GR5065 Homework 1

Ben Goodrich

Due February 7, 2022 at 6PM

1 Poker

1.1 Probability of rare five-card combinations

Before the hand starts (i.e. irrespective of any betting and presuming the player does not fold)

- What is the probability of a player ending up with a "flush", which is defined as a five-card combination where all five cards are of the same suit? You can ignore the tiny probability of obtaining a "straight flush", where all five cards have adjacent values in addition to being of the same suit.
- What is the probability of a player ending up with a "full house" — which is defined as a five-card combination where three of the cards are of the same value and the other two cards have the same value — but not four-of-a-kind?

For the flush,

```
( dhyper(x = 5, m = 13, n = 52 - 13, k = 7) +
  dhyper(x = 6, m = 13, n = 52 - 13, k = 7) +
  dhyper(x = 7, m = 13, n = 52 - 13, k = 7) ) * 4
```

```
## [1] 0.03056577
```

The first line calculates the probability of exactly five out of seven cards are of the same suit. But it is also a flush in the unlikely event that all six or seven cards of the same suit (only five of which count), so we use the General Addition Rule with disjoint events on the next two lines. Finally, we multiply by four because there are four suits that each have the same probability of comprising a flush (and ignore a few strait flushes).

For the full house, there are three ways the seven cards could be allocated:

(A) 3 of a one type of card, 3 of a second type of card (one of which does not count toward the five), and 1 of a third type of card (otherwise it would be four-of-a-kind). Thus, there is a $\binom{13}{2}$ factor below, because there are 13 values of cards that we are selecting two of to comprise the full house, and a factor of 44 because there are 44 cards of the third type.

(B) 3 of one type of card, 2 of a second type of card, and 2 of a third type of card (neither of which counts toward the five). Thus, there is a factor of 13 below because there are 13 values of cards to obtain three of and then a $\binom{12}{2}$ factor because there 12 other values of cards that we need two pairs of.

(C) 3 of one type of card, 2 of a second type of card, and 2 of distinct types of cards (neither of which counts toward the five). By the same reasoning as previously, there is a factor of 13 below, a factor of 12, and a factor of $\binom{11}{2}$

Thus, we have

```
total <- choose(52, 7)
A <- choose(4, 3)^2 * 44 * choose(13, 2) / total
B <- choose(4, 3) * choose(4, 2)^2  * 13 * choose(12, 2) / total
C <- choose(4, 3) * choose(4, 2) * 4^2   * 13 * 12 * choose(11, 2) / total
A + B + C
```

```
## [1] 0.02596102
```

The probability of a flush or a full house and other hands are listed at

https://en.wikipedia.org/wiki/Poker_probability#Frequency_of_7-card_poker_hands

which also provides some additional detail on the combinatorics of how they are calculated using the choose function.

1.2 Pre-flop action

Before the hand starts, what number is the probability of Selbst being dealt two Aces as hole cards? Then, for each of the following explain your thinking (but exact calculations are not required):

- From Selbst's perspective when she made the decision to bet the first time, is the probability that she has two Aces higher or lower than the probability of being dealt two Aces?
- From Baumann's perspective when she made the decision to call the first time, is the probability that Selbst has two Aces higher or lower than the probability of being dealt two Aces?
- From Rosen's perspective when he made the decision to fold, is the probability that Selbst has two Aces higher or lower than the probability of being dealt two Aces?

The probability of being dealt two Aces is

```r
4 / 52 * 3 / 51 # = dhyper(x = 2, m = 4, n = 52 - 4, k = 2)
```

```
## [1] 0.004524887
```

or $\frac{1}{13} \times \frac{1}{17} = \frac{1}{221}$.

From Selbst's perspective the "probability" that she has two Aces is 1, which is to say that she knows it with certainty because she sees the A♠A♦ with her eyes. This seems trivial but it is actually an important point in Bayesian analysis: A realization of a random process is data to condition on if you observe it.

Baumann cannot see that Selbst has two Aces but does observe that Selbst bets, which is consistent with Selbst having two Aces but is also consistent with Selbst having about $\frac{1}{7}\binom{52}{2}$ other good hands. By Bayes Rule, the conditional probability that Selbst has a pair of Aces given that Selbst bets is

$$\Pr\left(\text{Aces} \mid \text{bets}\right) = \frac{\frac{1}{221} \times 1}{\frac{1}{7}} = \frac{7}{221} \approx 0.032,$$

which reads as the prior probability from above that Selbst is dealt a pair of Aces, times the conditional "probability" that Selbst bets given that she has a pair of Aces (which is guaranteed), divided by the probability that Selbst bets irrespective of whether she has a pair of Aces, which was said to be roughly $\frac{1}{7}$. Technically, the calculation should condition on the fact that Baumann has two sevens, so there are four Aces among the 50 cards left in the deck, and Selbst would have bet with a few hands involving Baumann's 7♥7♦, but those considerations would not change the posterior probability very much, and it would still be considerably greater than the prior probability of $\frac{1}{221}$.

From Rosen's perspective, the fact that he has the Ace of hearts is much more consequential to the calculation. Now Bayes Rule implies

$$\Pr\left(\text{Aces} \mid \text{bets}, \text{A♥}\right) = \frac{\frac{3}{51} \times \frac{2}{50} \times 1}{\frac{1}{7}} \approx 0.016,$$

where we have recomputed the prior probability that Selbst is dealt two Aces from a 51-card deck that is missing the A♥ (but ignoring the fact that it is also missing the 6♣) but again multiplied by the conditional "probability" that Selbst would bet with a pair of Aces and divided by the probability that Selbst would bet irrespective of whether she has a pair of Aces (and ignoring that Selbst would have bet with several hands involving Rosen's A♥). This posterior probability is still higher than the probability of Selbst being dealt a pair of Aces but is considerably lower than the probability from Baumann's perspective.

1.3 Schwartz's call

Explain in words why it is a good decision (exact calculations are not required on this subproblem) for Schwartz to call pre-flop.

By betting 250 more chips, Schwartz gives himself a chance to win the pot which at that moment is $1,025$ chips. The expected change in chips would be

$$1025 \times \Pr\left(\text{wins} \mid \dots\right) - 250 \times \left(1 - \Pr\left(\text{wins} \mid \dots\right)\right)$$

So, Schwartz only needs about a $\frac{1}{5}$ chance of winning the hand in order for this expected change to be positive and thus exceed the value of zero chips gained if Schwartz were to fold pre-flop. Even though both Selbst and Baumann both presumably have very good hands, almost any two decent hole cards would give Schwartz at least a $\frac{1}{5}$ chance of winning the hand. It is more complicated because there will be more betting post-flop, but presuming Schwartz will make optimal decisions post-flop, it also makes sense for him to call pre-flop.

1.4 Flop action

If Schwartz's utility function were equal to his chips at the conclusion of the hand and Schwartz would be indifferent between folding and calling Selbst's bet of 700 chips (which Baumann called, bringing the pot to $2,675$ chips) if he had the Q♣ and 8♥, what must the conditional probability be that either Selbst or Baumann has the K♣? To simplify this subproblem, you can assume that Schwartz would definitely win the pot if he makes a club flush and no one else has the K♣ (although that turned out not to be true in this video) and that Schwartz would certainly lose if he does not make a flush. Also, you can assume that there would be no more betting.

Since Schwartz can see his two hole cards and three clubs in the middle, if either of the next two cards is the K♣, then Schwartz would know he has the best possible flush with the Q♣. The probability of that is

```r
dhyper(x = 1, m = 1, n = 52 - 2 - 3, k = 2)
```

```
## [1] 0.04166667
```

Alternatively, if either (or both) of the next two cards is a club but not the K♣, then Schwartz would know he has the second-best possible flush. The probability of that is

```r
dhyper(x = 1, m = 13 - 5, n = 52 - 8 - 1, k = 2) +
dhyper(x = 2, m = 13 - 5, n = 52 - 8 - 1, k = 2)
```

```
## [1] 0.2917647
```

If you add those three terms together, ignore the chance that either Selbst or Baumann ends up with a hand that is better than a flush, and round, you obtain the 33% chance of Schwartz winning that you see on the video at 0:57.

But Schwartz does not know that neither Selbst nor Baumann has the K♣, so the last two terms need to be multiplied by the conditional probability, say C, that neither Selbst nor Baumann has the K♣ when calculating the expected change in chips as

$$2675 \times 0.0417 + 2675 \times 0.2917 \times C - 700 \times (1 - 0.0417 - 0.2917 \times C)$$

Set that equal to zero because Schwartz is supposedly indifferent between calling and folding and solve for $C \approx 0.568$. Thus, the conditional probability that either Selbst or Baumann has the K♣ is about 0.432 or perhaps a little lower if we were to take into account that either Selbst or Baumann could end up with a hand that is better than a flush. Intuitively, Selbst or Baumann would be betting after the flop if either had the K♣, but they would also be betting with a somewhat larger collection of hands that do not involve the K♣.

1.5 Pre-turn calculation

After Schwartz folds at 1:02, the percentage chances (in the bottom left of the video) switch to 91% for Selbst to win and 4% for Baumann to win, with an implicit 5% chance that they tie if the remaining two cards in the middle are both clubs (all percentages were rounded). In the event that both Selbst and Baumann end up with three-of-a-kind, Selbst would win because Aces are better than sevens. How did the video arrive at these percentages?

The probability that Baumann wins is the probability that either of the next two cards in the middle is the 7♠ and the other is not the A♥, which is about 4%

```
dhyper(x = 1, m = 1, n = 52 - 7, k = 2) *
(1 - dhyper(x = 1, m = 1, n = 52 - 8, k = 1))
```

```
## [1] 0.04251208
```

The probability that both of the next two cards are clubs is about 5% when there are 10 clubs left in a deck of 45 cards

```
10 / 45 * 9 / 44 # = dhyper(x = 2, m = 10, n = 35, k = 2)
```

```
## [1] 0.04545455
```

And if anything else happens, then Selbst would win which is the remaining 91%.

1.6 Turn

The next card revealed in the middle (called the "turn") is 7♠ and the commentators go crazy because Selbst now has a full house but Baumann has four-of-a-kind. Before the hand started, irrespective of any betting and presuming neither player folds, what was the probability of one player at the table ending up with four-of-a-kind and another player ending up with a full house?

We can break down the probability of someone getting four of one type of card

```
dhyper(x = 4, m = 4, n = 52 - 4, k = 7)
```

```
## [1] 0.0001292825
```

as

(A) The probability that the player has a pair and the board has that same pair
(B) The probability that the board has three of some type card and the player has the other one
(C) The probability that the board has four of that type of card

```
A <- dhyper(x = 2, m = 4, n = 48, k = 2) * dhyper(x = 2, m = 2, n = 48, k = 5)
B <- dhyper(x = 1, m = 4, n = 48, k = 2) * dhyper(x = 3, m = 3, n = 47, k = 5)
C <- dhyper(x = 0, m = 4, n = 48, k = 2) * dhyper(x = 4, m = 4, n = 46, k = 5)
A + B + C
```

```
## [1] 0.0001292825
```

In scenario (C), it is impossible for another player to have a full house because both players have four-of-a-kind. In scenario (A), the other player would need to have three of a kind among their two hole cards and the three other cards on the board. In scenario (B), the other player would need to have a pair among their two hole cards and the two other cards on the board. Thus, the joint probability is

```
D <- A * dhyper(x = 3, m = 4, n = 44, k = 5) +
     B * dhyper(x = 2, m = 4, n = 44, k = 4)
```

Furthermore, `D` must be multiplied by 13, 12, and 2 because there are 13 ways for the first player to make four-of-a-kind, 12 ways each for the second player to make a full house, and 2 orderings of a pair of players. The probability of one player making a four-of-a-kind and the other making a full house is

```r
D * 13 * 12 * 2
```

```
## [1] 0.0006978258
```

which is perhaps greater than one might think, but much of that involves the less dramatic scenario (B) rather than the scenario (A) that happens in the video.

Taking into account that there are eight players at the table would be more difficult, but usually all but two players fold pre-flop anyway.

It is possible to check your answer with a simulation, but you need to draw many times in order to estimate small probabilities with high relative accuracy.

```r
library(dplyr)
library(parallel)
deck <- rep(1:13, each = 4)
S <- 10^6
mclapply(1:S, mc.cores = detectCores(), FUN = function(s) {
  cards <- sample(deck, size = 2 * 2 + 5, replace = FALSE)
  board <- tail(cards, n = 5)
  if (n_distinct(board) == 5) return(FALSE) # full house is impossible
  hole_cards <- matrix(head(cards, n = -5), nrow = 2)
  four <- apply(hole_cards, MARGIN = 2, FUN = function(hand) {
    any(table(c(hand, board)) == 4)
  })
  if (sum(four) != 1) return(FALSE) # no isolated four-of-a-kind
  full_house <- apply(hole_cards, MARGIN = 2, FUN = function(hand) {
    tab <- table(c(hand, board))
    all(2:3 %in% tab) | sum(tab == 3) == 2
  })
  if (!any(full_house)) return(FALSE) # no full house
  # Are the four-of-a-kind and the full house not the same player?
  !identical(which(four), which(full_house))
}) %>% unlist %>% mean
```

```
## [1] 0.000746
```

1.7 River action

Explain with reference to Bayes Rule why Selbst's decision to call Baumann's raise was justifiable. To do so, calculate the conditional probability that Selbst wins the pot given the seven cards that Selbst can see and all the previous betting. Note that Selbst's decision to call can be justified even if the probabilities that you use are not exactly correct.

Conditioning on the fact that Baumann raised all-in eliminates a lot of hands that she could have from Selbst's consideration, such as a flush. In other words, $\Pr(\text{All in} \mid \text{flush}) = 0$ because from Baumann's perspective it would be too likely that Selbst has a hand that is better than a flush and too unlikely that Selbst would call with a hand that is worse than a flush. Although Baumann might have raised all-in with A♥7♦, conditioning on the fact that Baumann called pre-flop makes it very unlikely that Baumann has an Ace and a seven of different suits.

If Baumann would definitely raise all-in with A♥7♥, then the conditional probability that Baumann has A♥7♥ would be the same as the conditional probability that Baumann has 7♥7♦. In other words, both hands are equally consistent with the available information. Bayes Rule implies

$$\Pr(\text{Selbst wins} \mid \dots) = \frac{\Pr(\text{Baumann has A♥7♥ and raises all-in})}{\Pr(\text{Baumann has A♥7♥ and raises all-in}) + \Pr(\text{Baumann has 7♥7♦ and raises all-in})} = \frac{1}{2}$$

If there were any other hole cards that Baumann would raise all-in with, the probability that Selbst would win would be even higher than $\frac{1}{2}$.

Thus, if Selbst has at least a $\frac{1}{2}$ chance to win $66,975$ chips by risking losing another $20,300$ chips, the expected change in chips if Selbst's calls is very positive. In order for the expected change in chips to be negative, Selbst would have to believe that the probability that Baumann would raise all-in with A♥7♥ is much lower than the "probability" that Baumann would raise all-in with 7♥7♦, which is a certainty. It seems from the video that Selbst is not entirely sure whether Baumann would raise all-in with A♥7♥ but that probability would have to be lower than $\frac{1}{3}$ to make calling a bad decision.

1.8 Baumann's strategy on the river

Given that the five cards in the middle are A♣7♣5♣7♠4♦, show that if Baumann's strategy were to raise all-in only if she either had two sevens or two Aces as hole cards, *and if Selbst knows that* then Baumann's strategy would have a worse expectation than a strategy of calling with either two sevens or two Aces. This implies that Baumann would also have to raise all-in with a third hand that is even worse (a bluff) in order for the all-in strategy as a whole to have a positive expectation.

If Selbst knows that this is Baumann's strategy, given that Baumann raises all-in, then Baumann must have either two sevens or two Aces. Thus, if Selbst has two sevens, then Baumann must have two Aces, in which case Selbst would win the pot by calling. Otherwise, Selbst would fold and not lose any more chips. In short, if Selbst knows that this is Baumann's strategy, Baumann will never get called by a worse hand and would only be called by a better hand. Thus, Baumann would be be better off calling in both cases rather than raising all-in in both.

2 Surnames

2.1 Notation

Rewrite the denominator in equation (2) from the paper using the crossed-out notation like that we used for bowling, $\Pr\left(\cancel{x_1} \bigcap x_2 \mid n = 10\right)$.

$$\Pr\left(R_i = r \mid S_i = s, G_i = g\right) = \frac{\Pr\left(G_i = g \mid R_i = r\right)\Pr\left(R_i = r \mid S_i = s\right)}{\sum_{r' \in \mathcal{R}} \Pr\left(G_i = g \mid R_i = r'\right)\Pr\left(R_i = r' \mid S_i = s\right)} = \frac{\Pr\left(G_i = g \bigcap R_i = r \mid S_i = s\right)}{\Pr\left(G_i = g \bigcap \cancel{R_i = r} \mid S_i = s\right)}$$

2.2 Frequentist Perspective

Would Fisher approve or disapprove of this use of Bayes' Rule in the paper? Why?

On one hand, there are no theoretically unobservable parameters in this model, and this paper does not actually use any Bayesian *estimation* methods that we will learn later in this semester. However, Fisher would still find this use of Bayes Rule to be objectionable. What race person i is analogous to the question of whether a huge odd integer is prime? There is no repeatable experiment in which person i is white some proportion of the time and not white other times.

One could ask the question what is the probability of randomly *drawing* a white person from a population, which Frequentist probability could be applied to, but that is not what is going on here. The people who register or turn out to vote in an election are not a random sample from a well-defined population. This is basically a missing data problem because the race of the voter is not recorded in most states.

2.3 You

```
NYS <- readRDS("NYS.rds")
my_df <- data.frame(surname = c("Goodrich", # Ben
```

```r
                                 "Chan",      # Monica
                                 "Perez"),    # Abel
                     state = "NY", county = "061",
                     age = c(42, 27, 26), sex = c(0, 1, 0))
library(wru)
me <- predict_race(my_df, census.geo = "county", census.data = NYS, age = TRUE, sex = TRUE)
```

```
## [1] "Proceeding with Census geographic data at county level..."
## [1] "Using Census geographic data from provided census.data object..."
## [1] "State 1 of 1: NY"
```

```r
me
```

```
##    surname state county age sex   pred.whi    pred.bla   pred.his   pred.asi
## 1 Goodrich    NY    061  42   0 0.86721839 0.050607518 0.03710415 0.00827340
## 2     Chan    NY    061  27   1 0.01408600 0.002441564 0.02301289 0.94725784
## 3    Perez    NY    061  26   0 0.04075648 0.002559527 0.93417929 0.01765106
##      pred.oth
## 1 0.036796539
## 2 0.013201710
## 3 0.004853639
```

The posterior probability than Ben is white is about 0.87, which is the much greater than the probability that he is of any other race. Similarly for Monica and Abel, their last names result in posterior probabilities that are even closer to 1, despite the fact that there are much fewer Asians and Hispanics in Manhattan than whites.

2.4 Granularity

Researchers can use this method with Census geolocation data at the county-level or at smaller geographical units, such as the Census tract (which contains about $4,000$ contiguous people). How would you anticipate the probabilities in `me` changing if we were to use the Census tract data (and the tract that you live in now) rather than the county-level data? Why?

Ben lives in West Harlem, which has a smaller percentage of white people and a larger percentage of black people than New York county (i.e. Manhattan) as a whole (although West Harlem has become more white in recent years, this trend would not have been picked up as much in the 2010 Census). Also, knowing that Ben's last name is Goodrich does not tell as much about whether he is white or black as it does tell you that he is neither Hispanic, Asian, nor "other". Thus, if the analysis were conducted at the Census tract level, a little bit of the posterior probability that Ben is white would shift to black.