

# Project Part 2

Prithvi Raj, Akhil Rachakonda, Shritan Goki

Link to Github Repository: <https://github.com/whojpj/DataToKnowledgeFinalProject---Prithvi-Raj-Akhil-Rachakonda-Shritan-Goki>

## Introduction

In Part 1 of this project, exploratory analysis revealed a clear upward trend in the proportion of total points scored from three-point field goals in the NBA from 2019 to 2022. The proportion increased from approximately 31.6% in 2019 to 33.5% in 2022, representing a 1.9 percentage point increase. However, exploratory analysis alone cannot determine whether this observed change represents a statistically significant shift in offensive strategy or could reasonably be attributed to natural variation.

### **Has the proportion of total points scored from three-point field goals increased in the NBA from 2019 to 2022?**

This research question was selected for hypothesis testing because it addresses a fundamental strategic shift in modern basketball and showed a clear directional trend in the exploratory analysis. The question is well-suited for statistical inference because it compares two distinct time periods (2019 vs. 2022) and involves a proportion that can be precisely measured from the available data. While the exploratory analysis suggested an increase, hypothesis testing will determine whether this change is statistically significant or could have occurred by chance.

The data used in this analysis comes from Basketball Reference (<https://www.basketball-reference.com/>), a comprehensive repository of NBA statistics. The dataset includes player-level per-game statistics for the 2018-19 season (labeled 2019) and the 2021-22 season (labeled 2022). These data represent all NBA players who appeared in games during these seasons, with official statistics recorded by the NBA including field goals made and attempted, three-pointers made and attempted, and points scored.

The data are appropriate for answering this research question because they provide complete league-wide coverage of all scoring events across both seasons. By aggregating player-level statistics, we can calculate the total number of points scored from three-pointers and the total number of points scored overall for each season, enabling a direct comparison of proportions. The data have been cleaned to remove players with missing values for critical variables (points, three-pointers) and players with zero field goal attempts, ensuring accurate calculations.

## Methods

### Test Selection

A **two-sample z-test for proportions** was selected to answer the research question. This test is appropriate because:

1. **Parameter tested:** The test compares two population proportions ( $p$  for 2019 and  $p$  for 2022), where each proportion represents the proportion of total points scored from three-point field goals in that season.
2. **Test type:** This is a two-sample test comparing proportions from two independent populations (the 2019 NBA season and the 2022 NBA season).
3. **Research question alignment:** The test directly addresses whether the proportion of points from three-pointers in 2022 is greater than in 2019, which matches the research question.

### Test Assumptions

The two-sample z-test for proportions requires the following assumptions, each of which must be verified:

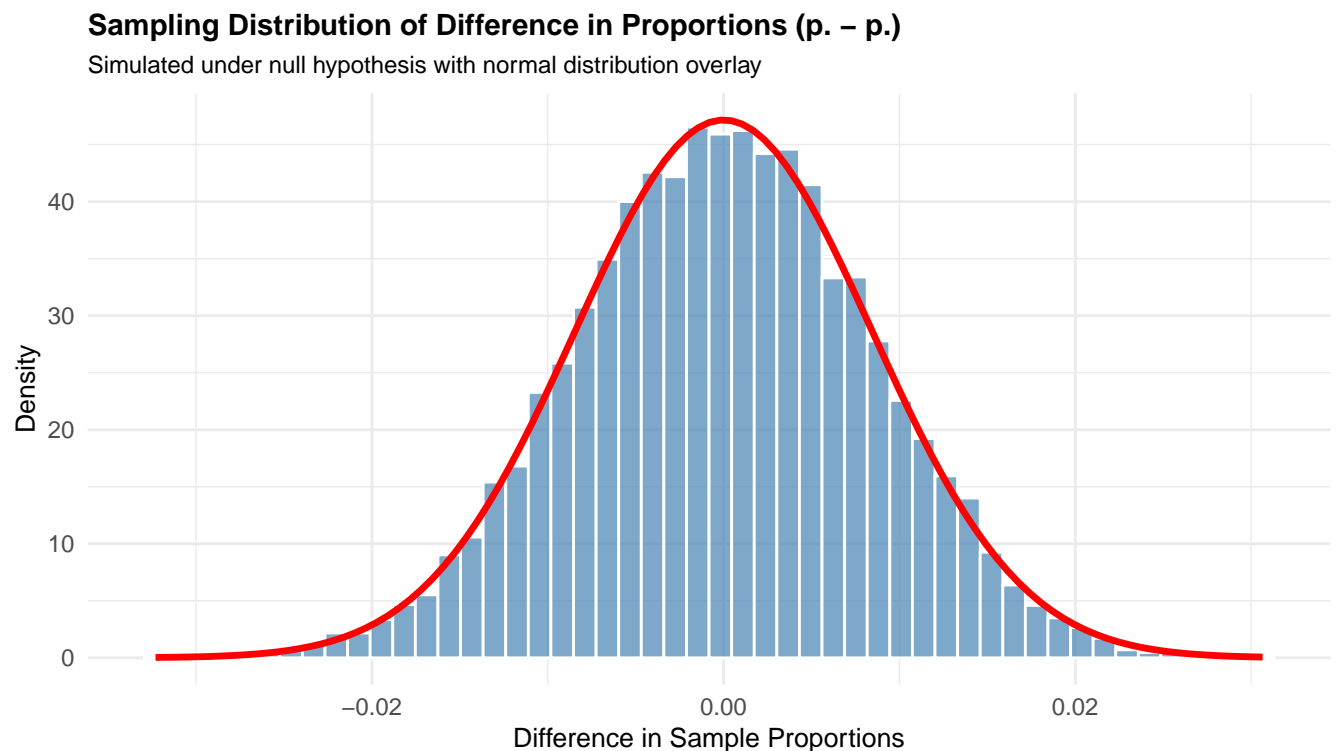
1. **Independent Samples** The 2019 and 2022 NBA seasons represent independent samples. These are distinct time periods with different rosters, different game schedules, and different competitive contexts. The outcomes of games and scoring patterns in 2019 do not influence those in 2022, satisfying the independence assumption.

**2. Large Sample Sizes (Normality Assumption)** The sampling distribution of the difference in sample proportions is approximately normal when both samples are sufficiently large. According to standard statistical theory (e.g., Agresti & Franklin, 2013; Moore et al., 2017), the sampling distribution of the difference in sample proportions is approximately normal when both samples satisfy  $np \geq 10$  and  $n(1-p) \geq 10$ , where  $n$  is the sample size and  $p$  is the proportion.

For the 2019 season: - Total points ( $n$ ): 5,936 - Total 3P points ( $x$ ): 1,878 - Sample proportion ( $p$ ): 0.3164 -  $np = 1,878 \geq 10$  -  $n(1-p) = 4,058 \geq 10$

For the 2022 season: - Total points ( $n$ ): 6,340 - Total 3P points ( $x$ ): 2,123 - Sample proportion ( $p$ ): 0.3348 -  $np = 2,123 \geq 10$  -  $n(1-p) = 4,218 \geq 10$

Both samples easily satisfy the large sample size requirement. To verify the normality assumption, a simulation of the sampling distribution of the difference in proportions was conducted, as shown in the histogram below:



The histogram shows that the simulated sampling distribution of the difference in proportions closely follows a normal distribution (red curve), confirming that the normality assumption is satisfied. The large sample sizes ( $np$ ,  $n(1-p)$ ,  $np$ , and  $n(1-p)$ ) all well exceed 10) ensure

that the Central Limit Theorem applies, making the normal approximation appropriate for this test.

**3. Random Sampling** While the data represent a population of all NBA players in these seasons, we treat the seasons as representative samples from the broader population of NBA seasons. The data collection process (official NBA statistics) ensures comprehensive and unbiased recording of all scoring events, satisfying the random sampling assumption for practical purposes.

## Statistical Hypotheses

The statistical hypotheses corresponding to the research question are:

- $H_0: p_{2019} = p_{2022}$  (The proportion of points from three-pointers in 2019 equals the proportion in 2022)
- $H_a: p_{2019} < p_{2022}$  (The proportion of points from three-pointers in 2019 is less than the proportion in 2022)

where  $p_{2019}$  represents the true proportion of total points scored from three-point field goals in the 2019 NBA season, and  $p_{2022}$  represents the true proportion in the 2022 NBA season.

This is a one-sided (upper-tailed) test because the research question specifically asks whether the proportion has *increased* from 2019 to 2022. The alternative hypothesis reflects the directional expectation from the exploratory analysis.

## Results

The two-sample z-test for proportions was conducted using R to compare the proportion of points from three-pointers between the 2019 and 2022 seasons:

```
# Conduct two-sample z-test for proportions
# H0: p1 = p2, Ha: p1 < p2 (one-sided upper-tailed test)

prop_test_result <- prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "less", correct = FALSE)
prop_test_result
```

```
##
## 2-sample test for equality of proportions without continuity correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 4.7297, df = 1, p-value = 0.01482
## alternative hypothesis: less
## 95 percent confidence interval:
## -1.0000000000 -0.004495399
## sample estimates:
##      prop 1      prop 2
## 0.3163693 0.3347792
```

The test output provides the following key components:

- **Sample proportions:** The estimated proportion for 2019 is 0.3164 (31.6%) and for 2022 is 0.3348 (33.5%).
- **Test statistic:** The z-statistic is 4.7297, calculated as the difference in sample proportions divided by the standard error of the difference.
- **P-value:** The p-value for this one-sided test is 0.0148227, representing the probability of observing a difference this large or larger if the true proportions were equal.
- **95% confidence interval:** The confidence interval for the difference  $p_1 - p_2$  is (-1, -0.004495).

**Statistical result:** At a significance level of  $\alpha = 0.05$ , we reject the null hypothesis ( $p = 0.0148227 < 0.05$ ). The p-value of 0.0148227 provides strong evidence against the null hypothesis that the proportions are equal. This indicates that the observed increase in the proportion of points from three-pointers from 2019 to 2022 is statistically significant and unlikely to have occurred by chance alone.

## Conclusions

The hypothesis test provides strong statistical evidence that the proportion of total points scored from three-point field goals has significantly increased in the NBA from 2019 to 2022. The test result ( $p = 0.0148227$ , which is less than 0.05) indicates that the observed

increase from 31.6% to 33.5% represents a real shift in offensive strategy rather than random variation. This finding confirms that NBA teams have meaningfully increased their reliance on three-point shooting as a central component of their offensive systems over this time period.

The statistical significance of this change has important implications for understanding the evolution of professional basketball. The increase reflects a league-wide strategic shift toward perimeter-oriented basketball, driven by the mathematical advantage of three-pointers (worth 50% more than two-pointers) and the influence of analytics in team decision-making. This trend represents a fundamental change in how the game is played at the highest level.

## Limitations

Several limitations should be considered when interpreting these results:

1. **Temporal scope:** The analysis compares only two seasons (2019 and 2022), providing a limited view of long-term trends. A more comprehensive analysis would examine multiple seasons or a longer time period to understand the full trajectory of this strategic shift.
2. **COVID-19 impact:** The 2020 and 2021 seasons were affected by the COVID-19 pandemic, which may have influenced player performance, game conditions, and statistical patterns. While this analysis compares 2019 (pre-pandemic) to 2022 (post-pandemic), the intervening years' disruptions may have affected the natural evolution of offensive strategies.
3. **Aggregation level:** The analysis aggregates player-level data to league-wide proportions, which may obscure important team-specific or player-specific variations in three-point strategy. Some teams may have increased their three-point emphasis more dramatically than others.
4. **Causal inference:** While the test establishes a statistically significant increase, it does not establish causality or identify the specific factors driving this change (e.g., rule changes, coaching strategies, player skill development, analytics adoption).
5. **Population vs. sample:** The data represent the complete population of NBA players in these seasons, but we treat them as samples for statistical inference. This approach is valid for generalizing to the broader population of NBA seasons, but the interpretation should acknowledge this framework.

These limitations could be addressed in future work by: (1) expanding the analysis to include more seasons to examine longer-term trends, (2) conducting team-level or player-level analyses to identify which teams or players are driving the change, (3) incorporating additional variables (e.g., rule changes, coaching changes) to explore causal mechanisms, and (4) using time series methods to model the evolution of three-point emphasis over time.

## Generalization

The conclusion from this hypothesis test generalizes beyond the specific 2019 and 2022 seasons used in the analysis. The statistically significant increase suggests that the trend toward greater three-point emphasis is a real phenomenon in modern NBA basketball, not merely a feature of these two particular seasons. This finding supports the broader narrative that the NBA has undergone a fundamental strategic transformation toward perimeter-oriented offense.

The results can be generalized to: - **Future NBA seasons:** The trend suggests that three-point emphasis will likely continue or increase in subsequent seasons, though the specific magnitude cannot be predicted from this analysis. - **Similar competitive contexts:** The findings may apply to other professional basketball leagues or competitive contexts where similar strategic considerations (point value of shots, defensive adaptations) are relevant. - **Strategic decision-making:** The statistical significance of the increase validates that teams' increased focus on three-point shooting represents a meaningful strategic shift rather than random variation, which has implications for coaching, player development, and roster construction.

However, generalization should be tempered by recognition that the NBA is a dynamic system where strategic equilibria can shift. Defensive adaptations, rule changes, or other factors could alter the trajectory of three-point emphasis in ways not captured by this two-season comparison.

## References

Agresti, A., & Franklin, C. (2013). *Statistics: The Art and Science of Learning from Data* (3rd ed.). Pearson.

Basketball Reference. (n.d.). *NBA & ABA League Index*. Sports Reference LLC. Retrieved from <https://www.basketball-reference.com/>

Basketball Reference. (n.d.). *NBA Per Game Stats*. Retrieved from [https://www.basketball-reference.com/leagues/NBA\\_2019\\_per\\_game.html](https://www.basketball-reference.com/leagues/NBA_2019_per_game.html)

Moore, D. S., Notz, W. I., & Fligner, M. A. (2017). *The Basic Practice of Statistics* (8th ed.). W. H. Freeman.

*Data files:* NBA2019.csv, NBA2020.csv, NBA2021.csv, NBA2022.csv (included in GitHub repository)

*AI tools were used to assist with code development and report writing.*