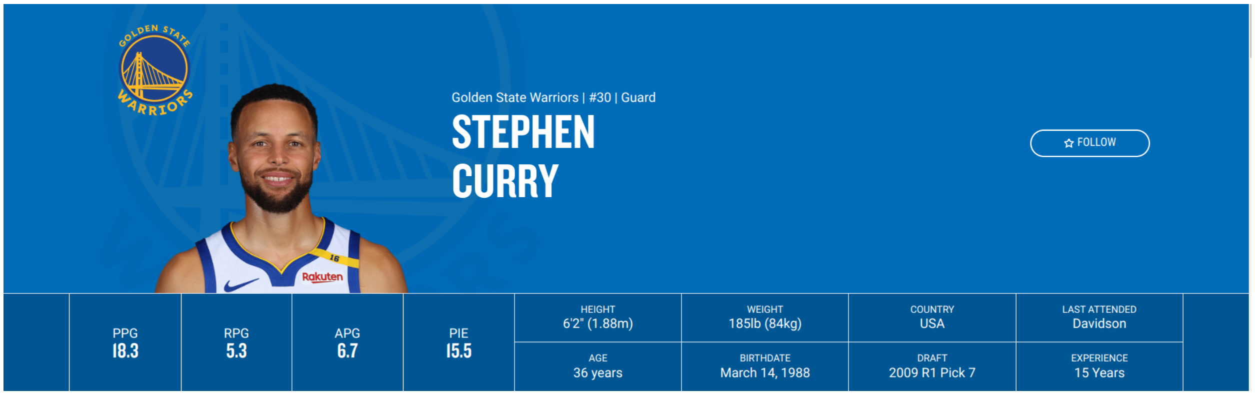
Portfolio

Richard K. Wainaina

2024-10-30

Stephen Curry - NBA

The data I am analyzing is of Stephen Curry; one of my favorite players in the NBA. He plays for the Golden State Warriors and has been for his whole career. The data looks at his performance throughout the seasons since he started playing: 2009 - 2024.



Golden State Warriors | #30 | Guard

STEPHEN CURRY

PPG	RPG	APG	PIE	HEIGHT	WEIGHT	COUNTRY	LAST ATTENDED
18.3	5.3	6.7	15.5	6'2" (1.88m)	185lb (84kg)	USA	Davidson
				AGE	**BIRTHDATE**	**DRAFT**	**EXPERIENCE**
				36 years	March 14, 1988	2009 R1 Pick 7	15 Years

Figure 1: Stephen Curry Profile

The data set consists of 26 rows and 6 columns. Here is a preview of the data set.

season	games_played	minutes_played	points	field_goals_made	field_goals_attempted
2023-24	74	32.7	26.4	8.8	19.5
2022-23	56	34.7	29.4	10.0	20.2
2021-22	64	34.5	25.5	8.4	19.1
2020-21	63	34.2	32.0	10.4	21.7
2019-20	5	27.9	20.8	6.6	16.4
2018-19	69	33.8	27.3	9.2	19.4

For reference, the data is available on the NBA website https://www.nba.com/stats/player/201939

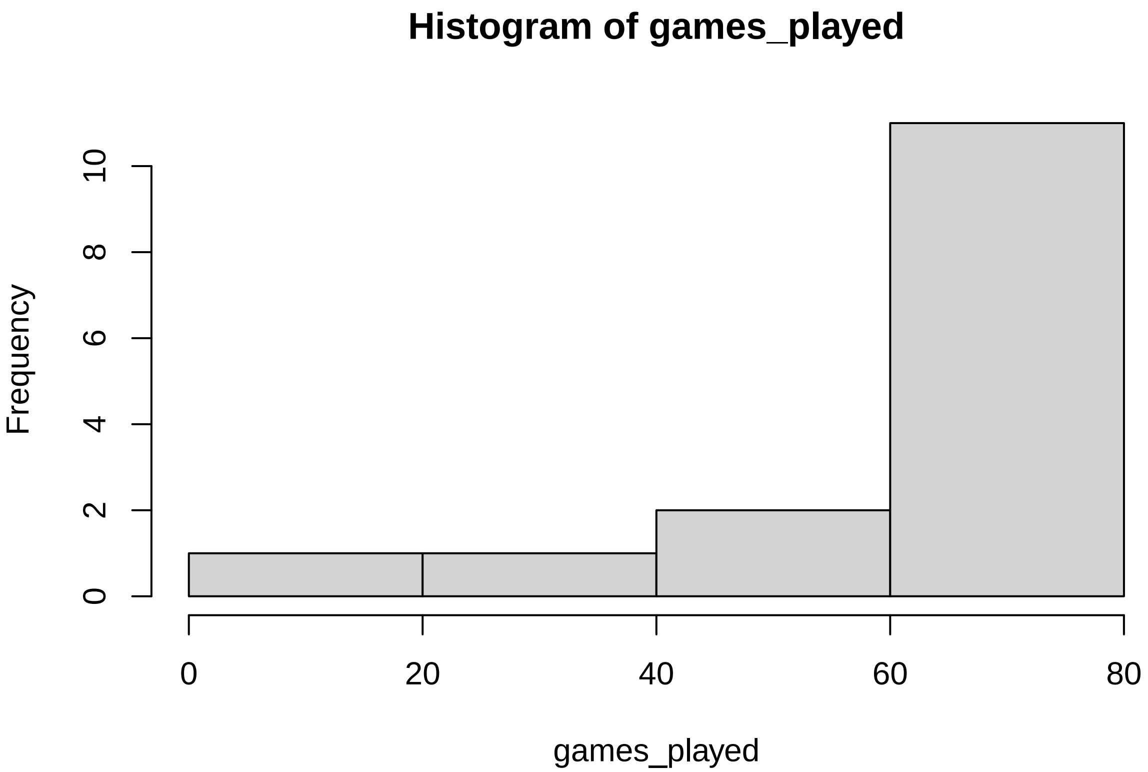
Summary of the data.

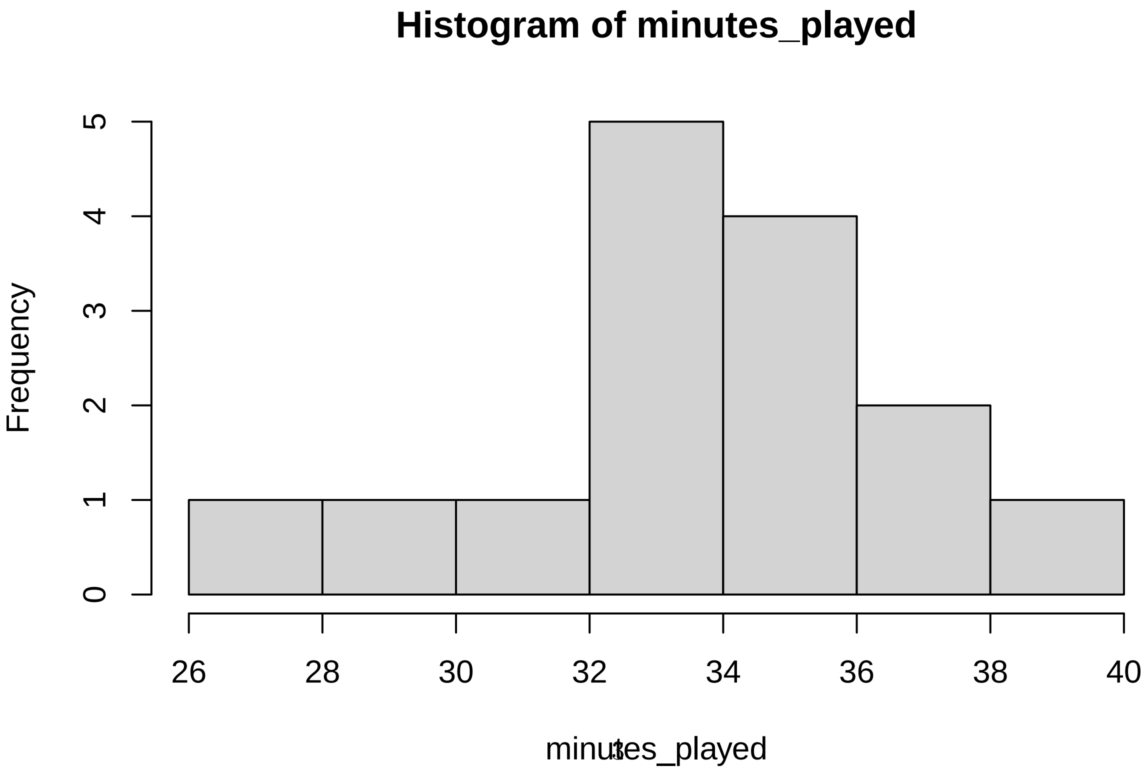
```r
kable(summary(stephen.curry[2:6]))
```

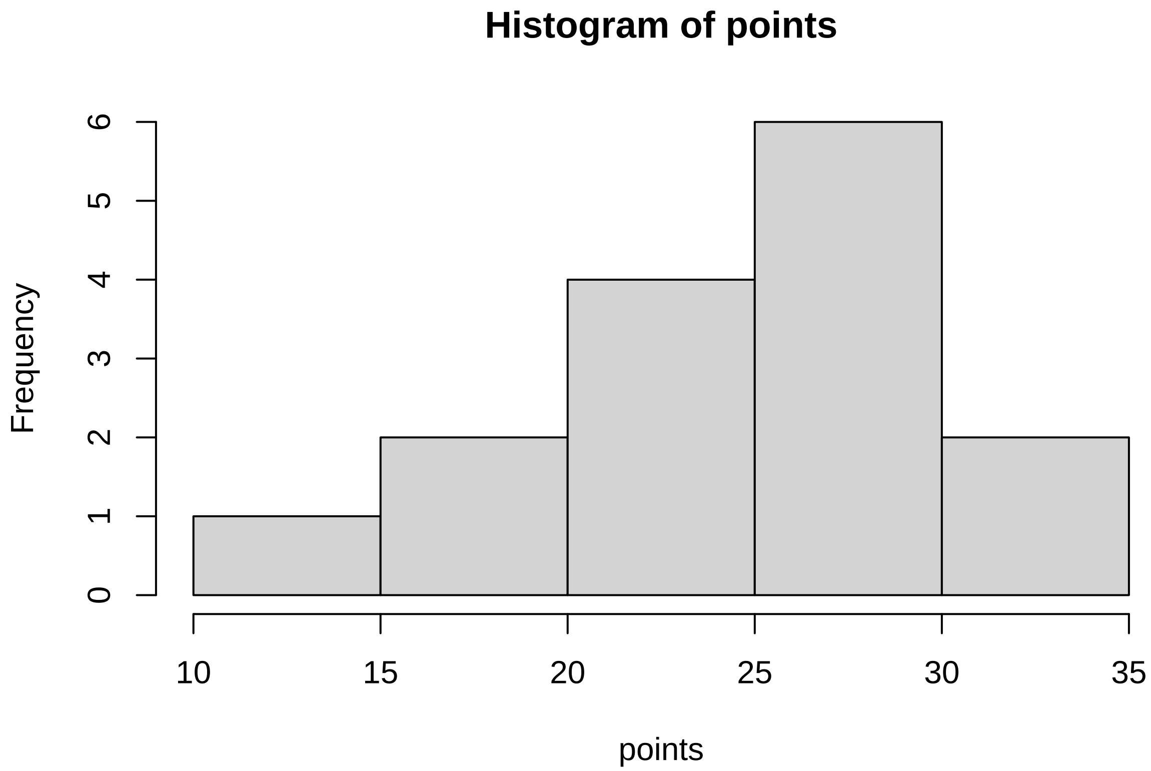
	games_played	minutes_played	points	field_goals_made	field_goals_attempted
	Min. : 5.00	Min. :27.90	Min. :14.70	Min. : 5.600	Min. :11.40
	1st Qu.:59.50	1st Qu.:32.70	1st Qu.:21.85	1st Qu.: 7.400	1st Qu.:16.60
	Median :74.00	Median :33.80	Median :25.30	Median : 8.400	Median :17.80
	Mean :63.73	Mean :33.51	Mean :24.31	Mean : 8.273	Mean :17.59
	3rd Qu.:78.50	3rd Qu.:34.60	3rd Qu.:26.85	3rd Qu.: 9.000	3rd Qu.:19.45
	Max. :80.00	Max. :38.20	Max. :32.00	Max. :10.400	Max. :21.70

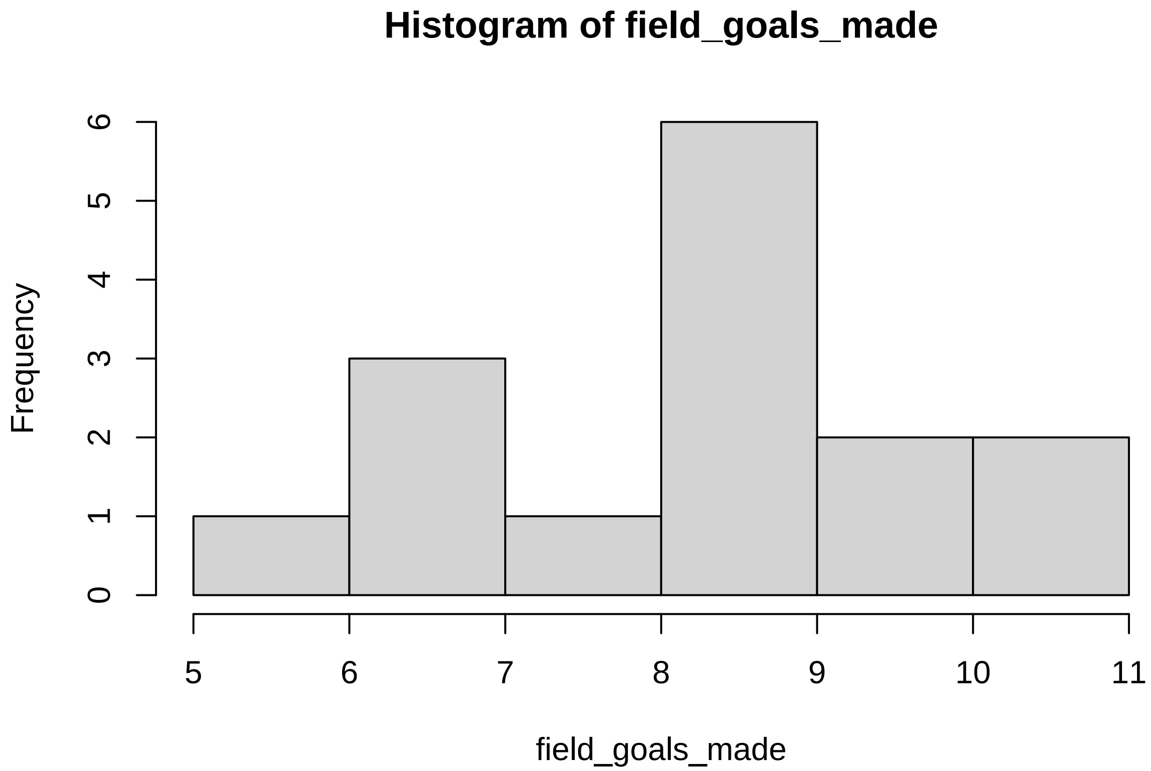
Visualization

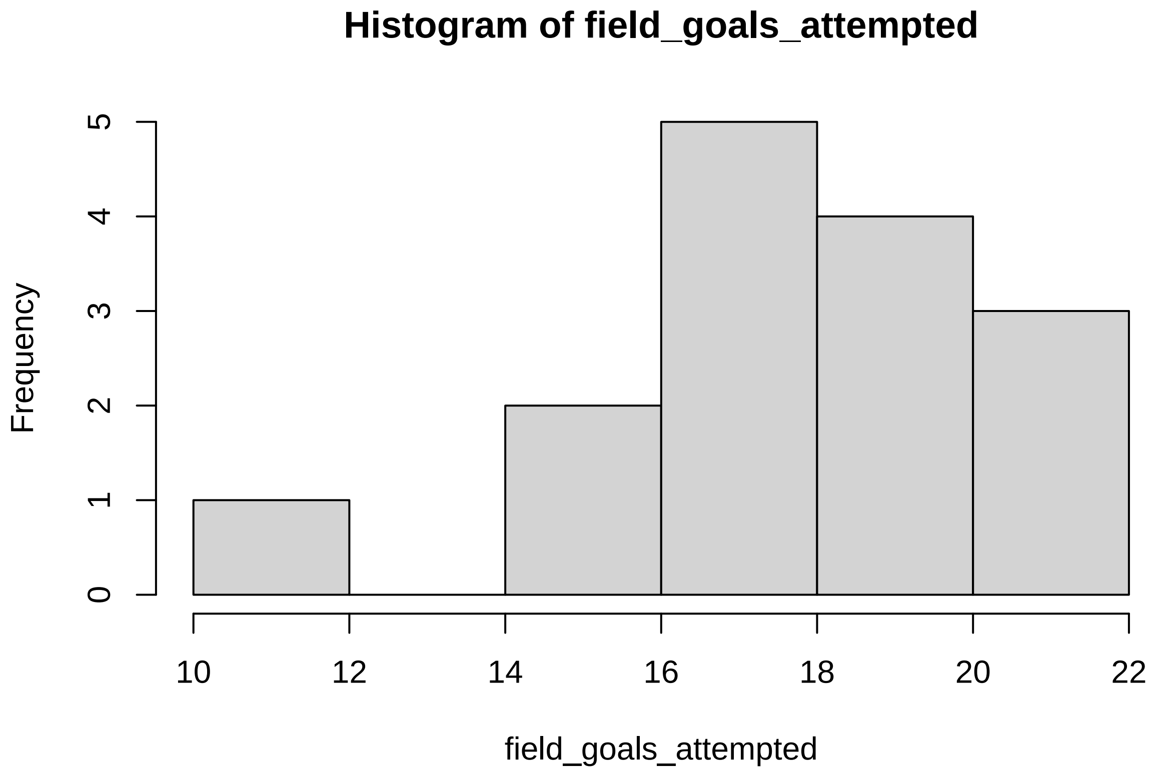
Histogram of each column






Histogram of points
Frequency
0
1
2
3
4
5
6
10
15
20
25
30
35
points


Histogram of field_goals_made
Frequency
0
1
2
3
4
5
6
5
6
7
8
9
10
11
field_goals_made



Student t-test

Test 1

Stephen Curry's lowest scoring season is the 2011-12 season where he averaged 14.7 points per game. We test whether this performance was statistically significant from his averages over the years. Could there have been something wrong with his performance that season, or was this deviation statistically insignificant.

Null: Stephen Curry's performance in 2011-12 is the same as his average over the years.

Alternative: Stephen Curry under-performed in year 2011-12

```
lowest_scoring_season_and_points <- subset(stephen.curry, points == min(stephen.curry$points), select =
kable(lowest_scoring_season_and_points)
```

	season	points
13	2011-12	14.7

```
# average score over the years
mean(stephen.curry$points)
```

```
## [1] 24.31333
```

```
scores_excluding_lowest <- subset(stephen.curry, points != min(stephen.curry$points), select = "points"
t.test(scores_excluding_lowest, mu = min(stephen.curry$points), alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  scores_excluding_lowest
## t = 9.2509, df = 13, p-value = 2.203e-07
## alternative hypothesis: true mean is greater than 14.7
## 95 percent confidence interval:
##  23.02823      Inf
## sample estimates:
## mean of x
##        25
```

Assuming a Gaussian Distribution of his average scores over the years, we use a 1-sample student t-test to check the probability of Stephen Curry averaging 14.7 points or lower if the null is true. The p-value in this case is very small and we can reject the null hypothesis. There is strong evidence to reject the null hypothesis.

Test 2

Stephen Curry is consistently touted as one of the best players in his team. We can compare his field_goal_percentage values to those of the player who won MVP last season (Nikola Jokic) to get an idea of whether they are far apart.

Field_goal_percentage is calculated as field_goals_made/field_goals_attempted. It is an accuracy score.

Null: Stephen Curry's field_goal_percentages are just as good as Nikola Jokic's. No difference between them.

Alternative: Stephen Curry's field_goal_percentages_are not the same as Nikola Jokic's. There is a difference.

Jokic data source: https://www.nba.com/stats/player/203999

```
curry_fgp = stephen.curry$field_goals_made/stephen.curry$field_goals_attempted * 100
jokic_fgp = c(57.3, 58.3, 63.2, 58.3, 56.6, 52.8, 51.1, 49.9, 57.8, 51.2)

t.test(curry_fgp, jokic_fgp)
```

```
##
##  Welch Two Sample t-test
##
## data:  curry_fgp and jokic_fgp
## t = -5.7428, df = 13.761, p-value = 5.441e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.864996  -5.404698
## sample estimates:
## mean of x mean of y
##  47.01515  55.65000
```

The field goal percentages of the two players result in a small p-value of less then 5 (5.441e-05). Therefore, we reject the null hypothesis and infer that Curry's and Jokic's percentages are not the same. A plausible reason might be because they play different positions and serve different roles in their respective teams.

Likelihood Function

I want to assess the probability that Stephen Curry has an average of 10.0 made field goals or more in a given year.

To recap, these are the made field goals.

```
stephen.curry$field_goals_made
```

```
##  [1]  8.8 10.0  8.4 10.4  6.6  9.2  8.4  8.5 10.2  8.2  8.4  8.0  5.6  6.8  6.6
```

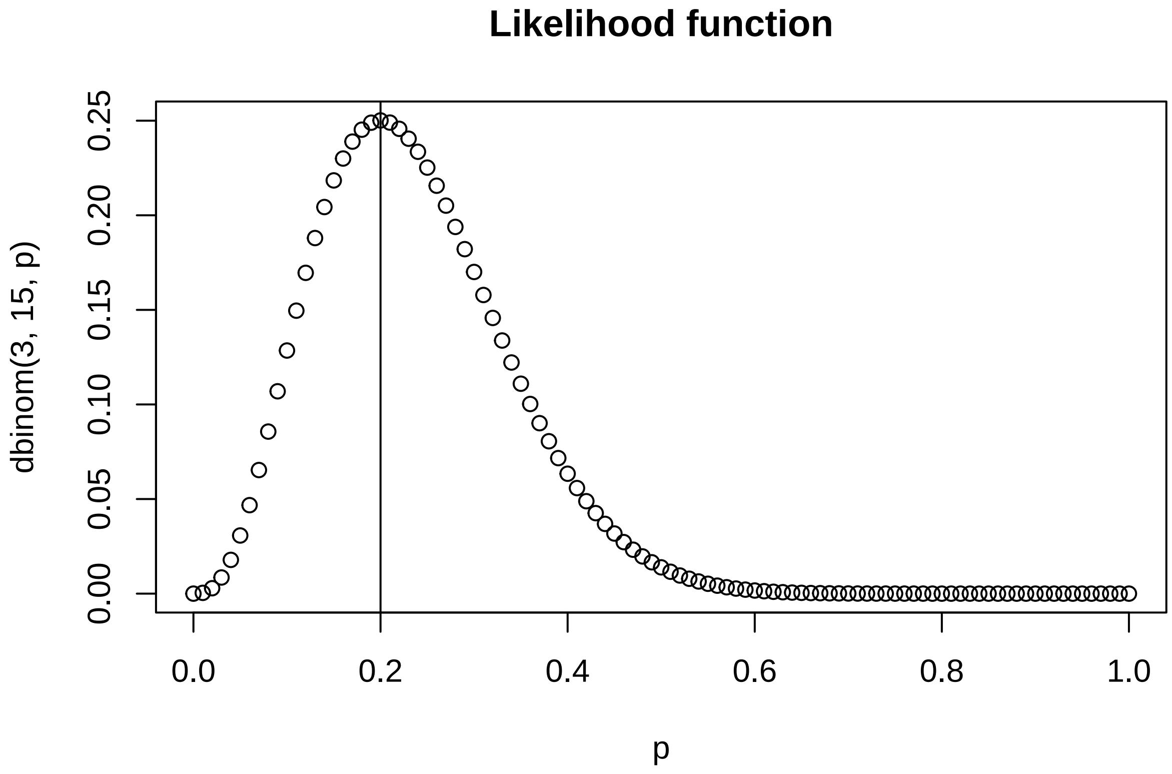
Here is a table of the total number of years he doesn't average over 10 field_goals_made against the total of those he does.

```
fg_table <- table(stephen.curry$field_goals_made<10)
names(fg_table) <- c("made", "not_made")
fg_table
```

```
##     made not_made
##        3       12
```

We plot this as

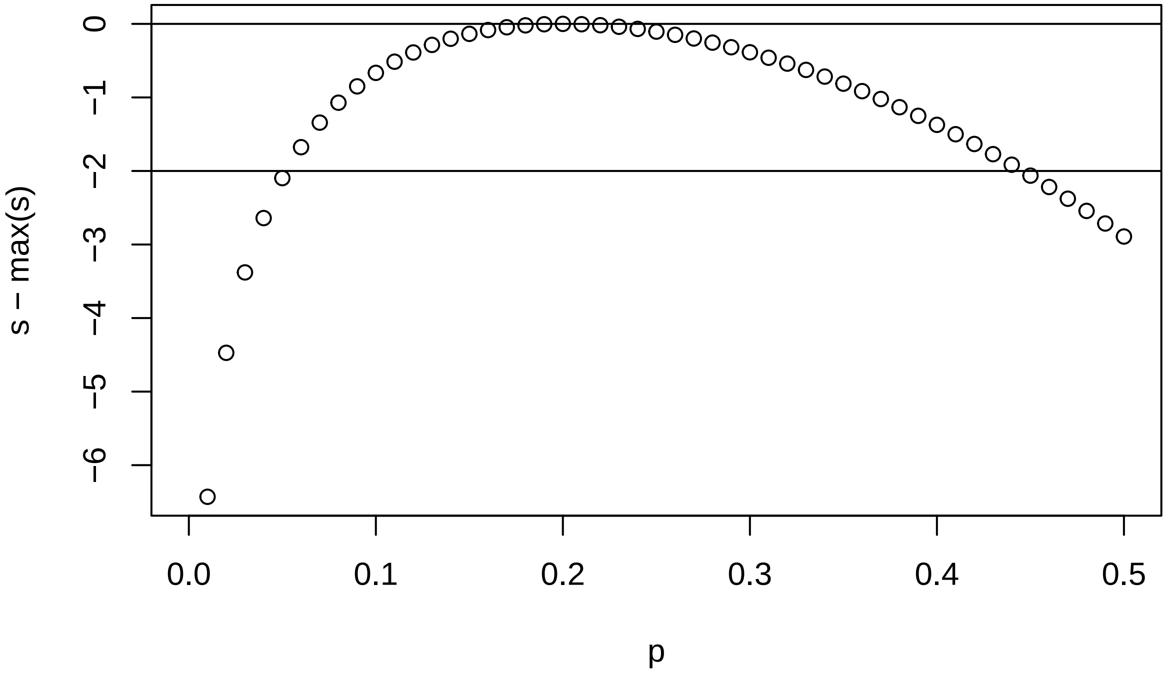
```
p <- seq(0, 1, 0.01)
plot(p, dbinom(3, 15, p), main = "Likelihood function")
abline(v=3/15)
```



The above plot shows the maximum likelihood of Stephen Curry making over 10 field goals in a given year is maximized at p=0.2

We plot the support function

```
p <- seq(0, 0.5, 0.01)
s <- dbinom(3, 15, p, log = TRUE)
plot(p, s - max(s))
abline(h=c(0, -2))
```



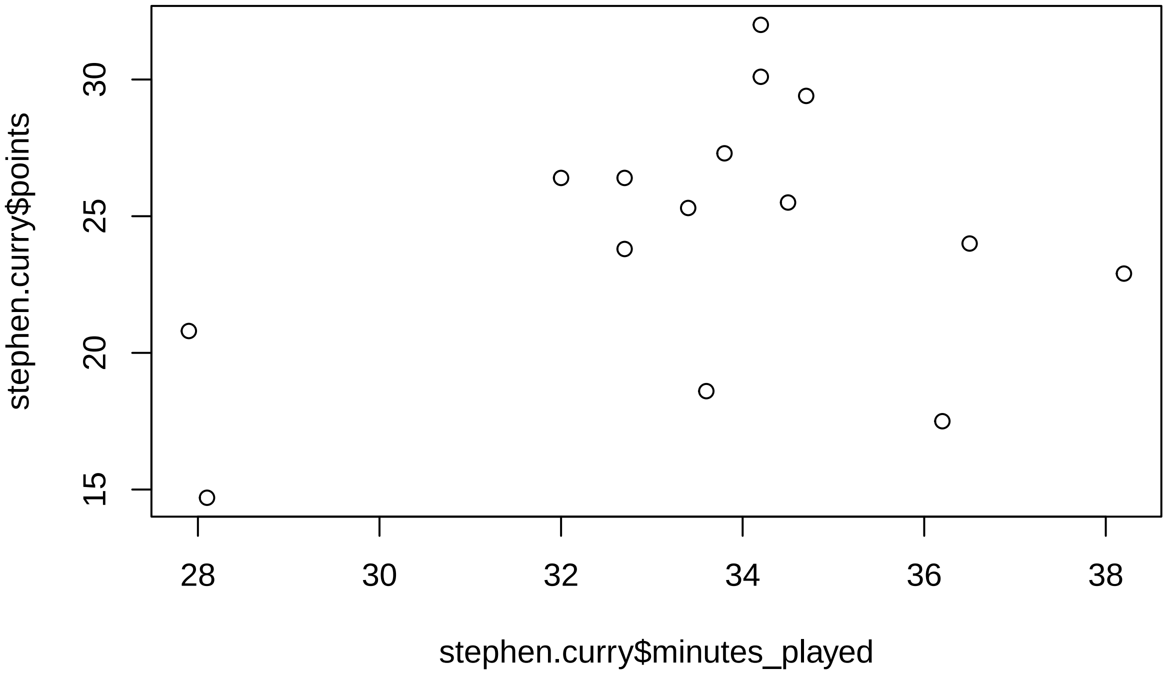
From the support curve we see that the maximum likelihood estimate is 0.2 and the credible interval is around 0.05 to 0.45.

Regression

We want to see what are the most likely factors leading to higher points scored. Some of the obvious ones that come to mind are time spent playing, and field goals made.

Regress points against minutes_played

```
plot(stephen.curry$points~stephen.curry$minutes_played)
```



```
lm(points~minutes_played, stephen.curry)
```

```
##
## Call:
## lm(formula = points ~ minutes_played, data = stephen.curry)
##
## Coefficients:
##    (Intercept)  minutes_played
##          5.916           0.549
```

```
summary(lm(points~minutes_played, stephen.curry))
```

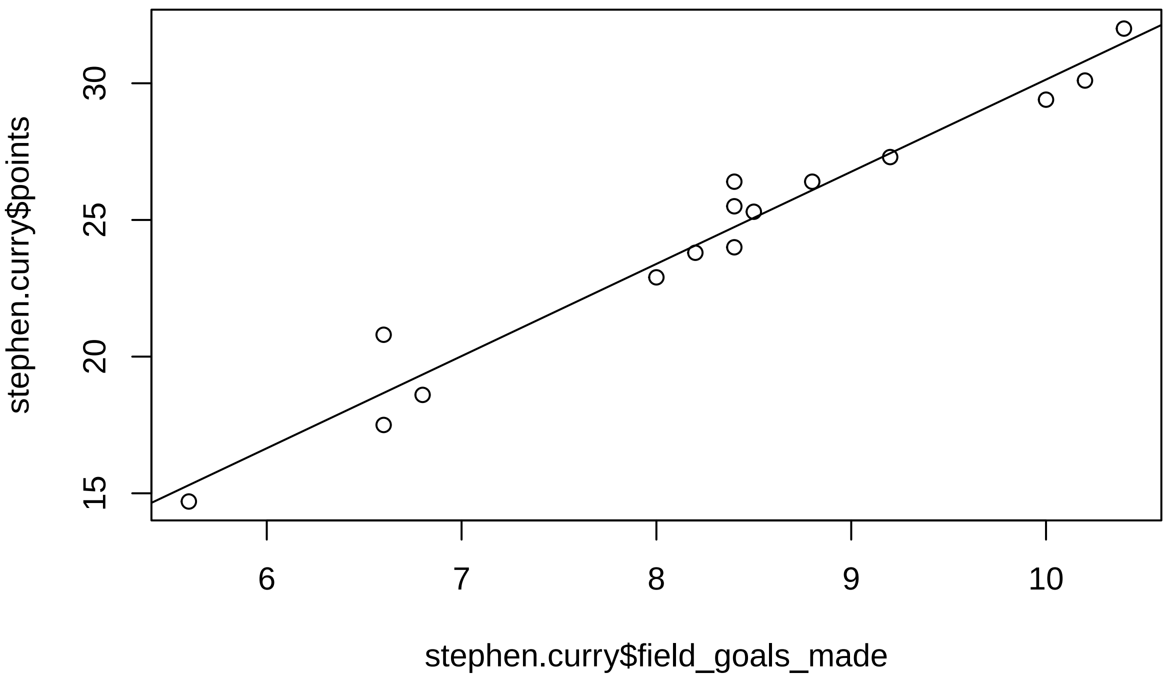
```
##
## Call:
## lm(formula = points ~ minutes_played, data = stephen.curry)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.288 -2.970  0.645  2.873  7.310
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.915     15.497   0.382    0.709
## minutes_played    0.549      0.461   1.191    0.255
```

```
##
## Residual standard error: 4.745 on 13 degrees of freedom
## Multiple R-squared:  0.09836,    Adjusted R-squared:  0.02901
## F-statistic: 1.418 on 1 and 13 DF,  p-value: 0.255
```

From the scatter plot we see that there isn't a clear linear relationship between the minutes played and points scored. The correlation coefficient also confirms this. In this case, it is very low at 0.02901. Lastly, we get a high p-value of 0.255, which is greater than 0.05. We can't reliably use the average minutes played by Stephen Curry to predict the number of points he'll score using this data set.

Regress points against field_goals_made

```
plot(stephen.curry$points~stephen.curry$field_goals_made)
abline(lm(lm(points~field_goals_made, stephen.curry)))
```



```
lm(points~field_goals_made, stephen.curry)
```

```
##
## Call:
## lm(formula = points ~ field_goals_made, data = stephen.curry)
##
## Coefficients:
##      (Intercept)  field_goals_made
##           -3.586             3.372
```

```
summary(lm(points~field_goals_made, stephen.curry))
```

```
##
## Call:
## lm(formula = points ~ field_goals_made, data = stephen.curry)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1705 -0.7233 -0.2660  0.4128  2.1295
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.5862     1.5681  -2.287   0.0396 *
## field_goals_made   3.3722     0.1871  18.028  1.4e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.98 on 13 degrees of freedom
## Multiple R-squared:  0.9615, Adjusted R-squared:  0.9586
## F-statistic:   325 on 1 and 13 DF,  p-value: 1.402e-10
```

Unlike the previous model, this one shows there is a positive relationship between field_goals_made and points scored. It's rather obvious in sports that the more you score, the more points you get. It also evident in the chart.

There is a correlation coefficient of 0.9586 showing the degree to which the field_goals_made explain the points scored. There is also a very low p-value of 1.402e-10 showing that the effect of field_goals_made on points is statistically significant.

Multiple regression of points against field_goals_made and field_goals_attempted

```
lm(points~field_goals_made+field_goals_attempted, stephen.curry)
```

```
##
## Call:
## lm(formula = points ~ field_goals_made + field_goals_attempted,
##     data = stephen.curry)
##
## Coefficients:
##           (Intercept)       field_goals_made  field_goals_attempted
##               -4.9663                 2.3360                 0.5657
```

```
summary(lm(points~field_goals_made + field_goals_attempted, stephen.curry))
```

```
##
## Call:
## lm(formula = points ~ field_goals_made + field_goals_attempted,
##     data = stephen.curry)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.0413 -0.3867 -0.1887  0.1211  2.1830
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -4.9663     1.5090  -3.291  0.00645 **
## field_goals_made       2.3360     0.4952   4.717  0.00050 ***
## field_goals_attempted  0.5657     0.2551   2.217  0.04666 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8591 on 12 degrees of freedom
## Multiple R-squared:  0.9727, Adjusted R-squared:  0.9682
## F-statistic: 213.9 on 2 and 12 DF,  p-value: 4.125e-10
```

From the p-values we see that field_goals_made and field_goals_attempted are relevant to the outcome of points.

The model is quite good. What if Curry made 1 field goal attempt which resulted in 1 field goal made, what would be the predicted points? From our model:

$points = -4.9663 + 2.3360 field\_goals\_made + 0.5657 field\_goals\_attempted$

```
-4.9663 + 2.3360 * 1 + 0.5657 * 1
```

```
## [1] -2.0646
```

The model says that, such an event would result in him having -2.0646 points. **Impossible.**

We can introduce log to our variables to make the model more realistic.

```
lm(log(points) ~ log(field_goals_made) + log(field_goals_attempted))
```

```
##
## Call:
## lm(formula = log(points) ~ log(field_goals_made) + log(field_goals_attempted))
##
## Coefficients:
##                (Intercept)       log(field_goals_made)
##                     0.1950                      0.7329
## log(field_goals_attempted)
##                     0.5034
```

```
summary(lm(log(points) ~ log(field_goals_made) + log(field_goals_attempted), stephen.curry))
```

```
##
## Call:
## lm(formula = log(points) ~ log(field_goals_made) + log(field_goals_attempted),
##     data = stephen.curry)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055227 -0.013749 -0.005867  0.004999  0.095076
```

```
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.1950     0.2156   0.905  0.38349
## log(field_goals_made)       0.7329     0.1731   4.233  0.00116 **
## log(field_goals_attempted)  0.5034     0.1828   2.754  0.01748 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.03787 on 12 degrees of freedom
## Multiple R-squared:  0.9733, Adjusted R-squared:  0.9688
## F-statistic: 218.3 on 2 and 12 DF,  p-value: 3.663e-10
```

The new model has a correlation coefficient of 0.9688 which is slightly higher than the previous linear model. It can be written as:

$log(points) = 0.1950 + 0.7329 log(field\_goals\_made) + 0.5034 log(field\_goals\_attempted)$

Using the same scenario to test the model:

```
log_points <- 0.1950 + 0.7329*log(1) + 0.5034*log(1)
predicted_points <- 10**log_points # convert from log
predicted_points
```

```
## [1] 1.566751
```

The new model, which uses log, is closer to reality. If a basketball player makes a field goal, they are usually awarded 2 points, while the model says they would get 1.566751 points.

Fisher’s exact test

Stephen Curry, while playing for the Golden State Warriors, has won 4 championships: 2014-15, 2016-17, 2017-18, 2021-22.

We will use Fisher’s exact test to see whether there is evidence that the team is more likely to win a championship when he plays over 60 matches in the regular season.

Null: Stephen Curry is likely to win or lose regardless of the matches played

Alternative: Matches played have an effect on whether Stephen Curr wins or loses a championship

Here is how the table looks like:

```
fisher_table <- matrix(c(3, 8, 1, 2), 2,2, byrow = TRUE)
dimnames(fisher_table) <- list(played_matches=c("over_60", "under_60"), championship=c("won", "lost"))
fisher_table
```

```
##                championship
## played_matches won lost
##       over_60    3    8
##       under_60   1    2
```

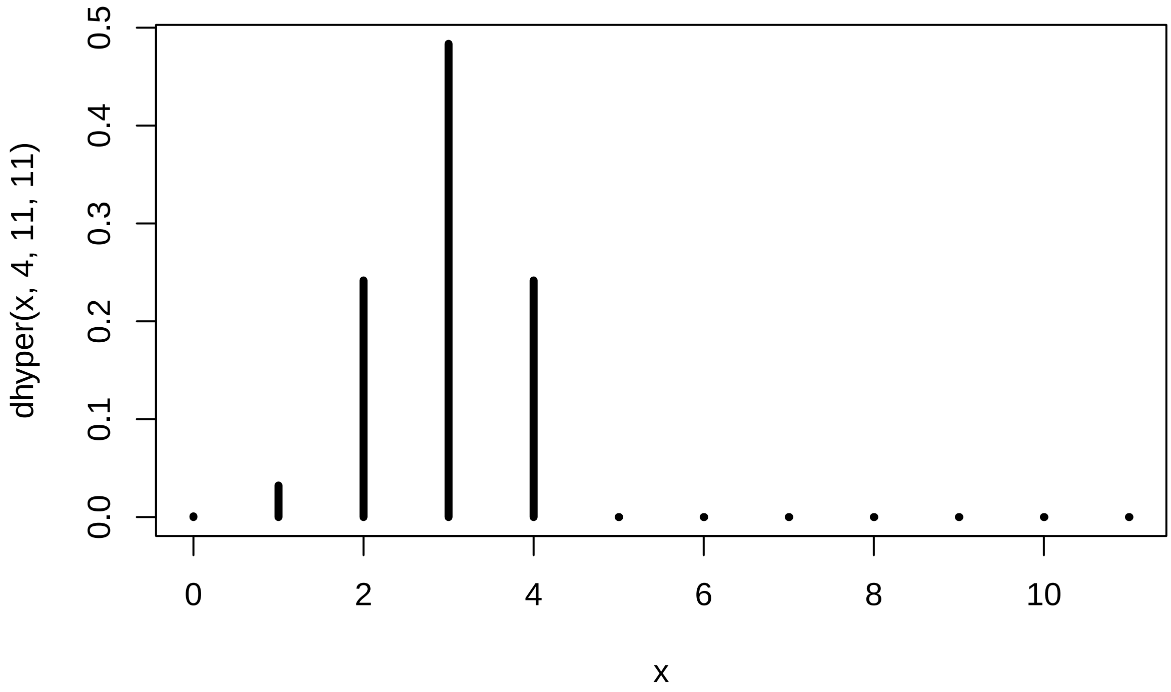
```
dhyper(3,4,11, 11)
```

```
## [1] 0.4835165
```

```
# using fisher test
fisher.test(fisher_table, alternative = "l")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  fisher_table
## p-value = 0.6703
## alternative hypothesis: true odds ratio is less than 1
## 95 percent confidence interval:
##   0.00000 28.83028
## sample estimates:
## odds ratio
##  0.7661432
```

```
x <- 0:11 # wins
plot(x,dhyper(x,4,11,11),type='h',lwd=4)
```



The p-values are significantly greater than 0.05: 0.6703. Therefore, we lack significant evidence to reject the null hypothesis. It stands that Stephen Curry's championship run is independent of whether he plays over or under 60 matches in a regular season.

Conclusion

This has been an interesting exercise for me as a basketball fan. It gave me the opportunity to scrutinize some of the assumptions and foregone conclusions fans make.

I found it shocking that minutes played have little predictive ability on the average points. My assumption has always been that the more time a player gets, the more they can score. However, the scatter plot shows otherwise. Sometimes, players can be quite effective even when they play a short amount of time. Other times, they could be heavily guarded which could result in fewer points though they played the entire match.

It seems I have been over-estimating the impact of Stephen Curry on the success of his team. It's easy to think that the more a star player is on the court, the more likely the team is to win. The Fisher's exact test has done a good job of proving there isn't enough evidence to make this statement.

Overall, I now have a deep appreciation for statistics. I've gained an understanding of the value of data-driven opinions and independent verification of these opinions. It is also scary to think that there exists ideas that I hold to be true but are in actuality wrong and at best, unjustified.