Tisha Harnlasiri
Dr. Nikki Meshkat
Math 123
12 June 2017

The Warriors and the Cavs: Exploring 3-Point Field Goal Percentages and Game Outcomes Using 2016-2017 NBA Statistics

A. Introduction

With the 2017 NBA Finals in full swing and as a supporter of the Golden State Warriors, I believed it would be interesting to analyze game statistics relevant to both the Warriors and the Cleveland Cavaliers, the team which they are facing in the finals. Using the first hypothesis test below, I analyze Stephen Curry's mean 3-point field goal percentage in home versus away games in the 2016-2017 Regular Season. Switching focus to the Cavs, my second hypothesis test determines if there is a dependence relationship between the number of rest days before a game and the game outcome in terms of wins and losses for the Cavs.

B. Test 1: Home Field Advantage and 3-Point Shooting

A two-time winner of the NBA's Most Valuable Player Award, Curry is considered one of the top 3-point shooters in the league with a record of thirteen 3-pointers made in a single NBA game [1, 2]. I am interested in analyzing whether Curry's mean 3-point percentage in home games differs from his mean 3-point percentage in away games.

(i) Data Collection

The data for my test was collected from Fox Sports' online database of individual player statistics for the Warriors during the 2016-2017 Regular Season.[3] The season consisted of 82 games with 41 games played at the Warriors' home court and the other 41 played away from home. The data I extracted from Fox Sports' website was the 3-point Field Goal Percentage (which I will call 3P% for short) for each of the games in the season. This statistic gives the ratio of the number of 3-point shots successfully made divided by the number of overall attempts. I separated the home-game data and the away-game data into two sample populations of "Home" and "Away." Because the games in the season are recent and are assumed to be independent of one another, the data comes from random samples that are fairly representative of Curry's recent performance.

Table 1: *Stephen Curry's 3-Point Field Goal Percentage in Home v. Away Games (2016-2017 Regular Season with the Warriors)* [3]

Home	Away	Home	Away
0.300	0.400	0.333	0.538
0.333	0.625	0.417	0.333
0.765	0.500	0.333	0.500
0.500	0.000	0.600	0.000
0.556	0.538	0.733	0.300
0.583	0.333	0.250	0.571
0.500	0.200	0.250	0.333
0.364	0.091	0.600	0.364
0.308	0.333	0.500	0.091
0.714	0.400	0.222	0.000
0.286	0.000	0.385	0.222
0.250	0.333	0.286	0.182
0.625	0.273	0.750	0.385
0.444	0.400	0.455	0.400
0.500	0.556	0.417	0.125
0.500	0.231	0.444	0.000
0.333	0.556	0.643	0.583
0.385	0.286	0.375	0.333
0.385	0.455	0.000	0.273
0.364	0.455	0.750	0.500
		0.357	0.615

Mean 3P%: Home (0.441317), Away (0.332032)

(ii) Hypotheses

Considering this data, I want to know: Is there sufficient statistical evidence to claim that Curry's mean 3-point percentage for home games is greater than his mean 3-point percentage for away games?

Since my test compares two sets of data, I use the null hypothesis that there is no difference between the population means for Curry's 3-point percentages. My alternative hypothesis is that the population mean for the 3-point percentages will be greater for Curry's home games. I chose this based on my belief that most players or teams will likely perform better if they are surrounded by a supportive, eager crowd. More formally:

Let Y_1 be the 3P% (3-point field goal percentage) for Curry in a home game during the 2016-2017 Regular Season, and Y_2 be the 3P% for Curry in an away game during the same season.

H_o: $\mu_1 = \mu_2$ (The mean 3P%s for home and away games are the same.)
H_a: $\mu_1 > \mu_2$ (The mean 3P%s for home and away games differ.)

(iii) Test

To test my hypothesis, I will perform a Z-test for the differences of the means. Using the Z-test is possible because the sample sizes of both my data sets are sufficiently large ($n_1 = n_2 = 41 > 30$). By the Central Limit Theorem, the sample means $\overline{Y_1}$ and $\overline{Y_2}$ have normal distributions, which allows me to use the Z standard random variable.

For my two samples, I know the following:

$$Home: n_1 = 41, \overline{y_1} = 0.441317, s_1^2 = 0.0287540$$
$$Away: n_2 = 41, \overline{y_2} = 0.332032, s_2^2 = 0.0347882$$

The sample means above are found by summing the 3-point percentages for the home and away games and then dividing by the sample sizes, respectively.

$$\overline{y_1} = \frac{0.3 + 0.333 + 0.765 + 0.5 + \ldots + 0.375 + 0 + 0.750 + 0.357}{41} = 0.441317$$

$$\overline{y_2} = \frac{0.4 + 0.625 + 0.5 + 0 + \cdots + 0.333\ 0.273 + 0.5 + 0.615}{41} = 0.0347882$$

I focus on the mean of the 3-point ratios rather than the ratio of the means because the former gives a better idea of Curry's 3-point shooting success rate for a single game; the latter would place more weight on the number of attempts made during the season.[4]

My sample standard deviations above are computed using the formula $S^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y}_i)^2$, which is an unbiased estimator of σ^2.

$$s_1^2 = \frac{1}{41-1}\ (\ (0.3 - 0.441317)^2 + (0.333 - 0.441317)^2 + (0.765 - 0.441317)^2 + \cdots + (0.750 - 0.441317)^2 + (0.357 - 0.441317)^2) = 0.0287540$$

$$s_2^2 = \frac{1}{41-1}\ ((0.4 - 0.332032)^2 + (0.625 - 0..332032)^2 + (0.5 - 0.332032)^2 + \cdots + (0.5 - 0.332032)^2 + (0.615 - 0.332032)^2) = 0.0347882$$

Assuming H$_0$: $\mu_1 - \mu_2 = 0$, the Z test statistic is:

$$z = \frac{(y_1 - y_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(0.441317 - 0.332032) - 0}{\sqrt{\frac{0.0287540}{41} + \frac{0.347882}{41}}} = \frac{0.109285}{0.0393676} = 2.776014$$

At $\alpha = 0.05, z_\alpha = 1.645$. Since $2.776014 > 1.645$, I can reject the null hypothesis and accept the alternative hypothesis that Curry's average 3-point percentage in his home games is higher than the 3-point percentage in his away-games.

The p-value is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected. My p-value for this test is $P(z > Z) = P(z > 2.776014) = 0.00275136$. Thus, for $\alpha > 0.00275136$, I can reject the null hypothesis that the mean 3-point percentage for the home games is the same as the mean 3-point percentage for the away games.

The result is *statistically significant*, which means that it is unlikely that we would have observed a value as extreme as the one obtained if the null hypothesis were true. In relation to the p-value, there is only a 0.27514% chance (which is very, very small) that z would equal 2.77601 if there was actually no difference between Curry's mean 3-point percentages in his home games and his away-games.

Although I can accept H$_a$, because my test only analyzes the 3P% statistic in a single season, it may not entirely be reflective of Curry's performance in past seasons in home and away games. There are also several confounding factors or "lurking variables" that may influence his 3-point performance in a game beyond a home court advantage. For example, travel fatigue, level of opponent competitiveness, and team injuries are factors which should be examined or held constant to get a more accurate statistical analysis of Curry's mean 3P% on different courts.

C. Test 2: Rest Days & Game Performance

Given that last year the Warriors had the greatest travel-distance between their games than any other team in the league and that the Cavs had the least, I was interested in whether the distance traveled or the amount of rest received before a game affects game outcome.[5] For my hypothesis test, I chose to focus on the relationship between the number of rest days and game wins or losses for the Cavs.

(i) Data Collection

I used Fox Sports to collect data on the Cavs' 2016-2017 regular season.[7] This data included information about the teams that the Cavs played against and if they won or loss. I combined this data with information I found on Tableau that provided figures for the number of rest days the Cavs had before each game. The combined data is shown below in Table 2a. Assuming that all the games are independent of one another, the data provides an accurate representation of the Cavs' playing season as all games are considered and were not biasedly selected.

Table 2a: Number of Rest Days v. Win/Loss **Note: The Cavs did not have rest days of 3 or 5.*

	Number of Rest Days					
Outcome	**0**	**1**	**2**	**4**	**6**	**7**
Win	11	30	11	1	1	1
Loss	7	15	5	0	0	0

(ii) Hypotheses:

I want to know: Does the above data show sufficient evidence to indicate a dependence relationship between the number of rest days that the Cavs had before a game and the game's outcome? I believe that players who are well-rested will likely perform better and therefore, will have a higher chance of winning their games. This is the reason why I chose the alternative hypothesis below, that the Cavs winning their games is dependent on the number of rest days they had before a game.

H_o: Game outcome is independent of the number of rest days before a game.
H_a: Game outcome is dependent on the number of rest days before a game.

(iii) Test

I will use the chi-square test for contingency tables to analyze my hypotheses. The chi-square test involves using Pearson's chi-square test statistic, X^2. To attain an adequate approximation of the χ^2 distribution for X^2, the test requires that no more than 20% of the *expected* cell counts of a contingency table be greater than or equal to 5 [8]. Due to this, I modified the data table above by combining the cells for Rest Days 2 to 7 with the cell for Rest Day 1 to attain larger cell counts. My modified table is shown below in *Table 2b.*

***Table 2b**: Modified Version of Table 2a*

	Number of Rest Days			
Outcome	**0**	**1**	**2+**	**Totals**
Win	11	30	14	55
Loss	7	15	5	27
Totals	18	45	19	82

I assume H_0 which implies the independence of rows and columns in the table. The expected cell counts $\widehat{E_{ij}} = \widehat{E(n_{ij})} = \frac{r_i c_j}{n}$ for each cell in the table are:

$$\widehat{E_{11}} = \frac{(55)(18)}{82} = 12.07317 \qquad \widehat{E_{21}} = \frac{(27)(18)}{82} = 5.926829$$

$$\widehat{E_{12}} = \frac{(55)(45)}{82} = 30.18293 \qquad \widehat{E_{22}} = \frac{(27)(45)}{82} = 14.81707$$

$$\widehat{E_{13}} = \frac{(55)(19)}{82} = 12.74390 \qquad \widehat{E_{23}} = \frac{(27)(19)}{82} = 6.256098$$

Then,

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{\left(n_i - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}}$$

$$= \frac{(11-12.07317)^2}{12.07317} + \frac{(30-30.18293)^2}{30.18293} + \frac{(14-12.74390)^2}{12.74390} + \frac{(7-5.926829)^2}{5.926829} + \frac{(14-14.81707)^2}{14.81707} + \frac{(5-6.256098)^2}{6.256098}$$

$$= 0.669085$$

The degrees of freedom are $df = (\#of\ table\ rows - 1) \times\ (\#of\ table\ columns - 1)$
$= (2-1)\ \times (3-1) = 2\ \times 1 = 2$

At the significance level $\alpha = 0.05$ and with df = 2, $\chi^2_{0.050} = 5.99147$.

Since 0.669085 < 5.99147, I cannot reject the null hypothesis. That is, it is possible that there is no dependence between the number of rest days that the Cavs had before a game and the outcome of that game when it was played.

The p-value of this test is $P(\chi^2 > 0.669085) = 0.715665$ which means that for $\alpha > 0.715665$, we can reject the null hypothesis and accept the alternative hypothesis that there is a dependence between the number of rest days and the win or loss of a game for the Warriors during the 2016-2017 regular season. This p-value is large! Because of this, the result X^2 is *not* statistically significant since there is a 71.57% chance that the result would have occurred if the null hypothesis was true.

Although X^2 is so small that one could argue that there is absolutely no dependence relationship between number of rest days and game outcome for the Cavs, I cannot draw an absolute conclusion that this is the case for all seasons for the Cavs. The same factors I listed in my first hypothesis test (team injury, competition, etc.), distance traveled between games or even the number of rest days that opponents had, can all affect game outcome. That being said, I am surprised that the X^2 statistic I obtained was not larger, since it seems to me that well-rested players would more often win their games.

Conclusion & Final Reflection

To conclude, my first hypothesis test allowed me to reject the null hypothesis that Curry's mean 3-point field-goal percentage in home games is no different than his mean 3-point field-goal percentage in away games. Thus, I was able to accept my alternative hypothesis that Curry's mean 3P% for home games is greater than the mean 3P% in away games. This result, however, might be confounded by external factors such as team travel, team injuries, and opponent competitiveness, among many other variables.

In my second hypothesis test, I was unable to reject the null hypothesis that for the Cavs, there is no dependence relationship between the number of rest days before a game and game outcome. Although this was a bit surprising to me, it makes sense considering that the number of rest days alone is only a single factor contributing to the win or loss of a game.

(Go Warriors!)

Sources

1. http://www.nba.com/players/stephen/curry/201939
2. http://www.nba.com/article/2016/11/08/curry-record-3-pointers
3. http://www.foxsports.com/nba/stephen-curry-player-game-log
4. http://www.wormbook.org/chapters/www_statisticalanalysis/statisticalanalysis.html#sec5-2
5. https://qz.com/698136/the-golden-state-warriors-are-the-nba-champions-of-traveling-too/
6. https://public.tableau.co/profile/positive.residual2774#!/vizhome/2016-17NBAScheduleTeamDashboard/2016-17NBAScheduleTeamDashboard
7. http://www.foxsports.com/nba/cleveland-cavaliers-team-game-log
   http://www.indiana.edu/~educy520/sec5982/week_12/chi_sq_summary011020.pdf
8. http://www.socr.ucla.edu/Applets.dir/Normal_T_Chi2_F_Tables.htm (p-value calculator used)