Austin Paczkowski

5/11/23

STATS516

The Three-Pointer and its Effects on Field Goal Efficiency

In an ordinary NBA game, there are only two types of baskets that a player can take: the two-pointer and the three-pointer. With the recent analytical movements in sports, including the NBA, the gradual increase in the number of three-pointers attempted per game has been a major talking point among statisticians. While implementing more threes was originally beneficial as the points per game have risen roughly 20 points leaguewide since 2000, there is now a growing argument among statisticians that the oversaturation of three-point attempts has decreased shot decision quality in both teams and players. With this in mind, one of the most important statistical NBA questions today is finding the proper balance between taking twos and threes. As such, this paper will attempt to answer whether players better at making threes contribute more to individual success than those who focus more on making twos by creating appropriate estimators for the samples, finding proper confidence intervals for such estimators, and comparing the two samples to converge on a hypothesis using 2022–23 NBA data from Basketball Reference. Before going into the data, we need to first discuss our data, what we are looking for, the assumptions we are making about the data, and the likelihood of the assumptions being held up in more complex analyses.

To start, we need to specify some parameters for the type of data we want to critique. As such, our data will require any given NBA player to have a minimum average of 1.5 three-pointers attempted per game, as many centers and scheme-specific role players focus on rebounds, attacking the paint, and short-range twos. In addition, we will only include players that have played at least one minute in at least 70% of regular season games (58 games), as we want to view players with at least a reasonable duration of playtime to analyze. With these two parameters in place, we have 201 players that meet our requirements. As such, we now need to find a viable statistic to compare the players to. While offensive rating (oRTG) and true shooting percentage (TS%) are both satisfactory statistics to use for comparison, both have too many external factors for the context of this paper. For one, offensive rating examines all factors of a player's game and how many points they would be expected to generate per 100 possessions, while true shooting percentage solely looks at free throws, two-pointers, and three-pointers. Ultimately, effective field goal percentage (eFG%) will be the best statistic to choose for our analysis as it averages the percentage of twos made with 1.5 times the percentage of three-pointers made, making it a good choice for our two vs. three-pointer argument. Finally, we need to address several assumptions about our data. First, we may assume our data is randomly chosen, as each player in our data has played a sufficient number of games during the season. Second, we can assume normality in our data and samples as player statistics are randomly generated by means of the variance of individual player season performance, meaning we can assume our data and subsets of our data will contain the same means and variances. Finally, we will assume player statistics are independent of one another. While this is false as players play off of one another, we will assume that player statistics are entirely individual-based for simplicity and to remain within the scope of this paper. With our parameters and objectives in place, we will begin by creating estimators for our samples.

To begin creating estimators for our population, we first need to decide what we want to compare and figure out what population parameters are of interest to us. As such, we will first find the average three-point attempts per game (3PA) within our 201 NBA players and separate them into two samples: those above the average attempts (sample 1) and those below (sample 2). For our main population, the mean of three-point attempts will be $\mu_{3PA} = \frac{1}{201}\sum_{i=1}^{201} 3PA(x_i)$, which comes out to an average of 4.43 three-point attempts per game among qualifying players. Comparing this average to the data, we have $n_1 = 87$ players in our first sample of interest (above average three-point attempts) and $n_2 = 114$ players in our second sample of interest (below average three-point attempts). With our population properly separated, we can create estimators for the sample mean of the eFG% for both samples. As such, we will have $\overline{Y_1} = \frac{1}{87}\sum_{i=1}^{87} eFG\%(x_i)$ = 0.5448 and $\overline{Y_2} = \frac{1}{114}\sum_{i=1}^{114} eFG\%(x_{i+87})$ = 0.5413. In addition, we will need the overall eFG% population mean to compare with later, which will be $\mu = \frac{1}{201}\sum_{i=1}^{201} eFG\%(x_i)$ = 0.5428. With our sample and population means generated, we can now find the standard deviation and variance of each. As such, $\sigma = \sqrt{\frac{\sum_{i=1}^{201}(eFG\%(x_i)-\mu)^2}{201}}$ = 0.0413, making the variance $\sigma^2 = 0.00171$ for our overall population. For our sample standard deviations, we will have $s_1 = \sqrt{\frac{\sum_{i=1}^{87}(eFG\%(x_i)-\overline{Y_1})^2}{87-1}}$ = 0.0396 and $s_2 = \sqrt{\frac{\sum_{i=1}^{114}(eFG\%(x_{i+87})-\overline{Y_2})^2}{114-1}}$ = 0.0427 with sample variances of s_1^2 = 0.00157 and s_2^2 = 0.00182 respectively. While discussing the sample means and variances, it is worth noting that both the sample means and sample variances are unbiased as $E(\overline{Y_1}) = \overline{Y_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} eFG(x_i)$, $E(\overline{Y_2}) = \overline{Y_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} eFG(x_{i+n_1})$, $s_1^2 = \frac{\sum_{i=1}^{n_1}(eFG\%(x_i)-\overline{Y_1})^2}{n_1-1}$, and $s_2^2 = \frac{\sum_{i=1}^{n_2}(eFG\%(x_{i+n_1})-\overline{Y_2})^2}{n_2-1}$, all of which are the commonly known unbiased estimators for a normal distribution, which our assumingly independent and identical samples follow. As a corollary, the MSE of these samples will equal s_1 and s_2 respectively. Finally, both of these estimators for our samples will be consistent, as n_1 and n_2 only remain in the denominator in the sample variances, meaning $\lim_{n_1\to\infty} s_1^2$ and $\lim_{n_2\to\infty} s_2^2$ will equal zero. After all of the above analysis, we now have unbiased and consistent estimators for the sample mean of the eFG% for each sample to evaluate with a confidence interval.

To begin, we will use our estimators to compare different values of α for our confidence interval. As such, we know that a two-tailed 1 – α confidence interval will have the range from the formula $\overline{Y} \pm z_{0.5\alpha}\frac{s}{\sqrt{n}}$. With this, we will give 95% confidence intervals (α = 0.05) for both of our samples and the ranges of the true eFG% sample means for both samples. Before plugging in our estimators into the confidence interval, we need to look for $z^{-1}(.025)$ as we are evaluating the two-tailed range for the mean. Using an external calculator, we find that the value of $z^{-1}(.025)$ = 1.960. Using this, we can now solve the 95% confidence interval range, which will be $\overline{Y_1} \pm z_{0.025}\frac{s_1}{\sqrt{n_1}} = 0.5448 \pm 1.960\frac{0.0396}{\sqrt{87}}$ = (0.5365, 0.5531) for the first sample confidence interval range (above average three-point attempts) and $\overline{Y_2} \pm z_{0.025}\frac{s}{\sqrt{n_2}} = 0.5413 \pm 1.960\frac{0.0427}{\sqrt{114}}$ = (0.5335, 0.5491) for the second sample confidence interval range (below average three-point attempts). In context to our problem, the significance of these confidence intervals is that there is a 95% chance for any given player's eFG% to lie between the confidence interval range for their given samples. Consequentially, this means our confidence interval has the exact coverage as our points are static and the sample means are known and unchanging. With this, we will now compare the

two samples and hypothesize if players who take more three-pointers have a different mean eFG% than those who take fewer threes.

To start, we will have our null hypothesis H_0 where $\overline{Y_1} - \overline{Y_2} = 0$ and our alternate hypothesis H_a where $\overline{Y_1} - \overline{Y_2} \neq 0$. In addition, we will test the null hypothesis on the z-statistic as our sample sizes are relatively large, meaning the t-distribution will closely represent the z-distribution as it is, corresponding to exact type I error control. Finally, we will make our rejection region for the null hypothesis the resulting p-value for our two-tailed z-statistic compared to $\alpha = 0.05$, which is the range $\{|z| > z_{0.5\alpha}\}$ or $\{|z| > 1.96\}$. As such, our z-statistic of interest for the two samples involved will come as a result of the formula $|z| = |\frac{\overline{Y_1} - \overline{Y_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}|$. Plugging in the corresponding values, we get $|z| = |\frac{0.5448 - 0.5413}{\sqrt{\frac{0.00157}{87} + \frac{0.00182}{114}}}| = 0.600$, which results in a one-tailed p-value of 0.274 or a two-tailed p-value of 0.548. Because $p = 0.548 > 0.05$, which corresponds with the absolute value of z being under the rejection region, we will fail to reject H_0. As a result, there is not enough evidence to indicate a difference between the true eFG% sample mean of NBA players who take more three-point attempts than average compared to those who take fewer based on the data collected.

Using estimators for the data and samples, the creation of confidence intervals for our estimators, and testing the two samples against one another, we concluded that taking more threes does not affect effective field goal percentage in the NBA. Even though a conclusion was reached, there were several assumptions we made that would not hold up in a more in-depth analysis. First, effective field goal percentage does not reward or punish how well an NBA player is able to draw fouls, which is a critical component in a player's contribution to the team. Second, the distribution of any scoring statistic among any sample of players in the NBA is not independent, as any scoring statistic is also heavily dependent on how well their teammates assist, score themselves, and any defensive metric that translates into easier offensive possessions such as steals and turnovers. Third, effective field goal percentage does not address the difficulty of shots a player takes or reward defense in any way. If a more complex analysis of the question were to be performed, two statistics that could be better for us to analyze are net rating (netRTG), which is the offensive rating minus the defensive rating of an NBA player, and plus-minus (+/-), which is the total impact on the score an NBA player has when on the court. While properly addressing the two-pointer vs. three-pointer debate is beyond the scope of this paper, the purpose of this paper is to ultimately discuss the complexity of the question, construct a simplistic opinion on the question, and explain why such a question is both important and difficult to answer.