Lab 6

Kyle Mukire 505832075

2022-05-15

Contents

Examples **1**

Binomial Model 1

Normal Model 3

The Normal Approximation to The Binomial Model 4

MLB 2021 Season — Are Some Teams Actually Better than Others? 7

Your Work **10**

```
## Date last run: 2022-05-15

## Hello World!
```

Note that included data sets were made by processing data obtained from MLB and the NHL.

Examples

Binomial Model

Imagine a baseball team, call them the Chattanooga P-Values. This upcoming season, this imaginary team will play 40 home games, and, for each home game, will have the same probability of winning, 70%.

The binomial distribution can be used here to model the number of season home game wins.

```
xdomain <- I(0:40)

hg_win_prop <- dbinom(xdomain, size=40, prob=0.70)
hg_win_prop
```

```
##  [1] 1.215767e-21 1.134715e-19 5.162955e-18 1.525940e-16 3.293487e-15 5.533059e-14
##  [7] 7.531108e-13 8.535256e-12 8.215184e-11 6.815560e-10 4.929921e-09 3.137223e-08
## [13] 1.769045e-07 8.890585e-07 4.000763e-06 1.618087e-05 5.899274e-05 1.943290e-04
## [19] 5.793884e-04 1.565365e-03 3.835144e-03 8.522543e-03 1.717422e-02 3.136161e-02
## [25] 5.183378e-02 7.740510e-02 1.041992e-01 1.260681e-01 1.365738e-01 1.318644e-01
## [31] 1.128173e-01 8.491625e-02 5.572629e-02 3.152194e-02 1.514289e-02 6.057157e-03
## [37] 1.962968e-03 4.951630e-04 9.121424e-05 1.091452e-05 6.366806e-07
```

```
par(mfrow=c(1,1), lend=1, cex=0.65)
plot(xdomain, hg_win_prop, type="h", lwd=3,
     xlab="Number of Home Game Wins", ylab="Probability")
```

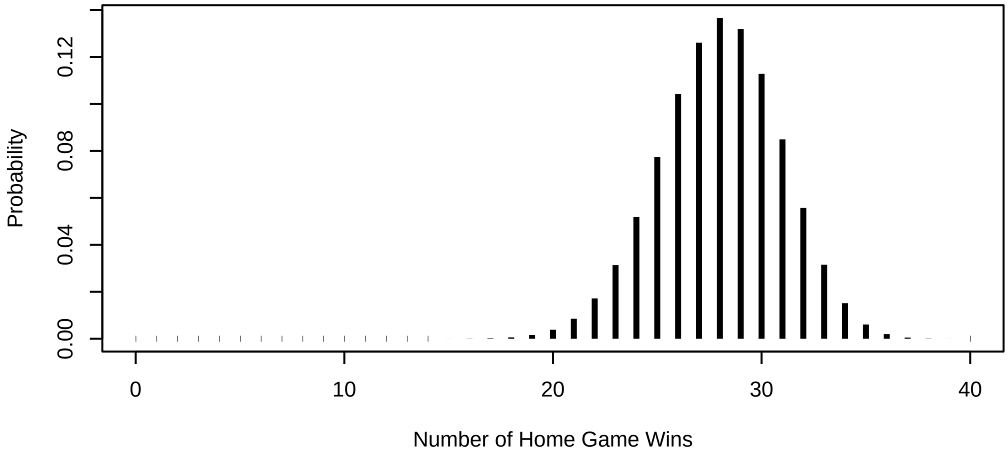


Figure 1: Distribution of Home Game Wins.

The expectation by definition of binomial PMF is $n \cdot p = 28$

Using the general definition for the expectation of a PMF, $\sum_i X_i \cdot \Pr[X_i] = 28$.

Same answer.

What's the probability team will win 30 or more home games?

```
sum( dbinom(I(30:40), size=40, prob=0.70) )
```

```
## [1] 0.3087427
```

Using the cumulative R function:

```
1 - pbinom(29, size=40, prob=0.70)
```

```
## [1] 0.3087427
```

What's the probability team will lose half or more of their home games?

```
sum( dbinom(I(0:20), size=40, prob=0.70) )
```

```
## [1] 0.006254504
```

Using the cumulative R function

```
pbinom(20, size=40, prob=0.70)
```

```
## [1] 0.006254504
```

Normal Model

The normal, or Gaussian probability distribution is a PDF — its domain is over the continuum of the real numbers.

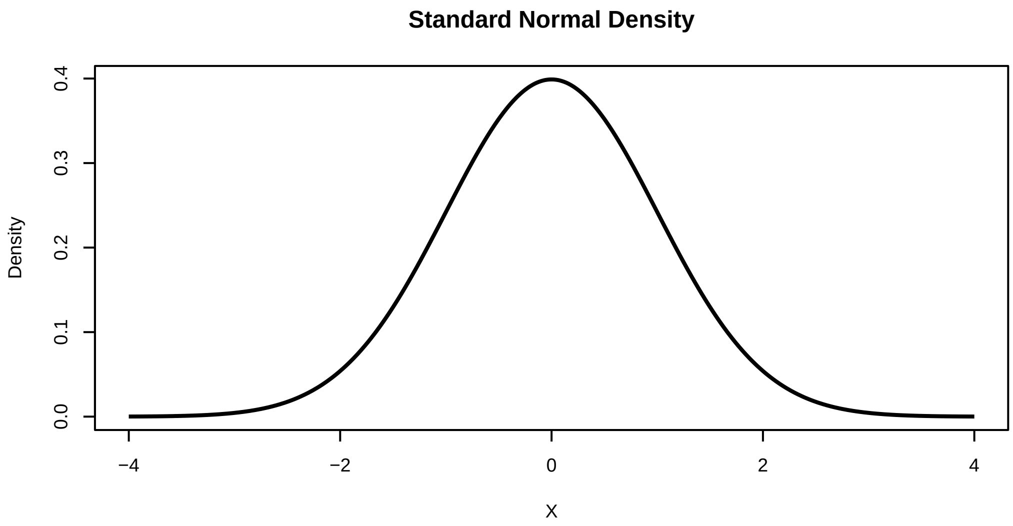
A normal distribution is uniquely defined by two parameters, the mean (the expectation) and the standard deviation (or the variance).

We'll use a path to show density.

```
xdom <- seq(-4, 4, length=500)

xdensity <- dnorm(xdom, 0, 1)

par(mfrow=c(1,1), lend=1, cex=0.60)
plot(xdom, xdensity, type="l", lwd=2,
     xlab="X", ylab="Density", main="Standard Normal Density")
```



The Normal Approximation to The Binomial Model

The normal model is rather unique as it is the limiting distribution of many estimators, along with other distributions.

The normal model can be used to model the binomial model.

Let's illustrate an example.

If X is normally distributed, the probability that X will be one or more standard deviation greater than the mean is

```
1 - pnorm(1, 0, 1)
```

```
## [1] 0.1586553
```

For increasing binomial sample size (i.e., number of trials), were going to calculate the probability of each respective binomial random variable being more than one standard deviation from the mean.

```
p_success <- 0.5

xtrialsTry <- seq(5, 2000, by=5)

pout_vec <- numeric(length(xtrialsTry))

for(i in 1:length(xtrialsTry)) {
  xthis_numTrials <- xtrialsTry[ i ]
  xthis_mean <- p_success * xthis_numTrials
  xthis_sd <- sqrt( (1 - p_success) * p_success * xthis_numTrials )
  xdom <- I(0:xthis_numTrials)

  xdom_prob <- xdom[ xdom > (xthis_mean + 1 * xthis_sd) ]
  pout_vec[ i ] <- sum(dbinom(xdom_prob, size=xthis_numTrials, prob=p_success))

}
```

```
par(mfrow=c(1,1), lend=1, cex=0.60)
plot(xtrialsTry, pout_vec, type="l", lwd=1,
     xlab="Sample size", ylab="Probability X > mu + 1*sd",
     main="Normal Approximation of Binomial Model, Example")

abline(h=1 - pnorm(1, 0, 1), lwd=2, col="#22BB5577")
```

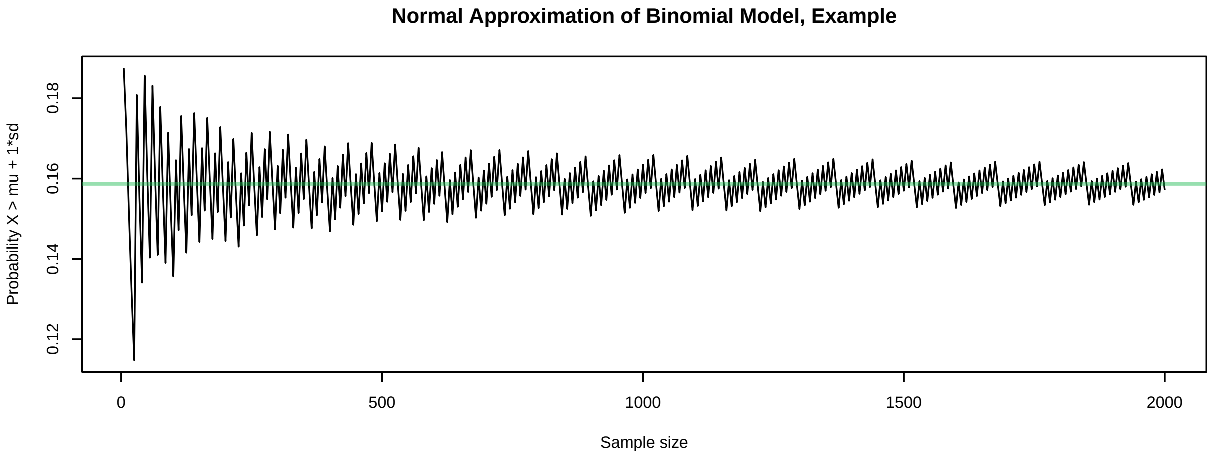


Figure 2: Binomial probability that number of successes will be greater than 1 standard deviation above the mean for increasing number of binomial trials. Grean line shows probability under normal distribution

To make the convergence more pronounced:

```
p_success <- 0.5

xtrialsTry <- 1 * 2^(I(2:15))

pout_vec <- numeric(length(xtrialsTry))

for(i in 1:length(xtrialsTry)) {
  xthis_numTrials <- xtrialsTry[ i ]
  xthis_mean <- p_success * xthis_numTrials
  xthis_sd <- sqrt( (1 - p_success) * p_success * xthis_numTrials )
  xdom <- I(0:xthis_numTrials)

  xdom_prob <- xdom[ xdom > (xthis_mean + 1 * xthis_sd) ]
  pout_vec[ i ] <- sum(dbinom(xdom_prob, size=xthis_numTrials, prob=p_success))

}
```

```
par(mfrow=c(1,1), lend=1, cex=0.60)
plot(xtrialsTry, pout_vec, type="l", lwd=1,
     xlab="Sample size", ylab="Probability X > mu + 1*sd",
     main="Normal Approximation of Binomial Model, Example")

abline(h=1 - pnorm(1, 0, 1), lwd=2, col="#22BB5577")
```

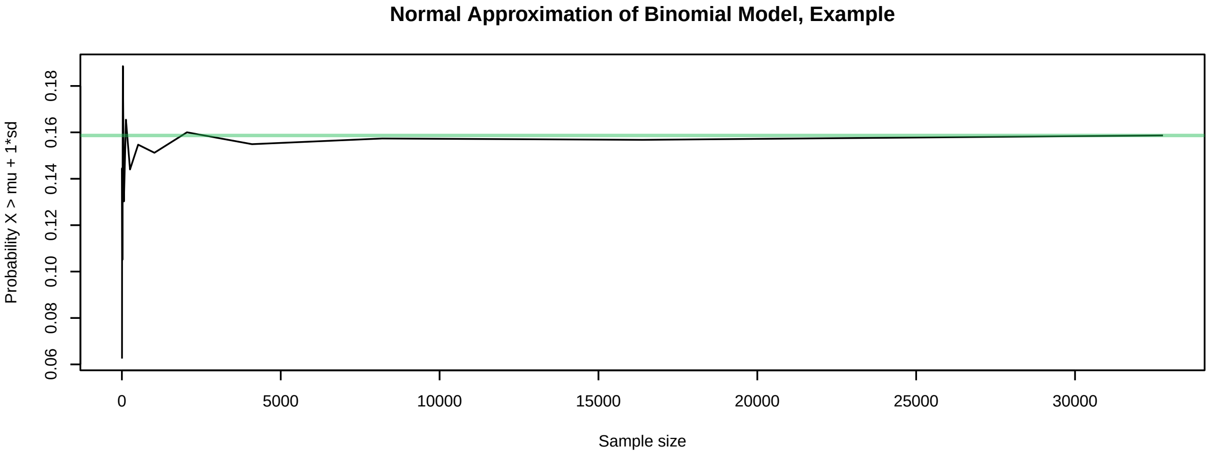


Figure 3: Binomial probability that number of successes will be greater than 1 standard deviation above the mean for increasing number of binomial trials. Grean line shows probability under normal distribution

MLB 2021 Season — Are Some Teams Actually Better than Others?

Suppose a friend says they've been to many MLB games, and they believe that there's no difference between the teams, the outcome of the game is pure chance, and that the probability the home team will win is always 50%.

The experiment that follows uses empiric probabilities, and requires some creative thinking.

```
## Read in our data
xdf <- read.csv("MLB_team_2021.csv", header=TRUE)
head(xdf, n=6)
```

```
##       date gameID                  team VorH bat_runs bat_homeRuns bat_strikeOuts
## 1 20210401 634615   Los Angeles Dodgers    V        5            0              6
## 2 20210401 634615      Colorado Rockies    H        8            0              4
## 3 20210401 634618  Arizona Diamondbacks    V        7            4             12
## 4 20210401 634618      San Diego Padres    H        8            2             10
## 5 20210401 634622        Atlanta Braves    V        2            1             10
## 6 20210401 634622 Philadelphia Phillies    H        3            0             13
##   bat_baseOnBalls pitch_runs pitch_homeRuns pitch_strikeOuts pitch_baseOnBalls
## 1               8          8              0                4                 3
## 2               3          5              0                6                 8
## 3               1          8              2               10                 5
## 4               5          7              4               12                 1
## 5               2          3              0               13                 4
## 6               4          2              1               10                 2
```

Let's look at the distribution of total home game wins for each of the thirty MLB teams.

```
WorL <- xdf[ , "bat_runs"] > xdf[ , "pitch_runs"]

xdf_HT <- xdf[ xdf[ , "VorH"] == "H", ]
dim(xdf_HT)
```

```
## [1] 2429   12
```

```
xWinTH <- WorL[ xdf[ , "VorH"] == "H" ]

xagg <- aggregate(xWinTH, by=list(xdf_HT[ , "team"]), sum)

xnumberHGwins <- xagg$x

xbrks <- seq(21.5, 65.5, by=4)
xbrks
```

```
##  [1] 21.5 25.5 29.5 33.5 37.5 41.5 45.5 49.5 53.5 57.5 61.5 65.5
```

```
par(cex=0.65)
hist(xnumberHGwins, breaks=xbrks, main="Total Home Game Wins for Each Team over MLB 2021 Season")
```

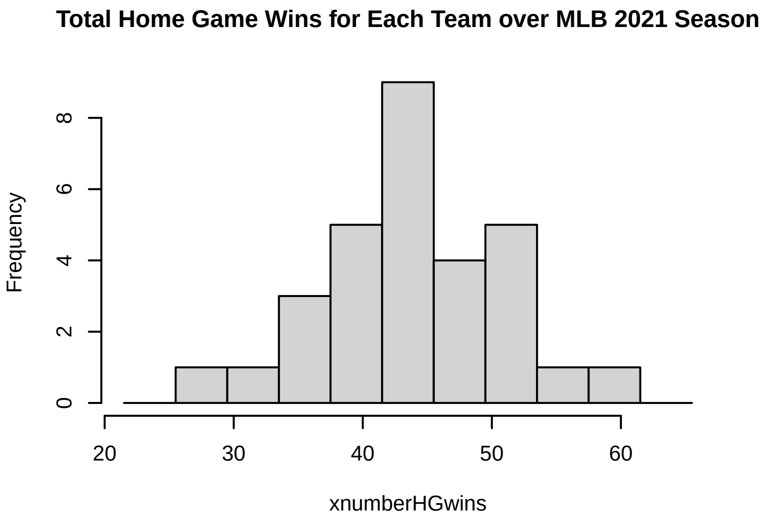


Figure 4: Total Home Games Wins

If all teams are actually the same, we would not expect to see much variation in the number of home game wins between the 30 teams.

What is the observed standard deviation for the 2021 Season?

It is 6.6079889.

So, let's use the binomial model to simulate our friend's claim.

```
set.seed(777)

nn <- 5000 ### number of simulations

#xsim_range <- integer(nn)
#xsim_max <- integer(nn)
xsim_sd <- integer(nn)
#xsim_IQR <- integer(nn)

for(j in 1:nn) {

    xsim_Win <- rbinom(length(xWinTH), 1, prob=1/2)

    xagg_sim <- aggregate(xsim_Win, by=list(xdf_HT[ , "team"]), sum)
    #xsim_range[j] <- max(xagg_sim[ , "x"]) - min(xagg_sim[ , "x"])
    #xsim_max[j] <- max(xagg_sim[ , "x"])
    xsim_sd[j] <- sd(xagg_sim[ , "x"])
    #xsim_IQR[j] <- IQR(xagg_sim[ , "x"])
}
```

```
par(mfrow=c(1,1), cex=0.65)

hist(xsim_sd, xlim=c(2, 9))
abline(v=sd(xnumberHGwins), lwd=2, col="#33AA33")
```

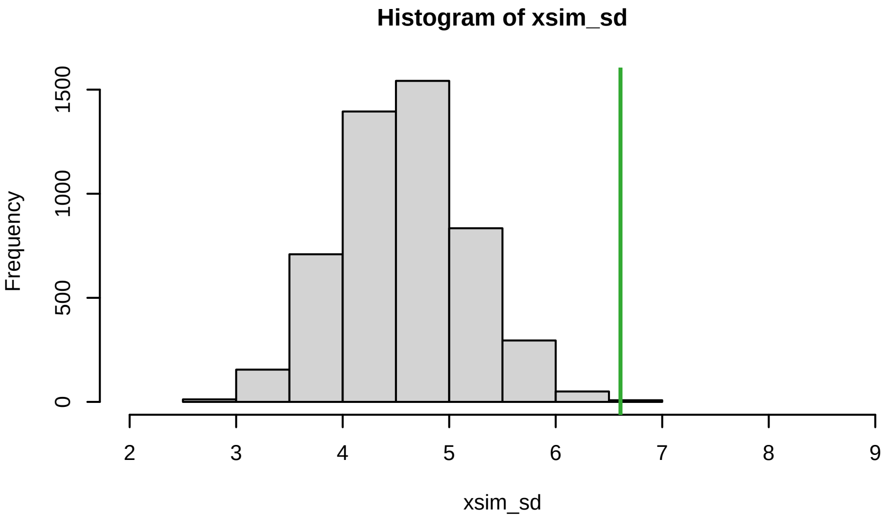


Figure 5: Simulation: Histogram of standard deviation of total home games won assuming our friend is correct

```
sum(xsim_sd >= sd(xagg[ , "x"])) / nn
```

```
## [1] 4e-04
```

Your Work

Make sure to edit the “author” information in the YAML header near the top to include your name and UID.

Complete/answer the following.

1 — Suppose the Chattanooga P-Values play only 30 home games. Keeping the probability of win at 70%, what is the probability they will lose half or more of their home games? How does this compare with the example we looked at above where they play 40 home games? Comment on the difference.

```
sum( dbinom(I(0:15), size=30, prob=.70) )
```

```
## [1] 0.01693731
```

The probability of 40 home game is 0.006254504 which is smaller than 0.01693731. which is the probability that they will lose half or more of their home games. This means that when we have more home games we are more likely to close to be close to the expectation, such as the center.

2 — Consider the example where we illustrated the binomial probabilities converging to that produced by the normal distribution. Run this experiment yourself, except change the following: Have the binomial probability of success be only 10% (instead of the 50% we used above), and also look at the probability our respective random variable will be more than 2 standard deviations above the mean. Comment on your results.

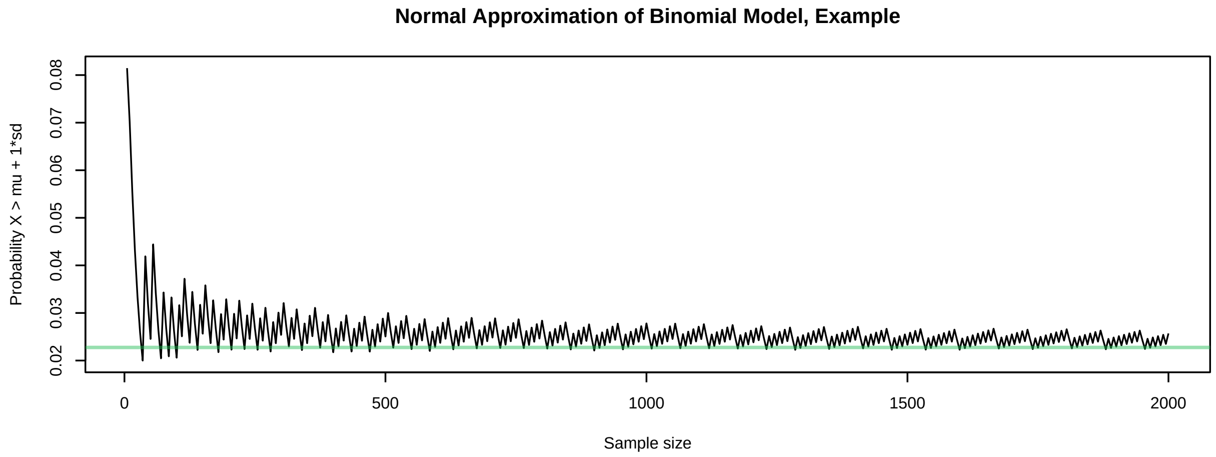
```
1 - pnorm(2, 0, 1)
```

```
## [1] 0.02275013
```

```
 p_success <- 0.1
xtrialsTry <- seq(5, 2000, by=5)
pout_vec <- numeric(length(xtrialsTry))
for(i in 1:length(xtrialsTry)) {
xthis_numTrials <- xtrialsTry[ i ]
xthis_mean <- p_success * xthis_numTrials #mean= n*P
xthis_sd <- sqrt( (1 - p_success) * p_success * xthis_numTrials ) #sd=sqrt(n*p*(1-p)) xdom <- I(0:xthis
  xdom_prob <- xdom[ xdom > (xthis_mean + 2 * xthis_sd) ]
  pout_vec[ i ] <- sum(dbinom(xdom_prob, size=xthis_numTrials, prob=p_success))
}
```

```
par(mfrow=c(1,1), lend=1, cex=0.60)
plot(xtrialsTry, pout_vec, type="l", lwd=1,
     xlab="Sample size", ylab="Probability X > mu + 1*sd",
     main="Normal Approximation of Binomial Model, Example")

abline(h=1 - pnorm(2, 0, 1), lwd=2, col="#22BB5577")
```



In increasing the binomial, it is getting closer and closer to the normal distribution.

3 — Interpret the simulated MLB results from the above Examples Section.

Our results are not from the distribution of the results simulated from the proposed binomial distribution. This is due to the green vertical line in the Histogram of xsim_sd being far from the distribution.

4 — Perform the same analysis we looked at with the MLB data with the NHL data. Are the results more or less convincing? Why?

```r
### here's a head start for you
xdf <- read.table( "NHL_20202021_game.tsv", sep="\t", header=TRUE )
tail(xdf)
```

```
##         date   season            startTime              endTime status VTabbr HTabbr
## 863 20210513 20202021 2021-05-14T00:00:00Z 2021-05-14T02:27:32Z  Final    MIN    STL
## 864 20210514 20202021 2021-05-15T00:00:00Z 2021-05-15T02:21:48Z  Final    TOR    WPG
## 865 20210515 20202021 2021-05-15T19:30:00Z 2021-05-15T21:53:17Z  Final    VAN    EDM
## 866 20210516 20202021 2021-05-17T02:30:00Z 2021-05-17T05:12:09Z  Final    CGY    VAN
## 867 20210518 20202021 2021-05-18T20:00:00Z 2021-05-18T22:39:15Z  Final    CGY    VAN
## 868 20210519 20202021 2021-05-19T19:30:00Z 2021-05-19T22:04:59Z  Final    VAN    CGY
##                      VT                HT periods VTgoals HTgoals VTfinal HTfinal
## 863      Minnesota Wild   St. Louis Blues       3       3       7       3       7
## 864 Toronto Maple Leafs     Winnipeg Jets       3       2       4       2       4
## 865   Vancouver Canucks   Edmonton Oilers       3       4       1       4       1
## 866      Calgary Flames Vancouver Canucks       4       6       5       6       5
## 867      Calgary Flames Vancouver Canucks       3       2       4       2       4
## 868   Vancouver Canucks    Calgary Flames       3       2       6       2       6
##                                                                           officials
## 863                       Dean Morton;Peter MacDougall;Jesse Marquis;Bryan Pancich
## 864                 Chris Schlenker;Graham Skilliter;Scott Cherrey;David Brisebois
## 865                      Kendrick Nicholson;Brad Meier;Derek Nansen;Kiel Murchison
## 866 Chris Schlenker;Graham Skilliter;Kendrick Nicholson;Derek Nansen;Kiel Murchison
## 867                 Chris Schlenker;Kendrick Nicholson;Derek Nansen;Kiel Murchison
## 868                 Chris Schlenker;Kendrick Nicholson;Derek Nansen;Kiel Murchison
##                                  official_type
## 863         Referee;Referee;Linesman;Linesman
## 864         Referee;Referee;Linesman;Linesman
## 865         Referee;Referee;Linesman;Linesman
## 866 Referee;Referee;Referee;Linesman;Linesman
## 867         Referee;Referee;Linesman;Linesman
## 868         Referee;Referee;Linesman;Linesman
```

```
dim(xdf)
```

```
## [1] 868  16
```

```
N <- nrow(xdf)

WorL <- xdf[ , "HTfinal"] > xdf[ , "VTfinal"]

sum(xdf[ , "HTfinal"] == xdf[ , "VTfinal"]) ### no ties
```

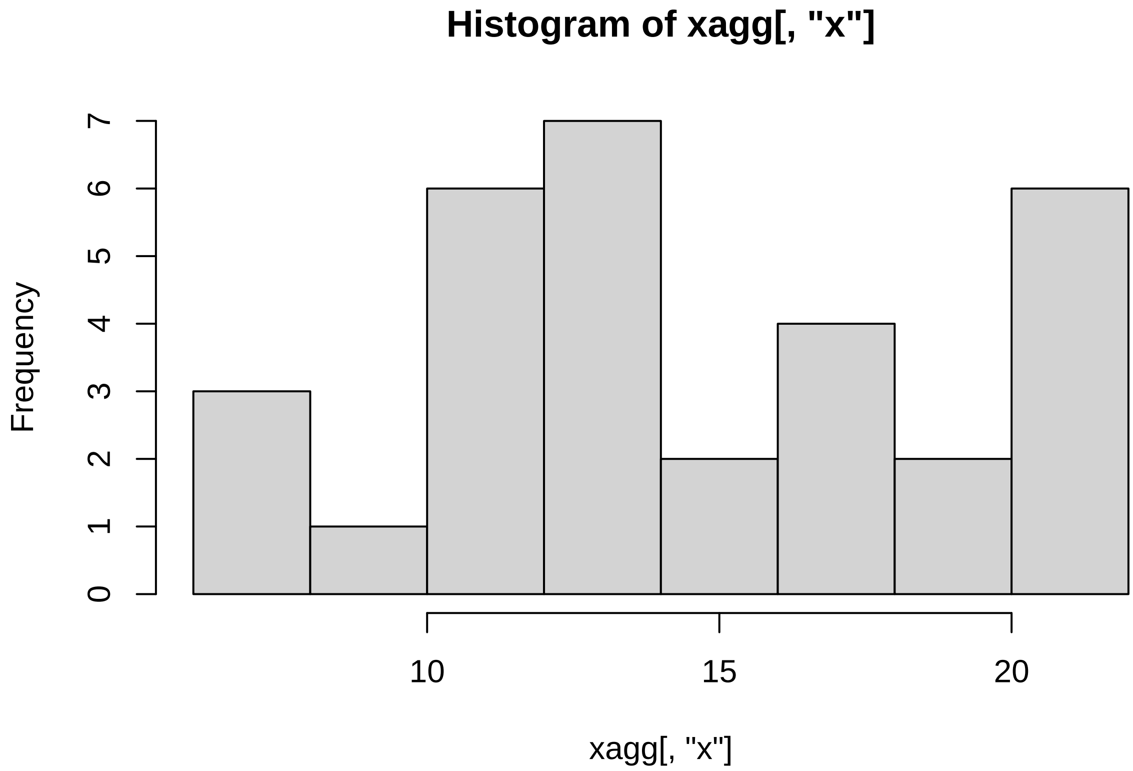
```
## [1] 0
```

```
xagg <- aggregate(WorL, by=list(xdf[ , "HT"]), sum)

sd(xagg[ ,"x"])
```

```
## [1] 4.632749
```

```
hist(xagg[ , 'x'])
```



```r
set.seed(777)
nn <- 5000 ### number of simulations
#xsim_range <- integer(nn) #xsim_max <- integer(nn) xsim_sd <- integer(nn) #xsim_IQR <- integer(nn)
for(j in 1:nn) {
xsim_Win <- rbinom(N, 1, prob=1/2)
xagg_sim <- aggregate(xsim_Win, by=list(xdf[ , "HT"]), sum) #xsim_range[j] <- max(xagg_sim[ , "x"]) - m
xsim_sd[j] <- sd(xagg_sim[ , "x"])
    #xsim_IQR[j] <- IQR(xagg_sim[ , "x"])
}
```

```r
par(mfrow=c(1,1), cex=0.65)
hist(xsim_sd, xlim=c(2, 9))
abline(v=sd(xagg[ ,"x"]), lwd=2, col="#33AA33")
```

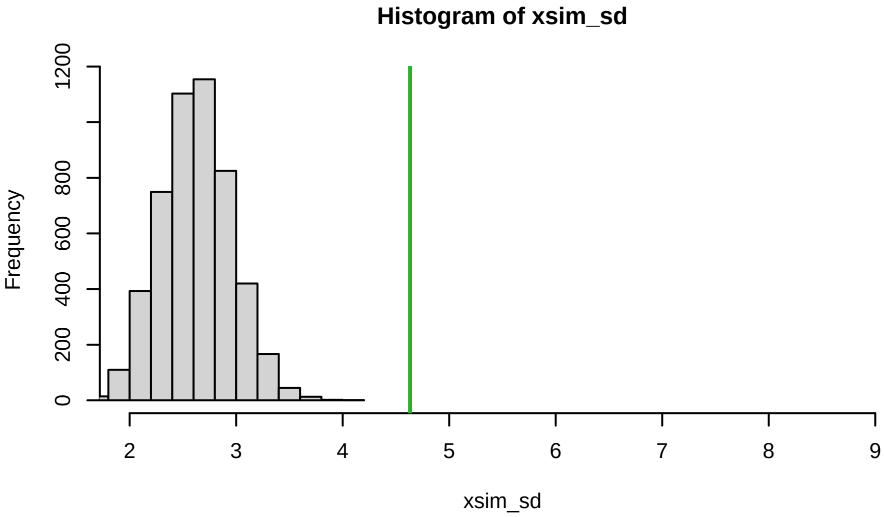


Figure 6: Simulation: Histogram of standard deviation of total home games won assuming our friend is correct

The results are more convincing because the green line strays even further from the distribution.