discrete probability calculations

Data Science 5620 — Deliverable 03

Probability Calculations

Rastko Stojsin

Many sports have series (particularly during playoffs), where two teams play each other some number of times to determine a winner. Series are intended to mitigate "accidental wins" and better determine the truly better team. Many variables go into determining the winner of a series. I will examine the relationship of series length and team strength in determining the winner of a series.

In order to conduct this analysis I have to define important variables. The variables that will be needed are as follows . . . — p: probability of team A winning a single head to head match up against team B — N: the number of games possible or maximum series length — win_series: this is calculated from the above two variables and is the probability that team A wins the series (in any number of possible games)

Below is the code I wrote to take the variables p and N and return win_series. I designed the code so in order to minimize the number of functions (as opposed to having functions for each series length). It is also able to handle a series of 1 game, for benchmark purposes.

```
# building the function to take in variables p and N and return win_series
calc_win_prob <- function(p, N) {
  # helper variables --- allow for variable manipulation within loop
  helper_01 <- 0
  helper_02 <- N
  # for series of 1
  if(N == 1)
  {
    helper_01 = p
  }
  # for series of > 1
  else{
  x_all <- for (i in 1:(N / 2))
  {
    # probability team wins in any but last possible game
    helper_01 = dbinom((N / 2) - 0.5, size = helper_02 - 1, prob = p) * p + helper_01
    helper_02 = helper_02 - 1
  }
  # probability team wins in last possible game
  helper_01 = dbinom((N / 2) + 0.5, size = (N / 2) + 0.5, prob = p) + helper_01
  return(helper_01)
  }
}
```

Now that I have the function to calculate win_series, I want to have a grid that can be used to provide many combinations of N and p.

This grid should be edited if you want to look at win_series probability for combinations of N and p not included in my analysis.

```r
# build and populate grid with N and p combinations that I want to analyse
odds <- c(1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21)
full_grid <- expand.grid(N = odds
                         , p = seq(0.5, 1, by = 0.001))
```

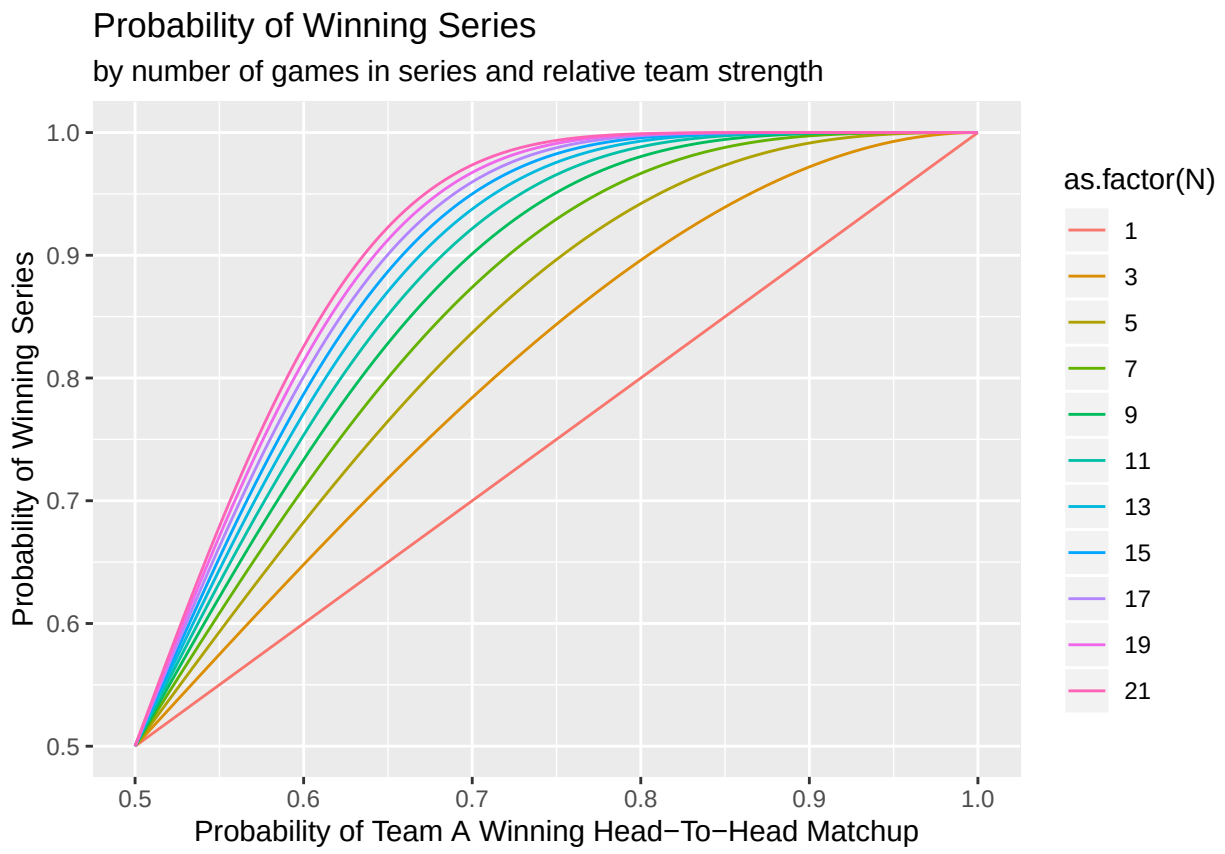
Now that I have the values of p and N, and the formula to calculate wins_series, I will use the function to create and populate a win_series column in the grid.

```r
# calculate the win_series from each combinations of N and p and insert into grid
for(i in 1:nrow(full_grid)) {
  full_grid[i, c("win_series")] <-
    calc_win_prob(full_grid[i, "p"], full_grid[i, "N"])
}
```

If we plot the probability of winning a series with compared to relative team strength and for series of lengths 1 to 21 (odds only), we can see that the more games in a series, the more likely the better team is to win. This makes sense intuitively as they have more opportunities to prove themselves and "flukes" are mitigated. Also obviously, the better team A is compared to B, the higher their chance of winning the series.

It is interesting to not that the improvement for the better team when series length increases becomes weaker and weaker the higher the series length gets — i.e. the better teams chances improve much more when moving from a 5 series finals to a 7 series one, than they do changing from a 19 series finals to a 21.

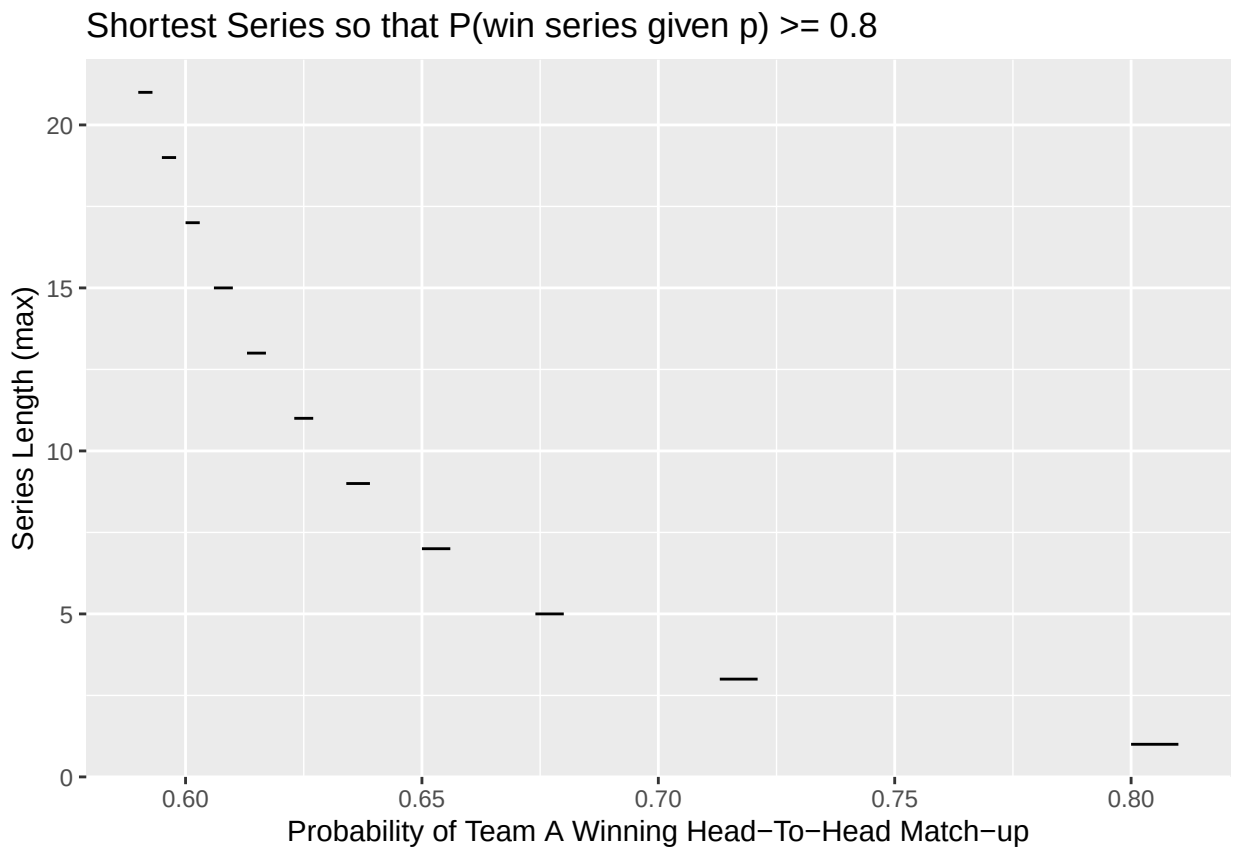
```r
ggplot(
  data = full_grid,
  mapping = aes(
    x = p,
    y = win_series,
    group = interaction(N),
    color = as.factor(N)
  )) +
  geom_line() +
  xlab("Probability of Team A Winning Head-To-Head Matchup") +
  ylab("Probability of Winning Series") +
  labs(title = "Probability of Winning Series",
       subtitle = "by number of games in series and relative team strength")
```



In the graph below we look at how good team A needs to be compared to team B in order to guarantee a probability of winning the series of 0.8 by the series length. The one series is a nice benchmark here and it shows us that to win a 1 game series 0.8 of the time, team A must win 0.8 of the time - makes sense! We see that as the series length gets longer, team A need not be as good as for a shorter series length.

Also the improvement from increasing series length decreases the higher the series number gets, which is consistent with the last graph.

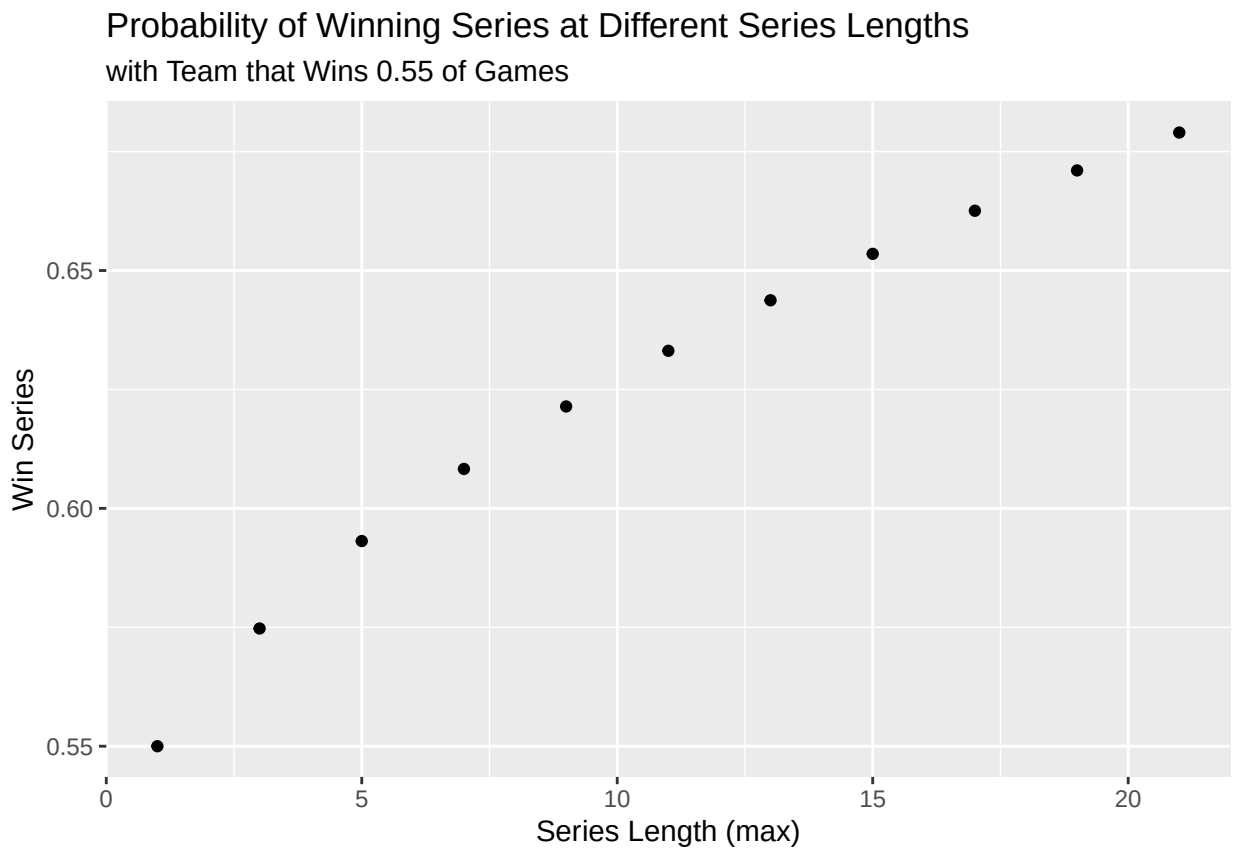
```
reduced_grid <- full_grid %>%
  filter(win_series >= 0.80 & win_series <= 0.81)
ggplot(data = reduced_grid,
       mapping = aes(
         x = p,
         y = N,
         group = interaction(N)
       )) +
  geom_line() +
  xlab("Probability of Team A Winning Head-To-Head Match-up") +
  ylab("Series Length (max)") +
  labs(title = "Shortest Series so that P(win series given p) >= 0.8")
```



Questions

—What is the probability that the Team A wins the series given that win probability = 0.55?

```
q1_grid <- full_grid %>%
  filter(p == 0.55)
ggplot(q1_grid, aes(y = win_series, x = N)) +
  geom_point() +
  xlab("Series Length (max)") +
  ylab("Win Series") +
  labs(title = "Probability of Winning Series at Different Series Lengths",
       subtitle = "with Team that Wins 0.55 of Games")
```




```
calc_win_prob(0.55, 3)
```

```
## [1] 0.57475
```

```
calc_win_prob(0.55, 21)
```

```
## [1] 0.6790034
```

The longer the series the better the teams odds of winning the series ranging from 0.57 at a 3 game series to 0.68 in a 21 game series.

—Suppose a series is best-of-9 or some other best-of-X series. What is the shortest series length so that P(win series|p = 0.55)>=0.8?

```
calc_win_prob(0.55,69)
```

```
## [1] 0.7983594
```

```
calc_win_prob(0.55,71)
```

```
## [1] 0.8017017
```

The shortest the series length needs to be in order for to guarantee a 0.8 series win percentage for a team that wins games 0.55 of the times is 71 games! As you can see above, the 0.8 is somewhere between 69 and 71 games!

—What is the shortest series length so that P(win series|p=x)>=0.8?

Please refer to second graph for shortest series lengths so win percentage is >= 0.8 by different series lengths. You can see that in a one game series the p of individual win must be 0.8 for the series win probability to also be 0.8. At 17 game series the win percentage must be around 0.6 in order to guarantee a series win percentage of 0.8.

— Calculate P(p = 0.55|Team A wins in game 7) under the assumption that either p = 0.55 or p = 0.45. Explain your solution.

```
prob_win_0.55 <- dbinom(3, size = 6, 0.55) * 0.55
prob_win_0.45 <- dbinom(3, size = 6, 0.45) * 0.45
prob_win_0.55 / (prob_win_0.45 + prob_win_0.55)
```

```
## [1] 0.55
```

Here I calculated the total area of possibilities that team a wins in 7 games. They can win in 7 as either a team that wins 0.55 of the times (prob_win_0.55) or a team that wins 0.45 of the times (prob_win_0.45). Then I simply found what proportion of that total the prob_win_0.55 is and it turns out to be 0.55.

Thus we can say that if team a wins games at either 0.55 or 0.45 of the time, and they won a series of 7 in game 7. 0.55 of the time they were a team that wins 0.55 of the time.