CHAPTER 4

Advanced Math and Statistics

4.1 INTRODUCTION

In this chapter we provide an overview of probability and statistics and their use in sports modeling applications. The chapter begins with an overview of the mathematics required for probability and statistics modeling and a review of essential probability distribution functions required for model construction and parameter estimation. The chapter concludes with an introduction to different sampling techniques that can be used to test the accuracy of sports prediction models and to correct for data limitation problems. These data limitation issues are often present in sports modeling problems due to limited data observations and/or not having games across all pairs of teams.

The sampling techniques include with and without replacement, Monte Carlo techniques, bootstrapping, and jackknife techniques. These techniques are useful for sports such as soccer, basketball, football, and hockey, as well as baseball when we are evaluating the best mix of players to use and best lineup based on the opposing team's starting pitcher.

Finally, these techniques serve as the building blocks for the advanced applications and sports models that are discussed in Chapter 12, Fantasy Sports Models and Chapter 13, Advanced Modeling Techniques. A summary of these important and essential techniques include:

- Probability and Statistics
- Probability Distribution Function (PDF) and Cumulative Distribution Function (CDF)
- Sampling Techniques
    - With Replacement
    - Without Replacement
    - Monte Carlo Distribution
    - Bootstrapping
    - Jackknife Sampling

Optimal Sports Math, Statistics, and Fantasy.
DOI: http://dx.doi.org/10.1016/B978-0-12-805163-4.00004-9

4.2 PROBABILITY AND STATISTICS

A *random variable* is defined as a variable that can take on different values. These values are determined from its underlying probability distribution, and the actual distribution is characterized by a mean and standard deviation term (such as a normal distribution) also a skewness and a kurtosis measure. The value of the random variable is also often subject to random variations due to noise or chance.

A random variable can represent many different items such as expected daily temperature at a location in the middle of July, the expected attendance at a sporting event, a sports team's strength rating, as well as the probability that a team will win a game or score a specified number of points.

A random variable can also be the parameter of a model used to predict the outcome of the sports game. The goal of the analyst in this case is to compute an accurate estimate of this random variable parameter.

Random variables can be either discrete or continuous values. A discrete random variable can take on only a specific finite value or a countable list of values. For example, a discrete random variable in sports is the number of points that a team scores or the number difference between the home team points scored and away team points scored. A continuous random variable can take on any numerical value in an interval (and theoretically, have an infinite number of decimal places). For example, a continuous random variable in sports could be the team's strength rating or a performance metric such as batting average (which can both have an infinite number of decimals).

Probability Distributions

Mathematicians utilize *probability distribution* functions in many different ways. For example, probability distribution functions can be used to "quantify" and "describe" random variables, they can be used to determine statistical significance of estimated parameter values, they can be used to predict the likelihood of a specified outcome, and also to calculate the likelihood that an outcome falls within a specified interval (i.e., confidence intervals). As mentioned, these probability distribution functions are described by their mean, variance, skewness, and kurtosis terms.

A *probability mass function* (pmf) is a function used to describe the probability associated with the discrete variable. A *cumulative mass function* (cmf) is a function used to determine the probability that the observation will be less than or equal to some specified value.

In general terms, if x is a discrete random variable and x^* is a specified value, then the pmf and cmf functions are defined as follows:

Probability Mass Function (pmf):

$$f(x) = Prob(x = x^*)$$

Cumulative Mass Function (cmf):

$$F(x) = Prob(x \leq x^*)$$

Probability distribution functions for continuous random variables are similar to those for discrete random variables with one exception. Since the continuous random variable can take on any value in an interval the probability that the random variable will be equal to a specified value is thus zero. Therefore, the probability distribution function (pdf) for a continuous random variable defines the probability that the variable will be within a specified interval (say between a and b) and the cumulative distribution function for a continuous random variable is the probability that the variable will be less than or equal to a specified value x^*.

A *probability distribution function* (pdf) is used to describe the probability that a continuous random variable and will fall within a specified range. In theory, the probability that a continuous value can be a specified value is zero because there are an infinite number of values for the continuous random value. The *cumulative distribution function* (cdf) is a function used to determine the probability that the random value will be less than or equal to some specified value. In general terms, these functions are:

Probability Distribution Function (pdf):

$$Prob(a \leq X \leq b) = \int_a^b f(x)dx$$

Cumulative Distribution Function (cdf):

$$F(x) = Prob(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

Going forward, we will use the terminology "pdf" to refer to probability distribution function and probability mass function, and we will use the terminology "cdf" to refer to cumulative distribution function and cumulative mass function.

Example 4.1 Discrete Probability Distribution Function

Consider a scenario where a person rolls two dice (die) and adds up the numbers rolled. Since the numbers on dice range from 1 to 6, the set of possible outcomes is from 2 to 12. A pdf can be used to show the probability of realizing any value from 2 to 12 and the cdf can be used to show the probability that the sum will be less than or equal to a specified value.

Table 4.1 shows the set of possible outcomes along with the number of ways of achieving the outcome value, the probability of achieving each outcome value (pdf), and the probability that the outcome value will be less than or equal to the outcome value (cdf). For example, there were 6 different ways to roll a 7 from two dice. These

Table 4.1 Discrete Random Variable: Rolling Die

Value	Count	Pdf	Cdf
2	1	3%	3%
3	2	6%	8%
4	3	8%	17%
5	4	11%	28%
6	5	14%	42%
7	6	17%	58%
8	5	14%	72%
9	4	11%	83%
10	3	8%	92%
11	2	6%	97%
12	1	3%	100%
Total	36	100%	

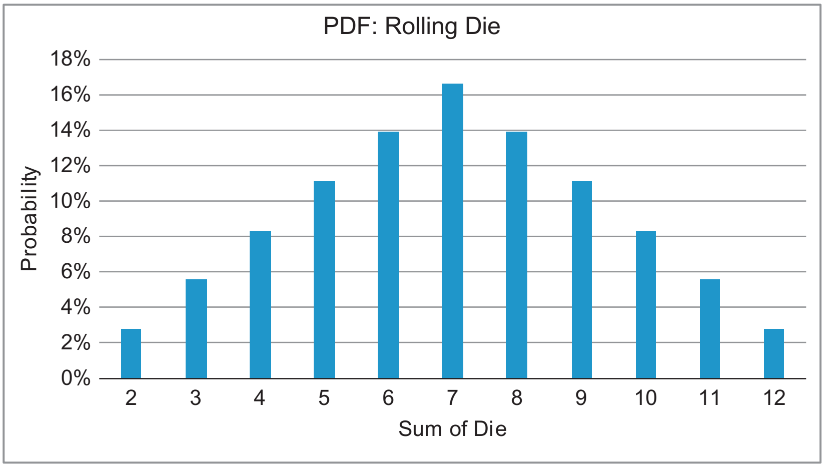


Figure 4.1 PDF: Rolling Die.

combinations are (1,6), (2,5), (3,4), (4,3), (5,2), and (6,1). Since there are 36 different combinations of outcomes from the two die, the probability of rolling a seven is 6/36 = 1/6, and thus, the pdf of 7 is 16.7%. Additionally, there are 21 ways that we can roll our die and have a value that is less than or equal to 7. Thus, the cdf is 21/36 = 58%. The pdf and cdf graphs for this example are shown in Figs. 4.1 and 4.2 respectively.

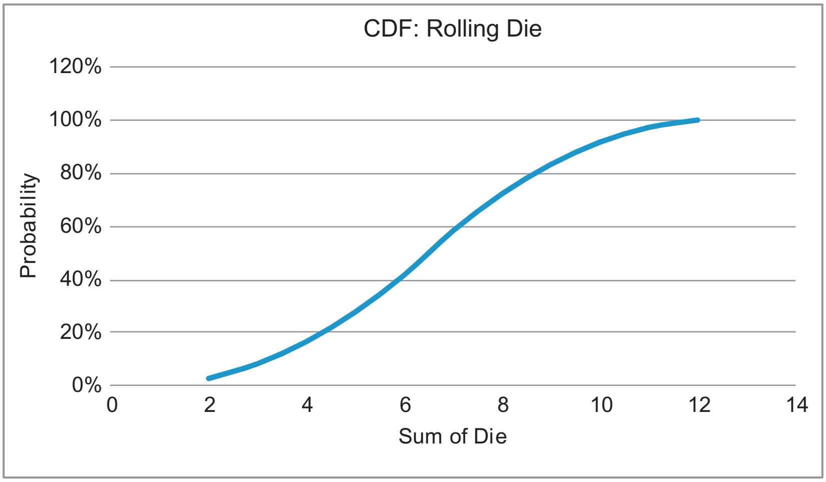


Figure 4.2 CDF: Rolling Die.

Example 4.2 Continuous probability distribution function

An example of a continuous probability distribution function can be best shown via the familiar standard normal distribution. This distribution is also commonly referred to as the Gaussian distribution as well as the bell curve.

Table 4.2 provides a sample of data for a standard normal distribution. The left-hand side of the table has the interval values *a* and *b*. The corresponding probability to the immediate right in this table shows the probability that the standard normal distribution

Table 4.2 Standard Normal Distribution

a	*b*	Pdf	*z*	Cdf
−1	1	68.3%	−3	0.1%
−2	2	95.4%	−2	2.3%
−3	3	99.7%	−1	15.9%
−inf	−1	15.9%	0	50.0%
−inf	−2	2.3%	1	84.1%
1	inf	15.9%	2	97.7%
2	inf	2.3%	3	99.9%

will have a value between *a* and *b*. That is, if *x* is a standard normal variable, the probability that *x* will have a value between *a* and *b* is shown in the probability column.

For a standard normal distribution, the values shown in column "*a*" and column "*b*" can also be thought of as the number of standard deviations where 1 = plus one standard deviation and −1 = minus one standard deviation (and the same for the other values). Readers familiar with probability and statistics will surely recall that the probability that a standard normal random variable will be between −1 and +1 is 68.3%, the probability that a standard normal variable will be between −2 and +2 is 95.4%, and the probability that a standard normal variable will be between −3 and +3 is 99.7%.

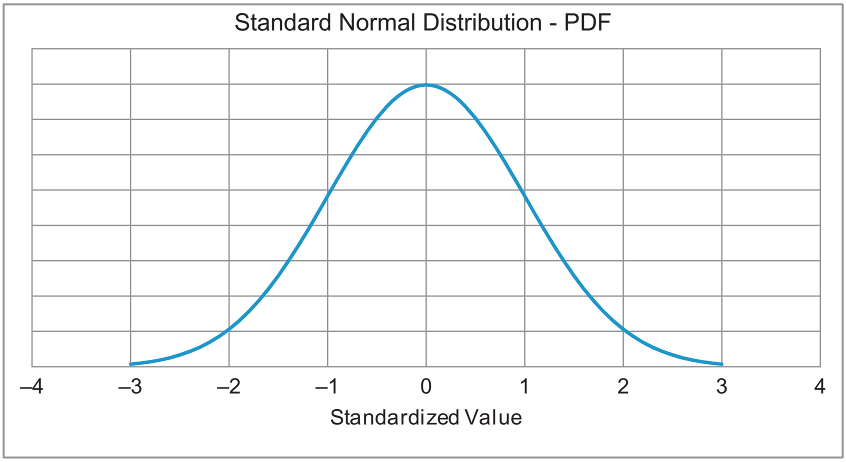


Figure 4.3 Standard Normal Distribution: PDF.

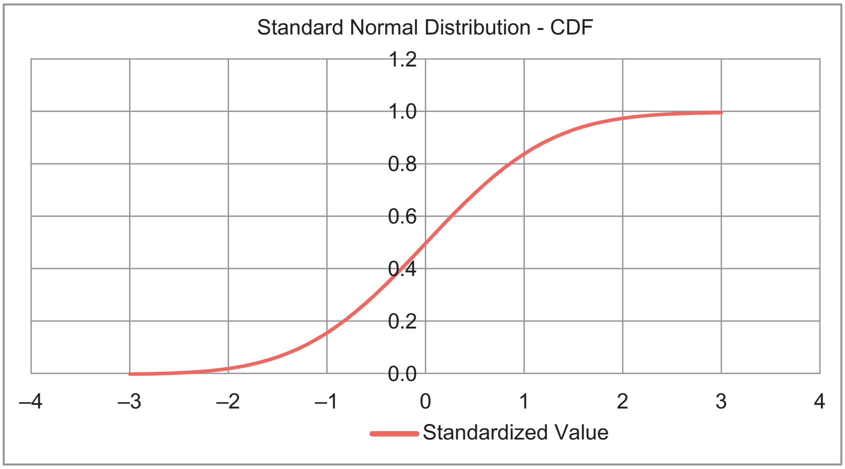


Figure 4.4 Standard Normal Distribution: CDF.

The data on the right-hand side of the table corresponds to the probability that a standard normal random value will be less than the value indicated in the column titled *z*. Readers familiar with probability and statistics will recall that the probability that a normal standard variable will be less than 0 is 50%, less than 1 is 84%, less than 2 is 97.7%, and less than 3 is 99.9%.

Fig. 4.3 illustrates a standard normal pdf distribution curve and Fig. 4.4 illustrates a standard normal cdf distribution curve. Analysts can use the pdf curves to determine the probability that an outcome event will be within a specified range and can use the cdf curves to determine the probability that an outcome event will be less than or equal to a specified value. For example, we utilize these curves to estimate the probability that a team will win a game and/or win a game by more than a specified number of points. These techniques are discussed in the subsequent sports chapters.

Important Notes:

- One of the most important items regarding computing probabilities such as the likelihood of scoring a specified number of points, winning a game, or winning by at least a specified number of points is using the proper distribution function to compute these probabilities.
- Different distribution functions will have different corresponding probability values for the same outcome value.
- It is essential that analysts perform a thorough review of the outcome variable they are looking to estimate and determine the correct underlying distribution.
- While there are many techniques that can be used to determine the proper distribution functions, analysts can gain important insight using histograms, p-p plots, and q-q plots as the starting points.
- We provide information about some of the more useful distributions below and analysts are encouraged to evaluate a full array of these distributions to determine which is most appropriate before drawing conclusions about outcomes, winning teams, scores, etc.

Descriptive Statistics

Each probability distribution has a set of descriptive statistics that can be used in analysis. The more important descriptive statistics for sports models are:

Mean: The arithmetic mean, also known as the simple mean or equal weighted mean. The mean of a data series is a unique value. The mean is also known as the first moment of the data distribution.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mode: The value(s) of a data series that occurs most often. The mode of a data series is not a unique value.

Median: The value of a data series such that one-half of the observations are lower or equal and one-half the observations are higher or equal value. The median value is not a unique number. For example, in the series 1, 2, 3 the median is the value 2. But in the series 1, 2, 3, 4 there is not a unique value. Any number $2 < x < 3$ is the median of this series since exactly 50% of the data values are lower than x and exactly 50% of the data points are higher than x. A general rule of thumb is that if there are an odd number of data points, the middle value is the median, and if there is an even number of data points, the median is selected as the mean of the two middle points. In our example, 1, 2, 3, 4 the median would be taken as 2.5. However, any value x such that $2 < x < 3$ would also be correct.

Standard Deviation: The amount of dispersion around the mean. A small standard deviation indicates that the data are all close to the mean and a high standard deviation indicates that the data could be far from the mean. The standard deviation $\sigma(x)$ is the square root of the variance $V[x]$ of the data. The variance is also known as the second moment about the distribution mean.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x-\mu)^2$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x-\mu)^2}$$

Coefficient of Variation: A measure of the standard deviation divided by the mean. The coefficient of variation serves as a normalization of the data for a fair comparison of data dispersion across different values (e.g., as a measure of data dispersion of daily or monthly stock trading volumes).

$$COV = \frac{\sigma}{\overline{x}}$$

Skewness: A measure of the symmetry of the data distribution. A positively skewed data distribution indicates that the distribution has more data on the right tail (data is positively skewed). A negatively skewed data distribution indicates that the distribution has more data on the left tail (data is negatively skewed). A skewness measure of zero indicates that the data is symmetric. Skewness is also known as the third moment about the mean.

$$\text{Skewness} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(x-\mu)^3}{\sigma}}$$

Kurtosis: A measure of the peakedness of the data distribution. Data distributions with negative kurtosis are called platykurtic distributions and data distributions with positive kurtosis are called leptokurtic distributions.

$$\text{Kurtosis} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(x-\mu)^3}{\sigma^2}}$$

Probability Distribution Functions

In this section we provide a description of the important probability distribution functions that are used in sports modeling. Readers interested in a more thorough investigation of these distributions are referred to Meyer (1970), Dudewicz and Mishra (1988), Pfeiffer (1978), DeGroot (1986).

Our summary table of the distribution statistics and moments is based on and can also be found at: www.mathworld.wolfram.com, www.wikipedia.org/, www.statsoft.com/textbook/, and www.mathwave.com/articles/distribution_fitting.html. These are excellent references and are continuously being updated with practical examples. These probability and distribution functions below are also a subset of those presented in Glantz and Kissell (2013) and used for financial risk modeling estimation.

Continuous Distribution Functions

Normal Distribution

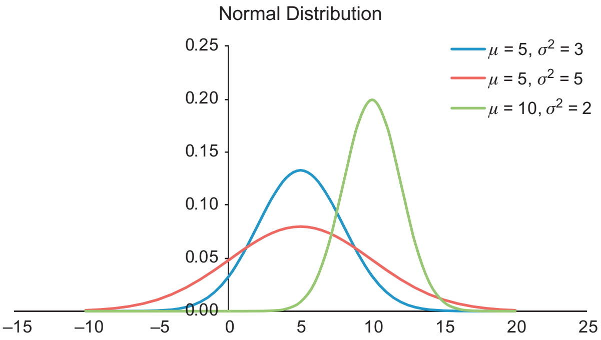
A normal distribution is the workhorse of statistical analysis. It is also known as the Gaussian distribution and the bell curve (for the distribution's resemblance to a bell). It is one of the most used distributions in statistics and is used for several different applications. The normal distribution also provides insight into issues where the data is not necessarily normal, but can be approximated by a normal distribution. Additionally, by the central limit theorem of mathematics we find that the mean of a sufficiently large number of data points will be normally distributed. This is extremely useful for parameter estimation analysis such as with our regression models.

Normal Distribution Statistics[1]	
Notation	$N(\mu, \sigma^2)$
Parameter	$-\infty < \mu < \infty$ $\sigma^2 > 0$
Distribution	$-\infty < x < \infty$
Pdf	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
Cdf	$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{2\sigma^2}\right)\right]$
Mean	μ
Variance	σ^2
Skewness	0
Kurtosis	0

where erf is the Gauss error function, i.e.,

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)$$

Normal Distribution Graph



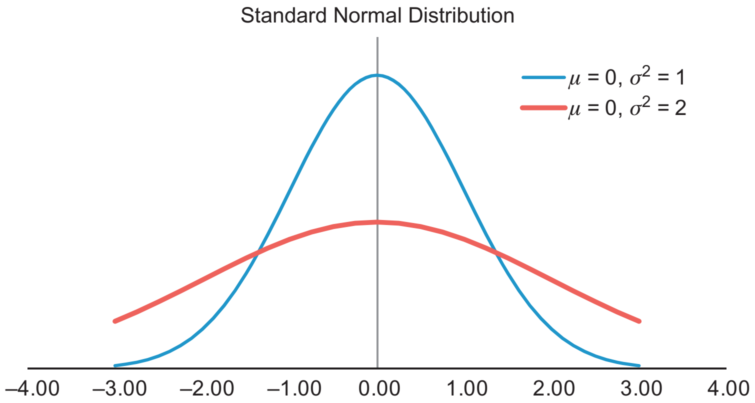
Standard Normal Distribution

The standard normal distribution is a special case of the normal distribution where $\mu = 0$, $\sigma^2 = 1$. If is often essential to normalize data prior to the analysis. A random normal variable with mean μ and standard deviation μ can be normalized via the following:

$$z = \frac{x - \mu}{\sigma}$$

Standard Normal Distribution Statistics[1]	
Notation	$N(0, 1)$
Parameter	n/a
Distribution	$-\infty < z < \infty$
Pdf	$\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} z^2\right\}$
Cdf	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{2}\right)\right]$
Mean	0
Variance	1
Skewness	0
Kurtosis	0

Standard Normal Distribution Graph

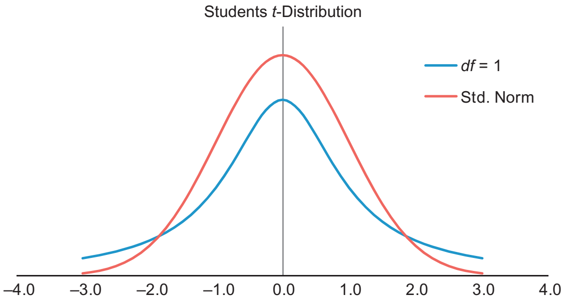


Student's t-*Distribution*

Student's *t*-distribution (a.k.a. *t*-distribution) is used when we are estimating the mean of normally distributed random variables where the sample size is small and the standard deviation is unknown. It is used to perform hypothesis testing around the data to determine if the data is within a specified range. The *t*-distribution is used in hypothesis testing of regression parameters (e.g., when developing risk factor models). The *t*-distribution looks very similar to the normal distribution but with fatter tails. But it also converges to the normal curve as the sample size increases.

Student's *t*-Distribution[1]	
Notation	$t\text{-dist}(\nu)$
Parameter	$\nu > 0$
Distribution	$-\infty < x < \infty$
Pdf	$\dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\dfrac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
Cdf	
Mean	$= \begin{cases} 0 & \nu > 1 \\ \mathit{undefined} & o.w. \end{cases}$
Variance	$= \begin{cases} \dfrac{\nu}{\nu+1} & \nu > 2 \\ \infty \quad 1 & < \nu \leq 2 \\ \mathit{undefined} & o.w. \end{cases}$
Skewness	$= \begin{cases} 0 & \nu > 3 \\ \mathit{undefined} & o.w. \end{cases}$
Kurtosis	$= \begin{cases} \dfrac{6}{\nu-4} & \nu > 4 \\ \infty & 2 < \nu \leq 4 \\ \mathit{undefined} & o.w. \end{cases}$

Student's *t*-Distribution Graph

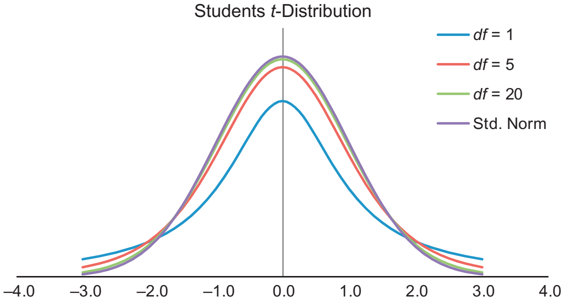


Student's *t*-Distribution: Interesting Notes

Have you ever wondered why many analysts state that you need to have at least 20 data points to compute statistics such as average or standard deviation? The reason is that once there are 20 data points, Student's *t*-distribution converges to a normal distribution. Then analysts could begin to use the simpler distribution function.

Where did the name "Student's t-distribution" come from? In many academic textbook examples, the Student's *t*-distribution is used to estimate their performance from class tests (e.g., midterms and finals, standardized tests, etc.). Therefore, the *t*-distribution is the appropriate distribution since it is a small sample size and the standard deviation is unknown. But the distribution did not arise from evaluating test scores. The Student's *t*-distribution was introduced to the world by William Sealy Gosset in 1908. The story behind the naming of the Student's *t*-distribution is as follows: William was working at the Guinness Beer Brewery in Ireland and published a paper on the quality control process they were using for their brewing process. And to keep their competitors from learning their processing secrets, Gosset published the test procedure he was using under the pseudonym Student. Hence, the name of the distribution was born.

Student's Distribution Graph
(with $k = 10, 20, 100$ and normal curve)



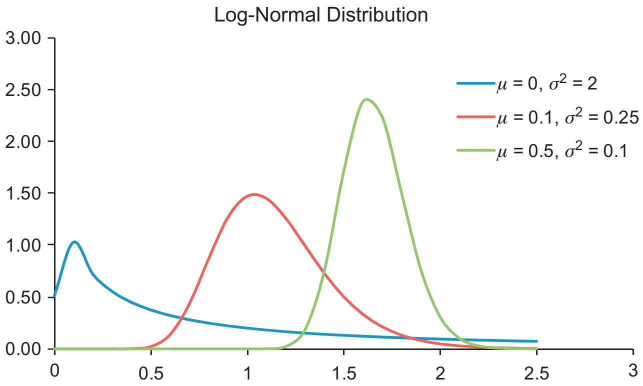
Log-Normal Distribution

A log-normal distribution is a continuous distribution of random variable y whose natural logarithm is normally distributed. For example, if random variable $y = \exp\{y\}$ has log-normal distribution then $x = \log(y)$ has normal distribution. Log-normal distributions are most often used in finance to model stock prices, index values, asset returns, as well as exchange rates, derivatives, etc.

Log-Normal Distribution Statistics[1]	
Notation	$\ln N(\mu, \sigma^2)$
Parameter	$-\infty < \mu < \infty$ $\sigma^2 > 0$
Distribution	$x > 0$
Pdf	$\frac{1}{\sqrt{2\pi}\sigma x}\exp\left\{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right\}$
Cdf	$\frac{1}{2}\left[1 + \text{erf}\left(\frac{\ln(x-\mu)}{\sigma}\right)\right]$
Mean	$e^{\left(\mu+\frac{1}{2}\sigma^2\right)}$
Variance	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$
Skewness	$(e^{\sigma^2}+2)\sqrt{(e^{\sigma^2}-1)}$
Kurtosis	$e^{4\sigma^2}+2e^{3\sigma^2}+3e^{2\sigma^2}-6$

where erf is the Gaussian error function.

Log-Normal Distribution Graph

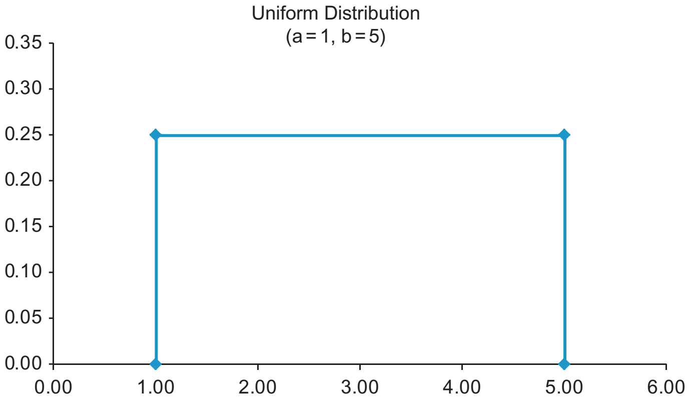


Uniform Distribution

The uniform distribution is used when each outcome has the same likelihood of occurring. One of the most illustrated examples of the uniform distribution is rolling a die where each of the six numbers has equal likelihood of occurring, or a roulette wheel where (again) each number has an equal likelihood of occurring. The uniform distribution has constant probability across all values. It can be either a discrete or continuous distribution.

Uniform Distribution Statistics[1]	
Notation	$U(a, b)$
Parameter	$-\infty < a < b < \infty$
Distribution	$a < x < b$
Pdf	$\frac{1}{b-a}$
Cdf	$\frac{x-a}{b-a}$
Mean	$\frac{1}{2}(a+b)$
Variance	$\frac{1}{12}(b-a)^2$
Skewness	0
Kurtosis	$-\frac{6}{5}$

Uniform Distribution Graph

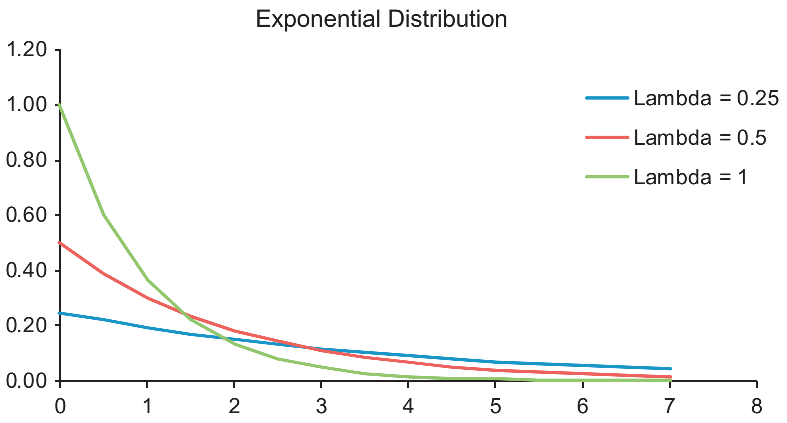


Exponential Distribution

The exponential distribution is a continuous distribution that is commonly used to measure the expected time for an event to occur. For example, in physics it is often used to measure radioactive decay, in engineering it is used to measure the time associated with receiving a defective part on an assembly line, and in finance it is often used to measure the likelihood of the next default for a portfolio of financial assets. It can also be used to measure the likelihood of incurring a specified number of defaults within a specified time period.

Exponential Distribution Statistics[1]	
Notation	$Exponential(\lambda)$
Parameter	$\lambda > 0$
Distribution	$x > 0$
Pdf	$\lambda e^{-\lambda x}$
Cdf	$1 - e^{-\lambda x}$
Mean	$1/\lambda$
Variance	$1/\lambda^2$
Skewness	2
Kurtosis	6

Exponential Distribution Graph



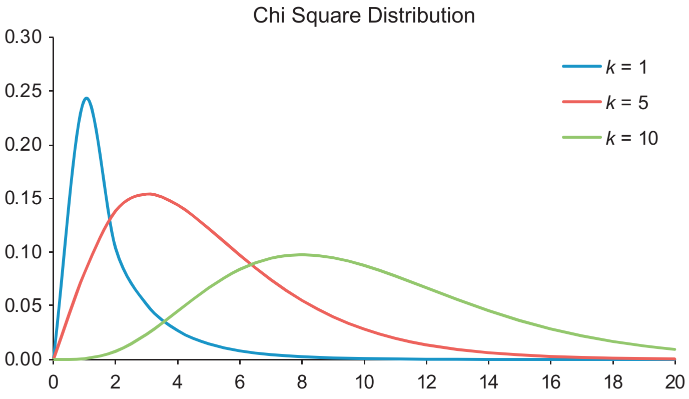
Chi-Square Distribution

A chi-square distribution is a continuous distribution with k degrees of freedom. It is used to describe the distribution of a sum of squared random variables. It is also used to test the goodness of fit of a distribution of data, whether data series are independent, and for estimating confidences surrounding variance and standard deviation for a random variable from a normal distribution. Additionally, chi-square distribution is a special case of the gamma distribution.

Chi-Square Distribution Statistics[1]	
Notation	$\chi(k)$
Parameter	$k = 1, 2, \ldots$
Distribution	$x \geq 0$
Pdf	$\left(x^{\frac{k}{2}-1}e^{-\frac{x}{2}}\right)\Big/\left(2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)\right)$
Cdf	$\gamma\left(\frac{k}{2},\frac{x}{2}\right)/\Gamma\left(\frac{k}{2}\right)$
Mean	k
Variance	$2k$
Skewness	$\sqrt{8/k}$
Kurtosis	$12/k$

where $\gamma\left(\frac{k}{2},\frac{x}{2}\right)$ is known as the incomplete Gamma function (www.mathworld.wolfram.com).

Chi-Square Distribution Graph

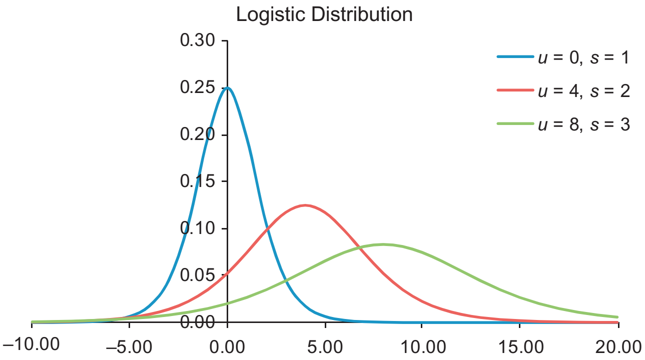


Logistic Distribution

The logistic distribution is a continuous distribution function. Both its pdf and cdf functions have been used in many different areas such as logistic regression, logit models, neural networks. It has been used in the physical sciences, sports modeling, and recently in finance. The logistic distribution has wider tails than a normal distribution so it is more consistent with the underlying data and provides better insight into the likelihood of extreme events.

Logistic Distribution Statistics[1]	
Notation	$Logistic(\mu, s)$
Parameter	$0 \leq \mu \leq \infty$ $s > 0$
Distribution	$0 \leq x \leq \infty$
Pdf	$\frac{\exp\left(-\frac{x-\mu}{s}\right)}{s\left(1+\exp\left(-\frac{x-\mu}{s}\right)\right)^2}$
Cdf	$\frac{1}{1+\exp\left(-\frac{x-\mu}{s}\right)}$
Mean	μ
Variance	$\frac{1}{3}s^2\pi^2$
Skewness	0
Kurtosis	$6/5$

Logistic Distribution Graph

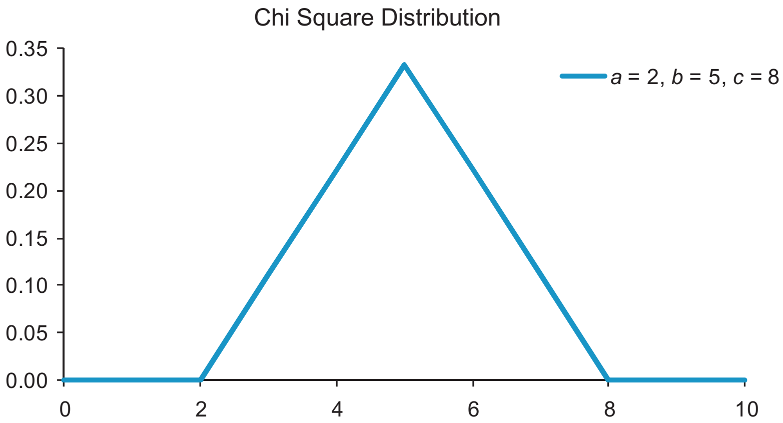


Triangular Distribution

The triangular distribution is when there is a known relationship between the variable data but when there is relatively little data available to conduct a full statistical analysis. It is often used in simulations when there is very little known about the data-generating process and is often referred to as a "lack of knowledge" distribution. The triangular distribution is an ideal distribution when the only data on hand are the maximum and minimum values, and the most likely outcome. It is often used in business decision analysis.

Triangular Distribution[1]	
Notation	Triangular(a, b, c)
Parameter	$-\infty \le a \le \infty$ $b > a$ $a < c < b$
Distribution	$a < x < b$
Pdf	$= \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)} & a \le x \le c \\ \dfrac{2(x-a)}{(b-a)(b-c)} & c \le x \le b \end{cases}$
Cdf	$= \begin{cases} \dfrac{2(x-a)^2}{(b-a)(c-a)} & a \le x \le c \\ 1 - \dfrac{(b-x)^2}{(b-a)(b-c)} & c \le x \le b \end{cases}$
Mean	$\dfrac{a+b+c}{3}$
Variance	$\dfrac{a^2+b^2+c^2-ab-ac-bc}{18}$
Skewness	$\dfrac{\sqrt{2}(a+b-2c)(2a-b-c)(a-2b+c)}{5(a^2+b^2+c^2-ab-ac-bc)^{\frac{3}{2}}}$
Kurtosis	$-\dfrac{3}{5}$

Triangular Distribution Graph



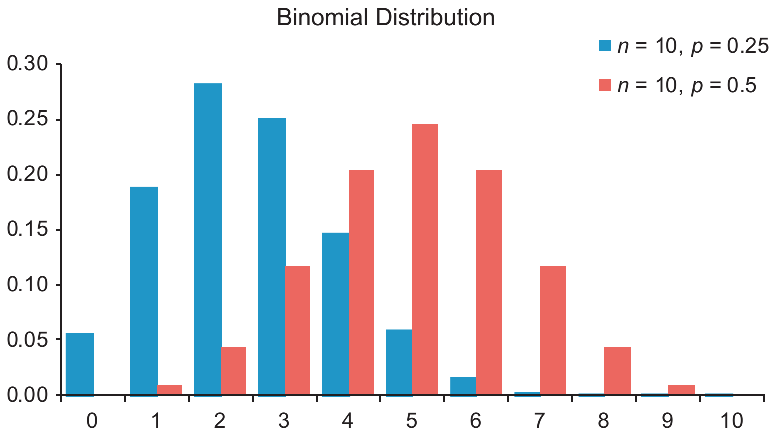
Discrete Distributions

Binomial Distribution

The binomial distribution is a discrete distribution used for sampling experiments with replacement. In this scenario, the likelihood of an element being selected remains constant throughout the data-generating process. This is an important distribution in finance in situations where analysts are looking to model the behavior of the market participants who enter reserve orders to the market. Reserve orders are orders that will instantaneously replace if the shares are transacted. For example, if an investor who has 1000 shares to buy entered at the bid may be showing 100 shares to the market at a time. Once those shares are transacted the order immediately replenishes (but the priority of the order moves to the end of the queue at that trading destination at that price). These order replenishments could occur with a reserve or iceberg type of order or via high-frequency trading algorithms where once a transaction takes place the market participant immediately submits another order at the same price and order size thus giving the impression that the order was immediately replaced.

Binomial Distribution Statistics[1]	
Notation	Binomial(n, p)
Parameter	$n \geq 0 \quad 0 \leq p \leq 1$
Distribution	$k = 1, 2, \ldots, n$
Pdf	$\left(\frac{n}{k}\right) p^k (1-p)^{n-k}$
Cdf	$\sum_{i=1}^{k} \binom{n}{i} p^i (1-p)^{n-i}$
Mean	np
Variance	$np(1-p)$
Skewness	$\frac{1-2p}{\sqrt{np(1-p)}}$
Kurtosis	$\frac{1-6p(1-p)}{np(1-p)}$

Binomial Distribution Graph

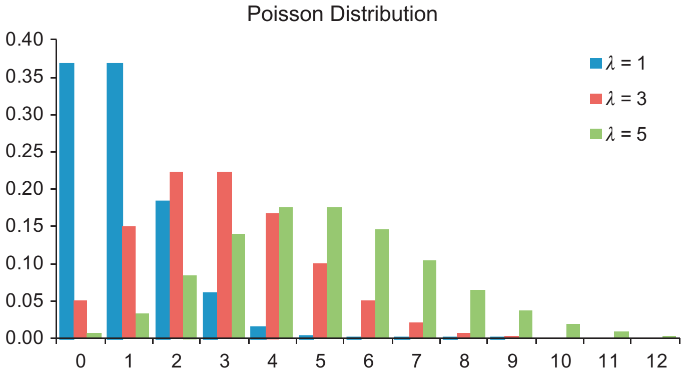


Poisson Distribution

The Poisson distribution is a discrete distribution that measures the probability of a given number of events happening in a specified time period. In finance, the Poisson distribution could be used to model the arrival of new buy or sell orders entered into the market or the expected arrival of orders at specified trading venues or dark pools. In these cases, the Poisson distribution is used to provide expectations surrounding confidence bounds around the expected order arrival rates. Poisson distributions are very useful for smart order routers and algorithmic trading.

Poisson Distribution Statistics[1]	
Notation	Poisson(λ)
Parameter	$\lambda > 0$
Distribution	$k = 1, 2, \ldots,$
Pdf	$\frac{\lambda^k e^{-\lambda}}{k!}$
Cdf	$\sum_{i=1}^{k} \frac{\lambda^k e^{-\lambda}}{k!}$
Mean	λ
Variance	λ
Skewness	$\lambda^{-1/2}$
Kurtosis	λ^{-1}

Poisson Distribution Graph



4.3 SAMPLING TECHNIQUES

What is data sampling? Data sampling is a statistical technique that is used to ascertain information about an outcome event such as a predicted score or probability of winning, or information about a specified model including the significance of the explanatory factors and parameters, or information about the underlying probability distribution using a subset of data rather than the entire data universe.

Data sampling is required when:

1. We are unable to observe and collect all data across all possible outcomes;
2. The collection of all data outcomes is not easily manageable;
3. We need to understand the accuracy of the model including significance of the parameters and distribution of the data;
4. We do not have a sufficient number of data points for a complete and thorough analysis.

For example, during a presidential election it is not possible to poll all voters to determine their favorite candidate and likely election winter. Thus, statisticians seek to draw conclusions about the likely winner using a smaller subset of data, known as a sample.

Furthermore, in sports modeling problems, we very often do not have enough observations across all teams and possible pairs of games to incorporate into our models. And in many types of sports competitions we do not have observations across all potential sets of teams to be able make a prediction based on actual data or historical data. This is the case with MLB, NFL, and college sports, as well as in international soccer or FIFA competitions, we do not have games or observations between all pairs of teams, thus making it difficult to draw conclusions. In all of these situations, we are left with making inferences and constructing models using a subset or limited amount of data.

Data sampling helps analysts resolve data limitation problems and generate outcome predictions. It allows modelers to utilize smaller data sets and/or incomplete data sets and build and test models efficiently. Data sampling, however, is associated with uncertainty and sampling error. It is required that the analyst understands the statistical error and uncertainty when making predictions about an upcoming game. As it turns out, understanding the statistical accuracy of the model and the underlying distribution of the error term is one of the most important functions of the data modeling process. In many situations, sampling of the data sample is needed to generate these error terms and to understand the distribution of these error terms. Many of the more important probability distribution functions for sports modeling problems are described above.

The remainder of this chapter will discuss different types of data sampling techniques and their use in sports modeling problems. These techniques include:

- Random Sampling
- Sampling with Replacement
- Sampling without Replacement
- Monte Carlo Techniques
- Bootstrapping Techniques
- Jackknife Sampling Techniques

4.4 RANDOM SAMPLING

Random sampling is a statistical technique that selects a data sample based upon a predefined probability that each data point may be selected for analysis. The probability levels are determined in a manner such that the underlying data subset will be most appropriate for the data modeling needs. In many cases these probability levels are specified such that each data point will have the same chance of being included and in other cases the probability levels are specified such that the expected data set will have consistent and/or similar characteristics as the data universe.

Nonrandom sampling is another sampling technique. In this case, the actual data samples are selected based on availability or ease of the data collection process. Data points are not selected based on any probability level, and thus, the likelihood of any data item being included in the subset sample will differ. This makes it difficult to make inferences about the larger data universe and introduces additional error into the modeling process. However, there are techniques that analysts can use to account for these biases. Many of these nonrandom sampling techniques are used in qualitative surveys where a surveyor stands at the front of a mall, supermarket, train station, or some other location and asks questions to people walking by. Thus, only the people who would be visiting these sites at these times could become part of the sample. These types of nonrandom sampling techniques include convenience sampling, consecutive sampling, and quota sampling techniques. These sampling techniques are not appropriate sampling techniques for sports modeling problems and will not be discussed in the text.

Resampling is a statistical technique that consists of performing an analysis, running a model, or estimating parameter values for many different data sets where these data sets are selected from the larger data universe. Resampling is an appropriate technique for many different statistical applications and can be used to estimate parameter values and probability distributions. In many situations, as mentioned above, we may not have enough data points or data observations to be able to use these metrics directly due to data limitation issues, and/or the underlying mathematical model may be too complex to calculate error terms due to data limitations.

Resampling allows analysts to estimate parameter values and probability distributions using the data samples. This then allows analysts to evaluate, test, and critique modeling approaches to determine the best and most appropriate model for problem. Resampling allows analysts to make proper statistical inferences and conclusions about future outcome events using only the data at hand.

4.5 SAMPLING WITH REPLACEMENT

Sampling with replacement is a resampling technique where each data item can be selected for and included in the data sample subset more than once. For example, suppose we have a bag of ping pong balls with numbers written on each ball. If we are interested in learning the average number written on the ping pong ball using a sampling with replacement approach, we would pick a ball at random, write down the number, and then put the ball back in the bag. Then we would pick another ball at random, write down the number, and then put the ball back in the bag. The selection of balls would be repeated for a specified number of times. Once completed, we would calculate the average across all numbers written down. In this analysis, it is quite possible to pick the same ball multiple times.

Sampling with replacement is similar to many lotto games where the player picks four numbers from 1 to 10 and where each number can be selected more than once. In this scenario, there would be 4 machines with 10 ping pong balls each numbered from 1 to 10. Then the machines would select one ball from each machine. The four numbers selected could consist of all different numbers such as 1-2-8-4 or have some or all repeated numbers such as 5-2-5-1 or 9-9-9-9.

- If a data item can be selected more than once it is considered sampling with replacement.

4.6 SAMPLING WITHOUT REPLACEMENT

Sampling without replacement is a resampling technique where each data item can be selected and used on our data sample subset only once. For example, using the same ping pong ball example where we are interested in learning the average value of the numbers on the ping pong balls the sampling without replacement would consist of picking a ball from the bag at random and writing down its value, but leaving the ball outside of the bag, and then picking another ball from the bag, writing down its value, and leaving that ball outside the bag, and repeating this process for a specified number of draws. In this case, each ball can only be selected one single time.

Sampling without replacement is similar to a Powerball type of contest where a player is asked to pick 6 numbers from 1 to 44 (or variations of this type of selection). In this scenario, each number can only be selected a single time.

- If a data item can only be selected one time than it is considered sampling without replacement.

4.7 BOOTSTRAPPING TECHNIQUES

Bootstrapping is a statistical technique that refers to random sampling of data with replacement. One of the main goals of bootstrapping is to allow analysts to estimate parameter values, corresponding standard errors, and to gain an understanding of the probability distribution of the model's error term.

In sports modeling problems, bootstrapping sampling techniques are essential for being able to calculate a statistically accurate team strength rating and also to be able to accurately predict the outcome of an event or game. This is especially true in situations where we may not have a large enough number of observations of games across all teams and/or situations where all teams may not play against each other during the season. Thus, it is important for all professional sports and college sports modeling problems.

For a bootstrapping sample, analysts could simply select a specified sample size, such as 25% of the actual data. Thus, if there are 1000 observations each sample could consist of 250 data points. Bootstrapping techniques use sampling with replacement so each data point can be selected for a sample more than one time. The model is then solved repeatedly. With today's computing power we can set the number of actual repeated samples to be quite large such as $N \gg 1000$ to allow accurate parameter estimates and confidence levels. That is, we can sample the data and solve the model 1000 times or more.

Consider the power function presented in Chapter 2, Regression Models, where we seek to maximize the following:

$$Max: \ \log L = \sum_{i=1}^{G} \ln\left(\frac{b_0 + b_h}{b_0 + b_h + b_a}\right)$$

where G is the total number of games in the sample and b_k represents the model parameters.

If we solve this optimization once we only have the parameter estimates (e.g., team strength rating) for each team and the home-field advantage term. But we do not have any estimates surrounding the standard errors of the parameters.

However, by performing bootstrapping sampling using say 25% of the games for each optimization solution and repeating this sampling technique 1000 times or more, we can calculate both the team rating parameter and the standard error of the parameter value, thus allowing us to statistically evaluate the model and mark comparisons across teams.

Using bootstrapping techniques, the expected parameters value is taken as the average value across all samples and the confidence interval or standard error can be computed using either the standard deviation of parameter estimates or computed from a specified middle percentile interval such as middle 50% or middle 68% (to be consistent with the standard deviation) of data points. It is important to note that using the standard deviation of results to compute standard errors in this case may be inaccurate in times of small sizes. Analysts will need to understand how the sample size affects the parameters estimates for their particular sports models or application.

4.8 JACKKNIFE SAMPLING TECHNIQUES

Jackknife sampling is another type of resampling technique that is used to estimate parameter values and corresponding standard deviations similar to bootstrapping. The sampling method for the jackknife technique requires that the analyst omit a single observation in each data sample. Thus, if there are n data points in the sample, the jackknife sampling technique will consist of n samples each with $n-1$ data points in each sample subset analysis. Thus, in this case, the analyst would solve the model n times each with $n-1$ data point. This would allow the analyst to estimate both parameter value and corresponding standard error.

Our research into sports modeling problems, however, finds that this jackknife technique may yield inconsistent results across teams and parameters especially in situations where a team wins all of its games or a high percentage of games, in situations where a team loses all of its games or a high percentage of its games, and/or where a team plays both very strong and very weak opponents—which is very common across college sports and also in many international tournaments such as FIFA soccer and other World Cup tournaments.

An appropriate adjustment to the jackknife sampling technique in these situations is to entirely leave out a team and all of its games in each data sample. For example, if there are 100 games across 10 teams where each team played 10 games, our jackknife sampling technique would consist of 10 samples each with 90 games. So if team A played 10 games and we are leaving team A out of this same run we would omit the 10 records with team A. While both variations of the jackknife sampling techniques have advantages and disadvantages, analysts will need to determine from the data which is the most appropriate technique to use and in which types of situations it is most appropriate and accurate.

Therefore, if we are looking to estimate team rating parameter values using the power function described in Chapter 2, Regression Models, for a scenario with 25 teams, we would solve the following optimization problem 25 times and in each optimization sample we would leave out one team and all of its games.

Thus, we would maximize the following:

$$Max: \ \log L = \sum_{i=1}^{G_k} \ln\left(\frac{b_0 + b_h}{b_0 + b_h + b_a}\right)$$

where G_k consists of all the games that did not involve team K. If there are M teams in total, we would repeat this optimization M times. Here, b_k represents the model parameters. Additionally, it is important to note that each team will have $M-1$ parameter values, one value for each scenario that included their team.

The estimated parameter values, or in this case team rating values, are computed from the optimization results across all M samples. The expected parameter value is the average across all $M-1$ results for each team. The standard error term is computed as the middle percentile values such as the middle 50% values or middle 68% values (to be consistent with standard deviation). The standard error can also be computed as the standard deviation across all $M-1$ parameter estimates.

Again, it is important for analysts to understand the effect of small samples and data limitations on their model and error terms. Analysts need to investigate the actual model error term to determine which is the most appropriate technique to quantify standard errors.

Finally, once we determine team rating parameters and corresponding standard errors, we can use this information to make statistically accurate rankings and comparisons across teams, and predict outcome events with a high level of accuracy, e.g., predict the expected winner of a game, calculate the expected winning margin, compute the winning probability.

4.9 MONTE CARLO SIMULATION

Monte Carlo simulation is a statistical technique that predicts outcomes based on probability estimates and other specified input values. These input values are often assumed to have a certain distribution or can take on a specified set of values.

Monte Carlo simulation is based on repeatedly sampling the data and calculating outcome values from the model. In each sample, the input factor and model parameters can take on different values. These values are simulated based on the distribution of the input factor and parameter values. For example, if X is an input factor for our model and X is a standard normal random variable, each simulation will sample a value of X from a standard normal distribution. Thus, each sample scenario will have a different value of X. Analysts then run repeated simulations where they can allow both the parameter values and input factors to vary based on their mean and standard error. Analysts then use these the results of these simulations to learn about the system and make better-informed future decisions.

Another important use of Monte Carlo simulation is to evaluate the performance and accuracy of a model, and also to evaluate whether or not a model or modeling methodology is appropriate for certain situation. For example, we discussed in Chapter 2, Regression Models, the power function and how it can be used in sports modeling as the basis for predicting the winning team, the winning score, and probability of winning. We can use these same Monte Carlo simulation techniques to determine whether or not this technique is appropriate for the sport we are looking to model. The process is as follows:

Suppose we want to determine if the power function and optimization process is an appropriate modeling technique to rank college football teams and to predict the winning team. Here, we apply Monte Carlo simulation as follows:

Step 1: Assign each team a rating score b_i that indicates the team's overall strength. These ratings can be assigned via a completely random process, or they can be assigned based on the team's previous years' winning records, based on conferences, etc. We suggest running multiple trials where the ratings are assigned via various methods in order to best analyze the model. It is important to note here that once we assign the team rating score b_i to each team we then know the exact ranking of teams and the exact probability that any one team will beat any other team.

Step 2: Run a simulation of game outcomes based on an actual schedule of games. Similar to Step 1, we suggest repeating this experiment using the schedules from different years in order to fully test the process and specified model.

Step 3: Determine the winner of a game based on the team's rating and power function, and based on a simulated random value (between 0 and 1). For example, if home team A is playing away team B, the probability that home team A will win the game is determined from the power function as follows:

$$P(A > B)\frac{b_0 + b_A}{b_0 + b_A + b_B}$$

If x is the randomly generated value (with a value between 0 and 1), we assign team A as the winner of the game if $x \leq P$ and we assign team B as the winner of the game if $x > P$.

Step 4: Simulate the winner of every game on the schedule across all teams during a season based on Step 3.

Step 5: Solve for each team's estimated rating value based on the outcomes of the simulated season.

Step 6: Compare the rating results obtained from the simulated season to the actual rating values used to simulate the season results. If the model is appropriate for the sport we should find a high correlation between actual ratings and estimated ratings. And, the rankings of the teams from the simulated results should be consistent with the actual rankings used to simulate the results. If either of these is found to be inconsistent with actual values used in the simulation, then the model would not be appropriate for the sport in question.

Step 7: Repeat this simulation test for various scenarios where teams are given different rating values and using schedules from different seasons.

4.10 CONCLUSION

In this chapter we provided readers with an overview of the essential mathematics required for probability and statistics modeling. The chapter included insight into different probability distribution functions and the important mathematical metrics used to describe these functions. We also provided readers with an overview of different sampling techniques and how these techniques can be used to evaluate, test, and critique sports prediction models.

ENDNOTE

1. www.mathworld.wolfram.com/topics/ProbabilityandStatistics.html

REFERENCES

DeGroot, M. H. (1986). *Probability and Statistic* (2nd ed.). New York, NY: Addison Wesley.

Dudewicz, E., & Mishra, S. (1988). *Modern Mathematical Statistics*. New York, NY: John Wiley & Sons.

Glantz, M., & Kissell, R. (2013). *Multi-Asset Risk Modeling: Techniques for a global economy in an electronic and algorithmic trading era*. Elsevier.

Meyer, P. (1970). *Introductory Probability and Statistical Applications* (2nd ed.). Addison-Wesley Publishing Company.

Pfeiffer, P. (1978). *Concepts of Probability Theory* (Second Revised Edition). Mineola, NY: Dover Publications, Inc.