SCIENTIFIC TEST & ANALYSIS TECHNIQUES
CENTER OF EXCELLENCE

Sizing Test Designs with Binary Versus Continuous Responses: Improving the Major League Baseball Playoffs

April 2023

Corey Thrush, Ctr
Dr. Lenny Truett, Ctr

To develop and apply independent, tailored Scientific Test & Analysis Techniques solutions to Test and Evaluation that deliver insight to inform better decisions.

About this Publication:
This work was conducted by the Scientific Test & Analysis Techniques Center of Excellence under contract FA8075-18-D-0002, Task FA8075-21-F-0074.

For more information:
Visit, www.AFIT.edu/STAT
Email, AFIT.ENS.STATCOE@us.af.mil
Call, 937-255-3636 x4736

Technical Reviewers:
Mr. Christopher Kershner, Ctr
Ms. Brittany Fischer, Ctr

Scientific Test & Analysis Techniques Center of Excellence
2950 Hobson Way
Wright-Patterson Air Force Base, Ohio

Version: 1, FY23

Abstract

Binary responses are problematic in Department of Defense testing because they require many test runs. Previous Best Practices suggest converting them to a continuous response to keep test size manageable and practical. This Best Practice will use simulations comparing win percentage and run differential from the Major League Baseball (MLB) 2021 regular season to compare test sizing for binary and continuous responses. The reduction in test size for similar probabilities of winning are dramatic after conversion. General discussion of how to change a binary to continuous is explained with simpler-to-more-complex examples not related to the MLB. Then details are provided on why run differential was chosen among other continuous alternatives. Noise is often considered to be fixed with no control, but in the context of the MLB, we show potential benefits of reducing noise. Throughout the paper the structure of playoffs games is manipulated, but it is done to provide a solution to determine the best team. This highlights that different test objectives will require tailoring of the test strategy and create test designs to answer different questions.

Keywords: MLB, baseball, simulation, binary, continuous

Table of Contents

Abstract ..... **i**

Introduction **1**

Use of a Binary Response in a Seven-Game Series–Large Difference to Detect **1**

What if There Were No Constraints on the Number of Games in a Series? *3*

Use of a Binary Response in a Seven-Game Series–Small Difference to Detect **4**

What if There Were No Constraints on the Number of Games in a Series? *5*

Converting Binary to Continuous–Discussion **7**

Converting Binary to Continuous for MLB Teams **7**

Run Differential–Braves vs Giants *8*

Run Differential–Braves vs Astros *8*

Smaller Variance–Continuous **9**

Different Types of Tests for a Different Question **10**

Conclusion **10**

References **11**

Introduction

Binary responses, like pass-fail or hit-miss, are often considered first in Department of Defense (DOD) testing because they are intuitive. However, binary responses are often not suitable in the DOD due to typical resource constraints such as budget, time, etc. However, continuous responses contain more information than binary responses and should be utilized whenever possible—even if they are not intuitively obvious. This is because when a response is continuous, the test size can be drastically reduced, and you can get more insight from the analysis. To demonstrate this efficiency, this Best Practice will utilize the 2021 Major League Baseball (MLB) playoffs and World Series as practical examples to compare the use of binary and continuous responses when sizing test designs. First, we will utilize a binary response, such as a large and small difference in win percentage, to help infer which team is better. Next, we will discuss converting a binary response (win-loss) to a continuous response (run differential). Then, the original example will be re-analyzed using a continuous response. Finally, we will demonstrate how reducing noise can affect test size.

Use of a Binary Response in a Seven-Game Series–Large Difference to Detect

The 2021 MLB playoffs were exciting, but were they a good enough test to decide who the best team was? We can begin to evaluate this test by inspecting the difference between each team's win percentage for the 2021 playoff teams. All data used in this Best Practice was collected and modified from Baseball Reference. Starting with a binary response such as win percentage is an intuitive way to start because the potential to detect a large difference should make it easier to demonstrate who the better team is. The best and worst team to qualify for the 2021 MLB Playoffs were the San Francisco Giants (SFG) and Atlanta Braves (ATL). The Giants' won 107 of their 162 games and the Braves won 88 of their 161 games. The 19-win difference, a rather large difference to detect, would imply that the Giants were the better team. We then took their winning percentages, 66% and 54.7%, and used them to simulate a seven-game series. In this hypothesis test, our null hypothesis was:

"H_o: The Giants have an equal or smaller win percentage than the Braves."

And our alternative hypothesis was:

"H_a: The Giants have a larger win percentage than the Braves."

The results of this simulated seven-game series (Figure 1) showed San Francisco winning the series 60.1% of the time. This outcome was surprisingly low considering the Giants were tied for the 14th most wins in MLB history, while the Braves only had seven wins above a 50% record that season. Based off their initial win percentages, we had inferred that the more historically great team, the Giants, would do well in the post-season; however, this simulation demonstrated that our assumption, or expectation, was wrong. The Giants only had a *slightly* better chance of advancing from the series than the Braves. This demonstrates how our binary responses of win-loss were misleading because they did not contain enough information for such a small test.

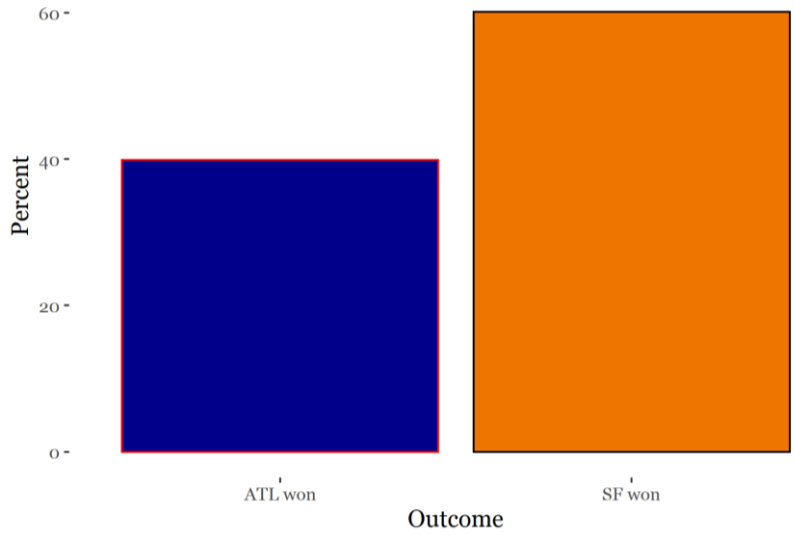

Figure 1

Outcomes of a Simulated Series between the San Francisco Giants and Atlanta Braves

Since we constructed a hypothesis test that utilized binary responses it is customary for us to need to find overwhelming evidence to conclude our alternative hypothesis (i.e., that the Braves have a higher win percentage). Originally, we had assumed that the Giants having a higher win percentage, based on their regular season win percentages, would result in the Giants winning our simulated series. However, if the Braves had won our simulated series, our test conclusion would be that they are better than the Giants. This would result in a statistical error, which is a Type 2 error or beta (β), where we incorrectly conclude the null when the alternative is true. We previously found beta to be about 0.4, which is not an acceptable amount of risk in any domain (or setting). Typically, a minimum power of 80% is considered sufficient which corresponds to a beta of 0.20. And, unfortunately, in this case we have a beta two times larger than the maximum allowed.

Table 1

Table of Error Types for SFG v ATL playoff series

		Truth	
		$SF \leq ATL$	$SF > ATL$
Result of Series (Conclusion)	$SF \leq ATL$	Correct	β
	$SF > ATL$	α	Correct

**For this best practice, we are only concerned with the second column.*

Let's further examine our simulated series results by looking at the probability distribution of the outcomes. Figure 2 plots the probabilities of each team winning the series in four, five, six, or seven games. On the x-axis of Figure 2 SF – 4 means that San Francisco won the series in four games, SF – 5 means San Francisco won in five games, etc.

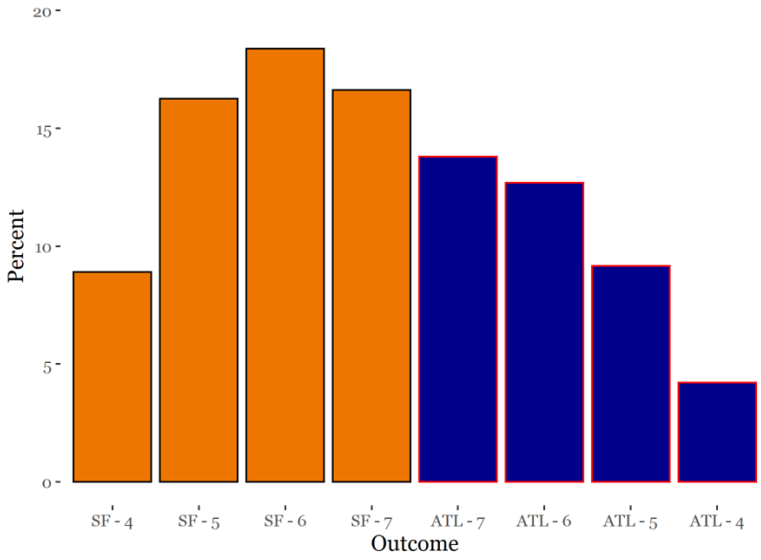

Figure 2

Distribution of Simulated Outcomes in a Seven-Game Series Between the Atlanta Braves and San Francisco Giants

An interesting result is that about 61.48% of the simulations have the series progressing to six or seven games. This result deviates from our initial hypothesis, which was based off the Giants' more favorable win percentage, that the Giants would win in four or five games. Instead, our simulated results show that the Giants would likely win in six games, 18.38% of the time, or in five or seven games.

Consequently, the simulation showed that the Braves were most likely to be victorious in seven games with a probability of 13.79%, and almost equally as likely to win in six games. Atlanta even had a 4.21% chance of sweeping the Giants (winning the series in four total games, with the opponent not winning any game). Despite having a large difference in their initial win percentages, the simulation resulted in the Giants winning the series about 60% of the time—further proving that our binary responses were misleading and did not gather enough information for such a small test. Our results also revealed some negligible differences between the teams in the probability distribution of games played. In summary, with the negligible differences in the probabilities of either team winning, we should not be surprised by any result.

What if There Were No Constraints on the Number of Games in a Series?

Let's assume a statistician within the Giants organization did the same simulation and pointed out the 40% chance of squandering the series by playing a best of seven. The same person suggested that there needed to be more games in the series to give them 80%, 90% or a 95% chance of advancing to the next round. As a result, the executives within the Giants organization wanted to know how many additional games it would take to reduce their beta to 0.20, 0.10, and 0.05. The statistician simulated the series and returned with the answer of 81 games for a beta of 0.20, 185 games for a beta of 0.10, and 293 games for a beta of 0.05. As we can see it required almost half a season to almost two full seasons for San Francisco to accept a manageable amount of risk, even against the weakest opponent in the MLB playoffs. It is clear using win percentage, which is binary, that our response was not reasonable due to the large number of games.

Use of a Binary Response in a Seven-Game Series–Small Difference to Detect

Our first example included a 19-win difference, or a large difference to detect, between the Giants and the Braves. What if our difference to detect was even smaller? The 2021 World Series teams, the Atlanta Braves (ATL), and the Houston Astros (HOU), had a seven-win difference. The World Series is the final series in the playoffs where the victor is considered the MLB champions for the season. To set up a hypothesis test using the 2021 World Series as our test design, our null hypothesis would be:

"H_o: The Astros have an equal or smaller win percentage than the Braves."

And our alternative hypothesis would be:

"H_a: The Astros have a larger win percentage than the Braves."

Because of the previous simulation between the Giants and Braves with a larger win difference, we should expect this matchup with a smaller win difference to be even closer to 50/50 for a team to win. Using each team's regular season win percentage, the Astros were simulated to have won 53.7% of those series, which is shown in Figure 3. This result makes our previous example's result even more confusing because the Giants had an almost three times larger win difference than the Astros did, but the Astros chances of winning against the Braves were only six percentage points lower. Beta in this result, 0.463, is almost as large as the previous simulated match up. Compared to the preferred 0.20 for beta, we can see this test would not be sufficient. This reiterates how binary responses can be misleading as they do not contain enough information for smaller tests.

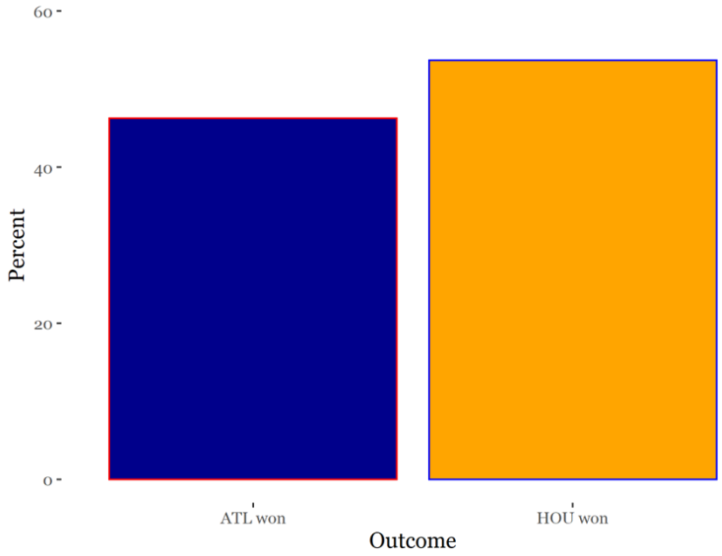

Figure 3

Outcomes of a Simulated Series between the Atlanta Braves and Houston Astros

Again, if the Braves had won our simulated series, our test conclusion would be that they are better than the Astros. But because we have assumed that the Astros have a higher win percentage is true then this error is also beta (Table 2). We found beta to be about 0.463, which is too large of a risk in any domain (or setting). Beta in this situation is over two times larger than typically preferred.

Table 2

Table of Error Types for HOU vs. ATL Playoff Series

		Truth	
		$HOU \leq ATL$	$HOU > ATL$
Result of Series (Conclusion)	$HOU \leq ATL$	Correct	β
	$HOU > ATL$	α	Correct

**For this best practice, we are only concerned with the second column.*

Based on the results of a 50/50 chance for either the Astros or Braves of winning, we can presume that our simulated test did not give us enough evidence to conclude which team is better. A seven-game difference out of 162 season games is miniscule (about 4%). As a result, this means we would require several more test runs to gain more evidence and minimize beta to support the claim that the Astros are better than the Braves.

Like our first example, Figure 4 plots the probabilities of HOU or ATL winning the series in four, five, six, or seven games. With the teams so closely matched, an initial expectation is for the series to end in six or seven games more often than the SFG and ATL series. This expectation was correct, but marginally. This simulated series ended in six or seven games 62.38% of the time compared to 61.48% in our previous example. As anticipated, the team with the higher win percentage (HOU) winning in six games is the most likely outcome at 16.75%, but HOU – 7, ATL – 7, and ATL – 6 outcomes trail slightly with about 15%.

The actual 2021 World Series resulted in the Braves winning the city's first championship in 26 years in six games. According to our simulation, there was a 14.68% chance of that outcome happening. In this situation, with the teams having a small difference of win percentage, there is a considerable amount of risk when concluding that a team is much better than the other by winning in four or five games, when really, they were equally matched.

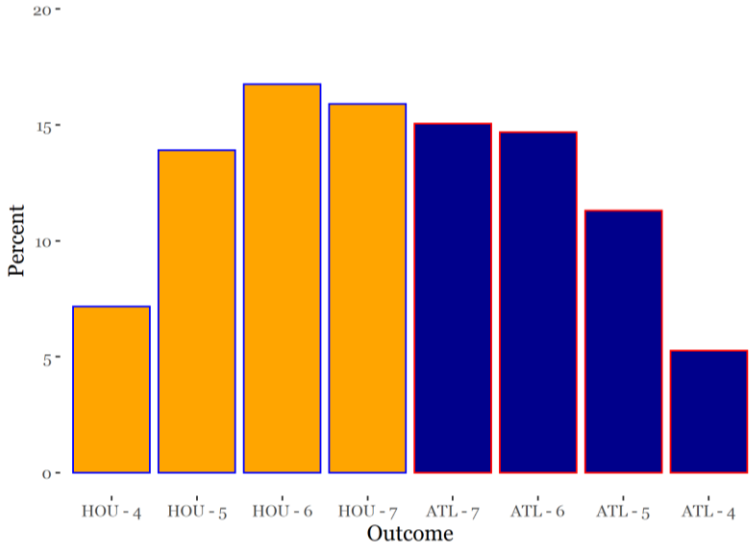

Figure 4

Distribution of Simulated Outcomes in a Seven-game Series Between the Atlanta Braves and Houston Astros

What if There Were No Constraints on the Number of Games in a Series?

With a smaller difference to detect, the Astros' statistician overheard the problem the Giants

were facing and presented the problem to the executives. The executives saw a beta of 0.463 in seven games, so they wondered how many games it would take to satisfy the 80%, 90%, and 95% requirements of winning a series against the Braves. The statistician found that the previous scenario for an 80% chance for the Giants to win it took about half of a regular season slate of games. But how many games would it take for a smaller difference to detect? 571 games were needed for the Astros to have a beta of 20%. Approximately 1,369 games were needed for a beta of 0.10 and over 3,000 games for a beta of 0.05. In the DOD, having the resources to do more than 50 test runs on a system is rare, let alone over 3,000 runs. With a large difference to detect, the number of games for each specified beta grows linearly, but still infeasibly. On the other hand, when the difference to detect is smaller it grows exponentially for each specified beta. Which would be impossible for testing in the DOD. These results are summarized in both Table 3 and Figure 5.

Table 3

Growth of Test Runs for Specified Beta Using Win Percentage

	p_1	p_2	β	Games
SF v ATL	0.66	0.547	39.9%	7
HOU v ATL	0.586	0.547	46.3%	7
SF v ATL	0.66	0.547	20%	81
HOU v ATL	0.586	0.547	20%	571
SF v ATL	0.66	0.547	10%	185
HOU v ATL	0.586	0.547	10%	1369
SF v ATL	0.66	0.547	5%	293
HOU v ATL	0.586	0.547	5%	3000+

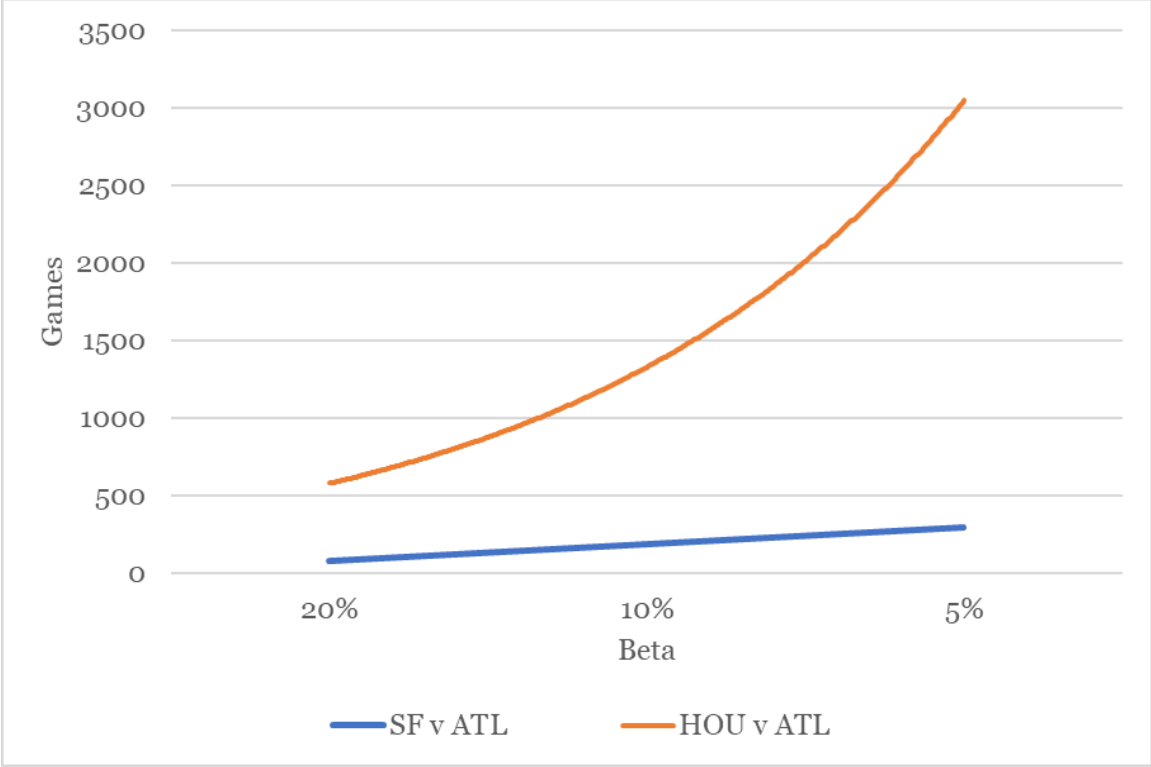

Figure 5

Growth of Number of Games for Specified Beta Using Win Percentage

Converting Binary to Continuous–Discussion

The MLB playoffs use a seven-game series to determine the better team. Is there a more effective method to consider, and how does this example translate to testing in the DOD? Converting a binary response into a continuous response typically captures more data information to support a hypothesis and reduce beta. Sometimes this process is straight forward, but sometimes this process can be challenging. A simple example is whether a weapon system hit or missed its target. You could measure distance from impact of the missile to the center of the target for all shots, or the miss distance. Starting with miss distance can be changed back into a proportion of runs that hit or miss for a target of any size, but only recording hit or miss cannot be changed into a miss distance. When analyzing miss distance, you may determine a system that never hits the target is consistently missing in one direction. Thus, by measuring the miss distance and direction, you can learn more about the system's performance.

Another example where a continuous response could provide more detail than a binary one is a smoke alarm. The sensor can detect whether a baked good was left in the oven too long to a more serious grease fire. Both situations had a threshold that was crossed, but only one required evacuation. The measured particles per million from the alarm could be analyzed to better sense the severity of the fire and provide the occupant with additional information on the level of threat.

Another well-known conversion between continuous and binary response is the body mass index (BMI): a formula using height and weight as inputs. BMI is used to classify someone as overweight, but many argue that it's too simplistic. BMI is generally correct at the extremes. But BMI is an attempt to develop a more representative continuous measure to classify overweight.

In your test, if you have many runs available or a large difference to detect, then a binary response may be reasonable. For example, in the medical device industry a test can be inexpensive, and although it can take a very large number of tests to prove the device has 99% accuracy, it is reasonable to record each test as pass-fail. However, this is rarely the case in DOD testing. Creative thought is often required to find an appropriate continuous response in your test. For help translating a binary into a continuous response for your test feel free to contact the STAT COE at AFIT.ENS.STATCOE@us.af.mil.

Converting Binary to Continuous for MLB Teams

As shown earlier, the differences to detect were not large enough to gather enough evidence without a large beta to conclude the Giants or Astros were better than the Braves in seven games. In this example, we want to change win percentage of an MLB team to an appropriate and representative continuous response. For baseball teams a combination of runs scored and runs allowed comes to mind. Separately, runs scored and runs allowed, do not properly characterize how good a team is because some teams are offensive juggernauts and others have great pitching and defense. However, the better teams often have a combination of good batting, pitching, and fielding. The simplest and most applicable metric is run differential, which is runs scored minus runs allowed. Run differential by itself is numeric, not continuous. However, run differential per game is continuous.

Using run differential per game as our response in a Simple Linear Regression (SLR) model,

win percentage is our predictor to estimate how strong of a linear relationship there is between the two. To evaluate the strength of the linear relationship a couple of metrics can be used. First, graphically the data points cluster around the line of best fit (Figure 6). Although, the points do not lie exactly on the line in the scatter plot, the vertical distance between the points and line is small. R^2, which measures the percentage of total variation estimated in the model, was 88%. Thus, we have a suitable substitute for a continuous variable from win percentage.

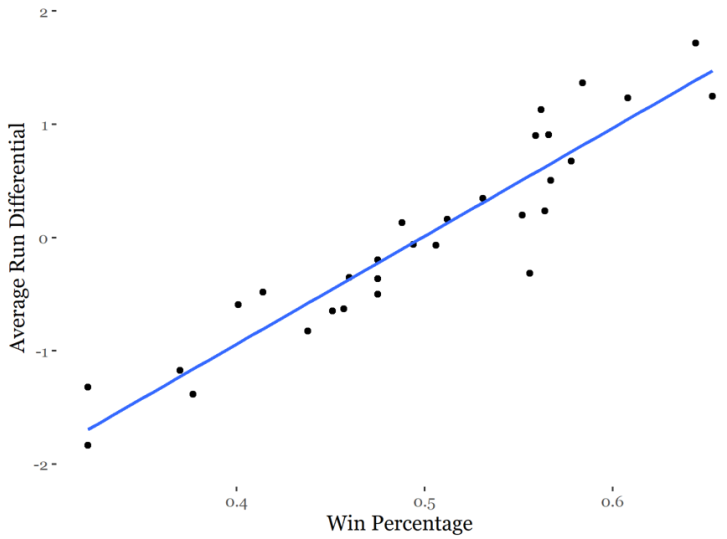

Figure 6

Simple Linear Regression on Run Differential per Game by Win Percentage

Run Differential–Braves vs Giants

After deciding that run differential is an appropriate measure for characterizing a team, a model was built to use average run differential. The formula for the line of best fit is $\widehat{Run\ Differential} = -4.7609 + 9.5480 \times Win\ Percentage$. Therefore, if there was a one percent increase in win percentage, there is an expected increase of 0.09548 in run differential. By plugging in win percentage for the Giants and Braves into the regression equation above, the expected run differential per game is, 1.54078 and 0.452308, respectively.

How the playoffs are currently constructed is an example of a bad test design. What if we created a new design with a smaller number of samples and compared the two teams—all with less risk of making an incorrect conclusion? A more representative test is having the teams play all seven games and sum the score of all the games. For example, in a three-game series the final scores were 3-2, 5-7, and 10-3; the first team wins 18-12.

If we played seven games and combined all the scores, based on expected run differentials, the Giants would win the series 88.1% of the time. Therefore, our beta decreased from 39.9% to 11.9%. There was an enormous gain of information in the data by developing a better test strategy and using a continuous response.

Run Differential–Braves vs Astros

When the teams are closer in run differential, the continuous response is even more advantageous. The Astros had an expected run differential of 0.834228 compared to the Braves 0.452308. In a seven-game series, the Astros win 66% of the time, and consequently the risk of losing the series is 34%. Which is a relatively large gain compared to earlier, 53.7% and 46.3%,

respectively. Another advantage of incorporating run differential (Figure 7), is that the number of test runs grows linearly for both situations. If the Astros wanted a 95% chance of advancing, a binary response required over 3,000 games, but the continuous response requires only 29. This is a 99% reduction in test size! With a smaller difference in expected run differential, the slope is steeper than that of the larger difference, but the total number of games required is much more reasonable.

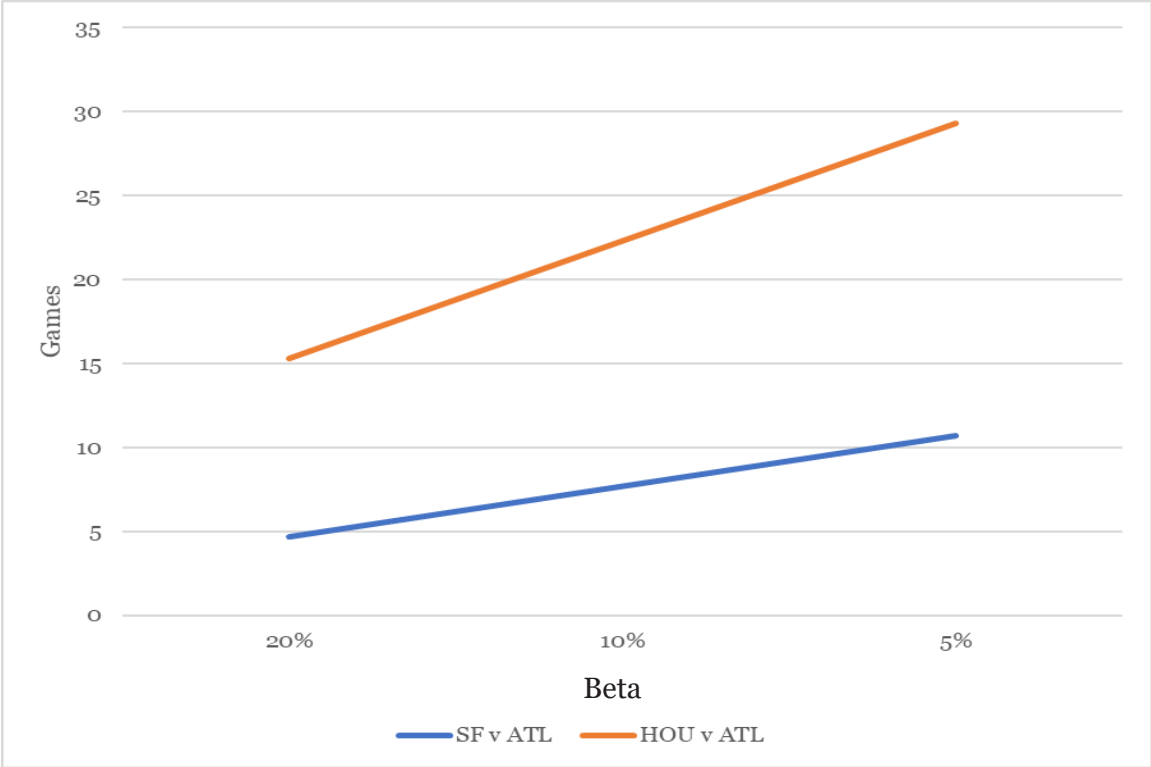

Figure 7
Growth of Test Runs for Specified Beta Using Run Differential

Smaller Variance–Continuous

There are additional things to consider in testing to become more efficient. When designing tests there are five metrics that test designer's trade-off:

1. number of test runs (sample size),
2. alpha (Type 1 error),
3. beta (Type 2 error),
4. delta (signal or difference), and
5. sigma (variance or noise).

In experimental data there are many sources of noise. This is problematic because it becomes more difficult to identify a significant effect as the noise is large relative to the difference to detect. This ratio is called the Signal-to-Noise ratio. For more details on the Signal-to-Noise ratio check out *Understanding the Signal-to-Noise Ratio in Design of Experiments* (Ramert, 2019).

In the 2021 regular season, the standard deviation of run differential was 4.503153. Since the distribution of run differential is approximately Normal the empirical rule is a rule of thumb describing the percentage of data that falls within a certain number of standard deviations. Which tells us for run differential in the MLB we expect to see 68% of games to result within 4.5

runs, 95% of games within 9 runs, and 99.7% of games within 13.5 runs. What if we decided to implement a mercy rule to reduce variation? Theoretically, the standard deviation could be reduced from 4.5 to 2.5. How different would the simulated 2021 World Series look with the reduction of standard deviation? The Astros are successful in 77% of those series compared to almost 66% earlier. Thus, because of reducing noise we have better estimates of those effects. Why did we investigate all these changes to the playoffs? It's because of the specific question that we wanted to answer, who is the best team in the MLB?

Different Types of Tests for a Different Question

In our simulation, we changed how the score was kept, but that test setting answered who is the best team. There are different test designs for each requirement. If you wanted to answer who would win a respective series, different information and different tests are needed.

The playoffs are situationally managed and played. Managers make decisions they think will increase their team's chances to win each game and the series. The best starting pitcher for a team does not pitch every game of the series, but neither does the worst. How often should a team use their ace starting pitcher? How much rest do they require? What about the bullpen when the starting pitcher pitches poorly? All these questions require different test designs.

In our simulation, we disregarded the situational aspects of baseball. A short list of outcomes could be that the Braves' best starting pitcher matches up against the Giants worst starting pitcher. The simulation's implicitly assumed team composition was held constant throughout the entire series. In this case, a tester has an interest in how large of an influence each of the starting and relief pitchers had on run differential or winning a game. To do this we would have modify the design by using a split-plot. In this case each pitcher would be assigned a whole plot, and each pitcher would add an additional error structure to the overall model.

Moreover, we as fans know where and when the game was played matters. For example, some stadiums are shorter which means more home runs are expected. Some stadiums are played at higher elevations, which reduces drag on the baseball. But we did not care about the effect of stadium on run differential, we want to know who is better. How should we proceed? The randomized block design would be a potential candidate for being able to remove noise that we as testers are not interested in. These were a few examples of different test designs that are often used.

Conclusion

In the DOD, binary responses are usually not practical; therefore, there should always be an attempt to convert to a continuous response. Although continuous responses are not perfect, such as our run differential from win percentage example, they still contain more information to help detect differences at different levels of beta, in far fewer samples. Additionally, understanding your test question will require different designs to account for different factors and assumptions of the test.

References

Adapted from Baseball Reference. (2021). 2021 Major League Baseball Standings. https://www.baseball-reference.com/leagues/majors/2021-standings.shtml

Ramert, A. (2019). Understanding the Signal to Noise Ratio in Design of Experiments. STAT COE, Department of Defense. Air Force Institute of Technology.