Lab3_probability_KC

Keith Colella

2023-02-25

```
library(tidyverse)
library(openintro)
library(scales)
library(reshape2)
kobe_streak <- calc_streak(kobe_basket$shot)
set.seed(1234)
```

Exercise 1

Question

What does a streak length of 1 mean, i.e. how many hits and misses are in a streak of 1? What about a streak length of 0?

Response

A streak of length 1 refers to a single hit followed by a miss. A streak of 0 refers to no hits, bu rather a single miss.

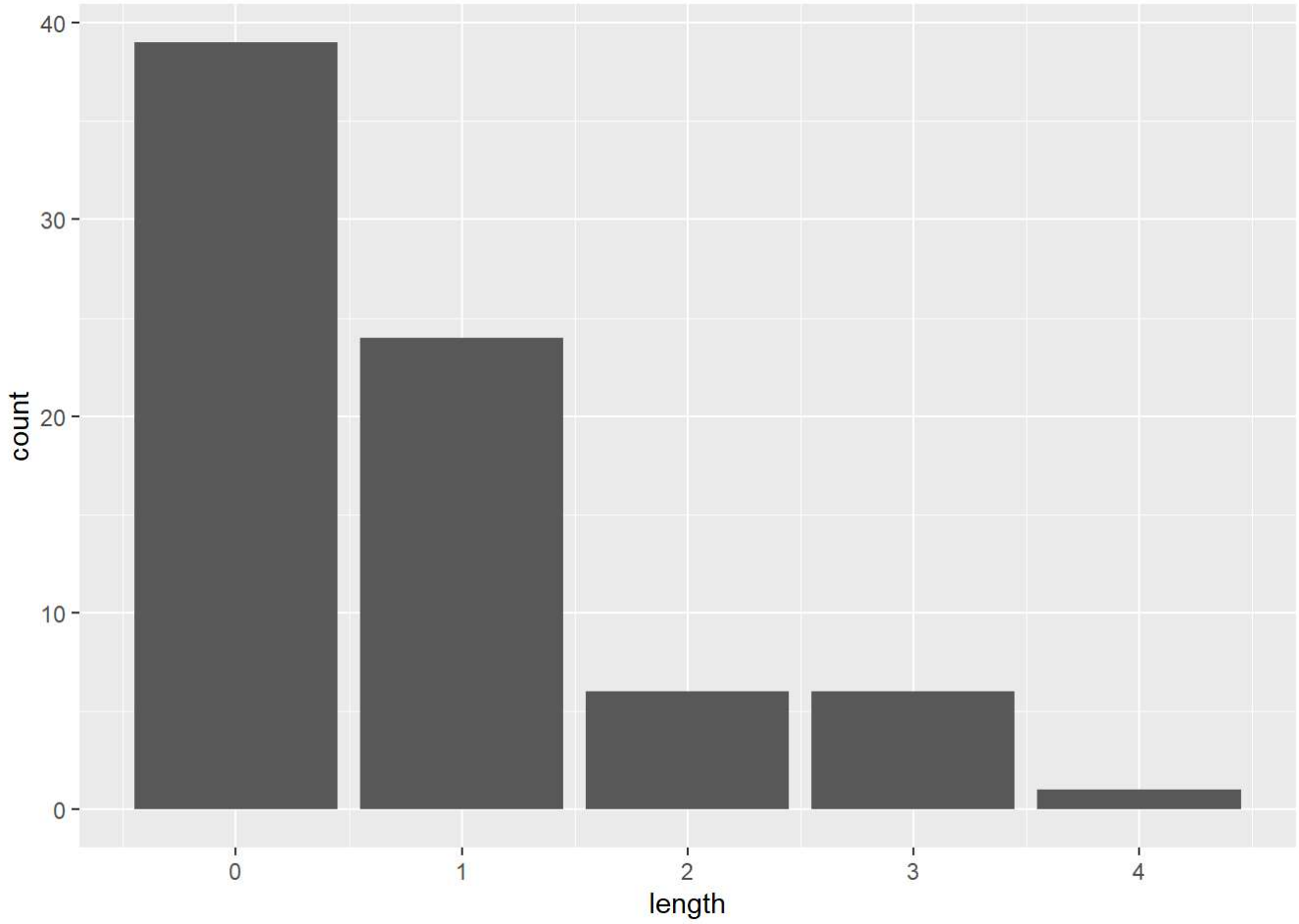
Exercise 2

Question

Describe the distribution of Kobe's streak lengths from the 2009 NBA finals. What was his typical streak length? How long was his longest streak of baskets? Make sure to include the accompanying plot in your answer.

Response

```
ggplot(data = kobe_streak, aes(x = length)) +
  geom_bar()
```




```
as.data.frame(table(kobe_streak)) %>%
  mutate(Proportion = percent(Freq / sum(Freq), 0.01))
```

```
##   length Freq Proportion
## 1      0   39     51.32%
## 2      1   24     31.58%
## 3      2    6      7.89%
## 4      3    6      7.89%
## 5      4    1      1.32%
```

Kobe's typical streak was length 0, if by typical we mean most frequent. He had 24 1-hit streaks, represent 32% of all streaks. His longest was a streak of 4, but he only hit that once.

Exercise 3

Question

In your simulation of flipping the unfair coin 100 times, how many flips came up heads? Include the code for sampling the unfair coin in your response. Since the markdown file will run the code, and generate a new sample each time you *Knit* it, you should also "set a seed" **before** you sample. Read more about setting a seed below.

Response

```
coin_outcomes <- c("heads", "tails")
unfair_coin <- c(0.2, 0.8)

## Note that the random seed was set above as part of setup
sim_unfair_coin <- sample(coin_outcomes, size = 100,
                          replace = TRUE,
                          prob = unfair_coin)

as.data.frame(table(sim_unfair_coin))
```

```
##   sim_unfair_coin Freq
## 1           heads   15
## 2           tails   85
```

Tails is sampled 85 times, while heads is sampled 15. This roughly conforms to expectations, given the unfair 20/80 probability provided for the simulation. The simulation does not, however, produce results the precisely conform to the provided probabilities, as the sample size is relatively small. If we increase the number of samples, however, see results converge to the provided probabilities.

```
sim_unfair_coin10k <- sample(coin_outcomes, size = 10000,
                             replace = TRUE,
                             prob = unfair_coin)

as.data.frame(table(sim_unfair_coin10k))
```

```
##   sim_unfair_coin10k Freq
## 1              heads 1989
## 2              tails 8011
```

With 10,000 samples, we observe much clearer convergence to our expected values (2000 and 8000 for heads and tails, respectively).

Exercise 4

Question

What change needs to be made to the sample function so that it reflects a shooting percentage of 45%? Make this adjustment, then run a simulation to sample 133 shots. Assign the output of this simulation to a new object called sim_basket.

Response

```
shot_outcomes <- c("H", "M")
sim_basket <- sample(shot_outcomes, size = 133, replace = TRUE)

as.data.frame(table(sim_basket)) %>%
  mutate(Proportion = percent(Freq / sum(Freq), 0.01))
```

```
##   sim_basket Freq Proportion
## 1          H   80     60.15%
## 2          M   53     39.85%
```

Without assigning any probability, the sample function defaults to assign equal (or uniform) probabilities to each potential outcome. So, we would expect to see roughly equal numbers of hits and misses. Surprisingly, however, we see a pretty unequal distribution, with significantly more hits than misses. I suspect the difference is due to the "luck of the draw". With a relatively small sample size, it's possible our observations deviate significantly from expected values.

To explicitly account for Kobe's shooting percentage of 45%, we add the `prob` argument as follows.

```
shooting_percentage <- c(0.45,0.65)

sim_basket <- sample(shot_outcomes,
                     size = 133,
                     replace = TRUE,
                     prob = shooting_percentage)

as.data.frame(table(sim_basket)) %>%
  mutate(Proportion = percent(Freq / sum(Freq),0.01))
```

```
##   sim_basket Freq Proportion
## 1          H   47     35.34%
## 2          M   86     64.66%
```

By specifying the `prob` parameter, the simulation more closely reflects the 45% shooting percentage. Still, we don't see clear convergence to expected values, due to the size of our sample. In Exercise #7, I'll explore whether this deviation from expected values remains when increasing sample size.

Exercise 5

Question

Using calc_streak, compute the streak lengths of sim_basket, and save the results in a data frame called sim_streak.

Response

```
sim_streak <- calc_streak(sim_basket)
```

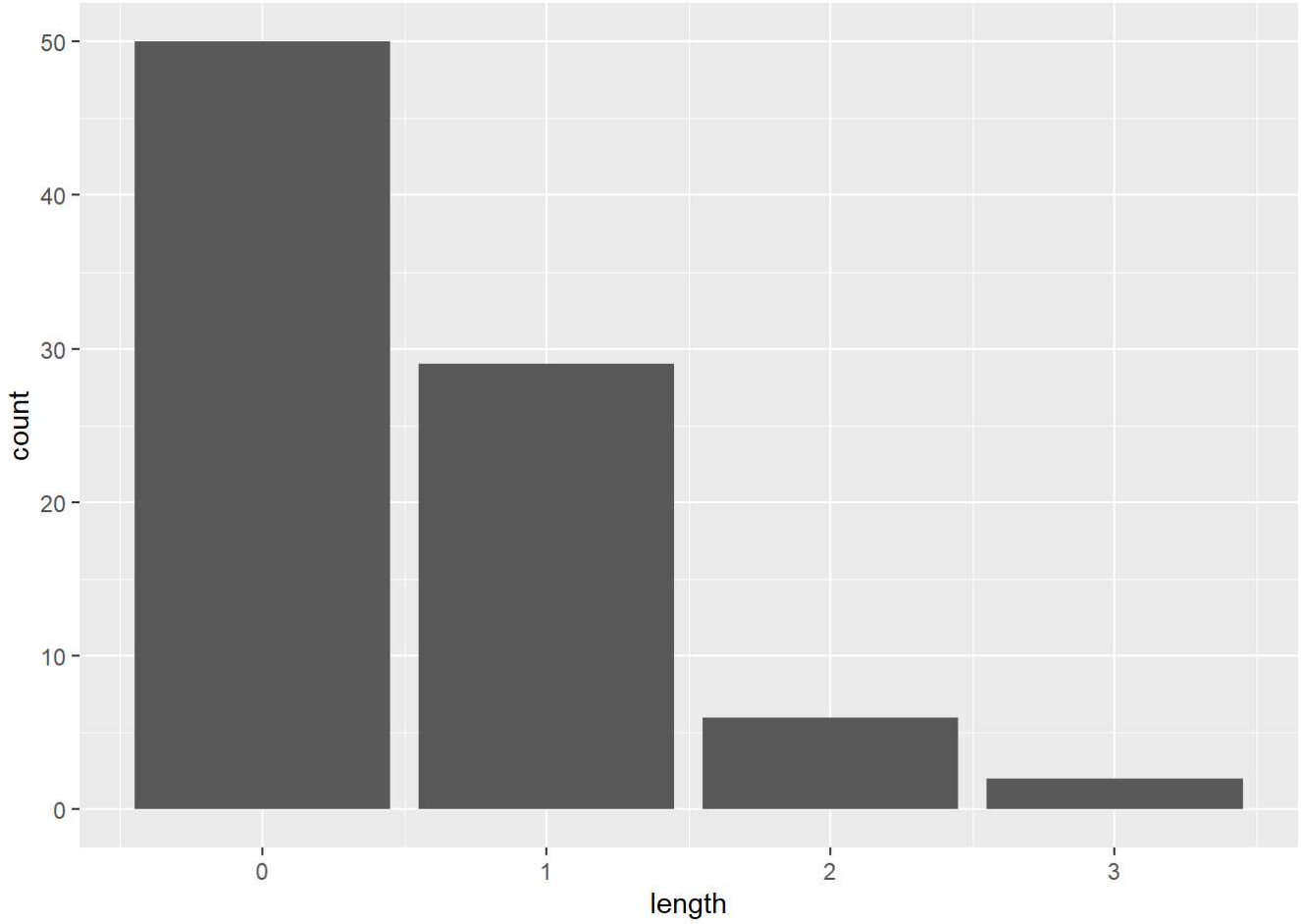
Exercise 6

Question

Describe the distribution of streak lengths. What is the typical streak length for this simulated independent shooter with a 45% shooting percentage? How long is the player's longest streak of baskets in 133 shots? Make sure to include a plot in your answer.

Response

```
ggplot(data = sim_streak, aes(x = length)) +
  geom_bar()
```




```
as.data.frame(table(sim_streak)) %>%
  mutate(Proportion = percent(Freq / sum(Freq), 0.01))
```

```
##   length Freq Proportion
## 1      0   50     57.47%
## 2      1   29     33.33%
## 3      2    6      6.90%
## 4      3    2      2.30%
```

As with Kobe, our simulated player mostly shoots 0 length streaks (that is, he has multiple sequential misses). Their single hit streaks are similar in length, as well, as at 29 versus 24 for Kobe. Our shooter do not, however, have as many long streaks. Their longest streak is only three hits, and they only achieved that twice. Kobe, on the other hand, got six three-shot streaks, along with a single four-shot streak.

Do these results provide support for the “hot hand” phenomenon? Or is it simply due to chance? I suspect the latter is true, but I’ll explore that more below.

Exercise 7

Question

If you were to run the simulation of the independent shooter a second time, how would you expect its streak distribution to compare to the distribution from the question above? Exactly the same? Somewhat similar? Totally different? Explain your reasoning.

Response

```
sim_basket2 <- sample(shot_outcomes,
                      size = 133,
                      replace = TRUE,
                      prob = shooting_percentage)

as.data.frame(table(sim_basket2)) %>%
  mutate(Proportion = percent(Freq / sum(Freq), 0.01))
```

```
##   sim_basket2 Freq Proportion
## 1           H   55     41.35%
## 2           M   78     58.65%
```

```
sim_streak2 <- calc_streak(sim_basket2)

as.data.frame(table(sim_streak2)) %>%
  mutate(Proportion = percent(Freq / sum(Freq), 0.01))
```

```
##   length Freq Proportion
## 1      0   44     55.70%
## 2      1   22     27.85%
## 3      2    7      8.86%
## 4      3    5      6.33%
## 5      4    1      1.27%
```

Our second simulation produces a roughly similar streak distribution, but significant differences emerge. We now observe a greater number of longer streaks, and a four-hit streak similar to Kobe's. It would seem that our first simulation resulted in fewer long streaks only by chance, rather than because of any dependence across trials in Kobe's sample (i.e. because of "hot hand").

It is hard to make confident conclusions with such small sample sizes. Of course, I would expect larger sample sizes to exhibit greater consistency across simulations. But why guess? Let's put it to the test!

WARNING: Tangent ahead!

The function below performs repeated simulations based on the framework established above. It takes a sample size and number of iterations as inputs. The streak frequencies obtained from each simulation are gathered in a dataframe.

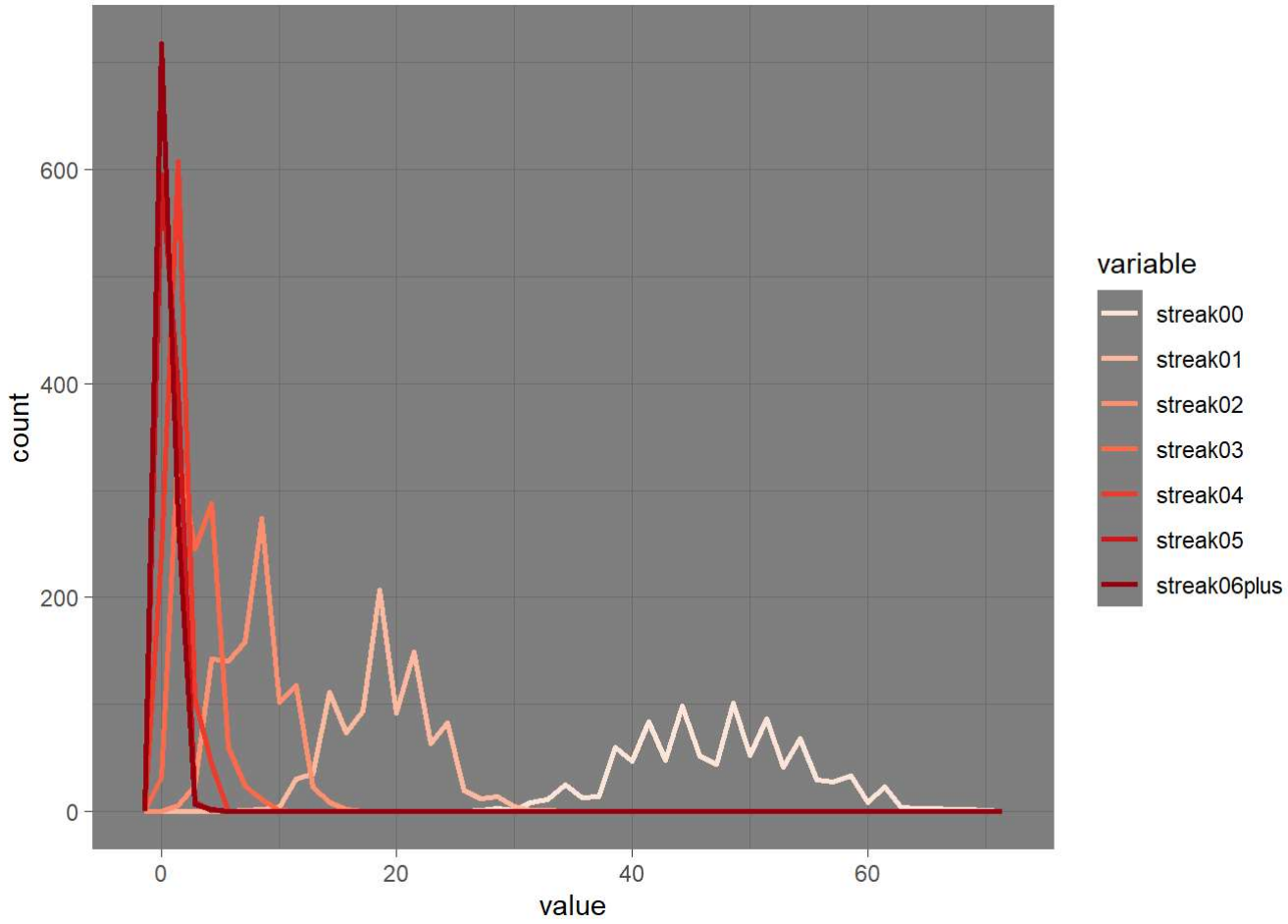
```
repeat_sims <- function (samples, iterations) {
  ## initiate variables
  results <- data.frame()

  ## loop through simulations
  for (i in 1:iterations) {
    ## simulate shots and calculate streaks
    sim_basket <- sample(shot_outcomes,
                                size = samples,
                                replace = TRUE,
                                prob = shooting_percentage)
    sim_streak <- calc_streak(sim_basket)

    ## format results and add to dataframe
    sim_streak <- as.data.frame(table(sim_streak))
    cols <- sim_streak$length
    sim_streak <- data.frame(t(sim_streak$Freq))
    colnames(sim_streak) <- cols
    results <- bind_rows(results, sim_streak)
  }

  ## final cleaning on dataframe
  cols <- c()
  for (i in colnames(results)) {
    cols <- c(cols, paste0('streak',
                           str_pad(i, 2, side="left", pad="0")
                           )
              )
  }
  colnames(results) <- cols

  col_order <- sort(colnames(results))
  results <- results[,col_order]

  results[is.na(results)] <- 0

  ## return results
  return (results)
}
```

We use this function to run two tests. First, we run 1000 simulations, each with 133 samples (similar to our original simulation). Then, we run 1000 simulations with 13,000 samples. We divide the resulting frequencies by 100 so that we have comparable frequencies while retaining relative proportions.

```
results133 <- repeat_sims(133, 1000)
results13300 <- repeat_sims(13300, 1000) / 100
```

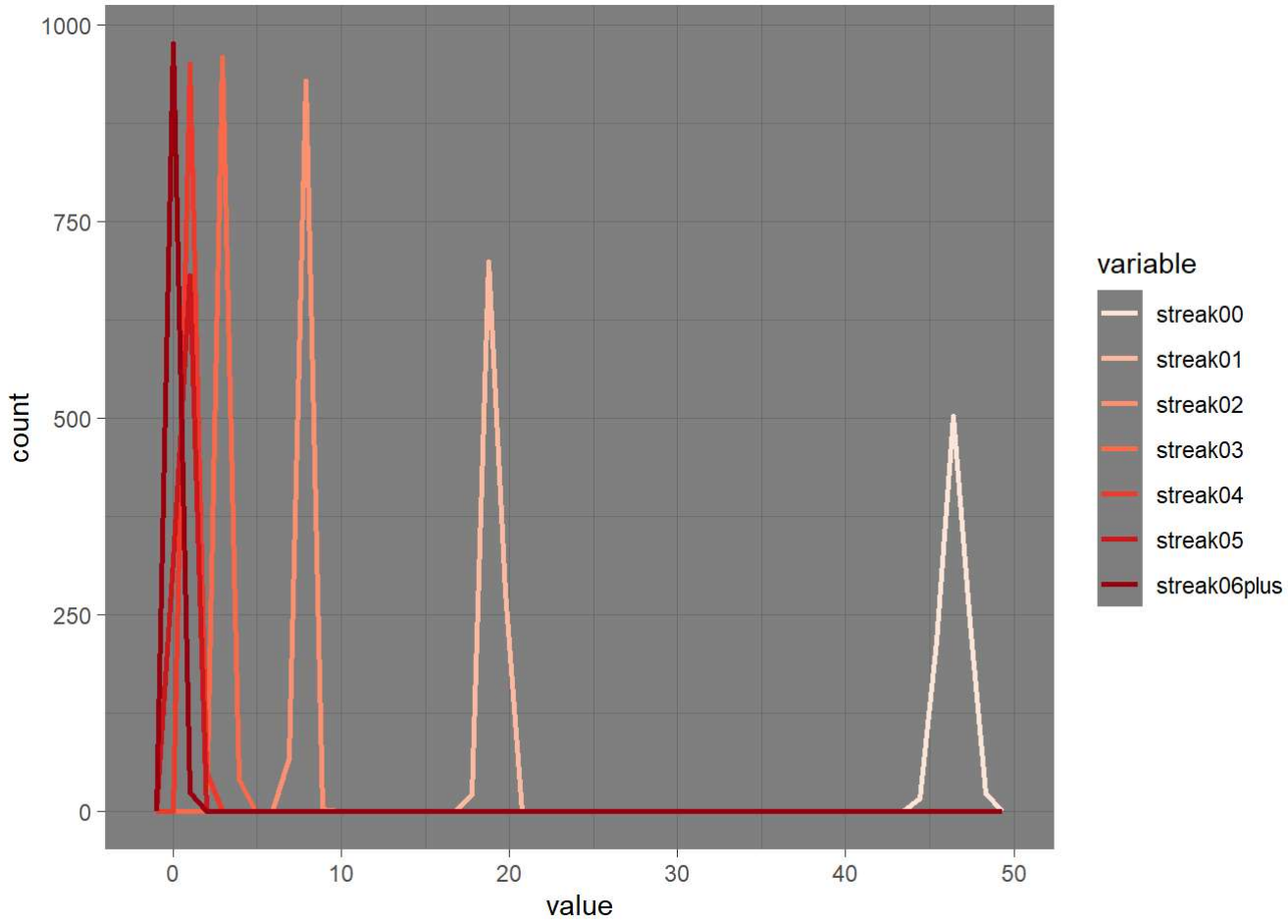
Below, we plot our results using the freqpoly geom to observe the distributions of streaks of various length. With 133 samples, we observe a significant amount of spread in the distribution of observations under each streak length. For example, the number of zero-length streaks per simulation ranges from ~30 to ~60 across all 1000 runs.

```
results133 <- results133 %>%
  mutate(streak06plus = rowSums(across(
    colnames(results133[1:10,7:ncol(results133)])))) %>%
  select(streak00:streak05,streak06plus)

results133_reshape <- melt(results133, id.vars = NULL)

ggplot(data = results133_reshape, aes(x = value, color = variable)) +
  geom_freqpoly(bins = 50, linewidth = 1) +
  scale_color_brewer(palette = 'Reds') +
  theme_dark()
```



By contrast, when using a sample size of 13,000, the spread for each streak length is drastically reduced. For all 1000 simulations, the number of zero-length streaks is limited to a narrow band between ~45 and ~55, a ~66% reduction from the band observed above with the smaller sample size.

```
results13300 <- results13300 %>%
  mutate(streak06plus = rowSums(across(
    colnames(results13300[1:10,7:ncol(results13300)])))) %>%
  select(streak00:streak05,streak06plus)

results13300_reshape <- melt(results13300, id.vars = NULL)

ggplot(data = results13300_reshape, aes(x = value, color = variable)) +
  geom_freqpoly(bins = 50, linewidth = 1) +
  scale_color_brewer(palette = 'Reds') +
  theme_dark()
```



This comparison shows that, by increasing the sample size, we can reduce some of the “noise” obtained with smaller samples and more clearly identify expected values.

Finally, we view the relative frequency of streaks of each length. The distributions are mostly similar, with one key difference.

```
results133_reshape %>%
  group_by(variable) %>%
  summarize(length_total = sum(value)) %>%
  mutate(frequency = percent(length_total / sum(length_total)))
```

```
## # A tibble: 7 × 3
##   variable     length_total frequency
##   <fct>               <dbl> <chr>
## 1 streak00            47319 59.44%
## 2 streak01            19098 23.99%
## 3 streak02             7840 9.85%
## 4 streak03             3188 4.00%
## 5 streak04             1329 1.67%
## 6 streak05              506 0.64%
## 7 streak06plus          331 0.42%
```

```
results13300_reshape %>%
  group_by(variable) %>%
  summarize(length_total = sum(value)) %>%
  mutate(frequency = percent(length_total / sum(length_total)))
```

```
## # A tibble: 7 × 3
##   variable     length_total frequency
##   <fct>               <dbl> <chr>
## 1 streak00           46439. 59.09%
## 2 streak01           19010. 24.19%
## 3 streak02            7766. 9.88%
## 4 streak03            3179. 4.05%
## 5 streak04            1304. 1.66%
## 6 streak05             530. 0.67%
## 7 streak06plus         368. 0.47%
```

In contrast with the comparison above, these two tables show that, while smaller sample sizes will produce more "noise" in each individual iterations, we can eliminate a lot of that noise with a sufficient number of simulations. When aggregating results across all 1000 iterations, we see very similar frequencies emerge.

NOTE: Tangent over!

Exercise 8

Question

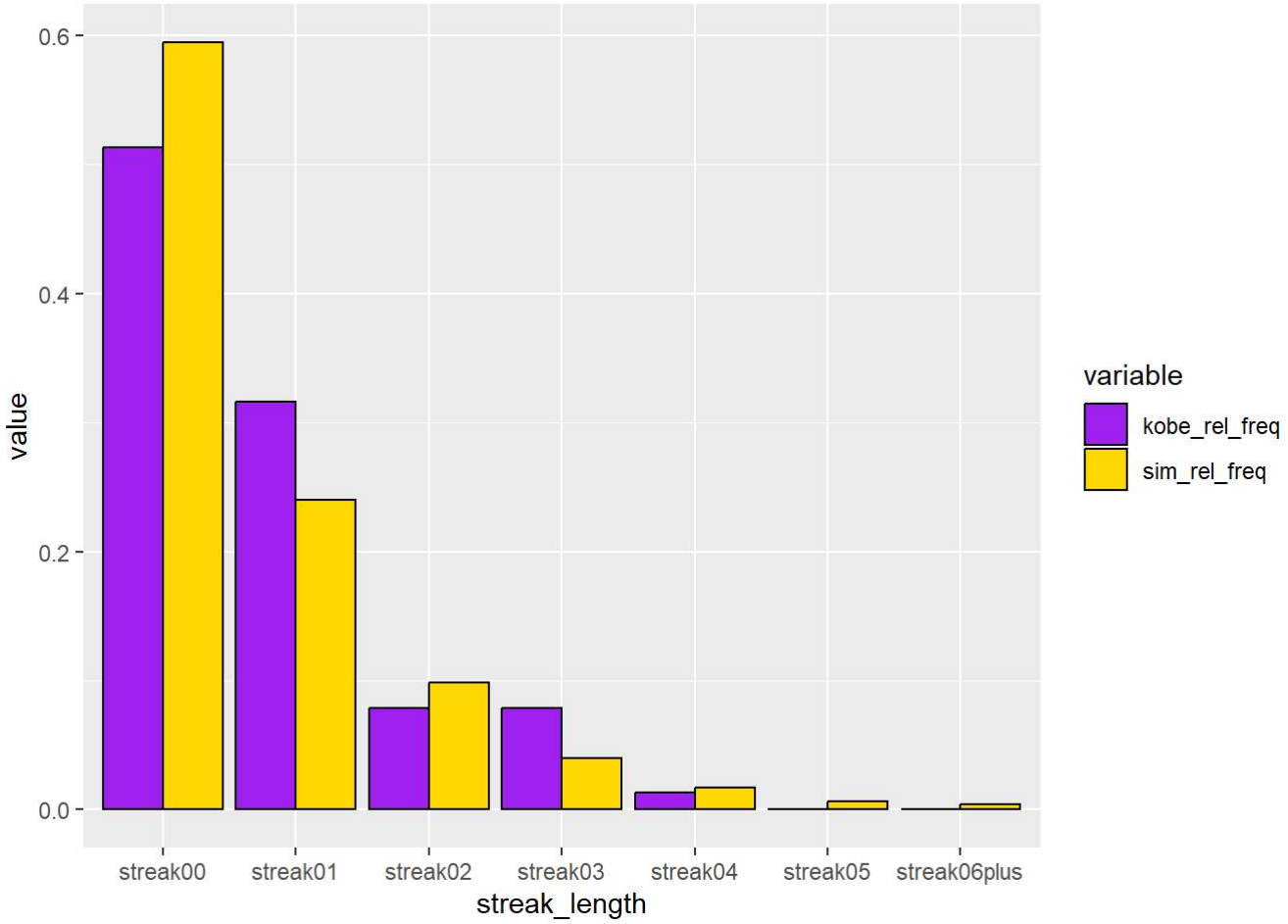
How does Kobe Bryant's distribution of streak lengths compare to the distribution of streak lengths for the simulated shooter? Using this comparison, do you have evidence that the hot hand model fits Kobe's shooting patterns? Explain.

Response

```
kobe_final <- as.data.frame(table(kobe_streak)) %>%
  mutate(kobe_rel_freq = Freq / sum(Freq), 0.01) %>%
  select(kobe_rel_freq) %>%
  as.vector()

sim_final <- results133_reshape %>%
  group_by(variable) %>%
  summarize(length_total = sum(value)) %>%
  mutate(sim_rel_freq = length_total / sum(length_total)) %>%
  select(variable, sim_rel_freq) %>%
  as.vector()

length(kobe_final$kobe_rel_freq) <- length(sim_final$sim_rel_freq)

comparison <- data.frame(streak_length = sim_final$variable,
                         kobe_rel_freq = kobe_final$kobe_rel_freq,
                         sim_rel_freq = sim_final$sim_rel_freq)

comparison[is.na(comparison)] <- 0

comparison %>%
  mutate(kobe_rel_freq = percent(kobe_rel_freq),
         sim_rel_freq = percent(sim_rel_freq))
```

```
##   streak_length kobe_rel_freq sim_rel_freq
## 1      streak00         51.3%       59.44%
## 2      streak01         31.6%       23.99%
## 3      streak02          7.9%        9.85%
## 4      streak03          7.9%        4.00%
## 5      streak04          1.3%        1.67%
## 6      streak05          0.0%        0.64%
## 7  streak06plus          0.0%        0.42%
```

```
ggplot(data = melt(comparison, id.vars = 'streak_length'),
       aes(x = streak_length, y = value, fill = variable)) +
  geom_col(color = 'black', position='dodge') +
  scale_fill_manual(values=c('purple','gold'))
```



The simulations above provide little support for the hot hand theory. The aggregated results of 1000 simulations with 133 samples and reveal generally comparable frequencies for long streaks. The simulated player hits slightly more 2-shot and 4-shot+ streaks, whereas Kobe hits more 3-shot streaks. The differences, however, appear minimal enough to discount any generalizable trend.