Introduction to Probability and Statistics with Applications in Sports Betting

Larry Winner
University of Florida
Department of Statistics

November 15, 2022

Contents

Chapter 1

Introduction to Sports Betting

A wide variety of research in the fields of Economics, Finance, and Statistics has been applied to betting on sporting events. Two particular streams of research involve betting on horse racing and betting on team sports. For an excellent literature review, see Sauer (1998) [**?**]. Betting on horse racing involves betting the odds on which horse among a group racing will win the race. Various other bets can be made including place (finish first or second) and show (finish first, second, or third) among others. Betting on team sports typically involve odds (money line) for each team winning, point spreads allowing for differences in scoring abilities and Over/Under for total points scored by both teams. In this book, the focus will be on team sports although some examples involving horse racing will be included. The data are mostly obtained from the website www.covers.com with other sources being used to supplement when needed.

The point spread is a number of points that one team is favored to win by over another team. In Game 6 of the 2019 NBA Finals, the Golden State Warriors were favored over the Toronto Raptors by 2.5 points in the game. If the Warriors won by 3 or more points, a bet on the Warriors won. If the Raptors won, or lost by 2 or fewer points, a bet on the Raptors won. The Raptors won by 4, so a bet on them would have won, a bet on the Warriors would have lost. Note that there is no possibility of a tie with that point spread line. While in many newspapers, publications, and internet sites, the favorite is listed with a negative spread (Warriors -2.5, in this case), we will use the opposite convention as saying the favorite has a positive advantage as is typically (but not universally) done in the sports betting literature. This makes interpreting results of various statistical tests more intuitive.

The Over/Under is the total number of points for the two teams. In the same game, the Over/Under was 211.5 points. The final score was Raptors 114, Warriors 110. The total points were 224, so the Over bet would have won, and the Under bet would have lost. Note that there is no possibility of a tie on that Over/Under line.

The Moneyline (odds) makes adjustments to payoffs when betting on teams to win a game "straight up." This is widely used in lower scoring sports such as soccer, hockey, and baseball. In Game 7 of the 2019 World Series, the Money Line was Houston Astros -136, Washington Nationals +126. These numbers are typically (but not always) one positive, the other negative. When teams are very closely matched they can both be negative. The way to interpret these are as follow. You must bet $136 on Houston to win $100 if the Astros win the game. If you bet $100 on Washington to win, you will win $126 if the Nationals win. There are two interpretations of a "unit bet," the other way of looking at the bet on the favorite is: Bet $100

on Houston and win \$100(100/136)=\$73.53. This second interpretation makes more sense when considering "unit bets" (see Woodland and Woodland (1994) [**?**], Gandar, et al (2002) [**?**], and Berkowitz, Depken, and Gandar (2018) [**?**]). We will discuss the calculations of odds and their conversions to probabilities in the chapter on probability.

In Major League Baseball (MLB) and the National Hockey League (NHL), there are Over/Under lines combined with Money Lines. For instance, in the Atlanta Braves/New York Mets baseball game on July 24, 2020, the Over/Under is 7.5 runs with a Money Line of -115 for Over and -110 for Under. That is, a bet of \$115 wins \$100 for Over (\$100 bet wins \$86.96) and a bet of \$110 wins \$100 for Under bet (\$100 bet wins \$90.91). These details were first described (to my knowledge) among a series of a paper, comment, reply, and discussion in Brown and Abraham (2002) [**?**], Paul and Weinbach (2004) [**?**], Brown and Abraham (2004) [**?**], and Gandar and Zuber (2004a) [**?**]. The original paper did not include the Money Line regarding the Over/Under betting strategies and the comment pointed out the omission.

When betting through a legal sportsbook, the "11-10" rule is widely described in the sports betting literature (see e.g. Vergin and Scriabin (1978) [**?**]) to determine success probability needed for profitability. A bettor wages \$11 to win \$10. If the bettor's bet is successful, she wins back her \$11 bet, plus the \$10 from the sportsbook. If the bet is a "tie" (better known as "push" in betting parlance), she wins back her \$11 bet (gains \$0). If her bet is unsuccessful, she loses the \$11 she wagered. Let π be the probability she wins the bet and $1-\pi$ be the probability she loses the bet (ignoring the chance of a push, since in that case, there was virtually no bet made). Then for a betting strategy to be profitable, we need the expected return $E\{R\}$ to be positive. That is, we need the following result, which is often simply reported as .524 in the literature.

$$E\{R\} = \pi(10) + (1-\pi)(-11) = \pi(21) - 11 > 0 \quad \Rightarrow \quad \pi > \frac{11}{21} = .5238$$

Example 1.1: Examples of Betting Lines for NFL and MLB

Table **??** gives typical betting lines and results for a subset of a week of NFL games (four games from Week 6 of 2019 Regular Season). Generally there will be multiple sportsbooks included, each with their own adjustments on the lines. Only closing lines (at gametime) are included, but opening lines and line histories are often available as well. Table **??** gives typical betting lines and results for a day of MLB games (four games from Tuesday, July 16, 2019). NBA would be similar to the NFL and NHL would be similar to MLB. Note that negative spreads mean that the "home" team is favored, thus, we are still using the common convention in posting lines.

Note that the Carolina/Tampa Bay NFL game was played at a neutral site at the Tottenham Hotspur Stadium in London. The spread is still with respect to the "home" team Tampa Bay. For the Thursday night New York Giants @ New England game, New England won by 35-14=21 points, covering the 16.5 point spread, and the total points 35+14=49 exceeded the 43 point Over/Under line.

The LA Dodgers @ Philadelphia Phillies MLB game was won by the Phillies 9-8. As they were +180, a \$100 bet on Philadelphia would have paid out \$180. The total runs scored was 17, well exceeding the 10.0 Over/Under. As the Money Line for the Over bet was -105, a \$105 Over bet would have paid out \$100 (a \$100 bet would have paid out \$95.24).

∇

Day	Away Team	Home Team	Neutral	Home Spread	Over/Under	Away Score	Home Score
Thursday, Oct. 10	NYG	NE	0	-16.5	43	14	35
Sunday, Oct. 13	CAR	TB	1	+2	47.5	37	26
Sunday, Oct. 13	WAS	MIA	0	+6	42	17	16
Sunday, Oct. 13	PHI	MIN	0	-3.5	44.5	20	38

Table 1.1: Betting lines for National Football League Games

Day	Away Team	Home Team	Away ML	Home ML	Over/Under	Over ML	Under ML	Away Score	Home Score
Tuesday, July 16	LAD	PHI	-220	+180	10.0	-105	-116	8	9
Tuesday, July 16	WAS	BAL	-185	+170	11.5	-103	-120	8	1
Tuesday, July 16	TB	NYY	+130	-152	10.5	100	-120	3	8
Tuesday, July 16	DET	CLE	+210	-250	10.5	-115	-106	0	8

Table 1.2: Betting lines for Major League Baseball Games

Chapter 2

Describing Data

Once data have been collected, it is typically described via graphical and numeric methods. The methods used to describe the data will depend on its type (nominal, ordinal, or numeric).

Nominal variables have levels that are categorical with no inherent ordering. If we took a game at random from a season, and observed the home team for that game, that variable would be nominal.

Ordinal variables have levels that are categorical, but do have an inherent ordering. In soccer matches, the home team can Win, Draw (Tie), or Lose the match, which represents an ordinal outcome.

Numeric variables can be either **discrete** or **continuous**. A discrete variable may be the number of points scored by the home team in a game. A continuous variable may be the time for a horse to complete a race. When discrete random variables have many possible levels (such as points by a team in a pro basketball game), they are often treated as if they are continuous.

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. However, they can be difficult to read when there are many groups and/or low frequency groups.

Bar charts are used to describe nominal or ordinal data. The variable levels are arrayed on the bottom

(or left side) of the plot and bars above (or beside) the levels represent the frequency or relative frequency of the number of observations belonging to the various categories.

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

Density plots are used for numeric variables. They represent smoothed histograms describing the location of the measurements.

Boxplots are used for numeric variables. They identify particular percentiles of a distribution and are useful in detecting outlying observations and spread in the distribution. A variation on this theme are violin plots, which vary in width depending on the density of the data.

Example 2.1: Major League Baseball Over/Under Bet Outcomes - 2015-2019

Of 12152 MLB regular season games played during 2015-2019, the "Over" bet won 5759 (.4739), the "Under" won 5819 (.4789), and there were 574 (.0472) "Pushes." Figure **??** gives a Bar Chart representing the frequencies of the three possible outcomes. Clearly there is no tendency for the "Over" or "Under" to occur more frequently than the other over this period. The percentages by season are given in Figure **??**. There is very little fluctuation over the years.

∇

Example 2.2: Women's NBA Home Team Spread Differentials - 2010-2019 - Histogram and Density Plot

For 2037 WNBA games played during the 2010-2019 regular seasons, we obtain the home team's spread differential. This is obtained my taking the difference between the home and away teams' score and subtracting the home team's "advantage." Note that this is the same as taking the difference and adding the home team's printed line. Positive values imply that the home team covered the spread, while negative values imply the away team covered, and values of 0, imply a "Push." Interestingly, the median (defined below) is 0. A histogram and a smooth density plot are given in Figure **??**. The histogram is approximately symmetric and mound-shaped. A scaled normal distribution with mean 0 and standard deviation equal to that of the differentials is super-imposed.

∇

Example 2.3: Women's NBA Over/Under Differentials - 2010-2019 - Boxplot

Figure **??** gives a boxplot for the WNBA Over/Under differentials for the 2010-2019 regular seasons, where the differential is defined as the difference between total points scored and the Over/Under betting line. Figure **??** displays side-by-side box plots by season. Note the outliers at the bottom and top of each plot. The extreme positive differentials tend to be games that went into overtime. The extreme negative differentials represent games where one or both teams under-performed offensively. The side-by-side boxplots show that the distribution is fairly consistent over the ten seasons, and that the median hovers around 0.

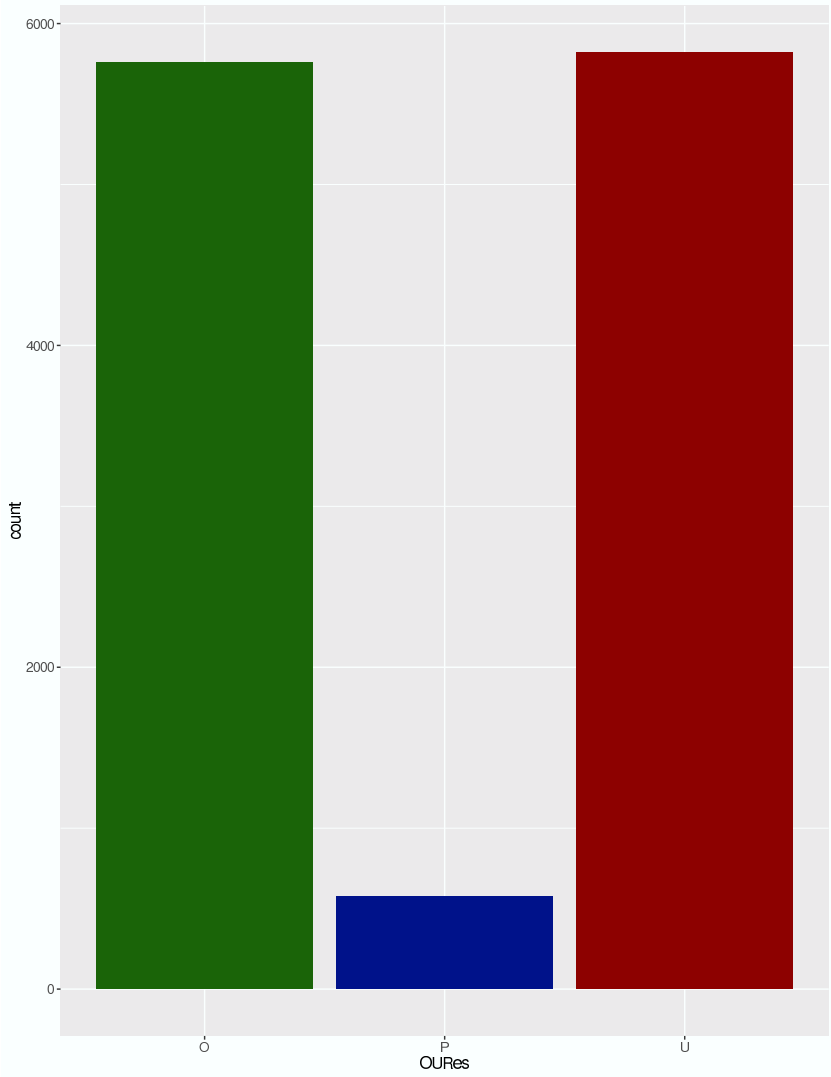


Figure 2.1: Bar Chart of Over/Under MLB results for 2015-2019
box

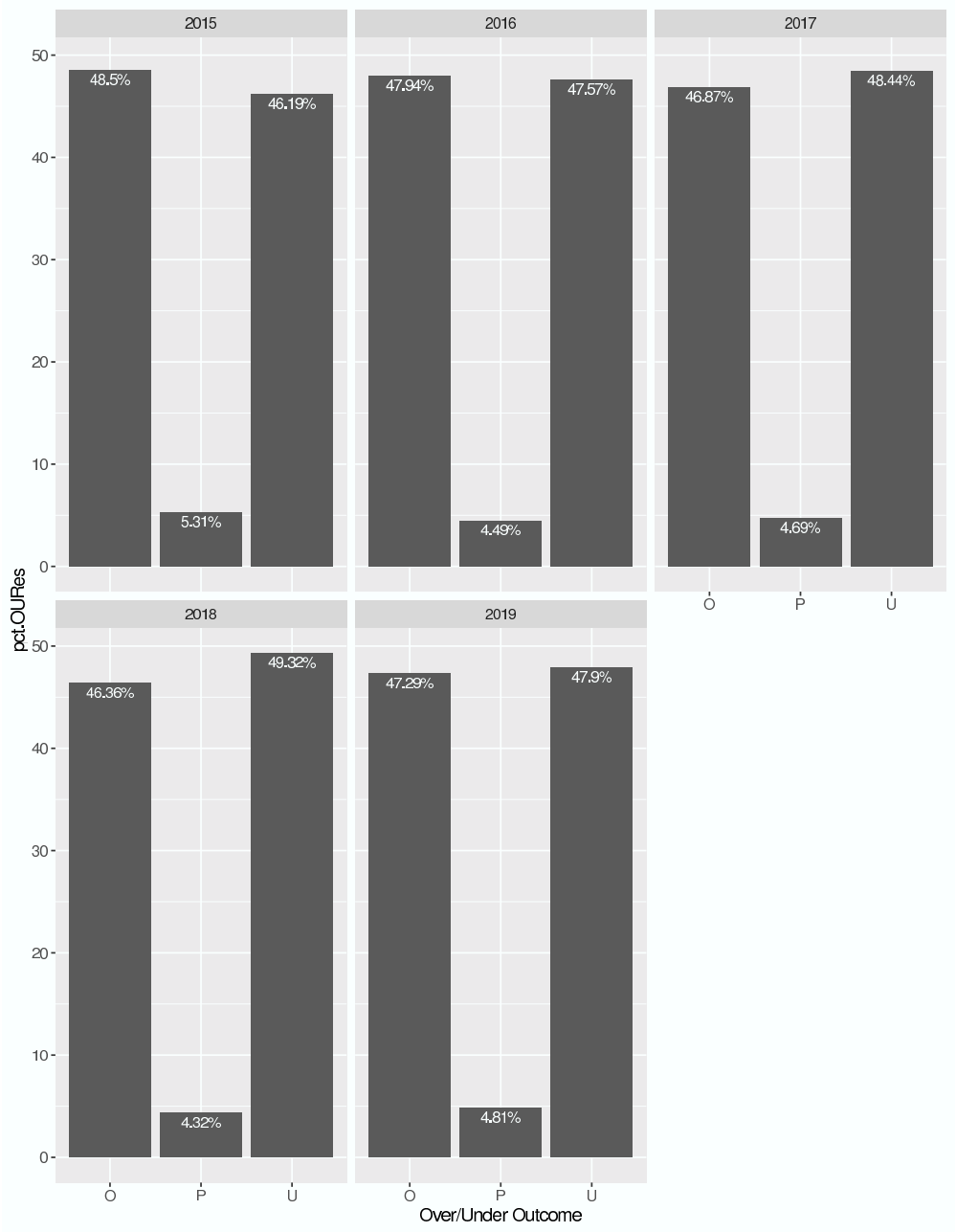


Figure 2.2: Bar Chart of Over/Under MLB results by season for 2015-2019

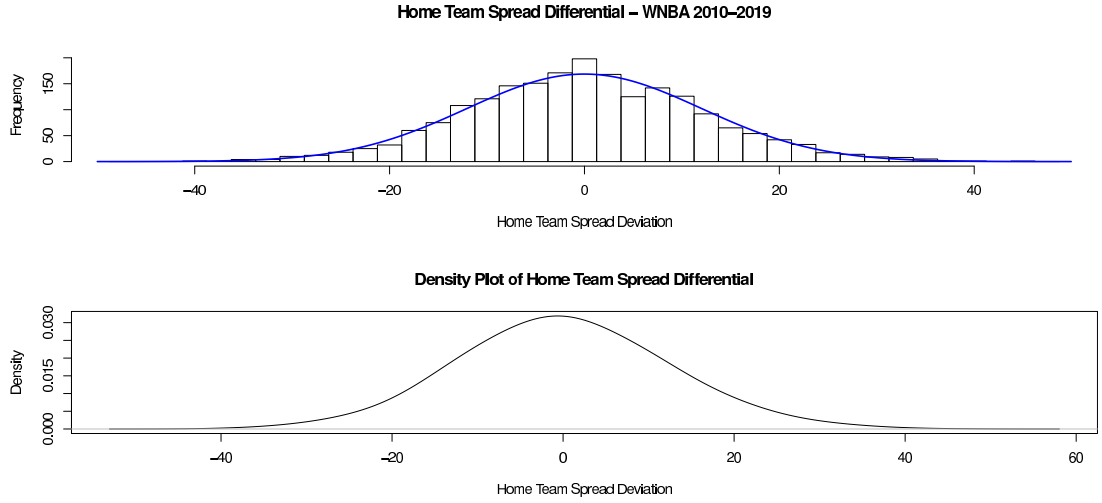


Figure 2.3: Histogram and Smooth Density of Home Team Spread Differentials for WNBA for 2010-2019

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

1. Minimum: Bottom of line at bottom of plot or lowest circle below the line if present.
2. Range for lowest 25% of measurements: Line extends to either the Minimum or 1.5 times the distance between 75th and 25th percentiles (height of the box), whichever is lowest. Circles represent outlying measurements.
3. 25th percentile: Bottom line of box
4. Range for the 25th to 50th percent of measurements: Between bottom of box and second horizontal line
5. Median (50th percentile): Second horizontal line
6. Range for the 50th to 75th percent of measurements: Between second horizontal line and top of box
7. 75th percentile: Top line of the box
8. Range for 75th to 100th percent of measurements: Line extends to either the Maximum or 1.5 times the distance between 75th and 25th percentiles (height of the box), whichever is lowest. Circles represent outlying measurements.

A smooth version of a boxplot, which does not separate the measurements into quantiles is a **violin plot**. For the Over/Under differential data, one is displayed in Figure **??**.

∇

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

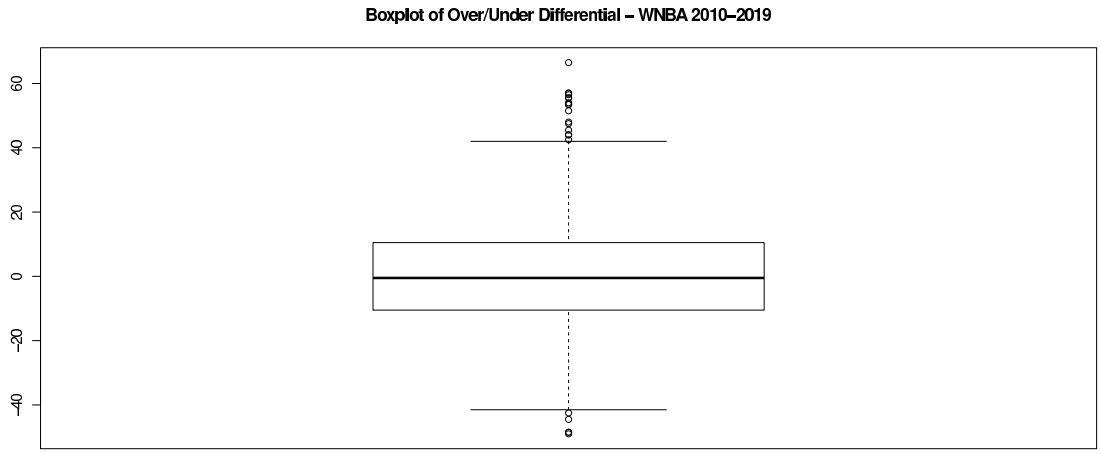


Figure 2.4: Boxplot of Over/Under Differential for WNBA 2010-2019

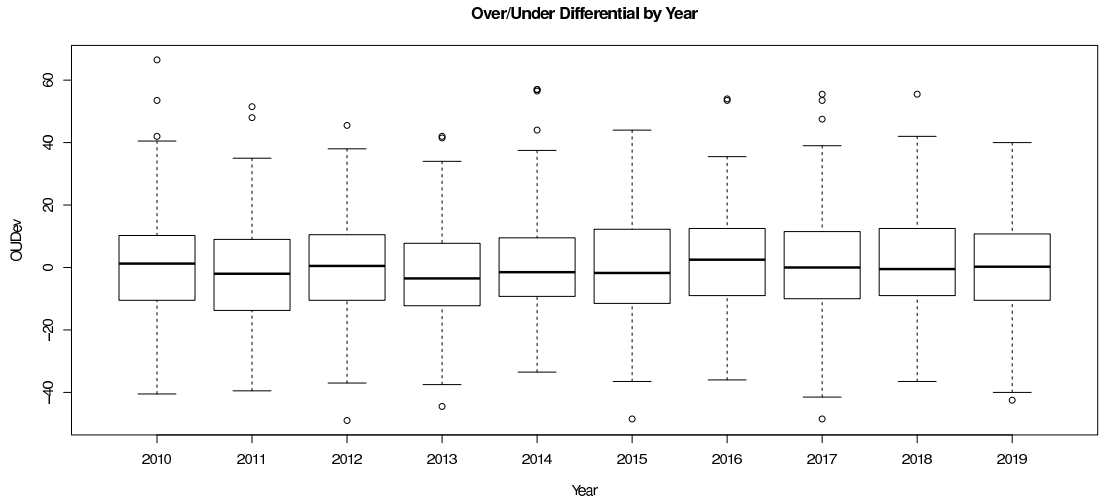


Figure 2.5: Side-by-side Boxplots of Over/Under Differential by year for WNBA 2010-2019

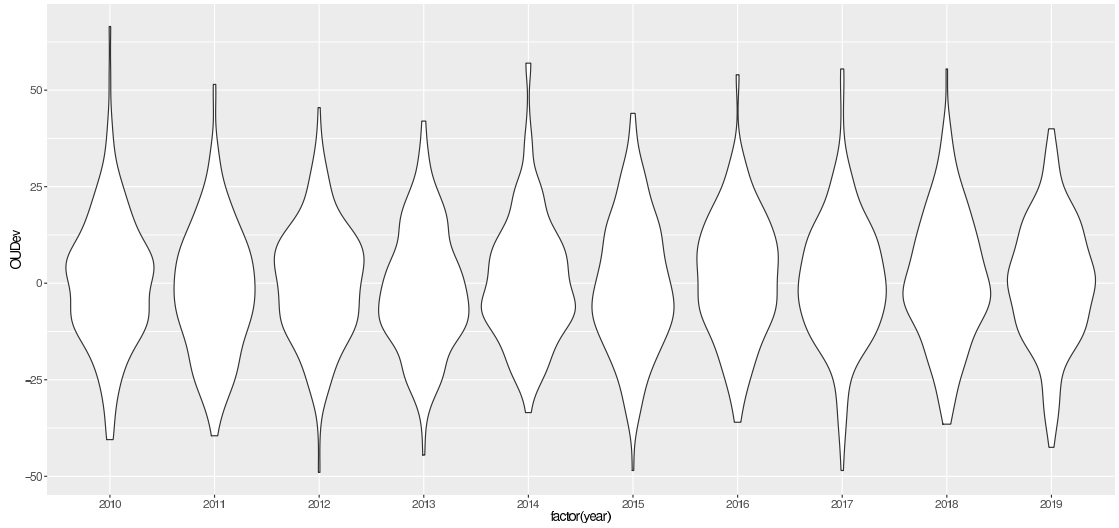


Figure 2.6: Violin plot of Over/Under Differential by year for WNBA 2010-2019

can be based on virtually any scale from nanoseconds to millenia. They can be used to detect trend and cyclical patterns over time. Figure **??** shows the weekly average spread deviations for NFL home teams by week, separately for the 2010-2019 seasons. The way the response is measured is to take the home team spread (again, using the convention that the spread is positive for the home team when it is favored) and subtracting it from the home team's point differential in the final score. For instance, if the home team's spread is $+7$ and it wins the game by the score $24 - 17$, then home spread deviation is $(24 - 17) - (+7) = 0$ and the game is a push. Of interest is whether betting lines become more accurate over the course of the seasons (closing in on 0). Another measure (described in detail below) is the standard deviation, a measure of the spread in a set of measurements. A plot of the weekly standard deviations is given in Figure **??**.

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.

Example 2.4: NFL Home Team Point Spread Differential

For the NFL 2006-2019 seasons, there were 3584 games played. The home team covered the spread in 1690 of the games, lost against the spread in 1797 games, and pushed in 97 of the games. The counts and proportions are given below.

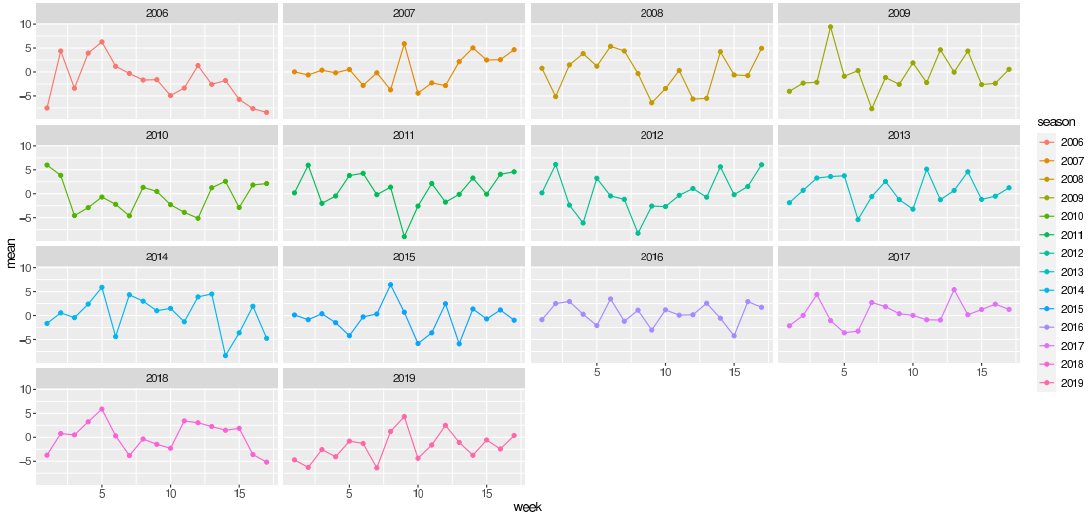


Figure 2.7: NFL 2006-2019 weekly average home spread deviations

Figure 2.8: NFL 2006-2019 weekly standard deviations of home spread deviations

R Commands and Output are given below. The table function counts the number of cases that are of each category, and dividing by their sum converts them to proportions.

R Commands and Output

```
### R Commands/Output (using previous dataset)
table(nfl_home$home_cover1)
table(nfl_home$home_cover1) / sum(table(nfl_home$home_cover1))

Lose Push  Win
1797   97 1690

      Lose       Push        Win
0.50139509 0.02706473 0.47154018
```

∇

2.2.1 Measures of Central Tendency

There are two commonly reported measures of central tendency, or location for a set of measurements. The **mean** is the sum of all measurements divided by the number of measurements, and is reported often as "per capita" in economic reports. The mean is the "balance point" of a set of measurements in a physical sense. The **median** is the point where half of the measurements fall at or below it, and half of the measurements fall at or above it. It is also the 50th percentile of the set of measurements. Many economic reports state median values. A third, less reported measure is the **mode** which really is only appropriate for discrete variables, and is the value that occurs most often. 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 2.5: MLB Over/Under Lines

For the 2015-2019 Major League Baseball seasons, there were a total of $N = 12152$ games with an Over/Under total of 103983.5 runs. The (population) mean of the Over/Under lines is $\mu = 103983.5/12152 =$

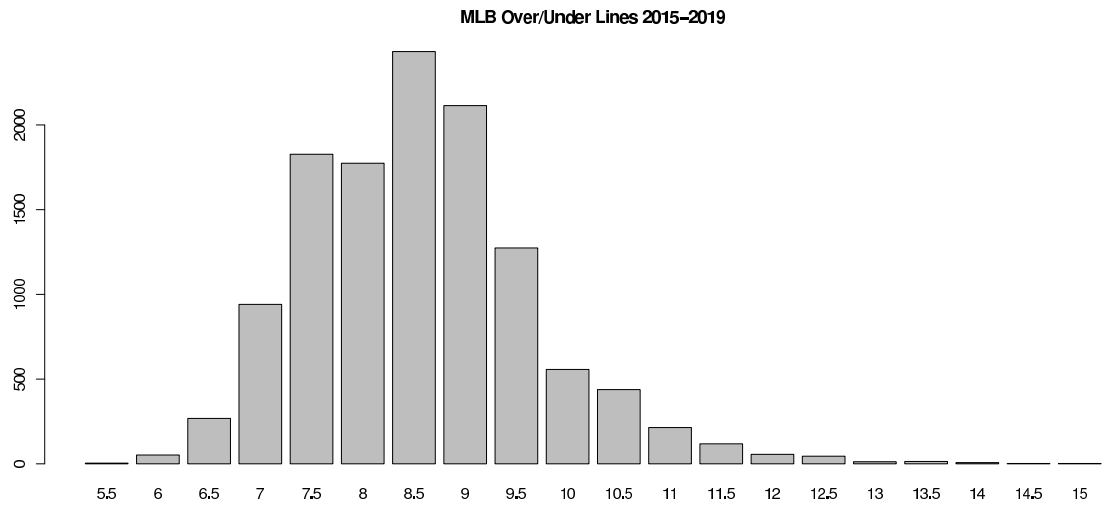


Figure 2.9: MLB Over/Under Lines 2015-2019

8.557 runs per game. The median is 8.5 runs per game. The results are given below, using functions in R, as well as a bar chart of the Over/Under lines in Figure **??**.

R Commands and Output

```
sum(mlb_home$OU)
length(mlb_home$OU)
mean(mlb_home$OU)
median(mlb_home$OU)
table(mlb_home$OU)

[1] 103983.5
[1] 12152
[1] 8.556904
[1] 8.5

 5.5    6  6.5    7  7.5    8  8.5    9  9.5   10 10.5   11 11.5   12 12.5
   4   52  268  941 1827 1774 2433 2114 1274  557  438  214  118   56   45
  13 13.5   14 14.5   15
  12   14    7    2    2
```

Note that the mean (8.557) and median (8.5) are very close, as would be expected for a symmetric distribution, although this distribution is somewhat skewed to the right.

∇

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 largest 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 computing 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 all possible samples, the "average" of the sample variance will be the population variance. The standard deviation is the positive square root of the variance and is in the same units as the measurements. The population standard deviation is denoted as σ, the sample standard deviation is denoted as s. For many (but certainly not all) distributions, approximately 2/3 of the measurements lie within one standard deviation of the mean and approximately 19/20 lie within two standard deviations of the mean.

Example 2.6: MLB Over/Under Lines and Total Run Outcomes

We compute the range, interquartile range, variance, and standard deviations for the MLB Over/Under lines and game total runs (outcome). 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. Bar charts of the Over/Under lines and total runs are given in Figure **??**. Note that the means and medians are similar (slightly higher for the actual total runs than the lines) but the variation is much higher for the actual total runs.

R Commands and Output

```
> mean_MLB <- rbind(mean(OU), mean(totRuns))
> median_MLB <- rbind(median(OU), median(totRuns))
> range_MLB <- rbind((max(OU)-min(OU)), (max(totRuns)-min(totRuns)))
> IQR_MLB <- rbind((quantile(OU,.75)-quantile(OU,.25)),
+                  (quantile(totRuns,.75)-quantile(totRuns,.25)))
> var_MLB <- rbind((N.MLB-1)*var(OU)/N.MLB,(N.MLB-1)*var(totRuns)/N.MLB)
> sd_MLB <- sqrt(var_MLB)
```

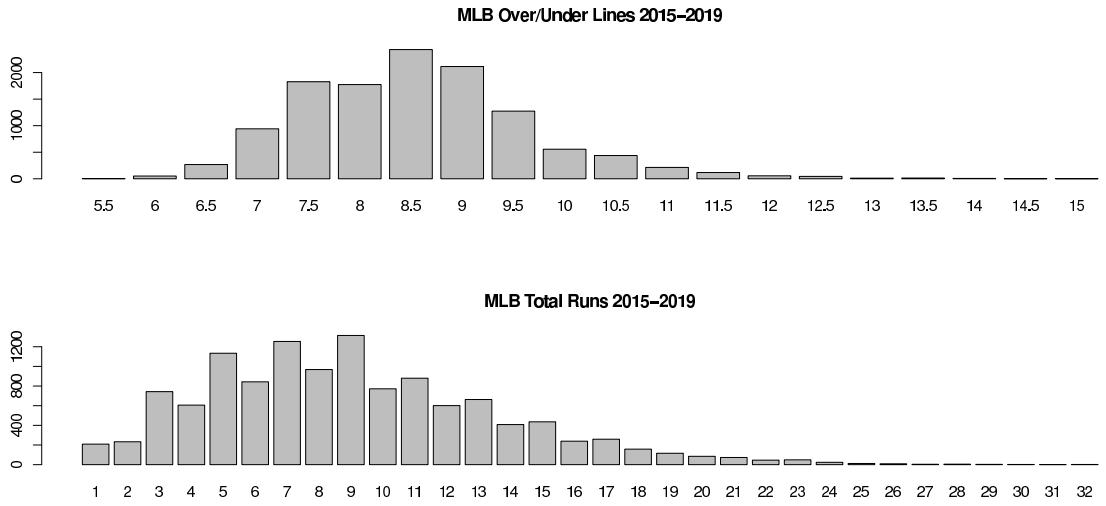


Figure 2.10: MLB Over/Under lines and total runs 2015-2019

```
> MLB.out1 <- cbind(mean_MLB, median_MLB, range_MLB, IQR_MLB, var_MLB, sd_MLB)
> colnames(MLB.out1) <- c("mean", "median", "range", "IQR", "variance", "std dev")
> rownames(MLB.out1) <- c("Over/Under Line", "Total Runs")
> round(MLB.out1, 4)
                 mean median range IQR variance std dev
Over/Under Line 8.5569    8.5   9.5 1.5   1.2637  1.1241
Total Runs      9.0638    9.0  31.0 6.0  20.8377  4.5648
```

∇

Two other measures of variation are the following. The **median absolute deviation** (MAD), which is the median of the absolute deviations from the median, 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).

Example 2.7: MLB Over/Under Lines and Total Run Outcomes

Here MAD and CV are computed for the both the Over/Under lines and the total runs. Note that the MAD for the total runs, when divided by 0.645 is 4.651, while $\sigma = 4.565$, so they are very consistent. This is not so for the Over/Under lines. Also, the coefficient of variation is much higher for actual total runs than for the O/U lines.

R Commands and Output

```
> MAD_MLB <- rbind(median(abs(OU-median(OU))),
+                  median(abs(totRuns-median(totRuns))))
> MAD_645_MLB <- MAD_MLB / 0.645
> CV_MLB <- sd_MLB / mean_MLB
>
> MLB.out2 <- cbind(MAD_MLB, CV_MLB, MAD_645_MLB, sd_MLB)
> colnames(MLB.out2) <- c("MAD", "CV", "MAD/0.645", "std dev")
```

```
> rownames(MLB.out2) <- c("Over/Under Line", "Total Runs")
> round(MLB.out2, 4)
                MAD     CV MAD/0.645 std dev
Over/Under Line 0.5 0.1314    0.7752  1.1241
Total Runs      3.0 0.5036    4.6512  4.5648
```

∇

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. Note that technically the kurtosis is the first part in the following formulas, when we subtract the 3, we have the "excess kurtosis" which should be around 0 if the data are normally distributed.

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

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

Example 2.8: MLB Over/Under Lines and Total Run Outcomes

Skewness and kurtosis are computed for the two variables here.

R Commands and Output

```
> mu3_MLB <- rbind(sum((OU-mean(OU))^3)/N.MLB,
+                  sum((teamRuns-mean(teamRuns))^3)/N.MLB)
> skew_MLB <- mu3_MLB / sd_MLB^3
> mu4_MLB <- rbind(sum((OU-mean(OU))^4)/N.MLB,
+                  sum((teamRuns-mean(teamRuns))^4)/N.MLB)
> kurt_MLB <- (mu4_MLB / sd_MLB^4) - 3
> MLB.out3 <- cbind(mu3_MLB, skew_MLB, mu4_MLB, kurt_MLB)
> colnames(MLB.out3) <- c("mu3", "skewness", "mu4", "kurtosis")
> rownames(MLB.out3) <- c("Over/Under Line", "Total Runs")
> round(MLB.out3, 4)
                  mu3 skewness      mu4 kurtosis
Over/Under Line  1.0541   0.7421   7.0313   1.4033
Total Runs      32.7771   0.3446 458.4699  -1.9441
```

		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

Skewness is positive for both variables, and is higher for the O/U lines than the total runs (presumably more "mass" in the upper tail). The kurtosis is positive for O/U lines and negative for total runs, these are difficult to interpret too closely.

∇

2.3 Describing More than One Variable

So far, we have looked at cases one variable at a time. 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 are based on results of a **contingency table** also known as a **crosstabulation**. These are tables where rows are the levels of one categorical variable, columns are levels of another variable, and numbers within the table are counts of the number of units falling in that cell (combination of variable levels). Often these are converted into proportions either overall (cell probabilities sum to 1), or within rows or columns. A contingency table is typically of the form in Table **??**.

Example 2.9: WNBA Home Team Spread and Over/Under Outcomes

For the WNBA 2010-2019 seasons, the outcomes of how the home team fares against the spread and the Over are ordinal with levels "Lose", "Push", and "Win". Group and Stacked bar charts for the proportions of games falling in the 9 combinations of Home/Over outcomes are given in Figure **??**.

R Commands and Output

```
> OUDiff = teamPts + oppPts - OU
> OverRes = ifelse(OUDiff > 0, "Win", ifelse(OUDiff < 0, "Lose", "Push"))
```

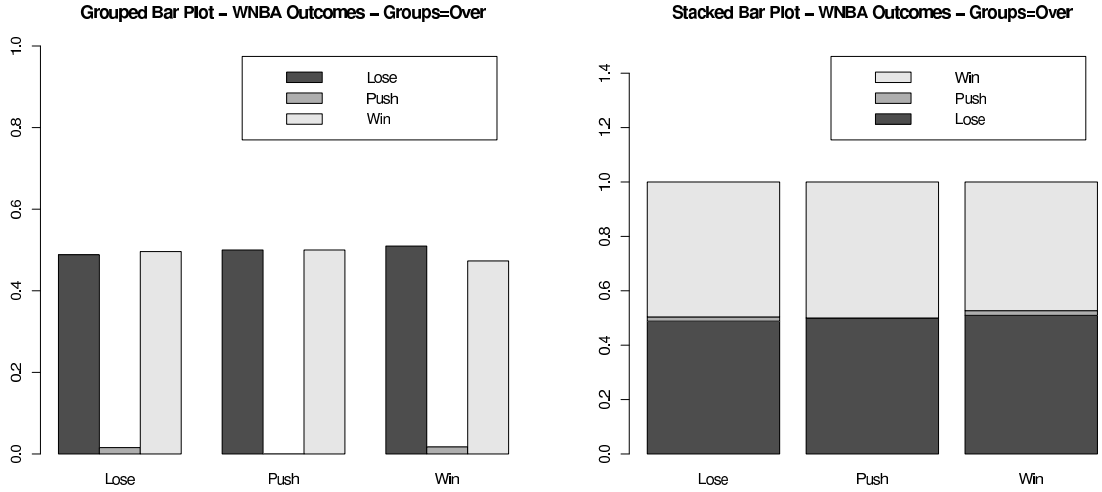


Figure 2.11: Outcomes for Over/Home wagers for WNBA 2010-2019 seasons

```
> HomeDiff = (teamPts - oppPts) - teamSprd
> HomeRes = ifelse(HomeDiff > 0, "Win", ifelse(HomeDiff < 0, "Lose", "Push"))
>
> ## Obtain Table of Counts (Row=Over, Column=Home)
> (OverHome <- table(OverRes,HomeRes))
       HomeRes
OverRes Lose Push Win
   Lose  501   16 509
   Push   14    0  14
   Win   502   17 466
> ## Obtain Row (1) and Column (2) Marginal Totals
> margin.table(OverHome,1)
OverRes
Lose Push  Win
1026   28  985
> margin.table(OverHome,2)
HomeRes
Lose Push  Win
1017   33  989
> ## Obtain Proportions across all Cells
> OverHome/sum(OverHome)
       HomeRes
OverRes        Lose        Push         Win
   Lose 0.245708681 0.007846984 0.249632173
   Push 0.006866111 0.000000000 0.006866111
   Win  0.246199117 0.008337420 0.228543404
> ## Obtain Row Proportions (Home w/in Over)
> prop.table(OverHome,1)
       HomeRes
OverRes       Lose       Push        Win
   Lose 0.48830409 0.01559454 0.49610136
   Push 0.50000000 0.00000000 0.50000000
   Win  0.50964467 0.01725888 0.47309645
> ## Obtain Column Proportions (Over w/in Home)
> prop.table(OverHome,2)
       HomeRes
OverRes       Lose       Push        Win
   Lose 0.49262537 0.48484848 0.51466127
   Push 0.01376598 0.00000000 0.01415571
   Win  0.49360865 0.51515152 0.47118301
```

∇

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

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

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

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: Relation Between Total Points and Over/Under Line - NFL 2010-2019

The correlation between the actual total score and the Over/Under line is $r = .2952$. There is a positive, but not particularly strong relation between the two variables. The formula for the ordinary least squares regression is given below. For this data, we obtain $\hat{y} = 1.8481 + 0.9682x$. Thus, for a game with an Over/Under line of $x = 40$ points, the predicted total score is 1.8481+0.9682(40)=40.58 points. Later, we consider tests of the regression coefficients for an "efficient market."

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

A plot of the data and the fitted regression equation are given in Figure **??**.

R Commands and Output

```
totPts <- teamPts + oppPts
cor(totPts, OU)
```

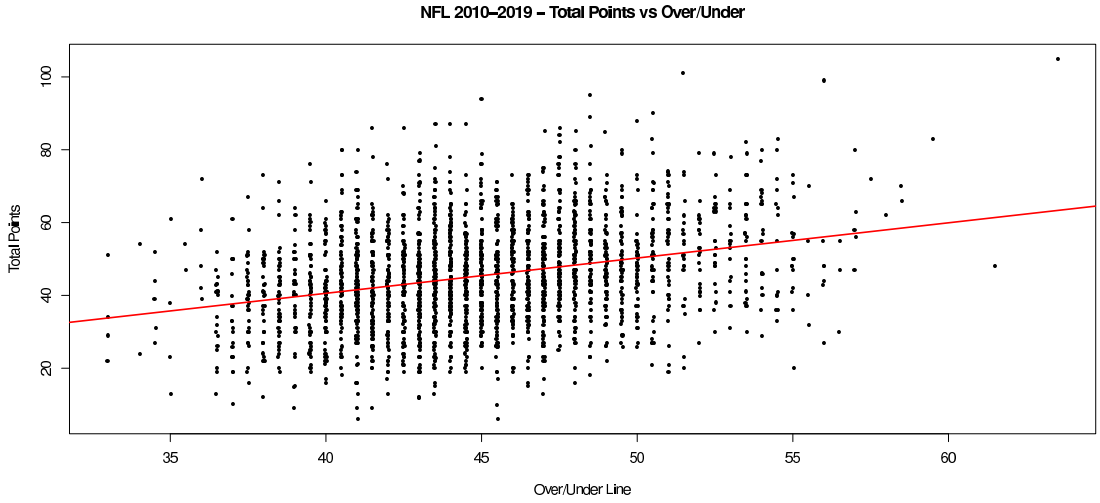


Figure 2.12: Plot of Total Points versus Over/Under Line - NFL 2010-2019

```
OU.mod1 <- lm(totPts ~ OU)
summary(OU.mod1)

plot(jitter(totPts,0.3) ~ jitter(OU,0.3), pch=16, cex=.5,
    xlab="Over/Under Line", ylab="Total Points",
    main="NFL 2010-2019 - Total Points vs Over/Under")
abline(OU.mod1,col="red", lwd=2)

> cor(totPts, OU)
[1] 0.2951634
>
> OU.mod1 <- lm(totPts ~ OU)
> summary(OU.mod1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.84812    2.79255   0.662    0.508
OU           0.96818    0.06197  15.625   <2e-16 ***


Residual standard error: 13.31 on 2558 degrees of freedom
Multiple R-squared:  0.08712,   Adjusted R-squared:  0.08676
F-statistic: 244.1 on 1 and 2558 DF,  p-value: < 2.2e-16
```

∇

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: WNBA 2010-2019 - Point Spread, Over/Under Line and Game Outcomes

Here we consider the following four variables for the WNBA 2010-2109 seasons: Home team spread (positive means they are favored to win), Over/Under Line, Home team point differential (Home-Visitor), and total points scored in the game. A scatterplot matrix is given in Figure **??** and the pairwise correlations

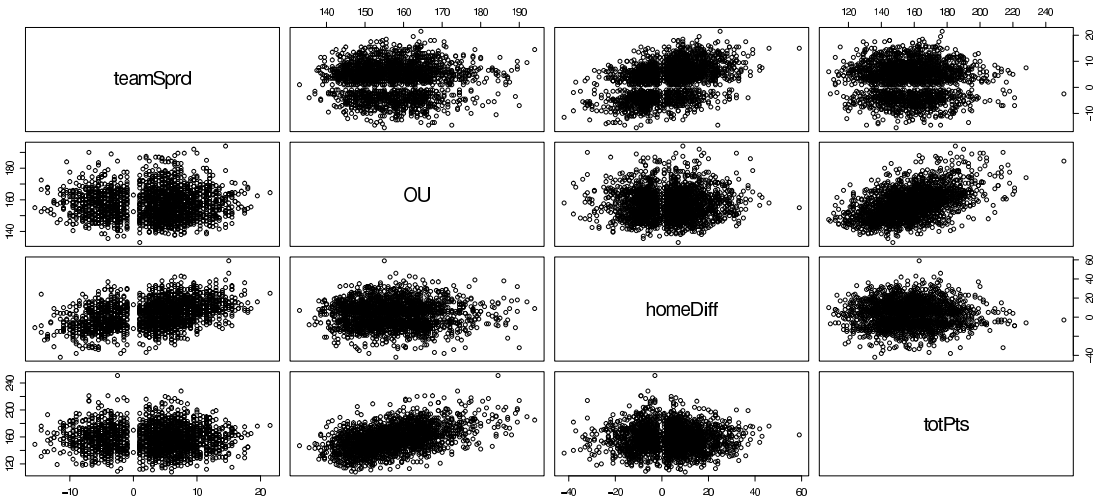


Figure 2.13: Scatterplot matrix of Home team spread, Over/Under line, Home team point differential and Total points - WNBA 2010-2019

are given below. The correlations between Home team spread and home team point differential (.4277) and between Over/Under Line and Total points (.4973) are the only values of any reasonable magnitude.

R Commands and Output

```
homeDiff <- teamPts - oppPts
totPts <- teamPts + oppPts

odds.df <- data.frame(teamSprd, OU, homeDiff, totPts)
plot(odds.df)
cor(odds.df)
```

```
> cor(odds.df)
             teamSprd          OU     homeDiff      totPts
teamSprd  1.000000000 0.001365649  0.427666583 -0.02335489
OU        0.001365649 1.000000000  0.001845519  0.49727595
homeDiff  0.427666583 0.001845519  1.000000000 -0.01109277
totPts   -0.023354890 0.497275950 -0.011092769  1.00000000
```

∇

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

Many series (particularly when measured over time) display **spurious correlations**, particularly when both variables tend to increase or decrease together with no **causal** reason that the two (or more) variables

move in tandem. For instance, the correlation between annual U.S. internet users (per 100 people) and electrical power consumption (kWh per capita) for the years 1994-2010 is .7821 (data source: The World Bank). Presumably increasing internet usage isn't leading to large increases in electrical consumption, or vice versa.

Chapter 3

Probability

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

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." The set of all possible outcomes is referred to as the **sample space**. 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.

	Over Result			
Home Spread Result	Lose	Push	Win	Total
Lose	3861 (.2452)	94 (.0060)	3900 (.2476)	7855 (.4988)
Push	128 (.0081)	5 (.0003)	135 (.0086)	268 (.0170)
Win	3849 (.2444)	96 (.0061)	3681 (.2337)	7626 (.4842)
Total	7838 (.4977)	195 (.0124)	7716 (.4899)	15749 (1.0000)

Table 3.1: Counts (proportions) of home team covering spread and Over winning - NBA 2006/07-2018/19 seasons

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) [**?**], for a large collection of research on the topic. Pre-game probabilities of teams or individuals winning a sporting event would best be thought of as subjective probabilities based on recent performances and when based on closing lines, also the wagers made by the betting public.

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

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

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

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

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

Example 3.1: NBA Bets, Horse Racing

For a professional basketball game, the home (or away) team can be bet based on the point spread as well as the over/under. Each of these wagers can end in one of three possible outcomes (Lose, Push, Win). Thus, there are $3(3) = 9$ possible outcomes. Based on the 15749 NBA regular season games during the 2006/07-2018/19 regular seasons, the frequencies (proportions) based on betting on the home team and the over are given in Table **??**.

A trifecta bet in horse racing involves selecting the horses that will finish first, second, and third in the correct order. In a race with 12 horses starting, there are $P_3^{12} = \frac{12!}{(12-3)!} = 12(11)(10) = 1320$ ways in which the horses can finish in the first (win), second (place), and third (show) positions, where order matters.

A quinella bet involves betting on the two horses that will finish first and second place, where order does not matter. In this case, again with 12 horses racing, there are $C_2^{12} = \frac{12!}{2!(12-2)!} = 12(11)/2 = 66$ possible combinations that can occur.

∇

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 **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 occurred is referred to as its **conditional** probability. If the unconditional probability of A and its conditional probability given B has occurred are equal, then the events A and B are said to be **independent**. The rules for obtaining conditional probabilities (assuming $P(A) > 0$ and $P(B) > 0$) are given below, as well as probabilities under independence.

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

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

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: NBA Home Team and Over Bet Spread Outcomes

Referring back to Table **??**, we define A to be the event that the home team covers the point spread and event B that the Over bet wins for this "population."

$$P(A) = \frac{7626}{15749} = .4842 \qquad P(B) = \frac{7716}{15749} = .4899 \qquad P(AB) = \frac{3681}{15749} = .2337$$

$$P(A \cup B) = .4842 + .4899 - .2337 = .7364$$

$$P(A|B) = \frac{.2337}{.4899} = \frac{3681}{7716} = .4771 \qquad P\left(A|\overline{B}\right) = \frac{3849+96}{7838+195} = .4911 \qquad P(B|A) = \frac{.2337}{.4842} = \frac{3681}{7626} = .4827$$

Note that the event that the Home team covers the spread (A) is not independent of whether the Over bet wins (B). However, there are only small differences among $P(A)$, $P(A|B)$, and $P\left(A|\overline{B}\right)$. Similarly, $P(B)$ and $P(B|A)$ are very close as well.

∇

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.

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

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

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

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

Example 3.3: WNBA Over/Under Results

Consider the WNBA Over/Under bets during the 2010-2019 regular seasons. There were 2039 games played over those 10 seasons, with 28 pushes, which we remove (where the total points equal the Over/Under line). We consider the 2011 remaining games. The partition is based on grouping the games based on their game of the season (based on the home team) into three portions of the 34 game seasons. Thus, A_1 is the event that games are 1-11, A_2 is that games are 12-23, and A_3 is that games are 24-34. Note that there will be more games in A_2 as it has 12 positions, while A_1 and A_3 each have 11. Each game falls into exactly one category. We define C as the event that the total points scored in the game exceeds the Over/Under line, that is that the "Over" bet wins. The counts and proportions (probabilities) are given in Table **??**.

Week Grp (A_i)	Over Wins (C)	Over Loses ($\overline{C}$)	Total Games	$P(A_i)$	$P(C\|A_i)$	$P(A_i \cap C)$	$P(A_i\|C)$
1-11 (A_1)	339	316	655	.3257	.5176	.1686	.3442
12-23 (A_2)	333	375	708	.3521	.4703	.1656	.3381
24-34 (A_3)	313	335	648	.3222	.4830	.1556	.3177
Total	985	1026	2011	1.0000	N/A	.4898	1.0000

Table 3.2: Counts and probabilities of Week Groups and Over/Under results and Bayes' Rule calculations

Note that the "prior" probabilities of the week groups and the conditional probability that the Over/Under bet wins, given week group are computed as follows. Then this leads to the computations of the updated "posterior" probabilities of the week groups, given knowledge of Over/Under bet winning.

$$P(A_1) = \frac{655}{2011} = .3257 \qquad P(C|A_1) = \frac{339}{655} = .5176 \qquad P(A_1 \cap C) = .3257(.5176) = .1686$$
$$P(A_2) = \frac{708}{2011} = .3521 \qquad P(C|A_2) = \frac{333}{708} = .4703 \qquad P(A_2 \cap C) = .3521(.4703) = .1656$$
$$P(A_3) = \frac{648}{2011} = .3222 \qquad P(C|A_3) = \frac{313}{648} = .4830 \qquad P(A_3 \cap C) = .3222(.4830) = .1556$$
$$P(C) = .1686 + .1656 + .1556 = .4898$$
$$P(A_1|C) = \frac{.1686}{.4898} = .3442 \qquad P(A_2|C) = \frac{.1656}{.4898} = .3381 \qquad P(A_3|C) = \frac{.1556}{.4898} = .3177$$

These calculations can be obtained much simpler based on the contingency table, however in many practical settings the relative base rates $P(A_i)$ and conditional probabilities $P(C|A_i)$ are reported as direct numbers and not counts (although they can be converted in such a way).

The primary result in terms of betting strategy is that (over the course of this data) the Over bet tends to win with the highest frequency early in the season, lowest frequency in the middle, and moderate frequency at the end. Of interest would be if this occurs regularly within individual seasons. We begin with a base rate (prior probability) for early season as $P(A_1) = .3257$ and update it with knowledge that the game won the Over bet to a posterior probability of $P(A_1|C) = .3442$.

∇

3.2 Random Variables and Probability Distributions

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

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

3.3 Discrete Random Variables

Discrete random variables can take on a finite, or countably infinite, set of outcomes. We label the random variable as Y, and its specific outcomes as $y_1, y_2, \ldots, y_k$. Note that in some 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.

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

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

Example 3.4: NBA Point Spread Bets

Suppose a gambler wishes to bet on a NBA team to cover the point spread in a particular game. Further, suppose that the point spread has a .5 in it (e.g. the team is favored by 7.5 in the game), which rules out a push for the bet. Then, if the bettor is facing a bookmaker using the 11-10 rule described in Chapter 1, he will win 100 units if the team wins and will win -110 if the team loses. These are the only two possibilities, say $y_1 = 100$ and $y_2 = -110$. Then the probabilities are $p(y_1)$ and $p(y_2) = 1 - p(y_1)$.

Suppose (for some random reason), he chose a strategy where he always bet on the home favorite to win, when the home team is favored by $\{1.5, 2.5, \ldots, 9.5, 10.5\}$. During the 2006/7-2018/19 NBA regular seasons, there were 4722 such games, of which the home favorite covered in 2403 of the games and failed to cover in 2319 of them. For this "population" of games, $p(y_1) = 2403/4722 = .5089$ and $p(y_2) = 2319/4722 = .4911$.

∇

Population Numerical Descriptive Measures

Three widely used numerical descriptive measures corresponding to a population are the **population mean**, μ, the **population variance**, σ^2, and the **population standard deviation**, σ. While we have previously covered these based on a population of individual 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 3.5: NBA Point Spread Bets - 2006/07-2018/19

For the NBA point spread betting strategy given in Example 3.4, we have the following mean and variance if we repeatedly sampled games and their betting outcomes from this probability distribution.

$$\mu_Y = \sum_y yp(y) = 100(.5089) + (-110)(.4911) = -3.13$$

$$\sigma_Y^2 = \sum_y (y - \mu_Y)^2 p(y) = \sum_y y^2 p(y) - \mu_Y^2 = 11031.31 - (-3.13)^2 = 11021.5131$$

$$\sigma_Y = +\sqrt{11021.5131} = 104.98$$

A sample of 10000 games is taken from this population (equivalently done by taking 10000 integers between 1 and 4722 WITH replacement), and observing the payout for each game (100 if Home team covered, -110 if not). Then the mean and standard deviation of those numbers are computed.

R Commands and Output

```
## Commands/Output
> payout <- ifelse(teamCover == 1, 100, ifelse(teamCover == -1, -110, 0))
> n.games <- length(payout)
> set.seed(12345)
> sample.game <- sample(1:n.games, size=10000, replace = TRUE)
>
> mean(payout[sample.game])
[1] -3.362
> sd(payout[sample.game])
[1] 104.9925
```

Note that the mean of the 10000 sampled games is close to the population mean (-3.36 vs -3.13) and sample standard deviation is close to the corresponding population value (104.99 vs 104.98). If a different (or no) seed had been used, the samples, and thus their means and standard deviations would change as well although the population values remain the same for this population of games.

∇

3.3.1 Converting Money Lines and Decimal Odds to Subjective Probabilities

The (subjective) probability that the home team wins a game can be obtained from the Money Lines and/or Decimal Odds for the two teams in the game. There are various ways to convert them, for an excellent description see Berkowitz, Depken, and Gandar (2017, [**?**]). Two methods when there are two teams playing are **basic** and **Shin** normalizations. We will describe them through Major League Baseball Money Lines, in particular the 2019 World Series Game 7 with Houston and Washington described in Chapter 1 and reviewed here.

Example 3.6: MLB Favored Team Outcomes

When the Money Line (ML) is positive, if a bet of 100 is placed on the team and it wins, the bettor wins ML, and she wins -100 if the team loses. If $ML < -100$, then if a bet of 100 is placed on the team and it wins, the bettor wins $100(100/|ML|)$ and he wins -100 if the team loses. These are referred to as "unit bets" as in each case, the bettor risks 100 units. Going back to Game 7 of the 2019 World Series described in Chapter 1, The Houston Astros had a Money Line of -136 and the Washington Nationals' Money Line was 126. Thus, if you wagered 100 on Houston and they won, you would win $100(100/|-136|) = 73.53$ and if they lost, you would win -100. If Washington was bet and they won, the bettor would win 126 and if they lost, the bettor would win -100.

Let W be the random amount won on a bet. Let η_F represent the predicted (subjective) probability the favored team wins the game and η_U be the predicted probability the underdog wins. Note that we will use η to represent predicted probabilities based on Money Lines and Decimal Odds, and use π to represent actual observed probabilities throughout these notes.

The expected winnings, $E\{W\}$ are obtained as follow, depending on whether the ML is positive or negative, based on a 100 unit wager. The goal is to solve for the expected winnings to be 0 (a fair bet).

$$ML > 100 \quad \Rightarrow \quad E\{W\} = \eta_U(ML) + (1-\eta_U)(-100) = 0 \quad \Rightarrow \quad \eta_U = \frac{100}{100+ML}$$

$$ML < -100 \quad \Rightarrow \quad E\{W\} = \eta_F(100(100/|ML|)) + (1-\eta_F)(-100) = 0 \quad \Rightarrow \quad \eta_F = \frac{100}{100+100(100/|ML|)}$$

For the Houston/Washington World Series Game 7, this translates as follows.

$$\text{Washington: } \eta_W = \frac{100}{100+126} = 0.4425 \qquad \text{Houston: } \eta_H = \frac{100}{100+73.53} = .5763$$

Note that these η values sum to more than 1, representing the "bookie commission." Basic normalization simply divides each of these η values by their sum, which then are scaled to sum to 1.

$$\eta_F^* = \frac{\eta_F}{\eta_F+\eta_U} \qquad \eta_U^* = \frac{\eta_U}{\eta_F+\eta_U} \qquad \eta_F^* + \eta_U^* = 1$$

$$\text{Houston (F): } \eta_H^* = \frac{.5763}{.5763+.4425} = \frac{.5763}{1.0188} = .5657 \qquad \text{Washington (U): } \eta_W^* = \frac{.4425}{1.0188} = .4343$$

Berkowitz, Depken, and Gandar (2017, [**?**]), reporting derivations from Gandar et al (2002, [**?**]) and Gandar et al (2004, [**?**]) give this result in one step in terms of $\beta_F = |ML_F|/100$ and $\beta_U = ML_U/100$. This is Conversion Method 4 in Berkowitz, Depken, and Gandar (2017, [**?**]). It ends up giving the same subjective probabilities of favorite and underdog winning as basic normalization.

$$\eta_F^* = \frac{\beta_F + \beta_F\beta_U}{2\beta_F + \beta_F\beta_U + 1} \qquad \eta_U^* = \frac{1+\beta_F}{2\beta_F + \beta_F\beta_U + 1}$$

This makes use of the notion that the bookmaker would like to have equivalent net revenues, regardless of who wins the game. Let W_F and W_U be the number of units wagered on the favorite and underdog,

respectively through the bookmaker for a total "handle" of $W_F + W_U$. If the favorite wins, her revenue ("hold") is $W_U - W_F/\beta_F$; if the underdog wins, she earns $W_F - W_U\beta_U$. If the bookmaker chooses to balance her revenue so that her revenue is the same, regardless who wins, she has the following results.

$$W_U - W_F/\beta_F = W_F - W_U\beta_U \quad \Rightarrow \quad W_U(1+\beta_U) = W_F(1+1/\beta_F) \quad \Rightarrow \quad \frac{W_F}{W_U} = \frac{1+\beta_U}{1+1/\beta_F}$$

$$\Rightarrow \quad W_U = \frac{W_F(1+1/\beta_F)}{(1+\beta_U)}$$

Then her "commission" (c) defined as hold divided by handle is given below, as well as the probabilities that equate expected revenues for bettors to the negative commission, where R_F and R_U represent revenues for betting on the favorite and underdog, respectively. Again let η_F^* and η_U^* be the unknown subjective probabilities of the favorite and underdog winning, respectively.

$$c = \frac{W_U - W_F/\beta_F}{W_F + W_U} = \frac{W_F - W_U\beta_U}{W_F + W_U} = \frac{\frac{W_F(1+1/\beta_F)}{(1+\beta_U)} - W_F/\beta_F}{W_F\left(1 + \frac{(1+1/\beta_F)}{(1+\beta_U)}\right)} = \frac{\frac{(1+1/\beta_F)-(1+\beta_U)/\beta_F}{1+\beta_U}}{\frac{(1+\beta_U)+(1+1/\beta_F)}{1+\beta_U}} =$$

$$= \frac{1-\beta_U/\beta_F}{2+\beta_U+1/\beta_F} = \frac{\beta_F-\beta_U}{2\beta_F+\beta_F\beta_U+1}$$

$$E\{R_F\} = \eta_F^*(1/\beta_F) + (1-\eta_F^*)(-1) = -c = -\frac{\beta_F-\beta_U}{2\beta_F+\beta_F\beta_U+1} \quad \Rightarrow \quad \eta_F^*(1+1/\beta_F) - 1 = -\frac{\beta_F-\beta_U}{2\beta_F+\beta_F\beta_U+1}$$

$$\Rightarrow \quad \eta_F^*\left(\frac{1+\beta_F}{\beta_F}\right) = \frac{2\beta_F+\beta_F\beta_U+1-\beta_F+\beta_U}{2\beta_F+\beta_F\beta_U+1} = \frac{\beta_F+\beta_U+\beta_F\beta_U+1}{2\beta_F+\beta_F\beta_U+1} = \frac{\beta_F(1+\beta_U)+(\beta_U+1)}{2\beta_F+\beta_F\beta_U+1} =$$

$$\frac{(1+\beta_F)(1+\beta_U)}{2\beta_F+\beta_F\beta_U+1} \quad \Rightarrow \quad \pi_F^* = \frac{\beta_F(1+\beta_U)}{2\beta_F+\beta_F\beta_U+1} = \frac{\beta_F+\beta_F\beta_U}{2\beta_F+\beta_F\beta_U+1}$$

$$\Rightarrow \quad \eta_U^* = 1-\eta_F^* = \frac{(2\beta_F+\beta_F\beta_U+1)-(\beta_F+\beta_F\beta_U)}{2\beta_F+\beta_F\beta_U+1} = \frac{\beta_F+1}{2\beta_F+\beta_F\beta_U+1}$$

For the 2018 World Series Game 7, we obtain the following values with $\beta_F = 1.36$ and $\beta_U = 1.26$.

$$c = \frac{1.36-1.26}{2(1.36)+1.36(1.26)+1} = \frac{0.10}{5.4336} = 0.0184$$

$$\eta_F^* = \frac{1.36+1.36(1.26)}{2(1.36)+1.36(1.26)+1} = \frac{3.0736}{5.4336} = .5657 \qquad \eta_U^* = 1-.5657 = .4343$$

As stated previously, this method and basic normalization provide the same subjective probabilities.

Related to the method of basic normalization, in many betting markets, **decimal odds** are used for wagering. In these cases, the decimal odds are always greater than 1 and can be obtained directly from the Money Line. When a team has $ML < 100$ as in the favorite in most games and both teams in very evenly matched games, the decimal odds o are $o_F = 1 + 1/\beta_F$ and if the underdog has $ML > 100$, $o_U = 1 + \beta_U$. If both teams have negative Money Lines, $o = 1 + 1/\beta$ for each team. The unstandardized probabilities η_F

and η_U are $1/o_F$ and $1/o_U$, respectively. Thus, in the 2018 World Series Game 7, with Money Lines -136 for Houston and 126 for Washington, we obtain the following decimal odds.

$$\text{Houston (F): } ML_H = -136 \quad \beta_H = 1.36 \quad o_H = 1 + 1/1.36 = 1.7353 \quad \eta_H = 1/1.7353 = .5763$$

$$\text{Washington (U): } ML_W = 126 \quad \beta_W = 1.26 \quad o_W = 1 + 1.26 = 2.26 \quad \eta_W = 1/2.26 = .4425$$

In much of the betting literature the sum $\eta_F + \eta_U$ is referred to as the **booksum** (see e.g. Strumbelj (2014), [**?**]). This notion can extend to events such as horse races with $n > 2$ competitors where the decimal odds for the i^{th} entrant is o_i, unstandardized probability $\eta_i = 1/o_i$ and booksum equal to $\eta_1 + \cdots + \eta_n$. A method to convert decimal odds to standardized probabilities (Shin (1993), [**?**]) uses the notion that a fraction, z, of traders (bettors in gambling) have "inside information," and bookmakers wish to estimate z and exploit the other bettors to make up for losses to the insiders.

The general case of $n > 2$ contestants involves an iterative fixed-point computing algorithm with $z = 0$ set as a starting value (Strumbelj (2014), [**?**]). However, the author shows that in the case of $n = 2$ contestants in a contest, z can be obtained in closed form and then used to obtain the standardized η_F^* and η_U^* from z and the unstandardized probabilities $\eta_F = 1/o_F$ and $\eta_U = 1/o_U$.

The Shin standardized probabilities are obtained as follows for the case with $n = 2$.

$$\eta_i^* = \sqrt{\frac{z^2}{4(1-z)^2} + \frac{\eta_i^2}{(1-z)(\eta_F + \eta_U)}} - \frac{z}{2(1-z)} \qquad i = F, U \quad 0 \le z \le 1$$

$$\text{where: } z = \frac{[(\eta_F + \eta_U) - 1]\left[(\eta_F - \eta_U)^2 - (\eta_F + \eta_U)\right]}{(\eta_F + \eta_U)\left[(\eta_F - \eta_U)^2 - 1\right]}$$

For the 2018 World Series Game 7, we obtain the following values with $\beta_F = 1.36$ and $\beta_U = 1.26$.

$$z = \frac{[(.5763 + .4425) - 1]\left[(.5763 - .4425)^2 - (.5763 + .4425)\right]}{(.5763 + .4425)\left[(.5763 - .4425)^2 - 1\right]} = \frac{.0188(-1.0009)}{1.0188(-.9821)} = \frac{-0.01882}{-1.00056} = .01881$$

$$\eta_F^* = \sqrt{\frac{.01881^2}{4(1-.01881)^2} + \frac{.5763^2}{(1-.01881)(.5763+.4425)}} - \frac{.01881}{2(1-.01881)} = \sqrt{.00009 + .33224} - .00959 =$$

$$= .57648 - .00959 = .5669 \qquad \eta_U^* = 1 - .5669 = .4331$$

Finally, a comparison of basic normalization and Shin's method is made by comparing the distributions of predicted probabilities of the winning teams as well as a table of agreement between favorite/winning teams for all 2430 MLB 2017 regular season games (of which 2397 games had a favorite, that is different Money Lines). Basic normalization and Shin's method are always in agreement on whether the home team or visitor is favored based on whether $\eta_H^* > 0.5$ or $\eta_H^* < 0.5$ (where η_H^* is the home team's subjective probability of winning). The home team has the higher η^* for $1604/2397 = .6692$ of the games. However, the home

team actually won in only 1295/2397 = .5403 of the games. Thus, the Money Lines tend to give too much of an advantage to the home team. Much debate in the baseball betting literature has been given to the "Favorite/Longshot bias (FLB)," going back to the seminal paper on the topic (Woodland and Woodland (1994), [**?**]). However, that paper took into account the magnitudes of the pre-game subjective probabilities, not just Favorite/Underdog.

Among the games the home team won, the average subjective probabilities of winning were .5509 for basic and .5519 for Shin normalization. Among the games the home team lost, their average subjective probabilities were .5189 for basic and .5193 for Shin normalization. The correlation between the basic and Shin probabilities across the 2397 games is .9999998. They give very similar results (at least when there are two contenders in the match).

Table **??** gives a cross-tabulation of games where the home team was favored ($\eta_H^* > 0.5$) or underdog ($\eta_H^* < 0.5$) (rows) and whether the home team won the game or not (columns). Some interesting conditional probabilities are obtained below where F implies the home team was favored and W implies the home team won he game. The table can easily be used to obtain similar quantities for visiting teams, since a home team being favored means the visiting team was underdog; and a home team winning means the visiting team lost.

$$P(W|F) = \frac{939}{1604} = \frac{.3917}{.6692} = .5854 \qquad P(F|W) = \frac{939}{1295} = \frac{.3917}{.5403} = .7251$$

$$P(W|\overline{F}) = \frac{356}{793} = \frac{.1485}{.3308} = .4489 \qquad P(F|\overline{W}) = \frac{665}{1102} = \frac{.2774}{.4597} = .6035$$

The last probability is probably the strangest, conditioning that the home team lost the game, there's a 60.35% chance it was favored. Again, we are not taking into account the magnitude of the Money Line and thus η_H^*. That will be considered when logistic regression is covered. The "overall accuracy" of the Money Line can be obtained from the proportions of games where the home team was both favored and won, and where the home team was both underdog and lost.

$$\text{Overall Accuracy: } P(F \cap W) + P(\overline{F} \cap \overline{W}) = \frac{939 + 437}{2397} = \frac{1376}{2397} = .3917 + .1823 = .5741$$

These last calculations are very similar to those made in diagnostic testing for disease or other characteristics. If the home team winning (W) is considered the characteristic/disease is present and the home team being favorite (F) is considered the diagnostic test result, then $P(F|W)$ represents the **sensitivity** of the test, and $P(\overline{F}|\overline{W})$ represents the **specificity** of the test. Further, $P(W|F)$ represents the **positive predictive value (PPV)** and $P(\overline{W}|\overline{F})$ the **negative predictive value (NPV)** of the test. Further, $P(W)$ represents the **prevalence** of the characteristic/disease. Just to complete the analogy, we have the following results (where these calculations can be thought of as special cases of Bayes' Rule).

$$\text{Prevalence: } P(W) = .5403 \quad \text{Sens: } P(W|F) = .5854 \quad \text{Spec: } P(\overline{F}|\overline{W}) = 1 - .6035 = .3905$$

$$\text{PPV: } P(W|F) = .5854 \quad \text{NPV: } P(\overline{W}|\overline{F}) = 1 - .4489 = .5511$$

∇

Home Team	Home Wins	Home Loses	Total
Favorite ($\eta_H^* > 0.5$)	939 (.3917)	665 (.2774)	1604 (.6692)
Underdog ($\eta_H^* < 0.5$)	356 (.1485)	437 (.1823)	793 (.3308)
Total	1295 (.5403)	1102 (.4597)	2397 (1.0000)

Table 3.3: Counts and joint/marginal probabilities of Home Team being Favorite/Underdog and Winning/Losing Game

3.3.2 Linear Functions of Random Variables

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

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

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

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

Example 3.7: NBA Point Spread Bets - Dollars vs Euros

In Examples 3.4 and 3.5 we considered making 100 unit bets on home favorites with spreads of $\{1.5, 2.5, \ldots, 9.5, 10.5\}$, based on the 11-10 rule. Based on that strategy, if we had bet \$110 on every such game, winning \$100 whenever the team covered the spread and losing \$110 when they did not, we found that $\mu_Y = -\$3.11$, $\sigma_Y^2 = 11021.43$ and $\sigma_Y = \$104.98$. As of July 22, 2021 @ 5:51PM UTC, the value of 1 dollar was 0.85 euros. Suppose instead of betting 110 dollars, 1) we bet 110 euros, or 2) we bet the equivalent of 110 dollars in euros.

What are the mean and standard deviations of these two bets. Let $X1$ be the random outcome (in dollars) for case 1, and $X2$ be the random outcome (in euros) for case 2. Note here that 1 dollar = 0.85 euros means that 1 euro = 1/0.85 = 1.176 dollars.

$$\text{Case 1: 110 euros} = 129.36 \text{ dollars} : X1 = 1.2936Y$$

$$\Rightarrow \quad \mu_{X1} = 1.2936(-3.11) = -4.02 \quad \sigma_{X1} = |1.2936|(104.98) = 135.80$$

$$\text{Case 2: 110 dollars} = 93.5 \text{ euros} : X2 = 0.935Y$$

$$\Rightarrow \quad \mu_{X2} = 0.935(-3.11) = -2.91 \quad \sigma_{X2} = |0.935|(104.98) = 98.16$$

Thus the effect of increasing the bet to 110 euros (to win 100) has the effect of increasing bet to 129.36 dollars (to win 117.6). When converting units from dollars to euros, the bet of 110 dollars (to win 100) has the effect of "reducing" the bet to 93.5 euros (to win 85). Note that in the first case the bet has been increased, and in the second we have simply changed units.

∇

3.3.3 Multivariate Random Variables

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 $a_1, \ldots, a_n$, $E\{Y_i\} = \mu_i$, $V\{Y_i\} = \sigma_i^2$, and $\text{COV}\{Y_i, Y_j\} = \sigma_{ij}$. In the definition below, we define $P(Y_i = y_i, Y_j = y_j) = p(y_i, y_j)$ as the joint probabilities of the outcomes y_i and y_j for the random variables Y_i and Y_j, respectively (calculations are shown in Example 3.8).

$$\text{COV}\{Y_i, Y_j\} = \sigma_{ij} = E\{(Y_i - \mu_i)(Y_j - \mu_j)\} = \sum_{y_i}\sum_{y_j}(y_i - \mu_i)(y_j - \mu_j)p(y_i, y_j) = \sum_{y_i}\sum_{y_j} y_i y_j p(y_i, y_j) - \mu_i\mu_j$$

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

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

If the Y_i are independent, we have some nice properties involving their distributions and expected values of functions of them (when the expectations exist).

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = p(y_1, \ldots, y_n) = p_1(y_1)\cdots p_n(y_n) = \prod_{i=1}^{n} p_i(y_i)$$

$$E\{g_1(Y_1)\cdots g_n(Y_n)\} = E\{g_1(Y_1)\}\cdots E\{g_n(Y_n)\} = \prod_{i=1}^{n} E\{g_i(Y_i)\}$$

In general, we have the following results (whether the random variables are independent or not).

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

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

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

In financial portfolio analysis, it is often considered to try to optimize a portfolio consisting of two or more securities. Typically, you would like to maximize the expected gain or reduce the variance (volatility) of the portfolio. Suppose there are two stocks with random returns X and Y, with mean, variance, and correlation structure given above. Suppose we set up the portfolio as given below and wish to minimize the variance of its return.

$$P = wX + (1 - w)Y \qquad 0 \leq w \leq 1 \qquad E\{P\} = \mu_P = w\mu_X + (1 - w)\mu_Y$$
$$V\{P\} = \sigma_P^2 = w^2\sigma_X^2 + (1 - w)^2\sigma_Y^2 + 2w(1 - w)\sigma_{XY} = w^2\sigma_X^2 + (1 - w)^2\sigma_Y^2 + 2w(1 - w)\rho_{XY}\sigma_X\sigma_Y =$$
$$w^2\left(\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y\right) - 2w\left(\sigma_Y^2 - \rho_{XY}\sigma_X\sigma_Y\right) + \sigma_Y^2$$

Taking the derivative of the portfolio variance with respect to the weight assigned to stock X (w), and setting equal to 0 leads to the optimal weights for the two stocks (in terms of minimizing its variance).

$$\frac{d\sigma_P^2}{dw} = 2w\left(\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y\right) - 2\left(\sigma_Y^2 - \rho_{XY}\sigma_X\sigma_Y\right) = 0 \quad \Rightarrow \quad w^* = \frac{\sigma_Y^2 - \rho_{XY}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y}$$

Technically, to show that this is a minimum, we need to show that the second derivative is positive.

$$\frac{d^2\sigma_P^2}{dw^2} = 2\left(\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y\right)$$

If $\rho_{XY} < 1$, this will be strictly positive. If $\rho_{XY} = 1$, it is $2\left(\sigma_X - \sigma_Y\right)^2$ and the best would be to choose the stock with the smaller variance. If they have the same variance, the portfolio variance is the same for all portfolios.

Example 3.8: NBA Spread/OU Betting Results - 2006/07-2018/19

Consider a betting strategy of always betting against the spread on the home (away) team and on the over (under) among 15749 NBA games played over the regular seasons 2006/07-2018/19. Assume the bettor is always wagering 110 based on the 11/10 rule. Pushes are included with payouts of 0 for spread and/or Over/Under.

	Over Line			
Home Spread	Lose ($y_O = -110$)	Push ($y_O = 0$)	Win ($y_O = 100$)	Total
Lose ($x_H = -110$)	3864 (.24535)	94 (.00597)	3897 (.24744)	7855 (.49876)
Push ($x_H = 0$)	126 (.00800)	5 (.00032)	135 (.00857)	266 (.01689)
Win ($x_H = 100$)	3851 (.24452)	97 (.00616)	3680 (.23367)	7628 (.48435)
Total	7841 (.49787)	196 (.01245)	7712 (.48968)	15749 (1.0000)

Table 3.4: Counts and joint/marginal probabilities of Home Team spread and Over line bet outcomes

We consider X_H to be the payoff when betting on the home team and X_A when betting on the away team. Let Y_O be the payoff when betting Over and Y_U be the payoff when betting Under. The joint and marginal counts and probabilities for the random variables X_H and Y_O are given in Table **??**. Note that when the home team covers, $x_H = 100$, when the away team covers, $x_H = -110$, and if the game is a push, $x_H = 0$. A similar definition is applied to y_O outcomes.

$$E\{X_H\} = \mu_{X_H} = (-110)(.49876) + 0(.01689) + 100(.48435) = -6.42 \quad E\{X_H^2\} = 10878.50$$

$$\Rightarrow \quad \sigma^2_{X_H} = 10878.50 - (-6.42)^2 = 10837.28 \quad \sigma_{X_H} = \sqrt{10837.28} = 104.10$$

$$E\{Y_O\} = \mu_{Y_O} = (-110)(.49787) + 0(.01245) + 100(.48968) = -5.80 \quad E\{Y_O^2\} = 10921.03$$

$$\Rightarrow \quad \sigma^2_{Y_O} = 10921.03 - (-5.80)^2 = 10887.39 \quad \sigma_{Y_O} = \sqrt{10887.39} = 104.34$$

$$E\{X_H Y_O\} = (-110)(-110)(.24535) + \cdots + 100(100)(.23367) = -106.13 \quad \sigma_{X_H Y_O} = -106.13 - (-6.42)(-5.80) = -143.36$$

$$\rho_{X_H Y_O} = \frac{-143.36}{104.10(104.34)} = -0.0132$$

We obtain the population mean, variance, and standard deviations of the sum of Home and Over bets $(X_H + Y_O)$ and the difference between Home and Over bets $(X_H - Y_O)$.

$$E\{X_H + Y_O\} = (-6.42) + (-5.80) = -12.22 \quad V\{X_H + Y_O\} = 10837.28 + 10887.39 + 2(-143.36) = 21437.95$$

$$\sigma_{X_H + Y_O} = 146.42$$

$$E\{X_H - Y_O\} = (-6.42) - (-5.80) = -0.62 \quad V\{X_H - Y_O\} = 10837.28 + 10887.39 - 2(-143.36) = 22011.39$$

$$\sigma_{X_H - Y_O} = 148.36$$

Suppose we were interested in the "portfolio" between Home and Over that minimized the variance over this period.

$$w^* = \frac{\sigma_Y^2 - \rho_{XY}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y} = \frac{10887.39 - (-0.0132)(104.10)(104.34)}{10837.28 + 10887.39 - 2(-0.0132)(104.10)(104.34)} = \frac{11030.77}{22011.42} = .5011$$

Thus, the weights would be virtually the same since the variances are very similar and correlation close to 0.

∇

3.3.4 Moment-Generating Functions

Moment-generating functions (MGFs) have two important uses in probability when they exist. For the first use, consider the following infinite sum for e^x, with $i! \equiv 0$ and its first two derivatives with respect to x.

$$e^x = \sum_{i=1}^{\infty} \frac{x^i}{i!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

$$\frac{de^x}{dx} = 0 + 1 + \frac{2x}{2!} + \frac{3x^2}{3!} + \cdots = e^x$$

$$\frac{d^2e^x}{dx^2} = 0 + 0 + 1 + \frac{6x}{3!} + \cdots = e^x$$

Setting $x = tY$ leads to the following infinite series and its first two derivatives with respect to t.

$$e^{tY} = \sum_{i=1}^{\infty} \frac{(tY)^i}{i!} = 1 + tY + \frac{(tY)^2}{2!} + \frac{(tY)^3}{3!} + \frac{(tY)^4}{4!} + \cdots = 1 + tY + \frac{t^2Y^2}{2} + \frac{t^3Y^3}{6} + \frac{t^4Y^4}{24} + \cdots$$

$$\frac{de^{tY}}{dt} = 0 + Y + \frac{2tY^2}{2!} + \frac{3t^2Y^3}{3!} + \frac{4t^3Y^4}{4!} + \cdots = Y + tY^2 + \frac{t^2Y^3}{2!} + \frac{t^3Y^4}{3!} + \cdots$$

$$\frac{d^2e^{tY}}{dt^2} = 0 + Y^2 + \frac{2tY^3}{2!} + \frac{3t^2Y^4}{3!} + \cdots = Y^2 + tY^3 + \frac{t^2Y^4}{2!} + \cdots$$

Evaluating the last two derivatives at $t = 0$ yields the following results.

$$\frac{de^{tY}}{dt}|_{t=0} = Y \qquad \frac{d^2e^{tY}}{dt^2}|_{t=0} = Y^2$$

Now, we define the moment-generating function as $M(t) = E\{e^{tY}\}$, for a discrete random variable Y. The function must exist in a neighborhood of t around 0 to be able to obtain the following non-central moments of the random variable.

$$M(t) = E\{e^{tY}\} \sum_y e^{ty} p(y) = \sum_y \left[\sum_{i=1}^{\infty} \frac{(ty)^i}{i!} \right] p(y)$$

$$\Rightarrow \quad \frac{dM(t)}{dt}|_{t=0} = M'(t)|_{t=0} = \sum_y y p(y) = E\{Y\}$$

$$M''(t)|_{t=0} = \sum_y y^2 p(y) = E\{Y^2\} \qquad M^{(k)}(t)|_{t=0} = \sum_y y^k p(y) = E\{Y^k\}$$

The moment-generating function is also useful in determining the distributions of functions of random variables, as seen for some of the families of random variables described in subsequent sections.

3.4 Common Families of Discrete Probability Distributions

Here we consider some commonly used families of probability distributions for discrete random variables. These are the binomial, Poisson, and negative binomial families. They are used in many situations of data being counts of numbers of particular events occurring in an experiment.

3.4.1 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

The case when there is only $n = 1$ trial is referred to as a Bernoulli experiment and the probability distribution is called the Bernoulli distribution.

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

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

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

$$\text{Binomial Expansion: } (a+b)^n = \sum_{y=0}^{n}\binom{n}{y}a^y b^{n-y} \quad \Rightarrow \quad \sum_{y=0}^{n}\binom{n}{y}\pi^y(1-\pi)^{n-y} = (\pi + (1-\pi))^n = 1^n = 1$$

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

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

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

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

The moment-generating function for the binomial distribution is obtained below and used to obtain the mean and variance as well.

$$M(t) = E\{e^{tY}\} = \sum_{y=0}^{n} e^{ty}\binom{n}{y}\pi^y(1-\pi)^{n-y} = \sum_{y=0}^{n}\binom{n}{y}\left(e^t\right)^y\pi^y(1-\pi)^{n-y} = \sum_{y=0}^{n}\binom{n}{y}\left(\pi e^t\right)^y(1-\pi)^{n-y} = \left(\pi e^t + (1-\pi)\right)^n$$

$$M'(t) = n\left(\pi e^t + (1-\pi)\right)^{n-1}\pi e^t$$

$$M''(t) = n(n-1)\left(\pi e^t + (1-\pi)\right)^{n-2}(\pi e^t)^2 + n\left(\pi e^t + (1-\pi)\right)^{n-1}\pi e^t$$

$$E\{Y\} = M'(0) = n\left(\pi(1) + (1-\pi)\right)^{n-1}\pi(1) = n(1)^{n-1}\pi = n\pi$$

$$E\{Y^2\} = M''(0) = n(n-1)\left(\pi(1) + (1-\pi)\right)^{n-2}(\pi(1))^2 + n\left(\pi(1) + (1-\pi)\right)^{n-1}\pi(1) = n(n-1)\pi^2 + n\pi$$

$$VE\{Y\} = E\{Y^2\} - (E\{Y\})^2 = n(n-1)\pi^2 + n\pi - (n\pi)^2 = n\pi - n\pi^2 = n\pi(1-\pi)$$

Consider a case where a binomial experiment is replicated m times with the i^{th} experiment having n_i trials and common probability of success, π. For he i^th experiment, Y_i is the random number of successes. This model could be used by a bettor who makes n_i wagers in week i for a sport, with presumably $\pi = 0.5$ probability of winning on each wager. If we assume the weeks (and wagers within weeks) are independent

y	$6!/(y!(6-y)!)$	$.5010^y.4990^{(6-y)}$	$p(y)$
0	1	.01544	.01544
1	6	.01550	.09300
2	15	.01556	.23344
3	20	.01562	.31250
4	15	.01569	.23531
5	6	.01575	.09450
6	1	.01581	.01581
Sum	64	N/A	1

Table 3.5: Probability Distribution for number of Winning Over bets out of $n = 6$ wagers - NFL 2010-2019

and define $W = Y_1 + \cdots + Y_m$, we have the following moment-generating function for W, the total number of winning bets during the m weeks.

$$M_W(t) = E\{e^{tW}\} = E\{e^{t(Y_1+\cdots+Y_m)}\} = E\{e^{tY_1}\}\cdots E\{e^{tY_m}\} = \prod_{i=1}^{n} E\{e^{tY_i}\} = \prod_{i=1}^{n} \left(\pi e^t + (1-\pi)\right)^{n_i} = \left(\pi e^t + (1-\pi)\right)^{\sum_{i=1}^{m} n_i}$$

That is, the total winnings is Binomial with $n_\bullet = \sum_{i=1}^{m} n_i$ trials and probability of success π.

Example 3.9: NFL Over/Under Bets 2010-2019 Seasons

During the 2010-2019 NFL seasons, there were 2560 games played. After removing the 27 games in which the Over/Under bet was a push, there were 2533 games in which the Over/Under wager either won or lost. Here we consider having bet Over, so if the teams combine for more points than the Over/Under line, the outcome is considered a success. The Over bet won 1269 times ($\pi = .5010$)) and lost 1264 times ($(1-\pi) = .4990$).

Suppose you placed Over bets on a random sample of $n = 6$ games during this period. Then the number of wins Y can be modeled as a binomial distribution with $\pi = .5010$. The number of winning games of the 6 can take on the values $y = 0, 1, \ldots, 5, 6$. The calculations to obtain $p(y)$ are given in Table **??**.

The mean, variance, and standard deviation of the number of successful wagers in the $n = 6$ bets under this model are as follow.

$$\mu_Y = n\pi = 6(0.5010) = 3.00600 \qquad \sigma_Y^2 = n\pi(1-\pi) = 6(0.5010)(1-0.5010) = 1.49999 \qquad \sigma_Y = \sqrt{1.49999} = 1.22474$$

Suppose that a bettor wagers 110 to win 100 on each of the 6 bets. If she wins the bet, she receives 110+100=210, if she loses the bet she receives 0; she pays 110 in advance for each bet. Then her winnings are $W = -660 + 210Y$, and her mean, standard deviation, and probability of nonnegative winnings are obtained below.

$$E\{W\} = \mu_W = -660 + 210\mu_Y = -660 + 210(3.006) = -28.74 \quad \sigma_W = |210|\sigma_Y = 210(1.22474) = 257.1954$$

$$P(W \geq 0) = P(Y \geq 4) = .23531 + .09450 + .01581 = .34562$$

We take 100000 random samples of 6 games from this population of games (Pushes were removed). The average winnings was -27.74, the standard deviation was 257.1878, and the fraction of winning samples was .34641. All of these empirical values are in strong agreement with the theoretical values (as they should be).

R Commands and Output

```
## Commands/Output
num.samp <- 100000
set.seed(123)
win.games <- numeric(num.samp)
N.games <- length(OU)
n.games <- 6
OU.result <- ifelse(teamPts + oppPts > OU, 1, 0)
sum(OU.result)

for (i1 in 1:num.samp) {
  samp.games <- sample(N.games, n.games, replace=F)
  win.games[i1] <- sum(OU.result[samp.games])
}

table(win.games)

winnings <- -660+210*win.games
mean(winnings)
sd(winnings)
mean(winnings > 0)

> table(win.games)
win.games
    0     1     2     3     4     5     6
 1502  9287 23224 31346 23528  9487  1626
>
> winnings <- -660+210*win.games
> mean(winnings)
[1] -27.7404
> sd(winnings)
[1] 257.1878
> mean(winnings > 0)
[1] 0.34641
```

∇

3.4.2 Multinomial Distribution

A multinomial experiment is an extension of the binomial experiment with the caveat that each of n trials can end in one of k possible outcomes or categories. The probability of outcome j is π_j, and the count of the number of trials falling in category j is Y_j. The following restrictions must hold.

$$\pi_1 + \cdots + \pi_k = 1 \qquad Y_1 + \cdots + Y_k = n \quad \Rightarrow \quad Y_k = n - Y_1 - \cdots - Y_{k-1}$$

Note that for the binomial case, we have previously labeled $Y_1 = Y$ and $Y_2 = n - Y$ where category 1

represents "Success" and category 2 represents "Failure." The probability of the experiment resulting in the observed counts $(y_1, \ldots, y_k)$ is as follows.

$$p(y_1, \ldots, y_k) = \frac{n!}{y_1! \cdots y_k!} \pi_1^{y_1} \cdots \pi_k^{y_k} \qquad \pi_1 + \cdots + \pi_k = 1 \quad y_1 + \cdots + y_k = n \quad \pi_i \geq 0 \quad y_i \geq 0$$

To obtain $E\{Y_j\}$, $V\{Y_j\}$, and $\text{COV}\{Y_j, Y_{j'}\}$, we make use of the following dummy variables, extending the Binomial case.

$$U_i = 1 \text{ if Trial } i \text{ is in category } j, \text{ 0 otherwise} \qquad W_i = 1 \text{ if Trial } i \text{ is in category } j', \text{ 0 otherwise}$$

$$E\{U_i\} = 1(\pi_j) + 0(1 - \pi_j) = \pi_j \qquad E\{U_i^2\} = 1^2(\pi_j) + 0^2(1 - \pi_j) = \pi_j \quad \Rightarrow \quad V\{U_i\} = \pi_j - (\pi_j)^2 = \pi_j(1 - \pi_j)$$

$$E\{W_i\} = \pi_{j'} \qquad V\{W_i\} = \pi_{j'}(1 - \pi_{j'}) \qquad U_i W_i = E\{U_i W_i\} = 0 \quad \Rightarrow \quad \text{COV}\{U_i, W_i\} = 0 - \pi_j \pi_{j'} = -\pi_j \pi_{j'}$$

Trials are independent, so that $\text{COV}\{U_i, U_{i'}\} = \text{COV}\{U_i, W_{i'}\} = \text{COV}\{W_i, W_{i'}\} = 0$.

$$Y_j = \sum_{i=1}^{n} U_i \quad \Rightarrow \quad E\{Y_j\} = n\pi_j \qquad V\{Y_j\} = n\pi_j(1 - \pi_j)$$

$$Y_{j'} = \sum_{i=1}^{n} W_i \quad \Rightarrow \quad E\{Y_{j'}\} = n\pi_{j'} \qquad V\{Y_{j'}\} = n\pi_{j'}(1 - \pi_{j'})$$

$$\text{COV}\{Y_j, Y_{j'}\} = \sum_{i=1}^{n} \text{COV}\{U_i, W_i\} + \sum_{i=1}^{n} \sum_{i' \neq i} \text{COV}\{U_i, W_{i'}\} = -n\pi_j \pi_{j'}$$

Example 3.10: EPL Game Outcome Bets 2000/01-2018/19 Seasons

For each English Premier Football (Soccer) League game over the 2000/01-2018/19 seasons, odds for the Home team, Draw (tie), and Away team were obtained and averaged over various bookmakers from the website https://www.football-data.co.uk/. These odds were converted to probabilities by basic normalization for the 19(380) = 7220 games played over that period.

Overall, the Home team won 3556 of the games, 1751 were Draws, and the Away team won 1913 of the games. These correspond to the following probabilities.

$$\pi_H = \frac{3556}{7220} = 0.4925 \qquad \pi_D = \frac{1751}{7220} = 0.2425 \qquad \pi_A = \frac{1913}{7220} = 0.2650$$

Next, we consider the predicted probabilities of the Home, Draw, and Away that are obtained for each game. For instance, in the Charlton/Manchester City match from Week 1 of 2000/01 season played at

	Outcome			
Favored	Home	Draw	Away	Total
Home	2820 (.5696)	1186 (.2395)	945 (.1909)	4951 (1.0000)
Draw	18 (.1593)	32 (.2832)	63 (.5575)	113 (1.0000)
Away	718 (.3330)	533 (.2472)	905 (.4198)	2156 (1.0000)
Total	3556	1751	1913	7220

Table 3.6: Counts (conditional probabilities) of game outcomes by predicted outcomes - EPL 2000/01-2018/19

Charlton, the average odds for the Home team was 2.14, for the Away team was 2.96, and for the Draw was 3.12. We can convert these to unstandardized probabilities by taking their inverses, and to standardized probabilities by basic normalization. Recall that we use η to represent predicted (subjective) probabilities.

$$\eta_H = \frac{1}{2.14} = 0.4673 \quad \eta_D = \frac{1}{3.12} = 0.3205 \quad \eta_A = \frac{1}{2.96} = 0.3418 \quad \Rightarrow \quad \eta_H + \eta_D + \eta_A = 1.1296$$

$$\eta_H^* = \frac{0.4673}{1.1296} = 0.4137 \qquad \eta_D^* = \frac{0.3205}{1.1296} = 0.2837 \qquad \eta_A^* = \frac{0.3418}{1.1296} = 0.3026$$

The outcome with the highest predicted probability is treated as the predicted outcome for the game based on odds posted by oddsmakers. Table **??** contains the distributions of outcomes when the Home, Draw, and Away are the predicted outcomes.

∇

3.4.3 Poisson Distribution

In many applications, researchers observe the counts of a random process in some fixed amount of time or space. The random variable Y is a count that can take on any non-negative integer. One important aspect of the Poisson family is that the mean and variance are the same. This is one aspect that does not work for all applications. We use the notation: $Y \sim Poi(\lambda)$. The probability distribution is given below. It can be derived from taking the limit of the Binomial distribution as the number of trial goes to ∞ and the probability of success on each trial goes to 0, with $n\pi = \lambda$ being the mean.

$$p(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \qquad n\pi = \lambda \quad \Rightarrow \quad \pi = \frac{\lambda}{n}$$

$$\Rightarrow \quad p(y) = \binom{n}{y}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} \qquad \text{Letting } n \to \infty:$$

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

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

$$\text{for fixed } y: \quad \lim_{n\to\infty}\left(\frac{n}{n-\lambda}\right) = \cdots = \left(\frac{n-y+1}{n-\lambda}\right) = 1 \qquad \text{for fixed } a: \quad \lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a$$

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

$$\text{Series expansion for } e^a: \quad e^a = \sum_{i=0}^{\infty}\frac{a^i}{i!} \qquad \Rightarrow \qquad \sum_{y=0}^{\infty} p(y) = e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = e^{-\lambda}e^{\lambda} = 1$$

Thus the probability mass function sums to 1. We give the probability mass function and derive the mean and variance directly below.

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

$$E\{Y\} = \mu_Y = \sum_{y=0}^{\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = 0 + \sum_{y=1}^{\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = e^{-\lambda}\sum_{y=1}^{\infty}\frac{\lambda^{(y-1)+1}}{(y-1)!} = e^{-\lambda}\lambda\sum_{y=1}^{\infty}\frac{\lambda^{(y-1)}}{(y-1)!} =$$

$$e^{-\lambda}\lambda\sum_{z=0}^{\infty}\frac{\lambda^z}{z!} = e^{-\lambda}\lambda e^{\lambda} = \lambda \qquad \text{where: } z = y-1 = 0,1,2,\ldots$$

$$E\{Y(Y-1)\} = E\{Y^2\} - E\{Y\} = \sum_{y=0}^{\infty} y(y-1)\frac{e^{-\lambda}\lambda^y}{y!} = 0 + 0 + \sum_{y=2}^{\infty} y(y-1)\frac{e^{-\lambda}\lambda^y}{y!} =$$

$$e^{-\lambda}\sum_{y=2}^{\infty}\frac{\lambda^{(y-2)}+2}{(y-2)!} = e^{-\lambda}\lambda^2\sum_{y=2}^{\infty}\frac{\lambda^{(y-2)}}{(y-2)!} =$$

$$e^{-\lambda}\lambda^2\sum_{z=0}^{\infty}\frac{\lambda^z}{z!} = e^{-\lambda}\lambda^2 e^{\lambda} = \lambda^2 \qquad \text{where: } z = y-2 = 0,1,2,\ldots$$

$$\Rightarrow \quad E\{Y^2\} = E\{Y(Y-1)\} + E\{Y\} = \lambda^2 + \lambda \quad \Rightarrow \quad V\{Y\} = \sigma_Y^2 = E\{Y^2\} - [E\{Y\}]^2 = \lambda^2 + \lambda - (\lambda)^2 = \lambda$$

The Poisson model 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.

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.

The moment-generating function for the Poisson distribution is obtained as follows.

$$M(t) = E\{e^{tY}\} = \sum_{y=0}^{\infty} e^{ty}\frac{e^{-\lambda}\lambda^y}{y!} = \sum_{y=0}^{\infty}\left(e^t\right)^y\frac{e^{-\lambda}\lambda^y}{y!} = \sum_{y=0}^{\infty}\frac{e^{-\lambda}\left(\lambda e^t\right)^y}{y!} = e^{-\lambda}\sum_{y=0}^{\infty}\frac{\left(\lambda e^t\right)^y}{y!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda\left(e^t-1\right)}$$

We can also obtain $E\{Y\}$, $E\left\{Y^2\right\}$, and $V\{Y\}$ making use of the MGF.

y	$p(y)$	Expected	Observed	y	$p(y)$	Expected	Observed
0	0.004427	68.8864	102	7	0.120695	1878.01	1808
1	0.023995	373.3643	296	8	0.081771	1272.352	1324
2	0.065027	1011.817	1123	9	0.049244	766.2386	722
3	0.117482	1828.016	1538	10	0.02669	415.3013	411
4	0.159188	2476.962	2684	11	0.013151	204.6303	210
5	0.17256	2685.027	2590	12	0.00594	92.42468	74
6	0.155879	2425.475	2637	≥ 13	0.003952	61.49423	41

Table 3.7: Probability Distribution for Number of goals during regulation playing time - NHL 2006/7-2018/9

$$M'(t) = \lambda e^t e^{\lambda\left(e^t-1\right)} \quad \Rightarrow \quad E\{Y\} = M'(0) = \lambda e^0 e^{\lambda\left(e^0-1\right)} = \lambda(1)(1) = \lambda$$

$$M''(t) = \lambda e^t e^{\lambda\left(e^t-1\right)} + \lambda e^t \lambda e^t e^{\lambda\left(e^t-1\right)} = \lambda e^t \left(1+\lambda e^t\right) e^{\lambda\left(e^t-1\right)}$$

$$\Rightarrow \quad E\left\{Y^2\right\} = M''(0) = \lambda(1+\lambda)(1) = \lambda + \lambda^2$$

$$\Rightarrow \quad V\{Y\} = E\left\{Y^2\right\} - (E\{Y\})^2 = \lambda + \lambda^2 - (\lambda)^2 = \lambda$$

If we observe independent Poisson random variables $Y_1, \ldots, Y_n$ from n experiments, where $Y_i \sim Poi(\lambda_i)$ and we obtain the sum $W = Y_1 + \cdots + Y_n$, then the moment-generating function for W is obtained below.

$$M_W(t) = E\{e^{tW}\} = E\{e^{t\sum_{i=1}^n Y_i}\} = E\{e^{tY_1}\cdots e^{tY_n}\} = E\{e^{tY_1}\}\cdots E\{e^{tY_n}\}e^{\lambda_1\left(e^t-1\right)}\cdots e^{\lambda_n\left(e^t-1\right)} = e^{\left(\sum_{i=1}^n \lambda_i\right)\left(e^t-1\right)}$$

That is, the sum of the independent Poisson random variables is Poisson with mean equal to the sum of their means.

Example 3.10: NHL Total Goals in Regulation Time

During the 2006/07-2018/19 NHL seasons, games are based on 3 20-minute periods of regulation play. If the game is tied, a 5-minute sudden death overtime is played with 3 skaters and a goalie for each team. If still tied after the 5-minute OT, then a shootout is held, with 3 attempts per team to begin, then sudden death. In any event, there will be one extra assigned goal if the game goes to OT. We consider the numbers of goals scored in the regulation period of the 15560 NHL games played during these seasons and use a Poisson distribution to model it. The mean and variance are $\mu_Y = 5.42$ and $\sigma^2_Y = 5.23$, respectively. While they are not exactly the same, they are reasonably close. The probability distribution, Expected, and Observed counts are given in Table **??**. The Expected counts are obtained by multiplying the probabilities by the total number of games (15560). The Observed and Expected Counts are similar, but not a perfect match. In a later chapter a goodness of fit test is described that can be used to assess the reasonability of the Poisson approximation to the distribution.

∇

3.4.4 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) [**?**] and Cameron and Trivedi (2005) [**?**]).

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

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

Example 3.11: Home and Away Team Scores - NFL 2010-2019

The number of points scored by the home team and the away team was studied by Cain, Law, and Peel (2000), [**?**]. They considered various distributions, including the Poisson and negative binomial for modeling the predictions based on the point spread and the Over/Under line. We simply use the negative binomial distribution to model the outcomes here, regression models are considered later.

The number of points for the home team ranged from 0 to 62 while the points for the away team ranged from 0 to 59. The mean and variance are given below, along with "method of moments" estimates for μ and α for the negative binomial distribution.

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

$$\mu_H = 23.75 \qquad \sigma_H^2 = 106.88 \qquad \alpha_H = \frac{106.88/23.75 - 1}{23.75} = 0.1474 \qquad \alpha_H^{-1} = 6.78$$

$$\mu_A = 21.54 \qquad \sigma_A^2 = 97.79 \qquad \alpha_A = \frac{97.79/21.54 - 1}{21.54} = 0.1643 \qquad \alpha_A^{-1} = 6.08$$

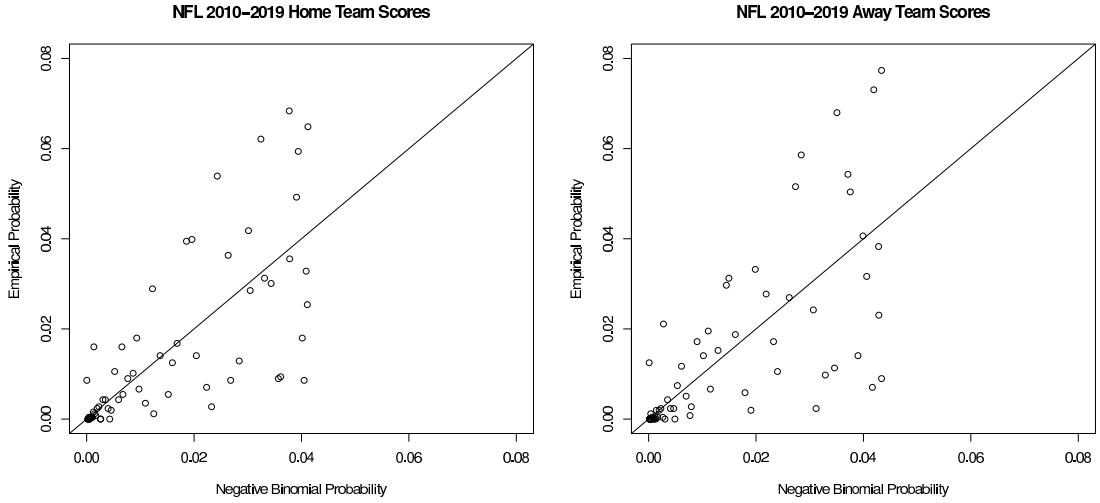


Figure 3.1: Probabilities of Home and Away team points versus negative binomial based probabilities - NFL 2010-2019

Plots of the observed probabilities for 0-65 versus the negative binomial based probabilities are given in Figure **??**. There is clearly an association between the probabilities, but they tend to fall a ways from the line of equality.

∇

3.5 Common Families of 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 denote 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. The gamma distribution also has two parameters associated with it. The beta distribution is often used to model data that are proportions (or can be extended to any finite length interval). The beta distribution also has two parameters. All of these families can take on a wide range of shapes by changing parameter values.

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

3.5.1 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 $Y \sim N\left(\mu, \sigma^2\right)$, but note that many statistics textbooks use the notation $Y \sim N\left(\mu, \sigma\right)$. The probability density function, the mean and variance are given below. Note that when using functions in most software packages and spreadsheets, including R (dnorm for density, pnorm for cumulative probability, qnorm for quantiles, and rnorm for random samples), you specify the standard deviation σ, not the variance σ^2. That is, if you want to generate a (pseudo) random sample of size $n = 100$ from a normal distribution with mean $\mu = 200$ and variance $\sigma^2 = 25$, use the command **rnorm(100,200,5)**.

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

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:

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

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:

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

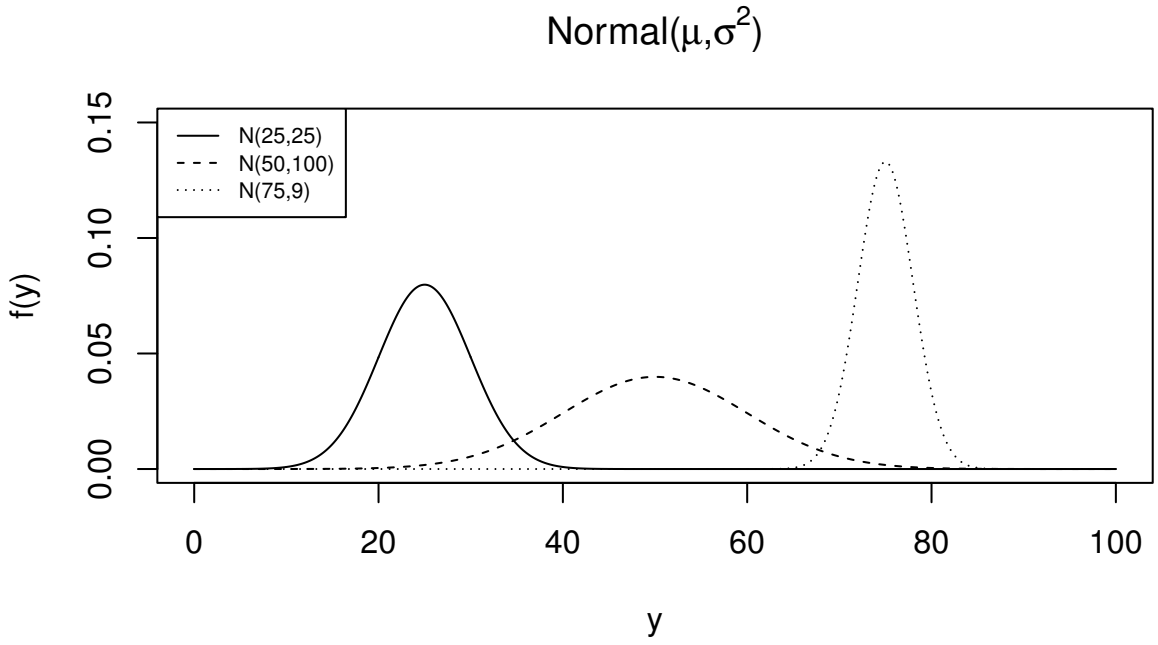


Figure 3.2: Three Normal Densities

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

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

Example 3.12: Point Spread Differential - WNBA 2010-2019

While point spread differential for the home team is technically discrete, the actual outcomes take on a wide range of values (-39.0 to 44.0) and is approximately normally distributed with $\mu = 0.0289$ and standard deviation $\sigma = 12.0472$ ($\sigma^2 = 145.1350$). A frequency histogram and super-imposed N(0.0289, 145.1350) is given in Figure **??**. Notice that in the center, the bar is higher than the normal curve, and in the bins to the right are below; otherwise the distribution is a reasonably good fit to the data.

Consider the following quantiles (.10, .25, .50, .75, .90) for the WNBA data and the corresponding N(0.0289, 145.1350) distribution. Also consider the probabilities of the following ranges ($< 0.0289 - 2(12.0472) = -24.0655, > 0.0289 + 2(12.0472) = 24.1233$, and $(-12.0183 = 0.0289 - 12.0472, 0.0289 + 12.0472 = 12.0761)$) for the WNBA home point spread differential data and the normal distribution. The $.10^{\text{th}}$ quantile for the data (-14.6) is higher than that for the normal distribution (-15.4), but the other quantiles are very close. The proportions more than 2 standard deviations above and below the mean are very similar (.0226 below versus .0228) and (.0260 above versus .0228). The proportions within a standard deviation of the mean are also close (.6964 versus .6827).

R Output

```
### R Output

> ## Quantiles: Theoretical Normal, Actual Distribution
> norm.q <- qnorm(c(.10,.25,.50,.75,.90), mean.sD, sd.sD)
> sprdDiff.q <- quantile(sprdDiff, c(.10,.25,.50,.75,.90))
> quant.out <- cbind(norm.q, sprdDiff.q)
> colnames(quant.out) <- c("Theoretical", "Actual")
> round(quant.out,2)
    Theoretical Actual
10%      -15.41  -14.6
25%       -8.10   -8.0
50%        0.03    0.0
75%        8.15    8.0
90%       15.47   15.5
>
> ## Probabilities: Theoretical Normal, Actual Distribution
> # Theoretical
> tprob <- cbind(
+ pnorm(mean.sD-2*sd.sD, mean.sD, sd.sD),
+ 1-pnorm(mean.sD+2*sd.sD, mean.sD, sd.sD),
+ pnorm(mean.sD+sd.sD, mean.sD, sd.sD) -
+     pnorm(mean.sD-sd.sD, mean.sD, sd.sD))
> # Actual
> aprob <- cbind(
+ mean(sprdDiff <= mean.sD-2*sd.sD),
+ mean(sprdDiff >= mean.sD+2*sd.sD),
+ mean(sprdDiff >= mean.sD-sd.sD & sprdDiff <= mean.sD+sd.sD))
>
> prob.out <- rbind(tprob.out, aprob.out)
> colnames(prob.out) <- c("P(Y < mu-2*sigma)", "P(Y > mu+2*sigma)",
+                "P(mu-sigma < Y < mu+sigma)")
> rownames(prob.out) <- c("Theoretical", "Actual")
> round(prob.out,4)
           P(Y < mu-2*sigma) P(Y > mu+2*sigma) P(mu-sigma < Y < mu+sigma)
Theoretical            0.0228            0.0228                     0.6827
Actual                 0.0226            0.0260                     0.6964
```

∇

The moment-generating function for the normal family is obtained below and involves "completing the square" inside the exponential in the integral.

$$M(t) = E\{e^{tY}\} = \int_{-\infty}^{\infty} e^{ty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^2+\mu^2-2\mu y\right)^2}{2\sigma^2} + \frac{ty(2\sigma^2)}{2\sigma^2}\right) dy =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^2+\mu^2-2\mu y-2ty\sigma^2\right)}{2\sigma^2}\right) dy =$$

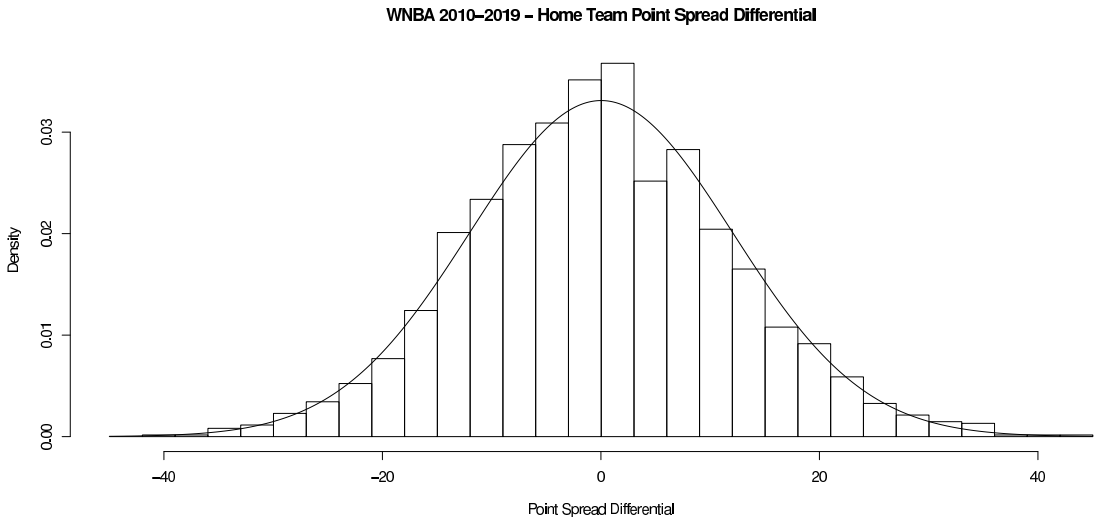


Figure 3.3: Home team spread differential and N(μ=.0289, σ^2=145.1350) distribution - WNBA 2010-2019

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^2+\mu^2-2(\mu+t\sigma^2)y\right)}{2\sigma^2}\right) dy =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^2+(\mu+t\sigma^2)^2-2(\mu+t\sigma^2)y-2\mu t\sigma^2-(t\sigma^2)^2\right)}{2\sigma^2}\right) dy =$$

$$\exp\left(\frac{2\mu t\sigma^2+(t\sigma^2)^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-(\mu+t))^2}{2\sigma^2}\right) dy =$$

$$\exp\left(\mu t+\frac{t^2\sigma^2}{2}\right) (1)$$

The final integral is 1 since, the integrand is the $N(\mu+t,\sigma^2)$ density.

3.5.2 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.

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} \quad y \geq 0, \alpha > 0, \beta > 0 \qquad \int_0^\infty w^a e^{-y/b} dy = \Gamma(a) b^{a+1}$$

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

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

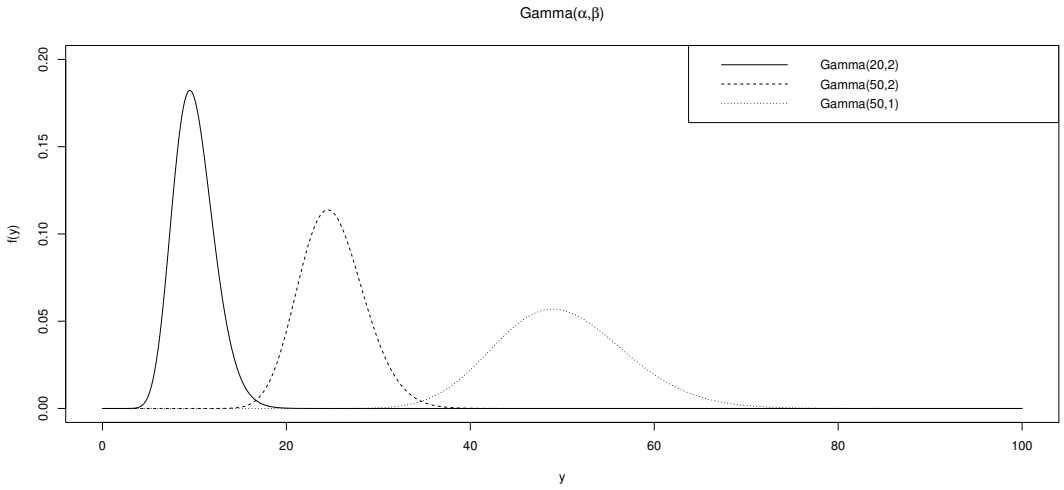


Figure 3.4: Three Gamma Densities

Thus, if α is an integer, $\Gamma(\alpha) = (\alpha - 1)!$. We give the mean and variance as well as their derivation below.

$$E\{Y\} = \mu_Y = \alpha\beta \qquad V\{Y\} = \sigma_Y^2 = \alpha\beta^2$$

$$E\{Y\} = \mu_Y = \int_0^\infty y\frac{1}{\Gamma(\alpha)\beta^\alpha}y^{\alpha-1}e^{-y/\beta}dy = \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty y^\alpha e^{-y/\beta}dy = \frac{1}{\Gamma(\alpha)\beta^\alpha}\Gamma(\alpha+1)\beta^{\alpha+1} =$$

$$\frac{\Gamma(\alpha+1)\beta}{\Gamma(\alpha)} = \frac{\alpha\Gamma(\alpha)\beta}{\Gamma(\alpha)} = \alpha\beta$$

$$E\left\{Y^2\right\} = \frac{\Gamma(\alpha+2)\beta^{\alpha+2}}{\Gamma(\alpha)\beta^\alpha} = (\alpha+1)\alpha\beta^2 \quad \Rightarrow \quad V\{Y\} = \sigma_Y^2 = (\alpha+1)\alpha\beta^2 - (\alpha\beta)^2 = \alpha\beta^2$$

The second version given here is in terms of *shape* and *rate* parameters.

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

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

Example 3.13: Total Points - WNBA 2010-2019

When considering the Total Points scored during regular season WNBA 2010-2019 games we see they are all positive, and skewed to the right. A histogram and smooth gamma density are given in Figure **??**. The mean is 157.57 and the variance is 350.07. 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.

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

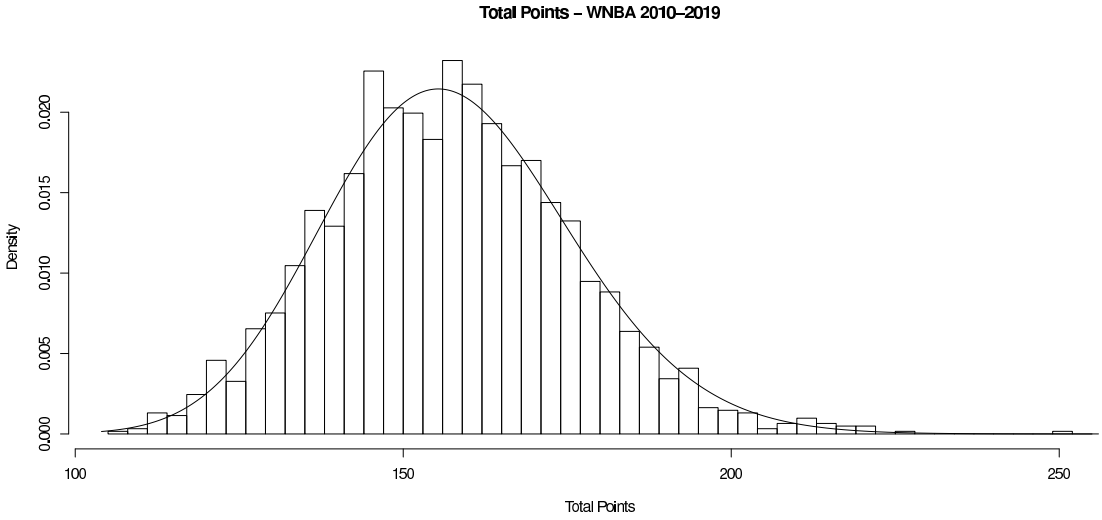


Figure 3.5: Total points scored per game WNBA 2010-2019

$$\alpha = \frac{157.57^2}{350.07} = 70.93 \qquad \beta = \frac{157.57}{350.07} = 0.45$$

Similar to what was done for the WNBA point spread differentials for the normal distribution, we compare the theoretical quantiles for the total points with the actual quantiles, and compare theoretical probabilities with observed probabilities. There is very good agreement between the quantiles as well as the probabilities.

R Output

```
## R Output

> ## Quantiles: Theoretical Gamma, Actual Distribution
> gamma.q <- qgamma(c(.10,.25,.50,.75,.90), alpha.TP, beta.TP)
> totPts.q <- quantile(totPts, c(.10,.25,.50,.75,.90))
> quant.out <- cbind(gamma.q, totPts.q)
> colnames(quant.out) <- c("Theoretical", "Actual")
> round(quant.out,2)
    Theoretical Actual
10\%      134.12    134
25\%      144.58    145
50\%      156.84    157
75\%      169.76    170
90\%      181.98    182
> ## Probabilities: Theoretical Gamma, Actual Distribution
> # Theoretical
> tprob.out <- cbind(
+ pgamma(mean.TP-2*sd.TP, alpha.TP, beta.TP),
+ 1-pgamma(mean.TP+2*sd.TP, alpha.TP, beta.TP),
+ pgamma(mean.TP+sd.TP, alpha.TP, beta.TP) -
+     pgamma(mean.TP-sd.TP, alpha.TP, beta.TP))
> # Actual
```

```
> aprob.out <- cbind(
+ mean(totPts <= mean.TP-2*sd.TP),
+ mean(totPts >= mean.TP+2*sd.TP),
+ mean(totPts >= mean.TP-sd.TP & totPts <= mean.TP+sd.TP))
>
> prob.out <- rbind(tprob.out, aprob.out)
> colnames(prob.out) <- c("P(Y < mu-2*sigma)", "P(Y > mu+2*sigma)",
+                "P(mu-sigma < Y < mu+sigma)")
> rownames(prob.out) <- c("Theoretical", "Actual")
> round(prob.out,4)
            P(Y < mu-2*sigma) P(Y > mu+2*sigma) P(mu-sigma < Y < mu+sigma)
Theoretical            0.0160            0.0288                      0.6838
Actual                 0.0162            0.0294                      0.6984
```

∇

To obtain the moment-generating function, we use the following result from calculus, then derive it directly using the result. Note that this based on the second formulation of the model.

$$\int_0^\infty y^{a-1}e^{-by}dy = \Gamma(a)b^{-a} \qquad \text{see normalizing constant for the gamma density}$$

$$\Rightarrow \qquad M(t) = E\{e^{tY}\} = \int_0^\infty e^{ty}\frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-y\beta}dy = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty y^{\alpha-1}e^{-y(\beta-t)}dy =$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)}\Gamma(\alpha)(\beta-t)^{-\alpha} = \frac{\beta^\alpha}{(\beta-t)^\alpha} = \left(1-\frac{t}{\beta}\right)^{-\alpha}$$

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, the symbol β is often replaced by θ.

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

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

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

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

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

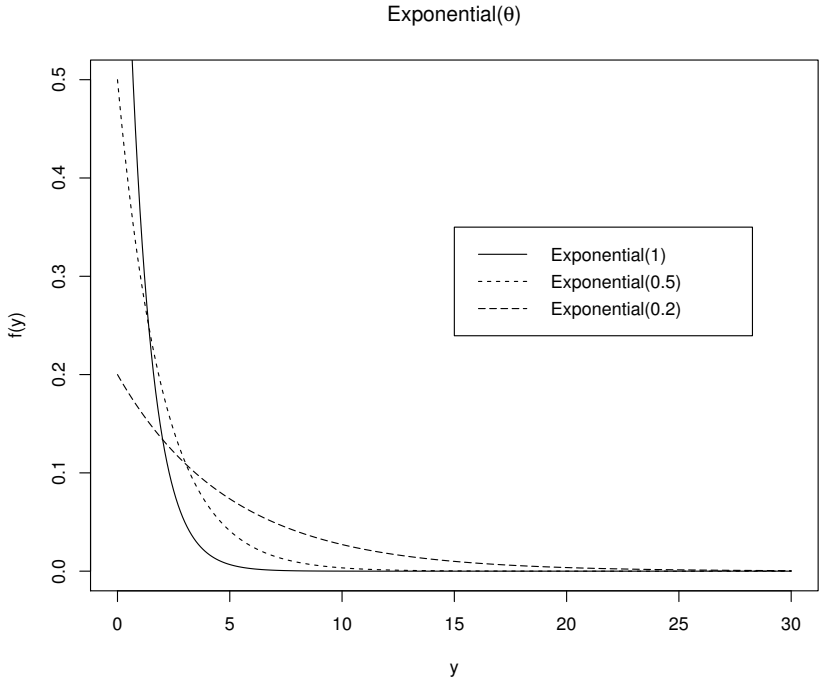


Figure 3.6: Three Exponential Densities

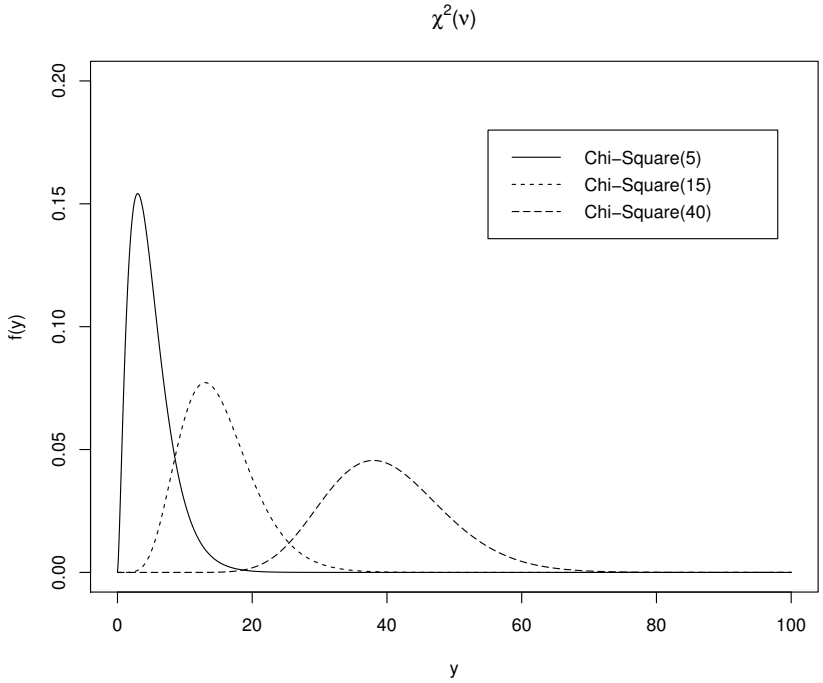


Figure 3.7: Three Chi-Square Densities

3.5.3 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 along with their derivation.

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

$$E\{Y\}=\mu_Y=\int_0^1 y\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}\,dy=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 y^{\alpha}(1-y)^{\beta-1}dy=$$

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}=\frac{\alpha\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)(\alpha+\beta)\Gamma(\alpha+\beta)}=\frac{\alpha}{\alpha+\beta}$$

$$E\left\{Y^2\right\}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)}=\frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$$

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

An alternative formulation of the distribution involves the following re-parameterization.

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

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

Figure **??** gives three Beta distributions.

Example 3.14: Home Team Subjective Win Probabilities - MLB 2015-2019

During the MLB 2015-2019 regular seasons there were 12152 games played. Based on basic normalization, we computed the home team subjective probability of winning the game based on the Money Lines. The mean and standard deviation are 0.5341 and 0.0934, respectively. These lead to the following parameters based on the method of moments.

$$\phi=\frac{0.5341(1-0.5341)}{0.0934^2}-1=27.50\qquad\alpha=27.50(.5341)=14.69\qquad\beta=27.50(1-.5341)=12.81$$

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

R Output

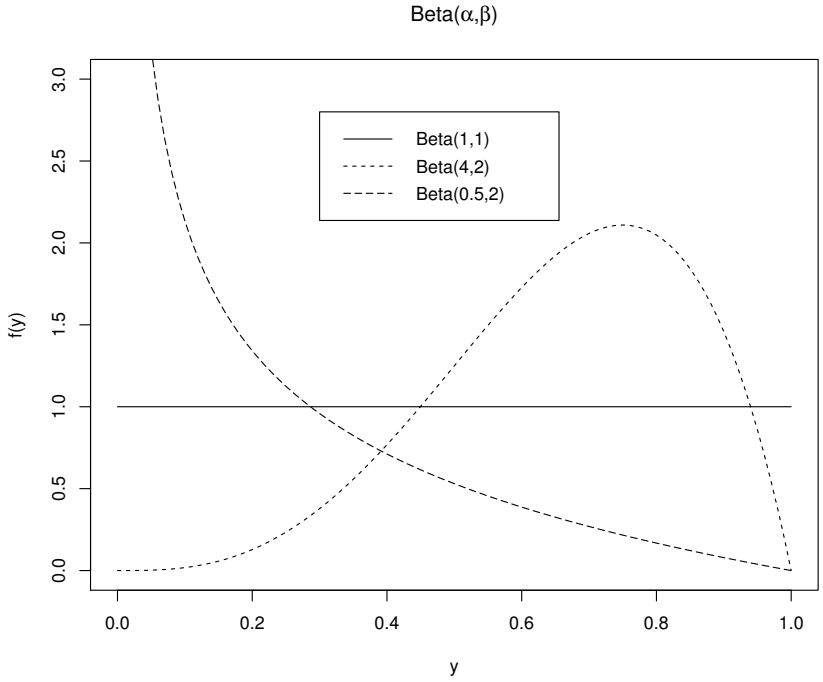


Figure 3.8: Three Beta Densities

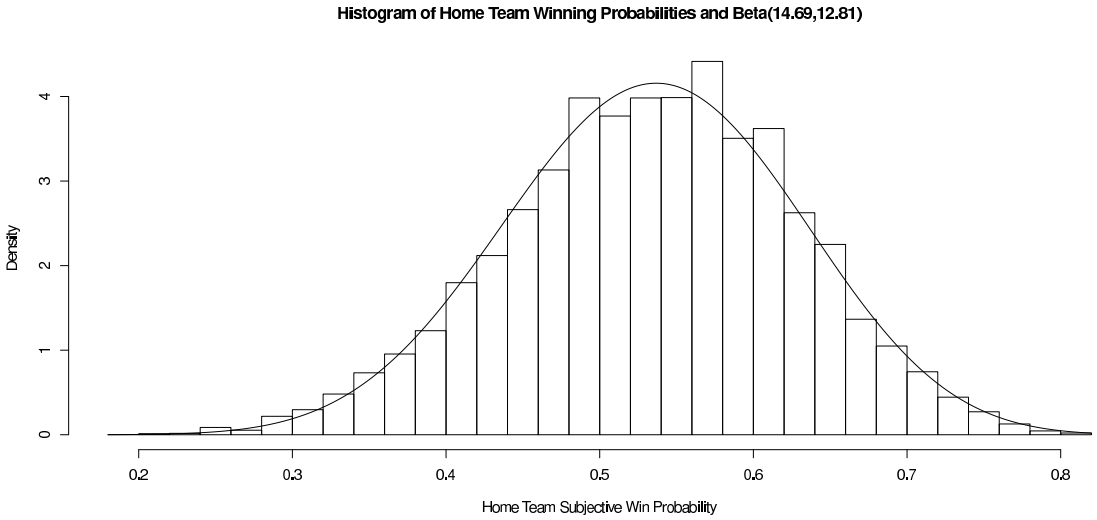


Figure 3.9: Home team subjective probabilities of winning and Beta(14.69,12.81) density - MLB 2015-2019

```
## R Commands

> ## Quantiles: Theoretical Gamma, Actual Distribution
> beta.q <- qbeta(c(.10,.25,.50,.75,.90), alpha.TP, beta.TP)
> teamprob.q <- quantile(teamprob, c(.10,.25,.50,.75,.90))
> quant.out <- cbind(beta.q, teamprob.q)
> colnames(quant.out) <- c("Theoretical", "Actual")
> round(quant.out,4)
    Theoretical Actual
10%      0.4124 0.4119
25%      0.4701 0.4714
50%      0.5350 0.5372
75%      0.5990 0.6010
90%      0.6547 0.6512
>
>
> ## Probabilities: Theoretical Beta, Actual Distribution
> # Theoretical
> tprob.out <- cbind(
+ pbeta(mean.TP-2*sd.TP, alpha.TP, beta.TP),
+ 1-pbeta(mean.TP+2*sd.TP, alpha.TP, beta.TP),
+ pbeta(mean.TP+sd.TP, alpha.TP, beta.TP) -
+     pbeta(mean.TP-sd.TP, alpha.TP, beta.TP))
> # Actual
> aprob.out <- cbind(
+ mean(teamprob <= mean.TP-2*sd.TP),
+ mean(teamprob >= mean.TP+2*sd.TP),
+ mean(teamprob >= mean.TP-sd.TP & teamprob <= mean.TP+sd.TP))
>
> prob.out <- rbind(tprob.out, aprob.out)
> colnames(prob.out) <- c("P(Y < mu-2*sigma)", "P(Y > mu+2*sigma)",
+                "P(mu-sigma < Y < mu+sigma)")
> rownames(prob.out) <- c("Theoretical", "Actual")
> round(prob.out,4)
           P(Y < mu-2*sigma) P(Y > mu+2*sigma) P(mu-sigma < Y < mu+sigma)
Theoretical            0.0232            0.0201                      0.6740
Actual                 0.0277            0.0179                      0.6796
```

∇

3.5.4 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 $Y_i \sim N(\mu, \sigma)$ 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.

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

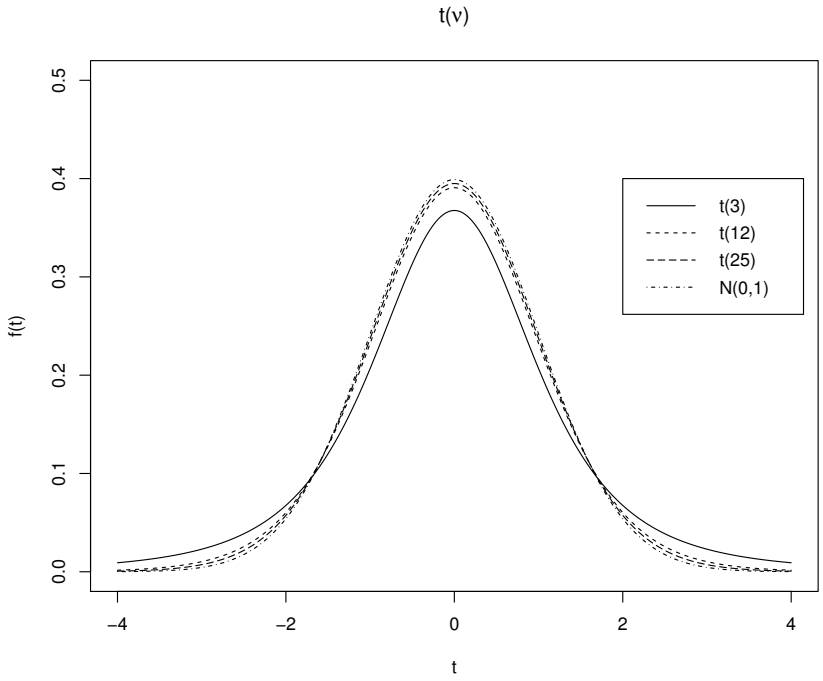


Figure 3.10: Three t-densities and z

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 **??**.

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

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

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

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

This result permits us to make inference concerning mean(s) even when the population variance is unknown and must be estimated.

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 **??**.

Critical values for the t, χ^2, and F-distributions are given in statistical textbooks and webpages. Probabilities, quantiles, densities, and random samples 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.

3.5.5 Bivariate Normal Distribution

The bivariate normal distribution is used to model pairs of variables, we will us X and Y to represent the variables with the following notation and joint probability density function.

$$E\{X\} = \mu_X \quad V\{X\} = \sigma^2_X = \sigma_{XX} \quad E\{Y\} = \mu_Y \quad V\{Y\} = \sigma^2_Y = \sigma_{YY} \quad \text{COV}\{X,Y\} = \sigma_{XY} = \rho\sigma_X\sigma_Y$$

$$f(x,y) = \frac{1}{2\pi\sqrt{\sigma^2_X\sigma^2_Y(1-\rho^2)}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma^2_X} - \frac{2(x-\mu_X)(y-\mu_y)\rho}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma^2_Y}\right]\right\}$$
$$-\infty < x, y, \mu_X, \mu_Y < \infty \quad \sigma_X, \sigma_Y > 0 \quad -1 \le \rho \le 1$$

We will derive the marginal density for X by integrating over Y, and then obtain the conditional density of $Y|X = x$ by taing the joint density of X, Y and dividing it by the marginal density of X.

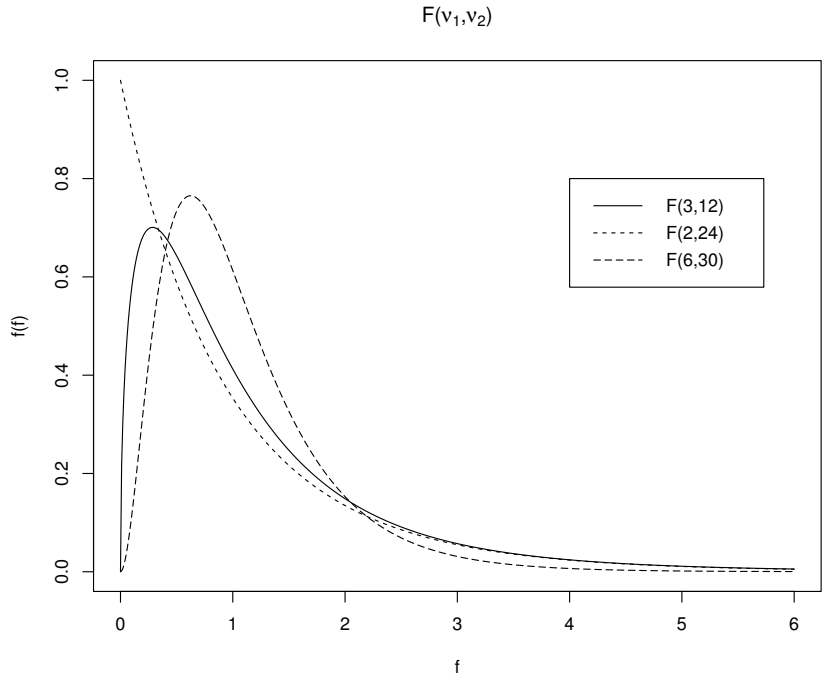


Figure 3.11: Three F-densities

$$\int_{-\infty}^{\infty} f(x,y)dy = \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)}} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2(x-\mu_X)(y-\mu_y)\rho}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\} dy$$

We label the normalizing constant as C and complete the square within the exponent. Note that the last two terms below do not involve y.

$$= C\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{\rho^2(x-\mu_X)^2}{\sigma_X^2} - \frac{2(x-\mu_X)(y-\mu_y)\rho}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} + \frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{\rho^2(x-\mu_X)^2}{\sigma_X^2}\right]\right\} dy$$

$$= \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)}} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{\rho^2(x-\mu_X)^2}{\sigma_X^2}\right]\right\} \times$$

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{\rho^2(x-\mu_X)^2}{\sigma_X^2} - \frac{2(x-\mu_X)(y-\mu_y)\rho}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\} dy$$

We first simplify the constant, then terms within the integrand.

$$\frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)}} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{\rho^2(x-\mu_X)^2}{\sigma_X^2}\right]\right\} = \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)}} \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}\right]\right\}$$

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{\rho^2(x-\mu_X)^2}{\sigma_X^2}-\frac{2(x-\mu_X)(y-\mu_y)\rho}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}dy =$$

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(y-\mu_Y)}{\sigma_Y}-\frac{\rho(x-\mu_X)}{\sigma_X}\right]^2\right\}dy =$$

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_Y^2}\left[y-\left(\mu_Y+\rho(x-\mu_X)\frac{\sigma_Y}{\sigma_X}\right)\right]^2\right\}dy$$

The integrand represents the exponential portion (kernel) of a normal density for Y with the following mean and variance.

$$E\{Y\}=\mu_Y+\rho(x-\mu_X)\frac{\sigma_Y}{\sigma_X} \qquad V\{Y\}=\left(1-\rho^2\right)\sigma_Y^2$$

Thus the integral is equal to the following constant (the reciprocal of the normalizing constant for the normal density).

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_Y^2}\left[y-\left(\mu_Y+\rho(x-\mu_X)\frac{\sigma_Y}{\sigma_X}\right)\right]^2\right\}dy=\sqrt{2\pi\left(1-\rho^2\right)\sigma_Y^2}$$

Putting together these two terms, we have the following result.

$$f_X(x)=\frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)}}\exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}\right]\right\}\times\sqrt{2\pi\left(1-\rho^2\right)\sigma_Y^2}=\frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left\{-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\right\}$$

This is the normal density for a random variable X with mean μ_X and variance σ_X^2. Note that the marginal distribution of Y is normal with mean μ_Y and variance σ_Y^2.

Finally, we divide the joint density of X and Y by the marginal density of X to obtain the conditional density of Y given $X=x$.

$$\frac{f(x,y)}{f_X(x)}=\frac{\frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2(1-\rho^2)}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}-\frac{2(x-\mu_X)(y-\mu_y)\rho}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}}{\frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left\{-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\right\}}=$$

$$\frac{1}{\sqrt{2\pi\sigma_Y^2\left(1-\rho^2\right)}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}-\frac{2(x-\mu_X)(y-\mu_y)\rho}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]-\left[-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\right]\right\}=$$

$$\frac{1}{\sqrt{2\pi\sigma_Y^2\left(1-\rho^2\right)}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}-\frac{2(x-\mu_X)(y-\mu_y)\rho}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-\frac{\left(1-\rho^2\right)(x-\mu_X)^2}{\sigma_X^2}\right]\right\}=$$

$$\frac{1}{\sqrt{2\pi\sigma_Y^2\left(1-\rho^2\right)}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(y-\mu_Y)^2}{\sigma_Y^2}-\frac{2(x-\mu_X)(y-\mu_y)\rho}{\sigma_X\sigma_Y}+\frac{(x-\mu_X)^2}{\sigma_X^2}\left(1-\left(1-\rho^2\right)\right)\right]\right\}=$$

$$\frac{1}{\sqrt{2\pi\sigma_Y^2\left(1-\rho^2\right)}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_Y^2}\left[(y-\mu_Y)^2-\frac{2(x-\mu_X)(y-\mu_y)\rho\sigma_Y}{\sigma_X}+\frac{\rho^2(x-\mu_X)^2\sigma_Y^2}{\sigma_X^2}\right]\right\}=$$

$$\frac{1}{\sqrt{2\pi\sigma_Y^2\left(1-\rho^2\right)}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_Y^2}\left[(y-\mu_Y)-\frac{\rho(x-\mu_X)\sigma_Y}{\sigma_X}\right]^2\right\}=$$

$$\frac{1}{\sqrt{2\pi\sigma_Y^2\left(1-\rho^2\right)}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_Y^2}\left[\left(y-\left(\mu_Y+\rho(x-\mu_X)\frac{\sigma_Y}{\sigma_X}\right)\right]^2\right\}$$

Thus, conditional on $X=x$, Y is normal with the following conditional distribution.

$$Y|X=x\sim N\left(E\{Y|X=x\}=\mu_Y+\rho(x-\mu_X)\frac{\sigma_Y}{\sigma_X},V\{Y|X=x\}=\left(1-\rho^2\right)\sigma_Y^2\right)$$

Example 3.15: WNBA Team Point Predictions and Outcomes

Based on the point spread and Over/Under, team predicted points were obtained for the population of WNBA games for the 2010-2019 regular seasons. Labeling the pre-game prediction as X and the team's actual points as Y, we obtain the following parameters.

$$\mu_X=78.74\quad\sigma_X=5.86\quad\mu_Y=78.79\quad\sigma_Y=11.57\quad\rho=0.49$$

Note that the means are virtually identical, but the standard deviation is almost twice as large for actual points than predicted (the variance is almost four times as high). Histograms of the predicted and actual team points are given in Figure **??**. A scatterplot and ordinary least squares regression line is given in Figure **??**. The equation for the regression line is given below, based on the conditional density of Y.

$$E\{Y|X=x\}=\mu_Y+\rho(x-\mu_X)\frac{\sigma_Y}{\sigma_X}=\left(\mu_Y-\rho\mu_X\frac{\sigma_Y}{\sigma_X}\right)+\rho\frac{\sigma_Y}{\sigma_X}x=$$

$$\left(78.79-0.49(78.74)\frac{11.57}{5.86}\right)+0.49\frac{11.57}{5.86}x=2.61+0.97x$$

A 3D plot of the bivariate normal density for these parameters is given in Figure **??** and contour plot is given in given in Figure **??**.

Conditional densities for Y with $X=60$ to 110 and the marginal density of Y are given in Figure **??** and an ellipsoid containing 95% of the joint density and the observed pairs (X,Y) are given in Figure **??**.

∇

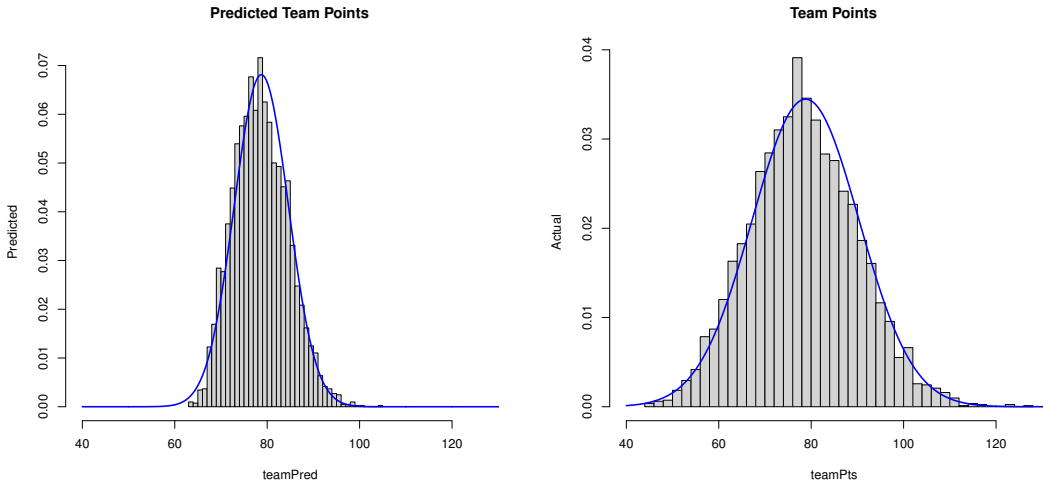


Figure 3.12: Histograms of Predicted and Actual Team Points - WNBA 2010-2019 Regular Seasons with super-imposed Normal Densities

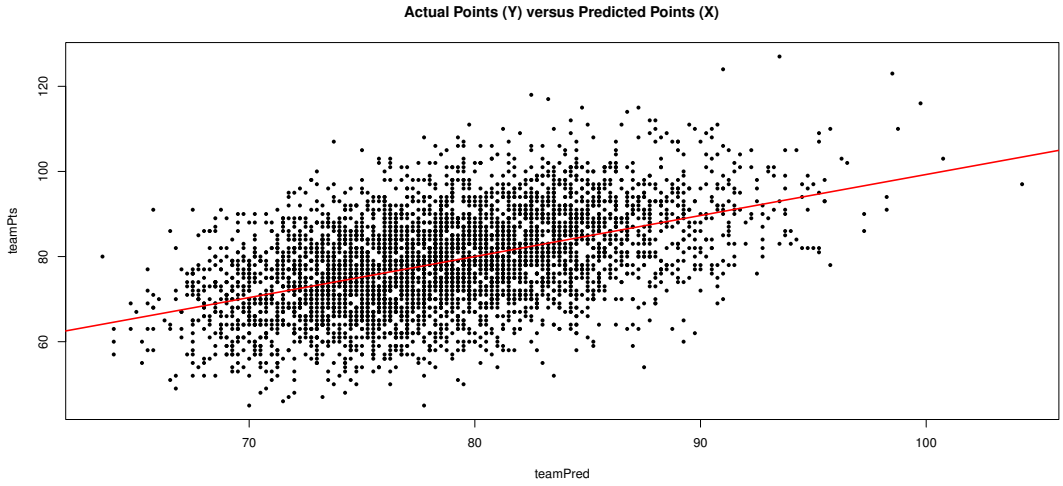


Figure 3.13: Plot of Actual versus Predicted Team Points - WNBA 2010-2019 Regular Seasons with Conditional Mean of Actual Points

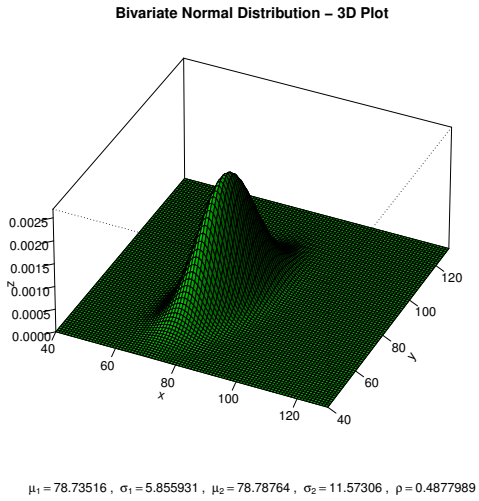


Figure 3.14: Bivariate Normal Density (3D Plot) - WNBA 2010-2019 Regular Seasons

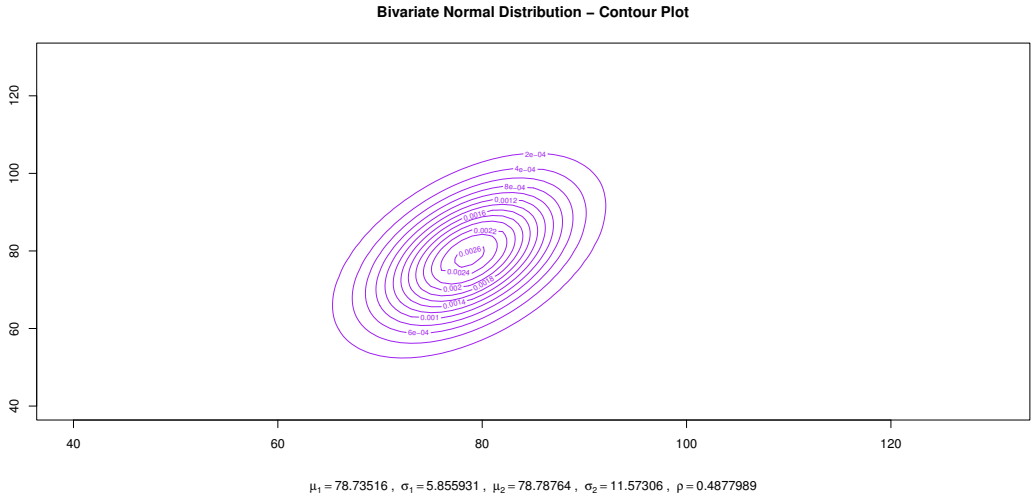


Figure 3.15: Bivariate Normal Density (Contour Plot) - WNBA 2010-2019 Regular Seasons

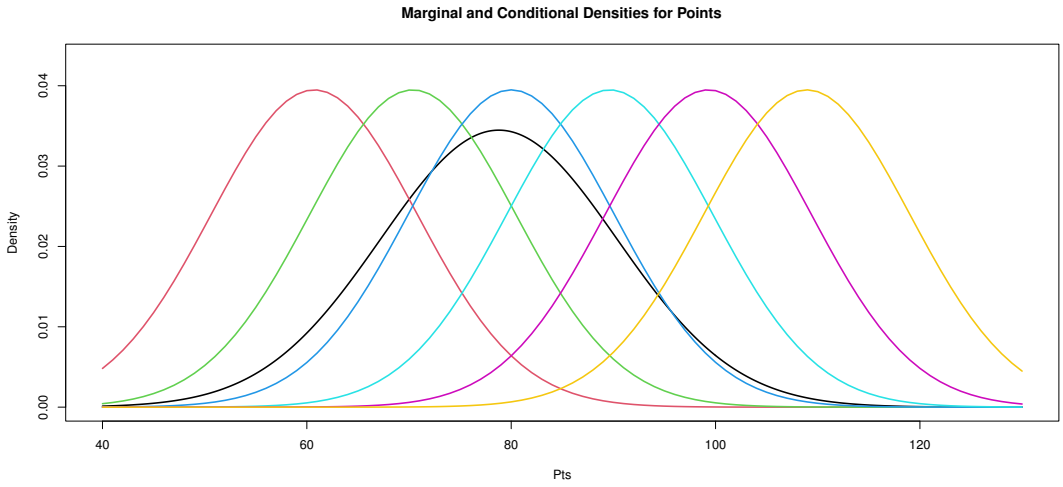


Figure 3.16: Conditional Densities of Y given X=60,70,...,100,110 and Marginal Density of Y - WNBA 2010-2019 Regular Seasons

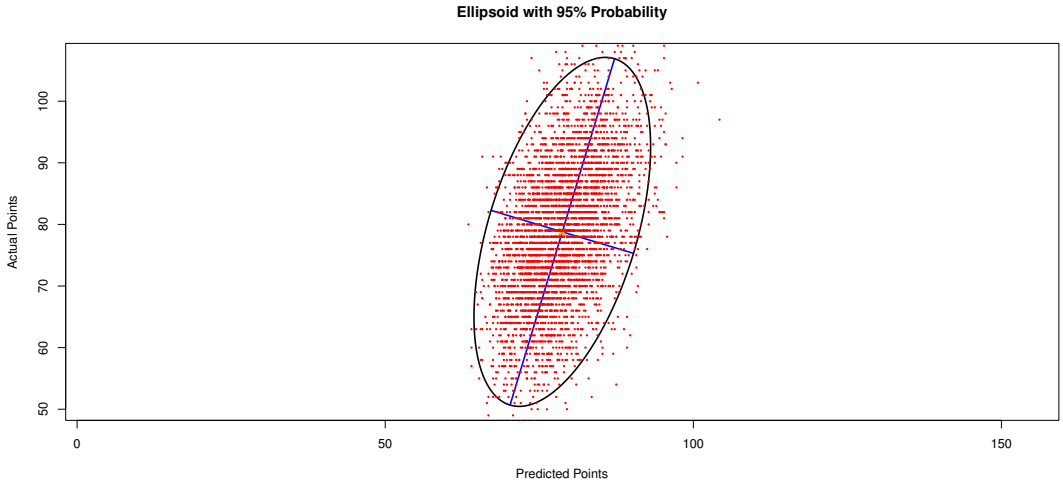


Figure 3.17: Data and Ellipsoid containing 95% of Joint Density - WNBA 2010-2019 Regular Seasons

3.6 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 deviation 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**.

$$\text{Sample Mean: } \overline{Y} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{Sample Proportion: } \hat{\pi} \dot{\sim} N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Example 3.15: Sampling Distribution of a Mean - NFL Team Point Predictions and Outcomes

Based on the NFL point spread and the Over/Under line, we are able to obtain a prediction for each team's score in a game. When the team is favored, again we use the convention that its spread is positive. Let S be the team's point spread for a game and O represent the Over/Under line for the game. Further, let P represent the predicted number of points for the team. Note that the dataset has a row for each team as the "focal team" with twice as many rows as there are games. That is, the dataset is based on "team games." The predicted score for the focal team is obtained by taking $P = (O + S)/2$. For instance if the focal team is favored by $S = +7$ and the Over/Under line is $O = 42$ points, the total score is predicted

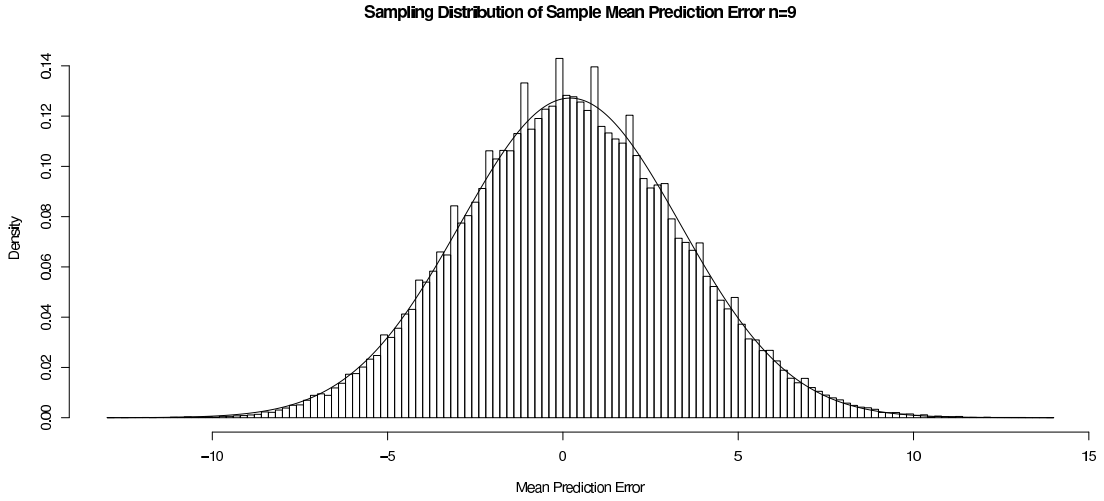


Figure 3.18: Sampling distribution of mean Point prediction errors based on Team spread and Over/Under line - NFL 2010-2019

to be 42, with the focal team expected to win by 7, making the predicted score for the focal team to be $P = (42 + 7)/2 = 24.5$ and their opponent's predicted score is $(42 - 7)/2 = 17.5$. We are interested in the difference between the focal teams' actual and predicted scores. This can be thought of as the "prediction error" of the betting lines.

The distribution has a minimum of -26 where the team scored much less than predicted and a maximum of 41. The median is -0.25 and the mean is 0.21. The lower and upper quartiles are -6.06 and 6.00, respectively and the standard deviation is 9.41. The distribution is somewhat skewed to the right.

We take 100000 random samples of size $n = 9$, computing and saving the sample mean for each sample. The theoretical and empirical (based on the 100000 random samples) mean and standard error of the sample means are given below and a histogram with the normal density are shown in Figure **??**.

$$\text{Theory: } \mu_{\overline{Y}} = \mu = 0.2102 \quad \sigma_{\overline{Y}} = \frac{9.4099}{\sqrt{9}} = 3.1366 \qquad \text{Empirical: } \overline{\overline{y}} = 0.2289 \quad s_{\overline{y}} = 3.1164$$

The mean and standard deviation are quite close to the corresponding theoretical values, as expected and the sampling distribution is well approximated by the $N(0.2102, 3.1366^2)$ distribution.

∇

Example 3.16: Sampling Distribution of a Proportion - NHL Favorites' Probability of Winning Game

Next we consider the proportions of NHL favored teams that win the game, based on 100000 random samples of $n = 200$. For the population, there were 15320 games with a favorite (different Money Lines for the two teams) and the favored team won 8972 of the games, so $\pi = 8972/15320 = .5856$. The theoretical and empirical results are given below, and the histogram is given in Figure **??**.

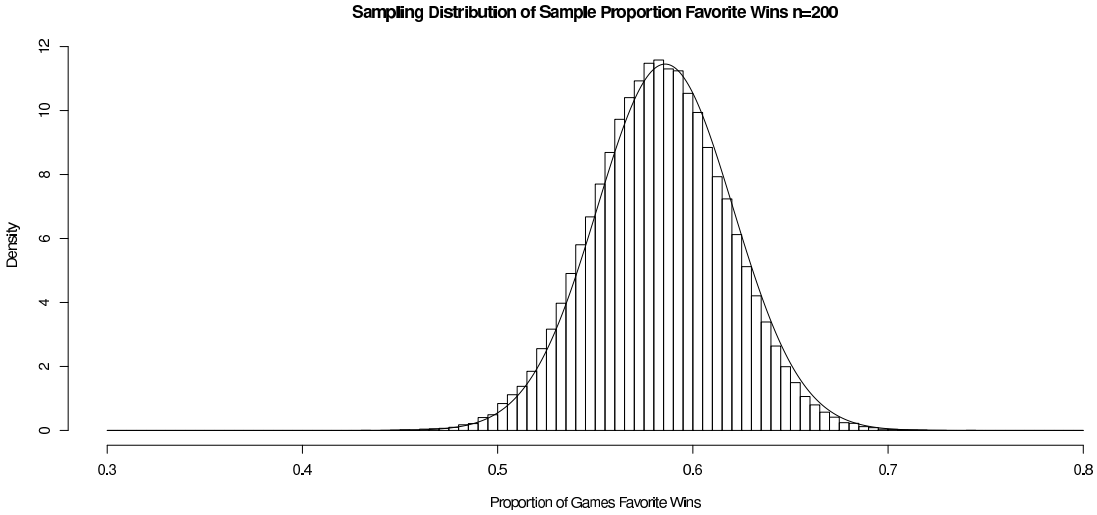


Figure 3.19: Sampling distribution for proportion of games favorite wins - n=200, NHL 2006/07-2018/19

$$\text{Theory: } \mu_{\hat{\pi}} = \pi = .5856 \quad \sigma_{\hat{\pi}} = \sqrt{\frac{.5856(1-.5856)}{200}} = .0348 \qquad \text{Empirical: } \overline{\hat{\pi}} = .5856 \quad s_{\hat{\pi}} = .0346$$

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 0.000, .005, .010, ..., .990, .995, 1.000, as $n = 200$. The histogram is clearly bell-shaped like a normal distribution. As n gets larger $\hat{\pi}$ becomes more "continuous."

∇

3.7 Introduction to Bayesian Statistics

In frequentist statistics, parameters such as the mean μ, variance σ^2, or proportion π are treated as fixed, typically unknown values. In Bayesian statistics, parameters are treated as random variables with probability distributions assigned prior to observing the data (prior distributions) which are updated once data have been observed (posterior distributions). In this section, we consider a binomial proportion π, a Poisson mean λ, and a normal mean μ and variance σ^2. We consider **conjugate** prior distributions, which allow for the the posterior and prior distributions to be of the same family. There are many alternatives to using conjugate priors and software exists to obtain samples from more complicated posterior distributions.

In general, we have the following structure, where θ is a parameter or a vector of parameters, y is data, $p(\theta)$ is the prior distribution for θ and $p(y|\theta)$ is the probability model for the random variable Y, given θ.

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \propto p(\theta)p(y|\theta) \qquad \theta \text{ discrete: } p(y) = \sum p(y|\theta)p(\theta) \quad \text{continuous: } p(y) = \int p(y|\theta)p(\theta)d\theta$$

3.7.1 Bayesian Model for a Binomial Proportion

The beta distribution provides a conjugate prior distribution for a binomial proportion. Also, the beta family has a lot of flexibility for the shape of the prior distribution based on various belief states corresponding to π. Consider the following prior distribution for π and the subsequent probability distribution for a binomial random variable Y where $\pi \sim Beta\left(\alpha_\pi, \beta_\pi\right)$.

$$p\left(\pi\right) = \frac{\Gamma\left(\alpha_\pi + \beta_\pi\right)}{\Gamma\left(\alpha_\pi\right)\Gamma\left(\beta_\pi\right)}\pi^{\alpha_\pi - 1}(1-\pi)^{\beta_\pi - 1} \quad 0 < \pi < 1 \quad \alpha_\pi, \beta_\pi > 0 \qquad p\left(y|n,\pi\right) = \binom{n}{y}\pi^y\left(1-\pi\right)^{n-y} \quad y = 0, 1, \ldots, n$$

$$p\left(\pi|y\right) \propto p\left(\pi\right)p\left(y|n,\pi\right) \propto \pi^{\alpha_\pi - 1}(1-\pi)^{\beta_\pi - 1}\pi^y(1-\pi)^{n-y} = \pi^{\alpha_\pi + y - 1}(1-\pi)^{\beta_\pi + n - y - 1}$$

Thus, the posterior distribution for π, given $Y = y$ is a beta distribution with parameters $\alpha_\pi + y$ and $\beta_\pi + n - y$, respectively.

Example 3.17: Probability that a WNBA Game Total Points Exceeds Over/Under Line

Suppose we are interested in the probability a WNBA game exceeds the Over/Under Line, which will be denoted as π. We will consider only games that are not "pushes," as they imply no bet was placed. Consider the following priors.

- We have no idea, except that $0 < \pi < 1$ and place a uniform prior between 0 and 1 which is a beta distribution with $\alpha_\pi = \beta_\pi = 1$
- We believe that there is a very good chance (say 90%) that π is between 0.40 and 0.60 and that the distribution is symmetric around 0.5

For the uniform prior, $P(0.40 \leq \theta \leq 0.60) = .20$. Using the **pbeta** function in R, we can use trial and error to obtain values $\alpha_\pi = \beta_\pi$ such that $P(0.40 \leq \pi \leq 0.60) = .90$. Based on integer values, when $\alpha_\pi = \beta_\pi = 34$, $P(0.40 \leq \pi \leq 0.60) = .9031$. Figure **??** gives the two prior distributions described here.

Once a sample of n games is obtained and y, the number of games which cover the Over bet is observed, we update the prior for π to obtain its posterior distribution. In a pseudo-random sample of $n = 100$ games, the over bet won $y = 45$ times and lost $n - y = 55$ times. For the first (uniform) prior, the posterior distribution for π is beta with $\alpha_1 = 1 + 45 = 46$ and $\beta_1 = 1 + 55 = 56$. For the second (concentrated) prior, the posterior is beta with $\alpha_2 = 34 + 45 = 79$ and $\beta_2 = 34 + 55 = 89$. Figure **??** gives the two posterior distributions described here. It should be noted that under the first posterior distribution, $P(0.40 \leq \pi \leq 0.60) = .8486$, and for the second it is .9666.

∇

3.7.2 Bayesian Model for a Poisson Mean

The gamma distribution provides a conjugate prior distribution for a Poisson mean. The gamma family can take on many forms for the shape of the prior distribution based on various belief states corresponding to

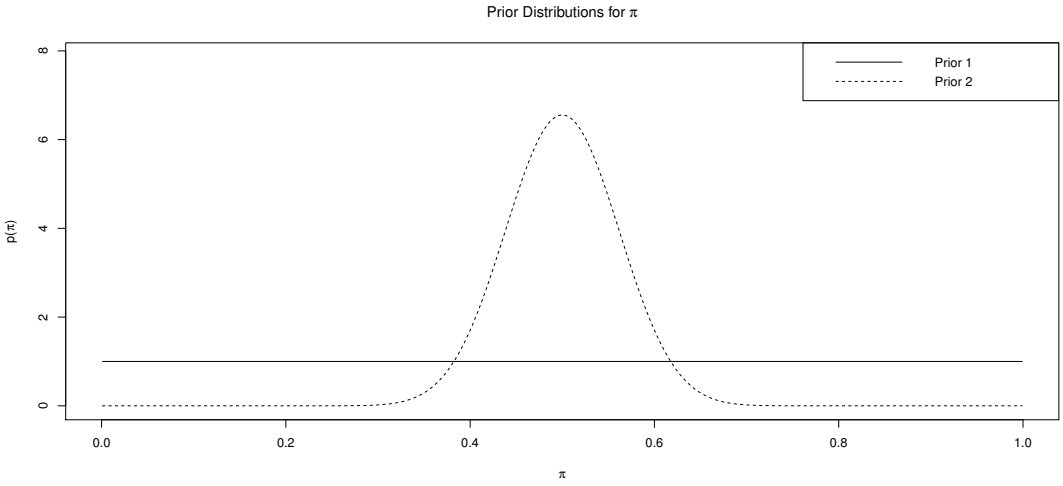


Figure 3.20: Prior Distributions for π - WNBA Over Probability

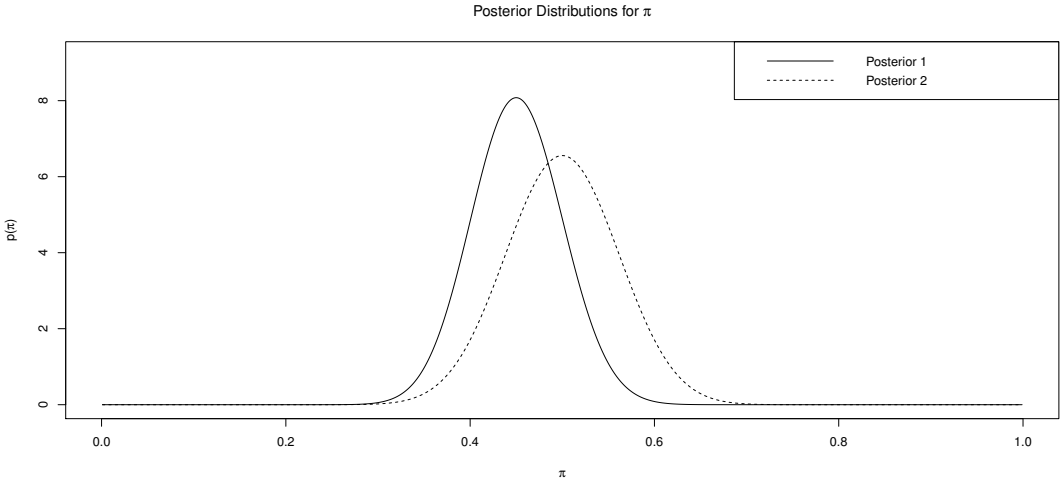


Figure 3.21: Posterior Distributions for π - WNBA Over Probability

λ. Consider the following prior distribution for λ and the subsequent probability distribution for a Poisson random variable Y where $\lambda \sim Gamma\,(\alpha_\lambda, \beta_\lambda)$. Suppose we take a random sample of size n from a Poisson distribution with mean λ and observe $y_1, \ldots, y_n$.

$$p\,(\lambda) = \frac{\beta_\lambda^{\alpha_\lambda}}{\Gamma\,(\alpha_\lambda)}\lambda^{\alpha_\lambda - 1}e^{-\lambda\beta_\lambda} \quad \lambda > 0 \quad \alpha_\lambda, \beta_\lambda > 0$$

$$p\,(y_1, \ldots, y_n|\lambda) = \prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \quad y_i = 0, 1, \ldots$$

$$p\,(\lambda|y_1, \ldots, y_n) \propto p\,(\lambda)\,p\,(y_1, \ldots, y_n|\lambda) \propto \lambda^{\alpha_\lambda - 1}e^{-\lambda\beta_\lambda}e^{-\lambda n}\lambda^{\sum_{i=1}^{n} y_i} = \lambda^{\alpha_\lambda + n\overline{y} - 1}e^{-\lambda(\beta_\lambda + n)}$$

Thus, the posterior distribution for λ, given $Y_1 = y_1, \ldots, Y_n = y_n$ is a gamma distribution with parameters $\alpha_\lambda + n\overline{y}$ and $\beta_\lambda + n$, respectively.

Example 3.18: Average Total Goals in English Premier League Football Matches 2002-2018 Seasons

In the English Premier League, all matches have an Over/Under total of 2.5 goals. The odds are assigned to the Over/Under line, depending presumably on the offensive/defensive strengths and weather conditions. We consider the games where the converted odds (by basic normalization) have probabilities of Over (and thus Under) between 0.47 and 0.53. Note that the Over and Under probabilities sum to 1, with the Over probability ranging from 0.3248 to 0.7732. We will place a prior on λ, the mean total number of goals in such games that is gamma with mean of 2.5 ($\alpha_\lambda = 2.5\beta_\lambda$) such that the probability that λ falls between 1 and 4 is approximately 0.70. For $\alpha_\lambda = 2.5$ and $\beta_\lambda = 1.0$, the $P(1 \leq \lambda \leq 4) = 0.6929$.

Of the games beginning in the 2002/03 season, there were 6458 games with Over/Under odds. Of those games, 2376 games had Over (and Under) probabilities between 0.47 and 0.53. We take a pseudo-random sample of $n = 50$ of these games and observe the mean $\overline{y} = 2.76$. This leads to a posterior distribution for λ as a gamma distribution with $\alpha = 2.5 + 50(2.76) = 140.5$ and $\beta = 1 + 50 = 51$. A plot of the prior and posterior distributions is given in Figure **??**

3.7.3 Bayesian Model for Normal Mean and Variance

If we have a data model for Y being normal with mean μ and variance σ^2, we can place priors on μ and σ^2 and sample from their posterior distributions iteratively with the **Gibbs sampler**. We consider a prior normal distribution on μ that is normal with mean μ_μ and variance σ_μ^2. For the variance σ^2, we consider a prior Inverse Gamma distribution with parameters a_{σ^2} and b_{σ^2}. Note that if a random variable has an Inverse Gamma distribution with parameters a and b, its reciprocal is distributed as gamma with parameters a and b. That is, the **precision**, $\tau^2 = 1/\sigma^2$ is distributed as gamma. The Inverse Gamma density is given below.

$$f(x|a, b) = \frac{b^a}{\Gamma(a)}x^{-a-1}\exp\left(\frac{-b}{x}\right) \quad x > 0 \quad a, b > 0 \quad E\{X\} = \frac{b}{a-1} \quad (a > 1) \quad V\{X\} = \frac{b^2}{(a-1)^2(a-2)} \quad (a > 2)$$

We have the following priors and data model.

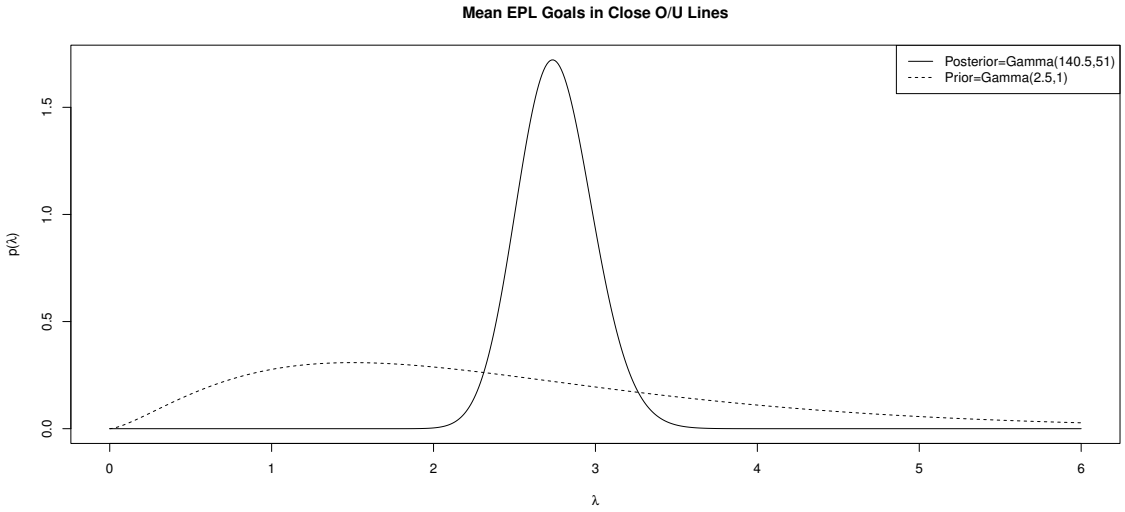


Figure 3.22: Prior and Posterior of Mean Total Goals in EPL games with subjective probability of Over (Under) between 0.47 and 0.53

$$\mu \sim N\left(\mu_\mu, \sigma_\mu^2\right) \qquad p(\mu) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left(-\frac{\left(\mu-\mu_\mu\right)^2}{2\sigma_\mu^2}\right)$$

$$\sigma^2 \sim IG\left(a_{\sigma^2}, b_{\sigma^2}\right) \qquad p\left(\sigma^2\right) = \frac{\left(b_{\sigma^2}\right)^{a_{\sigma^2}}}{\Gamma\left(a_{\sigma^2}\right)} \left(\sigma^2\right)^{-a_{\sigma^2}-1} \exp\left(\frac{-b_{\sigma^2}}{\sigma^2}\right)$$

$$p\left(y_1, \ldots, y_n | \mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n \left(y_i-\mu\right)^2}{2\sigma^2}\right)$$

The joint posterior density of μ and σ^2 given observed data $y_1, \ldots, y_n$ is proportional to the products of their priors (assuming independence) and the likelihood of the data.

$$p\left(\mu, \sigma^2 | y_1, \ldots, y_n\right) \propto p(\mu) p\left(\sigma^2\right) p\left(y_1, \ldots, y_n | \mu, \sigma^2\right)$$

Now, suppose σ^2 is known, and we want to obtain the posterior of μ given σ^2 and the data. Further, let σ_μ^2 be written in terms of the data variance: $\sigma_\mu^2 = \sigma^2/m$.

$$p\left(\mu | \sigma^2, y_1, \ldots, y_n\right) \propto \exp\left(-\frac{\left(\mu-\mu_\mu\right)^2}{2\sigma^2/m}\right) \exp\left(-\frac{\sum_{i=1}^n \left(y_i-\mu\right)^2}{2\sigma^2}\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left[m\left(\mu-\mu_\mu\right)^2 + \sum_{i=1}^n \left(y_i-\mu\right)^2\right]\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}\left[m\mu^2 - 2m\mu_\mu\mu + m\mu_\mu^2 + \sum_{i=1}^n y_i^2 - 2\mu\sum_{i=1}^n y_i + n\mu^2\right]\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left[\mu^2(m+n) - 2\mu\left(m\mu_\mu + \sum_{i=1}^n y_i\right)\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left[\frac{m+n}{\sigma^2}\mu^2 - 2\frac{m\mu_\mu + n\overline{y}}{\sigma^2}\mu\right]\right) = \exp\left(-\frac{1}{2}\left[\frac{1}{V_\mu}\mu^2 - 2M_\mu\mu\right]\right) \qquad V_\mu = \frac{\sigma^2}{m+n} \quad M_\mu = \frac{m\mu_\mu + n\overline{y}}{\sigma^2}$$

$$\Rightarrow \qquad p\left(\mu|\sigma^2, y_1, \ldots, y_n\right) \propto \exp\left(-\frac{1}{2}\left[\frac{1}{V_\mu}\mu^2 - 2\frac{M_\mu V_\mu}{V_\mu}\mu\right]\right) \propto \exp\left(-\frac{1}{2}\left[\frac{\left(\mu - M_\mu V_\mu\right)^2}{V_\mu}\right]\right)$$

Thus, given σ^2 and the data $y_1, \ldots, y_n$, μ has a normal posterior distribution with mean and variance given below.

$$E\{\mu|\sigma^2, y_1, \ldots, y_n\} = M_\mu V_\mu = \frac{m\mu_\mu + n\overline{y}}{\sigma^2}\frac{\sigma^2}{m+n} = \frac{m\mu_\mu + n\overline{y}}{m+n} \qquad V\{\mu|\sigma^2, y_1, \ldots, y_n\} = V_\mu = \frac{\sigma^2}{m+n}$$

In general, if we σ_μ^2 is not parameterized in terms of σ^2, we have the following result.

$$E\{\mu|\sigma^2, y_1, \ldots, y_n\} = \frac{\frac{\mu_\mu}{\sigma_\mu^2} + \frac{n\overline{y}}{\sigma^2}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}} \qquad V\{\mu|\sigma^2, y_1, \ldots, y_n\} = \frac{1}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}}$$

Thus, the posterior mean is a weighted average the prior mean and the data mean with weights based on the prior variance (m) and the data sample size (n). The posterior variance is the data variance σ^2 divided by the sum of the weights.

Now treating μ as known, we obtain the posterior distribution of σ^2 and the data $y_1, \ldots, y_n$. This takes the product of the prior for σ^2 and the likelihood, keeping only terms involving σ^2.

$$p\left(\sigma^2|\mu, y_1, \ldots, y_n\right) \propto \left(\sigma^2\right)^{-a_{\sigma^2}-1}\exp\left(\frac{-b_{\sigma^2}}{\sigma^2}\right)\left(\sigma^2\right)^{-n/2}\exp\left(-\frac{\sum_{i=1}^n\left(y_i-\mu\right)^2}{2\sigma^2}\right)$$

$$\propto \left(\sigma^2\right)^{-a_{\sigma^2}-\frac{n}{2}-1}\exp\left(-\frac{b_{\sigma^2} + \frac{1}{2}\sum_{i=1}^n\left(y_i-\mu\right)^2}{\sigma^2}\right)$$

Thus given μ and and the data $y_1, \ldots, y_n$, the posterior distribution is distributed Inverse Gamma with parameters given below.

$$a_{\text{post}} = a_{\sigma^2} + \frac{n}{2} \qquad\qquad b_{\text{post}} = b_{\sigma^2} + \frac{1}{2}\sum_{i=1}^n\left(y_i-\mu\right)^2$$

Once the prior parameters have been selected, and the data obtained, the Gibbs Sampler involves the following steps after setting $\sigma^2 = \sigma_0^2$ as an initial value where σ_0^2 is the sample variance of the data.

$$\text{Sample } \mu\text{: } \mu|\sigma^2, y_1, \ldots, y_n \sim N\left(\frac{m\mu_\mu + n\overline{y}}{m+n}, \frac{\sigma^2}{m+n}\right)$$

	μ				σ			
Year	Mean	2.5%	50%	97.5%	Mean	2.5%	50%	97.5%
2006	4.389	-1.877	4.409	10.700	22.527	18.500	22.338	27.632
2007	1.321	-3.593	1.319	6.257	17.567	14.440	17.408	21.524
2008	-1.271	-5.201	-1.269	2.589	13.995	11.496	13.871	17.213
2009	-2.311	-7.500	-2.322	2.949	18.805	15.444	18.632	23.161
2010	2.173	-1.939	2.178	6.292	14.704	12.101	14.569	18.069
2011	-0.739	-5.288	-0.751	3.834	16.343	13.446	16.191	20.063
2012	5.124	-0.292	5.121	10.439	19.214	15.818	19.047	23.650
2013	3.319	-1.711	3.323	8.341	17.971	14.793	17.817	22.037
2014	-1.010	-6.654	-1.030	4.760	20.324	16.720	20.131	25.009
2015	0.595	-4.540	0.593	5.696	18.151	14.893	18.003	22.198
2016	-1.315	-5.529	-1.304	2.886	15.002	12.349	14.884	18.374
2017	2.473	-3.369	2.479	8.221	20.654	17.008	20.472	25.221
2018	2.279	-4.025	2.304	8.464	22.217	18.250	22.028	27.273

Table 3.8: Posterior means and quantiles for NBA Total Points - Over/Under differentials for 2006/07-2018/19 seasons based on samples of $n = 50$ games per year

$$\text{Sample } \sigma^2\text{: } \sigma^2|\mu, y_1, \ldots, y_n \sim IG\left(a_{\sigma^2} + \frac{n}{2}, b_{\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

Note that for sampling the variance, it is generally simpler to sample the precision ($1/sigma^2$) from the gamma distribution with parameters a_{post} and b_{post}.

Example 3.19: NBA Actual and Predicted Point Totals from Over/Under Lines - 2006/07-2018/19

In this example, we consider the difference between the actual combined scores and Over/Under betting lines for samples of NBA games during the 2006/07 through 20018/19 seasons. We will take samples of size $n = 50$ from each season and obtain 25000 MCMC samples from the Gibbs Sampler. We will use non-informative priors for the mean and variance.

$$\mu_\mu = 0 \qquad \sigma^2_\mu = 100\sigma^2 = \frac{\sigma^2}{0.01} \qquad\qquad a_{\sigma^2} = b_{\sigma^2} = 0.01$$

The mean and quantiles(.025, .5, .975) for μ and σ are given in Table **??**. Note that the distributions shift quite a bit by season. Samples from the joint posterior distribution of μ and σ and histograms for μ and σ for 2006/07, 2012/13, and 2018/19 are given in Figure **??**.

∇

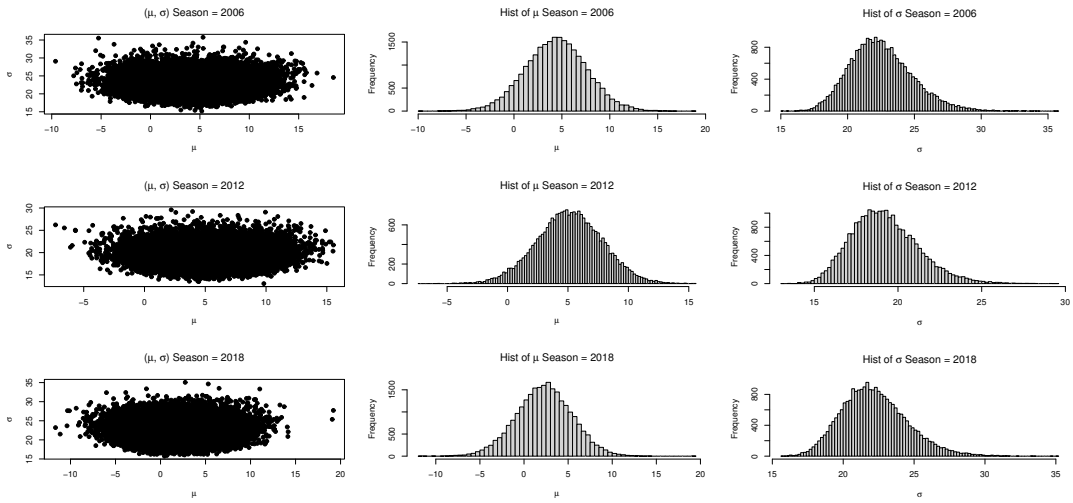


Figure 3.23: Posterior samples of (μ, σ), histograms of samples from μ and σ for 2006/07, 2012/13, and 2018/19 seasons for Total Points - Over/Under differential based on samples of $n = 50$ games per year

Chapter 4

Inferences Concerning a Single Population

Researchers often are interested in making statements regarding unknown population means, medians, and proportions 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 distributions of the sample mean and proportion when making statements regarding the population mean and proportion. Inferences regarding a population median are based on the binomial distribution.

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

4.1 Inference Concerning a Population Mean

4.1.1 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.

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

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

Some commonly used 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.

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

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

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

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

Example 4.1: WNBA Team Predicted and Actual Points

The differences between the actual points scored and predicted points based on the point spread and Over/Under for 4078 WNBA regular season team-games during 2010-2019 are obtained by taking (Point Spread + Over/Under)/2, where again the favored team has a positive point spread. We will call these differences (Actual-Predicted) "team point prediction errors." These differences are approximately normally distributed with mean $\mu = 0.0525$ and standard deviation $\sigma = 10.1062$. The P-value for the Shapiro-Wilk test for normality is $p = .3548$, implying no evidence of non-normality of the differences.

Year	n	Mean	SD	Lower Bound	Upper Bound
2010	408	0.2684	10.2038	-0.7247	1.2614
2011	408	-0.9975	9.8182	-1.9531	-0.0420
2012	408	0.1275	9.6650	-0.8132	1.0681
2013	408	-0.8480	10.1696	-1.8378	0.1417
2014	408	0.2990	9.7827	-0.6531	1.2511
2015	408	-0.2892	10.4884	-1.3100	0.7315
2016	408	0.9350	9.9229	-0.0307	1.9008
2017	408	0.2880	10.3358	-0.7179	1.2939
2018	406	0.6613	10.4139	-0.3547	1.6773
2019	408	0.0833	10.1730	-0.9067	1.0734

Table 4.1: Summary statistics and 95% Confidence Intervals for mean team prediction errors by year - WNBA 2010-2019

Note that the actual scores have mean $\mu_A = 78.7876$ and standard deviation $\sigma_A = 11.5745$, and the predicted scores have mean $\mu_A = 78.7352$ and standard deviation $\sigma_P = 5.8566$, and the correlation between actual and predicted scores is $\rho_{AP} = .4878$. This leads to the following direct computation of the mean, variance, and standard deviations among the team point prediction errors $Y = A - P$ based on rules for linear functions of random variables.

$$E\{Y\} = \mu_Y = E\{A-P\} = \mu_A - \mu_P = 78.7876 - 78.7352 = 0.0524$$

$$V\{Y\} = \sigma^2_Y = V\{A-P\} = \sigma^2_A + \sigma^2_P - 2\rho_{AP}\sigma_A\sigma_P = 11.5745^2 + 5.8566^2 - 2(.4878)(11.5745)(5.8566) = 102.1356$$

$$\Rightarrow \quad \sigma_Y = 10.1062$$

First, we consider this population of team games as a sample from a conceptual population of all possible team games that could be played. We obtain a 95% Confidence Interval for the population mean as follows with $\overline{y} = 0.0525$, $s = 10.1062$, $n = 4078$, and $t_{.025,4078-1} = 1.9605$.

$$\overline{y} \pm t_{.025,n-1}\frac{s}{\sqrt{n}} \quad \equiv \quad 0.0525 \pm 1.9605\frac{10.1062}{\sqrt{4078}} \quad \equiv \quad 0.0525 \pm 0.3103 \quad \equiv \quad (-0.2578, 0.3628)$$

The interval contains 0, showing no evidence of a systematic "bias" in team point predictions. Results by year are given in Table **??**. Only for 2011, when the average team point prediction error was close to -1, does the 95% Confidence Interval not contain 0.

Now we treat this as a population of games, and consider taking (pseudo) random samples and observing the properties of the Confidence Intervals in repeated sampling. We take 10000 random samples of size $n = 12$, implying a standard error of $\sigma_{\overline{Y}} = 10.1062/\sqrt{12} = 2.9174$. 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, 9044 (90.44%) lied within $\mu \pm 1.645(2.9174)$, 9539 (95.39%) within $\mu \pm 1.96(2.9174)$, and 9913 (99.13%) within $\mu \pm 2.576(2.9174)$. Had we constructed intervals of the form $\overline{y} \pm z_{\alpha/2}(2.9174)$ for each sample mean, the coverage rates for μ would have been the same values (90.44%, 95.39%, 99.13%).

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 87.90%, 92.77%, and 97.38%, 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: 90.58%, 95.04%, and 98.88%, respectively.

For the first random sample of the 10000 generated, we observe $\overline{y} = 3.6875$ and $s = 13.8929$. The 95% Confidence Interval for μ based on the first sample is obtained as follows.

$$\overline{y} \pm t_{.025,n-1}\frac{s}{\sqrt{n}} \quad \equiv \quad 3.6875 \pm 2.201\left(\frac{13.8929}{\sqrt{12}}\right) \quad \equiv \quad 3.6875 \pm 8.8272 \quad \equiv \quad (-5.1397, 12.5147)$$

Thus, this interval does contain $\mu = 0.0525$.

∇

Controlling the Error in Estimation with a Fixed Confidence Level

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 of the population mean with confidence level $1-\alpha$. Note that this makes the width of a $(1-\alpha)100\%$ Confidence Interval be $2E$. This can be done in one calculation based on using the z distribution, or more conservatively, by trivial iteration based on the t-distribution. Either way, we must have an approximation of σ based on previous research or a pilot study.

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

Example 4.2: Estimating Population Mean Team Point Prediction Error

Suppose we want to estimate the population mean of the WNBA point prediction errors within $E = 2$ points with 95% confidence. We treat the standard deviation as known, $\sigma = 10.1062$. 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.

$$z: \quad z_{.025} = 1.96 \quad n = \left(\frac{1.96(10.1062)}{2}\right)^2 = 98.09 \approx 99$$

R Commands and Output

```
## Commands

E <- 2.0
sigma <- 10.1062
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,] 101 1.995091
```

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 (99 and 101, respectively).

∇

4.1.2 Hypothesis Testing

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

Null Hypothesis (H_0) Statement regarding a parameter that is to be tested. 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 the **significance level** and is often set at 0.05).

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

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.

$$\overline{Y} \dot{\sim} N\left(\mu_0, \frac{\sigma^2}{n}\right) \qquad \frac{\overline{Y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

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.

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

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

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

The form of the rejection regions are given for 2-tailed, Upper and Lower tailed tests in Figure **??**. 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.

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

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

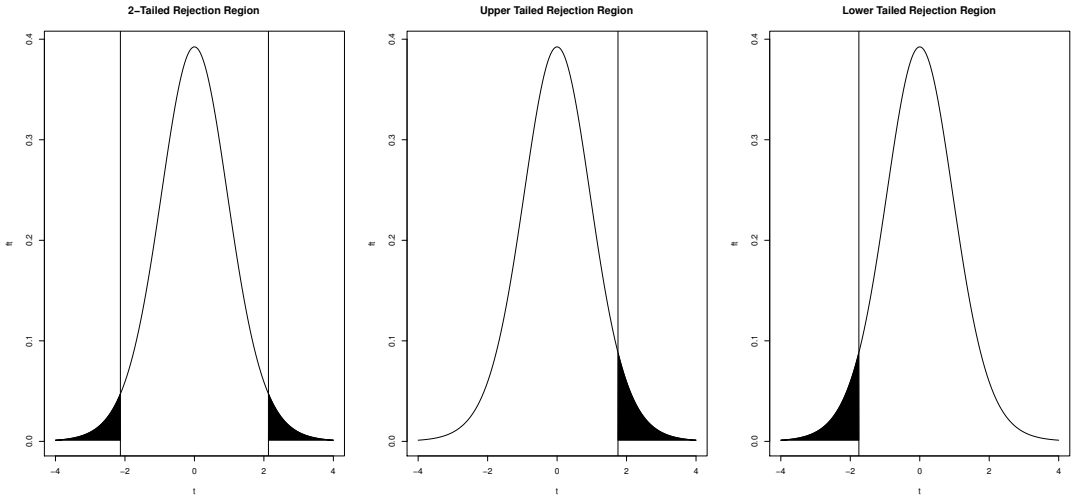


Figure 4.1: Rejection Regions for 2-tailed, Upper and Lower tailed tests, with $\alpha = 0.05$ and $n = 16$

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

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

As 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 the WNBA team point prediction errors (which are paired differences).

Example 4.3: WNBA Team Point Prediction Errors

For the WNBA team point prediction errors, the population mean was $\mu = 0.0524$ points per game with standard deviation of $\sigma = 10.1062$. Note that when conducting the tests in the repeated samples, we will use the sample standard deviation s, not the population standard deviation $\sigma = 10.1062$ in the tests below. We will demonstrate hypothesis testing regarding a single mean by first testing $H_0 : \mu = 0.0525$ versus $H_A : \mu \neq 0.0525$, 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 486 (.0486) 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 **??**. The two vertical bars on the t-statistic plot are at $\pm t_{.025,39} = \pm 2.023$.

Next we consider cases where the null hypothesis is not true. We consider $H_{01} : \mu = 0$ versus $H_{A1} : \mu \neq 0$ and $H_{02} : \mu = 2.0$ versus $H_{A2} : \mu \neq 2.0$. Since the null value for H_{01} is closer to the true value $\mu_A = 0.0524$ than the null value for H_{02}, we will expect that we reject H_{01} less often for tests based on the same sample size. That is, the power is higher for H_{02} than H_{01}. The non-centrality parameters and the corresponding

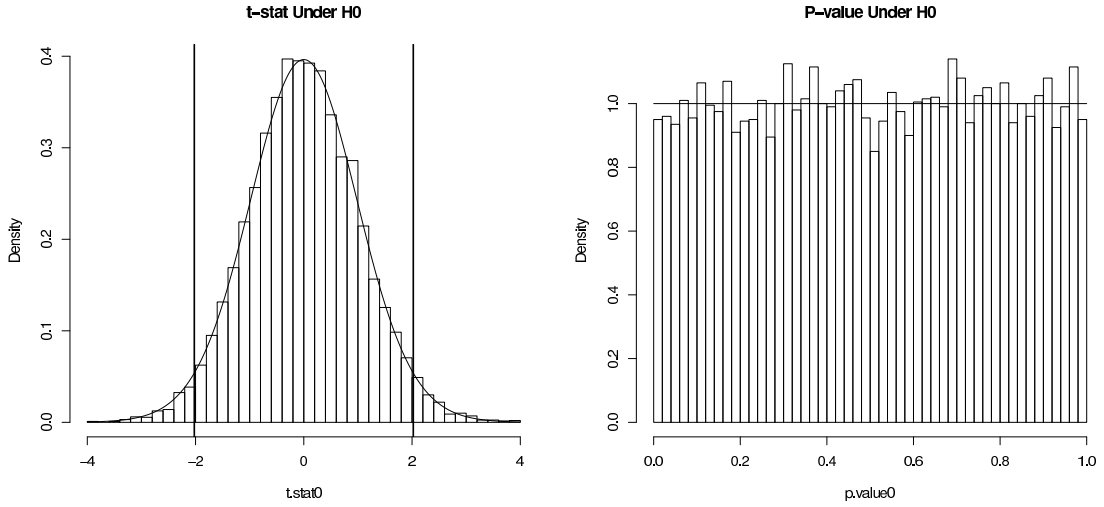


Figure 4.2: t-statistics and P-values for testing $H_0 : \mu = 0.0524$

power values are given below, based on samples of $n = 40$. Note that $\mu_{01} = 0$ is only 0.0328 standard errors from $\mu = 0.0524$, so the power is only slightly larger than $\alpha = 0.05$.

$$\Delta_1 = \frac{0.0524 - 0.0}{10.1062/\sqrt{40}} = 0.0328 \qquad \pi_1 = .0501 \qquad\qquad \Delta_2 = \frac{0.0524 - 2.0}{1.058/\sqrt{40}} = -1.2188 \qquad \pi_2 = .2212$$

Based on 10000 random samples from the WNBA team point prediction errors, 4.93% rejected $H_0 : \mu = 0$, and for another set of 10000 random samples, 22.79% rejected $H_0 : \mu = 2.0$. The histogram of the test statistics and the non-central t-distribution are given in Figure **??** for testing $H_0 : \mu = 2$.

∇

Choosing Sample Size for Fixed Power for an Alternative

Once an important difference $\mu_A - \mu_0$ 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 $\mu_A - \mu_0$ and appropriate σ. Alternatively, the difference can be in units of σ: $(\mu_A - \mu_0)/\sigma$.

2. Start with a 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 $\Delta = (\mu_0 - \mu_A)/\left(\sigma/\sqrt{n}\right) = \sqrt{n}\left(\frac{\mu_0 - \mu_A}{\sigma}\right)$.

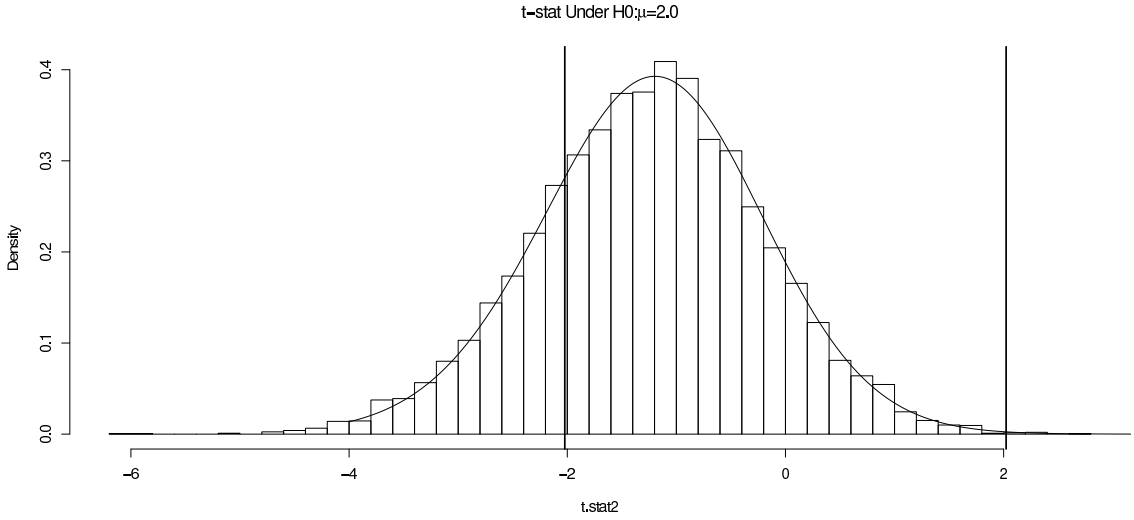


Figure 4.3: t-statistics and non-central t-distribution for testing $H_0 : \mu = 2.0$

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 higher).

Example 4.4: WNBA Team Point Prediction Errors

Suppose we would like to be able to detect a difference between μ_A and μ_0 of 1.25 points with power of $\pi = 0.8$ when the test is conducted at $\alpha = 0.05$. In this case, recall $\sigma = 10.1062$. Start with $n = 3$.

$$t_{.025,3-1} = 4.303 \qquad \Delta = \frac{1.25}{10.1062/\sqrt{3}} = \sqrt{3}\frac{1.25}{10.1062} = 0.214$$

$$\pi = P\left(t_{3-1,0.214} \leq -4.303\right) + P\left(t_{3-1,0.0.214} \geq 4.303\right) = .0177 + .0344 = .0521$$

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 $n = 515$ to meet the power requirement. The target difference is very small (1.25) relative to the standard deviation (10.1062) which is why such a large sample would be needed. A plot of the central and non-central t-distributions for n=25, 100, 300, and 550 is given in Figure **??**. The vertical bars give the critical values for the $\alpha = 0.05$ level test.

R Commands and Output

```
## Commands

alpha <- 0.05
```

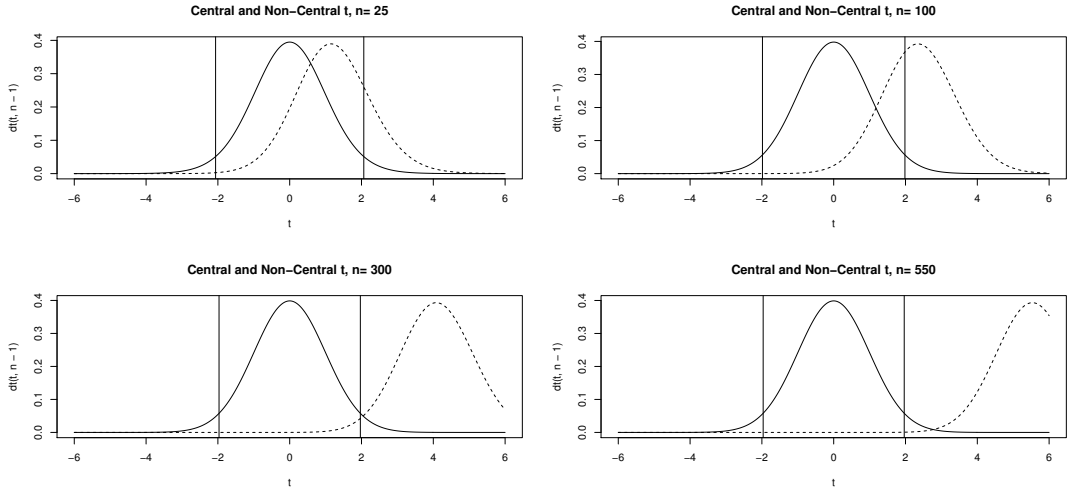


Figure 4.4: Central and non-Central t-distributions for n=25, 100, 300, 550, $\mu_0 - \mu_A = 1.25$, and $\sigma = 10.1062$

```
power.target <- 0.80
sigma <- sd(teamPts.dev)
n <- 3
mu_diff <- 1.25
Delta <- sqrt(n) * mu_diff / sigma

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 <= power.target) {
   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)

win.graph(height=5.5, width=7.0)
par(mfrow=c(2,2))
t <- seq(-6,6,0.01)
mu_diff <- 0.25
sigma <- 1.058
for (n in c(25, 100, 300, 550)) {
  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(alpha/2,n-1))
  abline(v=qt(1-alpha/2,n-1))
}

## Output

> (power <- pt(CV_LO,n-1,Delta) + (1-pt(CV_HI,n-1,Delta)))
[1] 0.05212313

> cbind(n, power)
```

```
      n     power
[1,] 515 0.8000167
```

∇

4.2 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 $P\left(Y \leq B_{\alpha/2,n} | Y \sim Bin(n, 0.5)\right) \leq \alpha/2$. 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 $(n/2) \pm \sqrt{n}$. 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.

Example 4.5: NFL Total Points Over/Under Differential

For the NFL regular seasons in 2010-2019, there were a total of 5120 games. We compute the difference between the actual total points scored and the over/under betting line. Note the following summary statistics for the actual total points and the over/under betting lines. Note that the medians (44.0 vs 44.5) and means (45.29 vs 44.87) are very similar, but the actual scores have much more variation, with standard deviations (13.93 vs 4.25), and the correlation is only 0.2952. It should be noted that the mean difference is equal to difference of the means (45.29-44.87=0.42), but the median difference is not necessarily equal to difference in the medians. In fact, the median difference for the 5120 games is 0.0.

```
> summary(teamPts+oppPts); sd(teamPts+oppPts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   36.00   44.00   45.29   54.00  105.00
[1] 13.92679
> summary(OU); sd(OU)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  33.00   42.00   44.50   44.87   47.50   63.50
[1] 4.245796
> cor(teamPts+oppPts, OU)
[1] 0.2951634

> totPts.dev <- teamPts + oppPts - OU
> summary(totPts.dev)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
```

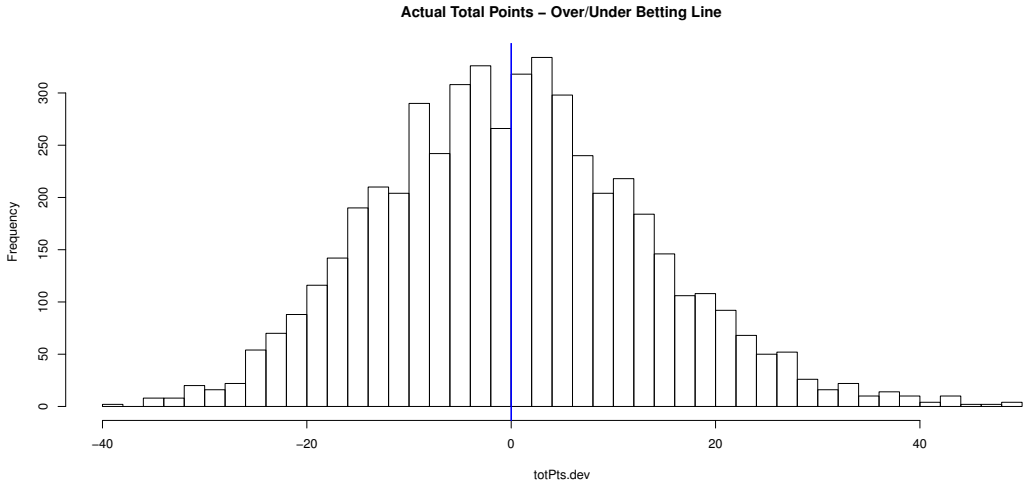


Figure 4.5: Total points and Over/Under differences - NFL 2010-2019

```
-39.5000  -9.0000   0.0000   0.4203   9.0000  49.5000
> sd(totPts.dev)
[1] 13.30699
```

A histogram of the differences is given in Figure **??**. 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 games, we would take the range encompassed by the 6^{th} through the 15^{th} games.

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

```
> (totPts.dev.sample.order <- sort(totPts.dev.sample)) ## Sample values sorted
 [1] -23.0 -17.0 -15.5 -12.0 -11.0 -11.0  -8.5  -8.0  -7.0  -5.5  -3.0  -3.0
[13]  -2.0   0.0   1.5   5.0   5.5  15.5  16.5  17.5
> cbind(totPts.dev.sample.order[6],
+       totPts.dev.sample.order[15])  ## 6th and 15th selected
     [,1] [,2]
[1,]  -11  1.5
```

For this sample, we obtain the 95% Confidence Interval: (-11, 1.5), which does contain the population median (0.0). We now obtain 10000 random samples of size $n = 20$, and count the number that contain 0.0.

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 0.0. 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 Commands and Output

```
## Commands
pbinom(0:20,20,0.5)

totPts.dev <- teamPts + oppPts - OU

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

set.seed(7654)
num.sim <- 10000
num.samp <- 20
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(totPts.dev[sample])[6]
    med.ci[i,2] <- sort(totPts.dev[sample])[15]
}

med.pop <- median(totPts.dev)
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

> (totPts.dev.sample.order <- sort(totPts.dev.sample)) ## Sample values sorted
 [1] -23.0 -17.0 -15.5 -12.0 -11.0 -11.0  -8.5  -8.0  -7.0  -5.5  -3.0  -3.0
[13]  -2.0   0.0   1.5   5.0   5.5  15.5  16.5  17.5
> cbind(totPts.dev.sample.order[6],
+       totPts.dev.sample.order[15])  ## 6th and 15th selected
     [,1] [,2]
[1,]  -11  1.5

> med.pop <- median(totPts.dev)
> sum(med.ci[,1] <= med.pop & med.ci[,2] >= med.pop) / num.sim
[1] 0.9611
```

∇

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. Any observations that are exactly equal to the median are removed and the sample size is adjusted to the number of cases strictly above or below the median.

2-tailed tests: $H_0 : M = M_0 \quad H_A : M \neq M_0 \quad T.S. : B_{obs} \quad R.R. : B_{obs} \leq B_{\alpha/2,n}$ 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 4.6: NFL Total Points Over/Under Differential

Suppose we wanted to test whether the population median total point differential, M, differs from $M_0 = 0$ points based on a sample of games. In this instance, the null hypothesis is true. Based on the sample of $n = 20$ games obtained previously, we have the following point differential values.

```
> (totPts.dev.sample.order <- sort(totPts.dev.sample)) ## Sample values sorted
 [1] -23.0 -17.0 -15.5 -12.0 -11.0 -11.0  -8.5  -8.0  -7.0  -5.5  -3.0  -3.0
[13]  -2.0   0.0   1.5   5.0   5.5  15.5  16.5  17.5
```

Note that for one game, the point differential was 0, with the actual total points being equal to the over/under betting line. Now, we have $n = 19$, and for the Binomial distribution, with $n = 19$ and $\pi = 0.5$, we obtain the following (partial) cumulative probability distribution. For a 2-tailed test with $\alpha = 0.05$, we reject the null hypothesis if $B_{obs} \leq 4$ or if $B_{obs} \geq 15$, with actual α level of $.0096 + (1 - .9904) = .0192$. If we chose to extend the rejection region to include 5 and 14, the actual α level would be $.0318 + (1 - .9682) = .0636$.

```
 y P(Y<=y)
 4 0.0096
 5 0.0318
 6 0.0835
   ...
12 0.9165
13 0.9682
14 0.9904
```

The test statistic is $B_{obs} = 6$. Note that the 2-sided P-value is $P = 2P(Y \leq 6|Y \sim Bin(19, 0.5)) = 2(.0835) = .1670$.

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

$$z_{obs} = \frac{6 - (19/2)}{0.5\sqrt{19}} = \frac{-3.5}{2.179} = -1.606 \qquad \text{2-tailed } P\text{-value: } P = 2P(Z \geq 1.606) = 2(.0541) = .1082$$

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 adding of the 0.5 is to get all the area under 6 for binomial, since 6 is below its expected value. This results in virtually the same P-value as the exact Binomial. As n gets large, the correction makes little difference.

$$z_{obs} = \frac{6 + 0.5 - (19/2)}{0.5\sqrt{19}} = \frac{-3.0}{2.179} = -1.376 \qquad \text{2-tailed } P\text{-value: } P = 2P(Z \geq 1.376) = 2(.0843) = .1686$$

4.3 Inference Concerning a Population Proportion

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

4.3.1 Variables with Two Possible Outcomes

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

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

In small samples, it has been seen that making an adjustment to the success count and the sample size performs well. This is referred to as the **Wilson-Agresti-Coull** method (Agresti and Coull (1998), [**?**]). Let y be the observed number of successes in the n trials, then the Confidence Interval is obtained as follows.

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

For $\alpha = 0.05$, $z_{\alpha/2} = 1.96 \approx 2$ which leads to a simpler way of adding 2 successes and 2 failures to the sample data, and $\tilde{y} = y + 2$ and $\tilde{n} = n + 4$.

Exact $(1-\alpha)100\%$ Confidence Intervals can be obtained by determining the range of values π for which $P(Y \leq y|Y \sim Bin(n,\pi)) \geq \alpha/2$ and $P(Y \geq y|Y \sim Bin(n,\pi)) \geq \alpha/2$. For instance, if we have $n = 50$ trials (say games) and have $y = 22$ successes (say the favored team covers the spread), we can compute $P(Y \leq 22|Y \sim Bin(n,\pi))$ and $P(Y \geq 22|Y \sim Bin(n,\pi)) = 1 - P(Y \leq 21|Y \sim Bin(n,\pi))$ for a range of pi values (say .01 to .99 by .01). For these values we obtain in R the following values for $n = 50$ and $y = 22$.

$\pi = 0.29$: $P(Y \leq 22) = .9920$ $P(Y \geq 22) = .0170$ $\pi = 0.30$: $P(Y \leq 22) = .9877$ $P(Y \geq 22) = .0251$

$\pi = 0.58$: $P(Y \leq 22) = .0320$ $P(Y \geq 22) = .9836$ $\pi = 0.59$: $P(Y \leq 22) = .0229$ $P(Y \geq 22) = .9887$

So for 0.30 and 0.58, both of the "tail probabilities" exceed $\alpha/2 = 0.05/2 = .0250$, and for 0.29 and 0.59, one of the probabilities is below .025. Using the **prop.test** function in R, we obtain the following results, including a P-value for a hypothesis test of whether $\pi = 0.5$ (described below). Note that R uses a finer grid for the possible π values, and gives a 95% CI for π as $(0.3027, 0.5865)$.

```
n <- 50
y <- 22
pi_0 <- 0.5
prop.test(y, n, pi_0)

## Output

> prop.test(y, n, pi_0)

        1-sample proportions test with continuity correction

data:  y out of n, null probability pi_0
X-squared = 0.5, df = 1, p-value = 0.4795
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3026605 0.5865007
```

```
sample estimates:
   p
0.44
```

A Large-sample (Wald) test of whether $\pi = \pi_0$ can also be conducted. For instance, a test may be whether a majority of people favor a political candidate or referendum, or whether a defective rate is below some tolerance level. Note that in the z-statistic, we use the standard error of $\hat{\pi}$ evaluated at the null value π_0, and not the estimated standard error that was used for the Confidence Interval.

$$\hat{SE}\left\{\hat{\pi}\right\} = \sqrt{\frac{\hat{\pi}\left(1-\hat{\pi}\right)}{n}} \qquad SE\left\{\hat{\pi}\right\}_{H_0} = \sqrt{\frac{\pi_0\left(1-\pi_0\right)}{n}}$$

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

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

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

An alternative large sample test that can be conducted is a **likelihood ratio test**. This was used in the football betting literature by Even and Noble (1992) [**?**] to test whether the probability that the actual point spread for a game exceeds the Vegas Line is equal to 1/2. This paper was discussed further by Gander, Zuber, and Russo (1993) [**?**].

The likelihood ratio test involves evaluating the log of the **likelihood function** of the parameter π at the value of π under the null hypothesis (π_0) and at the value that maximizes the likelihood ($\hat{\pi}$), where the likelihood is the probability function for the random variable Y (the number of successes in n trials) as a function of π and the observed number of successes y.

$$L(\pi|y) = P(Y = y|n, \pi) = p(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$$

The estimator $\hat{\pi}$ can be derived by taking the derivative of $L(\pi|y)$ with respect to π, setting it equal to zero, and solving for $\hat{\pi}$. However it is easier in the case of the Binomial distribution (and many others) to work with $l = \log(L)$ in obtaining the **Maximum Likelihood Estimator (MLE)**. In the case of the Binomial distribution, we have the following results.

$$l(\pi|y) = \log(L(\pi|y)) = \log\left(\binom{n}{y}\right) + y\log(\pi) + (n-y)\log(1-\pi) \quad \Rightarrow \quad \frac{dl}{d\pi} = \frac{y}{\pi} - \frac{n-y}{1-\pi}$$

Setting the derivative to zero leads to the MLE for π, which we have used directly before in the large-sample confidence intervals and Wald test.

$$\frac{dl}{d\pi} = \frac{y}{\pi} - \frac{n-y}{1-\pi} = 0 \quad \Rightarrow \quad \frac{y}{\hat{\pi}} - \frac{n-y}{1-\hat{\pi}} \quad \Rightarrow \quad \hat{\pi} = \frac{y}{n}$$

The likelihood ratio test takes the difference between the log likelihood evaluated at π_0 and evaluated at $\hat{\pi}$ and multiplies the difference by -2. For large samples, under the null hypothesis the statistic is approximately Chi-square with 1 degree of freedom. We use the natural logarithm when we numerically take logs (the default in R).

$$H_0 : \pi = \pi_0 \quad H_A : \pi \neq \pi_0 \quad \text{TS: } X^2_{LR} = -2[l(\pi_0) - l(\hat{\pi})] \quad \text{RR: } X^2_{LR} \geq \chi^2_{\alpha,1} \quad P = P(\chi^2_1 \geq X^2_{LR})$$

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

Example 4.7: NBA Point Spread and Over/Under Outcomes

For the 2006/7-2018/9 NBA regular seasons games, the home team covered the spread in 7626 games, failed to cover the spread in 7855 games, and "pushed" the spread in 268 games. We treat these games as a sample of the infinite population of games that could be played among NBA teams, and eliminate the 268 "pushes." The test is whether the true underlying probability that the home team covers is 0.50. Otherwise bettors could have an advantage over bookmakers. $H_0 : \pi = 0.50$ versus $H_A : \pi \neq 0.50$. We will conduct the large-sample Wald Z-test, the likelihood ratio test, and obtain the exact Binomial P-value.

$$y = 7626 \qquad n = 7626 + 7855 = 15481 \qquad \hat{\pi} = \frac{7626}{15481} = 0.4926$$

$$SE\left\{\hat{\pi}\right\}_{H_0} = \sqrt{\frac{\pi_0\left(1-\pi_0\right)}{n}} = \sqrt{\frac{0.5(1-0.5)}{15481}} = 0.0040$$

$$z_{obs} = \frac{0.4926 - 0.5}{0.0040} = -1.85 \qquad P = 2P(Z \geq 1.85) = 2(0.03216) = 0.0643$$

$$l(\pi) = \log\left(\binom{n}{y}\right) + y\log(\pi) + (n-y)\log(1-\pi) \quad \Rightarrow$$

$$l(\pi_0) = \log\left(\binom{15481}{7626}\right) + 7626\log(0.5) + (7855)\log(1-0.5) = c + (-5285.94) + (-5444.67) = c - 10730.61$$

$$l(\hat{\pi}) = c + 7626\log(0.4926) + (7855)\log(1-0.4926) = c + (-5399.65) + (-5329.27) = c - 10728.92$$

This leads to the likelihood ratio test given below.

TS: $\chi^2_{LR} = -2[(c-10730.61)-(c-10728.92)] = -2(-1.69) = 3.38$ RR: $\chi^2_{LR} \geq \chi^2_{.05,1} = 3.84$ $P = P(\chi^2_1 \geq 3.38) = .0660$

There is no evidence of a "bias" in terms of the home team covering the spread (at the $\alpha = 0.05$ significance level). Note that this is a tremendous "sample," and it would be quite easy to detect even a very small difference for $H_0 : \pi = 0.50$. An exact test is given here. Under the null hypothesis, the expected value of Y is $n\pi_0 = 15481(0.5) = 7740.5$. The observed y is 7626, which is 114.5 below its expected value. Had y been 7855, it would have been 114.5 above its expected value. The exact 2-tailed P-value is obtained as follows.

$$P = P\left(Y \leq 7626 | Y \sim Bin(15481, 0.5)\right) + P\left(Y \geq 7855 | Y \sim Bin(15481, 0.5)\right) = 0.03344 + 0.03344 = .0669$$

A similar test can be done for the "Over/Under" bet. For the Over/Under bet for those seasons, Under won 7838 times, Over won 7716 times, and there were 195 Pushes. Again, we eliminate the Pushes, and test $H_0 : \pi = 0.50$ versus $H_A : \pi \neq 0.50$, where π is the probability Over wins.

$$y = 7716 \qquad n = 7838{+}7716 = 15554 \qquad \hat{\pi} = \frac{7716}{15554} = 0.4961 \qquad SE\left\{\hat{\pi}\right\}_{H_0} = \sqrt{\frac{\pi_0\left(1-\pi_0\right)}{n}} = \sqrt{\frac{0.5(1-0.5)}{15554}} = 0.0040$$

$$z_{obs} = \frac{0.4961 - 0.5}{0.0040} = -0.975 \qquad P = 2P(Z \geq 0.975) = 2(.1648) = 0.3296$$

$$l(\pi_0) = c - 10781.21 \quad l(\hat{\pi}) = c - 10780.73 \qquad \text{TS: } \chi^2_{LR} = -2(-0.48) = 0.96 \quad P = .3272$$

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

$$P = P\left(Y \leq 7716 | Y \sim Bin(15554, 0.5)\right) + P\left(Y \geq 7838 | Y \sim Bin(15554, 0.5)\right) = 0.16597 + 0.16597 = .3319$$

R Commands and Output

```
### Commands

nbaou <- read.csv("http://www.stat.ufl.edu/~winner/data/nbaodds201415.csv")
attach(nbaou); names(nbaou)

#### Point Spread Analysis
table(TeamCov)
TeamCov01 <- subset(TeamCov, TeamCov != 0)
table(TeamCov01)
```

```
(Y.Cov <- sum(TeamCov01[TeamCov01 == 1]))
(n.Cov <- length(TeamCov01))
pi.H0 <- 0.50
(pihat.Cov <- Y.Cov / n.Cov)
(se.pihat.Cov.CI <- sqrt(pihat.Cov * (1-pihat.Cov) / n.Cov))
(se.pihat.Cov.H0 <- sqrt(pi.H0 * (1-pi.H0) / n.Cov))
(Z.Cov.H0 <- (pihat.Cov - pi.H0) / se.pihat.Cov.H0)
(p.Cov.H0 <- 2*(1-pnorm(abs(Z.Cov.H0),0,1)))
(pihat.Cov.CI <- pihat.Cov + c(-1.96, 1.96) * se.pihat.Cov.CI)

### Exact Tests
binom.test(Y.Cov,n.Cov,p=0.5,alternative="two.sided")

### Output
> table(TeamCov)
TeamCov
 -1   0   1
615  27 588
> TeamCov01 <- subset(TeamCov, TeamCov != 0)
> table(TeamCov01)
TeamCov01
 -1   1
615 588
> (Y.Cov <- sum(TeamCov01[TeamCov01 == 1]))
[1] 588
> (n.Cov <- length(TeamCov01))
[1] 1203
> (pihat.Cov <- Y.Cov / n.Cov)
[1] 0.4887781
> (se.pihat.Cov.CI <- sqrt(pihat.Cov * (1-pihat.Cov) / n.Cov))
[1] 0.01441212
> (se.pihat.Cov.H0 <- sqrt(pi.H0 * (1-pi.H0) / n.Cov))
[1] 0.01441575
> (Z.Cov.H0 <- (pihat.Cov - pi.H0) / se.pihat.Cov.H0)
[1] -0.7784504
> (p.Cov.H0 <- 2*(1-pnorm(abs(Z.Cov.H0),0,1)))
[1] 0.4363035
> (pihat.Cov.CI <- pihat.Cov + c(-1.96, 1.96) * se.pihat.Cov.CI)
[1] 0.4605303 0.5170258
> binom.test(Y.Cov,n.Cov,p=0.5,alternative="two.sided")

        Exact binomial test

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

4.3.2 Variables with $k > 2$ Possible Outcomes

A Multinomial experiment is an extension of the Binomial experiment with the caveat that each of n trials can end in one of k possible outcomes or categories. The probability of outcome i is π_i, and the random count of the number of trials falling in category i is Y_i. The following restrictions must hold.

$$\pi_1 + \cdots + \pi_k = 1 \qquad Y_1 + \cdots + Y_k = n \qquad \pi_i \geq 0 \quad y_i \geq 0$$

Note that for the Binomial case, we have previously labeled $Y_1 = Y$ and $Y_2 = n - Y$ where category 1 represents "Success" and category 2 represents "Failure." The probability of the experiment resulting in the observed counts $(y_1, \ldots, y_k)$ is as follows.

$$P(Y_1 = y_1, \ldots, Y_k = y_k) = p(y_1, \ldots, y_k) \qquad \text{where:}$$

$$p(y_1, \ldots, y_k) = \frac{n!}{y_1! \cdots y_k!} \pi_1^{y_1} \cdots \pi_k^{y_k} \qquad \pi_1 + \cdots + \pi_k = 1 \quad y_1 + \cdots + y_k = n \quad \pi_i \geq 0 \quad y_i \geq 0$$

A test can be conducted for $H_0 : \pi_1 = \pi_{10}, \ldots, \pi_k = \pi_{k0}$, for some specified set of k probabilities $\pi_{10}, \ldots, \pi_{k0}$ that sum to 1. Again this simply extends the binomial test of $H_0 : \pi = \pi_0$. Once the observed counts $y_1, \ldots, y_k$ are obtained, the test is conducted as follows.

$$E_i = n\pi_{i0} \qquad TS : X^2_{obs} = \sum_{i=1}^{k} \frac{(y_i - E_i)^2}{E_i} \qquad RR : X^2_{obs} \geq \chi^2_{\alpha,k-1} \qquad P = P\left(\chi^2_{k-1} \geq X^2_{obs}\right)$$

Example 4.8: Game Outcomes of EPL Home Favorites 2000/01-2018/19

For each English Premier League game for the seasons beginning in 2000-2018, based on odds converted to probabilities using basic normalization, we consider the games, where the home team's (subjective) probability of winning is at least 1/2. We are now treating these games as a sample of all possible games that could have been played among EPL teams during these seasons. There are probabilities for home win, away win, and draw. Another way of declaring that the home team is favored is if its probability was the highest of the three. Thus, we are choosing a higher threshold for the home team to be considered the favorite here. We consider the following (somewhat arbitrary) probabilities for the game outcomes (Home team Wins (1) is 1/2, Draws (2) is 1/3, Loses (3) 1/6). Note that these probabilities were "conjured" before looking at the data!

The null hypothesis and observed and expected counts are given below, where the categories with respect to the home team are (Win=1, Draw=2, Lose=3). There were $n = 2429$ games where the home team's probability of winning was at least 1/2.

$$H_0 : \pi_1 = 1/2 = 3/6 \quad \pi_2 = 1/3 = 2/6 \quad \pi_3 = 1/6$$

$$E_1 = 2429(1/2) = 1214.5 \quad E_2 = 2429(1/3) = 809.7 \quad E_3 = 2429(1/6) = 404.8 \qquad y_1 = 1635 \quad y_2 = 502 \quad y_3 = 292$$

Category (i)	π_{i0}	Expected #	Observed #	X^2
Win (1)	3/6	1214.5	1635	145.591
Draw (2)	2/6	809.7	502	116.931
Lose (3)	1/6	404.8	292	31.432
Total	1	2429	2429	293.41

Table 4.2: Numbers of Wins/Draws/Losses for EPL home favorites 2000-2018 seasons

The test statistic is $X^2_{obs} = 293.95$ with $k - 1 = 3 - 1 = 2$ degrees of freedom (based on R commands given below). The critical Chi-square value is $\chi^2_{.05,2} = 5.991$, and the P-value for the test is .0000. There is strong evidence that the hypothesized distribution is not the true distribution. Calculations are shown in Table **??** (using hand calculator and different rounding). The observed proportions are given below.

$$\hat{\pi}_1 = \frac{1635}{2429} = 0.673 \qquad \hat{\pi}_2 = \frac{502}{2429} = 0.207 \qquad \hat{\pi}_3 = \frac{292}{2429} = 0.120$$

R Commands and Output

```
### Commands
epl <- read.csv("http://www.stat.ufl.edu/~winner/data/EPL_2000_2018a.csv")
attach(epl); names(epl)

## Determine if focal team has prob of win >= 1/2 (1 if Yes, 0 if No)
FAVtm <- ifelse(BNtm >= 0.5,1,0)
## Determine game outcome for focal team (1=Win,2=Draw,3=Lose)
OTCMtm <- ifelse(FTGtm>FTGop,1,ifelse(FTGtm==FTGop,2,3))

## Game outcomes for Home Favorites and number of such games
table(OTCMtm[tmHome==1 & FAVtm ==1])
sum(table(OTCMtm[tmHome==1 & FAVtm ==1]))

## Chi-square test for the hypothesis and data
chisq.test(c(1635,502,292),p=c(1/2,1/3,1/6))

### Output
> table(OTCMtm[tmHome==1 & FAVtm ==1])
   1    2    3
1635  502  292
> sum(table(OTCMtm[tmHome==1 & FAVtm ==1]))
[1] 2429

> chisq.test(c(1635,502,292),p=c(1/2,1/3,1/6))

        Chi-squared test for given probabilities
data:  c(1635, 502, 292)
X-squared = 293.95, df = 2, p-value < 2.2e-16
```

∇

This test is often used when testing whether data come from a particular family of probability distributions.

y	Expected	Observed	χ^2	y	Expected	Observed	χ^2
0	68.89	102	15.92	7	1878.01	1808	2.61
1	373.36	296	16.03	8	1272.35	1324	2.10
2	1011.82	1123	12.22	9	766.24	722	2.55
3	1828.02	1538	46.01	10	415.30	411	0.04
4	2476.96	2684	17.31	11	204.63	210	0.14
5	2685.03	2590	3.36	12	92.42	74	3.67
6	2425.48	2637	18.45	≥ 13	61.49	41	6.83
				Sum	15560	15560	147.24

Table 4.3: Goodness of fit test for Poisson distribution of Number of goals during regulation playing time - NHL 2006/7-2018/9

- A family of distributions (e.g. Poisson, Negative Binomial, Normal, Gamma, Beta) is considered for the data.
- Data are sampled and used to estimate the m parameter(s) of the distribution.
- The data are placed in k mutually exclusive and exhaustive ranges of values.
- The observed counts n_i and the expected counts under the hypothesized distribution are obtained (expected counts should be ≥ 5)
- The chi-square statistic is computed and has $k-1-m$ degrees of freedom under the null hypothesis that the distribution is appropriate for the data.

Example 4.9: NHL Total Goals in Regulation Time

Referring back to Exercise 3.10, we considered total goals in regulation time for the 15560 games played during the 2006/07-2018/19 NHL seasons. The mean and variance are $\mu_Y = 5.42$ and $\sigma^2_Y = 5.23$, respectively. While they are not exactly the same, they are reasonably close. The probability distribution, Expected, and Observed counts are given in Table **??**. The Expected counts are obtained by multiplying the probabilities by the total number of games (15560). The Observed and Expected Counts are similar, but not a perfect match.

Here we conduct a goodness of fit test is described that can be used to assess the reasonability of the Poisson approximation to the distribution. Table **??** gives the counts, and their expected counts ($15560p(y)$) under the Poisson distribution with $\lambda = 5.42$ for the occurrences of 0 goals, 1 goal, ..., ≥ 13 goals. We do not include the probabilities that are given in Table **??**. The test statistic is $X^2_{obs} = 147.24$ with degrees of freedom $k-1-m = 14-1-1 = 12$. The critical value for $\alpha = 0.05$ is $\chi^2_{.05,12} = 21.026$, and the P-value is $P\left(\chi^2_{12} \geq 147.24\right) = .0000$. There is strong evidence that the Poisson distribution is **not** a good fit for the distribution of total goals.

R Commands and Output

```
### Commands
nhldata <- read.csv("data/nhl20062018_opponent_ML.csv")
attach(nhldata); names(nhldata)

## Games where the focal team was home team
```

```
nhl_home <- nhldata[home == 1,]
detach(nhldata); attach(nhl_home)

## Total Score and TS in Regulation
totScore <- teamScore + oppScore
totScoreReg <- ifelse(ot==1,totScore-1,totScore)

## Table of counts of Total Scores in Reg and mean, variance, n
(tScoreTable <- table(totScoreReg))   # freq dist of total goals in reg
(lambda <- mean(totScoreReg))
var(totScoreReg)
n.games <- length(totScoreReg)

## Poisson probs for 0-12 and 13+
pois.dist13p <- c(dpois(0:12, lambda),1-sum(dpois(0:12, lambda)))

# Observed and Expected counts for 0-12 and 13+
tScoreObs <- c(tScoreTable[1:13],n.games-sum(tScoreTable[1:13]))
tScoreExp <- n.games*pois.dist13p

# Chi-square Goodness of fit test
k <- length(tScoreObs)
m <- length(lambda)
X2.df <- k-1-m
X2.obs <- sum((tScoreObs-tScoreExp)^2/tScoreExp)
X2.05 <- qchisq(.95,X2.df)
X2.p <- 1-pchisq(X2.obs, X2.df)
X2.out <- cbind(X2.obs, X2.df, X2.05, X2.p)
colnames(X2.out) <- c("X2 stat", "DF", "X2(.05)", "P-value")
round(X2.out, 4)

### Output
> (lambda <- mean(totScoreReg))
[1] 5.415938

> round(X2.out, 4)
     X2 stat DF X2(.05) P-value
[1,] 147.027 12 21.0261       0
```

∇

Example 4.10: Point Spread Differential - WNBA 2010-2019

In Example 3.12, the normal distribution was used to model the point spread differential with respect to the home team for WNBA games for the 2010-2019 regular seasons. The mean of the differentials ((Home Pts - Away Pts) - Spread) is 0.0289 points and the standard deviation is 12.0472. Again, we are treating these data as a sample from a conceptual population of games. We will consider bins in terms of the mean ($\mu = 0.03$) and standard deviation ($\sigma = 12.05$).

$$(-\infty, \mu - 2.4\sigma], (\mu - 2.4\sigma, \mu - 2.0\sigma], \ldots, (\mu - 0.4\sigma, \mu], (\mu, \mu + 0.4\sigma), \ldots, [\mu + 2.0\sigma, \mu + 2.4\sigma), [\mu + 2.4\sigma, \infty)$$

The range of point spread differentials is broken into the following $k = 14$ categories:

$$(\infty, -28.88], (-28.88, -24.07], \ldots, (24.12, 28.94], (28.94, \infty).$$

Range	Observed	Expected	X^2
$(-\infty) - (-28.88)$	19	16.71478	0.3124
$(-28.88) - (-24.07)$	27	29.67274	0.2407
$(-24.07) - (-19.25)$	58	65.34824	0.8263
$(-19.25) - (-14.43)$	117	122.8913	0.2824
$(-14.43) - (-9.61)$	201	197.3461	0.0677
$(-9.61) - (-4.79)$	284	270.6219	0.6613
$(-4.79) - 0.029$	344	316.9049	2.3166
$0.029 - 4.85$	308	316.9049	0.2502
$4.85 - 9.67$	269	270.6219	0.0097
$9.67 - 14.49$	177	197.3461	2.0977
$14.49 - 19.30$	113	122.8913	0.7961
$19.30 - 24.12$	69	65.34824	0.2041
$24.12 - 28.94$	28	29.67274	0.0943
$28.94 - \infty$	25	16.71478	4.1068
Total	2039	2039	12.2664

Table 4.4: Goodness-of-fit test for WNBA 2010-2019 home spread point differential by a normal distribution

Table **??** gives the observed and expected counts, and computations for the chi-square goodness-of-fit test. The degrees of freedom are 14-1-2=11, with critical chi-square value of $\chi^2_{.05,11} = 19.675$ and P-value $P\left(\chi^2_{11} \geq 12.2664\right) = .3440$. The data are consistent with a normal distribution. R commands and output are given below.

R Commands

```
### Commands
wnba <- read.csv("data\\wnba_20102019_focal.csv")
attach(wnba); names(wnba)

wnba_home <- wnba[home == 1,]
detach(wnba); attach(wnba_home)

## teamSpDev = (Home Points - Away Points) - Home Spread
summary(teamSpDev)
(mn.hsd <- mean(teamSpDev))
(sd.hsd <- sd(teamSpDev))

## Create the bin cutoffs
z.bin <- c(seq(-2.4,2.4,0.4),10)  # The 10 is to pick up > 2.4
y.bin <- mn.hsd + sd.hsd*z.bin    # convert to units of y
k <- length(y.bin)
n <- length(teamSpDev)

## Obtain cumulative counts below bin cut-offs and convert to bin counts
y.obs.cum <- y.exp.cum <- rep(0,k)
for (i1 in 1:k) {
  y.obs.cum[i1] <- sum(teamSpDev <= y.bin[i1])
  y.exp.cum[i1] <- n * (pnorm(z.bin[i1]))
}
y.obs <- c(y.obs.cum[1], diff(y.obs.cum))
y.exp <- c(y.exp.cum[1], diff(y.exp.cum))

## Compute Chi-square statistic and obtain critical value, P-value
X2.obs.exp <- (y.obs-y.exp)^2/y.exp
X2.DF <- k-1-2
```

```
X2.TS <- sum(X2.obs.exp)
X2.CV <- qchisq(.95,X2.DF)
X2.PV <- 1-pchisq(X2.TS,X2.DF)

## Set up and print output
table.out <- cbind(y.bin,y.obs,y.exp,X2.obs.exp)
colnames(table.out) <- c("cell top", "observed", "expected", "chi-square")

test.out <- cbind(X2.DF,X2.TS,X2.CV,X2.PV)
colnames(test.out) <- cbind("df", "Test Stat", "X2(.05,df)", "P(>TS)")

round(table.out,4)
round(test.out,4)

### Output

> round(table.out,4)
      cell top observed expected chi-square
 [1,] -28.8843       19  16.7148     0.3124
 [2,] -24.0654       27  29.6727     0.2407
 [3,] -19.2466       58  65.3482     0.8263
 [4,] -14.4277      117 122.8913     0.2824
 [5,]  -9.6088      201 197.3461     0.0677
 [6,]  -4.7899      284 270.6219     0.6613
 [7,]   0.0289      344 316.9049     2.3166
 [8,]   4.8478      308 316.9049     0.2502
 [9,]   9.6667      269 270.6219     0.0097
[10,]  14.4856      177 197.3461     2.0977
[11,]  19.3044      113 122.8913     0.7961
[12,]  24.1233       69  65.3482     0.2041
[13,]  28.9422       28  29.6727     0.0943
[14,] 120.5008       25  16.7148     4.1068
> round(test.out,4)
     df Test Stat X2(.05,df) P(>TS)
[1,] 11   12.2664    19.6751  0.344
>
```

∇

4.4 Estimation and Testing for a Single Variance

When the data are normal (and independent), then a multiple of the sample variance follows a Chi-square distribution with $n-1$ degrees of freedom. That is, we have the following results.

$$Y_1, ..., Y_n \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \Rightarrow \quad P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

This leads to a rule for a $(1-\alpha)100\%$ Confidence Interval for σ^2 (and thus for σ), as well as a test of whether σ^2 is equal to some null value σ^2_0. Consider a Confidence Interval, then a test, where S^2 is a random variable (sample variance), and s^2 is a particular value from an observed sample.

$$1-\alpha = P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right)$$

$$\Rightarrow \quad (1-\alpha)100\% \text{ Confidence Interval for } \sigma^2\text{: } \left[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right]$$

$$\text{2-Tailed test: } H_0 : \sigma^2 = \sigma_0^2 \qquad H_A : \sigma^2 \neq \sigma_0^2 \qquad TS : X^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$$

$$RR : \left\{X^2_{obs} \leq \chi^2_{1-\alpha/2,n-1}\right\} \cup \left\{X^2_{obs} \geq \chi^2_{\alpha/2,n-1}\right\} \quad P = 2\min\left[P\left(\chi^2_{n-1} \leq X^2_{obs}\right), P\left(\chi^2_{n-1} \geq X^2_{obs}\right)\right]$$

$$\text{Upper-Tail test: } H_0 : \sigma^2 \leq \sigma_0^2 \qquad H_A : \sigma^2 > \sigma_0^2 \qquad TS : X^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$$

$$RR : X^2_{obs} \geq \chi^2_{\alpha,n-1} \qquad P = P\left(\chi^2_{n-1} \geq X^2_{obs}\right)$$

$$\text{Lower-Tail test: } H_0 : \sigma^2 \geq \sigma_0^2 \qquad H_A : \sigma^2 < \sigma_0^2 \qquad TS : X^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$$

$$RR : X^2_{obs} \leq \chi^2_{1-\alpha,n-1} \qquad P = P\left(\chi^2_{n-1} \leq X^2_{obs}\right)$$

Clearly, as in the case of a single mean, most practical situations will involve estimation rather than testing unless there is some focal null value σ_0^2 of interest. A plot of the Chi-Square distribution with 15 degrees of freedom along with $\chi^2_{.975,15} = 6.262$ and $\chi^2_{.025,15} = 27.488$ is given in Figure **??**.

Example 4.11: NBA Home Spread Differential 2006/07-2018/19

We now consider the Home Spread Differential for the population of $N = 15749$ NBA regular season games during the 2006/07-2018/19 seasons. Positive values imply the home team covered the spread, negative values imply the away team covered, and 0 implies a push. The population mean and variance are -0.1701 and $11.8362^2 = 140.096$, respectively. A histogram of the Home Spread Differentials is given in Figure **??** with the normal distribution superimposed.

Then 10000 random samples of size $n = 16$ are obtained and s^2 is obtained for each sample. The 95% Confidence Interval for σ^2 is computed for each sample. A histogram of the quantity $(n-1)s^2/\sigma^2$ is given in Figure **??** along with the density for the Chi-square distribution with 15 degrees of freedom. There are fewer values under the peak and more in the tails than we would expect if Home Spread Differentials were exactly normally distributed, however the fit still seems quite good. The first sample yielded a sample variance of $s^2 = 96.0573$. This leads to the following Confidence Interval computed below, and it does include the population variance $\sigma^2 = 140.105$.

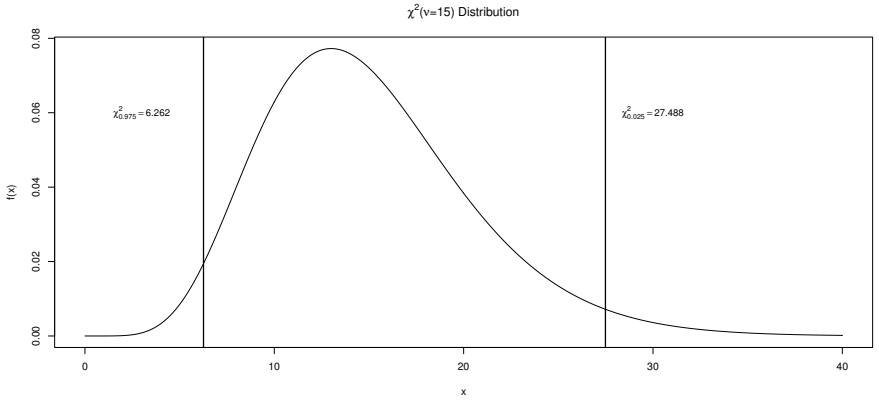


Figure 4.6: Chi-Square Distribution with 15 degrees of freedom

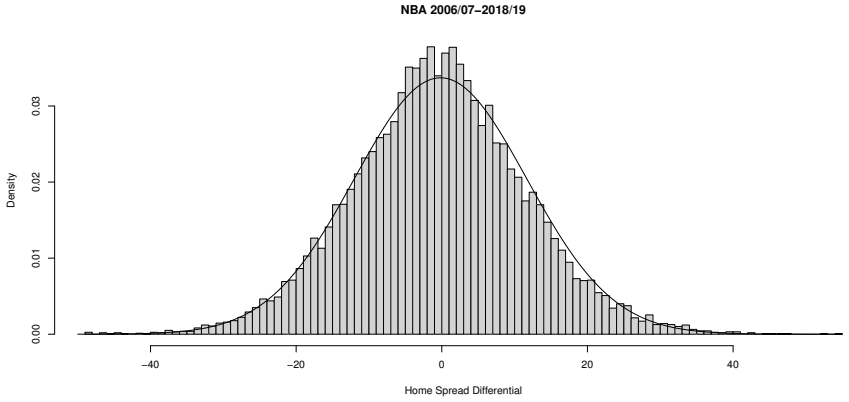


Figure 4.7: Histogram of NBA Home Spread Differentials - 2006/07-2018/19 Regular Seasons

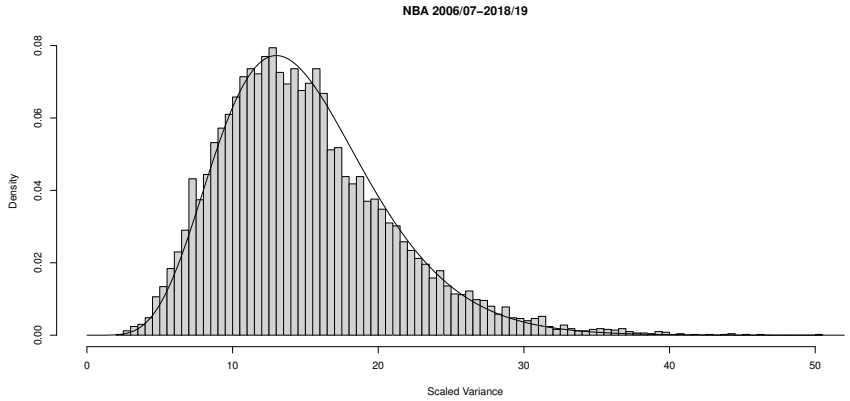


Figure 4.8: Histogram of Scaled Variance and Chi-Square(15) Density - NBA Home Spread Differential 2006/07-2018/19

$$n = 16 \quad s^2 = 96.0573 \quad \chi^2_{.975,15} = 6.262 \quad \chi^2_{.025,15} = 27.488 \quad \left[\frac{15(96.0573)}{27.488}, \frac{15(96.0573)}{6.262}\right] \equiv [52.418, 230.096]$$

The coverage rate of the 10000 Confidence Intervals is 93.16%, not so far from the nominal 95%. Of the 6.84% intervals that did not include σ^2, 3.54% fell entirely above 140.105 and 3.30% fell entirely below 140.105.

R Commands and Output

```
## Commands

nba.data <- read.csv("http://www.stat.ufl.edu/~winner/data/nba20062018.csv")
attach(nba.data); names(nba.data)

nba.home <- nba.data[home == 1,]
detach(nba.data); attach(nba.home)

## Create Home Spread Differential and summary stats (spread is coded as negative means favored)
hmSprdDiff <- teamPts-oppPts+spread
(N.HSD <- length(hmSprdDiff))
(mu.HSD <- mean(hmSprdDiff))
### Population SD (Uses N as denominator, not N-1)
(sigma.HSD <- sd(hmSprdDiff)*sqrt((N.HSD-1)/N.HSD))
summary(hmSprdDiff)

## Histogram of HSD
hsd.seq <- seq(-50,55,length=10000)

win.graph(height=5.5, width=7.0)
hist(hmSprdDiff, breaks=100, freq=FALSE, main="NBA 2006/07-2018/19",
   xlab="Home Spread Differential", ylab="Density")
lines(hsd.seq,dnorm(hsd.seq,mu.HSD,sigma.HSD))


### Begin taking 10000 samples of size n=16
```

```
num.samp <- 10000   # Number of samples
n.samp <- 16        # Number of games per sample
var.samp <- rep(0, num.samp)  # Holder for sample variances
set.seed(12345)

for (i1 in 1:num.samp) {
  samp.games <- sample(N.HSD, n.samp, replace=FALSE)
  var.samp[i1] <- var(hmSprdDiff[samp.games])
}

var.samp[1]
scale.var.samp <- (n.samp-1)*var.samp/(sigma.HSD^2)
range(scale.var.samp)

# Histogram and Chi-square Density
scv.seq <- seq(0,52,length=10000)
win.graph(height=5.5, width=7.0)
hist(scale.var.samp, breaks=100, freq=FALSE, xlim=c(0,52),
   main="NBA 2006/07-2018/19",
   xlab="Scaled Variance", ylab="Density")
lines(scv.seq,dchisq(scv.seq,n.samp-1))

# Coverage Rate
X2.LO <- qchisq(.025,n.samp-1)
X2.HI <- qchisq(.975,n.samp-1)
mean((n.samp-1)*var.samp/X2.HI <= sigma.HSD^2 &
     (n.samp-1)*var.samp/X2.LO >= sigma.HSD^2)

mean((n.samp-1)*var.samp/X2.HI >= sigma.HSD^2)
mean((n.samp-1)*var.samp/X2.LO <= sigma.HSD^2)


## Output

> (N.HSD <- length(hmSprdDiff))
[1] 15749
> (mu.HSD <- mean(hmSprdDiff))
[1] -0.170106
> ### Population SD (Uses N as denominator, not N-1)
> (sigma.HSD <- sd(hmSprdDiff)*sqrt((N.HSD-1)/N.HSD))
[1] 11.83618
> summary(hmSprdDiff)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-48.5000  -8.0000   0.0000  -0.1701   7.5000  55.0000

> var.samp[1]
[1] 96.05729
> mean((n.samp-1)*var.samp/X2.HI <= sigma.HSD^2 &
+      (n.samp-1)*var.samp/X2.LO >= sigma.HSD^2)
[1] 0.9316
>
> mean((n.samp-1)*var.samp/X2.HI >= sigma.HSD^2)
[1] 0.0354
> mean((n.samp-1)*var.samp/X2.LO <= sigma.HSD^2)
[1] 0.033
```

4.5 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 is unknown (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 in terms of its cumulative distribution function (CDF)) to approximate the sampling distribution of the estimator of interest. The classic reference is Efron and Tibshirani (1993) [**?**], and for an introduction to Mathematical Statistics based on resampling methods, see Chihara and Hesterberg (2011) [**?**].

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, the size of the original sample. 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.

4.5.1 Bootstrap Inferences Concerning the Population Mean and Standard Deviation

When trying to estimate a population mean (particularly with nonnormal data with a small sample size), a bootstrap confidence 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. Chihara 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 around the sample mean, unless the sample data are highly symmetric. We can follow the same approach regarding the standard deviation.

Example 4.12: NFL Betting Home Underdogs Strategy

Suppose we are interested in the returns of the following betting strategy of betting on NFL home underdogs and (and pick 'ems). We define the following strategy based on the "11-10" betting rule, with formulas given below. We are treating these weekly returns as a sample of a conceptual population of weekly games.

- Each week, we will place 1100 units worth of bets on the strategy.
- We will identify all games where the home team is not favored, with k_i being the number of such games for week i.
- We will use weights that have higher weights for teams "catching the most points," where S_{ij} is the number of points received in game j of week i.
- We will observe the returns as a percentage of our bet for that week, where R_i is the return for week i.

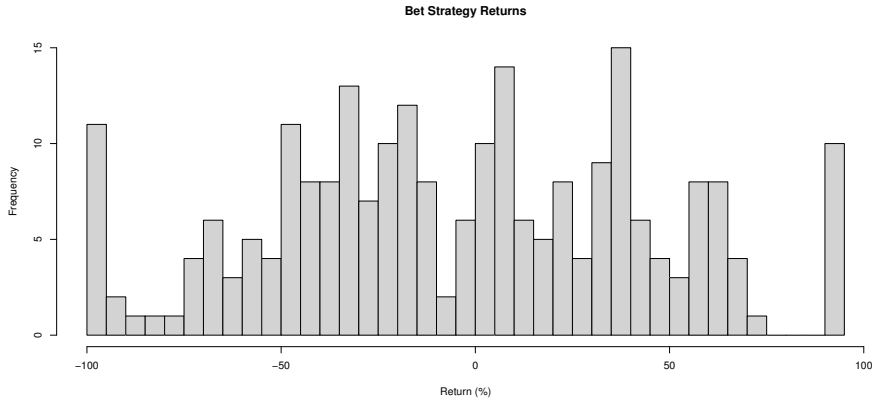


Figure 4.9: Histogram of Weekly Returns for Home Underdog Betting Strategy - NFL 2006-2019 regular seasons

$$\text{Betting weight on game } j \text{ on week } i\text{: } w_{ij} = \frac{S_{ij}+1}{\sum_{j=1}^{k_i}(S_{ij}+1)} \qquad \sum_{j=1}^{k_i} w_{ij} = 1$$

$$\text{Bet on game } j \text{ on week } i\text{: } B_{ij} = 1100w_{ij}$$

$$\text{Return on game } j \text{ on week } i\text{: If Win: } R_{ij} = 2100w_{ij} \qquad \text{If Lose: } R_{ij} = 0 \qquad \text{If Push: } R_{ij} = 1100w_{ij}$$

$$\text{Return percentage for week } i\text{: } R_i = 100\left(\frac{\sum_{j=1}^{k_i} R_{ij} - 1100}{1100}\right)$$

A histogram of the returns for $n = 238$ weeks for the 2006-2019 regular seasons is given Figure **??**. The returns are somewhat uniformly distributed between -100% and 90% with mean of -4.197% and standard deviation of 47.926%. The number of games per week meeting the strategy's requirement ranged from 1 (2 weeks) to 11 (2 weeks). Thus, the variance and standard deviation of the individual measurements should not be constant.

Suppose we wish to estimate the population mean of the returns of this strategy. While the "sample" is large ($n = 238$) which implies sampling distribution for the sample mean should be approximately normal. However, the sampling distribution of the scaled variance may not be approximately distributed as Chi-square.

As a reference point, we first compute "normal" based confidence interval for μ and σ using the traditional methods. Note that $t_{.025,237} = 1.970$, $\chi^2_{.975,237} = 196.253$, and $\chi^2_{.025,237} = 281.533$.

$$95\%\text{ CI for } \mu\text{: } -4.197 \pm 1.970\sqrt{\frac{47.926^2}{238}} \quad \equiv \quad -4.197 \pm 1.970(3.107) \quad \equiv \quad -4.197 \pm 6.120 \quad \equiv \quad (-10.317, 1.923)$$

$$(n-1)s^2 = (238-1)47.926^2 = 544365.65 \quad \Rightarrow \quad 95\%\text{ CI for } \sigma^2\text{: } \left(\frac{544365.65}{281.533}, \frac{544365.65}{196.253}\right) \quad \equiv \quad (1933.577, 2773.795)$$

$$\Rightarrow \quad 95\% \text{ CI for } \sigma\text{: } \left(\sqrt{1933.577}, \sqrt{2773.795}\right) \quad \equiv \quad (43.972, 52.667)$$

We then draw $B = 10000$ random resamples with replacement from the 238 returns, and compute the sample mean and standard deviation for each resample, labeled $\overline{y}_i^*$ and s_i^*, respectively for the i^{th} resample. Finally we obtain the 2.5%-ile and 97.5%-ile from the resample means, for intervals that we can be approximately 95% confident will contain μ and σ.

R Commands and Output

```
## Commands
#############################################################
### Bootstrap sample mean and standard deviation

# Set the seed, n, number of bootstrap samples, and holders for mean, sd
set.seed(123)
B <- 10000
n.wk <- length(R.week)
boot.mn <- rep(0,B)
boot.sd <- rep(0,B)
data.mn <- mean(R.week)     # Original sample mean
data.sd <- sd(R.week)       # Original sample sd

# Normal based Confidence Intervals
norm.mn.ci <- data.mn + qt(c(.025,.975),n.wk-1) * data.sd/sqrt(n.wk)
norm.sd.ci <- c(sqrt((n.wk-1)*data.sd^2/qchisq(.975,n.wk-1)),
                sqrt((n.wk-1)*data.sd^2/qchisq(.025,n.wk-1)))

# Take samples and obtain means and SDs
for (i1 in 1:B) {
  sample1 <- sample(1:n.wk,n.wk,replace=TRUE)
  boot.mn[i1] <- mean(R.week[sample1])
  boot.sd[i1] <- sd(R.week[sample1])
}

# Bootstrap Quantile Confidence Intervals
boot.mn.ci.q <- quantile(boot.mn, c(.025,.975))
boot.sd.ci.q <- quantile(boot.sd, c(.025,.975))

# Set-up and print output
mn.out <- cbind(data.mn,norm.mn.ci[1],norm.mn.ci[2],
          mean(boot.mn),boot.mn.ci.q[1],boot.mn.ci.q[2], sd(boot.mn))
sd.out <- cbind(data.sd,norm.sd.ci[1],norm.sd.ci[2],
          mean(boot.sd),boot.sd.ci.q[1],boot.sd.ci.q[2], sd(boot.sd))
mn.sd.out <- rbind(mn.out, sd.out)
colnames(mn.sd.out) <- c("Data", "Norm LB", "Norm UB",
                "Boot Mean", "Boot LB", "Boot UB", "Boot SD")
rownames(mn.sd.out) <- c("Mean", "Std Dev")
round(mn.sd.out, 3)

# Histograms of Bootstrap Samples
win.graph(height=5.5, width=7.0)
hist(boot.mn, breaks=50, xlab="Mean Return (%)",
  main="Bootstrap Means")
abline(v=boot.mn.ci.q[1], lty=2)
abline(v=boot.mn.ci.q[2], lty=3)
abline(v=mean(boot.mn), lty=4)
abline(v=data.mn, lty=5)
legend("topleft", c("Boot LB", "Boot UB", "Boot Mean", "Data Mean"),
   lty=2:5)
```

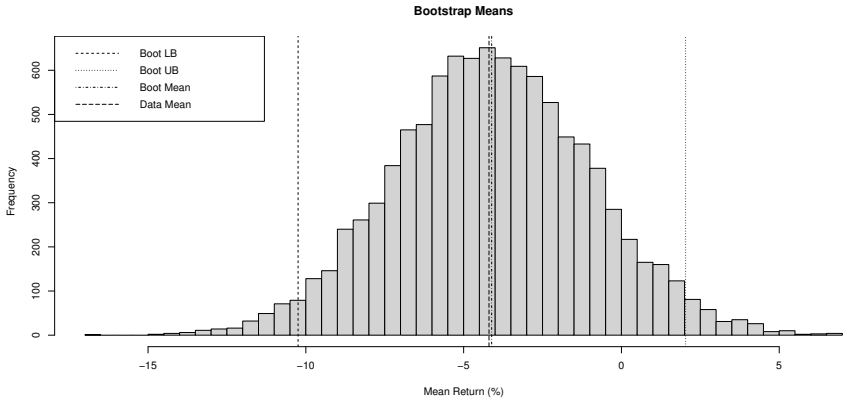


Figure 4.10: NFL home underdog betting strategy returns - Bootstrap means

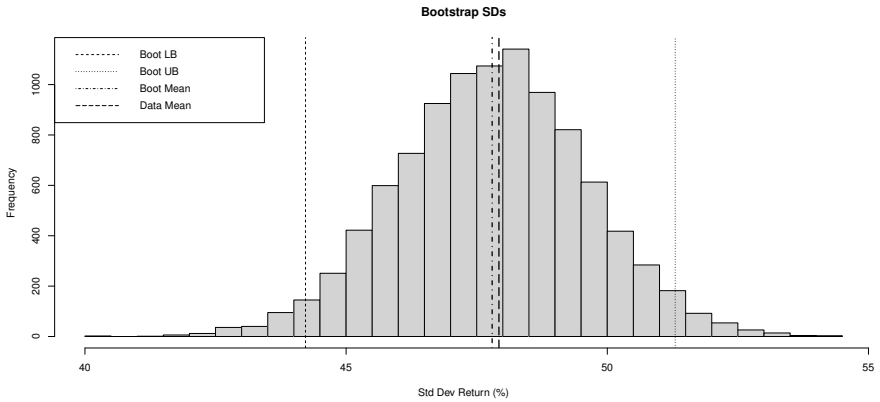


Figure 4.11: NFL home underdog betting strategy returns - Bootstrap standard deviations

```
win.graph(height=5.5, width=7.0)
hist(boot.sd, breaks=50, xlab="Std Dev Return (%)",
  main="Bootstrap SDs")
abline(v=boot.sd.ci.q[1], lty=2)
abline(v=boot.sd.ci.q[2], lty=3)
abline(v=mean(boot.sd), lty=4, lwd=2)
abline(v=data.sd, lty=5, lwd=2)
legend("topleft", c("Boot LB", "Boot UB", "Boot Mean", "Data Mean"),
   lty=2:5)


## Output
> round(mn.sd.out, 3)
          Data Norm LB Norm UB Boot Mean Boot LB Boot UB Boot SD
Mean    -4.197 -10.317   1.923    -4.112 -10.246   2.035   3.113
Std Dev 47.926  43.972  52.666    47.796  44.222  51.299   1.803
```

A histogram of the bootstrap means is given in Figure **??** and a histogram of the bootstrap standard

deviations is given in Figure **??**. The sample mean for the original sample is $\overline{y} = -4.197$. The mean of the $B = 10000$ resample means is $\overline{\overline{y}^*} = -4.112$, which is very close to $\overline{y}$. The sample standard deviation for the original sample is $s = 47.926$, while the mean of the resample standard deviations is $\overline{s^*} = 47.796$, again very close to the observed value. The approximate 95% confidence interval for μ is $(-10.246, 2.035)$. The approximate 95% confidence interval for the standard deviation is $(44.222, 51.299)$. Both of the bootstrap percentile confidence intervals are very similar to their normal based counterparts.

The standard deviation of the resample means and standard deviations are referred to as the bootstrap standard errors. Note that the confidence interval for the mean is not of the form $\overline{\overline{y}^*} \pm t_{.025,238-1} s_{\overline{y}*}$, which is of the following form, where $t_{.025,237} = 1.970$.

$$-4.112 \pm 1.970(3.113) \quad \equiv \quad -4.112 \pm 6.133 \quad \equiv \quad (-10.245, 2.021)$$

Of the 10000 sample means, 2.51% of the sample means fall below the lower bound -10.245, and 2.53% fall above the upper bound 2.021. The "t-type" interval goes inside both of the lower and upper bounds of the bootstrap interval. The lower bound of the bootstrap percentile interval is 1.970 bootstrap standard errors below the mean of the resample means, and the upper bound is 1.975 standard errors above it. In some cases the asymmetry and discrepancy 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 observations, the median is the "middle" of the n observed values. When we take bootstrap samples, the median will always be one of the n observations in the original sample when n is odd. Thus, there are only n possible values the sample median for each resample can take on in the case of odd n. For even n, where the sample median is the average of the middle two observations will also have relatively few possible outcomes.

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), [**?**] 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.

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

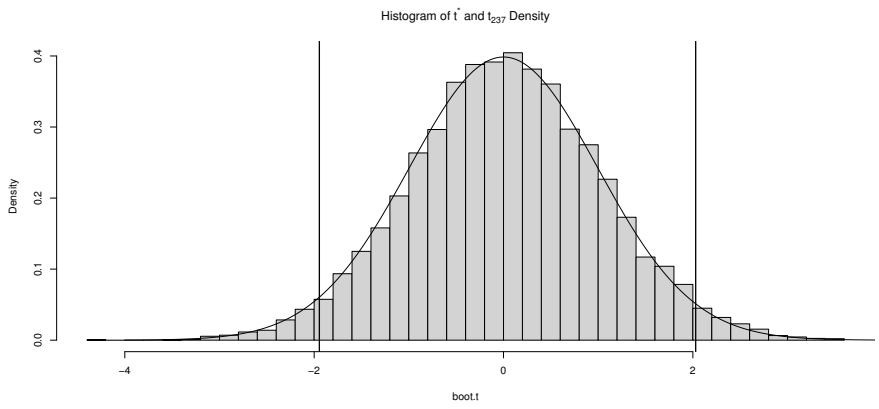


Figure 4.12: Histogram of t^* values and t_{24} Density - NFL home underdog betting strategy Data

Example 4.12: NFL Betting Home Underdogs Strategy

We apply this method to the same sample and resamples used previously for the NFL home underdog betting strategy returns.

R Commands and Output

```
## Commands

boot.t <- (boot.mn - data.mn)/(boot.sd / sqrt(n.wk))

Q_L <- quantile(boot.t,0.025)
Q_U <- quantile(boot.t,0.975)
mu_L <- data.mn - Q_U*data.sd/sqrt(n.wk)
mu_U <- data.mn - Q_L*data.sd/sqrt(n.wk)

q.out <- cbind(Q_L, Q_U, mu_L, mu_U)
colnames(q.out) <- c("Q_L", "Q_U", "mu_L", "mu_U")
round(q.out, 3)

win.graph(height=5.5, width=7.0)
hist(boot.t,breaks=40,freq=F,
   main=expression(paste("Histogram of ",t^"*"," and ",t[237]," Density")))
t.seq <- seq(-4,4,.01)
lines(t.seq,dt(t.seq,n.wk-1))
abline(v=c(Q_L,Q_U),lwd=2)

## Output

> round(q.out, 3)
        Q_L  Q_U    mu_L  mu_U
2.5% -1.945 2.03 -10.503 1.846
```

A histogram of the t^* values and the t_{237} density are given in Figure **??**. The distribution of the t^* values is fairly symmetric, with Q_L^* and Q_U^* being $-1,945$ and 2.030, respectively for $\alpha = 0.05$. The original sample mean and standard deviation are -4.197 and 47.926 respectively, leading to the following 95% Confidence Interval for μ.

$$\left(-4.197 - 2.030\frac{47.926}{\sqrt{238}}, 8.188 - (-1.945\frac{47.926}{\sqrt{238}}\right) \quad \equiv \quad (-4.197 - 6.306, -4.197 + 6.042) \quad \equiv \quad (-10.476, 1.845)$$

Note that the interval is not symmetric about $\overline{y}$, it adds a smaller term to the upper end than the term it subtracts from the lower end. This reflects the fact that the data are slightly left-skewed. The bootstrap estimate of the **bias** is the difference from the average of the resample means and the overall sample mean: $\overline{\overline{y}^*} - \overline{y} = 0.085$. This bias is small relative to the standard error of the bootstrap estimator: 0.085/3.113=0.0273.

∇

Chapter 5

Comparing Two Populations: Means, Medians, Proportions and Variances

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.

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

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

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 of samples are needed depends on the shape of the underlying distributions.

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

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

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

Example 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 **??**. 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 **??**. 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).

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

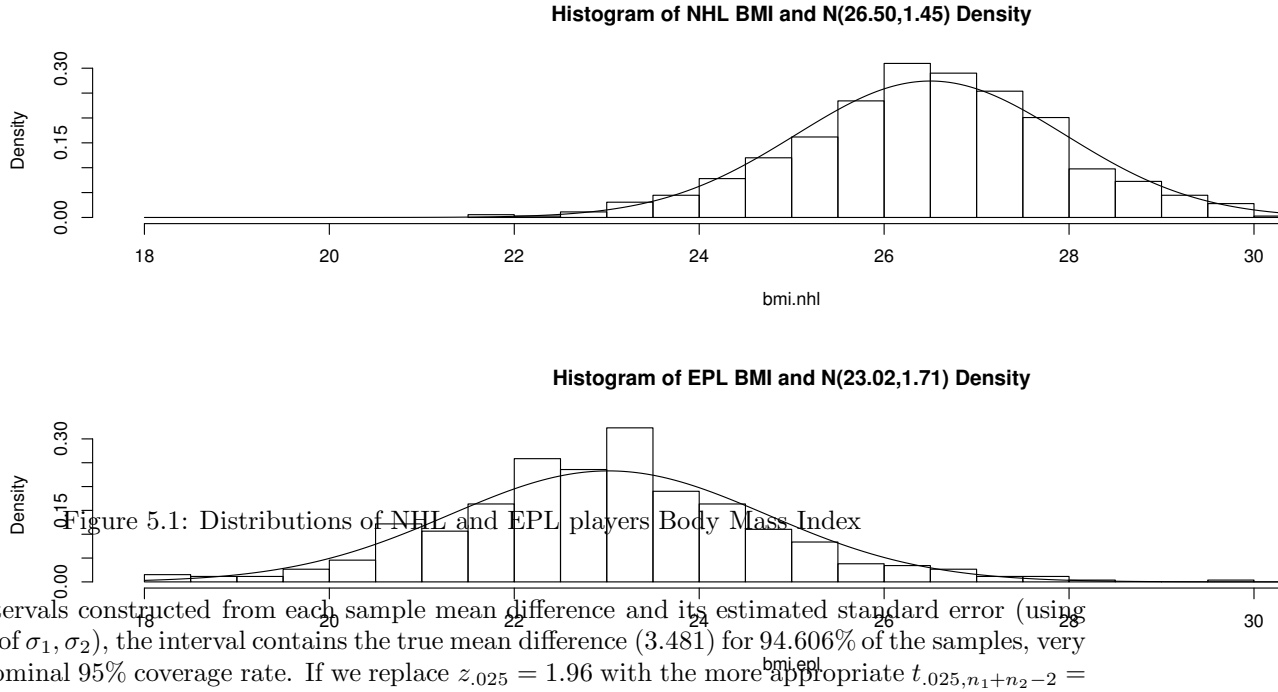


Figure 5.1: Distributions of NHL and EPL players Body Mass Index

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 $t_{.025,n_1+n_2-2} = t_{.025,38} = 2.0244$, 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))
```

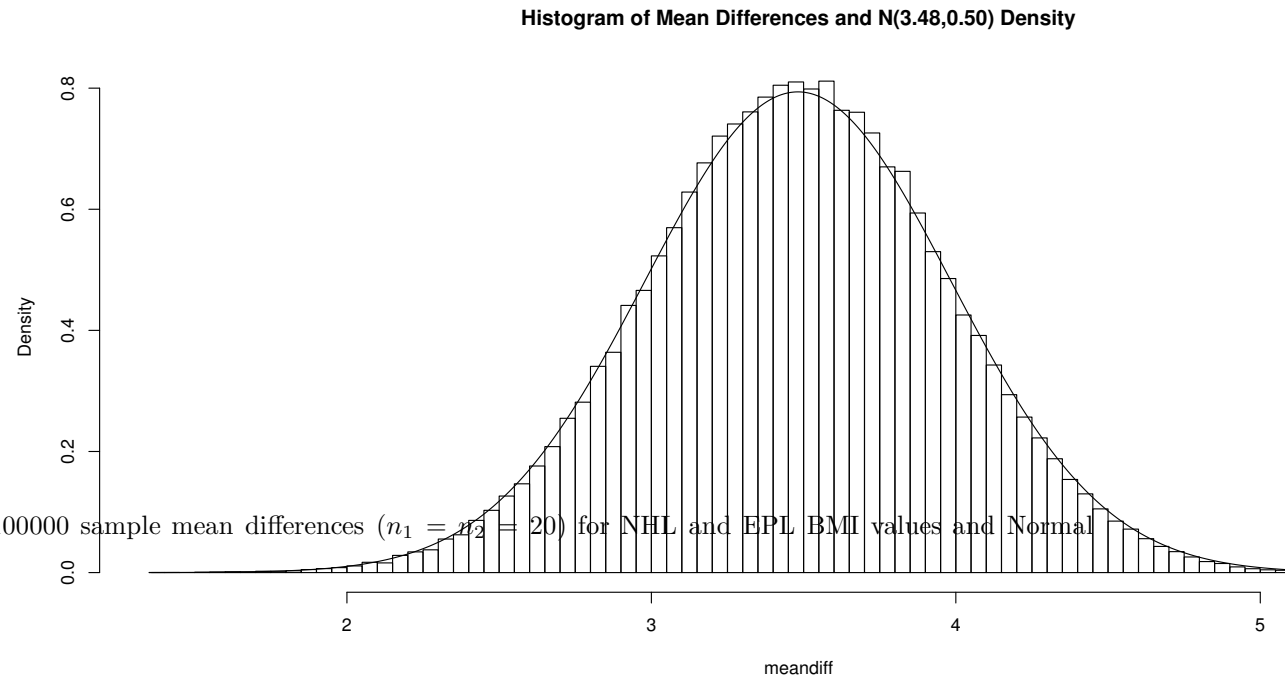


Figure 5.2: 100000 sample mean differences ($n_1 = n_2 = 20$) for NHL and EPL BMI values and Normal Density

```
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")
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), [**?**]). The measurements were length and breadth of glenoid cavity (LGC and BGC, in mm, respectively). Summary data for the two samples for BGC are given below.

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

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

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

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

∇

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 an independent samples (parallel groups) or paired samples (crossover) design? The appropriate test for each situation is given in Table **??**. 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) [**?**]) to compare two instructional methods: multimedia (treatment 1) and traditional (treatment 2) for teaching physics to undergraduate students in Rwanda. Subjects were assigned at random to the two treatments. Each subject received only one of the two methods. The numbers of subjects who completed the courses and took two exams were $n_1 = 13$ for the multimedia course and $n_2 = 19$ for the traditional course. The primary response was the post-course score on an

examination. We will conduct the test $H_0 : \mu_1 - \mu_2 = 0$ vs $H_A : \mu_1 - \mu_2 \neq 0$, where the null hypothesis is no difference in the effects of the two methods. The summary statistics are given below.

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

First, compute s_p^2, the pooled variance:

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

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

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_A : \mu_1 - \mu_2 \neq 0$
- T.S.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(11.10-8.35)}{\sqrt{8.42\left(\frac{1}{13}+\frac{1}{19}\right)}} = \frac{2.75}{1.04} = 2.633$
- R.R.: $|t_{obs}| \geq t_{\alpha/2,n_1+n_2-2} = t_{.05/2,13+19-2} = t_{.025,30} = 2.042$
- P-value: $2P(T \geq |t_{obs}|) = 2P(t_{30} \geq 2.633) = 0.0132$

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

Below we 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
rp <- read.csv("http://www.stat.ufl.edu/~winner/data/rwanda_physics.csv")
attach(rp); names(rp)

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

## Output

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

        Two Sample t-test

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

∇

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

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

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 Drainage in Breast Reconstruction Surgery

A study considered the effect of abdominal suture quilting on abdominal drainage during breast reconstruction surgery (Liang, et al, (2016), [**?**]). 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 abdominal drainage during the surgery (in ml). The summary data are given below, note that the sample standard deviations are substantially different, and these are relatively large sample sizes. Side-by-side box plots are given in Figure **??**.

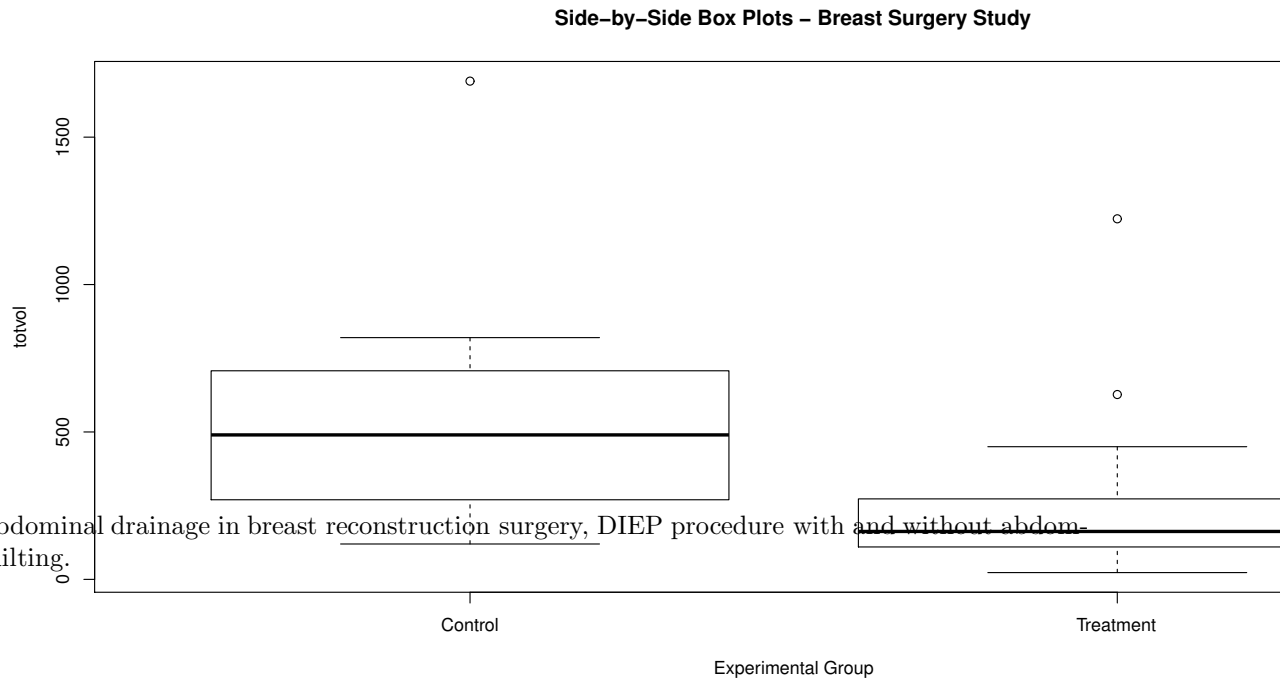


Figure 5.3: Abdominal drainage in breast reconstruction surgery, DIEP procedure with and without abdominal suture quilting.

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

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

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

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

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

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

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

R Commands and Output

```
## Commands

quilt <- read.csv("http://www.stat.ufl.edu/~winner/data/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. 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, where to be able to use the table on the webpage, $n_1 \geq n_2$.

1. H_0 : The two population medians are equal ($M_1 = M_2$)
2. H_A : The medians are not equal ($M_1 \neq M_2$)
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), [**?**]). Based on a small trial, with $n_1 = 19$ treatment subjects and $n_2 = 10$ control subjects, Table **??** 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 **??**, there is evidence of outlying cases in each group.

For a 2-tailed test, based on sample sizes of $n_1 = 19$ 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 can 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

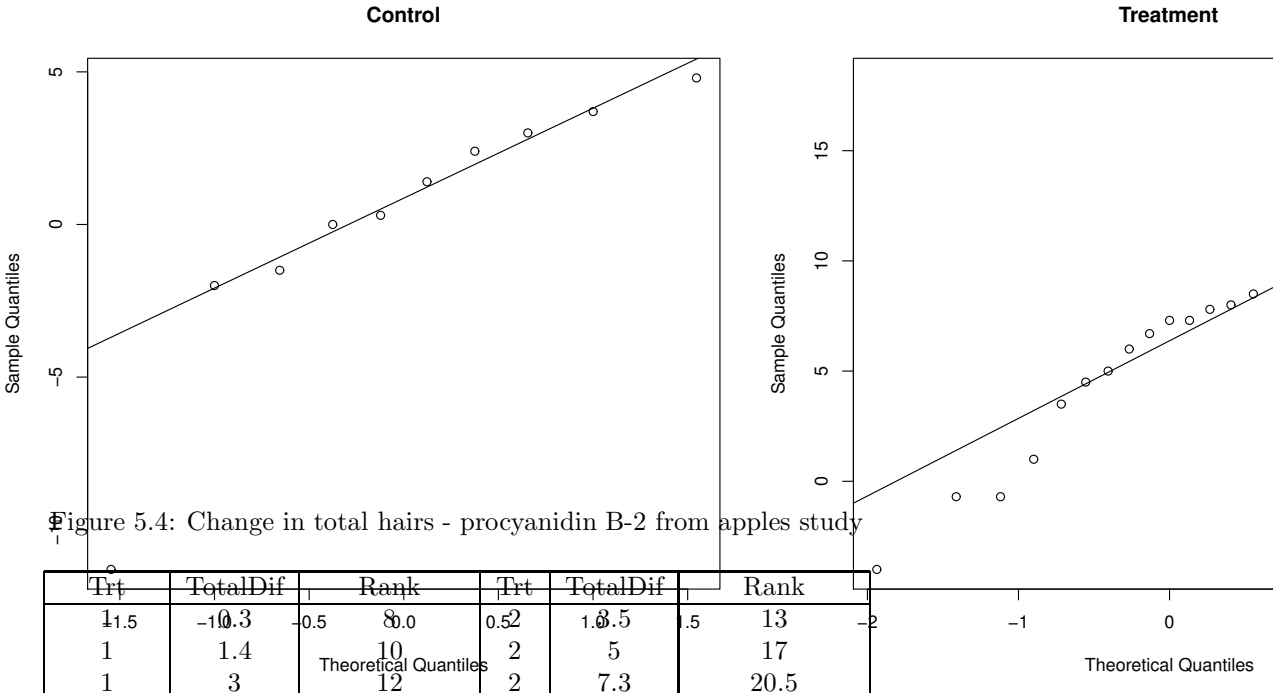


Figure 5.4: Change in total hairs - procyanidin B-2 from apples study

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

```
## Commands

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.

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

∇

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.

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

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.

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

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).

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

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

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.

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

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

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

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), [**?**]). 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 **??**.

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 of 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 (pair). 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 step 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) [**?**]). 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 **??**, along with ranks. We will test whether there is evidence that 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 **??**, 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), [**?**]). 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 (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.

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

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

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. The 95% Confidence Interval for μ_D is (.0287,.0671), an advantage in aggregate return of about 2.9% to 6.7%.

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

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

∇

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 $\overline{Y}_n - \overline{Y}_e$. 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 **??** for a histogram of 100000 random samples' mean differences of $n_1 = n_2 = 20$.

$$\text{BMI: } \mu_n = 26.50 \quad \sigma_n = 1.45 \quad \mu_e = 23.02 \quad \sigma_e = 1.71 \quad E\{\overline{Y}_n - \overline{Y}_e\} = 26.50 - 23.02 = 3.48$$

$$V\{\overline{Y}_n - \overline{Y}_e\} = \frac{1.45^2}{n_n} + \frac{1.71^2}{n_e}$$

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 $n_n = n_e = 10$. 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 $H_0 : \mu_n - \mu_e = 0$, 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 $\overline{y}_n$, s_n, $\overline{y}_e$, s_e
3. Compute the sample mean difference $\overline{y}_n - \overline{y}_e$ and its estimated standard error $\sqrt{\frac{s_n^2}{10} + \frac{s_e^2}{10}}$
4. Compute the approximate degrees of freedom for the unequal variance case (Satterthwaite's approximation)
5. Obtain the 95% Confidence Intervals for $\mu_n - \mu_e$
6. Determine whether the Confidence Intervals contain 3.48 (true value) and whether they contain 0 (Testing $\mu_n - \mu_e = 0$)
7. Obtain the width of the intervals

The equal variance Confidence Intervals contained $\mu_n - \mu_e = 3.48$ 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 $H_0 : \mu_n - \mu_e = 0$ in favor of $H_A : \mu_n - \mu_e \neq 0$ 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 **??**. The population means and standard deviations are given below, along with the mean and standard error of the sampling distribution of the sample mean $\overline{Y}_f - \overline{Y}_m$.

$$\mu_f = 5.840 \quad \sigma_f = 0.831 \qquad \mu_m = 6.337 \quad \sigma_f = 1.058$$

$$E\{\overline{Y}_f - \overline{Y}_m\} = -0.497 \qquad SE\{\overline{Y}_f - \overline{Y}_m\} = \sqrt{\frac{0.831^2}{n_f} + \frac{1.058^2}{n_m}}$$

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 in 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.

In the case of 1-tailed tests, the Rejection Region is in only 1-tail, with area α and only one of the tail area probabilities is computed.

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 $n_f = n_m = n = 6$, 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 6+6-2=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 are the R Commands and Output 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 The Bootstrap

The bootstrap method is the same principle as in the one-sample case. In terms of independent samples, take resamples within each group with replacement, then take the difference between the two group means 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.

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

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) [**?**]). The author reported on 7 characteristics among each subject. Consider the variable cubit (cm), given in Table **??**. 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 **??**.

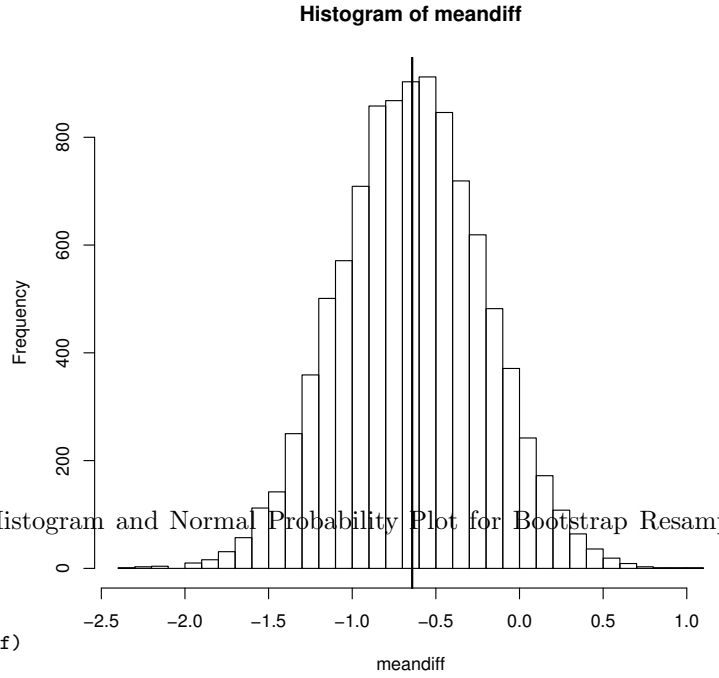
R Commands and Output

```
## 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
```



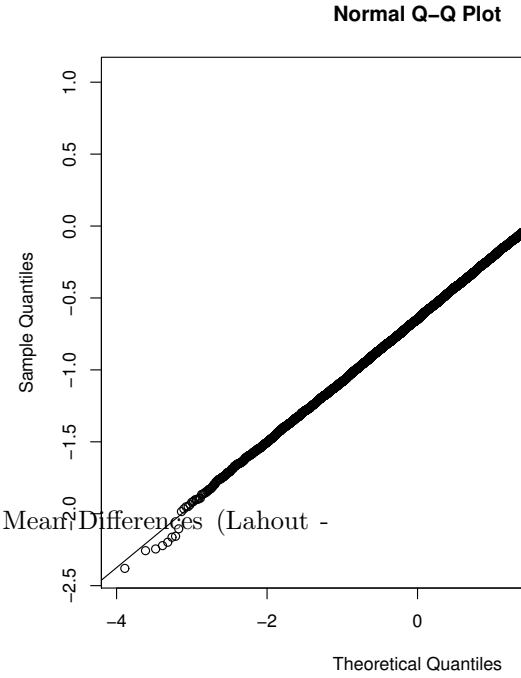


Figure 5.5: Histogram and Normal Probability Plot for Bootstrap Resample Mean Differences (Lahout - Kulu)

```
> 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 $\overline{y}_L - \overline{y}_K = -0.641$. 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 $\mu_L = \mu_K$. A histogram of the permutation mean differences with a vertical line at the observed mean difference is given in Figure **??**.

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
```

∇

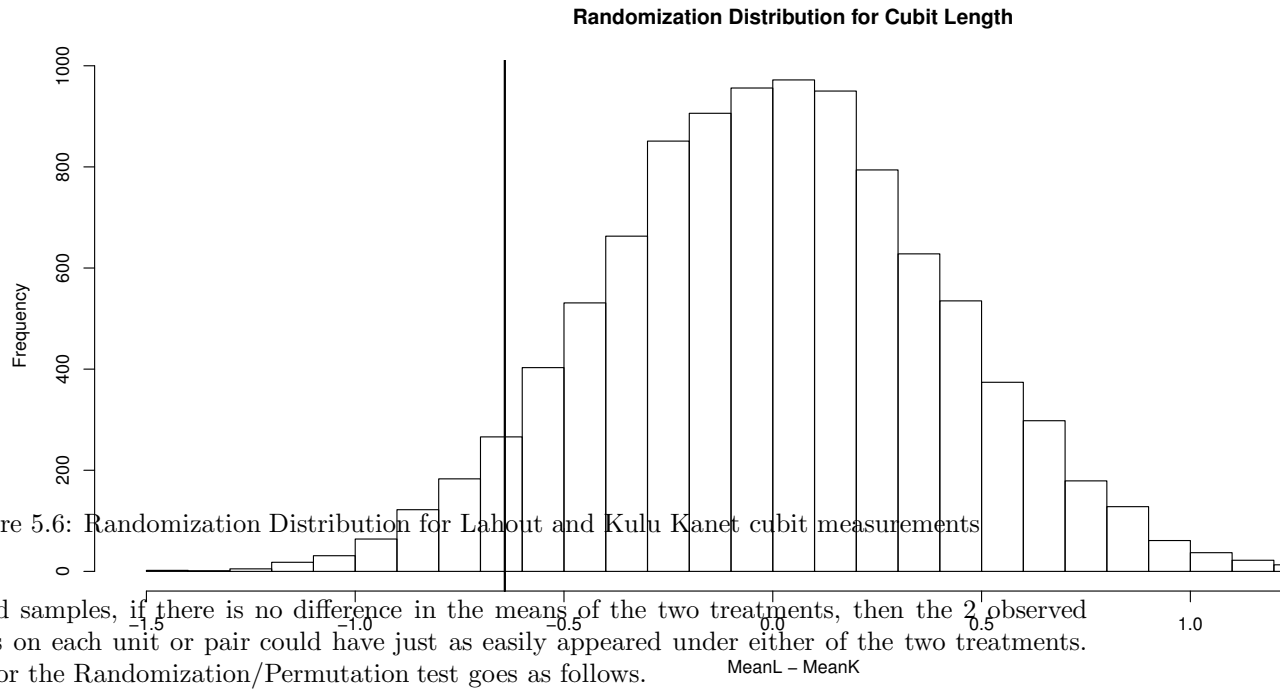


Figure 5.6: Randomization Distribution for Lahout and Kulu Kanet cubit measurements

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 **??**.

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
```

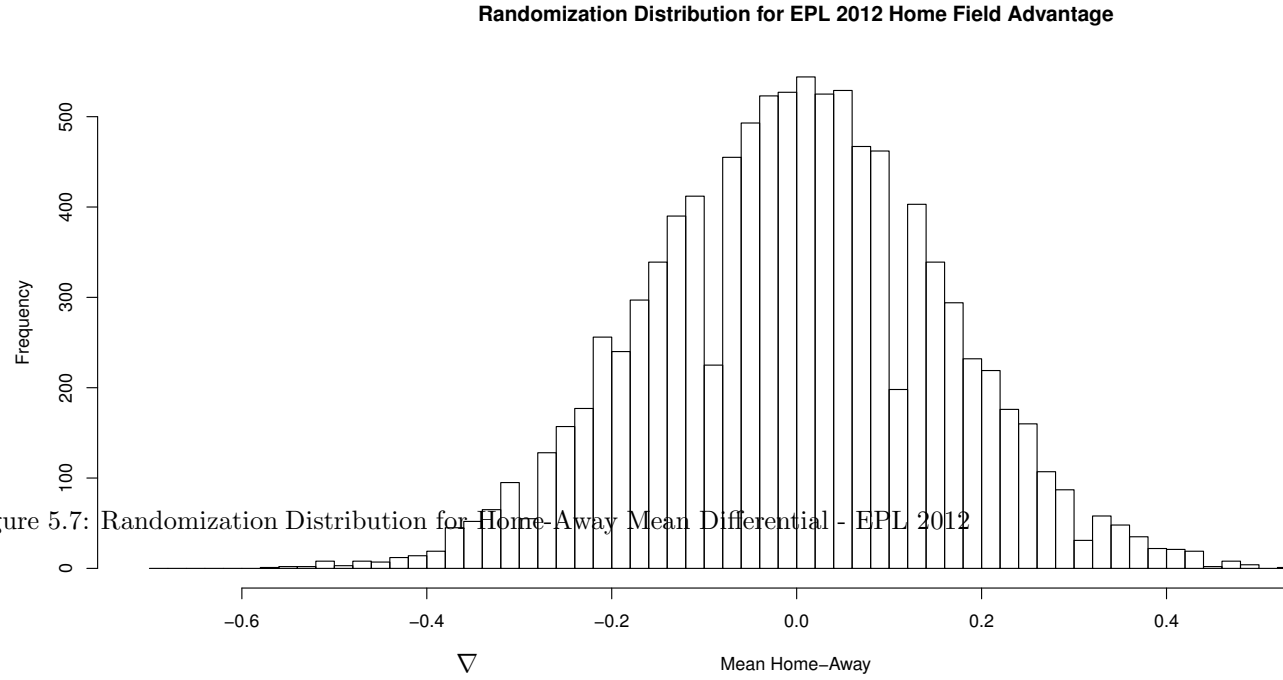


Figure 5.7: Randomization Distribution for Home-Away Mean Differential - EPL 2012