Introduction to Applied Statistical Methods

Larry Winner
University of Florida
Department of Statistics

August 24, 2017

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.

1. Define the problem/research question of interest, including what to measure and all relevant conditions or groups to study.
2. Collect the data by means of a controlled experiment, observational study, or sample survey.
3. Summarize the data numerically in tabular form and/or graphically.
4. Analyze, interpret, and communicate the studies 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.

- Political polls report the sample proportion (typically as a percentage) of voters who favor a candidate along with a measure of uncertainty with respect to the population proportion.
- A study compared 3 electronic reader models, each at 4 illumination levels in a sample of 60 subjects, measuring the times to read a document. The goal was to compare the effects of the models and illumination levels in the general population [6].
- Many studies have been conducted involving extrasensory perception (ESP). In a typical study, there are 4 choices of what target the sender is viewing and the receiver must identify which target was being viewed. Researchers wish to determine whether the true proportion of successful trials exceeds 1/4 from a sample of trials [28].
- Studies are conducted to measure general consistency within and between evaluators when assessing common items (fingerprints, x-rays, foods/beverages) based on sampled judges and targets [9].

Note that populations can be "fixed", a well defined and identified population of units (e.g. all National Hockey League players for the 2014-2015 season) or "conceptual" (e.g. all people with a particular condition currently or in the near future). In our work, we will often make use of taking random samples from fixed populations to understand the properties of statistical procedures as they are applied to different samples from a given population.

1.2 Data Collection

Once a research question has been made, then data is 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 and its sequel. 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 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.

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: Electronic Reader Reading Task Times by Model and Illumination

An experiment was conducted to compare reading times for a long duration reading task (Chang, Chou, and Shieh (2013) [6]). There were two factors: e-reader model with 3 levels (Sony PRS 700, Amazon Kindle DX, iRex 1000s) and 4 illumination levels (200 lx, 500, 1000, 1500). Thus there were 12 treatments (combinations of e-reader and illumination level). There were a total of 60 subjects, who were randomly assigned so that 5 subjects were assigned to each treatment (each subject read only 1 reader under only 1 illumination level). The response was the time to read the document in seconds.

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) [30].

In many research areas, data are collected through **Sample Surveys**. In particular, they are often used in Public Opinion, by Government Bureaus, Business, and Recreational Services. Unless surveys are based

on some sort of sampling based method, they are generally not reliable for making inferences regarding a population.

It should be noted that certain problems tend to arise with surveys. The primary problem is **nonresponse**. If the individuals who do not respond tend to be different form those who do respond, then any estimates of population based quantities will be biased. Also, when the questions are "sensitive" such as illegal behavior, there will tend to be **response bias**. **Recall bias** occurs when some sampled elements are more likely to recall a previous experience than others. This can effect observed associations in retrospective surveys. Needless to say wording of questions can have a large impact on responses.

Some commonly used sampling methods are as follow.

Simple Random Sampling All possible samples of size n from a population of size N are equally likely. This needs a frame listing all elements of the population and a random number generator.

Stratified Random Sampling Elements of the population are classified by group (strata) and simple random samples are taken within each group.

Cluster Sampling Elements of the population are classified by cluster (possibly physical location) and a random sample of clusters is taken. Elements within the sampled clusters are the sampled units.

Systematic Sampling When elements of the population are in a sequence, a random starting point is selected, and every k^{th} subsequent element is sampled.

Note that these techniques are often applied in combination in many government/business/political surveys. Also, these techniques generalize to taking samples of individuals or elements from any population to be observed and measured. For instance, in quality control, items may be sampled and tested from an assembly line by systematic sampling.

All methods covered in this course are based on simple random sampling. Some adjustments for estimates and standard errors are used for the other sampling plans. For a detailed and accessible coverage of sampling, see e.g. Scheaffer, Mendenhall, and Ott (1990) [26].

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

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

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 1.2: 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) [25]. Table 1.1 contains information on the 30 subjects in the study. These include: study 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, it is 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 on numeric variables, with 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 represent smoothed histograms describing the proportion of measurements within some distance of each point on the continuum.

Example 2.1: Charlotte, NC Traffic Stops - 2016

Data for a population of 79884 traffic stops in Charlotte, North Carolina in 2016 were obtained from Data.gov. There were 10 possible reasons for the traffic stops (including a category 'Other.' A pie chart (Figure 2.1) and a bar chart (Figure 2.2) are displayed. Note that the pie chart does a very poor job with the categories "DWI" and "Check Point." Pie charts should generally be avoided. It is clear that Registration and Speed violations are the most often occuring reasons.

The R commands are given below.

```
### Commands

## Read data off web page, attach file as data frame, and list variable names
clt2016 <- read.csv("http://www.stat.ufl.edu/~winner/data/trafficstop.csv")
attach(clt2016); names(clt2016)

head(clt2016)    ## Print first 6 observations

## Assign labels to the Catgories of Reasons for Stop
labels.RsnStop <- c("ChkPnt","DWI","Invstgtn","Other",
   "SafeMove","SeatBelt","Speed","StopLgtSgn","VhclMove","Rgstrtn")

## Obtain and print frequency table for Reasons for Stop
(table.RsnStop <- table(RsnStop))

## Pie chart based on Table and Labels from above
pie(table.RsnStop, labels.RsnStop, main="Pie Chart - CLT Traffic Stops")

## Bar chart based on Table and Labels from above (cex shrinks size of levels)
barplot(table.RsnStop, names.arg=labels.RsnStop,
   main="Bar Chart - CLT Traffic Stops", xlab="Reason", ylab="Frequency",
   cex.names=0.6)

### Output

> (table.RsnStop <- table(RsnStop))
RsnStop
    1     2     3     4     5     6     7     8     9    10
  286   114  1992  1926  4827   631 22222  7946  7535 32405
```

∇

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. As the height range is not particularly wide, a uniform random number between -0.5 and +0.5 is added to each player's height to make the variable more continuous. Due to a larger spread in weights, no such adjustment is made. A histogram is given in Figure 2.3. The histogram is symmetric and mound-shaped, centered between 26 and 28.

The R Program is given below.

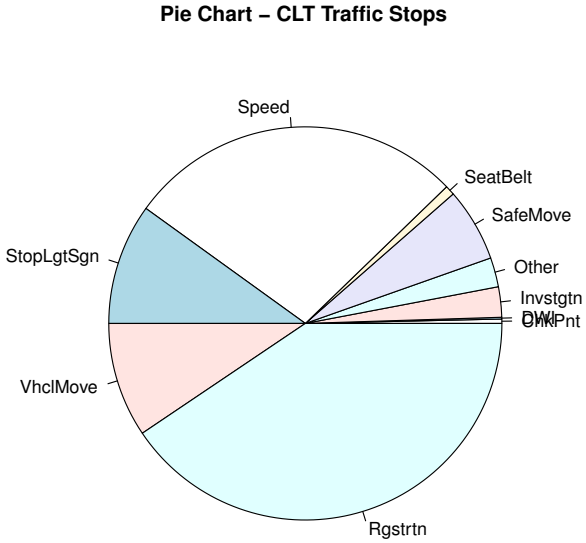

Figure 2.1: Pie Chart for Charlotte, NC traffic stops by Reason for Stop

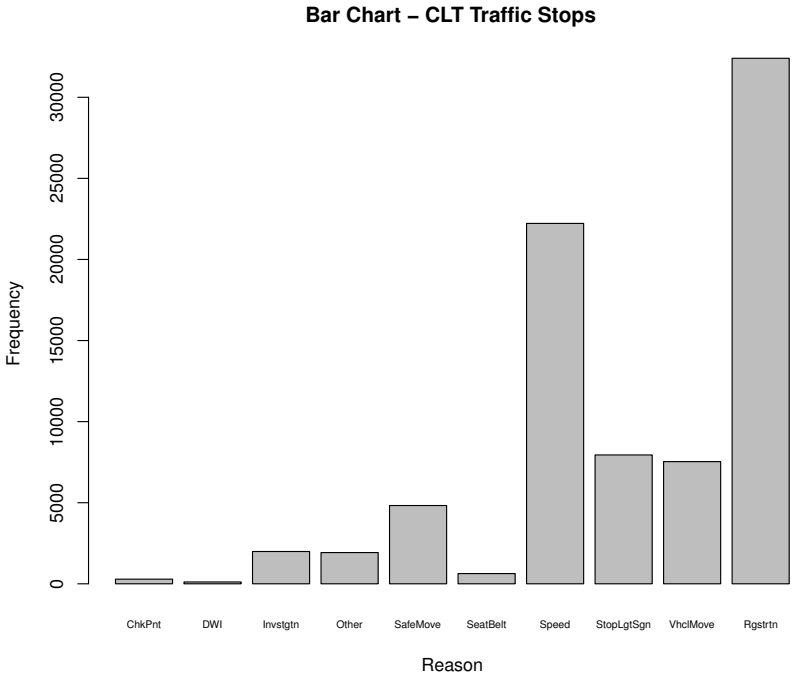

Figure 2.2: Bar Chart for Charlotte, NC traffic stops by Reason for Stop

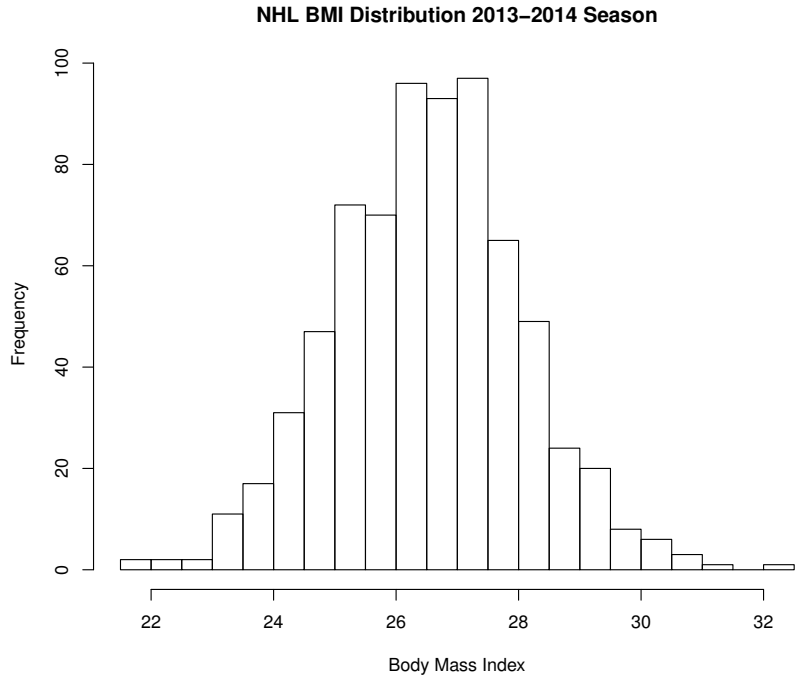

Figure 2.3: Body Mass Index for 2013/2014 season National Hockey League Players

```
### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Generate random values (-0.5 to 0.5) to add to Height
set.seed(1234)
N <- NROW(nhl)
Height.dev <- 0.5 - runif(N)
Height <- Height + Height.dev

### Compute BMI
bmi.nhl <- 703 * Weight / (Height^2)

### Obtain histogram
hist(bmi.nhl, breaks=30, xlab="Body Mass Index",
main="NHL BMI Distribution 2013-2014 Season")
```

∇

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

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**. The box-plot identifies from bottom to top the following elements.

1. Minimum: Bottom of line at bottom of plot
2. Range for slowest 25% of participants: Line below box
3. 25th percentile: Bottom line of box
4. Range for the 25th to 50th percent of participants: 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: Between second horizontal line and top of box
7. 75th percentile: Top line of the box
8. Range for 75th to 100th percent of participant: Line extends to either the Maximum speed or 1.5 times the distance between 75th and 25th percentiles (height of the box), whichever is lowest. Stars represent outlying measurements (very fast runners).

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.

The R commands are given below.

```
### Commands

## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)

## Obtain mean and standard deviation by gender
tapply(mph,Gender,mean)
tapply(mph,Gender,sd)

## Obtain the densities (for plotting) of mph by gender
d.F <- density(mph[Gender=="F"])
d.M <- density(mph[Gender=="M"])

## Set up a 2x2 grid for plots
par(mfrow=c(2,2))
## Histograms for Female and Male mph
hist(mph[Gender=="F"],breaks=25,main="Histogram of Female Speeds",
   xlab="Female Speeds")
hist(mph[Gender=="M"],breaks=25,main="Histogram of Male Speeds",
   xlab="Male Speeds")
## Density Plots for Female and Male mph
plot(d.F,
  main="Kernel Density Plot of Female Speeds")
plot(d.M,
  main="Kernel Density Plot of Male Speeds")
```

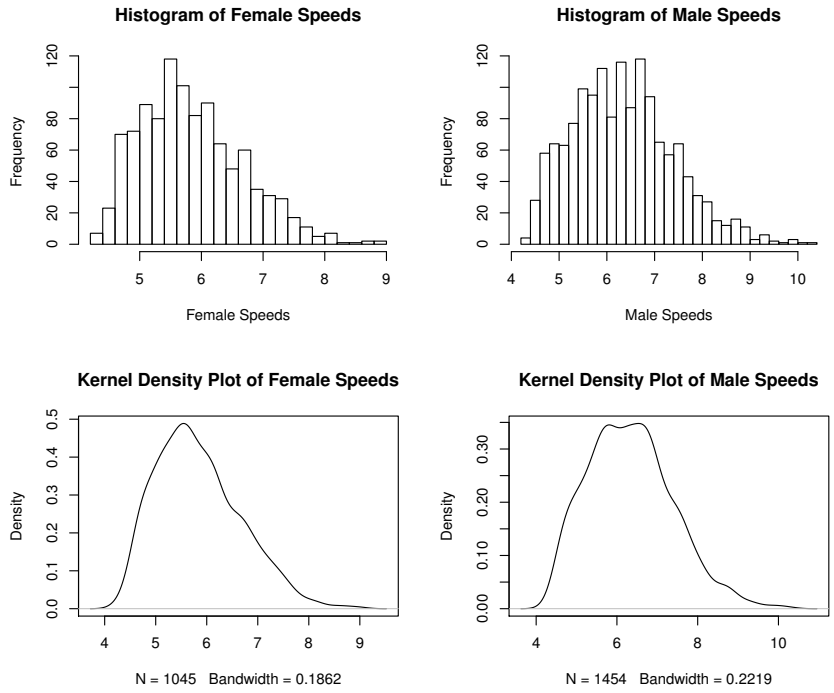

Figure 2.4: Histograms and density plots of Rock and Roll marathon speeds by gender

```
## Reset Plot to 1 per page and obtain side-by-side boxplots
## Gender is a factor vatiable (on the x-axis)
par(mfrow=c(1,1))
plot(Gender, mph, main="Box Plots of Speed(mph) by Gender")

## Obtain a "violin plot" - a "smoothed density" version of boxplot
require(ggplot2)
ggplot(rr.mar, aes(y=mph, x=Gender)) + geom_violin()

### Output

> ## Obtain mean and standard deviation by gender
> tapply(mph,Gender,mean)
       F        M
5.839839 6.336979
> tapply(mph,Gender,sd)
        F         M
0.8310405 1.0576868
```

∇

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 daily, weekly, monthly, quarterly, or annual data. 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

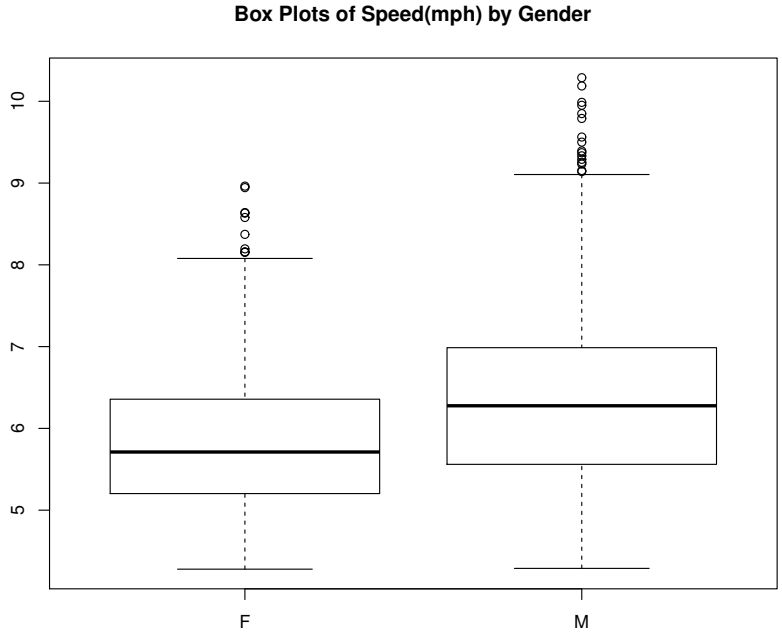

Figure 2.5: Side-by-side box plots of Rock and Roll marathon speeds by gender

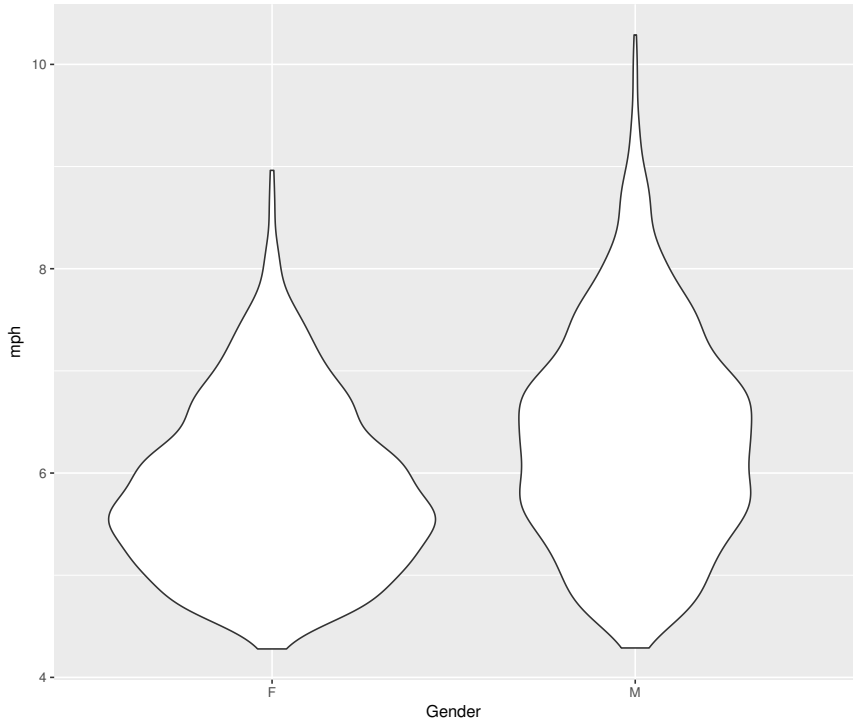

Figure 2.6: Side-by-side violin plots of Rock and Roll marathon speeds by gender

series, there certainly appears to be evidence of an increasing trend over approximately the second half of the series (after about 1970).

The R commands are given below.

```
### Commands

## Read data and set up data frame
mw1 <- read.csv("http://www.stat.ufl.edu/~winner/data/miami_weather.csv")
attach(mw1); names(mw1)

## Obtain mean temperature by year
(yearMeanTemp <- aggregate(meantemp ~ year, mw1, mean))

## Stack Monthly and Annual plots
par(mfrow=c(2,1))

## Monthly Plot gives only "y", not "x", this is a line plot
## type="l" draws lines meeting points
plot(meantemp, type="l", main="Miami Monthly Mean Temp (F) 1949-2014",
  xlab="Month", ylab="Mean Temperature")

## Plot "x"=Year (first column of yearMeanTemp) and
##  "y"=mean temp (second column of yearMeanTemp0
plot(yearMeanTemp[,1], yearMeanTemp[,2],
  type="l", main="Miami Yearly Mean Temp (F) 1949-2014",
  xlab="Year", ylab="Mean Temperature")

### Output (condensed)

> (yearMeanTemp <- aggregate(meantemp ~ year, mw1, mean))
   year meantemp
1  1949 76.11667
2  1950 75.15833
3  1951 75.24167
...
64 2012 77.30833
65 2013 77.88333
66 2014 77.54167
```

Data maps are very popular as more and more spatial datasets are available. Figure 2.8 displays Bigfoot sightings for the 50 United States.

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.

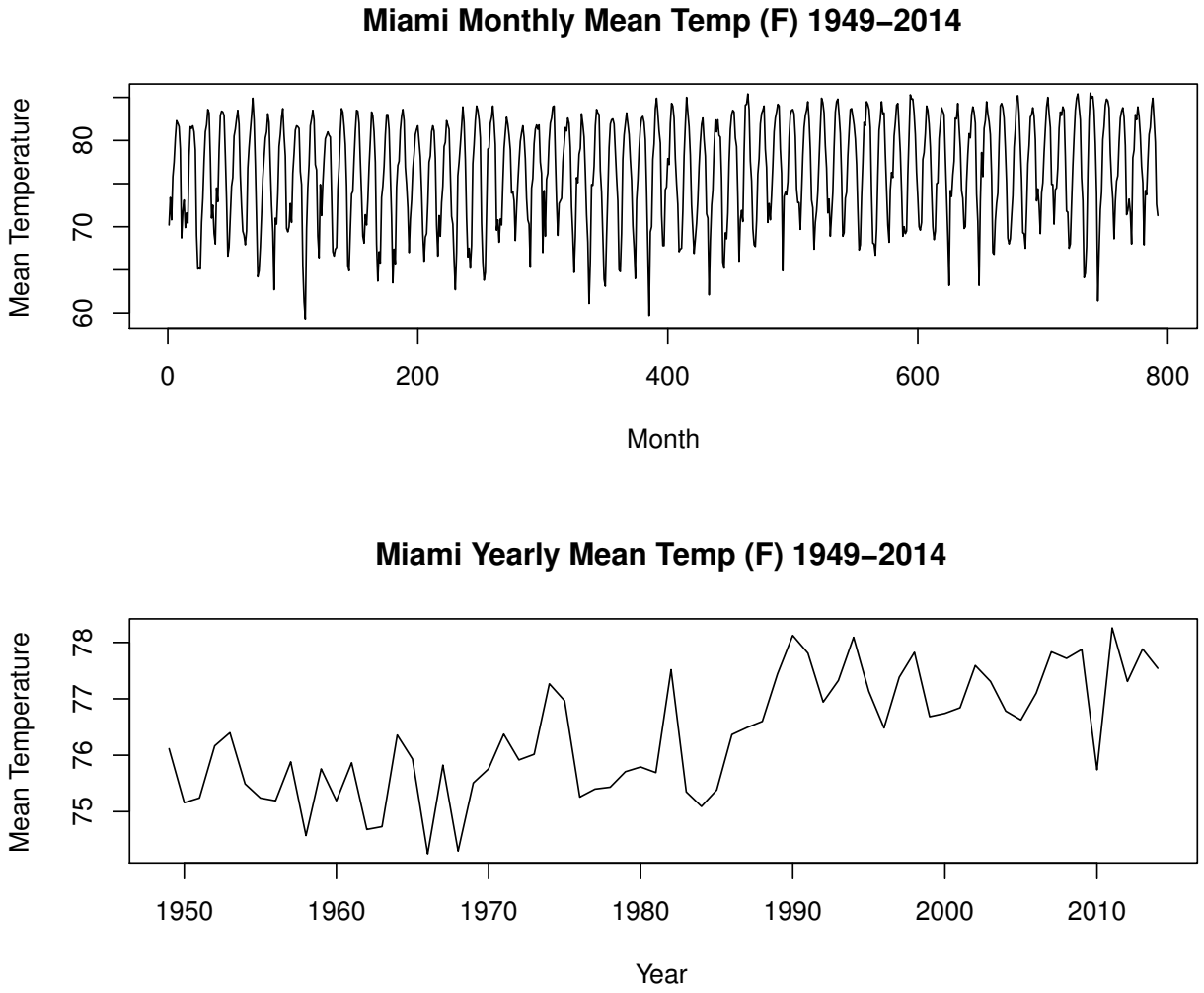

Figure 2.7: Monthly Mean Temperature in Miami, FL (January 1949 - December 2014)

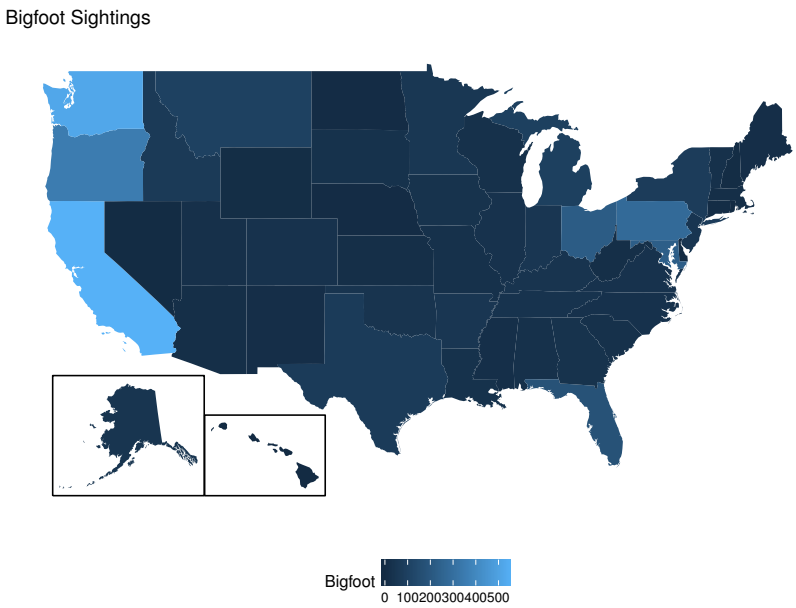

Figure 2.8: Bigfoot sightings by US state

Examples - CLT Traffic Stops and the VALI Laryngeal Study

For the Charlotte traffic stops, there were 10 categories for the reason for the stop. These reasons are treated as nominal, as there is no inherent ordering of the levels.

R Commands and Output are given below. The table function counts the number of cases (traffic stops) that are of each category, and dividing by their sum turns them into proportions.

```
### R Commands/Output (using previous dataset)
(table.RsnStop <- table(RsnStop))
round(table.RsnStop / sum(table.RsnStop), 5)

> (table.RsnStop <- table(RsnStop))
RsnStop
    1     2     3     4     5     6     7     8     9    10
  286   114  1992  1926  4827   631 22222  7946  7535 32405
> round(table.RsnStop / sum(table.RsnStop), 5)
RsnStop
      1       2       3       4       5       6       7       8       9      10
0.00358 0.00143 0.02494 0.02411 0.06043 0.00790 0.27818 0.09947 0.09432 0.40565
```

For the VALI study, the ordinal dysphonia rating had levels: 0, 1, 2, 3. There were 3, 6, 9, and 12 cases for those categories (total of 30 subjects). The proportions for the categories are:

$$0 : 3/30 = .10 \quad 1 : 6/30 = .20 \quad 2 : 9/30 = .30 \quad 3 : 12/30 = .40$$

The cumulative proportions (at or below that score) are:

$$0 : .10 \quad 1 : .10 + .20 = .30 \quad 2 : .30 + .30 = .60 \quad 3 : .60 + .40 = 1.00$$

In these examples, the traffic stop data can be thought of as a population (all traffic stops in Charlotte, N.C. in 2016), and the VALI dysphonia data is most certainly a sample.

∇

2.2.1 Measures of Central Tendency

There are 2 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. If you obtain 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 affected 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.

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

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.

Example 2.4: 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 Program and Output (NHL BMI)

```
### obtain the population size from number of rows of data frame
(N <- NROW(nhl))
### obtain the total of the BMI values
(sum.BMI <- sum(bmi.nhl))
### mean = sum / N
sum.BMI/N
### Use built-in mean function
mean(bmi.nhl)

### Obtain sorted bmi's
bmi.nhl.sort <- sort(bmi.nhl)
### Print first few cases to confirm ordered
head(bmi.nhl.sort)
### If N is even, average middle 2 cases, otherwise take middle case
ifelse(N%%2==0,(bmi.nhl.sort[N/2]+bmi.nhl.sort[N/2+1])/2,
      bmi.nhl.sort[(N+1)/2])
### Use built-in median function
median(bmi.nhl)
```

```
> (N <- NROW(nhl))
[1] 717
> (sum.BMI <- sum(bmi.nhl))
[1] 19016.05
> sum.BMI/N
[1] 26.52169
> mean(bmi.nhl)
[1] 26.52169
>
> bmi.nhl.sort <- sort(bmi.nhl)
> head(bmi.nhl.sort)
[1] 21.73927 21.95328 22.15455 22.42773 22.56005 22.92265
> ifelse(N%%2==0,(bmi.nhl.sort[N/2]+bmi.nhl.sort[N/2+1])/2,
+       bmi.nhl.sort[(N+1)/2])
[1] 26.55679
> median(bmi.nhl)
[1] 26.55679
```

Note that the mean (26.52) and median (26.56) are very close, as is expected for a (approximately) symmetric distribution.

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

R Program and 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).

∇

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, 90% trimmed mean is the mean of the middle 90% of the ordered measurements (removing the smallest 5% and largerst 5%).

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.

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

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 many 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 2.5: 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 Commands and Output

```
> ### Compute BMI
```

```
> bmi.nhl <- 703 * Weight / (Height^2)
>
> (bmi.max <- max(bmi.nhl))    # Highest BMI
[1] 32.05132
> (bmi.min <- min(bmi.nhl))    # Lowest BMI
[1] 21.65316
> (range <- bmi.max - bmi.min) # Compute Range
[1] 10.39816
> (bmi.75 <- quantile(bmi.nhl,.75))    # BMI 75%-ile
     75%
27.38411
> (bmi.25 <- quantile(bmi.nhl,.25))    # BMI 25%-ile
     25%
25.50129
> (IQR <-  bmi.75 - bmi.25)    # Comute IQR
     75%
1.882814
> (N <- length(bmi.nhl))     # Use "length" function to get N
[1] 717
> (mu <- mean(bmi.nhl))      # Use "mean" function to get mu
[1] 26.5097
> (sum.dev2 <- sum((bmi.nhl - mu)^2)) # Numerator of Variance
[1] 1559.084
> (sigma2 <- sum.dev2/N)     # Population Variance
[1] 2.174454
> (s2 <- sum.dev2/(N-1))     # Sample Variance
[1] 2.177491
> (sigma <- sqrt(sigma2))    # Population Standard Deviation
[1] 1.474603
> (s <- sqrt(s2))            # Sample Standard Deviation
[1] 1.475633
> var(bmi.nhl)               # Sample Variance with "var" function
[1] 2.177491
> (N-1)*var(bmi.nhl)/N       # Pop variance with "var" function
[1] 2.174454
> sd(bmi.nhl)                # Sample Std Dev with "sd" function
[1] 1.475633
> sqrt((N-1)/N)*sd(bmi.nhl)  # Population Std Dev with "sd" function
[1] 1.474603
> sum(bmi.nhl >= mu-sigma & bmi.nhl <= mu+sigma) / N     # Proportion w/in 1 sigma of mu
[1] 0.6987448
> sum(bmi.nhl >= mu-2*sigma & bmi.nhl <= mu+2*sigma) / N  # Proportion w/in 2 sigma of mu
[1] 0.9497908
```

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 Commands and Output

```
> f.mph <- mph[Gender=="F"]
> (N.f <- length(f.mph))
[1] 1045
> (mean.f <- mean(f.mph))
[1] 5.839839
> (var.f <- var(f.mph))
[1] 0.6906284
> (sd.f <- sd(f.mph))
[1] 0.8310405
> sum(f.mph >= mean.f - sd.f & f.mph <= mean.f + sd.f) / N.f
```

```
[1] 0.662201
> sum(f.mph >= mean.f - 2*sd.f & f.mph <= mean.f + 2*sd.f) / N.f
[1] 0.9636364
>
> m.mph <- mph[Gender=="M"]
> (N.m <- length(m.mph))
[1] 1454
> (mean.m <- mean(m.mph))
[1] 6.336979
> (var.m <- var(m.mph))
[1] 1.118701
> (sd.m <- sd(m.mph))
[1] 1.057687
> sum(m.mph >= mean.m - sd.m & m.mph <= mean.m + sd.m) / N.m
[1] 0.6650619
> sum(m.mph >= mean.m - 2*sd.m & m.mph <= mean.m + 2*sd.m) / N.m
[1] 0.9635488
```

We see that 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.

∇

Two other measures of variation are the following. The **median absolute deviation** (MAD), which should be clear by its name how it is computed, and when data are from a normal (Gaussian) distribution, should be approximately 0.645σ. The other is the **coefficient of variation** (CV), which is the ratio of the standard deviation to the mean (and is often reported as a percentage). The coefficient of variation is often reported measure the accuracy of laboratory equipment.

Example 2.6: NHL BMI's and Rock and Roll Marathon Speeds

Here we compute MAD and CV for the three datasets. Note that the MAD for the NHL BMI's, when divided by 0.645 is 1.461, while we saw above that $\sigma = 1.475$, so they are very consistent, as expected as the BMI distribution is well approximated by a normal distribution.

R Commands and Output

```
### NHL
> (mad <- median(abs(bmi.nhl - mu))) # Median absolute deviation
[1] 0.9425232
> mad/0.645                          # Approximating sigma
[1] 1.461276
> (cv <- sigma/mu)                   # Coefficient of Variation
[1] 0.05562504

### Rock and Roll Marathon
> (cv.f <- sd.f/mean.f)
[1] 0.1423054
> (mad.f <- median(abs(f.mph-mean.f)))
[1] 0.5925911
> (cv.m <- sd.m/mean.m)
```

```
[1] 0.1669071
> (mad.m <- median(abs(m.mph-mean.m)))
[1] 0.7392048
```

2.2.3 Higher Order Moments

Two other measures are occasionally reported: **skewness** and **kurtosis**. Skewness is used to measure the symmetry of the distribution, and kurtosis measures the heaviness of the tails of the distribution. Positive values for skewness correspond to right-skewed distributions, while negative values correspond to left-skewed distributions. Negative values of kurtosis imply a distribution has fewer extreme values (lighter tails) than a normal distribution, while positive values imply more extreme values (heavier tails) than a normal distribution. These measures are reported in many fields, and are especially important in financial modeling. For a set of measurements, the skewness and kurtosis are computed as follow.

Population Skewness: $\frac{\mu_3}{\sigma^3}$ where $\mu_3 = \frac{\sum_{i=1}^{N}(y_i-\mu)^3}{N}$ Sample Skewness: $\frac{m_3}{s^3}$ where $m_3 = \frac{\sum_{i=1}^{n}(y_i-\overline{y})^3}{n}$

Population Kurtosis: $\frac{\mu_4}{\sigma^4}-3$ where $\mu_4 = \frac{\sum_{i=1}^{N}(y_i-\mu)^4}{N}$ Sample Kurtosis: $\frac{m_4}{s^4}-3$ where $m_4 = \frac{\sum_{i=1}^{n}(y_i-\overline{y})^4}{n}$

Example 2.7: NHL BMI's and Rock and Roll Marathon Speeds

We compute the skewness and kurtosis for the three datasets here.

R Commands and Output

```
## NHL BMI

> (mu3 <-  (sum((bmi.nhl-mu)^3)/N))
[1] 0.01015497
> (skew <- mu3/(sigma^3))
[1] 0.003167039
>
> (mu4 <-  (sum((bmi.nhl-mu)^4)/N))
[1] 15.96356
> (kurt <- mu4/(sigma^4)-3)
[1] 0.3762083

### Rock and Roll Marathon

> (m3.f <- sum((f.mph-mean.f)^3)/N.f)
[1] 0.3616936
> (skew.f <- m3.f / sd.f^3)
[1] 0.6301939
> (m4.f <- sum((f.mph-mean.f)^4)/N.f)
[1] 1.483532
> (kurt.f <- (m4.f/sd.f^4)-3)
```

		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

```
[1] 0.1103411

> (m3.m <- sum((m.mph-mean.m)^3)/N.m)
[1] 0.5792011
> (skew.m <- m3.m / sd.m^3)
[1] 0.4895061
> (m4.m <- sum((m.mph-mean.m)^4)/N.m)
[1] 3.833391
> (kurt.m <- (m4.m/sd.m^4)-3)
[1] 0.0630552
```

Skewness is very close to 0 for the NHL BMI data, as expected from the histogram. The skewnesses for the Female and Male marathon speeds are positive, and well away from 0, again consistent with their histograms. The kurtosis for the NHL BMI data is greater than 0, corresponding to heavier tails than a normal distribution; the measures for marathon speeds are very close to 0.

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 (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 can 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. A contingency table is typically of the form in Table 2.1.

Example 2.8: Charlotte, NC Traffic Stops

For the Charlotte traffic stop data, each stap was classified by whether the Officer was Male or Female,

and whether the Driver was Male or Female. Suppose we are interested whether there is a difference in the proportions of Male (and thus Female) drivers stopped by Male and female officers. Among Female officers, there were 6709 traffic stops, of which 3789 (.5648) were Male and 2920 (.4352) were Female. Among Male officers, there were 73175 traffic stops, of which 42505 (.5809) were Male and 30670 (.4191) were Female. There are very small differences among the Genders of drivers stopped by Male and Female officers. Figure 2.9 shows the proportions of driver genders by officer genders.

R Commands and Output

```
### Commands
## Read data off web page, attach file as data frame, and list variable names
clt2016 <- read.csv("http://www.stat.ufl.edu/~winner/data/trafficstop.csv")
attach(clt2016); names(clt2016)

head(clt2016)    ## Print first 6 observations

## Make OffMale and DrvMale "factor variables" and give labels for 0/1
OffMale <- factor(OffMale, labels=c("OfficerF","OfficerM"))
DrvMale <- factor(DrvMale, labels=c("DriverF","DriverM"))

## Obtain Table of Counts (Row=Officer, Column=Driver)
(omdm <- table(OffMale,DrvMale))
## Obtain Row (1) and Column (2) Marginal Totals
margin.table(omdm,1)
margin.table(omdm,2)
## Obtain Proportions across all Cells
omdm/sum(omdm)
## Obtain Row Proportions (Driver Genders w/in Officer Gender)
prop.table(omdm,1)
## Obtain Column Proportions (Officer Genders w/in Driver Gender)
prop.table(omdm,2)

## Obtain Cluster (Grouped) and Stacked Bar Plots
## t(prop.table(omdm,1)) takes transpose so that group var is Officer
par(mfrow=c(1,2))
barplot(t(prop.table(omdm,1)),beside=T,legend=colnames(omdm),ylim=c(0,1),
   main="Grouped Bar Plot - CLT Traffic Stops")
barplot(t(prop.table(omdm,1)),beside=F,legend=colnames(omdm),
   main="Stacked Bar Plot - CLT Traffic Stops")

### Output

> (omdm <- table(OffMale,DrvMale))
          DrvMale
OffMale    DriverF DriverM
  OfficerF    2920    3789
  OfficerM   30670   42505
> ## Obtain Row (1) and Column (2) Marginal Totals
> margin.table(omdm,1)
OffMale
OfficerF OfficerM
    6709    73175
> margin.table(omdm,2)
DrvMale
DriverF DriverM
  33590   46294
> ## Obtain Proportions across all Cells
> omdm/sum(omdm)
          DrvMale
OffMale       DriverF    DriverM
```

```
  OfficerF 0.03655300 0.04743128
  OfficerM 0.38393170 0.53208402
> ## Obtain Row Proportions (Driver Genders w/in Officer Gender)
> prop.table(omdm,1)
          DrvMale
OffMale      DriverF   DriverM
  OfficerF 0.4352362 0.5647638
  OfficerM 0.4191322 0.5808678
> ## Obtain Column Proportions (Officer Genders w/in Driver Gender)
> prop.table(omdm,2)
          DrvMale
OffMale       DriverF    DriverM
  OfficerF 0.08693063 0.08184646
  OfficerM 0.91306937 0.91815354
```

If there are three or more categorical variables, then tables of higher order dimensions and **mosaic plots** can be constructed. Here we consider the three variables: Reason for Stop, Officer Male, and Driver Male. The mosaic plot is constructed within the **vcd** (visualizing categorical data) package and is shown in Figure 2.10.

R program and Output

```
 ### Commands
library(vcd)
table(RsnStop,OffMale,DrvMale, data=clt2016)
mosaic(~RsnStop+OffMale+DrvMale, data=clt2016, shade=TRUE, legend=TRUE)

### Output
> table(RsnStop,OffMale,DrvMale)
, , DrvMale = DriverF

       OffMale
RsnStop OfficerF OfficerM
     1        13       81
     2         0       28
     3        67      581
     4        44      666
     5       141     1556
     6        18      176
     7       509     9407
     8       259     3021
     9       272     2612
     10     1597    12542

, , DrvMale = DriverM

       OffMale
RsnStop OfficerF OfficerM
     1        18      174
     2         3       83
     3       129     1215
     4        56     1160
     5       325     2805
     6        34      403
     7       601    11705
     8       376     4290
     9       376     4275
```

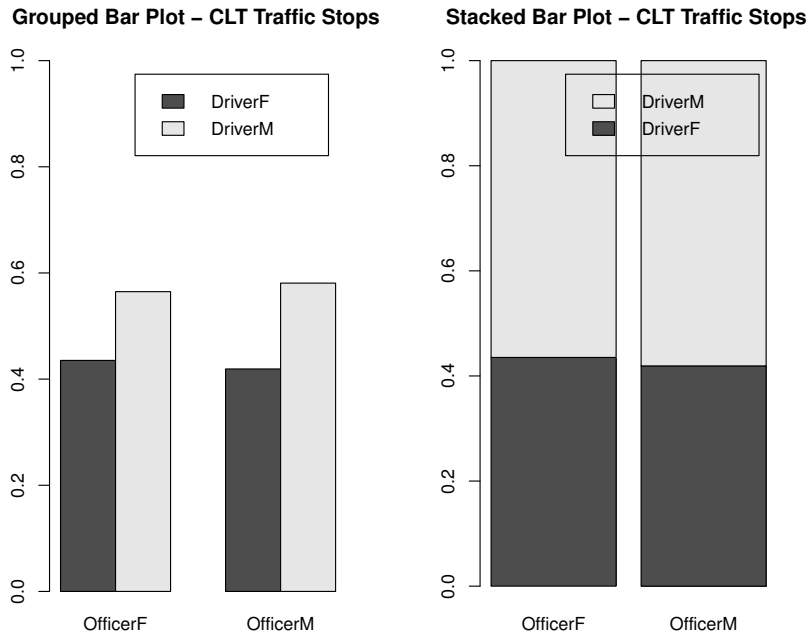

Figure 2.9: Stacked and Cluster (Grouped) Bar Charts - Charlotte NC Traffic Stops - Officer and Driver Genders

```
10     1871    16395
```

∇

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 (see Figure 2.4), box plots (see Figure 2.5), or violin plots (see Figure 2.6). Histograms and densities can also be placed into single plots with different colors or patterns.

Example 2.9: Rock and Roll Marathon Speeds by Gender

A density plot using basic plotting functions in R is displayed in Figure 2.11, and a combined box plot using the **ggplot2** package is given in Figure 2.12.

R Program

```
## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
```

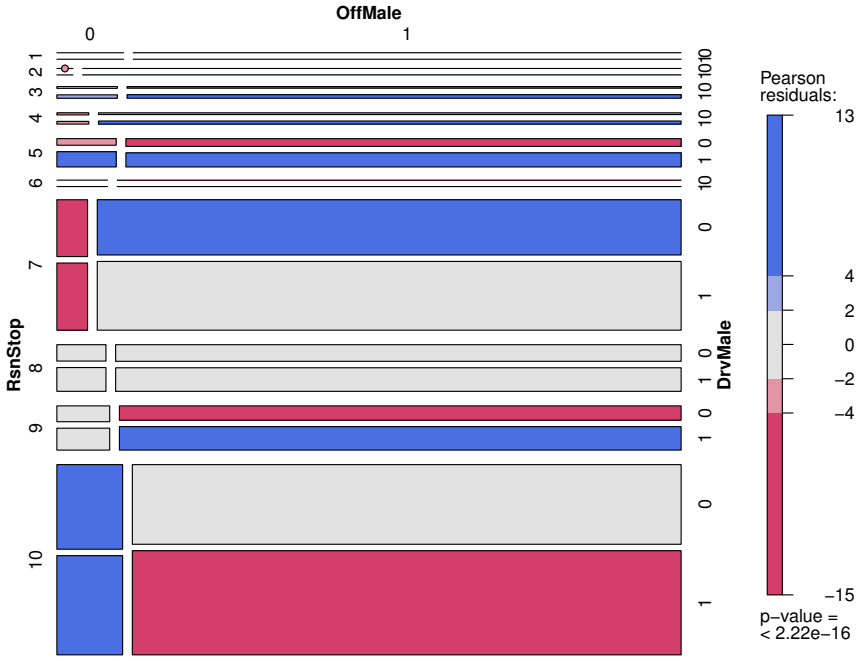

Figure 2.10: Mosaic plot for Charlotte Traffic Data - Reason for Stop is on Left Axis, Officer Gender on To Axis, Driver Gender on Right Axis

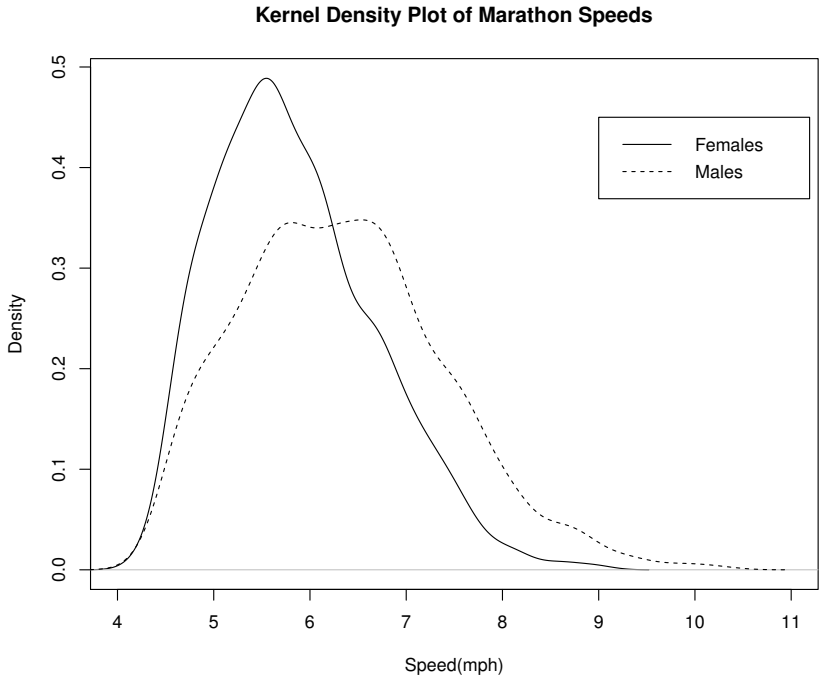

Figure 2.11: Density plot fro Females and Males - Rock and Roll Marathon Speeds

```
attach(rr.mar); names(rr.mar)


## Obtain the densities (for plotting) of mph by gender
d.F <- density(mph[Gender=="F"])
d.M <- density(mph[Gender=="M"])


## Density Plots for Female and Male mph
plot(d.F,xlim=c(4,11),xlab="Speed(mph)",ylab="Density",
  main="Kernel Density Plot of Marathon Speeds")
lines(d.M,lty=2)
legend(9,0.45,c("Females","Males"),lty=c(1,2))

## Combined histogram
library(ggplot2)
ggplot(rr.mar, aes(x=mph,fill=Gender)) +
   geom_histogram(binwidth=0.1)
```

∇

When two variables (labeled x and y) are both numeric, one numeric descriptive measure that is widely reported is the **correlation** between the two variable. 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

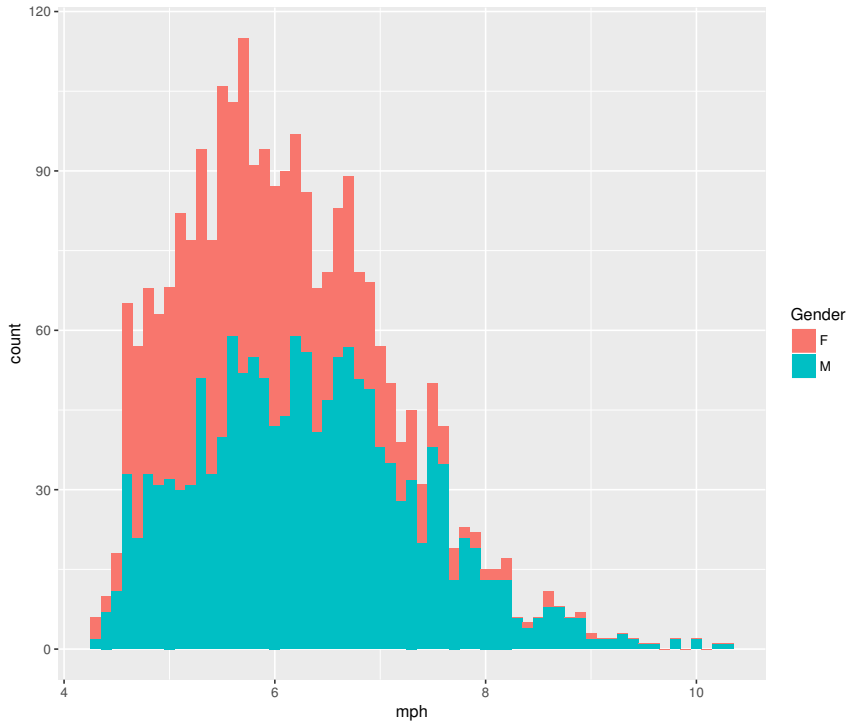

Figure 2.12: Combined Female/Male Histogram for Rock and Roll marathon speeds

variables. Unlike in Regression (described later), the variables are not necessarily (but can be) identified as an independent and or variable. The formula for this measure (population and sample) are given below.

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

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

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.

Example 2.10: Software Project Development - Size and Effort of Projects

A pair of studies considered the size (number of function points) and the effort needed for completion (hours) for 17 software development projects (Jeffery and Stathis, 1996 [15] and Jorgensen, Indahl, and Sjoberg, 2003 [16]). The data are given in Table 2.2. Note that the Project 17 is much larger than the others

and was not used in the Jorgensen paper. We consider data with and without that case, and also data based on natural logarithms of size (x) and effort (y). For the full dataset, based on the original scale, we obtain a correlation of $r = .9752$, see calculations in Table 2.2, based on an Excel spreadsheet. Also, for the full dataset, based on natural logarithms of size and effort (which often helps meet model assumptions when data are skewed with extreme case(s), as here), we find the correlation to be $r = .8791$. This was obtained using the **correl** built-in function in Excel. Plots of the four cases (original/log scale and with/without Project 17) are given in Figure 2.13, along with the "least squares regression line", which minimizes the error sum of squares (SSE), obtained as follows.

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

The plots were obtained in R, and the correlations for the 4 cases were obtained using the **cor** function. The **abline** command after each **plot** command adds the least squares regression line described above.

R Program and Output

```
### Commands

sw1 <- read.csv("http://www.stat.ufl.edu/~winner/data/software1.csv")
attach(sw1); names(sw1)

cor(sizeProj,effortProj)
cor(sizeProj[1:16],effortProj[1:16])

par(mfrow=c(2,2))
plot(sizeProj,effortProj,xlab="size",ylab="effort",
   main="Original Scale, All Projects")
abline(lm(effortProj~sizeProj))
plot(sizeProj[1:16],effortProj[1:16],xlab="size",ylab="effort",
   main="Original Scale, Project 17 Removed")
abline(lm(effortProj[1:16]~sizeProj[1:16]))

cor(log(sizeProj),log(effortProj))
cor(log(sizeProj[1:16]),log(effortProj[1:16]))

plot(log(sizeProj),log(effortProj),xlab="ln(size)",ylab="ln(effort)",
   main="Log Scale, All Projects")
abline(lm(log(effortProj) ~ log(sizeProj)))
plot(log(sizeProj[1:16]),log(effortProj[1:16]),xlab="ln(size)",ylab="ln(effort)",
   main="Log Scale, Project 17 Removed")
abline(lm(log(effortProj[1:16]) ~ log(sizeProj[1:16])))

### Text Output

> cor(sizeProj,effortProj)
[1] 0.9752405
> cor(sizeProj[1:16],effortProj[1:16])
[1] 0.9261634
> cor(log(sizeProj),log(effortProj))
[1] 0.8791134
> cor(log(sizeProj[1:16]),log(effortProj[1:16]))
[1] 0.8131933
```

projID	size(x)	effort(y)	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})(y-\overline{y})$	x^*=ln(x)	y^*=ln(y)
1	1164	3777	612.76	1683.71	1031715.54	7.06	8.24
2	1834	4389	1282.76	2295.71	2944850.48	7.51	8.39
3	388	1647	-163.24	-446.29	72850.95	5.96	7.41
4	336	1318	-215.24	-775.29	166870.66	5.82	7.18
5	116	529	-435.24	-1564.29	680836.01	4.75	6.27
6	182	691	-369.24	-1402.29	517776.48	5.20	6.54
7	65	291	-486.24	-1802.29	876339.01	4.17	5.67
8	160	448	-391.24	-1645.29	643697.13	5.08	6.10
9	185	262	-366.24	-1831.29	670684.54	5.22	5.57
10	168	415	-383.24	-1678.29	643181.54	5.12	6.03
11	422	2070	-129.24	-23.29	3010.42	6.05	7.64
12	296	1947	-255.24	-146.29	37339.42	5.69	7.57
13	129	1500	-422.24	-593.29	250509.72	4.86	7.31
14	143	1114	-408.24	-979.29	399782.42	4.96	7.02
15	38	362	-513.24	-1731.29	888561.25	3.64	5.89
16	89	921	-462.24	-1172.29	541875.72	4.49	6.83
17	3656	13905	3104.76	11811.71	36672567.54	8.20	9.54
Mean	551.24	2093.29		Sum/(n-1)	2940153.05		
SD	923.09	3265.97		Correlation	0.9752	Correlation	0.8791

Table 2.2: Software Projects Sizes and Effort Levels and Correlation Calculations

Note that the extreme Size of Project 17 had the impact of pulling the regression line toward its Effort level and tended to increase the correlation. That project has high "leverage" on the calculated regression line.

∇

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 2.11: Compressive Strength and Microfabric Properties of Amphibolites

A study reported the relationship between Uniaxial Compression Strenght (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 2.14. 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 Code and Output

```
### R Commands
rs1 <- read.csv("http://www.stat.ufl.edu/~winner/data/rockstrength.csv")
attach(rs1); names(rs1)

## Obtain Scatterplot matrix of UCS, hb, gs, ga (columns 2,6,7,8 of rs1)
plot(rs1[,c(2,6,7,8)])
```

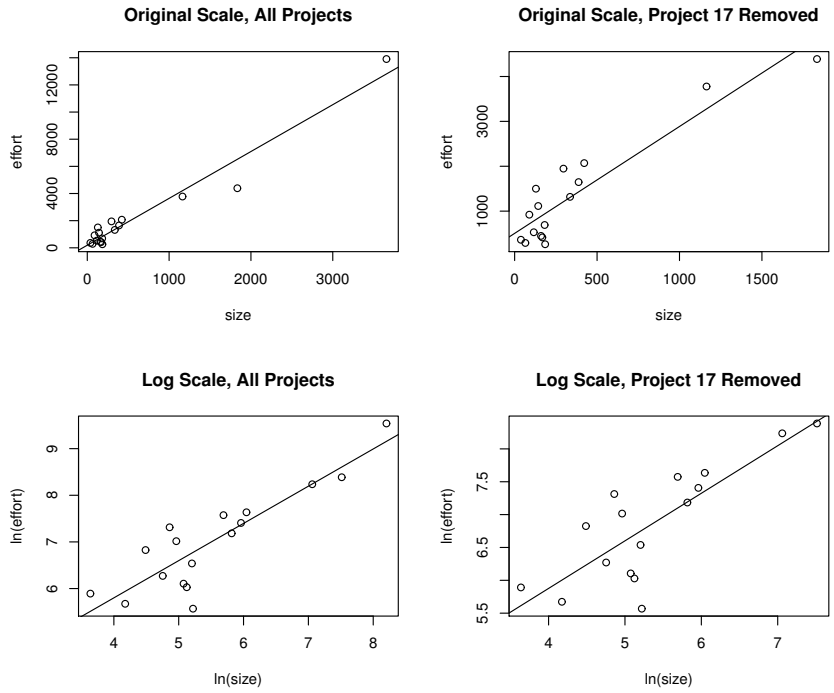

Figure 2.13: Plots of Effort (y) versus Size (x) for Original/log scales and with/without Project 17

```
## Obtain correlation matrix of UCS, hb, gs, ga  (columns 2,6,7,8 of rs1)
cor(rs1[,c(2,6,7,8)])

### 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, as in the software development example, 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, extereme 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 2.12: NASCAR Start and Finish Positions 1975-2003

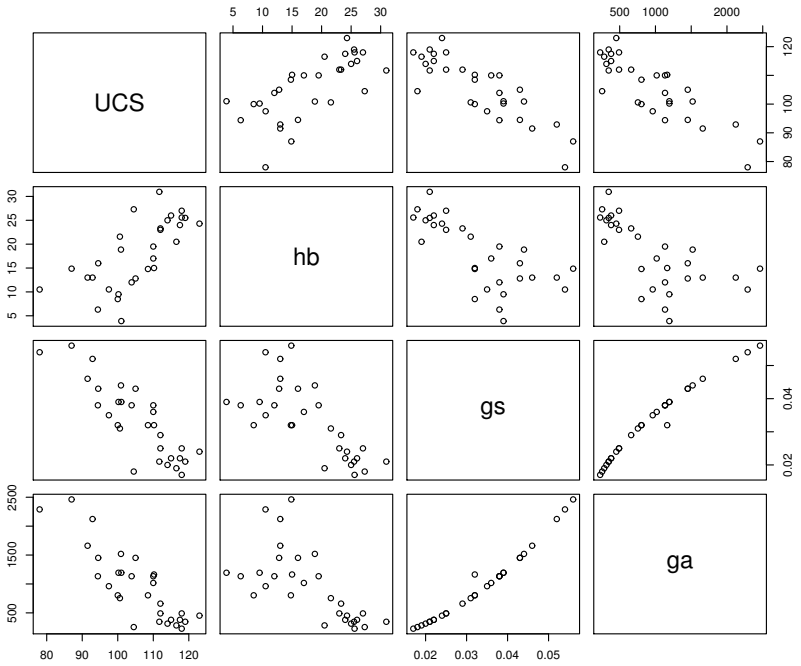

Figure 2.14: Bivariate Plots of Uniaxial Compression Strength (UCS), Percent Hornblende (hb), Grain Size (gs), and Grain Area (ga)

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 [31]). 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 2.15.

R Commands and Output

```
### Commands

nasRace <- read.fwf("http://www.stat.ufl.edu/~winner/data/nascarr.dat",
  widths=c(3,6,4,4,9,7,9,9,7,5,3,4,4,9,7,8,5,38),col.names=c("seriesRace",
  "year","yearRace","numCar","payout","cpiU","spearman","kendall","trkLength",
  "lapsComp","roadRace","cautionFlag","leadChange","winTime","trkLat","trkLong",
  "trkCode","trkName"))
attach(nasRace)

length(spearman)
summary(spearman)

hist(spearman,breaks=seq(-.5,1,.02),prob=T,
main="NASCAR Start/Finish - Spearman's rho")
lines(density(spearman))

### Output

> length(spearman)
[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.

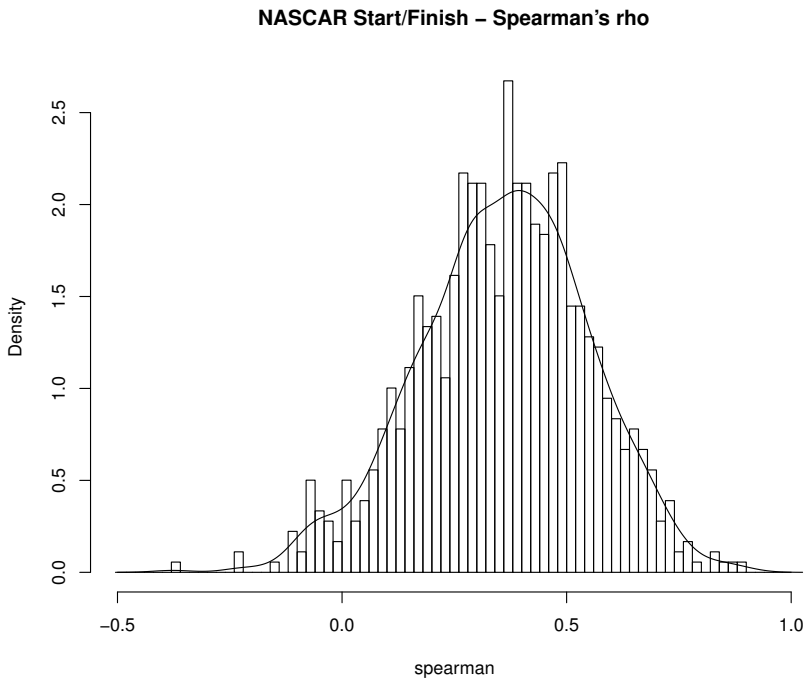

Figure 2.15: NASCAR Races 1975-2003 - Spearman's rank correlation coefficient for start/finish positions

Chapter 3

Probability

In this chapter, we describe the concepts of probability, random variables, and probability 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 and its sequel.

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 [17], 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 each 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 ar $10(10)(10)(10) = 10{,}000$ 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 is referred to as its **conditional** probability. If the unconditional probability of A and its conditional probability given B has occurred, 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.

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

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

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).

	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

∇

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

For the Rock and Roll marathon speeds considered previously, we class events as follow. Event F is that the runner is Female, event S_5 is the event that a runner's speed is less that 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 my 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 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.

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

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.

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

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

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 [2], 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 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 case 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.

$$0 \leq p(y_i) \leq 1 \qquad \sum_{i=1}^{k} p(y_i) = 1 \qquad F(y) = P(Y \leq y) = \sum_{i=-\infty}^{y} p(i)$$

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 [31], 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 stared 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 Program to Construct Probability Distribution

```
## Commands
## Read Driver Level Data into "nascard"
nascard <- read.fwf("http://www.stat.ufl.edu/~winner/data/nascard.dat",
   width=c(3,6,4,4,4,5,9,4,11,32), col.names=c("serRace", "year",
   "yrRace", "finPos", "strtPos", "lapsComp", "winnings", "numCars",
   "carMake", "driver"))

## Subset rows of nascard w/ driver starting in 1st 10 positions in "start10"
## Save only columns: series Race, start and finish positions
start10 <- nascard[nascard$strtPos <= 10,c("serRace","strtPos","finPos")]

(nraces <- length(unique(start10$serRace)))   ### number races=898
strt10Fin3 <- rep(0, nraces)    ### Initialize Top 3 Fins per race

### Count # of top 10 starters finish in top 3 for each race
for (i in 1:nraces) {
  strt10Fin3[i] <- sum(start10[start10$serRace==i,]$finPos <=3)
}

(t.strt10Fin3 <- table(strt10Fin3))  ### Count 0,1,2,3 Top 3 finishers
t.strt10Fin3 / sum(t.strt10Fin3)     ### Turn counts to proportions

## 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 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.

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

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

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

Example 3.7: NASCAR Race Finish Positions - 1975-2003

If we repeatedly sampled a race from this population, observed and saved the number of the top 10 starters who finished in the top 3, the long run mean would be μ_Y, and a "typical" distance from the mean would be σ_Y. From Table 3.4, we have the necessary calculations to compute μ_Y, σ_Y^2, and σ_Y.

$$\mu_Y = \sum_y y p(y) = 2.0991 \qquad \sigma_Y^2 = \sum_y (y-\mu_Y)^2 p(y) = \sum_y y^2 p(y) - \mu_Y^2 = 5.1213 - 2.0991^2 = 0.7151$$

$$\sigma_Y = +\sqrt{0.7151} = 0.8456$$

We now sample 10000 races from this population (equivalently done by taking 10000 integers between 1 and 898 WITH replacement), and observing the number of top 3 finishers for each race. Then we compute the mean and standard deviation of those numbers.

R Program and Output

```
## Commands (added to previous commands)

set.seed(12345)
sample.race <- sample(x=1:nraces, size=10000, replace=TRUE)
mean(strt10Fin3[sample.race])
sd(strt10Fin3[sample.race])

## Output

> mean(strt10Fin3[sample.race])
[1] 2.0994
> sd(strt10Fin3[sample.race])
[1] 0.8403513
```

Note that the mean of the 10000 sampled races is very close to the population mean (2.0994 vs 2.0991) and sample standard deviation is close to the corresponding population value (0.8404 vs 0.8456). If we had use a different seed, the samples, and thus their means and standard deviations would change as well.

∇

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^2_Y, respectively. Further, suppose that a and b are constants (not random). Then we have the following results.

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

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

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 3.8: NHL Hockey Player Weights 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.

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

The population mean and variance of heights are 73.25 inches and 4.29 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.

$$\mu_C = 2.54\mu_I = 2.54(73.25) = 186.06 \qquad \sigma^2_C = (2.54)^2\sigma^2_I = (2.54)^2(4.29) = 27.68 \qquad \sigma_C = \sqrt{27.68} = 5.26$$

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 $\text{COV}\{Y_i, Y_j\} = \sigma_{ij}$.

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

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, 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 3.9: 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 probability distributions, namely the Binomial, Poisson, and Negative Binomial families. These are used in many situations of data being 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 Tversy, 1985, [13].

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 $Y \sim Bin(n, \pi)$.

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

All 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 3.10: 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, [27]). 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 Commands and Output

```
### Commands

y <- 0:6    ## Values that Y can take on: y=0,1,...,6
(p_y <- dbinom(y, 6, 0.5))  ## Obtain p(y) for y=0,1,...,6
### Plot probabilities (type="h" is histogram)
plot(y, p_y, type="h", lwd=5, ylab="p(y)",
   main="Probabilty Distribution for Binomial(6,0.5)")


### 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, [27], 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

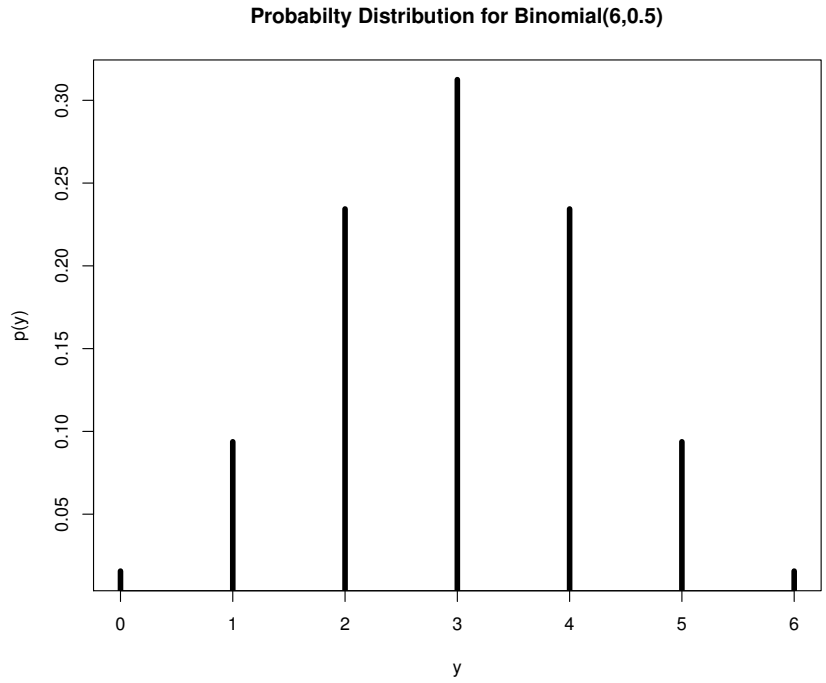

Figure 3.1: Probability Distribution for Mobile Telephone Telepathy experiment assuming random guessing, Y Bin(6,0.5)

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: $Y \sim \text{Poi}(\lambda)$. The probability distribution, mean and variance of Y are:

$$p(y) = \frac{e^{-\lambda}\lambda^y}{y!} \qquad E\{Y\} = \mu_Y = \lambda \qquad V\{Y\} = \sigma_Y^2 = \lambda$$

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 3.11: E Coli Bacterial Cell Counts

A study considered the distribution of bacterial cell counts for various bacteria strains in single-cell studies (Koyama, et al, 2016 [19]). There were 8 strains, and the authors observed counts for 96 cells under target means of $\lambda = 1$ and $\lambda = 2$ in an experimental study. They found that the observed counts were highly consistent with the Poisson models. The theoretical probability distributions are given as follow.

$$\lambda = 1: \quad p(y) = \frac{e^{-1}1^y}{y!} = \frac{e^{-1}}{y!} \quad y = 0, 1, 2, \ldots \qquad \lambda = 2: \quad p(y) = \frac{e^{-2}2^y}{y!} \quad y = 0, 1, 2, \ldots$$

∇

Example 3.12: 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 [8] also reported in Feller, 1950 [12] . 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).

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

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.

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

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) [4]).

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

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

Example 3.13: 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, [5] and included in a large number of datasets by Thorndike, 1926, [29]. 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.

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

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

y	# comets	$p(y)$	$yp(y)$	$y^2p(y)$	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

The Negative Binomial appears to fit better than a Poisson distribution with mean 2.58, this will be quantified later.

∇

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.

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

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

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 be 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.

3.4.1 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 $Y \sim N\left(\mu, \sigma^2\right)$. The probability density function, the mean and variance are:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \qquad E\{Y\} = \mu_Y = \mu \qquad V\{Y\} = \sigma_Y^2 = \sigma^2$$

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:

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

Where Z is **standard normal**, a normal distribution with mean 0, and variance (standard deviation) 1. Here $\Phi(z^*)$ 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:

$$Y \sim N\left(\mu, \sigma^2\right) \qquad \Rightarrow \qquad Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$$

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

Approximately 68% of measurements lie within 1 standard deviation from the mean, 95% lie within 2 standard deviations, and virtually all lie within 3 standard deviations.

Example 3.14: 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 26.51 snd standard deviation 1.47. Figure 3.3 gives a

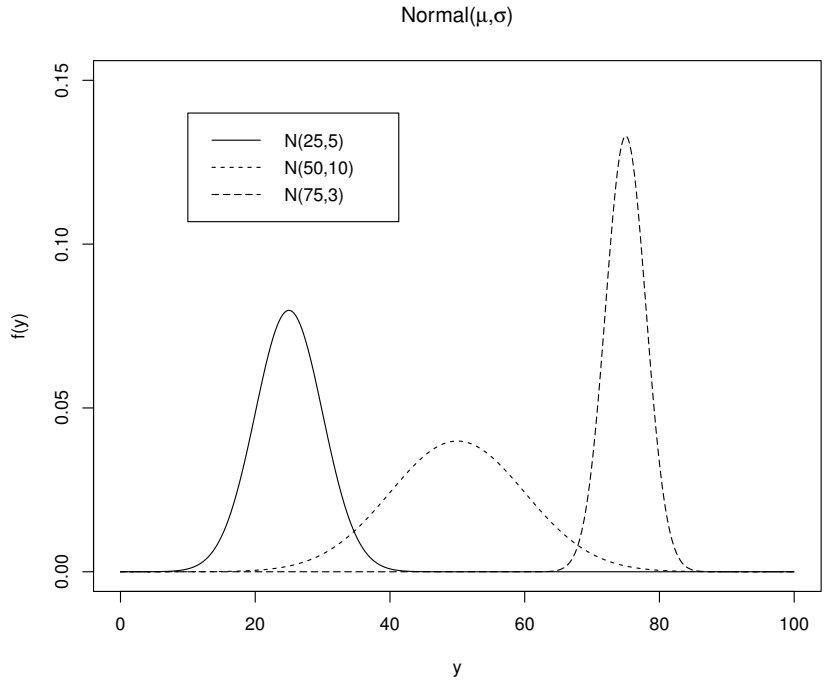

Figure 3.2: Three Normal Densities

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(26.51, 1.47) distribution. Also consider the probabilities of the following ranges ($< 26.51 - 2(1.47) = 23.57, > 26.51 + 2(1.47) = 29.45$, and $(25.04 = 26.51 - 1.47, 26.51 + 1.47 = 27.98)$) for the NHL data and the normal distribution.

R Program and Output

```
## Commands

### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Generate random values (-0.5 to 0.5) to add to Height
set.seed(1234)
N <- NROW(nhl)
Height.dev <- 0.5 - runif(N)
Height <- Height + Height.dev

### Compute BMI
bmi.nhl <- 703 * Weight / (Height^2)

(mean.bmi.nhl <- mean(bmi.nhl))
```

```
(sd.bmi.nhl <- sd(bmi.nhl)*sqrt((N-1)/N))
bmi <- seq(21,33,.01)

### Obtain histogram
hist(bmi.nhl, breaks=seq(21,33,0.2), xlab="Body Mass Index", freq=FALSE,
main="NHL BMI Distribution 2013-2014 Season")
lines(bmi, dnorm(bmi, mean.bmi.nhl, sd.bmi.nhl))

## Quantiles: Theoretical Normal, Actual Distribution
qnorm(c(.10,.25,.50,.75,.90), mean.bmi.nhl, sd.bmi.nhl)
quantile(bmi.nhl, c(.10,.25,.50,.75,.90))

## Probabilities: Theoretical Normal, Actual Distribution
# Theoretical
pnorm(mean.bmi.nhl-2*sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl)
1-pnorm(mean.bmi.nhl+2*sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl)
pnorm(mean.bmi.nhl+sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl) -
    pnorm(mean.bmi.nhl-sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl)
# Actual
sum(bmi.nhl <= mean.bmi.nhl-2*sd.bmi.nhl)/N
sum(bmi.nhl >= mean.bmi.nhl+2*sd.bmi.nhl)/N
sum(bmi.nhl >= mean.bmi.nhl-sd.bmi.nhl & bmi.nhl <= mean.bmi.nhl+sd.bmi.nhl)/N

### Output

> qnorm(c(.10,.25,.50,.75,.90), mean.bmi.nhl, sd.bmi.nhl)
[1] 24.61992 25.51510 26.50970 27.50430 28.39948
> quantile(bmi.nhl, c(.10,.25,.50,.75,.90))
     10%      25%      50%      75%      90%
24.65826 25.50129 26.57766 27.38411 28.29861
>
> ## Probabilities: Theoretical Normal, Actual Distribution
> # Theoretical
> pnorm(mean.bmi.nhl-2*sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl)
[1] 0.02275013
> 1-pnorm(mean.bmi.nhl+2*sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl)
[1] 0.02275013
> pnorm(mean.bmi.nhl+sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl) -
+     pnorm(mean.bmi.nhl-sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl)
[1] 0.6826895
> # Actual
> sum(bmi.nhl <= mean.bmi.nhl-2*sd.bmi.nhl)/N
[1] 0.0264993
> sum(bmi.nhl >= mean.bmi.nhl+2*sd.bmi.nhl)/N
[1] 0.0237099
> sum(bmi.nhl >= mean.bmi.nhl-sd.bmi.nhl & bmi.nhl <= mean.bmi.nhl+sd.bmi.nhl)/N
[1] 0.6987448
```

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

∇

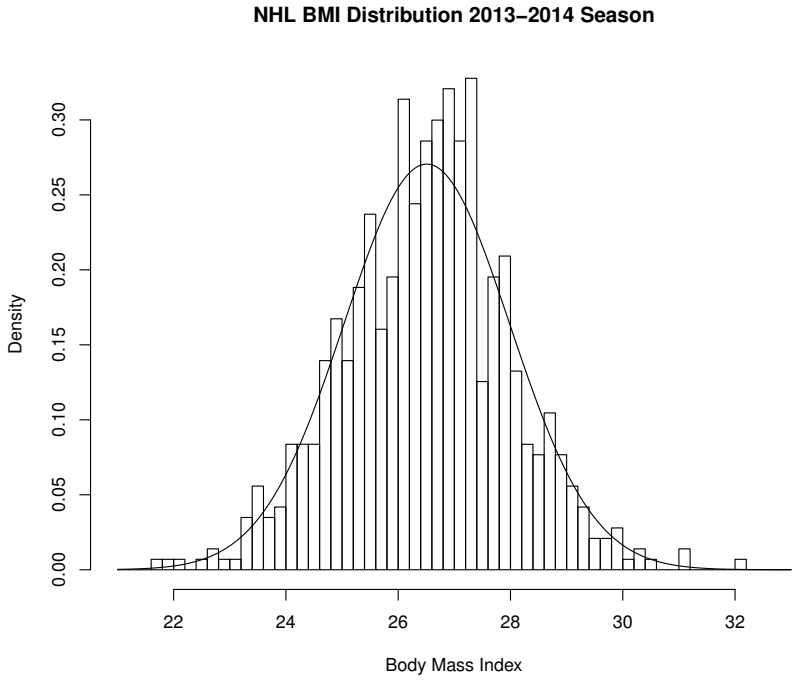

Figure 3.3: NHL Body Mass Indices and Normal Distribution

3.4.2 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.

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

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

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

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 3.15: Rock and Roll Marathon Speeds

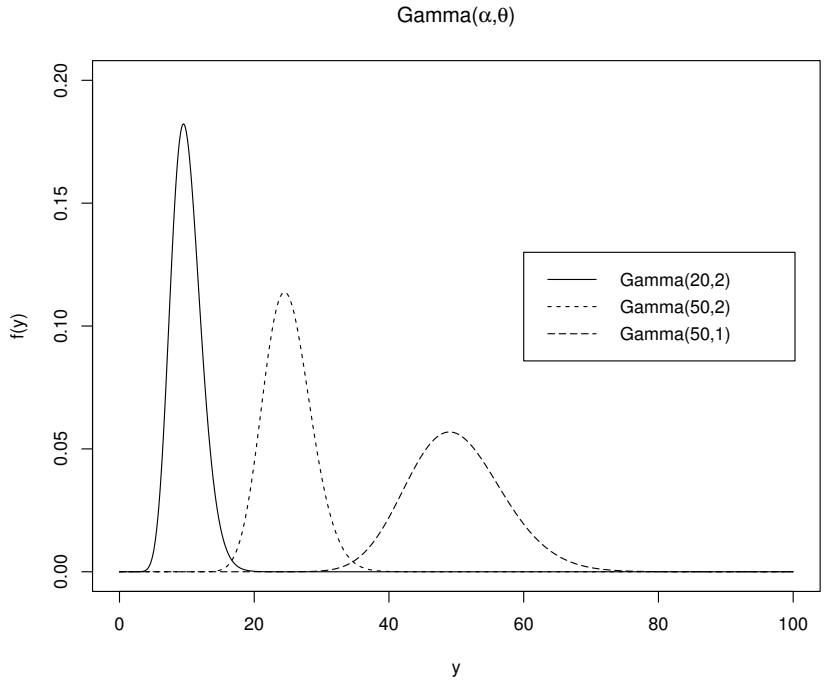

Figure 3.4: Three Gamma Densities

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.

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

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

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

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

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 Commands and Output

```
## R Commands

## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)

## Obtain mean and standard deviation by gender
tapply(mph,Gender,mean)
tapply(mph,Gender,median)
tapply(mph,Gender,var)
tapply(mph,Gender,sd)

## Obtain the Gamma parameters (for plotting) of mph by gender
(alpha.f <- mean(mph[Gender=="F"])^2 / var(mph[Gender=="F"]))
(alpha.m <- mean(mph[Gender=="M"])^2 / var(mph[Gender=="M"]))
(beta.f <- mean(mph[Gender=="F"]) / var(mph[Gender=="F"]))
(beta.m <- mean(mph[Gender=="M"]) / var(mph[Gender=="M"]))

## Set up a 1x2 grid for plots
par(mfrow=c(1,2))
## Histograms for Female and Male mph
hist(mph[Gender=="F"],breaks=25,main="Histogram of Female Speeds",
   xlab="Female Speeds", xlim=c(4,11), freq=FALSE)
x.seq <- seq(4,11,.01)
lines(x.seq, dgamma(x.seq, alpha.f, beta.f))

hist(mph[Gender=="M"],breaks=25,main="Histogram of Male Speeds",
   xlab="Male Speeds", xlim=c(4,11), freq=FALSE)
lines(x.seq, dgamma(x.seq, alpha.m, beta.m))

## Quantiles: Theoretical Gamma, Actual Distribution
qgamma(c(.10,.25,.50,.75,.90), alpha.f, beta.f)
quantile(mph[Gender=="F"], c(.10,.25,.50,.75,.90))

## Quantiles: Theoretical Gamma, Actual Distribution
qgamma(c(.10,.25,.50,.75,.90), alpha.m, beta.m)
quantile(mph[Gender=="M"], c(.10,.25,.50,.75,.90))

## Probabilities: Theoretical Normal, Actual Distribution
# Theoretical Female
(mean.mph.female <- alpha.f / beta.f)
(sd.mph.female <- sqrt(alpha.f) / beta.f)
(N.female <- length(mph[Gender=="F"]))
pgamma(mean.mph.female-2*sd.mph.female, alpha.f, beta.f)
1-pgamma(mean.mph.female+2*sd.mph.female, alpha.f, beta.f)
pgamma(mean.mph.female+sd.mph.female, alpha.f, beta.f) -
    pgamma(mean.mph.female-sd.mph.female, alpha.f, beta.f)
# Actual Female
sum(mph[Gender=="F"] <= mean.mph.female-2*sd.mph.female)/N.female
sum(mph[Gender=="F"] >= mean.mph.female+2*sd.mph.female)/N.female
sum(mph[Gender=="F"] >= mean.mph.female-sd.mph.female &
    mph[Gender=="F"] <= mean.mph.female+sd.mph.female)/N.female

# Theoretical Male
(mean.mph.male <- alpha.m / beta.m)
```

```
(sd.mph.male <- sqrt(alpha.m) / beta.m)
(N.male <- length(mph[Gender=="M"]))
pgamma(mean.mph.male-2*sd.mph.male, alpha.m, beta.m)
1-pgamma(mean.mph.male+2*sd.mph.male, alpha.m, beta.m)
pgamma(mean.mph.male+sd.mph.male, alpha.m, beta.m) -
    pgamma(mean.mph.male-sd.mph.male, alpha.m, beta.m)
# Actual Male
sum(mph[Gender=="M"] <= mean.mph.male-2*sd.mph.male)/N.male
sum(mph[Gender=="M"] >= mean.mph.male+2*sd.mph.male)/N.male
sum(mph[Gender=="M"] >= mean.mph.male-sd.mph.male &
    mph[Gender=="M"] <= mean.mph.male+sd.mph.male)/N.male

## R Output

> ## Quantiles: Theoretical Gamma, Actual Distribution
> qgamma(c(.10,.25,.50,.75,.90), alpha.f, beta.f)
[1] 4.803339 5.259926 5.800466 6.376852 6.926963
> quantile(mph[Gender=="F"], c(.10,.25,.50,.75,.90))
     10%      25%      50%      75%      90%
4.811270 5.202706 5.711109 6.356917 7.015054
>
> ## Quantiles: Theoretical Gamma, Actual Distribution
> qgamma(c(.10,.25,.50,.75,.90), alpha.m, beta.m)
[1] 5.024897 5.595198 6.278232 7.014761 7.724619
> quantile(mph[Gender=="M"], c(.10,.25,.50,.75,.90))
     10%      25%      50%      75%      90%
4.969699 5.560547 6.276599 6.986362 7.717513

> # Probabilities: Theoretical Normal, Actual Distribution
> # Theoretical Female
> pgamma(mean.mph.female-2*sd.mph.female, alpha.f, beta.f)
[1] 0.01460903
> 1-pgamma(mean.mph.female+2*sd.mph.female, alpha.f, beta.f)
[1] 0.02984681
> pgamma(mean.mph.female+sd.mph.female, alpha.f, beta.f) -
+     pgamma(mean.mph.female-sd.mph.female, alpha.f, beta.f)
[1] 0.6843405
> # Actual Female
> sum(mph[Gender=="F"] <= mean.mph.female-2*sd.mph.female)/N.female
[1] 0
> sum(mph[Gender=="F"] >= mean.mph.female+2*sd.mph.female)/N.female
[1] 0.03636364
> sum(mph[Gender=="F"] >= mean.mph.female-sd.mph.female &
+     mph[Gender=="F"] <= mean.mph.female+sd.mph.female)/N.female
[1] 0.662201
>
> # Theoretical Male
> pgamma(mean.mph.male-2*sd.mph.male, alpha.m, beta.m)
[1] 0.01314067
> 1-pgamma(mean.mph.male+2*sd.mph.male, alpha.m, beta.m)
[1] 0.03094947
> pgamma(mean.mph.male+sd.mph.male, alpha.m, beta.m) -
+     pgamma(mean.mph.male-sd.mph.male, alpha.m, beta.m)
[1] 0.68497
> # Actual Male
> sum(mph[Gender=="M"] <= mean.mph.male-2*sd.mph.male)/N.male
[1] 0
> sum(mph[Gender=="M"] >= mean.mph.male+2*sd.mph.male)/N.male
[1] 0.03645117
> sum(mph[Gender=="M"] >= mean.mph.male-sd.mph.male &
+     mph[Gender=="M"] <= mean.mph.male+sd.mph.male)/N.male
[1] 0.6650619
```

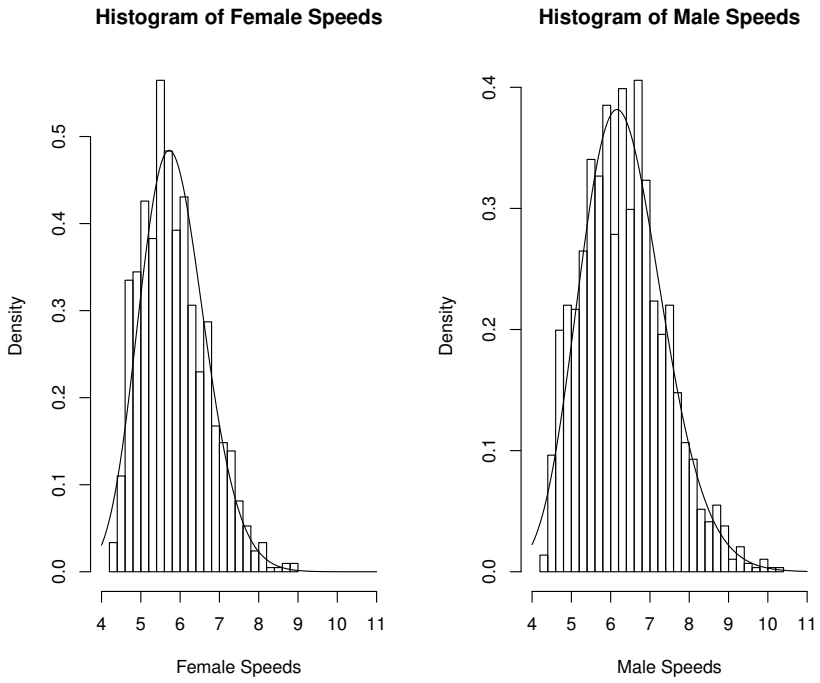

Figure 3.5: Rock and Roll Marathon speeds and Gamma Distributions for Females and Males

∇

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

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

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:

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

Here, ν is the **degrees of freedom** and we denote the distribution as: $Y \sim \chi^2_\nu$. Upper and lower critical values of the chi-square distribution are available in tabular form, and in statistical packages and spreadsheets. Probabilities and quantiles 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.

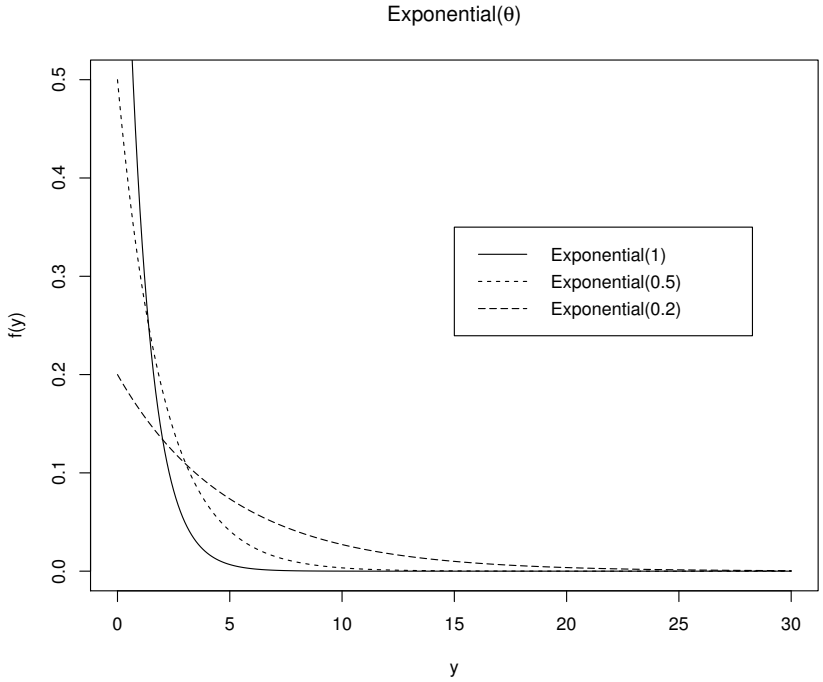

Figure 3.6: Three Exponential Densities

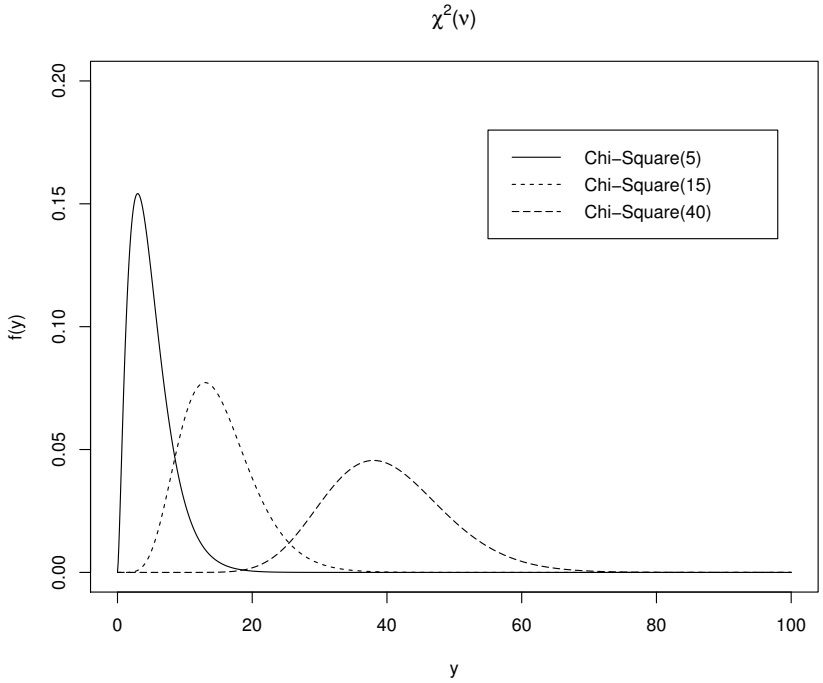

Figure 3.7: Three Chi-Square Densities

3.4.3 Beta Distribution

The Beta distribution can be used to model data that are proportions (or percentages divided by 100). The traditional model for the Beta distribution is given below.

$$f(y;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1} \qquad 0<y<1; \quad \alpha>0,\beta>0 \qquad \int_0^1 w^a(1-w)^b\,dw=\frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$$

Note that the Uniform distribution is a special case, with $\alpha=\beta=1$. The mean and variance of the Beta distribution are given here.

$$E\{Y\}=\frac{\alpha}{\alpha+\beta} \qquad V\{Y\}=\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

An alternative formulation of the distribution involves setting re-parameterizing as follows.

$$\mu=\frac{\alpha}{\alpha+\beta} \qquad \phi=\alpha+\beta \quad \Rightarrow \quad \alpha=\mu\phi \qquad \beta=(1-\mu)\phi$$

$$V\{Y\}=\sigma^2=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}=\frac{\mu(1-\mu)\phi^2}{\phi^2(\phi+1)}=\frac{\mu(1-\mu)}{\phi+1} \quad \Rightarrow \quad \phi=\frac{\mu(1-\mu)}{\sigma^2}-1$$

Figure 3.8 gives three Beta distributions.

Example 3.16: NBA 3-Point Field Goal Proportion by Team/Game - 2016/2017 Regular Season

During the NBA 2016/2017 regular season, each of the 30 teams played 82 games, for a total of 2460 team/games. For each team/game, the 3-Point field goal proportion is obtained by dividing the the number made by the number attempted. The number attempted per team/game ranged from 7 to 61, with mean and median of 27, and standard deviation of 6.7. Among the proportions made, the mean and standard deviation are 0.3566 and 0.0947, respectively. These lead to the following parameters based on the method of moments.

$$\phi=\frac{0.3566(1-0.3566)}{0.0947^2}-1=24.60 \qquad \alpha=24.60(.3566)=8.77 \qquad \beta=24.60(1-.3566)=15.83$$

A histogram of the data and the corresponding Beta density are given in Figure 3.9. As with the previous examples, we compare the theoretical quantiles and probabilities for the beta denstities with the actual values for this population. They show considerable agreement.

R Commands and Output

```
## R Commands
```

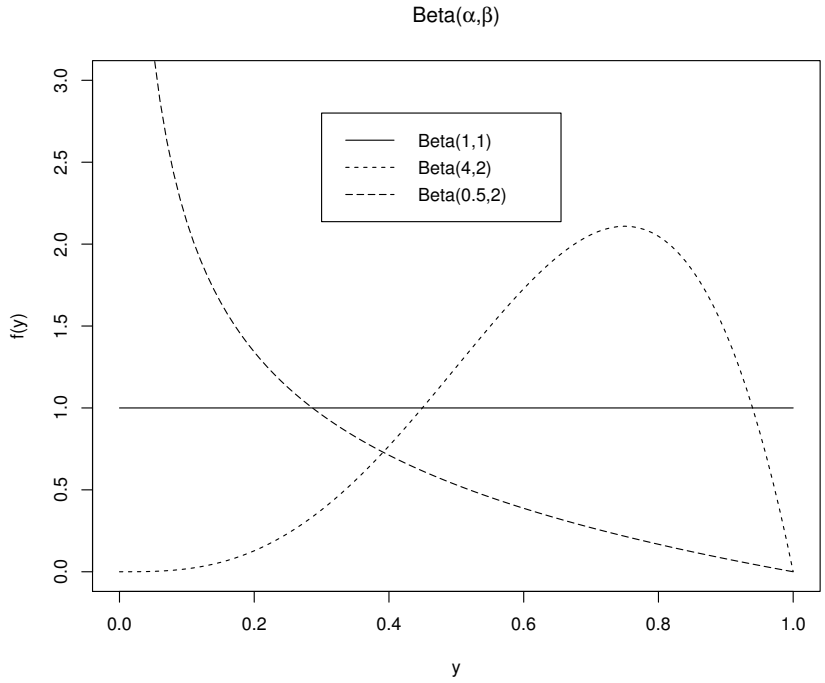

Figure 3.8: Three Beta Densities

```
nba2017 <- read.csv("http://www.stat.ufl.edu/~winner/data/nba_teamgame_20167.csv")
attach(nba2017); names(nba2017)

# Regular Season Games Only (GameType=1)
fg3prop <- fg3m[GameType==1]/fg3a[GameType==1]
summary(fg3a[GameType==1])
sd(fg3a[GameType==1])

# Function to compute phi, alpha, beta from mean, sd
betaShRtMeanSD <- function(mean, sd) {
  if (mean <= 0 | sd <= 0) return("FAIL")
  phi <- (mean*(1-mean) / sd^2) - 1
  alpha <- mean*phi
  beta <- (1-mean) * phi
  return(list(phi=phi, alpha=alpha, beta=beta))
}

(fg3ab <-  betaShRtMeanSD(mean(fg3prop),
   sd(fg3prop)))

(mean.fg3 <- mean(fg3prop))
(sd.fg3 <- sd(fg3prop))
(N.fg3 <- length(fg3prop))
hist(fg3prop, xlim=c(0,1), freq=FALSE, breaks=35,
   main="Histogram of 3-Point Field Goal Proportions and Beta(8.77,15.83)")
x <- seq(0,1.0,0.01)
lines(x,dbeta(x, fg3ab$alpha, fg3ab$beta))

## Quantiles: Theoretical Beta, Actual Distribution
qbeta(c(.10,.25,.50,.75,.90), fg3ab$alpha, fg3ab$beta)
```

```
quantile(fg3prop, c(.10,.25,.50,.75,.90))

## Probabilities: Theoretical Beta, Actual Distribution
# Theoretical
pbeta(mean.fg3-2*sd.fg3, fg3ab$alpha, fg3ab$beta)
1-pbeta(mean.fg3+2*sd.fg3, fg3ab$alpha, fg3ab$beta)
pbeta(mean.fg3+sd.fg3, fg3ab$alpha, fg3ab$beta) -
    pbeta(mean.fg3-sd.fg3, fg3ab$alpha, fg3ab$beta)
# Actual
sum(fg3prop <= mean.fg3-2*sd.fg3)/N.fg3
sum(fg3prop >= mean.fg3+2*sd.fg3)/N.fg3
sum(fg3prop >= mean.fg3-sd.fg3 &
    fg3prop <= mean.fg3+sd.fg3)/N.fg3

## Output

> ## Quantiles: Theoretical Normal, Actual Distribution
> qbeta(c(.10,.25,.50,.75,.90), fg3ab$alpha, fg3ab$beta)
[1] 0.2366108 0.2892122 0.3526696 0.4198314 0.4819008
> quantile(fg3prop, c(.10,.25,.50,.75,.90))
      10%       25%       50%       75%       90%
0.2380952 0.2941176 0.3548387 0.4146341 0.4782609
>
> ## Probabilities: Theoretical Normal, Actual Distribution
> # Theoretical
> pbeta(mean.fg3-2*sd.fg3, fg3ab$alpha, fg3ab$beta)
[1] 0.01394928
> 1-pbeta(mean.fg3+2*sd.fg3, fg3ab$alpha, fg3ab$beta)
[1] 0.02821849
> pbeta(mean.fg3+sd.fg3, fg3ab$alpha, fg3ab$beta) -
+     pbeta(mean.fg3-sd.fg3, fg3ab$alpha, fg3ab$beta)
[1] 0.6741803
> # Actual
> sum(fg3prop <= mean.fg3-2*sd.fg3)/N.fg3
[1] 0.02357724
> sum(fg3prop >= mean.fg3+2*sd.fg3)/N.fg3
[1] 0.0296748
> sum(fg3prop >= mean.fg3-sd.fg3 &
+     fg3prop <= mean.fg3+sd.fg3)/N.fg3
[1] 0.6829268
```

∇

3.4.4 Functions of Normal Random Variables

First, note that if $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$. Many software packages present Z-tests as (Wald) χ^2-tests.

Suppose $Y_1, ..., Y_n$ are independent with $Y_i \sim N\left(\mu, \sigma^2\right)$ for $i = 1, \ldots, n$. Then the sample mean and sample variance are computed as follow.

$$\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n} \qquad S^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}$$

In this case, we obtain the following sampling distributions for the mean and a function of the variance.

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{n-1} \qquad \overline{Y}, \quad \frac{(n-1)S^2}{\sigma^2} \text{ are independent.}$$

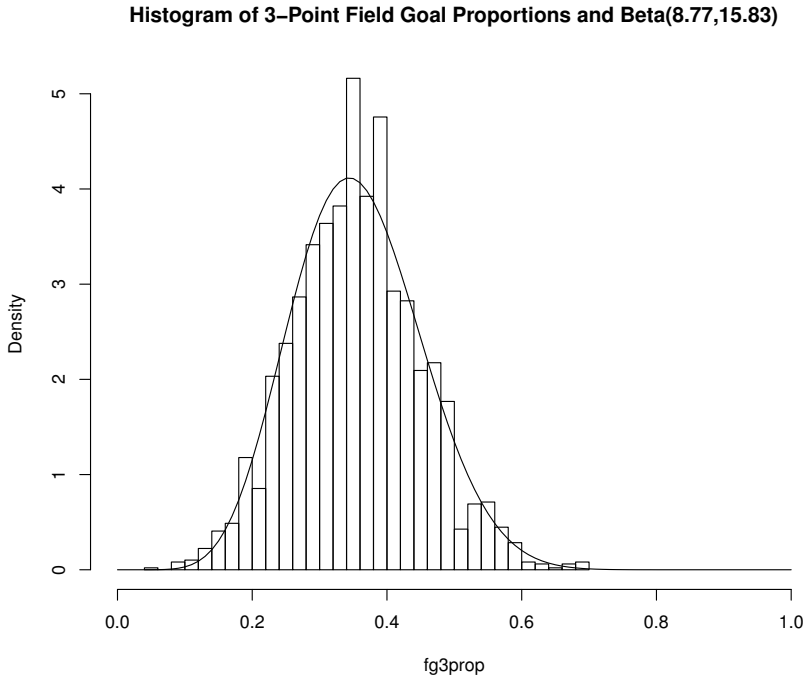

Figure 3.9: Three Point Field Goal proportions by team/game - NBA 2016/2017 regular season

Note that in general, if $Y_1, ..., Y_n$ are normally distributed (and not necessarily with the same mean and/or variance), any linear function of them will be normally distributed, with mean and variance given previously in the section with linear functions of random variables.

Two distributions associated with the normal and chi-squared distributions are **Student's** t and F. Student's t-distribution is similar to the standard normal ($N(0,1)$), except that it is indexed by its degrees of freedom and that is has heavier tails than the standard normal. As its degrees of freedom approach infinity, its distribution converges to the standard normal. Let $Z \sim N(0,1)$ and $W \sim \chi^2_\nu$, where Z and W are independent. Then, we have the following result.

$$Y \sim N\left(\mu, \sigma^2\right) \quad \Rightarrow \quad Z = \frac{Y-\mu}{\sigma} \sim N(0,1) \qquad\qquad T = \frac{Z}{\sqrt{W/\nu}} \sim t_\nu$$

where the probability density, mean, and variance for Student's t-distribution are:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad E\{T\} = \mu_T = 0 \qquad V\{T\} = \frac{\nu}{\nu-2} \quad \nu > 2$$

and we use the notation $T \sim t_\nu$. Three t-distributions, along with the standard normal (z) distribution are shown in Figure 3.10.

Now consider the sample mean and variance, and the fact they are independent.

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad Z = \frac{\overline{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n}\frac{\overline{Y}-\mu}{\sigma} \sim N(0,1)$$

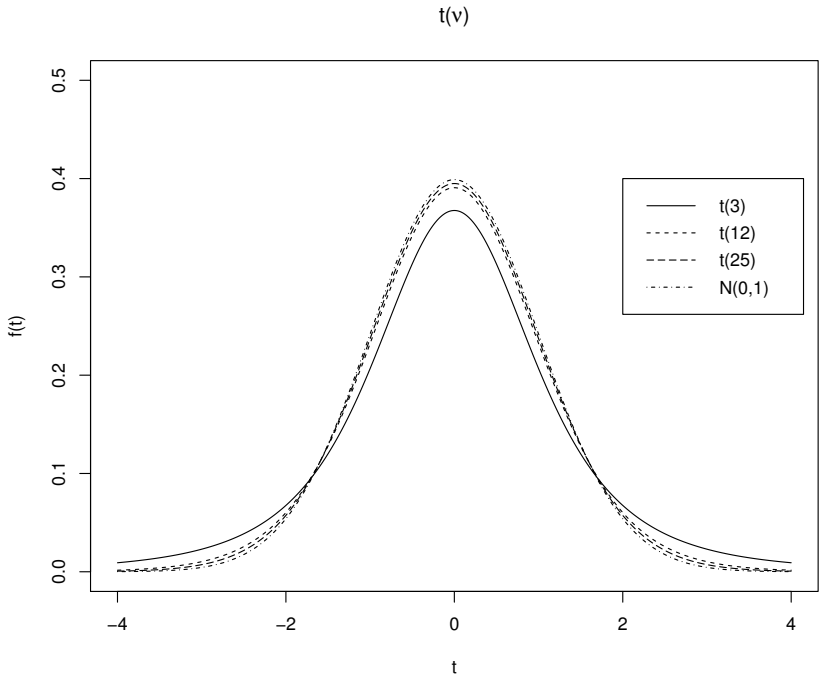

Figure 3.10: Three t-densities and z

$$W = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{n-1} \quad \Rightarrow \quad \sqrt{\frac{W}{\nu}} = \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = \frac{S}{\sigma}$$

$$\Rightarrow \quad T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}\frac{\overline{Y}-\mu}{\sigma}}{\frac{S}{\sigma}} = \sqrt{n}\frac{\overline{Y}-\mu}{S} \sim t_{n-1}$$

The F-distribution arises often in Regression and Analysis of Variance applications. If $W_1 \sim \chi^2(\nu_1)$, $W_2 \sim \chi^2(\nu_2)$, and W_1, W_2 are independent, then:

$$F = \frac{\left[\frac{W_1}{\nu_1}\right]}{\left[\frac{W_2}{\nu_2}\right]} \sim F_{\nu_1,\nu_2}.$$

where the probability density, mean, and variance for the F-distribution are given below as a function of the specific point $F = f$.

$$f(f) = \left[\frac{\Gamma\left(\left(\frac{\nu_1+\nu_2}{2}\right)\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\right]\left[\frac{f^{\nu_1/2-1}}{(\nu_1 f+\nu_2)^{(\nu_1+\nu_2)/2}}\right]$$

$$E\{F\} = \mu_F = \frac{\nu_1}{\nu_2-2} \quad \nu_2 > 2 \qquad V\{F\} = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)(\nu_2-4)} \quad \nu_2 > 4$$

Three F-distributions are given in Figure 3.11.

Critical values for the t, χ^2, and F-distributions are given in statistical textbooks and webpages. Probabilities and quantiles can be obtained from many statistical packages and spreadsheets. Technically, the t, χ^2, and F distributions described here are **central** t, **central** χ^2, and **central** F distributions. These will be made use of repeatedly when we make inferences regarding population parameters.

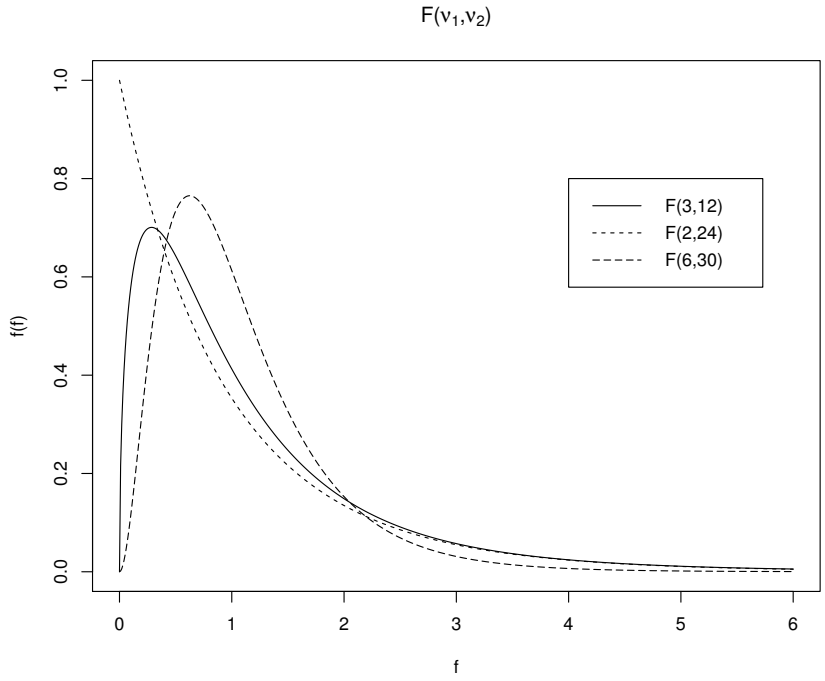

Figure 3.11: Three F-densities

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.

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

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

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

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

The standard distribution of the sampling distribution of a sample statistic (aka estimator) is referred to as its **standard error**. Thus $\sigma_{\overline{Y}}$ is the standard error of the sample mean, and $\sigma_{\hat{\pi}}$ 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 3.17: Sampling Distributions - NHL BMI, Female Marathon Speeds, Charlotte Traffic Stops

We consider the sampling distributions of sample means for the NHL player Body Mass Indices, Female Rock and Roll Marathon Speeds, and Charlotte, N.C. traffic stops (proportion of stops due to speed violations, category 7). For the NHL BMI data, the population mean is $\mu = 26.51$ and standard deviation is $\sigma = 1.48$. 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 he normal density are shown in Figure 3.12.

$$\text{Theory: } \mu_{\overline{Y}} = \mu = 26.510 \quad \sigma_{\overline{Y}} = \frac{1.476}{\sqrt{9}} = 0.492 \qquad \text{Empirical: } \overline{\overline{y}} = 26.508 \quad s_{\overline{y}} = 0.493$$

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 he normal density are shown in Figure 3.13.

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

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.

Finally, we consider the proportions of traffic stops due to speeding for the Charlotte, NC traffic stops, based on 10000 random samples of $n = 50$. For the population, $\pi = 22222/79884 = .2782$. The theoretical and empirical results are given below, and the histogram is given in Figure 3.14.

$$\text{Theory: } \mu_{\hat{\pi}} = \pi = .2782 \quad \sigma_{\hat{\pi}} = \sqrt{\frac{.2782(1-.2782)}{50}} = .0634 \qquad \text{Empirical: } \overline{\hat{\pi}} = .2776 \quad s_{\hat{\pi}} = .0631$$

The empirical mean and standard error, again, are in strong agreement with their theoretical values. Note that the sample proportion is discrete as it can only take on values .00, .02, ..., .98, 1.00, as $n = 50$. The histogram is clearly bell-shaped like a normal distribution. As n gets larger $\hat{\pi}$ becomes more "continuous."

R Program

```
### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Generate random values (-0.5 to 0.5) to add to Height
set.seed(1234)
N <- NROW(nhl)
Height.dev <- 0.5 - runif(N)
Height <- Height + Height.dev

### Compute BMI
bmi.nhl <- 703 * Weight / (Height^2)

mean(bmi.nhl)
sd(bmi.nhl)

num.sim <- 10000
num.sample <- 9
sampmean.bmi <- rep(0,num.sim)
for (i in 1:num.sim) {
  sample <- sample(1:N, num.sample, replace=F)
  sampmean.bmi[i] <- mean(bmi.nhl[sample])
}

mean(sampmean.bmi)
sd(sampmean.bmi)

hist(sampmean.bmi, breaks=50, xlim=c(24,29), freq=F,
main="Sampling Distribution of Sample Mean, n=9")
bmi.seq <- seq(24,29,0.01)
lines(bmi.seq,dnorm(bmi.seq,mean(bmi.nhl),sd(bmi.nhl)/sqrt(num.sample)))

detach(nhl)


## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)

f.mph <- mph[Gender == "F"]
mean(f.mph)
sd(f.mph)
N <- length(f.mph)

num.sim <- 10000
num.sample <- 16
```

```
sampmean.fmph <- rep(0,num.sim)
for (i in 1:num.sim) {
  sample <- sample(1:N, num.sample, replace=F)
  sampmean.fmph[i] <- mean(f.mph[sample])
}

mean(sampmean.fmph)
sd(sampmean.fmph)

hist(sampmean.fmph, breaks=100, xlim=c(4.80,7.00), freq=F,
main="Sampling Distribution of Sample Mean, n=16")
fmph.seq <- seq(4.80,7.00,0.01)
lines(fmph.seq,dnorm(fmph.seq,mean(f.mph),sd(f.mph)/sqrt(num.sample)))

detach(rr.mar)

## Read data off web page, attach file as data frame, and list variable names
clt2016 <- read.csv("http://www.stat.ufl.edu/~winner/data/trafficstop.csv")
attach(clt2016); names(clt2016)

table(RsnStop)
N <- length(RsnStop)
table(RsnStop)/N

num.sim <- 10000
num.sample <- 50
sampprop.cltspd <- rep(0,num.sim)
for (i in 1:num.sim) {
  sample <- sample(1:N, num.sample, replace=F)
  sampprop.cltspd[i] <- sum(RsnStop[sample] == 7) / num.sample
}

mean(sampprop.cltspd)
sd(sampprop.cltspd)

hist(sampprop.cltspd, breaks=100, xlim=c(0,0.50), freq=F,
main="Sampling Distribution of Sample Mean, n=50")
```

∇

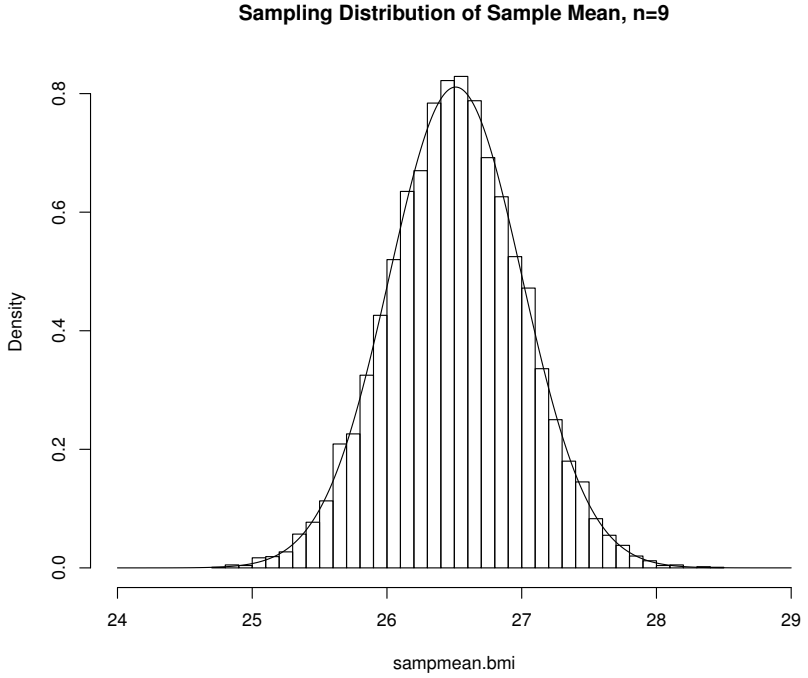

Figure 3.12: Sampling distribution for sample means (n=9) for NHL Body Mass Index

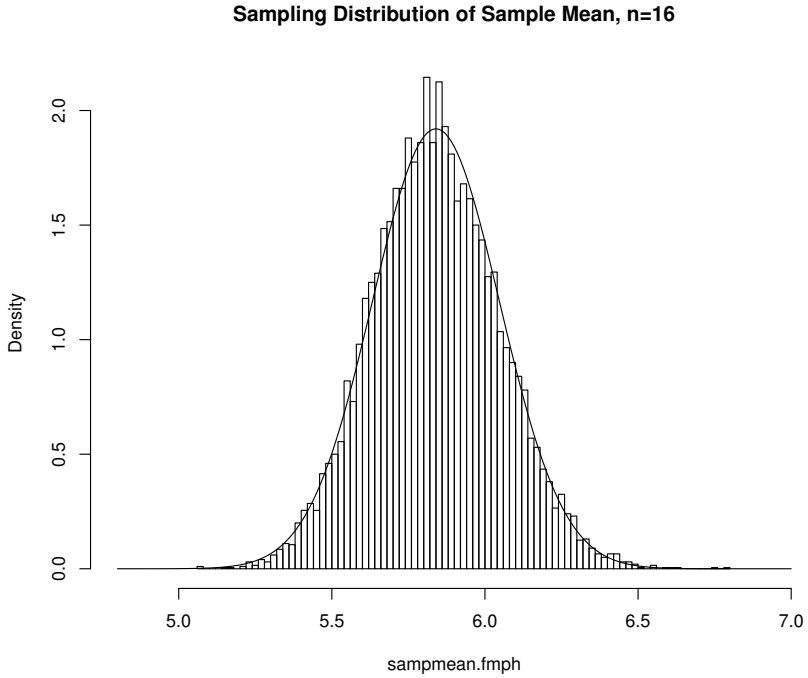

Figure 3.13: Sampling Distribution for sample means (n=16) for Female Rock and Roll Marathon speeds

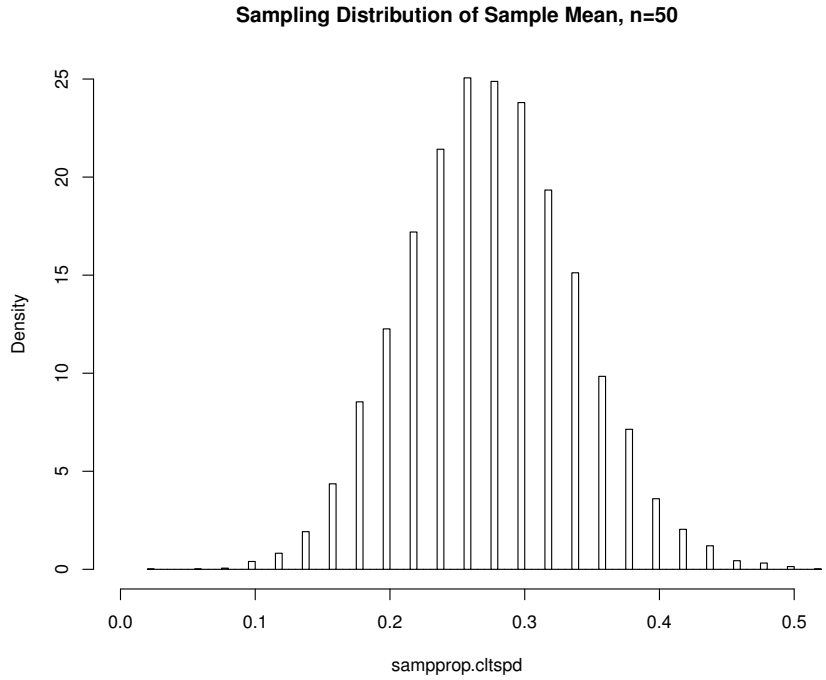

Figure 3.14: Sampling Distribution for sample proportions (n=50) for Charlotte traffic stops for speeding

Chapter 4

Inferences Concerning Population Means and Medians

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.

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

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

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

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

Note that in the probability statements above, μ is a fixed, unknown in practice constant, and $\overline{Y}$ is a random variable that changes from sample to sample. The probability refers to the fraction of the samples that will provide sample means so 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.

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

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

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

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 must use the z-distribution.

Example 4.1: NHL Players' BMI

The Body Mass Indices for the NHL players are approximately normally distributed with mean $\mu = 26.510$ and standard deviation $\sigma = 1.476$. We take 10000 random samples of size $n = 12$, implying a standard error of $\sigma_{\overline{Y}} = 1.476/\sqrt{12} = 0.426$. 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, 9019 (90.19%) lied within $\mu \pm 1.645(.426)$, 9513 (95.13%) within $\mu \pm 1.96(.426)$, and 9822 (98.22%) within $\mu \pm 2.576(.426)$. Had we constructed intervals of the form $\overline{y} \pm z_{\alpha/2}(.426)$ for each sample mean, the coverage rates for μ would have been the same values (90.19%, 95.13%, 98.22%).

When the population standard error $\sigma_{\overline{Y}} = \sigma/\sqrt{n}$ is replaced by the estimated standard error $s_{\overline{Y}} = s/\sqrt{n}$, 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 87.42%, 92.61%, and 96.14%, 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: 89.99%, 95.25%, and 99.21%, respectively.

For the first random sample of the 10000 generated, we observe $\overline{y} = 25.974$ and $s = 1.296$. The 95% Confidence Interval for μ based on the first sample is obtained as follows.

$$\overline{y} \pm t_{.025,n-1}\frac{s}{\sqrt{n}} \quad \equiv \quad 25.974 \pm 2.201\left(\frac{1.296}{\sqrt{12}}\right) \quad \equiv \quad 25.974 \pm 0.823 \quad \equiv \quad (25.151, 26.797)$$

Thus, this interval does contain $\mu = 26.510$.

R Program and Output

```
## Commands


### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Generate random values (-0.5 to 0.5) to add to Height
set.seed(1234)
N <- NROW(nhl)
Height.dev <- 0.5 - runif(N)
Height <- Height + Height.dev

### Compute BMI
bmi.nhl <- 703 * Weight / (Height^2)

set.seed(98765)
num.sim <- 10000
n.sample <- 12
samp.mean <- rep(0, num.sim)
samp.sd <- rep(0, num.sim)
mu.bmi <- mean(bmi.nhl)
std.err.bmi <- sd(bmi.nhl)/sqrt(n.sample)

for (i in 1:num.sim) {
  sample <- sample(1:N, n.sample, replace=FALSE)
  samp.mean[i] <- mean(bmi.nhl[sample])
  samp.sd[i] <- sd(bmi.nhl[sample])
}
cbind(samp.mean[1], samp.sd[1])

z_050 <- qnorm(.95, 0, 1)
z_025 <- qnorm(.975, 0, 1)
z_005 <- qnorm(.99, 0, 1)
(cover10a <- sum(samp.mean >= mu.bmi - z_050*std.err.bmi &
               samp.mean <= mu.bmi + z_050*std.err.bmi) / num.sim)
```

```
(cover05a <- sum(samp.mean >= mu.bmi - z_025*std.err.bmi &
                 samp.mean <= mu.bmi + z_025*std.err.bmi) / num.sim)
(cover01a <- sum(samp.mean >= mu.bmi - z_005*std.err.bmi &
                 samp.mean <= mu.bmi + z_005*std.err.bmi) / num.sim)

samp.se <- samp.sd / sqrt(n.sample)
(cover10b <- sum(samp.mean >= mu.bmi - z_050*samp.se &
                 samp.mean <= mu.bmi + z_050*samp.se) / num.sim)
(cover05b <- sum(samp.mean >= mu.bmi - z_025*samp.se &
                 samp.mean <= mu.bmi + z_025*samp.se) / num.sim)
(cover01b <- sum(samp.mean >= mu.bmi - z_005*samp.se &
                 samp.mean <= mu.bmi + z_005*samp.se) / num.sim)

t_050 <- qt(.95,11); t_025 <- qt(.975,11); t_005 <- qt(.995,11)
(cover10c <- sum(samp.mean >= mu.bmi - t_050*samp.se &
                 samp.mean <= mu.bmi + t_050*samp.se) / num.sim)
(cover05c <- sum(samp.mean >= mu.bmi - t_025*samp.se &
                 samp.mean <= mu.bmi + t_025*samp.se) / num.sim)
(cover01c <- sum(samp.mean >= mu.bmi - t_005*samp.se &
                 samp.mean <= mu.bmi + t_005*samp.se) / num.sim)

### Output
> cbind(samp.mean[1], samp.sd[1])
         [,1]     [,2]
[1,] 25.97359 1.296183
> (cover10a <- sum(samp.mean >= mu.bmi - z_050*std.err.bmi &
+                  samp.mean <= mu.bmi + z_050*std.err.bmi) / num.sim)
[1] 0.9069
> (cover05a <- sum(samp.mean >= mu.bmi - z_025*std.err.bmi &
+                  samp.mean <= mu.bmi + z_025*std.err.bmi) / num.sim)
[1] 0.9511
> (cover01a <- sum(samp.mean >= mu.bmi - z_005*std.err.bmi &
+                  samp.mean <= mu.bmi + z_005*std.err.bmi) / num.sim)
[1] 0.9796

> (cover10b <- sum(samp.mean >= mu.bmi - z_050*samp.se &
+                  samp.mean <= mu.bmi + z_050*samp.se) / num.sim)
[1] 0.8742
> (cover05b <- sum(samp.mean >= mu.bmi - z_025*samp.se &
+                  samp.mean <= mu.bmi + z_025*samp.se) / num.sim)
[1] 0.9261
> (cover01b <- sum(samp.mean >= mu.bmi - z_005*samp.se &
+                  samp.mean <= mu.bmi + z_005*samp.se) / num.sim)
[1] 0.9614

> (cover10c <- sum(samp.mean >= mu.bmi - t_050*samp.se &
+                  samp.mean <= mu.bmi + t_050*samp.se) / num.sim)
[1] 0.8999
> (cover05c <- sum(samp.mean >= mu.bmi - t_025*samp.se &
+                  samp.mean <= mu.bmi + t_025*samp.se) / num.sim)
[1] 0.9525
> (cover01c <- sum(samp.mean >= mu.bmi - t_005*samp.se &
+                  samp.mean <= mu.bmi + t_005*samp.se) / num.sim)
[1] 0.9921
```

∇

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 mean to be within E with

confidence level $1-\alpha$. Note that 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.

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

Example: 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 Commands and Output

```
## Commands

E <- 0.20
sigma <- 1.058
alpha <- 0.05
n <- 1
E.t <- E+1
# Keep increasing $n$ until E.t < E
while (E.t >= E) {
   n <- n+1
   E.t <- qt(1-alpha/2,n-1)*sigma/sqrt(n)
}
cbind(n, E.t)

## 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 obtain 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. 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. 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 (typically $\alpha = P(\text{Type I Error})$ is 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 of ten selected so that it has a particular small probability for an important difference. $\beta = P(\text{Type II Error})$.

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 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.

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

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.

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

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

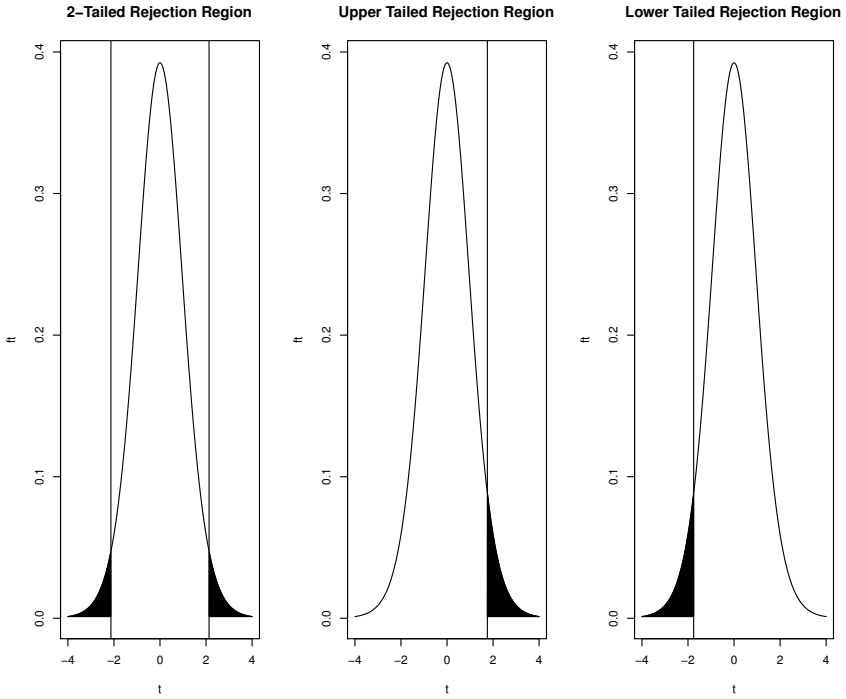

Figure 4.1: Rejection Regions for 2-tailed, Upper and Lower tailed tests, with $\alpha = 0.05$ and $n = 16$

Lower tailed: $H_0 : \mu \geq \mu_0 \qquad H_A : \mu < \mu_0 \qquad$ T.S.: $t_{obs} = \dfrac{\overline{y} - \mu_0}{s/\sqrt{n}} \quad$ 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.

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

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

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

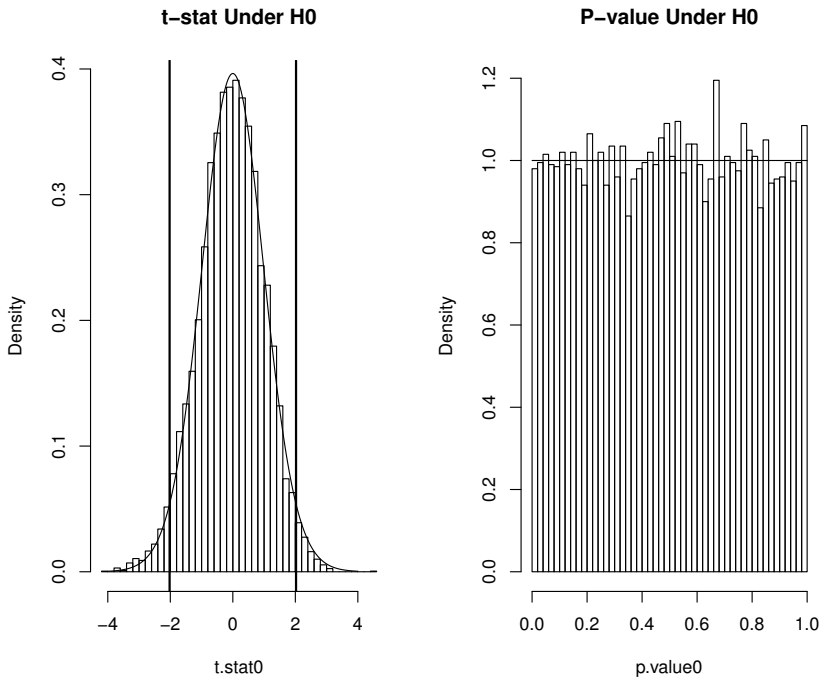

Figure 4.2: t-statistics and P-values for testing $H_0 : \mu = 6.337$

Lower tailed tests: $\pi = P\left(t_{n-1,\Delta} \leq -t_{\alpha,n-1}\right)$ 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 individuals), we will demonstrate the procedure based on male Rock and Roll marathon speeds.

Example 4.2: 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. We find that 475 (4.75%) 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 is given in Figure 4.2. The two vertical bars on the t-statistic plot at $\pm t_{.025,39} = \pm 2.023$.

Next we consider cases where the null hypothesis is not true. We 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 $\mu = 6.337$ than the null value for H_{01}, we will expect that we reject H_{02} less often for tests based on the same sample

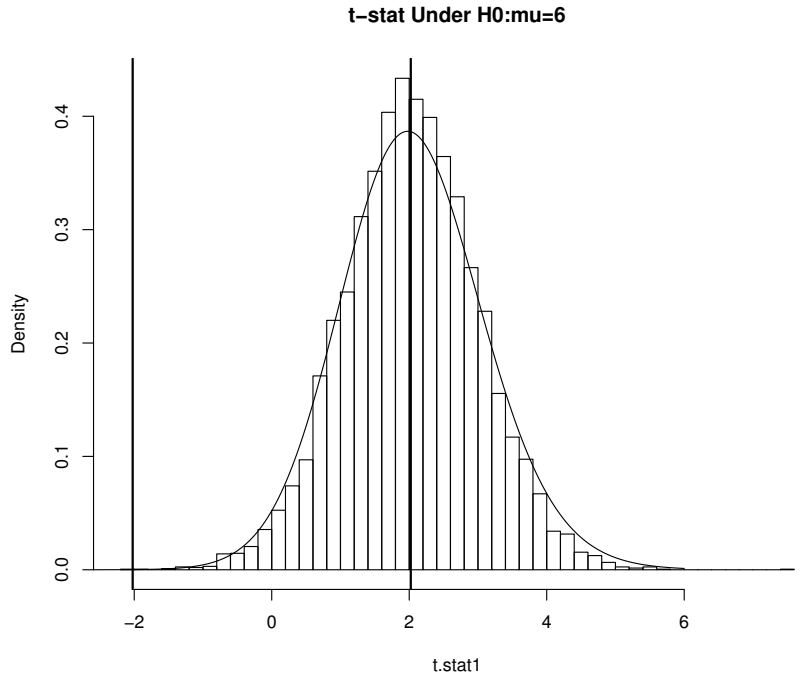

Figure 4.3: t-statistics and non-central t-distribution for testing $H_0 : \mu = 6.0$

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 $n = 40$.

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

Based on 10000 random samples from the male marathon speeds, 49.28% rejected $H_0 : \mu = 6$, and for another set of 10000 random samples, 16.83% rejected $H_0 : \mu = 6.5$. The histogram of the test statistics and the non-central t-distribution are given in Figure 4.3 for testing $H_0 : \mu = 6$.

R Commands and Output

```
## R Commands

## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)

male.mph <- mph[Gender == "M"]

N <- length(male.mph)
mu0 <- mean(male.mph)
```

```
sigma <- sd(male.mph)
set.seed(13579)
num.sim <- 10000
num.samp <- 40
cv.lo <- qt(.025,num.samp-1)
cv.hi <- qt(.975,num.samp-1)
t.stat0 <- rep(0, num.sim)
p.value0 <- rep(0, num.sim)

for (i in 1:num.sim) {
   sample <- sample(1:N, num.samp, replace=FALSE)
   ybar <- mean(male.mph[sample])
   s <- sd(male.mph[sample])
   t.stat0[i] <- (ybar - mu0) / (s / sqrt(num.samp))
   p.value0[i] <- 2*(1-pt(abs(t.stat0[i]),num.samp-1))
}

sum(p.value1 <= 0.05) / num.sim
par(mfrow=c(1,2))
hist(t.stat0, breaks=50, freq=FALSE, main="t-stat Under H0")
xt <- seq(-4,4,.01)
lines(xt, dt(xt,num.samp-1))
abline(v=cv.lo,lwd=2)
abline(v=cv.hi,lwd=2)


hist(p.value0, breaks=50, freq=FALSE, main="P-value Under H0")
xp <- seq(0,1,0.01)
lines(xp,dbeta(xp,1,1))

mu01 <- 6.0
(Delta1 <- (mu0 - mu01) / (sigma/sqrt(num.samp)))
(power1 <- pt(cv.lo, num.samp-1, Delta1) + (1-pt(cv.hi, num.samp-1, Delta1)))

mu02 <- 6.5
(Delta2 <- (mu0 - mu02) / (sigma/sqrt(num.samp)))
(power2 <- pt(cv.lo, num.samp-1, Delta2) + (1-pt(cv.hi, num.samp-1, Delta2)))

set.seed(1234)
t.stat1 <- rep(0, num.sim)
p.value1 <- rep(0, num.sim)

for (i in 1:num.sim) {
   sample <- sample(1:N, num.samp, replace=FALSE)
   ybar <- mean(male.mph[sample])
   s <- sd(male.mph[sample])
   t.stat1[i] <- (ybar - mu01) / (s / sqrt(num.samp))
   p.value1[i] <- 2*(1-pt(abs(t.stat1[i]),num.samp-1))
}

sum(t.stat1 <= cv.lo | t.stat1 >= cv.hi) / num.sim

par(mfrow=c(1,1))
hist(t.stat1, breaks=50, freq=FALSE, main="t-stat Under H0:mu=6")
xt <- seq(-4,6,.01)
lines(xt, dt(xt,num.samp-1,Delta1))
abline(v=cv.lo,lwd=2)
abline(v=cv.hi,lwd=2)

set.seed(5678)
t.stat2 <- rep(0, num.sim)
p.value2 <- rep(0, num.sim)

for (i in 1:num.sim) {
```

```
   sample <- sample(1:N, num.samp, replace=FALSE)
   ybar <- mean(male.mph[sample])
   s <- sd(male.mph[sample])
   t.stat2[i] <- (ybar - mu02) / (s / sqrt(num.samp))
   p.value2[i] <- 2*(1-pt(abs(t.stat2[i]),num.samp-1))
}

sum(t.stat2 <= cv.lo | t.stat2 >= cv.hi) / num.sim

## Output

> sum(p.value0 <= 0.05) / num.sim
[1] 0.0475
> (Delta1 <- (mu0 - mu01) / (sigma/sqrt(num.samp)))
[1] 2.015005
> (power1 <- pt(cv.lo, num.samp-1, Delta1) + (1-pt(cv.hi, num.samp-1, Delta1)))
[1] 0.5021642
> mu02 <- 6.5
> (Delta2 <- (mu0 - mu02) / (sigma/sqrt(num.samp)))
[1] -0.9748006
> (power2 <- pt(cv.lo, num.samp-1, Delta2) + (1-pt(cv.hi, num.samp-1, Delta2)))
[1] 0.1582836
> sum(t.stat1 <= cv.lo | t.stat1 >= cv.hi) / num.sim
[1] 0.5013
> sum(t.stat2 <= cv.lo | t.stat2 >= cv.hi) / num.sim
[1] 0.1683
```

∇

4.2.1 Choosing Sample Size for Fixed Power for an Alternative

Once an important difference $\mu_0 - \mu_A$ is determined, and an estimate of σ is obtained, the functions involving the non-central t-distribution can be used iteratively to find the n that makes the power large enough. The algorithm goes as follows for 2-tailed tests.

1. Choose an important difference $\mu_0 - \mu_A$ and σ. Or alternatively make the difference in units of σ: $(\mu_0 - \mu_a)/\sigma$.
2. Start with small value for n, and compute the critical values for the t-test: $CV_{LO} = -t_{\alpha/2,n-1}$, $CV_{HI} = t_{\alpha/2,n-1}$.
3. Compute $\Delta = (\mu_0 - \mu_A)/\left(\sigma/\sqrt{n}\right)$.
4. Obtain the probability the test statistic falls in the Rejection Region, based on the non-central t-distribution, with $n-1$ degrees of freedom, and non-centrality parameter Δ: Power = pt(CV_{LO},$n-1$,Δ) + (1-pt(CV_{HI},$n-1$,Δ))
5. Continue increasing n until Power exceeds some specified value (typically 0.80 or 0.90)

Example 4.3: Male Rock and Roll Marathon Speeds

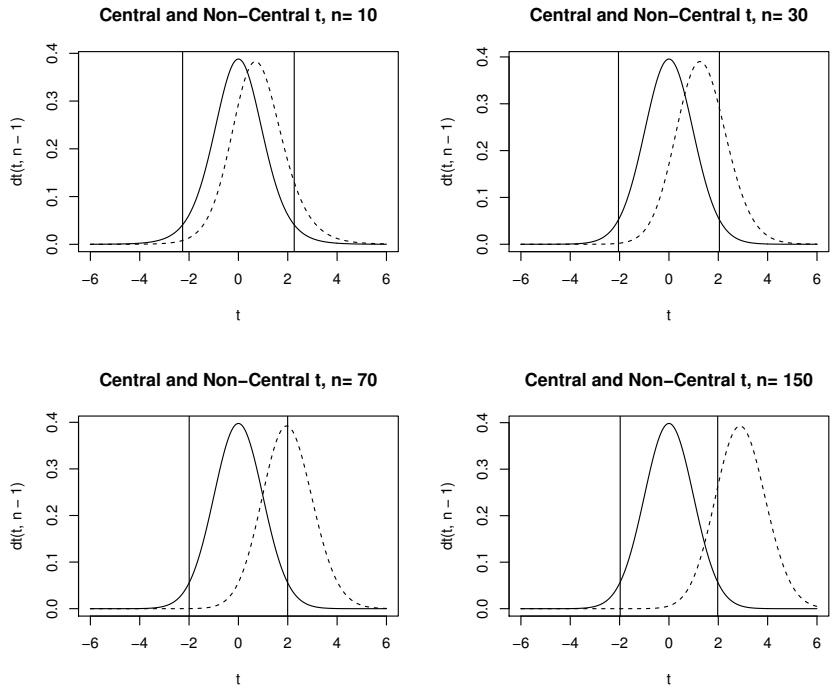

Figure 4.4: Central and non-Central t-distributions for n=10, 30, 70, 150, $\mu_0 - \mu_A = 0.25$, and $\sigma = 1.058$

Suppose we would like to be able to detect a difference between μ_0 and μ_A of 0.25 with power of $\pi = 0.8$ when the test is conducted at $\alpha = 0.05$. In this case, recall $\sigma = 1.058$. Start with $n = 3$.

$$t_{.025,3-1} = 4.303 \qquad \Delta = \frac{0.25}{1.058/\sqrt{3}} = \sqrt{3}\frac{0.25}{1.058} = 0.409$$

$$\pi = P\left(t_{3-1,0.409} \leq -4.303\right) + P\left(t_{3-1,0.409} \geq 4.303\right) = .0577$$

Keep increasing n, which affects the critical t-values (making them smaller in absolute value) and increasing Δ, thus increasing the power of the test, until $\pi \geq 0.80$. It ends up that we would need a sample of $n = 143$ to meet the power requirement. The target difference is very small (0.25) relative to the standard deviation (1.058) which is why such a large sample would be needed. A plot of the central and non-central t-distributions for n=10, 30, 70, and 150 is given in Figure 4.4. The vertical bars give the critical values for the $\alpha = 0.05$ level test.

R Commands and Output

```
## Commands

n <- 3
alpha <- 0.05
mu_diff <- 0.25
```

```
sigma <- 1.058
Delta <- mu_diff/(sigma/sqrt(n))
CV_LO <- qt(alpha/2,n-1)
CV_HI <- qt(1-alpha/2,n-1)
(power <- pt(CV_LO,n-1,Delta) + (1-pt(CV_HI,n-1,Delta)))

while (power <= 0.80) {
   n <- n+1
   Delta <- mu_diff/(sigma/sqrt(n))
   CV_LO <- qt(alpha/2,n-1)
   CV_HI <- qt(1-alpha/2,n-1)
   power <- pt(CV_LO,n-1,Delta) + (1-pt(CV_HI,n-1,Delta))
}

cbind(n, power)

par(mfrow=c(2,2))
t <- seq(-6,6,0.01)
mu_diff <- 0.25
sigma <- 1.058
for (n in c(10, 30, 70, 150)) {
  Delta <- mu_diff/(sigma/sqrt(n))
  plot(t,dt(t,n-1),type="l",main=paste("Central and Non-Central t, n=",n))
  lines(t,dt(t,n-1,Delta),lty=2)
  abline(v=qt(.025,n-1))
  abline(v=qt(.975,n-1))
}

## Output

> (power <- pt(CV_LO,n-1,Delta) + (1-pt(CV_HI,n-1,Delta)))
[1] 0.05772603
> cbind(n, power)
       n     power
[1,] 143 0.8013787
```

∇

4.3 Inferences Concerning the Population Median

The population median represents the 50^{th} percentile of the distribution. For each sampled observation, there is a 0.5 probability that]it is larger or smaller than the median. The number of observations of a random sample of size n that are above (or below) the median is binomial with n trials, and probability of success $\pi = 0.5$. Let $B_{\alpha/2,n}$ be the smallest number such that $P\left(Y \leq B_{\alpha/2,n} | Y\ Bin(n, 0.5)\right) \leq \alpha/2$. Then the probability that the number of sample observations falling above or below the median will lie in the range $\left(L_{\alpha/2} = B_{\alpha/2,n} + 1, U_{\alpha/2} = n - B_{\alpha/2,n}\right)$ will be greater than or equal to $1 - \alpha$. This leads to a $(1 - \alpha)100\%$ Confidence Interval for the population median to be the range encompassed by the $(L_{\alpha/2})^{th}$ ordered observation to the $U^{th}_{\alpha/2}$ ordered observation.

A large-sample approximation based on the normal distribution involves taking the range encompassed by the observations within ranks $(n/2) \pm n^{1/2}$. This is a result of the standard error of Y being $\sqrt{n(0.5)(1 - 0.5)}$, and using mean plus/minus 2 standard errors for approximate 95% confidence.

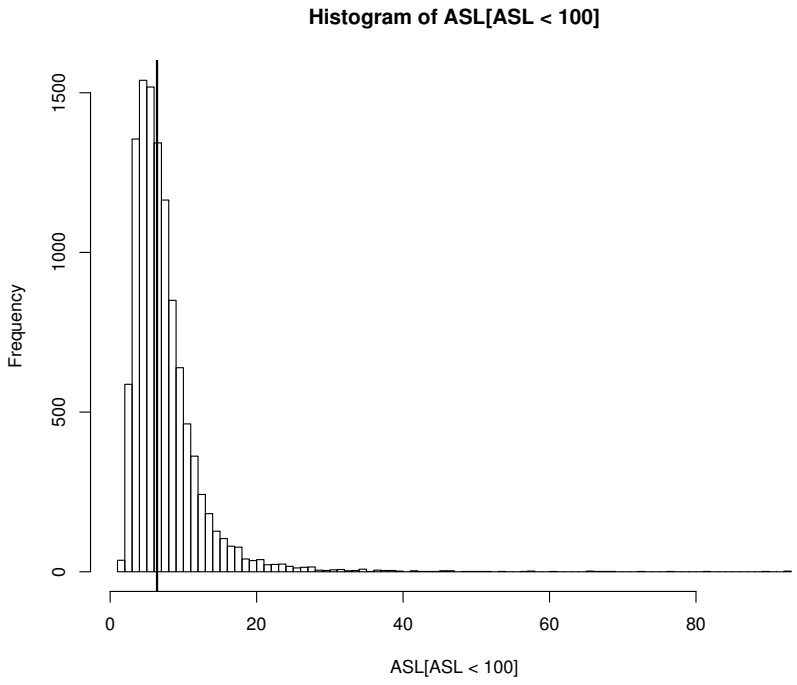

Figure 4.5: Average Shot Length (ASL) for a population of 11001 films.

Example 4.4: Movie Average Shot Lengths

Barry Sands has compiled a population of 11001 films and their average shot length (ASL, in seconds). The distribution of ASL is highly skewed to the right, with a population median of 6.4. A histogram of the ASL's, truncated at 100 are given in Figure 4.5, it has been truncated at 100, with 8 cases falling between 100 and 1000. The thick vertical line is the population median.

We consider samples of $n = 20$. For the Bin(20,0.5) distribution, we have the following cumulative probabilities.

$$P(Y \leq 4) = .0059 \qquad P(Y \leq 5) = .0207 \qquad P(Y \leq 6) = .0577$$

$$\Rightarrow \qquad B_{\alpha/2,n} = 5 \quad L_{\alpha/2} = 5 + 1 = 6 \quad U_{\alpha/2} = 20 - 5 = 15$$

Thus, once we order the the 20 sampled films, we would take the range encompassed by the 6^{th} through the 15^{th} films.

The following random sample (ordered) was obtained in R.

```
> (ASL.sample.order <- sort(ASL.sample))
 [1]  2.24  2.60  2.60  3.06  3.80  3.90  5.20  5.68  6.10  7.67  8.27  9.10
```

```
[13]  9.20 10.00 11.50 11.80 12.90 16.00 21.10 24.00
> cbind(ASL.sample.order[6],ASL.sample.order[15])
     [,1] [,2]
[1,]  3.9 11.5
```

For this sample, we obtain the 95% Confidence Interval: (3.9, 10), which does contain the population median (6.4). We now obtain 10000 random samples of size $n = 20$, and count the number that contain 6.4. Note that due to the "discreteness" of the distribution, $\alpha = 2(.0207) = .0414$, so we expect slightly more than 95% of the intervals to contain 6.4. Based on the 10000 random samples, 9645 (96.45%) contain the population mean.

Had we used the large-sample approximation here, which is questionable, with $n = 20$, we would have $n/2 = 10$, and $\sqrt{n} = 4.47$, and $L_{.025} \approx 10 - 4.47 = 5.53 = 5$ and $U_{.025} \approx 10 + 4.47 = 15$. We would still be selecting the 6^{th} and 15^{th} ordered values.

R Program and Output

```
## Commands

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

median(ASL)
sum(ASL > 100)
hist(ASL[ASL < 100], breaks=100)
abline(v=median(ASL),lwd=2)

pbinom(0:20,20,0.5)

set.seed(4321)
N <- length(ASL)
sample1 <- sample(1:N, 20, replace=FALSE)
ASL.sample <- ASL[sample1]
(ASL.sample.order <- sort(ASL.sample))          ## Sample values sorted
cbind(ASL.sample.order[6],ASL.sample.order[15]) ## 6th and 15th selected

set.seed(7654)
num.sim <- 10000
num.samp <- 100
med.ci <- matrix(rep(0, 2*num.sim),ncol=2)

for (i in 1:num.sim) {
   sample <- sample(1:N, 20, replace=FALSE)
    med.ci[i,1] <- sort(ASL[sample])[6]
    med.ci[i,2] <- sort(ASL[sample])[15]
}

med.pop <- median(ASL)
sum(med.ci[,1] <= med.pop & med.ci[,2] >= med.pop) / num.sim

## Output

> pbinom(0:20,20,0.5)
 [1] 9.536743e-07 2.002716e-05 2.012253e-04 1.288414e-03 5.908966e-03
 [6] 2.069473e-02 5.765915e-02 1.315880e-01 2.517223e-01 4.119015e-01
```

```
[11] 5.880985e-01 7.482777e-01 8.684120e-01 9.423409e-01 9.793053e-01
[16] 9.940910e-01 9.987116e-01 9.997988e-01 9.999800e-01 9.999990e-01
[21] 1.000000e+00

> (ASL.sample.order <- sort(ASL.sample))
 [1]  2.24  2.60  2.60  3.06  3.80  3.90  5.20  5.68  6.10  7.67  8.27  9.10
[13]  9.20 10.00 11.50 11.80 12.90 16.00 21.10 24.00
> cbind(ASL.sample.order[6],ASL.sample.order[15])
     [,1] [,2]
[1,]  3.9 11.5
>
> sum(med.ci[,1] <= med.pop & med.ci[,2] >= med.pop) / num.sim
[1] 0.9645
```

∇

For a hypothesis test of whether the population median is some particular value (as with the mean, this is rare except in paired data experiments), we can use the **sign test**. The test makes use of the count of the number of observations exceeding the null value of the median being a binomial random variable with n trials, and probability of success $\pi = 0.5$ under the null hypothesis $H_0 : M = M_0$. There can be 2-tailed or Upper/Lower tailed alternatives. In each case, let B_{obs+} be the count of the number of observations above M_0.

2-tailed tests: $H_0 : M = M_0 \quad H_A : M \neq 0 \quad T.S. : B_{obs} \quad R.R. : B_{obs} \leq B_{\alpha/2,n}$ or $B_{obs} \geq n - B_{\alpha/2,n}$

Upper tailed tests: $H_0 : M \leq M_0 \quad H_A : M > M_0 \quad T.S. : B_{obs} \quad R.R. : B_{obs} \geq n - B_{\alpha,n}$

Lower tailed tests: $H_0 : M \geq M_0 \quad H_A : M < M_0 \quad T.S. : B_{obs} \quad R.R. : B_{obs} \leq B_{\alpha,n}$

For large-samples, the approximate normality of the Binomial can be used, and under the null hypothesis, the number of observations exceeding M_0 is approximately normal with mean $n/2$ and standard deviation $\sqrt{n(0.5)(1-0.5)} = 0.5\sqrt{n}$. Then we can obtain a z-statistic for the tests.

$$T.S. : z_{obs} = \frac{B_{obs} - (n/2)}{0.5\sqrt{n}} \quad R.R.(2) : |z_{obs}| \geq z_{\alpha/2} \quad R.R.(U) : z_{obs} \geq z_{\alpha/2} \quad R.R.(L) : z_{obs} \leq z_{1-\alpha/2} = -z_{\alpha/2}$$

Example 4.5: Movie Average Shot Lengths

Suppose we wanted to test whether the population median average shot length (ASL),M, differs from $M_0 = 5$ seconds (for some reason). Based on the sample of $n = 20$ films obtained previously, we have the following ASL values.

```
> (ASL.sample.order <- sort(ASL.sample))
 [1]  2.24  2.60  2.60  3.06  3.80  3.90  5.20  5.68  6.10  7.67  8.27  9.10
[13]  9.20 10.00 11.50 11.80 12.90 16.00 21.10 24.00
```

The test statistic is $B_{obs} = 14$. Depending on whether we want a 2-tailed or 1-tailed test we have that $P(Y \leq 5) = .0207$ and $P(Y \leq 6) = .0577$ for $Y \sim B(n = 20, \pi = .5)$. Thus, for a 2-tailed test, we reject the null $H_0 : M = M_0$ for a 2-tailed test (with $\alpha = 0.05$), if $B_{obs} \leq 5$ or if $B_{obs} \geq 20 - 5 = 15$. Because the probability that $Y \leq 6$ exceeds $\alpha = 0.05$, the Upper tail rejection region would be $R.R. : B_{obs} \geq 15$ and the Lower tail rejection region would be $R.R. : B_{obs} \leq 5$. Note that the 2-sided P-value is $P = 2P(Y \geq 14|Y \sim Bin(20, 0.5)) = 2P(Y \leq 6) = 2(.0577) = .1154$.

The large-sample z-statistic would be computed as follows.

$$z_{obs} = \frac{14 - (20/2)}{0.5\sqrt{20}} = \frac{4}{2.236} = 1.789 \qquad \text{2-tailed } P\text{-value: } P = 2P(Z \geq 1.789) = 2(.0378) = .0756$$

The reason for the discrepancy between the P-values is the discreteness of the binomial and the continuity of the normal approximation. Some authors suggest the following continuity correction. The subtracting of the 0.5 is to get all the area over 14 for binomial, since 14 is above its expected value. This results in virtually the exact same P-value. As n gets large, the correction makes little difference.

$$z_{obs} = \frac{14 - (20/2) - 0.5}{0.5\sqrt{20}} = \frac{3.5}{2.236} = 1.565 \qquad \text{2-tailed } P\text{-value: } P = 2P(Z \geq 1.565) = 2(.0588) = .1176$$

4.4 The Bootstrap

In many applications, individual measurements are not normally distributed and the sample size is not large enough to justify the use of the Central Limit Theorem. Further, in many practical settings, the sampling distribution of an estimator (such as a median or coefficient of variation). The bootstrap makes use of the sample that is obtained (and is assumed to be representative of the population of measurements) to approximate the sampling distribution of the estimator of interest. The classic reference is Efron and Tibshirani (1993) [11], and for an introduction to Mathematical Statistics based on resampling methods, see Chihara and Hesterberg (2011) [7].

The process involves resampling from the sample data, with replacement many times, and computing the estimate for each resample, and saving the values. The samples are each of size n. Note that when estimating the sampling distribution of the sample mean, the mean of the resampled means will be very close to the sample mean of the original sample. That implies that the bootstrap will not directly estimate the mean of the sampling distribution (which is the population mean). The spread, bias, and skewness of the bootstrap distribution do reflect those of the target sampling distribution, where bias refers to the difference between the mean of the bootstrap distribution and the population mean.

4.4.1 Bootstrap Inferences Concerning the Population Mean

When trying to estimate a population mean (particularly with nonnormal data with a small sample size), a bootstrap prediction interval for the population mean μ can be obtained from the central $(1-\alpha)100\%$ values of the bootstrap sample estimates. This is a very simple approach, as all that is needed to be computed and saved are the sample means from each of the resamples (see e.g. Chiara and Hesterberg (2011), [**?**] Section 5.3. Once the means are obtained, the $\alpha/2$ and $1-\alpha/2$ quantiles are identified. Note that this interval will not typically be symmetric, unless the sample data are highly symmetric.

Example 4.6: Movie Average Shot Lengths

Suppose we wish to estimate the population mean of movie average shot lengths (ASL). The distribution is highly skewed, refer back to Figure 4.5. We first take a sample of $n = 25$ films, then draw $B = 10000$ random resamples with replacement from the 25 sampled films, and compute the sample mean for each resample, labeled $\overline{y}_i^*$ for the i^{th} resample. Finally we obtain the 2.5%-ile and 97.5%-ile from the resample means, for an interval that we can be approximately 95% confident will contain μ.

R Commands and Output

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

(N <- length(ASL))
(mu <- mean(ASL))
(sigma <- sd(ASL))

## Obtain the original random sample of n=25
set.seed(34567)
samp.size <- 25
sample1 <- sample(1:N,samp.size,replace=F)

ASL.sample1 <- ASL[sample1]
ASL.sample1
mean(ASL.sample1)

qqnorm(ASL.sample1); qqline(ASL.sample1)
shapiro.test(ASL.sample1)

### Method 1 - Chihara/Hesterberg Section 5.3, pp. 113-114
set.seed(24680)
num.boot.inner <- 10000
ASL.mean <- rep(0,num.boot.inner)
for (i2 in 1:num.boot.inner) {
    x <- sample(ASL.sample1, samp.size, replace=T)
    ASL.mean[i2] <- mean(x)
}

quantile(ASL.mean,c(.025,.975))
mean(ASL.mean)
sd(ASL.mean)
mean(ASL.mean) - qt(.975,samp.size-1)*sd(ASL.mean)
mean(ASL.mean) + qt(.975,samp.size-1)*sd(ASL.mean)
sum(ASL.mean < mean(ASL.mean) - qt(.975,samp.size-1)*sd(ASL.mean)) /
            num.boot.inner
```

```
sum(ASL.mean > mean(ASL.mean) + qt(.975,samp.size-1)*sd(ASL.mean)) /
            num.boot.inner

## Output
> (mu <- mean(ASL))
[1] 7.739382
> ASL.sample1
 [1]  4.40 14.98  7.80  9.50  9.50  6.70  7.50  9.20  3.70  8.04  4.47  9.40
[13]  8.40  8.88  5.50 16.30  6.70  3.65  4.27 11.60  9.30  3.40  2.90 12.00
[25] 16.60
> mean(ASL.sample1)
[1] 8.1876
> quantile(ASL.mean,c(.025,.975))
  2.5%  97.5%
6.7444 9.7113
> mean(ASL.mean)
[1] 8.189939
> sd(ASL.mean)
[1] 0.7620199
> mean(ASL.mean) - qt(.975,samp.size-1)*sd(ASL.mean)
[1] 6.617208
> mean(ASL.mean) + qt(.975,samp.size-1)*sd(ASL.mean)
[1] 9.762671
> sum(ASL.mean < mean(ASL.mean) - qt(.975,samp.size-1)*sd(ASL.mean)) /
+             num.boot.inner
[1] 0.0148
> sum(ASL.mean > mean(ASL.mean) + qt(.975,samp.size-1)*sd(ASL.mean)) /
+             num.boot.inner
[1] 0.0216
```

The sample mean for the original sample is $\overline{y} = 8.188$ which exceeds the population mean $\mu = 7.740$, although different samples could be below, close to, or above μ due to sampling error. The mean of the $B = 10000$ resample means is $\overline{\overline{y}^*} = 8.190$, which is very close to $\overline{y}$, but not to μ, as would be expected due to the sampling process of the bootstrap. The approximate 95% prediction interval for μ is (6.744,9.711) which does include $\mu = 7.740$. The standard deviation of the resample means is referred to as the bootstrap standard error. Note that the prediction interval is not of the form $\overline{\overline{y}^*} \pm t_{.025,25-1} s_{\overline{y}^*}$, which is of the following form, where $t_{.025,24} = 2.064$.

$$8.190 \pm 2.064(0.762) \quad \equiv \quad 8.190 \pm 1.573 \quad \equiv \quad (6.617, 9.763)$$

Of the 10000 sample means, 1.48% of the sample means fall below the lower bound 6.617, and 2.16% fall above the upper bound 9.763. The "t-type" interval goes outside both of the lower and upper bounds of the bootstrap interval. The lower bound is 1.898 bootstrap standard errors below the mean of the resample means, and the upper bound is 1.996 standard errors above it. In some cases the asymmetry will be much larger.

∇

This approach of obtaining an approximate Confidence Interval for a parameter works well for many types of estimators/parameters. It is particularly useful when the bootstrap sample estimators have an approximately continuous distribution. When the distribution of bootstrap sample estimators have a discrete sampling distribution, the method does not work well. Consider estimating the population median in the average shot length example. Once we have our sample of $n = 25$ films, the median is the "middle" ASL of the 25 (13th) ordered film. When we take bootstrap samples, the median will always be one of the 25

ASL's in the original sample. Thus, there are only 25 possible values the sample median for each resample can take on.

A second approach that is specific to estimating a population mean makes use of a t-type statistic computed for each resample. This is referred to as **Bootstrap t Confidence Intervals**, (see e.g. Chihara and Hesterberg (2011), [7] Section 7.5). In this method, once the original sample of size n is taken, obtain the sample mean $\overline{y}$ and standard deviation s. Then for each of B resamples, compute the mean $\overline{y}_i^*$ and standard deviation s_i^*, where i represents the i^{th} resample. Then compute a t-type statistic for each resample, making use of the original sample mean as follows.

$$t_i^* = \frac{\overline{y}_i^* - \overline{y}}{s_i^*/\sqrt{n}} = \sqrt{n}\left(\frac{\overline{y}_i^* - \overline{y}}{s_i^*}\right) \quad i = 1, \ldots, B$$

Once the B values of t_i^* are computed, obtain the $\alpha/2$ quantile and the $(1-\alpha/2)$ quantiles, say (Q_L^*, Q_U^*). Note that Q_L^* will be negative and Q_U^* will be positive, and not necessarily of the same magnitude. The $(1-\alpha)100\%$ Confidence Interval for μ will be of the following form.

$$\text{Lower Bound: } \overline{x} - Q_U^* \frac{s}{\sqrt{n}} \qquad \text{Upper Bound: } \overline{x} - Q_L^* \frac{s}{\sqrt{n}}$$

Example 4.7: Movie Average Shot Lengths

We apply this method to the same sample and resamples used previously.

R Commands and Output

```
## Commands

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

(N <- length(ASL))
(mu <- mean(ASL))
(sigma <- sd(ASL))

## Obtain the original random sample of n=25
set.seed(34567)
samp.size <- 25
sample1 <- sample(1:N,samp.size,replace=F)
ASL.sample1 <- ASL[sample1]
mean(ASL.sample1)
sd(ASL.sample1)

### Method 2 - Chihara/Hesterberg Section 7.5, pp. 195-198
# ASL.t computes and saves t*
set.seed(24680)
ybar <- mean(ASL.sample1)
  num.boot.inner <- 10000
  ASL.t <- rep(0,num.boot.inner)
```

```
  ASL.mean <- rep(0,num.boot.inner)
  for (i2 in 1:num.boot.inner) {
    x <- sample(ASL.sample1, samp.size, replace=T)
    ASL.t[i2] <- (mean(x) - ybar) /
                 (sd(x) / sqrt(samp.size))
    ASL.mean[i2] <- mean(x)
  }
Q_L <- quantile(ASL.t,0.025)
Q_U <- quantile(ASL.t,0.975)
mu_L <- ybar - Q_U*sd(ASL.sample1)/sqrt(samp.size)
mu_U <- ybar - Q_L*sd(ASL.sample1)/sqrt(samp.size)
cbind(Q_L, Q_U, mu_L, mu_U)
mean(ASL.mean) - ybar
sd(ASL.mean)


hist(ASL.t,breaks=40,freq=F,
   main=expression(paste("Histogram of ",t^"*"," and ",t[24]," Density")))
t.seq <- seq(-4,4,.01)
lines(t.seq,dt(t.seq,samp.size-1))
abline(v=c(Q_L,Q_U),lwd=2)

## Output

> (mu <- mean(ASL))
[1] 7.739382
> mean(ASL.sample1)
[1] 8.1876
> sd(ASL.sample1)
[1] 3.894509
> cbind(Q_L, Q_U, mu_L, mu_U)
          Q_L      Q_U     mu_L     mu_U
     -2.335245 1.849595 6.746947 10.00653
> mean(ASL.mean) - ybar
[1] 0.00233944
> sd(ASL.mean)
[1] 0.7620199
```

A plot of the t^* values and the t_{24} density is given in Figure 4.6. The distribution of the t^* values is skewed left, with Q_L^* and Q_U^* being -2.335 and 1.850, respectively for $\alpha = 0.05$. The original sample mean and standard deviation are 8.188 and 3.895 respectively leading to the following 95% Confidence Interval for μ.

$$\left(8.188 - 1.855\frac{3.895}{\sqrt{25}}, 8.188 - (-2.335\frac{3.895}{\sqrt{25}}\right) \quad \equiv \quad (8.188 - 1.445, 8.188 + 1.819) \quad \equiv \quad (6.743, 10.007)$$

Note that the interval is not symmetric about $\overline{y}$, it adds a larger term to the upper end than the term it subtracts from the lower end. This reflects the fact that the data are right-skewed. The bootstrap estimate of the **bias** is the difference from the average of the resample means and the overall sample mean: $\overline{\overline{y}^*} - \overline{y} = 0.0023$. This bias is very small relative to the standard error of the bootstrap estimator: .00234/.76202=.00307. The 95% Confidence Interval using just the original sample mean, standard deviation, and the t-distribution is given for comparison.

$$8.188 \pm 2.064\frac{3.895}{\sqrt{25}} \quad \equiv \quad 8.188 \pm 1.608 \quad \equiv \quad (6.580, 9.796)$$

Finally, 1000 random samples were taken from the ASL data. The two Bootstrap methods were performed on 1000 resamples from each (original) random sample and their 95% Confidence Intervals were obtained, as was the 1000 t-based intervals from the (original) samples. The first bootstrap method (middle 95% of the resample means) contained $\mu = 7.739$ in 849 of the 1000 original samples (84.9% coverage). The second bootstrap method (based on constructed t-statistics around the sample mean) contained μ in 903 of the 1000 original samples (90.3% coverage). The normal based t-interval contained μ in 869 of the 1000 original samples (86.9%). All three performed below the nominal 95% level. This is due to the very large amount of skew in the data (largest ASL is 1000, while the population mean is less than 8), as well as the relativel small sample size ($n = 25$).

R Program and Output

```
## Commands

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

(N <- length(ASL))
(mu <- mean(ASL))
(sigma <- sd(ASL))

## Initialize mean/sd and CI holders - 1000 outer (original) samples
set.seed(13579)
num.boot.outer <- 1000
ASL.mean.sd <- matrix(rep(0,2*num.boot.outer),ncol=2)
ASL.boot1 <- matrix(rep(0,2*num.boot.outer),ncol=2)
ASL.boot2 <- matrix(rep(0,2*num.boot.outer),ncol=2)
ASL.tnorm <- matrix(rep(0,2*num.boot.outer),ncol=2)
samp.size <- 25
sqrt.n <- sqrt(samp.size)
t.24 <- qt(c(.025,.975),samp.size-1)

### Begin outer loop
for (i1 in 1:num.boot.outer) {
  sample1 <- sample(1:N,samp.size,replace=F)
  ASL.sample1 <- ASL[sample1]               ### Original Samples
  ASL.mean.sd[i1,1] <- mean(ASL.sample1) ### Save mean in column 1
  ASL.mean.sd[i1,2] <- sd(ASL.sample1)   ### Save sd in column 2

### Begin inner (bootstrap) loop
  num.boot.inner <- 1000
  ASL.mean <- rep(0,num.boot.inner)
  ASL.t <- rep(0,num.boot.inner)
  for (i2 in 1:num.boot.inner) {
    x <- sample(ASL.sample1, samp.size, replace=T)
    ASL.mean[i2] <- mean(x)
    ASL.t[i2] <- (mean(x) - ASL.mean.sd[i1,1]) /
                 (sd(x) /sqrt.n)

  }   ### Close inner loop
ASL.boot1[i1,] <- quantile(ASL.mean,c(.025,.975))
ASL.boot2[i1,] <-  ASL.mean.sd[i1,1] -
    quantile(ASL.t,c(.975,.025)) * ASL.mean.sd[i1,2]/sqrt.n
ASL.tnorm[i1,] <- ASL.mean.sd[i1,1] +  t.24 *
        ASL.mean.sd[i1,2]/sqrt.n
}     ### Close outer loop

## Obtain coverage probabilities
```

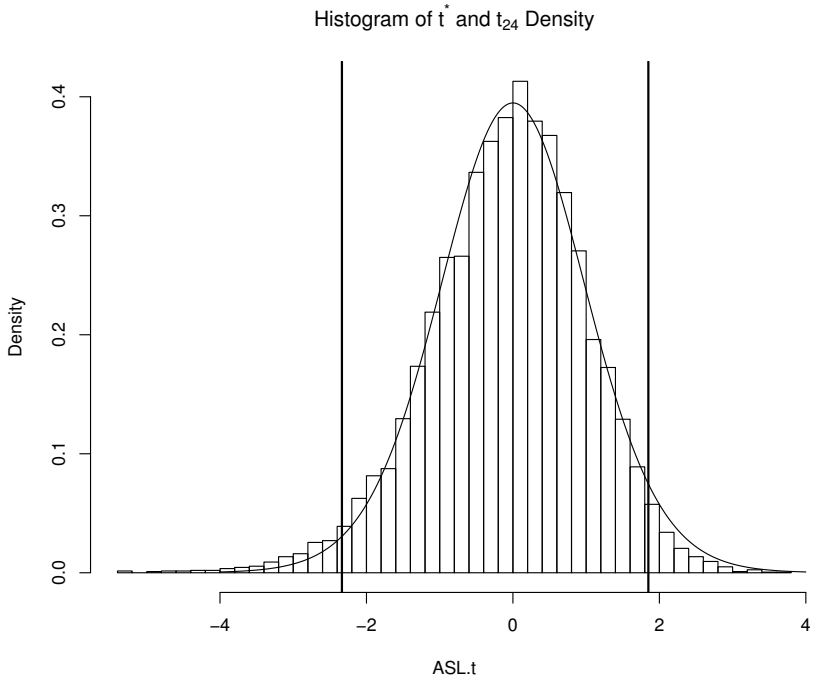

Figure 4.6: Histogram of t^* values and t_{24} Density - ASL Data

```
sum(ASL.boot1[,1] <= mu & ASL.boot1[,2] >= mu)/num.boot.outer
sum(ASL.boot2[,1] <= mu & ASL.boot2[,2] >= mu)/num.boot.outer
sum(ASL.tnorm[,1] <= mu & ASL.tnorm[,2] >= mu)/num.boot.outer

### Output

> sum(ASL.boot1[,1] <= mu & ASL.boot1[,2] >= mu)/num.boot.outer
[1] 0.849
> sum(ASL.boot2[,1] <= mu & ASL.boot2[,2] >= mu)/num.boot.outer
[1] 0.903
> sum(ASL.tnorm[,1] <= mu & ASL.tnorm[,2] >= mu)/num.boot.outer
[1] 0.869
```

∇

Chapter 5

Comparing Two Population's Central Values

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.

5.1 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_{1i} is the i^{th} unit (replicate) from sample 1, and Y_{2i} is the i^{th} unit (replicate) from sample 2. In the case of independent samples, these two random variables are independent.

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

$$E\{\overline{Y}_1-\overline{Y}_2\}=E\{\overline{Y}_1\}-E\{\overline{Y}_2\}=\mu_1-\mu_2 \qquad V\{\overline{Y}_1-\overline{Y}_2\}=V\{\overline{Y}_1\}+V\{\overline{Y}_2\}-2\text{COV}\{\overline{Y}_1,\overline{Y}_2\}=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}+0$$

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 a sample is 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 Z.

$$\frac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\dot{\sim}N(0,1) \quad \Rightarrow \quad P\left((\overline{Y}_1)+z_{1-\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\leq\mu_1-\mu_2\leq(\overline{Y}_1)+z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)\approx 1-\alpha$$

Example 5.1: 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. Note the NHL differs slightly from previous examples as the heights have not been "tweaked" to make them less discrete here.

$$N_1=717 \quad \mu_1=26.500 \quad \sigma_1=1.455 \qquad N_2=526 \quad \mu_2=23.019 \quad \sigma_2=1.713$$

A plot of the two population histograms, along with normal densities is given in Figure 5.1. 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 0.503 (calculation given below) is shown in Figure 5.2. The mean of the 100000 mean differences $\overline{y}_1-\overline{y}_2$ is 3.482 with standard deviation (standard error) 0.494. 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).

$$(\overline{y}_1-\overline{y}_2)\pm 1.96\sqrt{\frac{s_1^2}{20}+\frac{s_2^2}{20}} \qquad \mu_1-\mu_2=26.500-23.019=3.481 \quad \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}=\sqrt{\frac{1.455^2}{20}+\frac{1.713^2}{20}}=0.503$$

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 94.606% of the samples, very close to the nominal 95% coverage rate. If we replace $z_{.025}=1.96$ with the more appropriate $t_{.025,n_1+n_2-2}=$

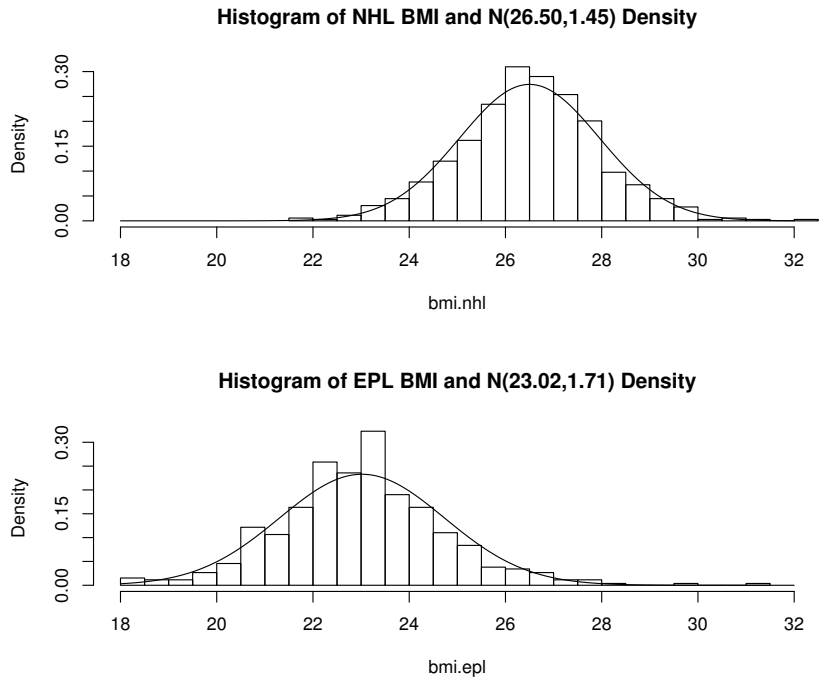

Figure 5.1: Distributions of NHL and EPL players Body Mass Index

$t_{.025,38} = 2.0244$, the coverage rate increases to 95.349%. 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 0.2524, while its theoretical value is $\sigma_1^2/n_1 + \sigma_2^2/n_2 = 0.2525$. Note that the variance of the estimated difference is unbiased, not so for the standard error.

R Commands and Output

```
## Commands

bmi.sim <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_nba_ebl_bmi.csv",
   header=TRUE)
attach(bmi.sim); names(bmi.sim)


N.nhl <- 717     #  # of NHL players
N.epl <- 526     #  # of EPL players
bmi.nhl <- NHL_BMI[1:N.nhl]
bmi.epl <- EPL_BMI[1:N.epl]
(mu.nhl <- mean(bmi.nhl)); (sigma.nhl <- sd(bmi.nhl))
(mu.epl <- mean(bmi.epl)); (sigma.epl <- sd(bmi.epl))

par(mfrow=c(2,1))
hist(bmi.nhl, breaks=30, xlim=c(18,32), freq=F,
  main="Histogram of NHL BMI and N(26.50,1.45) Density")
```

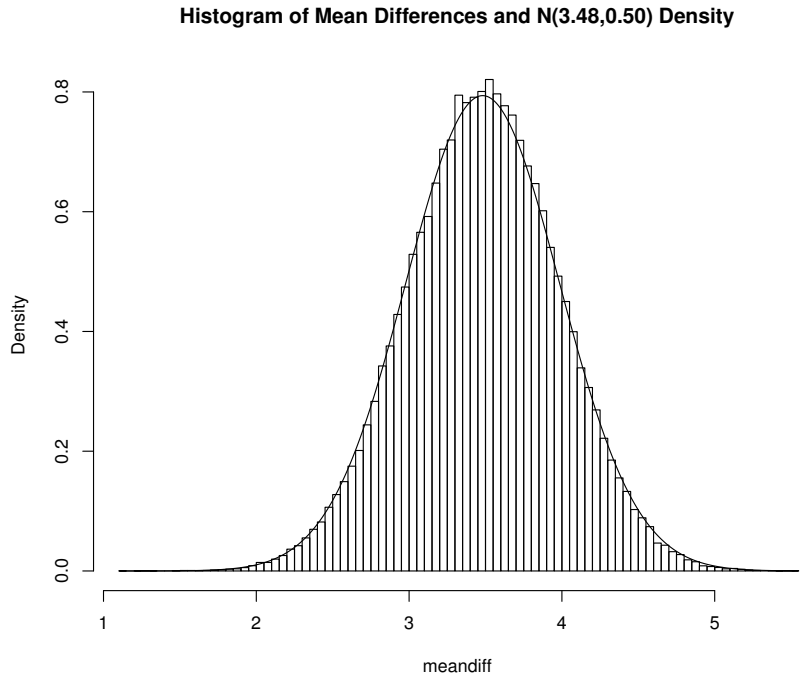

Figure 5.2: 100000 sample mean differences ($n_1 = n_2 = 20$) for NHL and EPL BMI values and Normal Density

```
bmi.x <- seq(18,32,.01)
lines(bmi.x, dnorm(bmi.x, mu.nhl, sigma.nhl))

hist(bmi.epl, breaks=30, xlim=c(18,32), freq=F,
  main="Histogram of EPL BMI and N(23.02,1.71) Density")
lines(bmi.x, dnorm(bmi.x, mu.epl, sigma.epl))

num.sim <- 100000
n.nhl <- 20
n.epl <- 20
(mu.meandiff <- mu.nhl - mu.epl)
(sigma.meandiff <- sqrt(sigma.nhl^2/n.nhl + sigma.epl^2/n.epl))
set.seed(6677)
ybar.s.nhl <- matrix(rep(0,2*num.sim),ncol=2)
ybar.s.epl <- matrix(rep(0,2*num.sim),ncol=2)

for (i in 1:num.sim) {
y1 <- sample(bmi.nhl,n.nhl,replace=F)
y2 <- sample(bmi.epl,n.nhl,replace=F)

ybar.s.nhl[i,1] <- mean(y1)
ybar.s.nhl[i,2] <- sd(y1)
ybar.s.epl[i,1] <- mean(y2)
ybar.s.epl[i,2] <- sd(y2)
}

meandiff <- ybar.s.nhl[,1] - ybar.s.epl[,1]
mean(meandiff)
sd(meandiff)
par(mfrow=c(1,1))
hist(meandiff, breaks=100, xlim=c(min(meandiff)-0.01, max(meandiff)+0.01),
  freq=F, main="Histogram of Mean Differences and N(3.48,0.50) Density")
diff.x <- seq(min(meandiff)-0.01, max(meandiff)+0.01,length.out=1000)
lines(diff.x, dnorm(diff.x, mu.meandiff, sigma.meandiff))

se.meandiff <- sqrt(ybar.s.nhl[,2]^2/n.nhl + ybar.s.epl[,2]^2/n.epl)
mean(se.meandiff^2)
sigma.meandiff^2
diff.lo.z <- meandiff + qnorm(.025,0,1) * se.meandiff
diff.hi.z <- meandiff + qnorm(.975,0,1) * se.meandiff
sum(diff.lo.z <= mu.meandiff & diff.hi.z >= mu.meandiff) / num.sim

diff.lo.t <- meandiff + qt(.025,n.nhl+n.epl-2) * se.meandiff
diff.hi.t <- meandiff + qt(.975,n.nhl+n.epl-2) * se.meandiff
sum(diff.lo.t <= mu.meandiff & diff.hi.t >= mu.meandiff) / num.sim

### Output

> (mu.nhl <- mean(bmi.nhl)); (sigma.nhl <- sd(bmi.nhl))
[1] 26.50015
[1] 1.454726
> (mu.epl <- mean(bmi.epl)); (sigma.epl <- sd(bmi.epl))
[1] 23.01879
[1] 1.713098
> (mu.meandiff <- mu.nhl - mu.epl)
[1] 3.481361
> (sigma.meandiff <- sqrt(sigma.nhl^2/n.nhl + sigma.epl^2/n.epl))
[1] 0.5025401
> mean(meandiff)
[1] 3.482383
> sd(meandiff)
[1] 0.4944524
> mean(se.meandiff^2)
[1] 0.2524164
```

```
> sigma.meandiff^2
[1] 0.2525466
> sum(diff.lo.z <= mu.meandiff & diff.hi.z >= mu.meandiff) / num.sim
[1] 0.94606
> sum(diff.lo.t <= mu.meandiff & diff.hi.t >= mu.meandiff) / num.sim
[1] 0.95349
```

∇

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).

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

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

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

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

Example 5.2: 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), [24]). 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.

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

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

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

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

The interval is very far away from 0, 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 $\alpha = 0.05$.

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

∇

5.2 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 a independent samples (parallel groups) or paired samples (crossover) design? The appropriate test for each situation is given in Table 5.1. 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	
	Parallel Groups	Crossover
Normally Distributed Data	2–Sample t–test	Paired t–test
Non–Normally Distributed Data	Wilcoxon Rank Sum test (Mann–Whitney U–Test)	Wilcoxon Signed–Rank Test

Table 5.1: Statistical Tests for small–sample 2 group situations

5.2.1 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 identical to the large–sample test, except the critical values for the rejection regions are based on the t–distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. We will assume the two population variances are equal in the 2–sample t–test. If they are not, simple adjustments can be made to obtain 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: $2P(t_{n_1+n_2-2} > |t_{obs}|)$ or $P(t_{n_1+n_2-2} > t_{obs})$ or $P(t_{n_1+n_2-2} < t_{obs})$ (again, depending on which alternative you are using).

Example 5.3: Comparison of Two Instructional Methods A study was conducted (Rusanganwa (2013) [22]) 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, we compute s_p^2, the pooled variance:

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

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

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_A : \mu_1 - \mu_2 \neq 0$
- T.S.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(11.10-8.35)}{\sqrt{8.42\left(\frac{1}{13}+\frac{1}{19}\right)}} = \frac{2.75}{1.04} = 2.633$
- 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: $2P(T \geq |t_{obs}|) = 2P(t_{30} \geq 2.633) = 0.0132$

Based on this test, we reject H_0, 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 generate 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
n1 <- 13; ybar1 <- 11.10; sd1 <- 3.47   # Give sample sizes, means and sd's
n2 <- 19; ybar2 <-  8.35; sd2 <- 2.45

set.seed(1234)
z1 <- rnorm(n1, 0, 1)              # Generate sample n N(0,1) rv's
z2 <- rnorm(n2, 0, 1)
z1a <- (z1 - mean(z1)) / sd(z1)    # standardize to mean 0, sd 1
z2a <- (z2 - mean(z2)) / sd(z2)

samp1 <- ybar1 + sd1*z1a           # Generate sample to have mean ybar and sd
samp2 <- ybar2 + sd2*z2a

mean(samp1); sd(samp1)             # Check
mean(samp2); sd(samp2)

t.test(samp1, samp2, var.equal=T)  # t-test with 2 sets of variables

sample.y <-c(samp1, samp2)         # Combine samples (t=transpose)
trt.y <- c(rep(1,n1),rep(2,n2))    # treatment id

cbind(sample.y, trt.y)             # print data

t.test(sample.y ~ trt.y, var.equal=T)   # t-test with single y-var and trt id

## Output

> t.test(samp1, samp2, var.equal=T)
```

```
        Two Sample t-test

data:  samp1 and samp2
t = 2.6333, df = 30, p-value = 0.01324
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6172446 4.8827554
sample estimates:
mean of x mean of y
    11.10      8.35

> cbind(sample.y, trt.y)
       sample.y trt.y
 [1,]  8.198004     1
         ...
[13,]  9.889514     1
[14,]  8.827296     2
         ...
[32,]  7.389956     2
>
> t.test(sample.y ~ trt.y, var.equal=T)   # t-test with single y-var and trt id

        Two Sample t-test

data:  sample.y by trt.y
t = 2.6333, df = 30, p-value = 0.01324
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6172446 4.8827554
sample estimates:
mean in group 1 mean in group 2
          11.10            8.35
```

∇

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.

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

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.

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

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

Example 5.4: Abdominal Quilting to Reduce Blood Loss in Breast Reconstruction Surgery

A study considered the effect of abdominal suture quilting on blood loss during breast reconstruction surgery. A group of $n_1 = 27$ subjects (controls) received the standard DIEP procedure, while a group of $n_2 = 26$ subjects (study) received the DIEP procedure along with the suture quilting. The response measured was the amount of blood loss 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 5.3.

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

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

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

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

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.

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

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

R Commands and Output

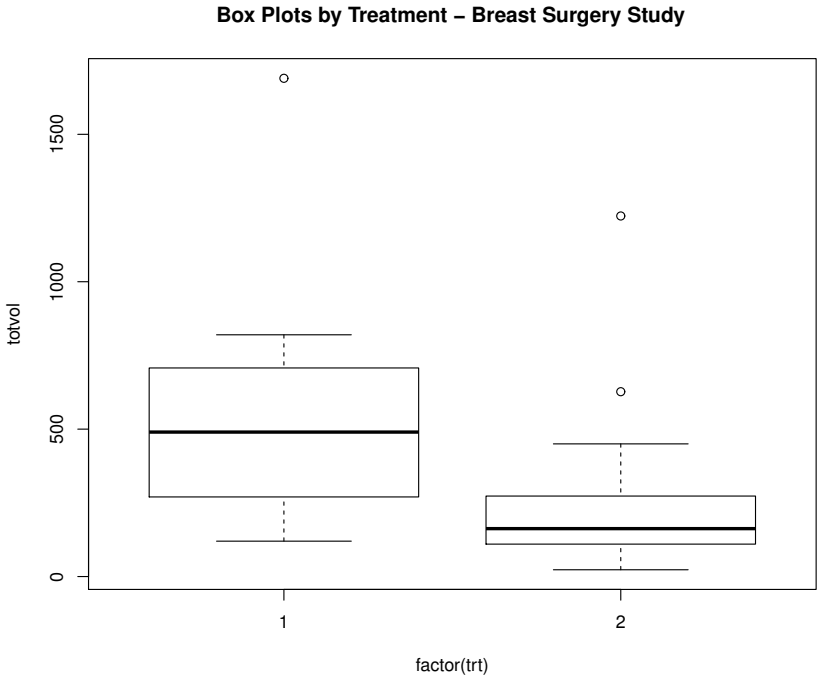

Figure 5.3: Caption for breast_diep

```
## Commands

quilt <- read.csv("F:\\orangedrive\\sta6166\\breast_diep.csv")
attach(quilt); names(quilt)

plot(totvol ~ factor(trt), main="Box Plots by Treatment - Breast Surgery Study")
t.test(totvol ~ trt, 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
```

∇

Wilcoxon Rank-Sum Test for Non-Normally Distributed Data

The idea behind this test is as follows. We have samples of n_1 measurements from population 1 and n_2 measurements from population 2 (Wilcoxon,1945). We rank the $n_1 + n_2$ measurements from 1 (smallest) to $n_1 + n_2$ (largest), adjusting for ties by averaging the ranks the measurements would have received if they were different. We then compute T_1, the rank sum for measurements from population 1, and T_2, the rank sum for measurements from population 2. This test is mathematically equivalent to the Mann–Whitney U–test. To test for differences between the two population distributions, we use the following procedure:

1. H_0 : The two population medians are equal ($M_1 = M_2$)
2. H_A : The medians are not equal ($M_1 \neq M_2$) ($\mu_1 \neq \mu_2$)
3. T.S.: $T = \min(T_1, T_2)$
4. R.R.: $T \leq T_0$, where values of T_0 given in tables in many statistics texts and on the web for various levels of α and sample sizes.

For one-sided tests to show that the distribution of population 1 is shifted to the right of population 2 ($M_1 > M_2$), we use the following procedure (simply label the distribution with the suspected higher mean as population 1):

1. H_0 : The median for population 1 is less than or equal the median for population 2 ($M_1 \leq M_2$)
2. H_A : The median for population 1 is larger than the median for population 2 ($M_1 > M_2$)
3. T.S.: $T = T_2$
4. R.R.: $T \leq T_0$, where values of T_0 are given in tables in many statistics texts and on the web for various levels of α and various sample sizes.

Example 5.5: Apple Procyanidin B-2 for Hair Growth

A study was conducted to determine whether procyanidin B-2 from apples is effective in hair growth (Kamimura, Takahishi, and Watanabe (2000), [18]). Based on a small trial, with $n_1 = 19$ treatment subjects and $n_2 = 10$ control subjects, Table 5.2 gives the 6 month change in total hairs, along as their ranks from smallest (most negative) to largest. Note that the ranks sum to 1+2+...+29=29(29+1)/2=435. Normal probability plots are given in Figure 5.4, there is evidence of outlying cases in each group.

For a 2-tailed test, based on sample sizes of n_19 and $n_2 = 10$, we will reject H_0 for $T = min(T_c, T_t) \leq 107$. Since $T = min(86, 349) = 86$, we reject H_0, and cannot conclude that the median hair growth differs between the treatment and control conditions. The table used is available on the class web site and necessitates that $n_1 \geq n_2$.

R Commands and Output

```
## Commands
```

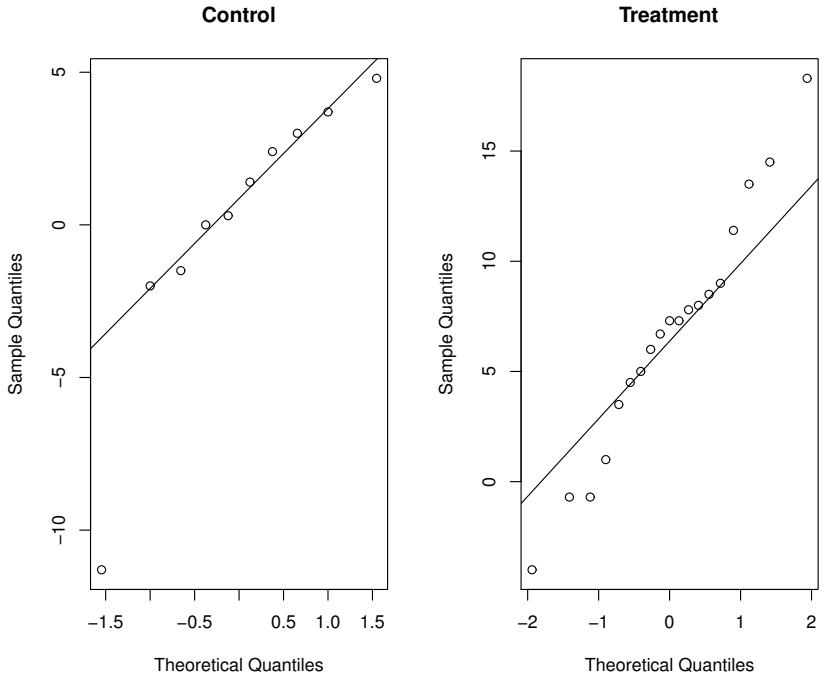

Figure 5.4: Caption for apple_hair1

Trt	TotalDif	Rank	Trt	TotalDif	Rank
1	0.3	8	2	3.5	13
1	1.4	10	2	5	17
1	3	12	2	7.3	20.5
1	3.7	14	2	18.3	29
1	-1.5	4	2	14.5	28
1	-2	3	2	6.7	19
1	0	7	2	9	25
1	4.8	16	2	-0.7	5.5
1	2.4	11	2	7.8	22
1	-11.3	1	2	-4	2
			2	6	18
			2	4.5	15
			2	8	23
			2	11.4	26
			2	1	9
			2	7.3	20.5
			2	8.5	24
			2	-0.7	5.5
			2	13.5	27
Total		$T_c = 86$			$T_t = 349$
Average		$T_c/n_c = 8.60$			$T_t/n_t = 18.37$

Table 5.2: Total Growth measurements (and ranks) for Procyanidin B-2 from Apple Hair Growth Experiment

```
apphair <- read.table("http://www.stat.ufl.edu/~winner/data/apple_hair.dat",
   header=F, col.names=c("hair.trt","total0","total6","totaldiff",
   "term0", "term6", "termdiff"))
attach(apphair)

hair.trt.f <- factor(hair.trt, levels=1:2, labels=c("placebo", "PC2"))

plot(totaldiff ~ hair.trt.f)

par(mfrow=c(1,2))
qqnorm(totaldiff[hair.trt==1],main="Control")
qqline(totaldiff[hair.trt==1])
qqnorm(totaldiff[hair.trt==2],main="Treatment")
qqline(totaldiff[hair.trt==2])

wilcox.test(totaldiff ~ hair.trt)

## Output

> wilcox.test(totaldiff ~ hair.trt)

        Wilcoxon rank sum test with continuity correction

data:  totaldiff by hair.trt
W = 31, p-value = 0.003565
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(0.3, 1.4, 3, 3.7, -1.5, -2, 0, 4.8,  :
  cannot compute exact p-value with ties
```

Note that the W represents the difference between the Rank Sum for each group and its minimum (low average rank group) or maximum (high average rank group) possible value. In this example, if all of the Controls fell below all of the Treatments, the rank sums would be as follow.

$$\text{Control: } 1{+}2{+}\ldots{+}10 = \frac{10(11)}{2} = 55 \quad \text{Treatment: } 11{+}\ldots{+}29 = 435{-}55 = 380 \quad W = 86{-}55 = 380{-}349 = 31$$

$$\nabla$$

For large samples, it's difficult to find tables that contain the critical values (this example pushed the limits, in fact). The rank sums are approximately normal in large samples, so a normal approximation can be used. Let T be the rank sum for group 1 (the test is symmetric, so the statistic will have the same absolute value, no matter which group gets labeled as 1). The expected value and standard deviation of T under the null hypothesis $M_1 = M_2$ and the test statistic are given here.

$$N = n_1 + n_2 \quad T = T_1 \quad \mu_T = \frac{n_1(N+1)}{2} \quad \sigma_T = \sqrt{\frac{n_1 n_2 (N+1}{12}} \quad z_{obs} = \frac{T - \mu_T}{\sigma_T}$$

The critical values for the Rejection Region are based on whether the test is 2-tailed or upper tailed and α, as in other large-sample z-tests.

$H_A : M_1 \neq M_2 \quad R.R.|z_{obs}| \geq z_{\alpha/2} \quad P = 2P(Z \geq |z_{obs}|) \qquad H_A : M_1 > M_2 \quad R.R.z_{obs} \geq z_{\alpha/2} \quad P = P(Z \geq z_{obs})$

Example 5.5: Apple Procyanidin B-2 for Hair Growth

To use the large-sample approximation, let the treatment group be treatment 1 (again, the conclusions do not depend on this for a 2-tailed test).

$$n_1 = 19 \quad n_2 = 10 \quad N = 29 \quad T = 349 \quad \mu_T = \frac{19(30)}{2} = 285 \quad \sigma_T = \sqrt{\frac{19(10)(30)}{12}} = 21.79$$

$$z_{obs} = \frac{349 - 285}{21.79} = 2.937 \quad P = 2P(Z \geq 2.937) = .0033$$

The rank sum for the treatment group is much larger than we would have expected under the null hypothesis of no treatment effect.

∇

5.2.2 Paired Sample Designs

In paired samples (aka crossover) 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.

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.

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

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

$$\overline{D} = \frac{\sum_{i=1}^n D_i}{n} \quad E\{\overline{D}\} = \mu_D \quad V\{\overline{D}\} = \frac{\sigma_D^2}{n} \quad \text{For large } n\text{: } \overline{D} \dot{\sim} N\left(\mu_D, \frac{\sigma_D^2}{n}\right)$$

Normality holds for any sample sizes 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, we will always have equal sized samples when the experiment is conducted properly. We will always be looking at the $n = n_1 = n_2$ differences, and will have n differences, even though there were $2n = n_1 + n_2$ measurements made. From the n differences obtained in a sample, we will compute the mean and standard deviation, which we will label as $\overline{d}$ and s_d.

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

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

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

The test procedure is conducted as follows.

1. $H_0 : \mu_1 - \mu_2 = \mu_D = 0$
2. $H_A : \mu_D \neq 0$ or $H_A : \mu_D > 0$ or $H_A : \mu_D < 0$ (which alternative is appropriate should be clear from the setting).
3. T.S.: $t_{obs} = \frac{\overline{d}}{s_{\overline{D}}} = \frac{\overline{d}}{\left(\frac{s_d}{\sqrt{n}}\right)}$
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 5.6: 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), [10]). These are used in geographic tracing of wine. The two methods

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 5.3: $^{87}SR/^{86}SR$ Isotope ratios for 18 wine samples by Microwave and Low Temperature Methods

are microwave (method 1) and low temperature (method 2). The data, and the differences (microwave - lowtemp) are given in Table 5.3.

The 95% Confidence Interval for μ_D is computed below, where $t_{.025,18-1} = 2.110$. First we multiply the mean and standard deviations of the differences by 100000 (remove first 5 0s after decimal). This is legitimate as they are od the same units. This leads to $\overline{d}^* = 0.3667$ and $s^*_{\overline{D}} = 2.46466$.

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

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.

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

1. $H_0 : \mu_1 - \mu_2 = \mu_D = 0$
2. $H_A : \mu_D \neq 0$
3. T.S.: $t_{obs} = \frac{0.3667}{\left(\frac{2.4646}{\sqrt{18}}\right)} = \frac{0.3667}{0.5809} = 0.631$
4. R.R.: $t_{obs} > t_{\alpha/2,n-1} = t_{.025,17} = 2.110$

5. P-value: $2P(t_{17} \geq 0.631) = .5364$

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)

mean(microwave); sd(microwave)
mean(lowtemp);   sd(lowtemp)
cor(microwave,lowtemp)

## Brute Force Computations
diff <- microwave - lowtemp              ## Obtain differences
n.diff <- length(diff)                   ## Obtain n of diffs
mean.diff <- mean(diff)                  ## Obtain mean of diffs
sd.diff <- sd(diff)                      ## Obtain SD of diffs
se.diff <- sd.diff/sqrt(length(diff))    ## Obtain Std Error of mean
t.diff <- mean.diff/se.diff              ## t-statistic
pt.diff <- 2*(1-pt(abs(t.diff),n.diff-1))## P-value
t.025 <- qt(.975,n.diff-1)               ## Critical t-value
muD.LO <- mean.diff-t.025*se.diff        ## Lower Bound CI
muD.HI <- mean.diff+t.025*se.diff        ## Upper Bound CI

diff.out <- cbind(mean.diff, sd.diff, se.diff, t.diff, pt.diff, muD.LO,
                  muD.HI)
colnames(diff.out) <- c("mean","SD","Std Err", "t", "P(>|t|)","LB","UB")
round(diff.out,9)

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

## Output

> mean(microwave); sd(microwave)
[1] 0.7089294
[1] 0.0002870958
> mean(lowtemp);   sd(lowtemp)
[1] 0.7089257
[1] 0.0002878604
> cor(microwave,lowtemp)
[1] 0.9963286

> 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
```

∇

Wilcoxon Signed–Rank Test for Paired Data

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

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

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

1. H_0 : The two population distributions have equal Medians ($M_1 = M_2$)
2. H_A : The Medians Differ ($M_1 \neq M_2$)
3. T.S.: $T = \min(T^+, T^-)$
4. R.R.: $T \leq T_0$, where T_0 is a function of n and α and given in tables in many statistics texts and on the web.

For a one-sided test, if you wish to show that the distribution of population 1 is shifted to the right of population 2 ($M_1 > M_2$), the procedure is as follows:

1. H_0 : The two population distributions have equal Medians ($M_1 = M_2$)
2. H_A : Distribution 1 is shifted to the right of distribution 2 ($M_1 > M_2$)
3. T.S.: $T = T^-$
4. R.R.: $T \leq T_0$, where T_0 is a function of n and α and given in tables in many statistics texts and on the web.

Note that if you wish to use the alternative $M_1 < M_2$, use the above procedure with T^+ replacing T^-. The idea behind this test is to determine whether the differences tend to be positive ($M_1 > M_2$) or negative ($M_1 < M_2$), where differences are 'weighted' by their magnitude.

Example 5.7: Water Consumption by Cats under Still and Flowing Sources

A small pilot study was conducted to compare the daily amount of water consumed (mL) when presented with still or flowing water (Pachel and Neilson (2010) [23]). Each of $n = 9$ cats was observed 2 days each under each condition, and the mean for each condition was computed for each cat. Data are given in Table 5.4, along with ranks. We will test whether there is evidence if the true medians differ (even though this is clearly a very small sample).

Cat (i)	still	flowing	d_i=flowing-still	$\|d_i\|$	rank($\|d_i\|$)
1	157.5	164.5	7	7	2
2	84.5	51.5	-33	33	6
3	134.0	250.0	116	116	9
4	74.0	139.0	65	65	7
5	108.0	113.0	5	5	1
6	107.5	124.5	17	17	4
7	106.0	95.5	-10.5	10.5	3
8	163.0	70.5	-92.5	92.5	8
9	54.0	30.5	-23.5	23.5	5

Table 5.4: Average daily water consumed by cats in still and flowing conditions

Based on Table 5.4, we get T^+ (the sum of the ranks for positive differences) and T^- (the sum of the ranks of the negative differences), as well as the test statistic T, as follows:

$$T^+ = 2 + 9 + 7 + 1 + 4 = 23 \qquad T^- = 6 + 3 + 8 + 5 = 22 \qquad T = min(T^+, T^-) = min(23, 22) = 22$$

Note that short of there having been a tie, this is the closest T^+ and T^- could be. We can then use the previously given steps to test for differences in the medians of the true distributions for the 2 water conditions.

1. H_0 : The two population medians ($M_1 = M_2$)
2. H_A : One distribution is shifted to the right of the other ($M_1 \neq M_2$)
3. T.S.: $T = \min(T^+, T^-) = 22$
4. R.R.: $T \leq T_0$, where $T_0 = 5$ is based on 2-sided alternative , $\alpha = 0.05$, and $n = 9$.

Since $T = 22$ does not fall in the rejection region, we cannot reject H_0, and we fail to conclude that the medians differ. Note that the P-value is thus larger than 0.05, since we fail to reject H_0 (in fact it is 1).

R Commands and Output

```
## Commands
```

```
still <- c(157.5, 84.5, 134, 74, 108, 107.5, 106, 163, 54)
flowing <- c(164.5, 51.5, 250, 139, 113, 124.5, 95.5, 70.5, 30.5)

wilcox.test(still, flowing, paired=TRUE)

## Output

> wilcox.test(still, flowing, paired=TRUE)

        Wilcoxon signed rank test

data:  still and flowing
V = 22, p-value = 1
alternative hypothesis: true location shift is not equal to 0
```

∇

In large-samples, the rank-sums T^+ and T^- have approximately normal sampling distributions. By definition, $T^+ + T^- = 1 + \cdots + n = \frac{n(n+1)}{2}$. Under the null hypothesis $H_0 : M_1 = M_2$, we have the following mean and variance for T^+ and T^-.

$$\mu_T = \frac{n(n+1)}{4} \qquad \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \qquad z_obs = \frac{T - \mu_T}{\sigma_T}$$

The usual rules for rejection regions and P-values apply. If the alternative is $H_A : M_1 > M_2$ use $T = T^+$ and reject H_0 if $z_{obs} \geq z_\alpha$. If the alternative is $H_A : M_1 > M_2$ use $T = T^-$ and reject H_0 if $z_{obs} \geq z_\alpha$. For $H_A : M_1 \neq M_2$, use either T^+ or T^-, and reject if $|z_{obs}| \geq z_{\alpha/2}$.

Example 5.8: Efficiency Comparison of Recreational and Professional Bettors

An economic study was conducted, comparing recreational and professional bettors' efficiencies (Bruce, Johnson, and Peirson (2012), [3]). They considered race attendees as Recreational bettors and remote (online) bettors as Professional bettors. The authors had aggregate returns (amount won divided by amount bet) data for both groups on $n = 2057$ races. The difference for each race (remote - attendee) was obtained for each race. There were 963 negative differences (attendees outperformed remote bettors) and 1094 positive differences. The rank sum information is given below.

$$T^+ = 1167023.5 \quad T^- = 949629.5 \quad T^+T^- = 2116653 = 1+\cdots+2057 = \frac{2057(2058)}{2} \quad \mu_T = \frac{2057(2058)}{4} = 1058326.5$$

$$\sigma_T = \sqrt{\frac{2057(2057+1)(2(2057)+1)}{24}} = 26941.34 \qquad z_{obs} = \frac{1167023.5 - 1058326.5}{26941.34} = 4.03$$

There is strong evidence of a difference in the two groups. Note the authors also present the mean and the standard deviation of the differences.

$$\overline{y}_r = 0.8659 \quad \overline{y}_a = 0.8180 \quad \overline{d} = 0.0479 \quad s_d = 0.4448 \quad \frac{s_d}{\sqrt{n}} = \frac{0.4448}{\sqrt{2057}} = 0.0098$$

$$0.0479 \pm 1.96(0.0098) \equiv (0.0287, 0.0671)$$

∇

5.3 Power and Sample Size Considerations

In this section, issues of power and sample size are considered in the 2-Sample Location problem. Power refers to the probability of rejecting the null hypothesis. When H_0 is true, it should be α, and when the alternative is true, it will depend on the magnitude of the difference, the variability and the sample sizes. Once power has been considered empirically, sample size computations will be made based on distributional results.

5.3.1 Empirical Study of Power

To compare the power of the independent sample t-test and the Wilcoxon Rank-Sum test, we return to the populations of NHL/EPL players' BMI and the Female and Male marathon runner's speeds. The BMI distributions were approximately normal, while the marathon speeds were right skewed.

Example 5.9: Small-Sample Inference Comparing BMI for NHL and EPL Players

The means and standard deviations of the BMI levels for NHL and EPL players are given below, along with the mean and variance of the sampling distribution of $\overline{Y}_{NHL} - \overline{Y}_{EPL}$. Note that as each distribution is approximately normal, its sampling distribution will be very close to a normal distribution, even with relatively small samples. Further, the variances are not equal, although they are not too far apart. Refer back to Figure 5.2 for a histogram of 100000 random samples' mean differences of $n_1 = n_2 = 20$.

BMI: $\mu_{NHL} = 26.50 \quad \sigma_{NHL} = 1.45 \quad \mu_{EPL} = 23.02 \quad \sigma_{EPL} = 1.71 \quad E\{\overline{Y}_{NHL} - \overline{Y}_{EPL}\} = 26.50 - 23.02 = 3.48$

$$V\{\overline{Y}_{NHL} - \overline{Y}_{EPL}\} = \frac{1.45^2}{n_1} + \frac{1.71^2}{n_2}$$

We compare the coverage rates of small sample Confidence Intervals based on equal variance and unequal variance assumptions, as well as their widths for samples of $n_1 = n_2 = 10$. The unequal variance case will always be wider, as the sample mean difference and estimated standard error will be the same as the equal variance case, but will have fewer degrees of freedom. Due to the equivalence of the 2-tailed test and Confidence Interval for testing $H_0 : \mu_1 - \mu_2 = 0$, we can also observe the empirical power of the two methods. The process is conducted as follows.

1. Sample 10 players from NHL and 10 players from EPL
2. Compute $\overline{y}_{NHL}$, s_{NHL}, $\overline{y}_{EPL}$, s_{EPL}

3. Compute the sample mean difference $\overline{y}_{NHL} - \overline{y}_{EPL}$ and its estimated standard error $\sqrt{\frac{s^2_{NHL}}{10} + \frac{s^2_{EPL}}{10}}$
4. Compute the approximate degrees of freedom for the unequal variance case (Satterthwaite's approximation)
5. Obtain the 95% Confidence Intervals for $\mu_{NHL} - \mu_{EPL}$
6. Determine whether the Confidence Intervals contain 3.48 (true value) and whether they contain 0 (Testing $\mu_{NHL} - \mu_{EPL} = 0$)
7. Obtain the width of the intervals

The equal variance Confidence Intervals contained $\mu_1 - \mu_2 = 3.48$ in 95.18% of the samples, the unequal variance CI's covered in 95.36% of the samples. Based on equal sample sizes, (and will typically always be the case) the unequal case will always have wider intervals and thus higher coverage rates at the cost of being wider. The average width of the equal variance CI's was 2.9291 versus 2.9597 for the unequal case. The unequal case was only about 1% wider on average due to how similar the population standard deviations are. The equal variance case rejected $H_0 : \mu_1 - \mu_2 = 0$ in favor of $H_A : \mu_1 - \mu_2 \neq 0$ in 99.08% of the samples, while the unequal variance case did so in 98.91%. Neither ever rejected with a negative t-statistic. The mean difference was very large, so it's not surprising to have such high power.

R Commands and Output

```
## Commands

bmi.sim <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_nba_ebl_bmi.csv",
   header=TRUE)
attach(bmi.sim); names(bmi.sim)

## Obtain populations and mu and sigma for each
N.nhl <- 717     #  # of NHL players
N.epl <- 526     #  # of EPL players
bmi.nhl <- NHL_BMI[1:N.nhl]
bmi.epl <- EPL_BMI[1:N.epl]
(mu.nhl <- mean(bmi.nhl)); (sigma.nhl <- sd(bmi.nhl))
(mu.epl <- mean(bmi.epl)); (sigma.epl <- sd(bmi.epl))

## Set up and run samples and ybar and s arrays
num.sim <- 100000
n.nhl <- 10
n.epl <- 10
(mu.meandiff <- mu.nhl - mu.epl)
(sigma.meandiff <- sqrt(sigma.nhl^2/n.nhl + sigma.epl^2/n.epl))
set.seed(1122)
ybar.s.nhl <- matrix(rep(0,2*num.sim),ncol=2)
ybar.s.epl <- matrix(rep(0,2*num.sim),ncol=2)

for (i in 1:num.sim) {
y1 <- sample(bmi.nhl,n.nhl,replace=F)
y2 <- sample(bmi.epl,n.nhl,replace=F)

ybar.s.nhl[i,1] <- mean(y1)
ybar.s.nhl[i,2] <- sd(y1)
ybar.s.epl[i,1] <- mean(y2)
ybar.s.epl[i,2] <- sd(y2)
}
## End of sampling
```

```
## Generate sample mean differences SE's and CI's
## ev=equal variances, uv=unequal variances
meandiff <- ybar.s.nhl[,1] - ybar.s.epl[,1]
se.meandiff <- sqrt(ybar.s.nhl[,2]^2/n.nhl + ybar.s.epl[,2]^2/n.epl)
df.uv1 <- (ybar.s.nhl[,2]^2/n.nhl + ybar.s.epl[,2]^2/n.epl)^2
df.uv2 <- ((ybar.s.nhl[,2]^2/n.nhl)^2/(n.nhl-1)) +
          ((ybar.s.epl[,2]^2/n.epl)^2/(n.epl-1))
df.uv <- df.uv1 / df.uv2
df.ev <- n.nhl + n.epl - 2
meandiff.LB.ev <- meandiff + qt(.025,df.ev) * se.meandiff
meandiff.UB.ev <- meandiff + qt(.975,df.ev) * se.meandiff
meandiff.LB.uv <- meandiff + qt(.025,df.uv) * se.meandiff
meandiff.UB.uv <- meandiff + qt(.975,df.uv) * se.meandiff

## Obtain Coverage rates, widths, power (H0:mu1-mu2=0)
sum(meandiff.LB.ev <= mu.meandiff & meandiff.UB.ev >= mu.meandiff) / num.sim
sum(meandiff.LB.uv <= mu.meandiff & meandiff.UB.uv >= mu.meandiff) / num.sim
mean(meandiff.UB.ev-meandiff.LB.ev)
mean(meandiff.UB.uv-meandiff.LB.uv)
sum(meandiff.LB.ev >= 0) / num.sim
sum(meandiff.LB.uv >= 0) / num.sim
sum(meandiff.UB.ev <= 0) / num.sim
sum(meandiff.UB.uv <= 0) / num.sim

## Output

> (mu.nhl <- mean(bmi.nhl)); (sigma.nhl <- sd(bmi.nhl))
[1] 26.50015
[1] 1.454726
> (mu.epl <- mean(bmi.epl)); (sigma.epl <- sd(bmi.epl))
[1] 23.01879
[1] 1.713098
> (mu.meandiff <- mu.nhl - mu.epl)
[1] 3.481361
> (sigma.meandiff <- sqrt(sigma.nhl^2/n.nhl + sigma.epl^2/n.epl))
[1] 0.7106991
> sum(meandiff.LB.ev <= mu.meandiff & meandiff.UB.ev >= mu.meandiff) / num.sim
[1] 0.95178
> sum(meandiff.LB.uv <= mu.meandiff & meandiff.UB.uv >= mu.meandiff) / num.sim
[1] 0.95362
> mean(meandiff.UB.ev-meandiff.LB.ev)
[1] 2.929125
> mean(meandiff.UB.uv-meandiff.LB.uv)
[1] 2.959715
> sum(meandiff.LB.ev >= 0) / num.sim
[1] 0.9903
> sum(meandiff.LB.uv >= 0) / num.sim
[1] 0.98914
> sum(meandiff.UB.ev <= 0) / num.sim
[1] 0
> sum(meandiff.UB.uv <= 0) / num.sim
[1] 0
```

∇

Example 5.10: Small-Sample Inference for Female and Male Marathon Speeds

Comparisons among Female and Male marathon speeds are now made. Unlike the NHL/EPL Body Mass Indices, these speeds are not approximately normally distributed, but are rather skewed to the right,

refer to Figure 3.5. The population means and standard deviations are given below, along with the mean and standard error of the sampling distribution of the sample mean $\overline{Y}_F - \overline{Y}_M$.

$$\mu_F = 5.840 \quad \sigma_F = 0.831 \qquad \mu_M = 6.337 \quad \sigma_F = 1.058$$

$$E\{\overline{Y}_F - \overline{Y}_M\} = -0.497 \qquad V\{\overline{Y}_F - \overline{Y}_M\} = \sqrt{\frac{0.831^2}{n_F} + \frac{1.058^2}{n_2}}$$

We will consider fairly small samples, $n_F = n_M = 6$, and first repeat the comparisons made in BMI example, and further compare the t-tests with the Wilcoxon Rank-Sum test in terms of power for testing $H_0 : \mu_F - \mu_M \geq 0$ vs $H_A : \mu_F - \mu_M < 0$. The equal variance Confidence Interval covered $\mu_F - \mu_M - 0.497$ in 94.86% of samples, the unequal case covered in 95.33%, so even with these small samples, and the skewed distributions, the t-based Confidence Intervals performed well. In terms of concluding $H_A : \mu_F - \mu_M < 0$, the equal variance t-test correctly rejected H_0 in 20.67% of samples, the unequal variance t-test in 19.58%, and the Wilcoxon Rank-Sum test in 19.04%.

R Program and Output

```
## Commands
## Read data from website and attach data frame and obtain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)
f.mph <- mph[Gender=="F"]
m.mph <- mph[Gender=="M"]
(mu.f <- mean(f.mph)); (sigma.f <- sd(f.mph))

(mu.f <- mean(f.mph)); (sigma.f <- sd(f.mph))
(mu.m <- mean(m.mph)); (sigma.m <- sd(m.mph))

num.sim <- 100000
n.f <- 6; n.m <- 6
(mu.meandiff <- mu.f - mu.m)
(sigma.meandiff <- sqrt(sigma.f^2/n.f + sigma.m^2/n.m))
set.seed(3344)
ybar.s.f <- matrix(rep(0,2*num.sim),ncol=2)
ybar.s.m <- matrix(rep(0,2*num.sim),ncol=2)
ranksum.fm <- matrix(rep(0,2*num.sim),ncol=2)


for (i in 1:num.sim) {
y1 <- sample(f.mph,n.f,replace=F)
y2 <- sample(m.mph,n.m,replace=F)

ybar.s.f[i,1] <- mean(y1)
ybar.s.f[i,2] <- sd(y1)
ybar.s.m[i,1] <- mean(y2)
ybar.s.m[i,2] <- sd(y2)
ranksum.fm [i,1] <- sum(rank(c(y1,y2))[1:n.f])
ranksum.fm [i,2] <- sum(rank(c(y1,y2))[(n.f+1):(n.f+n.m)])
}

meandiff <- ybar.s.f[,1] - ybar.s.m[,1]
se.meandiff <- sqrt(ybar.s.f[,2]^2/n.f + ybar.s.m[,2]^2/n.m)
df.uv1 <- (ybar.s.f[,2]^2/n.f + ybar.s.m[,2]^2/n.m)^2
```

```
df.uv2 <- ((ybar.s.f[,2]^2/n.f)^2/(n.f-1)) +
          ((ybar.s.m[,2]^2/n.m)^2/(n.m-1))
df.uv <- df.uv1 / df.uv2
df.ev <- n.f + n.m - 2
meandiff.LB.ev <- meandiff + qt(.025,df.ev) * se.meandiff
meandiff.UB.ev <- meandiff + qt(.975,df.ev) * se.meandiff
meandiff.LB.uv <- meandiff + qt(.025,df.uv) * se.meandiff
meandiff.UB.uv <- meandiff + qt(.975,df.uv) * se.meandiff

## Obtain Coverage rates, widths, power (H0:mu1-mu2=0  HA:mu1-mu2<0)
sum(meandiff.LB.ev <= mu.meandiff & meandiff.UB.ev >= mu.meandiff) / num.sim
sum(meandiff.LB.uv <= mu.meandiff & meandiff.UB.uv >= mu.meandiff) / num.sim
mean(meandiff.UB.ev-meandiff.LB.ev)
mean(meandiff.UB.uv-meandiff.LB.uv)

t.uv.ev <- meandiff / se.meandiff
rr.t.uv <- qt(.05,df.uv)
rr.t.ev <- qt(.05,df.ev)
rr.t1.w <- 28                        ## From Wilcoxon Rank-sum w/ n1=n2=6
sum(t.uv.ev <= rr.t.uv) / num.sim
sum(t.uv.ev <= rr.t.ev) / num.sim
sum(ranksum.fm[,1] <= rr.t1.w) / num.sim

## Output

> (mu.f <- mean(f.mph)); (sigma.f <- sd(f.mph))
[1] 5.839839
[1] 0.8310405
> (mu.m <- mean(m.mph)); (sigma.m <- sd(m.mph))
[1] 6.336979
[1] 1.057687
> (mu.meandiff <- mu.f - mu.m)
[1] -0.49714
> (sigma.meandiff <- sqrt(sigma.f^2/n.f + sigma.m^2/n.m))
[1] 0.5491402
> sum(meandiff.LB.ev <= mu.meandiff & meandiff.UB.ev >= mu.meandiff) / num.sim
[1] 0.94856
> sum(meandiff.LB.uv <= mu.meandiff & meandiff.UB.uv >= mu.meandiff) / num.sim
[1] 0.95331
> mean(meandiff.UB.ev-meandiff.LB.ev)
[1] 2.383104
> mean(meandiff.UB.uv-meandiff.LB.uv)
[1] 2.45255
> sum(t.uv.ev <= rr.t.uv) / num.sim
[1] 0.19584
> sum(t.uv.ev <= rr.t.ev) / num.sim
[1] 0.20669
> sum(ranksum.fm[,1] <= rr.t1.w) / num.sim
[1] 0.19042
```

∇

5.3.2 Power Computations

To obtain the sample sizes needed to detect an important difference means, the non-central t-distribution can be used in a similar manner to what was done for the one-sample problem. The only difference is that instead of looking for an important difference from some pre-specified null mean, we are interested in the

difference between two population means. First, consider the case of independent samples. This is generally done under the assumption of equal variances.

$$H_0 : \mu_1 - \mu_2 = 0 \quad \mu_1 - \mu_2 = (\mu_1 - \mu_2)_A \neq 0 \quad \Delta = \frac{(\mu_1 - \mu_2)_A}{\sigma\sqrt{\frac{2}{n}}} \qquad t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{S_p^2\left(\frac{2}{n}\right)}} \sim t_{2(n-1),\Delta}$$

If σ is known (or well approximated), researchers can choose an important difference $(\mu_1 - \mu_2)_A$, and determine the sample size that gives a reasonable power π to detect it based on a test with significance level α. In other situations, an important **effect size** $\delta = (\mu_1 - \mu_2)_A / \sigma$ can be obtained, which measures the difference in means in standard deviation units. Once the important effect size is chosen, beginning with small n, the power π is determined and the process continues until the desired power is obtained. The process works as follows for a 2-tailed test.

1. Determine important effect size $\delta = (\mu_1 - \mu_2)_A / \sigma$ and set the significance level α and desired power π.
2. Starting with (say) $n_1 = n_2 = n = 2$, obtain the degrees of freedom $2(n-1)$ and critical value $t_{\alpha/2,2(n-1)}$.
3. Compute the non-centrality parameter $\Delta = \frac{\delta}{\sqrt{2/n}}$.
4. Obtain π_n: the probability the non-central t is greater than $t_{\alpha/2,2(n-1)}$ or less than $-t_{\alpha/2,2(n-1)}$.
5. If π_n exceeds the desired π, stop. Otherwise, increment n by 1 and repeat the process.

Example 5.11: Power Calculation for Comparison of Female and Male Marathon Speeds

Using numbers similar to those observed in the populations of marathon runners, suppose we want to be able to detect a difference $(\mu_F - \mu_M)_A = -0.5$ and that $\sigma_F = \sigma_M = \sigma = 0.94$ (we are just averaging the true standard deviations for computational purposes). We then obtain the following results. We will start with $n_1 = n_2 = n = 6$, since the power was so low (approximately 0.20) for the lower-tailed t-test in Example 5.10.

$$\delta = \frac{-0.50}{0.94} = -0.532 \quad \Delta_6 = \frac{-0.532}{\sqrt{2/6}} = -0.921 \quad df = 2(6-1) = 10 \quad -t_{.05,10} = -1.812$$

For the lower-tailed test $H_A : \mu_F - \mu_M < 0$, for these sample sizes, reject the null of no difference if $t_{obs} \leq -1.812$. Now we find the probability under the non-central t-density with 10 degrees of freedom and non-centrality parameter -0.921 that is below -1.812. The power turns out to be 0.216 (see R output below). Using the R functions **qt** for quantiles and **pt** for lower tail probabilities (cumulative distribution function), the relevant probabilities (powers) can be obtained. Samples of size $n_F = n_M = 45$ would be needed for the power to reach 0.8.

R Commands and Output

```
## Commands

## Set parameters, alpha, chosen power, for starting sample size (n0)
m1_m2_A <- -0.50
sigma <- 0.94
n0 <- 6
alpha <- 0.05
power.star <- 0.80
(delta <- m1_m2_A / sigma)
(crit_val <- qt(.05,2*(n0-1)))
(power.lt <- pt(crit_val,2*(n0-1),delta/sqrt(2/n0)))

## Set up holders for power and sample size and row and sample size start values
power.out <- numeric()
n.out <- numeric()
i <- 0
n <- n0

## Loop until power exceeds chosen power
while (power.lt < power.star) {
i <- i+1
n <- n+1
crit_val <- qt(alpha,2*(n-1))
power.lt <- pt(crit_val,2*(n-1),delta/sqrt(2/n))
power.out[i] <- power.lt
n.out[i] <- n
}

## Print Sample sizes and corresponding powers
cbind(n.out, power.out)

## Output

> (delta <- m1_m2_A / sigma)
[1] -0.5319149
> (crit_val <- qt(.05,2*(n0-1)))
[1] -1.812461
> (power.lt <- pt(crit_val,2*(n0-1),delta/sqrt(2/n0)))
[1] 0.2161749

> cbind(n.out, power.out)
      n.out power.out
 [1,]     7 0.2402697
 [2,]     8 0.2636261
        ...
[38,]    44 0.7968277
[39,]    45 0.8047651
```

Had this been a 2-tailed test with $H_A : \mu_F - \mu_M \neq 0$, the Rejection Region would be $|t_{obs}| \geq t_{\alpha/2,2(n-1)}$. Below is the R Commands and that computes the power for the 2-tailed test (it only contains the initial calculation, the loop part is similar to the lower-tail test). Samples of $n = 57$ females and males would be needed for the power to reach 0.80.

R Commands and Output

```
## Commands
####### 2-Tailed Test
## Set parameters, alpha, chosen power, for starting sample size (n0)
m1_m2_A <- -0.50
sigma <- 0.94
```

```
n0 <- 6
alpha <- 0.05
power.star <- 0.80
(delta <- m1_m2_A / sigma)
(crit_val_lo <- qt(.05/2,2*(n0-1)))
(crit_val_hi <- qt(1-.05/2,2*(n0-1)))
(power.2t <- pt(crit_val_lo,2*(n0-1),delta/sqrt(2/n0)) +
          (1-pt(crit_val_hi,2*(n0-1),delta/sqrt(2/n0))))

## Output
> (delta <- m1_m2_A / sigma)
[1] -0.5319149
> (crit_val_lo <- qt(.05/2,2*(n0-1)))
[1] -2.228139
> (crit_val_hi <- qt(1-.05/2,2*(n0-1)))
[1] 2.228139
> (power.2t <- pt(crit_val_lo,2*(n0-1),delta/sqrt(2/n0)) +
+           (1-pt(crit_val_hi,2*(n0-1),delta/sqrt(2/n0))))
[1] 0.1329802
> cbind(n.out, power.out)
      n.out power.out
 [1,]     7 0.1505426
      ...
[50,]    56 0.7967349
[51,]    57 0.8037961
```

∇

In terms of the paired t-test, when testing $H_0 : \mu_D = 0$ vs $H_A : \mu_D \neq 0$, there may be a specific difference μ_{DA} that would like to be detected with a specified power π. This is very similar to the 1-sample problem in the previous chapter. Define the following terms, where μ_{DA} is the mean difference under H_A and σ_D is the standard deviation of the differences.

$$t_{obs} = \frac{\overline{d}}{s_d/\sqrt{n}} = \sqrt{n}\frac{\overline{d}}{s_d} \qquad \delta = \frac{\mu_{DA}}{\sigma_D} \qquad \Delta = \sqrt{n}\delta$$

Again δ is the effect size and Δ is the non-centrality parameter. The degrees of freedom for the paired t-test is $n - 1$. The process generalizes directly from the independent samples method described above.

Example 5.12: Water Consumption by Cats under Still and Flowing Sources

In the pilot study of cats drinking flowing versus still water, the standard deviation of the differences was approximately 60 ml. Suppose the researchers would like to detect a true mean difference of $\mu_{DA} = 30$ mL with power $\pi = 0.75$. In this setting $\delta = 30/60 = 0.5$ and $\Delta = \sqrt{n}(0.5)$. Beginning with the authors original sample of $n = 9$, we obtain the power then iterate until $\pi \geq 0.75$. The R program and output are given below, for $n = 9$, $\pi = 0.263$. A sample of $n = 30$ would be needed to reach $\pi = 0.75$.

R Commands and Output

```
## Commands
```

```
mu_DA <- 30
sigma_D <- 60
n0 <- 9
alpha <- .05
power.star <- 0.75
(delta <- mu_DA / sigma_D)
(crit_val_lo <- qt(alpha/2,n0-1))
(crit_val_hi <- qt(1-alpha/2,n0-1))
(power.2t <- pt(crit_val_lo,n0-1,sqrt(n0)*delta) +
          (1-pt(crit_val_hi,n0-1,sqrt(n0)*delta)))

power.out <- numeric()
n.out <- numeric()
i <- 0
n <- n0

## Loop until power exceeds chosen power
while (power.2t < power.star) {
i <- i+1
n <- n+1
crit_val_lo <- qt(.05/2,n-1)
crit_val_hi <- qt(1-.05/2,n-1)
power.2t <- pt(crit_val_lo,n-1,sqrt(n)*delta) +
          (1-pt(crit_val_hi,n-1,sqrt(n)*delta))

power.out[i] <- power.2t
n.out[i] <- n
}

## Print Sample sizes and corresponding powers
cbind(n.out, power.out)

## Output
> (power.2t <- pt(crit_val_lo,n0-1,sqrt(n0)*delta) +
+           (1-pt(crit_val_hi,n0-1,sqrt(n0)*delta)))
[1] 0.2627461
> cbind(n.out, power.out)
      n.out power.out
 [1,]    10 0.2931756
      ...
[20,]    29 0.7386963
[21,]    30 0.7539647
```

∇

5.4 Methods Based on Resampling

In this section, two methods for comparing two means are considered. These are the **Bootstrap** and **Randomization/Permutation Tests**.

5.4.1 Bootstrap

The bootstrap method is the same principle as in the one-sample case. In terms of independent samples, take resamples within each group, then take the difference between the two groups in each subsample. This will be illustrated below. In terms of paired samples, the one-sample methods are used on the observed paired differences from the original sample.

For the Bootstrap t Intervals, for each resample, compute $\overline{y}^*_{1i}, s^*_{1i}, \overline{y}^*_{2i}, s^*_{2i}$ for the i^{th} resample, and compute t^*_i as below, where $n_1, \overline{y}_1, s_1, n_2, \overline{y}_2, s_2$ are the sizes, means, and standard deviations of the original samples.

$$t^*_i = \frac{(\overline{y}^*_{1i} - \overline{y}^*_{2i}) - (\overline{y}_1 - \overline{y}_2)}{\sqrt{\frac{s^{*2}_{1i}}{n_1} + \frac{s^{*2}_{2i}}{n_2}}} \quad i = 1, \ldots, B$$

Once the B t_i statistics are obtained the $\alpha/2$ and $1 - |alpha/2$ quantiles are obtained and labeled Q^*_L and Q_U, respectively. The $(1-\alpha)100\%$ Bootstrap t CI for $\mu_1 - \mu_2$ is of the following form.

$$(\overline{y}_1 - \overline{y}_2) - Q^*_U\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}, (\overline{y}_1 - \overline{y}_2) - Q^*_L\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}$$

Example 5.13: Anthropometric Measurements of Lahoul and Kulu Kanets in Punjab

A study sampled 30 Lahoul Kanet adults and 60 Kulu Kanet adults, making various physical measurements (Holland (1902) [14]). The author reported on 7 characteristics among each subject. Consider the variable cubit (cm), given in Table 5.5. The summary statistics from the samples are given below.

$$n_L = 30 \quad \overline{y}_L = 44.657 \quad s_L = 2.056 \qquad n_K = 60 \quad \overline{y}_K = 45.298 \quad s_K = 1.692 \qquad \overline{y}_L - \overline{y}_K = -0.641$$

We take 10000 resamples of 30 Lahoul and 60 Kulu Kanets, obtaining the means for each group and the difference. Then, obtaining the bootstrap mean and standard error for the differences, along with the bootstrap percentile intervals from the 2.5 and 97.5 percentiles of the resampled mean differences. The mean of the 10000 mean differences is -0.639, the bootstrap standard error is 0.427, and the 95% bootstrap percentile Confidence Interval is (-1.458, 0.212). A histogram of the resample mean differences and a normal probability plot are given in Figure 5.5.

R Commands and Output

```
## Commands

kanet <- read.fwf("http://www.stat.ufl.edu/~winner/data/kanet.dat",
     width=c(18,2,rep(8,7)), col.names=c("name","kgroup","age","stature",
     "armspan","sitheight","knlheight","cubit","leftfoot"))
attach(kanet)
```

Lahoul	Lahoul	Kulu	Kulu	Kulu	Kulu
45.2	44.3	44.8	46.6	44.9	43.2
46.9	46.6	45.7	43.3	46.1	45.7
44.7	42.4	44.4	44.9	47.5	46.4
46.3	42.7	45.8	44.6	44.9	49.3
43.4	44.9	44.6	45.3	49.2	46.1
43.3	42.3	44.3	44.6	43.7	44.7
39.6	43.5	45.4	47.8	46.0	45.1
45.6	42.9	44.3	44.0	43.7	43.4
43.6	46.8	44.8	47.8	45.4	45.6
44.2	46.2	43.2	47.8	45.0	47.7
47.4	43.9	46.5	44.9	42.8	50.3
48.2	46.8	45.0	45.1	47.1	42.1
45.0	43.3	46.8	44.3	45.7	46.2
45.4	42.5	41.9	45.5	45.2	44.1
42.9	48.9	44.9	43.8	42.5	45.6

Table 5.5: Cubit lengths (cm) for samples of 30 Lahoul Kanets and 60 Kulu Kanets

```
cubit
tapply(cubit,kgroup,mean)
tapply(cubit,kgroup,sd)

L.cubit <- cubit[kgroup==1]
K.cubit <- cubit[kgroup==2]

set.seed(97531)
num.boot <- 10000
boot.ybar1 <- rep(0,num.boot)
boot.ybar2 <- rep(0,num.boot)
n.L <- length(L.cubit)
n.K <- length(K.cubit)
for (i in 1:num.boot) {
y1 <- sample(L.cubit, n.L, replace=T)
y2 <- sample(K.cubit, n.K, replace=T)
boot.ybar1[i] <- mean(y1); boot.ybar2[i] <- mean(y2)
}

meandiff <- boot.ybar1-boot.ybar2
mean(meandiff)
sd(meandiff)
quantile(meandiff,c(.025,.975))

par(mfrow=c(1,2))
hist(meandiff,breaks=30)
abline(v=(mean(L.cubit)-mean(K.cubit)),lwd=2)
qqnorm(meandiff); qqline(meandiff)

## Output

> tapply(cubit,kgroup,mean)
       1        2
44.65667 45.29833
> tapply(cubit,kgroup,sd)
       1        2
2.055889 1.692305
```

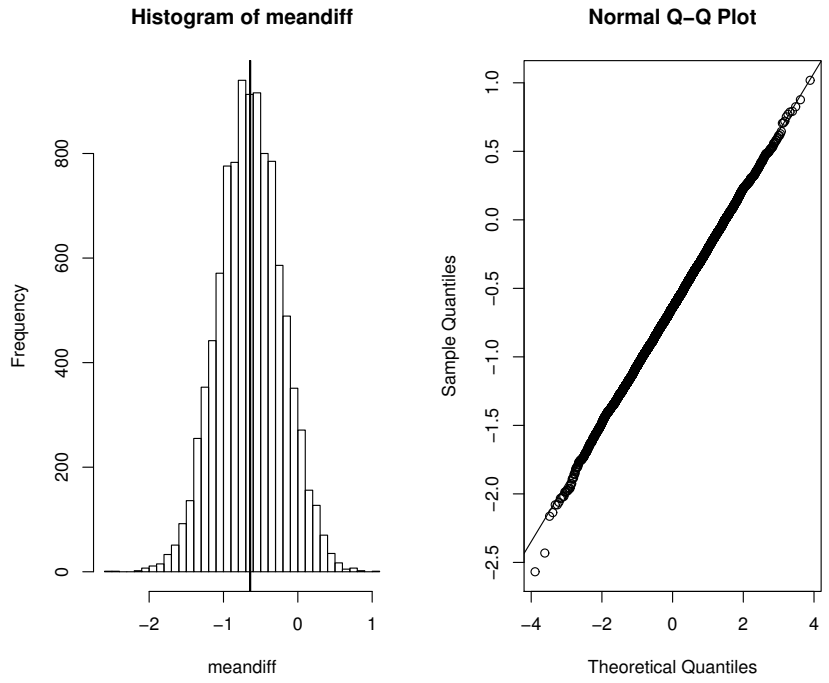

Figure 5.5: Histogram and Normal Probability Plot for Bootstrap Resample Mean Differences (Lahout - Kulu)

```
> mean(meandiff)
[1] -0.6386542
> sd(meandiff)
[1] 0.4273809
> quantile(meandiff,c(.025,.975))
      2.5%      97.5%
-1.4583750  0.2116667
```

For the 95% Bootstrap t Confidence Interval, the .025 quantile of t^* is $Q_L = -1.715$ and the .975 quantile is $Q_U = 1.830$ and the resulting 95% Confidence Interval is (-1.437, 0.103).

R Commands and Output

```
## Commands

## Bootstrap t CIs - Chihara and Hesterberg, Sec.7.5, p.198-200
set.seed(97531)
num.boot <- 10000
boot.ybar.s.L <- matrix(rep(0,2*num.boot),ncol=2)
boot.ybar.s.K <- matrix(rep(0,2*num.boot),ncol=2)
n.L <- length(L.cubit)
n.K <- length(K.cubit)
mean.L <- mean(L.cubit)
mean.K <- mean(K.cubit)
sd.L <- sd(L.cubit)
```

```
sd.K <- sd(K.cubit)

for (i in 1:num.boot) {
y1 <- sample(L.cubit, n.L, replace=T)
y2 <- sample(K.cubit, n.K, replace=T)
boot.ybar.s.L[i,1] <- mean(y1); boot.ybar.s.K[i,1] <- mean(y2)
boot.ybar.s.L[i,2] <- sd(y1);   boot.ybar.s.K[i,2] <- sd(y2)
}

t.star <- ((boot.ybar.s.L[,1]-boot.ybar.s.K[,1])-(mean.L-mean.K)) /
          sqrt((boot.ybar.s.L[,2]^2/n.L)+(boot.ybar.s.K[,2]^2/n.L))

(Q_L <- quantile(t.star, 0.025))
(Q_U <- quantile(t.star, 0.975))

((mean.L - mean.K) - Q_U * sqrt((sd.L^2/n.L) + (sd.K^2/n.K)))
((mean.L - mean.K) - Q_L * sqrt((sd.L^2/n.L) + (sd.K^2/n.K)))

## Output

> (Q_L <- quantile(t.star, 0.025))
     2.5%
-1.715131
> (Q_U <- quantile(t.star, 0.975))
   97.5%
1.830131
> ((mean.L - mean.K) - Q_U * sqrt((sd.L^2/n.L) + (sd.K^2/n.K)))
-1.436502
> ((mean.L - mean.K) - Q_L * sqrt((sd.L^2/n.L) + (sd.K^2/n.K)))
0.1032235
```

∇

5.4.2 Randomization/Permutation Tests

Randomization/Permutation tests consider the observed responses as being made up of a treatment/population mean and a random error term. That is, $Y_{ij} = \mu_i + \epsilon_{ij}, i = 1, 2; j = 1, \ldots, n_{ij}$. The random error term is unique to the experimental unit that it corresponds to, and could be due to any number of factors. If there are no differences in the treatment/population means ($\mu_1 = \mu_2$), then all of the observed values could have come from either treatment/population on any number of randomizations by the experimenter or nature. The process of randomization and permutation tests is as follows for the independent sample t-test.

1. Compute a statistic from the original data that measures a discrepancy between the sample data and the null hypothesis, such as $\overline{y}_1 - \overline{y}_2$.
2. Generate many permutations (N) of the original samples to the two groups and compute and save the statistic for each permutation.
3. Count the number of permutations for which the statistic is as or more extreme than the original sample's value.
4. The P-value is (Count+1)/(N+1) the proportion of the statistics as or more extreme than the original (including the original).

Example 5.14: Cubit Lengths of Lahout and Kulu Kanets

To illustrate the test, consider the lengths of the cubits of the Lahout and Kulu Kanets. In Example 5.14, the mean difference from the original samples was $\overline{y}_1 - \overline{y}_2 = -0.641$. Suppose there is no difference in the two cultures' tendencies to generate different cubit lengths and they are due to randomness among individuals who "nature" randomized to the cultures. Then consider 9999 permutations of these 90 cubit lengths to the n_L Lahouts and n_K Kulus. Of $N = 9999$ permutation samples, 1207 were as large as the observed difference in absolute value, for a P-value of (1207+1)/(9999+1) = .1208. Thus, there is no evidence to reject the null hypothesis that $\mu_L \neq \mu_L$. A histogram of the permutation mean differences with a vertical line at the observed mean difference is given in Figure 5.6.

R Commands and Output

```
## Commands

kanet <- read.fwf("http://www.stat.ufl.edu/~winner/data/kanet.dat",
     width=c(18,2,rep(8,7)), col.names=c("name","kgroup","age","stature",
     "armspan","sitheight","knlheight","cubit","leftfoot"))
attach(kanet)
L.cubit <- cubit[kgroup==1]
K.cubit <- cubit[kgroup==2]
(TS.obs <- mean(L.cubit) - mean(K.cubit))

## Set up and obtain Permutation Samples
set.seed(24680)
num.perm <- 9999
TS <- rep(0,num.perm)
n.L <- length(L.cubit)
n.K <- length(K.cubit)
n.LK <- n.L + n.K
for (i in 1:num.perm) {
perm <- sample(1:n.LK,n.LK,replace=F)       # Permutation of 1:90
ybar1 <- mean(cubit[perm[1:n.L]])           # First 30 assigned L
ybar2 <- mean(cubit[perm[(n.L+1):n.LK]])     # Last 60 assigned K
TS[i] <- ybar1 - ybar2
}

## Count # permutations where |TS| >= |TS.obs| and obtain 2-tail P-value
(num.exceed <- sum(abs(TS) >= abs(TS.obs)))
(p.val.2tail <- (num.exceed+1) / (num.perm+1))

hist(TS,breaks=30, xlab="MeanL - MeanK",
   main="Randomization Distribution for Cubit Length")
abline(v=TS.obs,lwd=2)

## Output

> (TS.obs <- mean(L.cubit) - mean(K.cubit))
[1] -0.6416667
> (num.exceed <- sum(abs(TS) >= abs(TS.obs)))
[1] 1207
> (p.val.2tail <- (num.exceed+1) / (num.perm+1))
[1] 0.1208
```

∇

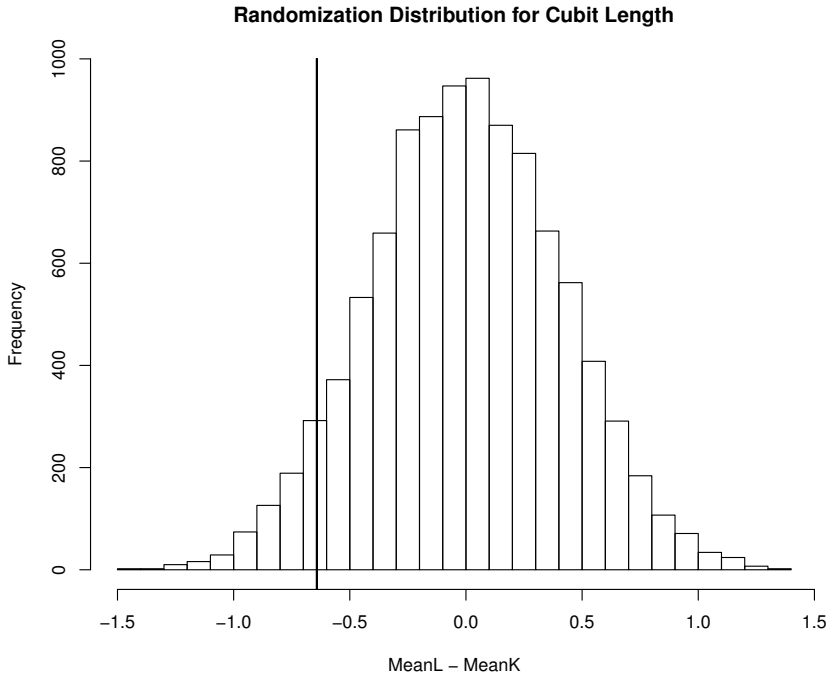

Figure 5.6: Randomization Distribution for Lahout and Kulu Kanet cubit measurements

For paired samples, if there is no difference in the means of the two treatments, then the 2 observed measurements on each unit or pair could have just as easily appeared under either of the two treatments. The process for the Randomization/Permutation test goes as follows.

1. Compute a statistic from the original data that measures a discrepancy between the sample data and the null hypothesis, such as $\overline{d}$.
2. Generate many permutations (N) of the signs of the observed differences, where for each unit, its sign is changed with probability 0.5 (in effect switching the observed scores for the two treatments). Compute and save the mean difference $\overline{d}^*$.
3. Count the number of permutations for which the statistic is as or more extreme than the original sample's value.
4. The P-value is (Count+1)/(N+1) the proportion of the statistics as or more extreme than the original (including the original).

Example 5.15: Home Field Advantage in English Premier League Football (2012)

The English Premier League has 20 football clubs. Each club plays the remaining 19 clubs twice each season (once at home, once away). If clubs are labeled in alphabetical order from 1:20, then let $y_{1jk} = H_j - A_k$ $j < k$ be the score differential (Home-Away) when club j played at home versus club k. Further, let $y_{2jk} = A_j - H_k$ $j < k$ be the score differential (Away-Home) when club j played away versus club k. Then:

$$d_jk = y_{1jk} - y_{2jk} = (H_j - A_k) - (A_j - H_k) = (H_j + H_k) - (A_j + A_k)$$

That is, d_{jk} represents the total home versus away differential for the two matches played between clubs j and k. There are $\begin{pmatrix} 20 \\ 2 \end{pmatrix} = 190$ pairs of clubs. If there is no home field differential, then $\mu_D = 0$. Here we conduct a 2-tailed permutation test for a home field differential. There is overwhelming evidence of a home field advantage. None of the permutation means is close to the observed mean $\overline{d} = 0.6368$. A histogram of the randomization distribution and observed mean differential (vertical line) is given in Figure 5.7.

R Commands and Output

```
## Commands

epl2012 <- read.csv("http://www.stat.ufl.edu/~winner/data/epl_2012_home_perm.csv",
      header=T)
attach(epl2012); names(epl2012)

### Obtain Sample Size and Test Statistic (Average of d.jk)
(n <- length(d.jk))
(TS.obs <- mean(d.jk))

### Choose the number of samples and initialize TS, and set seed
N <- 9999; TS <- rep(0,N); set.seed(86420)

### Loop through samples and compute each TS
for (i in 1:N) {
ds.jk <- d.jk                       # Initialize d*.jk = d.jk
u <- runif(n)-0.5                   # Generate n U(-0.5,0.5)'s
u.s <- sign(u)                      # -1 if u.s < 0, +1 if u.s > 0
ds.jk <- u.s * ds.jk
TS[i] <- mean(ds.jk)                # Compute Test Statistic for this sample
}
summary(TS)

(num.exceed1 <- sum(TS >= TS.obs))   # Count for 1-sided (Upper Tail) P-value
(num.exceed2 <- sum(abs(TS) >= abs(TS.obs)))  # Count for 2-sided P-value
(p.val.1sided <- (num.exceed1 + 1)/(N+1))     # 1-sided p-value
(p.val.2sided <- (num.exceed2 + 1)/(N+1))     # 2-sided p-value

### Draw histogram of distribution of TS, with vertical line at TS.obs
hist(TS,breaks=seq(-.7,.7,.02), xlab="Mean Home-Away",
   main="Randomization Distribution for EPL 2012 Home Field Advantage")

## Output

> (n <- length(d.jk))
[1] 190
> (TS.obs <- mean(d.jk))
[1] 0.6368421
> (num.exceed1 <- sum(TS >= TS.obs))   # Count for 1-sided (Upper Tail) P-value
[1] 0
> (num.exceed2 <- sum(abs(TS) >= abs(TS.obs)))  # Count for 2-sided P-value
[1] 0
> (p.val.1sided <- (num.exceed1 + 1)/(N+1))     # 1-sided p-value
[1] 1e-04
> (p.val.2sided <- (num.exceed2 + 1)/(N+1))     # 2-sided p-value
[1] 1e-04
```

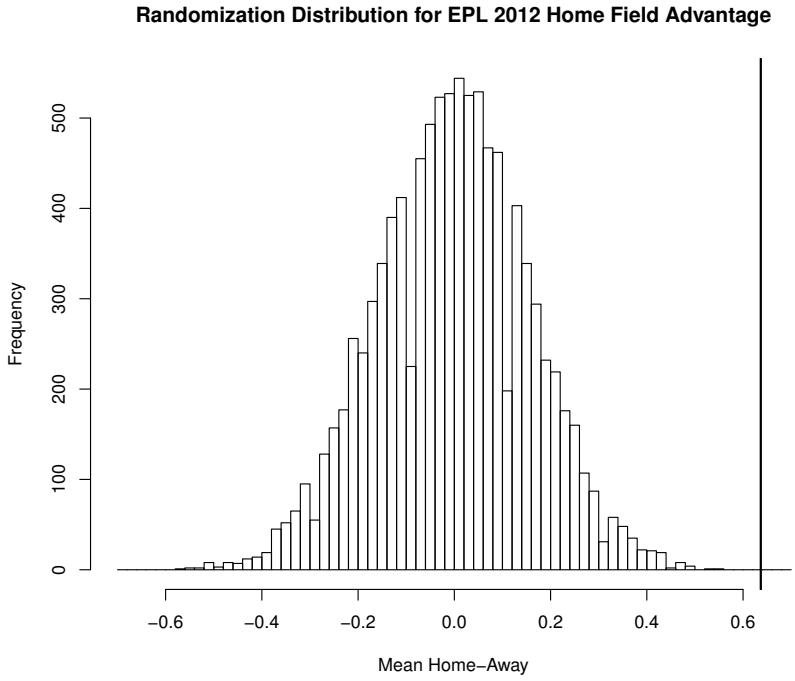

Figure 5.7: Randomization Distribution for Home-Away Mean Differential - EPL 2012

∇

Bibliography

[1] Agresti, A. (2002). *Categorical Data Analysis. 2nd Ed.* Wiley, New York.

[2] 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.

[3] 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.

[4] Cameron, A.C. and P.K. Trivedi (2005). *Microeconometrics: Methods and Applications.* Cambridge, Cambridge.

[5] Chambers, G.F. (1889). *Handbook of Astronomy, 4th Ed.* Oxford.

[6] Chang, P.-C., S.-Y. Chou, and K.-K. Shieh (2013). "Reading Performance and Visual Fatigue When Using Electronic Paper Displays in Long-Duration Reading Tasks Under Various Lighting Conditions," *Displays*, Vol. 34, pp:208-214.

[7] Chihara, L. and T. Hesterberg (2011). *Mathematical Statistics with Resampling and R.* Wiley, Hoboken, NJ.

[8] Clarke, R.D. (1946). "An Application of the Poisson Distribution," *Journal of the Institute of Actuaries*, Vol. 72, p. 481.

[9] 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.

[10] 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.

[11] Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

[12] Feller, W. (1968). *An Introduction to Probability Thory and Its Applications, Vol.1, 3rd. Ed.* Wiley, New York.

[13] 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.

[14] 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.

[15] Jeffery, F.R. and J. Stathis (1996). "Function Point Sizing: Structure, Validity, and Applicability," *Empirical Software Engineering*, Vol. 1, #1, pp. 11-30.

[16] Jorgensen, M., U. Indahl, D.Sjoberg (2003). "Software Effort Estimation by Analogy and "Regression to the Mean"," *The Journal of Systems and Software*, Vol. 68, pp. 253-262.

[17] Kahneman, D., P. Slovic, and A. Tversky (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, UK.

[18] 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.

[19] 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.

[20] 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.

[21] Ott, R.L. and M. Longnecker (2016). *Statistical Methods & Data Analysis, 7th Ed.* Cengage Learning, Boston.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Scheaffer, R.L., W. Mendenhall, and L. Ott (1990). *Elementary Survey Sampling, 4th Ed.* PWS-KENT, Boston.

[27] Sheldrake, R., P. Smart, and L. Avraamides (2015). "Automated Tests for Telephone Telepathy Using Mobile Phones," *Explore*, Vol. 11, #4, pp. 310-319.

[28] 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.

[29] Thorndike, F. (1926). "Applications of Poisson's Probability Summation," *Bell System Technical Journal*, Vol. 5, pp. 604-624.

[30] 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.

[31] Winner, L. (2006). "NASCAR Winston Cup Race Results for 1975-2003," *Journal of Statistical Education*, Vol. 14, #3.