165	625	790

Table 5.19: Observed cell counts for antihypertensive drug/MI data

26. A Phase III clinical trial generated the following results in terms of efficacy of the cholesterol reducing drug pravastatin in men with high cholesterol levels prior to treatment (Shepherd, et al., 1995). A sample of $n = 6595$ men from ages 45 to 64 were randomized to receive either pravastatin or placebo. The men were followed for an average of 4.9 years, and were classified by presence or absence of the primary endpoint: nonfatal MI or death from CHD. The results are given in Table 5.20. Compute the relative risk of suffering nonfatal MI or death from CHD (pravastatin relative to placebo), and the corresponding 95% CI. Does pravastatin appear to reduce the risk of nonfatal MI or death from CHD? Give a point estimate, and 95% confidence interval for the percent reduction in risk.

	Nonfatal MI or Death from CHD		
	Present (D)	Absent ($\overline{D}$)	
Pravastatin (E)	174	3128	3302
Placebo ($\overline{E}$)	248	3045	3293
	422	6173	6595

Table 5.20: Observed cell counts for pravastatin efficacy trial

27. In Lister's study of the effects of antiseptic in amputations, he stated that amputations in the upper limb were quite different, and that in these cases "if death does occur, it is commonly the result of the wound assuming unhealthy characters" (Lister, 1870). Thus, he felt that the best way do determine antiseptic's efficacy was to compare the outcomes of upper limb surgeries separately. The results are given in Table 5.21.

(a) Given there were 7 deaths, and 12 people in the antiseptic group and 12 in the control group,

		Surgical Outcome: Death	Surgical Outcome: No Death	
Treatment Group	Antiseptic (E)	1	11	12
	Control ($\overline{E}$)	6	6	12
		7	17	24

Table 5.21: Observed cell counts for antiseptic data – Upper limb cases

write out the two tables that provide as strong or stronger evidence of antiseptic's effect (hint: this table is one of them).

(b) Under the hypothesis of no antiseptic effect, the combined probability of the correct two tables from part a) being observed is .034. If we use Fisher's exact test with $\alpha = 0.05$, do we conclude that there is an antiseptic effect in terms of reducing risk of death from amputations? What is the lowest level of α for which we will reject H_0?

28. A study was conducted to compare the detection of genital HIV-1 from tampon eluents with cervicovaginal lavage (CVL) and plasma specimens in women with HIV-1 (Webber, et al (2001)). Full data were obtained from 97 women. Table 5.22 has the numbers of women testing positive and negative based on tampon eluents and CVL (both tests are conducted on each of the women). Test whether the probabilities of detecting HIV-1 differ based on tampons versus CVL at the $\alpha = 0.05$ significance level.

		Tampon: Positive	Tampon: Negative	
CVL	Positive	23	19	42
	Negative	10	45	55
		33	64	97

Table 5.22: Detection of HIV-1 via CVL and Tampons in Women with HIV-1

29. A case–control study was reported on a population–based sample of renal cell carcinoma patients (cases), and controls who did not suffer from the disease (McLaughlin, et al.,1984). Among the factors reported was the ethnicity of the individuals in the study. Table 5.23 contains the numbers of cases and controls by each of 7 ethnicities (both parents from that ethnic background). Use Pearson's chi–squared test to determine whether or not there is an association between ethnic background and occurrence of renal cell carcinoma (first, complete the table by computing the expected cell counts under the null hypothesis of no association for the Scandinavians).

30. A survey was conducted among pharmacists to study attitudes toward shifts from prescription to over–the–counter status (Madhavan, 1990). Pharmacists were asked to judge the appropriateness of switching to OTC dor each of the three drugs: promethazine, terfenadine, and naproxen. Results were operationalized to classify each pharmacist into one of two switch judgment groups (yes/no).

Ethnicity	Cancer	No Cancer	Total
German	60 (59.0) $(60-59.0)^2/59.0 = .017$	64 (65.0) $(64-65.0)/65.0 = .015$	124
Irish	17 (14.7) $(17-14.7)^2/14.7 = .360$	14 (16.3) $(14-16.3)^2/16.3 = .325$	31
Swedish	22 (25.7) $(22-25.7)^2/25.7 = .533$	32 (28.3) $(32-28.3)^2/28.3 = .484$	54
Norwegian	23 (20.9) $(23-20.9)^2/20.9 = .211$	21 (23.1) $(21-23.1)^2/23.1 = .191$	44
Czech	6 (6.2) $(6-6.2)^2/6.2 = .006$	7 (6.8) $(7-6.8)^2/6.8 = .006$	13
Russian	4 (4.8) $(4-4.8)^2/4.8 = .133$	6 (5.2) $(6-5.2)^2/5.2 = .123$	10
Scandinavian	63 ( )	71 ( )	134
Total	195	215	410

Table 5.23: Observed (expected) values of numbers of subjects within each ethnicity/cancer status group and chi–square test stat contribution

Results are given in Table 5.24. Conduct a chi–square test ($\alpha = 0.05$) to determine whether there is an association between experience($\leq 15/ \geq 16$ years). If an association exists, which group is has a higher fraction of pharmacists favoring the switch to OTC status.

Experience	No OTC Switch	OTC Switch	Total
≤ 15 years	28 (38.7)	50 (39.3)	78
≥ 16 years	46 (35.3)	25 (—)	71
Total	75	74	149

Table 5.24: Observed (expected) values of numbers of subjects within each experience/OTC switch status group

31. In a review of studies relating smoking to drug metablism, the side effect of drowsiness (absent/present) and smoking status (non/light/heavy) were reported in a study of 1214 subjects receiving diazepam (Dawson and Vestal,1982). The numbers of subjects falling into each combination of these ordinal variables is given in Table 5.25.

Treating each variable as ordinal, we can obtain the numbers of concordant pairs (where one person scores higher on both variables than the other) and discordant pairs (where one scores higher on smoking, and the other scores higher on drowsiness) of subjects. The numbers of concordant and discordant pairs are:

$$C = 5(359 + 593) + 30(593) = 22550 \qquad D = 176(30 + 51) + 359(51) = 32565$$

Smoking Status	Drowsiness		
	Absent	Present	Total
Nonsmokers	593	51	644
Light Smokers	359	30	389
Heavy Smokers	176	5	181
Total	86	1128	1214

Table 5.25: Numbers of subjects within each smoking/drowsy status combination

(a) Compute $\hat{\gamma}$.

(b) The std. error of $\hat{\gamma}$ is $\hat{\sigma}_{\hat{\gamma}} = .095$. Test $H_0 : \gamma = 0$ vs $H_A : \gamma \neq 0$ at $\alpha = 0.05$ significance level. Does there appear to be an association between drowsiness and smoking status?

(c) $\hat{\tau}_b = -0.049$ and $\hat{\sigma}_{\hat{\tau}_b} = 0.025$. Compute a 95% CI for the population measure of association and interpret it.

32. A randomized trial was conducted to study the effectiveness of intranasal ipratropium bromide against the common clod (Hayden, et al, 1996). Patients were randomized to receive one of three treatments: intranasal ipratropium, vehicle control, or no treatment. Patients assessed the overall treatment effectiveness as one of three levels: much better, better, or no difference/worse. Outcomes for day 1 are given in Table 5.26. We will treat both of these variables as ordinal.

Treatment Group	Effectiveness			
	No Diff/Worse	Better	Much Better	Total
No Treatment	58	73	5	136
Vehicle Control	37	82	18	137
Ipratropium	18	86	33	137
Total	113	244	56	

Table 5.26: Numbers of subjects within each treatment/effectiveness combination

(a) Compute the number of concordant and discordant pairs. Treat the ipratropium group as the high level for treatment and much better for the high level of effectiveness.

(b) Compute $\hat{\gamma}$.

(c) The estimated standard error of $\hat{\gamma}$ is .059. Can we conclude that there is a positive association between treatment group and effectiveness at the $\alpha = 0.05$ significance level based on this measure?

(d) Compute $\hat{\tau}_B$.

(e) The estimated standard error of $\hat{\tau}_B$ is .039. Compute a 95% confidence interval for the population value. Based on the interval, can we conclude that there is a positive association between treatment group and effectiveness at the $\alpha = 0.05$ significance level?

33. A study designed to determine the effect of lowering cholesterol on mood state was conducted in a placebo controlled parallel groups trial (Wardle, et al.,1996). Subjects between the ages of 40 and 75 were assigned at random to receive simvastatin, an HMG CoA reductase inhibitor, or a placebo. Subjects were followed an average of 3 years, and asked to complete the profile of mood states (POMS) questionnaire. Since some previous studies had shown evidence of an association between low cholesterol, they compared the active drug group with the placebo group in POMS scores for several scale. Table 5.27 gives the numbers of subjects for each treatment group falling in $k = 7$ ordinal categories for the fatigue/inertia scale (high scores correspond to high fatigue). Compute the rank sums, and test whether or not the distributions of the POMS scores differ among the treatment groups ($\alpha = 0.05$), using the Kruskal–Wallis test. Is there any evidence that subjects with lower cholesterol (simvastatin group) tend to have higher levels of fatigue than the control group?

	Profile of Mood States (POMS) Score							
Trt Group	0	1–4	5–8	9–12	13–16	17–20	21–24	Rank Sum (T_i)
Simvastatin ($n_1 = 334$)	23	98	98	58	40	10	7	
Placebo ($n_2 = 157$)	8	43	56	28	8	11	3	
Ranks	1–31	32–172	173–326	327–412	413–460	461–481	482–491	
Avg. Rank	16.0	102.0	249.5	369.5	436.5	471.0	486.5	

Table 5.27: Data and ranks for cholesterol drug/fatigue data ($n = n_1 + n_2 = 491$)

34. In the paper, studying agreement among movie reviewers, the following results were obtained for Michael Medved and Jeffrey Lyons, formerly of *Sneak Previews* (Agresti and Winner,1997). The following table gives the observed frequencies, observed proportions, and expected proportions under chance. Compute and interpret Cohen's κ for Table 5.28.

Lyons	Medved Rating			
Rating	Con	Mixed	Pro	Total
Con	22	7	8	37
	(.179)	(.057)	(.065)	(.301)
	(.117)	(.078)	(.105)	—
Mixed	5	7	7	19
	(.041)	(.057)	(.057)	(.154)
	(.060)	(.040)	(.054)	—
Pro	21	18	28	67
	(.171)	(.146)	(.228)	(.545)
	(.213)	(.142)	(.191)	—
Total	48	32	43	123
	.390	.260	.350	1.00

Table 5.28: Ratings on $n = 123$ movies by Michael Medved and Jeffrey Lyons – raw counts, observed proportions, and proportions expected under chance

Chapter 6

Experimental Design and the Analysis of Variance

In previous chapters, we have covered methods to make comparisons between the means of a numeric response variable for two treatments. We have seen the case where the experiment was conducted as a parallel groups design, as well as a crossover design. Further, we have used procedures that assume normally distributed data, as well as nonparametric methods that can be used when data are not normally distributed.

In this chapter, we will introduce methods that can be used to compare more than two groups (that is, when the explanatory variable has more than two levels). In this chapter, we will refer to explanatory variables as **factors**, and their levels as **treatments**. We will cover the following situations:

- 1–Factor, Parallel Groups Designs
- 1–Factor, Crossover Design Designs
- 2–Factor, Parallel Groups Designs
- Crossover Designs With Sequence and Period Effects
- Parallel Groups Repeated Measures Designs

In all situations, we will have a numeric response variable, and at least one categorical (or numeric, with several levels) independent variable. The goal will always be to compare mean (or median) responses among several populations.

6.1 Completely Randomized Design (CRD) For Parallel Groups Studies

In the Completely Randomized Design, we have one factor that we are controlling. This factor has k levels (which are often treatment groups), and we measure n_i units on the i^{th} level of the factor.

We will define the observed responses as Y_{ij}, representing the measurement on the j^{th} experimental unit (subject), receiving the i^{th} treatment. We will write this in model form as follows:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}.$$

Here, μ is the overall mean measurement across all treatments, α_i is the effect of the i^{th} treatment ($\mu_i = \mu + \alpha_i$), and ε_{ij} is a random error component that has mean 0 and variance σ^2. This ε_{ij} can be thought of as the fact that there will be variation among the measurements of different subjects receiving the same treatment.

We will place a condition on the effects α_i, namely that they sum to zero. Of interest to the experimenter is whether or not there is a **treatment effect**, that is do any of the levels of the treatment provide higher (lower) mean response than other levels. This can be hypothesized symbolically as $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ (no treatment effect) against the alternative $H_A :$ Not all $\alpha_i = 0$ (treatment effects exist).

As with the case where we had two treatments to compare, we have a test based on the assumption that the k populations are normal (mound–shaped), and a second test (based on ranks) that does not assume that the k populations are normal. However, these methods do assume common spreads (standard deviations) within the k populations.

6.1.1 Test Based on Normally Distributed Data

When the underlying populations of measurements that are to be compared are approximately normal, we conduct the F–test. To conduct this test, we partition the total variation in the sample data to variation **within** and **among** treatments. This partitioning is referred to as the **analysis of variance** and is an important tool in many statistical procedures. First, we will define the following items:

$$\begin{aligned}
\overline{y}_i &= \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \\
s_i &= \sqrt{\frac{\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2}{n_i - 1}} \\
n &= n_1 + \cdots + n_k \\
\overline{y} &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}{n} \\
TotalSS &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 \\
SST &= \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2 \\
SSE &= \sum_{i=1}^{k}(n_i - 1)s_i^2
\end{aligned}$$

Here, $\overline{y}_i$ and s_i are the mean and standard deviation of measurements in the i^{th} treatment group, and $\overline{y}$ and n are the overall mean and total number of *all* measurements. $TotalSS$ is the total

variability in the data (ignoring treatments), SST measures the variability in the sample means among the treatments, and SSE measures the variability within the treatments. In these last terms, SS represents *sum of squares*.

Note that we are trying to determine whether or not the population means differ. If they do, we would expect SST to be large, since that sum of squares is picking up differences in the sample means. We will be able to conduct a test for treatment effects after setting up an Analysis of Variance table, as shown in Table 6.1. In that table, we have the *sums of squares* for treatments (SST), for error (SSE), and total ($TotalSS$). Also, we have *degrees of freedom*, which represents the number of "independent" terms in the sum of squares. Then, we have *mean squares*, which are sums of squares divided by their degrees of freedom. Finally, the F–statistic is computed as $F = MST/MSE$. This will serve as our test statistic. While this may look daunting, it is simply a general table that can be easily computed and used to test for treatment effects. Note that MSE is an extension of the pooled variance we computed in Chapter 3 for two groups, and often we see that $MSE = s^2$.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
TREATMENTS	$SST = \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2$	$k-1$	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE}$
ERROR	$SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2$	$n-k$	$MSE = \frac{SSE}{n-k}$	
TOTAL	$TotalSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$	$n-1$		

Table 6.1: The Analysis of Variance Table for the Completely Randomized (Parallel Groups) Design

Recall the model that we are using to describe the data in this design:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}.$$

The effect of the i^{th} treatment is α_i. If there is no treatment effect among any of the levels of the factor under study, that is if the population means of the k treatments are the same, then each of the parameters α_i are 0. This is a hypothesis we would like to test. The alternative hypothesis will be that not all treatments have the same mean, or equivalently, that treatment effects exist (not all α_i are 0). If the null hypothesis is true (all k population means are equal), then the statistic $F_{obs} = \frac{MST}{MSE}$ follows the F-distribution with $k-1$ numerator and $n-k$ denominator degrees of freedom. Large values of F_{obs} are evidence against the null hypothesis of no treatment effect (recall what SST and SSE are). The formal method of testing this hypothesis is as follows.

1. $H_0 : \alpha_1 = \cdots = \alpha_k = 0 \quad (\mu_1 = \cdots = \mu_k)$ (No treatment effect)
2. $H_A :$ Not all α_i are 0 (Treatment effects exist)
3. T.S. $F_{obs} = \frac{MST}{MSE}$
4. R.R.: $F_{obs} > F_{\alpha,k-1,n-k}$ critical values of the F–distribution are given in Table A.4.

5. p-value: $P(F \geq F_{obs})$

Example 6.1 A randomized clinical trial was conducted to observe the safety and efficacy of a three–drug combination in patients with HIV infection (Collier, et al., 1996). Patients were assigned at random to one of three treatment groups: saquinavir, zidovudine, zalcitabine (SZZ); saquinivir, zidovudine (SZ), or zidovudine, zalcitabine (ZZ). One of the numeric measures made on patients was their normalized area under the log–transformed curve for the $CD4+$ count from day 0 to day 24. Positive values imply increasing $CD4+$ counts (relative to baseline), and negative values imply decreasing $CD4+$ counts. We would like to compare the three treatments, and in particular show that the three–drug treatment is better than either of the two two–drug treatments. First, however, we will simply test whether there are treatment effects. Summary statistics based on the normalized area under the log–transformed $CD4+$ count curves at week 24 of the study are given in Table 6.2. The Analysis of Variance is given in Table 6.3. Note that we have $k = 3$ treatments and $n = 270$ total measurements (90 subjects per treatment). We will test whether or not the three means differ at $\alpha = 0.05$.

	Trt 1 (SZZ)	Trt 2 (SZ)	Trt 3 (ZZ)
Mean	$\overline{y}_1 = 12.2$	$\overline{y}_2 = 5.1$	$\overline{y}_3 = -0.3$
Std Dev	$s_1 = 18.97$	$s_2 = 19.92$	$s_3 = 20.87$
Sample Size	$n_1 = 90$	$n_2 = 90$	$n_3 = 90$

Table 6.2: Sample statistics for sequanavir study in HIV patients

ANOVA				
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
TREATMENTS	7074.9	2	3537.5	$F = \frac{3537.5}{397.5} = 8.90$
ERROR	106132.5	267	397.5	
TOTAL	113207.4	269		

Table 6.3: The Analysis of Variance table for the sequanavir study in HIV patients

1. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad (\mu_1 = \mu_2 = \mu_3)$ (No treatment effect)
2. $H_A :$ Not all α_i are 0 (Treatment effects exist)
3. T.S. $F_{obs} = \frac{MST}{MSE} = 8.90$
4. R.R.: $F_{obs} > F_{\alpha,k-1,n-k} = F_{0.05,2,267} = 3.03$
5. p-value: $P(F \geq F_{obs}) = P(F \geq 8.90) = .0002$

Since, we do reject H_0, we can conclude that the means differ, now we will describe methods to make pairwise comparisons among treatments.

Comparison of Treatment Means

Assuming that we have concluded that treatment means differ, we generally would like to know which means are significantly different. This is generally done by making either pre–planned or all pairwise comparisons between pairs of treatments. We will look at how to make comparisons for each treatment with a control, and then how to make all comparisons. The three methods are very similar.

Dunnett's Method for Comparing Treatments With a Control

In many situations, we'd like to compare each treatment with the control (when there is a natural control group). Here, we would like to make all comparisons of treatment vs control ($k-1$, in all) with an overall confidence level of $(1-\alpha)100\%$. If we arbitrarily label the control group as treatment 1, we want to obtain simultaneous confidence intervals for $\mu_i - \mu_1$ for $i = 2, \ldots, k$. Based on each confidence interval, we can determine whether the treatment differs from the control by determining whether or not 0 is included in the interval. The general form of the confidence intervals is:

$$(\overline{y}_i - \overline{y}_1) \pm d_{\alpha,k-1,n-k}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_1}\right)},$$

where $d_{\alpha,k-1,n-k}$ is given in tables of various statistical texts (see Montgomery (1991)). We will see an application of Dunnett's method in Chapter 10.

Bonferroni's Method of Multiple Comparisons

Bonferroni's method is used in many situations and is based on the following premise: If we wish to make c comparisons, and be $(1-\alpha)100\%$ confident they are all correct, we should make each comparison at a higher level of confidence (lower probability of type I error). If we make each comparison at α/c level of significance, we have an overall error rate no larger than α. This method is conservative and can run into difficulties (low power) as the number of comparisons increases. The general procedure is to compute the c intervals as follows:

$$(\overline{y}_i - \overline{y}_j) \pm t_{\alpha/2c,n-k}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

where $t_{\alpha/2c,n-k}$ is obtained from the t–table. When exact values of $\alpha/2c$ are not available from the table, the next lower α (higher t) value is used.

Tukey's Method for All Pairwise Comparisons

The previous method described works well when comparing various treatments with a control. Various methods have been developed to handle all possible comparisons and keep the overall error rate at α, including the widely reported Bonferroni procedure described above. Another commonly used procedure is Tukey's method, which is more powerful than the Bonferroni method (but more limited in its applicability). Computer packages will print these comparisons automatically. Tukey's method involves setting up confidence intervals for all pairs of treatment means simultaneously. If there are k treatments, their will be $\frac{k(k-1)}{2}$ such intervals. The general form, allowing for different sample sizes for treatments i and j is:

$$(\overline{y}_i - \overline{y}_j) \pm q_{\alpha,k,n-k}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)/2},$$

where $q_{\alpha,k,n-k}$ is called the studentized range and is given in tables in many text books (see Montgomery (1991)). When the sample sizes are equal ($n_i = n_j$), the formula can be simplified to:

$$(\overline{y}_i - \overline{y}_j) \pm q_{\alpha,k,n-k}\sqrt{MSE\left(\frac{1}{n_i}\right)}.$$

Example 6.2 In the sequanavir study described in Example 6.1, we concluded that treatment effects exist. We can now make pairwise comparisons to determine which pairs of treatments differ. There are three comparisons to be made: SZZ vs SZ, SZZ vs ZZ, and SZ vs ZZ. We will use Bonferroni's and Tukey's methods to obtain 95% CI's for each difference in mean area under the log–transformed $CD4+$ curve. The general form for Bonferroni's simultaneous 95% CI's is (with $c = 3$):

$$(\overline{y}_i - \overline{y}_j) \pm t_{\alpha/2c,n-k}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \quad (\overline{y}_i - \overline{y}_j) \pm t_{.0083,267}\sqrt{397.5\left(\frac{1}{90} + \frac{1}{90}\right)}$$

$$(\overline{y}_i - \overline{y}_j) \pm 2.41(2.97) \quad (\overline{y}_i - \overline{y}_j) \pm 7.16$$

For Tukey's method, the confidence intervals are of the form:

$$(\overline{y}_i - \overline{y}_j) \pm q_{\alpha,k,n-k}\sqrt{MSE\left(\frac{1}{n_i}\right)} \quad (\overline{y}_i - \overline{y}_j) \pm q_{0.05,3,267}\sqrt{397.5\left(\frac{1}{90}\right)}$$

$$(\overline{y}_i - \overline{y}_j) \pm 3.32(2.10) \quad (\overline{y}_i - \overline{y}_j) \pm 6.97$$

The corresponding confidence intervals are given in Table 6.4.

		Simultaneous 95% CI's	
Comparison	$\overline{y}_i - \overline{y}_j$	Bonferroni	Tukey
SZZ vs SZ	$12.2 - 5.1 = 7.1$	$(-0.06, 14.26)$	$(0.13, 14.07)$
SZZ vs ZZ	$12.2 - (-0.3) = 12.5$	$(5.34, 19.66)$	$(5.53, 19.47)$
SZ vs ZZ	$5.1 - (-0.3) = 5.4$	$(-1.76, 12.56)$	$(-1.57, 12.37)$

Table 6.4: Bonferroni and Tukey multiple comparisons for the sequanavir study in HIV patients

Based on the intervals in Table 6.4, we can conclude that patients under the three–drug treatment (SZZ) have higher means than those on either of the two two–drug therapies (SZ and ZZ), although technically, Bonferroni's method does contain 0 (just barely). This is a good example that Bonferroni's method is less powerful than Tukey's method. No difference can be detected between the two two–drug treatments. The authors did not adjust α for multiple comparisons (see p. 1012, **Statistical Analysis** section). This made it 'easier' to find differences, but increases their risk of declaring an ineffective treatment as being effective.

6.1.2 Test Based on Non–Normal Data

A nonparametric test for the Completely Randomized Design (CRD), where each experimental unit receives only one treatment, is the **Kruskal-Wallis Test** (Kruskal and Wallis,1952). The idea behind the test is similar to that of the Wilcoxon Rank Sum test. The main difference is that instead of comparing 2 population distributions, we are comparing $k > 2$ distributions. Sample measurements are ranked from 1 (smallest) to $n = n_1 + \cdots + n_k$ (largest), with ties being replaced with the means of the ranks the tied subjects would have received had they not tied. For each treatment, the sum of the ranks of the sample measurements are computed, and labelled T_i. The sample size from the i^{th} treatment is n_i, and the total sample size is $n = n_1 + \cdots + n_k$. We have previously seen this test in Chapter 5.

The hypothesis we wish to test is whether the k population distributions are identical against the alternative that some distribution(s) is (are) shifted to the right of other(s). This is similar to the hypothesis of no treatment effect that we tested in the previous section. The procedure is as follows:

1. H_0 : The k population distributions are identical ($\mu_1 = \mu_2 = \cdots = \mu_k$)
2. H_A : Not all k distributions are identical (Not all μ_i are equal)
3. T.S.: $H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$.
4. R.R.: $H > \chi^2_{\alpha,k-1}$
5. p–value: $P(\chi^2 \geq H)$

Note that each of the sample sizes n_i must be at least 5 for this procedure to be used.

If we do reject H_0, and conclude treatment differences exist, we could run the Wilcoxon Rank Sum test on all pairs of treatments, adjusting the individual α levels by taking α/c where c is the number of comparisons, so that the overall test (on all pairs) has a significance level of α. This is an example of Bonferroni's procedure.

Example 6.2 The use of thalidomide was studied in patients with HIV–1 infection (Klausner, et al.,1996). All patients were HIV–1^+, and half of the patients also had tuberculosis infection (TB^+). There were $n = 32$ patients at the end of the study, 16 received thalidomide and 16 received placebo. Half of the patients in each drug group were TB^+ (the other half TB^-), so we can think of this study having $k = 4$ treatments: TB^+/thalidomide, TB^+/placebo, TB^-/thalidomide, and TB^-/placebo. One primary measure was weight gain after 21 days. We would like to test whether or not the weight gains differ among the 4 populations. The weight gains (negative values are losses) and their corresponding ranks are given in Table 6.5, as well as the rank sum for each group.

We can test whether or not the weight loss distributions differ among the four groups using the Kruskal–Wallis test. We will conduct the test at the $\alpha = 0.05$ significance level.

1. H_0 : The 4 population distributions are identical ($\mu_1 = \mu_2 = \mu_3 = \mu_4$)
2. H_A : Not all 4 distributions are identical (Not all μ_i are equal)

Group (Treatment)			
TB^+/Thal ($i=1$)	TB^-/Thal ($i=2$)	TB^+/Plac ($i=3$)	TB^-/Plac ($i=4$)
9.0 (32)	2.5 (23)	0.0 (9)	–0.5 (7)
6.0 (31)	3.5 (26.5)	1.0 (15.5)	0.0 (9)
4.5 (30)	4.0 (28.5)	–1.0 (6)	2.5 (23)
2.0 (20.5)	1.0 (15.5)	–2.0 (4)	0.5 (12)
2.5 (23)	0.5 (12)	–3.0 (1.5)	–1.5 (5)
3.0 (25)	4.0 (28.5)	–3.0 (1.5)	0.0 (9)
1.0 (15.5)	1.5 (18.5)	0.5 (12)	1.0 (15.5)
1.5 (18.5)	2.0 (20.5)	–2.5 (3)	3.5 (26.5)
$T_1 = 195.5$	$T_2 = 173.0$	$T_3 = 52.5$	$T_4 = 107.0$

Table 6.5: 21–day weight gains in kg (and ranks) for thalidomide study in HIV–1 patients

3. T.S.: $H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1) = \frac{12}{32(33)} \left(\frac{(195.5)^2}{8} + \frac{(173.0)^2}{8} + \frac{(52.5)^2}{8} + \frac{(107.0)^2}{8} \right) - 3(33) = 116.98 - 99 = 17.98$.
4. R.R.: $H \geq \chi^2_{\alpha,k-1} = \chi^2_{.05,3} = 7.815$
5. p–value: $P(\chi^2_3 \geq 17.98) = .0004$

We reject H_0, and conclude differences exist. Based on the high rank sums for the thalidomide groups, the drug clearly increases weight gain. Pairwise comparisons could be made using the Wilcoxon Rank Sum test. We could also combine treatments 1 and 2 as a thalidomide group and treatments 3 and 4 as a placebo group, and compare them using the Wilcoxon Rank Sum test.

6.2 Randomized Block Design (RBD) For Crossover Studies

In crossover designs, each subject receives each treatment. In these cases, subjects are referred to as **blocks**. The notation for the RBD is very similar to that of the CRD, with only a few additional elements. The model we are assuming here is:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} = \mu_i + \beta_j + \varepsilon_{ij}.$$

Here, μ represents the overall mean measurement, α_i is the effect of the i^{th} treatment, β_j is the effect of the j^{th} block, and ε_{ij} is a random error component that can be thought of as the variation in measurments if the same experimental unit received the same treatment repeatedly. Note that just as before, μ_i represents the mean measurement for the i^{th} treatment (across all blocks). The general situation will consist of an experiment with k treatments being received by each of b blocks.

6.2.1 Test Based on Normally Distributed Data

When the block effects (β_j) and random error terms (ε_{ij}) are independent and normally distributed, we can conduct an F–test similar to that described for the Completely Randomized Design. The

notation we will use is as follows:

$$\begin{aligned}
\overline{y}_{i.} &= \frac{\sum_{j=1}^{b} y_{ij}}{b} \\
\overline{y}_{.j} &= \frac{\sum_{i=1}^{k} y_{ij}}{k} \\
n &= b \cdot k \\
\overline{y} &= \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}}{n} \\
TotalSS &= \sum_{i=1}^{k} \sum_{j=1}^{b} (y_{ij} - \overline{y})^2 \\
SST &= \sum_{i=1}^{k} b(\overline{y}_{i.} - \overline{y})^2 \\
SSB &= \sum_{j=1}^{b} k(\overline{y}_{.j} - \overline{y})^2 \\
SSE &= TotalSS - SST - SSB
\end{aligned}$$

Note that we simply have added items representing the block means ($\overline{y}_{.j}$) and variation among the block means (SSB). We can further think of this as decomposing the total variation into differences among the treatment means (SST), differences among the block means (SSB), and random variation (SSE).

Source of Variation	Sum of Squares	ANOVA Degrees of Freedom	Mean Square	F
TREATMENTS	SST	$k-1$	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE}$
BLOCKS	SSB	$b-1$	$MSB = \frac{SSB}{b-1}$	
ERROR	SSE	$(b-1)(k-1)$	$MSE = \frac{SSE}{(b-1)(k-1)}$	
TOTAL	$TotalSS$	$bk-1$		

Table 6.6: The Analysis of Variance Table for the Randomized Block Design

Once again, the main purpose for conducting this type of experiment is to detect differences among the treatment means (treatment effects). The test is very similar to that of the CRD, with only minor adjustments. We are rarely interested in testing for differences among blocks, since we expect there to be differences among them (that's why we set up the design this way), and they were just a random sample from a population of such experimental units. The treatments are the items we chose specifically to compare in the experiment. The testing procedure can be described as follows:

1. $H_0 : \alpha_1 = \cdots = \alpha_k = 0 \quad (\mu_1 = \cdots = \mu_k)$ (No treatment effect)
2. $H_A :$ Not all α_i are 0 (Treatment effects exist)
3. T.S. $F_{obs} = \frac{MST}{MSE}$
4. R.R.: $F_{obs} \geq F_{\alpha,k-1,(b-1)(k-1)}$
5. p-value: $P(F \geq F_{obs})$

Not surprisingly, the procedures to make comparisons among means are also very similar to the methods used for the CRD. In each formula described previously for Dunnett's, Bonferroni's, and Tukey's methods, we replace n_i with b, when making comparisons among treatment means.

Example 6.3 In Example 1.5, we plotted data from a study quantifying the interaction between theophylline and two drugs (famotidine and cimetidine) in a three–period crossover study that included receiving theophylline with a placebo control (Bachmann, et al.,1995). We would like to compare the mean theophylline clearances when it is taken with each of the three drugs: cimetidine, famotidine, and placebo. Recall from Figure 1.5 that there was a large amount of subject–to–subject variation. In the RBD, we control for that variation when comparing the three treatments. The raw data, as well as treatment and subject (block) means are given in Table 6.7. The Analysis of Variance is given in Table 6.8. Note that in this example, we are comparing $k = 3$ treatments in $b = 14$ blocks.

	Interacting Drug			Subject
Subject	Cimetidine	Famotidine	Placebo	Mean
1	3.69	5.13	5.88	4.90
2	3.61	7.04	5.89	5.51
3	1.15	1.46	1.46	1.36
4	4.02	4.44	4.05	4.17
5	1.00	1.15	1.09	1.08
6	1.75	2.11	2.59	2.15
7	1.45	2.12	1.69	1.75
8	2.59	3.25	3.16	3.00
9	1.57	2.11	2.06	1.91
10	2.34	5.20	4.59	4.04
11	1.31	1.98	2.08	1.79
12	2.43	2.38	2.61	2.47
13	2.33	3.53	3.42	3.09
14	2.34	2.33	2.54	2.40
Trt Mean	2.26	3.16	3.08	2.83

Table 6.7: Theophylline clearances (liters/hour) when drug is taken with interacting drugs

We can now test for treatment effects, and if necessary use Tukey's method to make pairwise comparisons among the three drugs ($\alpha = 0.05$ significance level).

1. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad (\mu_1 = \mu_2 = \mu_3)$ (No drug effect on theophylline clearance)

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
TREATMENTS	7.01	2	3.51	10.64
BLOCKS	71.81	13	5.52	
ERROR	8.60	26	0.33	
TOTAL	87.42	41		

Table 6.8: Analysis of Variance table for theophylline interaction data (RBD)

2. H_A : Not all α_i are 0 (Drug effects exist)

3. T.S. $F_{obs} = \frac{MST}{MSE} = 10.64$

4. R.R.: $F_{obs} \geq F_{\alpha,k-1,(b-1)(k-1)} = F_{0.05,2,26} = 3.37$

5. p-value: $P(F \geq F_{obs}) = P(F \geq 10.64) = 0.0004$

Since we do reject H_0, and conclude differences exist among the treatment means, we will use Tukey's method to determine which drugs differ significantly. Recall that for Tukey's method, we compute simultaneous confidence intervals of the form given below, with k being the number of treatments (k=3), n the total number of observations ($n = bk$=3(14)=42), and n_i the number of measurements per treatment ($n_i = b = 14$).

$$(\overline{y}_i - \overline{y}_j) \pm q_{\alpha,k,n-k}\sqrt{MSE(\frac{1}{n_i})} \quad \Longrightarrow \quad (\overline{y}_i - \overline{y}_j) \pm 3.514\sqrt{0.33(\frac{1}{14})} \quad \Longrightarrow (\overline{y}_i - \overline{y}_j) \pm 0.54$$

The corresponding simultaneous 95% confidence intervals and conclusions are given in Table 6.9. We conclude that theophylline has a significantly lower clearance when taken with cimetidine than

Comparison	$\overline{y}_i - \overline{y}_j$	CI	Conclusion
Cimetidine vs Famotidine	$2.26 - 3.16 = -0.90$	$(-1.44, -.36)$	$C < F$
Cimetidine vs Placebo	$2.26 - 3.08 = -0.82$	$(-1.36, -.28)$	$C < P$
Famotidine vs Placebo	$3.16 - 3.08 = 0.08$	$(-0.46, 0.62)$	$F = P$

Table 6.9: Tukey's simultaneous 95% CI's for theophylline interaction data (RBD)

when taken with famotidine or placebo. No difference appear to exist when theophylline is taken with famotidine or with placebo. While cimetidine appears to interact with theophylline, famotidine does not appear to interact with it in patients with chronic obstructive pulmonary disease.

6.2.2 Friedman's Test for the Randomized Block Design

A nonparametric procedure that can be used to analyze data from the Randomized Block Design (RBD), where each subject receives each treatment is Friedman's Test. The idea behind Friedman's Test is to rank the measurements corresponding to the k treatments within each block. We then compute the rank sum corresponding to each treatment. This test can also be used when the data consists of preferences (ranks) among k competing items.

Once the measurements are ranked within each block from 1 (smallest) to k (largest), and the rank sums $T_1, T_2, \ldots, T_k$ are computed for each treatment, the test is conducted as follows (assume b blocks are used in the experiment):

1. H_0 : The k population distributions are identical ($\mu_1 = \mu_2 = \cdots = \mu_k$)
2. H_A : Not all k distributions are identical (Not all μ_i are equal)
3. T.S.: $F_r = \frac{12}{bk(k+1)} \sum_{i=1}^{k} T_i^2 - 3b(k+1)$.
4. R.R.: $F_r \geq \chi^2_{\alpha,k-1}$.
5. p–value:$P(\chi^2 \geq F_r)$

Either k (the number of treatments) or b (the number of blocks) must be larger than 5 for this test to be appropriate.

If we do reject H_0, and conclude treatment effects exist, we can conduct Wilcoxon's Signed–Rank Test on all pairs of treatments (adjusting α for the number of comparisons being made, as in Bonferroni's method), to determine which pairs differ significantly. Other, more powerful methods are available that need extensive tables (see Hollander and Wolfe (1974), p.151).

Example 6.4 A crossover study was conducted to compare the absorption characteristics of a new formulation of valproate – depakote sprinkle in capsules (Carrigan, et al.,1990). There were $b = 11$ subjects, and each received the new formulation (capsule) in both fasting and non–fasting conditions. They also received an enteric–coated tablet. Each drug was given to each subject three times. Among the pharmacokinetic parameters measured was t_{max}, the time to maximum concentration. The mean t_{max} for each treatment (capsule–fasting, capsule–nonfasting, enteric–coated–fasting) is given for each subject in Table 6.10, as well as the within subject ranks. We will test for treatment effects using Friedman's test ($\alpha = 0.05$).

1. H_0 : The 3 distributions of t_{max} are identical for the three treatments ($\mu_1 = \mu_2 = \mu_3$)
2. H_A : The 3 distributions of t_{max} are not identical (Not all μ_i are equal)
3. T.S.: $F_r = \frac{12}{bk(k+1)} \sum_{i=1}^{k} T_i^2 - 3b(k+1) = \frac{12}{11(3)(4)}[(19.0)^2 + (32.5)^2 + (14.5)^2] - 3(11)(4) = 147.95 - 132 = 15.95$.
4. R.R.: $F_r \geq \chi^2_{\alpha,k-1} = \chi^2_{.05,2} = 5.99$.

	Formulation/Fasting State		
Subject	Capsule (fasting)	Capsule (nonfasting)	Enteric–Coated (fasting)
1	3.5 (2)	4.5 (3)	2.5 (1)
2	4.0 (2)	4.5 (3)	3.0 (1)
3	3.5 (2)	4.5 (3)	3.0 (1)
4	3.0 (1.5)	4.5 (3)	3.0 (1.5)
5	3.5 (1.5)	5.0 (3)	3.5 (1.5)
6	3.0 (1)	5.5 (3)	3.5 (2)
7	4.0 (2.5)	4.0 (2.5)	2.5 (1)
8	3.5 (2)	4.5 (3)	3.0 (1)
9	3.5 (1.5)	5.0 (3)	3.5 (1.5)
10	3.0 (1)	4.5 (3)	3.5 (2)
11	4.5 (2)	6.0 (3)	3.0 (1)
Rank sum	$T_1 = 19.0$	$T_2 = 32.5$	$T_3 = 14.5$

Table 6.10: Mean t_{max} and (ranks) for valproate absorption study

We reject H_0, and conclude that treatment effects exist. Clearly, the capsule taken nonfasting has the highest times to maximum concentration (lowest rate of absorption). We could conduct the Wilcoxon signed–rank test on all pairs of treatments to determine which pairs are significantly different. The results would be that the nonfasting/capsule had a higher mean than both fasting/capsule and fasting/enteric–coated tablet, and that the fasting/capsule and fasting/enteric–coated tablet were not significantly different.

6.3 Other Frequently Encountered Experimental Designs

In this section, we will introduce three commonly used designs through examples. We will proceed through their analyses without spending much time on the theoretical details. Hopefully these examples will assist you if you ever encounter them in practice. The three designs are:

1. Factorial Designs
2. Crossover Designs With Sequence and Period Effects
3. Repeated Measures Designs

6.3.1 Factorial Designs

Many times we have more than one set of treatments that we'd like to compare simultaneously. For instance, drug trials are generally run at different medical centers. In this case, the drug a subject receives would be factor A (active or placebo), while the center he/she is located at would be factor B. Then we might test for drug effects or center effects.

An interaction would exist if the drug effects differ among centers. That is an undesirable situation, but one we should test for. If we wish to measure the interaction we will have to have

more than one measurement (replicate) corresponding to each combination of levels of the 2 factors. In this situation, that would mean having multiple subjects receiving each treatment at each center.

Denoting the k^{th} measurement observed under the i^{th} level of factor A and the j^{th} level of factor B, the model is written as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk},$$

where μ is the overall mean, α_i is the effect of the i^{th} level of factor A, β_j is the effect of the j^{th} level of factor B, $\alpha\beta_{ij}$ is the effect of the interaction of the i^{th} level of factor A and the j^{th} level of factor B, and ε_{ijk} is the random error term representing the fact that subjects within each treatment combinations will vary, as well as if the same subject were measured repeatedly, his/her measurements would vary. As before, we assume that ε_{ijk} is normally distributed with mean 0 and variance σ^2. Some interesting hypotheses to test are as follows:

1. $H_0 : \alpha\beta_{11} = \cdots = \alpha\beta_{ab} = 0$ (No interaction effect).
2. $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ (No effects among the levels of factor A)
3. $H_0 : \beta_1 = \cdots = \beta_b = 0$ (No effects among the levels of factor B)

The total variation in the set of observed measurements can be decomposed into four parts: variation in the means of the levels of factor A, variation in the means of the levels of factor B, variation due to the interaction of factors A and B, and error variation. The formulas for the sums of squares are given in many statistics textbooks and will not be given here. Note that this type of analysis, is almost always done on a computer. The analysis of variance can be set up as shown in Table 6.11, assuming r measurements are made at each combination of levels of the two factors.

Source of Variation	Sum of Squares	ANOVA Degrees of Freedom	Mean Square	F
FACTOR A	SSA	$a-1$	$MSA = \frac{SSA}{a-1}$	$F = \frac{MSA}{MSE}$
FACTOR B	SSB	$b-1$	$MSB = \frac{SSB}{b-1}$	$F = \frac{MSB}{MSE}$
INTERACTION AB	$SSAB$	$(a-1)(b-1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F = \frac{MSAB}{MSE}$
ERROR	SSE	$ab(r-1)$	$MSE = \frac{SSE}{ab(r-1)}$	
TOTAL	$TotalSS$	$abr-1$		

Table 6.11: The Analysis of Variance Table for a 2-Factor Factorial Design

The tests for interactions and for effects of factors A and B involve the three F–statistics, and can be conducted as follow.

1. $H_0 : \alpha\beta_{11} = \cdots = \alpha\beta_{ab} = 0$ (No interaction effect).
2. H_A : Not all $\alpha\beta_{ij} = 0$ (Interaction effects exist)

3. T.S. $F_{obs} = \frac{MSAB}{MSE}$
4. R.R.: $F_{obs} \geq F_{\alpha,(a-1)(b-1),ab(r-1)}$
5. p-value: $P(F \geq F_{obs})$

Assuming no interaction effects exist, we can test for differences among the effects of the levels of factor A as follows.

1. $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ (No factor A effect).
2. H_A : Not all $\alpha_i = 0$ (Factor A effects exist)
3. T.S. $F_{obs} = \frac{MSA}{MSE}$
4. R.R.: $F_{obs} \geq F_{\alpha,(a-1),ab(r-1)}$
5. p-value: $P(F \geq F_{obs})$

Assuming no interaction effects exist, we can test for differences among the effects of the levels of factor B as follows.

1. $H_0 : \beta_1 = \cdots = \beta_b = 0$ (No factor B effect).
2. H_A : Not all $\beta_j = 0$ (Factor B effects exist)
3. T.S. $F_{obs} = \frac{MSB}{MSE}$
4. R.R.: $F_{obs} \geq F_{\alpha,(b-1),ab(r-1)}$
5. p-value: $P(F \geq F_{obs})$

Note that if we conclude interaction effects exist, we usually look at the individual combinations of factors A and B separately (as in the Completely Randomized Design), and don't conduct the last two tests.

Example 6.5 Various studies have shown that interethnic differences in drug–metabolizing enzymes exist. A study was conducted to determine whether differences exist in pharmacokinetics of the tricyclic antidepressant nortriptyline between Hispanics and Anglos (Gaviria, et al.,1986). The study consisted of five males and five females from each ethnic group. We would like to determine whether ethnicity or sex differences exist, or whether there is an interaction between the two variables on the outcome variable total clearance ($ml/min/kg$). The data (by ethnic/sex group) are given in Table 6.12 and in Figure 6.1. The model is written as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk},$$

where μ is the overall mean, α_i is the effect of the i^{th} level of factor A (ethnicity: 1=Hispanic, 2=Anglo), β_j is the effect of the j^{th} level of factor B (sex: 1=Female, 2=Male), $\alpha\beta_{ij}$ is the effect of the interaction of the i^{th} of ethnicity and the j^{th} level of sex.

Hispanics		Anglos	
Females	Males	Females	Males
10.5	5.4	7.1	5.7
8.3	7.1	10.8	3.8
8.5	6.1	12.3	7.8
6.4	10.8	7.0	4.4
6.5	4.1	7.9	9.9

Table 6.12: Total clearance data for nortriptyline study

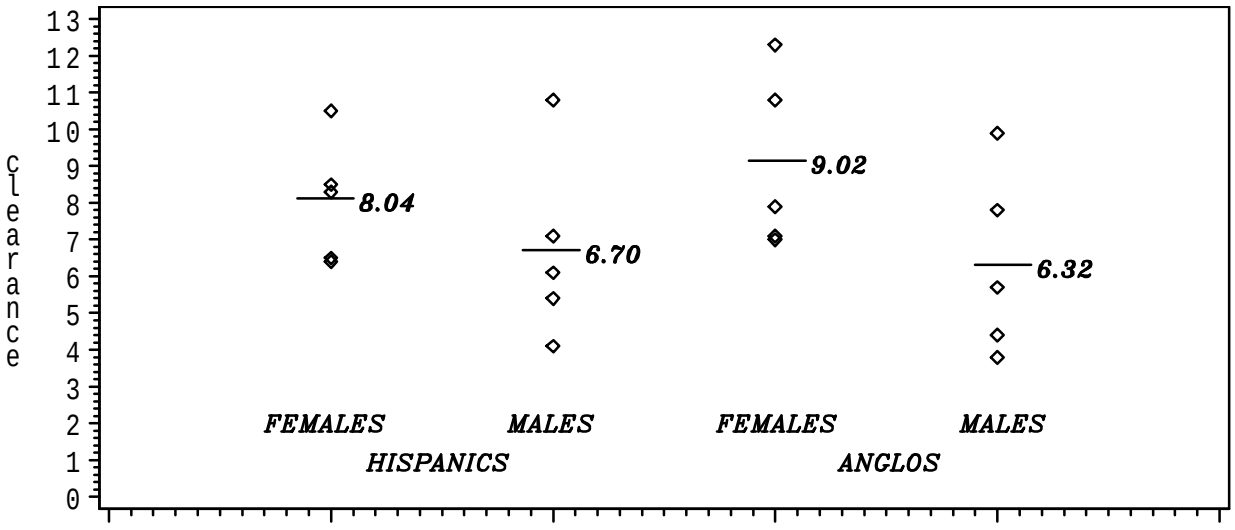

Figure 6.1: Total clearances and group means for nortriptyline ethnicity study

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Ethnicity (A)	0.450	1	0.450	$\frac{0.450}{5.347} = 0.084$
Sex (B)	20.402	1	20.402	$\frac{20.402}{5.347} = 3.816$
Ethnic×Sex (AB)	2.312	1	2.312	$\frac{2.312}{5.347} = 0.432$
ERROR	85.548	16	5.347	
TOTAL	108.712	19		

Table 6.13: The Analysis of Variance table for the nortriptyline data

We can compute the Analysis of Variance table, by obtaining the total sum of squares and partitioning that variation into parts attributable to: ethnicity differences, sex differences, ethnicity/sex interaction, and random (within ethnic/sex group) variation. The Analysis of Variance is given in Table 6.13.

The tests for interactions and for effects of factors A and B involve the three F–statistics, and can be conducted as follow (each test at $\alpha = 0.05$).

1. $H_0 : \alpha\beta_{11} = \cdots = \alpha\beta_{22} = 0$ (No interaction effect).
2. H_A : Not all $\alpha\beta_{ij} = 0$ (Interaction effects exist)
3. T.S. $F_{obs} = \frac{MSAB}{MSE} = 0.432$
4. R.R.: $F_{obs} \geq F_{.05,1,16} = 4.49$
5. p-value: $P(F \geq 0.432) = .5203$

Since no interaction effects exist, we can test for differences among the effects of the levels of factor A (ethnicity) as follows.

1. $H_0 : \alpha_1 = \alpha_2 = 0$ (No ethnicity effect).
2. H_A : Not all $\alpha_i = 0$ (Ethnicity effects exist)
3. T.S. $F_{obs} = \frac{MSA}{MSE} = 0.084$
4. R.R.: $F_{obs} \geq F_{.05,1,16} = 4.49$
5. p-value: $P(F \geq 0.084) = .7757$

Again, since no interaction effects exist, we can test for differences among the effects of the levels of factor B (sex) as follows.

1. $H_0 : \beta_1 = \beta_2 = 0$ (No sex effect).

2. H_A : Not all $\beta_j = 0$ (Sex effects exist)
3. T.S. $F_{obs} = \frac{MSB}{MSE} = 3.816$
4. R.R.: $F_{obs} \geq F_{.05,1,16} = 4.49$
5. p-value: $P(F \geq 3.816) = .0685$

A few things are worth noting here:

- An interaction would have meant that ethnicity effects differed between the sexes (and that sex effects differed between the ethnicities).
- Since there is no interaction, we can test for main effects between ethnicities and then between sexes, seperately. If there had been an interaction, we would have had to treat all four ethnic/sex combinations as individual groups, and compared those four groups (like in the CRD).
- There clearly is no ethnic effect here, we have a large p–value for that test, and also see Figure 6.1.
- While the sex effects are not significant at $\alpha = 0.05$ (p–value=.0685), it is close. These are fairly small samples given the subject–to–subject variability. This is not a very powerful test. Based on the plots, and the consistency across ethnicities, women do appear to eliminate nortriptyline more rapidly (higher clearance) than men do.

6.3.2 Crossover Designs With Sequence and Period Effects

In Chapter 3, we saw how to compare two treatments in a crossover design (paired t–test), and earlier in this chapter we compared three or more treatments in a crossover study (RBD). These two methods assumed that there were no sequence (order of treatments received) effects or period effects. However, sometimes there may be such effects, and we would like to remove them when comparing treatments. In fact, in most studies they are removed since computationally it is no more difficult to conduct this analysis than it is to conduct the paired t–test or the randomized block design (on a computer, anyway). This method of analysis is a major component in determining pharmaceutical bioequivalence (Chapter 10).

Although this analysis looks more formidable than the previous two methods, it is important to remember that the goal is still the same, namely to compare two or more treatments in a crossover study. We will consider only the two treatment case, but the method extends easily to any general number of treatments, although the number of sequences grows rapidly (if there are k treatments, there are k periods, and $k!$ sequences). We will label the treatments as A and B; they may be: new drug/control, new drug/old drug, formulation 1/formulation 2, etc. The experiment is typically conducted in a 2–period crossover, with subjects being randomly assigned into one of two sequences (A followed by B or B followed by A). Then, we partition the total variation of the observed measurements into variation due to: treatments, periods, sequences, subjects within sequences, and random error. The analysis of variance can be formed (on a computer) and is given in Table 6.14. We assume that n_1 subjects received sequence 1 and n_2 subjects received sequence

2, for a total of n subjects and $2n$ measurements. Further, we will denote $\overline{TRT}_i$, $\overline{PER}_i$, $\overline{SEQ}_i$, and $\overline{SUBJ}_{j(i)}$ as the means of the i^{th} treatment, period, sequence, and j^{th} subject (within i^{th} sequence), respectively in the sums of squares formulas. Further $\overline{y}$ is the overall mean. To test for

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Treatments	$SS_{TRT} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(\overline{TRT}_i - \overline{y})^2$	1	$MS_{TRT} = SS_{TRT}$
Periods	$SS_{PER} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(\overline{PER}_i - \overline{y})^2$	1	$MS_{PER} = SS_{PER}$
Sequences	$SS_{SEQ} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(\overline{SEQ}_i - \overline{y})^2$	1	$MS_{SEQ} = SS_{SEQ}$
Subjects(Sequences)	$SS_{SUBJ(SEQ)} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(\overline{SUBJ}_{j(i)} - \overline{SEQ}_i)^2$	$n-2$	$MS_{SUBJ(SEQ)} = \frac{SS_{SUBJ(SEQ)}}{n-2}$
Error	By Subtraction	$n-2$	$MSE = \frac{SSE}{n-2}$
TOTAL	$TotalSS = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y - \overline{y})^2$	$2n-1$	

Table 6.14: The Analysis of Variance Table for a 2–Period Crossover Study

treatment effects, we first denote the means for treatments A and B as μ_A and μ_B, respectively. Then, we conduct the following test:

1. $H_0 : \mu_A = \mu_B$ (No treatment effect).
2. $H_A : \mu_A \neq \mu_B$ (Treatment effects exist)
3. T.S. $F_{obs} = \frac{MS_{TRT}}{MSE}$
4. R.R.: $F_{obs} \geq F_{\alpha,1,n-2}$
5. p-value: $P(F \geq F_{obs})$

Example 6.6 A 2–period crossover study was conducted to compare pharmacokinetics of two transdermal nicotine delivery systems (Gupta, et al.,1995). The two treatments (Nicoderm and Habitrol) were given to $n = 24$ male smokers in random order, with a six day washout period. Pharmacokinetic measurements were made for nicotine and cotinine concentrations at first application and at steady state (fifth day of application). Among the pharmacokinetic parameters measured was $AUC_{0-24}(ng \cdot hr/mL)$ for nicotine at steady state. From the data reported, we get the Analysis of Variance in Table 6.15. We report only the treatment, error, and total sums of squares, since we cannot break down the remaining variation into components due to sequence, period, and subject within sequence – which are unnecessary to the analysis.

We now test for treatment effects, which represents differences in the mean AUC for the two nicotine delivery systems. Note that the sample means are $\overline{y}_N = 441.0$ and $\overline{y}_H = 386.0$ for Nicoderm and Habitrol, respectively.

1. $H_0 : \mu_N = \mu_H$ (No Brand differences)

Source of Variation	ANOVA Sum of Squares	Degrees of Freedom	Mean Square
Treatments	36300.0	1	36300.0
Periods	—	1	—
Sequences	—	1	—
Subjects(Sequences)	—	22	—
Error	107205.1	22	4873.0
TOTAL	343478.8	47	

Table 6.15: The Analysis of Variance table for 2–period crossover transdermal nicotine study

2. $H_A : \mu_A \neq \mu_B$ (Brand differences exist)
3. T.S. $F_{obs} = \frac{MS_{TRT}}{MSE} = \frac{36300.0}{4873.0} = 7.45$
4. R.R.: $F_{obs} \geq F_{\alpha,1,n-2} = F_{.05,1,22} = 4.30$
5. p-value: $P(F \geq F_{obs}) = P(F \geq 7.45) = .0122$

Thus, we can conclude that the means differ, and since the sample mean is higher for Nicoderm than Habitrol. The Nicoderm system has a higher level of bioavailability than the Habitrol system, as measured by AUC.

6.3.3 Repeated Measures Designs

In some experimental situations, subjects are assigned to treatments, and measurements are made repeatedly over some fixed period of time. This can be thought of as a CRD, where more than one measurement is being made on each experimental unit. We would still like to detect differences among the treatment means (effects), but we must account for the fact that measurements are being made over time. Previously, the error was differences among the subjects within the treatments (recall that $SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2$). Now we are observing various measurements on each subject within each treatment, and have a new error term. The measurement Y_{ijk}, representing the outcome for the i^{th} treatment on the j^{th} subject (who receives that treatment) at the k^{th} time point, can be written as:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \varepsilon_{ijk},$$

where:

- μ is the overall mean
- α_i is the effect of the i^{th} treatment

- $\beta_{j(i)}$ is the effect of the j^{th} subject who receives the i^{th} treatment
- γ_k is the effect of the k^{th} time point
- $\alpha\gamma_{ik}$ is the interaction of the i^{th} treatment and the k^{th} time point
- ε_{ijk} is the random error component that is assumed to be $N(0, \sigma^2)$.

The Analysis of Variance is given in Table 6.16 (this is always doneon a computer). The degrees of freedom are based on the experiment consisting of a treatments, b subjects receiving each treatment, and measurements being made at r points in time. Note that if the number of subjects per treatment differ (b_i subjects receiving treatment i), we replace $a(b-1)$ with $\sum_{i=1}^{a}(b_i - 1)$.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Treatments	SSA	$a-1$	$MSA = \frac{SSA}{a-1}$	$F = \frac{MSA}{MSB(A)}$
Subjects(Trts)	$SSB(A)$	$a(b-1)$	$MSB(A) = \frac{SSB(A)}{a(b-1)}$	
Time	$SSTime$	$r-1$	$MSTime = \frac{SSTime}{r-1}$	
Trt * Time	$SSATi$	$(a-1)(r-1)$	$MSATi = \frac{SSATi}{(a-1)(r-1)}$	
Error	SSE	$a(b-1)(r-1)$	$MSE = \frac{SSE}{a(b-1)(r-1)}$	
TOTAL	$TotalSS$	$abr-1$		

Table 6.16: The Analysis of Variance Table for a Repeated Measures Design

The main hypothesis we would test is for a treatment effect. This test is of the form:

1. $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ (No treatment effect)
2. H_A : Not all $\alpha_i = 0$ (Treatment effects)
3. T.S.: $F_{obs} = \frac{MSA}{MSB(A)}$
4. R.R.: $F_{obs} \geq F_{\alpha,a-1,a(b-1)}$
5. p-value: $P(F \geq F_{obs})$

Example 6.7 A study was conducted to determine the safety of long–term use of sulfacytine on the kidney function in men (Moyer, et al.,1972). The subjects were 34 healthy prisoners in Michigan, where the prisoners were assigned at random to receive one of: high dose (500 mg, 4 times daily), low dose (250 mg, 4 times daily), or no dose (placebo, 4 times daily). Measurements of creatinine clearance were taken once weekly for 13 weeks, along with a baseline (pre–R_x) reading.

The goal was to determine whether or not long–term use of sulfacytine affected renal function, which was measured by creatinine clearance.

Note the following elements of this study:

Treatments These are the dosing regimens (high, low, none)

Subjects 34 prisoners, each prisoner being assigned to one treatment (parallel groups). 12 received each drug dose, 10 received placebo.

Time Periods There were 14 measurements made on each prisoner, one at baseline, then one each week for 13 weeks.

The means for each treatment/week combination are given in Table 6.17. Again, recall our goal is to determine whether the overall means differ among the three treatment groups.

Treatment	Week														Trt
Grp	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Mean
H.D.	100.0	102.2	102.3	105.2	94.3	104.8	90.8	96.8	93.3	93.6	85.7	91.8	93.6	98.0	96.6
L.D.	87.6	96.1	94.7	105.2	91.9	102.5	98.5	95.5	99.7	106.3	93.3	103.0	94.7	94.9	97.4
Plac	103.9	108.7	99.4	105.6	89.0	97.3	101.1	97.1	94.7	101.7	100.4	102.6	91.1	90.8	98.8
Mean	96.8	102.0	98.8	105.3	91.9	101.8	96.5	96.4	96.0	100.5	92.7	98.9	93.3	94.8	97.6

Table 6.17: Mean Creatinine clearances for each treatment/week combination – sulfacytine data

Note that in this problem, we have $a = 3$ treatments and $r = 14$ time points, but that b, the number of subjects within each treatment varies ($b_1 = b_2 = 12$, $b_3 = 10$). This will cause no problems however, and the degrees of freedom for the analysis of variance table will be adjusted so that $a(b-1)$ will be replaced by $b_1 + b_2 + b_3 - a = 34 - 3 = 31$. Based on the data presented, and a reasonable assumption on the subject–to–subject variability, we get the analysis of variance given in Table 6.18.

ANOVA				
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Treatments	376.32	2	188.16	$\frac{188.16}{3381.36} = 0.0556$
Subjects(Trts)	104822.16	31	3381.36	
Time	2035.58	13	156.58	
Trt * Time	6543.34	26	251.67	
Error	45282.14	403	112.36	
TOTAL	159059.54	475		

Table 6.18: The Analysis of Variance Table for Sulfacytine Example

Now, we can test whether the mean creatinine clearances differ among the three treatment groups (at $\alpha = 0.05$ significance level):

1. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ (No treatment effect)
2. H_A : Not all $\alpha_i = 0$ (Treatment effects)
3. T.S.: $F_{obs} = \frac{MSA}{MSB(A)} = 0.0556$
4. R.R.: $F_{obs} \geq F_{\alpha,a-1,a(b-1)} = F_{0.05,2,31} = 3.305$
5. p-value: $P(F \geq F_{obs}) = P(F \geq 0.0556) = .9460$

We fail to reject H_0, and we conclude that there is no treatment effect. Long–term use of sulfacytine does not appear to have any effect on renal function (as measured by creatinine clearance).

6.4 Exercises

35. A study to determine whether of not patients who had suffered from clozapine–induced agranulocytosis had abnormal free radical scavenging enzyme activity (FRESA), compared $k = 4$ groups: post–clozapine agranulocytosis (PCA), clozapine no agranulocytosis (CNA), West Coast controls (WCC), and Long Island Jewish Medical Center controls (LIJC) (Linday, et al.,1995). One measure FRESA was the glutathione peroxidase level in Plasma. Table 6.19 gives the summary statistics for each group, Table 6.20 has the corresponding Analysis of Variance, and Table 6.21 has Bonferroni's and Tukey's simultaneous 95% confidence intervals comparing each pair of groups.

(a) Test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ vs H_A : Not all μ_i are equal.

(b) Assuming you reject H_0 in part a), which groups are significantly different? In particular, does the PCA group appear to differ from the others?

(c) Which method, Bonferroni's or Tukey's, gives the most precise confidence intervals?

	Group 1 (PCA)	Group 2 (CNA)	Group 3 (WCC)	Group 4 (LIJC)
Mean	$\overline{y}_1 = 34.3$	$\overline{y}_2 = 44.5$	$\overline{y}_3 = 45.3$	$\overline{y}_4 = 46.4$
Std Dev	$s_1 = 6.9$	$s_2 = 7.4$	$s_3 = 4.6$	$s_4 = 8.7$
Sample Size	$n_1 = 9$	$n_2 = 12$	$n_3 = 14$	$n_4 = 12$

Table 6.19: Sample statistics for glutathione peroxidase levels in four patient groups

36. A study was conducted to compare the sexual side effects among four antidepressants: bupropion, fluoxetine, paroxetine, and sertraline (Modell, et al., 1997). Psychiatric outpatients were asked to anonymously complete a questionaire regarding changes in the patients' sexual functioning relative to that before the onset of the patients' psychiatric illnesses.

One of the questions asked was: "Compared with your previously normal state, please rate how your *libido (sex drive)* has changed since you began taking this mediation. The range of outcomes

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
TREATMENTS	917.6	3	305.9	
ERROR	2091.0	43	48.6	
TOTAL	3008.6	46		

Table 6.20: The Analysis of Variance table for glutathione peroxidase levels in four patient groups

		Simultaneous 95% CI's	
Comparison	$\overline{y}_i - \overline{y}_j$	Bonferroni	Tukey
PCA vs CNA	$34.3 - 44.5 = -10.2$	$(-18.7, -1.7)$	$(-17.7, -2.7)$
PCA vs WCC	$34.3 - 45.3 = -11.0$	$(-19.3, -2.7)$	$(-18.3, -3.7)$
PCA vs LIJC	$34.3 - 46.4 = -12.1$	$(-20.6, -3.6)$	$(-19.6, -4.6)$
CNA vs WCC	$44.5 - 45.3 = -0.8$	$(-8.4, 6.8)$	$(-7.5, 5.9)$
CNA vs LIJC	$44.5 - 46.4 = -1.9$	$(-9.8, 6.0)$	$(-8.8, 5.0)$
WCC vs LIJC	$45.3 - 46.4 = -1.1$	$(-8.7, 6.5)$	$(-7.8, 5.6)$

Table 6.21: Bonferroni and Tukey multiple comparisons for the glutathione peroxidase levels in four patient groups

ranged from -2 (very much decreased) to $+2$ (very much increased). Although the scale was technically ordinal, the authors treated it as interval scale (as is commonly done).

The overall mean score was $\overline{y} = -0.38$. The group means and standard deviations are given below in Table 6.22.

Drug (i)	n_i	$\overline{y}_i$	s_i	$n_i(\overline{y}_i - \overline{y})^2$	$(n_i - 1)s_i^2$
Buproprion (1)	22	0.46	0.80	$22(0.46 - (-0.38))^2 = 15.52$	$(22 - 1)(0.80)^2 = 13.44$
Fluoxetine (2)	37	-0.49	0.97		
Paroxetine (3)	21	-0.90	0.73		
Sertraline (4)	27	-0.49	1.25		

Table 6.22: Summary statistics and sums of squares calculations for sexual side effects of antidepressant data.

(a) Complete Table 6.22.

(b) Set up the Analysis of Variance table.

(c) Denoting the population mean change for drug i as μ_i, test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_A : Not all μ_i are equal at the $\alpha = 0.05$ significance level.

(d) Use Table 6.23 to make all pairwise comparisons among treatment means at the $\alpha = 0.06$ significance level (this strange level allows us to make each comparison at the $\alpha = 0.01$ level). Use Bonferroni's procedure.

Trts (i,j)	$\overline{y}_i - \overline{y}_j$	$t_{.01,103}\sqrt{MSE(1/n_i + 1/n_j)}$	Conclude
(1,2)	$0.46 - (-0.49) = 0.95$	0.624	$\mu_1 > \mu_2$
(1,3)		0.708	
(1,4)		0.666	
(2,3)		0.634	
(2,4)		0.587	
(3,4)		0.675	

Table 6.23: Pairwise comparison of antidepressant formulations on sexual side effects.

(e) Does any formulation appear to be better than all the others? If so, which is it.

(f) The F–test is derived from the assumption that all populations being compared are approximately normal with common variance σ^2, which is estimated by $\hat{\sigma}^2 = MSE$. Based on this estimate of the variance, as well as the estimates of the individual means, sketch the probability distributions of the individual measurements (assuming individuals scores actuall fall along a continuous axis, not just at the discrete points $-2, -1, \ldots, 2$.

37. A Phase III clinical trial compared the efficacy of fluoxetine with that of imipramine in patients with major depressive disorder (Stark and Hardison, 1985). The mean change from baseline for each group, as well as the standard deviations are given in Table 6.24. Obtain the analysis of variance, test for treatment effects, and use Bonferroni's procedure to obtain 95% simultaneous confidence intervals for the differences among all pairs of means.

	Group 1 (Fluoxetine)	Group 2 (Imipramine)	Group 3 (Placebo)
Mean	$\overline{y}_1 = 11.0$	$\overline{y}_2 = 12.0$	$\overline{y}_3 = 8.2$
Std Dev	$s_1 = 10.1$	$s_2 = 10.1$	$s_3 = 9.0$
Sample Size	$n_1 = 185$	$n_2 = 185$	$n_3 = 169$

Table 6.24: Sample statistics for change in Hamilton depression scores in three treatment groups

38. An intranasal monoclonal antibody (HNK20) was tested against respiratory syncytial virus (RSV) in rhesus monkeys (Weltsin, et al.,1996). A sample of $n = 24$ monkeys were given RSV, and randomly assigned to receive one of $k = 4$ treatments: placebo, 0.2 mg/day, 0.5 mg/day, or 2.5 mg/day HNK20. The monkeys, free of RSV, received the treatment intranasally once daily for two days, then given RSV and given treament daily for four more days. Nasal swabs were collected daily to measure the amount of RSV for 14 days. Table 6.25 gives the peak RSV titer (log_{10}/mL) for the 24 monkeys by treatment, and their corresponding ranks. Note that low RSV titers correspond to more effective treatment. Table 6.26 gives the Wilcoxon rank sums for each pair of treatments.

(a) Use the Kruskal–Wallis test to determine whether or not treatment differences exist.

(b) Assuming treatment differences exist, use the Wilcoxon Rank Sum test to compare each pair of treatments. Note that 6 comparisons are being made, so that if each is conducted at $\alpha = .01$ the overall error rate will be no higher than 6(.01)=.06, which is close to the .05 level. For each comparison, conclude $\mu_1 \neq \mu_2$ if $T = min(T_1, T_2) \leq 23$, this gives an overall error rate of $\alpha = .06$.

Treatment			
Placebo ($i = 1$)	HNK20 ($0.2mg/day$) ($i = 2$)	HNK20 ($0.5mg/day$) ($i = 3$)	HNK20 ($2.5mg/day$) ($i = 4$)
5.5 (19)	3.5 (9)	4.0 (11.5)	2.5 (6)
6.0 (22.5)	5.5 (19)	3.0 (7.5)	≤ 0.5 (2.5)
4.5 (14.5)	6.0 (22.5)	4.0 (11.5)	≤ 0.5 (2.5)
5.5 (19)	4.0 (11.5)	3.0 (7.5)	1.5 (5)
5.0 (16.5)	6.0 (22.5)	4.5 (14.5)	≤ 0.5 (2.5)
6.0 (22.5)	5.0 (16.5)	4.0 (11.5)	≤ 0.5 (2.5)
$T_1 = 114.0$	$T_2 = 101.0$	$T_3 = 64.0$	$T_4 = 21.0$
$T_1^2/n_1 = 2166.0$	$T_2^2/n_2 = 1700.2$	$T_3^2/n_3 = 682.7$	$T_4^2/n_4 = 73.5$

Table 6.25: Peak RSV titer in HNK20 monoclonal antibody study in $n = 24$ rhesus monkeys

Trt Pairs	T_1	T_2
Placebo/0.2 mg/day	42.5	35.5
Placebo/0.5 mg/day	56.5	21.5
Placebo/2.5 mg/day	57.0	21.0
0.2/0.5	50.5	27.5
0.2/2.5	57.0	21.0
0.5/2.5	57.0	21.0

Table 6.26: Wilcoxon Rank Sums for each pair of treatment in HNK20 monoclonal antibody study

39. Pharmacokinetics of $k = 5$ formulations of flurbiprofen were compared in a crossover study (Forland, et al.,1996). Flurbiprofen is commercially available as a racemic mixture, with its pharmacologic effect being attributed to the S isomer. The drug was delivered in toothpaste in $k = 5$ concentration/R:S ratio combinations. These were:

1. 1% 50:50 (commercially available formulation)
2. 1% 14:86
3. 1% 5:95
4. 0.5% 5:95
5. 0.25% 5:95

Data for mean residence time ($ng \cdot hr^2/mL$) for S–flurbiprofen are given in Table 6.27, as well as the block means and treatment means. The Analysis of Variance is given in Table 6.28. Test whether or not the treatment means differ in terms of the variable mean residence time ($\alpha = 0.05$). Does there appear to be a formulation effect in terms of the length of time the drug is determined to be in the body? Which seems to vary more, the treatments or the subjects?

	Formulation					
Subject	1	2	3	4	5	Mean
1	13.3	8.0	8.2	10.2	9.3	9.80
2	10.8	13.7	9.5	8.9	10.5	10.68
3	3.0	5.9	6.9	10.3	3.3	5.88
4	2.9	5.7	7.2	6.3	7.8	5.98
5	0.7	4.7	8.0	4.8	8.2	5.28
6	3.4	4.3	7.4	4.0	4.1	4.64
7	16.1	11.9	7.6	8.4	8.5	10.50
8	9.8	7.2	8.5	8.3	3.7	7.50
Mean	7.5	7.7	7.9	7.7	6.9	7.5

Table 6.27: S–flurbiprofen mean residence times

ANOVA				
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
TREATMENTS	5.50	4	1.38	
BLOCKS	243.71	7	34.82	
ERROR	200.72	28	7.17	
TOTAL	449.93	39		

Table 6.28: Analysis of Variance table for flurbiprofen mean residence time data (RBD)

40. In the previously described study (Forland, et al.,1996), the authors also reported the S isomer area under the concentration–time curve (AUC, $ng \cdot hr/mL$). These data have various outlying observations, thus normality assumptions won't hold. Data and some ranks are given in Table 6.29. Complete the rankings and test whether or not differences exist among the formulations with Friedman's test ($\alpha = 0.05$).

Assuming you determine treatment effects exist, we would like to compare all 10 pairs of treatments, simultaneously. We will make each comparison at the $\alpha = 0.01$ significance levels, so that the overall significance level will be no higher than 10(0.01)=0.10. For the Wilcoxon Signed–Rank test, with $n = 8$ pairs, we conclude that $\mu_1 \neq \mu_2$ if $T = min(T^+, T^-) \leq 2$ when we conduct a 2–sided test at $\alpha = 0.01$. Which pairs of treatments differ significantly, based on this criteria? Within each

	Formulation				
Subject	1	2	3	4	5
1	2249 (2)	3897 (4)	3938 (5)	3601 (3)	2118 (1)
2	1339 (4)	2122 (5)	649 (1)	834 (3)	815 (2)
3	282 (2)	694 (4)	1200 (5)	583 (3)	228 (1)
4	319 (1)	1617 (5)	982 (3)	1154 (4)	419 (5)
5	10 (1)	192 (3)	263 (4)	487 (5)	165 (2)
6	417 (3)	536 (4)	685 (5)	285 (2)	257 (1)
7	1063 ( )	1879 ( )	1404 ( )	1302 ( )	642 ( )
8	881 ( )	1433 ( )	1795 ( )	2171 ( )	619 ( )

Table 6.29: *AUC* and (ranks) for flurbiprofen crossover study

significantly different pair, which treatment has a higher mean *AUC*? For each comparison, T^+ and T^- are given in Table 6.30.

Comparison	T^+	T^-
1 vs 2	0	36
1 vs 3	5	31
1 vs 4	6	30
1 vs 5	30	6
2 vs 3	20	16
2 vs 4	26	10
2 vs 5	36	0
3 vs 4	21	15
3 vs 5	34	2
4 vs 5	36	0

Table 6.30: Wilcoxon signed–rank statistics for each pair of formulations for flurbiprofen crossover study

41. The effects of long–term zidovudine use in patients with HIV–1 was studied in terms of neurocognitive functions (Baldeweg, et al.,1995). Patients were classified by zidovudine status (long–term user or non–user), and by their disease state (asymtomatic, symptomatic (non–AIDS), or AIDS). We would like to determine if there is a drug effect, a disease state effect, and/or an interaction between drug and disease state for the response Hospital Anxiety Scale scores. Note that this experiment is unbalanced, in the sense that the sample sizes for each of the six groups vary. The mean, standard deviation, and sample size for each drug/disease state group is given in Table 6.31, and the Analysis of Variance (based on partial, or Type III sums of squares) is given in Table 6.32.

(a) Is the interaction between drug and disease state significant at $\alpha = 0.05$?

(b) Assuming the interaction is not significant, test for drug effects and disease state effects at $\alpha = 0.05$. Does long–term use of zidovudine significantly reduce anxiety scores?

ZDV Group	Asymptomatic	Symptomatic (non–AIDS)	AIDS
Off ZDV	$\overline{y} = 6.6$ $s = 4.3$ $n = 35$	$\overline{y} = 10.7$ $s = 4.7$ $n = 21$	$\overline{y} = 9.6$ $s = 5.1$ $n = 5$
Off ZDV	$\overline{y} = 6.2$ $s = 3.2$ $n = 19$	$\overline{y} = 5.9$ $s = 3.4$ $n = 11$	$\overline{y} = 8.3$ $s = 2.4$ $n = 7$

Table 6.31: Summary statistics for anxiety scale scores for each drug/disease state combination for long–term zidovudine study

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Zidovudine (A)	77.71	1	77.71	
Disease State (B)	105.51	2	52.76	
ZDV×Disease (AB)	89.39	2	44.70	
ERROR	1508.98	92	16.40	
TOTAL	—	97		

Table 6.32: The Analysis of Variance table for the zidovudine anxiety data (Partial or Type III sums of squares)

42. Two oral formulations of loperamide (Diarex Lactab and Imodium capsules) were compared in a two–period crossover study in 24 healthy male volunteers (Doser, et al.,1995). Each subject received each formulation, with half the subjects taking each formulation in each study period. The raw data for $log(AUC)(ng \cdot hr/ml)$ is given in Table 6.33. Note that sequence A means the subject received Diarex Lactab in the first time period and the Imodium capsule in the second time period. Sequence B means the drugs were given in opposite order. The Analysis of Variance is given in Table 6.34. Test whether or not the mean $log(AUC)$ values differ for the two formulations at $\alpha = 0.05$ significance level. (Note that sequence A and B means are 4.0003 and 4.2002, respectively; and period 1 and 2 means are 4.0645 and 4.1360, respectively; if you wish to reproduce the Analysis of Variance table).

Subject	Sequence	Diarex	Imodium	Mean
1	A	3.6881	3.7842	3.7362
2	B	4.0234	3.8883	3.9559
3	B	3.7460	4.0932	3.9196
4	A	3.7593	3.6981	3.7287
5	B	4.3720	4.0042	4.1881
6	B	4.2772	4.1279	4.2026
7	B	4.1382	4.6751	4.4066
8	A	4.1171	4.3944	4.2558
9	A	3.9806	3.8567	3.9187
10	A	3.8040	4.1528	3.9784
11	A	4.2466	4.5474	4.3970
12	B	3.7858	3.8910	3.8384
13	A	4.0193	3.9202	3.9697
14	A	3.7381	3.8073	3.7727
15	A	4.4042	4.5736	4.4889
16	A	3.5888	3.9705	3.7796
17	B	4.6881	4.2060	4.4471
18	B	5.0342	4.7617	4.8980
19	A	4.2502	4.4703	4.3602
20	B	4.0596	4.5567	4.3081
21	B	3.6633	3.7537	3.7085
22	B	4.1735	3.9967	4.0851
23	B	4.3594	4.5285	4.4440
24	A	3.4689	3.7672	3.6180
Mean	—	4.0577	4.1427	4.1002

Table 6.33: $Log(AUC_{0-\infty})$ for Diarex Lactab and Imodium Capsules from bioequivalence study

43. The effects of a hepatotropic agent, malotilate, were observed in rats (Akahane, et al.,1987). In the experiment, it was found that high doses of malotilate were associated with anemia, and reduced red blood cell counts. A sample of 30 rats were taken, and assigned at random to receive either: control, 62.5, 125, 250, 500, or 1000 mg/kg malotilate. Five rats were assigned to each of the $k = 6$ treatments. Measurements of anemic response were based on, among others, red blood cell count $RBC(\times 10^4/mm^3)$, which was measured once a week for 6 weeks. Mean RBC is given

Source of Variation	ANOVA Sum of Squares	Degrees of Freedom	Mean Square
Formulations	0.0867	1	0.0867
Periods	0.0613	1	0.0613
Sequences	0.4792	1	0.4792
Subjects(Sequences)	4.3503	22	0.1977
Error	0.7826	22	0.0356
TOTAL	5.7602	47	

Table 6.34: The Analysis of Variance table for 2–period crossover loperamide bioequivalence study

in Table 6.35 for each treatment group at each time period. The repeated–measures Analysis of Variance is given in Table 6.36. Test whether or not differences in mean RBC exist among the $k = 6$ treatment groups at $\alpha = 0.05$.

Dose	Week 1	Week 2	Week 3	Week 4	Week 5
Control	636	699	716	732	744
62.5	647	708	753	762	748
125	674	722	790	844	760
250	678	694	724	739	704
500	617	668	662	722	645
1000	501	607	613	705	626

Table 6.35: Mean $RBC(\times 10^4/mm^3)$ counts for each treatment group at each time period

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Treatments	324360.5	5	64872.1	
Subjects(Trts)	68643.12	24	2680.1	
Time	244388.2	4	61097.1	
Trt * Time	67827.8	20	3391.4	
Error	155676.48	96	1621.63	
TOTAL	860896.1	149		

Table 6.36: The Analysis of Variance table for malotilate example

Chapter 7

Linear Regression and Correlation

In many situations, both the explanatory and response variables are numeric. We often are interested in determining whether or not the variables are linearly associated. That is, do subjects who have high measurements on one variable tend to have high (or possibly low) measurements on the other variable? In many instances, researchers will fit a linear equation relating the response variable as a function of the explanatory variable, while still allowing random variation in the response. That is, we may believe that the response variable Y can be written as:

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where x is the level of the explanatory variable, $\beta_0 + \beta_1 x$ defines the deterministic part of the response Y, and ε is a random error term that is generally assumed to be normally distributed with mean 0, and standard deviation σ. In this setting, β_0, β_1, and σ are population parameters to be estimated based on sample data. Thus, our model is that $Y|x \sim N(\beta_0 + \beta_1 x, \sigma)$.

More often, instead of reporting the estimated regression equation, investigators will report the correlation. The correlation is a measure of the strength of association between the explanatory and response variables. The correlation measures we will cover fall between -1 and 1, where values close to 1 (or -1) imply strong positive (or negative) association between the explanatory and response variables. Values close to 0 imply little or no association between the two variables.

In this chapter, we will cover estimation and inference for the simple regression model, measures of correlation, the analysis of variance table, an overview of multiple regression (when there is more than one explanatory variable). First, we give a motivating example of simple regression (models with one numeric explanatory variable), which we will build on throughout this chapter.

Example 7.1 A study was conducted to determine the effect of impaired renal function on the pharmacokinetics of gemfibrozil (Evans, et al.,1987). The primary goal was to determine whether modified dosing schedules are needed for patients with poor renal function. The explanatory variable of interest was serum creatinine clearance ($CL_{CR}(mg/dL)$), which serves as a measure of glomerular filtration rate. Patients with end stage renal disease were arbitrarily given a CL_{CR} of 5.0. Four pharmacokinetic parameters were endpoints (response variables) of interest. These were terminal elimination half–life at single and multiple doses ($t^s_{1/2}$ and $t^m_{1/2}$ in hr), and apparent

gemfibrozil clearance at single and multiple dose (CL_g^s and CL_g^m in mL/min). We will focus on the clearance variables.

Of concern to physicians prescribing this drug is whether people with lower creatinine clearances have lower gemfibrozil clearances. That is, does the drug tend to remain in the body longer in patients with poor renal function. As has been done with many drugs (see Gibaldi (1984) for many examples), it is of interest to determine whether or not there is an association between CL_{CR} and CL_g, either at single dose or multiple dose (steady state). The data are given in Table 7.1. Plots of the data and estimated regression equations are given in Figure 7.1 and Figure 7.2 for single and multiple dose cases, respectively. We will use the multiple dose data as the ongoing example throughout this chapter.

Subject	CL_{CR}	CL_g^s	CL_g^m
1	5	122	278
2	5	270	654
3	5	50	355
4	5	103	581
5	21	806	484
6	23	183	204
7	28	124	255
8	31	452	415
9	40	61	352
10	44	459	338
11	51	272	278
12	58	273	260
13	67	248	383
14	68	114	376
15	69	264	141
16	70	136	236
17	70	461	122
Mean	38.8	258.7	336.0
Std Dev	25.5	194.3	142.3

Table 7.1: Clearance data for patients with impaired renal function for gemfibrozil study

7.1 Least Squares Estimation of β_0 and β_1

We now have the problem of using sample data to compute estimates of the parameters β_0 and β_1. First, we take a sample of n subjects, observing values y of the response variable and x of the explanatory variable. We would like to choose as estimates for β_0 and β_1, the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that 'best fit' the sample data. If we define the **fitted equation** to be an equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

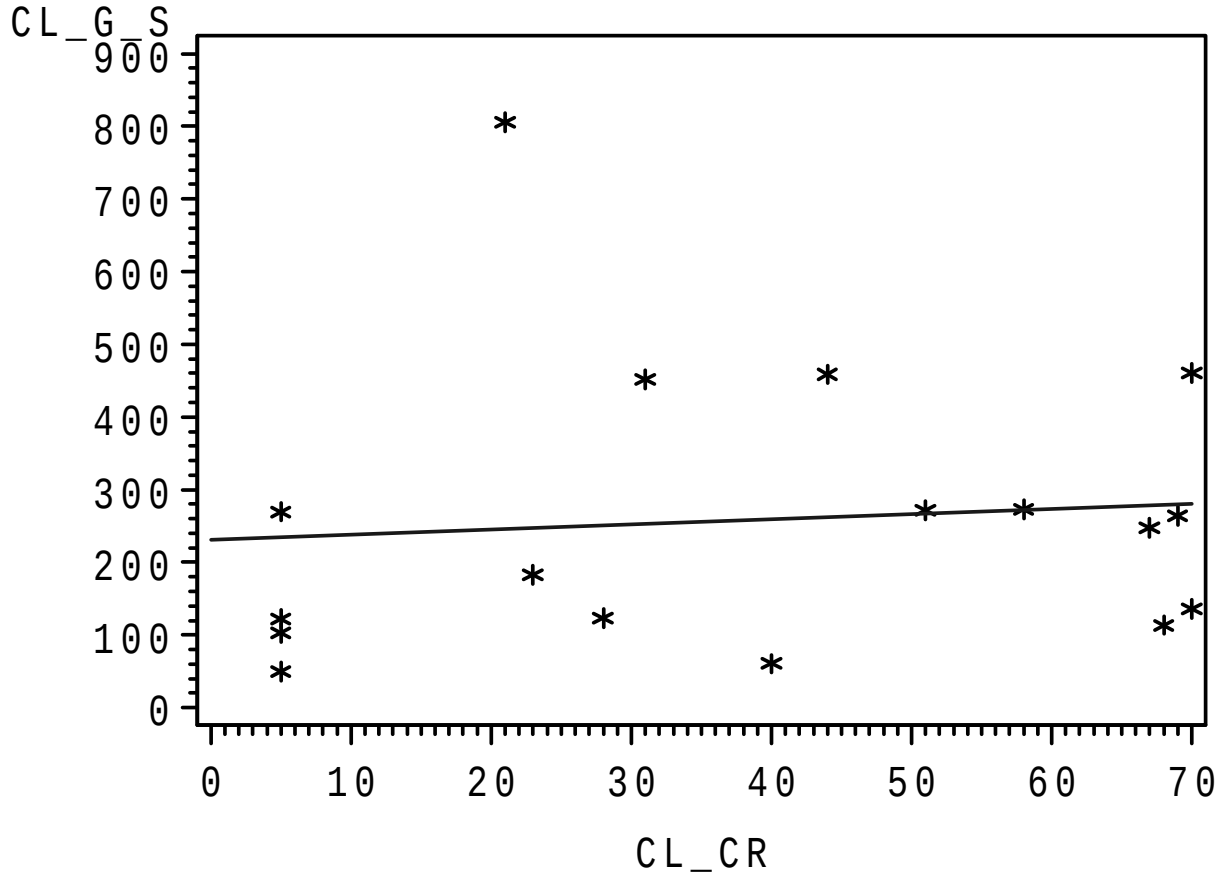

Figure 7.1: Plot of gemfibrozil vs creatininine clearance at single dose, and estimated regression line ($\hat{y} = 231.31 + 0.71x$)

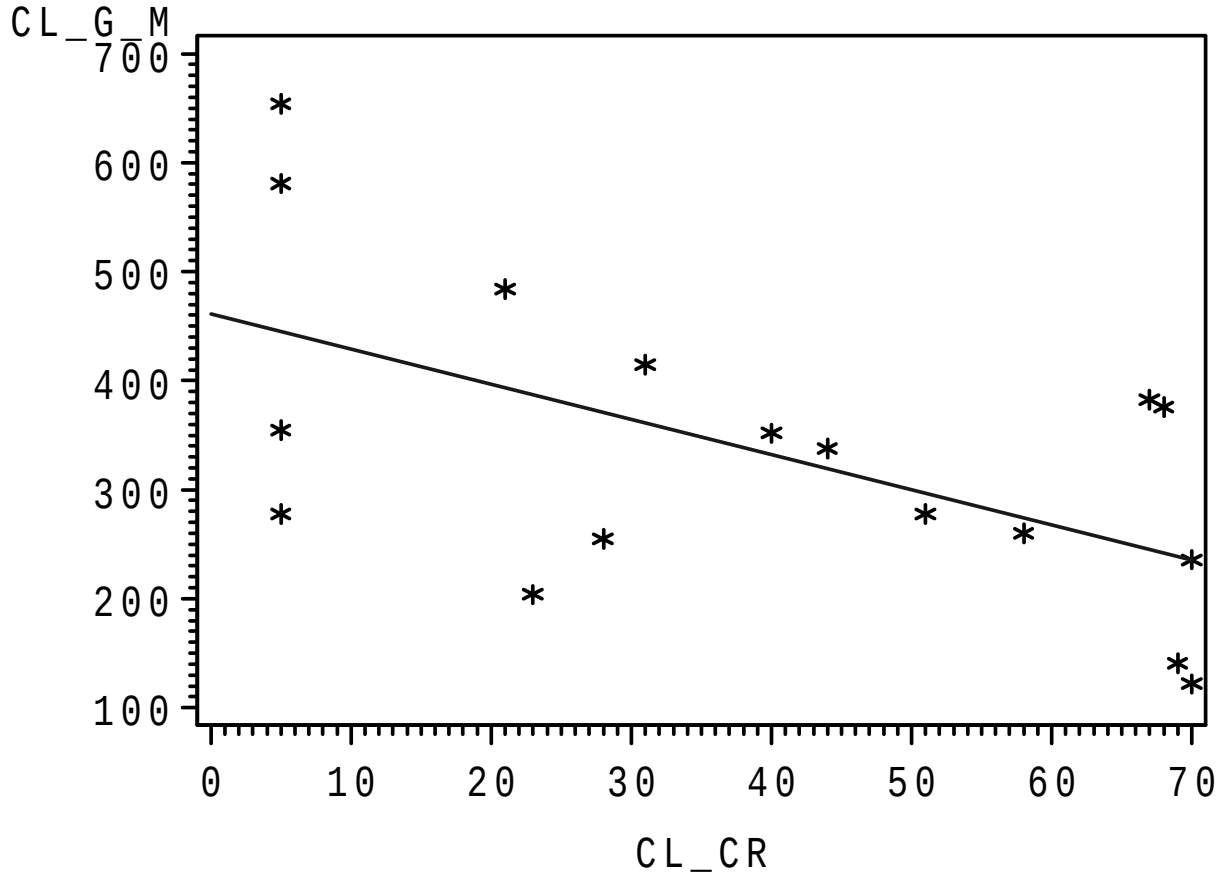

Figure 7.2: Plot of gemfibrozil vs creatininine clearance at multiple dose, and estimated regression line ($\hat{y} = 460.83 - 3.22x$)

we can choose the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to be the values that minimize the distances of the data points to the fitted line. Now, for each observed response y_i, with a corresponding predictor variable x_i, we obtain a **fitted value** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. So, we would like to minimize the sum of the squared distances of each observed response to its fitted value. That is, we want to minimize the **error sum of squares,** SSE, where:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

Three summary statistics are useful in computing regression estimates. They are:

$$S_{xx} = \sum(x - \overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{xy} = \sum(x - \overline{x})(y - \overline{y}) = \sum xy - \frac{(\sum x)(\sum y}{n}$$

$$S_{yy} = \sum(y - \overline{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

A little bit of calculus can be used to obtain the estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}},$$

and

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n}.$$

We have seen now, how to estimate β_0 and β_1. Now we can obtain an estimate of the variance of the responses at a given value of x. Recall from Chapter 1, we estimated the variance by taking the 'average' squared deviation of each measurement from the sample (estimated) mean. That is, we calculated $s^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$. Now that we fit the regression model, we know longer use $\overline{Y}$ to estimate the mean for each y_i, but rather $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ to estimate the mean. The estimate we use now looks similar to the previous estimate except we replace $\overline{Y}$ with $\hat{y}_i$ and we replace $n - 1$ with $n - 2$ since we have estimated 2 parameters, β_0 and β_1. The new estimate is:

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n-2} = \frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{n-2}.$$

This estimated variance s^2 can be thought of as the 'average' squared distance from each observed response to the fitted line.. The word average is in quotes since we divide by $n-2$ and not n. The closer the observed responses fall to the line, the smaller s^2 is and the better our predicted values will be.

Example 7.2 For the gemfibrozil example, we will estimate the regression equation and variance for the multiple dose data. In this example, our response)variable is multiple dose gemfibrozil

clearance (y), and the explanatory variable is creatinine clearance (x). Based on data in Table 7.1, we get the following summary statistics:

$$n = 17 \quad \sum x = 660.0 \quad \sum x^2 = 35990.0 \quad \sum y = 5712.0 \quad \sum y^2 = 2243266.0 \quad \sum xy = 188429.0$$

From this, we get:

$$S_{xx} = 35990.0 - \frac{(660.0)^2}{17} = 10366.5$$

$$S_{xy} = 188429.0 - \frac{(660.0)(5712.0)}{17} = -33331.0$$

$$S_{yy} = 2243266.0 - \frac{(5712.0)^2}{17} = 324034.0$$

From these computations, we get the following estimates:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-33331.0}{10366.5} = -3.22$$

$$\hat{\beta}_0 = \frac{\sum y}{n} - \hat{\beta}_1 \left(\frac{\sum x}{n} \right) = \frac{5712.0}{17} - (-3.22) \left(\frac{660.0}{17} \right) = 461.01$$

$$s^2 = \frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{n-2} = \frac{324034.0 - \frac{(-33331.0)^2}{10366.5}}{17-2} = 14457.7$$

So, we get the fitted regression equation $\hat{y} = 461.01 - 3.22x$. Patients with higher creatinine clearances tend to have lower multiple dose gemfibrozil clearance, based on the sample data. We will test whether the population parameter β_1 differs from 0 in the next section. Also, the standard deviation (from the fitted equation) is $s = \sqrt{14457.7} = 120.2$.

Note that the estimated y–intercept ($\hat{\beta}_0$) is slightly different than that given in Figure 7.2. That is due to round–off error in my hand calculations, compared to the computer values that carry many decimals throughout calculations. In most instances (except when data are very small – like decimals), these differences are trivial. The plot of the fitted equation is given in Figure 7.2.

7.1.1 Inferences Concerning β_1

Recall that in our regression model, we are stating that $E(Y|x) = \beta_0 + \beta_1 x$. In this model, β_1 represents the change in the mean of our response variable Y, as the predictor variable x increases by 1 unit. Note that if $\beta_1 = 0$, we have that $E(Y|x) = \beta_0 + \beta_1 x = \beta_0 + 0x = \beta_0$, which implies the mean of our response variable is the same at all values of x. This implies that knowledge of the level of the predictor variable does not help predict the response variable.

Under the assumptions stated previously, namely that $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$, our estimator $\hat{\beta}_1$ has a sampling distribution that is normal with mean β_1 (the true value of the parameter), and variance $\frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$. That is $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2})$. We can now make inferences concerning β_1, just as we did for $\mu, p, \mu_1 - \mu_2$, and $p_1 - p_2$ previously.

A Confidence Interval for β_1

Recall the general form of a $(1-\alpha)100\%$ confidence interval for a parameter θ (based on an estimator that is approximately normal). The interval is of the form:

$$\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}$$

for large samples, or

$$\hat{\theta} \pm t_{\alpha/2}\hat{\sigma}_{\hat{\theta}}$$

for small samples where the random error terms are approximately normal.

This leads us to the general form of a $(1-\alpha)100\%$ confidence interval for β_1. The interval can be written:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}_1} \quad \equiv \quad \hat{\beta}_1 \pm t_{\alpha/2,n-2}\frac{s}{\sqrt{S_{xx}}}.$$

Note that $\frac{s}{\sqrt{S_{xx}}}$ is the estimated standard error of $\hat{\beta}_1$ since we use $s = \sqrt{MSE}$ to estimate σ. Also, we have $n-2$ degrees of freedom instead of $n-1$, since the estimate s^2 has 2 estimated paramters used in it (refer back to how we calculate it above).

Example 7.3 For the data in Example 7.2, we can compute a 95% confidence interval for the population parameter β_1, which measures the change in mean multiple dose gemfibrozil clearance, for unit changes in creatinine clearance. Note that if $\beta_1 > 0$, then multiple dose gemfibrozil clearance is higher among patients with high creatinine clearance (lower in patients with impaired renal function). Conversely, if $\beta_1 < 0$, the reverse is true, and patients with impaired renal function tend to have higher clearances. Finally if $\beta_1 = 0$, there is no evidence of any association between creatinine clearance and multiple dose creatinine clearance. In this example, we have $n = 17$ patients, so $t_{.05/2,n-2} = t_{.025,15} = 2.131$. The 95% CI for β_1 is:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}_1} \equiv \hat{\beta}_1 \pm t_{\alpha/2,n-2}\frac{s}{\sqrt{S_{xx}}} \equiv -3.32 \pm 2.131\frac{120.2}{\sqrt{10366.5}} \equiv -3.32 \pm 2.52 \equiv (-5.84, -0.80).$$

Thus, we can conclude that multiple dose gemfibrozil clearance decreases as creatinine clearance increases. That is, the drug is removed quicker in patients with impaired renal function. Since the authors were concerned that the clearance would be **lower** in these patients, they stated that dosing schedules does not need to be altered for patients with renal insufficiency.

Hypothesis Tests Concerning β_1

Similar to the idea of the confidence interval, we can set up a test of hypothesis concerning β_1. Since the confidence interval gives us the range of 'believable' values for β_1, it is more useful than a test of hypothesis. However, here is the procedure to test if β_1 is equal to some value, say β_{10}. In virtually all real–life cases, $\beta_{10} = 0$.

1. $H_0 : \beta_1 = \beta_{10}$
2. $H_A : \beta_1 \neq \beta_{10}$ or $H_A : \beta_1 > \beta_{10}$ or $H_A : \beta_1 < \beta_{10}$ (which alternative is appropriate should be clear from the setting).

3. T.S.: $t_{obs} = (\hat{\beta}_1 - \beta_{10}) / \left(\frac{s}{\sqrt{S_{xx}}}\right)$
4. R.R.: $|t_{obs}| \geq t_{\alpha/2,n-2}$ or $t_{obs} \geq t_{\alpha,n-2}$ or $t_{obs} \leq -t_{\alpha,n-2}$ (which R.R. depends on which alternative hypothesis you are using).
5. p-value: $2P(T > |t_{obs}|)$ or $P(T > t_{obs})$ or $P(T < t_{obs})$ (again, depending on which alternative you are using).

Example 7.4 Although we've already determined that $\beta_1 \neq 0$ in Example 7.3, we will conduct the test ($\alpha = 0.05$) for completeness.

1. $H_0 : \beta_1 = 0$
2. $H_A : \beta_1 \neq 0$
3. T.S.: $t_{obs} = (\hat{\beta}_1 - 0) / \left(\frac{s}{\sqrt{S_{xx}}}\right) = -3.32 / \left(\frac{120.2}{\sqrt{10366.5}}\right) = -2.81$
4. R.R.: $|t_{obs}| \geq t_{.05/2,17-2} = t_{.025,15} = 2.131$
5. p-value: $2P(T \geq |t_{obs}|) = 2P(T \geq 2.81) = 2(.0066) = .0132$

Again, we reject H_0, and conclude that $\beta_1 \neq 0$. Also, since our test statistic is negative, and we conclude that $\beta_1 < 0$, just as we did based on the confidence interval in Example 7.3.

7.2 Correlation Coefficient

In many situations, we would like to obtain a measure of the strength of the linear association between the variables y and x. One measure of this association that is often reported in research journals from many fields is the **Pearson product moment coefficient of correlation**. This measure, denoted by r, is a number that can range from -1 to +1. A value of r close to 0 implies that there is very little association between the two variables (y tends to neither increase or decrease as x increases). A positive value of r means there is a positive association between y and x (y tends to increase as x increases). Similarly, a negative value means there is a negative association (y tends to decrease as x increases). If r is either +1 or -1, it means the data fall on a straight line ($SSE = 0$) that has either a positive or negative slope, depending on the sign of r. The formula for calculating r is:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

Note that the sign of r is always the same as the sign of $\hat{\beta}_1$. Also a test that the population correlation coefficient is 0 (no linear association between y and x) can be conducted, and is algebraically equivalent to the test $H_0 : \beta_1 = 0$. Often, the correlation coefficient, and its p–value for that test is reported.

Another measure of association that has a clearer physical interpretation than r is r^2, the coefficient of determination. This measure is always between 0 and 1, so it does not reflect whether

y and x are positively or negatively associated, and it represents the proportion of the total variation in the response variable that is 'accounted' for by fitting the regression on x. The formula for r^2 is:

$$r^2 = (r)^2 = \frac{S_{yy} - SSE}{S_{yy}}.$$

Note that $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ represents the total variation in the response variable, while $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ represents the variation in the observed responses about the fitted equation (after taking into account x). This is why we sometimes say that r^2 is "proportion of the variation in y that is 'explained' by x."

When the data (or errors) are clearly not normally distributed (large outliers generally show up on plots), a nonparametric correlation measure **Spearman's coefficient of correlation** can be computed. Spearman's measure involves ranking the x values from 1 to n, and the y values from 1 to n, then computing r as in Pearson's measure, but replacing the raw data with the ranks.

Example 7.5 For the multiple dose gemfibrozil clearance data, we compute the following values for r and r^2:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-33331.0}{\sqrt{(10366.5)(324034.0)}} = -.575 \quad r^2 = (-.575)^2 = .331$$

Note that r is negative. It will always be the same sign as $\hat{\beta}_1$. There is a moderate, negative correlation between multiple dose gemfibrozil clearance and creatinine clearance. Further, r^2 can be interpreted as the proportion of variation in multiple dose gemfibrozil clearance that is "explained" by the regression on creatinine clearance. Approximately one–third (33.1%) of the variance in gemfibrozil clearance is reduced when we use the fitted value (based on the subject's creatinine clearance) in place of the sample mean (ignoring the patient's creatinine clearance) to predict it.

7.3 The Analysis of Variance Approach to Regression

Consider the deviations of the individual responses, y_i, from their overall mean $\overline{y}$. We would like to break these deviations into two parts, the deviation of the observed value from its fitted value, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, and the deviation of the fitted value from the overall mean. This is similar in nature to the way we partitioned the total variation in the completely randomized design. We can write:

$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}).$$

Note that all we are doing is adding and subtracting the fitted value. It so happens that algebraically we can show the same equality holds once we've squared each side of the equation and summed it over the n observed and fitted values. That is,

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2.$$

These three pieces are called the **total**, **error**, and **model sums of squares**, respectively. We denote them as S_{yy}, SSE, and SSR, respectively. We have already seen that S_{yy} represents the

total variation in the observed responses, and that SSE represents the variation in the observed responses around the fitted regression equation. That leaves SSR as the amount of the total variation that is 'accounted for' by taking into account the predictor variable x. We can use this decomposition to test the hypothesis $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$. We will also find this decomposition useful in subsequent sections when we have more than one predictor variable. We first set up the **Analysis of Variance (ANOVA) Table** in Table 7.2. Note that we will have to make minimal calculations to set this up since we have already computed S_{yy} and SSE in the regression analysis.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
MODEL	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
ERROR	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	
TOTAL	$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$	$n-1$		

Table 7.2: The Analysis of Variance table for simple regression

The testing procedure is as follows:

1. $H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$ (This will always be a 2–sided test)
2. T.S.: $F_{obs} = \frac{MSR}{MSE}$
3. R.R.: $F_{obs} \geq F_{\alpha,1,n-2}$
4. p-value: $P(F \geq F_{obs})$

Note that we already have a procedure for testing this hypothesis (see the section on Inferences Concerning β_1), but this is an important lead–in to multiple regression.

Example 7.6 For the multiple dose gemfibrozil clearance data, we give the Analysis of Variance table, as well as the F–test for testing for an association between the two clearance measures. The Analysis of Variance is given in Table 7.3. The testing procedure ($\alpha = 0.05$) is as follows:

1. $H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$
2. T.S.: $F_{obs} = \frac{MSR}{MSE} = 7.41$
3. R.R.: $F_{obs} \geq F_{\alpha,1,n-2} = F_{.05,1,15} = 4.54$
4. p-value: $P(F \geq F_{obs}) = P(F \geq 7.41) = .0132$

The conclusion reached is identical to that given in Example 7.4.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
MODEL	107167.9	1	107167.9	$F = \frac{107167.9}{14457.7} = 7.41$
ERROR	216866.1	15	14457.7	
TOTAL	324034.0	16		

Table 7.3: The Analysis of Variance Table for multiple dose gemfibrozil clearance data

7.4 Multiple Regression

In most situations, we have more than one explanatory variable. While the amount of math can become overwhelming and involves matrix algebra, many computer packages exist that will provide the analysis for you. In this section, we will analyze the data by interpreting the results of a computer program. It should be noted that simple regression is a special case of multiple regression, so most concepts we have already seen apply here.

In general, if we have p explanatory variables, we can write our response variable as:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon.$$

Again, we are writing the random measurement Y in terms of its deterministic relationship to a set of p explanatory variables and a random error term, ε. We make the same assumptions as before in terms of ε, specifically that it is normally distributed with mean 0 and variance σ^2. Just as before, $\beta_0, \beta_1, \ldots, \beta_p$, and σ^2 are unknown parameters that must be estimated from the sample data. The parameters β_i represent the change in the mean response when the i^{th} explanatory variable changes by 1 unit and all other explanatory variables are held constant.

The Analysis of Variance table will be very similar to what we used previously, with the only adjustments being in the degrees of freedom. Table 7.4 shows the values for the general case when there are p explanatory variables. We will rely on computer outputs to obtain the Analysis of

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
MODEL	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
ERROR	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	
TOTAL	$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$	$n-1$		

Table 7.4: The Analysis of Variance table for multiple regression

Variance and the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ and their standard errors.

7.4.1 Testing for Association Between the Response and the Set of Explanatory Variables

To see if the set of predictor variables is useful in predicting the response variable, we will test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$. Note that if H_0 is true, then the mean response does not depend on the levels of the explanatory variables. We interpret this to mean that there is no association between the response variable and the set of explanatory variables. To test this hypothesis, we use the following procedure:

1. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad H_A$: Not every $\beta_i = 0$
2. T.S.: $F_{obs} = \frac{MSR}{MSE}$
3. R.R.: $F_{obs} \geq F_{\alpha,p,n-p-1}$
4. p-value: $P(F \geq F_{obs})$

Statistical computer packages automatically perform this test and provide you with the p-value of the test, so you really don't need to obtain the rejection region explicitly to make the appropriate conclusion. Recall that we reject the null hypothesis if the p-value is less than α.

7.4.2 Testing for Association Between the Response and an Individual Explanatory Variable

If we reject the previous null hypothesis and conclude that not all of the β_i are zero, we may wish to test whether individual β_i are zero. Note that if we fail to reject the null hypothesis that β_i is zero, we can drop the predictor x_i from our model, thus simplifying the model. Note that this test is testing whether x_i is useful **given that we are already fitting a model containing the remaining $p-1$ explanatory variables**. That is, does this variable contribute anything once we've taken into account the other explanatory variables. These tests are t-tests, where we compute $t = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$ just as we did in the section on making inferences concerning β_1 in simple regression. The procedure for testing whether $\beta_i = 0$ (the i^{th} explanatory variable does not contribute to predicting the response given the other $p-1$ explanatory variables are in the model) is as follows:

1. $H_0 : \beta_i = 0$
2. $H_A : \beta_i \neq 0$ or $H_A : \beta_i > 0$ or $H_A : \beta_i < 0$ (which alternative is appropriate should be clear from the setting).
3. T.S.: $t_{obs} = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$
4. R.R.: $|t_{obs}| \geq t_{\alpha/2,n-p-1}$ or $t_{obs} \geq t_{\alpha,n-p-1}$ or $t_{obs} \leq -t_{\alpha,n-p-1}$ (which R.R. depends on which alternative hypothesis you are using).
5. p-value: $2P(T \geq |t_{obs}|)$ or $P(T \geq t_{obs})$ or $P(T \leq t_{obs})$ (again, depending on which alternative you are using).

Computer packages print the test statistic and the p-value based on the two-sided test, so to conduct this test is simply a matter of interpreting the results of the computer output.

Example 7.7 A sample survey was conducted to investigate college students' perceptions of information source characteristics for OTC drug products (Portner and Smith,1994). Students at a wide range of colleges and universities were asked to assess five sources of OTC drug information (pharmacists, physicians, family, friends, and TV ads) on four characteristics (accuracy, convenience, expense, and time consumption). Each of these characteristics was rated from 1 (low (poor) rating) to 10 (high (good) rating) for each information source. Further, each student was given 18 minor health problem scenarios, and asked to rate the likelihood they would go to each source (pharmacist, physician, family, friends, TV ads) for information on an OTC drug. These likelihoods were given on a scale of 0 (would not go to source) to 100 (would definitely go to source), and averaged over the 18 scenarios for each source.

The authors treated the mean likelihood of going to the source as the response variable (one for each source for each student). The explanatory variables were the students' attitudinal scores on the four characteristics (accuracy, convenience, expense, and time consumption) for the corresponding source of information.

One of the regression models reported was the model fit for pharmacists. The goal is to predict the likelihood a student would use a pharmacist as a source for OTC drug information (this response, y, was the mean for the 18 scenarios), based on knowledge of the student's attitude toward the accuracy (x_1), convenience (x_2), expense (x_3), and time consumption (x_4) of obtaining OTC drug information from a pharmacist. The goal is to determine which, if any, of these attitudinal scores is related to the likelihood of using the pharmacist for OTC drug information.

The fitted equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 = 58.50 + .1494 x_1 + .2339 x_2 + .0196 x_3 - .0337 x_4$$

The Analysis of Variance for testing whether the likelihood of use is related to any of the attitudinal scores is given in Table 7.5. There were $n = 769$ subjects and $p = 4$ explanatory variables. Individual tests for the coefficients of each attitudinal variable are given in Table 7.6.

ANOVA				
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
MODEL	42503.4	4	10625.9	$F = \frac{10625.9}{513.8} = 20.68$
ERROR	392536.9	764	513.8	
TOTAL	435040.3	768		

Table 7.5: The Analysis of Variance Table for OTC drug information data

1. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad H_A$: Not every $\beta_i = 0$

2. T.S.: $F_{obs} = \frac{MSR}{MSE} = 20.68$
3. R.R.: $F_{obs} > F_{\alpha,p,n-p-1} = F_{.05,4,764} = 2.38$
4. p-value: $P(F \geq F_{obs}) = P(F \geq 20.68) < .0001$

Variable (x_i)	$\hat{\beta}_i$	$\hat{\sigma}_{\hat{\beta}_i}$	$t = \hat{\beta}_i / \hat{\sigma}_{\hat{\beta}_i}$	p–value
Accuracy (x_1)	.1494	.0359	4.162	$< .0001$
Convenience (x_2)	.2339	.0363	6.450	$< .0001$
Expense (x_3)	.0196	.0391	0.501	.6165
Time (x_4)	$-.0337$	.0393	-0.857	.3951

Table 7.6: Tests for individual regression coefficients ($H_0 : \beta_i = 0$ vs $H_A : \beta_i \neq 0$) for OTC drug information data

For the overall test, we reject H_0 and conclude that at least one of the attitudinal variables is related to likelihood of using pharmacists for OTC drug information. Based on the individual variables' tests, we determine that accuracy and convenience are related to the likelihood (after controlling all other independent variables), but that expense and time are not. Thus, pharmacists should focus on informing the public of the accuracy and convenience of their information, to help increase people's likelihood of using them for information on the widely expanding numbers of over–the–counter drugs.

Regression Models With Dummy Variables

All of the explanatory variables we have used so far were numeric variables. Other variables can also be used to handle categorical variables. There are two ways to look at this problem:

- We wish to fit separate linear regressions for seperate levels of the categorical explanatory variable (e.g. possibly different linear regressions of y on x for males and females).
- We wish to compare the means of the levels of the categorical explanatory variable after controlling for a numeric explanatory variable that differs among individuals (e.g. comparing treatments after adjusting for baseline scores of subjects).

The second situation is probably the most common in drug trials and is referred to as the **Analysis of Covariance**.

If a categorical variable has k levels, we create $k-1$ **indicator** or **dummy variables** as in the following example.

Example 7.4 A study was conducted in patients with HIV–1 to study the efficacy of thalidomide (Klausner, et al.,1996). Recall that this study was described in Example 6.2, as well. There were 16 patients who also had tuberculosis (TB^+), and 16 who did not have tuberculosis (TB^-). Among the measures reported was plasma HIV–1 RNA at day 0 (prior to drug therapy) and at

day 21 (after three weeks of drug therapy). For this analysis, we work with the natural logarithm of the values reported. This is often done to produce error terms that are approximately normally distributed when the original data are skewed. The model we will fit is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

Here, y is the log plasma HIV–1 RNA level at day 21, x_1 is the subject's baseline (day 0) log plasma HIV–1 RNA level,

$$x_2 = \begin{cases} 1 & \text{if subject received thalidomide} \\ 0 & \text{if subject received placebo} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if subject was } TB^+ \\ 0 & \text{If subject was } TB^- \end{cases}$$

We can write the deterministic portion (ignoring the random error terms) of the model for each treatment group as follows:

Placebo/TB^- $y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 x_1$

Thalidomide/TB^- $y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) = \beta_0 + \beta_1 x_1 + \beta_2$

Placebo/TB^+ $y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) = \beta_0 + \beta_1 x_1 + \beta_3$

Thalidomide/TB^+ $y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(1) = \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3$

Note that we now have a natural way to compare the efficacy of thalidomide and the effect of tuberculosis after controlling for differences in the subjects' levels of plasma HIV–1 RNA before the study (x_1). For instance:

- $\beta_2 = 0 \implies$ No thalidomide effect
- $\beta_3 = 0 \implies$ No tuberculosis effect

Estimates of the parameters of this regression function, their standard errors and tests are given in Table 7.7. Note that patients with higher day 0 scores tend to have higher day 21 scores ($\beta_1 > 0$),

Variable (x_i)	$\hat{\beta}_i$	$\hat{\sigma}_{\hat{\beta}_i}$	$t = \hat{\beta}_i / \hat{\sigma}_{\hat{\beta}_i}$	p–value
Intercept	2.662	0.635	4.19	0.0003
Day 0 RNA (x_1)	0.597	0.116	5.16	$< .0001$
Drug (x_2)	–0.330	0.258	–1.28	.2115
Tuberculosis (x_3)	–0.571	0.262	–2.18	.0379

Table 7.7: Tests for individual regression coefficients ($H_0 : \beta_i = 0$ vs $H_A : \beta_i \neq 0$) thalidomide study in HIV–1 patients

which is not surprising. Our goal is to compare the drug after adjusting for these baseline scores. We fail to reject $H_0 : \beta_2 = 0$, so after controlling for baseline score and tuberculosis state, we cannot conclude the drug significantly lowers plasma HIV–1 RNA. Finally, we do conclude that $\beta_3 < 0$, which means that after controlling for baseline and drug group, TB^+ patients tend to have lower HIV–1 RNA levels than TB^- patients.

7.5 Exercises

44. The kinetics of zidovudine in pregnant baboons was investigated in an effort to determine dosing regimens in pregnant women, with the goal to maintain AZT levels in the therapeutic range to prevent HIV infection in children (Garland, et al.,1996). As part of the study, $n = 25$ measurements of AZT concentration (y) were made at various doses (x). The values of AZT concentration ($\mu g/ml$) and dose ($mg/kg/hr$) are given in Table 7.8. Their appears to be a linear association between concentration and dose, as seen in Figure 7.3. For this data,

$$S_{xx} = \sum (x - \overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 76.7613 - \frac{(41.27)^2}{25} = 8.63$$

$$S_{xy} = \sum (x - \overline{x})(y - \overline{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n} = 27.02793 - \frac{(41.27)(14.682)}{25} = 2.79$$

$$S_{yy} = \sum (y - \overline{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 9.945754 - \frac{(14.682)^2}{25} = 1.32$$

AZT Conc.	Dose
0.169	0.67
0.178	0.86
0.206	0.96
0.391	0.6
0.387	0.94
0.333	1.12
0.349	1.4
0.437	1.17
0.428	1.35
0.597	1.76
0.587	1.92
0.653	1.43
0.66	1.77
0.688	1.55
0.704	1.51
0.746	1.82
0.797	1.91
0.875	1.89
0.549	2.5
0.666	2.5
0.759	2.02
0.806	2.12
0.83	2.5
0.897	2.5
0.99	2.5

Table 7.8: AZT concentration (y) and dose (x) in pregnant and nonpregnant baboons

(a) Computed the estimated regression equation, $\hat{y}$ and the estimated standard deviation.

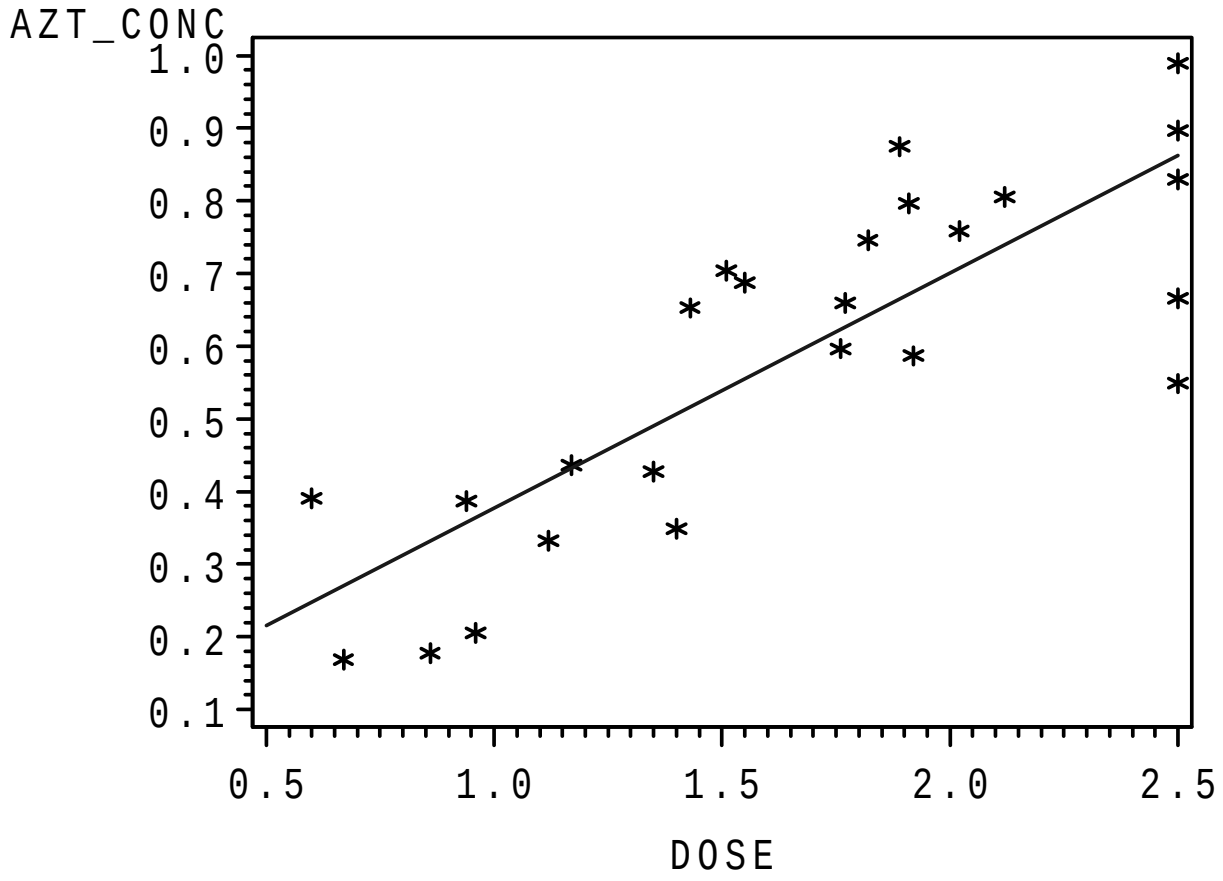

Figure 7.3: Plot of AZT concentration vs dose, and estimated regression equation for baboon study

(b) Compute a 95% CI for β_1. Can we conclude there is an association between dose and AZT concentration?

(c) Set up the Analysis of Variance table.

(d) Compute the coefficients of correlation (r) and determination (r^2).

45. A study reported the effects LSD on performance scores on a math test consisting of basic problems (Wagner, at al, 1968). The authors studied the correlation between tissue concentration of LSD (x) and performance score on arithmetic score as a percent of control (y). All measurements repesent the mean scores at seven time points among five males volunteers. A plot of the performance scores versus the tissue concentrations show a strong linear association between concentration at a non–plasma sight and pharmacological effect (Figure 7.4). The data are given in Table 7.9. For this data,

$$S_{xx} = \sum(x-\overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 153.89 - \frac{(30.33)^2}{7} = 22.47$$

$$S_{xy} = \sum(x-\overline{x})(y-\overline{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n} = 1316.66 - \frac{(30.33)(350.61)}{7} = -202.48$$

$$S_{yy} = \sum(y-\overline{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 19639.24 - \frac{(350.61)^2}{7} = 2078.19$$

Score (y)	Concentration (x)
78.93	1.17
58.20	2.97
67.47	3.26
37.47	4.69
45.65	5.83
32.92	6.00
29.97	6.41

Table 7.9: Performance scores (y) and Tissue LSD concentration (x) in LSD PK/PD study

(a) Compute the correlation coefficient between performance score and tissue concentration.

(b) What is the estimate for the change in mean performance score associated with a unit increase in tissue concentration?

(c) Do you feel the authors have demonstrated an association between performance score and tissue concntration?

(d) On the plot, identify the tissue concentration that is associated with a performance score of 50%.

46. An association between body temperature and stroke severity was observed in the Copenhagen Stroke Study (Reith, et al.,1996). In the study, the severity of the stroke (y) was measured using the Scandinavian Stroke Scale (low values correspond to higher severity) at admission. Predictor variables that were hypothesized to be associated with severity include:

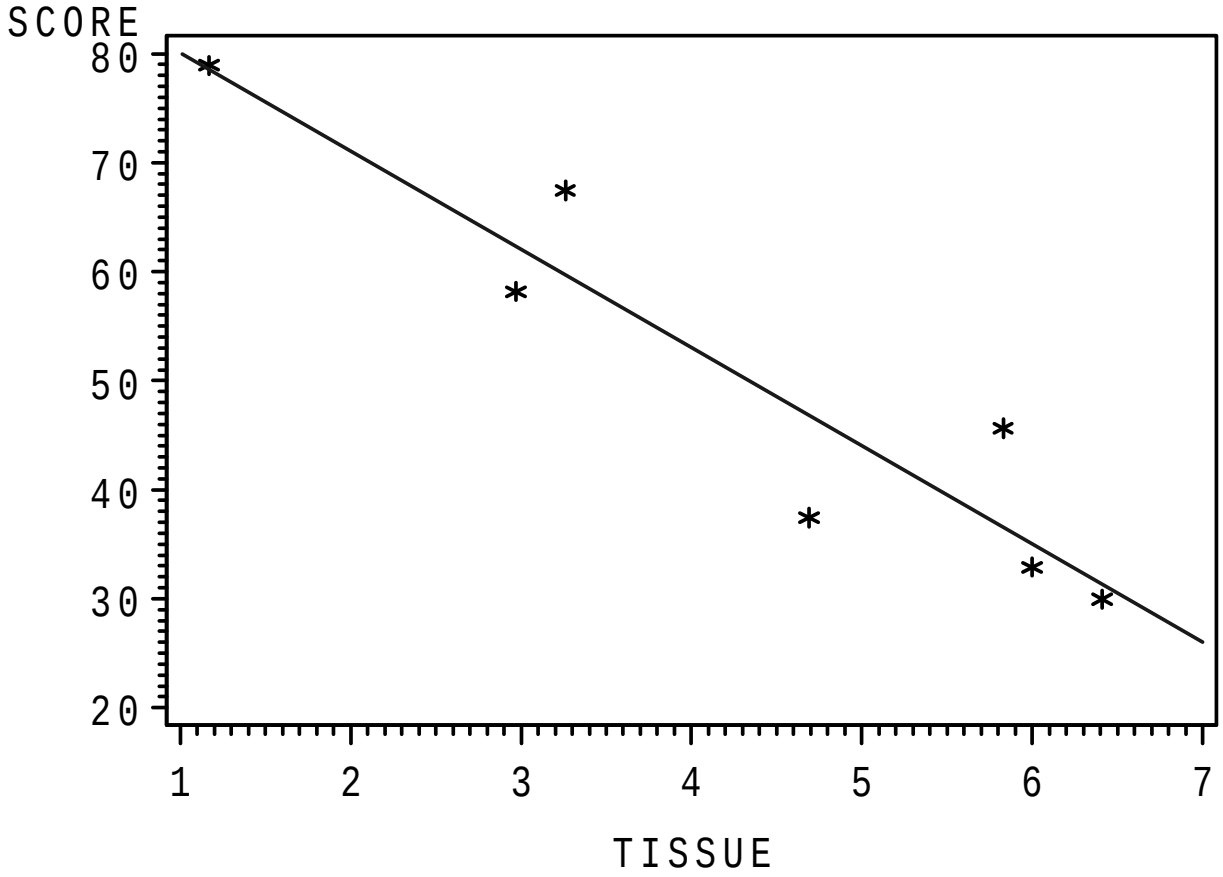

Figure 7.4: Plot of perrformance score vs tissue concentration, and estimated regression equation for LSD PK/PD study

- Body temperature (x_1=temp (celsius))
- Sex (x_2=1 if male, 0 if female)
- Previous stroke ($x_3 = 1$ if yes, 0 if no)
- Atrial fibrillation ($x_4 = 1$ if present on admission EKG, 0 if absent)
- Leukocytosis (x_5=1 if count at admission $\geq 9 \times 10^9/L$, 0 otherwise)
- Infections ($x_6 = 1$ if present at admission, 0 if not)

The regression coefficients and their corresponding standard errors are given in Table 7.10. The study was made of $n = 390$ stroke victims. Test whether or not each of these predictors is associated with stroke severity ($\alpha = 0.05$). Interpret each coefficient in terms of the direction of the association with severity (e.g. Males tend to have less severe (higher severity score) strokes than women).

Variable (x_i)	$\hat{\beta}_i$	$\hat{\sigma}_{\hat{\beta}_i}$
Intercept	171.95	
Body temp (x_1)	−3.70	1.40
Sex (x_2)	4.68	1.66
Previous stroke (x_3)	−4.56	1.91
Atrial fibrillation (x_4)	−5.07	2.05
Leucocytosis (x_5)	−1.21	0.28
Infections (x_6)	−10.74	2.43

Table 7.10: Regression coefficients and standard errors for body temperature in stroke patients data

47. Factors that may predict first–year academic success of pharmacy students at the University of Georgia were studied (Chisolm, et al,1995). The authors found that after controlling for the student's prepharmacy math/science GPA (x_1) and an indicator of whether or not the student had an undergraduate degree ($x_2 = 1$ if yes, 0 if no), the first–year pharmacy GPA (y) was not associated with any other predictor variables (which included PCAT scores). The fitted equation and coefficient of multiple determination are given below:

$$\hat{y} = 1.2619 + 0.5623x_1 + 0.3896x_2 \qquad R^2 = 0.2804$$

(a) Obtain the predicted first–year pharmacy GPA for a student with a prepharmacy math/science GPA of 3.25 who has an undergraduate degree.

(b) By how much does predicted first–year pharmacy GPA for a student with an undergraduate degree excced that of a student without a degree, after controlling for prepharmacy math/science GPA?

(c) What proportion of the variation in first–year pharmacy GPAs is "explained" by the model using prepharmacy math/science GPA and undergraduate degree status as predictors?

(d) Complete the analysis of variance in Table 7.11 for this data.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
MODEL				
ERROR				
TOTAL	12.4146	114		

Table 7.11: The Analysis of Variance Table for the first–year pharmacy GPA study

Chapter 8

Logistic and Nonlinear Regression

In this chapter we introduce two commonly used types of regression analysis. These methods are logistic and nonlinear regression.

Logistic regression is a method that is useful when the response variable is dichotomous (has two levels) and at least one of the the explanatory variable(s) is (are) continuous. In this situation, we are modeling the probability that the response variable takes on the level of interest (Success) as a function of the explanatory variable(s).

Nonlinear Regression is a method of analysis that involves fitting a nonlinear function between a numeric response variable and one or more numeric explanatory variables. In many biologic (and economic) situations, models between a response and explanatory variable(s) is nonlinear, and in fact has a functional form that is based on some theoretical model.

8.1 Logistic Regression

In many experiments, the endpoint, or outcome measurement, is dicotomous with levels being the presence or absence of a characteristic (e.g. cure, death, myocardial infarction). We have seen methods in a previous chapter to analyze such data when the explanatory variable was also dichotomous (2×2 contingency tables). We can also fit a regression model, when the explanatory variable is continuous. Actually, we can fit models with more that one explanatory variable, as we did in Chapter 7 with multiple regression. One key difference here is that probabilities must lie between 0 and 1, so we can't fit a straight line function as we did with linear regression. We will fit "S–curves" that are constrained to lie between 0 and 1.

For the case where we have one independent variable, we will fit the following model:

$$\pi(x) \quad = \quad \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Here $\pi(x)$ is the probability that the response variable takes on the characteristic of interest (success), and x is the level of the numeric explanatory variable. Of interest is whether or not $\beta = 0$. If $\beta = 0$, then the probability of success is independent of the level of x. If $\beta > 0$, then the probability of success increases as x increases, conversely, if $\beta < 0$, then the probability of success decreases

as x increases. To test this hypothesis, we conduct the following test, based on estimates obtained from a statistical computer package:

1. $H_0 : \beta = 0$
2. $H_A : \beta \neq 0$
3. T.S.: $X^2_{obs} = \left(\frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}\right)^2$
4. R.R.: $X^2_{obs} \geq \chi^2_{\alpha,1}$
5. p-value: $P(\chi^2_1 \geq X^2_{obs})$

In logistic regression, $e^{\hat{\beta}}$ is the change in the odds ratio of a success at levels of the explanatory variable one unit apart. Recall that the odds of an event occurring is:

$$o = \frac{\pi}{1-\pi} \quad \Longrightarrow \quad o(x) = \frac{\pi(x)}{1-\pi(x)}.$$

Then the ratio of the odds at $x+1$ to the odds at x (the odds ratio) can be written (independent of x) as:

$$OR(x+1, x) = \frac{o(x+1)}{o(x)} = \frac{e^{\alpha+\beta(x+1)}}{e^{\alpha+\beta x}} = e^{\beta}$$

An odds ratio greater than 1 implies that the probability of success is increasing as x increases, and an odds ratio less than 1 implies that the probability of success is decreasing as x increases. Frequently, the odds ratio, rather than $\hat{\beta}$ is reported in studies.

Example 8.1 A nonclinical study was conducted to study the therapeutic effects of individual and combined use of vinorelbine tartrate (Navelbine) and paclitaxel (Taxol) in mice given one million $P388$ murine leukemia cells (Knick, et al.,1995). One part of this study was to determine toxicity of the drugs individually and in combination. In this example, we will look at toxicity in mice given only Navelbine. Mice were given varying doses in a parallel groups fashion, and one primary outcome was whether or not the mouse died from toxic causes during the 60 day study. The observed numbers and proportions of toxic deaths are given in Table 8.1 by dose, as well as the fitted values from fitting the logistic regression model:

$$\pi(x) \quad = \quad \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

where $\pi(x)$ is the probability a mouse that received a dose of x dies from toxicity. Based on a computer analysis of the data, we get the fitted equation:

$$\hat{\pi}(x) = \frac{e^{-6.381+0.488x}}{1+e^{-6.381+0.488x}}$$

To test whether or not P(Toxic Death) is associated with dose, we will test $H_0 : \beta = 0$ vs

Navelbine Dose (mg/kg)	Total Mice	Observed Toxic Deaths	P(Toxic Death)	Fitted $\hat{\pi}(x)$
8	87	1	1/87=.012	.077
12	77	38	38/77=.494	.372
16	69	54	54/69=.783	.806
20	49	45	45/49=.918	.967
24	41	41	41/41=1.000	.995

Table 8.1: Observed and fitted (based on logistic regression model) probability of toxic death by Navelbine dose (individual drug trial)

$H_A : \beta \neq 0$. Based on computer analysis, we have:

$$\hat{\beta} = 0.488 \qquad \hat{\sigma}_{\hat{\beta}} = 0.0519$$

Now, we can conduct the test for association (at $\alpha = 0.05$ significance level):

1. $H_0 : \beta = 0$ (No association between dose and P(Toxic Death))
2. $H_A : \beta \neq 0$ (Association Exists)
3. T.S.: $X^2_{obs} = (\hat{\beta}/\hat{\sigma}_{\hat{\beta}})^2 = (0.488/0.052)^2 = 88.071$
4. R.R.: $X^2_{obs} \geq \chi^2_{0.05,1} = 3.84$
5. p-value: $P(\chi^2_1 \geq 88.071) < .0001$

A plot of the logistic regression and the observed proportions of toxic deaths is given in Figure 8.1. The plot also depicts the dose at which the probability of death is 0.5 (50% of mice would die of toxicity at this dose). This is often referred to as LD_{50}, and is 13.087 mg/kg based on this fitted equation.

Finally, the estimated odds ratio, the change in the odds of death for unit increase in dose is $OR = e^{\hat{\beta}} = e^{0.488} = 1.629$. The odds of death increase by approximately 63% for each unit increase in dose.

Multiple logistic regression can be conducted by fitting a model with more than one explanatory variable. It is similar to multiple linear regression in the sense that we can test whether or not one explanatory variable is associated with the dichotomous response variable after controlling for all other explanatory variables. We will demonstrate its use through an example.

Example 8.2 A study reported the relationship between risk of coronary heart disease and two variables: serum cholesterol level and systolic blood pressure (Cornfield, 1962). Subjects in a long–term follow–up study (prospective) in Framingham, Massachusetts were classified by their baseline serum cholesterol and systolic blood pressure. The endpoint of interest was whether or not the subject developed coronary heart disease (myocardial infarction or angina pectoris).

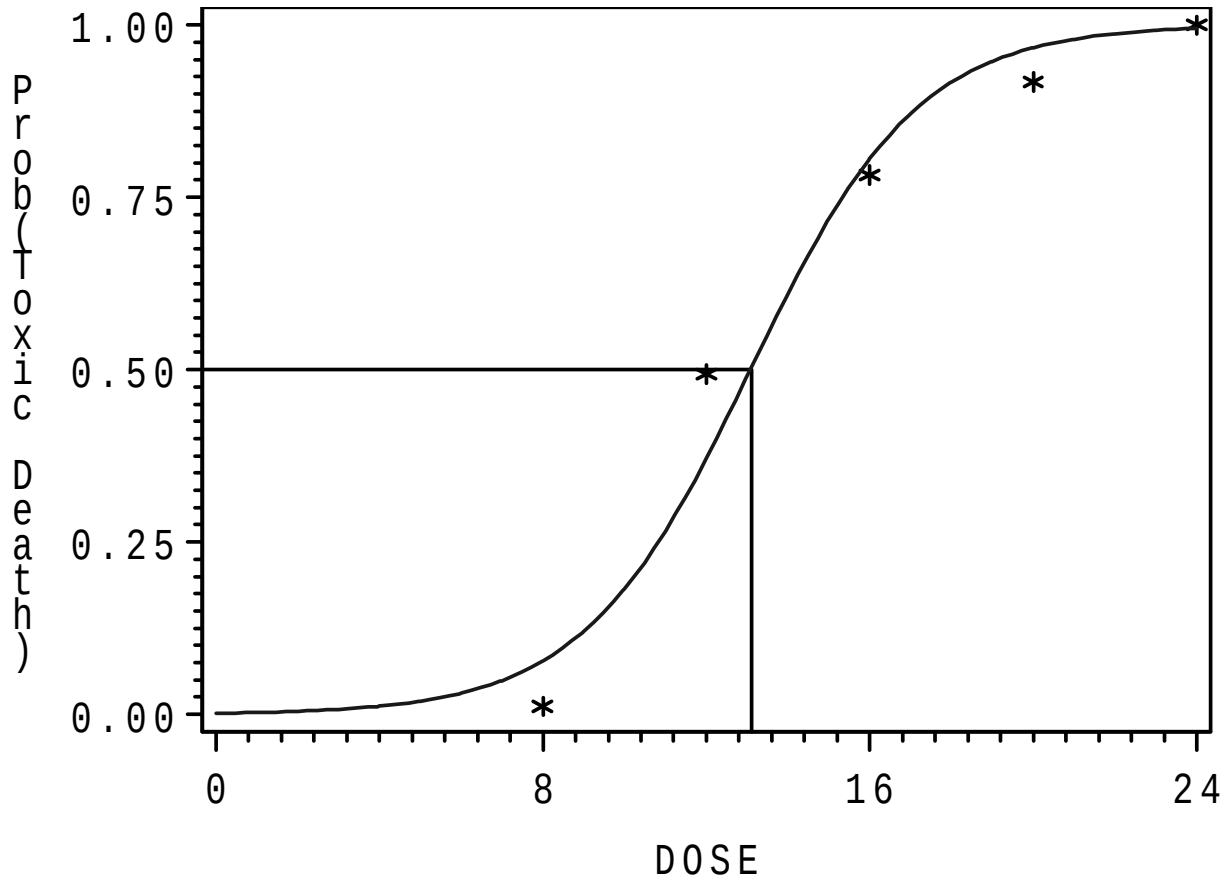

Figure 8.1: Plot of proportion of toxic deaths, estimated logistic regression curve ($\hat{\pi}(x) = \frac{e^{-6.381+.488x}}{1+e^{-6.381+.488x}}$), and LD_{50} (13.087)

Serum cholesterol levels were classified as $< 200, 200-209, 210-219, 220-244, 245-259, 260-284, > 285$. For each range, the midpoint was used as the cholesterol level for that group, and 175 and 310 were used for the lower and higher groups, respectively.

Systolic blood pressure levels were classified as $< 117, 117-126, 127-136, 137-146, 147-156, 157-166, 167-186, > 186$. For each range, the midpoint was used as the blood pressure level for that group, and 111.5 and 191.5 were used for the lower and higher groups, respectively.

The numbers of subjects and deaths for each combination of cholesterol and blood pressure are given in Table 8.2.

Blood	Serum Cholesterol						
Pressure	< 200	$200-209$	$210-219$	$220-244$	$245-259$	$260-284$	> 285
< 117	2/53	0/21	0/15	0/20	0/14	1/22	0/11
$117-126$	0/66	2/27	1/25	8/69	0/24	5/22	1/19
$127-136$	2/59	0/34	2/21	2/83	0/33	2/26	4/28
$137-146$	1/65	0/19	0/26	6/81	3/23	2/34	4/23
$147-156$	2/37	0/16	0/6	3/29	2/19	4/16	1/16
$157-166$	1/13	0/10	0/11	1/15	0/11	2/13	4/16
$167-186$	3/21	0/5	0/11	2/27	2/5	6/16	3/14
> 186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

Table 8.2: Observed CHD events/number of subjects for each serum cholesterol/systolic blood pressure group

The fitted equation is:

$$\hat{\pi} = \frac{e^{-24.50+6.57X_1+3.50X_2}}{1+e^{-24.50+6.57X_1+3.50X_2}} \qquad \hat{\sigma}_{\hat{\beta}_1} = 1.48 \quad \hat{\sigma}_{\hat{\beta}_2} = 0.84$$

where:

$$X_1 = log_{10}(\text{serum cholesterol}) \qquad X_2 = log_{10}(\text{blood pressure} - 75)$$

These transformations were chosen for theoretical considerations.

First note that we find that both serum cholesterol and systolic blood pressure levels are associated with probability of CHD (after controlling for the other variable):

1. $H_0 : \beta_1 = 0$ (No association between cholesterol and $P(\text{CHD})$)
2. $H_A : \beta_1 \neq 0$ (Association Exists)
3. T.S.: $X^2_{obs} = (\hat{\beta}_1/\hat{\sigma}_{\hat{\beta}_1})^2 = (6.57/1.48)^2 = 19.71$
4. R.R.: $X^2_{obs} \geq \chi^2_{0.05,1} = 3.84$
5. p-value: $P(\chi^2_1 \geq 19.71) < .0001$

1. $H_0 : \beta_2 = 0$ (No association between blood pressure and $P(\text{CHD})$)
2. $H_A : \beta_2 \neq 0$ (Association Exists)

3. T.S.: $X^2_{obs} = (\hat{\beta}_2/\hat{\sigma}_{\hat{\beta}_2})^2 = (3.50/0.84)^2 = 17.36$
4. R.R.: $X^2_{obs} \geq \chi^2_{0.05,1} = 3.84$
5. p-value: $P(\chi^2_1 \geq 17.36) < .0001$

Plots of the probability of suffering from CHD as a function of cholesterol level are plotted at low, middle, and high levels of blood pressure in Figure 8.2.

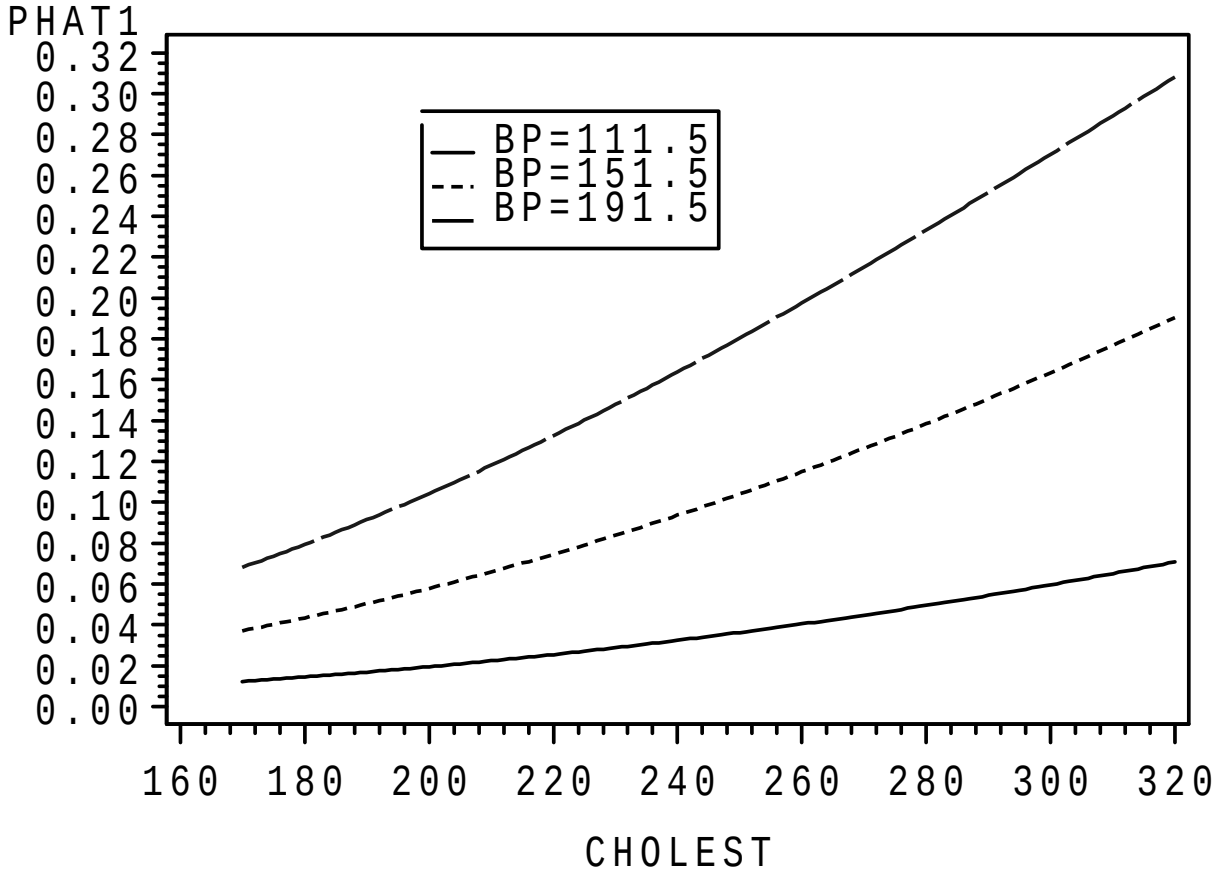

Figure 8.2: Plot of probability of CHD, as a function of cholesterol level

8.2 Nonlinear Regression

In many biologic situations, the relationship between a numeric response and numeric explanatory variables is clearly nonlinear. In many cases, a theory has been developed that describes the relationship in terms of a mathematical model. Examples of primary interest in pharmaceutics arise in the areas of pharmacokinetics anc pharmacodynamics.

In pharmacokinetics, compartmental models are theorized to describe the fate of a drug in the human body. Based on how many compartments are assumed (one or more), a mathematical model can be fit to describe the absorption and elimination of a drug. For instance, for a one–compartment model, with first–order absorption and elimination, the plasma concentration at time t ($C_p(t)$) at single dose can be written (Gibaldi (1984), p.7):

$$C_p(t) = \frac{k_a F D}{V(k_a - k_e)}[e^{-k_e t} - e^{-k_a t}]$$

where k_a is the absorption rate constant, F is the fraction of the dose (D) that is absorbed and reaches the bloodstream, V is the volume of distribution, and k_e is the elimination rate constant. Further, k_e can be written as the ratio of clearance (Cl) to volume of distribution (V), that is $k_e = Cl/V$. In this situation, experimenters often wish to estimate an individual's pharmacokinetic parameters: k_a, V, and Cl, based on observed plasma concentration measurements at various points in time.

In pharmacodynamics, it is of interest to estimate the dose–response relationship between the dose of the drug, and its therapeutic effect. It is well-known in this area of work that the relationship between dose and effect is often a "S–shape" function that is approximatley flat at low doses (no response), then takes a linear trend from a dose that corresponds to approximately 20% of maximum effect to a dose that yields 80% of maximum effect, then flattens out at higher doses (maximum effect). One such function is referred to as the sigmoid–E_{max} relationship (Holford and Sheiner, 1981):

$$E = \frac{E_{max} \cdot C^N}{EC_{50}^N + C^N}$$

where E is the effect, E_{max} is the maximum effect attributable to the drug, C is the drug concentration (typically in plasma, not at the effect site), EC_{50} is the concentration producing an effect 50% of E_{max}, and N is a parameter that controls the shape of the response function.

Example 8.3 A study was conducted in five AIDS patients to study pharmacokinetics, safety, and efficacy of an orally delivered protease inhibitor MK–639 (Stein, et al.,1996). In this phase I/II trial, one goal was to assess the relationship between effectiveness (as measured by changes in log_{10} HIV RNA copies/ml from baseline (y)) and drug concentration (as measured by $AUC_{0-6h}(x)$). High values of y correspond to high inhibition of HIV RNA generation. The sigmoid–E_{max} function can be written as:

$$y = \frac{\beta_0 x^{\beta_2}}{x^{\beta_2} + \beta_1^{\beta_2}}$$

Parameters of interest are: β_0 which is E_{max} (maximum effect), and β_1 which is the value of x that produces a 50% effect (ED_{50}). Note that this is a very small study, so that the estimates of these parameters will be very imprecise (large confidence intervals). The data (or at least a very close approximation) are given in Table 8.3.

A computer fit of the sigmoid–E_{max} function produces the following estimated equation:

$$\hat{y} = \frac{3.47 x^{35.23}}{x^{35.23} + 18270.0^{35.23}}$$

Subject	log_{10} RNA change (y)	$AUC_{0-6h}(nM \cdot h)$ (x)
1	0.000	10576.9
2	0.167	13942.3
3	1.524	18235.3
4	3.205	19607.8
5	3.518	22317.1

Table 8.3: Log_{10} HIV RNA change and drug concentrations (AUC_{0-6}) for MK639 efficacy trial

So, our estimate of the maximum effect is 3.47, and the estimate of the AUC_{0-6} value producing 50% effect is 18270.0. A plot of the data and the fitted curve are given in Figure 8.3.

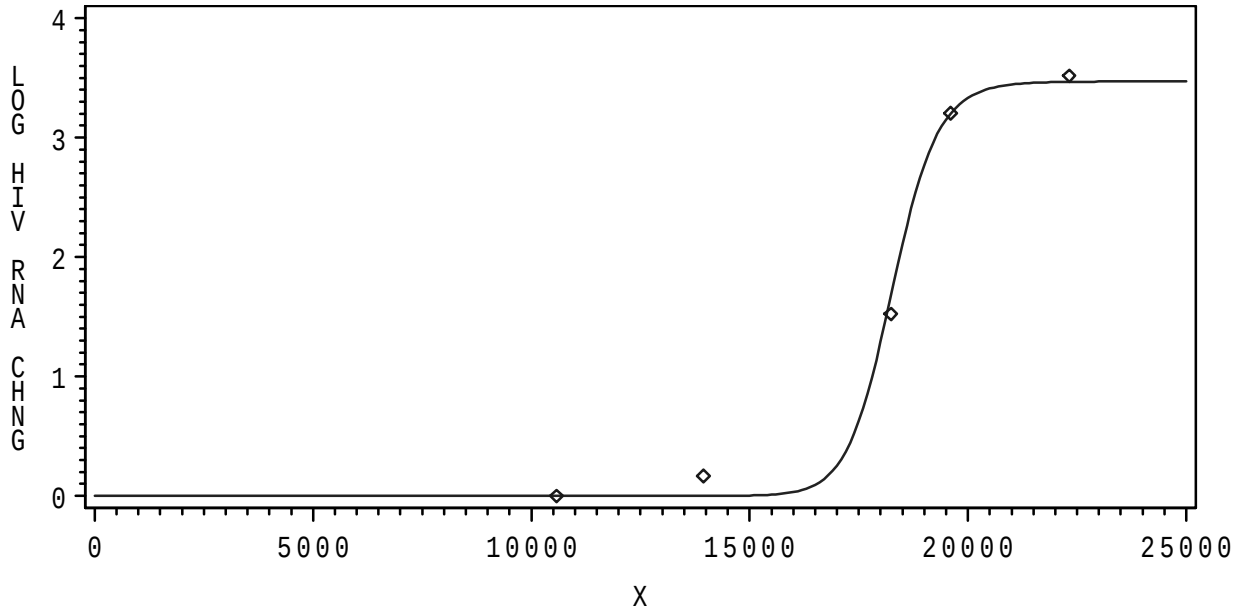

Figure 8.3: Plot of log_{10} HIV RNA change vs AUC_{0-6}, and estimated nonlinear regression equation $\hat{y} = \frac{3.47x^{35.23}}{x^{35.23}+18270.0^{35.23}}$

8.3 Exercises

48. Several drugs were studied in terms of their effects in inhibiting audiogenic seizures in rats (Corn, et al.,1955). Rats that were susceptible to audiogenic seizures were given several drugs at various dosage levels in an attempt to determine whether or not the drug inhibited seizure when audio stimulus was given. Rats were tested again off drug, to make certain that the rat had not become 'immune' to audiogenic seizures throughout the study. One drug studied was sedamyl, at doses

25,50,67 80, and 100 mg/kg. Table 8.4 gives the number of rats tested in each dosage group (after removing 'immunized' rats), the numbers that have no seizure (display drug inhibition), the sample proportion inhibited, and the fitted value from the logistic regression model. Figure 8.4 plots the sample proportion, the fitted equation and ED_{50} values. The estimated logistic regression equation is:

$$\hat{\pi}(x) = \frac{e^{-4.0733+0.0694x}}{1+e^{-4.0733+0.0694x}}$$

(a) Test $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$ at $\alpha = 0.05$. ($\hat{\sigma}_{\hat{\beta}} = 0.0178$)

(b) By how much does the (estimated) odds of no seizure for unit increase in dose?

(c) Compute the (estimated) dose that will provide inhibition in 50% of rats (ED_{50}). Hint: when $\hat{\alpha} + \hat{\beta}x = 0$, $\hat{\pi}(x) = 0.5$.

Sedamyl		Observed		Fitted
Dose (mg/kg)	Total Rats	# without seizure	P(No seizure)	$\hat{\pi}(x)$
25	11	0	0/11=0.000	.088
50	25	11	11/25=.440	.354
67	5	3	3/5=.600	.640
80	14	10	10/14=.714	.814
100	8	8	8/8=1.000	.946

Table 8.4: Observed and fitted (based on logistic regression model) probability of no audiogenic seizure by sedamyl dose

49. A bioequivalence study of two enalapril maleate tablet formulations investigated the pharmacodynamics of enalaprilat on angiotensin converting enzyme (ACE) activity (Ribeiro, et al.,1996). Two formulations of enalapril maleate (a pro–drug of enalaprilat) were given to 18 subjects in a two period crossover design (Eupressin tablets 10 mg, Biosentica (test) vs Renitec tablets 10 mg, Merck (reference)). In the pharmacodynamic part of the study, mean enalaprilat concentration ($ng \cdot ml^{-1}$) and mean % ACE inhibition were measured across patients at each of 13 time points where concentration measurements were made, for each drug. Thus, each mean is the average among subjects, and there are 13(2)=26 means.

A nonlinear model based on a single binding site Michaelis–Mentin relation of the following form was fit, where y is the mean % ACE inhibition, and x is the mean enalaprilat concentration.

$$y = \frac{\beta_1 x}{\beta_2 + x}$$

In this model, β_1 is the maximum effect (maximum attainable ACE inhibition) and β_2 is the EC_{50}, the concentration that produces 50% of the maximum effect. A plot of the data and estimated regression equation are given in Figure 8.5. The estimated equation is:

$$\hat{y} = \frac{\hat{\beta}_1 x}{\hat{\beta}_2 + x} = \frac{93.9809x}{3.8307 + x}$$

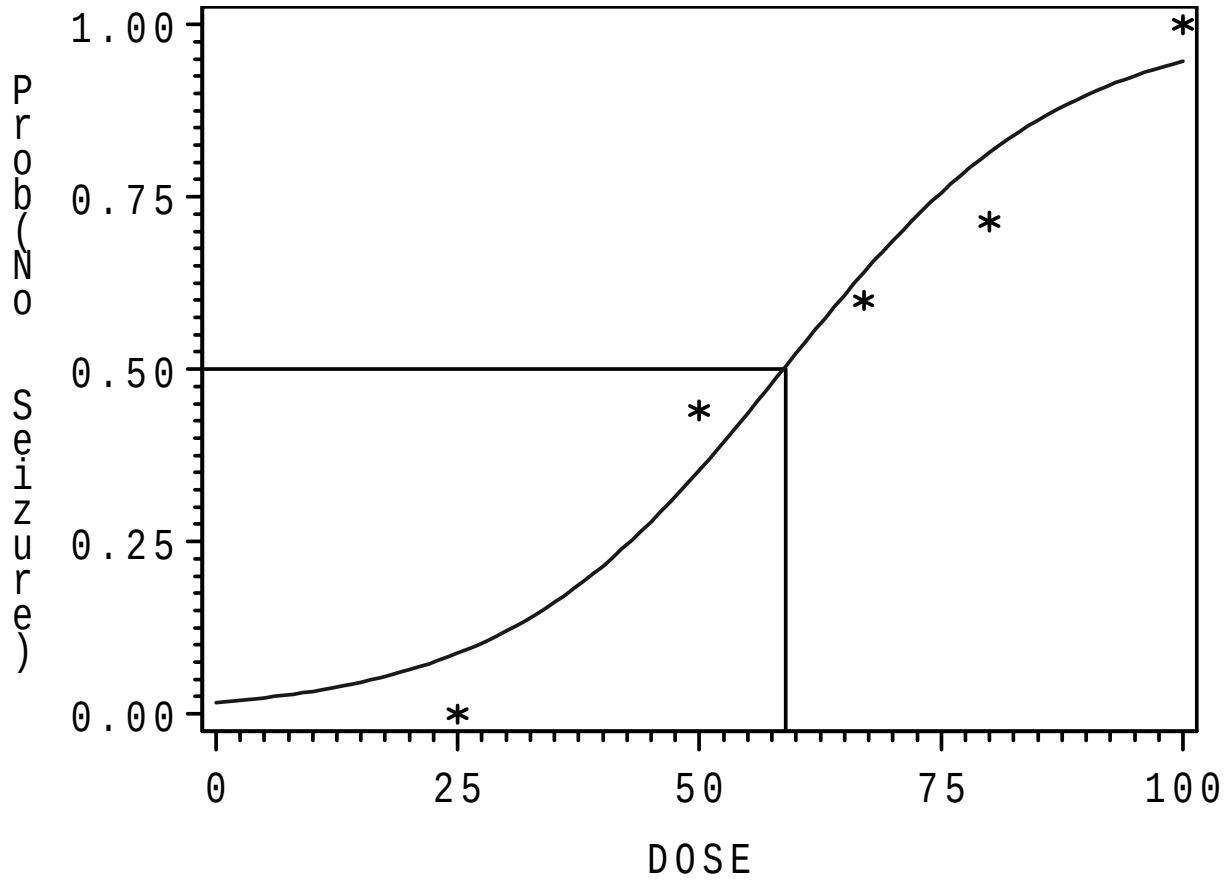

Figure 8.4: Plot of proportion of inhibited audiogenic seizures, estimated logistic regression curve ($\hat{\pi}(x) = \frac{e^{-4.0733+.0694x}}{1+e^{-4.0733+.0694x}}$), and ED_{50}

(a) Obtain the fitted % ACE mean inhibitions for mean concentrations of 3.8307, 10.0, 50.0, 100.0.

(b) The estimated standard error of $\hat{\beta}_2$ is $\hat{\sigma}_{\hat{\beta}_2}$=.2038. Compute a large–sample 95% CI for EC_{50}.

(c) Why might measurement of ACE activity be a good measurement for comparing bioavailability of drugs in this example?

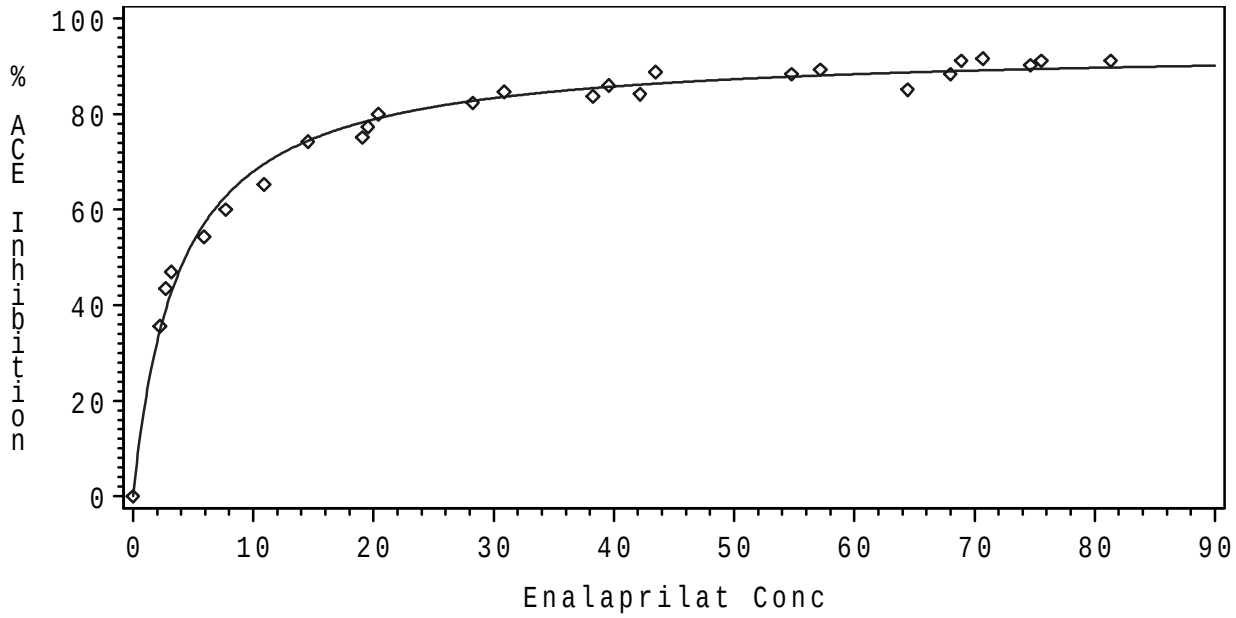

Figure 8.5: Plot of proportion of inhibited audiogenic seizures, estimated logistic regression curve ($\hat{y} = \frac{93.9809x}{3.8307+x}$).

50. A study combined results of Phase I clinical trials of orlistat, and inhibitor of gastric and pancreatic lipases (Zhi, et al.,1994). One measure of efficacy reported was fecal fat excretion (orlistat's purpose is to inhibit dietary fat absorption, so high fecal fat exretion is consistent with efficacy). The authors fit a simple maximum–effect (E_{max}) model, relating excretion, E, to dose, D, in the following formulation:

$$E = E_0 + \frac{E_{max} \cdot D}{ED_{50} + D} = \beta_0 + \frac{\beta_1 x}{\beta_2 + x}$$

where E is the intensity of the treatment effect, E_0 is the intensity of a placebo effect, E_{max} is the maximum attainable intensity of effect from orlistat, and ED_{50} is the dose that produces 50% of the maximal effect. The fitted equation (based on my attempt to read the data from their plot) is:

$$\hat{E} = 6.115 + \frac{27.620 \cdot D}{124.656 + D}$$

Note that all terms are significant. The data and the fitted equation are displayed in Figure 8.6.

(a) Obtain the fitted value for subjects receiving D = 50, 100, 200, and 400 mg/day of orlistat. Would you suspect that the effect will be higher at higher levels of D?

(b) How would you describe the variation among subjects? Focus on a given dose, do subjects at that dose tend to give similar responses?

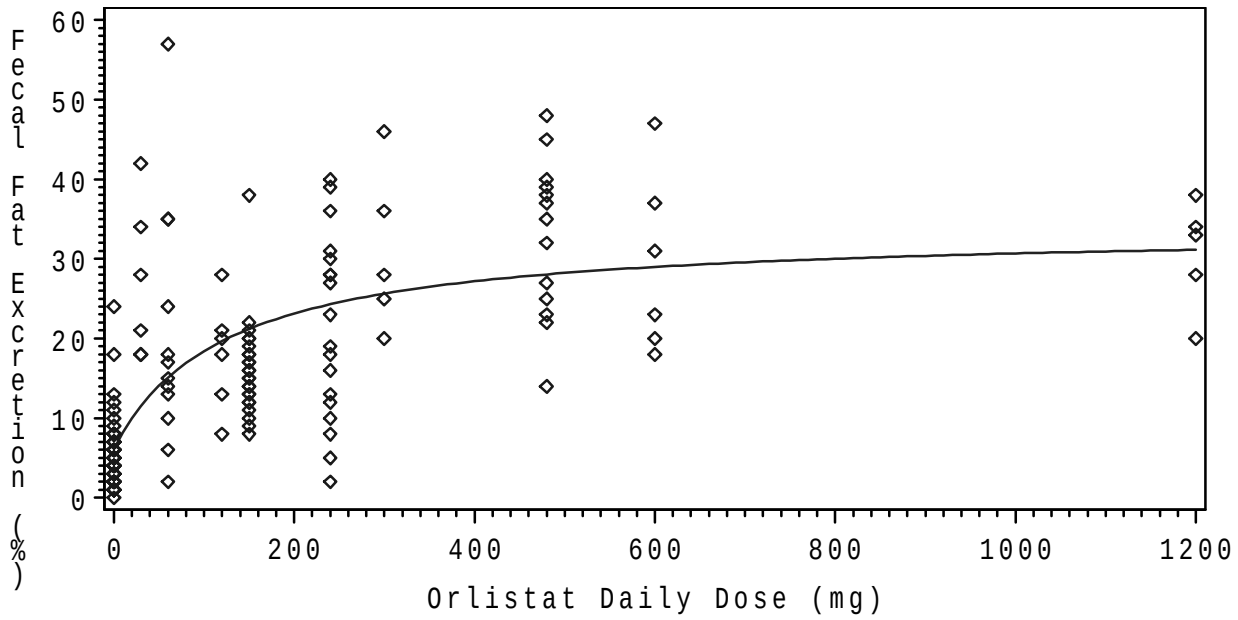

Figure 8.6: Plot of percent fecal fat excretion, estimated maximum effect (E_{max}) regression curve ($\hat{E} = 6.115 + \frac{27.620 \cdot D}{124.656 + D}$)

Chapter 9

Survival Analysis

In many experimental settings, the endpoint of interest is the time until an event occurs. Often, the event of interest is death (thus the name *survival analysis*), however it can be any event that can be observed. One problem that distinguishes survival analysis from other statistical methods is *censored* data. In these studies, people may not have the event of interest occur during the study period. However, we do have information from these subjects, so we don't simply discard their information. That is, if we have followed a subject for 3.0 years at the time the study ends, and he/she has not died, we know that the subject's survival time is greater than 3.0 years.

Their are several useful functions that describe populations of survival times. The first is the survival function ($S(t)$). In this chapter, we will call our random variable T, which is a randomly selected subject's survival time. The survival function can be written as:

$$S(t) = P(T > t) = \frac{\text{\# of subjects in population with } T > t}{\text{\# of subjects in population}}$$

This function is assumed to be continuous, with $S(0) = 1$ and $S(\infty) = 0$. A second function that defines a survival distribution is the hazard function ($\lambda(t)$). The hazard function can be thought of as the instantaneous failure rate at time t, among subjects who have survived to that point, and can be written as:

$$\lambda(t) = \frac{lim_{\Delta t \to 0} P\{T \in (t, t + \Delta t] | T > t\}}{\Delta t}$$

This function is very important in modelling survival times, as we will see in the section on proportional hazards models.

9.1 Estimating a Survival Function — Kaplan–Meier Estimates

A widely used method in the description of survival among individuals is to estimate and plot the survival distribution as a function of time. The most common estimation method is the product limit method (Kaplan and Meier,1958). Using notation given elsewhere (Kalbfleisch and Street (1990), pp.322–323), we define the following terms:

- Data: $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, where t_i is the i^{th} subject's observed time to failure (death) or censoring, and δ_i is an indicator (1=censored, 0=not censored (actual failure)).

- Observed Failure Times: $t_{(1)} < \cdots < t_{(k)}$, each failure time $t_{(i)}$ having associated with it d_i failures. Subjects who are censored at $t_{(i)}$ are treated as if they had been censored between $t_{(i)}$ and $t_{(i+1)}$
- Number of Items Censored in Time Interval: m_i, the number of censored subjects in the time interval $[t_{(i)}, t_{(i+1)})$. These subjects are all "at risk of failure" at time $t_{(i)}$, but not at $t_{(i+1)}$.
- Number of Subjects at Risk Prior to $t_{(i)}$: $n_i = \sum_{j=i}^{k}(d_j + m_j)$, the number of subjects with failure times or censored times of $t_{(i)}$ or greater.
- Estimated Hazard at Time $t_{(i)}$: $\hat{\lambda}_i = \frac{d_i}{n_i}$, the proportion of those at risk just prior to $t_{(i)}$ who fail at time $t_{(i)}$.
- Estimated Survival Function at Time t: $\hat{S}(t) = \prod_{i|t_{(i)} \leq t}(1 - \hat{\lambda}_i)$, the probability that a subject survives beyond time t.

Statistical computer packages can compute this estimate ($\hat{S}(t)$), as well as provide a graph of the estimated survival function as a function of time, even for large sample sizes.

Example 9.1 A nonclinical trial was conducted in mice who had received one million $P388$ murine leukemia cells (Knick, et al.,1995). The researchers discovered that by giving the mice a combination therapy of vinorelbine tartrate (Navelbine) and paclitaxel (Taxol), they increased survival and eliminated toxicity, which was high for each of the individual drug therapies (see Example 8.1).

Once this combination was found to be successful, a problem arises in determining the dosing regimen (doses and timing of delivery). Two of the more successful regimens were:

Regimen A 20 mg/kg Navelbine plus 36 mg/kg Taxol, concurrently.

Regimen B 20 mg/kg Navelbine plus 36 mg/kg Taxol, 1–hour later.

In regimen A, there were $n_A = 49$ mice, of which 9 died, on days 6,8,22,32,32,35,41,46,and 54, respectively. The other 40 mice from regimen A survived the entire 60 days and were 'censored'.

In regimen B, there were $n_B = 15$ mice, of which 9 died, on days 8,10,27,31,34,35,39,47, and 57, respectively. The other 6 mice from regimen B survived the entire 60 days and were 'censored'.

We will now construct the Kaplan–Meier estimates for each of the the drug delivery regimens, and plot the curves. We will follow the descriptions given above in completing Table 9.1. Note that $t_{(i)}$ is the i^{th} failure time, d_i is the number of failures at $t_{(i)}$, $n_{(i)}$ is the number of subjects at risk (with failure or censor times greater than $t_{(i)}$) at $t_{(i)}$, $\hat{\lambda}_i = d_i/n_i$ is the proportion dying at $t_{(i)}$ among those at risk, and $\hat{S}(t_{(i)})$ is the probability of surviving past time $t_{(i)}$.

A plot of these functions is given in Figure 9.1. Note that the curve for regimen A is 'higher' than that for regimen B. It appears that by delivering the Navelbine and Taxol concurrently, we improve survival as opposed to waiting 1–hour to deliver Taxol, when using these doses. We will conduct a test for equivalent survival distributions in the next section. For an interesting comparison, refer back to Example 8.1, to see the probability of suffering death from toxicity at the 20 mg/kg dose of Navelbine. Clearly, taking the two drugs in combination is improving survival.

Regimen A						Regimen B					
i	$t_{(i)}$	n_i	d_i	$\hat{\lambda}_i$	$\hat{S}(t_{(i)})$	i	$t_{(i)}$	n_i	d_i	$\hat{\lambda}_i$	$\hat{S}(t_{(i)})$
1	6	49	1	.020	.980	1	8	15	1	.067	.933
2	8	48	1	.021	.959	2	10	14	1	.071	.867
3	22	47	1	.021	.939	3	27	13	1	.077	.800
4	32	46	2	.043	.899	4	31	12	1	.083	.733
5	35	44	1	.023	.878	5	34	11	1	.091	.667
6	41	43	1	.023	.858	6	35	10	1	.100	.600
7	46	42	1	.024	.837	7	39	9	1	.111	.533
8	54	41	1	.024	.817	8	47	8	1	.125	.467
–	–	–	–	–	–	9	57	7	1	.143	.400

Table 9.1: Kaplan–Meier estimates of survival distribution functions for two dosing regimens

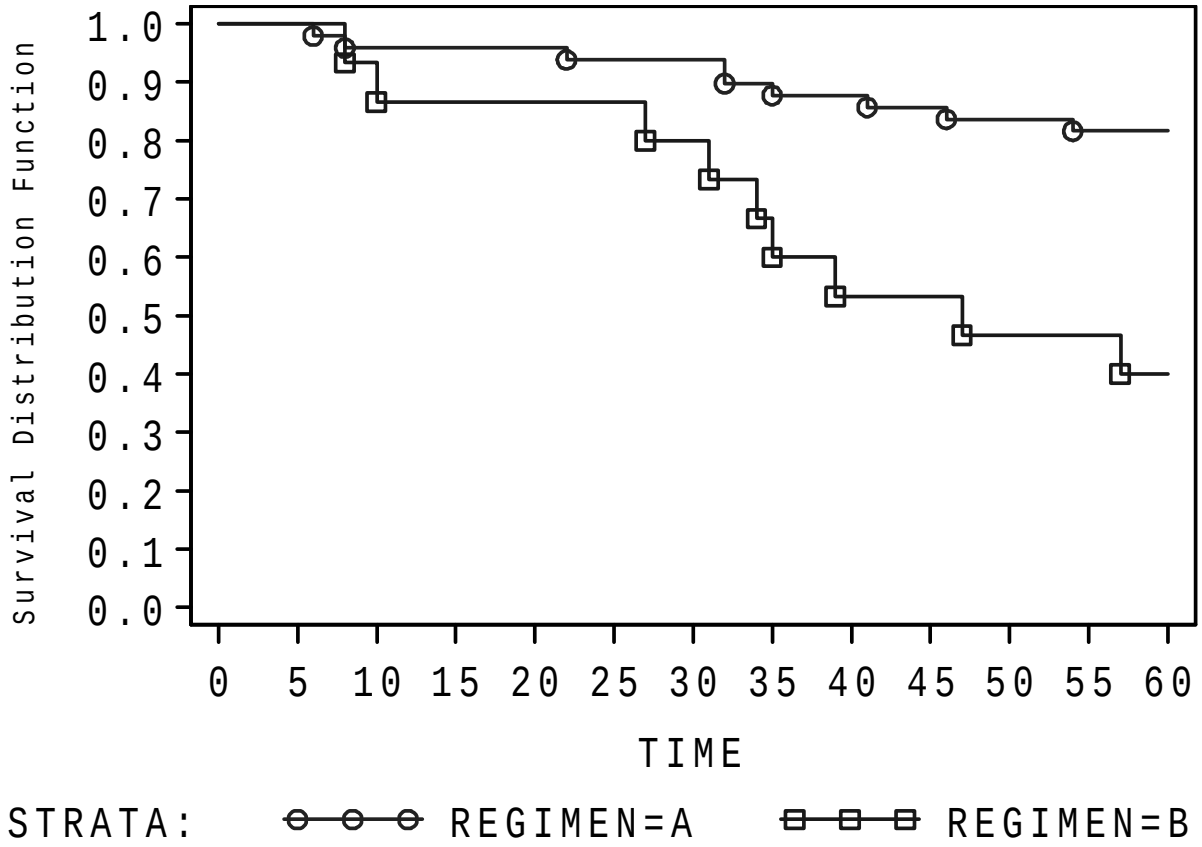

Figure 9.1: Kaplan–Meier estimates of survival functions for regimen A (concurrent) and regimen B (1–hour delay)

9.2 Log–Rank Test to Compare 2 Population Survival Functions

Generally, we would like to compare 2 (or more) survival functions. That is, we may like to compare the distribution of survival times among subjects receiving an active drug to that of subjects receiving a placebo. Note that this situation is very much like the comparisons we made between two groups in Chapter 3 (and comparing $k > 2$ groups in Chapter 6. Again, we will use the notation given elsewhere (Kalbfleisch and Street (1990), pp. 327–328). We will consider only the case where we have two groups (treatment and control). Extensions can easily be made to more than 2 groups.

We set up k 2×2 contingency tables, one at each failure time $t_{(i)}$ as described in the previous section. We will also use the same notation for subjects at risk (within each group) and subjects failing at the current time (again, for each group). At each failure time, we obtain a table like that given in Table 9.2.

	Failures	Survivals	At Risk
Treatment	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
Control	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Table 9.2: 2×2 table of Failures and Survivals at Failure Time $t_{(i)}$

We can then test whether or not the two survival functions differ by computing the following statistics, and conducting the log–rank test, described below:

$$e_{1i} = \frac{n_{1i}d_i}{n_i} \quad v_{1i} = \frac{n_{1i}n_{2i}d_i(n_i - d_i)}{n_i^2(n_i - 1)} \qquad O_1 - E_1 = \sum_{i=1}^{k}(d_{1i} - e_{1i}) \quad V_1 = \sum_{i=1}^{k} v_{1i}$$

1. H_0 : Treatment and Control Survival Functions are Identical (No treatment effect)
2. H_A : Treatment and Control Survival Functions Differ (Treatment effects)
3. T.S.: $T_{MH} = \frac{0_1 - E_1}{\sqrt{V_1}}$
4. R.R.: $|T_{MH}| \geq z_{\alpha/2}$
5. p–value: $2P(Z \geq |T_{MH}|)$
6. Conclusions: Significant positive test statistics imply that subjects receiving treatment fail quicker than controls, negative test statistics imply that controls fail quicker than those receiving treatment (treatment prolongs life in the case where failure is death).

Example 9.2 For the survival data in Example 9.1, we would like to formally test for differences in the survival distributions for the two dosing regimens. In this case, there are 15 distinct failure times (days 6,8,10,22,27,31,32,34,35,39,41,46,47,54,57). We will denote regimen A as treatment 1. All relevant quantities and computations are given in Table 9.3.

Failure	Regimen A		Regimen B			
Time (i)	d_{1i}	n_{1i}	d_{2i}	n_{2i}	e_{1i}	v_{1i}
6 (1)	1	49	0	15	0.766	.1794
8 (2)	1	48	1	15	1.524	.3570
10 (3)	0	47	1	14	0.770	.1768
22 (4)	1	47	0	13	0.783	.1697
27 (5)	0	46	1	13	0.780	.1718
31 (6)	0	46	1	12	0.793	.1641
32 (7)	2	46	0	11	1.614	.3059
34 (8)	0	44	1	11	0.800	.1600
35 (9)	1	44	1	10	1.630	.2961
39 (10)	0	43	1	9	0.827	.1431
41 (11)	1	43	0	8	0.843	.1323
46 (12)	1	42	0	8	0.840	.1344
47 (13)	0	41	1	8	0.837	.1366
54 (14)	1	41	0	7	0.854	.1246
57 (15)	0	40	1	7	0.851	.1268
Sum	9	–	9	–	14.512	2.7786

Table 9.3: Computation of observed and expected values for log–rank test to compare survival functions of two dosing regimens

We now test to determine whether or not the two (population) survival functions differ ($\alpha = 0.05$):

1. H_0 : Regimen A and Regimen B Survival Functions are Identical (No treatment effect)
2. H_A : Regimen A and Regimen B Survival Functions Differ (Treatment effects)
3. T.S.: $T_{MH} = \frac{0_1 - E_1}{\sqrt{V_1}} = \frac{9 - 14.512}{\sqrt{2.7786}} = -3.307$
4. R.R.: $|T_{MH}| \geq z_{\alpha/2} = z_{.025} = 1.96$
5. p–value: $2P(z \geq 3.307) = .0009$

We reject H_0, and since the test statistic is negative for regimen A, there were fewer combined deaths than expected for that treatment. Regimen A provides higher survival rates (at least up to 60 days) than regimen B.

It should be noted that some computer packages report a chi–squared statistic. That statistic is computed as follows:

$$O_2 = \sum d_{2i} \quad E_2 = O_1 + O_2 - E_2 \quad X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Then X^2 is compared with $\chi^2_{\alpha,1}$. This test can also be extended to compare $k > 2$ populations (see Kalbfleish and Street (1990)).

9.3 Relative Risk Regression (Proportional Hazards Model)

In many situations, we have additional information on individual subjects that we believe may be associated with risk of the event of interest occuring. For instance, age, weight, and sex may be associated with death, as well as which treatment a patient receives. As with multiple regression, the goal is to test whether a certain factor (in this case, treatment) is related to survival, after controlling for other factors such as age, weight, sex, etc.

For this model, we have p explanatory variables (as we did in multiple regression), and we will write the relative risk of a subject who has observed levels $x_1, \ldots, x_p$ (relative to a subject with each explanatory variable equal to 0) as:

$$RR(t; x_1, \ldots, x_p) = \frac{\lambda(t; x_1, \ldots, x_p)}{\lambda(t; 0, \ldots, 0)} = \frac{\lambda(t; x_1, \ldots, x_p)}{\lambda_0(t)}$$

Recall that the relative risk is the ratio of the probability of event for one group relative to another (see Chapter 5), and that the hazard is a probability of failure (as a function of time). One common model for the relative risk is to assume that it is constant over time, which is referred to as the **proportional hazards model**. A common model (log–linear model) is:

$$RR(t; x_1, \ldots, x_p) = e^{\beta_1 x_1 + \cdots + \beta_p x_p}$$

Consider this situation. A drug is to be studied for efficacy at prolonging the life of patients with a terminal disease. Patients are classified based on a scale of 1–5 in terms of the stage of the disease (x_1) (1–lowest, 5–highest), age (x_2), weight (x_3), and a dummy variable (x_4) indicating whether or not the patient received active drug ($x_4 = 1$) or placebo ($x_4 = 0$). We fit the following relative regression model:

$$RR(t; x_1, \ldots, x_p) = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4}$$

The implications of the following concusions (based on tests involving estimates and their estimated standard errors obtained from computer output) are:

$\beta_1 = 0$ After controlling for age, weight, and treatment group, risk of death is not associated with disease stage. Otherwise, they are associated.

$\beta_2 = 0$ After controlling for disease stage, weight, and treatment group, risk of death is not associated with age. Otherwise, they are associated.

$\beta_3 = 0$ After controlling for disease stage, age, and treatment group, risk of death is not associated with weight. Otherwise, they are associated.

$\beta_4 = 0$ After controlling for disease stage, age, and weight, risk of death is not associated with treatment group. Otherwise, they are associated.

Of particular interest in drug trials is the last test ($H_0 : \beta_4 = 0$ vs $H_A : \beta_4 \neq 0$). In particular, to show that the active drug is effective, you would want to show that $\beta_4 < 0$, since the relative risk (after controlling for the other three variables) of death for active drug group, relative to controls is e^{β_4}.

Example 9.3 In his landmark paper on the proportional hazards model, Professor D.R. Cox, analyzed remission data from the work of Freireich, et al,(1963) to demonstrate his newly developed model (Cox, 1972). The response of interest was the remission times of patients with acute leukemia who were given a placebo ($x = 1$) or 6–MP ($x = 0$). This data was used to describe sequential designs in Chapter 1 (Example 1.16) and is given in Table 1.8. The model fit and estimates were:

$$RR(t;x) = e^{\beta x} \quad \hat{\beta} = 1.60 \quad \hat{\sigma}_{\hat{\beta}} = 0.42$$

These numbers differ slightly from the values he report due to statistical software differences. Note that the risk of failure is estimated to be $e^{1.6} = 5.0$ times higher for those on placebo ($x = 1$) than those on 6–MP. An approximate 95% confidence interval for β and the relative risk of failure (placebo relative to 6–MP) are:

$$\hat{\beta} \pm 1.96\hat{\sigma}_{\hat{\beta}} \quad \equiv \quad 1.60 \pm 0.82 \quad \equiv \quad (0.78, 2.42) \qquad (e^{0.78}, e^{2.42}) \quad \equiv \quad (2.18, 11.25)$$

Thus, we can be 95% confident that the risk of failure is between 2.18 and 11.25 times higher for patients on placebo than patients on 6–MP. This can be used to confirm the effectiveness of 6–MP in prolonging remission among patients with leukemia. A plot of the Kaplan–Meier survival functions if given in Figure 9.2. In honor of his work in this area, the Proportional Hazards model is often referred to as the Cox regression model.

Example 9.4 A cohort study was conducted to quantify the long–term incidence of AIDS based on early levels of HIV–1 RNA levels, and the age at HIV–1 serioconversion (O'Brien, et al.,1996). Patients were classified based on their early levels of HIV–1 RNA ($< 1000, 1000 - 9999, \geq 10000$ copies/mL), and age at HIV–1 serioconversion (1–17,18–34,35–66). Dummy variables were created to represent these categories.

$$x_1 = \begin{cases} 1 & \text{if early HIV–1 RNA is 1000–9999} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if early HIV–1 RNA is} \geq 10000 \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if age at serioconversion is 18–34} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if age at serioconversion is 35–66} \\ 0 & \text{otherwise} \end{cases}$$

The model for the relative risk of developing AIDS (relative to baseline group – early HIV–1 RNA < 1000, age at serioconversion 1–17) is:

$$RR(t; x_1, x_2, x_3, x_4) = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4}$$

Note that the relative risk is assumed to be constant across time in this model. Parameter estimates and their corresponding standard errors are given in Table 9.4. Also, we give the adjusted relative risk (also referred to as relative hazard), and a 95% CI for the population relative risk. Recall that a relative risk of 1.0 can be interpreted as 'no association' between that variable and the event of interest. In this situation, we get the following interpretations:

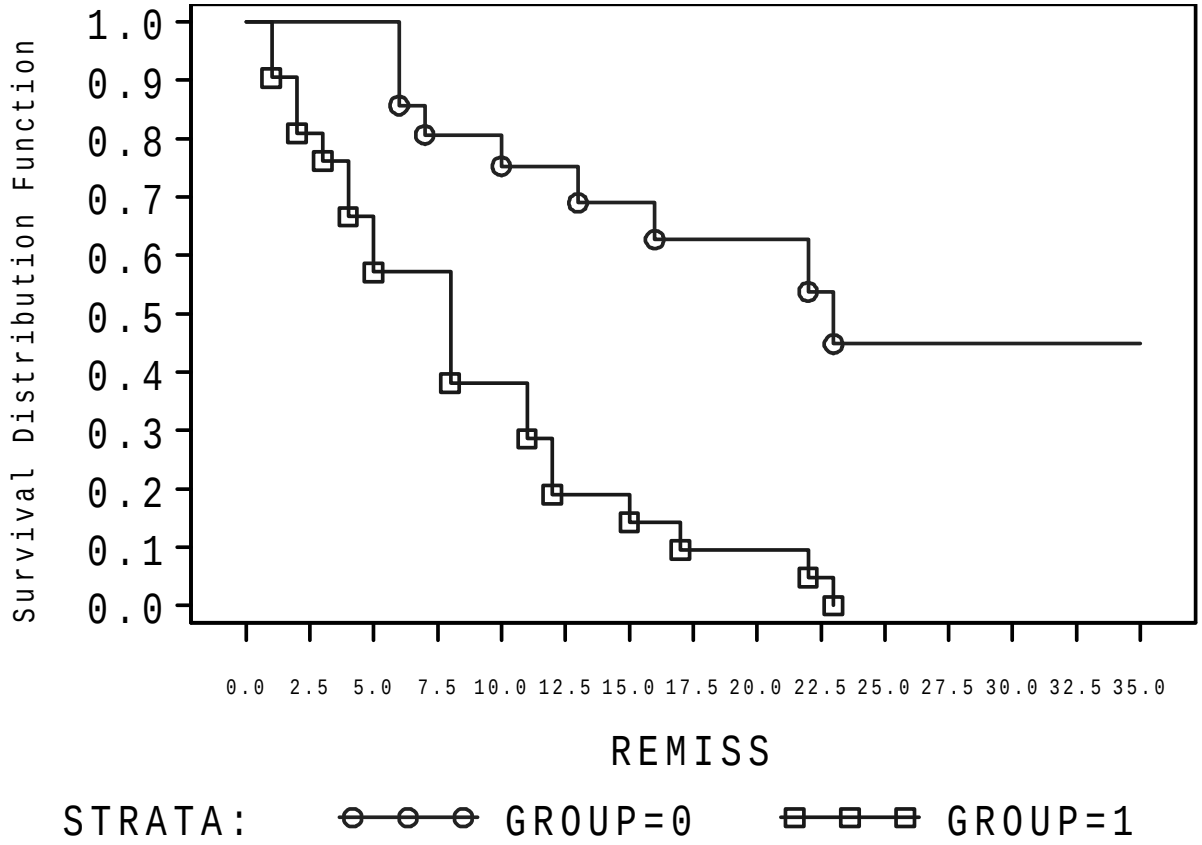

Figure 9.2: Kaplan–Meier estimates of survival functions for acute leukemia patients receiving 6–MP and placebo

Variable (x_i)	Estimate ($\hat{\beta}_i$)	Std. Error ($\hat{\sigma}_{\hat{\beta}}$)	Rel. Risk ($e^{\hat{\beta}_i}$)	95% CI
HIV–1 RNA 1000–9999 (x_1)	1.61	1.02	5.0	(0.7,36.9)
HIV–1 RNA $\geq$ 10000 (x_2)	2.66	1.02	14.3	(1.9,105.6)
Age 18–34 (x_3)	0.79	0.31	2.2	(1.2,4.0)
Age 35–66 (x_4)	1.03	0.31	2.8	(1.5,5.3)

Table 9.4: Parameter estimates for proportional hazards model relating survival to developing AIDS to early HIV–1 RNA levels and age at serioconversion

β_1 The CI contains 1. We cannot conclude that risk of developing AIDS is higher for subjects with HIV–1 RNA 1000–9999 than for subjects with HIV–1 RNA < 1000, after controlling for age.

β_2 The CI is entirely above 1. We can conclude that risk of developing AIDS is higher for subjects with HIV–1 RNA ≥ 10000 than for subjects with HIV–1 RNA < 1000, after controlling for age.

β_3 The CI is entirely above 1. We can conclude that patients whose age at serioconversion is 18–34 have higher risk of developing AIDS than patients whose age is 1–17 at serioconversion.

β_4 The CI is entirely above 1. We can conclude that patients whose age at serioconversion is 35–66 have higher risk of developing AIDS than patients whose age is 1–17 at serioconversion.

Finally, patients can be classified into one of 9 HIV–1 RNA level and age combinations. We give the estimated relative risk for each group, based on the fitted model in Table 9.5. Recall the baseline group is the lowest HIV–1 RNA level and lowest age group. The authors conclude that these two

HIV–1 RNA	Age	$RR = e^{1.61x_1+2.66x_2+0.79x_3+1.03x_4}$
< 1000 $(x_1 = 0, x_2 = 0)$	1–17 $(x_3 = 0, x_4 = 0)$	$e^0 = 1.0$
< 1000 $(x_1 = 0, x_2 = 0)$	18–34 $(x_3 = 1, x_4 = 0)$	$e^{0.79} = 2.2$
< 1000 $(x_1 = 0, x_2 = 0)$	35–66 $(x_3 = 0, x_4 = 1)$	$e^{1.03} = 2.8$
1000–9999 $(x_1 = 1, x_2 = 0)$	1–17 $(x_3 = 0, x_4 = 0)$	$e^{1.61} = 5.0$
1000–9999 $(x_1 = 1, x_2 = 0)$	18–34 $(x_3 = 1, x_4 = 0)$	$e^{1.61+0.79} = 11.0$
1000–9999 $(x_1 = 1, x_2 = 0)$	35–66 $(x_3 = 0, x_4 = 1)$	$e^{1.61+1.03} = 14.0$
≥ 10000 $(x_1 = 0, x_2 = 1)$	1–17 $(x_3 = 0, x_4 = 0)$	$e^{2.66} = 14.3$
≥ 10000 $(x_1 = 0, x_2 = 1)$	18–34 $(x_3 = 1, x_4 = 0)$	$e^{2.66+0.79} = 31.5$
≥ 10000 $(x_1 = 0, x_2 = 1)$	35–66 $(x_3 = 0, x_4 = 1)$	$e^{2.66+1.03} = 40.0$

Table 9.5: Relative risks (hazards) of developing AIDS for each HIV–1 RNA level and age at serioconversion combination

factors are strong predictors of long–term AIDS–free survival. In particular, they have shown that early levels of HIV–1 RNA is good predictor (independent of age) of AIDS development.

9.4 Exercises

51. A study of survival times of mice with induced subcutaneous sarcomas compared two carcinogens – methylcholanthrene and dibenzanthracene (Shimkin, 1941). Mice were assigned to receive either of the carcinogens at one of seveal doses, further, mice were eliminated from data analysis if they died from extraneous cause. This problem deals with the survival times of the 0.1 mg dose groups for methylcholanthrene (M) and dibenzanthracene (D). Mice were followed for 38 weeks, if they survived past 38 weeks, their survival time would be considered censored at time 38. The M group consisted of 46 mice, the D group had 26. Survival times for each group (where 39^* implies censored after week 38, and 16(5) implies five died at week 16) were:

$$M : 14(1), 15(2), 16(6), 17(4), 18(5), 19(10), 20(3), 21(6), 22(1), 23(1), 28(1), 31(1), 39^*(5)$$

$$D : 21(1), 23(3), 24(2), 25(2), 26(1), 28(1), 29(1), 31(4), 32(2), 33(1), 38(4), 39^*(4)$$

Table 9.6 sets up the calculations needed to obtain the Kaplan–Meier estimates of the survival function $\hat{S(t)}$.

Methylcholanthrene						Dibenzanthracene					
i	$t_{(i)}$	n_i	d_i	$\hat{\lambda}_i$	$\hat{S}(t_{(i)})$	i	$t_{(i)}$	n_i	d_i	$\hat{\lambda}_i$	$\hat{S}(t_{(i)})$
1	14	46	1	.022	.978	1	21	26	1	.038	.962
2	15	45	2	.044	.935	2	23	25	3	.120	.847
3	16	43	6	.140	.800	3	24	22	2	.091	.770
4	17	37	4	.108	.714	4	25	20	2	.100	.693
5	18	33	5	.152	.605	5	26	18	1	.056	.654
6	19	28	10	.357	.389	6	28	17	1	.059	.615
7	20	18	3	.167	.324	7	29	16	1	.063	.576
8	21	15	6	.400	.194	8	31	15	4	.267	.422
9	22	9	1	.111	.172	9	32	11	2	.182	.345
10	23	8	1			10	33	9	1		
11	28	7	1			11	38	8	4		
12	31	6	1			–	–	–	–	–	–

Table 9.6: Kaplan–Meier estimates of survival distribution functions for two carcinogens

Figure 9.3 gives the estimated survival functions for the two drugs. Table 9.7 sets up the calculations needed to perform the log–rank test, comparing the survival functions. Complete the table and test whether or not the survival functions differ. If they differ, which carcinogen causes the quickest deaths?

(a) Complete Table 9.6 for each drug group.

(b) On Figure 9.3, identify which curve belongs to which drug.

(c) Complete Table 9.7 and test whether or not the survival functions differ. If they differ, which carcinogen causes the quickest deaths?

52. A randomized, controlled clinical trial was conducted to compare the effects of two treatment regimens on survival in patients with acute leukemia (Frei, et al,1958). A total of 65 patients were randomized to receive one of two regimens of combination chemotherapy, involving methotrexate and 6–mercaptopurine. The first regimen involved receiving each drug daily (continuous), while the second regimen received 6–mercaptopurine daily, but methotrexate only once every 3 days (intermittent). The total doses were the same however (the continuous group received 2.5 mg/day of methotrexate, the intermittent group received 7.5 mg every third day). The survival (and death) information are given in Table 9.8. The survival curves are displayed in Figure 9.4.

(a) Complete the table, computing the survival function over the last months for the continuous group.

(b) Based on the graph, identify the curves representing the intermittent and continuous groups.

Failure	Carcinogen M		Carcinogen D			
Time (i)	d_{1i}	n_{1i}	d_{2i}	n_{2i}	e_{1i}	v_{1i}
14 (1)	1	46	0	26	0.639	0.231
15 (2)	2	45	0	26	1.268	0.458
16 (3)	6	43	0	26	3.739	1.305
17 (4)	4	37	0	26	2.349	0.923
18 (5)	5	33	0	26	2.797	1.147
19 (6)	10	28	0	26	5.185	2.073
20 (7)	3	18	0	26	1.227	0.691
21 (8)	6	15	1	26	2.561	1.380
22 (9)	1	9	0	25	0.265	0.195
23 (10)	1	8	3	25	0.967	0.666
24 (11)	0	7	2	22	0.483	0.353
25 (12)	0	7	2	20	0.519	0.369
26 (13)	0	7	1	18	0.280	0.202
28 (14)	1	7	1	17	0.583	0.395
29 (15)	0	6	1	16	0.273	0.198
31 (16)	1	6	4	15	1.429	0.816
32 (17)	0	5	2	11		0.401
33 (18)	0	5	1	9		0.230
38 (19)	0	5	4	8		
Sum	41	–	22	–		

Table 9.7: Computation of observed and expected values for log–rank test to compare survival functions of two carcinogens

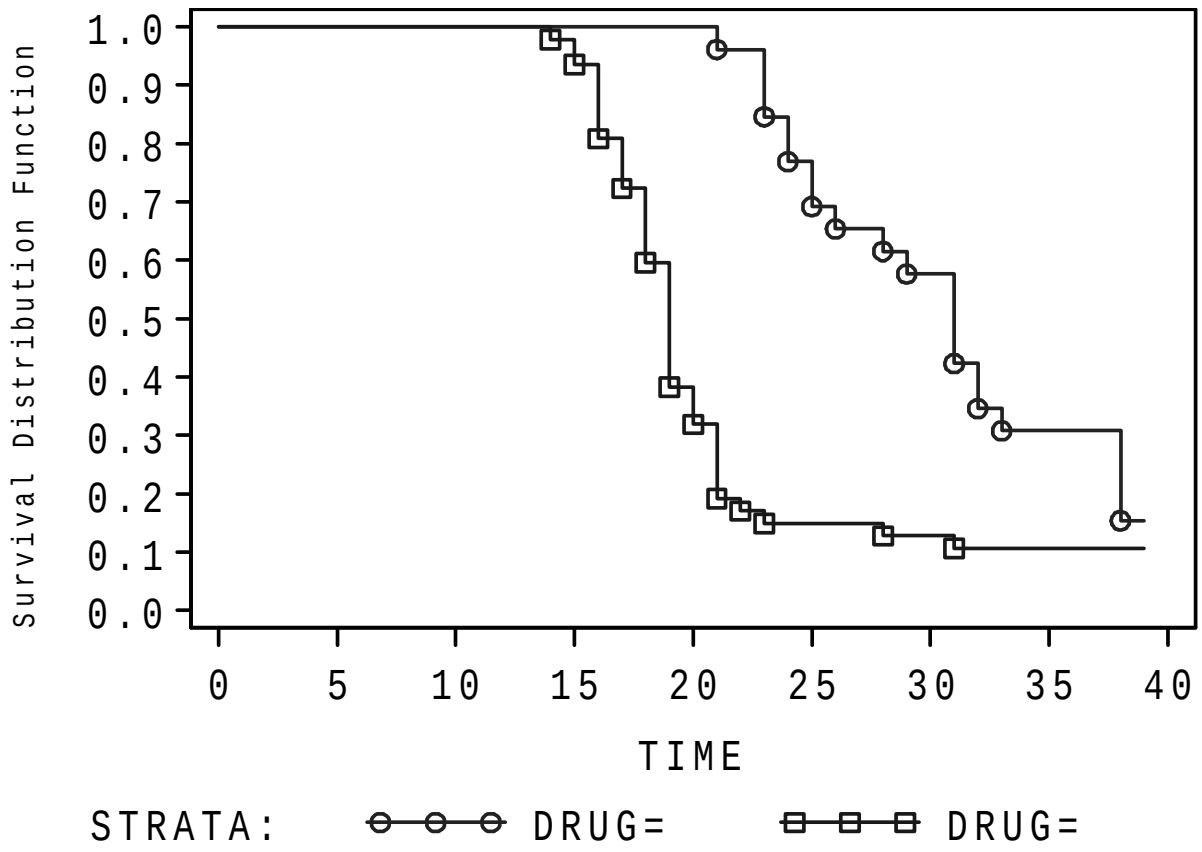

Figure 9.3: Kaplan–Meier estimates of survival functions for methylcholanthrene and dibenzanthracene

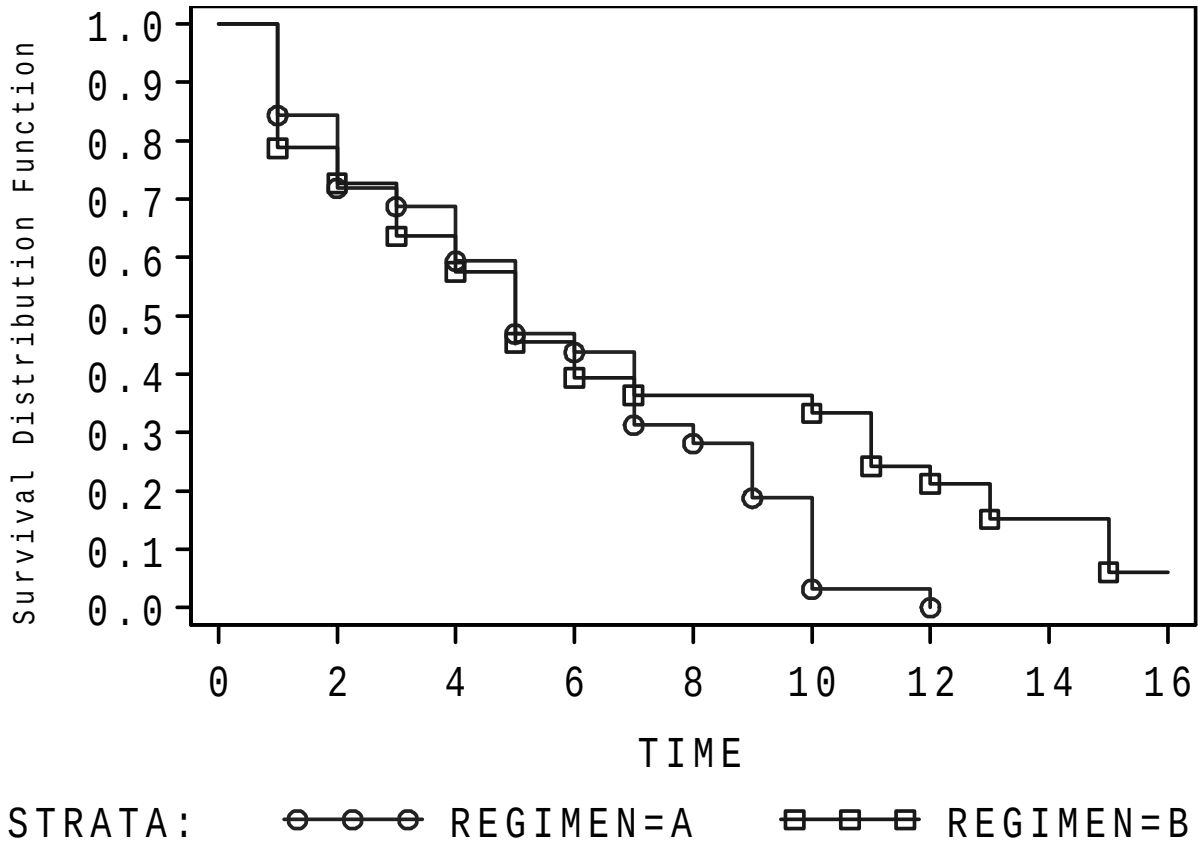

Figure 9.4: Kaplan–Meier estimates of survival functions for intermittent and continuous combination chemotherapy treatments in patients with acute leukemia

	Intermittent				Continuous			
Month (i)	n_i	d_i	$\hat{\lambda}_i$	$\hat{S}(t_{(i)})$	n_i	d_i	$\hat{\lambda}_i$	$\hat{S}(t_{(i)})$
1	32	5	.1563	.8437	33	7	.2121	.7879
2	27	4	.1481	.7187	26	2	.0769	.7273
3	23	1	.0435	.6875	24	3	.1250	.6364
4	22	3	.1364	.5938	21	2	.0952	.5758
5	19	4	.2105	.4688	19	4	.2105	.4530
6	15	1	.0667	.4375	15	2	.1333	.3926
7	14	4	.2857	.3125	13	1	.0769	.3624
8	10	1	.1000	.2813	12	0	.0000	.3624
9	9	3	.3333	.1877	12	0	.0000	.3624
10	6	4	.6667	.0625	12	1	.0833	.3322
11	2	0	.0000	.0625	11	3	.2727	.2416
12	2	2	1.000	.0000	8	1	.1250	.2114
13	—	—	—	.0000	7	2		
14	—	—	—	.0000	5	0		
15	—	—	—	.0000	5	3		
16	—	—	—	.0000	2	0		

Table 9.8: Kaplan–Meier estimates of survival distribution functions for two combination chemotherapy regimens

(c) Over the first half of the study, do the survival curves appear to differ significantly? What about over the second half of the study period?

53. After a waterborne outbreak of cryptosporidiosis in 1993 in Milwaukee, a group of $n = 81$ HIV–infected patients were classified by several factors, and were followed for one year (Vakil, et al.,1996). All of these subjects developed cryptosporidiosis during the outbreak, and were classified by:

- Age (x_1)
- Nausia/vomiting ($x_2 = 1$ if present, 0 if absent)
- Biliary disease from cryptosporidiosis ($x_3 = 1$ if present, 0 if absent)
- CD4 count ($x_4 = 1$ if $\leq 50/mm^3$, 0 if $> 50/mm^3$)

The response was the survival time of the patient (47 died during the year, the remaining 31 survived, and were censored at one year). The proportional hazards regression model was fit:

$$RR(t; x_1, x_2, x_3, x_4) = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4},$$

where $x_1, \ldots, x_4$ are described above. The estimated regression coefficients and their corresponding estimated standard errors are given in Table 9.9.

(a) Interpret each of the coefficients.

(b) Holding all other variables constant, how much higher is the risk (hazard) of death in patients with CD4 counts below $50/mm^3$ ($x_4 = 1$) than patients with higher CD4 counts ($x_4 = 0$).

Variable (x_i)	Estimate ($\hat{\beta}_i$)	Std. Error ($\hat{\sigma}_{\hat{\beta}}$)	Rel. Risk ($e^{\hat{\beta}_i}$)	95% CI
Age (x_1)	0.044	0.017	1.04	(1.01,1.08)
Naus/Vom (x_2)	0.624	0.304	1.87	(1.03,3.38)
Biliary (x_3)	0.358	0.311	1.43	(0.78,2.64)
CD4 (x_4)	1.430	0.719	4.10	(1.72,9.76)

Table 9.9: Parameter estimates for proportional hazards model relating death to age, nausia/vomiting status, biliary disease, and CD4 counts in HIV–infected patients with cryptosporidiosis

(c) Is presence of biliary disease associated with poorer survival after controlling for the other three explanatory variables?

54. Survival data were reported on a cohort of 1205 AIDS patients in Milan, Italy (Monforte, et al.,1996). The authors fit a proportional hazards regression model relating risk of death to such factors as age, sex, behavioral risk factor, infection date, opportunistic infection, $CD4+$ count, use of ZDV prior to AIDS, and PCP prophylaxis prior to AIDS. Within each factor, the first level acted as the baseline for comparisons. Estimated regression coefficients, standard errors and hazard ratios are given in Table 9.10.

(a) Describe the baseline group.

(b) Describe the group that has the highest estimated risk of death.

(c) Describe the group that has the lowest estimated risk of death.

(d) Does ZDV use prior AIDS appear to increase or decrease risk of AIDS after controlling all other variables? Test at $\alpha = 0.05$ significance level. (Hint: This can be done by a formal test or simply interpreting the confidence interval for the hazard ratio).

(e) Repeat part d) in terms of PCP prophylaxis before AIDS.

(f) Computed the estimated relative risks for the groups in parts b) and c), relative to the group in part a).

Variable	Level	Cases	Estimate ($\hat{\beta}$)	Std. Error ($\hat{\sigma}_{\hat{\beta}}$)	Rel. Risk ($e^{\hat{\beta}}$)	95% CI
Age	≤ 35	907	—	—	1	—
	> 35	298	0.231	.086	1.26	(1.06,1.49)
Sex	Male	949	—	—	1	—
	Female	256	−0.083	.086	0.92	(0.77,1.09)
Behavior	IDU	508	—	—	1	—
	Ex–IDU	267	−0.151	.082	0.86	(0.72,1.01)
	Homosexual	247	−0.073	.113	0.93	(0.74,1.16)
	Heterosexual	162	−0.128	.132	0.88	(0.68,1.14)
	Transfused	21	−0.030	.315	0.97	(0.52,1.80)
Date	1984–1987	185	—	—	1	—
	1988–1990	404	−0.223	.103	0.8	(0.69,0.98)
	1991–1994	616	−0.105	.116	0.9	(0.75,1.13)
Infection	PCP	292	—	—	1	—
	Candidiasis	202	0.039	.110	1.04	(0.84,1.29)
	TE	134	0.262	.119	1.3	(1.00,1.64)
	CMV	114	0.531	.115	1.7	(1.58,2.13)
	KS	109	−0.139	.147	0.87	(0.65,1.16)
	ADC	102	0.336	.156	1.4	(1.08,1.90)
	Other	375	0.470	.124	1.6	(1.31,2.04)
	Multiple	123	0.262	.109	1.3	(1.05,1.61)
$CD4 + i(\times 10^6/l)$	≤ 50	645	—	—	1	—
	50–100	182	−0.223	.123	0.8	(0.70,1.01)
	> 100	285	−0.693	.084	0.5	(0.41,0.59)
ZDV	No	762	—	—	1	—
	Yes	443	0.030	.086	1.03	(0.87,1.22)
PCP prophylaxis	No	931	—	—	1	—
	Yes	274	0.058	.100	1.06	(0.80,1.29)

Table 9.10: Parameter estimates for proportional hazards model relating risk of death to age, sex, risk behavior, year of infection, oppotunistic infection, CD4 counts, ZDV use prior to AIDS, and PCP prophylaxis use prior to AIDS in Italian AIDS patients

Chapter 10

Special Topics in Pharmaceutics

In this chapter, we will describe two additional types of statistical applications that are commonly used in clinical pharmacology. These methods are actually specific applications of previously described statistical models. These procedures are:

1. Assessment of Pharmaceutical Bioequivalence
2. Dose–Response Studies

Bioequivalence studies make use of the analysis of variance and the construction of confidence intervals for the difference between two population means. Dose–response studies make use of issues in experimental design and the analysis of variance, as well as nonlinear regression.

10.1 Assessment of Pharmaceutical Bioequivalence

When patents of popular drugs expire, rival manufacturers inevitably produce *generic* substitutes for the original (or *pioneer*). Besides being identical in formulation, the makers of the generic drug must show that its version is "equivalent" in terms of bioavailability to the pioneer version. A second situation that involves bioequivalence testing is when a manufacturer creates a new formulation of a current drug (e.g. $100mg$ tablets instead of $200mg$ tablets). An overview of statistical methods of determining bioequivalence can be found (Yuh,1995).

In this section we will refer to the new formulation as the **test** and the original (already approved) as the **reference**. The strategy is to demonstrate that the test's bioavailability (as measured by AUC, C_{max}, and t_{max}) and the reference's bioavailability are within 20% of each other. That is, for each pharmacokinetic parameter, we wish to demonstrate that population means differ by less than 20%. The analysis of AUC and C_{max} is generally conducted after taking logs of the original data, and then transformed back to the original scale. If we denote the reference mean μ_R and the test mean μ_T, we wish to show (for each of the three pharmacokinetic parameters):

$$0.80 \leq \frac{\mu_T}{\mu_R} \leq 1.25 \quad \Longrightarrow \quad -0.20 \leq \frac{\mu_T - \mu_R}{\mu_R} \leq 0.25$$

The experiment is typically conducted in a 2–period crossover, with subjects being randomly assigned into one of two sequences (test followed by reference or reference followed by test). Then,

we partition the total variation of the observed measurements into variation due to: formulations, periods, sequences, subjects within sequences, and random error. The analysis of variance can be formed (on a computer) and is given in Table 10.1. We assume that n_1 subjects received sequence 1 and n_2 subjects received sequence 2, for a total of n subjects and $2n$ measurements. Further, we will denote $\overline{FORM}_i$, $\overline{PER}_i$, $\overline{SEQ}_i$, and $\overline{SUBJ}_{j(i)}$ as the means of the i^{th} formulation, period, sequence, and j^{th} subject (within i^{th} sequence), respectively in the sums of squares formulas. Further $\overline{y}$ is the overall mean.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Formulations	$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\overline{FORM}_i - \overline{y})^2$	1	SS_{FORM}
Periods	$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\overline{PER}_i - \overline{y})^2$	1	SS_{PER}
Sequences	$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\overline{SEQ}_i - \overline{y})^2$	1	SS_{SEQ}
Subjects(Sequences)	$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\overline{SUBJ}_{j(i)} - \overline{SEQ}_i)^2$	$n-2$	$\frac{SS_{SUBJ(SEQ)}}{n-2}$
Error	By Subtraction	$n-2$	$\frac{SSE}{n-2}$
TOTAL	$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y - \overline{y})^2$	$2n-1$	

Table 10.1: The Analysis of Variance Table for a Bioequivalence Study

Once we compute the analysis of variance to obtain MSE, we compute an approximate 90% CI for (μ_T/μ_R) by completing the following steps (and denoting the sample means for the test and reference $\overline{y}_T$ and $\overline{y}_R$, respectively):

1. Obtain a 90% CI for $\mu_T - \mu_R$ by computing

$$(\overline{y}_T - \overline{y}_R) \pm t_{.05,n-2}\sqrt{MSE\left(\frac{2}{n}\right)} \quad \equiv \quad (LB_1, UB_1)$$

2. Obtain an approximate 90% CI for $(\mu_T - \mu_R)/\mu_R)$ by computing:

$$(\frac{LB_1}{\overline{y}_R}, \frac{UB_1}{\overline{y}_R}) \quad \equiv \quad (LB_2, UB_2)$$

3. Obtain an approximate 90% CI for μ_T/μ_R by computing:

$$(LB_2 + 1, UB_2 + 1) \quad \equiv \quad (LB_3, UB_3)$$

This procedure is conducted for all three pharmacokinetic parameters (AUC, C_{max}, t_{max}), and if all three 90% CI's are inside the range (0.80,1.25) the two formulations are are determined to be bioequivalent.

Example 10.1 A bioequivalence study was conducted between a $40mg$ famotidine wafer (test formulation) and a $40mg$ famotidine tablet (reference formulation) in a 2–period crossover study (Schwartz, et al.,1995). The wafer was considered a novel alternative to the tablet and considered to be more a more convenient means of delivery. The researchers computed the Analysis of Variance as described above for the log transformed AUC and C_{max} values and the original scale t_{max} values. The pertinent results are given in Table 10.2. Confidence intervals for $\mu_T - \mu_R$ are of the form $(n = 18)$:

$$(\overline{y}_T - \overline{y}_R) \pm t_{.10/2,n-2}\sqrt{MSE\left(\frac{2}{n}\right)} \quad \equiv \quad (\overline{y}_T - \overline{y}_R) \pm 1.746\sqrt{MSE\left(\frac{2}{18}\right)}$$

Variable	$\overline{y}_T$	$\overline{y}_R$	MSE	90%CI $(\mu_T - \mu_R)$	90%CI (μ_T/μ_R)	Orig. Scale
$Log(AUC)$	7.034124	6.992831	0.020704	$(-.042451, .125035)$	$\left(\frac{6.9504}{6.9928}, \frac{7.1179}{6.9928}\right)$	$\left(\frac{e^{6.9504}}{e^{6.9928}}, \frac{e^{7.1179}}{e^{6.9928}}\right) \equiv (0.958, 1.133)$
$Log(C_{max})$	5.129899	5.181222	0.064547	$(-.199186, .096540)$	$\left(\frac{4.9820}{5.1812}, \frac{5.2778}{5.1812}\right)$	$\left(\frac{e^{4.9820}}{e^{5.1812}}, \frac{e^{4.9820}}{e^{5.1812}}\right) \equiv (0.819, 1.101)$
t_{max}	3.0	2.1	0.5915	$(.497, 1.393)$	(1.237,1.663)	(1.237,1.663)

Table 10.2: Approximate 90% CI's for μ_T/μ_R (Test=Wafer,Reference=Tablet)

We see from Table 10.2 that in terms of AUC and C_{max}, the wafer meets U.S. bioequivalence criteria (entire 90% CI within 0.80–1.25). However, the rate of absorption is slower, since the t_{max} mean is higher (between 1.24 and 1.66 times as high). If time to maximum concentration is not very important, the manufacturer would probably consider this wafer equivalent to the tablet, and market it, particularly if it improves compliance to prescribed therapy.

10.2 Dose–Response Studies

Dose–response studies are generally conducted in the early stages of drug development. They are conducted as toxicity studies in pre–clinical trials (see Example 8.1), and are also used in phase I and II trials to obtain proper dosing regimens for the large–scale phase III comparative studies. Dose–response studies can be analyzed as either regression models or one–way Analysis of Variance models. They are generally conducted as parallel groups designs. Of primary interest is estimation of the minimum effective dose. For a recent statistical description of the design and analysis issues, see (Ruberg,1996a,1996b).

When analyzed as a regression model, studies tend to have observations at a relatively large number of doses, and a "S"–shaped function is fit, as in Example 8.2. Parameters that can be estimated include:

MED Minimum Effective Dose – The lowest dose that has a mean response significantly different from no dose (placebo).

ED_{50} Dose that produces a 50% (of maximum) effect.

ME Maximum Effect – The highest response that can be attained, across all doses.

In Example 8.2, we fit what was referred to as the sigmoid–E_{max} function. Another function that is widely used is the four–parameter logistic function:

$$y = \frac{\beta_0 - \beta_3}{1 + (x/\beta_2)^{\beta_1}} + \beta_3$$

Many studies are conducted at only a few doses (three or fewer). In these situations, the studies are typically analyzed as a one–way Analysis of Variance. The strategy is to first test for overall treatment effects using the F–test for the Completely Randomized Design (Section 6.1), or a more powerful linear trend test (Ruberg,1996b). If treatment effects exist, Dunnett's procedure can be used to compare treatments to the control. When a clinically meaningful response (CMR) is known, the minimum effective dose can be considered to be the smallest dose that: 1) is significantly different from the zero–dose, and 2) has a sample mean larger than the CMR (Ruberg,1996b). In cases where there is not a known CMR, the MED is considered to be the smallest dose that is significantly different from the zero–dose.

Example 10.2 The effects of nonsteroidal anti–inflammatory drugs (NSAID) on renal sodium excretion in rats were reported in a dose–response study (Kadokawa, et al.,1979). Rats were assigned at random to receive one of: vehicle control, $1mg/kg$ indomethacin, $3mg/kg$, $10mg/kg$. One response measured was the sodium electrolyte excretion collected in urine over a five–hour period. The means and standard deviations are given in Table 10.3. The corresponding Analysis of Variance table is given in Table 10.4. Six rats received each of the $k = 4$ treatments.

Treatment (i)	$\overline{y}_i$	s_i	n_i
Control (1)	2.40	0.37	6
$1mg/kg$ (2)	2.21	0.39	6
$3mg/kg$ (3)	1.87	0.27	6
$10mg/kg$ (4)	1.54	0.32	6

Table 10.3: Means and Std. Devs. of sodium electrolyte excretion — NSAID dose–respone study

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
TREATMENTS	2.60	3	0.87	$F = \frac{0.87}{0.12} = 7.25$
ERROR	2.32	20	0.12	
TOTAL	4.92	23		

Table 10.4: The Analysis of Variance table for the Indomethacin dose–response study in rats

We reject the null hypothesis of no treatment effect ($F_{.05,3,20} = 3.10$, p–value=.0018). We now use Dunnett's method to compare the 3 dose means with the vehicle control, using 2–sided

confidence intervals, and an overall error rate of $\alpha = 0.05$. The form of the confidence intervals is:

$$(\overline{y}_i - \overline{y}_1) \pm d_{\alpha,k-1,n-k}\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_1})},$$

where $d_{\alpha,k-1,n-k}$ is given in tables of various statistical texts (see Montgomery (1991)). For this example, $d_{\alpha,k-1,n-k} = d_{.05.3,20} = 2.54$. The confidence intervals for each dose versus control are given in Table 10.5.

Comparison	$\overline{y}_i - \overline{y}_j$	Simultaneous 95% CI's Dunnett
$1mg/kg$ vs Control	$2.21 - 2.40 = -0.19$	$(-0.70, 0.32)$
$3mg/kg$ vs Control	$1.87 - 2.40 = -0.53$	$(-1.04, -0.02)$
$10mg/kg$ vs Control	$1.54 - 2.40 = -0.86$	$(-1.37, -.35)$

Table 10.5: Dunnett's multiple comparisons for the indomethacin dose–response study in rats

We cannot conclude that the $1mg/kg$ dose is significantly different from the control (confidence interval contains 0). We can, however, conclude that doses 3 and 10 have significantly lower means than the control (both CI's are entirely below 0). The $3mg/kg$ dose appears to be the minimum effective dose (MED).

10.3 Exercises

55. A bioequivalence evaluation of two oral formulations of loperamide was conducted in a two–period crossover study in 24 healthy males (Doser, et al.,1995). Based on European standards (CPMP,1991), bioequivalence is confirmed if approximate 90% CI's for μ_T/μ_R are in the ranges (.80,1.25) and (.70,1.43), for AUC and C_{max}, respectively. Due to lack of normality in these variables, confidence intervals are computed after setting up the Analysis of Variance for a 2–period crossover study with sequence and period effects (Chapter 6) for $log(AUC)$ and $log(C_{max})$. The Analysis of Variance for $log(AUC)$ was given in Table 6.34 in Chapter 6. The pertinent results are given in Table 10.6 in Chapter 6. Complete the table, by converting the units back to the original scale. Does this Diarex meet the European criteria of bioequivalence? Does it meet U.S. criteria?

Variable	$\overline{y}_T$	$\overline{y}_R$	MSE	90%CI $(\mu_T - \mu_R)$	90%CI (μ_T/μ_R)	Orig. Scale
$Log(AUC)$	4.05774	4.14275	0.03557	$(-.17850, .00848)$	$\left(\frac{3.96425}{4.14275}, \frac{4.15123}{4.14275}\right)$	
$Log(C_{max})$	1.10979	1.29954	.05899	$(-.31013, -.06937)$	$\left(\frac{0.98941}{1.29954}, \frac{1.23017}{1.299514}\right)$	

Table 10.6: Approximate 90% CI's for μ_T/μ_R (Test=Diarex,Reference=Imodium)

56. A trial was conducted to measure the effect of orlistat on the pharmacokinetics and pharmacodynamics of warfarin (Zhi, et al,1996). In a two–period crossover study, healthy subjects were randomized to receive orlitat and placebo (in random order, and with a long washout period) for

17 days ($120mg$ t.i.d.), and receive a $30mg$ dose of racemic warfarin sodium on the 11^{th} day. Three variables measured and reported were $AUC_{0-\infty}$ for R–warfarin, net prothrombin time (PT) AUC and net Factor VII AUC. The first was a measure of the pharmacokinetics of warfarin, while the second two were measures of its pharmacodynamics. Each was measured when warfarin was taken with orlistat (O) and with placebo (P). Data are given in Table 10.7. Complete the table by converting the units to the original scale by taking the exponential of the endpoints of the CI for the difference between the means of the logs. Is there any reason to feel that there is a significant interaction between orlistat and warfarin?

Variable	$\overline{y}_O$	$\overline{y}_P$	MSE	90%CI ($\mu_T - \mu_R$)	Orig. Scale
$Log(AUC_{0-\infty})$	11.4377	11.4464	.0052	$(-0.0619, 0.0488)$	
$Log(NetPTAUC)$	5.2883	5.1120	.6768	$(-0.1625, 0.5128)$	
Log(Net Factor VII AUC)	8.2789	8.2822	.0398	$(-0.1508, 0.1484)$	

Table 10.7: Approximate 90% CI's for μ_T/μ_R (Test=Diarex,Reference=Imodium)

57. The use of intracavernosal alprostadil was studied in men suffering from erectile dysfunction (Linet and Ogring,1996). Patients were randomly assigned to receive placebo or one of four doses of alprostadil (2.5,5,10,20–μg). The response measured was duration of erection as measured by Rigiscan ($\geq$ 70% rigidity), with mean and standard deviations given in Table 10.8. The Analysis of Variance is given in Table 10.9, and the set–up of Dunnett's comparisons are given in Table 10.10. Note that since no one responded in the placebo group, we will have an underestimate of the within group varaiation (SSE).

(a) Complete the confidence intervals in Table 10.10.

(b) If there is no clinically meaningful response given, what is the minimum effective dose (MED).

(c) Suppose a group of sex researchers determined a clinically meaningful response as being 30 minutes. What is the MED?

Treatment (i)	$\overline{y}_i$	s_i	n_i
Control (1)	0	0.0	59
$2.5\mu g$ (2)	12	27.7	57
$5\mu g$ (3)	33	75.5	60
$10\mu g$ (4)	31	60.4	62
$20\mu g$ (5)	44	55.8	58

Table 10.8: Means and Std. Devs. of duration of erection (minutes) — alpostradil dose–response study

58. In Problem 2 of Chapter 6, determine the minimum effective dose of HNK20 in terms of reducing RSV in rhesus monkeys.

Source of Variation	Sum of Squares	ANOVA Degrees of Freedom	Mean Square	F
TREATMENTS	73286.2	4	18321.5	$F = \frac{18321.5}{2678.0} = 6.84$
ERROR	779298.2	291	2678.0	
TOTAL	852584.4	295		

Table 10.9: The Analysis of Variance table for the alpostradil dose–response study in men with erectile dysfunction

Comparison	$\overline{y}_i - \overline{y}_j$	Simultaneous 95% CI's Dunnett
$2.5\mu g$ vs Control	$12 - 0 = 12$	$12 \pm 2.44\sqrt{2678.0\left(\frac{1}{57} + \frac{1}{59}\right)}$
$5\mu g$ vs Control	$33 - 0 = 33$	$33 \pm 2.44\sqrt{2678.0\left(\frac{1}{60} + \frac{1}{59}\right)}$
$10\mu g$ vs Control	$31 - 0 = 31$	$31 \pm 2.44\sqrt{2678.0\left(\frac{1}{62} + \frac{1}{59}\right)}$
$20\mu g$ vs Control	$44 - 0 = 44$	$44 \pm 2.44\sqrt{2678.0\left(\frac{1}{58} + \frac{1}{59}\right)}$

Table 10.10: Dunnett's multiple comparisons for the alpostradil dose–response study in men with erectile dysfunction

59. A clinical trial was conducted in obese patients to determine the safety and efficacy of the lipase inhibitor Orlistat (Drent, et al,1995). Patients were randomized to receive one of the following four treatments: placebo, 30 mg/day, 180 mg/day, 360 mg/day Orlistat.

After a four–week placebo run–in, weight losses were measured in a 12 week trial, where patients were placed on similar diets. Weight loss summary statistics and sample sizes for the intent–to–treat analysis are given in Table 10.11. Test for a treatment effect in terms of mean weight loss, and perform Dunnett's method of multiple comparisons to compare each treatment to the control group. What is the minimum effective dose? Note that the critical value for using Dunnett's procedure is $d_{.05,3,182} \approx 2.35$ and the overall mean is $\overline{y} = 3.76$.

Treatment (i)	$\overline{y}_i$	s_i	n_i
Control (1)	2.98	2.58	46
$30mg/day$ (2)	3.61	2.63	48
$180mg/day$ (3)	3.69	2.62	45
$360mg/day$ (4)	4.74	2.61	47

Table 10.11: Means and Std. Devs. of weight loss (kg) — Orlistat dose–response study

60. In a clinical trial of orlistat among normal volunteers receiving diets of $60gram/day$ of fat, measurements of fecal fat (g/d) were reported (Hauptman, et al.,1992). The results are given in Table 10.12, for subjects receiving: placebo, $100mg/day$, $200mg/day$, and $400mg/day$. Test for a treatment effect in terms of mean fecal fat, and perform Dunnett's method of multiple comparisons to compare each treatment to the control group. What is the minimum effective dose? Note that the critical value for using Dunnett's procedure is $d_{.05,3,23} = 2.17$ and the overall mean is $\overline{y} = 13.7$. Are you comfortable with the equal variance assumption here? Do there appear to be any difference among the orlistat groups?

$$SST = 1649 \qquad SSE = 586.82$$

Treatment (i)	$\overline{y}_i$	s_i	n_i
Control (1)	2.7	0.79	9
$100mg/day$ (2)	19.7	7.3	6
$200mg/day$ (3)	18.8	6.8	6
$400mg/day$ (4)	19.3	4.1	6

Table 10.12: Means and Std. Devs. of fecal fat $(grams/day)$ — Orlistat dose–response study

Appendix A

Statistical Tables

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

Table A.1: Right–hand tail area for the standard normal (z) distribution. Values within the body of the table are the areas in the tail above the value of z corresponding to the row and column. For instance, $P(Z \geq 1.96) = .0250$

ν	$t_{.100,\nu}$	$t_{.050,\nu}$	$t_{.025,\nu}$	$t_{.010,\nu}$	$t_{.005,\nu}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
110	1.289	1.659	1.982	2.361	2.621
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Table A.2: Critical values of the t–distribution for various degrees of freedom (ν). $P(T \geq t_\alpha) = \alpha$. Values on the bottom line (df=∞) correspond to the cut–offs of the standard normal (Z) distribution

ν	$\chi^2_{.100,\nu}$	$\chi^2_{.050,\nu}$	$\chi^2_{.010,\nu}$	$\chi^2_{.001,\nu}$
1	2.706	3.841	6.635	10.828
2	4.605	5.991	9.210	13.816
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.467
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.458
7	12.017	14.067	18.475	24.322
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.528
14	21.064	23.685	29.141	36.123
15	22.307	24.996	30.578	37.697

Table A.3: Critical values of the χ^2–distribution for various degrees of freedom (ν).

ν_2	$F_{.05,1,\nu_2}$	$F_{.05,2,\nu_2}$	$F_{.05,3,\nu_2}$	$F_{.05,4,\nu_2}$	$F_{.05,5,\nu_2}$	$F_{.05,6,\nu_2}$	$F_{.05,7,\nu_2}$	$F_{.05,8,\nu_2}$	$F_{.05,9,\nu_2}$
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99
110	3.93	3.08	2.69	2.45	2.30	2.18	2.09	2.02	1.97
130	3.91	3.07	2.67	2.44	2.28	2.17	2.08	2.01	1.95
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94
170	3.90	3.05	2.66	2.42	2.27	2.15	2.06	1.99	1.94
190	3.89	3.04	2.65	2.42	2.26	2.15	2.06	1.99	1.93
210	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92
230	3.88	3.04	2.64	2.41	2.25	2.14	2.05	1.98	1.92
250	3.88	3.03	2.64	2.41	2.25	2.13	2.05	1.98	1.92

Table A.4: Critical values ($\alpha = 0.05$) of the F–distribution for various numerator and denominator degrees of freedom (ν_1, ν_2).

Appendix B

Bibliography

Agresti, A. (1990), *Categorical Data Analysis*, New York: Wiley.

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: Wiley.

Agresti, A., and Winner, L. (1997),"Evaluating Agreement and Disagreement Among Movie Reviewers," *Chance*, To appear.

Akahane, K., Furuhama, K., Inage, F., and Onodera, T. et al. (1987),"Effects of Malotilate on Rat Erythrocytes," *Japanese Journal of Pharmacology*, 45:15–25.

Amberson, J.B., McMahon, B.T., and Pinner, M. (1931),"A Clinical Trial of Sanocrysin in Pulmonary Tuberculosis," *American Review of Tuberculosis*, 24:401–435.

Anagnostopoulos, A., Aleman, A., Ayers, G., et al. (2004),"Comparison of High-Dose Melphalan with a More Intensive Regimen of Thiotepa, Busulfan, and Cyclophosphamide for Patients with Multiple Myeloma," *Cancer*, 100:2607–2612.

Aweeka, F.T., Tomlanovich, S.J., Prueksaritanont, T. et al. (1994),"Pharmacokinetics of Orally and Intravenously Administered Cyclosporine in Pre–Kidney Transplant Patients," *Journal of Clinical Pharmacology*, 34:60–67.

Bachmann, K., Sullivan, T.J., Reese, J.H., et al. (1995),"Controlled Study of the Putative Interaction Between Famotidine and Theophylline in Patients with Chronic Obstructive Pulmonary Disorder," *Journal of Clinical Pharmacology*, 35:529–535.

Baldeweg, T., Catalan, J., Lovett, E., et al. (1995),"Long–Term Zidovudine Reduces Neurocognitive Deficits in HIV–1 Infection," *AIDS*, 9:589–596.

Band, C.J., Band, P.R., Deschamps, M., et al. (1994),"Human Pharmacokinetic Study of Immediate–Release (Codeine Phosphate) and Sustained–Release (Codeine Contin) Codeine," *Journal of Clinical Pharmacology*, 34:938–943.

Berenson, J.R., Lichtenstein, A., Porter, L., et al. (1996),"Efficacy of Pamidronate in Reducing Skeletal Events in Patients with Advanced Multiple Myeloma," *New England Journal of Medicine*, 334:488–493.

Bergstrom, L., Yocum, D.E., Ampei, N.M., et al. (2004),"Increased Risk of Coccidioidomycosis in Patients Treated with Tumor Necrosis Factor α Antagonists," *Arthristis & Rheumatism*, 50:1959–1966.

Berry, D.A. (1990),"Basic Principles in Designing and Analyzing Clinical Studies". In *Statistical Methodology in the Pharmaceutical Sciences*, (D.A. Berry, ed.). New York: Marcel Dekker. pp.1–55.

Blanker, M.H., Bohnen, A.M., Groeneveld, F.P.M.J., et al. (2001),"Correlates for Erectile and Ejaculatory Dysfunction in Older Dutch Men: A Community-Based Study," *Journal of the American Geriatric Society*, 49:436–442.

Boner, A.L., Bennati, D., Valleta, E.A., et al. (1986),"Evaluation of the Effect of Food on the Absorption of Sustained–Release Theophylline and Comparison of Two Methods for Serum Theophylline Analysis," *Journal of Clinical Pharmacology*, 26:638–642.

Broders, A.C. (1920),"Squamous–Cell Epithelioma of the Lip," *Journal of the American Medical Association*, 74:656–664.

Brown, S.L., and Pennello, G. (2002),"Replacement Surgery and Silicone Gel Breast Implant Rupture: Self-Report by Women and Mammoplasty," *Journal of Women's Health & Gender-Based Medicine*, 11:255–264.

Bryant, H., and Brasher, P. (1995),"Breast Implants and Breast Cancer – Reanalysis of a Linkage Study," *New England Journal of Medicine*, 332:1535–1539.

Carr, A., Workman, C., Crey, D., et al. (2004),"No Effect of Rosiglitazone for Treatment of HIV-1 Lipoatrophy: Randomised, Double-Blind, Placebo-Controlled Trial," *Lancet*, 363:429-438.

Carrigan, P.J., Brinker, D.R., Cavanaugh, J.H., et al. (1990),"Absorption Characteristics of a New Valproate Formulation: Divalproex Sodium–Coated Particles in Capsules (Depakote Sprinkle)," *Journal of Clinical Pharmacology*, 30:743–747.

Carson, C.C., Burnett, A.L., Levine, L.A., and Nehra, A. (2002),"The Efficacy of Sildenafil Citrate (Viagra) in Clinical Populations: An Update," *Urology*, 60 (Suppl 2b):12–27.

Carson, C.C., Rajfer, J., Eardley, I., et al. (2004),"The Efficacy and Safety of Tadalafil: An Update," *BJU International*, 93:1276–1281.

Catalona, W.J., Smith, D.S., Wolfert, R.L., et al. (1995),"Evaluation of Percentage of Free Serum Prostate–Specific Antigen to Improve Specificity of Prostate Cancer Screening," *Journal of the American Medical Association*, 274:1214–1220.

Chan, R., Hemeryck, L., O'Regan, M., et al. (1995),"Oral Versus Intravenous Antibiotics for Community Acquired Respiratory Tract Infection in a General Hospital: Open, Randomised Controlled Trial," *British Medical Journal*, 310:1360–1362.

Chisolm, M.A., Cobb, H.H., and Kotzan, J.A. (1995),"Significant Factors for Predicting Academic Success of First–Year Pharmacy Students," *American Journal of Pharmaceutical Education*, 59:364–370.

Collier, A.C., Coombs, R.W., Schoenfeld, D.A., et al. (1996),"Treatment of Human Immunodeficiency Virus Infection with Saquinavir, Zidovudine, and Zalcitabine," *New England Journal of Medicine*, 334:1011–1017.

Cook, T.D. and Campbell, D.T. (1979), *Quasi–Experimentation: Design & Analysis Issues for Field Settings*, Boston: Houghton Mifflin.

Corn, M., Lester, D., and Greenberg, L.A. (1955),"Inhibiting Effects of Certain Drugs on Audiogenic Seizures in the Rat," *Journal of Pharmacology and Experimental Therapeutics*, 113:58–63.

Cornfield, J. (1962),"Joint Dependence of Risk of Coronary Heart Disease on Serum Cholesterol and Systolic Blood Pressure: A Discriminant Function Analysis," *Federation Proceedings*, 21, Supplement No. 11:58–61.

Cox, D.R. (1972),"Regression Models and Life–Tables," *Journal of the Royal Statistical Society* B, 34:187–202.

Dale, L.C., Hurt, R.D., Offord, K.P., et al. (1995),"High–Dose Nicotine Patch Therapy," *Journal of the American Medical Association*, 274:1353–1358.

Dawson, G.W., and Vestal, R.E. (1982),"Smoking and Drug Metabolism," *Pharmacology & Therapeutics*, 15:207–221.

De Mello, N.R., Baracat, E.C., Tomaz, G., et al. (2004),"Double-Blind Study to Evaluate Efficacy and Safety of Meloxicam 7.5mg and 15mg versus Mefenamic Acid 1500mg in the Treatment of Primary Dtsmenorrhea," *Acta Obstetricia et Gynecologica Scandinavica*, 83:667-673.

De Vita, V.T.Jr., Serpick, A.A., and Carbone, P.P. (1970),"Combination Chemotherapy in the Treatment of Advanced Hodgkin's Disease," *Annals of Internal Medicine*, 73:881–895.

Doll, R., and Hill, A.B. (1950),"Smoking and Carcinoma of the Lung," *British Medical Journal*, 2:739–748.

Doser, K., Meyer, B., Nitsche, V., and Binkert–Graper, P. (1995),"Bioequivalence Evaluation of Two Different Oral Formulations of Loperamide (Diarex Lactab vs Imodium Capsules,"*International Journal of Clinical Pharmacology and Therapeutics*, 33:431–436.

Drent, M.L., Larson, I., William–Olsson, T., et al. (1995),"Orlistat (RO 18–0647), a Lipase Inhibitor, in the Treatment of Human Obesity: A Multiple Dose Study,"*International Journal of Obesity*, 19:221–226.

Evans, J.R., Forland, S.C., and Cutler, R.E. (1987),"The Effect of Renal Function on the Pharmacokinetics of Gemfibrozil,"*Journal of Clinical Pharmacology*, 27:994–1000.

Feigelson, H.S., Criqui, M.H., Fronek, A., et al. (1994),"Screening for Peripheral Arterial Disease: The Sensitivity, Specificity, and Predictive Value of Noninvasive Tests in a Defined Population,"*American Journal of Epidemiology*, 140:526–534.

Fontaine, R., and Chouinard, G. (1986),"An Open Clinical Trial of Fluoxetine in the Treatment of Obsessive–Compulsive Disorder,"*Journal of Clinical Psychopharmacology*, 6:98–101.

Gaviria, M., Gil, A.A., and Javaid, J.I. (1986),"Nortriptyline Kinetics in Hispanic and Anglo Subjects,"*Journal of Clinical Psychopharmacology*, 6:227–231.

Falkner, B., Hulman, S., and Kushner, H. (2004),"Effect of Birth Weight on Blood Pressure and Body Size in Early Adolescence,"*Hypertension*, 43:203–207.

Flegal, K.M., Troiano, R.P., Pamuk, E.R., et al. (1995),"The Influence of Smoking Cessation on the Prevalence of Overweight in the United States,"*New England Journal of Medicine*, 333:1165–1170.

Foltin, G., Markinson, D., Tunik, M., et al. (2002),"Assessment of Pediatric Patients by Emergency Medical Technicians-Basic,"*Pediatric Emergency Care*, 18:81–85.

Forland, S.C., Wechter, W.J., Witchwoot, S., et al. (1996),"Human Plasma Concentrations of *R*, *S*, and Racemic Flurbiprofen Given as a Toothpaste,"*Journal of Clinical Pharmacology*, 36:546–553.

Frei, E.III, Holland, J.F., Schneiderman, M.A., et al. (1958),"A Comparative Study of Two Regimens of Combination Chemotherapy in Acute Leukemia,"*Blood*, 13:1126–1148.

Freireich, E.J., Gehan, E., Frei, E.III, et al. (1963),"The Effect of 6–Mercaptopurine on the Duration of Steroid–Induced Remissions in Acute Leukemia: A Model for Evaluation of Other Potentially Useful Therapy,"*Blood*, 21:699–716. Froehlich, F., Hartmann, D., Guezelhan, C., et al. (1996),"Influence of Orlistat on the Regulation of Gallbladder Contraction in Man,"*Digestive Diseases and Sciences* 41:2404–2408.

Frost, W.H. (1936),Appendix to *Snow on Cholera*, London, Oxford University Press.

Galton, F. (1889), *Natural Inheritance*, London: MacMillan and Co.

Garland, M., Szeto, H.H., Daniel, S.S., et al. (1996),"Zidovudine Kinetics in the Pregnant Baboon,"*Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11:117–127.

Gehan, E.A. (1984),"The Evaluation of Therapies: Historical Control Studies,"*Statistics in Medicine*, 3:315–324.

Gibaldi, M. (1984), *Biopharmaceutics and Clinical Pharmacokinetics, 3rd Ed.*, Philadelphia: Lea & Febiger.

Gijsmant, H., Kramer, M.S., Sargent, J., et al. (1997),"Double-Blind, Placebo-Controlled, Dose-Finding Study of Rizatriptan (MK-642) in the Acute Treatment of Migraine,"*Cephalagia*, 17:647–651.

Glaser, J.A., Keffala, V., and Spratt, K. (2004),"Weather Conditions and Spinal Patients,"*Spine*, 29:1369–1374.

Gostynski, M., Gutzwiller, F., Kuulasmaa, K., et al. (2004),"Analysis of the Relationship Between Total Cholesterol, Age, Body Mass Index Among Males and Females in the WHO MONICA Project,"*International Journal of Obesity*, advance online publication, 22 June 2004, 1–9.

Grønbæk, M., Deis, A., Sørensen, T.I.A., et al. (1995),"Mortality Associated With Moderate Intakes of Wine, Beer, or Spirits,"*British Medical Journal*, 310:1165–1169.

Gupta, S.K., Manfro, R.C., Tomlanovich, S.J., et al. (1990),"Effect of Food in the Pharmacokinetics of Cyclosporine in Healthy Subjects Following Oral and Intravenous Administration,"*Journal of Clinical Pharmacology*, 30:643–653.

Gupta, S.K., Okerholm, R.A., Eller, M., et al. (1995),"Comparison of the Pharmacokinetics of Two Nicotine Transdermal Systems: Nicoderm and Habitrol,"*Journal of Clinical Pharmacology*, 35:493–498.

Hammond, E.C., and Horn, D. (1954),"The Relationship Between Human Smoking Habits and Death Rates,"*Journal of the American Medical Association*, 155:1316–1328.

Hartmann, D., Güzelhan, C., Zuiderwijk, P.B.M., and Odink, J. (1996), "Lack of Interaction Between and Orlistat and Oral Contraceptives,"*European Journal of Clinical Pharmacology*, 50:421–424.

Hauptman, J.B., Jeunet, F.S., and Hartmann, D. (1992),"Initial Studies in Humans With the Novel Gastrointestinal Lipase Inhibitor Ro 18–0647 (Tetrahydrolipstatin),*American Journal of Clinical Nutrition*, 55:309S–313S.

Hausknecht, R.U. (1995),"Methotrexate and Misoprostol to Terminate Early Pregnancy," *New England Journal of Medicine*, 333:537–540.

Hayden, F.G., Diamond, L., Wood, P.B., et al. (1996),"Effectiveness and Safety of Intranasal Ipratropium Bromide in Common Colds," *Annals of Internal Medicine*, 125:89–97.

Hennekens, R.U., Buring, J.E., Manson, J.E. (1996),"Lack of Effect of Long–Term Supplementation with Beta Carotene on the Incidence of Malignant Neoplasms and Cardiovascular Disease," *New England Journal of Medicine*, 334:1145–1149.

Hermansson, U., Knutsson, A., Brandt, L., et al. (2003),"Screening for High-Risk and Elevated Alcohol Consumption in Day and Shift Workers by Use of AUDIT and CDT," *Occupational Medicine*, 53:518–526.

Hill, A.B. (1953),"Observation and Experiment," *New England Journal of Medicine*, 248:995–1001.

Holford, N.H.G., and Sheiner, L.B. (1981),"Understanding the Dose–Effect Relationship: Clinical Application of Pharmacokinetic–Pharmacodynamic Models," *Clinical Pharmacokinetics*, 6:429–453.

Hollander, A.A.M.J., van Rooij, J., Lentjes, E., et al. (1995),"The Effect of Grapefruit Juice on Cyclosporine and Prednisone Metabolism in Transplant Patients," *Clinical Pharmacology & Therapeutics*, 57:318–324.

Hunt, D., Young, P., Simes, J. (2001),"Benefits of Pravastatin on Cardiovascular Events and Mortality in Older Patients with Coronary Heart Disease are Equal to or Exceed Those Seen in Younger Patients: Results from the LIPID Trial," *Annals of Internal Medicine*, 134:931–940.

Ingersoll, B. (1997),"Hoffman–La Roche's Obesity Drug Advances," *The Wall Street Journal*, May 15:B10.

Kadokawa, T., Hosoki, K., Takeyama, K., et al. (1979),"Effects of Nonsteroidal Anti–Inflammatory Drugs (NSAID) on Renal Excretion of Sodium and Water, and on Body Fluid Volume in Rats," *Journal of Pharmacology and Experimental Therapeutics*, 209:219–224.

Kaitin, K.I., Dicerbo, P.A., and Lasagna, L. (1991),"The New Drug Approvals of 1987, 1988, and 1989: Trends in Drug Development," *Journal of Clinical Pharmacology*, 31:116–122.

Kaitin, K.I., Richard, B.W., and Lasagna, L. (1987),"Trends in Drug Development: The 1985–86 New Drug Approvals," *Journal of Clinical Pharmacology*, 27:542–548.

Kaitin, K.I., Manocchia, M., Seibring, M., and Lasagna, L. (1994),"The New Drug Approvals of 1990, 1991, and 1992: Trends in Drug Development," *Journal of Clinical Pharmacology*, 34:120–127.

Kalbfleisch, J.D., and Street, J.O. (1990),"Survival Analysis". In *Statistical Methodology in the Pharmaceutical Sciences*, (D.A. Berry, ed.). New York: Marcel Dekker. pp.313–355.

Kametas, N.A., Krampl, E., McAuliffe, F., et al. (2004),"Pregnancy at High Altitude: A Hyperviscosity State," *Acta Obstetricia et Gynecologica Scandinavica*, 83:627-633.

Kaplan, E.L., and Meier, P. (1958),"Nonparametric Estimation From Incomplete Observations,"*Journal of the American Statistical Association*, 53:457–481.

Khan, A., Dayan, P.S., Miller, S., et al. (2002),"Cosmetic Outcome of Scalp Wound with Staples in the Pediatric Emergency Department: A Prospective, Randomized Trial,"*Pediatric Emergency Care*, 18:171–173.

Klausner, J.D., Makonkawkeyoon, S., Akarasewi, P., et al. (1996),"The Effect of Thalidomide on the Pathogenesis of Human Immunodeficiency Virus Type 1 and *M. tuberculosis* Infection,"*Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11:247–257.

Knick, V.C., Eberwein, D.J., and Miller, C.G. (1995),"Vinorelbine Tartrate and Paclitaxel Combinations: Enhanced Activity Against In Vivo P388 Murine Leukemia Cells,"*Journal of the National Cancer Institute*, 87:1072–1077.

Kruskal, W.H., and Wallis, W.A. (1952),"Use of Ranks in One–Criterion Variance Analysis,"*Journal of the American Statistical Association*, 47:583–621.

Lee, D.W., Chan, K.W., Poon, C.M., et al. (2002),"Relaxation Music Decreases the Dose of Patient-Controlled Sedation During Colonoscopy: A Prospective Randomized Controlled Tial,"*Gastrointestinal Endoscopy*, 55:33–36.

Lewis, R., Bennett, C.J., Borkon, W.D., et al. (2001),"Patient and Partner Satisfaction with Viagra (Sildenafil Citrate) Treatment as Determined by the Erectile Dysfunction Inventory of Treatment Satisfaction Questionnaire,"*Urology*, 57:960–965.

Linday, L.A., Pippenger, C.E., Howard, A., and Lieberman, J.A. (1995),"Free Radical Scavenging Enzyme Activity and Related Trace Metals in Clozapine–Induced Agranulocytosis: A Pilot Study,"*Journal of Clinical Psychopharmacology*, 15:353–360.

Linet, O.I., and Ogrinc, F.G. (1996),"Efficacy and Safety of Intracavernosal Alprostadil in Men with Erectile Dysfunction," *New England Journal of Medicine*, 334:873–877.

Lister, J. (1870),"Effects of the Antiseptic System of Treatment Upon the Salubrity of a Surgical Hospital,"*The Lancet*, 1:4–6,40–42.

Lombard, H.L., and Doering, C.R. (1928),"Cancer Studies in Massachusetts. 2. Habits, Characteristics and Environment of Individuals With and Without Cancer,"*New England Journal of Medicine*, 198:481–487.

Madhavan, S. (1990),"Appropriateness of Switches from Prescription to Over–the–Counter Drug Status,"*Journal of Pharmacy of Technology*, 6:239–242.

Mantel, N., and Haenszel, W. (1959),"Statistical Aspects of the Analysis of Data From Retrospctive Studies of Disease,"*Journal of the National Cancer Institute*, 22:719–748.

McLaughlin, J.K., Mandel, J.S., Blot. W.J., et al. (1984),"A Population–Based Case–Control Study of Renal Cell Carcinoma,"*Journal of the National Cancer Institute*, 72:275–284.

Medical Research Council (1948),"Streptomycin Treatment of Pulmonary Tuberculosis,"*British Medical Journal*, 2:769–782.

Medical Research Council (1950),"Clinical Trials of Antihistaminic Drugs in the Prevention and Treatment of the Common Cold,"*British Medical Journal*, 2:425–429.

Melbye, M., Wohlfahrt, J., Olsen, J.H., et al. (1997),"Induced Abortion and the Risk of Breast Cancer,"*New England Journal of Medicine*, 336:81–85.

Mendels, J., Camera, A., and Sikes, C. (1995),"Sertraline Treatment for Premature Ejaculation,"*Journal of Clinical Psychopharmacology*, 15:341–346.

Mikus, G., Trausch, B., Rodewald, C., et al. (1997),"Effect of Codeine on Gastrointestinal Motility in Relation to CYP2D6 Phenotype,"*Clinical Pharmacology & Therapeutics*, 61:459–466.

Modell, J.G., Katholi, C.R., Modell, J.D., and DePalma, R.L. (1997),"Comparative Sexual Side Effects of Bupropion, Fluoxetine, Paroxetine, and Sertraline,"*Clinical Pharmacology & Therapeutics*, 61:476–487.

Monforte, A.d'A., Mainini, F., Moscatelli, et al. (1996),"Survival in a Cohort of 1205 AIDS Patients from Milan," (correspondence) *AIDS* 10:798–799.

Montgomery, D.C. (1991), *Design and Analysis of Experiments, 3rd Ed.*, New York: Wiley.

Moyer, C.E., Goulet, J.R., and Smith, T.C. (1972),"A Study of the Effects of Long–Term Administration of Sulfcytine, a New Sulfonamide, on the Kidney Function of Man,"*Journal of Clinical Pharmacology*, 12:254–258.

Nguyen, T.D., Spincemaille, P., and Wang, Y. (2004),"Improved Magnetization Preparation for Navigator Steady-State Free Precession 3D Coronary MR Angiography,"*Magnetic Resonance in Medicine*, 51:1297–1300.

O'Brien, T.R., Blattner, W.A., Waters, D., et al. (1996),"Serum HIV–1 RNA Levels and Time to Development of AIDS in the Multicenter Hemophilia Cohort Study," *Journal of the American Medical Association*, 276:105–110.

Pahor, M., Guralnik, J.M., Salive, M.E., et al. (1996),"Do Calcium Channel Blockers Increase the Risk of Cancer?," *American Journal of Hypertension*, 9:695–699.

Pauling, L. (1971),"The Significance of the Evidence about Ascorbic Acid and the Common Cold," *Proceedings of the National Academy of Sciences of the United States of America*, 11:2678–2681.

Pocock, S.J. (1976),"The Combination of Randomized and Historical Controls in Clinical Trials," *Journal of Chronic Diseases*, 29:175–188.

Portner, T.S., and Smith, M.C. (1994),"College Students' Perceptions of OTC Information Source Characteristics," *Journal of Pharmaceutical Marketing and Management*, 8:161–185.

Psaty, B.M., Heckbert, S.R., Koepsell, T.D., et al. (1995),"The Risk of Myocardial Infarction Associated with Antihypertensive Drug Therapies," *Journal of the American Medical Association*, 274:620–625.

Redelmeier, D.A., and Tibshirani, R.J. (1997),"Association Between Cellular–Telephone Calls and Motor Vehicle Collisions," *New England Journal of Medicine*, 336:453–458.

Reith, J., Jørgensen, S., Pedersen, P.M., et al. (1996),"Body Temperature in Acute Stroke: Relation to Stroke Severity, Infarct Size, Mortality and Outcome," *The Lancet*, 347:422–425.

Ribeiro, W., Muscará, M.N., Martins, A.R., et al. (1996),"Bioequivalence Study of Two Enalapril Maleate Tablet Formulations in Healthy Male Volunteers," *European Journal of Clinical Pharmacology*, 50:399–405.

Rickels, K., Smith, W.T., Glaudin, V., et al. (1985),"Comparison of Two Dosage Regimens of Fluoxetine in Major Depression," *Journal of Clinical Psychiatry*, 46[3,Sec.2]:38–41.

Ruberg, S.J. (1996a),"Dose Response Studies. I. Some Design Considerations," *Journal of Biopharmaceutical Statistics*, 5:1–14.

Ruberg, S.J. (1996b),"Dose Response Studies. II. Analysis and Interpretation," *Journal of Biopharmaceutical Statistics*, 5:15–42.

Rubinstein, M.L., Haplern-Felsher, B.L., and Irwin, C.E. (2004),"An Evaluation of the Use of the Transdermal Contraceptive Patch in Adolescents," *Journal of Adolescent Health*, 34:395–401.

Salzman, C., Wolfson, A.N., Schatzberg, A., et al. (1995),"Effects of Fluoxetine on Anger in Symptomatic Volunteers with Borderline Personality Disorder,"*Journal of Clinical Psychopharmacology*, 15:23–29.

Sasomsin, P., Mentré, F., Diquet, B., et al. (2002),"Relationship to Exposure to Zidovudine and Decrease of P24 Antigenemia in HIV-Infected Patients in Monotherapy,"*Fundamental & Clinical Pharmacology*, 16:347–352.

Schömig, A., Neumann, F., Kastrati, A., et al. (1996),"A Randomized Comparison of Antiplatelet and Anticoagulant Therapy After the Placement of Coronary–Artery Stents,"*New England Jounal of Medicine*, 334:1084–1089.

Schwartz, J.I., Yeh, K.C., Berger, M.L., et al. (1995),"Novel Oral Medication Delivery System for Famotidine,"*Journal of Clinical Pharmacology*, 35:362–367.

Scott, J., and Huskisson, E.C. (1976),"Graphic Represntation of Pain,"*Pain*, 2:175–184.

Shepherd, J., Cobbe, S.M., Ford, I., et al. (1995),"Prevention of Coronary Heart Disease With Pravastatin in Men With Hypercholsterolemia," *New England Journal of Medicine*, 333:1301–1307.

Shimkin, M.B. (1941),"Length of Survival of Mice With Induced Subcutaneous Sarcomas,"*Journal of the National Cancer Institute*, 1:761–765.

Singh, N., Saxena, A., and Sharma, V.P. (2002),"Usefulness of an Inexpensive, Paracheck Test in Detecting Asymptomatic Infectious Reservoir of Plasmodium Falciparum During Dry Season in an Inaccessible Terrain in Central India,"*Journal of Infection*, 45:165–168.

Sivak-Sears, N.R., Schwarzbaum, J.A., Miike, R., et al. (2004),"Case-Contro Study of Use of Nonsteroidal Antiinflamatory Drugs and Glioblastoma Multiforme,"*American Journal Of Epidemiology*, 159:1131–1139.

Skeith, K.J., Russell, A.S., Jamali, F. (1993),"Ketoprofen Pharmacokinetics in the Elderly: Influence of Rheumatic Disease, Renal Function, and Dose," *Journal of Clinical Pharmacology*, 33:1052–1059.

Sperber, S.J., Shah, L.P., Gilbert, R.D., et al. (2004),"*Echinacea purpurea* for Prevention of Experimental Rhinovirus Colds,"*Clinical Infectious Diseases*, 38:1367–1371.

Spitzer, R.L., Cohen, J., Fleiss, J.L., and Endicott, J. (1967),"Quantification of Agreement in Psychiatric Diagnosis,"*Archives of General Psychiatry*, 17:83–87.

Stark, P.L., and Hardison, C.D. (1985),"A Review of Multicenter Controlled Studies of Fluoxetine vs. Imipramine and Placebo in Outpatients with Major Depressive Disorder,"*Journal of Clinical Psychiatry*, 46[3,Sec.2]:53–58.

Stein, D.S., Fish, D.G., Bilello, J.A., et al. (1996),"A 24–Week Open–Label Phase I/II Evaluation of the HIV Protease Inhibitor MK–639 (Indinavir)," *AIDS*, 10:485–492.

Steiner, M., Steinberg, S., Stewart, D., et al. (1995),"Fluoxetine in the Treatment of Premenstrual Dysphoria," *New England Jounal of Medicine*, 332:1529–1534.

Student (1931),"The Lanarkshire Milk Experiment," *Biometrika*, 23:398–406.

Umney, C. (1864),"On Commercial Carbonate of Bismuth," *Pharmaceutical Journal*, 6:208–209.

Vakil, N.B., Schwartz, S.M., Buggy, B.P., et al. (1996),"Biliary Cryptosporidiosis in HIV–Infected People After the Waterborne Outbreak of Cryptosporidiosis in Milwaukee," *New England Jounal of Medicine*, 334:19–23.

Wagner, J.G., Agahajanian, G.K., and Bing, O.H. (1968),"Correlation of Performance Test Scores with 'Tissue Concentration' of Lysergic Acid Diethylamide in Human Subjects," *Clinical Pharmacology and Therapeutics*, 9:635–638.

Wang, L., Kuo, W., Tsai, S., and Huang, K. (2003),"Characterizations of Life-Threatening Deep Cervical Space Infections: A Revieww of One Hundred Ninety Six Cases," *American Journal of Otolaryngology*, 24:111-117.

Wardle, J., Armitage, J., Collins, R., et al. (1996),"Randomised Placebo Controlled Trial of Effect on Mood of Lowering Cholesterol Concentration," *British Medical Journal*, 313:75–78.

Webber, M.P., Schonbaum, E.E., Farzadegan, H., and Klein, R.S. (2001), "Tampons as a Self-Administered Collection Method for the Detection and Quantification of Genital HIV-1," *AIDS*, 15:1417-1420.

Weinberg, M., Hopkins, J., Farrington, L., et al. (2004),"Hepatitis A in Hispanic Children Who Live Along the United States-Mexico Border: The Role of International Travel and Food-Borne Exposures," *Pediatrics*, 114:e68–e73.

Weiser, M., Reichenberg, A., Grotto, I., et al. (2004),"Higher Rates of Cigarette Smoking in Male Adolescents Before the Onset of Schizophrenia: A Historical-Prospective Cohort Study," *American Journal of Psychiatry*, 161:1219–1223.

Weltzin, R., Traina–Dorge, V., Soike, K., et al. (1996),"Intranasal Monoclonal IgA Antibody to Respiratory Syncytial Virus Protects Rhesus Monkeys Against Upper and Lower Respiratory Tract Infection," *Journal of Infectious Diseases*, 174:256–261.

Wilcoxon, F. (1945),"Individual Comparisons By Ranking Methods," *Biometrics*, 1:80–83.

Wissel, J., Kanovsky, P., Ruzicka, E., et al. (2001),"Efficacy and Safety of a Standardised 500 Unit Dose of Dysport (Clostridium Botulinum Toxin Type A Haemaglutinun Complex)in a Heterogeneous Cervical Dystonia Population: Results of a Prospective, Multicentre, Randomized, Double-Blind, Placebo-Controlled, Parallel Group Study," *Journal of Neurology*, 248:1073–1078.

Yuh, L. (1995),"Statistical Considerations for Bioavailability/Bioequivalence Studies." In *Pharmacokinetics: Regulatory, Industrial, Academic Perspectives*, 2^{nd} ed. (P.G. Welling and F.L.S. Tse, eds.). New York: Marcel Dekker. pp.479–502

Zazgonik, J., Huang, M.L., Van Peer, A., et al. (1993),"Pharmacokinetics of Orally Administered Levocabastine in Patients with Renal Insufficiency," *Journal of Clinical Pharmacology*, 33:1214–1218.

Zhi, J., Melia, A.T., Guerciolini, R., et al. (1994),"Retrospective Population–Based Analysis of the Dose–Response (Fecal Fat Excretion) Relationship of Orlistat in Normal and Obese Volunteers," *Clinical Pharmacology & Therapeutics*, 56:82–85.

Zhi, J., Melia, A.T., Guerciolini, R., et al. (1996),"The Effect of Orlistat on the Pharmacokinetics and Pharmacodynamics of Warfarin in Healthy Volunteers," *Journal of Clinical Pharmacology*, 36:659–666.

Zuna, R.E., and Behrens, A. (1996),"Peritoneal Washing Cytology in Gynecologic Cancers: Long–Term Follow–Up of 355 Patients," *Journal of the National Cancer Institute*, 88:980–987.