BASEBALL COMPETITIONS - ARE ENOUGH GAMES PLAYED?

BY

JEAN D. GIBBONS, INGRAM OLKIN, and MILTON SOBEL

TECHNICAL REPORT NO. 135
OCTOBER 1978

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MPS 75-09450

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
BASEBALL COMPETITIONS – ARE ENOUGH GAMES PLAYED?

by

Jean D. Gibbons, Ingram Olkin, and Milton Sobel

TECHNICAL REPORT NO. 135

OCTOBER 1978

PREPARED UNDER THE AUPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MPS 75-09450

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Baseball Competitions—Are Enough Games Played?

JEAN D. GIBBONS, INGRAM OLKIN, AND MILTON SOBEL*

This article investigates the number of games of baseball that should be played (1) in a World Series competition, and (2) in a pennant race competition within each league, in order to have a reasonable level of confidence that the best team wins the competition. The current number of games played is found to be highly inadequate for the World Series and only barely sufficient for the pennant race.

KEY WORDS: Baseball competitions; Sports competitions; World Series; Pennant race; Round robin tournament.

1. Introduction

Statisticians have shown a considerable interest in the game of baseball since it originated in Cooperstown, New York, in 1839. The reason is obvious—few activities have been so thoroughly, even painstakingly, documented numerically as has baseball. Some recent statistical studies are of particular interest. Mosteller (1951) made an early study of the season wins and losses of the major league teams in 1948. Mosteller (1952) compared many aspects of the statistics on the major leagues and, based on this analysis, raised the question of determining the probability that the better team actually wins the conventional seven-game World Series, competition. Lindsey (1961) studied many questions related to the progress and pattern of scoring during a game, including the distribution of the number of runs per half-inning, the distribution of total score by a team, and the length of extra-inning games. Lindsey (1963) and Hooke (1972) discussed and analyzed records of league games in 1959 and 1960 in order to make recommendations concerning certain strategies of play (such as bunting). Efron and Morris (1975) used the James-Stein estimator to predict individual batting averages for 18 players from data on the first 45 times at bat in 1970 (some of these results are also given in Efron and Morris (1977)). Groeneveld and Meeden (1975) investigated probability modeling to describe the number of games played in a best-of-seven game series in sports, including, of course, the World Series. Simulation studies have been carried out and various other distributional data collected.

In the present article, we consider a different type of problem concerning baseball. The question relates to the number of games that should be played in order to have a high confidence that the best team wins. We consider the number of games needed in (1) the World Series competition, and (2) the pennant competition. Cook (1966) considers these questions in a nontechnical book, but his approach differs from ours.

2. The World Series Competition

Throughout the modern era of baseball, there have been two major leagues, the American and the National. The teams within each league compete during the season for the League Championship. In 1903 these two league champion teams met for the first World Series to determine the World Champion as the first team to win five out of a maximum of nine games. In 1905 (there was no World Series in 1904), the procedure was changed to a best-of-seven game series and has continued to the present (with the exception of 1919–1921, when at most nine games were played). Hence, for practical purposes, we can say that the World Series has consisted of a maximum of seven games throughout most of baseball history.

We assume that the two league champion teams are not equal in ability; hence one team must be better than the other. But the better team does not necessarily win each game, or even the Series, because of chance fluctuations.

Let us be specific. In any single game of the World Series, the winning team is of course the one with the higher score. However, the team that wins may be the better team or the poorer team. Let \( p > .5 \) denote the probability of the better team winning, and \( q = 1 - p < .5 \) denote the probability of the poorer team winning. We assume that the games are independent and identical in probability structure; in particular, that \( p \) is constant for each of the games in any one World Series (we call this the independence model). (The usefulness of the independence model has been questioned by some persons because an excessive number of World Series competitions in history have run to a full seven games; the question has not been settled.)

Our problem is to determine the number \( n \) of games that are required in order to state that the probability is at least some specified value \( P^* \) that the better team wins the Series, whenever \( p \) satisfies the relationship

\[
\delta = p/(1 - p) \geq \delta^* \tag{2.1}
\]

for some specified \( \delta^* > 1 \) (since \( p > .5 \)). Since \( P^* \) reflects our level of confidence that the winner of the Series is the better team, appropriate values for \( P^* \) might be .90 or .95. Suitable specification of \( \delta^* \) should depend on the average size of \( p \) relative to \( q \), and hence on an average value of the ratio of the relative abilities of the two teams (against each other, against common adversaries, and in overall objective scores).
2.1 Determination of an Appropriate n as a Constant Total Number of Games

Suppose for the present that World Series play always consists of a fixed total of \( n \) games, which may be odd or even. The procedure \( R \) for this kind of play is to observe the number of games won and lost as follows:

<table>
<thead>
<tr>
<th>Team</th>
<th>Won</th>
<th>Lost</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( a )</td>
<td>( n - a )</td>
</tr>
<tr>
<td>B</td>
<td>( n - a )</td>
<td>( a )</td>
</tr>
</tbody>
</table>

If \( a > n/2 \), then team A is declared the winner; if \( a < n/2 \), then team B is declared the winner; if \( a = n/2 \), then we randomize.

Under this procedure where a fixed total of \( n \) independent games are played, the binomial distribution with \( n \) trials applies. The two possible outcomes which occur with constant probability \( p \) and \( 1 - p \), respectively, for each trial (or game) are (1) the team that wins is the better team, and (2) the team that wins is the poorer team. The \( n \) games or trials are regarded as a random sample of size \( n \) from this distribution. Table 1 gives the sample size \( n \) needed to satisfy the \((\delta^*, P^*)\) requirement in the problem of selecting the outcome with the largest probability for the binomial distribution. In our present context, selecting the outcome with the largest probability in \( n \) trials corresponds to selecting the team with the larger probability of winning; that is, the better team wins the Series of \( n \) games.

For example, if we specify \( \delta^* = 1.6 \) and \( P^* = .90 \), then Table 1 shows that \( n = 21 \) games are required; for \( \delta^* = 1.6 \) and \( P^* = .95 \), \( n = 49 \) games are required. This table shows that \( n \) increases for fixed \( P^* \) as \( \delta^* \) increases (and hence the minimum probability that the better team wins the Series increases), and \( n \) decreases for fixed \( P^* \) as \( \delta^* \) increases (and hence the teams are more discrepant in relative ability to win a single game).

Since \( \delta^* \) is the threshold value of \( p/(1 - p) \), an estimate of \( p \) would be helpful in determining a reasonable value to specify for \( \delta^* \). The league champions in an arbitrary year should be close to evenly matched and so \( p \) is probably between .55 and .65. Hence we might choose \( \delta^* \) somewhere between 1.4 and 1.6. For the conservative choice \( P^* = .90 \), Table 1 gives \( n = 31 \) if \( \delta^* = 1.4 \), and \( n = 31 \) if \( \delta^* = 1.6 \). The latter result implies that at the very least, the World Series should continue until one team wins 16 out of at most 31 games. This conclusion implies that a Series based on at most seven games cannot provide us with a very high probability of a correct selection, and hence we can have little confidence that the winner is indeed the better of the two teams.

---

1 Actually, the World Series is carried out sequentially, so that the first team to win \( n/2 \) games or more is declared the winner. This avoids the possibility of a tied series when \( n \) is even, and also means that all \( n \) games need not be played if one team wins \( n/2 \) games or more before the \( n \)th game is played. This latter situation is called curtailment, and is discussed in Section 2.2.

---

<table>
<thead>
<tr>
<th>( \delta^* )</th>
<th>.750</th>
<th>.900</th>
<th>.950</th>
<th>.975</th>
<th>.990</th>
<th>( p(\delta^*) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>65</td>
<td>199</td>
<td>327</td>
<td>463</td>
<td>653</td>
<td>.545</td>
</tr>
<tr>
<td>1.4</td>
<td>47</td>
<td>97</td>
<td>137</td>
<td>193</td>
<td>.615</td>
<td>.583</td>
</tr>
<tr>
<td>1.6</td>
<td>31</td>
<td>49</td>
<td>71</td>
<td>101</td>
<td>.667</td>
<td>.615</td>
</tr>
<tr>
<td>1.8</td>
<td>19</td>
<td>33</td>
<td>45</td>
<td>65</td>
<td>.667</td>
<td>.643</td>
</tr>
<tr>
<td>2.0</td>
<td>15</td>
<td>23</td>
<td>33</td>
<td>47</td>
<td>.667</td>
<td>.667</td>
</tr>
<tr>
<td>2.2</td>
<td>11</td>
<td>19</td>
<td>25</td>
<td>37</td>
<td>.667</td>
<td>.688</td>
</tr>
<tr>
<td>2.4</td>
<td>9</td>
<td>15</td>
<td>21</td>
<td>29</td>
<td>.667</td>
<td>.706</td>
</tr>
<tr>
<td>2.6</td>
<td>7</td>
<td>13</td>
<td>17</td>
<td>21</td>
<td>.667</td>
<td>.722</td>
</tr>
<tr>
<td>2.8</td>
<td>7</td>
<td>11</td>
<td>15</td>
<td>21</td>
<td>.667</td>
<td>.737</td>
</tr>
<tr>
<td>3.0</td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>19</td>
<td>.667</td>
<td>.750</td>
</tr>
</tbody>
</table>

*To find intermediate values in this table, the authors recommend use of linear interpolation on \( \Delta \), \( \ln(1 - P^*) \), and \( 1/n \); this kind of interpolation gives accurate results.

\( p(\delta^*) \) is the value of \( p \) that corresponds to the least favorable configuration with \( \delta = \delta^* \).

In order to evaluate the current procedure of seven games, we obtain the operating characteristic (OC) curve, defined as the locus of different \((\delta^*, P^*)\) values that yield \( n = 7 \) games. Some points on this curve are as follows (\( p(\delta^*) \) is the value of \( p \) that corresponds to the least favorable configuration with \( \delta = \delta^* \) (see the following Technical Notes)).

\[
P^* \quad .50 \quad .60 \quad .75 \quad .90 \quad .95 \quad .975 \quad .99
\]

\[
\delta^* \quad 1.00 \quad 1.20 \quad 1.64 \quad 2.59 \quad 3.44 \quad 4.44 \quad 6.03
\]

\[
p(\delta^*) \quad .500 \quad .545 \quad .621 \quad .721 \quad .775 \quad .816 \quad .858
\]

The curve is graphed in Figure A. The tabulation shows that if the two teams are close in ability so that \( \delta^* \) is close to 1, then we achieve a very small probability of a
correct selection with only seven games—indeed, the probability is close to that of flipping a coin. The probability of a correct selection becomes reasonable only if the teams are very highly discrepant in ability (\( \rho \) at least 0.7), which would not usually be the case in the World Series. Figure A also shows the OC curve for a Series of \( n = 50 \) games; note how much more rapidly \( P^* \) increases as a function of \( \delta^2 \) for this larger \( n \).

**Technical Notes:** The entries in Table 1 were calculated under the configuration of parameter values where equality holds in (2.1); i.e., \( \delta^2 = \rho(1 - \rho) \), or equivalently, \( P(\delta^2) = \delta^2(\delta^2 + 1) \). This is known technically as the least favorable configuration, since it gives the smallest value of \( n \) that is needed in order to state with confidence level at least \( P^* \) that the better team wins the Series for any configuration where \( \delta \geq \delta^* \). For further explanation, examples of the use of this idea of a least favorable configuration, and further references, see, e.g., Gibbons, Olkin, and Sobel (1977).

The theory for the procedure of this section was developed in Bechhofer, Elmaghraby, and Morse (1959).

### 2.2 Expected Number of Games Played for a Series of at most \( n \) Games with Curtailment

As we noted earlier, the World Series is an example of a curtailed procedure in the sense that although the maximum number of games is always seven, the champion team may be determined before all seven games are played. The procedure \( R' \) that allows curtailment is to play only until one team wins four games. Thus the minimum number of games played is four and the maximum is seven. If \( N \) denotes the (random) number of games required by the curtailed procedure, then clearly the expected number of games \( E(N|R') \) under the curtailed procedure is smaller than the fixed number of games \( n \) needed for the corresponding procedure \( R \) without curtailment. The expected percent saved (PS) by curtailment is

\[
PS_n = \{[n - E(N|R')]/n\} \cdot 100.
\]

As \( n \) gets large, this percent saved due to curtailment approaches the limit (PS) given by

\[
PS = 100[1 - 1/(2\rho)] = 100(\delta - 1)/26.
\]

Since \( \rho \geq 0.5 \), the percent saved ranges from 0 to 50 percent. Exact values of \( E(N|R') \) are given in Table 2 for \( n = 1(1)20 \).

The limiting values of the percent saved as a result of curtailment are

<table>
<thead>
<tr>
<th>( \delta )</th>
<th>( \rho )</th>
<th>( PS )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.0</td>
<td>0.500</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0</td>
<td>0.545</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0</td>
<td>0.583</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0</td>
<td>0.615</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0</td>
<td>0.643</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0</td>
<td>0.667</td>
</tr>
<tr>
<td>1.2</td>
<td>0.0</td>
<td>0.688</td>
</tr>
<tr>
<td>1.4</td>
<td>0.0</td>
<td>0.706</td>
</tr>
<tr>
<td>1.6</td>
<td>0.0</td>
<td>0.722</td>
</tr>
<tr>
<td>1.8</td>
<td>0.0</td>
<td>0.731</td>
</tr>
<tr>
<td>2.0</td>
<td>0.0</td>
<td>0.750</td>
</tr>
</tbody>
</table>

Figure B shows a graph of these limiting values as a function of \( \delta = \rho(1 - \rho) \). When \( \delta \) is small, the expected saving is small. As \( \delta \) increases, the saving becomes considerable. In the case of the World Series under the independence model, the expected number of games played is between 4 and 5.8125 depending on the value of \( \rho \) (the latter value is for \( \rho = 0.5 \) and the former is for \( \rho = 1 \)).

**Technical Notes:** For any odd \( n \), the expected sample size with curtailment is

\[
E(N|R') = \frac{n + 1}{2} \left[ \frac{1}{\rho} I_{a}\left(\frac{n + 3}{2}, \frac{n + 1}{2}\right) \right] + \frac{1}{q} I_{a}\left(\frac{n + 3}{2}, \frac{n + 1}{2}\right)
\]

where \( I_{a}(a,b) \) is the standard incomplete beta function.

![Figure B. Limit of Percent Saved (PS) as \( \delta \to \infty \) for \( k = 2 \) Players.](image-url)
The entries in Table 2 were calculated under the independence model with \( \delta^* = p/(1-p) \); that is, with the same least favorable configuration that was introduced in the technical notes at the end of Section 2.1. Thus the first column where \( \delta^* = 1 \) corresponds to the case where \( p = .5 \) and hence the teams are evenly matched in ability.

3. The Pennant Race—Round Robin Tournaments

The number of teams per league, the total number of games played per season, and the method of determining the league champion have all varied over the history of baseball. Beginning in the early 1900s, there were eight teams per league and each team played every one of the other seven teams exactly 22 times for a total of 154 games per season. In 1961 the American League expanded to ten teams, each playing every one of the other nine teams exactly 18 times for a total of 162 games per season; the National League followed suit in 1962. Throughout this period, the team with the largest number of wins in the season declared the league champion. (In the event of ties, the tied contenders had a playoff game.) In 1969 each league changed to 12 teams divided into two divisions; since then, the total number of games played per season has remained at 162, but each team no longer plays every other team in its league the same number of times, and there are divisional playoffs to determine the league champion. For our present purposes, we consider only the period prior to 1969 so that we may model the baseball season play in terms of a round robin tournament.

In the general case of a round robin tournament, each of \( k \) teams plays every other team some fixed number, \( r \), of times, and these \( r \) replications are assumed independent. Hence each of the \( k \) teams plays \( r(k-1) \) games, and the total number of games in the tournament is \( rk(k-1)/2 \). A table of frequencies, \( f_{ij} \), is made, where \( f_{ij} \) is the number of times team \( T_i \) beats team \( T_j \) for all \( i \neq j \) (\( f_{ii} \) is left blank). The total number of wins for team \( T_i \) is then

\[
f_i = \sum_{j \neq i} f_{ij}.
\]

The procedure is to order all \( k \) of these \( f_{ij} \) values from smallest to largest. The team with the largest total number of wins is declared the winner of the tournament.

A natural conclusion then is that the winning team is the best team. However, this is not necessarily the case because there are chance fluctuations in each game. The question then arises, what confidence can we have in the assertion that the winner of the tournament is really the best team? The confidence depends on \( r \), the number of replications, because if there really is a best team and \( r \) is large enough, then there is a high probability that the best team will achieve the highest score. First of all, we must define what is meant by the best team.

For simplicity of notation we define \( \pi_{ij} \) as a relative measure of how much better team \( T_i \) is than team \( T_j \) on the average in pairwise competition (or in a large number of matches between these two teams). Then we say that \( T_i \) is better than (or is preferred to) \( T_j \) if \( \pi_{ij} > .5 \); \( T_i \) and \( T_j \) are said to be equal in ability if \( \pi_{ij} = .5 \).

How can we use these \( \pi_{ij} \) values to define the overall best team? One possibility is to define a team \( T_m \) as best if \( \pi_{mj} > .5 \) for all \( j \neq m \). Then there can be at most one best team since \( \pi_{mj} + \pi_{jm} = 1 \). If we relax the definition of best to \( \pi_{mj} \geq .5 \) for all \( j \neq m \), then several teams could be tied for best. With either of these definitions, it may happen that no team is best. The following tabulation shows a hypothetical set of values of \( \pi_{ij} \) (configuration of \( \pi \) values) for three teams where no team is best by the above definitions:

<table>
<thead>
<tr>
<th></th>
<th>( i )</th>
<th>( j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \ldots )</td>
<td>.7</td>
</tr>
<tr>
<td>2</td>
<td>.3</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>3</td>
<td>.7</td>
<td>.3</td>
</tr>
</tbody>
</table>

Indeed, in this example we have intrinsity in that \( T_1 \) is better than \( T_2 \), \( T_2 \) is better than \( T_3 \), and \( T_3 \) is better than \( T_1 \).

As we have seen, an unrestricted configuration of \( \pi \) values may not define a best team. Furthermore, with some configurations of \( \pi \) values, such as that in Table 3, we should not be that much concerned about which team is designated as best. The critical problem is to be able to identify the best team when there really is a best team, or, to be more specific, to identify the best team with a specified high probability when that team is distinctly better than each of the other teams in the league.

We now specify the zone of preference for selecting a particular team as best; however, the reader should not regard this as a restriction of the total possible set of configurations of \( \pi \) values. In fact, our formulation is compatible with two different definitions of best but we use only one common requirement. The definition already given is that \( T_m \) is best if \( \pi_{mj} > .5 \) for all \( j \neq m \). The other definition says that \( T_m \) is best if \( \sum_{j \neq m} \pi_{mj} > \sum_{i \neq m} \pi_{ij} \) for all \( i \neq m \). (Although this second definition is compatible with our single requirement, the computation of the associated operating characteristic curve (defined below) would be quite different and is not considered in this article; thus our main interest is in the first definition.)

The requirement is that the probability of selecting the best team, say \( T_m \), is to be at least \( P^* \) when \( \pi_{mj} \geq \pi \) for all \( j \neq m \). Here \( \pi^* > .5 \) and \( P^* > 1/k \) are both preassigned. The following tabulation shows a set of \( \pi \) values which lie in the preference zone for selecting team 2 as best. Note that \( \pi^* \) and \( P^* \) are not part of
the definition of a best team:

<table>
<thead>
<tr>
<th></th>
<th>i</th>
<th>j</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>.3</td>
</tr>
<tr>
<td>2</td>
<td>.7</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.4</td>
<td>.2</td>
</tr>
</tbody>
</table>

With this definition of a best team among \( k \) teams, we can determine the value of \( r \) that is required so that the probability that the procedure described in the preceding paragraph leads to a correct assertion is at least some specified value \( P^* \) for a specified value \( \pi^* \).

Table 3 gives these values of \( r \) for selected values of \( \pi^* \) and \( P^* \) in the special case where \( k = 10 \). (Tables for other values of \( k \) are available in, e.g., Gibbons, Olkin, and Subel (1977).) The case \( k = 10 \) is included here because each baseball league had \( k = 10 \) teams in the 1960s when the play corresponded exactly to a round robin tournament. Table 3 shows that if \( P^* = .90 \) and \( \pi^* = .55 \), then 89 replications are required, whereas if \( \pi^* = .60 \) (for the same \( P^* \)), then 22 replications are required. During a regular baseball season in the 1960s there were \( r = 18 \) replications. The following tabulation gives some points on the operating characteristic curve; that is, pairs \((\pi^*, P^*)\) which yield \( r = 18 \). A graph of the corresponding OC curve is given in Figure C.

3. Values of \( r \) Needed to Satisfy the \((\pi^*, P^*)\) Requirement Under the LFS Configuration When \( k = 10^k \)

<table>
<thead>
<tr>
<th>( P^* )</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
<th>.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi^* )</td>
<td>55</td>
<td>60</td>
<td>65</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>13</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>89</td>
<td>22</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>117</td>
<td>29</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>180</td>
<td>45</td>
<td>20</td>
<td>11</td>
</tr>
</tbody>
</table>

These results show that when each team plays every other team for only 18 games, \( \pi^* \) must be fairly large in order to achieve a \( P^* \) of moderate size. Note that \( P^* = .90 \) corresponds to about \( \pi^* = .611 \) (found by the method of interpolation recommended in the footnote to Table 3), and this means that if the best team has probability at least .611 of defeating any other team in a single game, then it has probability at least .90 of winning the tournament. Figure C also shows the OC curve for \( r = 50 \) games in the tournament; in this case \( P^* = .90 \) corresponds to about \( \pi^* = .566 \).

To select an appropriate value for \( \pi^* \), we could look at the summary data for all league games played during some "typical" season. Table 4 shows some data for 1965. The relative frequencies for the "best" American League team, Minnesota (not quite the best since the frequency \( f_{15} = .389 < .5 \), has an overall percentage of .630, while the best National League team, Los Angeles, has an overall percentage of .599. The data in Table 4 suggest that \( \pi^* \) should be taken to be about .60 for each league. Then the probability that the best team wins in such round robin league play is about .90, which seems reasonable. Interpolation in Table 3 (using the method suggested in the footnote) actually yields .867.

Technical Notes: The entries in Table 3 were calculated under the configuration where \( \pi_{mj} = \pi^* \) and all other \( \pi \) values are equal to .5 (all \( \pi \) values are .5 except when \( T_{ij} \) is involved); i.e.,

\[
\pi_{mj} = \pi^* \text{ for all } j \neq m, \\
\pi_{ij} = .5 \text{ for all } i \neq m, j \neq m, i \neq j.
\]

We call this configuration of \( \pi \) values the least favorable slippage (LFS) configuration because all the \( \pi \) values are equal except when \( i = m \), in which case \( \pi_{ij} \) has "slipped" to the right by the amount \( \pi^* - .5 \).

![Figure C. Graph of the OC Curve for \( r = 18 \) and \( r = 50 \) for \( k = 1 \) Teams.](image-url)
4. Proportion of 18 Games Won by the Team in the Left Column When Playing the Team in the Top Row in 1965

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop.</td>
<td>.411</td>
<td>.516</td>
<td>.556</td>
<td>.569</td>
<td>.722</td>
<td>.722</td>
<td>.500</td>
<td>.823</td>
<td>.846</td>
</tr>
</tbody>
</table>

The main source for the theory relating to the procedure of this section is David (1963).

4. What Do the Season Percentages Mean?

If we could assume that the 20 teams in the two leagues were randomly divided between the leagues, then we might expect the overall season percentage of wins for the two pennant winners to be some kind of measure of their respective ability in a match between them for world champion. Of course, the teams are not randomly divided and the percentages do not add to one; therefore the question arises, what do the season percentages mean with regard to (1) the number of games played in the World Series, and (2) ability to win the World Series? In order to investigate these questions empirically for the period when the round robin format was followed, the historical data for the 61 relevant years between 1905 and 1968 were studied (1919–1921 are omitted because the criterion for World Champion was the best of nine games in each of these years).

Over the indicated 61 years of World Series play, the observed modal number of games is 7, and the sample average number of games is 5.7377. The Spearman rank correlation coefficient (with correction for ties) between the number of games played and (a) the ratio of the season percentages of the two pennant-winning teams (with the larger percentage in the numerator), is −.273, with one-tailed (asymptotic) P-value .017; (b) the absolute value of the difference of the natural logarithms of the season percentages of the two pennant-winning teams, is −.283, with one-tailed P-value .014; and (c) the absolute value of the difference of the season percentages of the two pennant-winning teams, is −.290, with one-tailed P-value .012. (These P-values are based on the normal approximation.) Each correlation is significant at any one-sided level greater than .017. Thus there does appear to be a relationship between the season percentages and the number of games played. The negative values of these correlations imply that the closer the teams are to being evenly matched as regards season percentages, the more games are needed to determine the World Series champion. It should be pointed out that the season percentages apply only to games won within each team’s respective league and hence are computed independently; they are not based on information about the ability of the two pennant-winning teams relative to each other; i.e., about our parameter $p$. It would be desirable to know the relation between $p/(1 - p)$ and the number of games played, but there seems to be no reasonable method for obtaining a good estimate of $p$.

In order to investigate the relation between season percentages and ability to win the Series, we first note that the season percentages were equal in only two of the years under consideration. In the remaining 59 years, the winner of the World Series had the higher season percentage only 34 times; i.e., only 57.6 percent of the time (which is not highly significant). Hence it appears that the season percentages are not very useful in predicting the winner of the World Series, nor are they very useful in estimating $p$ or making a reasonable specification for $\delta^8$.

5. Conclusions

One conclusion from the analyses presented in this article is that a maximum of seven games is highly inadequate to give much confidence that the team declared the world champion of baseball is really the better team. Furthermore, for the round robin tournament format (which was in effect prior to 1969), the total number 162 of games played per season in each league is barely sufficient to determine with a high level of confidence that the team declared the pennant winner is really the best team in the league when one team is better than all the others and the other teams are evenly matched. The first conclusion that seven games are highly inadequate applies equally today. Whether the second conclusion that 162 games are barely sufficient still holds after the change in 1969 (to 12 teams in two divisions that do not play each other the same fixed number of times) requires a more complex analysis.

[Received January 26, 1977, Revised February 13, 1978.]
References


