Play-by-Play Analysis of Free Throws

Kevin Shain

Yale University
kevin.shain@yale.edu

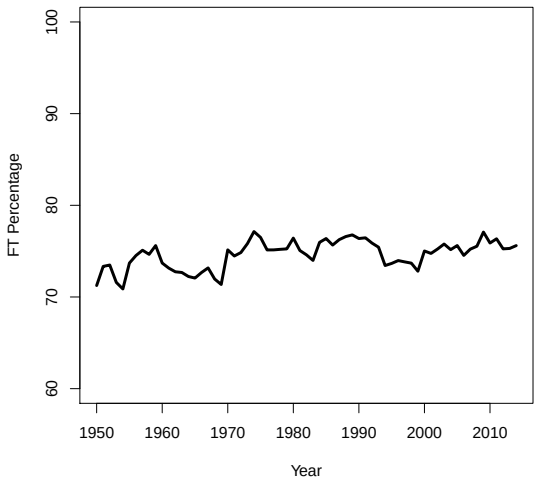
Abstract

It is often observed in statistical analyses of sports that free throw shooting, while simple to the fan, is a remarkable act. It seems to be one of the few acts that could be considered identical, independent, and distinguishable. The hypothesis of independence will be tested using a runs test and transitions analysis. The effect of game time, score differential, and venue will also be examined using contingency tables and graphical analysis. This article uses play-by-play unaggregated free throw data to present the most detailed analysis of streaks in free throws to date.

I. Introduction

The interest of statisticians in free throw shooting was sparked by one observation: free throw percentage is one of the most historically constant statistics in all of sports. The standard plot shows just the year and total league free throw percentage to demonstrate how little it changes over the years.

Figure 1: NBA Free Throw Percentage by Year



That is an interesting observation, but the data used for this analysis will be much less aggregated. Since the advent of real time score updates in 2001, written play-by-play commentary has been made for every NBA game. This record of every game is available through *www.basketball-reference.com*. Extracting free throw information from this repository gives unprecedented information about an individual free throw attempt in game context and in the context of other attempts. This type of analysis has been done for an individual player, but never across the entire league over many seasons. The data set compiled for this report enables the testing of previously inaccessible hypotheses.

II. Data Scrape

All the data for this report is from *www.basketball-reference.com* which makes player statistics and play-by-play accounts available for academic use. For the league average free throw percentage, I pulled the end of season statistics of every player. The URLs only varied by the year so I was able to loop through all available years, starting in 1950, and bind them together into one dataframe. The code for this can be found in Appendix A. The more involved data scrape came in pulling data from play-by-play reports. This is a much more difficult process which starts with generating a vector of IDs of all games with play-by-play data. I did this by pulling the game IDs that are used behind the scenes by *www.basketball-reference.com*. With the game IDs, I was able to loop through the URLs with play-by-play data. The data was originally in an awkward table that looks nice on the screen, but its format depends on the

action being performed in every row. Each entry had the action, game time, and current score; however, there were a number of difficulties with the data set. I had to change the game time from minutes and seconds within a quarter to fractions of minutes of total elapsed time. Also, I was able to use *grep* to find "made free throw" or "missed free throw" and code them as 1 or 0. Finally, I was able to extract player names from the event commentary and make that into a categorical variable.

III. Analysis

1. Serial Correlation

A popular topic in basketball statistics is the so called "Hot Hand". The general conclusion from many analyses is that there is no serial correlation in shots from the field. However, there has been much less research into the serial correlation of free throws. When surveyed, most fans believe in streaky shooting from the field, and I think that they would believe the same thing about free throws. I will test the hypothesis that there is positive serial correlation between free throw attempts.

My first test is a transistions analysis. I first split the data set by player to find their personal sequence of attempts. Then, I break the attempts into those taken after a made or a missed free throw. With this, I could then find the percentage after a make or a miss.

Table 1: *End of Game Shooter Percentage*

After Make	After Miss
76.3%	71.5%

This is a large difference in percentages given a data set with 800,000 attempts. The p-value from a χ^2 test is $<$ 2e-16, but there may be confounds. One example is that bad free throw shooters are more likely to miss their shots so they contribute more to the *After Miss* group than *After Make*.

To eliminate this effect, I then analyzed serial correlation on a player by player basis using a Runs test. This test compares the number of streaks in a series of attempts to the average number of streaks expected with no correlation. It is clear that if makes are more common after makes or misses more common after misses, there would be fewer runs in a player's sequence than if attempts were independent. The mean number of runs is given by

$$\bar{r} = \frac{2N_{made} \cdot N_{missed}}{N_{total}} + 1 \tag{1}$$

and the standard deviation is

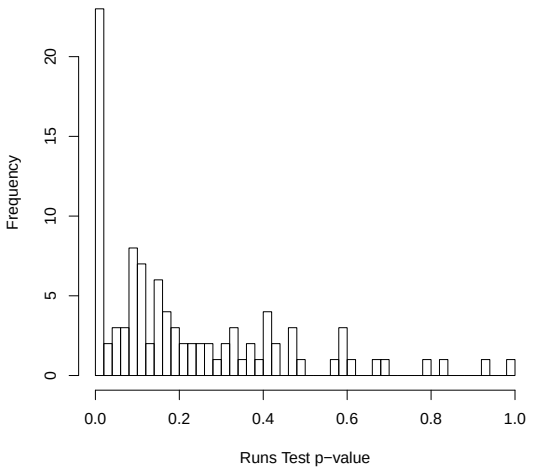
$$\sigma = \sqrt{\frac{2N_{made} \cdot N_{missed}(2N_{made} \cdot N_{missed} - N_{total})}{N_{total}^2(N_{total} - 1)}} \tag{2}$$

The test statistic is then

$$Z = \frac{r - \bar{r}}{\sigma} \tag{3}$$

I applied the Runs test to the top 100 free throw shooters by number of attempts. The p-values are plotted in the histogram.

Figure 3: Runs Test Histogram

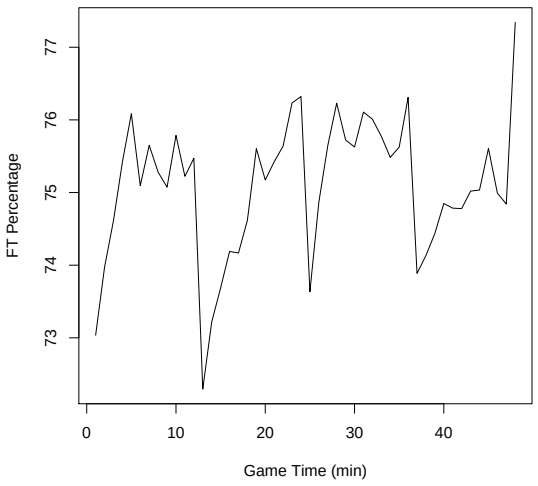


This histogram shows that the nearly a quarter of p-values are less than 0.02. This shows that there are significantly fewer runs than expected if there were no serial correlation. Therefore, the "Hot Hand" phenomenon appears to hold for free throws even if it doesn't hold for shots from the field. This may not be so surprising considering free throws are all exactly the same shot and are often taken in immediate succession.

2. End of Game Free Throws

Anyone who has watched a basketball game knows that the predominant strategy when losing near the end of the game is to foul the opposing team. Though the other team is likely to convert on many of their free throw attempts, the losing team gets the ball back with less time coming off the clock. Knowing this strategy, the winning coach will generally substitute in better free throw shooters and design plays to get the ball in their hands. On the other side, the losing team is trying to find the worst free throw shooter on the other team to foul. Which team gets the best of this game within the game? To answer this, I first made a plot of free throw percentage versus game time. I simply broke the dataset up by minute and calculated the percentage.

Figure 2: Free Throw Percentage vs. Game Time



This plot shows that the free throw percentage goes up around 2% from the average in the last minute of the game. This increase is considerably larger than the noise so it seems like the increase in free throw percentage in the last minute is significant. However, this may be due to players focusing more in the last minute. This plot does not really answer the question I posed earlier since it does not address whether better free throw shooters are fouled more at the end of games. To test the question more specifically, I created a data frame of players and their average free throw percentage. Then, I split the total dataset into two groups: free throws in the last minute and free throws at other times. I then took an average of the free throw percentage of the shooter from every attempt in those subsets.

Table 2: *End of Game Shooter Percentage*

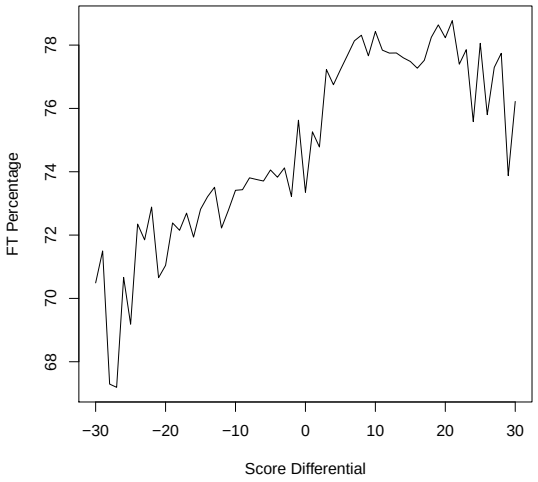
Shooter Percentage **before** Last Minute	Shooter Percentage **in** Last Minute
75.3%	78.0%

This result shows that the players the players who are fouled in the last minute of the game generally have a higher free throw percentage. Given that the data set is so large, the 2.7% difference seems significant. In the end, it seems like the winning team succeeds in making sure that better free throw shooters get fouled more often at the end of games.

3. Effect of Score Differential

Another possible factor of free throw percentage is momentum in the game. To the fan, it often seems like the team who is ahead, especially by more than just a couple of points, has the confidence necessary to make more shots. To test whether this perception appears in the data, we can use the fact that in the play-by-play, there is the current score recorded with every event. I went through and calculated a score differential from the perspective of the shooting player. This enabled me to calculate a free throw percentage aggregated by score differential. This shows that free throw percentage has a visually obvious positive correlation with score differential. There is the possible confound of previous free throws correlating with score and later free throws, so it is more difficult to determine the effect of score differential on free throw percentage. This could be a topic of further study as the data set has each free throw and score differential placed within the context of the game and other free throws.

Figure 3: Free Throw Percentage by Score Differential



4. Home vs. Away Free Throws

A popular topic of analysis in all sports is home field advantage. A free throw seems superficially like it would be the same in any venue, but fans clearly do not believe this as they make noise and wave foam tubes in an attempt to hamper the visiting team. I have encoded the home and away team for each free throw in the data set so the effect of venue can be tested. I do this by finding free throw percentage conditional upon whether the shooter was home or away.

Table 3: *Home vs. Away FT Percentage*

Home	Away
75.4%	75.3%

It is clear that there is not a significant difference between home and away free throw percentage. A difference of 0.1% is easily within the season to season variation. Despite the efforts of home fans, there is no evidence to believe that the game location affects free throw shooting.

IV. Conclusion

This paper demonstrates the insights play-by-play analysis can yield that season totals cannot. From the cleanest and most complete data set of its kind, several conclusions could be made about free throw shooting. A fascinating discovery is that unlike shots from the field, free throws show a significant positive serial correlation. Additionally, the leading team in the last minute of the game is more successful at getting a good shooter fouled than the losing team is at fouling a bad shooter. There is also a positive correlation between the score differential at an attempt and the free throw percentage. Finally, there is no evidence of a difference between home and away free throw percentage. Free throws have previously been studied because of their apparent uniformity and independence, but there is a wealth of nuance elucidated through play-by-play analysis.

Appendix A: Season Totals Data Scrape

```
> trim <- function (x) gsub("^\\s+|\\s+$", "", x)
> for (i in 1950:2014){
+   url <- sprintf("http://www.basketball-reference.com/leagues/NBA_%d_totals.html",i)
+   z <- scan(url, what="", sep="\n")
+   start <- grep("<col><col><col><col><col>", z, fixed=TRUE)+1
+   z <- z[start:length(z)]
+   z <- z[-c(32,33)]
+   end <- grep("</div><!-- div.table_container#div_xxxx -->", z, fixed=TRUE)-3
+   z <- z[1:end]
+   y <- gsub("<[^<>]*>", "", z)
+   colNames <- y[1:31]
+   colNames <- trim(colNames)
+   y <- y[32:length(y)]
+   y <- trim(y) # Remove whitespace
+   x <- matrix(y, length(y)/31, 31, byrow = TRUE, dimnames=list(NULL,colNames))
+   x <- x[,-1]
+   x <- x[,-1]
+   printName <- sprintf("%d_stats.csv",i)
+   write.csv(x, file = printName, row.names=FALSE) # Save each year separately as CSV
+ }
> for (i in 1950:2014){
+   filename <- sprintf("%d_stats.csv",i)
+   a <- read.csv(filename, as.is=TRUE, na.strings="")
+   a <- a[a$Rk!="Rk",] #remove Column Name rows
+   a <- a[!duplicated(a$Player),] # Players that swtich teams are listed twice
+   a$Year <- i
+   row.names(a)<-NULL
+   a <- a[c(30,1:29)]
+   if (exists('bb') && is.data.frame(get('bb'))){
+     bb <- rbind(bb, a)
+   }
+   else{
+     bb <- a
+   }
+ }
> k <- c(1,2,5,7:30)
> for (j in k){
+   bb[,j] <- as.numeric(bb[,j])
+ }
> write.csv(bb, file = "bb.csv",  row.names=FALSE)
```

Appendix B: Play-by-Play Data Scrape

```
> for (i in 2001:2013){ # Get game IDs
+   nn <- NULL
+   url <- sprintf("http://www.basketball-reference.com/leagues/NBA_%d_games.html",i)
```

```
+   if (class(m <- try(scan(url, what="", sep="\n"))) == "try-error") {
+     cat("ERROR on page", i, "\n")
+   }
+   boxscoreLines <- grep(">Box Score<", m)
+   n <- m[boxscoreLines]
+   IDs <- substr(n, 44, 55)
+   Season <- rep(i, length(IDs))
+   if (exists('IDFrame') && is.data.frame(get('IDFrame'))){
+     IDFrame <- rbind(IDFrame, data.frame(Season=Season, ID=IDs))
+   }
+   else{
+     IDFrame <- data.frame(Season=Season, ID=IDs)
+   }
+ }
> for (i in 1:length(IDFrame$ID)){
+   ID <- as.character(IDFrame$ID[i])
+   Season<- IDFrame$Season[i]
+   url <- sprintf("http://www.basketball-reference.com/boxscores/pbp/%s.html", ID)
+   if (class(z <- try(scan(url, what="", sep="\n"))) == "try-error") {
+     cat("ERROR on page", i, gameID[i], "\n")
+   }
+   start <- grep("Start of 1st quarter", z, fixed=TRUE)
+   end <- grep("<td class=\"valign_top\">", z, fixed=TRUE)
+
+   h <- z[(start-15):start]
+   timeLine <- grep("<th>Time</th>", h, fixed=TRUE)
+   h2 <- gsub("<[^<>]*>", "", h)
+
+   z <- z[start:end]
+   z <- gsub("</td><td[^<>]*>", ":", z)
+   y <- gsub("<[^<>]*>", "", z)
+   awayTeam <- h2[timeLine+1]
+   homeTeam <- h2[timeLine+3]
+   ##________________________Find Relevant lines_____________________________##
+   eventLines <- grep(" ", y, fixed=TRUE)
+   events <- y[eventLines]
+   time <- y[eventLines-1]
+   ##________________________Find Quarter Breaks_____________________________##
+   first <- grep("End of 1st quarter", y)
+   second <- grep("End of 2nd quarter", y)
+   third <- grep("End of 3rd quarter", y)
+   fourth <- grep("End of 4th quarter", y)
+   ot1 <- grep("End of 1st overtime", y)
+   ot2 <- grep("End of 2nd overtime", y)
+   ot3 <- grep("End of 3rd overtime", y)
+   ##_________________Make Time in fractions of a minute_____________________##
+   times <- strsplit(time,":")
+   times <- unlist(times)
```

```
+   times <- as.numeric(times)
+   time <- times[seq(1, length(times), by=2)] + times[seq(2, length(times), by=2)]/60
+   ##_____________Get actual gameTime adjusted by quarter_____________##
+   gameTime <- NA
+   if (class(gameTime <- try(ifelse(eventLines < first, 12-time,
+     ifelse(eventLines < second, 24-time,
+     ifelse(eventLines < third, 36-time,
+     ifelse(eventLines < fourth), 48-time,
+     ifelse(eventLines < ot1, 53-time,
+     ifelse(eventLines < ot2, 58-time,
+     ifelse(eventLines < ot3, 63-time,  NA)))))))) == "try-error") {
+     cat("ERROR separating quarters of", i, gameID[i], "\n")
+   }
+   ##_____________________Split event into parts_____________________##
+   events <- as.character(events)
+   Score <- NA
+   Action <- NA
+   Team <- NA
+   for (k in 1:length(events)){
+     pieces <- unlist(strsplit(events[k],":"))
+     if (pieces[1] == " "){
+       Action[k] <- pieces[5]
+       Team[k] <- "Home"
+     }
+     else {
+       Action[k] <- pieces[1]
+       Team[k] <- "Away"
+     }
+     Score[k] <- pieces[3]
+   }
+   ##_____________________Split score into parts_____________________##
+   Score1 <- NA
+   Score2 <- NA
+   for (m in 1:length(Score)){
+     parts <- unlist(strsplit(Score[m],"-"))
+     Score1[m] <- parts[1]
+     Score2[m] <- parts[2]
+   }
+   gameID <- rep(ID, length(Score1))
+   away <- rep(awayTeam, length(Action))
+   home <- rep(homeTeam, length(Action))
+
+   p2 <- NULL
+   p2 <- data.frame(ID=gameID, Away=away, Home=home, gameTime=gameTime, AwayScore=Score1,
+                    HomeScore=Score2, Action=Action, PerformedBy=Team)
+   p2$freeThrow <- NA
+   p2$freeThrow <- ifelse(grepl("makes free throw", p2$Action), 1,
+                          ifelse(grepl("misses free throw", p2$Action), 0, NA))
```

```
+   p2$FTname <- NA
+   for (l in 1:length(p2$freeThrow)){
+     if (!is.na(p2$freeThrow[l])){
+       names <- unlist(strsplit(as.character(p2$Action[l]), " "))
+       p2$FTname[l] <- paste(names[1],names[2])
+     }
+   }
+   p2$Season <- rep(Season, length(p2$gameTime))
+   p2 <- p2[c(11,1:10)]
+   if (exists('pp2') && is.data.frame(get('pp2'))){
+     pp2 <- rbind(pp2, p2)
+   }
+   else{
+     pp2 <- p2
+   }
+ }
> write.csv(pp2, file = "AllPlayByPlay2.csv",  row.names=FALSE)
> justFT <- pp[!is.na(pp$freeThrow),]
> justFT$AwayScore <- as.numeric(as.character(justFT$AwayScore))
> justFT$HomeScore <- as.numeric(as.character(justFT$HomeScore))
> justFT$Venue.f <- factor(justFT$Home)
> justFT$HomeAway.f <- factor(justFT$PerformedBy)
> justFT$Diff <- as.numeric(justFT$HomeScore) - as.numeric(justFT$AwayScore)
> justFT$ShooterDiff <- ifelse(justFT$PerformedBy=="Home", justFT$Diff, -1* justFT$Diff)
> justFT$PerformedByTeam <- ifelse(justFT$PerformedBy=="Home", justFT$Home, justFT$Away)
> justFT$Player.f <- factor(justFT$FTname)
> write.csv(justFT, file = "justFT2.csv",  row.names=FALSE)
```

Appendix C: Free Throws by Game Time

```
> pct <- rep(NA, 48)
> freq <- rep(NA, 48)
> for(min in 1:48){
+   total <- length(justFT$freeThrow[ceiling(justFT$gameTime)==min])
+   freq[min] <- total
+   made <- sum(justFT$freeThrow[ceiling(justFT$gameTime)==min])
+   pct[min] <- made/total*100
+ }
> plot(pct, type='l', xlab="Game Time (min)", ylab="FT Percentage",
+      main="Free Throw Percentage vs. Game Time")
> ##_____________________Better shooters in last minute?_____________________##
> playerList <- justFT$FTname[!is.na(justFT$FTname)]
> playerList <- unique(playerList)
> playerPct <- data.frame(Name = rep(NA, length(playerList)),
+                         Pct = rep(NA, length(playerList)))
> for (i in 1:length(playerList)){
+   playerPct$Name[i] <- playerList[i]
+   temp2 <- justFT$freeThrow[justFT$FTname==playerList[i]]
```

```
+   playerPct$Pct[i] <- sum(temp2)/length(temp2)
+ }
> lastMin <- justFT[ceiling(justFT$gameTime)==48,]
> lastMinPctSum <- 0
> for (i in 1:length(lastMin$gameTime)){
+   lastMinPctSum<-lastMinPctSum+playerPct$Pct[playerPct$Name==as.character(lastMin$Player.f[i])]
+ }
> lastMinPct <- lastMinPctSum/length(lastMin$gameTime)
> earlyMin <- justFT[ceiling(justFT$gameTime)!=48,]
> earlyMinPctSum <- 0
> for (i in 1:52838){
+   earlyMinPctSum<-earlyMinPctSum+playerPct$Pct[playerPct$Name==as.character(earlyMin$Player.f[i])]
+ }
> earlyMinPct <- earlyMinPctSum/52838
```

Appendix D: Score Differentials

```
> pct2 <- rep(NA, 61)
> freq2 <- rep(NA, 61)
> mades <- rep(NA, 61)
> j <- 1
> for(diff in -30:30){
+   paste(diff)
+   total <- length(justFT$freeThrow[justFT$ShooterDiff==diff])
+   made <- sum(justFT$freeThrow[justFT$ShooterDiff==diff])
+   pct2[j] <- made/total*100
+   j <- j+1
+ }
> x <- c(-30:30)
> plot(pct2 ~ x, type='l', xlab="Score Differential", ylab="Free Throw Percentage",
+      main="Figure 3: Free Throw Percentage by Score Differential")
```

Appendix E: Home vs. Away

```
> HAtable <- table(justFT$HomeAway.f, justFT$freeThrow)
> C <- matrix(c( HAtable[1,2],  HAtable[2,2],  HAtable[1,1],  HAtable[2,1]), nrow=2, ncol=2,
+             dimnames = list(c("Away", "Home"),
+                             c("Made FT", "Missed FT")))
> chisq.test(C)
> FTAway <- HAtable[1,2]/(HAtable[1,2]+HAtable[1,1])
> FTHome <- HAtable[2,2]/(HAtable[2,2]+HAtable[2,1])
```