Baseball: A Statistical Analysis

Jeff Patterson
Two main parts

1. Measure the importance of offense and defense on winning in the game of baseball.

2. Determine the greatest game ever pitched.
Operational Definition of Offense and Defense

- **Offense**
  - Average on-base percentage of a team in a season
  - Total runs scored by a team in a season

- **Definitions**
  - On-base percentage: percent chance in which a batter reaches base safely
  - Runs: total runs scored by a team
Operational Definition of Offense and Defense

- **Defense**
  - Total *earned run average* of all pitching on a team in a season
  - Average *fielding percentage* for all fielding positions for a team in a season

- **Definitions**
  - *Earned run average*: average number of runs pitchers are accountable for through nine innings
  - *Fielding percentage*: percent chance a fielder has to make successful play
Goals

- Determine if a team’s offense or defense is significantly important to getting to a World Series
- Determine if a team’s offense or defense is significantly important to winning a World Series
- Determine if runs, on-base percentage, earned run average, and fielding percentage are of the same importance to winning
- If not, determine which are most and least important
Research

- Winners and losers of the World Series
  - Rank them according to their on-base percentage, runs, era, and fielding percentage
- 1905-1960
  - Two leagues, 8 teams each
Means of Ranks

- **Standard mean** = 4.5

- **Winners of the World Series**
  - On-base percentage = 2.348214
  - Runs = 1.830357
  - Earned run average = 1.785714
  - Fielding percentage = 2.589286
Means of Ranks

- Standard mean = 4.5

- Losers of the World Series
  - On-base percentage = 2.214286
  - Runs = 1.982143
  - Earned run average = 1.919643
  - Fielding percentage = 2.901786
For all t-tests...

- \( H_0: \mu = 4.5 \)
- \( H_1: \mu < 4.5 \)
- \( \alpha = 0.05 \)
- \( n = 56 \)
- \( cv = -1.96, 1.96 \)

\[
  t = \frac{\bar{x} - 4.5}{\frac{s}{\sqrt{n}}}
\]
Results

- Winners of World Series
  - On-base percentage: $t = -10.48$
  - Runs scored: $t = -17.21$
  - Earned run average: $t = -20.33$
  - Fielding percentage: $t = -8.53$

- Reject all null hypotheses.
Results

- Losers of World Series
  - On-base percentage: $t = -11.61$
  - Runs scored: $t = -15.48$
  - Earned run average: $t = -14.62$
  - Fielding percentage: $t = -6.64$

- Reject all null hypotheses.
What does this mean?

- The better a team’s on-base percentage, runs scored, earned run average, and fielding percentage, the better chance the team has of getting to the World Series.
Means

- **Winners**
  - Runs = 1.830357
  - OB% = 2.348214
  - ERA = 1.785714
  - F% = 2.589286

- **Losers**
  - Runs = 1.982143
  - OB% = 2.214286
  - ERA = 1.919643
  - F% = 2.901786
T-test for two means

- For all t-tests...
  - $H_0: \mu_{\text{winner}} = \mu_{\text{loser}}$
  - $H_1: \mu_{\text{winner}} \neq \mu_{\text{loser}}$
  - $\alpha = 0.05$
  - $n = 56$
  - $cv = -1.96, 1.96$

\[
t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}
\]
Results

- On-base percentage: \( t = 0.47 \)
- Runs: \( t = -0.68 \)
- Earned run average: \( t = -0.61 \)
- Fielding percentage: \( t = -0.95 \)

- Not sufficient evidence to reject null hypotheses.
What does this mean?

- After a team gets to the World Series, their offense and defense are not significantly important to winning the World Series.
Means of winners and losers

- Runs = 1.90625
- On-base percentage = 2.28125
- Earned run average = 1.85268
- Fielding percentage = 2.74554
Analysis of Variance

- $H_0$: $\mu_{\text{runs}} = \mu_{\text{ob\%}} = \mu_{\text{era}} = \mu_{\text{f\%}}$
- $H_1$: at least one is different
- $\alpha = 0.05$
- $n = 112$
- Numerator d.f. = $k - 1 = 3$
- Denominator d.f. = $k(n - 1) = 444$
- Critical value = 2.6049
Analysis of Variance

\[ F = \frac{nS_{\bar{x}}^2}{S_p^2} \]

- \( nS_{\bar{x}}^2 \) = variance between samples = 112 (.17036) = 19.08
- \( S_p^2 \) = variance within samples = 2.0112
- \( F = 9.49 \)

Reject null hypothesis.
What does this mean?

- At least one of the categories has a greater or lesser impact on helping a team win in the game of baseball than the others.
Studentized Range

\[ w = q_\alpha (t, v) \sqrt{\frac{s_{w}^2}{n}}, \]

if \(|\bar{x}_1 - \bar{x}_2| \geq w\), then the two means are different.

- \(n = 112\)
- \(s_{w}^2 = 2.0112\)
- \(t = 4, v = \text{degrees of freedom} = 444\)
- \(Q_\alpha(t,v) = 3.63\)
- \(w = .486435\)
Studentized Range

if $|\bar{x}_1 - \bar{x}_2| \geq w$, then the two means are different.

- $w = 0.486435$
- Earned run average = 1.85268
- Runs = 1.90625
- On-base percentage = 2.28125
- Fielding percentage = 2.74554
What does this mean?

- Earned run average and team runs scored are significantly more important to winning in the game of baseball than is fielding percentage. However, on-base percentage is not significantly more important than fielding percentage.
Conclusions

- A team’s offense and defense are significantly important to getting to a World Series.
- A team’s offense and defense is not significantly important to winning over their opponent in the World Series.
- A team’s earned run average and runs scored are significantly more important to winning than is fielding percentage, however, on-base percentage is not significantly more important than fielding percentage.
Determining the greatest game ever pitched
The perfect game

- **Definition**: A game in which a pitcher does not allow any batter of the opposing team to reach base.

- “Words alone cannot describe pitching's top "club" and most desired goal - the masterpiece of any career, the pinnacle of the pitching aspect and one of the most difficult feats to achieve in the entire game of baseball.” –unknown (speaking of a perfect game)

- Fourteen perfect games pitched in Major League baseball history.
The perfect game

1904...Cy Young
1908...Addie Joss
1922...Charlie Robertson
1956...Don Larson
1964...Jim Bunning
1965...Sandy Koufax
1968...Catfish Hunter
The perfect game

1981…Len Barker
1984…Mike Witt
1988…Tom Browning
1991…Dennis Martinez
1994…Kenny Rogers
1998…David Wells
1999…David Cone
Method

- To find the greatest game ever pitched, find the probability of each perfect game:

\[ (1 - \text{on-base percentage})^{\text{number of at bats}} \times \]
\[ (1 - \text{on-base percentage})^{\text{number of at bats}} \times \]
\[ (1 - \text{on-base percentage})^{\text{number of at bats}} \times \]
\[ \vdots \]
Results

14. Sandy Koufax: 0.0005423
13. Len Barker: 0.0001592
12. Cy Young: 0.0000933
11. Tom Browning: 0.0000678
10. Jim Bunning: 0.0000594
 9. Addie Joss: 0.0000531
 8. Mike Witt: 0.0000462
Results

7. David Wells: 0.0000398
6. Catfish Hunter: 0.0000298
5. Dennis Martinez: 0.0000245
4. Kenny Rogers: 0.0000227
3. David Cone: 0.0000187
2. Don Larson: 0.0000089
1. Charlie Robertson: 0.0000051
Is this the greatest game ever pitched?

- Statistically?
  - Yes.
- Reality?
  - Probably not.

- Don Larson pitched his perfect game in game 5 of the 1956 World Series.
Conclusion

- A team’s offense and defense are significantly important to getting to a World Series, but not to beating their opponent in the World Series.
- A team’s earned run average and runs scored are significantly more important to winning than is fielding percentage, however, on-base percentage is not significantly more important than fielding percentage.
- Although it is hard to determine the “greatest game ever pitched” because of situational effects, we have found the greatest game ever pitched statistically, which is Charlie Robertson in 1922.