Course Notes for Statistics and the Physical World

Larry Winner
University of Florida
Department of Statistics

July 4, 2021

Contents

Chapter 1

Introduction

1.1 Basic Concepts of Statistical Analysis

Statistical tools and methods are used to describe data and make inferences regarding states of nature in a wide variety of areas of study. From simple graphs and numeric summaries provided in mainstream press to highly complex models used to describe measurements across a wide range of individuals or sampling units, we see reports making use of statistical tools and methods constantly. We will go through many of the commonly used methods in these notes.

After a brief introduction to **descriptive statistics**, making use of numeric and graphical summaries of variables, we will spend the remainder of the notes on **inferential statistics** that make use of information from a sample to make statements regarding a larger population of units. When conducting a study, researchers typically use the following strategy known as the **Scientific Method**.

1. Define the problem/research question of interest, including what to measure and all relevant conditions or groups to study.

2. Generate a hypothesis regarding the question of interest.

3. Construct one or more predictions based on the hypothesis.

4. Collect the data by means of a controlled experiment, observational study, or sample survey.

5. Summarize the data numerically in tabular form and/or graphically.

6. Analyze, interpret, and communicate the study's findings.

Many methods exist for the final part, data analysis, that we describe in detail in these notes. Many factors lead to the choice of the statistical methods to use for the analysis, including: data type(s), sampling method, and distributional assumptions regarding the measurements.

Populations will be thought of as the universe of units, while **samples** will refer to subsamples of the populations that are observed and measured. In practice, we observe the sample with the goal of making **inferences** regarding the corresponding population. Consider the following examples.

1.2 Data Collection

Once a research question has been posed, then data are collected to attempt to answer the question. Three common methods of collecting data are: controlled experiments, observational studies, and sample surveys.

In a **Controlled Experiment**, a sample of experimental units is obtained, and randomized to the various treatments or conditions to be compared. There are many ways that these can be conducted, and we will describe many variations of them throughout this course. Some elements of controlled experiments are given here.

Factors Variable(s) that are controlled by the experimenter (e.g. new drug vs placebo, 4 doses of a pesticide, 3 packages for food product)

Responses Measurements/Outcomes obtained during the experiment (e.g. change in blood pressure, weeds killed, consumer ratings for the product)

Treatments Conditions that are generated by the factor(s). When only 1 factor, these are the levels. With 2 or more factors, these are combinations of levels.

Experimental Unit Entity that is randomized to the Treatments. These can be individual items (patients in clinical trial, plants in botanical experiment) or groups of items (classrooms of students in an education experiment, pens of animals in a feed study).

Replications Treatments are assigned to more than one experimental unit, allowing for experimental error (variation) to be measured.

Measurement Unit Entity on which measurements are obtained. These can be experimental units when individuals are randomized, or subunits within the experimental units (students in a classroom, pigs in a pen).

Controlled experiments can be conducted in laboratories/hospitals/greenhouses, but can also be conducted in the "real world" where they are often referred to as "field studies" or "natural experiments."

There are many different treatment designs that are commonly applied. Some classes of designs are given below.

Single Factor Designs In these designs, there is a single factor to be studied with various levels.

Multi Factor Designs More than one factor is varied. Treatments correspond to combinations of factor levels.

Completely Randomized Designs Experimental units are randomly assigned to treatments with no restriction on randomization.

Height Dropped (H)	Distance Traveled (D)	Predicted Distance $\left(47.086\sqrt{H}\right)$
1000	1500	1489
828	1340	1355
800	1328	1332
600	1172	1153
300	800	816

Table 1.1: Measurements and predictions for Galileo's experiment with ramp and shelf on horizontal distance traveled as a function of height of ball drop

Randomized Block Designs Experimental units are grouped into homogeneous blocks, with treatments assigned so that each block receives each treatment.

Latin Square Designs Two or more blocking factors are available.

Repeated Measure Designs Units can be assigned to each treatment or be measured at multiple occasions on the same treatment.

Note that in designs with 2 or more factors, researchers are often interested in whether the effects of the levels of one factor depend on the levels of the other factor(s). When the effects do depend on the levels of the other factor, this is referred to as an **interaction**.

Example 1.1: Galileo's Experiments with Gravity

Experimental work by Galileo has been described and analyzed (Dickey and Arnold (1995) [19]). Two experiments involved rolling a ball down a ramp and measuring the horizontal distance traveled by the ball as a function of the height at which the ball was dropped. One set of measurements contained only a ramp, the second set of measurements had a ramp and a flat shelf at the bottom of the ramp.

One theory is that the horizontal distance traveled increases with the height at which the ball is dropped on the ramp. However, the rate of change should decrease with height. Another restriction is that the distance traveled should be 0 when the height it is dropped at is 0. One mathematical equation that could be used to relate Distance (D) to Height (H) is the following.

$$D = \alpha + \beta\sqrt{H}$$

In this formulation, it is expected that $\alpha = 0$, that is, that $D = 0$ when $H = 0$ and that $\beta > 0$. The authors fit a regression model and, first found no evidence that $\alpha \neq 0$. Then they fit a model without the intercept and found that $D = 47.086\sqrt{H}$. Table 1.1 contains the data and the predictions based on the equation for 5 observations. As seen in the table, the predictions are very close to the observed values.

$$\nabla$$

Example 1.2: Factors Effecting Color Strength of Dyes Applied to Modified Cotton

An experiment was conducted to measure the effects of 4 factors on color strength measured as K/S (Ben Ticha, Haddar, Meksi, Guesmi, and Mhenni (2016) [7]). Each factor was set at 2 levels and the experiment included all $2^4 = 16$ combinations of the factor levels. The factors studied and their levels were: Cationizing Agent Amount (5%, 10%), pH (5, 11), Dying Temperature (40C, 100C), and Drying Time (30min, 60min).

∇

In many settings, it is not possible or ethical to assign units to treatments. For instance, when comparing quality of products of various brands, you can take samples from the various brands, but not assign "raw materials" at random to the brands. Studies comparing residents of various parts of a country can only take samples of residents from the areas, not assign people to them. In studies of the effects of smoking or drinking, it is unethical to assign subjects to the conditions. In all of these cases, we refer to these as **Observational Studies**. Typically the method of analysis is the same for controlled experiments and observational studies, however the ability to imply "cause and effect" is more difficult in observational studies than controlled experiments. Researchers in such studies must try and control for any potential alternative explanations of the association. For an interesting discussion of various aspects of observational studies, including: external validity (generalizing results beyond the original study), causation, reliability of measurement, and inclusion of covariates, involving study of interruption and multitasking, see Walter, Dunsmuir, and Westbrook (2015) [59].

1.3 Variable Types

In most settings, researchers have one or more "output" variable(s) and one or more "input" variable(s). For instance, a study comparing salaries among males and females would have the output variable be salary and possible input variables: gender (1 if female, 0 if male), experience (years), and education (years). The output variables are often referred to as **dependent variables**, **responses**, or **end points**. The input variables are often referred to as **independent variables**, **predictors**, or **explanatory variables**.

Variables are measured on different scales, and the data analysis methods are determined by variable types. Variables can be **categorical** or **numeric**. Categorical variables can be **nominal** or **ordinal**, while numeric variables can be **discrete** or **continuous**.

Examples of nominal variables include gender, hair color, and automobile make. These are categories with no inherent ordering. Ordinal variables are categorical, but with an inherent ordering, such as: strongly disagree, disagree, neutral, agree, strongly agree. Discrete variables can take on only a finite or countably infinite set of values, these can be counts of number of occurrences of an event in a series of trials or in a fixed time or space, or the number facing up on a roll of a dice. Continuous variables can take on any value along a continuum, such as temperature, time, or blood pressure. When discrete variables take on many values, they are often treated as continuous, and continuous variables are often reported as discrete values.

Example 1.3: Consistency of Ratings Based on a Rating Scale for Videostroboscopy

A study was conducted to measure inter-rater and intra-rater reliability of the Voice-Vibratory Assessment with Laryngeal Imaging (VALI) rating form for assessing videostroboscopy and high-speed videoendoscopic (HSV) recordings (Poburka, Patel, and Bless (2017) [48]). Table 1.2 contains information on the 30

Subject	Age	Gender	Dysphonia	Subject	Age	Gender	Dysphonia	Subject	Age	Gender	Dysphonia
1	10	M	3	11	45	F	3	21	57	F	2
2	19	M	1	12	47	F	3	22	59	F	2
3	27	F	1	13	48	M	1	23	60	F	3
4	32	M	1	14	49	F	2	24	60	M	1
5	37	F	2	15	50	F	3	25	62	F	2
6	37	M	0	16	51	F	3	26	62	M	3
7	39	F	3	17	51	M	0	27	64	F	3
8	42	F	2	18	51	M	0	28	70	M	3
9	44	F	2	19	53	F	1	29	77	F	3
10	45	F	2	20	57	F	3	30	89	F	2

Table 1.2: Age, Gender, and Dysphonia Grade for 30 Subjects - VALI Study

subjects in the study. These include: subject ID, Age (continuous, reported as a discrete variable), gender (nominal), and an overall dysphonia grade (ordinal, with 0=normal, 1=mild, 2=moderate, 3=severe).

∇

Chapter 2

Describing Data

Once data have been collected, they are typically described via graphical and numeric means. The methods used to describe the data will depend on its type (nominal, ordinal, or numeric). We also need to distinguish whether the data corresponds to a sample or a population. In this chapter, we focus purely on describing a set of measurements, not making inferences. First we consider graphical and numeric descriptions of a single variable. Then we consider pairs of variables.

2.1 Graphical Description of a Single Variable

Depending on the type of measurement, common plots are **pie charts**, **bar charts**, **histograms**, **box plots**, and **density plots**.

Pie charts can be used to describe any variable type. Continuous numeric variables must be collapsed into "bins" or "buckets." The size of the sectors of the pie represent the relative frequency of each category.

Bar charts are used to describe nominal or ordinal data. The variable levels are arrayed on the bottom (or left side) of the plot and bars above (or beside) the levels represent the frequency or relative frequency of the number of observations belonging to the various categories.

Histograms are used for numeric variables, where the heights of the bars above the bins represent the frequency or relative frequency of the various bins.

Box plots are used on numeric variables. They identify particular percentiles of a distribution and are useful in detecting outlying observations and spread in the distribution.

Density plots are used for numeric variables. They offer a smoother description of the measurements than a histogram does and are simple to obtain with modern statistical software packages.

Example 2.1: Measurements of the Velocity of Light circa 1931-1933

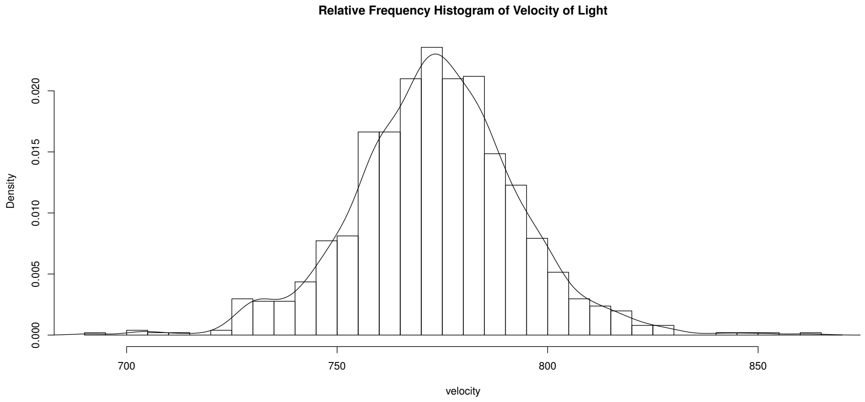

Figure 2.1: Relative Frequency Histogram and Smooth Density for 1010 measurements of the velocity of light made near Irvine, CA during 1931-1933. Data represent the measured value - 299000 km/sec

A.A. Michelson, F.G. Pease, and F. Pearson set up an approximately one mile long tube to make determinations of the speed of light near Irvine, CA in the early 1930s (Michelson, Pease, and Pearson, (1935), [43]). Without getting into the very detailed description given in the paper, we have 1010 determinations of the velocity of light after having removed some runs with anomalous values in the table. Further, we do not include weights that varied due to the experimental protocol as it evolved during the data collection process. Figure 2.1 provides a histogram of the $n = 1010$ measurements (approximated from their tabular information) as well as a smooth density function overlay on the graph. The values on the graph represent velocity - 299000 km/sec. The individual measurements are mound shaped around a center point with arithmetic mean of 299773.5 km/sec. Modern assessments of the velocity of light in a vacuum is 299792.5 km/sec.

Measurements were made in 4 groups of series: Series 1-54 (2/16/1931-7/14/1931), Series 55-110 (3/3/1932-5/13/1932), Series 111-158 (5/13/1932-8/4/1932), and Series 159-233 (12/3/1932-2/27/1933). Side-by-side box plots are given in Figure 2.2. The box plot identifies from bottom to top the following elements.

1. Minimum: Bottom of line at bottom of plot (or the lowest circle)
2. Range for lowest 25% of measurements: Distance from minimum observation to the bottom of the box
3. 25th percentile (Lower Quartile, aka LQ): Bottom line of box
4. Range for the 25th to 50th percent of participants: Distance between bottom of box and second horizontal line
5. Median (50th percentile): Second horizontal line
6. Range for the 50th to 75th percent of participants: Distance between second horizontal line and top of box
7. Interquartile Range (IQR): Distance between top (75th percentile) and bottom (25th percentile) of the box
8. 75th percentile (Upper Quartile, UQ): Top line of the box

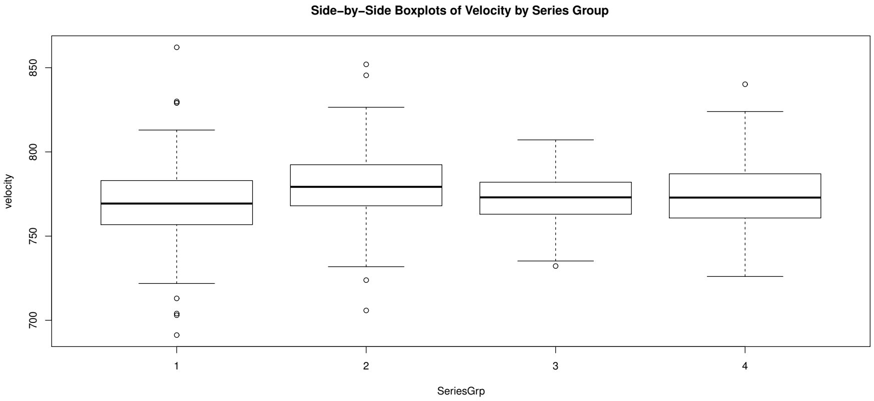

Figure 2.2: Side-by-side boxplots for velocities measured in the 4 groups of series

9. Range for 75th to 100th percent of participants: Distance from the top of the box to the maximum observation
10. Maximum: Top of line at the top of plot (or the highest circle)
11. Lower line extends either to the minimum or 1.5(IQR) below the LQ, whichever is shortest.
12. Upper line extends either to the maximum or 1.5(IQR) above the UQ, whichever is shortest.
13. Circles represent outlying measurements (very extreme measurements).

The precision of the measurements tend to improve slightly over the course of the study. Note that the average weights for the individual measurements included in this analysis were approximately 1.65 for the first series and approximately 3 for the remaining series.

∇

Example 2.2: Body Mass Index for National Hockey League Players - 2013/2014 Season

Body mass index (BMI) is a measure of body fat that is based on the the work of Adolphe Quetelet, a renowned Belgian researcher in astronomy and statistics and other areas, particularly social sciences. In terms of metric units, BMI is mass(kg)/height(m)^2; in the American system, BMI is 703×mass(lbs)/height(in)^2. Data for all National Hockey League (NHL) players are obtained, reported in pounds (lbs) and inches, discretely. A histogram is given in Figure 2.3. The histogram is approximately symmetric and mound-shaped, centered above 26.

∇

Example 2.3: Female and Male Speeds at Washington, DC Rock and Roll Marathon - 2015

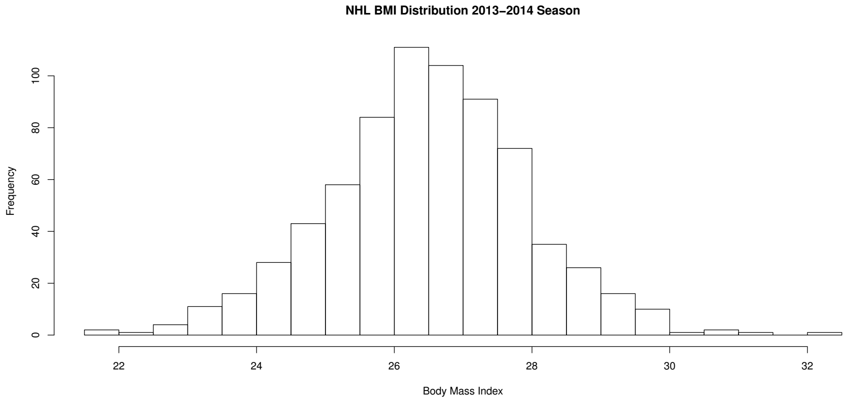

Figure 2.3: Body Mass Index for 2013/2014 season National Hockey League Players

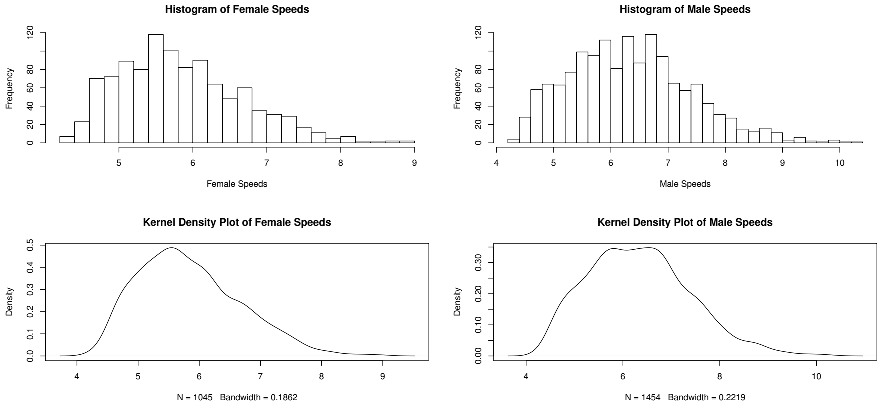

Figure 2.4: Histograms and density plots of Rock and Roll marathon speeds by gender

The 2015 Rock and Roll Marathon in Washington, D.C. was completed by 1045 female and 1454 male participants. Each participant's time to complete the marathon was converted to a speed (miles per hour). Histograms and kernel density plots for females and males are given in Figure 2.4, and side-by-side box plots are given in Figure 2.5. For both genders, there tend to be more cases at lower speeds with a few extreme cases with higher speeds. These distributions are **right-skewed**.

A smooth version of a boxplot, which does not separate the measurements into quantiles is a **violin plot**. For the marathon data, one is displayed in Figure 2.6.

∇

Time series plots are widely used in many areas including economics, finance, climatology, and biology. These graphs include one or more characteristics being observed in a sequential time order. These plots can be based on virtually any level of sampling interval.

Example 2.4: Miami Monthly and Annual Mean Temperature 1/1949-12/2014

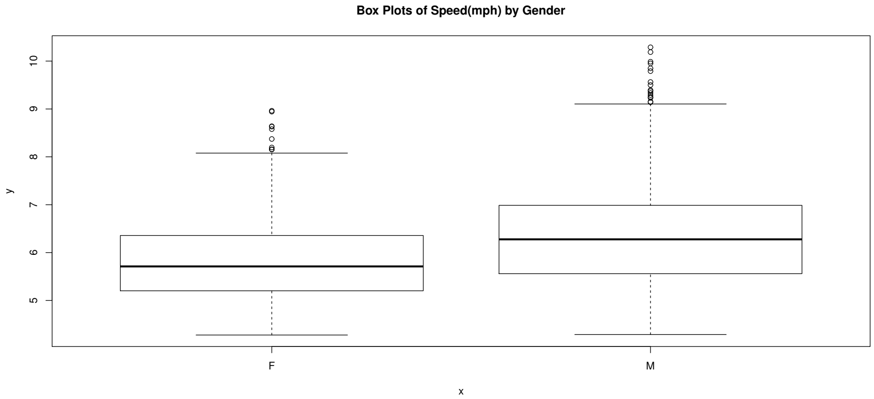

Figure 2.5: Side-by-side box plots of Rock and Roll marathon speeds by gender

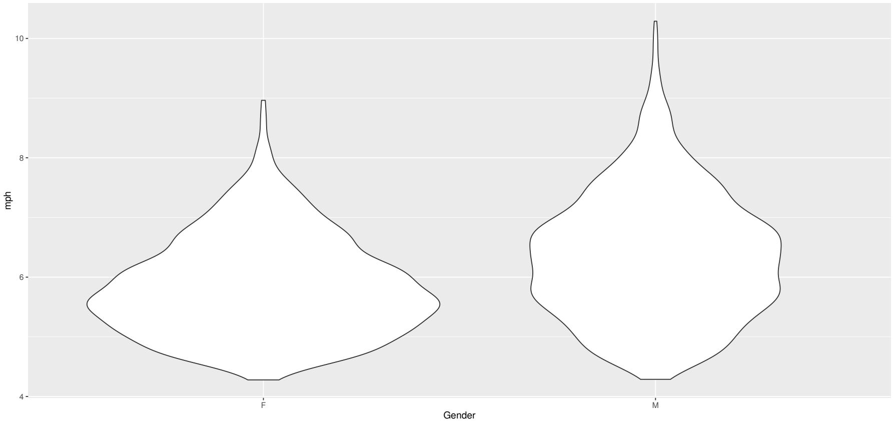

Figure 2.6: Side-by-side violin plots of Rock and Roll marathon speeds by gender

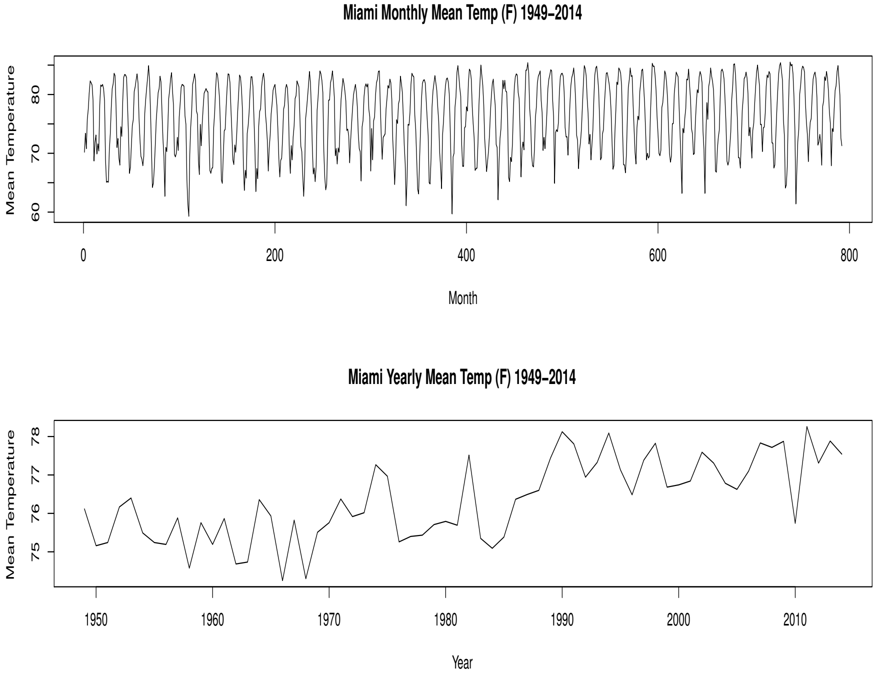

Figure 2.7: Monthly Mean Temperature in Miami, FL (January 1949 - December 2014)

They can be used to detect trend and cyclical patterns over time. Figure 2.7 shows the the monthly and annual mean temperature in Miami for the years 1949 through 2014. Clearly there is a cyclical pattern occurring within years, and after a flat early annual series, there certainly appears to be evidence of an increasing trend over approximately the second half of the series (after about 1970).

∇

2.2 Numerical Descriptive Measures of a Single Variable

Numerical descriptive measures describe a set of measurements in quantitative terms. When describing a **population** of measurements, they are referred to as **parameters**; when describing a **sample** of data, they are referred to as **statistics**.

In terms of nominal and ordinal data, **proportions** are generally the numeric measures of interest. These are simply the fraction of measurements falling into the various possible levels (and must sum to 1). For ordinal variables, the **cumulative proportions** are also of interest, representing the fraction of measurements falling in or below the various categories.

2.2.1 Measures of Central Tendency

There are two commonly reported measures of central tendency, or location for a set of measurements. The **mean** is the sum of all measurements divided by the number of measurements, and is reported often as "per capita" in economic reports. The mean is the "balance point" of a set of measurements in a physical sense. The **median** is the point where half of the measurements fall at or below it, and half of the measurements fall at or above it. It is also the 50th percentile of the set of measurements. Many economic reports state median values. A third, less reported measure is the **mode** which really is only appropriate for discrete variables, and is the value that occurs most often. For a histogram of discretely measured data, the mode is the level with the highest bar.

Note that the mean is affected by outlying measurements, as it is the sum of all measurements, evenly distributed among all of the measurements. The median is more "robust" as it is not effected by the actual values of individual measurements, only the center of them. The formulas for the population mean μ, based on a population of N items and the sample mean $\overline{y}$ for a sample of n items are given below.

$$\text{Population Mean: } \mu = \frac{\sum_{i=1}^{N} y_i}{N} \qquad\qquad \text{Sample Mean: } \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

To obtain the median, measurements are ordered from smallest to largest, and the middle observation (odd population/sample size) or the average of the middle two observations (even population/sample size) are identified.

Examples 2.2, 2.3 Continued: NHL BMI's and Rock and Roll Marathon Speeds

Using the **mean** and **median** functions in R, we obtain the population means for NHL BMI's and marathon speeds by gender for the Rock and Roll marathon.

R Output

```
### Output

> cbind(head(bmi.nhl.sort), tail(bmi.nhl.sort))
         [,1]     [,2]
[1,] 21.56757 29.98314
[2,] 21.75521 30.12259
[3,] 22.14871 30.51215
[4,] 22.64680 30.82813
[5,] 22.75987 31.39688
[6,] 22.75987 32.00386
> round(bmi.cent.out, 4)
       N      sum    mean  median
[1,] 717 19000.61 26.5002 26.5159
>
> ### Use built-in mean and median functions
> mean(bmi.nhl)
[1] 26.50015
> median(bmi.nhl)
[1] 26.51586
```

Note that the mean (26.50) and median (26.52) are very close, as is expected for an (approximately) symmetric distribution.

For the marathon speeds, we use the **tapply** function in R that will compute functions separately for different groups (gender).

R Output

```
> tapply(mph,Gender,mean)
       F        M
5.839839 6.336979
> tapply(mph,Gender,median)
       F        M
5.711109 6.276599
```

These distributions are skewed-right, with a few very fast runners in each gender. This causes the means (F=5.84, M=6.37) to be larger than the medians (F=5.71, M=6.28).

∇

Example 2.5: James Short's Measurements of the Sun's Parallax

The parallax is defined as (Merriam-Webster Dictionary):

> the apparent displacement or the difference in apparent direction of an object as seen from two different points not on a straight line with the object.
>
> especially : the angular difference in direction of a celestial body as measured from two points on the earth's orbit.

James Short reported $n = 158$ measurements of the parallax of the sun in seconds of a degree (Short (1763) [54], also reported in Stigler (1977) [55]). From the data on the class website, we obtain the following quantities:

$$n = 158 \qquad \sum_{i=1}^{n} y_i = 1360.31 \qquad \overline{y} = \frac{1360.31}{158} = 8.610 \qquad \text{Median} = 8.55$$

The true value is 8.798. A histogram of the data, the true value, and sample mean, as well as a box plot of the measurements are given in Figure 2.8.

R Output

```
> length(prlxsun)
```

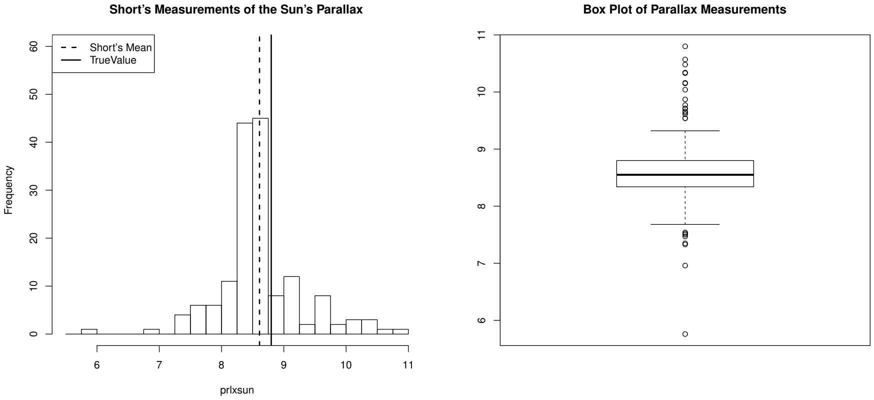

Figure 2.8: James Short's measurements of the sun's parallax. Left: Histogram, Right: Box plot

```
[1] 158
> sum(prlxsun)
[1] 1360.31
> mean(prlxsun)
[1] 8.609557
> median(prlxsun)
[1] 8.55
```

∇

Outliers are observations that lie "far" away from the others. These may be data that have been entered erroneously or just individual cases that are quite different from others. As stated above, means can be affected by outliers, while medians generally are not. A measure of the mean that is not affected by outliers is the **trimmed mean**. This is the mean of observations in the "middle" of the measurements. For instance, a 90% trimmed mean is the mean of the middle 90% of the ordered measurements (removing the smallest 5% and largest 5%).

Note that the Short parallax data has some extreme outliers in the box plot. The 90% trimmed mean is 8.594, which is not far from the sample mean as the data are still fairly symmetric despite the outliers.

2.2.2 Measures of Variability

Along with the "location" of a set of measurements, researchers are also interested in their variability (aka dispersion). The **range** is the distance between the largest and smallest measurements (note that this differs from the standard meaning which would just give the lowest and highest values). The **interquartile range** (IQR) is the distance between the 75th percentile (3/4 of measurements lie below it) and the 25th percentile (1/4 of the measurements lie below it). That is, the IQR measures the range for the middle half of the ordered measurements.

Measures that are more widely used in making inferences are the **variance** and its square root, the **standard deviation**. In terms of measurements, the variance is approximately the average squared distance of the individual measurements from the mean (for a population, it is the average). The formulas for the population and sample variance are given below. Note that unless stated otherwise specifically, software packages are reporting the sample version.

$$\text{Population Variance: } \sigma^2 = \frac{\sum_{i=1}^{N}\left(y_i-\mu\right)^2}{N} \qquad \text{Sample Variance: } s^2 = \frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{n-1}$$

The reason for dividing by $n-1$ in the sample variance is to make the estimator an unbiased estimator for the population variance. That is, when computed across all possible samples, the "average" of the sample variance will be the population variance. The standard deviation is the positive square root of the variance and is in the same units as the measurements. The population standard deviation is denoted as σ, the sample standard deviation is denoted as s. For many (but certainly not all) distributions, approximately 2/3 of the measurements lie within one standard deviation of the mean and approximately 19/20 lie within two standard deviations of the mean.

Example 2.2, 2.3 Continued: NHL BMI's and Rock and Roll Marathon Speeds

We compute the range, interquartile range, variance, and standard deviations for the NHL BMI's and the Rock and Roll mathon speeds by gender. Since we treat each of these as a population, we will make a slight adjustment to R's "built-in" functions **var** and **sd**, which compute the sample versions by default.

R Output

```
### Output
> var(bmi.nhl)                # Sample Variance with "var" function
[1] 2.116228
> (N-1)*var(bmi.nhl)/N        # Pop variance with "var" function
[1] 2.113277
> sd(bmi.nhl)                 # Sample Std Dev with "sd" function
[1] 1.454726
> sqrt((N-1)/N)*sd(bmi.nhl)   # Population Std Dev with "sd" function
[1] 1.453711
> round(bmi.var.out1, 3)
      min    max  range    LQ     UQ   IQR
   21.568 32.004 10.436 25.62 27.439 1.819
> round(bmi.var.out2, 3)
     mean sum(dev^2) sigma^2   s^2 sigma     s P(mu+/-1sigma) P(mu+/-2sigma)
[1,] 26.5   1515.219   2.113 2.116 1.454 1.455          0.706          0.946
```

For the marathon speeds, we will simply use the **var** and **sd** functions in R, applied separately to Females and Males. As both population sizes exceed 1000, the adjustment for population variances and standard deviations would be very small.

R Output

```
### Output
> round(rr.var.out, 3)
           N  mean sigma^2 sigma P(mu+/-1sigma) P(mu+/-2sigma)
Females 1045 5.840   0.691 0.831          0.662          0.964
Males   1454 6.337   1.119 1.058          0.665          0.964
```

Male speeds tend to be higher and more variable than Female speeds. All three distributions have approximately 2/3 of individuals lying with one standard deviation of the mean, and approximately 95% lying within two standard deviations from the mean.

∇

Example 2.5 Continued: James Short's Measurements of the Sun's Parallax

We compute the range, interquartile range, variance, and standard deviations for the sample of $n = 158$ sun parallax measurements.

R Output

```
> (n <- length(prlxsun))
[1] 158
> (range <- max(prlxsun) - min(prlxsun))
[1] 5.04
> (IQR <- quantile(prlxsun,0.75) - quantile(prlxsun,0.25))
0.445
> (var <- var(prlxsun))
[1] 0.4545164
> (sd <- sd(prlxsun))
[1] 0.6741783
> sum(prlxsun >= mean(prlxsun)-sd & prlxsun <= mean(prlxsun)+sd) / n
[1] 0.778481
> sum(prlxsun >= mean(prlxsun)-2*sd & prlxsun <= mean(prlxsun)+2*sd) / n
[1] 0.9367089
```

The full set of measurements lie within a range of 5.04 seconds of a degree, while the middle 50% lie within a range of 0.445. The variance is 0.455 and the standard deviation (a typical distance from an observation to the mean) is 0.674. Further, approximately 77.8% of measurements lie with one standard deviation and 93.7% lie within two standard deviations of the mean.

∇

2.3 Describing More than One Variable

So far, we have looked at cases one variable at a time, although the marathon speed data set has two variables: speed and gender. Now we consider describing relationships when two variables are observed on each sampling/experimental unit. These can be extended to more than two variables, but can be harder to visualize. We consider graphical techniques as well as numerical measures. Keep in mind that variable types (nominal, ordinal, and numeric) will dictate which method(s) is (are) appropriate.

When both variables are categorical (nominal or ordinal), two methods of plotting them are **stacked bar graphs** and **cluster bar graphs**. For the stacked bar graph, one variable is on the horizontal axis

		Column				
		1	2	$\cdots$	c	Total
Row	1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
	2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	$n_{..}$

Table 2.1: Contingency Table for Row Variable with r levels, and Column variable with c columns

(one slot for each level) and the other variable is displayed within the bars with subcategories for each of its levels. In a cluster (grouped) bar graph, one variable forms “major groupings,” while the second variable is plotted “side-by-side” within the groupings. Both methods are based on results of a **contingency table** also known as a **crosstabulation**. These are tables where rows are the levels of one categorical variable, columns are levels of another variable, and numbers within the table are counts of the number of units falling in that cell (combination of variable levels). Often these are converted into proportions either overall (cell probabilities sum to 1), or within rows or columns marginally. A contingency table is typically of the form in Table 2.1.

Example 2.6: Thumb Styles of Blues Guitarists by Region and Period

A study reported hand and thumb styles of Blues guitarists as well as the region they were from and when they were born (Cohen (1996) [16]). The regions are 1=East, 2=Delta, and 3=Texas. The thumb styles are 1=Alternating, 2=Utility, and 3=Dead. The birth period was labeled post1906 with 0=Born before 1906, 1=born after 1906. First, the association between region (row) and thumb style (column) is considered, then birth period is added. The crosstabulations are given below in the R code. Figure 2.9 gives the Stacked and Cluster Bar Graphs.

R Output

```
### Output

> (reg_ts <- table(region, thumbSty))
       thumbSty
region  Alternating Utility Dead
  East           20       8    7
  Delta           9      19   19
  Texas           1       2    8
> ## Obtain Row (1) and Column (2) Marginal Totals
> margin.table(reg_ts,1)
region
 East Delta Texas
   35    47    11
> margin.table(reg_ts,2)
thumbSty
Alternating     Utility        Dead
         30          29          34
> ## Obtain Proportions across all Cells
> reg_ts/sum(reg_ts)
       thumbSty
region  Alternating    Utility       Dead
  East   0.21505376 0.08602151 0.07526882
  Delta  0.09677419 0.20430108 0.20430108
```

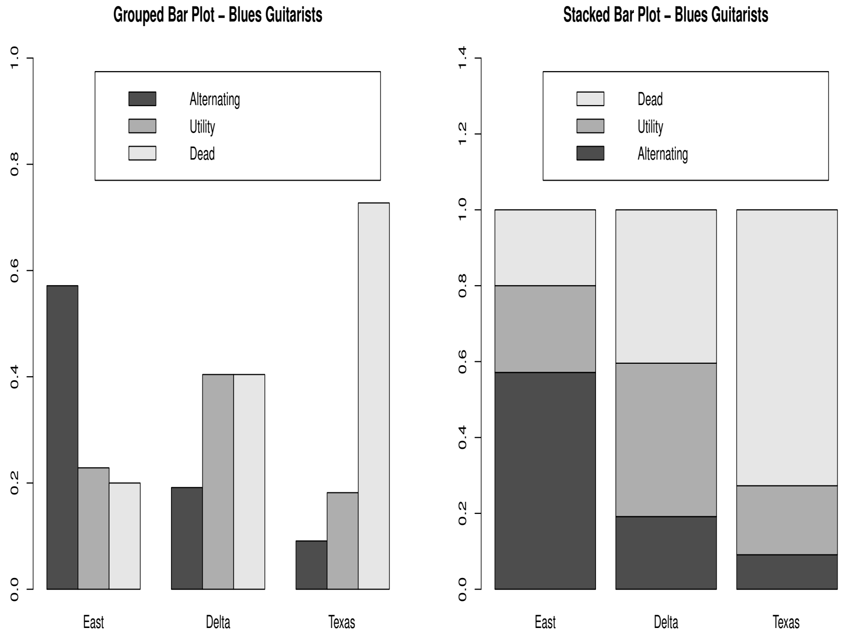

Figure 2.9: Stacked and Cluster (Grouped) Bar Charts - Blues Guitarists - Region and Thumb Style

```
  Texas  0.01075269 0.02150538 0.08602151
> ## Obtain Row Proportions (Thumb Style w/in Region)
> prop.table(reg_ts,1)
       thumbSty
region  Alternating    Utility       Dead
  East   0.57142857 0.22857143 0.20000000
  Delta  0.19148936 0.40425532 0.40425532
  Texas  0.09090909 0.18181818 0.72727273
> ## Obtain Column Proportions (Region w/in Thumb Style)
> prop.table(reg_ts,2)
       thumbSty
region  Alternating    Utility       Dead
  East   0.66666667 0.27586207 0.20588235
  Delta  0.30000000 0.65517241 0.55882353
  Texas  0.03333333 0.06896552 0.23529412
```

∇

When the independent variable is categorical (nominal or ordinal) and the response (dependent variable) is numeric, we can construct side-by-side histograms and density plots, or box plots (see Figure 2.2 for side-by-side box plots). Histograms and densities can also be placed into single plots with different colors or patterns.

When two variables (labeled x and y) are both numeric, one numeric descriptive measure that is widely reported is the **correlation** between the two variables. Technically, this is called the Pearson product moment coefficient of correlation. This measure is only for the **linear**, or “straight line” relation between the two variables. Unlike in Regression (described later), the variables are not necessarily (but can be) identified as an independent and or dependent variable. The formula for this measure (population and sample) are given below.

$$\text{Population Correlation: } \rho = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i-\mu_x\right)\left(y_i-\mu_y\right)}{\sigma_x\sigma_y} = \frac{\sum_{i=1}^{N}\left(x_i-\mu_x\right)\left(y_i-\mu_y\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i-\mu_x\right)^2\sum_{i=1}^{N}\left(y_i-\mu_y\right)^2}}$$

$$\text{Sample Correlation: } r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{s_xs_y} = \frac{\sum_{i=1}^{N}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i-\overline{x}\right)^2\sum_{i=1}^{N}\left(y_i-\overline{y}\right)^2}}$$

A **scatterplot** is a plot where each case’s x and y pairs are plotted in two dimensions. When one variable is the dependent variable, it is labeled y, and plotted on the vertical axis and the independent variable is labeled x, plotted on the horizontal axis. We are interested in any pattern (linear or possibly nonlinear, or none at all) between the variables. The formulas for the (ordinary) least squares regression line relating y to x are given below.

$$\hat{y} = \hat{\beta}_0+\hat{\beta}_1x \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \qquad \hat{\beta}_0 = \overline{y}-\hat{\beta}_1\overline{x} \qquad SSE = \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i-\left(\hat{\beta}_0+\hat{\beta}_1x_i\right)\right)^2$$

Example 2.7: Relation Between Temperature and Water Evaporation

An experiment was conducted that observed the temperature x (fahrenheit) and water evaporation y (grains of water) with measurements taken at 8:00AM daily from 11/10/1692-11/09/1693 (Halley (1694) [28]).

The plot of the data and the linear regression equation was obtained in R and is given in Figure 2.10. The correlation and regression equation were obtained using the **cor** and **lm** functions.

R Output

```
> cor(dayEvap, dayTemp)
[1] 0.7961281

> mod1 <- lm(dayEvap ~ dayTemp)
> summary(mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94556    1.09028   21.96   <2e-16 ***
dayTemp      0.62879    0.02509   25.07   <2e-16 ***
```

The sample correlation is $r = 0.7691$ and the fitted linear regression equation is $\hat{y} = 23.9456 + 0.6288x$.

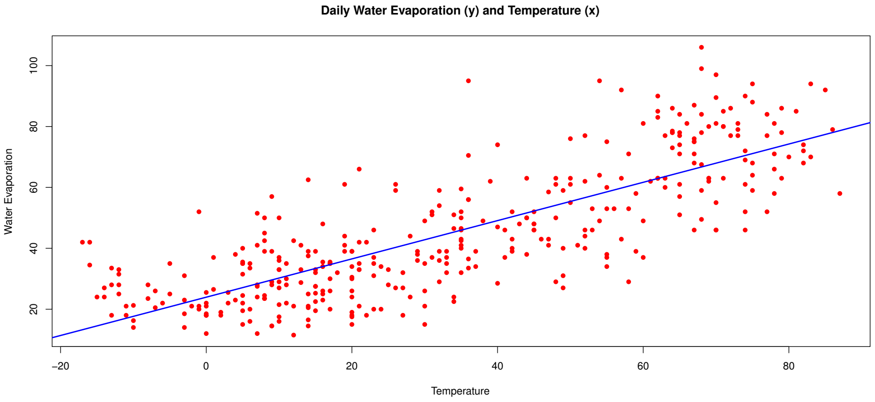

Figure 2.10: Daily Water Evaporation (y) and Temperature (x) for experiment conducted 11/10/1692-11/09/1693 reported by Edmund Halley

∇

Example 2.8 - Heights of Adult Children and Their Parents

Francis Galton measured many aspects of humans, plants, and animals during the late 1800s, some of which were presented in table form in his book *Natural Inheritance*. One analysis that had been published previously (Galton (1886) [25]) introduced the notion of linear regression. Galton reported the heights of adult children and their "mid-parents" which was the average height of the parents. Galton multiplied female heights for the adult children and the mothers by 1.08 to make the female and male heights "comparable." The individual data were obtained from Galton's notebooks and are available due to Professor James A. Hanley (Hanley (2004), [30]).

Histograms of the male and (unscaled) female heights is given in Figure 2.11. The histograms are approximately mound-shaped within gender. The plot of adult child height versus mid-parent height (with female heights scaled by 1.08) is given in Figure 2.12. The plot contains three lines, which are described below.

- Steepest: Line of equality $y = x$, which represents the case with the average adult child height equaling the mid-parent height.
- Flat: Constant line $y = \overline{x}$, which represents the case with the average adult child height equaling the average mid-parent height (no association between adult child and mid-parent height).
- Middle: Least squares regression line $y = 18.77 + 0.73x$

The fact the least squares line falls between the two reference lines showed that adult children of tall parents tended to be tall, but not as tall on average as their parents. Similarly, adult children of short parents tended to be short, but not as short on average as their parents. Galton referred to this phenomenon as "regression to mediocrity." Today it is more widely referred to as "regression to the mean."

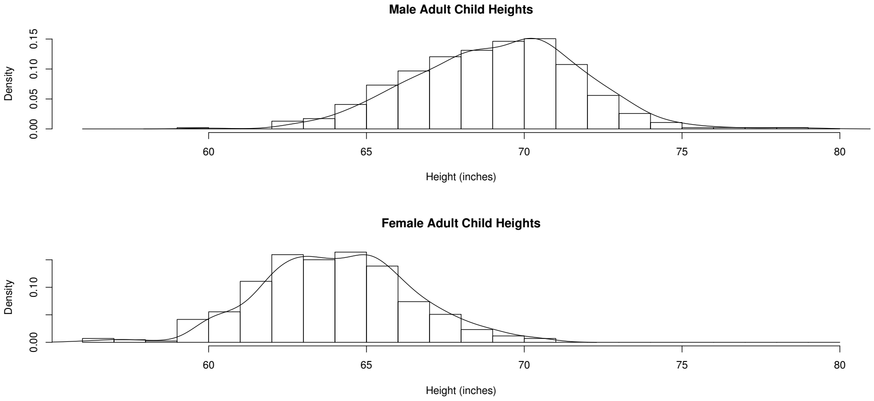

Figure 2.11: Histograms and smooth densities of adult child heights by gender in Frances Galton's data

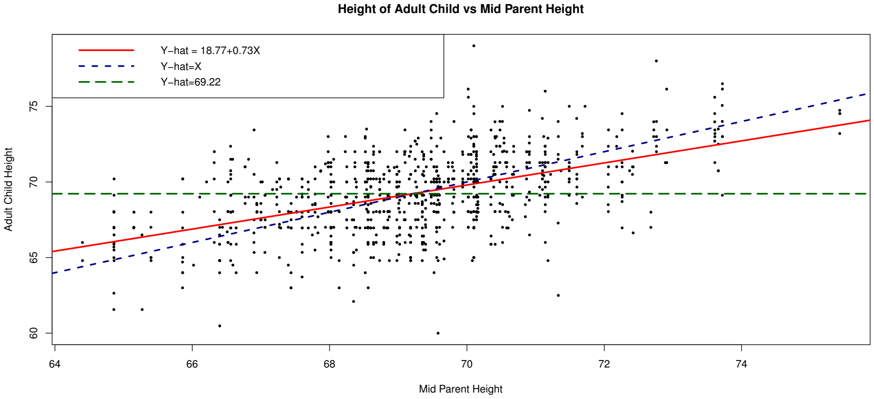

Figure 2.12: Adult child height (y), Mid-parent height (x), and three lines relating y to x

∇

We often are interested in relationships among more than two numeric variables. Scatterplot and correlation matrices can be constructed to demonstrate the bivariate association of all pairs of variables.

Example 2.9: Compressive Strength and Microfabric Properties of Amphibolites

A study (Ali, Guang, and Ibrahim (2014) [5]) reported the relationship between Uniaxial Compression Strentgh (UCS) and 8 predictor variables including: percent hornblende (hb), grain size (gs), and grain area (ga). A simple scatterplot matrix of plots of all pairs of these four variables is given in Figure 2.13. The correlation matrix is given along with R code below. Note that this can be extended to all pairs of variables, the plot just gets very difficult to focus on particular pairs of variables.

R Output

```
### Text Output

> cor(rs1[,c(2,6,7,8)])
           UCS         hb         gs         ga
UCS  1.0000000  0.6935996 -0.8535317 -0.8537215
hb   0.6935996  1.0000000 -0.7200409 -0.6641698
gs  -0.8535317 -0.7200409  1.0000000  0.9845240
ga  -0.8537215 -0.6641698  0.9845240  1.0000000
```

∇

When data are highly skewed, individual cases have the ability to have a large impact on the correlation coefficient. An alternative measure that is widely used is the Spearman Rank Correlation Coefficient (aka Spearman's rho). This coefficient is computed by ranking the x and y values from 1 (smallest) to n or N (largest), and applying the formula for Pearson's coefficient to the ranks. This way, extreme x or y values do not have as large of an impact on the coefficient. Also, in many situations, the natural measurements are the rankings or ordering themselves.

Example 2.10: NASCAR Start and Finish Positions 1975-2003

A study of NASCAR races for the years 1975-2003, considered the correlation between starting and finishing positions among drivers for the 898 races during those seasons (Winner (2006) [60]). As the data were orderings, it was natural to compute the correlation using Spearman's rank correlation. The summary of the correlations is given below, and a density plot and histogram are given in Figure 2.14.

R Output

```
### Output
> length(spearman)
```

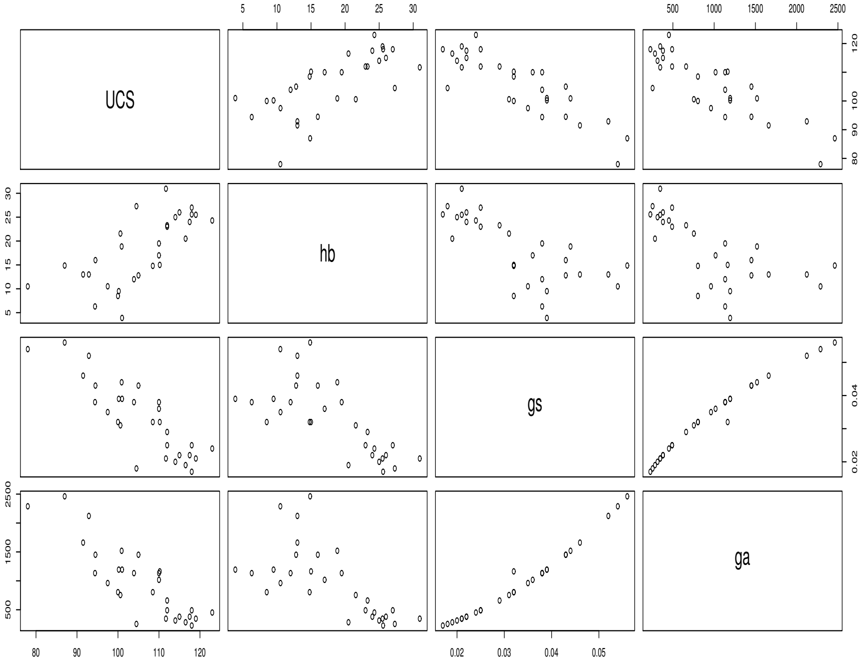

Figure 2.13: Bivariate Plots of Uniaxial Compression Strength (UCS), Percent Hornblende (hb), Grain Size (gs), and Grain Area (ga)

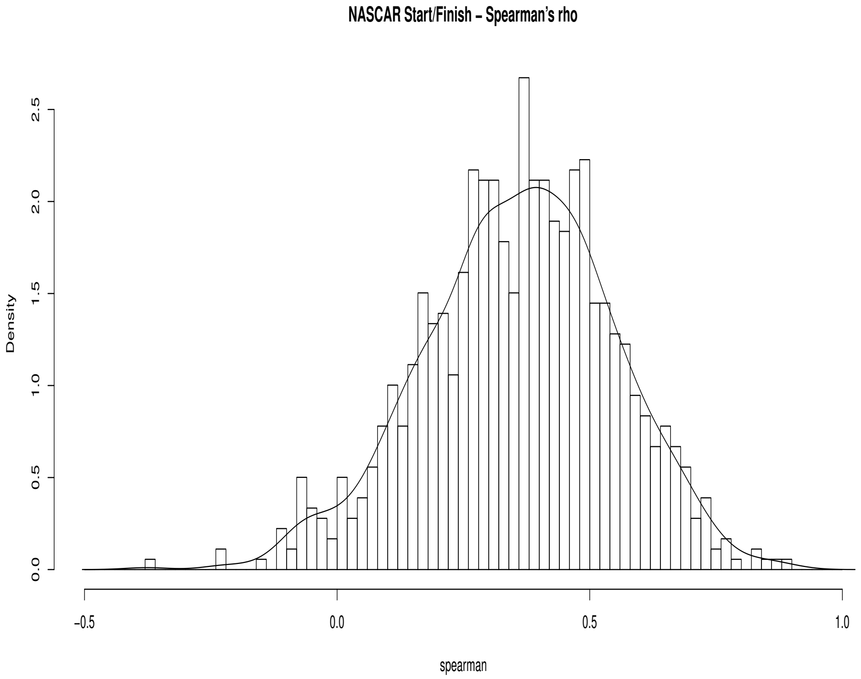

Figure 2.14: NASCAR Races 1975-2003 - Spearman's rank correlation coefficient for start/finish positions

```
[1] 898
> summary(spearman)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.3768  0.2399  0.3690  0.3590  0.4869  0.8977 
```

∇

Many series (particularly when measured over time) display **spurious correlations**, particularly when both variables tend to increase or decrease together with no **causal** reason that the two (or more) variables move in tandem. For instance, the correlation between annual U.S. internet users (per 100 people) and electrical power consumption (kWh per capita) for the years 1994-2010 is .7821 (data source: The World Bank). Presumably increasing internet usage isn't leading to large increases in electrical consumption, or vice versa.

Chapter 3

Probability

In this chapter, we describe the concepts of probability, random variables, probability distributions, and sampling distributions. There are three commonly used interpretations of probability: classical, relative frequency, and subjective. Probability is the basis of all methods of statistical inference covered in this course.

3.1 Terminology and Basic Probability Rules

The **classical** interpretation of probability involves listing (or using counting rules to quantify) all possible outcomes of a random process, often referred to as an "experiment." It is often (but not necessarily) assumed that each outcome is equally likely. If a coin is tossed once, it can land either "heads" or "tails," and unless there is reason to believe otherwise, we would assume the probability of each possible outcome is 1/2. If a dice is rolled, the possible numbers on the "up face" are $\{1,2,3,4,5,6\}$. Again, unless some external evidence leads us to believe otherwise, we would assume each side has a probability of landing as the "up face" is 1/6. When dealing a 5 card hand from a well shuffled 52 card deck, there are $\frac{52!}{5!(52-5)!} = 2,598,960$ possible hands. Clearly that would be impossible to enumerate, but with counting rules it is still fairly easy to assign probabilities to different types of hands.

An **event** is a pre-specified outcome of an experiment/random process. It can be made up of a single element or a group of elements of the sample space. If the sample space is made up of N elements and the event of interest constitutes N_E elements of the sample space, the probability of the event is $p_E = N_E/N$, when all elements are equally likely. If elements are not equally likely, p_E is the sum of the probabilities of the elements constituting the event (where the sum of all the N probabilities is 1).

The **relative frequency** interpretation of probability corresponds to how often an event of interest would occur if an experiment were conducted repeatedly. If an unbalanced dice were tossed a very large number of times, we could observe the fractions of times each number was the "up face." With modern computing power, simulations can be run to approximate probabilities of complex events, which could never be able to be obtained via a model of a sample space.

In cases where a sample space can not be enumerated or an experiment can not be repeated, individuals often resort to assessing **subjective** probabilities. For instance, in considering whether the price of a stock will increase over a specific time horizon, individuals may speculate on the probability based on any market information available at the time of the assessment. Different individuals may have different probabilities for the same event. Many studies have been conducted to assess people's abilities and heuristics used to assign probabilities to events (see e.g. Kahneman, Slovic, and Tversky (1982) [33]), for a large collection of research on the topic.

Three useful counting tools are the **multiplication rule**, **permutations** and **combinations**. The multiplication rule is useful when the experiment is made up of k stages, where stage i can end in one of m_i outcomes. Permutations are used when sampling k items from n items without replacement, and order matters. Combinations are similar to permutations with the exception that order does not matter. The total possible outcomes for each of these rules is given below.

$$\text{Multiplication Rule: } m_1 \times m_2 \times \cdots \times m_k = \prod_{i=1}^{k} m_i$$

$$\text{Permutations: } P_k^n = n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!} \qquad 0! \equiv 1$$

$$\text{Combinations: } C_k^n = \frac{n \times (n-1) \times \cdots \times (n-k+1)}{k \times (k-1) \cdots \times 1} = \frac{n!}{k!(n-k)!}$$

Note that there are $k!$ possible orderings of the k items selected from n items, which is why there are fewer combinations than permutations.

Example 3.1: Lotteries and Competitions

The Florida lottery has many "products" for consumers (flalottery.com). The Pick 4 game is conducted twice per day and pays out up to \$5000 per drawing. Participants choose 4 digits from 0-9 (digits can be repeated). Thus at each of $k = 4$ stages, there are $m = 10$ potential digits. Thus there are $10(10)(10)(10) =$ 10,000 possible sequences (order matters in payouts).

In a race among 10 "identical" mice of a given strain, there are $P_3^{10} = 10(9)(8) = 720$ possible orderings of 1st, 2nd, and 3rd place. In the 2017 Kentucky Derby, there were 22 horses in the race. Starting positions are taken by "pulling names out of a hat." Thus, there are $22! = 1.124 \times 10^{21}$ possible orderings of the horses to the starting positions. This is 10.4 billion times as many people who had ever lived on the earth as of 2011 according to the Population Reference Bureau (www.prb.com).

The Florida Lotto game, held every Wednesday and Saturday night, involves selecting 6 numbers without replacement from the integers 1,...,53; where order does not matter. There are $C_6^{53} = \frac{53!}{6!47!} = 22,957,480$ possible drawings.

3.1.1 Basic Probability

Let A and B be events of interest with corresponding probabilities $P(A)$ and $P(B)$, respectively. The **Union** of events A and B is the event that either A and/or B occurs and is denoted $A \cup B$. Events A and B are

	B	$\overline{B}$	Total
A	909	67	976
$\overline{A}$	2528	142	2670
Total	3437	209	3646

Table 3.1: Counts of UFO's by Shape Type and nation of sighting

mutually exclusive if they can not both occur as an experimental outcome. That is, if A occurs, B cannot occur, and vice versa. The **Complement** of event A, is the event that A does not occur and is denoted by $\overline{A}$ or sometimes A'. The **Intersection** of events A and B is the event that both A and B occur, and is denoted as $A \cap B$ or simply AB. In terms of probabilities, we have the following rules.

Union: $P(A \cup B) = P(A) + P(B) - P(AB)$ Mutually Exclusive: $P(AB) = 0$ Complement: $P\left(\overline{A}\right) = 1 - P(A)$

The probability of an event A or B, without any other information, is referred to as its **unconditional** or **marginal** probability. When information is known whether or not another event has (or has not) occurred it is referred to as its **conditional** probability. If the unconditional probability of A and its conditional probability given B has occurred are equal, then the events A and B are said to be **independent**. The rules for obtaining conditional probabilities (assuming $P(A) > 0$ and $P(B) > 0$) are given below, as well as probabilities under independence.

Prob. of A Given B: $P(A|B) = \dfrac{P(AB)}{P(B)}$ Prob. of B Given A: $P(B|A) = \dfrac{P(AB)}{P(A)}$

$$P(AB) = P(A)P(B|A) = P(B)P(A|B)$$

A and B independent: $P(A) = P(A|B) = P\left(A|\overline{B}\right)$ $P(B) = P(B|A) = P\left(B|\overline{A}\right)$ $P(AB) = P(A)P(B)$

Example 3.2: UFO Sightings

Based on 3646 UFO sightings on the UFO Research Database (www.uforesearchdb.com), we define A to be the event that a UFO is classified as being shaped as an orb/sphere or circular or a disk and event B that the sighting is in the USA. Table 3.1 gives a cross-tabulation of the counts for this "population."

$$P(A) = \frac{976}{3646} = .2677 \qquad P(B) = \frac{3437}{3646} = .9427 \qquad P(AB) = \frac{909}{3646} = .2493 \qquad P(A \cup B) = .2677 + .9427 - .2493 = .9611$$

$$P(A|B) = \frac{.2493}{.9427} = \frac{909}{2528} = .2645 \qquad P\left(A|\overline{B}\right) = \frac{67}{209} = .3206 \qquad P(B|A) = \frac{.2493}{.2677} = \frac{909}{976} = .9314$$

Note that the event that a UFO is classified as orb/sphere or circular or a disk is not independent of whether it was sighted in the USA. There is a higher probability for these types of shapes to be sighted outside the USA (.3206) than in the USA (.2645).

∇

Example 3.3: Women's and Men's Marathon Speeds

For the Rock and Roll marathon runner speeds, we can classify events as follow. Event F is that the runner is Female, event S_5 is the event that a runner's speed is less than or equal to 5 miles per hour, and S_7 is the event that the runner's speed is greater than or equal to 7 miles per hour. Counts of runners by gender and speed are given in Table 3.2. Note that the middle row represents the intersection of the compliments of events S_5 and S_7 and represents the runners with speeds between 5 and 7 miles per hour. We compute various probabilities below.

$$P(F) = \frac{1045}{2499} = .4182 \qquad P\left(\overline{F}\right) = 1-.4182 = \frac{1454}{2499} = .5818 \qquad P(S_5) = \frac{326}{2499} = .1305 \qquad P(S_7) = \frac{464}{2499} = .1857$$

$$P\left(\overline{S_5} \cap \overline{S_7}\right) = 1-.1305-.1857 = \frac{1709}{2499} = .6839 \qquad P(F \cap S_5) = \frac{172}{2499} = .0688 \qquad P\left(\overline{F} \cap S_5\right) = \frac{154}{2499} = .0616$$

$$P(F \cap S_7) = \frac{106}{2499} = .0424 \qquad P\left(\overline{F} \cap S_7\right) = \frac{358}{2499} = .1433 \qquad P\left(F \cap \overline{S_5} \cap \overline{S_7}\right) = \frac{767}{2499} = .3069$$

$$P\left(\overline{F} \cap \overline{S_5} \cap \overline{S_7}\right) = \frac{942}{2499} = .3770 \qquad P(S_5|F) = \frac{.0688}{.4182} = \frac{172}{1045} = .1646 \qquad P(S_7|F) = \frac{.0424}{.4182} = \frac{106}{1045} = .1014$$

$$\left(\overline{S_5} \cap \overline{S_7}|F\right) = \frac{.3069}{.4182} = \frac{767}{1045} = .7340$$

∇

3.1.2 Bayes' Rule

Bayes' rule is used in a wide range of areas to update probabilities (and probability distributions) in light of new information (data). In the case of updating probabilities of particular events, we start with a set

	F	$\overline{F}$	Total
S_5	172	154	326
$\overline{S_5} \cap \overline{S_7}$	767	942	1709
S_7	106	358	464
Total	1045	1454	2499

Table 3.2: Counts of Speeds (mph) by Gender - 2015 Rock and Roll Marathon

of events $A_1, \ldots, A_k$ that represent a **partition** of the sample space. That means that each element in the sample space must fall in exactly one A_i. In probability terms this means the following statements hold.

$$i \neq j: \quad P(A_i \cap A_j) = 0 \qquad\qquad P(A_1) + \cdots + P(A_k) = 1$$

The probability $P(A_i)$ is referred to as the **prior probability** of the i^{th} portion of the partition, and in some contexts are referred to as **base rates**. Let C be an event, such that $0 < P(C) < 1$, with known conditional probabilities $P(C|A_i)$. This leads to being able to "update" the probability that A_i occurred, given knowledge that C has occurred, the **posterior probability** of the i^{th} portion of the partition. This is simply (in this context) an application of conditional probability making use of formulas given above and the fact that there is a partition of the sample space.

$$P(A_i \cap C) = P(A_i)P(C|A_i) \qquad\qquad P(C) = \sum_{i=1}^{k} P(A_i \cap C) = \sum_{i=1}^{k} P(A_i)P(C|A_i)$$

$$\Rightarrow \qquad P(A_i|C) = \frac{P(A_i \cap C)}{P(C)} = \frac{P(A_i)P(C|A_i)}{\sum_{i=1}^{k} P(A_i)P(C|A_i)} \quad i = 1, ..., k$$

Example 3.4: Women's and Men's Marathon Speeds

Treating the three speed ranges ($A_1 \equiv \leq 5, \quad A_2 \equiv 5-7, \quad A_3 \equiv \geq 7$) as a partition of the sample space, we can update the probabilities of the runner's speed range, given knowledge of gender. The prior probabilities are $P(A_1) = 326/2499 = .1305$, $P(A_2) = 1709/2499 = .6839$, and $P(A_3) = 464/2499 = .1857$. The relevant probabilities are given below to obtain the posterior probabilities of the speed ranges, given the runner's gender.

$$P(A_1) = \frac{326}{2499} = .1305 \qquad P(F|A_1) = \frac{172}{326} = .5276 \qquad P(A_1 \cap F) = P(A_1)P(F|A_1) = \left(\frac{326}{2499}\right)\left(\frac{172}{326}\right) = .0688$$

$$P(A_2) = \frac{1709}{2499} = .6839 \qquad P(F|A_2) = \frac{767}{1709} = .4488 \qquad P(A_2 \cap F) = P(A_2)P(F|A_2) = \left(\frac{1709}{2499}\right)\left(\frac{767}{1709}\right) = .3069$$

$$P(A_3) = \frac{464}{2499} = .1857 \qquad P(F|A_3) = \frac{106}{464} = .2284 \qquad P(A_3 \cap F) = P(A_3)P(F|A_3) = \left(\frac{464}{2499}\right)\left(\frac{106}{464}\right) = .0424$$

$$P(F) = \sum_{i=1}^{3} P(A_i \cap F) = .0688 + .3069 + .0424 = .4182 \qquad P(A_1|F) = \frac{.0688}{.4182} = .1646$$

$$P(A_2|F) = \frac{.3069}{.4182} = .7340 \qquad P(A_3|F) = \frac{.0424}{.4182} = .1014$$

Note that these can be computed very easily from the counts in Table 3.2 by taking the cell counts over the column totals, as can be seen for the males.

$$P(M) = \frac{1454}{2499} = .5818 \qquad P(A_1|M) = \frac{154}{1454} = .1059 \qquad P(A_2|M) = \frac{942}{1454} = .6479 \qquad P(A_3|M) = \frac{358}{1454} = .2462$$

∇

Example 3.5: Drug Testing Accuracy

As a second example based on assessed probabilities, Barnum and Gleason (1964), [6], considered drug tests among workers. They had four sources of prevalence of recreational drug users based on published data sources (2.4% (.024), 3.1% (.031), 8.2% (.082), and 20.2% (.202)). Further, based on studies of test accuracy at the time, they had the probability that a drug user (correctly) tests positive is 0.80, and the probability a non-drug user (incorrectly) tests positive is 0.02. Let D be the event that a worker is a drug user, and T^+ be the event that a worker tests positive for drug use.

Consider the case where $P(D) = .024$. We are interested in the probability a worker who tests positive is a drug user. Note that we do not have this probability stated above. The relevant probabilities and calculations are given below.

$$P(D) = .024 \qquad P\left(\overline{D}\right) = 1 - .024 = .976 \qquad P\left(T^+|D\right) = .80 \qquad P\left(T^+|\overline{D}\right) = .02$$

$$P\left(D \cap T^+\right) = .024(.80) = .01920 \qquad P\left(\overline{D} \cap T^+\right) = .976(.02) = .01952 \qquad P\left(T^+\right) = .01920 + .01952 = .03872$$

$$P\left(D|T^+\right) = \frac{.01920}{.03872} = .4959 \qquad P\left(\overline{D}|T^+\right) = \frac{.01952}{.03872} = .5041$$

Thus a positive result on the test implies slightly less than a 50:50 chance the worker uses drugs. As the prevalence increases, this probability increases, see Table 3.3.

∇

$P(D)$	$P\left(D \cap T^{+}\right)$	$P\left(\overline{D} \cap T^{+}\right)$	$P\left(T^{+}\right)$	$P\left(D\|T^{+}\right)$
.024	.01920	.01952	.03872	.4959
.031	.02480	.01938	.04418	.5613
.082	.06560	.01836	.08396	.7813
.202	.16160	.01596	.17756	.9101

Table 3.3: Probability a Positive Drug test corresponds to a drug user as a function of Prevalence of Drug Use

3.2 Random Variables and Probability Distributions

When an experiment is conducted, or an observation is made, the outcome will not be known in advance, and is considered to be a **random variable**. Random variables can be qualitative or quantitative. Qualitative variables are generally modeled as a list of outcomes and their corresponding counts, as in contingency tables and cross-tabulations. Quantitative random variables are numeric outcomes and are classified as being either discrete or continuous, as described previously in describing data.

A **probability distribution** gives the values a random variable can take on and their corresponding probabilities (discrete case) or density (continuous case). Probability distributions can be given in tabular, graphic, or formulaic form. Some commonly used families of distributions are described below.

3.3 Discrete Random Variables

Discrete random variables can take on a finite, or countably infinite, set of outcomes. We label the random variable as Y, and its specific outcomes as $y_1, y_2, \ldots, y_k$. Note that in some cases there is no upper limit for k. We denote the probabilities of the outcomes as $P(Y = y_i) = p(y_i)$, with the following restrictions.

$$0 \leq p(y_i) \leq 1 \qquad \sum_{i=1}^{k} p(y_i) = 1 \qquad F(y_t) = P(Y \leq y_t) = \sum_{i=1}^{t} p(y_i) \quad t = 1, \ldots, k$$

Here $F(y)$ is called the **cumulative distribution function (cdf)**. This is a monotonic "step" function for discrete random variables, and ranges from 0 to 1.

Example 3.6: NASCAR Race Finish Positions - 1975-2003

For the NASCAR race data in Winner (2006) [60], each driver was classified by their starting position and their finishing position in the 898 races (34884 driver/races). For each race, we identify the number of racers who start in the top 10, that finish in the top 3. This random variable (Y) can take on the values $y =$ 0, 1, 2, or 3. That is, none of the people who start toward the front (top 10) finish in the top 3, or one, or two, or three. Table 3.4 gives the counts, probabilities, cumulative probabilities, and calculations used later to numerically describe the empirical population distribution. The probability of either 2 or 3 drivers who started in the top 10 finish in the top 3, is over 3/4 (.3987+.3708=.7695).

y	# races	$p(y)$	$F(y)$	$yp(y)$	$y^2p(y)$
0	37	.0412	.0412	0.0000	0.0000
1	170	.1893	.2305	0.1893	0.1893
2	358	.3987	.6292	0.7974	1.5948
3	333	.3708	1.0000	1.1124	3.3372
Total	898	1		2.0991	5.1213

Table 3.4: Probability Distribution for Number of Top 10 Starters finishing in Top 3 positions, NASCAR races 1975-2003

R Output

```
## Output
> (t.strt10Fin3 <- table(strt10Fin3))  ### Count 0,1,2,3 Top 3 finishers
strt10Fin3
  0   1   2   3 
 37 170 358 333 
> t.strt10Fin3 / sum(t.strt10Fin3)     ### Turn counts to proportions
strt10Fin3
         0          1          2          3 
0.04120267 0.18930958 0.39866370 0.37082405 
```

∇

Population Numerical Descriptive Measures

Three widely used numerical descriptive measures corresponding to a population are the **population mean**, μ, the **population variance**, σ^2, and the **population standard deviation**, σ. While we have previously covered these based on a population of measurements, we now base them on a probability distribution. Their formulas are given below.

$$\text{Mean: } E\{Y\} = \mu_Y = y_1p(y_1) + \cdots + y_kp(y_k) = \sum_y yp(y)$$

$$\text{Variance: } V\{Y\} = E\{(Y-\mu_Y)^2\} = \sigma_Y^2 = (y_1-\mu_Y)^2p(y_1) + \cdots + (y_k-\mu_Y)^2p(y_k) = \sum_y (y-\mu_Y)^2p(y) =$$

$$= \sum_y y^2p(y) - \mu_Y^2 \qquad \text{Standard Deviation: } \sigma_Y = +\sqrt{\sigma_Y^2}$$

Some useful rules among **linear** functions of random variables are given here. Suppose Y is a random variable with mean and variance μ_Y and σ_Y^2, respectively. Further, suppose that a and b are constants (not random). Then we have the following results.

$$E\{a+bY\} = \sum_y (a+by)p(y) = a\sum_y p(y) + b\sum_y yp(y) = a(1) + b\mu_Y = a + b\mu_Y$$

$$V\{a+bY\} = \sum_y ((a+by)-(a+b\mu_Y))^2 p(y) = b^2 \sum_y (y-\mu_Y)^2 p(y) = b^2\sigma_Y^2 \qquad \sigma_{a+bY} = |b|\sigma_Y$$

Examples where these can be applied involve transforming from inches to centimeters (1 inch = 2.54 cm, 1 cm = 1/2.54=0.3937 inch), from pounds to kilograms (1 kilogram = 2.204623 pounds) and from degrees Fahrenheit to Celsius ($\deg F = 32 + 1.8 \deg C$). These rules do not work for values raised to powers, exponentials, or logarithms, although some approximations exist.

Example 3.7: NHL Hockey Player BMI and Marathon Speeds

Previously, we obtained the population mean and variance for NHL player body mass indices. Now we obtain the mean, variance, and standard deviation of their weights (pounds) and heights (inches), and convert them to kilograms and centimeters, respectively. The mean weight is 202.42 pounds, and the variance is 228.60 pounds^2. To convert from pounds to kilos, we have to divide pounds by 2.2, that is $K = (1/2.204623)P = 0.453592P$. Thus, we obtain the following quantities.

$$\mu_K = 0.453592\mu_P = 0.453592(202.42) = 91.92 \qquad \sigma_K^2 = (0.453592)^2\sigma_P^2 = (0.453592)^2(228.60) = 47.03$$
$$\sigma_K = \sqrt{47.03} = 6.86$$

The population mean and variance of heights are 73.26 inches and 4.26 inches^2, respectively. To convert inches to centimeters, we have to multiply by 2.54, that is $C = 2.54I$. Thus, we obtain the following quantities.

$$\mu_C = 2.54\mu_I = 2.54(73.26) = 186.08 \qquad \sigma_C^2 = (2.54)^2\sigma_I^2 = (2.54)^2(4.26) = 27.48 \qquad \sigma_C = \sqrt{27.48} = 5.24$$

Note that in the metric system, the weights in kilograms are less variable than weights in pounds, while the heights in centimeters are more variable than than heights in inches.

For the female marathon runners, the mean and variance of their speeds were 5.84 mph and 0.69 mph^2, respectively. One mile represents 1.60394 kilometers, so that so that a person who runs M miles in 1 hour, runs $K = 1.60394M$ kilometers in one hour. This leads to the following quantities.

$$\mu_K = 1.60394(5.84) = 9.37 \qquad \sigma_K^2 = (1.60394)^2(0.69) = 1.78 \qquad \sigma_K = \sqrt{1.78} = 1.33$$

∇

In many settings, we are interested in linear functions of a sequence of random variables: $Y_1, \ldots, Y_n$. Typically, we have fixed coefficients $a_1, \ldots, a_n$, and $E\{Y_i\} = \mu_i$, $V\{Y_i\} = \sigma_i^2$, and $\mathrm{COV}\{Y_i, Y_j\} = \sigma_{ij}$.

$$\mathrm{COV}\{Y_i, Y_j\} = E\{(Y_i - \mu_i)(Y_j - \mu_j)\} = \sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$$

$$W = \sum_{i=1}^{n} a_i Y_i \qquad E\{W\} = \mu_W = \sum_{i=1}^{n} a_i \mu_i \qquad V\{W\} = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j \sigma_{ij}$$

If, as in many, but by no means all, cases, the Y_i values are independent ($\sigma_{ij} = 0$), the variance simplifies to $V\{W\} = \sum_{i=1}^{n} a_i^2 \sigma_i^2$. A special case is when we have two random variables: X and Y, and a linear function $W = aX + bY$ for fixed constants. We have means μ_X, μ_Y, standard deviations σ_X, σ_Y, covariance σ_{XY}, and correlation ρ_{XY}.

$$W = aX + bY \qquad E\{W\} = a\mu_X + b\mu_Y \qquad V\{W\} = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY} = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho_{XY}\sigma_X\sigma_Y$$

Some special cases include where we have: $a = 1, b = 1$ (sums), and $a = 1, b = -1$ (differences). This leads to the following results.

$$E\{X+Y\} = \mu_X + \mu_Y \qquad V\{X+Y\} = \sigma_X^2 + \sigma_Y^2 + 2\rho_{XY}\sigma_X\sigma_Y$$

$$E\{X-Y\} = \mu_X - \mu_Y \qquad V\{X-Y\} = \sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y$$

Example 3.8: Movie “Close Up” Scenes

Barry Salt has classified film shots along an ordinal scale for a “population” of 398 movies. The levels are (BCU=Big Close Up, CU=Close Up, MCU=Medium Close Up, MLS=Medium Long Shot, LS=Long Shot, and VLS=Very Long Shot). We consider X to be the number of Big Close Up’s and Y to be the number of Close Up’s in a film. For this population, $\mu_X = 28.84$, $\mu_Y = 79.23$, $\sigma_X = 31.48$, $\sigma_Y = 61.37$, and $\rho_{XY} = 0.51$. We obtain the population mean, variance, and standard deviations of the sum of Big Close Up’s and Close Up’s $(X + Y)$ and the difference between Big Close Up’s and Close Up’s $(X - Y)$.

$$E\{X+Y\} = 28.84+79.23 = 108.07 \quad V\{X+Y\} = 31.48^2+61.37^2+2(0.51)(31.48)(61.37) = 6727.83 \quad \sigma_{X+Y} = 82.02$$

$$E\{X-Y\} = 28.84-79.23 = -50.39 \quad V\{X-Y\} = 31.48^2+61.37^2-2(0.51)(31.48)(61.37) = 2786.70 \quad \sigma_{X-Y} = 52.79$$

Source: `http://www.cinemetrics.lv/salt.php`

∇

3.3.1 Common Families of Discrete Probability Distributions

Here we consider some commonly used families of discrete probability distributions, namely the Binomial, Poisson, and Negative Binomial families. These are used in many situations where data are counts of numbers of events occurring in an experiment.

Binomial Distribution

A binomial "experiment" is based on a series of Bernoulli trials with the following characteristics.

- The experiment consists of n trials or observations.
- Trial outcomes are independent of one another.
- Each trial can end in one of two possible outcomes, often labeled **S**uccess or **F**ailure.
- The probability of Success, π is constant across all trials.
- The random variable, Y, is the number of Successes in the n trials

Note that many experiments are well approximated by this model, and thus it has wide applicability. One problem that has been considered in great detail is the assumption of independence from trial to trial. A classic paper that looked at the "hot hand" in basketball shooting has led to many studies in sports involving the topic is Gilovich, Vallone, and Tversky (1985), [26].

The probability of any sequence of y Successes and $n-y$ Failures is $\pi^y(1-\pi)^{n-y}$ for $y=0,1,\ldots,n$. The number of ways to observe y successes in n trials makes use of combinations described previously. The number of ways of choosing y positions from $1,2,\ldots,n$ is $C_y^n = \frac{n!}{y!(n-y)!} = \binom{n}{y}$. For instance, there is only one way observing either 0 or n Successes, there are n ways of observing 1 or $n-1$ Successes, and so on. This leads to the following probability distribution for $Y \sim Bin(n,\pi)$.

$$P(Y=y) = p(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \quad y=0,1,\ldots,n \qquad \sum_{y=0}^{n} p(y) = (\pi + (1-\pi))^n = 1^n = 1$$

Statistical packages and spreadsheets have functions for computing probabilities for the Binomial (and all distributions covered in these notes). In R, the function **dbinom(*y*,*n*,π)** returns $P(Y=y)=p(y)$ (the probability "density") when $Y \sim Bin(n,\pi)$.

To obtain the mean and variance of the Binomial distribution, consider the n independent trials individually (these are referred to as **Bernoulli** trials). Let $S_i = 1$ if trial i is a success, and $S_i = 0$ if it is a failure. Then Y, the number of Successes is the sum of the independent S_i values, leading to the following results.

$$E\{S_i\} = 1\pi + 0(1-\pi) = \pi \qquad E\{S_i^2\} = 1^2\pi + 0^2(1-\pi) = \pi \qquad V\{S_i\} = E\{S_i^2\} - (E\{S_i\})^2 = \pi - \pi^2 = \pi(1-\pi)$$

$$Y = \sum_{i=1}^{n} S_i \quad \Rightarrow \quad E\{Y\} = \mu_Y = \sum_{i=1}^{n} E\{S_i\} = n\pi \qquad V\{Y\} = \sigma_Y^2 = \sum_{i=1}^{n} V\{S_i\} = n\pi(1-\pi) \qquad \sigma_Y = \sqrt{n\pi(1-\pi)}$$

Example 3.9: Experiments of Mobile Phone Telepathy

A set of experiments was conducted to determine whether people displayed evidence of telepathy in receiving mobile phone calls (Sheldrake, Smart, and Avraamides (2015), [53]). Each subject received 6 calls from one of two potential callers. Each subject predicted which caller was calling. Assuming random guessing, the number of successful predictions should be Binomial, with $n = 6$ trials, and probability of Success $\pi = 0.5$, since there were two potential callers. The probabilities of 0,1,2,...,6 successes for a subject in the experiment are given below. A plot of the probability distribution is given in Figure 3.1.

$$\frac{6!}{0!(6-0)!} = \frac{6!}{6!(6-6)!} = 1 \quad \frac{6!}{1!(6-1)!} = \frac{6!}{5!(6-5)!} = 6 \quad \frac{6!}{2!(6-2)!} = \frac{6!}{4!(6-4)!} = 15 \quad \frac{6!}{3!(6-3)!} = 20$$

$$.5^y(1-.5)^{6-y} = .5^6 = .015625$$

$$p(0) = p(6) = .015625 \quad p(1) = p(5) = .09375 \quad p(2) = p(4) = .234375 \quad p(3) = .3125$$

R Output

```
### Output
> (p_y <- dbinom(y, 6, 0.5))  ## Obtain p(y) for y=0,1,...,6
[1] 0.015625 0.093750 0.234375 0.312500 0.234375 0.093750 0.015625
```

The mean, variance, and standard deviation of the number of Successful predictions in the $n = 6$ trials under this model are as follow.

$$\mu_Y = n\pi = 6(0.5) = 3 \qquad \sigma_Y^2 = n\pi(1-\pi) = 6(0.5)(1-0.5) = 1.5 \qquad \sigma_Y = \sqrt{1.5} = 1.2247$$

For the Sheldrake, et al study, [53], 110 subjects completed 6 trials each (660 total trials). There were a total of 369 hits (there appears to be a typo saying 370 in their Table 3). This corresponds to a proportion of 369/660=.559, in other words, these subjects in aggregate showed better than expected success in predicting callers. Table 3.5 gives the probability distributions for $\pi = 0.50$ and $\pi = 0.56$, along with expected counts under the two models and the observed counts ($N = 110$ subjects).

∇

y	$\pi = 0.50 : p(y)$	$\pi = 0.56 : p(y)$	$\pi = 0.50$: Expected #	$\pi = 0.56$: Expected #	Observed #
0	.015625	.007256	1.72	0.80	1
1	.093750	.055412	10.31	6.10	5
2	.234375	.176310	25.78	19.39	18
3	.312500	.299193	34.38	32.91	37
4	.234375	.285594	25.78	31.42	31
5	.093750	.145393	10.31	15.99	15
6	.015625	.030841	1.72	3.39	3
Total	1	1	110	110	110

Table 3.5: Probability Distribution for Number of successful prediction for mobile telephone telepathy study

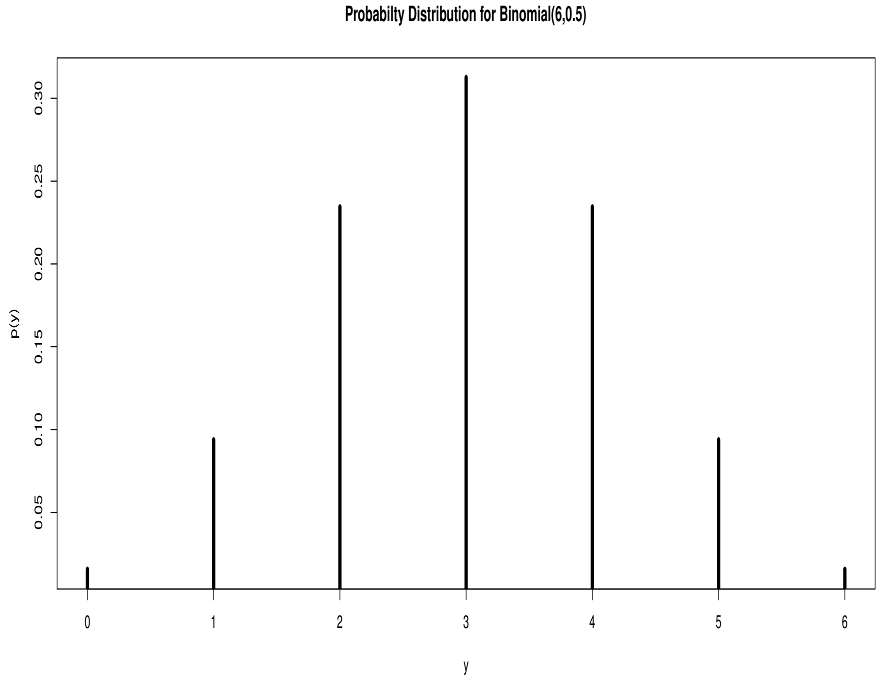

Figure 3.1: Probability Distribution for Mobile Telephone Telepathy experiment assuming random guessing, $Y \sim$Bin(6,0.5)

y	$p(y)$	Expected #	Observed #
0	.3936	226.71	229
1	.3670	211.39	211
2	.1711	98.55	93
3	.0532	30.64	35
4	.0124	7.14	7
≥ 5	.0027	1.56	1
Total	1	576	576

Table 3.6: Probability Distribution for Number of bombs hitting within 576 areas on a grid in the south of London during World War II

Poisson Distribution

In many applications, researchers observe the counts of a random process in some fixed amount of time or space. The random variable Y is a count that can take on any non-negative integer. One important aspect of the Poisson family is that the mean and variance are the same. This is one aspect that does not work for all applications. We use the notation: $Y \sim Poi\,(\lambda)$. The probability distribution, mean and variance of Y are:

$$p(y) = \frac{e^{-\lambda}\lambda^y}{y!} \quad y = 0, 1, \ldots; \quad \lambda > 0 \qquad E\{Y\} = \mu_Y = \lambda \qquad V\{Y\} = \sigma_Y^2 = \lambda$$

Note that $\lambda > 0$. The Poisson arises by dividing the time/space into n "infinitely" small areas, each having either 0 or 1 Success, with Success probability $\pi = \lambda/n$. Then Y is the number of areas having a success.

$$p(y) = \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} = \frac{n(n-1)\cdots(n-y+1)}{y!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} =$$

$$= \frac{1}{y!}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-y+1}{n}\right)\lambda^y\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-y}$$

The limit as n goes to ∞ is:

$$\lim_{n\to\infty} p(y) = \frac{1}{y!}(1)(1)\cdots(1)\lambda^y e^{-\lambda}(1) = p(y) = \frac{e^{-\lambda}\lambda^y}{y!} \quad y = 0, 1, 2...$$

The mean and variance for the Poisson distribution are both λ. This restriction can be problematic in many applications, and the Negative Binomial distribution (described below) is often used when the variance exceeds the mean.

Example 3.10: London Bomb Hits in World War II

A widely reported application of the Poisson Distribution involves the counts of the number of bombs hitting among 576 areas of $0.5km^2$ in south London during WWII (Clarke (1946), [15], also reported in Feller (1950), [24]). There were a total of 537 bombs hit with a mean of 537/576 = .9323. Table 3.6 gives the counts, and their expected counts ($576p(y)$) for the occurrences of 0 bombs, 1 bomb, ..., ≥ 5 bombs (the last cell involves 1 area which was hit 7 times).

Negative Binomial Distribution

The negative binomial distribution is used in two quite different contexts. The first is where a binomial type experiment is being conducted, except instead of having a fixed number of trials, the experiment is completed when the r^{th} success occurs. The random variable Y is the number of trials needed until the r^{th} success, and can take on any integer value greater than or equal to r. The probability distribution, its mean and variance are given below.

$$p(y) = \binom{y-1}{r-1}\pi^r \left(1-\pi\right)^{y-r} \qquad E\left\{Y\right\} = \mu_Y = \frac{r}{\pi} \qquad V\left\{Y\right\} = \sigma_Y^2 = \frac{r\left(1-\pi\right)}{\pi^2}.$$

A second use of the negative binomial distribution is as a model for count data. It arises from a mixture of Poisson models. In this setting it has 2 parameters and is more flexible than the Poisson (which has the variance equal to the mean), and can take on any non-negative integer value. In this form, the negative binomial distribution and its mean and variance can be written as follow (see e.g. Agresti (2002) [1] and Cameron and Trivedi (2005) [12]).

$$f\left(y;\mu,\alpha\right) = \frac{\Gamma\left(\alpha^{-1}+y\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y} \qquad \Gamma(w) = \int_0^\infty x^{w-1}e^{-x}dx = \left(w-1\right)\Gamma\left(w-1\right).$$

$$E\left\{Y\right\} = \mu \qquad V\left\{Y\right\} = \mu\left(1+\alpha\mu\right).$$

Example 3.11: Number of Comets Observed per Year - 1789-1888

The number of comets observed per year for the century 1789-1888 inclusive were reported by Chambers (1889), [13] and included in a large number of datasets by Thorndike (1926), [58]. The annual number of comets ranged from 0 (19 years) to 9 (1 year), with frequency counts and computations for the mean and variance given in Table 3.7, treating this as a population of years. The mean and variance are given below, along with "method of moments" estimates for μ and α for the Negative Binomial distribution.

$$\mu_Y = \sum_y yp(y) = 2.58 \qquad \sigma_Y^2 = \sum_y y^2 p(y) - \mu_Y^2 = 11.36 - 2.58^2 = 4.70$$

$$\sigma^2 = \mu(1+\alpha\mu) \quad \Rightarrow \quad \alpha = \frac{\sigma^2/\mu - 1}{\mu} = \frac{4.70/2.58-1}{2.58} = 0.32$$

The Negative Binomial appears to fit better than a Poisson distribution with mean 2.58, based on observed and expected counts.

∇

y	# years	$p(y)$	$yp(y)$	$y^2p(y)$	Exp(Poi)	Exp(NegBin)
0	19	.19	0.00	0.00	7.58	15.22
1	19	.19	0.19	0.19	19.55	21.54
2	17	.17	0.34	0.68	25.22	20.11
3	14	.14	0.42	1.26	21.69	15.54
4	13	.13	0.52	2.04	13.99	10.76
5	8	.08	0.40	2.00	7.22	6.93
6	4	.04	0.24	1.44	3.10	4.24
7	2	.02	0.14	0.98	1.14	2.50
8	3	.03	0.24	1.92	0.37	1.43
≥ 9	1	.01	0.09	0.81	0.14	1.73
Total	100	1	2.58	11.36	100	100

Table 3.7: Probability Distribution for Number of Comets Observed for years 1789-1888

3.4 Continuous Random Variables

Continuous random variables can take on any values along a continuum. Their distributions are described as densities, with probabilities being assigned as areas under the curve. Unlike discrete random variables, individual points have no probability assigned to them. While discrete probabilities and means and variances make use of summation, continuous probabilities and means and variances are obtained by integration. The following rules and results are used for continuous random variables and probability distributions. We use $f(y)$ to denote a probability density function and $F(y)$ to dentote the cumulative distribution function.

$$f(y) \geq 0 \qquad \int_{-\infty}^{\infty} f(y)dy = 1 \qquad P(a \leq Y \leq b) = \int_a^b f(y)dy \qquad F(y) = \int_{-\infty}^{y} f(t)dt$$

$$E\{Y\} = \mu_Y = \int_{-\infty}^{\infty} yf(y)dy \qquad V\{Y\} = \sigma_Y^2 = \int_{-\infty}^{\infty} (y-\mu_Y)^2 f(y)dy = \int_{-\infty}^{\infty} y^2 f(y)dy - \mu_Y^2 \qquad \sigma_Y = +\sqrt{\sigma_Y^2}$$

3.4.1 Common Families of Continuous Probability Distributions

Three commonly applied families of distributions for describing populations of continuous measurements are the **normal**, **gamma**, and **beta** families, although there are many other families also used in practice.

The normal distribution is symmetric and mound-shaped. It has two parameters: a mean and variance (the standard deviation is often used in software packages). Many variables have distributions that are modeled well by the normal distribution, and many estimators have **sampling distributions** that are approximately normal. The gamma distribution has a density over positive values that is skewed to the right. There are many applications where data are skewed with a few extreme observations, such as the marathon running times observed previously. The gamma distribution also has two parameters associated with it. The beta distribution is often used to model data that are proportions (or can be extended to any finite length interval). The beta distribution also has two parameters. All of these families can take on a wide range of shapes by changing parameter values.

Probabilities, quantiles, densities, and random number generators for specific distributions and parameter values can be obtained from many statistical software packages and spreadsheets such as EXCEL. We will use R throughout these notes.

Normal Distribution

The normal distributions, also known as the Gaussian distributions, are a family of symmetric mound-shaped distributions. The distribution has 2 parameters: the mean μ and the variance σ^2, although often it is indexed by its standard deviation σ. We use the notation $Y \sim N(\mu, \sigma)$. The probability density function, the mean and variance are:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad -\infty < y < \infty, -\infty < \mu < \infty, \sigma > 0 \quad E\{Y\} = \mu_Y = \mu \quad V\{Y\} = \sigma_Y^2 = \sigma^2$$

The mean μ defines the center (median and mode) of the distribution, and the standard deviation σ is a measure of the spread ($\mu - \sigma$ and $\mu + \sigma$ are the inflection points). Despite the differences in location and spread of the different distributions in the normal family, probabilities with respect to standard deviations from the mean are the same for all normal distributions. For $-\infty < z_1 < z_2 < \infty$, we have:

$$P(\mu + z_1\sigma \leq Y \leq \mu + z_2\sigma) = \int_{\mu+z_1\sigma}^{\mu+z_2\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy = \int_{z1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi(z_2) - \Phi(z_1).$$

Here Z is **standard normal**, a normal distribution with mean 0, and variance (standard deviation) 1. $\Phi(z^*)$ is the cumulative distribution function of the standard normal distribution, up to the point z^*:

$$\Phi(z^*) = \int_{-\infty}^{z^*} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

These probabilities and critical values can be obtained directly or indirectly from standard tables, statistical software, or spreadsheets. Note that:

$$Y \sim N(\mu, \sigma) \qquad \Rightarrow \qquad Z = \frac{Y - \mu}{\sigma} \sim N(0,1).$$

This makes it possible to use the standard normal table to obtain probabilities and quantiles for any normal distribution. Plots of three normal distributions are given in Figure 3.2.

Approximately 68% (.6826) of the probability lies within 1 standard deviation from the mean, 95% (.9544) lies within 2 standard deviations, and virtually all (.9970) lies within 3 standard deviations.

Example 3.12: NHL Player Body Mass Indices

Previously, we saw that the Body Mass Indices (BMI) of National Hockey League players for the 2013-2014 season were mound shaped with a mean of 26.50 and standard deviation 1.45. Figure 3.3 gives a histogram along with the corresponding normal density. There is a tendency to observe more actual BMI's in the center than the normal distribution would imply, but the normal model seems to be reasonable.

Consider the following quantiles (.10, .25, .50, .75, .90) for the NHL data and the corresponding N(26.50, 1.45) distribution. Also consider the probabilities of the following ranges ($< 26.50 - 2(1.45) = 23.60, >$

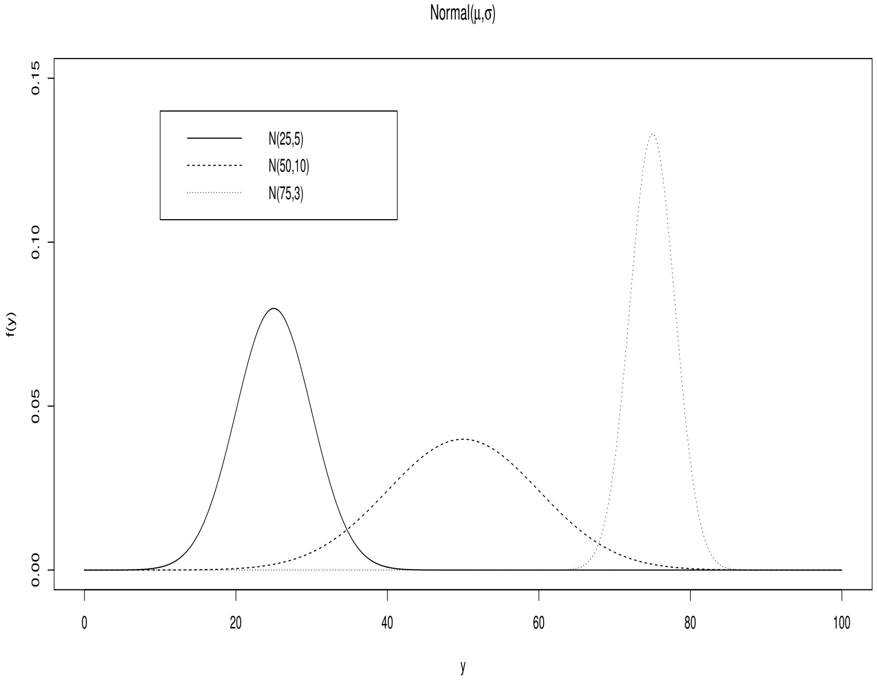

Figure 3.2: Three Normal Densities

$26.50 + 2(1.45) = 29.40$, and $(25.05 = 26.50 - 1.45, 26.50 + 1.45 = 27.95))$ for the NHL data and the normal distribution.

R Output

```
### Output
> round(q.out, 3)
              10%   25%     50    75%    90%
Theoretical 24.637 25.52 26.500 27.481 28.363
Empirical   24.702 25.62 26.516 27.439 28.342
>
> round(p.out, 4)
           <mu-2sigma (mu-sigma,mu+sigma) >mu+2sigma
Theoretical     0.0228             0.6827     0.0228
Empirical       0.0265             0.7057     0.0279
```

The quantiles and probabilities are very similar, showing the normal model is a reasonable approximation to the distribution of NHL BMI values.

∇

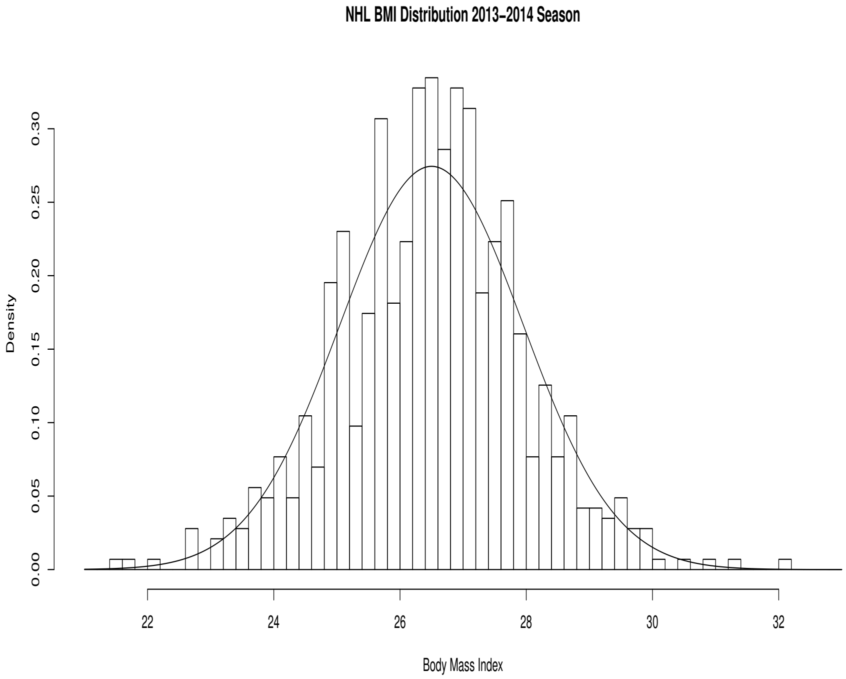

Figure 3.3: NHL Body Mass Indices and Normal Distribution

Gamma Distribution

The gamma family of distributions are used to model non-negative random variables that are often right-skewed. There are two widely used parameterizations. The first given here is in terms of *shape* and *scale* parameters.

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-y/\beta} \quad y \geq 0, \alpha > 0, \beta > 0 \qquad E\{Y\} = \mu_Y = \alpha\beta \qquad V\{Y\} = \sigma_Y^2 = \alpha\beta^2$$

Here, $\Gamma(\alpha)$ is the gamma function $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}dy$ and is built-in to virtually all statistical packages and spreadsheets. It also has two simple properties.

$$\alpha > 1: \quad \Gamma(\alpha) = (\alpha-1)\,\Gamma(\alpha-1) \qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Thus, if α is an integer, $\Gamma(\alpha) = (\alpha-1)!$. The second parameterization given here is in terms of *shape* and *rate* parameters.

$$f(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-y\beta} \quad y \geq 0, \alpha > 0, \beta > 0 \qquad E\{Y\} = \mu_Y = \frac{\alpha}{\beta} \qquad V\{Y\} = \sigma_Y^2 = \frac{\alpha}{\beta^2}$$

Note that different software packages use the different parameterizations in generating samples and giving tail-areas and critical values. For instance, EXCEL uses the first parameterization and R uses the second. Figure 3.4 displays three gamma densities of various shapes.

Example 3.13: Rock and Roll Marathon Speeds

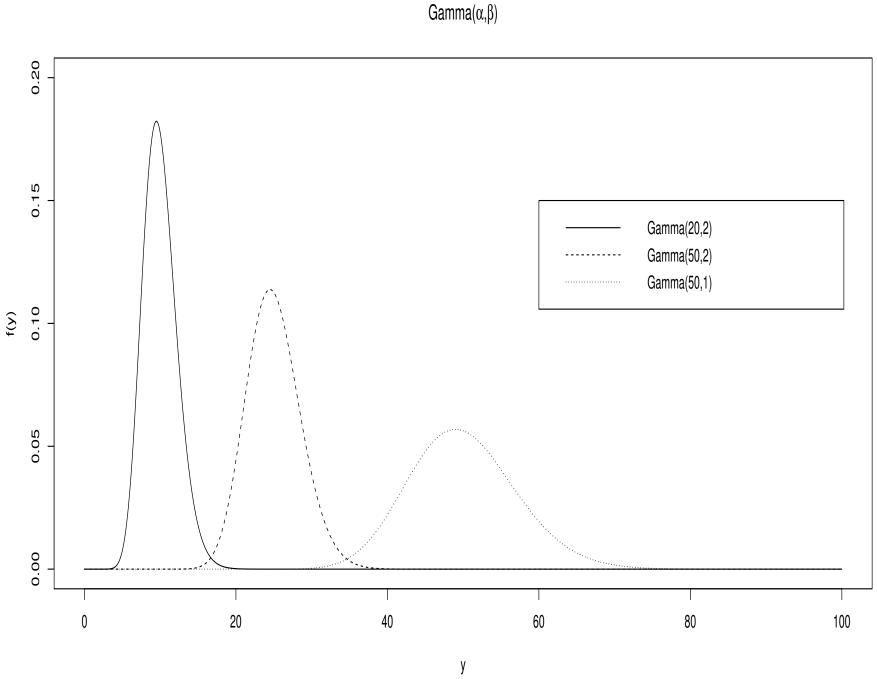

Figure 3.4: Three Gamma Densities

As seen previously, when considering females and males separately, the distributions of running speeds are all positive, and skewed to the right. The means for females and males were 5.8398 and 6.3370, respectively; and the variances were 0.6906 and 1.1187, respectively. Using the second formulation of the gamma distribution, with $\mu = \alpha/\beta$ and $\sigma^2 = \alpha/\beta^2$, we obtain the following parameter values for the two distributions based on the method of moments.

$$\frac{\mu^2}{\sigma^2} = \frac{(\alpha/\beta)^2}{\alpha/\beta^2} = \alpha \qquad \frac{\mu}{\sigma^2} = \frac{\alpha/\beta}{\alpha/\beta^2} = \beta$$

$$\text{Females: } \alpha_F = \frac{5.8398^2}{0.6906} = 49.38 \qquad \beta_F = \frac{5.8398}{0.6906} = 8.46$$

$$\text{Males: } \alpha_M = \frac{6.3370^2}{1.1187} = 35.90 \qquad \beta_M = \frac{6.3370}{1.1187} = 5.66$$

Histograms of the actual speeds and the corresponding Gamma densities are given in Figure 3.5. Similar to what was done for the NHL BMI measurements, we compare the theoretical quantiles for the female and male speeds with the actual quantiles, and compare theoretical probabilities for females and males with observed probabilities. There is very good agreement between the quantiles. The extreme probabilities do

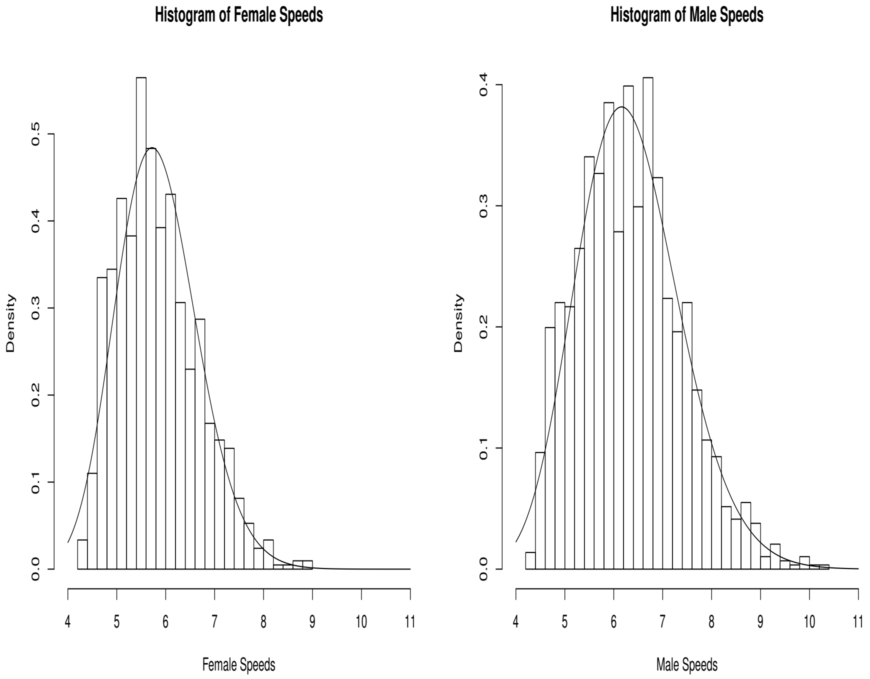

Figure 3.5: Rock and Roll Marathon speeds and Gamma Distributions for Females and Males

not match up as well, but still show fairly good agreement, with exception of no actual cases falling more than 2 standard deviations below the means.

R Output

```
## Output
> round(q.out, 3)
                     10%   25%    50   75%   90%
Theoretical/Female 4.803 5.260 5.800 6.377 6.927
Empirical/Female   4.811 5.203 5.711 6.357 7.015
Theoretical/Male   5.025 5.595 6.278 7.015 7.725
Empirical/Male     4.970 5.561 6.277 6.986 7.718

> round(p.out, 4)
                   <mu-2sigma (mu-sigma,mu+sigma) >mu+2sigma
Theoretical/Female     0.0146              0.6843     0.0298
Empirical/Female       0.0000              0.6622     0.0364
Theoretical/Male       0.0131              0.6850     0.0309
Empirical/Male         0.0000              0.6651     0.0365
```

∇

Two special cases are the exponential family, where $\alpha = 1$ and the Chi-square family, with $\alpha = \nu/2$ and $\beta = 2$ for integer valued ν. For the exponential family, based on the second parameterization, the symbol β

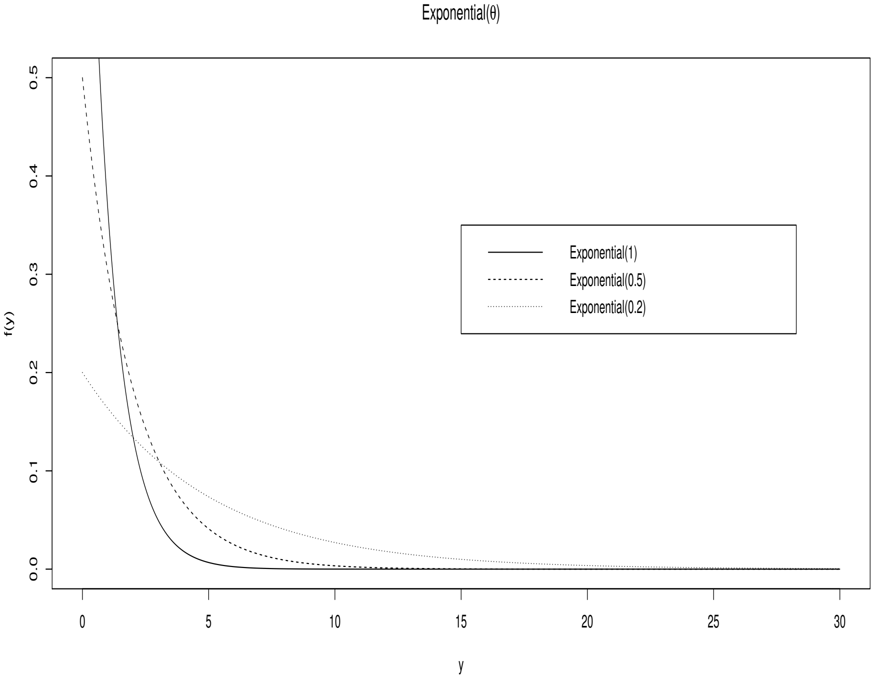

Figure 3.6: Three Exponential Densities

is often replaced by θ.

$$f(y) = \theta e^{-y\theta} \qquad E\{Y\} = \mu_Y = \frac{1}{\theta} \qquad V\{Y\} = \sigma_Y^2 = \frac{1}{\theta^2}.$$

Probabilities for the exponential distribution are trivial to obtain as $F(y^*) = 1 - e^{-y^*\theta}$. Figure 3.6 gives three exponential distributions.

For the chi-square family, based on the first parameterization, we have the following.

$$f(y) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right) 2^{\nu/2}} y^{\frac{\nu}{2}-1} e^{-y/2} \qquad E\{Y\} = \mu_Y = \nu \qquad V\{Y\} = \sigma_Y^2 = 2\nu$$

Here, ν is the **degrees of freedom** and we denote the distribution as: $Y \sim \chi_\nu^2$. Upper and lower critical values of the chi-square distribution are available in tabular form, and in statistical packages and spreadsheets. Probabilities, quantiles, densities, and random samples can be obtained with statistical packages and spreadsheets. The chi-square distribution is widely used in statistical testing as will be seen later. Figure 3.7 gives three Chi-square distributions.

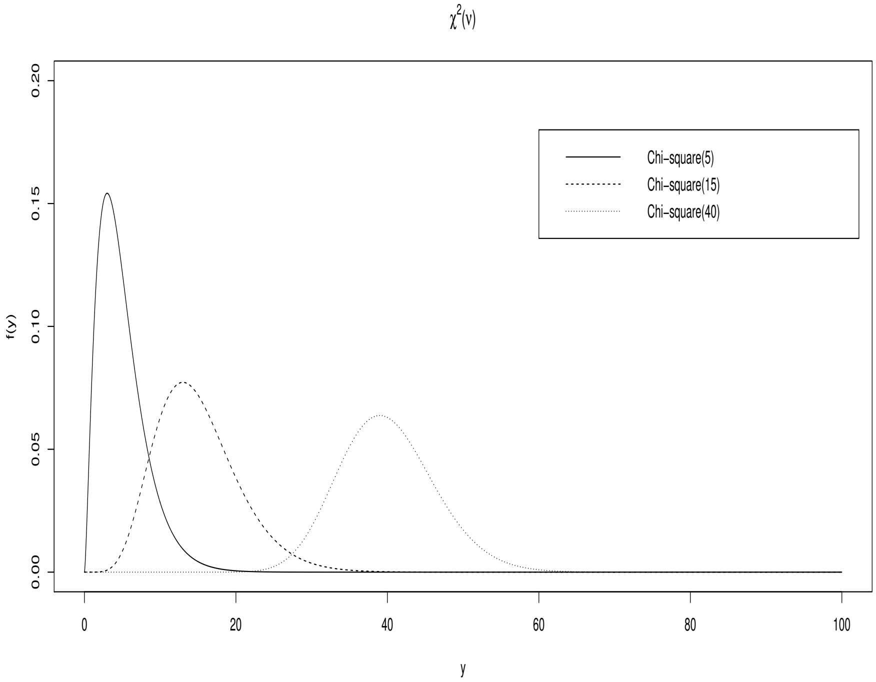

Figure 3.7: Three Chi-Square Densities

3.5 Sampling Distributions and the Central Limit Theorem

Sampling distributions are the probability distributions of sample statistics across different random samples from a population. That is, if we take many random samples, compute the statistic for each sample, then save that value, what would be the distribution of those saved statistics? In particular, if we are interested in the sample mean $\overline{Y}$, or the sample proportion with a characteristic $\hat{\pi}$, we know the following results, based on independence of elements within a random sample.

$$\text{Sample Mean: } E\{Y_i\} = \mu \quad V\{Y_i\} = \sigma^2 \quad E\{\overline{Y}\} = E\left\{\sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i\right\} = n\left(\frac{1}{n}\right)\mu = \mu$$

$$V\{\overline{Y}\} = V\left\{\sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i\right\} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 V\{Y_i\} = n\left(\frac{1}{n}\right)^2\sigma^2 = \frac{\sigma^2}{n}$$

$$SE\{\overline{Y}\} = \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

$$\text{Sample Proportion: } E\{Y_i\} = \pi \quad V\{Y_i\} = \pi(1-\pi) \quad E\{\hat{\pi}\} = E\left\{\sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i\right\} = n\left(\frac{1}{n}\right)\pi = \pi$$

$$V\{\hat{\pi}\} = V\left\{\sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i\right\} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 V\{Y_i\} = n\left(\frac{1}{n}\right)^2\pi(1-\pi) = \frac{\pi(1-\pi)}{n}$$

$$SE\{\hat{\pi}\} = \sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

The standard deviation of the sampling distribution of a sample statistic (aka estimator) is referred to as its **standard error**. Thus $SE\{\overline{Y}\} = \sigma_{\overline{Y}}$ is the standard error of the sample mean, and $SE\{\hat{\pi}\} = \sigma_{\hat{\pi}}$ is the standard error of the sample proportion.

When the data are normally distributed, the sampling distribution of the sample mean is also normal. When the data are not normally distributed, as the sample size increases, the sampling distribution of the sample mean or proportion tends to normality. The "rate" of convergence to normality depends on how "non-normal" the underlying distribution is. The mathematical arguments for these results are **Central Limit Theorems**.

$$\text{Sample Mean: } \overline{Y} \dot{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad \text{Sample Proportion: } \hat{\pi} \dot{\sim} N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Example 3.14: Sampling Distributions - NHL BMI, Female Marathon Speeds

We consider the sampling distributions of sample means for the NHL player Body Mass Indices, and Female Rock and Roll Marathon Speeds. For the NHL BMI data, the population mean is $\mu = 26.500$ and standard deviation is $\sigma = 1.454$. As the underlying distribution is approximately normal, the sampling distribution of the mean is approximately normal, regardless of the sample size. We take 10000 random samples of size $n = 9$, computing and saving the sample mean for each sample. The theoretical and empirical (based on the 10000 random samples) mean and standard error of the sample means are given below and a histogram with the normal density are shown in Figure 3.8.

$$\text{Theory: } \mu_{\overline{Y}} = \mu = 26.500 \quad \sigma_{\overline{Y}} = \frac{1.454}{\sqrt{9}} = 0.485 \qquad \text{Empirical: } \overline{\overline{y}} = 26.504 \quad s_{\overline{y}} = 0.485$$

The mean and standard deviation are very close to the corresponding theoretical values (they won't always be this close, as sampling error exists).

For the female marathon speeds, we saw that the distribution was skewed to the right, and well modeled by a gamma distribution with mean $\mu = 5.84$ and standard deviation $\sigma = 0.83$. We take 10000 random samples of $n = 16$ from this population, computing and saving the sample mean from each sample. The theoretical and empirical (based on the 10000 random samples) mean and standard error of the sample means are given below and a histogram with the normal density are shown in Figure 3.9.

$$\text{Theory: } \mu_{\overline{Y}} = \mu = 5.840 \quad SE\{\overline{Y}\} = \frac{0.831}{\sqrt{16}} = 0.208 \qquad \text{Empirical: } \overline{\overline{y}} = 5.839 \quad SE\{\overline{y}\} = 0.206$$

Again, we see very strong agreement between the empirical and theoretical values (as we should). Also, note that the sampling distribution is very well approximated by the N(5.840,0.208) in the graph.

∇

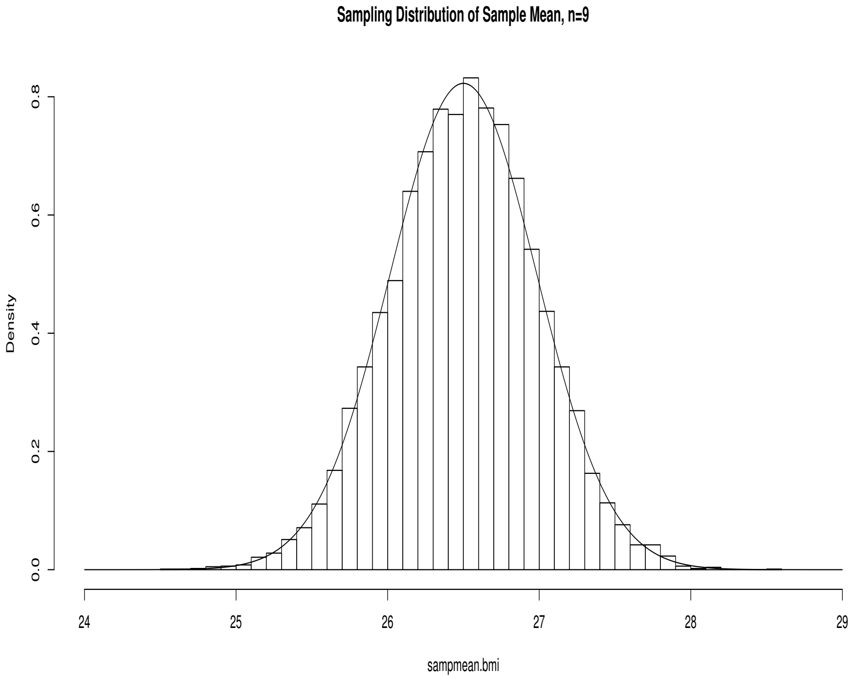

Figure 3.8: Sampling distribution for sample means (n=9) for NHL Body Mass Index

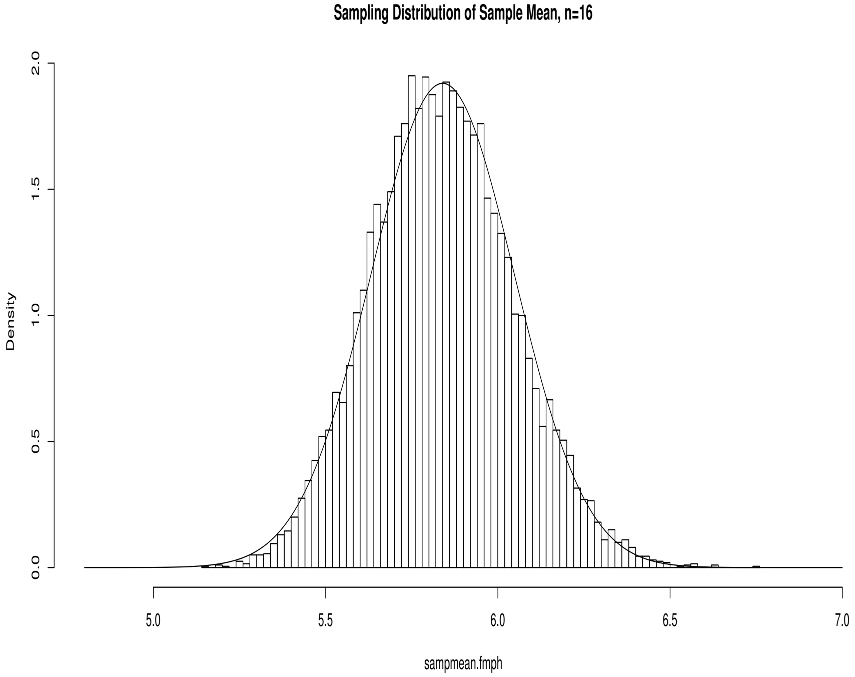

Figure 3.9: Sampling Distribution for sample means (n=16) for Female Rock and Roll Marathon speeds

Chapter 4

Inferences for Population Means

Researchers often are interested in making statements regarding unknown population means and medians based on sample data. There are two common methods for making inferences: **Estimation** and **Hypothesis Testing**. The two methods are related and make use of the sampling distribution of the sample mean when making statements regarding the population mean.

Estimation can provide a single "best" prediction of the population mean, a **point estimate**, or it can provide a range of values that hopefully encompass the true population mean, an **interval estimate**. Hypothesis testing involves setting an a priori (null) value for the unknown population mean, and measuring the extent to which the sample data contradict that value. Note that a confidence interval provides a credible set of values for the unknown population mean, and can be used to test whether or not the population mean is the null value. Both methods involve uncertainty as we are making statements regarding a population based on sample data.

4.1 Estimation

For large samples, the sample mean has an approximately normal sampling distribution centered at the population mean, μ, and a standard error $\sigma/\sqrt{n}$. When the data are normally distributed, the sampling distribution is normal for all sample sizes. For normal distributions, 95% of its density lies in the range (mean +/- 1.96 SD). Thus, when we take a random sample, we obtain the following probability statement regarding the sample mean.

$$\overline{Y} \dot{\sim} N\left(\mu, SE\{\overline{Y}\} = \frac{\sigma}{\sqrt{n}}\right) \quad \Rightarrow \quad P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{Y} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1-\alpha \qquad P\left(Z \geq z_a\right) = a$$

$$\Rightarrow \quad 1-\alpha \approx P\left(-z_{\alpha/2} \leq \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Some commonly used coverage probabilities $(1-\alpha)$ are given here, along with the corresponding z values.

$$1-\alpha = .90 \Rightarrow \alpha = .10 \Rightarrow \frac{\alpha}{2} = .05 \Rightarrow z_{.05} = 1.645 \quad 1-\alpha = .95 \Rightarrow z_{.025} = 1.96 \quad 1-\alpha = .99 \Rightarrow z_{.005} = 2.576$$

Note that in the probability statements above, μ is a fixed, unknown constant in practice, and $\overline{Y}$ is a random variable that varies from sample to sample. The probability refers to the fraction of the samples that will provide sample means such that the lower and upper bounds "cover" μ. Also, in practice, σ will be unknown and need to be replaced by the sample standard deviation.

A Large-Sample $(1-\alpha)100\%$ Confidence Interval for a Population Mean μ is given below, where $\overline{y}$ and s are the observed mean and standard deviation from a random sample of size n and $\hat{SE}\{\overline{Y}\}$ represents the **estimated standard error** .

$$\overline{y} \pm z_{\alpha/2}\hat{SE}\{\overline{Y}\} \qquad \overline{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

When the data are normally distributed, for small samples (although this has shown to work well for other distributions), replace $z_{\alpha/2}$ with $t_{\alpha/2,n-1}$.

$$\overline{y} \pm t_{\alpha/2,n-1}\hat{SE}\{\overline{Y}\} \qquad \overline{y} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Any software package or spreadsheet that is used to obtain a confidence interval for a mean (or difference between two means) will always use the version based on the t-distribution. There will be settings, when making confidence intervals for parameters, that there is no justification for using the t-distribution, and we will make use the z-distribution, as does statistical software packages.

Example 4.1: NHL Players' BMI

The Body Mass Indices for the NHL players are approximately normally distributed with mean $\mu = 26.500$ and standard deviation $\sigma = 1.454$. We take 10000 random samples of size $n = 12$, implying a standard error of $\sigma_{\overline{Y}} = 1.454/\sqrt{12} = 0.420$. We count the number of the 10000 sample means that lie in the ranges $\mu \pm z_{\alpha/2}\sigma_{\overline{Y}}$ for the three values of $1-\alpha$ given above.

Of the 10000 sample means, 8975 (89.75%) lied within $\mu \pm 1.645(.420)$, 9512 (95.12%) within $\mu \pm 1.96(.420)$, and 9902 (99.02%) within $\mu \pm 2.576(.420)$. Had we constructed intervals of the form $\overline{y} \pm z_{\alpha/2}(.420)$ for each sample mean, the coverage rates for μ would have been the same values (89.75%, 95.12%, 99.02%).

When the population standard error $SE\{\overline{Y}\} = \sigma/\sqrt{n}$ is replaced by the estimated standard error $\hat{SE}\{\overline{Y}\} = s/\sqrt{n}$, which varies from sample to sample, we find the coverage rates of the intervals decrease. When constructing intervals of the form $\overline{y} \pm z_{\alpha/2}s/\sqrt{n}$, the coverage rates fall to 86.78%, 92.29%, and 97.58%, respectively. This is a by-product of the fact that the sampling distribution of the standard deviation is skewed right, and its median is below its mean. Whenever the sample standard deviation is small, the width of the constructed interval is shortened. When using the estimated standard error, replace $z_{\alpha/2}$ with the corresponding critical value for the t-distribution, with $n-1$ degrees of freedom: $t_{\alpha/2,n-1}$. For this case, with $n = 12$, we obtain $t_{.05,11} = 1.796$, $t_{.025,11} = 2.201$, and $t_{.005,11} = 3.106$. When z is replaced by the corresponding t values, the coverage rates for the constructed intervals with the estimated standard errors reach their nominal rates: 89.79%, 95.22%, and 99.15%, respectively.

For the first random sample of the 10000 generated, we observe $\overline{y} = 25.838$ and $s = 1.717$. The 95% Confidence Interval for μ based on the first sample is obtained as follows.

$$\overline{y} \pm t_{.025,n-1}\frac{s}{\sqrt{n}} \quad \equiv \quad 25.838 \pm 2.201\left(\frac{1.717}{\sqrt{12}}\right) \quad \equiv \quad 25.838 \pm 1.091 \quad \equiv \quad (24.747, 26.929)$$

Thus, this interval does contain $\mu = 26.500$.

R Output

```
### Output
> round(cover.out,4)
                  90% Confidence 95% Confidence 99% Confidence
Z - True SE               0.8975         0.9512         0.9902
Z - Estimated SE          0.8678         0.9229         0.9758
t - Estimated SE          0.8979         0.9522         0.9915
```

∇

Often, researchers choose the sample size so that the **margin of error** will not exceed some fixed level E with high confidence. That is, we want the difference between the sample and population means to be within E with confidence level $1-\alpha$. This means the width of a $(1-\alpha)100\%$ Confidence Interval will be $2E$. This can be done in one calculation based on using the z distribution, or more conservatively, by trivial iteration based on the t-distribution. Either way, we must have an approximation of σ based on previous research or a pilot study.

$$z: \quad E_z = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \left(\frac{z_{\alpha/2}\sigma}{E_z}\right)^2 \qquad\qquad t: \text{Smallest } n \text{ such that } E_t \leq t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}$$

Example 4.2: Estimating Population Mean Male Marathon Speed

Suppose we want to estimate the population mean of the male Rock and Roll marathon running speeds within $E = 0.20$ miles per hour with 95% confidence. We treat the standard deviation as known, $\sigma = 1.058$. The calculation for the sample size based on the z-distribution is given below, followed by R commands that iteratively solve for n based on the t-distribution.

$$z: \quad z_{.025} = 1.96 \quad n = \left(\frac{1.96(1.058)}{0.20}\right)^2 = 107.5 \approx 108$$

R Output

```
## Output

> cbind(n, E.t)
       n       E.t
[1,] 110 0.1999336
```

Since n was needed to be so large, $z_{.025}$ and $t_{.025,n-1}$ are very close, and both methods give virtually the same n (108 and 110).

4.2 Hypothesis Testing

In hypothesis testing, a sample of data is used to determine whether a population mean is equal to some pre-specified level μ_0. It is rare, except in some situations to test whether the mean is some specific value based on historical level, or government or corporate specified level to have a null value to test. These tests are more common when comparing two or more populations or treatments and determining whether their means are equal. The elements of a hypothesis test are given below.

Null Hypothesis (H_0) Statement regarding a parameter that is to be tested. It always includes an equality, and the test is conducted assuming its truth.

Alternative (Research) Hypothesis (H_A) Statement that contradicts the null hypothesis. Includes "greater than" ($>$), "less than" ($<$),or "not equal too" ($\neq$)]

Test Statistic (T.S.) A statistic measuring the discrepancy between the sample statistic and the parameter value under the null hypothesis (where the equality holds).

Rejection Region (R.R.) Values of the Test Statistic for which the Null Hypothesis is rejected. Depends on the significance level of the test.

***P*-value** Probability under the null hypothesis (at the equality) of observing a Test Statistic as extreme or more extreme than the observed Test Statistic. Also known as the observed significance level.

Type I Error Rejecting the Null Hypothesis when in fact it is true. The Rejection Region is chosen so that this has a particular small probability ($\alpha = P$(Type I Error) is the **significance level** and is often set at 0.05).

Type II Error Failing to reject the Null Hypothesis when it is false. Depends on the true value of the parameter. Sample size is often selected so that it has a particular small probability for an important difference. $\beta = P$(Type II Error).

Power The probability the Null Hypothesis is rejected. When H_0 is true the power is $\pi = \alpha$, when H_A is true, it is $\pi = 1 - \beta$.

The testing procedure for a mean is based on the sampling distribution of $\overline{Y}$ being approximately normal with mean μ_0 under the null hypothesis. Also, when the data are normal the difference between the sample mean and μ_0 divided by its estimated standard error is distributed as t with $n-1$ degrees of freedom under the null hypothesis.

$$\overline{Y} \stackrel{\cdot}{\sim} N\left(\mu_0, SE\{\overline{Y}\} = \frac{\sigma}{\sqrt{n}}\right) \qquad \frac{\overline{Y}-\mu_0}{\hat{SE}\{\overline{Y}\}} = \frac{\overline{Y}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

When the absolute value of the t-statistic is large, there is evidence against the null hypothesis. Once a sample is taken (observed), and the sample mean $\overline{y}$ and sample standard deviation s are observed, the test is conducted as follows for 2-tailed, upper tailed, and lower tailed alternatives.

$$\text{2-tailed: } H_0:\mu=\mu_0 \quad H_A:\mu\neq\mu_0 \quad \text{T.S.: } t_{obs}=\frac{\overline{y}-\mu_0}{s/\sqrt{n}} \quad \text{R.R.: } |t_{obs}|\geq t_{\alpha/2,n-1} \quad P=2P\left(t_{n-1}\geq|t_{obs}|\right)$$

$$\text{Upper tailed: } H_0:\mu\leq\mu_0 \quad H_A:\mu>\mu_0 \quad \text{T.S.: } t_{obs}=\frac{\overline{y}-\mu_0}{s/\sqrt{n}} \quad \text{R.R.: } t_{obs}\geq t_{\alpha,n-1} \quad P=P\left(t_{n-1}\geq t_{obs}\right)$$

$$\text{Lower tailed: } H_0:\mu\geq\mu_0 \quad H_A:\mu<\mu_0 \quad \text{T.S.: } t_{obs}=\frac{\overline{y}-\mu_0}{s/\sqrt{n}} \quad \text{R.R.: } t_{obs}\leq -t_{\alpha,n-1} \quad P=P\left(t_{n-1}\leq t_{obs}\right)$$

The form of the rejection regions are given for 2-tailed, Upper and Lower tailed tests in Figure 4.1. These are based on $\alpha = 0.05$, and $n = 16$. The vertical lines lie at $t_{.975,15} = -t_{.025,15} = -2.131$ and $t_{.025,15} = 2.131$ for the 2-tailed test, $t_{.05,15} = 1.753$ for the Upper tailed test, and $t_{.95,15} = -t_{.05,15} = -1.753$ for the Lower tailed test.

When the Null Hypothesis is false, the test statistic is distributed as non-central t with non-centrality parameter given below.

$$H_0:\mu=\mu_0 \qquad \text{In reality: } \mu=\mu_A\neq\mu_0 \qquad \Delta=\frac{\mu_A-\mu_0}{\sigma/\sqrt{n}} \qquad t=\frac{\overline{Y}-\mu_0}{S/\sqrt{n}} \stackrel{\cdot}{\sim} t_{n-1,\Delta}$$

Power probabilities, which depend on whether the test is 2-tailed or 1-tailed can be obtained from statistical software packages, such as R, but not directly in EXCEL.

$$\text{2-tailed tests: } \pi = P\left(t_{n-1,\Delta}\leq -t_{\alpha/2,n-1}\right) + P\left(t_{n-1,\Delta}\geq t_{\alpha/2,n-1}\right)$$

$$\text{Lower tailed tests: } \pi = P\left(t_{n-1,\Delta}\leq -t_{\alpha,n-1}\right) \qquad \text{Upper tailed tests: } \pi = P\left(t_{n-1,\Delta}\geq t_{\alpha,n-1}\right)$$

While it is rare to use hypothesis testing regarding a single mean (except in the case where data are paired differences within individual units), the procedure is demonstrated based on male Rock and Roll marathon speeds with several values of μ_0.

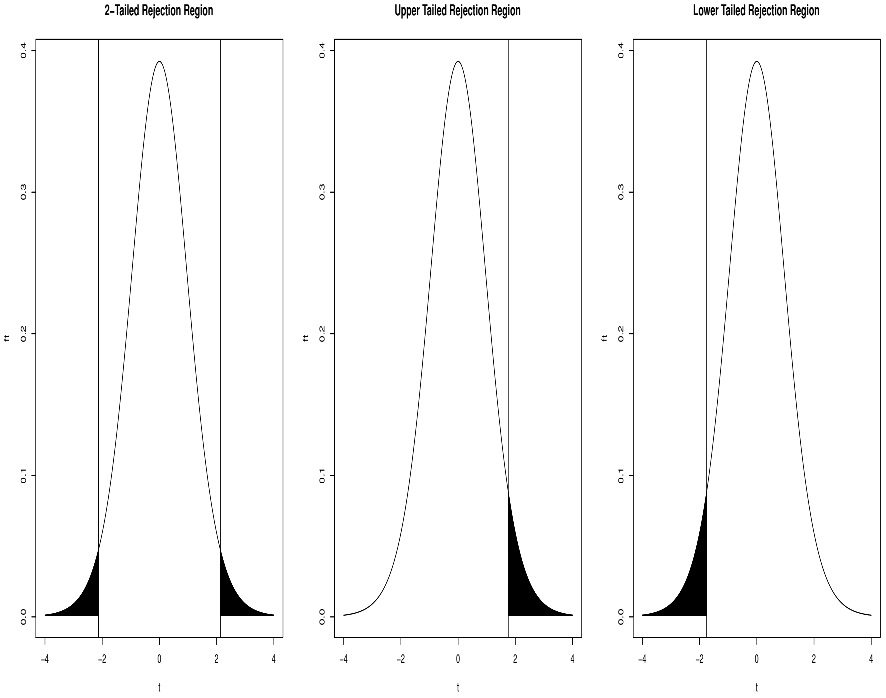

Figure 4.1: Rejection Regions for 2-tailed, Upper and Lower tailed tests, with $\alpha = 0.05$ and $n = 16$

Example 4.3: Male Rock and Roll Marathon Speeds

For the males participating in the Rock and Roll marathon, the population mean speed was $\mu = 6.337$ miles per hour with standard deviation of $\sigma = 1.058$. We will demonstrate hypothesis testing regarding a single mean by first testing $H_0 : \mu = 6.337$ versus $H_A : \mu \neq 6.337$, based on random samples of $n = 40$. Since the null hypothesis is true, if the test is conducted with a Type I Error rate of $\alpha = 0.05$, the test should reject the null in approximately 5% of samples. The distribution of the test statistic is t with $n - 1 = 39$ degrees of freedom. Further, the P-values should approximate a Uniform distribution between 0 and 1. Note that 482 (4.82%) of the 10000 samples reject the null hypothesis, in agreement with what is to be expected. A histogram of the observed test statistics, along with the t-density, and the P-values and the the Uniform density are given in Figure 4.2. The two vertical bars on the t-statistic plot are at $\pm t_{.025,39} = \pm 2.023$.

Next consider cases where the null hypothesis is not true. Consider $H_{01} : \mu = 6$ versus $H_{A1} : \mu \neq 6$ and $H_{02} : \mu = 6.5$ versus $H_{A2} : \mu \neq 6.5$. Since the null value for H_{02} is closer to the true value $\mu_A = 6.337$ than the null value for H_{01}, we expect that we will reject H_{02} less often for tests based on the same sample size. That is, the power is higher for H_{01} than H_{02}. The non-centrality parameters and the corresponding power values are given below, based on samples of $n = 40$.

$$\Delta_1 = \frac{6.337 - 6.0}{1.058/\sqrt{40}} = 2.015 \qquad \pi_1 = .5022 \qquad\qquad \Delta_2 = \frac{6.337 - 6.5}{1.058/\sqrt{40}} = -0.974 \qquad \pi_2 = .1583$$

Based on 10000 random samples from the male marathon speeds, 49.93% rejected $H_0 : \mu = 6$, and for another set of 10000 random samples, 17.05% rejected $H_0 : \mu = 6.5$. The histogram of the test statistics and

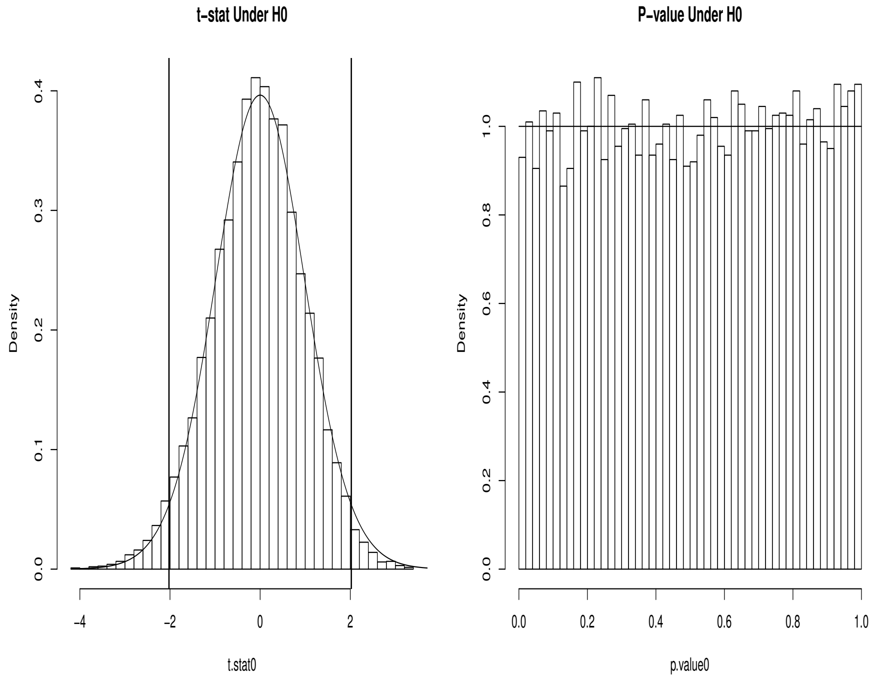

Figure 4.2: t-statistics and P-values for testing $H_0 : \mu = 6.337$

the non-central t-distribution are given in Figure 4.3 for testing $H_0 : \mu = 6$.

R Output

```
## Output

> round(power.out, 4)
           Delta Theoretical Power Empirical Power
mu0=6.33  0.0000            0.0500          0.0482
mu0=6.00  2.0150            0.5022          0.4993
mu0=6.50 -0.9748            0.1583          0.1705
```

∇

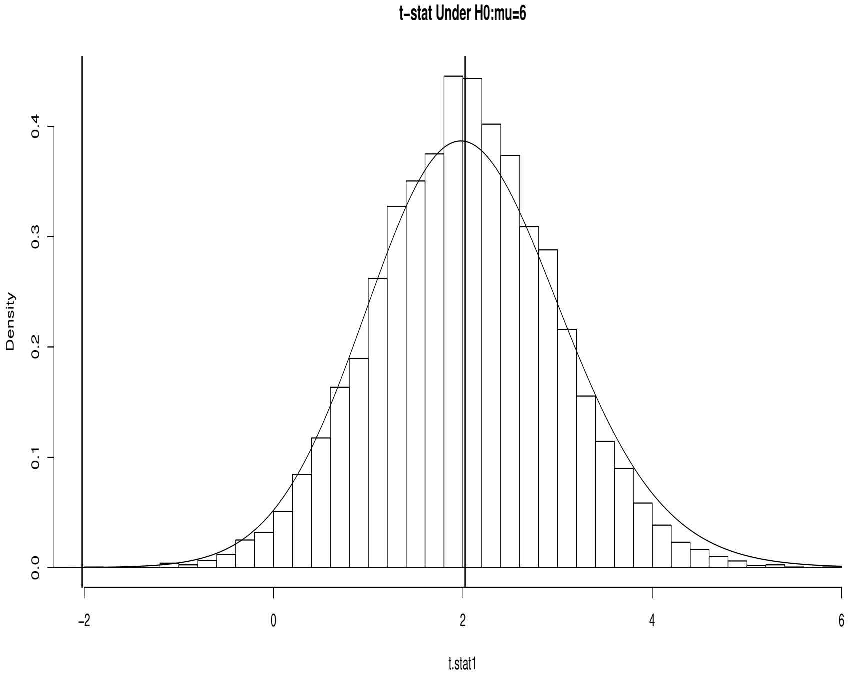

Figure 4.3: t-statistics and non-central t-distribution for testing $H_0 : \mu = 6.0$

Chapter 5

Introduction to Experimentation

This chapter will briefly introduce the following models which will be described in detail in subsequent chapters.

- Observational Studies vs Controlled Experiments
- Completely Randomized Design
- Randomized Block Design
- Factorial Designs
- Chi-Square Tests for Categorical Variables
- Regression Models

Studies can be described as **Observational** or as **Controlled Experiments**. Observational studies occur when experimental/sampling units are obtained from existing populations. These could be different brands of a product, animals from different species, or people who do or do not practice a particular habit. In controlled experiments, a sample of units is obtained, and the individual elements of the sample are randomly assigned to the various conditions/treatments. This could involve batches of raw material being assigned to various machines, mice being assigned to various doses of a chemical compound, or humans assigned to various advertising campaigns.

The **Completely Randomized Design (CRD)** simply randomizes the experimental units to the various treatments in the most basic manner with no restrictions on randomization. The **Randomized Block Design (RBD)** first creates "blocks" of units that are similar based on some external criteria (e.g. age or skill) and then assigns units to treatments within blocks. In many experiments, each individual unit may receive each treatment, and the units are treated as blocks.

In **Factorial Designs**, there are multiple treatment factors that are simultaneously controlled. These can be structured as a CRD or a RBD. In some cases, one or more of the factors may be controlled, while others may be observed. In many engineering applications, a set of k factors, each at two or more levels, may be observed to determine which variables have the largest impacts on the response, or to optimize the response.

When the independent and dependent variables are categorical, **Chi-Square Tests** can be conducted to determine whether the response variable is associated with the predictor, or grouping variable.

When the independent and dependent variables are both numeric, **Regression Models** can be applied to measure the associations among variables. These models can be extended in many ways to various types of predictor and response variables.

Chapter 6

Comparing Two Population Means

While estimating the mean or median of a population is important, many more applications involve comparing two or more treatments or populations. There are two commonly used designs: **independent samples** and **paired samples**. Independent samples are used in controlled experiments when a sample of experimental units is obtained, and randomly assigned to one of two treatments or conditions. That is, each unit receives only one of the two treatments. These are often referred to as **Completely Randomized** or **Parallel Groups** or **Between Subjects** designs in various fields of study. Paired samples can involve the same experimental unit receiving each treatment, or units being matched based on external criteria, then being randomly assigned to the two treatments within pairs. These are often referred to as **Randomized Block** or **Crossover** or **Within Subjects** designs.

In observational studies, independent samples can be taken from two existing populations, or elements within two populations can be matched based on external criteria and observed. In each case, the goal is to make inferences concerning the difference between the two means or medians based on sample data.

6.1 Independent Samples

In the case of independent samples, assume we sample n_1 units or subjects in treatment 1 which has a population mean response μ_1 and population standard deviation σ_1. Further, a sample of n_2 elements from treatment 2 is obtained where the population mean is μ_2 and standard deviation is σ_2. Measurements within and between samples are independent. Regardless of the distributions of the individual measurements, we have the following results based on linear functions of random variables, in terms of the means of the two random samples. The notation used is Y_{1j} is the j^{th} unit (replicate) from sample 1, and Y_{2j} is the j^{th} unit (replicate) from sample 2. In the case of independent samples, these two random variables are independent.

$$\overline{Y}_1 = \frac{\sum_{j=1}^{n_1} Y_{1j}}{n_1} = \sum_{j=1}^{n_1}\left(\frac{1}{n_1}\right)Y_{1j} \quad \Rightarrow \quad E\{\overline{Y}_1\} = \mu_1 \quad V\{\overline{Y}_1\} = \frac{\sigma_1^2}{n_1} \quad E\{\overline{Y}_2\} = \mu_2 \quad V\{\overline{Y}_2\} = \frac{\sigma_2^2}{n_2}$$

$$E\{\overline{Y}_1 - \overline{Y}_2\} = E\{\overline{Y}_1\} - E\{\overline{Y}_2\} = \mu_1 - \mu_2$$

$$V\{\overline{Y}_1 - \overline{Y}_2\} = \sigma^2_{\overline{Y}_1 - \overline{Y}_2} = V\{\overline{Y}_1\} + V\{\overline{Y}_2\} - 2\mathrm{COV}\{\overline{Y}_1, \overline{Y}_2\} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + 0 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$SE\{\overline{Y}_1 - \overline{Y}_2\} = \sigma_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If the data are normally distributed, $\overline{Y}_1 - \overline{Y}_2$ is also normally distributed. If the data are not normally distributed, $\overline{Y}_1 - \overline{Y}_2$ will be approximately normally distributed in large samples. As in the case of a single mean, how large of samples are needed depends on the shape of the underlying distributions.

The problem arises again that the variances will be unknown and must be estimated. For large sample sizes n_1 and n_2, we have the following approximation for the sampling distribution of the following quantity, where the sample variances replace the true population variances.

$$\frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \mathrel{\dot\sim} N(0,1)$$

$$\Rightarrow \quad P\left(\left(\overline{Y}_1 - \overline{Y}_2\right) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\overline{Y}_1 - \overline{Y}_2\right) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) \approx 1 - \alpha$$

Example 6.1: NHL and EPL Players' BMI

Body Mass Indices for all National Hockey League (NHL) and English Premier League (EPL) football players for the 2013/4 season were obtained. Identifying the NHL as league 1 and EPL as league 2 we have the following population parameters.

$$N_1 = 717 \quad \mu_1 = 26.500 \quad \sigma_1 = 1.454 \qquad N_2 = 526 \quad \mu_2 = 23.019 \quad \sigma_2 = 1.711$$

A plot of the two population histograms, along with normal densities is given in Figure 6.1. Both distributions are well approximated by the normal distribution, with the NHL having a substantially higher mean and EPL having a slightly higher standard deviation.

We take 100000 independent random samples of sizes $n_1 = n_2 = 20$ from the two populations, each time computing and saving $\overline{y}_1, s_1, \overline{y}_2, s_2$. A histogram of the 100000 sample mean differences and the superimposed Normal density with mean $\mu_1 - \mu_2 = 3.481$ and standard error 0.502 (calculation given below) is shown in Figure 6.2. The mean of the 100000 mean differences $\overline{y}_1 - \overline{y}_2$ is 3.482 with standard deviation (standard error) 0.493. Both are very close to their theoretical values (as they should be). Then we compute the following quantity (and interval), counting the number of samples for which it contains $\mu_1 - \mu_2$, and its average estimated variance (squared standard error).

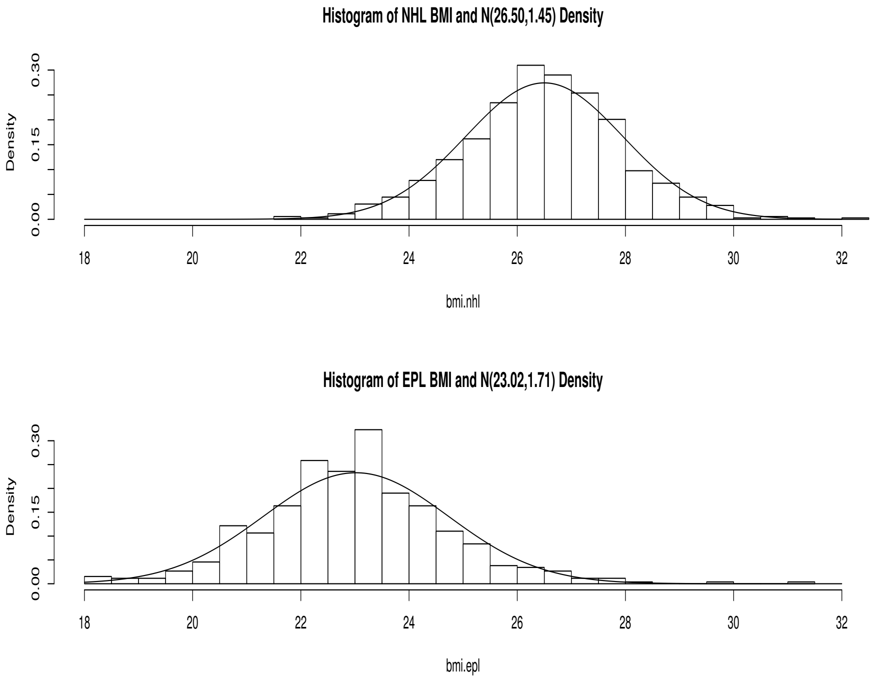

Figure 6.1: Distributions of NHL and EPL players Body Mass Index

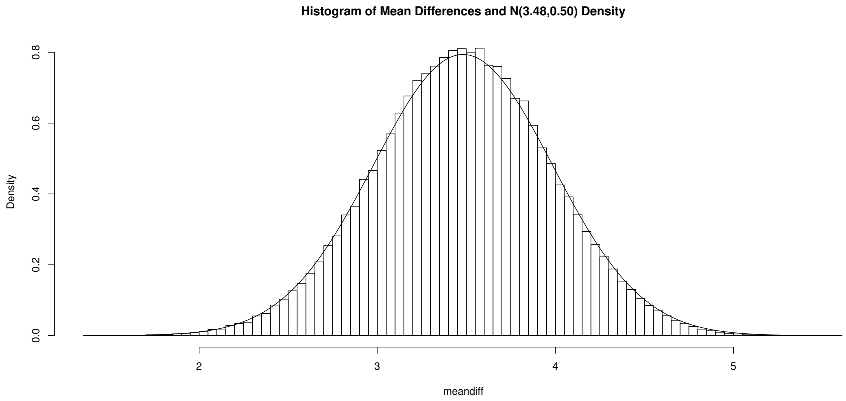

Figure 6.2: 100000 sample mean differences ($n_1 = n_2 = 20$) for NHL and EPL BMI values and Normal Density

$$(\overline{y}_1 - \overline{y}_2) \pm 1.96\sqrt{\frac{s_1^2}{20} + \frac{s_2^2}{20}} \qquad \mu_1 - \mu_2 = 26.500 - 23.019 = 3.481$$

$$SE\left\{\overline{Y}_1 - \overline{Y}_2\right\} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{1.454^2}{20} + \frac{1.711^2}{20}} = 0.502$$

The mean of the 100000 sample mean differences is 3.479 compared to the theoretical mean difference of 3.481. The standard deviation of the sample mean differences is 0.493, compared to the theoretical standard error of 0.502.

Of the intervals constructed from each sample mean difference and its estimated standard error (using s_1, s_2 in place of σ_1, σ_2), the interval contains the true mean difference (3.481) for 94.698% of the samples, very close to the nominal 95% coverage rate. If we replace $z_{.025} = 1.96$ with the more appropriate $t_{.025,n_1+n_2-2} = t_{.025,38} = 2.0244$, the coverage rate increases to 95.395%. Note that virtually all software packages will automatically use t in place of z, however, there are various statistical methods that always use the z case.

The average of the estimated variance of $\overline{y}_1 - \overline{y}_2$: $s_1^2/n_1 + s_2^2/n_2$ is 0.2527, while its theoretical value is $\sigma_1^2/n_1 + \sigma_2^2/n_2 = 0.2521$. Note that the variance of the estimated difference is unbiased, not so for the standard error.

R Output

```
### Output

> round(md.out, 3)
      mu1    mu2 sigma1 sigma2  n mu1-mu2 SE{Yb1-Yb2} Mean(yb1-yb2) SD(yb1-yb2) cover(z) cover(t)
[1,] 26.5 23.019  1.454  1.711 20   3.481       0.502         3.479       0.493    0.947    0.954
```

∇

This logic leads to a large-sample test and Confidence Interval regarding $\mu_1-\mu_2$ once estimates $\overline{y}_1, s_1, \overline{y}_2, s_2$ have been observed in an experiment or observational study. The Confidence Interval and test are given below. Typically, $z_{\alpha/2}$ is replaced with $t_{\alpha/2,\nu}$, where ν is the degrees of freedom, which depends on assumptions involving the variances (see below).

Large Sample $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$: $(\overline{y}_1 - \overline{y}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

2-tail: $H_0 : \mu_1-\mu_2 = \Delta_0 \quad H_A : \mu_1-\mu_2 \neq \Delta_0 \quad TS : z_{obs} = \frac{(\overline{y}_1 - \overline{y}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad RR : |z_{obs}| \geq z_{\alpha/2} \quad P = 2P(Z \geq |z_{obs}|)$

Upper tail: $H_0 : \mu_1 - \mu_2 \leq \Delta_0 \quad H_A : \mu_1 - \mu_2 > \Delta_0 \quad TS : z_{obs} = \dfrac{(\overline{y}_1 - \overline{y}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad RR : z_{obs} \geq z_\alpha \quad P = P(Z \geq z_{obs})$

Lower tail: $H_0 : \mu_1 - \mu_2 \geq \Delta_0 \quad H_A : \mu_1 - \mu_2 < \Delta_0 \quad TS : z_{obs} = \dfrac{(\overline{y}_1 - \overline{y}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad RR : z_{obs} \leq z_\alpha \quad P = P(Z \leq z_{obs})$

Example 6.2: Gender Classification from Physical Measurements

A study in forensics used measurements of the length and breadth of the scapula from samples of 95 male and 96 female Thai adults (Peckmann, Scott, Meek, Mahakkanukrauh (2017), [47]). The measurements were length and breadth of glenoid cavity (LGC and BGC, in mm), respectively. Summary data for the two samples for BGC are given below.

$$n_m = 95 \quad \overline{y}_m = 27.87 \quad s_m = 2.04 \qquad n_f = 96 \quad \overline{y}_f = 23.77 \quad s_f = 1.85$$

$$\overline{y}_m - \overline{y}_f = 27.87 - 23.77 = 4.10 \qquad \hat{SE}\{\overline{Y}_m - \overline{Y}_f\} = \sqrt{\frac{2.04^2}{95} + \frac{1.85^2}{96}} = 0.282$$

A 95% Confidence Interval for the population mean difference, $\mu_m - \mu_f$ is given below.

$$\left(\overline{y}_m - \overline{y}_f\right) \pm z_{.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \equiv \quad 4.10 \pm 1.960(0.282) \quad \equiv \quad 4.10 \pm 0.553 \quad \equiv \quad (3.55, 4.65)$$

The interval is very far away from 0, making us very confident that the population mean is higher for males than females. To test whether the population means differ (which they clearly do from the Confidence Interval), we conduct the following 2-tailed test with $\alpha = 0.05$.

$$H_0 : \mu_m - \mu_f = 0 \quad H_A : \mu_m - \mu_f \neq 0 \quad T.S. : z_{obs} = \frac{4.10 - 0}{0.282} = 14.54 \quad R.R. : |z_{obs}| \geq 1.960 \quad P = 2P(Z \geq 14.54) \approx 0$$

∇

6.2 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 an independent samples (parallel groups) or paired samples (crossover) design? The appropriate test for each situation is given in Table 6.1. 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.

	Design Type	
	Completely Randomized	Randomized Block
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 6.1: Statistical Tests for small–sample 2 group situations

6.2.1 Independent Samples (Completely Randomized Designs)

Completely Randomized 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 drug or placebo), or possibly selected at random from one of two populations (as in Example 5.1, where we had NHL and EPL players and in Example 5.2 where they measured males and females). In the case where the two populations of measurements are normally distributed, the 2–sample t–test is used. Note that it also works well for reasonably large sample sizes when the measurements are not normally distributed. This procedure is very similar to the large–sample test from the previous section, where only the critical values for the rejection region changes. In the case where the populations of measurements are not approximately normal, the Wilcoxon Rank–Sum test (or, equivalently the Mann–Whitney U–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 similar to the large–sample test, except the critical values for the rejection regions and Confidence Intervals are based on the t–distribution with $\nu = n_1+n_2-2$ degrees of freedom and the variances are "pooled" (see below). 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 an appropriate test, which will be given below. 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.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
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: $2P\left(t_{n_1+n_2-2} > |t_{obs}|\right)$ or $P\left(t_{n_1+n_2-2} > t_{obs}\right)$ or $P\left(t_{n_1+n_2-2} < t_{obs}\right)$ (again, depending on which alternative you are using).

Example 6.3: Comparison of Two Instructional Methods

A study was conducted (Rusanganwa (2013) [49]) to compare two instructional methods: multimedia (treatment 1) and traditional (treatment 2) for teaching physics to undergraduate students in Rwanda. Subjects were assigned at random to the two treatments. Each subject received only one of the two methods. The numbers of subjects who completed the courses and took two exams were $n_1 = 13$ for the multimedia course and $n_2 = 19$ for the traditional course. The primary response was the post-course score on an examination. We will conduct the test $H_0 : \mu_1 - \mu_2 = 0$ vs $H_A : \mu_1 - \mu_2 \neq 0$, where the null hypothesis is no difference in the effects of the two methods. The summary statistics are given below.

$$n_1 = 13 \quad \overline{y}_1 = 11.10 \quad s_1 = 3.47 \qquad n_2 = 19 \quad \overline{y}_2 = 8.35 \quad s_2 = 2.45$$

First, compute s_p^2, the pooled variance:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(13-1)(3.47)^2 + (19-1)(2.45)^2}{13+19-2} = \frac{252.54}{30} = 8.42 \quad (s_p = 2.90)$$

Now 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.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(11.10-8.35)}{\sqrt{8.42\left(\frac{1}{13}+\frac{1}{19}\right)}} = \frac{2.75}{1.04} = 2.633$
- R.R.: $|t_{obs}| \geq t_{\alpha/2,n_1+n_2-2} = t_{.05/2,13+19-2} = t_{.025,30} = 2.042$
- P-value: $2P\left(t_{30} \geq |t_{obs}|\right) = 2P\left(t_{30} \geq 2.633\right) = 0.0132$

Based on this test, reject H_0 (for any $\alpha \geq .0132$), and conclude that the population mean post course scores differ under these two conditions. The 95% Confidence Interval for $\mu_1 - \mu_2$ is $2.75 \pm 2.042(1.04) \equiv (0.62, 4.88)$ which does not contain 0.

Below we use generated samples that have the same means and standard deviation and use **t.test** function in R to conduct the 2-sample t-test.

R Commands and Output

```
## Commands

rp <- read.csv("http://www.stat.ufl.edu/~winner/data/rwanda_physics.csv")
attach(rp); names(rp)
t.test(score ~ trt.y, var.equal=T) # t-test with single y-var and trt id

## Output

> t.test(score ~ trt.y, var.equal=T)

        Two Sample t-test

data:  score by trt.y
t = 2.6323, df = 30, p-value = 0.01327
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6163295 4.8826179
sample estimates:
mean in group 1 mean in group 2
      11.100000        8.350526
```

∇

When the population variances are not equal, there is no justification for pooling the sample variances to better estimate the common variance σ^2. In this case the estimated standard error of $\overline{Y}_1 - \overline{Y}_2$ is $\sqrt{s_1^2/n_1 + s_2^2/n_2}$. An adjustment is made to the degrees of freedom for an approximation to a t-distribution of the t-statistic.

$$\frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \stackrel{\cdot}{\sim} t_\nu \qquad\qquad \nu = \frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\left[\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}\right]}$$

The test is referred to as **Welch's Test**, and the degrees of freedom **Satterthwaite's Approximation**. Statistical software packages automatically compute the approximate degrees of freedom. The approximation extends to more complex models as well. Once the samples are obtained, and the sample means and standard deviations are computed, the $(1-\alpha)100\%$ Confidence Interval for $\mu_1 - \mu_2$ is computed as follows.

$$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad\qquad \nu = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\left[\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}\right]}$$

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).

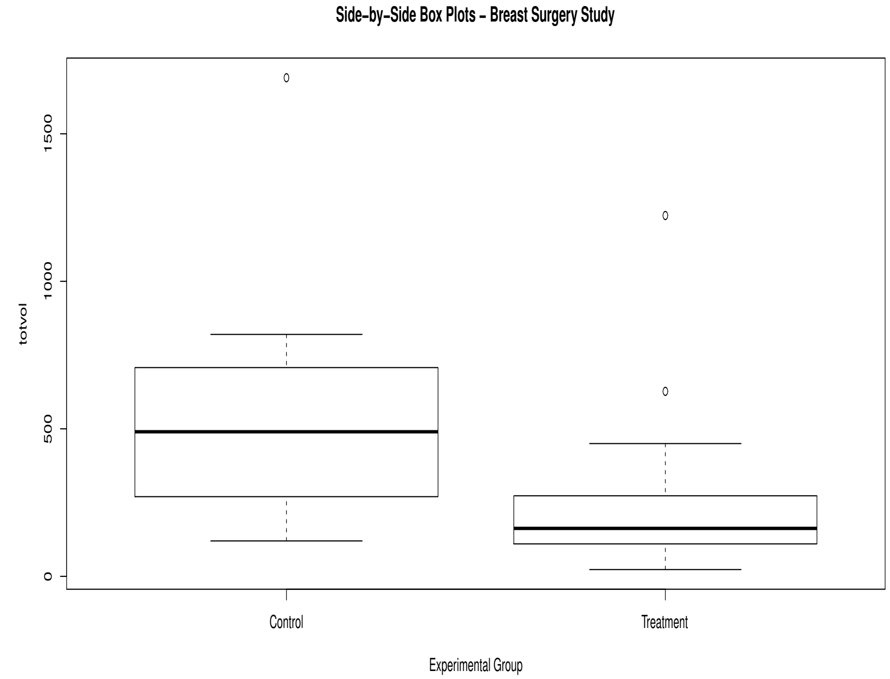

Figure 6.3: Abdominal drainage in breast reconstruction surgery, DIEP procedure with and without abdominal suture quilting.

3. T.S.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

4. R.R.: $|t_{obs}| \geq t_{\alpha/2,\nu}$ or $t_{obs} \geq t_{\alpha,\nu}$ or $t_{obs} \leq -t_{\alpha,\nu}$ (which R.R. depends on which alternative hypothesis you are using).

5. p-value: $2P\left(t_\nu \geq |t_{obs}|\right)$ or $P\left(t_\nu \geq t_{obs}\right)$ or $P\left(t_\nu \leq t_{obs}\right)$ (again, depending on which alternative you are using).

Example 6.4: Abdominal Quilting to Reduce Drainage in Breast Reconstruction Surgery

A study considered the effect of abdominal suture quilting on abdominal drainage during breast reconstruction surgery (Liang, et al, (2016), [37]). A group of $n_1 = 27$ subjects (controls) received the standard DIEP procedure, while a group of $n_2 = 26$ subjects (treatment) received the DIEP procedure along with the suture quilting. The response measured was the amount of abdominal drainage during the surgery (in ml). The summary data are given below, note that the sample standard deviations are substantially different, and these are relatively large sample sizes. Side-by-side box plots are given in Figure 6.3.

$$n_1 = 27 \quad \overline{y}_1 = 527.78 \quad s_1 = 322.07 \qquad n_2 = 26 \quad \overline{y}_2 = 238.31 \quad s_2 = 242.66$$

The estimated mean difference, standard error, and degrees of freedom are computed below.

$$\overline{y}_1 - \overline{y}_2 = 527.78 - 238.31 = 289.47 \qquad \hat{SE}\{\overline{Y}_1 - \overline{Y}_2\} = \sqrt{\frac{322.07^2}{27} + \frac{242.66^2}{26}} = 78.14$$

$$\nu = \frac{\left[\frac{322.07^2}{27} + \frac{242.66^2}{26}\right]^2}{\left[\frac{(322.07^2/27)^2}{27-1} + \frac{(242.66^2/26)^2}{26-1}\right]} = 48.25 \qquad t_{.025,48.25} = 2.010$$

The 95% Confidence Interval for $\mu_1 - \mu_2$ and test statistic and P-value for testing $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$ are given below. There is strong evidence that the suture quilting reduces blood loss during surgery.

$$95\% \text{ CI for } \mu_1 - \mu_2\text{: } 289.47 \pm 2.010(78.14) \quad \equiv \quad 289.47 \pm 157.06 \quad \equiv \quad (132.41, 446.53)$$

$$\text{T.S.: } t_{obs} = \frac{289.47}{78.14} = 3.705 \qquad P\left(t_{48.25} \geq 3.705\right) = .0005$$

R Commands and Output

```
## Commands

quilt <- read.csv("http://www.stat.ufl.edu/~winner/data/breast_diep.csv")
attach(quilt); names(quilt)

trt.f <- factor(trt)
levels(trt.f) <- c("Control", "Treatment")
t.test(totvol ~ trt.f, var.equal=F)

## Output

> t.test(totvol ~ trt, var.equal=F)

        Welch Two Sample t-test

data:  totvol by trt
t = 3.7043, df = 48.25, p-value = 0.0005452
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 132.3707 446.5695
sample estimates:
mean in group 1 mean in group 2
       527.7778        238.3077
```

∇

6.2.2 Paired Sample Designs

In paired samples (aka crossover or within subjects) designs, subjects receive each treatment, thus acting as their own control. They may also have been matched based on some characteristics. Procedures based on these designs take this into account, and are based in determining differences between treatments after "removing" variability in the subjects (or pairs). When it is possible to conduct them, paired sample designs are more powerful than independent sample 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 and when measurements within subjects or pairs are positively correlated.

Paired t–test for Normally Distributed Data

In paired sample designs, each subject (or pair) receives each treatment. In the case of two treatments being compared, we compute the difference in the two measurements within each subject (or pair), 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. Then this is simply a 1-sample problem on the differences.

Let Y_1 be the score in condition 1 for a randomly selected subject, and Y_2 be the score in condition 2 for the subject. Let $D = Y_1 - Y_2$ be the difference. Further, suppose the following assumptions and their corresponding results. Note that the differences across subjects (or pairs) are considered to be independent.

$$E\{Y_1\} = \mu_1 \qquad V\{Y_1\} = \sigma_1^2 \qquad E\{Y_2\} = \mu_2 \qquad V\{Y_2\} = \sigma_2^2 \qquad \text{COV}\{Y_1, Y_2\} = \sigma_{12}$$

$$\Rightarrow \quad E\{D\} = \mu_1 - \mu_2 = \mu_D \qquad V\{D\} = \sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$$

$$\overline{D} = \frac{\sum_{i=1}^n D_i}{n} \qquad E\{\overline{D}\} = \mu_D \qquad V\{\overline{D}\} = \sigma_{\overline{D}}^2 = \frac{\sigma_D^2}{n} \qquad SE\{\overline{D}\} = \sigma_{\overline{D}} = \frac{\sigma_D}{\sqrt{n}}$$

$$\text{For large } n\text{: } \overline{D} \dot{\sim} N\left(\mu_D, SE\{\overline{D}\} = \frac{\sigma_D}{\sqrt{n}}\right)$$

Normality holds for any sample size if the individual measurements (or the differences) are normally distributed.

It should be noted that in the paired case $n_1 = n_2$ by definition. That is, there will always be equal sized samples when the experiment is conducted properly. There will be $n = n_1 = n_2$ differences, even though there were $2n = n_1 + n_2$ measurements made. From the n differences obtained in a sample, the mean and standard deviation are obtained, and will labeled as $\overline{d}$ and s_d.

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

A $(1-\alpha)100\%$ Confidence Interval for the population mean difference μ_D is given below.

$$\overline{d} \pm t_{\alpha/2,n-1}\hat{SE}\{\overline{D}\} \qquad \equiv \qquad \overline{d} \pm t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}}$$

The test 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.: $t_{obs} = \frac{\overline{d}}{\hat{SE}\{\overline{D}\}} = \frac{\overline{d}}{\left(\frac{s_d}{\sqrt{n}}\right)}$
4. R.R.: $|t_{obs}| \geq t_{\alpha/2,n-1}$ or $t_{obs} \geq t_{\alpha,n-1}$ or $t_{obs} \leq -t_{\alpha,n-1}$ (which R.R. depends on which alternative hypothesis you are using).
5. p-value: $2P(t_{n-1} \geq |t_{obs}|)$ or $P(t_{n-1} \geq t_{obs})$ or $P(t_{n-1} \leq t_{obs})$ (again, depending on which alternative you are using).

Example 6.5: Comparison of Two Analytic Methods for Determining Wine Isotope

A study was conducted to compare two analytic methods for determining $^{87}Sr/^{86}Sr$ isotope ratios in wine samples (Durante, et al (2015), [21]). These are used in geographic tracing of wine. The two methods are microwave (method 1) and low temperature (method 2). The data, and the differences (microwave - lowtemp) are given in Table 6.2.

As there are $n = 18$ differences, the degrees of freedom are $n - 1 = 17$. The 95% Confidence Interval for μ_D is computed below, where $t_{.025,17} = 2.110$. First, the mean and standard deviation of the differences are multiplied by 100000 (remove first 5 0s after decimal) to reduce the risk of calculation error. This is legitimate as the mean and standard deviation are of the same units. This leads to $\overline{d}^* = 0.3667$ and $s_d^* = 2.46466$.

$$0.3667 \pm 2.110\frac{2.4646}{\sqrt{18}} \quad \equiv \quad 0.3667 \pm 2.110(0.5809) \quad \equiv \quad 0.3667 \pm 1.2257 \quad \equiv \quad (-0.8590, 1.5924)$$

In the original units the interval is of the form of (-.00000859,.000015924). Since the interval contains 0, there is no evidence that one method tends to score higher (or lower) than the other on average.

The test of whether there is a difference in the true mean determinations between the two methods (with $\alpha = 0.05$) is conducted by completing the steps outlined below.

1. $H_0 : \mu_1 - \mu_2 = \mu_D = 0$
2. $H_A : \mu_D \neq 0$
3. T.S.: $t_{obs} = \frac{0.3667}{\left(\frac{2.4646}{\sqrt{18}}\right)} = \frac{0.3667}{0.5809} = 0.631$

sample id	microwave	lowtemp	diff(m-l)
1	0.70866	0.70861	0.000050000
2	0.708762	0.708792	-0.00003000
3	0.708725	0.708734	-0.00000900
4	0.708668	0.708662	0.000006000
5	0.708675	0.70867	0.000005000
6	0.708702	0.708713	-0.00001100
7	0.708647	0.708661	-0.00001400
8	0.708677	0.708667	0.000010000
9	0.709145	0.709176	-0.00003100
10	0.709017	0.709024	-0.00000700
11	0.70882	0.708814	0.000006000
12	0.709402	0.709364	0.000038000
13	0.709374	0.709378	-0.00000400
14	0.709508	0.709517	-0.00000900
15	0.70907	0.709063	0.000007000
16	0.709061	0.709079	-0.00001800
17	0.709096	0.709039	0.000057000
18	0.70872	0.7087	0.000020000
Mean	0.708929	0.708926	0.000003667
SD	0.000287	0.000288	0.000024646

Table 6.2: $^{87}SR/^{86}SR$ Isotope ratios for 18 wine samples by Microwave and Low Temperature Methods

4. R.R.: $t_{obs} > t_{\alpha/2,n-1} = t_{.025,17} = 2.110$
5. P-value: $2P\left(t_{17} \geq 0.631\right) = .5364$

There is definitely no evidence that the two methods differ in terms of determinations of wine isotope ratios.

R Commands and Output

```
## Commands

wine1 <- read.csv("http://www.stat.ufl.edu/~winner/data/wine_isotope.csv")
attach(wine1); names(wine1)

## t.test Function
t.test(microwave, lowtemp, paired=TRUE)

## Output

> round(wine.out, 6)
        ybar1       s1    ybar2       s2 cor(y1,y2)
[1,] 0.708929 0.000287 0.708926 0.000288   0.996329

> round(diff.out,9)
          mean         SD   Std Err         t   P(>|t|)         LB         UB
[1,] 3.667e-06 2.4646e-05 5.809e-06 0.6311987 0.5363058 -8.589e-06 1.5923e-05

> t.test(microwave, lowtemp, paired=TRUE)
```

```
        Paired t-test

data:  microwave and lowtemp
t = 0.6312, df = 17, p-value = 0.5363
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.589364e-06  1.592270e-05
sample estimates:
mean of the differences
           3.666667e-06
```

∇

6.3 Nonparametric Tests

When data are highly skewed, the extreme measurements can have large impacts on the group means and standard deviations. Two rank-based tests that are not effected by outliers are the **Wilcoxon Rank-Sum Test** for independent samples and the **Wilcoxon Signed-Rank Test** for paired samples. Note that for independent samples there is an alternative, but mathematically equivalent **Mann-Whitney U-Test** that is often reported. These tests require special tables for small samples and have normal approximations for larger samples. We briefly describe them here and use R for the computations.

6.3.1 Independent Samples - Rank-Sum Test

For the Rank-Sum test, let the sample sizes for groups 1 and 2 be n_1 and n_2, respectively. Let the combined sample size be $n_. = n_1 + n_2$.

1. Rank the measurements across treatments from 1 (smallest) to $n_.$ (largest), adjusting for ties by giving the average rank for tied cases.
2. Obtain the rank sums for each treatment: T_1 and T_2 with $T_1 + T_2 = 1 + 2 + \ldots + n_. = \frac{n_. \times (n_. + 1)}{2}$
3. The test involves looking for discrepancies between T_1 and T_2 with what would be expected under the hypothesis of equal medians, namely $E\{T_i\} = \frac{n_i \times (n_. + 1)}{2}$
4. Special tables or statistical software packages can be used for the tests.

Example 6.6: Abdominal Quilting to Reduce Drainage in Breast Reconstruction Surgery

With sample sizes of $n_1 = 27$ and $n_2 = 26$ being well above the limits of standard tables, we will use R for the test. The rank sums are $T_1 = 962.5$ and $T_2 = 468.5$, respectively, with expected values under the hypothesis of equal medians being 729 and 702, respectively. These actual values are much higher than expected for the control group and much lower than expected for the treatment group. The approximate p-value is very small, implying a higher median for Controls than Treated patients.

```
## Commands

quilt <- read.csv("http://www.stat.ufl.edu/~winner/data/breast_diep.csv")
attach(quilt); names(quilt)

trt.f <- factor(trt)
levels(trt.f) <- c("Control", "Treatment")
wilcox.test(totvol ~ trt.f)

## Output

> wilcox.test(totvol ~ trt.f)

        Wilcoxon rank sum test with continuity correction

data:  totvol by trt.f
W = 584.5, p-value = 3.384e-05
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(550L, 160L, 250L, 600L, 720L, 680L,  :
  cannot compute exact p-value with ties
```

∇

6.3.2 Paired Samples - Signed Rank Test

When the data are paired, differences are taken within the pairs as in the paired t-test. Then the absolute values of the differences are ranked from smallest (1) to largest (n), again with tied differences receiving average ranks. The rank sum for positive differences (T^+) and negative differences (T^-) are obtained with $T^+ + T^- = 1 + \cdots + n = n(n+1)/2$. Then T^+ and T^- can be compared with their expected values which are both $n(n+1)/4$. Again, special tables are available, or statistical software packages can be used for the test.

Example 6.7: Comparison of Two Analytic Methods for Determining Wine Isotope

For the wine samples, there were $n = 18$ pairs analyzed by the microwave and low temperature methods. The ranks sums for the positive and negative differences were $T^+ = 87.5$ and $T^- = 83.5$, respectively, each with expected value equal to $18(19)/4 = 85.5$. The observed values are very close to their expected values under the hypothesis of equal medians. The p-value is .9479, implying no evidence of difference in location for the two analytic methods.

```
## Commands

wine1 <- read.csv("http://www.stat.ufl.edu/~winner/data/wine_isotope.csv")
attach(wine1); names(wine1)

wilcox.test(microwave, lowtemp, paired=TRUE)

## Output

> wilcox.test(microwave, lowtemp, paired=TRUE)
```

```
        Wilcoxon signed rank test with continuity correction

data:  microwave and lowtemp
V = 87.5, p-value = 0.9479
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(microwave, lowtemp, paired = TRUE) :
  cannot compute exact p-value with ties
```

∇

Chapter 7

Experimental Design and the Analysis of Variance

Chapter 6 covered methods to make comparisons between the means of a numeric response variable for two treatments or groups. The cases were considered where the experiment was conducted as an independent samples (aka parallel groups, between subjects) design, as well as a paired (aka crossover, within subjects) design.

This chapter 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 variable as a **factor**, and their levels as **treatments**. The following situations will be covered.

- 1–Factor, Independent Samples Designs (Completely Randomized Design)
- 1– Treatment Factor, Paired Designs (Randomized Block Design)

In all situations, there will be a numeric response variable, and at least one categorical (or possibly numeric, with several levels) independent variable. The goal will always be to compare mean (or median) responses among several populations. When all factor levels for a factor are included in the experiment, the factor is said to be **fixed**. When a sample of a larger population of factor levels are included, the factor is said to be **random**. Only fixed effects designs are considered here.

7.1 Completely Randomized Design (CRD) For Independent Samples

In the Completely Randomized Design, there is one factor that is controlled. This factor has k levels (which are often treatment groups), and n_i units are measured for the i^{th} level of the factor. Observed responses are defined 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 based on random responses as follows where the factor is **fixed** (all levels of interest are included in the experiment).

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \ldots, k; \quad j = 1, \ldots, n_i$$

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} allows for the fact that there will be variation among the measurements of different subjects (units) receiving the same treatment. A common parameterization that has nice properties is to assume $\sum n_i \alpha_i = 0$.

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). Note that if $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ then $\mu_1 = \cdots = \mu_k$.

As with the case where there are two treatments to compare, tests based on the assumption that the k populations are normal (mound–shaped) will be used, either assuming equal or unequal variances. Also, alternative tests (based on ranks) that do not assume that the k populations are normal can be used to compare population medians. These are the **Kruskal-Wallis Test** for the CRD and **Friedman's Test** for the RBD. These tests will not be covered in these notes.

7.1.1 Tests Based on Normally Distributed Data

When the underlying populations of measurements that are to be compared are approximately normal, with equal variances, the F–test is appropriate. To conduct this test, 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, define the following items, based on random outcomes Y_{ij} where i indexes treatment and j represents the replicate number, with n_i observations for treatment i and $n_. = n_1 + \cdots + n_k$.

$$Y_{ij} \sim N(\mu_i, \sigma)$$

$$\overline{Y}_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \qquad \overline{Y}_{i.} \sim N\left(\mu_i, \frac{\sigma}{\sqrt{n_i}}\right)$$

$$\overline{Y}_{..} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}}{n_.} = \frac{\sum_{i=1}^{k} n_i \overline{Y}_{i.}}{n_.}$$

$$\text{Total (Corrected) Sum of Squares: } TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{..}\right)^2 \qquad df_{Total} = n_. - 1$$

$$\text{Between Treatment Sum of Squares: } SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 = \sum_{i=1}^{k} n_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 \qquad df_T = k - 1$$

Within Treatment (Error) Sum of Squares: $SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i.}\right)^2 = \sum_{i=1}^{k}\left(n_i - 1\right)S_i^2 \qquad df_E = n_. - k$

Under the null hypothesis of no treatment effects ($\mu_1 = \cdots = \mu_k = \mu$), or equivalently ($\alpha_1 = \cdots = \alpha_k = 0$) the following results are obtained, where MST and MSE are mean squares for treatments and error, respectively.

$$E\left\{MST\right\} = E\left\{\frac{SST}{k-1}\right\} = \sigma^2$$

$$E\left\{MSE\right\} = E\left\{\frac{SSE}{n_. - k}\right\} = \sigma^2$$

Under the null hypothesis of no treatment effects, $E\left\{MST\right\} = E\left\{MSE\right\} = \sigma^2$ and the ratio MST/MSE follows the F-distribution with $k-1$ numerator and $n_. - k$ denominator degrees of freedom. When the null is not true and not all $\alpha_i = 0$, then the ratio follows the non-central F-distribution with parameter λ given below.

$$\frac{MST}{MSE} \sim F_{\nu_1,\nu_2,\lambda} \qquad \lambda = \frac{\sum_{i=1}^{k} n_i\alpha_i^2}{\sigma^2} \qquad \nu_1 = k-1 \qquad \nu_2 = n_. - k$$

Once samples have been obtained and the y_{ij} are observed, the F-test is conducted as follows.

$$\begin{aligned}
\overline{y}_{i.} &= \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \\
s_i &= \sqrt{\frac{\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2}{n_i - 1}} \\
n_. &= n_1 + \cdots + n_k \\
\overline{y}_{..} &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}{n_.} \\
TSS &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{..}\right)^2 \\
SST &= \sum_{i=1}^{k} n_i\left(\overline{y}_{i.} - \overline{y}_{..}\right)^2 \\
SSE &= \sum_{i=1}^{k}(n_i - 1)s_i^2
\end{aligned}$$

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. TSS is the total variability in the data (ignoring treatments), SST measures the variability in the sample means among the treatments (weighted by the sample sizes), and SSE measures the variability within the treatments.

Note that the goal is 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 measuring differences in the sample means. A test for treatment effects is conducted after constructing an Analysis of Variance table, as shown in Table 7.1. In that table, there are *sums of squares* for treatments (SST), for error (SSE), and total (TSS). Also, there are *degrees of freedom*, which represent the number of "independent" terms in the sum of squares. Then, the *mean squares*, are sums of squares divided by their degrees of freedom. Finally, the F–statistic is computed as $F = MST/MSE$. This will serve as the test statistic. Note that MSE is an extension of the pooled variance computed in Chapter 6 for two groups, and often it is written as $MSE = s^2$.

ANOVA

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

Table 7.1: The Analysis of Variance Table for the Completely Randomized (Parallel Groups) Design

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.: $F_{obs} \geq F_{\alpha,k-1,n_{.}-k}$
5. p-value: $P(F_{k-1,n_{.}-k} \geq F_{obs})$

Example 7.1: Body Mass Indices of NHL, NBA, and EPL, Players

Consider an extension of the Body Mass Index analysis to include National Basketball Association players. The populations are NHL ($i = 1$), NBA ($i = 2$), and EPL ($i = 3$). Histograms for the three populations are given in Figure 7.1. The population sizes, means, and standard deviations are given below.

$$N_1 = 707 \quad N_2 = 505 \quad N_3 = 526 \qquad \mu_1 = 26.50 \quad \mu_2 = 24.74 \quad \mu_3 = 23.02 \qquad \sigma_1 = 1.45 \quad \sigma_2 = 1.72 \quad \sigma_3 = 1.71$$

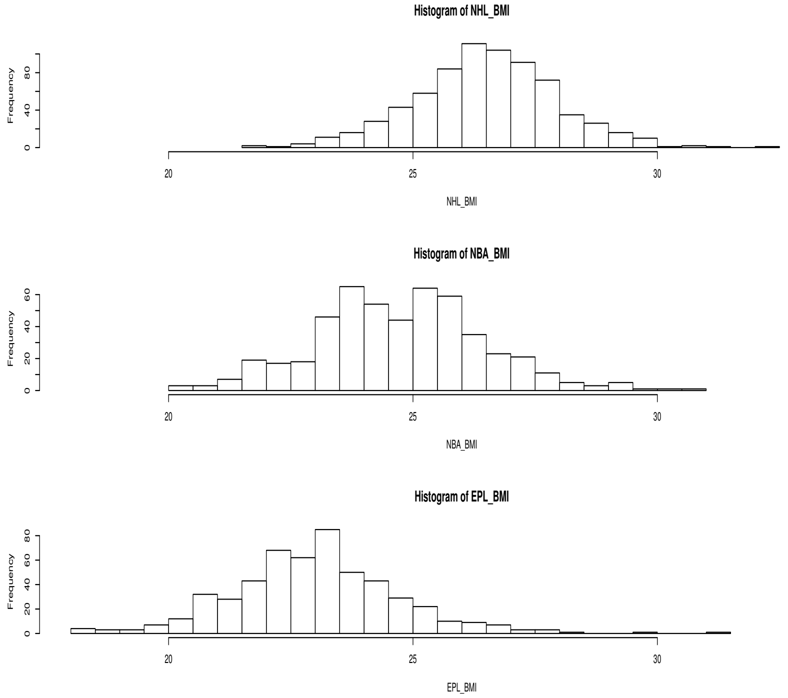

Figure 7.1: Histograms of NHL, NBA, and EPL Body Mass Indices

While the population standard deviations (and thus variances) are not all equal, a "pooled" variance is used for computational purposes. Also, μ and α_i are computed.

$$\sigma^2 = \frac{717\left(1.45^2\right) + 505\left(1.72^2\right) + 526\left(1.71^2\right)^2}{717 + 505 + 526} = 2.60 \qquad \mu = \frac{717(26.50) + 505(24.74) + 526(23.02)}{717 + 505 + 526} = 24.94$$

$$\alpha_1 = 26.50 - 24.94 = 1.56 \qquad \alpha_2 = 24.74 - 24.94 = -0.20 \qquad \alpha_3 = 23.02 - 24.94 = -1.92$$

Note that these α_i are obtained under the assumption $\sum N_i \alpha_i = 0$. If samples of sizes $n_1 = n_2 = n_3 = 4$ and $n_1 = n_2 = n_3 = 12$ are taken, the following F-distributions for the ratio MST/MSE are obtained.

$$n_i = 4: \quad \frac{MST}{MSE} \sim F_{\nu_1,\nu_2,\lambda_1} \quad \lambda_1 = \frac{4\left(1.56^2 + (-0.20)^2 + (-1.92)^2\right)}{2.60} = 9.48 \quad \nu_1 = 3-1 = 2 \quad \nu_2 = 12-3 = 9$$

$$n_i = 12: \quad \frac{MST}{MSE} \sim F_{\nu_1,\nu_2,\lambda_2} \quad \lambda_2 = \frac{12\left(1.56^2 + (-0.20)^2 + (-1.92)^2\right)}{2.60} = 28.43 \quad \nu_1 = 3-1 = 2 \quad \nu_2 = 36-3 = 33$$

When $n_1 = n_2 = n_3 = 4$, the critical value for testing $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ at $\alpha = 0.05$ significance level is $F_{.05,2,9} = 4.256$. The power of the F-test under this configuration is $\pi_1 = .636$. When $n_1 = n_2 = n_3 = 12$, the critical value for testing $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ at $\alpha = 0.05$ significance level is $F_{.05,2,33} = 3.285$.

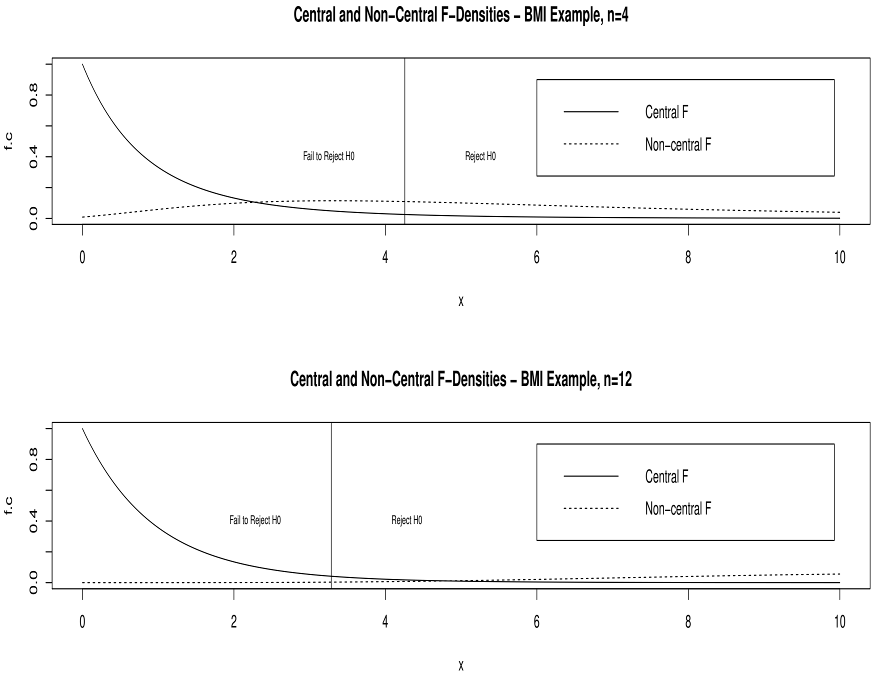

Figure 7.2: Central and non-central F-distributions for Body Mass Index example

The power of the F-test under this configuration is $\pi_2 = .997$. The central F-densities and the non-central F-densities with $\lambda_1 = 9.48$ and $\lambda_2 = 28.43$ for the denominator degrees of freedom of 9 and 33 are given in Figure 7.2.

Based on 100000 random sample of size $n_i = 4$ from each league, the F-test rejected the null hypothesis of no league differences in 63.4% of the samples. With samples of size $n_i = 12$, 99.7% of the F-tests rejected the null hypothesis. Despite the fact that the populations of measurements are not exactly normally distributed with equal variances, the test performs as expected. Computations for the first samples of size $n_1 = n_2 = n_3 = 12$ are given below.

$$\overline{y}_{1.} = 26.666 \quad \overline{y}_{2.} = 24.986 \quad \overline{y}_{3.} = 22.449 \quad \overline{y}_{..} = 24.701 \qquad s_1 = 1.968 \quad s_2 = 1.762 \quad s_3 = 1.149$$

$$SST = 12\left[(26.666 - 24.701)^2 + (24.986 - 24.701)^2 + (22.449 - 24.701)^2\right] = 108.167$$

$$df_T = 3 - 1 = 2 \qquad MST = \frac{108.167}{2} = 54.084$$

$$SSE = (12 - 1)\left[1.968^2 + 1.762^2 + 1.149^2\right] = 91.277 \quad df_E = 3(12) - 3 = 33 \quad MSE = \frac{91.277}{33} = 2.766$$

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad TS : F_{obs} = \frac{54.084}{2.766} = 19.55 \quad RR : F_{obs} \geq 3.285 \quad P = P\left(F_{2,33} \geq 19.55\right) < .0001$$

R Output

```
### Output

> round(ftest.out, 4)
     df_T df_E F(>05) P(F_obs>F(.05))
[1,]    2   33 3.2849          0.9942
> F[1]
[1] 19.55004
> cbind(ybar1[1], ybar2[1], ybar3[1], ybar[1], sd1[1], sd2[1], sd3[1])
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
[1,] 26.66637 24.98606 22.44932 24.70058 1.968428 1.762007 1.148883
```

∇

Example 7.2: Comparison of 5 Mosquito Repellents

A study compared $k = 5$ mosquito repellent patches on fabric for soldiers in military operations (Bhatnagar and Mehta (2007), [9]). The 5 treatments were: Odomos (1), Deltamethrin (2), Cyfluthrin (3), Deltamethrin+Odomos (4), and Cyfluthrin+Odomos (5), with $n_i = 30$ subjects per treatment, and a total of $n_. = 150$ measurements. The response observed was the "Per Man-Hour Mosquito Catch." Sample statistics are given in Table 7.2, and the Analysis of Variance is given in Table 7.3. Data that have been generated to match the means and standard deviations are plotted in Figure 7.3. The overall mean (long line) and individual treatment means (short lines) are included.

Treatment	n_i	$\overline{y}_{i.}$	s_i
Odomos (1)	30	7.900	3.367
Deltamethrin (2)	30	8.133	3.461
Cyfluthrin (3)	30	8.033	3.011
D+O(4)	30	6.333	3.122
C+O (5)	30	5.367	3.068

Table 7.2: Sample statistics for Mosquito Repellent study

		ANOVA				
Source of	Degrees of	Sum of	Mean			
Variation	Freedom	Squares	Square	F_{obs}	$F_{.05}$	P
TREATMENTS	4	184.650	46.163	4.478	2.434	.0019
ERROR	145	1494.680	10.308			
TOTAL	149	1679.334				

Table 7.3: The Analysis of Variance table for the Mosquito Repellent study

1. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0 \quad (\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5)$ (No treatment effect)
2. $H_A :$ Not all α_i are 0 (Treatment effects exist)
3. T.S. $F_{obs} = \frac{MST}{MSE} = 4.478$

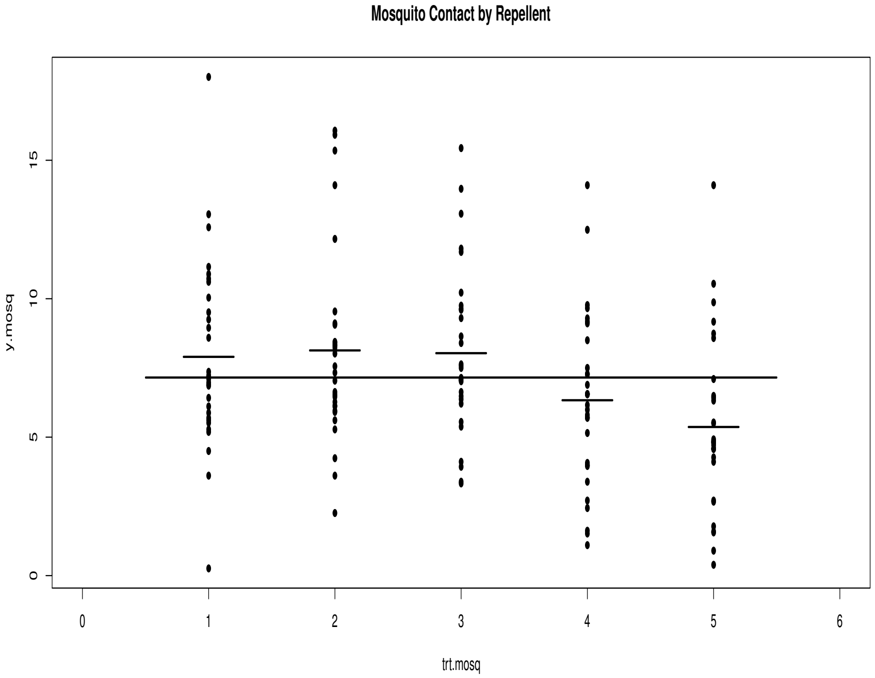

Figure 7.3: Mosquito catch by repellent treatment - data generated to match treatment means and standard deviations

4. R.R.: $F_{obs} > F_{\alpha,k-1,n-k} = F_{0.05,4,145} = 2.434$

5. P-value: $P(F_{k-1,n_{.}-k} \geq F_{obs}) = P(F_{4,145} \geq 4.478) = .0019$

The following R output gives the Analysis of Variance and the F-test.

R Output

```
### Output

> round(aov.out, 4)
           df        SS      MS      F F(.05)  P(>F)
Treatment   4  184.6501 46.1625 4.4782 2.4341 0.0019
Error     145 1494.6843 10.3082     NA     NA     NA
Total     149 1679.3345      NA     NA     NA     NA
```

The following R commands use the **aov** function to obtain the Analysis of Variance based on the raw data (not summary statistics).

R Commands and Output

```
## Commands
mp <- read.csv("http://www.stat.ufl.edu/~winner/data/mosquito_patch.csv")
attach(mp); names(mp)

trt.mosq <- factor(trt.mosq)
mosq.mod <- aov(y.mosq ~ trt.mosq)
summary(mosq.mod)

## Output

> summary(mosq.mod)
             Df Sum Sq Mean Sq F value  Pr(>F)
trt.mosq      4  184.6   46.16    4.48 0.00192 **
Residuals   145 1494.1   10.30
```

Since the F-statistic is sufficiently large, conclude that the means differ, then the following methods are used to make comparisons among treatments.

∇

Comparisons among Treatment Means

Assuming that it has been concluded that treatment means differ, we generally would like to know which means are significantly different. This is generally done by making contrasts among treatments. Special cases of contrasts include pre–planned or all pairwise comparisons between pairs of treatments.

A **contrast** is a linear function of treatment means, where the coefficients sum to 0. A contrast among population means can be estimated with the same contrast among sample means, and inferences can be made based on the sampling distribution of the contrast. Let C be the contrast among the population means, and $\hat{C}$ be its estimator based on means of the independent random samples.

$$C = a_1\mu_1 + \cdots + a_k\mu_k = \sum_{i=1}^{k} a_i\mu_i \text{ where } \sum_{i=1}^{k} a_i = 0 \qquad \hat{C} = a_1\overline{Y}_{1.} + \cdots + a_k\overline{Y}_{k.} = \sum_{i=1}^{k} a_i\overline{Y}_i$$

$$V\{\hat{C}\} = \sigma^2 \left[\frac{a_1^2}{n_1} + \cdots + \frac{a_k^2}{n_k}\right] = \sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}$$

When the sample sizes are balanced (all n_i are equal), the formula for the variance clearly simplifies. Contrasts are specific to particular research questions, so the general rules for tests and Confidence Intervals are given here, followed by an application to the Mosquito Repellent study. Since the coefficients sum to 0, we are virtually always testing $H_0 : C = 0$ in practice.

Once samples are obtained, obtain $\hat{c}$, the contrast based on the observed sample means among the treatments.

$$\hat{c} = a_1\overline{y}_{1.} + \cdots + a_k\overline{y}_{k.} = \sum_{i=1}^{k} a_i\overline{y}_{i.} \qquad \hat{SE}\{\hat{C}\} = \sqrt{MSE \sum_{i=1}^{k} \frac{a_i^2}{n_i}}$$

Testing whether a contrast is equal to 0 and obtaining a $(1-\alpha)100\%$ Confidence Interval for C are done as follow.

$$H_0 : C = 0 \quad H_A : C \neq 0 \quad TS : t_C = \frac{\hat{c}}{\hat{SE}\{\hat{C}\}} \quad RR : |t_C| \geq t_{\alpha/2,n_.-k} \quad P = 2P\left(t_{n_.-k} \geq |t_C|\right)$$

$$(1-\alpha)100\% \text{ Confidence Interval for } C : \hat{c} \pm t_{\alpha/2,n_.-k}\hat{SE}\{\hat{C}\}$$

The test can be conducted as upper or lower-tailed with obvious adjustments. An alternative approach is to compute the sums of squares for the contrast SSC, and use an F-test, comparing its Mean Square to MSE.

$$SSC = \frac{(\hat{c})^2}{\sum_{i=1}^{k} \frac{a_i^2}{n_i}} \qquad df_C = 1 \qquad MSC = \frac{SSC}{1} = SSC$$

$$H_0 : C = 0 \quad H_A : C \neq 0 \quad TS : F_C = \frac{MSC}{MSE} \quad RR : F_C \geq F_{\alpha,1,n_.-k} \quad P = P\left(F_{1,n_.-k} \geq F_C\right)$$

Example 7.3: Comparison of 5 Mosquito Repellents

Suppose the researchers are interested in comparing the two treatments that use Deltamethrin (2 and 4) with the two treatments that use Cyfluthrin (3 and 5). Then, the following calculations are made.

$$C_1 = (\mu_2 + \mu_4) - (\mu_3 + \mu_5) \qquad a_1 = 0 \quad a_2 = a_4 = 1 \quad a_3 = a_5 = -1 \qquad n_i = 30 \qquad MSE = 10.308$$

$$\overline{y}_{2.} = 8.133 \quad \overline{y}_{4.} = 6.333 \quad \overline{y}_{3.} = 8.033 \quad \overline{y}_{5.} = 5.367 \qquad \hat{c}_1 = (8.133 + 6.333) - (8.033 + 5.367) = 1.066$$

$$\hat{SE}\{\hat{C}_1\} = \sqrt{10.308\left(\frac{0^2 + 1^2 + (-1)^2 + 1^2 + (-1)^2}{30}\right)} = 1.172$$

For a test ($\alpha = 0.05$) of $H_0 : C_1 = 0$, the test statistic, rejection region and P-value, along with a 95% Confidence Interval for C are given below.

$$TS : t_{C_1} = \frac{1.066}{1.172} = 0.910 \qquad RR : |t_{C_1}| \geq t_{.025,145} = 1.976 \qquad P = 2P\left(t_{145} \geq |0.910|\right) = .364$$

$$95\% \text{ CI for } C : 1.066 \pm 1.976(1.172) \equiv 1.066 \pm 2.316 \equiv (-1.250, 3.382)$$

There is no evidence of any difference between the effects of these two types of repellents. Next, we conduct the F-test, knowing in advance that its conclusion and P-value will be equivalent to 2-tailed t-test performed above (the only difference due to rounding is in third decimal place).

$$SSC_1 = \frac{1.066^2}{\frac{4}{30}} = 8.523 = MSC_1 \quad TS : F_{C_1} = \frac{8.523}{10.308} = 0.827 \quad RR : F_{C_1} \geq F_{.05,1,145} = 3.906$$

$$P = P\left(F_{1,145} \geq 0.827\right) = .365$$

For a second contrast (C_2), without going through all calculations, consider comparing Deltamethrin and Cyfluthrin (each without Odomos: 2 and 3) with Deltamethrin and Cyfluthrin (each with Odomos: 4 and 5). This involves: $a_1 = 0, a_2 = a_3 = 1, a_4 = a_5 = -1$. Note that the standard error of the contrast will be exactly the same as that for contrast $\hat{c}_1$.

$$\hat{c}_2 = 4.466 \quad TS : t_{C_2} = \frac{4.466}{1.172} = 3.811 \quad P = 2P\left(t_{145} \geq |3.811|\right) = .0002 \qquad 95\% \text{ CI: } 4.466 \pm 2.316 \equiv (2.150, 6.782)$$

There is evidence that the combined mean is higher without Odomos than with Odomos. Since low values are better (mosquito contacts), Odomos as an additive to the two chemicals (individually) is better than no additive to the two chemicals individually. The F-test is given below. The R output that follows extends the calculations made in Example 7.2.

$$SSC_2 = \frac{4.466^2}{\frac{4}{30}} = 149.59 = MSC_2 \quad TS: F_{C_2} = \frac{149.59}{10.308} = 14.51 \quad P = P\left(F_{1,145} \geq 14.51\right) = .0005$$

R Output

```
## Output

> round(contrast.out, 4)
     Estimate Std Err      t 2P(>|t|)      LB     UB Sum Sq      F  P(>F)
[1,]    1.066  1.1724 0.9093   0.3647 -1.2511 3.3831 8.5227 0.8268 0.3647
> round(contrast.out, 4)
     Estimate Std Err      t 2P(>|t|)     LB     UB   Sum Sq       F P(>F)
[1,]    4.466  1.1724 3.8094    2e-04 2.1489 6.7831 149.5887 14.5117 2e-04
```

∇

A special class of contrasts are **orthogonal contrasts**. Two contrasts are orthogonal if the sum of the products of their a_i coefficients, divided by the sample sizes n_i, is 0. This concept is shown below.

$$C_1 = \sum_{i=1}^{k} a_{1i}\mu_i \qquad C_2 = \sum_{i=1}^{k} a_{2i}\mu_i \quad C_1 \text{ and } C_2 \text{ are orthogonal if } \quad \sum_{i=1}^{k} \frac{a_{1i}a_{2i}}{n_i} = 0$$

Note that if the sample sizes are all equal (balanced design), this simplifies to $\sum_{i=1}^{k} a_{1i}a_{2i} = 0$. The two contrasts in Example 7.3 are orthogonal (check this). If there are k treatments, and $k-1$ degrees of freedom for Treatments, any $k-1$ pairwise orthogonal contrasts' sums of squares will add up to the Treatment sum of squares. That is, SST can be decomposed into the sums of squares for the $k-1$ contrasts. The decomposition is not unique, there may be various sets of orthogonal contrasts.

Example 7.4: Comparison of 5 Mosquito Repellents

Consider these two other contrasts.

- (D versus C without O) vs (D versus C with Odomos): $C_3 = (\mu_2 - \mu_3) - (\mu_4 - \mu_5) = \mu_2 - \mu_3 - \mu_4 + \mu_5$
- Odomos only versus the four other treatments: $C_4 = 4\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5$

Table 7.4 gives the contrast coefficients for these four contrasts. For all six pairs, $\sum_{i=1}^{k} a_{ji}a_{j'i} = 0$, $j \neq j'$. Also given are the estimates, standard errors, t-tests, 95% Confidence Intervals, Sums of Squares and F-statistics.

Treatment (i)	C_1 ($j=1$)	C_2 ($j=2$)	C_3 ($j=3$)	C_4 ($j=4$)	$\overline{y}_{i.}$
1	0	0	0	4	7.900
2	1	1	1	-1	8.133
3	-1	1	-1	-1	8.033
4	1	-1	-1	-1	6.333
5	-1	-1	1	-1	5.367
$\hat{c}_j$	1.066	4.466	-0.866	3.734	
$\hat{SE}\{\hat{C}_j\}$	1.172	1.172	1.172	2.621	
t_{C_j}	0.909	3.809	-0.739	1.424	
P-value	.3642	.0002	.4613	.1565	
95% CI	(-1.251,3.383)	(2.149,6.783)	(-3.183,1.451)	(-1.447,8.915)	
SSC_j	8.523	149.589	5.625	20.914	
F_{C_j}	0.827	14.512	0.546	2.029	

Table 7.4: Orthogonal Contrasts for the Mosquito Repellent study

From Table 7.3, see that the Treatment sum of squares is $SST = 184.650$. As these four contrasts are pairwise orthogonal, their sums of squares add up to SST: $8.523 + 149.589 + 5.625 + 20.914 = 184.650$. Note that virtually all of the differences among the treatments (based on this set of contrasts) is contrast 2, comparing the average of D and C without O versus the average of D and C with O. The commands for Contrasts 3 and 4 are identical as that for Example 7.3 (with changes to the a^s), and are not included here.

∇

Bonferroni Method of Multiple Comparisons

The Bonferroni method is used in many situations and is based on the following premise: If there are c comparisons to be made simultaneously, and desire to be $(1-\alpha_E)\,100\%$ confident that all are correct, each comparison should be made at a higher level of confidence (lower probability of type I error). If individual comparisons are made at $\alpha_I = \alpha_E/c$ level of significance, there is an overall error rate no larger than α_E. This method is conservative and can run into difficulties (low power) as the number of comparisons gets very large. The general procedures for simultaneous tests and Confidence Intervals are as follow in terms of comparing pairs of treatment means.

$$\text{Define:} \qquad B_{ii'} = t_{\alpha_E/(2c),\nu}\hat{SE}\{\overline{Y}_{i.} - \overline{Y}_{i'.}\} = t_{\alpha_E/(2c),\nu}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)} \qquad i<i'$$

Conclude $\mu_i \neq \mu_{i'}$ if $|\overline{y}_{i.} - \overline{y}_{i'}| \geq B_{ii'}$ Simultaneous $(1-\alpha_E)\,100\%$ CI's for $\mu_i - \mu_{i'}$: $(\overline{y}_{i.} - \overline{y}_{i'.}) \pm B_{ii'}$ where $t_{\alpha_E/(2c),\nu}$, with ν being the error degrees of freedom, $\nu = n_. - k$ for the Completely Randomized Design, is obtained from the Bonferroni t–table (see chapter powerpoint slides) or from statistical packages or EXCEL.

Tukey Method for All Pairwise Comparisons

Various methods have been developed to handle all possible comparisons and keep the overall error rate at α_E, including the widely reported Bonferroni procedure described above. Another commonly used procedure is Tukey's Honest Significant Difference method, which is more powerful than the Bonferroni method (but more limited in its applicability). Statistical computer packages can make these comparisons automatically. Tukey's method can be used for tests and confidence intervals for all pairs of treatment means simultaneously. If there are k treatments, their will be $\frac{k(k-1)}{2}$ such tests or intervals. The general forms, allowing for different sample sizes for treatments i and i' are as follow (the unequal sample size case is referred to as the "Tukey-Kramer" method). The procedure makes use of the **Studentized Range Distribution** with critical values, $q_{\alpha_E,k,\nu}$, indexed by the number of treatments (k) and error degrees of freedom $\nu = n_. - k$ for the Completely Randomized Design. The R functions **qtukey** and **ptukey** in R give quantiles and probabilities for the distribution. A table of critical values for $\alpha_E = 0.05$ is given in this chapter's powerpoint slides.

$$\text{Define:} \qquad HSD_{ii'} = \frac{q_{\alpha_E,k,\nu}}{2}\hat{SE}\{\overline{Y}_{i.} - \overline{Y}_{i'.}\} = \frac{q_{\alpha_E,k,\nu}}{2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)} \quad i < i'$$

Conclude $\mu_i \neq \mu_{i'}$ if $|\overline{y}_{i.} - \overline{y}_{i'.}| \geq HSD_{ii'}$ Simultaneous $(1-\alpha_E)\,100\%$ CI's for $\mu_i - \mu_{i'}$: $(\overline{y}_{i.} - \overline{y}_{i'.}) \pm HSD_{ii'}$

When the sample sizes are equal ($n_i = n_{i'}$), the formula for $HSD_{ii'}$ can be simplified as follows.

$$HSD_{ii'} = q_{\alpha_E,k,\nu}\sqrt{MSE\left(\frac{1}{n_i}\right)} \quad i < i'$$

Example 7.7: Comparison of 5 Mosquito Repellents

The Bonferroni and Tukey methods are used to obtain simultaneous 95% CI's for each difference in mean mosquito contacts. The general form for the Bonferroni simultaneous 95% CI's (with $c = 5(4)/2 = 10$ and $\nu = 150 - 4 = 145$)) is given below. Recall that $MSE = 10.308$ and $n_i = 30$ for each treatment.

$$B_{ii'} = t_{.05/(2(10)),145}\sqrt{10.308\left(\frac{1}{30}+\frac{1}{30}\right)} = 2.851(0.829) = 2.363 \qquad \text{Simultaneous 95\% CIs: } (\overline{y}_{i.} - \overline{y}_{i'.}) \pm 2.363$$

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

$$HSD_{ii'} = q_{0.05,5,145}\sqrt{10.308\left(\frac{1}{30}\right)} = 3.907(0.586) = 2.290 \qquad \text{Simultaneous 95\% CIs: } (\overline{y}_{i.} - \overline{y}_{i'.}) \pm 2.290$$

The corresponding confidence intervals are given in Table 7.5.

Based on the intervals in Table 7.5, it can be concluded that treatments 1 (Odomos) and 5 (Cyfluthrin + Odomos) are significantly different, as are treatments 2 (Deltamethrin) and 5, and treatments 3 (Cyfluthrin) and 5.

While it is easy to write a function in R to conduct the Bonferroni method, there does not appear an easy "follow up" to the ANOVA. There is an easy one for Tukey's honest significant difference method, the **TukeyHSD** function. Note that R takes the mean with the higher subscript minus the mean with the lower subscript.

		Simultaneous 95% CI's	
Comparison	$\overline{y}_{i.} - \overline{y}_{i'.}$	Bonferroni	Tukey
1 vs 2	$7.900 - 8.133 = -0.233$	$(-2.596, 2.130)$	$(-2.523, 2.057)$
1 vs 3	$7.900 - 8.033 = -0.133$	$(-2.496, 2, 230)$	$(-2.423, 2.157)$
1 vs 4	$7.900 - 6.333 = 1.567$	$(-0.796, 3.930)$	$(-0.723, 3.857)$
1 vs 5	$7.900 - 5.367 = 2.533$	$(0.170, 4.896)$	$(0.243, 4.823)$
2 vs 3	$8.133 - 8.033 = 0.100$	$(-2.263, 2.463)$	$(-2.190, 2.390)$
2 vs 4	$8.133 - 6.333 = 1.800$	$(-0.563, 4.163)$	$(-0.490, 4.090)$
2 vs 5	$8.133 - 5.367 = 2.766$	$(0.403, 5.129)$	$(0.476, 5.056)$
3 vs 4	$8.033 - 6.333 = 1.700$	$(-0.663, 4.063)$	$(-0.590, 3.990)$
3 vs 5	$8.033 - 5.367 = 2.666$	$(0.303, 5.029)$	$(0.376, 4.956)$
4 vs 5	$6.333 - 5.367 = 0.966$	$(-1.397, 3.329)$	$(-1.324, 3.256)$

Table 7.5: Bonferroni and Tukey multiple comparisons for the mosquito repellent study

R Commands and Output

```
## Commands

### Tukey follow-up to 1-Way ANOVA
mp <- read.csv("http://www.stat.ufl.edu/~winner/data/mosquito_patch.csv")
attach(mp); names(mp)

trt.mosq <- factor(trt.mosq)
mosq.mod1 <- aov(y.mosq ~ trt.mosq)
anova(mosq.mod1)
TukeyHSD(mosq.mod1, "trt.mosq")

### Output
> TukeyHSD(mosq.mod1, "trt.mosq")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ trt.mosq)

$trt.mosq
      diff       lwr        upr     p adj
2-1  0.233 -2.056985  2.5229848 0.9986197
3-1  0.133 -2.156985  2.4229848 0.9998497
4-1 -1.567 -3.856985  0.7229848 0.3272067
5-1 -2.533 -4.822985 -0.2430152 0.0221285
3-2 -0.100 -2.389985  2.1899848 0.9999517
4-2 -1.800 -4.089985  0.4899848 0.1965250
5-2 -2.766 -5.055985 -0.4760152 0.0093768
4-3 -1.700 -3.989985  0.5899848 0.2474716
5-3 -2.666 -4.955985 -0.3760152 0.0136760
5-4 -0.966 -3.255985  1.3239848 0.7710691

> bon.ci(0.05, y.mosq, trt.mosq)
      Trt i Trt i'   Diff Lower Bound Upper Bound p adjusted
 [1,]     1      2 -0.232      -2.595       2.130      1.000
 [2,]     1      3 -0.132      -2.495       2.231      1.000
 [3,]     1      4  1.567      -0.795       3.930      0.606
 [4,]     1      5  2.534       0.171       4.896      0.027
 [5,]     2      3  0.100      -2.262       2.463      1.000
 [6,]     2      4  1.800      -0.563       4.162      0.315
 [7,]     2      5  2.766       0.403       5.129      0.011
 [8,]     3      4  1.699      -0.663       4.062      0.421
 [9,]    3     5  2.666       0.303       5.028      0.016
[10,]    4     5  0.966      -1.396       3.329      1.000
```

∇

7.2 Randomized Block Design (RBD) For Studies Based on Matched Units

In crossover designs (aka within subjects designs), each unit or subject receives each treatment. In these cases, units are referred to as **blocks**. In other studies, units or subjects may be matched based on external criteria. The notation for the Randomized Block Design (RBD) is very similar to that of the CRD, with additional elements. The model we are assuming here is written as follows.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} = \mu_i + \beta_j + \epsilon_{ij} \quad i = 1, \ldots, k; \quad j = 1, \ldots, b$$

Here, μ represents the overall mean measurement, α_i is the (**fixed**) effect of the i^{th} treatment, β_j is the (typically **random**) effect of the j^{th} block, and ϵ_{ij} is a random error component that can be thought of as the variation in measurements 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. Blocks can be fixed or random, typically they are random.

7.2.1 Test Based on Normally Distributed Data

When the (random) block effects (β_j) and random error terms (ϵ_{ij}) are independent and normally distributed, an F–test is conducted that is similar to that described for the Completely Randomized Design, but with an extra source of variation. If blocks are fixed, the analysis is the same. The notation used is as follows.

$$\begin{aligned}
\overline{y}_{i.} &= \frac{\sum_{j=1}^{b} y_{ij}}{b} \\
\overline{y}_{.j} &= \frac{\sum_{i=1}^{k} y_{ij}}{k} \\
n_{.} &= b \cdot k \\
\overline{y}_{..} &= \frac{\sum_{i=1}^{k} \sum_{j=1}^{b} y_{ij}}{bk} \\
TSS &= \sum_{i=1}^{k} \sum_{j=1}^{b} \left(y_{ij} - \overline{y}_{..}\right)^2 \\
SST &= \sum_{i=1}^{k} b \left(\overline{y}_{i.} - \overline{y}_{..}\right)^2 \\
SSB &= \sum_{j=1}^{b} k \left(\overline{y}_{.j} - \overline{y}_{..}\right)^2
\end{aligned}$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{b}\left(y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..}\right)^2$$

Note that the Analysis of Variance simply has added items representing the block means $\left(\overline{y}_{.j}\right)$ 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 not explained by either differences among treatment or block means (SSE). Also, note that $SSE = TSS - SST - SSB$.

		ANOVA		
Source of Variation	Sum of Squares	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	TSS	$bk-1$		

Table 7.6: 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 often are not interested in testing for differences among blocks, since we expect there to be differences among them (that's why the design was set up this way), and they were just a random sample from a population of such experimental units. However, in some cases, estimating the unit to unit (subject to subject) variance component is of interest. 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: $P(F_{k-1,(b-1)(k-1)} \geq F_{obs})$

The procedures to make comparisons among means are very similar to the methods used for the CRD. In each formula described previously for Bonferroni's, and Tukey's methods, n_i is replaced by b, when making comparisons among treatment means, and $\nu = (b-1)(k-1)$ is the error degrees of freedom.

Example 7.10: Comparison of 3 Methods for Estimating Value of Wood Logs

A study compared 3 methods of assessing the lumber value of logs (Lin and Wang (2012), [39]). The $k = 3$ treatments the authors compared was the actual sawmill value of the log, a value based on a heuristic programming algorithm, and a value based on a dynamic programming algorithm. Each "treatment" was measured on $b = 30$ logs (acting as the blocks). The goal was to compare the 3 treatments at valuating the logs. Data are given in Table 7.7. A crude interaction plot is given in Figure 7.4, which plots the valuation versus log ID, with seperate lines for the three methods.

Log ID	Actual	Heuristic	Dynamic	LogMean
1	17.67	20.83	21.03	19.8433
2	31.76	35.05	34.24	33.6833
3	30.77	33.60	34.87	33.0800
4	40.27	42.52	42.89	41.8933
5	33.51	35.06	36.48	35.0167
6	23.07	25.37	26.34	24.9267
7	21.33	21.95	23.00	22.0933
8	26.28	28.07	28.69	27.6800
9	28.89	31.94	32.49	31.1067
10	18.46	19.14	21.76	19.7867
11	35.61	38.18	39.87	37.8867
12	23.15	25.67	27.22	25.3467
13	18.03	19.58	20.70	19.4367
14	28.22	30.89	30.05	29.7200
15	20.33	21.36	21.62	21.1033
16	12.42	13.01	14.02	13.1500
17	21.90	24.52	25.06	23.8267
18	36.16	38.12	38.86	37.7133
19	13.73	14.74	15.12	14.5300
20	15.74	17.96	18.00	17.2333
21	19.22	20.69	20.83	20.2467
22	17.12	19.12	19.31	18.5167
23	15.21	16.42	16.63	16.0867
24	22.03	23.58	24.24	23.2833
25	31.22	32.66	32.90	32.2600
26	25.69	28.39	28.81	27.6300
27	29.25	31.63	30.72	30.5333
28	32.77	33.29	35.87	33.9767
29	31.88	34.79	34.82	33.8300
30	24.54	26.23	26.54	25.7700
Trt Mean	24.8743	26.8120	27.4327	26.3730

Table 7.7: Log Values for 3 Methods of Valuation

$$TSS = (17.67 - 26.3730)^2 + \cdots + (26.54 - 26.3730)^2 = 5170.073 \qquad df = 30(3) - 1 = 89$$

$$SST = 30\left[(24.8743 - 26.3730)^2 + (26.8120 - 26.3730)^2 + (27.4327 - 26.3730)^2\right] = 106.8536 \qquad df_T = 3{-}1 = 2$$

$$SSB = 3\left[(19.8433 - 26.3730)^2 + \cdots + (25.7700 - 26.3730)^2\right] = 5042.772 \qquad df_B = 30 - 1 = 29$$

$$SSE = (17.67 - 19.8433 - 24.8743 + 26.3730)^2 + \cdots + (26.54 - 25.7700 - 26.4327 + 26.3730)^2 =$$
$$5170.073 - 106.8536 - 5042.772 = 20.448 \qquad df_E = (30 - 1)(3 - 1) = 58$$

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

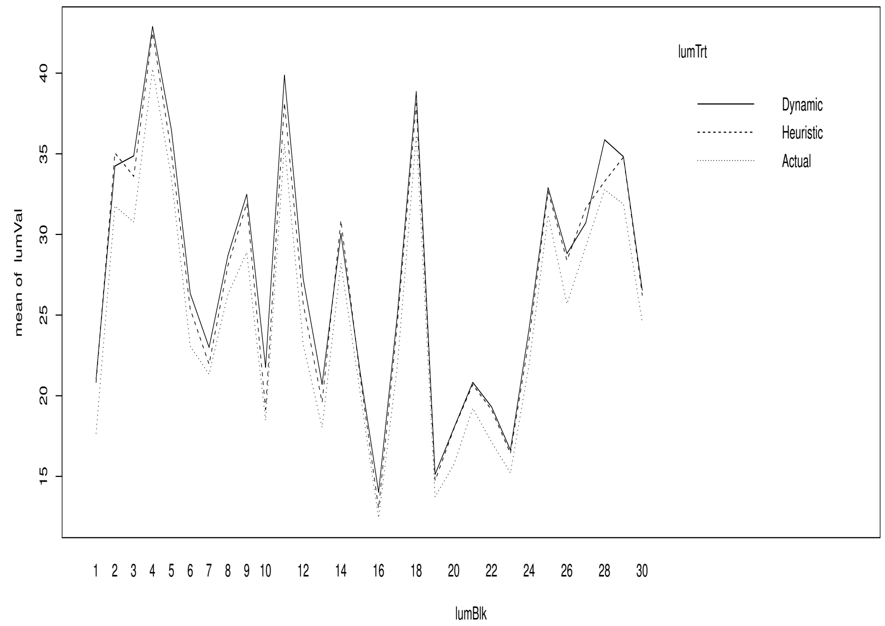

Figure 7.4: Plot of valuation versus log ID, with separate lines for valuation method

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_{obs}
TREATMENTS	106.854	2	53.427	151.546
BLOCKS	5042.772	29	173.889	
ERROR	20.448	58	0.353	
TOTAL	5170.073	89		

Table 7.8: Analysis of Variance table for log valuation data (RBD)

1. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad (\mu_1 = \mu_2 = \mu_3)$ (No differences among valuation method means)
2. $H_A :$ Not all α_i are 0 (Valuation differences exist)
3. T.S. $F_{obs} = \frac{MST}{MSE} = 151.546$
4. R.R.: $F_{obs} \geq F_{\alpha,k-1,(b-1)(k-1)} = F_{0.05,2,58} = 3.156$
5. P-value: $P(F_{2,58} \geq F_{obs}) = P(F_{2,58} \geq 151.546) = 0.0000$

Tukey's method is used to determine which valuation methods differ significantly. Recall that for Tukey's method, simultaneous confidence intervals of the form given below are computed, with k being the number of treatments (k=3), b being the number of blocks, and n_i the number of measurements per valuation method ($n_i = b = 30$).

$$(\overline{y}_{i.} - \overline{y}_{i'.}) \pm q_{\alpha,k,(b-1)(k-1)}\sqrt{MSE\left(\frac{1}{n_i}\right)} \quad \Longrightarrow \quad (\overline{y}_{i.} - \overline{y}_{i'.}) \pm 3.402\sqrt{0.353\left(\frac{1}{30}\right)} \quad \Longrightarrow (\overline{y}_{i.} - \overline{y}_{i'.}) \pm 0.369$$

The corresponding simultaneous 95% confidence intervals and conclusions are given in Table 7.9. Conclude that Actual sawmill valuation is significantly lower than Heuristic, which is significantly lower than Dynamic.

Comparison	$\overline{y}_{i.} - \overline{y}_{i'.}$	CI	Conclusion
Actual vs Heuristic	$24.874 - 26.812 = -1.938$	$(-2.307, -1.569)$	$\mu_A < \mu_H$
Actual vs Dynamic	$24.874 - 27.433 = -2.559$	$(-2.928, -2.190)$	$\mu_A < \mu_D$
Heuristic vs Dynamic	$26.812 - 27.433 = -0.621$	$(-0.990, -0.252)$	$\mu_H < \mu_D$

Table 7.9: Tukey's simultaneous 95% CI's for wood log valuation data (RBD)

Note that to run this in R, it is necessary to have a separate row for each observation, along with a treatment ID and block ID.

R Commands and Output

```
## Commands

saw <- read.csv("http://www.stat.ufl.edu/~winner/data/sawmill1.csv")
attach(saw); names(saw)
lumTrt <- factor(lumTrt)
lumBlk <- factor(lumBlk)
levels(lumTrt) <- c("Actual", "Heuristic", "Dynamic")

saw.mod1 <- aov(lumVal ~ lumTrt + lumBlk)
anova(saw.mod1)
TukeyHSD(saw.mod1, "lumTrt")

interaction.plot(lumBlk, lumTrt, lumVal)

## Output
> anova(saw.mod1)
Analysis of Variance Table
Response: lumVal
          Df Sum Sq Mean Sq F value    Pr(>F)
```

```
lumTrt     2  106.8  53.424  151.53 < 2.2e-16 ***
lumBlk    29 5042.8 173.889  493.21 < 2.2e-16 ***
Residuals 58   20.4   0.353
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> TukeyHSD(saw.mod1, "lumTrt")
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = lumVal ~ lumTrt + lumBlk)
$lumTrt
                       diff       lwr       upr    p adj
Heuristic-Actual  1.9376667 1.5689056 2.3064277 0.000000
Dynamic-Actual    2.5583333 2.1895723 2.9270944 0.000000
Dynamic-Heuristic 0.6206667 0.2519056 0.9894277 0.000449
```

7.3 Factorial Designs

In many cases, the research is interested in whether multiple factors have effects on responses, and whether the effects of the individual factor levels depend on the levels of the remaining factor(s). We will consider only the case of two treatment factors: A with a levels and B with b levels. Further, we will only consider the case where there are n observation within each of the ab treatments (the crossing of the levels of factors A and B).

In this section, models with two factors are considered. 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 follows.

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, n \quad \epsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

Here 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 units within each treatment combination will vary, as well as if the same unit were measured repeatedly, its measurements would vary. Here, we consider the model where both factors A and B are **fixed**, with all levels of interest present in the study.

Note that an interaction represents the fact that the effect of a particular level of factor A depends on the level of factor B, and vice versa. As before, we assume that ϵ_{ijk} is normally distributed with mean 0 and variance σ^2.

When factors A and B are fixed, the effects are unknown parameters to be estimated. One common way of parameterizing the model is as follows.

$$E\left\{Y_{ijk}\right\} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad V\left\{Y_{ijk}\right\} = \sigma^2 \quad \sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$$

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 here.

$$\begin{aligned}
\overline{y}_{ij.} &= \frac{\sum_{k=1}^{n} y_{ijk}}{n} \\
\overline{y}_{i..} &= \frac{\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}}{bn} \\
\overline{y}_{.j.} &= \frac{\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}}{an} \\
\overline{y}_{...} &= \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}}{abn} \\
n_{..} &= a \cdot b \cdot n \\
s_{ij}^2 &= \frac{\sum_{k=1}^{n}\left(y_{ijk} - \overline{y}_{ij.}\right)^2}{n-1} \\
TSS &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{...})^2 \\
SSA &= bn\sum_{i=1}^{a}(\overline{y}_{i..} - \overline{y}_{...})^2 \\
SSB &= an\sum_{j=1}^{b}(\overline{y}_{.j.} - \overline{y}_{...})^2 \\
SSAB &= n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 \\
SSE &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2
\end{aligned}$$

The error sum of squares can also be computed from the within cell standard deviations, which is helpful as many research articles provide the treatment means and standard deviations.

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2 = (n-1)\sum_{i=1}^{a}\sum_{j=1}^{b} s_{ij}^2$$

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 7.10, assuming that n measurements are made at each combination of levels of the two factors.

The tests for interactions and for effects of factors A and B involve the three F–statistics, and can be conducted as follow. Note that under each of the three null hypotheses, the corresponding expected mean square in the numerator simplifies to $\sigma^2 = E\{MSE\}$.

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$Pr(>F)$
FACTOR A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSE}$	$P\left(F \geq F_A\right)$
FACTOR B	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$	$P\left(F \geq F_B\right)$
INTERACTION AB	$(a-1)(b-1)$	$SSAB$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$	$P\left(F \geq F_{AB}\right)$
ERROR	$ab(n-1)$	SSE	$MSE = \frac{SSE}{ab(n-1)}$		
TOTAL	$abn-1$	TSS			

Table 7.10: The Analysis of Variance Table for a Balanced 2-Factor Factorial Design with Fixed Effects

1. $H_0^{AB} : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{ab} = 0$ (No interaction effect).
2. H_A^{AB} : Not all $(\alpha\beta)_{ij} = 0$ (Interaction effects exist)
3. T.S. $F_{AB} = \frac{MSAB}{MSE}$
4. R.R.: $F_{AB} \geq F_{\alpha,(a-1)(b-1),ab(n-1)}$
5. p-value: $P\left(F_{(a-1)(b-1),ab(n-1)} \geq F_{AB}\right)$

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

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

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

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

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 7.11: Espresso Foam Index by Temperature and Extraction Pressure

An experiment was conducted to measure foam index (Y) of espresso samples brewed at $a = 3$ levels of temperature (factor A: 75C, 85C, 95C) and $b = 2$ extraction pressures (factor B: 15bar, 20bar) with $n = 9$ replicates per treatment (Masella, et al (2015), [41]). The sample means (standard deviations), temperature means, pressure means, and overall mean are given in Table 7.11. Note that at 75C, Foam Index increases by $135.0 - 113.4 = 21.6$ when pressure increases from 15 to 20. Similarly, the changes are 11.4 and 26.5 for 85C and 95C, respectively. Those are fairly similar increases, but we will test formally for the interaction below.

	Pressure (B)		
Temperature (A)	15bar	20bar	Temp Mean
75C	113.4 (15.0)	135.0 (27.5)	124.20
85C	102.4 (12.2)	113.8 (20.5)	108.10
95C	91.1 (12.9)	117.6 (21.7)	104.35
Press Mean	102.30	122.13	112.22

Table 7.11: Espresso Foam Index Mean (SD) by Temperature and Pressure

The sums of squares are set-up below, as well as the Analysis of Variance in Table 7.12.

$$SSA = 2(9)\left[(124.20 - 112.22)^2 + (108.10 - 112.22)^2 + (104.35 - 112.22)^2\right] = 4004 \qquad df_A = 3 - 1 = 2$$

$$SSB = 3(9)\left[(102.30 - 112.22)^2 + (122.13 - 112.22)^2\right] = 5309 \qquad df_B = 2 - 1 = 1$$

$$SSAB = 9\left[(113.4 - 124.20 - 102.30 + 112.22)^2 + \cdots + (117.6 - 104.35 - 122.13 + 112.22)^2\right] = 534 \qquad df_{AB} = 2(1) = 2$$

$$SSE = (9 - 1)\left[15.0^2 + \cdots + 21.7^2\right] = 17501 \qquad df_E = 3(2)(9 - 1) = 48$$

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$Pr(> F)$
Temperature (A)	2	4004	$\frac{4004}{2} = 2002$	$\frac{2002}{365} = 5.485$	.0072
Pressure (B)	1	5309	$\frac{5309}{1} = 5309$	$\frac{5309}{365} = 14.545$	.0004
Interaction (AB)	2	534	$\frac{534}{2} = 267$	$\frac{267}{365} = 0.732$	.4862
ERROR	48	17501	$\frac{17501}{48} = 365$		
TOTAL	53	27348			

Table 7.12: Analysis of Variance Table for Espresso Foam Index experiment

Thus, there are significant main effects for Temperature and Pressure, but the interaction is not significant. The R code for the analysis and output (including Tukey's comparisons among Temperature and

Pressure effects individually is given below). Note, the R program uses more internal decimal places than the computations above. Foam Index is significantly higher at 75C than 85C and 95C. The Foam Index at 85C and 95C are not significantly different. It is significantly higher at 20bar than 15bar.

An interaction plot is given in Figure 7.5.

R Commands and Output

```
## Commands
espresso1 <- read.csv("http://www.stat.ufl.edu/~winner/data/espresso2.csv")
attach(espresso1); names(espresso1)

tempC <- factor(tempC)
prssBar <- factor(prssBar)

fi.mod <- aov(foamIndx ~ tempC * prssBar)
summary(fi.mod)
TukeyHSD(fi.mod, "tempC")
TukeyHSD(fi.mod, "prssBar")

interaction.plot(tempC, prssBar, foamIndx)

## Output
> fi.mod <- aov(foamIndx ~ tempC * prssBar)
> summary(fi.mod)
              Df Sum Sq Mean Sq F value   Pr(>F)
tempC          2   4004    2002   5.491 0.007123 **
prssBar        1   5310    5310  14.564 0.000388 ***
tempC:prssBar  2    534     267   0.732 0.486075
Residuals     48  17501     365
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> TukeyHSD(fi.mod, "tempC")
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = foamIndx ~ tempC * prssBar)
$tempC
            diff       lwr        upr     p adj
85-75 -16.100556 -31.49407 -0.7070449 0.0385233
95-75 -19.850000 -35.24351 -4.4564893 0.0084704
95-85  -3.749444 -19.14296 11.6440662 0.8266145

> TukeyHSD(fi.mod, "prssBar")
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = foamIndx ~ tempC * prssBar)
$prssBar
          diff      lwr      upr     p adj
20-15 19.83333 9.384174 30.28249 0.0003876
```

∇

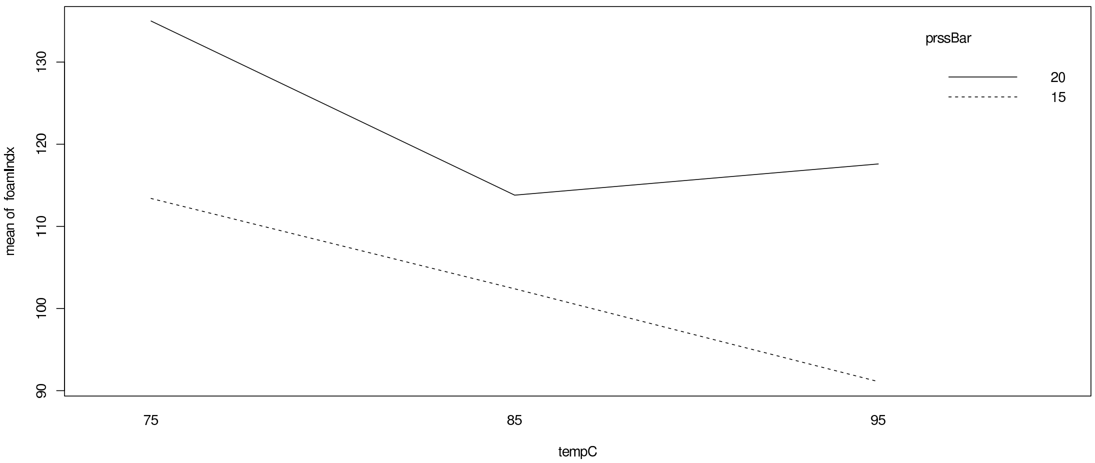

Figure 7.5: Mean Foam Index for espresso by Temperature and Pressure

Chapter 8

Categorical Data Analysis

Recall that variables can be categorical or numeric. Chapters 4, 6, and 7 dealt with making inferences for quantitative responses. In this chapter, methods commonly used to analyze data when the response variable is categorical are introduced. First, consider estimating and testing proportions corresponding to a single binomial (2 possible outcomes) or multinomial ($k > 2$ possible outcomes) variable. Then, cases of testing for associations among two or more categorical variables are covered.

8.1 Inference Concerning a Single Variable

A single variable can have two levels, and counts are modeled by the Binomial distribution, or it can have $k > 2$ levels and counts are modeled by the Multinomial distribution. Note that the Binomial is a special case of the Multinomial, however there are many methods that apply strictly to binary outcomes.

8.1.1 Variables with Two Possible Outcomes

In the case of a binary variable, the goal is typically to estimate the true underlying probability of success, π. The sample proportion $\hat{\pi} = Y/n$ from a binomial experiment with n trials and Y successes has a sampling distribution with mean π and standard error $\sqrt{\pi(1-\pi)/n}$. In large samples, the sampling distribution is approximately normal. One commonly used rule of thumb is that $n\pi \geq 5$ and $n(1-\pi) \geq 5$. When estimating π, the estimated standard error must be used, where π is replaced with $\hat{\pi}$. Note that the standard error is maximized for a given n when $\pi = 1 - \pi = 0.5$, so a conservative case uses $\pi = 0.5$ in the standard error. The large-sample $(1-\alpha)100\%$ Confidence Interval for π and the sample size needed for a given margin of error, E, are given below.

$$(1-\alpha)100\%\text{CI for } \pi : \hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}\,(1-\hat{\pi})}{n}} \qquad E = z_{\alpha/2}\sqrt{\frac{\pi\,(1-\pi)}{n}} \quad \Rightarrow \quad n = \frac{z_{\alpha/2}^2\pi(1-\pi)}{E^2}$$

In small samples, the large-sample normal approximation does poorly in terms of coverage rates for π. It has been seen that making an adjustment to the success count and the sample size performs well. This is referred to as the **Wilson-Agresti-Coull** method. Let y be the observed number of successes in the n trials, then the Confidence Interval is obtained as follows. Note that since $z_{.025} = 1.96 \approx 2$, for a 95% Confidence Interval, this can be thought of as adding 2 Successes and 2 Failures to the observed data (Agresti and Coull (1998), [2]).

$$\tilde{y} = y + 0.5z_{\alpha/2}^2 \quad \tilde{n} = n + z_{\alpha/2}^2 \quad \tilde{\pi} = \frac{\tilde{y}}{\tilde{n}} \qquad (1-\alpha)100\%\text{CI for } \pi : \tilde{\pi} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}\,(1-\tilde{\pi})}{\tilde{n}}}$$

Example 8.1: Estimating Shaquille O'Neal's Free Throw Success Probability

During Shaquille O'Neal's NBA regular season career, he took 11252 free throws, successfully making 5935, so that $\pi = 5935/11252 = .5275$. Stringing out his within game free throw attempts into a sequence of 1^s and 0^s, and taking 100000 random samples of size $n = 10$, the coverage rates for the two methods are 88.5% for the "traditional" large-sample method and 94.8% for the Wilson-Agresti-Coull method. For the small-sample case, the adjustment performs very well. When the samples are of size $n = 30$, the coverage rates are 93.2% and 95.9%, respectively. For samples of size $n = 100$, they are 94.3% and 95.3%, respectively.

R Output

```
## Output

> round(ft.out, 4)
           pi pi-hat cover pi-tilde cover pi-hat mean width pi-tilde mean width
n=10  0.5275       0.8836         0.9455            0.5842              0.5120
n=30  0.5275       0.9321         0.9588            0.3510              0.3319
n=100 0.5275       0.9436         0.9538            0.1946              0.1911
```

For the first sample of size $n = 10$, $y = 7$ free throws were successes and the following calculations are used to obtain the 95% Confidence Intervals for π.

$$\hat{\pi} = \frac{7}{10} = 0.7 \qquad S\hat{E}\{\hat{\pi}\} = \sqrt{\frac{0.7\,(1-0.7)}{10}} = 0.145 \qquad 0.70 \pm 1.96(0.145) \equiv 0.70 \pm 0.284 \equiv (0.416, 0.984)$$

$$\tilde{y} = 7 + 0.5(1.96)^2 = 8.92 \quad \tilde{n} = 10 + (1.96)^2 = 13.84 \quad \tilde{\pi} = \frac{8.92}{13.84} = 0.645 \quad \sqrt{\frac{0.645\,(1-0.645)}{13.84}} = 0.129$$

$$0.645 \pm 1.96(0.129) \equiv 0.645 \pm 0.253 \equiv (0.392, 0.898)$$

Both intervals contain $\pi = 0.5275$.

∇

A large-sample test of whether $\pi = \pi_0$ can also be conducted. For instance, a test may be whether a majority of people favor a political candidate or referendum, or whether a defective rate is below some tolerance level.

$$\text{2-tailed test: } H_0 : \pi = \pi_0 \quad H_A : \pi \neq \pi_0 \quad TS : z_{obs} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \quad RR : |z_{obs}| \geq z_{\alpha/2} \quad P = 2P\left(Z \geq |z_{obs}|\right)$$

$$\text{Upper-tailed test: } H_0 : \pi \leq \pi_0 \quad H_A : \pi > \pi_0 \quad TS : z_{obs} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \quad RR : z_{obs} \geq z_{\alpha} \quad P = P\left(Z \geq z_{obs}\right)$$

$$\text{Lower-tailed test: } H_0 : \pi \geq \pi_0 \quad H_A : \pi < \pi_0 \quad TS : z_{obs} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \quad RR : z_{obs} \leq -z_{\alpha} \quad P = P\left(Z \leq z_{obs}\right)$$

An exact test can be conducted by use of the binomial distribution and statistical packages by obtaining the exact probability that the count could be more extreme than the observed count y under the null hypothesis. See the examples below.

Example 8.2: NBA Point Spread and Over/Under Outcomes for 2014-2015 Regular Season

For each NBA game there is a "point spread" for bettors to wager on. If the home team is favored to win the game by 5 points, it must win by 6 or more points to "cover the spread," if it loses the game or wins by less than 5 points, the home team loses the bet, and if it wins by 5 points, the bet is a tie or "push." For the 2014-2015 regular season games, the home team covered the spread in 588 games, failed to cover the spread in 615 games, and "tied" the spread in 27 games. We treat these games as a sample of the infinite population of games that could be played among NBA teams, and eliminate the 27 "pushes." The test is whether the true underlying probability that the home team covers is 0.50. Otherwise bettors could have an advantage over bookmakers. $H_0 : \pi = 0.50$ versus $H_A : \pi \neq 0.50$.

$$y = 588 \qquad n = 615 + 588 = 1203 \qquad \hat{\pi} = \frac{588}{1203} = 0.4888 \qquad SE\left\{\hat{\pi}\right\}_{H_0} = \sqrt{\frac{0.5(1-0.5)}{1203}} = 0.0144$$

$$z_{obs} = \frac{0.4888 - 0.5}{0.0144} = -0.78 \qquad P = 2P(Z \geq 0.78) = 2(0.2177) = 0.4354$$

There is no evidence of a "bias" (positive or negative) in terms of the home team performance against the spread. An exact test is given here. Under the null hypothesis, the expected value of Y is $n\pi_0 = 1203(0.5) = 601.5$. The observed y is 588, which is 13.5 below its expected value. Had y been 615, it would have been 13.5 above its expected value. The exact 2-tailed P-value is obtained as follows.

$$P = P\left(Y \leq 588 | Y \sim Bin(1203, 0.5)\right) + P\left(Y \geq 615 | Y \sim Bin(1203, 0.5)\right) = 0.22675 + 0.22675 = .4535$$

A similar test can be done for the "Over/Under" bet. Bookmakers set a total score for the two teams, and if the combined points exceed this line the Over wins, if it falls short, the Under wins, and if it ties, it is a "Push." For the Over/Under bet for that season, Under won 633 times, Over won 583 times, and there were 14 Pushes. Again, we eliminate the Pushes, and test $H_0 : \pi = 0.50$ versus $H_A : \pi \neq 0.50$, where π is the probability Over wins.

$$y = 583 \qquad n = 633 + 583 = 1216 \qquad \hat{\pi} = \frac{583}{1216} = 0.4794 \qquad SE\left\{\hat{\pi}\right\}_{H_0} = \sqrt{\frac{0.5(1-0.5)}{1216}} = 0.0143$$

$$z_{obs} = \frac{0.4794 - 0.5}{0.0143} = -1.44 \qquad P = 2P(Z \geq 1.44) = 2(.0749) = 0.1498$$

Again there is no evidence of a bias. An exact P-value is obtained below.

$$P = P\left(Y \leq 583 | Y \sim Bin(1216, 0.5)\right) + P\left(Y \geq 633 | Y \sim Bin(1216, 0.5)\right) = 0.07997 + 0.07997 = 0.1599$$

R Output

```
### Output
> round(cov.out, 4)
     pi(H0)   y    n  pihat SE{H0}       Z   P(Z) P(Exact) SE{pihat}  Lower Upper
[1,]    0.5 588 1203 0.4888 0.0144 -0.7785 0.4363   0.4535    0.0144 0.4605 0.517
> ### Exact Tests
> binom.test(Y.Cov,n.Cov,p=0.5,alternative="two.sided")

        Exact binomial test

data:  Y.Cov and n.Cov
number of successes = 588, number of trials = 1203, p-value = 0.4535
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4601721 0.5174390
sample estimates:
probability of success
             0.4887781
```

∇

8.2 Introduction to Tests for Association for Two Categorical Variables

The data are generally counts of individuals or units, 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 8.1.

		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}	$n_{r.}$
		$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	$n_{..}$

Table 8.1: 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 gender (male, female) or hair color (black, blonde, brown, red). Ordinal variables have levels that do have a distinct ordering such as reviewer's assessment of a movie (negative opinion, mixed opinion, positive opinion).

In this chapter, the following cases are covered.

- 2×2 tables (both variables have two levels)
- Both variables are nominal.

8.3 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 (prospective) and case–control (retrospective) settings.

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 8.2.

		Disease State		
		D (Present)	$\overline{D}$ (Absent)	Total
Exposure	E (Present)	$n_{11} = y_1$	$n_{12} = n_1 - y_1$	$n_{1.} = n_1$
State	$\overline{E}$ (Absent)	$n_{21} = y_2$	$n_{22} = n_2 - y_2$	$n_{2.} = n_2$
Total		$n_{.1} = y_1 + y_2$	$n_{.2} = (n_1 - y_1) + (n_2 - y_2)$	$n_{..} = n_1 + n_2$

Table 8.2: 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 three measures of association commonly estimated and reported. These are the **absolute risk** (aka difference in proportions), 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.

8.3.1 Difference in Proportions: $\pi_1 - \pi_2$

In many studies, the goal is to compare the Success probabilities for two groups. These studies can be based on large samples or small samples, and can be based on independent or paired samples.

For the large sample case, based on independent samples, the estimators $\hat{\pi}_1 = Y_1/n_1$ and $\hat{\pi}_2 = Y_2/n_2$ for the two groups are independent and have sampling distributions that are approximately normal. The relevant results are given below.

$$E\left\{\hat{\pi}_1 - \hat{\pi}_2\right\} = \pi_1 - \pi_2 \quad SE\left\{\hat{\pi}_1 - \hat{\pi}_2\right\} = \sqrt{\frac{\pi_1\left(1-\pi_1\right)}{n_1} + \frac{\pi_2\left(1-\pi_2\right)}{n_2}} \quad \hat{SE}\left\{\hat{\pi}_1 - \hat{\pi}_2\right\} = \sqrt{\frac{\hat{\pi}_1\left(1-\hat{\pi}_1\right)}{n_1} + \frac{\hat{\pi}_2\left(1-\hat{\pi}_2\right)}{n_2}}$$

$$(1-\alpha)100\% \text{ CI for } \pi_1 - \pi_2 : (\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\hat{SE}\left\{\hat{\pi}_1 - \hat{\pi}_2\right\} \quad \equiv \quad (\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1\left(1-\hat{\pi}_1\right)}{n_1} + \frac{\hat{\pi}_2\left(1-\hat{\pi}_2\right)}{n_2}}$$

In terms of testing the hypothesis $H_0 : \pi_1 - \pi_2 = 0$, an adjustment is made to the standard error of $\hat{\pi}_1 - \hat{\pi}_2$. In this case the overall combined proportion of successes is obtained and used in the "pooled" standard error.

$$\hat{\pi} = \frac{y_1 + y_2}{n_1 + n_2} \qquad \hat{SE}_p\left\{\hat{\pi}_1 - \hat{\pi}_2\right\} = \sqrt{\hat{\pi}\left(1-\hat{\pi}\right)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

The test statistic for testing $H_0 : \pi_1 - \pi_2 = 0$ is given below with the usual rules for rejection regions and P-values for 2-tailed and 1-tailed tests.

$$TS : z_{obs} = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{SE}_p\left\{\hat{\pi}_1 - \hat{\pi}_2\right\}} = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}\left(1-\hat{\pi}\right)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

Example 8.6: Risk Taking After Large Financial Losses

An Australian natural experiment considered the effect of large losses on subsequent risk taking behavior (Page, Savage, and Torgler (2014), [45]). The study included a sample of $n_1 = 94$ people who had been effected by the flood in Brisbane during 2011 and a sample of $n_2 = 107$ people who had not been effected. The subjects in the experiment were given the choice between a certain \$10 and a scratch card valued at \$10, but with a maximum prize of \$500,000. The scratch card is considered the "high risk" choice. Of the effected participants, $y_1 = 75$ chose the scratch card, of the uneffected, $y_2 = 53$ chose the scratch card.

$$\hat{\pi}_1 = \frac{75}{94} = .7979 \quad \hat{\pi}_2 = \frac{53}{107} = 0.4953 \quad \hat{\pi}_1 - \hat{\pi}_2 = .7979 - .4953 = .3026 \qquad \hat{\pi} = \frac{75+53}{94+107} = \frac{128}{201} = 0.6368$$

$$\hat{SE}\left\{\hat{\pi}_1 - \hat{\pi}_2\right\} = \sqrt{\frac{.7979(.2021)}{94} + \frac{.4953(.5047)}{107}} = .0637 \qquad .3026 \pm 1.96(.0637) \equiv .3026 \pm .1248 \equiv (.1778, .4274)$$

$$\hat{SE}_p\left\{\hat{\pi}_1 - \hat{\pi}_2\right\} = \sqrt{.6368(.3632)\left[\frac{1}{94} + \frac{1}{107}\right]} = .0680 \qquad z_{obs} = \frac{.3026}{.0680} = 4.451 \quad P = 2P(Z \geq 4.451) \approx 0$$

This provides empirical evidence consistent with prospect theory that states that people adopt risk taking attitudes after losses.

R Commands and Output

```
### Commands

y1 <- 75; n1 <- 94 ## Successes and Total for Group 1 (Affected by Flood)
y2 <- 53; n2 <- 107 ## Successes and Total for Group 2 (Unaffected)

pihat.1 <- y1/n1
pihat.2 <- y2/n2
pihat <- (y1+y2)/(n1+n2)
se.pihat.12 <- sqrt((pihat.1*(1-pihat.1)/n1)+(pihat.2*(1-pihat.2)/n2))
se.pihat.12p <- sqrt(pihat*(1-pihat)*(1/n1+1/n2))
z025 <- qnorm(.975,0,1)

pi12.ci <- (pihat.1-pihat.2) + c(-z025,z025)*se.pihat.12    # 95%CI for pi1-pi2
pi12.z <- (pihat.1-pihat.2)/se.pihat.12p                    # Z_obs for H0:pi1-pi2=0
pi12.p <- 2 * (1-pnorm(abs(pi12.z)))                        # 2-sided P-value

pi12.out <- cbind(y1, y2, n1, n2, pihat.1, pihat.2, pihat, se.pihat.12, pi12.ci[1],
     pi12.ci[2], se.pihat.12p, pi12.z, pi12.p)
colnames(pi12.out) <- c("y1", "y2", "n1", "n2", "pihat1", "pihat2", "pooled",
    "SE{Diff}", "Lower",  "Upper", "SE{(H0)}", "Z", "P-value")
round(pi12.out, 4)

prop.test(c(y1,y2),c(n1,n2),correct=F)

### Output

> round(pi12.out, 4)
     y1 y2 n1  n2 pihat1 pihat2 pooled SE{Diff}  Lower  Upper SE{(H0)}      Z P-value
[1,] 75 53 94 107 0.7979 0.4953 0.6368   0.0637 0.1778 0.4273    0.068 4.4502       0
```

```
>
> prop.test(c(y1,y2),c(n1,n2),correct=F)

        2-sample test for equality of proportions without continuity correction

data:  c(y1, y2) out of c(n1, n2)
X-squared = 19.804, df = 1, p-value = 8.58e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1777846 0.4273059
sample estimates:
   prop 1    prop 2
0.7978723 0.4953271
```

Note that R presents the “Z-test” as a chi-square test (with 1 degree of freedom), $z_{obs}^2 = 4.4502^2 = 19.804$. The P-values are identical for a 2-tailed test.

∇

8.3.2 McNemar’s Test for Paired Designs

When the same units are being observed under both experimental treatments (or units have been matched based on some criteria), McNemar’s test can be used to test for treatment effects. The relevant subjects (pairs) are the ones who respond differently under the two conditions. Counts will appear as in Table 8.3.

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

Table 8.3: Notation for McNemar’s Test

Note that n_{11} is the number of units that 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 the difference in treatment effects. The units that provide information are the n_{12} cases that have the outcome present under treatment 1, and absent under treatment 2; and the n_{21} units that 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) $\Rightarrow$ No Trt effect
- H_A : The probabilities differ (Trt effects - This can be 1-sided also)
- $TS : z_{obs} = \frac{n_{12}-n_{21}}{\sqrt{n_{12}+n_{21}}}$
- $RR : |z_{obs}| \geq z_{\alpha/2}$ (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. The 2-sided z-test, and the chi-square test are mathematically equivalent.

An exact test is based on the binomial distribution. Under the null hypothesis of no treatment effect, the count n_{12} is distributed binomial with $n = n_{12} + n_{21}$ and $\pi = 0.5$. The P-value is computed as follows.

$$H_0 : \pi_1 = \pi_2 \quad H_A : \pi_1 \neq \pi_2 \quad P = 2\min\left[P\left(Y \leq n_{12}\right), P\left(Y \geq n_{12}\right)\right] \quad Y \sim Bin\left(n = n_{12} + n_{21}, \pi = 0.5\right)$$

If trying to demonstrate that $\pi_1 > \pi_2$, we would expect $n_{12} > n_{21}$ and $P = P\left(Y \geq n_{12}\right)$. If the goal is to demonstrate that $\pi_1 < \pi_2$, we would expect $n_{12} < n_{21}$ and $P = P\left(Y \leq n_{12}\right)$.

Example 8.11: Framing of Risky Outcomes

In one of many studies testing prospect theory, subjects were asked to make two decisions regarding risky gambles (Kahneman and Tversky (1984), [34]). The decision choices are given below.

- Decision 1: Choose between (A): a sure gain of $240 and (B): a 25% chance of winning $1000 and 75% chance of winning $0.
- Decision 2: Choose between (C): a sure loss of $750 and (D): a 75% chance of losing $1000 and a 25% chance of losing $0.

The results are given below. Decision 1 is a Positive frame, Decision 2 is Negative. Choices A and C are “sure thing” selections, B and D are “risky.”

- In 16 subjects, both sure things (A and C) were chosen.
- In 110 subjects, the Positive sure thing (A) and Negative risky bet (D) were chosen.
- In 4 subjects, the Positive risky bet (B) and Negative sure thing (C) were chosen.
- In 20 subjects, both risky bets (B and D) were chosen.

The data are summarized in Table 8.4.

We can test whether the tendency to choose between a sure thing and risky bet depends on whether the choice is framed positive (gain) or negative (loss) based on McNemar’s test, since both outcomes are being observed on the same subjects.

- H_0 : No differences in tendency to choose between sure thing and risky bet under the two frames

		Negative Frame		
		Sure Thing	Risky Bet	
Positive	Sure Thing	16	110	126
Frame	Risky Bet	4	20	24
		20	130	150

Table 8.4: Positive and Negative frames and subjects' selections between sure thing and risky bet

- H_A : The probabilities differ
- $TS : z_{obs} = \frac{110-4}{\sqrt{110+4}} = \frac{106}{10.6771} = 9.9278$
- $RR : |z_{obs}| \geq z_{.025} = 1.96$ (For 2-sided test, with $\alpha = 0.05$)
- P-value: $2P(Z \geq 9.9278) \approx 0$ (For 2-sided test)

Thus, we conclude that the tendencies differ. People tend to choose the sure thing when posed as a gain, and the risky bet when posed as a loss. The exact P-value is set-up below.

$$P = 2P\left(Y \geq 110 | Y \sim Bin(n = 114, \pi = 0.5)\right) \approx 0$$

R Commands and Output

```
## Commands

(bet <- matrix(c(16,110,4,20),byrow=T,ncol=2))

mcnemar.test(bet,correct=F)
z.stat <- (bet[1,2]-bet[2,1])/sqrt(bet[1,2]+bet[2,1])
z.p <- 2*(1-pnorm(abs(z.stat),0,1))
binom.p <- 2*(1-pbinom(max(bet[1,2],bet[2,1])-1,bet[1,2]+bet[2,1],0.5))

bet.out <- cbind(bet[1,2], bet[2,1], z.stat, z.stat^2, z.p, binom.p)
colnames(bet.out) <- c("n12=+R/-S", "n21=+S/-R", "z", "z^2", "P(z)", "P(exact)")
round(bet.out, 4)

### Output

> (bet <- matrix(c(16,110,4,20),byrow=T,,ncol=2))
     [,1] [,2]
[1,]   16  110
[2,]    4   20
>
>> mcnemar.test(bet,correct=F)

        McNemar's Chi-squared test

data:  bet
McNemar's chi-squared = 98.561, df = 1, p-value < 2.2e-16

> round(bet.out, 4)
     n12=+R/-S n21=+S/-R      z     z^2 P(z) P(exact)
[1,]       110         4 9.9278 98.5614    0        0
```

The chi-square statistic from **mcnemar.test** is the square of the z-statistic. They give identical P-values for a 2-tailed test.

∇

8.4 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 8.1. The n_{ij} values are referred to as the **observed** counts. 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 8.5.

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

The special case of 2×2 tables has already been covered. Now generalize to r groups (treatments) and c possible outcomes. To perform Pearson's Chi–square test, compute the **expected** values for each cell count under the hypothesis of independence, and obtain a statistic based on discrepancies between the observed and expected values.

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

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 (grouping) variable. They apply the marginal proportion of cases in column j, $n_{.j}/n_{..}$ to the number of units in row i, $n_{i.}$. The test is conducted as follows:

1. H_0 : No association between the explanatory and response variables (see Table 8.5).
2. H_A : Explanatory and response variables are associated
3. T.S.: $X^2_{obs} = \sum_{\text{all cells}} \frac{(\text{observed}-\text{expected})^2}{\text{expected}} = \sum_{i,j} \frac{(n_{ij}-E_{ij})^2}{E_{ij}}$

4. RR: $X^2_{obs} > \chi^2_{\alpha,(r-1)(c-1)}$
5. P–value: $P(\chi^2_{(r-1)(c-1)} \geq X^2_{obs})$

If the chi-square test rejects the null hypothesis, **standardized (adjusted) residuals** can be used to determine which cells are the "cause" of the association between the variables. These are like Z-statistics. Generally, standardized residuals larger than 2 or 3 in absolute values are considered to be evidence against independence in that cell.

$$R_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}\left(1 - \frac{n_{i.}}{n_{..}}\right)\left(1 - \frac{n_{.j}}{n_{..}}\right)}}$$

Example 8.13: Jury Decisions in Product Liability Cases

An experiment was conducted regarding jurors' decisions to award plaintiffs in product liability trials (Culp and Pollage (2002) [17]). The observed and expected values are given in Table 8.6. There were $r = 5$ treatments and $c = 2$ outcomes (award in favor of plaintiff, or not). The five conditions were as follows (all conditions included the jurors hearing the facts of the case).

1. Judge's instruction on strict liability and lawyer's oral arguments
2. Judge's instruction on negligence and lawyer's oral arguments
3. No judge's instruction or lawyer's oral arguments (Control)
4. Judge's instruction on strict liability but no lawyer's oral arguments
5. Judge's instruction on negligence but no lawyer's oral arguments

Jury Condition (i)	Award	No Award	Total
Strict Liability/Oral Argument (1)	15 (21.80)	43 (36.20)	58
Negligence/Oral Argument (2)	18 (17.66)	29 (29.34)	47
Control (3)	7 (14.66)	32 (24.34)	39
Strict Liability/No Oral Argument (4)	37 (28.19)	38 (46.81)	75
Negligence/No Oral Argument (5)	38 (32.70)	49 (54.30)	87
Total	115	191	306

Table 8.6: Observed (expected) values of numbers of jurors voting to award or not award plaintiff in product liability trial)

Overall, the proportion of jurors voting to award the plaintiff is 115/206 = .3758, and the proportion voting no award is .6242. These proportions are applied to the row totals to obtain the expected counts under the hypothesis of no association between juror condition and voting outcome.

$$E_{11} = \left(\frac{115}{306}\right)(58) = 21.80 \quad E_{12} = \left(\frac{191}{306}\right)(58) = 36.20 \cdots E_{51} = \left(\frac{115}{306}\right)(87) = 32.70 \quad E_{52} = \left(\frac{191}{306}\right)(87) = 54.30$$

The test of whether there is an association between jury condition and vote outcome is conducted below.

H_0:Jury condition and voting outcome are independent vs H_A: Jury condition and voting outcome are associated.

$$TS : X^2_{obs} = \sum_{i=1}^{5}\sum_{j=1}^{2}\frac{\left(n_{ij}-E_{ij}\right)^2}{E_{ij}} = \frac{(15-21.80)^2}{21.80} + \cdots + \frac{(49-54.30)^2}{54.30} = 2.121 + \cdots + 0.517 = 15.609$$

$$RR : X^2_{obs} \geq \chi^2_{.05,(5-1)(2-1)} = 9.488 \quad P = P\left(\chi^2_4 \geq 15.609\right) = .0036$$

The standardized residuals for the control treatment (Jury Condition 3) are −2.71 for Award and +2.71 for No Award, while those for Jury Condition 4 are +2.42 and −2.42, respectively. While these do not exceed 3 in absolute value, they are well above 2. Fewer jurors in the Control Group voted to award the plaintiff than expected under independence, and more voted to award the plaintiff in Jury Condition 4. The calculations for the Control Group are given below.

$$R_{31} = \frac{7-14.66}{\sqrt{14.66(1-39/306)(1-115/306)}} = \frac{-7.66}{2.83} = -2.71 \quad R_{32} = \frac{32-24.34}{\sqrt{24.34(1-39/306)(1-191/306)}} = \frac{7.66}{2.83} = 2.71$$

R Commands and Output

```
## Commands

pla <- read.csv("http://www.stat.ufl.edu/~winner/data/productliability_award.csv")
attach(pla); names(pla)

(jury_award <- table(jury,award))

X2_ja <- chisq.test(jury_award, correct=F)
X2_ja
X2_ja$stdres

## Output

> (jury_award <- table(jury,award))
    award
jury  0  1
   1 43 15
   2 29 18
   3 32  7
   4 38 37
   5 49 38
> X2_ja
        Pearson's Chi-squared test
data:  jury_award
X-squared = 15.608, df = 4, p-value = 0.003592
> X2_ja$stdres
    award
jury          0          1
   1  2.0470036 -2.0470036
   2 -0.1101878  0.1101878
   3  2.7100635 -2.7100635
   4 -2.4184629  2.4184629
   5 -1.3878162  1.3878162
```

∇

Chapter 9

Regression Models

Linear regression is used when there is a numeric response variable and numeric (and possibly categorical) predictor (explanatory) variable(s). The mean of the response variable is to be related to the predictor(s) with random error terms typically assumed to be independent and normally distributed with constant variance. The fitting of linear regression models is very flexible, allowing for fitting curvature, categorical predictors, and interactions between factors.

Logistic Regression is used when the outcome is binary, and there are one or more numeric (or possibly categorical) predictor variable(s). These models are used to determine whether there the probability the outcome of interest is associated with the predictor variable(s).

9.1 Simple Linear Regression

When there is a single numeric predictor, the model is referred to as **Simple Regression**. The response variable is denoted as Y and the predictor variable is denoted as X. The model is written as follows.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad\qquad \epsilon \sim N(0, \sigma) \text{ independent}$$

Here β_0 is the intercept (mean of Y when X=0) and β_1 is the slope (the change in the mean of Y when X increases by 1 unit). Of primary concern is whether $\beta_1 = 0$, which implies the mean of Y is constant (β_0), and thus Y and X are not associated.

9.1.1 Estimation of Model Parameters

A sample of pairs $(X_i, Y_i) \quad i = 1, \ldots, n$ is observed. The goal is to choose estimators of β_0 and β_1 that minimize the error sum of squares: $Q = \sum_{i=1}^{n} \epsilon_i^2$. The resulting **ordinary least squares** estimators are given below (the formulas are derived making use of calculus).

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad i = 1, \ldots, n \qquad \epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Once estimates have been computed, **fitted values** and **residuals** are obtained for each observation. The **error sum of squares (SSE)** is obtained as the sum of the squared residuals from the regression fit.

$$\text{Fitted Values: } \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad \text{Residuals: } e_i = Y_i - \hat{Y}_i \qquad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

The (unbiased) estimator of the error variance σ^2 is $s^2 = MSE = \frac{SSE}{n-2}$, where MSE is the **Mean Square Error**. The subtraction of 2 can be thought of as the fact two parameters have been estimated: β_0 and β_1.

The estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ are linear functions of $Y_1, \ldots, Y_n$ and thus using basic rules of mathematical statistics, their sampling distributions are as follow, assuming the error terms are normal, independent, with constant variance.

$$\hat{\beta}_1 \sim N\left(\beta_1, \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}\right) \qquad \hat{\beta}_0 \sim N\left(\beta_0, \sqrt{\sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}\right)$$

The estimated standard errors are the standard error with the unknown σ^2 replaced by MSE.

$$\hat{SE}\{\hat{\beta}_1\} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \qquad \hat{SE}\{\hat{\beta}_0\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}$$

Example 9.1: Bollywood Films' Revenues and Budgets 2013-2017

Box office data for $n = 190$ Bollywood films, as well as their approximate budgets (production and advertising) were obtained from bollywoodmoviereviewz.com. These films are being treated as a random sample of all movies that could have been made under similar conditions. Plots of gross revenues versus budget are given in Figure 9.1. As is often seen with this type of data, logarithmic transformations on Y and/or X can be helpful in linearizing the relationship. All four possibilities are considered.

Based on the plots, the model with both variables transformed to the logarithmic scale is fit. This is due to the linear relation with approximately constant variance. When both variables have been transformed this way, the slope can be interpreted as percent change in Y when X is increased by 1%. Calculations for the linear regression are given below.

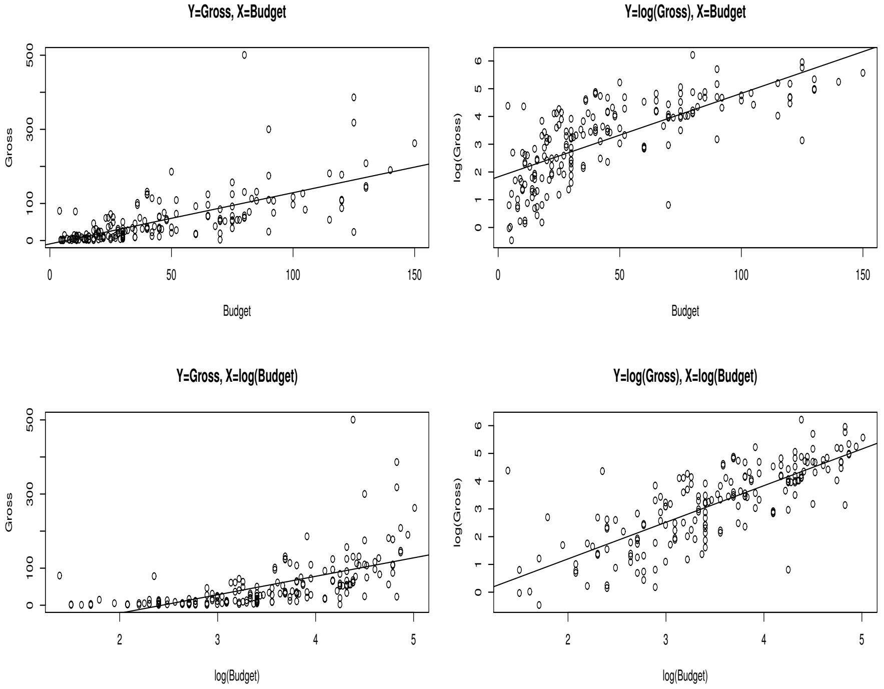

Figure 9.1: Bollywood Film Revenues and Budgets 2013-2017.

$$n = 190 \qquad \overline{X} = 3.5049 \qquad \overline{Y} = 3.1846$$

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = 131.043 \qquad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 381.436 \qquad \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = 172.9174$$

$$\hat{\beta}_1 = \frac{172.9174}{131.043} = 1.3195 \quad \hat{\beta}_0 = 3.1846 - 1.3195(3.5049) = -1.4401 \quad SSE = 381.436 - (1.3195)^2(131.043) = 153.2796$$

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{153.2796}{190-2} = 0.8153$$

$$\hat{SE}\{\hat{\beta}_1\} = \sqrt{\frac{0.8153}{131.043}} = 0.0789 \qquad \hat{SE}\{\hat{\beta}_0\} = \sqrt{0.8153\left[\frac{1}{190} + \frac{3.5049^2}{131.043}\right]} = 0.2841$$

R Output

```
## Output

> round(ss.out, 4)
        SSYY    SSXX     SSXY      SSE    MSE beta1-hat  b0-hat SE{b1} SE{b0}
[1,] 381.436 131.043 172.9174 153.2636 0.8152    1.3195 -1.4402 0.0789 0.2841
```

∇

9.1.2 Inference Regarding β_1 and β_0

Primarily of interest are inferences regarding β_1. Note that if $\beta_1 = 0$, Y and X are not linearly associated. We can test hypotheses and construct confidence intervals based on the estimate β_1 and its estimated standard error. The t-test is conducted as follows. Note that the null value β_{10} is almost always 0, and that software packages that report these tests always are treating β_{10} as 0.

$$H_0 : \beta_1 = \beta_{10} \quad H_A : \beta_1 \neq \beta_{10} \quad TS : t_{obs} = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{SE}\{\hat{\beta}_1\}} \quad RR : |t_{obs}| \geq t_{\alpha/2,n-2} \quad P = 2P\left(t_{n-2} \geq |t_{obs}|\right)$$

One-sided tests use the same test statistic, but the Rejection Region and P-value are changed to reflect the alternative hypothesis.

$$H_A^{+} : \beta_1 > \beta_{10} \qquad RR : t_{obs} \geq t_{\alpha,n-2} \qquad P = P\left(t_{n-2} \geq t_{obs}\right)$$

$$H_A^{-} : \beta_1 < \beta_{10} \qquad RR : t_{obs} \leq -t_{\alpha,n-2} \qquad P = P\left(t_{n-2} \leq t_{obs}\right)$$

A $(1-\alpha)100\%$ confidence interval for β_1 is obtained as:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\hat{SE}\{\hat{\beta}_1\}$$

Note that the confidence interval represents the values of β_{10} for which the two-sided test: $H_0 : \beta_1 = \beta_{10} \quad H_A : \beta_1 \neq \beta_{10}$ fails to reject the null hypothesis.

Inferences regarding β_0 are of less interest in practice, but can be conducted in analogous manner, using the estimate $\hat{\beta}_0$ and its estimated standard error $\hat{SE}\{\hat{\beta}_0\}$.

Example 9.2: Bollywood Films' Revenues and Budgets 2013-2017

Continuing with the Bollywood data with both Revenues and Budget on logarithmic scales, a test of $H_0 : \beta_1 = 0$ and a 95% Confidence Interval for β_1 are obtained.

$$H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0 \quad TS : t_{obs} = \frac{1.3195}{0.0789} = 16.72 \quad RR : |t_{obs}| \geq 1.973 \quad P \approx 0$$

$$95\% \text{ Confidence Interval for } \beta_1 : 1.3195 \pm 1.973(0.0789) \equiv 1.3195 \pm 0.1557 \equiv (1.1638, 1.4752)$$

There is strong evidence of an association between log(Revenue) and log(Budget). Similarly, inference regarding the intercept β_0 can be made as well (although is of less interest as no movies had log(Budget)=0).

$$H_0 : \beta_0 = 0 \quad H_A : \beta_0 \neq 0 \quad TS : t_{obs} = \frac{-1.4402}{0.2841} = -5.069 \quad RR : |t_{obs}| \geq 1.973 \quad P \approx 0$$

$$95\% \text{ Confidence Interval for } \beta_0 : -1.4402 \pm 1.973(0.2841) \equiv -1.4402 \pm 0.5605 \equiv (-2.0007, -0.8797)$$

R Commands and Output

```
## Commands
## Analysis using lm (linear model) function in R

bolly.mod1 <- lm(Y ~ X)
summary(bolly.mod1)
confint(bolly.mod1)

## Output

> round(b.out, 4)
            Estimate Std. Error       t P-Value Lower Bound Upper Bound
Intercept    -1.4402     0.2841 -5.0695       0     -2.0007     -0.8798
log(Budget)   1.3195     0.0789 16.7298       0      1.1640      1.4751

> summary(bolly.mod1)
Call:
lm(formula = Y ~ X)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.44023    0.28410  -5.069 9.51e-07 ***
X            1.31955    0.07887  16.730  < 2e-16 ***

Residual standard error: 0.9029 on 188 degrees of freedom
Multiple R-squared:  0.5982,    Adjusted R-squared:  0.5961
F-statistic: 279.9 on 1 and 188 DF,  p-value: < 2.2e-16

> confint(bolly.mod1)
                2.5 %    97.5 %
(Intercept) -2.000665 -0.879805
X            1.163955  1.475138
```

∇

9.1.3 Estimating a Mean and Predicting a New Observation @ $X = X^*$

There may be interest in estimating the mean response at a specific level X^*. The parameter of interest is $\mu^* = \beta_0 + \beta_1 X^*$. The point estimator, standard error, and $(1-\alpha)100\%$ Confidence Interval are given below.

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad \hat{SE}\left\{\hat{Y}^*\right\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}$$

$$(1-\alpha)100\% \text{ CI } : \hat{Y}^* \pm t_{\alpha/2,n-2}\hat{SE}\left\{\hat{Y}^*\right\}$$

To obtain a simultaneous $(1-\alpha)100\%$ Confidence Interval for the entire regression line (not just a single point), the Working-Hotelling method can be used.

$$\hat{Y}^* \pm \sqrt{2F_{\alpha,2,n-2}}\hat{SE}\left\{\hat{Y}^*\right\}$$

If the goal is to predict a new observation when $X = X^*$, uncertainty with respect to estimating the mean (as seen by the Confidence Interval above), and the random error for the new case (with standard deviation σ) must be taken into account. The point prediction is the same as for the mean. The prediction, standard error of prediction, and $(1-\alpha)100\%$ Prediction Interval are given below.

$$\hat{Y}^*_{\text{New}} = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad \hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\} = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}$$

$$(1-\alpha)100\% \text{ PI } : \hat{Y}^*_{\text{New}} \pm t_{\alpha/2,n-2}\hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\}$$

Note that the Prediction Interval will tend to be much wider than the Confidence Interval for the mean.

Example 9.3: Bollywood Films' Revenues and Budgets 2013-2017

Continuing with the Bollywood data with both Revenues and Budget on logarithmic scales, a 95% Confidence Interval for the mean log(Revenue) of all possible films with a Budget of 60 ($X^* = \log(60) = 4.0943$) is obtained. Also a Prediction Interval for a single new movie with a budget of 60 is computed. The predicted value is $\hat{Y}^* = -1.4401 + 1.3195(4.0943) = 3.9623$. A plot of the data, fitted equation, 95% Confidence and Prediction Intervals is given in Figure 9.2.

$$\hat{SE}\left\{\hat{Y}^*\right\} = \sqrt{0.8153\left[\frac{1}{190} + \frac{(4.0943 - 3.5049)^2}{131.043}\right]} = \sqrt{0.8153(0.0079)} = 0.0803$$

$$\hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\} = \sqrt{0.8153(1.0079)} = 0.9065$$

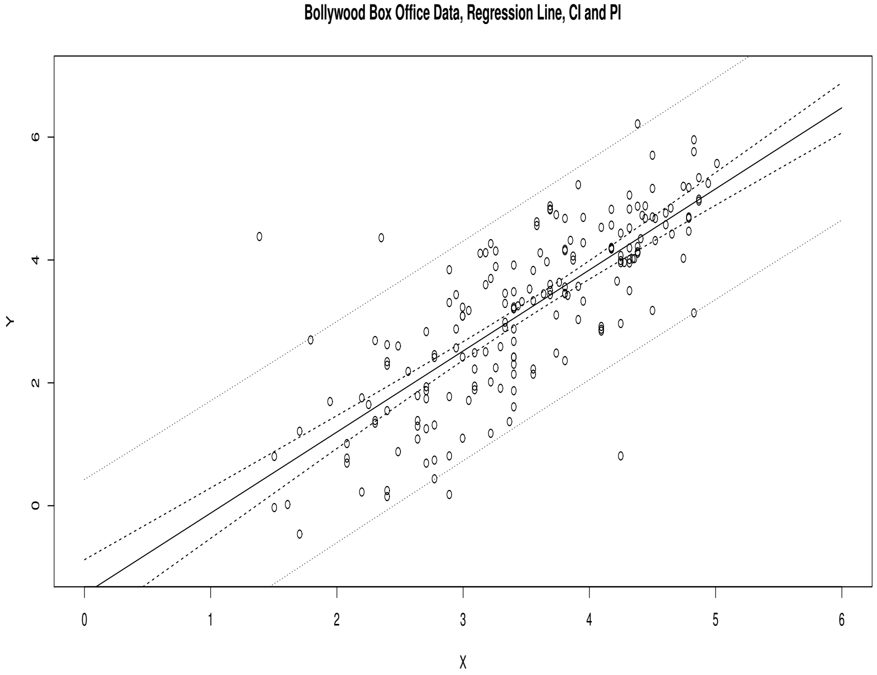

Figure 9.2: Bollywood Data, Fitted Equation 95% Confidence Interval for the mean and Prediction Interval for individual films

$$\text{95\% CI for Mean: } 3.9623 \pm 1.973(0.0803) \equiv 3.9623 \pm 0.1585 \equiv (3.8038, 4.1208)$$

$$\text{95\% PI for Individual: } 3.9623 \pm 1.973(0.9065) \equiv 3.9623 \pm 1.7885 \equiv (2.1738, 5.7508)$$

To convert back to the original units, the bounds of the Confidence and Prediction Intervals are exponentiated. The predicted revenue is $e^{3.9623} = 52.58$ and the 95% Confidence Interval and Prediction Interval are given below.

$$\text{95\% CI for Mean: } \left(e^{3.8038} = 44.87, e^{4.1208} = 61.61\right) \qquad \text{95\% PI for Individual: } \left(e^{2.1738} = 8.79, e^{5.7508} = 314.44\right)$$

R Commands and Output

```
## Commands
## Using predict function based on bolly.mod1 object with X*=log(60)
# CI for mean
ci.log60 <- predict(bolly.mod1, list(X=log(60)), interval="c")
# PI for individual movie
pi.log60 <- predict(bolly.mod1, list(X=log(60)), interval="p")

cipi.out1 <- rbind(ci.log60, pi.log60, exp(ci.log60), exp(pi.log60))
```

```
colnames(cipi.out1) <- c("Estimate", "Lower Bound", "Upper Bound")
rownames(cipi.out1) <- c("CI(log scale)", "PI(log scale)",
                         "CI(original scale)", "PI(original scale)")
round(cipi.out1, 4)

## Output

> round(cipi.out,4)
                     X*  Y-hat* CI Lower CI Upper PI Lower PI Upper
Log Scale       4.0943  3.9624   3.8040   4.1209   2.1743   5.7506
Original Scale 60.0000 52.5856  44.8797  61.6147   8.7959 314.3788

> round(cipi.out1, 4)
                   Estimate Lower Bound Upper Bound
CI(log scale)        3.9624      3.8040      4.1209
PI(log scale)        3.9624      2.1743      5.7506
CI(original scale)  52.5856     44.8797     61.6147
PI(original scale)  52.5856      8.7959    314.3788
```

∇

9.1.4 Analysis of Variance

When there is no association between Y and X ($\beta_1 = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing sum of squares of prediction errors). In this case, the total variation can be denoted as $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, the **Total Sum of Squares**.

When there is an association between Y and X ($\beta_1 \neq 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ (in terms of minimizing sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, the **Error Sum of Squares**.

The difference between TSS and SSE is the variation "explained" by the regression of Y on X (as opposed to having ignored X). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ the **Regression Sum of Squares**.

$$TSS = SSE + SSR \qquad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

A plot including the data (Y), the horizontal line at the mean response ($\overline{Y}$) and the fitted equation is given in Figure 9.3. The sum of the squared vertical distances from the data Y_i to $\overline{Y}$ is the Total Sum of Squares TSS. The sum of the squared vertical distances from Y_i to their fitted values $\hat{Y}_i$ is the Error Sum of Squares SSE. The sum of the squared vertical distances from $\hat{Y}_i$ to $\overline{Y}$ is the Regression Sum of Squares SSR.

Each sum of squares has a **degrees of freedom** associated with it. The **Total Degrees of Freedom** is $df_{\text{Total}} = n - 1$. The **Error Degrees of Freedom** is $df_{\text{Error}} = n - 2$ (for simple regression). The **Regression Degrees of Freedom** is $df_{\text{Regression}} = 1$ (for simple regression).

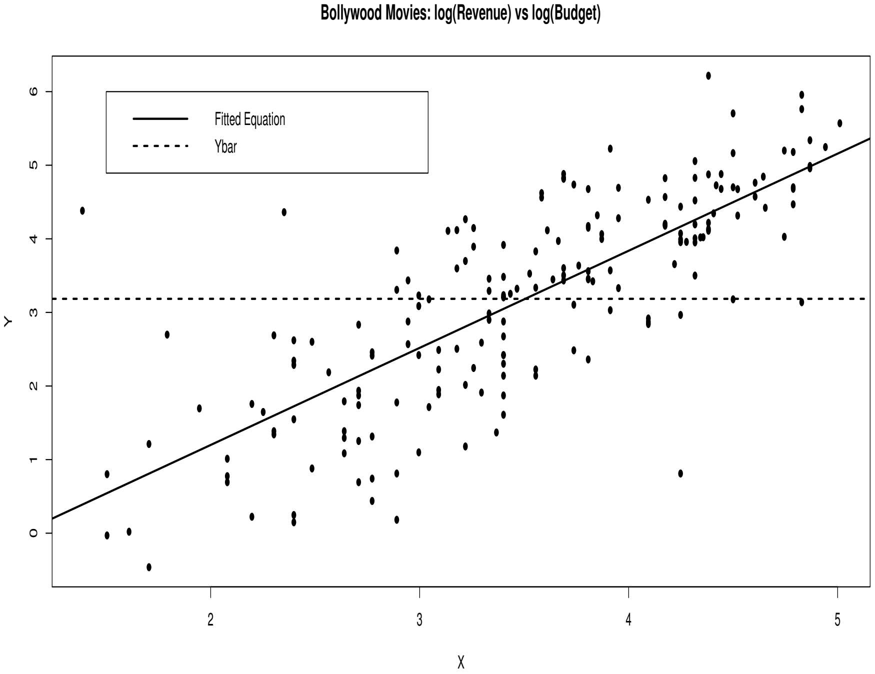

Figure 9.3: Plot of Data (points), Fitted Equation and Mean of Y - Bollywood movie regression with Y=log(Revenue) and X=log(Budget)

Source	*df*	*SS*	*MS*	F_{obs}	P-value
Regression (Model)	1	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{1}$	$F_{obs} = \frac{MSR}{MSE}$	$P\left(F_{1,n-2} \geq F_{obs}\right)$
Error (Residual)	$n-2$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-2}$		
Total (Corrected)	$n-1$	$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Table 9.1: Analysis of Variance Table for Simple Linear Regression

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n-1 = n-2+1$$

The Error and Regression sums of squares have **Mean Squares**, which are the sum of squares divided by their corresponding degrees of freedom: $MSE = SSE/(n-2)$ and $MSR = SSR/1$. It can be shown that these mean squares have the following **Expected Values**, average values in repeated sampling at the same observed X levels.

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

Note that when $\beta_1 = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A second way of testing whether $\beta_1 = 0$ is by the following F-test.

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0 \quad TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,1,n-2} \qquad P = P\left(F_{1,n-2} \geq F_{obs}\right)$$

The Analysis of Variance is typically set up in a table as in Table 9.1.

A measure often reported from a regression analysis is the **Coefficient of Determination** or r^2. This represents the variation in Y "explained" by X, divided by the total variation in Y.

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq r^2 \leq 1$$

The interpretation of r^2 is the proportion of variation in Y that is "explained" by X, and is often reported as a percentage ($100r^2$).

Example 9.4: Bollywood Films' Revenues and Budgets 2013-2017

Continuing with the Bollywood data with both Revenues and Budget on logarithmic scales, the Analysis of Variance and F-test are given Table 9.2. Note that the Total Sum of Squares and Error Sum of Squares were computed in Example 9.1. The Regression Sum of Squares is the difference $SSR = TSS - SSE = 381.4360 - 153.2636 = 228.1725$.

Source	df	SS	MS	F_{obs}	P-value
Regression (Model)	1	228.1725	228.1725	279.8667	≈ 0
Error (Residual)	188	153.2676	0.8152		
Total (Corrected)	189	381.4360			

Table 9.2: Analysis of Variance Table for Bollywood Box Office Data

The coefficient of determination, r^2, is 228.1725/381.4360=0.5982. Approximately 60% of the variation in log Revenue is "explained" by log Budget.

R Commands and Output

```
## Commands

bolly.mod1 <- lm(Y ~ X)
summary(bolly.mod1)
anova(bolly.mod1)

## Output

> round(aov.out,4)
         TSS      SSE      SSR    MSE    F_obs F(.05) P-value    R^2
[1,] 381.436 153.2636 228.1725 0.8152 279.8867 3.8914       0 0.5982

> summary(bolly.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.44023    0.28410  -5.069 9.51e-07 ***
X            1.31955    0.07887  16.730  < 2e-16 ***

Residual standard error: 0.9029 on 188 degrees of freedom
Multiple R-squared:  0.5982,    Adjusted R-squared:  0.5961
F-statistic: 279.9 on 1 and 188 DF,  p-value: < 2.2e-16

> anova(bolly.mod1)
Analysis of Variance Table
Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X           1 228.17 228.172  279.89 < 2.2e-16 ***
Residuals 188 153.26   0.815
```

∇

9.1.5 Correlation

The regression coefficient β_1 depends on the units of Y and X. It also depends on which variable is the dependent variable and which is the independent variable. A second widely reported measure is the **Pearson Product Moment Coefficient of Correlation**. It is invariant to linear transformations of Y and X, and does not distinguish which is the dependent and which is the independent variable. This makes it a widely reported measure when researchers are interested in how two random variables vary together in a population.

The population correlation coefficient is labeled ρ, and the sample correlation is labeled r, and its formula is given below.

$$r = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} = \left(\frac{s_X}{s_Y}\right)\hat{\beta}_1$$

where s_X and s_Y are the standard deviations of X and Y, respectively. While $\hat{\beta}_1$ can take on any value, r lies between -1 and $+1$, taking on the extreme values if all of the points fall on a straight line. The test of whether $\rho = 0$ is mathematically equivalent to the t-test for testing whether $\beta_1 = 0$. The 2-sided test is given below.

$$H_0:\rho=0 \qquad H_A:\rho\neq 0 \quad TS: t_{obs}=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad RR: |t_{obs}|\geq t_{\alpha/2,n-2} \qquad P=2P(t_{n-2}\geq|t_{obs}|)$$

To construct a large-sample confidence interval, **Fisher's z transform** is used to make the transformed r to have a sampling distribution that is approximately normal. A confidence interval is obtained on the transformed correlation, then "back transformed" to the end points in terms of ρ.

$$z'=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad (1-\alpha)100\% \text{ CI for } \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right): \quad z'\pm z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

Labeling the endpoints of the Confidence Interval as (a,b), the Confidence Interval for ρ is computed as follows.

$$(1-\alpha)100\% \text{ Confidence Interval for } \rho: \left(\frac{e^{2a}-1}{e^{2a}+1},\frac{e^{2b}-1}{e^{2b}+1}\right)$$

Example 9.5: Bollywood Films' Revenues and Budgets 2013-2017

Continuing with the Bollywood data with both Revenues and Budget on logarithmic scales, the sample correlation, a test of whether $\rho = 0$, and a 95% Confidence Interval for ρ are computed below.

$$r=\frac{172.9174}{\sqrt{131.0430(381.4360)}}=0.7734 \qquad t_{obs}=\frac{0.7734}{\sqrt{\frac{1-0.7734^2}{190-2}}}=16.73$$

$$z'=\frac{1}{2}\ln\left(\frac{1+0.7734}{1-0.7734}\right)=1.0287 \qquad 1.0287\pm1.96\sqrt{\frac{1}{190-3}}\equiv 1.0287\pm0.1433\equiv(0.8854,1.1720)$$

$$\Rightarrow \quad (1-\alpha)100\% \text{ CI for } \rho: \left(\frac{e^{2(0.8854)}-1}{e^{2(0.8854)}+1},\frac{e^{2(1.1720)}-1}{e^{2(1.1720)}+1}\right)\equiv\left(\frac{4.8756}{6.8756},\frac{9.4228}{11.4228}\right)\equiv(.7091,.8249)$$

R Commands and Output

```
## Commands
cor.test(X,Y)
## Output
> cor.test(X,Y)
        Pearson's product-moment correlation
data:  X and Y
t = 16.73, df = 188, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7091543 0.8249551
sample estimates:
      cor
0.7734296
```

9.2 Multiple Linear Regression

When there is more than one predictor variable, the model generalizes to multiple linear regression. The calculations become more complex, but conceptually, the ideas remain the same. We will use the notation of p as the number of predictors, and $p' = p+1$ as the number of regression coefficients in the model (including the intercept). The model can be written as follows with the same assumptions about the errors as in simple regression.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad\qquad \epsilon \sim N(0,\sigma^2) \text{ independent}$$

Least squares (and maximum likelihood) estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ minimize the error sum of squares. The fitted values, residuals, and error sum of squares are given below.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots \hat{\beta}_p X_{ip} \qquad e_i = Y_i - \hat{Y}_i \qquad SSE = \sum_{i=1}^{n} e_i^2$$

The degrees of freedom for error are now $n - p' = n - (p+1)$, as the model estimates $p' = p+1$ parameters. The degrees of freedom for regression is p.

In the multiple linear regression model, β_j represents the change in $E\{Y\}$ when X_j increases by 1 unit, with all other predictor variables being held constant. It is referred to as the **partial regression coefficient**.

9.2.1 Testing and Estimation for Partial Regression Coefficients

Once the model is fit, for each predictor variable, the estimated regression coefficient, its estimated standard error, t-statistic and confidence interval are obtained. Technically, the estimated variance-covariance matrix for the vector of regression coefficients is computed, with the standard errors being the square root of the variances of the individual coefficients.

To test whether Y is associated with X_j, after controlling for the remaining $p-1$ predictors, the test is whether $\beta_j = 0$. This is equivalent to the t-test from simple regression (in general, the test can be whether a regression coefficient is any specific number, although software packages are testing whether it is 0).

$$H_0 : \beta_j = \beta_{j0} \qquad H_A : \beta_j \neq \beta_{j0} \quad TS : t_{obs} = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{SE}\{\hat{\beta}_j\}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-p'} \qquad P = 2P(t_{n-p'} \geq |t_{obs}|)$$

One-sided tests make the same adjustments as in simple linear regression.

$$H_A^+ : \beta_j > \beta_{j0} \qquad RR : t_{obs} \geq t_{\alpha,n-p'} \qquad P = P(t_{n-p'} \geq t_{obs})$$

$$H_A^- : \beta_j < \beta_{j0} \qquad RR : t_{obs} \leq -t_{\alpha,n-p'} \qquad P = P(t_{n-p'} \leq t_{obs})$$

A $(1-\alpha)100\%$ confidence interval for β_j is obtained as:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p'} \hat{SE}\{\hat{\beta}_j\}$$

Note that the confidence interval represents the values of β_{j0} for which the two-sided test: $H_0 : \beta_j = \beta_{j0} \quad H_A : \beta_j \neq \beta_{j0}$ fails to reject the null hypothesis.

Example 9.7: How Stature (Height) Relates to Hand and Foot Length among Females

A regression model was fit, relating stature (Y, height, in mm) to hand length (X_1, mm) and foot length (X_2, mm) for a sample of $n = 75$ female adult Turks (Sanli, Kizilkanat, Boyan, et al. (2005), [51]). The data have been simulated to match means, standard deviations, and bivariate correlations. A matrix plot of the variables is given in Figure 9.4. The model, fitted equation, Error sum of squares and mean square are given below ($n = 75, p' = 2 + 1 = 3$).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad \hat{Y}_i = 743.970 + 2.375X_1 + 1.727X_2 \quad SSE = 68924.42 \quad MSE = 957.284$$

The estimated standard errors are 0.486 for $\hat{\beta}_1$ and 0.375 for $\hat{\beta}_2$, respectively. The t-tests and 95% Confidence Intervals for β_1 and β_2 are given below.

$$\text{Hand: } H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0 \quad TS : t_{obs} = \frac{2.375}{0.486} = 4.89 \quad RR : |t_{obs}| \geq t_{.025,72} = 1.993 \quad P = P(t_{72} \geq 5.63) \approx 0$$

$$\text{Foot: } H_0 : \beta_2 = 0 \quad H_A : \beta_2 \neq 0 \quad TS : t_{obs} = \frac{1.727}{0.375} = 4.61 \quad RR : |t_{obs}| \geq t_{.025,72} = 1.993 \quad P = P(t_{72} \geq 4.61) \approx 0$$

$$95\% \text{ CI for } \beta_1\text{: } 2.375 \pm 1.993(0.486) \equiv 2.375 \pm 0.969 \equiv (1.406, 3.344)$$

$$95\% \text{ CI for } \beta_2\text{: } 1.727 \pm 1.993(0.375) \equiv 1.727 \pm 0.747 \equiv (0.980, 2.474)$$

R Commands and Output

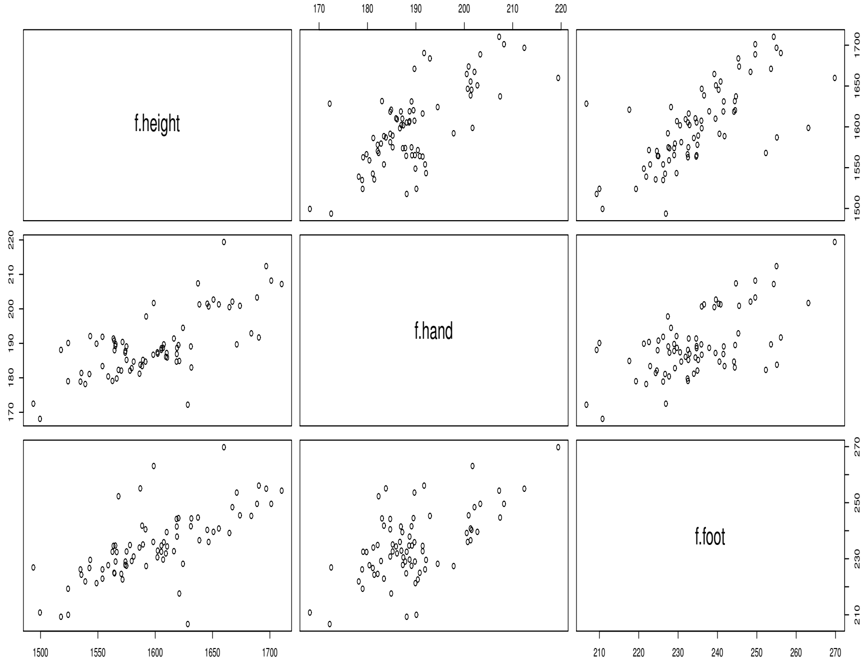

Figure 9.4: Heights, Hand Lengths and Foot Lengths among a Sample of 75 Adult Female Turks

```
### Commands
shf1 <- read.table("http://www.stat.ufl.edu/~winner/data/stature_hand_foot.dat",
    header=F, col.names=c("idnum", "gender", "height", "hand", "foot"))
attach(shf1)

f.height <- height[gender == 2]     ### Female Heights
f.hand <- hand[gender == 2]         ### Female Hand Lengths
f.foot <- foot[gender == 2]         ### Female Foot Lengths

f.stature <- data.frame(f.height, f.hand, f.foot)
plot(f.stature)

shf.mod1 <- lm(f.height ~ f.hand + f.foot)
summary(shf.mod1)
confint(shf.mod1)

#### Output
> summary(shf.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 743.9696    79.7499   9.329 5.12e-14 ***
f.hand        2.3748     0.4858   4.888 5.99e-06 ***
f.foot        1.7271     0.3745   4.611 1.69e-05 ***

Residual standard error: 30.94 on 72 degrees of freedom
Multiple R-squared:  0.6159,    Adjusted R-squared:  0.6053
F-statistic: 57.73 on 2 and 72 DF,  p-value: 1.093e-15

> confint(shf.mod1)
                  2.5 %     97.5 %
(Intercept) 584.9911070 902.948034
```

```
f.hand        1.4062645   3.343310
f.foot        0.9804939   2.473711
```

∇

9.2.2 Analysis of Variance

When there is no association between Y and $X_1, \ldots, X_p$ ($\beta_1 = \cdots = \beta_p = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing sum of squares of prediction errors). In this case, the total variation can be denoted as $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, the **Total Sum of Squares**, just as with simple regression.

When there is an association between Y and at least one of $X_1, \ldots, X_p$ (not all $\beta_i = 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$ (in terms of minimizing the sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, the **Error Sum of Squares**.

The difference between TSS and SSE is the variation "explained" by the regression of Y on $X_1, \ldots, X_p$ (as opposed to having ignored $X_1, \ldots, X_p$). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ the **Regression Sum of Squares**. Note that when there are $p > 1$ predictors, the fitted equation is no longer a straight line in 2-dimensions. This makes visualization more difficult, but the concept of distance from observed to predicted value is the same. For the stature example, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ represents a 2-dimensional plane in 3-dimensional space.

$$TSS = SSE + SSR \qquad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

The **Total Degrees of Freedom** remains $df_{\text{Total}} = n - 1$. The **Error Degrees of Freedom** is $df_{\text{Error}} = n - p'$. The **Regression Degrees of Freedom** is $df_{\text{Regression}} = p$. Note that when there is $p = 1$ predictor, this generalizes to simple regression.

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n - 1 = n - p' + p$$

The Mean Squares for Error and Regression are: $MSE = SSE/(n - p')$ and $MSR = SSR/p$. It can be shown that these mean squares have the following **Expected Values**, average values in repeated sampling at the same observed X levels.

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} \geq \sigma^2$$

Note that when $\beta_1 = \cdots \beta_p = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A way of testing whether $\beta_1 = \cdots \beta_p = 0$ is by the F-test.

Source	df	SS	MS	F_{obs}	$P(>F)$
Regression (Model)	p	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{p}$	$F_{obs} = \frac{MSR}{MSE}$	$P(F_{p,n-p'} \geq F_{obs})$
Error (Residual)	$n - p'$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-p'}$		
Total (Corrected)	$n - 1$	$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Table 9.3: Analysis of Variance Table for Multiple Linear Regression

$$H_0 : \beta_1 = \cdots \beta_p = 0 \qquad H_A : \text{ Not all } \beta_j = 0$$

$$TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,p,n-p'} \qquad P = P\left(F_{p,n-p'} \geq F_{obs}\right)$$

The Analysis of Variance is typically set up in a table as in Table 9.3.

The **Coefficient of Determination** is labeled R^2 for the multiple regression model. This represents the variation in Y "explained" by $X_1, \ldots, X_p$, divided by the total variation in Y. Note that the **summary** function in R reports "Multiple R-squared" even when there is only a single predictor.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq R^2 \leq 1$$

Example 9.8: Stature (Height) as Function of Hand and Foot Length among Females

In a continuation of the Turkish adult females' model relating stature to hand and foot lengths, the following sums of squares and F-test are computed.

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 179409 \quad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 68924 \quad SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = 110504$$

$$MSE = \frac{68924}{75-3} = 957.3 \quad MSR = \frac{110504}{2} = 55252$$

$$H_0 : \beta_1 = \beta_2 = 0 \quad TS : F_{obs} = \frac{55252}{957.3} = 57.72 \quad RR : F_{obs} \geq F_{.05,2,72} = 3.124 \quad P\left(F_{2,72} \geq 57.72\right) \approx 0$$

The Coefficient of Determination is $R^2 = 110504/179409 = .616$, approximately 62% of the variation in height is "explained" by hand and foot length.

R Commands and Output

```
### Commands

shf.mod1 <- lm(f.height ~ f.hand + f.foot)
```

```
summary(shf.mod1)
anova(shf.mod1)
drop1(shf.mod1, test="F")

### Output

> summary(shf.mod1)
Residual standard error: 30.94 on 72 degrees of freedom
Multiple R-squared:  0.6159,    Adjusted R-squared:  0.6053
F-statistic: 57.73 on 2 and 72 DF,  p-value: 1.093e-15

> anova(shf.mod1)
Analysis of Variance Table

Response: f.height
          Df Sum Sq Mean Sq F value    Pr(>F)
f.hand     1  90153   90153  94.203 1.027e-14 ***
f.foot     1  20351   20351  21.265 1.694e-05 ***
Residuals 72  68905     957
```

Note that $SSR = SSR(X_1) + SSR(X_2|X_1) = 90153 + 20351 = 110504$. The sums of squares for the **anova** function are the **Sequential Sums of Squares** and sum up to the Regression Sum of Squares.

∇

9.2.3 Testing a Subset of $\beta^s = 0$

The F-test from the Analysis of Variance and the t-tests represent extremes of model testing (all variables simultaneously versus one-at-a-time). Often interest lies in testing whether a group of predictors do not improve prediction, after controlling for the remaining predictors.

Suppose that after controlling for g predictors, we wish to test whether the remaining $p-g$ predictors are associated with Y. That is, we wish to test the following hypotheses.

$$H_0 : \beta_{g+1} = \cdots \beta_p = 0 \qquad H_A : \text{ Not all of } \beta_{g+1}, \ldots, \beta_p = 0$$

Note that, the t-tests control for all other predictors, while here, we want to control for only $X_1, \ldots, X_g$. To do this, fit two models: the **Complete** or **Full Model** with all p predictors, and the **Reduced Model** with only the g "control" variables. For each model, obtain the Regression and Error sums of squares, as well as R^2. Let (F) represent the Full model and (R) represent the Reduced model. This leads to the following test statistic and rejection region.

$$TS : F_{obs} = \frac{\left[\frac{SSE(R)-SSE(F)}{(n-g')-(n-p')}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{SSR(F)-SSR(R)}{p-g}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{R_F^2-R_R^2}{p-g}\right]}{\left[\frac{1-R_F^2}{n-p'}\right]}$$

$$RR : F_{obs} \geq F_{\alpha,p-g,n-p'} \qquad P = P\left(F_{p-g,n-p'} \geq F_{obs}\right)$$

Example 9.9: Energy Consumption of Luxury Hotels

A study considered factors relating to Energy Consumption (Y, in millions of kilowatt-hours) for a sample of $n = 19$ luxury hotels in Hainan Province, China (Xin, Lu, Xu, and Wu (2012), [61]). The model had 3 predictors: Area (X_1, in 1000s of square meters), Age (X_2, in years), and Effective number of guest rooms (X_3, # rooms times occupancy rate).

Consider two models: Model 1 with X_1, X_2, X_3 as predictors and Model 2 with only X_1 as a predictor. The goal is to determine whether age and/or effective guest rooms is associated with energy consumption, after controlling for the hotel's size (Area). The data, fitted values and residuals for Models 1 and 2 are given in Table 9.4. The fitted equations and Error Sums of Squares are given below ($n = 19, p = 3, p' = 4, g = 1$).

$$\text{Model 1: Full: } \hat{Y}_F = -2.1320 + 0.1540X_1 + 0.0959X_2 + 0.0075X_3 \quad SSE(F) = 67.846 \quad df_E(F) = n - p' = 19 - 4 = 15$$

$$\text{Model 2: Reduced: } \hat{Y}_R = -0.5380 + 0.1593X_1 \quad SSE(R) = 75.129 \quad df_E(R) = n - g' = 19 - 2 = 17 \quad p - g = 3 - 1 = 2$$

The test of $H_0 : \beta_2 = \beta_3 = 0$ versus $H_A : \beta_2$ and/or $\beta_3 \neq 0$ is given below.

$$TS : F_{obs} = \frac{\left[\frac{75.129-67.846}{17-15}\right]}{\left[\frac{67.846}{15}\right]} = \frac{3.642}{4.523} = 0.805 \quad RR : F_{obs} \geq F_{.05,2,15} = 3.682 \quad P(F_{2,15} \geq 0.805) = .4634$$

After controlling for Area, neither Age or Effective guest rooms are associated with Energy Consumption.

R Commands and Output

```
### Commands
hotel_ec <- read.csv("http://www.stat.ufl.edu/~winner/data/hotel_energy.csv")
attach(hotel_ec); names(hotel_ec)

enrgcons <- enrgcons/1000000
area <- area/1000

## Full Model
hec.mod1 <- lm (enrgcons ~ area + age + effrooms)
summary(hec.mod1)
anova(hec.mod1)

## Reduced Model
hec.mod2 <- lm (enrgcons ~ area)
summary(hec.mod2)
anova(hec.mod2)

## Full versus Reduced F-test
anova(hec.mod2, hec.mod1)

### Output

> summary(hec.mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.252767   1.781202  -1.265 0.225260
```

Hotel	Y	X_1	X_2	X_3	$\hat{Y}_1$	e_1	$\hat{Y}_2$	e_2
1	1.95	43.00	6.00	44.64	5.61	-3.66	6.31	-4.36
2	1.05	19.98	16.00	85.33	3.31	-2.26	2.64	-1.60
3	4.25	46.53	7.00	115.52	6.48	-2.24	6.87	-2.63
4	2.13	20.96	6.00	110.34	2.48	-0.32	2.80	-0.67
5	2.79	24.21	5.00	230.27	3.82	-1.04	3.32	-0.53
6	13.83	112.20	4.00	188.73	17.11	-3.28	17.33	-3.50
7	5.56	45.00	3.00	78.03	5.70	-0.14	6.63	-1.07
8	4.00	28.55	6.00	54.37	3.27	0.73	4.01	-0.01
9	4.67	32.87	8.00	89.75	4.58	0.09	4.70	-0.03
10	8.92	59.41	5.00	167.23	8.82	0.10	8.92	0.00
11	6.87	45.00	10.00	368.20	7.83	-0.96	6.63	0.24
12	6.01	37.44	13.00	197.29	6.44	-0.43	5.42	0.59
13	8.19	50.83	4.00	83.31	6.74	1.45	7.56	0.63
14	11.74	68.00	13.00	187.53	11.02	0.72	10.29	1.45
15	14.84	78.87	8.00	206.12	12.25	2.58	12.02	2.82
16	5.37	28.45	13.00	128.30	4.42	0.95	3.99	1.37
17	13.52	70.00	4.00	228.74	10.56	2.95	10.61	2.91
18	3.88	20.00	5.00	85.81	2.04	1.85	2.65	1.24
19	10.57	50.00	12.00	120.28	7.67	2.90	7.42	3.15

Table 9.4: Hotel Energy Consumption Data, Fitted Values, and Residuals for Model 1 and Model 2

```
area         0.148709   0.029066   5.116 0.000127 ***
age          0.113045   0.134527   0.840 0.413924
effrooms     0.005777   0.007096   0.814 0.428315

Residual standard error: 2.127 on 15 degrees of freedom
Multiple R-squared:  0.7946,    Adjusted R-squared:  0.7535
F-statistic: 19.35 on 3 and 15 DF,  p-value: 2.049e-05

> anova(hec.mod1)
Response: enrgcons
          Df  Sum Sq Mean Sq F value    Pr(>F)
area       1 255.218 255.218 56.4258 1.854e-06 ***
age        1   4.286   4.286  0.9475    0.3458
effrooms   1   2.998   2.998  0.6628    0.4283
Residuals 15  67.846   4.523

> summary(hec.mod2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.53804    1.08509  -0.496    0.626
area         0.15925    0.02096   7.599 7.29e-07 ***

Residual standard error: 2.102 on 17 degrees of freedom
Multiple R-squared:  0.7726,    Adjusted R-squared:  0.7592
F-statistic: 57.75 on 1 and 17 DF,  p-value: 7.294e-07

> anova(hec.mod2)
Response: enrgcons
          Df  Sum Sq Mean Sq F value    Pr(>F)
area       1 255.218 255.218   57.75 7.294e-07 ***
Residuals 17  75.129   4.419

> anova(hec.mod2, hec.mod1)
```

```
Analysis of Variance Table
Model 1: enrgcons ~ area
Model 2: enrgcons ~ area + age + effrooms
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     17 75.129
2     15 67.846  2    7.2834 0.8051 0.4654
```

∇

9.2.4 Models With Categorical (Qualitative) Predictors

Often, one or more categorical variables are included in a model. If a categorical variable has m levels, there will need to be $m-1$ **dummy** or **indicator variables** to reflect the effects of the variable's levels. The variable will take on 1 if the i^{th} observation is in that level of the variable, 0 otherwise. Note that one level of the variable will have 0^s for all $m-1$ dummy variables, making it the reference category. The β^s for the other groups (levels of the qualitative variable) reflect the difference in the mean for that group with the reference group, controlling for all other predictors.

Note that if the qualitative variable has 2 levels, there will be a single dummy variable, and we can test for differences in the effects of the 2 levels with a t-test, controlling for all other predictors. If there are $m-1>2$ dummy variables, the F-test can be used to test whether all $m-1$ β^s are 0, controlling for all other predictors. An example is given below.

9.2.5 Models With Interaction Terms

When the effect of one predictor depends on the level of another predictor (and vice versa), the predictors are said to **interact**. The way to model interaction(s) is to create a new variable that is the product of the 2 predictors. Suppose the model has Y, and 2 numeric predictors: X_1 and X_2. Create a new predictor $X_3 = X_1X_2$. Now, consider the following model.

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 = \beta_0 + \beta_2 X_2 + \left(\beta_1 + \beta_3 X_2\right) X_1$$

The slope with respect to X_1 depends on the level of X_2, unless $\beta_3 = 0$, which can be tested with a t-test of $H_0 : \beta_3 = 0$. This logic extends to qualitative variables as well. Create cross-product terms between numeric (or other categorical) predictors with the $m-1$ dummy variables representing the qualitative predictor. Then the t-test ($m-1=1$) or F-test ($m-1 \geq 2$) can be conducted to test for interactions among predictors. This is demonstrated by adding males to the stature data below.

Example 9.10: Heights, Hand and Foot Lengths in Males and Females

In the stature study (Sanli, Kizilkanat, Boyan, et al. (2005), [51]), there were also 80 males, for a total of $n = 75+80 = 155$ adults. For these models, Y is height, X_1 is hand length, and X_2 is foot length. Create the dummy (indicator) variable $X_3 = 1$ if male, $X_3 = 0$ if female. Then consider three models: Common slopes and intercept by gender (Model 1), Common slopes but different intercepts by gender (Model 2), and Different slopes and intercepts by gender (Model 3). The models are given below.

$$\text{Model 1: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\text{Model 2: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$\text{Females: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad \text{Males: } E\{Y\} = (\beta_0 + \beta_3) + \beta_1 X_1 + \beta_2 X_2$$

$$\text{Model 3: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3$$

$$\text{Males: } E\{Y\} = (\beta_0 + \beta_3) + (\beta_1 + \beta_4) X_1 + (\beta_2 + \beta_5) X_2$$

The fitted equations and their Error Sums of Squares are given below (the regression coefficients are taken from the R output given below).

$$\text{Model 1: } \hat{Y}_F = \hat{Y}_M = 372.64 + 3.32X_1 + 2.58X_2 \quad SSE_1 = 189029 \quad df_{E1} = 155 - 3 = 152$$

$$\text{Model 2: } \hat{Y}_F = 581.99 + 2.81X_1 + 2.06X_2 \quad \hat{Y}_M = 621.55 + 2.81X_1 + 2.06X_2 \quad SSE_2 = 165341 \quad df_{E2} = 155 - 4 = 151$$

$$\text{Model 3: } \hat{Y}_F = 743.97 + 2.38X_1 + 1.73X_2 \quad \hat{Y}_M = 439.27 + 3.29X_1 + 2.38X_2 \quad SSE_3 = 157360 \quad df_{E3} = 155 - 6 = 149$$

Tests comparing the different models include Model 2 versus Model 1, where the null hypothesis is common slopes and intercepts (Model 1) and the alternative is common slopes and different intercepts (Model 2). The null hypothesis is $H_0 : \beta_3 = 0$.

$$TS : F_{12} = \frac{\left[\frac{189029 - 165341}{152 - 151}\right]}{\left[\frac{165341}{151}\right]} = \frac{23688}{1095} = 21.63 \quad RR : F_{12} \geq F_{.05,1,151} = 3.904$$

A second test comparing the different models include Model 3 versus Model 2, where the null hypothesis is common slopes and different intercepts (Model 2) and the alternative is different slopes and intercepts (Model 3). The null hypothesis is $H_0 : \beta_4 = \beta_5 = 0$.

$$TS : F_{23} = \frac{\left[\frac{165341 - 157360}{151 - 149}\right]}{\left[\frac{157360}{149}\right]} = \frac{3990.5}{1056} = 3.78 \quad RR : F_{23} \geq F_{.05,2,149} = 3.057 \quad P = .0251$$

The “full model” allowing for different slopes and intercepts for males and females gives the best fit.

R Commands and Output

```
### Commands
shf1 <- read.table("http://www.stat.ufl.edu/~winner/data/stature_hand_foot.dat",
header=F, col.names=c("idnum", "gender", "height", "hand", "foot"))
attach(shf1)
```

```
male <- 2-gender ### male = 1 if male, 0 if female

## Model 1: Common slope/intercept
shf.mod1 <- lm(height ~ hand + foot)
summary(shf.mod1)
anova(shf.mod1)

## Model 2: Common slope/Different intercept
shf.mod2 <- lm(height ~ hand + foot + male)
summary(shf.mod2)
anova(shf.mod2)

## Model 3: Different slope/intercept
shf.mod3 <- lm(height ~ hand + foot + male + I(hand*male) + I(foot*male))
summary(shf.mod3)
anova(shf.mod3)

anova(shf.mod1,shf.mod2) ### Compare Models 1 and 2
anova(shf.mod2,shf.mod3) ### Compare Models 2 and 3
```

```
### Output
> ## Model 1: Common slope/intercept
> shf.mod1 <- lm(height ~ hand + foot)
> summary(shf.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 372.6378    43.2581   8.614 8.41e-15 ***
hand          3.3175     0.3461   9.586  < 2e-16 ***
foot          2.5816     0.2490  10.370  < 2e-16 ***

Residual standard error: 35.26 on 152 degrees of freedom
Multiple R-squared:  0.8608,    Adjusted R-squared:  0.859
F-statistic: 470.1 on 2 and 152 DF,  p-value: < 2.2e-16

> anova(shf.mod1)
Analysis of Variance Table
Response: height
           Df  Sum Sq Mean Sq F value    Pr(>F)
hand        1 1035412 1035412  832.59 < 2.2e-16 ***
foot        1  133728  133728  107.53 < 2.2e-16 ***
Residuals 152  189029    1244

> ## Model 2: Common slope/Different intercept
> shf.mod2 <- lm(height ~ hand + foot + male)
> summary(shf.mod2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 581.9858    60.6099   9.602  < 2e-16 ***
hand          2.8116     0.3425   8.210 9.11e-14 ***
foot          2.0643     0.2587   7.979 3.43e-13 ***
male         39.5640     8.5064   4.651 7.16e-06 ***

Residual standard error: 33.09 on 151 degrees of freedom
Multiple R-squared:  0.8783,    Adjusted R-squared:  0.8758
F-statistic: 363.1 on 3 and 151 DF,  p-value: < 2.2e-16

> anova(shf.mod2)
Analysis of Variance Table
Response: height
           Df  Sum Sq Mean Sq F value    Pr(>F)
hand        1 1035412 1035412 945.602 < 2.2e-16 ***
```

```
foot        1  133728  133728 122.128 < 2.2e-16 ***
male        1   23687   23687  21.633 7.157e-06 ***
Residuals 151  165341    1095

> ## Model 3: Different slope/intercept
> shf.mod3 <- lm(height ~ hand + foot + male + I(hand*male) + I(foot*male))
> summary(shf.mod3)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     743.9696    83.7772   8.880 1.98e-15 ***
hand              2.3748     0.5104   4.653 7.17e-06 ***
foot              1.7271     0.3934   4.390 2.14e-05 ***
male           -304.7039   125.5987  -2.426   0.0165 *
I(hand * male)    0.9120     0.6809   1.340   0.1824
I(foot * male)    0.6537     0.5162   1.266   0.2074

Residual standard error: 32.5 on 149 degrees of freedom
Multiple R-squared:  0.8841,    Adjusted R-squared:  0.8803
F-statistic: 227.4 on 5 and 149 DF,  p-value: < 2.2e-16

> anova(shf.mod3)
Analysis of Variance Table
Response: height
                Df  Sum Sq Mean Sq  F value    Pr(>F)
hand             1 1035412 1035412 980.4040 < 2.2e-16 ***
foot             1  133728  133728 126.6232 < 2.2e-16 ***
male             1   23687   23687  22.4289 5.035e-06 ***
I(hand * male)   1    6288    6288   5.9538   0.01586 *
I(foot * male)   1    1694    1694   1.6036   0.20737
Residuals      149  157360    1056
>
> anova(shf.mod1,shf.mod2) ### Compare Models 1 and 2
Analysis of Variance Table

Model 1: height ~ hand + foot
Model 2: height ~ hand + foot + male
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    152 189029
2    151 165341  1     23687 21.633 7.157e-06 ***

> anova(shf.mod2,shf.mod3) ### Compare Models 2 and 3
Analysis of Variance Table

Model 1: height ~ hand + foot + male
Model 2: height ~ hand + foot + male + I(hand * male) + I(foot * male)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    151 165341
2    149 157360  2    7981.4 3.7787 0.02507 *
```

∇

9.3 Logistic Regression

When the response variable is binary (presence/absence of a characteristic), one model that is often fit is the logistic regression model. It fits the probability of a “Success” as an S-shaped, logistic function. In this formulation, we let π represent the probability of Success, which is bounded between 0 and 1. Define the

odds of Success as $o = \pi/(1-\pi)$ which represents the number of successes for every failure. Note that if $\pi = 0.9$, then $o = .9/.1 = 9$, and we expect 9 successes for every failure. If $\pi = 0.1$, then $o = .1/.9 = 1/9$, and we expect 1/9 successes for every failure. Odds can range from 0 to ∞, and log odds, also known as **logit** can range from $-\infty$ to ∞. The model is given below.

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad \Rightarrow \qquad \pi = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Statistical software packages can be used to obtain maximum likelihood estimates of the β^s and their estimated standard errors. Unlike linear regression, there are no closed form solutions, and they are obtained iteratively. The predicted probability of success for the i^{th} observation with predictor variables $X_{i1}, \ldots, X_{ip}$ is computed as follows (recall that the observed y_i is 0 or 1).

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\cdots+\hat{\beta}_p X_{ip}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\cdots+\hat{\beta}_p X_{ip}}} \qquad i = 1, \ldots, n$$

Tests can be conducted as in linear regression for individual coefficients, all coefficients, and subsets of coefficients. They use different distributions for the tests (z-tests in place of t-tests and Chi-square tests in place of F-tests).

Wald tests are printed out automatically by any standard statistical software package and can be reported as a z-statistic (R and Stata) or as a Chi-square statistic (SAS and SPSS). These are similar to the t-statistics in linear regression for the individual regression coefficients.

$$z\text{-test: } H_0: \beta_j = 0 \quad H_A: \beta_j \neq 0 \qquad z_j = \frac{\hat{\beta}_j}{\hat{SE}\{\hat{\beta}_j\}} \quad RR: |z_j| \geq z_{\alpha/2} \quad P = 2P\left(Z \geq |z_j|\right)$$

$$\chi^2\text{-test: } H_0: \beta_j = 0 \quad H_A: \beta_j \neq 0 \qquad X_j^2 = \left(\frac{\hat{\beta}_j}{\hat{SE}\{\hat{\beta}_j\}}\right)^2 \quad RR: X_j^2 \geq \chi^2_{1,\alpha} \quad P = P\left(\chi^2_{1,\alpha} \geq X_j^2\right)$$

Tests that all coefficients (besides β_0) and that a subset of coefficients are 0 can be conducted as **Likelihood-Ratio tests**. Once the regression coefficients are computed for the various models, and used to obtain the predicted values $\hat{\pi}_i$, the likelihood and the log-likelihood for the model is obtained as follows by standard packages.

$$\text{Likelihood: } L = \prod_{i=1}^{n} \hat{\pi}^{y_i}\left(1-\hat{\pi}_i\right)^{1-y_i} \qquad \text{log-likelihood: } l = log(L)$$

Then for any Full and Reduced models, we can compute l_F and l_R, where necessarily $l_F \geq l_R$. To test $H_0: \beta_1 = \cdots = \beta_p = 0$, we fit a full model with all p predictors and a reduced model with only an intercept (probability of success is equal for all individual cases). The test is conducted as follows.

$$H_0: \beta_1 = \cdots = \beta_p = 0 \quad H_A: \text{ Not all } \beta_j = 0 \qquad TS: X^2 = -2\left(l_R - l_F\right) \quad RR: X^2 \geq \chi^2_{\alpha,p} \quad P = P\left(\chi^2_{\alpha,p} \geq X^2\right)$$

Flight	Temp(X)	O-Ring Fail (Y)	Flight	X	Y	Flight	X	Y
1	66	0	9	57	1	17	70	0
2	70	1	10	63	1	18	81	0
3	69	0	11	70	1	19	76	0
4	68	0	12	78	0	20	79	0
5	67	0	13	67	0	21	75	1
6	72	0	14	53	1	22	76	0
7	73	0	15	67	0	23	58	1
8	70	0	16	75	0			

Table 9.5: Challenger O-Ring Failure/Temperature Data for n=23 missions pre-disaster

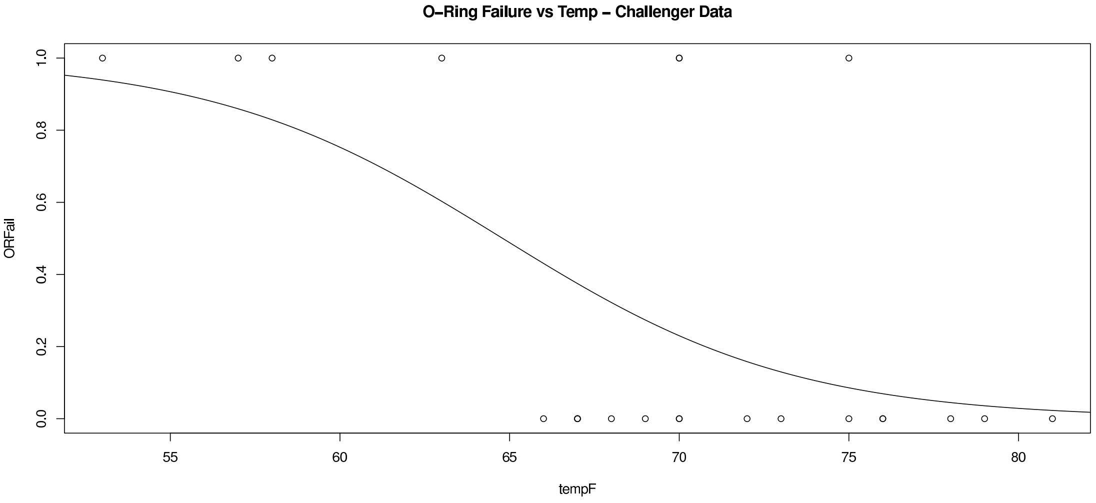

Figure 9.5: Data and fitted equation for Challenger O-Ring Failure/Temperature data

We will demonstrate these with two examples, one with a single predictor, the other with a set of predictors.

Example 9.11: Pre-Challenger O-Ring Failure and Temperature

Prior to the space shuttle Challenger's explosion after lift-off in January, 1986, the shuttle had flown on $n = 23$ successful flights (Dalal, et al (1989), [18]). The flights were classified by whether there had been field-joint O-ring failure in the connection of the shuttle to the solid rocket booster (Y=1 if yes, 0 if no) and the temperature at lift-off (X=degrees F). Data are given in Table 9.5. Note that the ”Success” in this case is actually an unfavorable event, and the model is relating the probability of the event occurring as a function of temperature. The R program and output are given below. The result is that there is a significant association between temperature and the event of O-ring failure. Also, since there is evidence that $\beta_1 < 0$, that the probability of O-ring failure decreases as temperature increases. A plot of the data and the fitted equation is given in Figure 9.5. The fitted equation is:

$$\hat{\pi} = \frac{e^{15.0429-0.2322X}}{1+e^{15.0429-0.2322X}}$$

R Commands and Output

```
### Commands
chal.data <- read.table("http://stat.ufl.edu/~winner/data/challenger.dat",
                        header=F,
                       col.names=c("flight","tempF","ORFnum","ORFail"))
attach(chal.data)

chal.mod <- glm(ORFail ~ tempF, binomial("logit"))
summary(chal.mod)
confint(chal.mod)

### Output
> summary(chal.mod)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
tempF        -0.2322     0.1082  -2.145   0.0320 *

> confint(chal.mod)
Waiting for profiling to be done...
                2.5 %      97.5 %
(Intercept)  3.3305848 34.34215133
tempF       -0.5154718 -0.06082076
```

$$\nabla$$

Example 9.12: Presence/Absence of Gold Deposits in India

A study was conducted (Sahoo and Pandalai (1999), [50]) to detemine whether the presence/absence of gold at locations in India is related to three predictors: Arsenic level (X_1), Antimony level (X_2) and lineament presence ($X_3 = 1$ if yes, 0 if no) based on a sample of $n = 64$ locations. We consider three models: one with only an intercept, one with only arsenic as a predictor, and one with all 3 predictors.

$$\text{Model 0: } \pi = \frac{e^{\beta_0}}{1+e^{\beta_0}} \quad \text{Model 1: } \pi = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}} \quad \text{Model 3: } \pi = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}$$

First, we will compare Models 0 and 3 to test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, that is that gold presence is not associated with any of the predictors. Second, we will compare Models 1 and 3 to test $H_0 : \beta_2 = \beta_3 = 0$, that is that gold presence is not associated with antimony or lineament, after controlling for arsenic. The log-likelihoods for the three models are $l_0 = -43.8601$, $l_1 = -11.3014$, and $l_3 = -7.0972$, respectively. Now, we conduct the two tests described above. Note that for the second test, the degrees of freedom is the number of restrictions under the null hypothesis, which is 2 ($\beta_2 = \beta_3 = 0$).

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \qquad TS : X^2 = -2\,((l_0 - l_3) = -2(-43.8601 - (-7.0972)) = 73.526$$

$$RR : X^2 \geq \chi^2_{.05,3} = 7.815 \quad P = P\left(\chi^2_{.05,3} \geq 73.526\right) < .0001$$

$$H_0 : \beta_2 = \beta_3 = 0 \qquad TS : X^2 = -2\left((l_1 - l_3\right) = -2(-11.3014 - (-7.0972)) = 8.408$$

$$RR : X^2 \geq \chi^2_{.05,2} = 5.991 \quad P = P\left(\chi^2_{.05,2} \geq 8.408\right) = .0149$$

The R program and output are given below. Note that the P-values for the Wald tests of $\beta_2 = 0$ ($P = .0516$) and $\beta_3 = 0$ ($P = .0909$) for Model 3 are both above .0500. Those tests are controlling for each other, while the Chi-square test above tests that they are both simultaneously 0, controlling only for Arsenic ($P = .0149$).

R Commands and Output

```
### Commands
gold.data <- read.table("http://stat.ufl.edu/~winner/data/gold_target1.dat",
                       header=F,
                      col.names=c("arsenic","antimony","lineament","gold"))
attach(gold.data)

gold.mod0 <- glm(gold ~ 1, binomial("logit"))
summary(gold.mod0)
(ll.mod0 <- logLik(gold.mod0))

gold.mod1 <- glm(gold ~ arsenic, binomial("logit"))
summary(gold.mod1)
(ll.mod1 <- logLik(gold.mod1))

gold.mod3 <- glm(gold ~ arsenic + antimony + lineament, binomial("logit"))
summary(gold.mod3)
(ll.mod3 <- logLik(gold.mod3))

anova(gold.mod0, gold.mod3)
anova(gold.mod1, gold.mod3)

### Output
> gold.mod0 <- glm(gold ~ 1, binomial("logit"))
> summary(gold.mod0)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2513     0.2520  -0.997    0.319

> (ll.mod0 <- logLik(gold.mod0))
'log Lik.' -43.86011 (df=1)
>
> gold.mod1 <- glm(gold ~ arsenic, binomial("logit"))
> summary(gold.mod1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.9460     0.9641  -4.093 4.26e-05 ***
arsenic       1.3456     0.3354   4.012 6.02e-05 ***

> (ll.mod1 <- logLik(gold.mod1))
'log Lik.' -11.30143 (df=2)

> gold.mod3 <- glm(gold ~ arsenic + antimony + lineament, binomial("logit"))
> summary(gold.mod3)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.6096     3.1661  -2.403   0.0162 *
arsenic       1.2046     0.4899   2.459   0.0139 *
```

```
antimony      1.4210     0.7301   1.946   0.0516 .
lineament     3.1973     1.8911   1.691   0.0909 .

> (ll.mod3 <- logLik(gold.mod3))
'log Lik.' -7.097155 (df=4)
>
> anova(gold.mod0, gold.mod3)
Analysis of Deviance Table

Model 1: gold ~ 1
Model 2: gold ~ arsenic + antimony + lineament
  Resid. Df Resid. Dev Df Deviance
1        63     87.720
2        60     14.194  3   73.526
> anova(gold.mod1, gold.mod3)
Analysis of Deviance Table

Model 1: gold ~ arsenic
Model 2: gold ~ arsenic + antimony + lineament
  Resid. Df Resid. Dev Df Deviance
1        62     22.603
2        60     14.194  2   8.4085
```

∇

Bibliography

[1] Agresti, A. (2002). *Categorical Data Analysis. 2nd Ed.* Wiley, New York.

[2] Agresti, A. and B. Coull (1998). “Approximate is Better than ‘Exact’ for Interval Estimation of Binomial Proportions,” *The American Statistician*, Vol. 52, #2 pp. 119-126.

[3] Agresti, A. and L. Winner (1997). “Evaluating Agreement and Disagreement Among Movie Reviewers,” *Chance*, Vol. 10, pp. 10-14.

[4] Ahonen, H., A.J. Stow, R.G. Harcourt, and I. Charrier (2014). “Adult Male Australian Sea Lion Barking Calls Reveal Clear Geographical Variations,” *Animal Behaviour*, Vol. 97, pp. 229-239.

[5] Ali, E., W. Guang, and A. Ibrahim (2014). “Empirical Relations Between Compressive Strength and Microfabric Properties of Amphibolites Using Multivariate Regression, Fuzzy Inference, and Neural Networks: A Comparative Study,” *Engineering Geology*, Vol. 183, pp. 230-240.

[6] Barnum, D.T. and J.M. Gleason (1994). “The Credibility of Drug Tests: A Multi-Stage Bayesian Analysis,” *Industrial and Labor Relations Review*, Vol. 47, #4, pp. 610-621.

[7] Ben Ticha, M., W. Haddar, N. Meksi, A. Guesmi, and M.F. Mhenni (2016). ”Improving dyeability of modified cotton fabrics by the natural aqueous extract from red cabbage using ultrasonic energy,” *Carbohydrate Polymers*, Vol. 154, pp. 287-295.

[8] Berenson, J.R., A. Lichtenstein, L. Porter, et al. (1996). “Efficacy of Pamidronate in Reducing Skeletal Events in Patients with Advanced Myeloma,” *New England Journal of Medicine*, Vol. 334, pp. 488-493.

[9] Bhatnagar, A. and V.K. Mehta (2007). “Efficacy of Deltamethrin and Cyfluthrin Impregnated Cloth Over Uniform Against Mosquito Bites,” *Medical Journal Armed Forces India*, Vol. 63, pp. 120-122.

[10] Broders, A.C. (1920). “Squamous-Cell Epithelioma of the Lip,” *Journal of the American Medical Association*, Vol. 74, pp. 656-664.

[11] Bruce, A.C., J.E.V. Johnson, and J. Peirson (2012). “Recreational versus Professional Bettors: Performance Differences and Efficiency Implications,” *Economic Letters*, Vol. 114, pp. 172-174.

[12] Cameron, A.C. and P.K. Trivedi (2005). *Microeconometrics: Methods and Applications.* Cambridge, Cambridge.

[13] Chambers, G.F. (1889). *Handbook of Astronomy, 4th Ed.* Oxford.

[14] Chihara, L. and T. Hesterberg (2011). *Mathematical Statistics with Resampling and R.* Wiley, Hoboken, NJ.

[15] Clarke, R.D. (1946). “An Application of the Poisson Distribution,” *Journal of the Institute of Actuaries*, Vol. 72, p. 481.

[16] Cohen, A.M. (1996). “The Hands of Blues Guitarists,” *American Music*, Vol. 14, #4, pp. 455-479.

[17] Culp Jr., R.L. and D. Pollage (2002). “The Rhetoric of Strict Products Liability versus Negligence: An Empirical Analysis,” *New York University Law Review*, Vol. 77, #4, pp. 874-961.

[18] Dalal, S.R., E.B. Fowlkes, B. Hoadley (1989).“”Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure,” *Journal of the American Statistical Association*, Vol. 84, #408, pp. 945-957.

[19] Dickey, D.A. and and J.T. Arnold (1995). “Teaching Statistics with Data of Historic Significance: Galileo’s Gravity and Motion Experiments,” *Journal of Statistics Education*, Vol. 3, #1.

[20] Dror, I.E., C. Champod, G. Langenburg, D. Charlton, H. Hunt, and R. Rosenthal (2011). “Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a ‘Target’ Comparison,” *Forensic Science International*, Vol. 208, pp. 10-17.

[21] Durante, C., C. Baschieri, L. Bertacchini, D. Bertelli, M. Cocchi, A. Marchetti, D. Manzini, G. Papotti, S. Sighinolfi (2015). “An Analytical Approach to Sr Isotope Ratio Detemination in *Lambrusco* Wines for Geographic Traceability Purposes,” *Food Chemistry*, Vol. 173, pp. 557-563.

[22] Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

[23] Eisenberg, T., S.P. Garvey and M.T. Wells (2001). “Forecasting Life and Death: Juror Race, Religion, and Attitude Toward the Death Penalty,” *The Journal of Legal Studies*, Vol. 30, pp. 277-311.

[24] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Vol.1, 3rd. Ed.* Wiley, New York.

[25] Galton, F. (1886). “Regression Towards Mediocrity in Hereditary Stature,” *The Journal of the Anthropological Institute of Great Britain and Ireland*, Vol. 15, pp. 246-263.

[26] Gilovich, T. R. Vallone, and A. Tvesky (1985). “The Hot Hand in Basketball: On the Misperception of Random Sequences,” *Cognitive Psychology*, Vol. 17, #3, pp. 295-314.

[27] Gueguen, N. and C. Jacob (2013). “Color and Cyber-Attractiveness: Red Enhances Men’s Attraction to Women’s Personal Ads,” *Color Research and Application*, Vol. 38, #4, pp. 309-312.

[28] Halley, E. (1694). “An Account of the Evaporation of Water, as It Was Experimented in Gresham Colledge in the Year 1693. With Some Observations Thereon,” *Philosophical Transactions*, Vol. 18, pp. 183-190.

[29] Hammond, E.C. and D. Horn (1954), “The Relationship Between Human Smoking Habits and Death Rates,” *Journal of the American Medical Association*, Vol. 155, pp. 1316-1328.

[30] Hanley, J.A. (2004). “Transmuting Women into Men: Galtons Family Data on Human Stature,” *The American Statistician*, Vol. 58, #3, pp. 237-243.

[31] Holland, T.H. (1902). “The Kanets of Kulu and Lahoul, Punjab: A Study in Contact-Metamorphism,” *The Journal of the Anthropological Institute of Great Britain and Ireland*, Vol. 32, pp.96-123.

[32] Hollander, M. and D.A. Wolfe (1999). *Nonparametric Statistical Methods, 2nd. Ed.*, Wiley, New York.

[33] Kahneman, D., P. Slovic, and A. Tversky (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, UK.

[34] Kahneman, D. and A. Tversky (1984). "Choices, Values, and Frames," *American Psychologist*, Vol. 35, #4, pp. 341-350.

[35] Kamimura, A., T. Takahishi, and Y. Watanabe (2000). "Investigation of Topical Application of Procyanidin B-2 from Apple to Identify its Potential Use as a Hair Growing Agent," *Phytomedicine*, Vol. 7, #6, pp. 529-536.

[36] Koyama, K., H. Hokunan, M. Hasegawa, S. Kawamura, and S. Koseki (2016). "Do Bacterial Cell Numbers Follow a Theoretical Poisson Distribution? Comparison of Experimentally Obtained Numbers of Single Cells with Random Number Generation via Computer Simulation," *Food Microbiology*, Vol. 60, pp. 49-53.

[37] Liang, D.G., J.R. Dusseldorp , C. van Schalkwyk , S. Hariswamy, S. Wood, V. Rose, P. Moradi (2016). "Running Barbed Suture Quilting Reduces Abdominal Drainage in Perforator-Based Breast Reconstruction," *Journal of Plastic, Reconstructive & Aesthetic Surgery*, Vol. 69, pp. 42-47.

[38] Lim, T.-S. and W.Y. Loh (1996). "A Comparison of Tests of Equality of Variances," *Computational Statistics and Data Analysis*, Vol. 22, pp. 287-301.

[39] Lin, W. and J. Wang (2012). "An Integrated 3D Log Processing Optimization System for Hardwood Sawmills in Central Appalachia, USA," *Computers and Electronics in Agriculture*, Vol. 82, pp. 61-74.

[40] Lister, J. (1870). "Effects of the Antiseptic System of Treatment on the Salubrity of a Surgical Hospital," *The Lancet*, Vol.1, pp. 4-6,40-42.

[41] Masella, P., Guerrini, S. Spinelli, L. Calamai, P. Spugnoli, F. Illy, A. Parenti (2015). "A New Espresso Brewing Method," *Journal of Food Engineering*, Vol. 146, pp. 204-208.

[42] Mehrgini, B., H. Memarian, M.B. Dusseault, A. Ghavidel, and M. Heydarizadeh (2016). "Geomechanical Characteristics of Common Reservoir Caprock in Iran (Gachsaran Formation), Experimental and Statistical Analysis," *Journal of Natural Gas Science and Engineering*, Vol. 34, pp. 898-907.

[43] Michelson, A.A., F.G. Pease, F. Pearson (1935). "Measurement of the Velocity of Light in a Partial Vacuum," *Astrophysical Journal*, Vol. 82, pp. 26-61.

[44] Ott, R.L. and M. Longnecker (2016). *Statistical Methods & Data Analysis, 7th Ed.* Cengage Learning, Boston.

[45] Page, L., D.A. Savage, and B. Torgler (2014). "Variation in Risk Seeking Behaviour Following Large Losses: A Natural Experiment," *European Economic Review*, Vol. 71, pp. 121-131.

[46] Pachel, C. and J. Neilson (2010). "Comparison of Feline Water Consumption Between Still and Flowing Water Sources: A Pilot Study," *Journal of Veterinary Behavior*, Vol. 5, pp. 130-133.

[47] Peckmann, T.R., S. Scott, S. Meek, and P. Mahakkanukrauh (2017). "Sex Estimation from the Scapula in a Contemporary Thai Population: Applications for Forensic Anthropology," *Science and Justice*, Vol. 57, pp. 270-275.

[48] Poburka, P.J., R.R. Patel, and D.M. Bless (2017). "Voice-Vibratory Assessment With Laryngeal Imaging (VALI) Form: Reliability of Rating Stroboscopy and High-speed Videoendoscopy," *Journal of Voice*, Vol. 31, No. 4, pp. 513.e1513.e14.

[49] Rusanganwa, J. (2013). "Multimedia as a Means to Enhance Teaching Technical Vocabulary to Physics Undergraduates in Rwanda," *English for Specific Purposes*, Vol. 32, pp. 36-44.

[50] Sahoo, N.R. and H.S. Pandalai (1999). "Integration of Sparse Geologic Information in Gold Targeting Using Logistic Regression Analysis in the Hutti-Maski Schist Belt, Raichur, Karnataka, India - A Case Study" *Natural Resources Research*, Vol. 8, #3, pp. 233-250.

[51] Sanli, G.S., E.D. Kizilkanat, N. Boyan, E.T. Ozsahin, M.G. Bozkir, R. Soames, H. Erol, and O. Oguz (2005). "Stature Estimation Based on Hand Length and Foot Length," *Clinical Anatomy*, Vol. 18, #8, pp. 589-596.

[52] Scheaffer, R.L., W. Mendenhall, and L. Ott (1990). *Elementary Survey Sampling, 4th Ed.* PWS-KENT, Boston.

[53] Sheldrake, R., P. Smart, and L. Avraamides (2015). "Automated Tests for Telephone Telepathy Using Mobile Phones," *Explore*, Vol. 11, #4, pp. 310-319.

[54] Short, J. (1763). "Second Paper Concerning the Parallax of the Sun Determined from the Observations of the Late Transit of Venus, in Which This Subject is Treated of More at Length, and the Quantity of the Parallax More Fully Ascertained," *Philosophical Transactions*, Vol. 53, pp. 300-345. (Tables on pages 310,316,325).

[55] Stigler, S.M. (1977). "Do Robust Estimators Work With Real Data?," *The Annals of Statistics*, Vol. 5, #6, pp. 1055-1098.

[56] Storm, L., P.E. Tressoldi, and L. Di Risio (2010). "Meta-Analysis of Free Response Studies, 1992:2008: Assessing the Noise Reduction Model in Parapsychology," *Psychological Bulletin*, Vol. 136, No. 4, pp. 471-485.

[57] Teerapatsakul, C., C. Bucke, R. Parra, T. Keshavarz, and L. Chitradon (2008). "Dye Decolorisation by Laccase Entrapped in Copper Alginate," *World Journal of Microbiology and Technology*, Vol. 24, pp. 1367-1374.

[58] Thorndike, F. (1926). "Applications of Poisson's Probability Summation," *Bell System Technical Journal*, Vol. 5, pp. 604-624.

[59] Walter, S.R., W.T.M. Dunsmuir, and J.I. Westbrook (2015). "Studying Interruptions and Multitasking in situ: The Untapped Potential of Quantitative Observational Studies," *International Journal of Human-Computer Studies*, Vol. 79, pp. 118-125.

[60] Winner, L. (2006). "NASCAR Winston Cup Race Results for 1975-2003," *Journal of Statistical Education*, Vol. 14, #3.

[61] Xin, Y., S. Lu, N. Zhu, and W. Wu (2012). "Energy Consumption Quota of Four and Five Star Luxury Hotels Buildings in Hainan Province, China," *Energy and Buildings*, Vol. 45, pp. 250-256.

[62] Zouid, I., R. Siret, F. Jourjon, E. Mehinagic, and L. Rolle (2013). "Impact of Grapes Heterogeneity According to Sugar Level on Both Physical and Mechanical Berries Properties and Their Anthocyanins Extractability at Harvest," *Journal of Texture Studies*, Vol. 44, pp. 95-103.