OVERVIEW

My name is Taneesh Amin, and I am a Freshman Student who is planning on majoring in Computer Science and/or Data Science with minors with Business Administration and Sport Management. Combining all my interests, I often look at NBA data, specifically on the website Basketball Reference. Basketball Reference does a phenomenal job keeping data up to date for the NBA while making it accessible in CSV form. While they do not make any models or analyses themselves, their raw data gives me the ability to do so. Given this, I knew for my Data Blog I wanted to focus on the NBA.

DECIDING MY DATASET

When analyzing NBA stats, there are two routes you can go, the player route or the team route. Because there are only 30 teams in the league, I decided to go by the player route. I wanted to look at something regarding scoring efficiency. This is where I landed on the controversial topic of free throws! Various articles talk about how players have a friendlier whistle and free throws aren't assigned at the same rate for everyone. Therefore, I decided I would explore the topic of free throw rates in comparison to various other variables. Additionally, I decided to get data from the latest full season, the 2021-22 season. Basketball reference had a data set with 450 rows and 28 columns (after some data cleaning) that would help me compare free throw rates to different possible factors.

DATA CLEANING

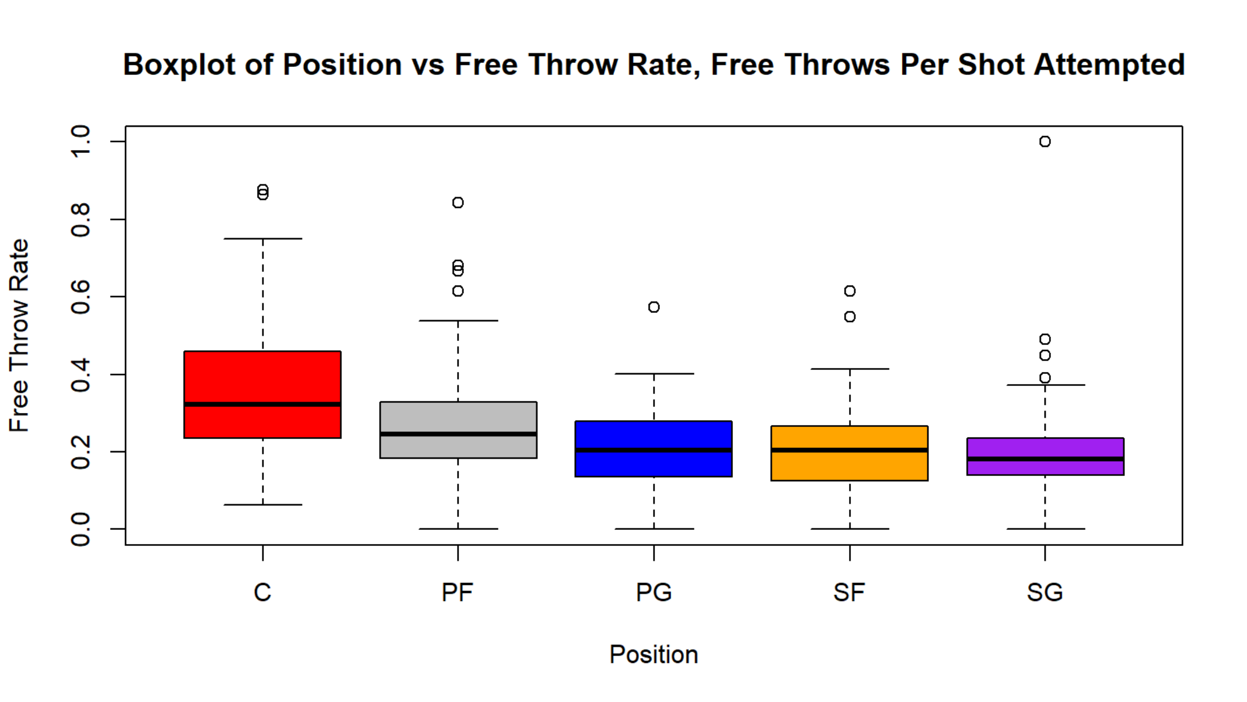
Unsurprisingly, there was a lot of data to clean from this raw, unmodified dataset. One of these issues is empty columns. There were three blank columns in my importation of the Dataset caused by spacing within the CSV file. I used the subset function to select=-(bad column) to remove all of the null columns out of my dataset. Next, I noticed, some players were duplicated within my Dataset. Because of that, I created another subset, that omitted rows with a value that all duplicates had in common (Team = 'TOT') . Lastly, I wanted to make sure my sample had players with a solid amount of playing time. I took the common n = 30 minimum, and I had all players I examined in this dataset play at least 30 minutes. Lastly, I wanted to add a variable of my own, conference. To do this I made a Boolean called "In East" which was false if a player was on a team not in the east and true if they were in the east. After this cleaning, my data went from 605 rows with 30 columns to 450 rows with 28 columns.

HYPOTHESIS 1: CENTERS VS OTHER POSITIONS

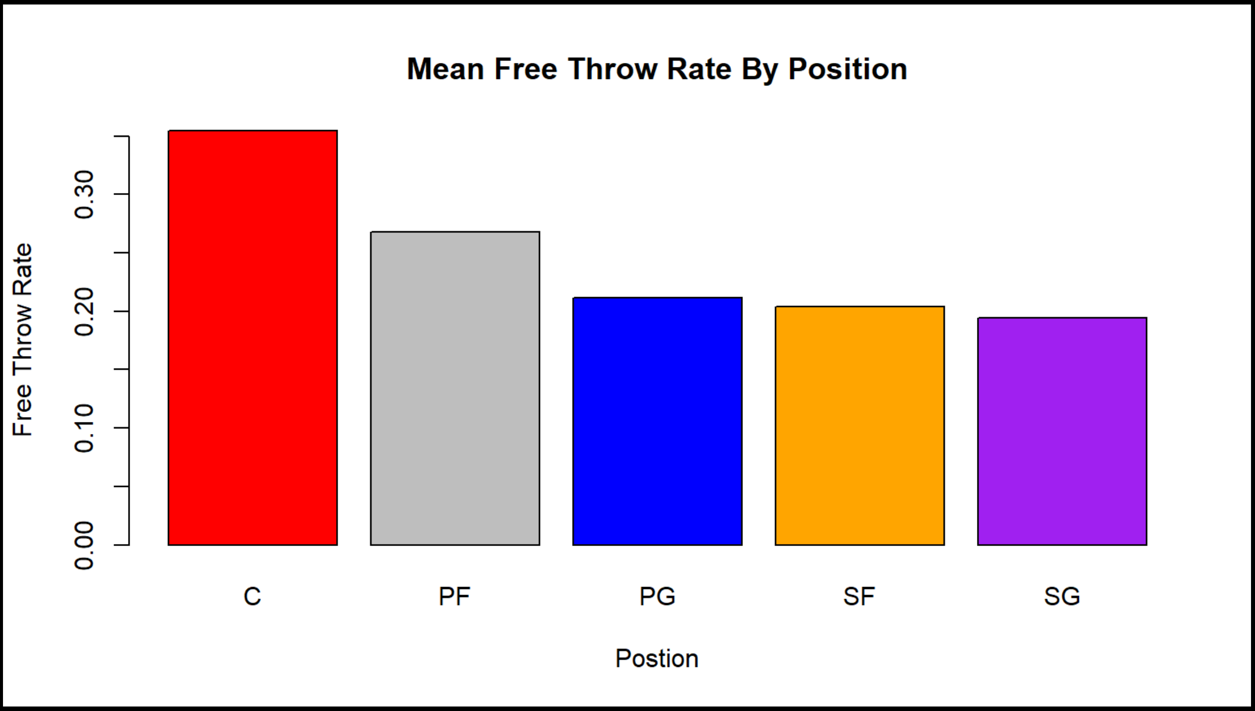
One area of free throw rate I wanted to examine is how many free throws each position in the NBA shoots. So, I went Hypothesis Hunting! My hypothesis was:

“CENTERS WILL HAVE A HIGHER FREE THROW RATE ON AVERAGE COMPARED TO ALL OTHER POSITIONS.”

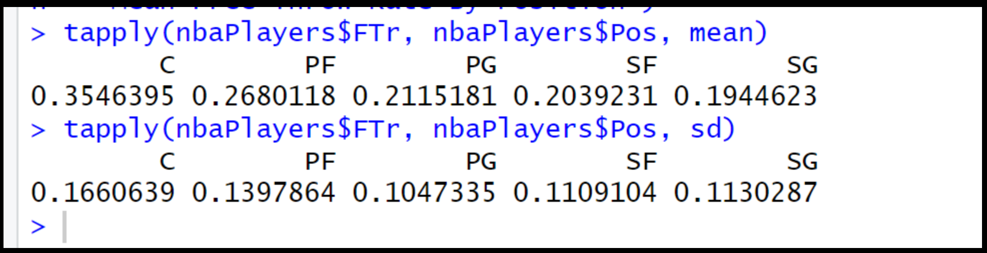
This was in clear contradiction of the null hypothesis of *All positions having the same free throw rate. Let’s look at what the eye test tells us about this data theory*.



JUST A THOUGHT: Most positions seem to have similar distributions apart from the center position. Even the Power Forward position which is the most like Center position in what player composition is has a distribution more like the smaller positions. While all positions have their respective outliers, it seems that the distribution of the Center position would lead to the assumption that they tend to get free throws at a higher rate. We see the medians here, but what does the mean distribution look like?



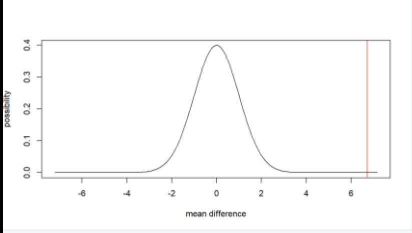
The mean also shows Centers far above the other positions. In order of height average, smallest to largest, we have: PG, SG, SF, PF, C. Except for the Point Guard position which had a higher mean rate than the SF and SG position, it seems that as the positions get taller, the free throw rate increases. This led me to my first hypothesis test where I tested every position against each other to see which ones had significant differences in Free Throw Distribution. Here are their respective means and standard deviations for context.



```
> tapply(nbaPlayers$FTr, nbaPlayers$Pos, mean)
        C        PF        PG        SF        SG 
0.3546395 0.2680118 0.2115181 0.2039231 0.1944623 
> tapply(nbaPlayers$FTr, nbaPlayers$Pos, sd)
        C        PF        PG        SF        SG 
0.1660639 0.1397864 0.1047335 0.1109104 0.1130287 
> 
```

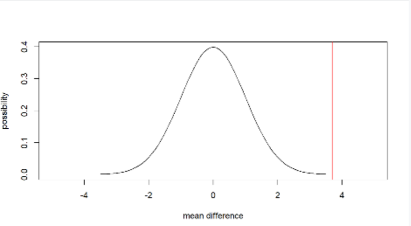
Centers also have the highest standard deviation, so I wanted to use ZTest from Data to see what the P Value is that my data could be explained from my null hypothesis. Here are the P-Values of each position, with graphs following on the next page:

Positions Compared	P-Value
C-PG	8.374 * 10^-12
C-PF	0.000111
C-SG	1.210143 * 10 ^ -14
C-SF	8.375522 * 10^-13



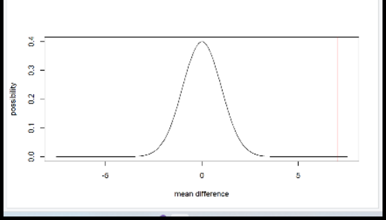
C-PG

ZTEST
GRAPH



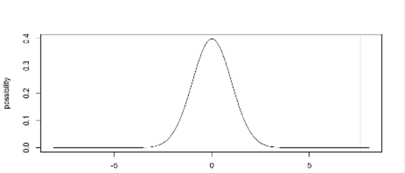
C-PF

ZTEST
GRAPH



C-SF

ZTEST
GRAPH



C-SG
ZTEST

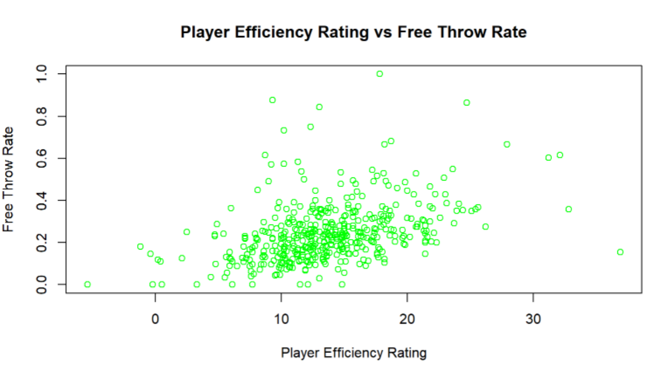
GRAPH

Using these P-Values, we see it is below the P-Value of 0.05. However, our level of significance is even lower at 0.005 by the Bonferroni correction. (0.05/ (5 choose 2)) = 0.005.

All values in our table fall below the 0.005 significance level, so by the multiple hypotheses correction, we still can reject the null hypothesis that the free throw rate of Centers in the NBA is the same as the mean of all other positions.

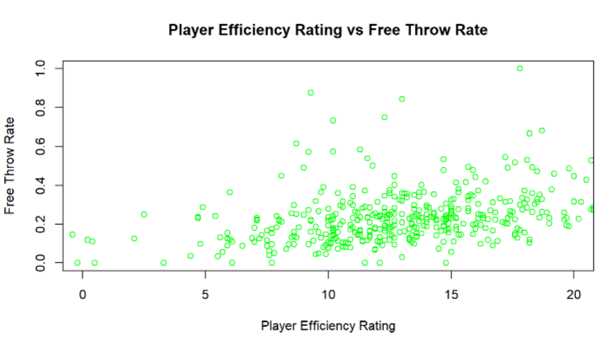
HYPOTHESIS 2: PER VS FTR

Next, I want to see if players who have a high Player Efficiency rating would have a high Free Throw Rate. The formula is calculated using league averages in points, rebounds, assists, turnovers, blocks, and other box score stats and seeing how they compare to player averages. Therefore, people with a PER better than the leagues median PER are in the top half of the league, and those with a PER below the league median are in the bottom half of the league. Using the summary() function, I was able to find a median PER of 13. Let's look at how PER corresponds to free throw rate using a box plot! The graph beside the text covers the whole league. However, it all looks clumped together, so I made a zoomed in version (Graph Right Below) to examine if there is just a clump or an actual relationship.

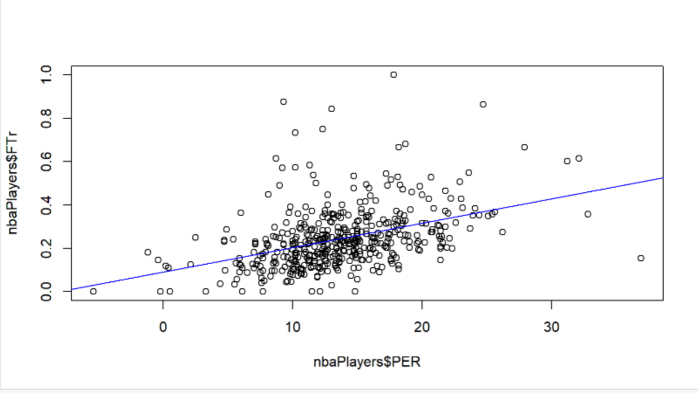


Using this graph I formulate the Alternative Hypothesis that: Players in the top half of the league in terms of PER will have a better FTR than players in the bottom half of the league.

My Null hypothesis is that regardless of PER ranking, FTR is the same.



Examining this graph, it seems like there is some relationship between PER and FTR. In fact, using a line of best fit predictor, you do see some increase in FTR as PER goes up, but is it signicant?

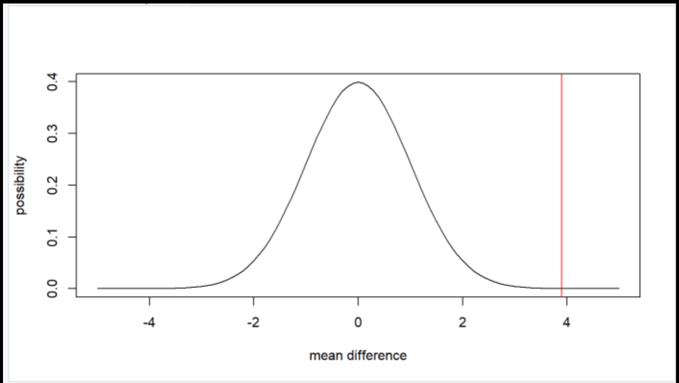


To test this, we are running a ZTest from aggregate data. We used the subset function to find the mean, standard deviations, and number of rows of players FTR depending on if they have a PER of 13+ or a PER below 13. When performing a ZTest on these values, we find that the P-Val is **4.770093 * 10^-5.**

Here is a table of the means, standard deviations, and the number of rows per subset.

PER	SD	ROWS	MEAN
>= 13	.133	230	.282
< 13	.133	215	.201

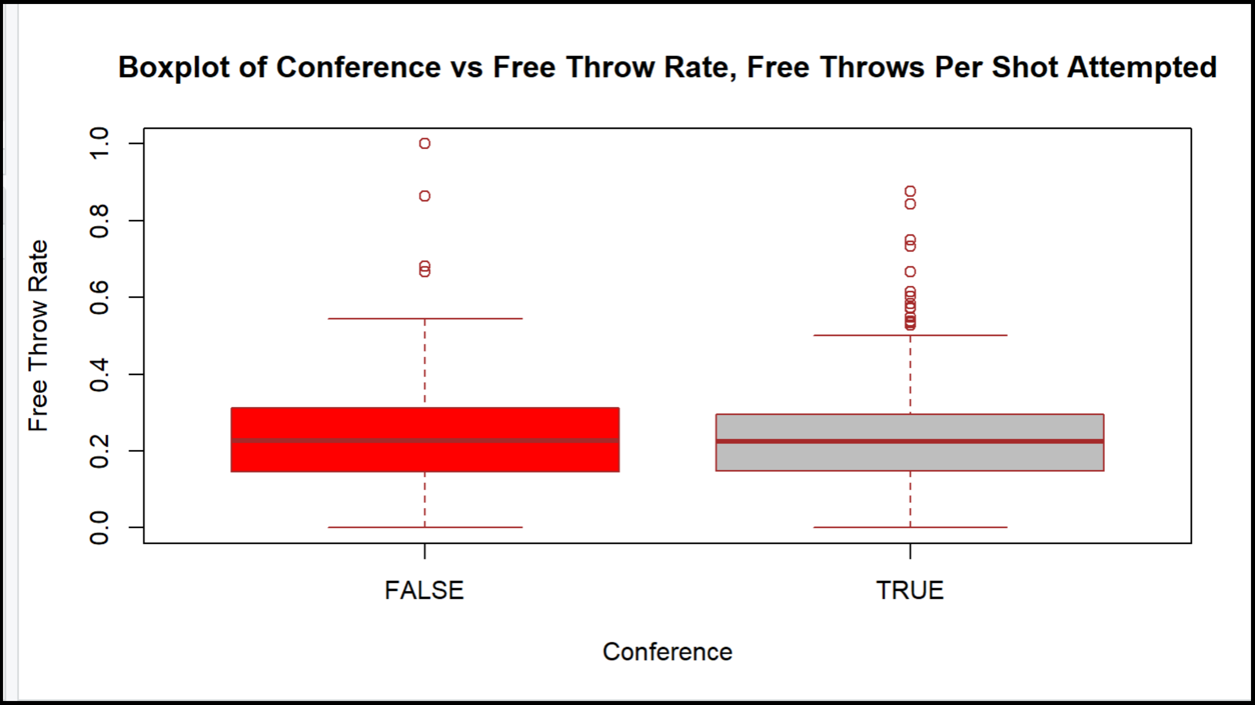
We find that there is similar standard deviation for both subsets! However, our Z-Test showed a significant difference in the means. This means players who are more efficient also get to the free throw line more per possession. A further test I want to do once more data comes out is whether the refs favor these players or if they are just skilled at getting to the free throw line. Below is the graph of my Z-Test on the hypothesis test conducted for my second hypothesis.



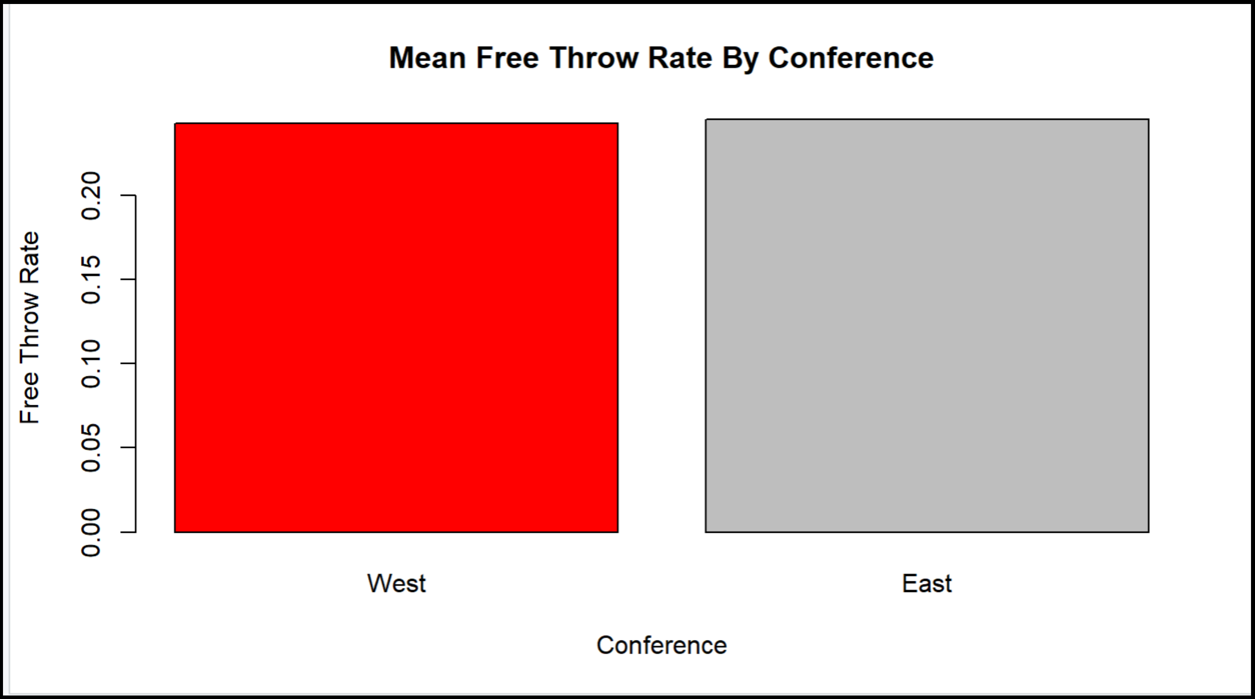
HYPOTHESIS 3: CONFERENCE VS FTR

Now, we are going to look at Conference to see if there is any differentiation within those. This column was self -created using the %in% function on R Studio. I separated teams into conferences based on what the NBA Considers them. Let's look at the distribution of Free Throw Rate by Conference:

IMPORTANT GRAPH NOTE: FALSE = WESTERN, TRUE = EASTERN



The distribution looks uniform with the East having slightly more outliers but the west's outliers being higher in value. We also see that the lows are identical, and the medians are as well. So far it is looking like my null hypothesis will fail to be rejected. While the medians look relatively the same, how do the means look?



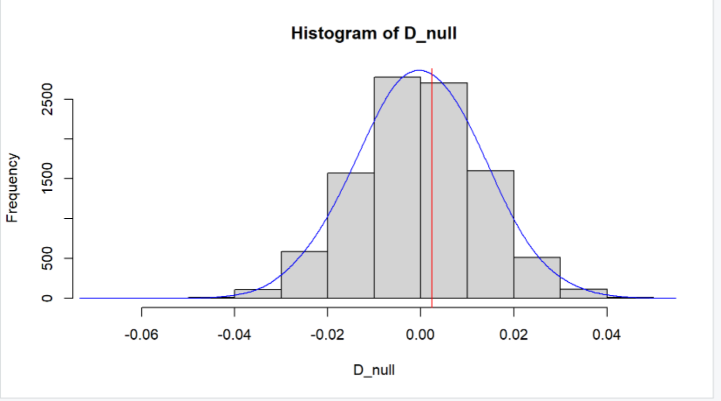
The means also look the same. It is hard to find too many differences in this data, as when I looked at the sample distribution of EAST vs WEST (TRUE vs FALSE) I founded there to 227 rows for False and 224 rows for True. Even in terms of positional value, we saw that there were a similar amount of players in each position as well. This means the datasets had a near even amount of results too. This made an alternative hypothesis difficult to choose but I landed on this.

“NBA teams in the EAST will have a higher mean FTR than teams in the WEST”

Here is some data that led to my conclusion:

	East	**West**
Mean	0.2427313	0.2451161
SD	0.1471549	0.1346813

Now Let’s Put it in the TESTER:



We find on the graph of a Permutation Test that the probability on average is around 0.42. Running it multiple times, the number is always above 0.4. Therefore, I fail to reject the null hypothesis that the East and West have players get to the free throw line at the same rate. We can essentially eliminate bias between the two conferences as a result which was never too hot of an issue in the first place.

This also means that we found one area where there is not a discrepancy in FTR. We can use this data in the future to conclude the conferences are treated equally.

CONCLUSION:

Through my hypothesis testing, I founded that two null hypotheses could be rejected. I also had a null hypothesis that was failed to be rejected. Because of the corrections I made for multiple hypothesis testing, I believe that no TYPE I errors were made. The P-Value being as high as it was for the null hypothesis, I failed to reject makes me believe that no Type II errors were present either. The NBA’s Free Throw disparity is no myth. Some positions clearly get more free throws than others, and better players do get a friendlier whistle overall. However, we cannot pinpoint a cause for that as correlation cannot equate to causation. The goal of this data blog was to show that free throws are not as random as we think, and the media has a good basis for questioning how they are called. My blog was meant to create a story on NBA free throws, and that is exactly what I did.

CREDITS:

I want to acknowledge Basketball Reference for providing me this dataset on their public website. I give all credit to Basketball Reference for their raw data which I analyzed today.

Any data I used can be publicly accessed here. It is important to give credit to dataset providers, as they provide us the tools for analysis and making hypotheses like I do in this blog right here.