Order Statistics and Applications to Basketball Data

by Nicholas Burke

A project submitted to the Department of Mathematical Sciences in conformity with the requirements for Math 5301 (Graduate Seminar)

Lakehead University Thunder Bay, Ontario

Abstract

This Graduate project will discuss the topic of order statistics. Order statistics can provide efficient linear unbiased estimates of parameters, such as mean and standard deviation. Thus it can be used various aspects of life such as health care, finance and sports. This project will include a historical review of order statistics, some definitions as well as some theoretical properties to help further elaborate on this topic. This project will also included an application with respect to basketball data, and a simulation.

Acknowledgements

I have had the pleasure to work with both Dr. Deli Li and Dr. Liping Liu on this project. Their expertise has been greatly appreciated and vital to the completion of this project. I also want to thank everyone who has supported me throughout my graduate school journey.

Contents

1 Introduction **4**

1.1 Historical Review 8

1.2 Outline of Project 8

2 Basic Notations and Definitions **9**

2.1 Definitions 9

2.2 Distributions 10

3 Theoretical Properties of Order Statistics **15**

3.1 Probability Distribution of Ordered Statistics 15

3.2 Special Case 16

4 Applications to Basketball Data **18**

4.1 Confidence intervals 18

5 Simulation **23**

5.1 Simulating Free Throws 24

1 Introduction

The objective of this project is to define order statistics, indulge into some theoretical aspects, and conclude with both an application and a simulation. This project will demonstrate the usefulness of order statistics in basketball data. Order statistic can be used in any type of data in which the values differ.

Review of Basic Probability

Firstly, we shall introduce some elementary concepts and definitions of probability that will help supplement the understanding of the terminology used throughout this project.

Definition 1.1. *A random variable is the outcome of a natural process that can not be predicted with certainty.*

Definition 1.2. *A sample space is the set of all possible outcomes for a random variable. Every point in the sample space has a corresponding probability, which is between 0 and 1.*

Definition 1.3. *A distribution is the sample space together with all probabilities. The sum of all probabilities must equal to 1.*

Definition 1.4. *A simple random sample is sample of size n and is independent and identically distributed if a random variable can be observed repeatedly and independently n times.*

Definition 1.5. *A probability mass function (pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value.*

$$p(x) = P(X = x)$$

where $p(x) \geq 0$ *and for all* $x \sum p(x) = 1$.

Definition 1.6. *A probability density function (pdf) is a function that describes the likelihood for a random variable to take on a given value.*

$$P\{X = x\} = f(x)$$

where $f(x) \leq 1$ *for all* x *and* $\sum f(x) = 1$.

Definition 1.7. *A cumulative distribution function (cdf) is the probability that the variable takes a value less than or equal to x.*

$$F(x) = Pr\{X \leq x\}$$

Definition 1.8. *A joint probability distribution shows a probability distribution for two or more random variables.*

$$f(x, y) = P(X = x, Y = y)$$

Definition 1.9. *A joint cumulative distribution function represents the probability that the random variable X takes on a values less than or equal to x and that Y takes on a values less than or equal to y.*

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

Definition 1.10. *A quantile function is the inverse of the cdf.*

$$F(Q(p)) = P(X \leq Q(p)) = p$$

the point $Q(p)$ in the sample space such that with probability p the observation will be less than or equal to $Q(p)$.

Definition 1.11. *The sample mean is the average of the all the random variable.*

$$\bar{X} = (X_1 + \cdots + X_n)/n$$

Definition 1.12. *The expected value or mean of X, denoted $E[X]$ or μ is given by,*

$$E[X] = \mu = \int x f(x) dx$$

where $f(x)$ is a density function.

Definition 1.13. *The variance of X, denoted by VarX,*

$$VarX = E[X^2] - E[X]^2$$

Definition 1.14. *The standard deviation assesses the level or variability in a distribution, denoted by σ.*

$$\sigma = \sqrt{\int (x - \mu)^2 f(x) dx}$$

where $f(x)$ is a density function.

Definition 1.15. *The sample standard deviation is denoted $\hat{\sigma}$,*

$$\hat{\sigma} = \sqrt{((X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2)/(n-1)}$$

Definition 1.16. *The variance is the square of the standard deviation denoted as σ^2.*

Definition 1.17. *A binomial experiment is a statistical experiment that consists of n repeated trails, with each trail having just two possible outcomes; success with probability p or failure with probability $q = (1 - p)$.*

Example 1.18. *Suppose a basketball players shoots a free throw n times and counting the number of made free throws. This is a binomial experiment with probability of success(Make) to be the player's free throw percentage, p.*

Definition 1.19. *Let X be a binomial random variable. The probability of exactly x successes is n trails is,*

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

where p is the probability of success and $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.

How are Order Statistics Used in Basketball Data?

In basketball, just as in any sport the goal is to win. In order to do so a team must create an advantage over the opposing team. This advantage can be gained long before the referee even tosses the ball up. Analyzing basketball data can create this advantage in many aspects such as coaching, player development and even in executive decision making. Order statistics are based on the ranking of observations, they are useful in non-parametric statistics.

From a coaching perspective, analytics can be used to determine the rotation of players throughout a game. The coaching staff must determine the amount of time a player should be on the court or which plays the team should run. Ranking specific on-court lineups can be used to allocate playing time accordingly. Also tracking the amount of times a play is successful throughout a game, can help aid in determining what sets to call through out the course of a game.

Order statistics can be used to rank players across various statistical categories such as points, assists and rebounds. The ranking of these statistics can use the average of each player over a season to determine how the stack up against their counterparts. This is important when it comes to end-of-season awards, and also can help increase fan involvement to see how good their favorite player is.

A lot can go into executive decision making, from a financial point of view there can be a lot at stake. Order statistics can make inferences on the projection of a player's value to a team. It can account for various factors, such efficiency position and even their age. Having this sort of information and being able to interpret is vital to the success of an organization. Since most team's having only a certain amount of money that it can spend on their player's, order statistic can be used to allocate funds efficiently among the roster.

The use of such parameters such a mean, standard deviation, median and quartiles can help drawn conclusions about the data. Using the median in relation to basketball data maybe be a better estimator than mean because due the occurrences of outline that will skew the data in a particular direction. Player's have an aberration of game or in a bit of a slump shouldn't be the determining factor in their value.

The construction of confidence intervals can be used as tool to make predictions of the outcome of a certain event. They are mainly used for relating sample data to population data. In basketball the entire game can be broken down into some statistical measurements and this observations can be used to determine the outcome of a game before it starts. This is helpful in determining betting odds on games as well as setting expectations for a team or player.

1.1 Historical Review

A Princeton University professor named Samuel Stanley Wilks coined the term ”order statistics” at the institution in 1942. Earlier that century related functions order statistics such as the median, the mid range and the symmetrically trimmed means were introduced. In the early 1800s, Pierre-Simon Laplace was able to derive the distribution of the r^{th} order statistic in random samples. A condition was also found such that the median was a more efficient estimator than the mean around the same time.

1.2 Outline of Project

Ordered statistics can be very useful in analyzing basketball data in various ways such as ranking players, determining rotations, awards, salaries and other decision making processes. In chapter 2, order statistics and other related jargon will be defined formally, different types of distributions will also be discussed. Chapter 3 will pertain to more theoretical properties which include special cases of ordered statistics, and ordered statistics derived from various distributions. Chapter 4 will use basketball data to demonstrate an application of techniques introduced. In chapter 5 with the use of the computer program R a simulation will be run.

2 Basic Notations and Definitions

This chapter will formally define an order statistic, as well as other terms that will be mentioned throughout. A brief synopsis of distributions will be introduce to help flow into the following chapter.

2.1 Definitions

We will be begin with a formal definition of an order statistic.

Definition 2.1. *Let $X_1, X_2, ..., X_n$ be i.i.d random variables taken from a discrete or continuous distribution, The order statistic of*

$$X_{(1)}, X_{(2)}, ..., X_{(n)}.$$

The order sample is the strict inequality

$$X_{(1)} < X_{(2)} < ... < X_{(n)}$$

- *The first order statistic is the minimum of the sample:*

$$X_{(1)} = min(X_1, X_2, ..., X_n)$$

- *The n^{th} order statistic is the maximum of the sample:*

$$X_{(n)} = max(X_1, X_2, ..., X_n)$$

- *The r^{th} order statistic is the r^{th} smallest $X_1, X_2, ..., X_n$ in the sample:*

$$X_{(r)}$$

Example 2.2. *Consider the following basketball scores for the Lakehead men's basketball scores for the last 9 games of the 2018-2019 regular season:*

72, 94, 75, 71, 125, 75, 79, 85, 93

Determine the order of the scores. Find the the first and n^{th} order statistics.

- *Ordering of the scores*

$$x_{(1)} = 71, x_{(2)} = 72, x_{(3)} = 75, x_{(4)} = 75, x_{(5)} = 79, x_{(6)} = 85, x_{(7)} = 93, x_{(8)} = 94, x_{(9)} = 125$$

- *$x_{(1)}$ is 71, $x_{(n)} =$ is 125*

Definition 2.3. *The sample range is the difference between the maximum and the minimum values*

$$R_n = Range(X_1, ..., X_n) = X_{(n)} - X_{(1)}$$

- *The median is the middle value of a data in which one half the values fall below the median and one half will be above the median.*

$$M = \begin{cases} X_{((n+1)/2)} \text{ if } n \text{ is odd} \\ (X_{(n/2)} + X_{(n/2+1)})/2 \quad \text{if } n \text{ is even} \end{cases}$$

- *The first quartile, Q_1 is the median of the lower half of the data set.*
- *The third quartile, Q_3 , is the median of the upper half of the data set.*
- *The Interquartile Range is the difference between the first and third quartiles $(Q_3 - Q_1)$.*

Definition 2.4. *The (100p)th percentile of a sample percentile has approximately np sample observation less than it, and has $n(1-p)$ sample observations greater than it , for $0 < p < 1$.*

Example 2.5. *Consider the set of scores from the previous example . Find the range, first and third quantiles, the interquantile range and the median.*

- *Range is 54*
- *Q_1 is 75; Q_3 is 93*
- *IQR is 18*
- *Median is 79*

2.2 Distributions

We will now look at some distributions and theorems of order statistics as well as how they are derived. These distributions are the key to probabilistic analysis of basketball data.

Density of $X_{(r)} : f_r(x)$

We will begin with the most general case of the order statistics, r^{th} order statistic. Below is a theorem in relation to its probability distribution.

Theorem 2.6. *Let $X_{(1)}, ..., X_{(n)}$ denote the order statistics of a random sample, $X_1, ..., X_n$ from a continuous population with cdf $F(x)$ and pdf $f(x)$. Then the pdf of $X_{(r)}$ denoted as $f_{(r)}(x)$ has density function*

$$f_{(r)}(x) = \tfrac{n!}{(r-1)!(n-r)!}(F^{r-1}(x))(1 - F(x))^{n-r} f(x)$$

Probability distribution of the two extreme order statistics

Now we will examine the probability densities of both the minimum $X_{(1)}$ and maximum $X_{(n)}$ order statistics. Assume for all $X_1, X_2, ..., X_n$ are iid continuous random variables with pdf f_r and cdf F_r

Probability distribution of $X_{(1)} : F_{(1)}(x)$

The cumulative distribution of the first-order statistic denoted as $F_{(1)}$, it can be derived as follows:

$$\begin{aligned} F_{(1)}(x) &= Pr\{X_{(1)} \leq x\} \\ &= Pr[min(X_1, X_2, ..., X_n) \leq x] \\ &= 1 - Pr[min(X_1, X_2, ..., X_n) > x]] \\ &= 1 - (Pr(X_1 > x, X_2 > x, ..., X_n > x)) \\ &= 1 - (Pr(X_1 > x)Pr(X_2 > x) \cdots Pr(X_n > x)) \\ &= 1 - (1 - F(x))^n \end{aligned} .$$

Probability distribution of $X_{(n)} : F_{(n)}(x)$

The cumulative distribution of the n^{th}-order statistic denoted as $F_{(n)}$, it can be derived as follows:

$$\begin{aligned} F_{(n)}(x) &= Pr\{X_{(n)} \leq x\} \\ &= Pr(max(X_1, X_2, ..., X_n) \leq x) \\ &= Pr(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) \\ &= Pr(X_1 \leq x)Pr(X_2 \leq x) \cdots Pr(X_n \leq x) \\ &= [Pr(X_1 \leq x)]^n \\ &= F^n(x) \end{aligned} .$$

Probability distribution of $X_{(r)} : F_r(x)$

Now consider both cases we can use them to together to produce a general formula for the cumulative distribution of the r^{th}-order statistics denoted F_r

$$\begin{aligned} F_r(x) &= Pr\{X_{(r)} \leq x\} \\ &= Pr[\text{ at least } r \text{ of the } X_i \text{ are less than or equal to } x] \\ &= \textstyle\sum_{r=1}^{n} \binom{n}{r} (F(x))^r (1 - F(x))^{n-r} \\ &= \textstyle\int_{-\infty}^{\infty} f_r(x)dx \end{aligned} .$$

The Joint Distribution of $X_{(r)}$ and $X_{(s)}$: $f_{rs}(x,y)$

It is important to examine the joint behavior of two order statistics.

Theorem 2.7. *Let $X_{(1)}, ..., X_{(n)}$ be the order statistics of a random sample, $X_1, .., X_n$ from a continuous population with cdf $F_X(x)$ and pdf $f_x(x)$. Then the joint pdf of $X_{(r)}$ and $X_{(s)}$, $1 \leq r \leq s \leq \infty$ is*

$$f_{r,s}(x,y) = C_{rs}F^{r-1}(x)f(x)[F(y)-F(x)]^{s-r-1}f(y)[1-F(y)]^{n-s}$$

where,

$$C_{rs} = \tfrac{n!}{(r-1)!(s-r-1)!(n-s)!}$$

for $-\infty < x < y < \infty$.

Here is a lemma for the special case of the joint distribution of the two extreme order statistics.

Lemma 2.8. *The joint density of the minimum $X_{(1)}$ and maximum $X_{(n)}$ order statistics is*

$$f_{1,n}(x,y) = n(n-1)(F(y)-F(x))^{n-2}f(x)f(y) \text{ for } x < y$$

Joint distribution of $X_{(r)}$ and $X_{(s)}$

The joint cumulative distribution of two order statistics, $X_{(r)}$ and $X_{(s)}$ for $1 \leq r < s \leq n$ can be derived as follows

$$\begin{aligned} F_{(r,s)}(x,y) &= Pr[\text{at least r of } X's \leq x \text{ and at least } s \text{ of } X's \leq y] \\ &= \textstyle\sum_{j=s}^{n}\sum_{i=r}^{j} Pr[\text{ exactly } i \text{ of } X's \leq x \text{ and exactly } j \text{ of } X's \leq y] \\ &= \textstyle\sum_{j=s}^{n}\sum_{i=r} \tfrac{n!}{i!(j-i)!(n-j)!}F^i(x)[F(y)-F(x)]^{j-i}[1-F(y)]^{n-j} \end{aligned}$$

for $x < y$

$$\begin{aligned} F_{(r,s)}(x,y) &= Pr[\text{at least } r \text{ of } X's \leq x \text{ and at least } s \text{ of } X's \leq y] \\ &= Pr[\text{at least s of } X's \leq y] \qquad . \\ &= F_{(s)}(y) \end{aligned}$$

for $x \geq y$

Joint Probability Distribution of all order statistics

The probability distribution of all order statistics is the most useful, its density is given below.

Theorem 2.9. *For all order statistics from a sample size n*

$$f_{X_{(1)},...,X_{(n)}}(x_1,...,x_n) = n!f(y_1)\cdots f(y_n) for -\infty < y_1 \leq y_2 \leq \cdots < y_n < \infty$$

Distribution of the range

The range, $R_n = X_{(n)} - X_{(1)}$was previously mentioned in definition 2.3

Theorem 2.10. *Given Lemma 2.8, the range R_n has a density as follows*

$$f_{R_n}(r) = n(n-1)\int_{-\infty}^{\infty}(F(r+u) - F(u))^{n-2} f(u) f(r+u) du \text{ for } r \geq 0$$

Definition 2.11. *The expected value of the r^{th}-order statistic, $X_{(r)}$, denoted by $E(X_{(r)})$ is given by,*

$$E_r = \int_{-\infty}^{\infty} x f_r(x) dx$$

$$E_r = \frac{n!}{(r-1)!(n-r)!}\int_{-\infty}^{\infty} x F^{r-1}(x) f(x)[1 - F(x)]^{n-r} dx$$

Example 2.12. *A basketball league tracks the number of free throw attempts of every player their respective careers. Over the course of this time a player can attempt thousands of free throws. For simplicity, suppose that free throw attempts are modeled by a continuous uniform distribution on the interval (0,10)*

Consider the following, let's select 5 players at random.

1. *What is the probability that the minimum number of free throw attempts is between 2000-6000?*
2. *What is the expected value for the maximum amount of free throw attempts?*

Since free throw attempts are modeled after the continuous uniform distribution the cumulative distribution for the common variable X is given by:

$$P(x) = \frac{x}{10} \text{ where } 0 < x < 10$$

1. *We need to find the cumulative distribution for the first order statistic.*

$$\begin{aligned} F_{(1)}(x) &= P(X_{(1)} \leq x) \\ &= 1 - (1 - F(x))^5 \quad . \\ &= 1 - (1 - \tfrac{x}{10})^5 \end{aligned}$$

Now this function can be used to calculate the probability of the minimum falling between 2000 and 6000.

$$\begin{array}{ll} P(X_1 \textit{ is between 2 and 6}) & = P(2 < X_1 < 6) \\ & = P(X_1 \leq 6) - P(X_1 \leq 2) \\ & = [1-(1-\frac{6}{10})^5] - [1-(1-\frac{2}{10})^5] \\ & = (1-\frac{2}{10})^5 - (1-\frac{6}{10})^5 \\ & = 0.8^5 - 0.4^5 \\ & = 0.32 \end{array}.$$

The 32% probability that minimum number of free throw attempts being in the desired range.

2. *Finding the cumulative distribution of the n^{th}-order statistic.*

$$\begin{array}{ll} F_{X_5}(x) & = P(X_5 \leq x) \\ & = F(x)^5 \\ & = (\frac{x}{10})^5 \end{array}.$$

To find the density, differentiate with respect to the variable

$$\begin{array}{ll} f_{X_5}(x) & = 5(\frac{x}{10})^4 \frac{1}{10} \\ & = \frac{1}{2*10^4} x^4 \end{array}.$$

Using the formula for expected value

$$\begin{array}{ll} E(X_5) & = \int_0^{10} \frac{1}{2*10^4} x^4 dx \\ & = 8.33 \end{array}.$$

The expected maximum number of free throw attempts in 8300.

3 Theoretical Properties of Order Statistics

3.1 Probability Distribution of Ordered Statistics

Uniform Distribution

Will be examining uniform order statistics, which pertain to order statistics sampled from a uniform distribution. And we will start with a definition.

Definition 3.1. *The uniform Distribution of the unit interval (0,1). The distribution function $F(x)$ of the uniform distribution is then given by*

$$F(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x \geq 1 \end{cases}$$

Example 3.2. *Let $X_1, ..., X_n$ be identically and independent distributed uniformly on the interval (0,1), so $f_X(x) = 1$ for $x \,\exists(0,1)$ and $F_X(x) = x$ for $x \,\exists(0,1)$. Thus the probability distribution function for the r^{th} order statistic is*

$$f_{X_{(r)}}(x) = \tfrac{n!}{(r-1)!(n-r)!} x^{r-1}(1-x)^{n-r}$$

for $x \,\exists(0,1)$.
Hence, $X_{(r)}$ $Beta(r, n-r+1)$. From this we can deduce that

$$EX_{(r)} = \tfrac{r}{n+1}$$

and

$$VarX_{(r)} = \tfrac{r(n-r+1)}{(n+1)^2(n+2)}$$

Exponential Distribution

Order statistics sampled from an exponential distribution

Definition 3.3. *A random variable X has exponential distribution with rate parameter λ if it has probability density function give by*

$$P(X > x) = e^{\lambda x} \text{ for all } x \geq 0$$

Example 3.4. *Let $x_{1:n}, x_{2:n}, ..., x_{n:n}$ be the order statistics of a random sample of size n from the exponential distribution with the following probability density function,*

$$f(x) = \{e^{-x}, x > 0$$

Clearly

$$F(x) = 1 - e^{-x}$$

Thus

$$f_{r:n} = \tfrac{n!}{(r-1)!(n-r)!}(1-e^{-x})^{r-1}e^{-(n-r+1)} \text{ x>0}$$

3.2 Special Case

For the most part, we have been dealing with continuous random variables. In this section we will investigate discrete order statistics. If the distribution of a random variable is discrete then the order statistics of a random sample of size n arising from such a distribution are known as the discrete order statistics of a random sample.

Definition 3.5. *Let $x_1, x_2, ..., x_n$ are discrete random variables, then the ordering of these random variables denoted $x_{1:n}, x_{2:n}, ..., x_{n:n}$ are discrete order statistics.*

Theorem 3.6. *Let $X_1, ..., X_n$ be a random sample from a discrete distribution with probability mass function $p(x) = p_i$, with cumulative distribution function $F(x)$ where $x_1 < x_2 < \cdots$ are possible values of X in ascending order. Define*

$$\begin{aligned} F_0 &= 0 \\ F_1 &= p_1 \\ F_2 &= p_1 + p_2 \\ &\vdots \\ F_i &= p_1 + p_2 + \cdots + p_i \\ &\vdots \end{aligned} .$$

Let $X_(1), ..., X_{(n)}$ denote the order statistics from the sample. Then

$$P(X_{(r)} \leq x) = \textstyle\sum_{k=r}^{n} \binom{n}{k} F_i^k (1 - F_i)^{n-k}$$

and

$$P(X_{(r)} = x) = \textstyle\sum_{k=r}^{n} \binom{n}{k} [F_i^k (1 - F_i)^{n-k} - F_{i-1}^k (1 - F_{i-1})^{n-k}].$$

Proof. Fix i, and let Y be a random variable that counts the number of $X_1, ..., X_n$ that are less than or equal to x. Consider the following binomial experiment, for $\{X_r \leq x\}$ one outcome say "success" and $\{X_r > x\}$ to be the other outcome say "failure". Then Y is the number of success in n trails. Thus Y binomial(n, F_i).

The event $\{X_r \leq x\}$ is the same as $Y \geq r$; that is, at least r of the sample values are less than or equal to x. □

Order statistics for discrete parents

Definition 3.7. *A parent distribution for a given measurement gives the probability of obtaining a particular result from a single measure*

Definition 3.8. *A distribution free statistic is computed without the knowledge of the parameters of the distribution from which observations are drawn*

Example 3.9. *Suppose that X is the medium value that satisfies the equation $F(x) = 1/2$. Then the probability that lands between two order statistics $X_{(s)}$ and $X_{(r)}$ is*

$$\begin{array}{ll} Pr[X_{(r)} \leq x < X_{(s)}] & = \sum_{i=0}^{x} \sum_{j=x+1}^{\infty} Pr[X_{(r)} = i \wedge X_{(s)} = j] \\ & = \sum_{i=0}^{x} \sum_{j=x+1}^{\infty} Pr[F(i-1) \leq U_{(r)} < F(i) \wedge F(j-1) \leq U_{(s)} < F(j)] \\ & = \sum_{i=0}^{x} Pr[F(i-1) \leq U_{(r)} < F(i) \wedge U_{(s)} \geq F(x)] \\ & = Pr[U_{(r} < F(x) \wedge U_{(s)} \geq F(x)] \\ & = Pr[U_{(r)} < \frac{1}{2} \leq U_{(s)}] \end{array} .$$

As a result the probability has nothing to do with the distribution of the function F.

4 Applications to Basketball Data

This section will introduce and discuss one of various applications of order statistic. Confidence intervals can provide an estimate an unknown parameters such as the mean, median and quartiles. Each parameter will be discussed and applied to basketball data. This will determine how precise the estimate is and how confident we are that the result we found are indeed correct.

Confidence intervals are meant as a range of values that act as good estimates of the unknown variable can help provide some answers to the following questions:

- Who is the better player?
- Which starting lineup is ideal?
- What is the probability of winning a game?
- What is the correct play to run given time and score?

4.1 Confidence intervals

We will begin with a formal definition

Definition 4.1. *A confidence interval for a population parameter is an interval of plausible values for that parameter.*

A confidence interval consists of:

- A point estimate
- A margin of error
- Confidence level

The range of the confidence interval is defined as:

$$\text{Confidence Interval} = \text{Point estimate} \pm \text{Margin of Error}$$

We will define each parameter briefly.

Definition 4.2. *A point estimate of an a single number based on sample data that represents a possible value for that unknown population parameter.*

Definition 4.3. *The margin of error is the range of values above and below the point estimate.*

$$\textit{Margin of error} = \textit{Critical value* Standard deviation}$$

Definition 4.4. *The confidence level states the probability that the method will give a correct result*

Example 4.5. *Suppose a team has an average margin of victory of 15 points per game. We want to know the minimum and minimum number of wins we would expect for at a certain confidence level. Let the confidence level be 95% now consider that the team is projected to win between 50 and 72 games in an 82 game season. This means that if the season were played 100 times, 95 of them would fall within 50 and 72 wins.*

Confidence Intervals for Quantiles

Now we will investigate how well the order statistics are as estimators of quantiles of an underlying distribution.

Definition 4.6. *Quantiles are points that divide the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample into equal-sized and adjacent subgroups.*

Definition 4.7. *The quantile of a random variable X can be represented as*

$$Q(p) = sup\{x \exists \Re : F(x) \leq p\}$$

$Q(p)$ can be used as a point estimate for a confidence interval.

The probability of $Q(p)$ being in between two order statistics $X_{(r)}$ and $X_{(s)}$ for $1 \leq r < s \leq n$

$$Pr[X_{(r)} \leq Q(p) < X_{(s)}]$$

Is independent of the distribution of X. This allows us to construct a distribution-free confidence interval.

Corollary 4.8. *Confidence intervals with the confidence coefficient $\geq 1 - \alpha$*

$$Pr[X_{(r)} \leq Q(p) < X_{(s)}] = \pi(r, s, n, p) = \textstyle\sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}$$

For a given n and p, make $(s - r)$ as small as possible subject to $\pi(r, s, n, p) \geq 1 - \alpha$. The confidence coefficient is distributed binomial.

Confidence interval for any percentile

Next we will examine a specific type of quantile, the percentile. As from definition 2.4, percentiles are quantiles that divide a distribution into 100 equal parts

Definition 4.9. *Percentile denoted π_p, the probability that X_i is less than π_p*

$$p = P(X_i < \pi_p)$$

Definition 4.10. *The confidence coefficient is*

$$1-\alpha = P(X_{(r)} < \pi_p < X_{(s)}) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}$$

For large samples of size $n \geq 20$, an approximate confidence coefficient

$$Z = \frac{W-np}{\sqrt{np(1-p)}}$$

Once the sample is observed and then ordered, the interval $(X_{(r)}, X_{(s)})$ forms a $100(1-\alpha)\%$ confidence interval for the unknown population percentile π_p

Example 4.11. *The height of 26 of top center's for the NBA Draft are recorded in inches below*

80.1 81.1 81.4 81.5 81.7 82.2 82.6 83.0 84.2 85.3 85.5 85.5 85.7

86.3 86.7 86.7 86.8 86.9 87.5 87.6 87.6 88.1 88.6 88.9 89.2 89.4

Determine a confidence interval for the 75th percentile, and find the confidence coefficient.

Solution

First we must find the order statistic that falls on the 75 percentile

(0.75)(26+1) = 20.25

Take the average of the two closest order statistics the 20^{th} and the 21^{st}. Since both are 87.6 it can be used as a point estimate of $\pi_{0.75}$.
In order to find a confidence interval for $\pi_{0.75}$, let's use $X_{(16)}$ and $X_{(24)}$, with the interval being

$(X_{(16)}, X_{(24)})$ = *(86.7, 88.9)*

We can use a binomial distribution with n = 26 and p = 0.75 to determine an exact confidence coefficient:

$$\begin{aligned} P(X_{(16)} < m < X_{(24)}) &= P(16 \leq W \leq 23) \\ &= P(W \leq 23) - P(W \leq 16) \\ &= 0.9742 - 0.0401 \\ &= 0.9341 \end{aligned}$$

We are 93.4% confident that the 75th percentile of the draft prospects' height is between 86.7 and 88.9 inches.

In this case since $n > 20$, we can use the normal approximation , the mean and variance of the binomial distribution are:

$$\mu = np = 0.75(26) = 19.5$$
$$\sigma^2 = np(1-p) = 26(0.75)(1-0.75) = 4.875$$

respectively. Therefore, the approximate confidence coefficient for the interval $(X_{(16)}, X_{(24)})$ *is:*

$$\begin{aligned} P(X_{(16)} < m < X_{(24)}) &= P(16 \leq W \leq 23) \\ &= P\left(\tfrac{15.5-19.5}{\sqrt{4.875}} < Z < \tfrac{23.5-19.5}{\sqrt{4.875}}\right) \\ &= P(-1.81 \leq Z \leq 1.81) \\ &= 0.9649 - 0.0359 \\ &= 0.929 \end{aligned}$$

The normal approximation is 0.929 compared to the exact probability of 0.934, which is a fairly accurate comparison.

Confidence intervals for median

Median is a measure of location that might be considered an alternative to the sample mean.The sample median is less affected by extreme observation as in comparison to the sample mean. We will examine a confidence interval for this specific percentile, 0.5, the median. The median is defined previously from definition 2.3.

To achieve a $(1-\alpha)$-confidence interval for the median, the following steps can be made:

- Getting n random samples
- Calculate $d = \sqrt{n}\Phi^{-1}(1 - \frac{\alpha}{2})$
- Let

  $$r = \lfloor \tfrac{n+1}{2} - \tfrac{d}{2} \rfloor$$

  and

  $$s = \lfloor \tfrac{n+1}{2} + \tfrac{d}{2} \rfloor$$

- The median should lie in between $X_{(r)}$ and $X_{(s)}$ with $(1-\alpha)$

  $$\Pr[X_{(r)} \leq median < X_{(s)} \geq 1 - \alpha]$$

Example 4.12. *Suppose there is 100 basketball players in a league, each of their scoring averages are recorded and ranked accordingly. Find a 95% confidence interval for the median of points per game.*

- *To get 100 random samples*
- *Calculate* $d = \sqrt{100}\Phi^{-1}(1 - \frac{0.05}{2}) = 10 * 1.96 = 19.6$
- *Let*

$$r = \lfloor \frac{101}{2} - \frac{19.6}{2} \rfloor = 40$$

and

$$s = \lfloor \frac{101}{2} + \frac{19.6}{2} \rfloor = 60$$

- $Pr[X_{(40)} \leq median < X_{(60)} \geq 0.95]$

With 95% confidence the median will fall between the 40^{th} *and* 60^{th} *values.*

Confidence intervals for mean

The last type of confidence interval, we will discuss is for the sample mean. The mean can be defined as the average of the given data.

Definition 4.13. *The confidence level at least* $(1 - \alpha)$*, where*

$$Pr[\frac{X_1+\cdots+X_n}{n} - \varepsilon \leq m < \frac{X_1+\cdots+X_n}{n} + \varepsilon] \geq 1 - \alpha$$

m is the true mean

A confidence interval for the mean of a sample of size $n \geq 30$ can be constructed as:

$$\bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

Where z is the appropriate value from the Z table for the standard normal distribution.

A confidence interval for the mean of a sample of size $n < 30$ can be constructed as:

$$\bar{X} \pm t\frac{\sigma}{\sqrt{n}}$$

Where t is the appropriate value from the t table for degrees of freedom equal to $n - 1$.

5 Simulation

In this section we will introduce another one of various applications of order statistic. A simulation uses computer programming to generate random numbers to create a random process to determine a certain outcome.

We will begin with a formal definition.

Definition 5.1. *A simulation is an attempt to estimate the properties of a process by using a random variable to represent that process.*

A simulation can be used to identify where randomness enters the problem. A simulation can be run to predict the outcome of an event. For example, predicting the winner of a series of games where each game is random.

Here are some random values that can be drawn from a simulation:

- Discrete random variable that can take on n possible values.
- A continuous uniform random variable, meaning one that can take on all possible values between a minimum and a maximum.
- Normally distributed random variable

A basic outline of a simulation:

1. Define the outcomes of the experiment
2. Define the possible outcomes of the simulation that represent the real outcomes of the experiment.
3. Label each outcome with a specific digit or phrase
4. Define a success in the performance of a trail
5. Run the simulation repeatedly
6. Draw conclusion from the results

To make an estimate of the probability based on a simulation

$$Pr = \frac{\text{number of successes in } n \text{ trails}}{n}$$

A simulation can use an artificial process such as tossing a fair coin to represent the outcomes of a real process that provides information about the probability of events.

5.1 Simulating Free Throws

With the use of computer program R, , we will now run a simple simulation to demonstrate it's usefulness. Here is an example of basketball player

Example 5.2. *Suppose a basketball player shoots 50% from the free throw line. Each shot independent of one another. We will is use a simulation to determine how many makes the player will have if they shot 100 free throws.*

Example 5.3. *Here is a R output of the simulation when run 100 times*

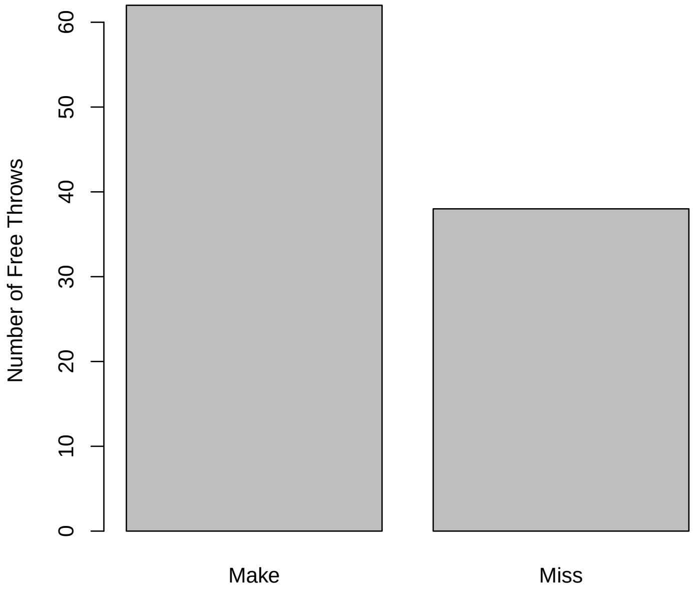
```
> outcomes = c("Make", "Miss")
> sim_basket = sample(outcomes, size=100, replace=T, prob=c(0.5, 0.55))
> sim_basket

  [1] "Miss" "Make" "Miss" "Miss" "Miss" "Make" "Miss" "Miss" "Make" "Make"
 [11] "Make" "Make" "Miss" "Make" "Miss" "Make" "Make" "Miss" "Make" "Make"
 [21] "Make" "Miss" "Make" "Miss" "Make" "Make" "Miss" "Miss" "Make" "Miss"
 [31] "Miss" "Make" "Make" "Make" "Miss" "Make" "Miss" "Miss" "Make" "Miss"
 [41] "Make" "Make" "Make" "Make" "Miss" "Miss" "Miss" "Miss" "Make" "Miss"
 [51] "Miss" "Miss" "Make" "Make" "Miss" "Make" "Miss" "Make" "Make" "Make"
 [61] "Make" "Make" "Make" "Miss" "Miss" "Make" "Make" "Make" "Make" "Make"
 [71] "Miss" "Make" "Make" "Make" "Make" "Make" "Make" "Make" "Make" "Make"
 [81] "Miss" "Miss" "Make" "Make" "Make" "Make" "Make" "Make" "Make" "Miss"
 [91] "Make" "Make" "Make" "Miss" "Make" "Miss" "Make" "Miss" "Miss" "Make"

> table(sim_basket)

sim_basket
Make Miss
  62   38

> sim_table = table(sim_basket)
> barplot(sim_table, ylab="Number of Free Throws")
>
>
```



As we can see from the bar graph and the summary the simulation has 45 makes and 55 misses. Based on this simulation the players free throw percentage is

$$Pr = \frac{45}{100} = 45\%$$

The more repetitions a simulation is run for the closer the result's will be to the true likelihood. The result of one trail has no effect on the next trail, thus showing independence.

Further Discussion

I will discuss two additional applications of order statistics, hypothesis testing and rank order statistics. I will discuss both briefly and their use in basketball data.

Decision makers can use statistical analysis to make different kinds of choices. Some choices are continuous, meaning that you can adjust a quantity over a range. Other choices are discrete, meaning that it is an either-or choice. An example of a continuous choice is the amount by team pay players for different teams and years of experience. During contract negotiations the price that you pay for a player depends on the team's statistical evaluation of the player's projection of both their career and value to franchise. An example of a discrete choice is a team willingness to cut or sign a specific player. This discrete choice can relate to the concept of hypothesis testing Hypothesis testing can be used to compares players or evaluate a player's performance of two different teams. The team can use the z-test compare the difference in means among players and scenarios.

In basketball, individuals or teams are given rankings, by their respective governing body. Teams are ranked among st their opposition in their respective leagues, as well as players individually among each other. Rankings are very important in decision making and can also be used to engage fans. Another mechanism is the rating percentage index, known as RPI, is a value generated used to rank teams based upon wins and losses and their respective schedule. This tool compares team's that play different competition.

Order statistics can provide a solid logical based for basketball analytics and its applications are broad of enough to be used in different aspects of the game.

Glossary of Basketball Terminology

General Team and Individual Statistics

GM,GP,GS: Games played, Games started

PTS Points

FGM,FGA,FG% Field goals made, attempted and percentage

FTM,FTA,FT% Free throws made, attempted and percentage

3FGM,3FGA,3FG% Three pointers made, attempted and percentage

REB,OREB,DREB Rebounds, offensive rebounds, defensive rebounds

AST Assists

STL Steals

BLK Blocks

TO Turnovers

PF Personal fouls

MIN Minutes

AST/TO Assist to turnover ratio

EFF Efficiency

*PG Per game averages

Bibliography

1. Order Statistics and Applications by Rosemary Smith
2. Applications of Order Statistics in Health Data by Bernard G. Greenberg and Ahmed E. Sarhan
3. Computational Order Statistics by Colin Rose and Murray D. Smith
4. Order Statistics by H.A. Davis
5. Basketball on Paper by Dean Oliver
6. Basic Theories on On Order Statistics by Po-Ning Chen