GR5065 Homework 1

Ben Goodrich

Due February 7, 2022 at 6PM

1 Poker

This problem is about one instance of the game poker, specifically the most popular form of poker, which is known as No Limit (on betting) Texas Hold 'Em. To understand poker in general requires years of dedication, but to analyze one instance only requires that you apply the principles of probability because you will be told the relevant rules and facts. And you can ask on Ed Discussion about anything you do not undertand.

A deck consists of 52 shuffled cards, of which there are 13 cards (2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace) ordered from lowest to highest for each of four suits (Spades, Hearts, Diamonds, Clubs). In No Limit Texas Hold 'Em, each player is dealt two cards (known as "hole cards" or a "hand") face down so that only they can see (or use) them. Five cards eventually get placed face up (known collectively as the "board") in the middle of the table. In between, there are several rounds of betting of poker chips. A person wins all of the poker chips that have been previously bet (known as the "pot") if either all of the other players fold (i.e. give up) or by using their two hole cards and the five cards in the middle to form the best five-card combination that is better than the best five-card combination of any other non-folding player at the table.

The rules of No Limit Texas Hold 'Em are explained in more (and excessive for this problem) detail at

https://en.wikipedia.org/wiki/Texas_hold_%27em

Unfortunately, poker involves a lot of jargon and words that do not even make sense in English. If you prefer an explanation in Chinese, you could look at

https://zh.wikipedia.org/zh/%E5%BE%B7%E5%B7%9E%E6%92%B2%E5%85%8B

The instance of poker we are considering in this problem was filmed in 2017

https://youtu.be/Dd_VJ6l29-k

which you should watch (ignoring any commercials that may pop up) and took place primarily between Vanessa Selbst (with short hair), Gaelle Baumann (with long hair), and Noah Schwartz (wearing a black cap). This poker tournament consisted of $7,221$ entrants (spread over more than a thousand tables) who each paid $\$10,000$ to enter the poker tournament. Each entrant was given $50,000$ worth of plastic poker chips. Players are eliminated from the tournament when they have zero chips left, and a winner is crowned when all other players are eliminated, although there are (large) cash prizes for finishing in second place, third place, etc. As can be seen on the top line in the video, after playing poker for not quite one hour, the number of chips that each player had deviated little from $50,000$, although Selbst had a bit less and Baumann had a bit more.

1.1 Probability of rare five-card combinations

Assuming neither player folds, more rare five-card combinations beat more common five-card combinations (and then there are some tie-breakers if two or more players have equally rare hands). The applicable probability distribution when dealing with a deck of cards (or when drawing uniformly from any finite set without replacement) is called the hypergeometric distribution and its probabilities are defined by the `dhyper` function in R, which computes

$$\Pr\left(x \mid m,n,k\right) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$$

where

- x is the number of "good" elements that you want the probability of drawing

- m is the number of "good" elements in the set being drawn from
- n is the number of "bad" elements in the set being drawn from
- k is the number of times you draw without replacement
- $\binom{a}{b} = \frac{a!}{b!(a-b)!}$ is the choose function, which is defined as zero if $a < b$, and `!` indicates the factorial function

Thus, before the hand starts (i.e. irrespective of any betting and presuming the player does not fold) the probability of ending up with four-of-a-kind (also known as "quads") is given by

```
dhyper(x = 4, m = 4, n = 52 - 4, k = 2 + 5) * 13
```

```
## [1] 0.001680672
```

because there are $m = 4$ of each value of card in a deck, you need all $x = 4$ of them to make four-of-a-kind, there are $n = 48$ other cards in the deck, and you have $k = 7$ opportunities between your two hole cards and the five cards in the middle. We multiply by 13 as a shortcut for adding the same probability 13 times because there are 13 distinct card values in a deck. In other words, the probability of ending up with four sevens is

```
dhyper(x = 4, m = 4, n = 52 - 4, k = 2 + 5)
```

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

so the probability of ending up with all four of some card is thirteen times that.

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?

There are several other possibilities for a five-card combination, but the relevant ones for this video are four-of-a-kind, a full house, a flush, and three-of-a-kind (which is also known as a "set" when the two hole cards are the same value and there is exactly one card with the same value in the middle).

1.2 Pre-flop action

Each of the eight players at the table was required to bet 50 poker chips (known as the "ante") before receiving their hole cards. In addition, Schwartz was required to bet an additional 150 chips (known as the "big blind") and Rosen, who was the person to Schwartz's immediate right, was required to bet 75 chips (known as the "small blind") before receiving their hole cards so that the initial pot was 625 chips.

Before the (highly edited) video starts, the person to Schwartz's immediate left folded, which is not too surprising because the optimal poker strategy — assuming the utility function is equal to the number of poker chips at the end of the hand — is to fold about seven out of eight times (and bet one of eight times) when you are the person who has to act first (which rotates when new hole cards are dealt). The conditional probability of winning with each combination of hole cards against exactly one other random hand — dealt from a deck with the 50 remaining cards — assuming no one folds is given at

https://wizardofodds.com/games/texas-hold-em/2-player-game/
although you should keep in mind that players tend to fold bad (or even mediocre) hands so the probability of winning against a player who does not fold is somewhat less than what is stated in the table (although the relative ranking is about the same). In the table, "suited" means that both hole cards are of the same suit, which means a flush can be obtained if three of the five cards in the middle are also that suit, and a person with a flush will usually win the pot (although not in this particular video).

The players proceed clockwise. Selbst was the second player to act and was dealt A♠A♦, which gives a greater probability of winning against any other hand (except A♥A♣, which has the same probability). For a person who is second to act after the first person has folded, optimal poker strategy is to bet with about one out of seven hands (and fold with six out of seven hands), and Selbst bet 400 chips, which is fairly standard and the amount Selbst would have bet with any hand among the top seventh.

The next three players all fold, which is not too surprising given that Selbst presumably has a hand among the top seventh and they would have needed a similarly good hand to "call" (i.e. match Selbst's bet of 400 chips). Baumann's hole cards are 7♥7♦, which is quite good, so Baumann calls, bringing the pot up to $1,025$ chips. Rosen, who is the player to Baumann's left (and Schwartz's right) folds with A♥6♣, which is not too surprising given that both Selbst and Baumann presumably have very good hands and A♥6♣ is only a little better than average. Note that no one else at the table knows that Rosen had A♥ (which gets superimposed onto the video for people watching at home).

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?

1.3 Schwartz's call

The video essentially starts with Schwartz who calls with J♣8♥. Since Schwartz already had to bet 150 chips as the big blind before receiving hole cards, he only has to bet 250 additional chips in order to call Selbst's bet of 400 chips. J♣8♥ is a bit worse than average, both Selbst and Baumann presumably have hands that are well above average, and to make matters even worse, Schwartz will have to act first (followed by Selbst and then Baumann) for any and all subsequent bets as cards in the middle are revealed. And yet, Schwartz's decision to call was a good (and very standard) decision. **Explain in words why it is a good decision (exact calculations are not required on this subproblem)**.

1.4 Flop action

After Schwartz calls and the pot is up to $1,275$ chips, three cards (collectively known as the "flop") are turned over in the middle and are all clubs, specifically A♣7♣5♣. Schwartz bets zero chips (known as "checking"), which is what he would do with all of his hands (that had not previously folded). Selbst now has a set between the A♠ and A♦ in her hand along with the A♣ in the middle and bets 700 chips. However, optimal poker strategy would dictate that Selbst bet that amount with all hands (that had not previously folded), so the fact that Selbst bet 700 chips right after the flop does not really tell you anything about her hole cards beyond what was implied by Selbst's bet of 400 chips before the flop.

Baumann also has a set (of sevens) and calls Selbst's bet of 700, which conveys some information because Baumann would have folded if the flop were sufficiently unfavorable in conjunction with her hole cards. But given that Selbst would bet 700 chips with every hand (that had not previously folded), it is rational for Baumann to call with a reasonably large proportion of hands (that had not previously folded).

At about 1:00 into the video, Schwarz folds despite having J♣, which means if either of the two remaining cards to be revealed in the middle were a club, he would have a flush. However, in the event that two or more players have a club flush, the player with the highest club would win the pot. As one of the commentators (who is also a professional poker player) subsequently says

> If [Schwartz] had the King of clubs [or] maybe the Queen of clubs, we could see a call, but you don't really want to call when you draw to the [Jack].

Assume it is a good decision for Schwartz to fold with the the J♣ and a largely irrelevant card like 8♥ and further assume it would have been a good decision to call with the K♣ and the 8♥. **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.

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?

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?

1.7 River action

Selbst checked immediately after the 7♠, which is not a terrible decision but is uncommon (Baumann refers to it as "kind of freaky" in the video linked in the last subproblem). Checking right after the turn would tend to indicate that Selbst either has a very strong hand or a very weak hand (among the hands that had not previously folded). Baumann then takes the opportunity to bet 1700 chips, which she would do with most of her hands (that had not previously folded). Selbst then raises to 5800 chips and Baumann calls the difference of 4100 chips. By Selbst checking and then raising Baumann's bet, it either indicates that Selbst has at least a flush or the K♣ with a Queen that is not the Q♣ (in which case the raise would be considered a "bluff" but with a chance to improve to the best flush if the last card in the middle is a club). Moreover, Baumann is fine with those few possible hole cards that Selbst could have, so Baumann presumably also has something like that.

The last card (known as the "river") is the 4♦, which is irrelevant except for the fact that it is not a club. Selbst bets and at 2:30 in the video Baumann raises by more chips that what Selbst has left $(20,300)$, bringing the pot to $66,975$ chips. At which point, Selbst starts talking in an attempt to induce some sort of a reaction out of Baumann but none is forthcoming. Although the audio is quiet, at 2:50 in the video, Selbst says about Baumann's potential hole cards

> Ace (of hearts since that is the only Ace Selbst cannot see) seven of diamonds is not a thing

By this, Selbst means that if Baumann had A♥7♦, then Baumann would have folded *before* the flop since an Ace and a seven of different suits is only a little bit better than average and is decidedly worse than the top seventh hand Selbst presumably has. Recall that Rosen actually did fold with an Ace and a six of different suits. Selbst also repeatedly speculates that Baumann could have A♥7♥, which is good enough to not fold at any round of betting and would mean that Baumann now has a full house with three sevens and two Aces, in which case Selbst would win the pot by virtue of the fact that three Aces and two sevens is better.

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.

1.8 Baumann's strategy on the river

Baumann was interviewed later that day, which can be seen at

https://youtu.be/TKGZyYknJr0

The interviewer asked Baumann if she would have raised Selbst's river bet for all the remaining chips (also known as a "shove") if her hole cards were A♥7♥ and Baumann says at 1:03 in the video

> I said yes in [the other video] but after thinking about the hand, I don't think so because [Selbst] makes such a large [bet] sizing, 16K in a 14K pot[1] ... so I would definitely just call that.

By this, Baumann means that because Selbst bet so much right after the 4♦ appeared on the river (and raised after initially checking after the 7♠ appeared on the turn), the conditional probability that Selbst had a better full house with A♠A♦ would be too large for Baumann to raise all-in but too small for Baumann to fold with A♥7♥.

If Baumann's strategy were to raise all-in only with the best possible hand, given the cards that she can see, *and if Selbst knows that*, then Selbst would fold and Baumann would win no more chips than if Baumann had called. Of course, it is possible that is Baumann's strategy but Selbst does not know that, in which case Selbst might call.

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.

2 Surnames

Read the short paper and appendix linked at

https://imai.fas.harvard.edu/research/race.html

which attempts to predict a registered voter's race given their surname (i.e. family name, which is the last name in the United States) and geolocation. The Census, which is conducted in the United States every ten years, collects data on essentially all residents, including their (self-identified) race and their last name. Thus, it is possible to calculate with negligible uncertainty, the proportion of people in each racial category that have a particular surname (s). It is also possible to condition on other demographic variables in the Census, as well as the political party the citizen registers with (which is collected by most states).

This is relevant because in most states, the race of voters is not recorded but their name, voting location, and perhaps a few other variables are recorded. Thus, in order to analyze voting behavior by race, researchers have to consider the probable race of voters, which can be higher or lower depending on their surname, geolocation, etc.

[1] Selbst actually bet $16,200$ chips on the river when the pot had $14,275$ chips in it but a discrepancy of a couple hundred chips is very negligible to the decision-making

2.1 Notation

Rewrite the denominator in equation (2) from the paper using the crossed-out notation like that we used for bowling, $\Pr\left(\cancel{x_1} \bigcap x_2 \mid n = 10\right)$. The RMarkdown syntax to strike a line through text (like ~~this~~) is to surround that text with two tildes on each side.

2.2 Frequentist Perspective

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

2.3 You

Install the wru package from CRAN, which implements the methods described in the paper. There is a dataset in the Assignments/HW1 folder that you pulled from GitHub that has the necessary county-level data for the state of New York that is needed by the wru package. You can load it into R via

```r
NYS <- readRDS("NYS.rds")
```

Predict the probability that you fall in each of the five racial categories by calling

```r
library(wru)
me <- predict_race(my_df, census.geo = "county", census.data = NYS, age = TRUE, sex = TRUE)
```

where `my_df` is a data.frame with exactly one row that contains the following columns:

- `surname`: Your family name, as a character string
- `state`: "NY" (If you do not live in New York state, pretend that you live in Staten Island)
- `county`: A three-digit character string that indicates what county you live in now; see below but note that you must include any leading zeros
- `age`: Your age in years
- `sex`: Your sex at birth as `0` for male and `1` for female

The code for the county that you live in can be obtained by executing

```r
library(dplyr)
select(NYS$NY$county, county, county_name) %>%
  arrange(county_name)
```

and looking at

https://en.wikipedia.org/wiki/List_of_counties_in_New_York#/media/File:New_York_Counties.svg

Is the mode among the last five columns of `me` the correct racial category for you?

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?