Central Limit Theorem and Its Applications to Baseball

by
Nicole Anderson

A project submitted to the Department of
Mathematical Sciences in conformity with the requirements
for Math 4301 (Honours Seminar)

Lakehead University
Thunder Bay, Ontario, Canada

Abstract

This honours project is on the Central Limit Theorem (CLT). The CLT is considered to be one of the most powerful theorems in all of statistics and probability. In probability theory, the CLT states that, given certain conditions, the sample mean of a sufficiently large number or iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed.

In this project, a brief historical review of the CLT is provided, some basic concepts, two proofs of the CLT and several properties are discussed. As an application, we discuss how to use the CLT to study the sampling distribution of the sample mean and hypothesis testing using baseball statistics.

Acknowledgements

I would like to thank my supervisor, Dr. Li, who helped me by sharing his knowledge and many resources to help make this paper come to life. I would also like to thank Dr. Adam Van Tuyl for all of his help with Latex, and support throughout this project. Thank you very much!

Contents

CHAPTER 1

Introduction

1. Historical Review of Central Limit Theorem

The Central Limit Theorem, CLT for short, has been around for over 275 years and has many applications, especially in the world of probability theory. Many mathematicians over the years have proved the CLT in many different cases, therefore provided different versions of the theorem. The origins of the Central Limit Theorem can be traced to *The Doctrine of Chances* by Abraham de Moivre in 1738. Abraham de Moivre's book provided techniques for solving gambling problems, and in this book he provided a statement of the theorem for Bernoulli trails as well as gave a proof for $p = \frac{1}{2}$. This was a very important discovery at the time which inspired many other mathematicians years later to look at de Moivre's previous work and continue to prove it for other cases. [**7**]

In 1812, Pierre Simon Laplace published his own book titled *Theorie Analytique des Probabilities*, in which he generalized the theorem for $p \neq \frac{1}{2}$. He also gave a proof, although not a rigorous one, for his finding. It was not until around 1901-1902 did the Central Limit Theorem become more generalized and a complete proof was given by Aleksandr Lyapunov.

A more general statement of the Central Limit Theorem did appear in 1922 when Lindeberg gave the statement, "the sequence of random variables need not be identically distributed, instead the random variables only need zero means with individual variances small compared to their sum" [**3**].

Many other contributions to the statement of the theorem, as well as many different ways to prove the theorem began to surface around 1935, when both Levy and Feller published their own independent papers regarding the Central Limit Theorem.

The Central Limit Theorem has had, and continues to have, a great impact in the world of mathematics. Not only was the theorem used in probability theory, but it was also expanded and can be used in topology, analysis and many other fields in mathematics.

2. Central Limit Theorem in Practice

The Central Limit Theorem is a powerful theorem in statistics that allows us to make assumptions about a population and states that a normal distribution will occur regardless of what the initial distribution looks like for a sufficiently large sample size n. Many applications, such as hypothesis testing, confidence intervals and estimation, use the Central Limit Theorem to make reasonable assumptions concerning the population

since it is often difficult to make such assumptions when it is not normally distributed and the shape of the distribution is unknown.

The goal of this project is to focus on the Central Limit Theorem and its applications in statistics, as well as answer the questions, "Why is the Central Limit Theorem Important?", "How can we prove the theorem?" and "How can we apply the Central Limit Theorem in baseball?"

Our paper is structured as follows. In Chapter 2 we will first give key definitions that are important in understanding the Central Limit Theorem. Then we will give three different statements of the Central Limit Theorem. Chapter 3 will answer the second problem posed by proving the Central Limit Theorem. We will first give a proof using moment generating functions, and then we will give a proof using characteristic functions. In Chapter 4 we will answer the third problem and show that the Central Limit Theorem can be used to answer the question, "Is there such thing as a home-field advantage in baseball?" by using an important application known as hypothesis testing. Finally, Chapter 5 will summarize the results of the project and discuss future applications.

CHAPTER 2

Preliminaries

This chapter will provide some basic definitions, as well as some examples, to help understand the various components of the Central Limit Theorem. Since the Central Limit Theorem has strong applications in probability and statistics, one must have a good understanding of some basic concepts concerning random variables, probability distribution, mean and variance, and the like.

1. Definitions

There are many definitions that must first be understood before we give the statement of the Central Limit Theorem.

The following definitions can be found in [**12**].

Definition 2.1. A **population** consists of the entire collection of observations in which we are concerned.

Definition 2.2. An **experiment** is a set of positive outcomes that can be repeated.

Definition 2.3. A **sample** is a subset of the population.

Definition 2.4. A **random sample** is a sample of size n in which all observations are taken at random and assumes independence.

Definition 2.5. A **random variable**, denoted by X, is a function that associates a real number with every outcome of an experiment. We say X is a **discrete random variable** if it can assume at most a finite or a countably infinite number of possible values. A random variable is **continuous** if it can assume any value in some interval or intervals of real numbers and the probability that it assumes any specific value is 0.

Example 2.6. Consider if we wish to know how well a baseball player performed this season by looking at how often they got on base. Define the random variable X by

$$X = \begin{cases} 1, & \text{if the hitter got on base,} \\ 0, & \text{if the hitter did not get on base.} \end{cases}$$

This is an example of a random variable with a **Bernoulli distribution**.

DEFINITION 2.7. The **probability distribution** of a discrete random variable X is a function f that associates a probability with each possible value of x if it satisfies the following three properties,

1. $f(x) \geq 0$,
2. $\sum_x f(x) = 1$,
3. $P(X = x) = f(x)$.

where $P(X = x)$ refers to the probability that the random variable X is equal to a particular value, denoted by x.

DEFINITION 2.8. A **probability density function** for a continuous random variable X, denoted $f(x)$, is a function such that

1. $f(x) \geq 0$, for all x in $\mathbb{R}$,
2. $\int_{-\infty}^{+\infty} f(x)\,dx = 1$,
3. $P(a < X < b) = \int_a^b f(x)\,dx$ for all $a < b$.

DEFINITION 2.9. Let X be a discrete random variable with probability distribution function $f(x)$. The **expected value** or **mean** of X, denoted μ or $E(X)$ is

$$\mu = E(X) = \textstyle\sum_x f(x).$$

EXAMPLE 2.10. We are interested in finding the expected number of home runs that Jose Bautista will hit next season based on his previous three seasons. To do this, we can compute the expected value of home runs based on his last three seasons.

TABLE 1. Jose Bautista's Yearly Home Runs

Year	Home Runs
2011	43
2012	27
2013	28

$$\begin{aligned}\mu &= E(X)\\ &= 43f(43) + 27f(27) + 28f(28)\\ &= 43\left(\frac{1}{3}\right) + 27\left(\frac{1}{3}\right) + 28\left(\frac{1}{3}\right)\\ &= \frac{98}{3} \approx 33.\end{aligned}$$

This tells us that based on the past three seasons, Jose Bautista is expected to hit approximately 33 home runs in the 2014 season. These statistics are taken from [**5**].

Definition 2.11. Let X be a random variable with mean μ. The **variance** of X, denoted *Var(x)* or σ^2, is

$$\sigma^2 = E[X - E(X)]^2 = E(X^2) - (E(X))^2 = E(X^2) - \mu^2.$$

Definition 2.12. The **standard deviation** of a random variable X, denoted σ, is the positive square root of the variance.

Example 2.13. Using Alex Rodriguez's yearly triples from Table 2 below, compute the variance and standard deviation.

$E(X^2) = \frac{\sum X^2}{n} = \frac{\sum X^2}{20} = \frac{0^2+2^2+1^2+\cdots+0^2+1^2+0^2}{20} = \frac{96}{20} = 4.8$

$E(X) = \frac{\sum X}{n} = \frac{\sum X}{20} = \frac{0+2+1+3+\cdots+0+1+0}{20} = \frac{30}{20} = 1.5$

$\sigma^2 = E(X^2) - E(X)^2 = 4.8 - (1.5)^2 = 2.55$

$\sigma = \sqrt{2.55} = 1.5968719422671 \approx 1.6$

These statistics are taken from [**5**].

Definition 2.14. A **sampling distribution** is the probability distribution of a statistic.

Definition 2.15. A continuous random variable X is said to follow a **Normal Distribution** with mean μ and variance σ^2 if it has a probability density function

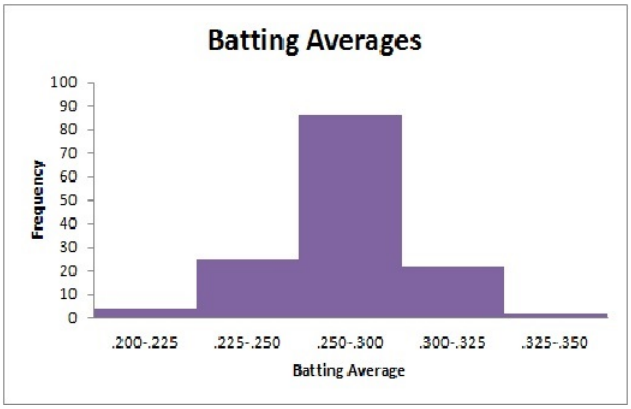
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad -\infty < x < \infty \ .$$

We write $X \backsim N(\mu, \sigma^2)$.

Example 2.16. Consider the batting averages of Major League Baseball Players in the 2013 baseball season.

TABLE 2. Alex Rodriguez Stats 1994 - 2013

Year	AVG	Triples	Home Runs
1994	.204	0	0
1995	.232	2	5
1996	.358	1	36
1997	.300	3	23
1998	.310	5	42
1999	.285	0	42
2000	.316	2	41
2001	.318	1	52
2002	.300	2	57
2003	.298	6	47
2004	.286	2	36
2005	.321	1	48
2006	.290	1	35
2007	.314	0	54
2008	.302	0	35
2009	.286	1	30
2010	.270	2	30
2011	.276	0	16
2012	.272	1	18
2013	.244	0	7



These statistics are taken from [**5**].

Taking all of their batting averages, we can see in the graph that the averages follow a "bell curve", which is unique to normal distribution. We see that the majority of players have an average between .250 and .300, and that few players have an average between .200 and .225, and .325 and .350. This gives a perfect example of how normal distribution

can help approximate even discrete random variables. Just by looking at the graph we can make some inferences about the population.

2. Central Limit Theorem

Over the years, many mathematicians have contributed to the Central Limit Theorem and its proof, and therefore many different statements of the theorem are accepted.

The first statement of the theorem is widely known as the de Moivre-Laplace Theorem, which was the very first statement of the Central Limit Theorem.

THEOREM 2.17. [**3**] *Consider a sequence of Bernoulli trials with probability p of success, where* $0 < p < 1$. *Let* S_n *denote the number of successes in the first n trials,* $n \geq 1$. *For any* $a, b \in \mathbb{R} \cup \{\pm\infty\}$ *with* $a < b$,

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{\frac{-z^2}{2}} dz.$$

Another statement of the Central Limit Theorem was given by Lyapunov which states:

THEOREM 2.18. [**8**] *Suppose* $X_n, n \geq 1$, *are independent random variables with mean 0 and* $\sum_{k=1}^{n} \frac{E(|X_k|^\delta)}{s_n^\delta} \to 0$ *for some* $\delta > 2$, *then*

$$\frac{S_n}{s_n} \xrightarrow{distr} N(0,1),$$

where $S_n = X_1 + X_2 + ... + X_n$, $s_n = \sum_{k=1}^{n} E(X_k^2), n \geq 1$ *and where* $\xrightarrow{distr}$ *represents convergence in distribution.*

Before giving the final statement of the Central Limit Theorem, we must define what it means for random variables to be independent and identically distributed.

DEFINITION 2.19. A sequence of random variables is said to be **independent and identically distributed** if all random variables are mutually independent, and if each random variable has the same probability distribution.

We will now give the final statement of the Central Limit Theorem, a special case of the Lindeberg-Feller theorem. This statement is the one we will use throughout the rest of the paper.

THEOREM 2.20. [**8**] *Suppose* $X_1, X_2, \cdots, X_n$ *are independent and identically distributed with mean* μ *and variance* $\sigma^2 > 0$. *Then,*

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{distr} N(0,1),$$

where $S_n = X_1 + X_2 + ... + X_n, n \geq 1$ *and* $\xrightarrow{distr}$ *represents convergence in distribution.*

CHAPTER 3

Proofs of Central Limit Theorem

There are many ways to prove the Central Limit Theorem. In this chapter we will provide two proofs of the Central Limit Theorem. The first proof uses moment generating functions, and the second uses characteristic functions.

We will first prove the Central Limit Theorem using moment generating functions.

1. Proof of Central Limit Theorem Using Moment Generating Functions

Before we give the proof of the Central Limit Theorem, it is important to discuss some basic definitions, properties and remarks concerning moment generating functions. First, we will give the definition of a moment generating function as follows:

DEFINITION 3.1. The **moment-generating function** (MGF) of a random variable X is defined to be

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} f(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{+\infty} e^{tx} f(x) dx, & \text{if } X \text{ is continuous.} \end{cases}$$

Moments can also be found by differentiation.

THEOREM 3.2. *Let X be a random variable with moment-generating function $M_X(t)$. We have*

$$\frac{d^r M_X(t)}{dt^r}|_{t=0} = \mu_r^{'},$$

where $\mu_r^{'} = E(X^r)$.

REMARK 3.3. $\mu_r^{'} = E(X^r)$ describes the *rth moment about the origin* of the random variable X. We can see then that $\mu_1^{'} = E(X)$ and $\mu_2^{'} = E(X^2)$ which therefore allows us to write the mean and variance in terms of moments.

Moment generating functions also have the following properties.

THEOREM 3.4. $M_{a+bX}(t) = E(e^{t(a+bX)}) = e^{at} M_X(bt)$.

PROOF. $M_{a+bX}(t) = E[e^{t(a+bX)}] = E(e^{at}) \cdot E(e^{t(bX)}) = e^{at} E(e^{(bt)X}) = e^{at} M_X(bt)$. □

THEOREM 3.5. *Let X and Y be random variables with moment-generating functions $M_X(t)$ and $M_Y(t)$ respectively. Then*

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

PROOF. $M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX} \cdot e^{tY}) = E(e^{tX}) \cdot E(e^{tY})$ (by independence of random variables) $= M_X(t) \cdot M_Y(t)$. □

COROLLARY 3.6. *Let $X_1, X_2, ..., X_n$ be random variables, then*

$$M_{X_1+X_2+...+X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdots M_{X_n}(t).$$

The proof is nearly identical to the proof of the previous theorem.

To prove the Central Limit Theorem, it is necessary to know the moment generating function of the normal distribution:

LEMMA 3.7. *The moment generating function (MGF) of the normal random variable X with mean μ and variance σ^2, (i.e., $X \sim N(\mu, \sigma^2)$) is*

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

PROOF. First we will find the MGF for the normal distribution with mean 0 and variance 1, i.e, $N(0,1)$. If $Y \backsim N(0,1)$, then

$$\begin{aligned}
M_Y(t) &= E(e^{tY}) \\
&= \int_{-\infty}^{+\infty} e^{ty} f(y) dx \\
&= \int_{-\infty}^{+\infty} e^{ty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \right) dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{ty} e^{-\frac{1}{2}y^2} dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{(ty - \frac{1}{2}y^2)} dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{(\frac{1}{2}t^2 + [-\frac{1}{2}(y^2+2ty+t^2)])} dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\frac{1}{2}t^2} e^{-\frac{1}{2}(y^2-2ty+t^2)} dy \\
&= e^{\frac{1}{2}t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(y-t)^2} dy.
\end{aligned}$$

But note that by Definition 2.14, $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(y-t)^2} dy$ is just the probability distribution function of normal distribution. So

$$M_Y(t) = e^{\frac{1}{2}t^2}.$$

Now, if $X \backsim N(\mu, \sigma^2)$, and by Theorem 3.3,

$$\begin{aligned} M_X(t) &= M_{\mu+\sigma Y}(t) \\ &= e^{\mu t} M_Y(\sigma t) \\ &= e^{\mu t} e^{(\frac{1}{2}\sigma^2 t^2)} \\ &= e^{(\mu t + \frac{\sigma^2 t^2}{2})}. \end{aligned}$$

□

Before we begin the proof of the Central Limit Theorem, we must recall the following remark from calculus:

LEMMA 3.8. $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$

Now we are ready to prove the Central Limit Theorem. We will prove a special case of where $M_X(t)$ exists in a neighbourhood of 0.

PROOF. (of Theorem 2.20) Let $Y_i = \frac{X_i - \mu}{\sigma}$ for $i = 1, 2, 3, ...$ and $R_n = Y_1 + Y_2 + ... + Y_n$. So we have

$$\begin{aligned} \frac{S_n - n\mu}{\sqrt{n\sigma^2}} &= \frac{Y_1 + Y_2 + ... + Y_n}{\sqrt{n}} \\ &= \frac{R_n}{\sqrt{n}}. \end{aligned}$$

So

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{R_n}{\sqrt{n}} = Z_n.$$

Since R_n is the sum of independent random variables, we see that its moment generating function is

$$\begin{aligned} M_{R_n}(t) &= M_{Y_1}(t) M_{Y_2}(t) \cdots M_{Y_n}(t) \\ &= [M_Y(t)]^n \end{aligned}$$

by Corollary 3.5. We note that this is true because each Y_i is independent and identically distributed. Now,

$$M_{Z_n}(t) = M_{\frac{R_n}{\sqrt{n}}}(t) = E\left(e^{\frac{R_n}{\sqrt{n}}(t)}\right) = E\left(e^{(R_n)(\frac{t}{\sqrt{n}})}\right) = M_{R_n}\left(\frac{t}{\sqrt{n}}\right) = \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

Taking the natural logarithm of each side,

$$lnM_{Z_n}(t) = nlnM_Y(\tfrac{t}{\sqrt{n}}).$$

But note along with using Remark 3.7 that,

$$\begin{aligned} M_Y\left(\frac{t}{\sqrt{n}}\right) &= E\left(e^{\frac{t}{\sqrt{n}}Y}\right) \\ &= E\left(1 + \frac{tY}{\sqrt{n}} + \frac{(\frac{t^2Y}{\sqrt{n}})^2}{2} + O\left(\frac{1}{n^{\frac{3}{2}}}\right)\right) \\ &= 1 + \frac{t^2}{n}\frac{E(Y^2)}{2} + O\left(\frac{1}{n^{\frac{3}{2}}}\right) \\ &= 1 + \frac{t^2}{2n} + O\left(\frac{1}{n^{\frac{3}{2}}}\right). \end{aligned}$$

where $O\left(\frac{1}{n^\alpha}\right)$ stands for $\limsup_{n\to\infty} \frac{\left|O\left(\frac{1}{n^\alpha}\right)\right|}{\frac{1}{n^\alpha}} < \infty$. Then

$$\begin{aligned} lnM_{Z_n}(t) &= nln\left(1 + \frac{t^2}{2n} + O\left(\frac{1}{n^{\frac{3}{2}}}\right)\right) \\ &= n\left(\frac{t^2}{2n} + O\left(\frac{1}{n^{\frac{3}{2}}}\right)\right) \\ &= \frac{t^2}{2} + O\left(\frac{1}{n^{\frac{1}{2}}}\right). \end{aligned}$$

We see that

$$lnM_{Z_n}(t) = \tfrac{t^2}{2} + O\left(\tfrac{1}{n^{\frac{1}{2}}}\right),$$

So we have,

$$M_{Z_n}(t) \to e^{\frac{t^2}{2}} \text{ as } n \to \infty.$$

Thus, $Z_n \to N(0,1)$, i.e, $\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \to N(0,1)$. □

2. Proof of Central Limit Theorem Using Characteristic Functions

Now we will prove the Central Limit Theorem another way by looking at characteristic functions.

Moment generating functions do not exist for all distributions. This is because some moments of the distributions are not finite. In these instances, we look at another general function known as the characteristic function.

DEFINITION 3.9. The **characteristic function** of a continuous random variable X is

$$C_X(t) = E(e^{itX}) = \int_{-\infty}^{+\infty} e^{itx} f(x)dx,$$

where t is a real valued function, and $i = \sqrt{-1}$.

$C_X(t)$ will always exist because e^{itx} is a bounded function, that is, $|e^{itx}| = 1$ for all $t, x \in \mathbb{R}$, and so the integral exists. The characteristic function also has many similar properties to moment generating functions.

To prove the central limit theorem using characteristic functions, we need to know the characteristic function of the normal distribution.

LEMMA 3.10. *Let $R_n, n \geq 1$ be a sequence of random variables. If, as $n \to \infty$,*

$$C_{R_n}(t) = E\left(e^{iR_n t}\right) \to e^{\frac{-t^2}{2}}$$

for all $t \in (-\infty, \infty)$, then $R_n \to N(0, 1)$.

We can now prove the Central Limit Theorem using characteristics functions.

PROOF. (Of Theorem 2.20) Similar to the proof using moment generating functions, let $Y_i = \frac{X_i - \mu}{\sigma}$ for $i = 1, 2, 3...$ and let $R_n = Y_1 + Y_2 + ... + Y_n$ so,

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{R_n}{\sqrt{n}} = Z_n,$$

where $S_n = X_1 + X_2 + ... + X_n$.
Note that R_n is the sum of independent random variables, so we see that the characteristic function of R_n is

$$\begin{aligned} C_Y(t) &= C_{Y_1}(t)C_{Y_2}(t)\cdots C_{Y_n}(t) \\ &= [C_Y(t)]^n \end{aligned}$$

since all Y_i's are independent and identically distributed. Now,

$$\begin{aligned} C_{Z_n}(t) &= C_{\frac{R_n}{\sqrt{n}}}(t) \\ &= E[e^{i\frac{R_n}{\sqrt{n}}t}] \\ &= E[e^{i}(R_n)\left(\frac{t}{\sqrt{n}}\right)] \\ &= C_{R_n}\left(\frac{t}{\sqrt{n}}\right) \\ &= [C_Y\left(\frac{t}{\sqrt{n}}\right)]^n. \end{aligned}$$

Taking the natural logarithm of each side,

$$lnC_{Z_n}(t) = nlnC_Y\left(\tfrac{t}{\sqrt{n}}\right).$$

We can note from the previous proof with some modifications that

$$C_Y\left(\tfrac{t}{\sqrt{n}}\right) = 1 - \tfrac{t^2}{2n} + O\left(\tfrac{1}{n^{\frac{3}{2}}}\right).$$

Then,

$$lnC_{Z_n}(t) = nln(1 - \tfrac{t^2}{2n} + O(\tfrac{1}{n^{\frac{3}{2}}})).$$

Using Remark 3.8, we see that

$$lnC_{Z_n}(t) = -\tfrac{t^2}{2} + O(\tfrac{1}{n^{\frac{1}{2}}}),$$

So, as $n \to \infty, lnC_{Z_n}(t) \to -\frac{t^2}{2}$ and

$$C_{Z_n}(t) \to -e^{\frac{t^2}{2}} \text{ as } n \to \infty.$$

Thus by Lemma 3.10, we conclude that

$$Z_n = \tfrac{S_n - n\mu}{\sqrt{n\sigma^2}} \to N(0,1).$$

□

CHAPTER 4

Applications of the Central Limit Theorem in Baseball

The Central Limit Theorem has many applications in probability theory and statistics, but one very interesting application is known as *hypothesis testing*. This chapter will focus on the application of hypothesis testing, and in particular, answer the following question:

PROBLEM 4.1. *Is there such thing as a home-field advantage in Major League Baseball?*

Before we begin, there are a few definitions that must be understood.

DEFINITION 4.2. A conjecture concerning one or more populations is known as a **statistical hypothesis.**

DEFINITION 4.3. A **null hypothesis** is a hypothesis that we wish to test and is denoted H_0.

DEFINITION 4.4. An **alternative hypothesis** represents the question to be answered in the hypothesis test and is denoted by H_1.

REMARK 4.5. The *null hypothesis* H_0 opposes the *alternative hypothesis* H_1. H_0 is commonly seen as the complement of H_1. Concerning our problem, the null hypothesis and the alternative hypothesis are:

H_0: There is no home-field advantage,
H_1: There is a home-field advantage.

When we do a hypothesis test, the goal is to determine if we will reject the null hypothesis or if we fail to reject the null hypothesis. If we **reject** H_0, we are in favour of H_1 because of sufficient evidence in the data. If we **fail to reject** H_0, then we have insufficient evidence in the data.

DEFINITION 4.6. A **test statistic** is a sample that is used to determine whether or not a hypothesis is rejected or not.

DEFINITION 4.7. A **critical value** is a cut off value that is compared to the test statistic to determine whether or not the null hypothesis is rejected.

DEFINITION 4.8. The **level of significance** of a test statistic is the probability that H_0 is rejected, although it is true.

DEFINITION 4.9. A **z-score** or **z-value** is a number that indicates how many standard deviations an element is away from the mean.

Definition 4.10. A **confidence interval** is an interval that contains an estimated range of values in which an unknown population parameter is likely to fall into.

Remark 4.11. If the test statistic falls into the interval, then we *fail to reject* H_0, but if the test statistic is not in the interval, then we *reject* H_0.

Definition 4.12. A **p-value** is the lowest level of significance in which the test statistic is significant.

Remark 4.13. We reject H_0 if the p-value is very small, usually less than 0.05.

Now to return to our problem, is there such thing as a home-field advantage? How can we test this notion? In the 2013 Major League Baseball season, there were 2431 games played, and of those games, 1308 of them were won at home. This indicates that approximately 53.81% of the games played were won at home. We will let our observed value be this value, so $\hat{p} = 0.5381$. It seems as though there is such thing as a home-field advantage, but we must test this notion to be certain. To do this, we will test the hypothesis that there is no such thing as a home-field advantage, so our *null hypothesis* will be

$$H_0 : p = 0.50$$

That is, 50% of the Major League Baseball games are won at home, hence, there is no home-field advantage.

Our *alternative hypothesis* will be $H_1 : p > 0.50$.

If there is no home-field advantage, then we would expect our proportion to be 0.50, since half of the games would be won at home and the other half on the road.

Before we begin to compute if there is such thing as a home-field advantage we must first satisfy four conditions; independence assumption, random condition, 10% condition, and the success/failure condition. These conditions will assure that we can test our hypothesis.

Each game is *independent* of one another and one game does not effect how another game is played. Although in some cases when a key batter or pitcher is injured, the team may not do as well in the immediate upcoming games, but roughly speaking, the games played are generally independent of one another, and so our independence condition holds.

Since there have been many games played over the years, each year having roughly 2430 games, it can be seen that taking just one year to observe the data will account for

our *randomization* condition.

Also, as stated above, we can see that the 2431 games played in the 2013 season, are less than 10% of the total games played over the years that Major League Baseball has been around, so our 10% *condition* also holds true, that is, the sample size is no more than 10% of the population.

Finally we must check that the number of games multiplied by our proportion of 0.50, is larger than 10. So we have

$$np = 2431(0.50) = 1215.5$$

which is larger than 10, so our *success/failure* condition holds as well. Since all of these conditions are met, we are now able to use the Normal Distribution model to help us test our hypothesis. We will test our hypothesis using two different methods: the first by using a confidence interval, and the second using a p-value.

First, we will test our hypothesis using a confidence interval. For testing

$$H_0 : p = 0.50 \text{ vs. } H_1 : p > 0.50$$

at the 0.05 level of significance, we may construct a *right-sided* 95% confidence interval for p. If our test statistic of $p = 0.50$ is in the interval, then we fail to reject H_0 at the 0.05 level of significance. If $p = 0.50$ is not in the interval, we reject H_0. The *right-sided* $100(1-\alpha)\%$ *confidence interval* for p for a large sample is given by

$$\hat{p} - z_\alpha \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} < p \leq 1$$

where α is the level of significance.

Since $n = 2431, \hat{p} = 0.5381$, and $\alpha = 0.05$, we see from the Normal Distribution table in the Appendix that $z_{0.05} = 1.645$. So a right-sided 95% confidence interval for p is

$$\begin{gathered} 0.5381 - 1.645\sqrt{\tfrac{(0.5381)(1-0.5381)}{2431}} < p \leq 1 \\ 0.5381 - 1.645(0.0101114) < p \leq 1 \\ 0.5215 < p \leq 1. \end{gathered}$$

Since $0.50 \notin (0.5215, 1]$, we reject $H_0 : p = 0.50$ in favour of $H_1 : p > 0.50$ at the 0.05 level of significance, that is, we have enough evidence to support that there is a home-field advantage, and the home team wins more than 50% of the games played at home.

Now we will use the p-value approach to test our hypothesis. We must find the **z-value** for testing our observed value. We use the following equation to do so;

$$z = \frac{(\hat{p}-p_o)}{\sqrt{\frac{p_o q_o}{n}}}$$

Now, with $p = 0.50, \hat{p} = 0.5381$, and $n = 2431$, we have

$$z = \frac{(\hat{p}-p_o)}{\sqrt{\frac{p_o q_o}{n}}} = \frac{0.5381-0.5}{\sqrt{\frac{0.5 \cdot 0.5}{2431}}} = \frac{0.0381}{0.010140923} = 3.76$$

This results in a p-value < 0.0001.

So we can conclude, since the p-value < 0.0001 is less than 0.05, we reject H_0. That is, the data seems to support that the home field team wins more than 50% of the time, and hence there is such thing as a home-field advantage in Major League Baseball.

We have shown that taking all of the games played in the 2013 Major League Baseball season, that there is a home-field advantage, but is there a difference between the American League and the National League? Do both leagues have a home-field advantage? We will test this notion using a $100(1-\alpha)\%$ confidence interval at the 0.01 level of significance. This will allow us to be 99% confident of our results.

In the 2013 season, the National League played 1211 games, and won 660 of those games at home. So this indicates that approximately 54.5% of the games were won at home. As we calculated above, we will let the observed value be $\hat{p} = 0.545$ and we will test the same hypothesis, that is,

$$H_0 : p = 0.50 \text{ vs. } H_1 : p > 0.50$$

Since $n = 1211, \hat{p} = 0.545$ and $\alpha = 0.01$, we can see from the Normal Distribution table in the Appendix that $z_{0.01} = 2.33$. So a right-sided 99% confidence interval for p is

$$\begin{gathered} 0.545 - 2.33\sqrt{\frac{(0.545)(1-0.545)}{1211}} < p \leq 1 \\ 0.545 - 2.33(0.014309744) < p \leq 1 \\ 0.5117 < p \leq 1. \end{gathered}$$

Since $0.50 \notin (0.5117, 1]$, we reject $H_0 : p = 0.50$ in favour of $H_1 : p > 0.50$. So we can conclude that the National League has a home-field advantage. Will the same be true for the American League? We will again test the same hypothesis, using a 99% confidence interval for the American League.

In the 2013 season, the American League played slightly more games than the National League. They played 1220 games and of those games, 648 of them were won at home. So this indicates that approximately 53.11% of the games played were won at home. Once again, let our observed value be $\hat{p} = 0.5311$, and testing the same hypothesis above, we

see that a 99% confidence interval for p is

$$\begin{gathered} 0.5311 - 2.33\sqrt{\tfrac{(0.5311)(1-0.5311)}{1220}} < p \leq 1 \\ 0.5311 - 2.33(0.01428724) < p \leq 1 \\ 0.4978 < p \leq 1. \end{gathered}$$

Since $0.50 \in (0.4978, 1]$, we fail to reject $H_0 : p = 0.50$. That is, we do not have enough evidence to support that there is a home-field advantage in the American League.

We can see that by testing these hypotheses for the National League and the American League, that we can confidently state that there is a home-field advantage in the National League, but we cannot say the same thing for the American League based on the 2013 Major League Baseball season.

CHAPTER 5

Summary

The Central Limit Theorem is very powerful in the world of mathematics and as numerous applications in probability theory as well as statistics. In this paper, we have stated the Central Limit Theorem, proved the theorem two different ways, one using moment generating functions and another using characteristic functions, and finally showed an application of the Central Limit Theorem by using hypothesis testing to answer the question, "Is there such thing as a home-field advantage?"

We proved that we could express normal distribution in terms of a moment generating function, and used this to prove the Central Limit Theorem, by showing that the moment generating function converges to the normal distribution model. We then applied our results from the first proof using moment generating functions to characteristic functions, noting that moment generating functions are not always defined, and once again arrived at the same conclusion and proving the Central Limit Theorem. In our final chapter, we successfully proved by taking statistics from the 2013 baseball season and using confidence intervals, as well as a p-value, to show that there is indeed such thing as a home-field advantage in Major League Baseball. We also showed that we can come to the same conclusion about the National League, but we do not have enough evidence to show that there is a home-field advantage in the American League.

In the future, it may be interesting to use my application on other sports such as hockey, or football, although we must make sure that we have a sufficiently large sample size to have accurate results. Other applications of the Central Limit Theorem, as well as other properties such as convergence rates may also be interesting areas of study for the future.

CHAPTER 6

Appendix

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9031	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Bibliography

[1] Albert, Jim. *Teaching Statistics Using Baseball*. Washington, DC: The Mathematical Association of America, 2003.

[2] *Characteristic Functions and the Central Limit Theorem*. University of Waterloo. Chapter 6. Web. http://sas.uwaterloo.ca/ dlmcleis/s901/chapt6.pdf.

[3] Dunbar, Steven R. The de Moivre-Laplace Central Limit Theorem. *Topics in Probability Theory and Stochastic Processes*. http://www.math.unl.edu 1, 7

[4] Emmanuel Lesigne. *Heads or Tails: An Introduction to Limit Theorems in Probability*, Vol **28** of *Student Mathematical Library*. American Mathematical Society, 2005.

[5] *ESPN.com*. 2013. ESPN Internet Ventures. Web. http://espn.go.com/mlb. 5, 6

[6] Filmus, Yuval. Two Proofs of the Central Limit Theorem. Jan/Feb 2010. Lecture. www.cs.toronto.edu/ yuvalf/CLT.pdf

[7] Grinstead, Charles M., and J. Laurie Snell. Central Limit Theorem. *Introduction to Probability*. Dartmouth College. 325-364. Web. http://www.dartmouth.edu/chance/teaching_aids/books_articles/probability_book/Chapter9.pdf. 1

[8] Hildebrand, A.J. The Central Limit Theorem. Lecture. http://www.math.uiuc.edu/hildebr/370/408clt.pdf 7

[9] Introduction to The Central Limit Theorem. *The Theory of Inference*. NCSSM Statistics Leadership Institute Notes. Web. http://courses.ncssm.edu/math/Stat_Inst/PDFS/SEC_4_f.pdf

[10] Krylov, N.V. An Undergraduate Lecture on The Central Limit Theorem. Lecture. www.math.umn.edu/∼krylov/CLT1.pdf

[11] *Moment Generating Functions*. Chapter 6. Web. http://www.am.qub.ac.uk/users/g.gribakin/sor/Chap6.pdf.

[12] Walpole, Ronald E, Raymond H. Myers, Sharon L. Myers, and Keying Ye. *Probability & Statistics For Engineers & Scientists.* Prentice Hall. 2012. 3