STAC58 Group Project Part 2

2025-04-10

Preliminary Data Analysis

In order to fully understand what kind of data would be useful for us to incorporate into a Bayesian Network, we performed a preliminary data analysis on a dataset compiled by the University of Alberta's Computer Poker Research Group. The specific dataset used was from the Internet Relay Chat (IRC) poker server. We used stratified sampling on **three** groups: *holdem*, *holdem*2, and *holdem*3 and examined all of the data from May 2000 for these groups. IRC groups were split by experience (moderated through having win-rate requirements) and betting pool sizes. *holdem* is a group that contains new but enthusiastic players who are familiar with the rules of Texas Hold'em poker, but aren't quite proficient yet. *holdem*2 contains players who have met the win-rate requirements and are willing to bet more money. *holdem*3 has the least amount of players, but everyone who plays in those games have high skill and the starting bets are the highest among the IRC games.

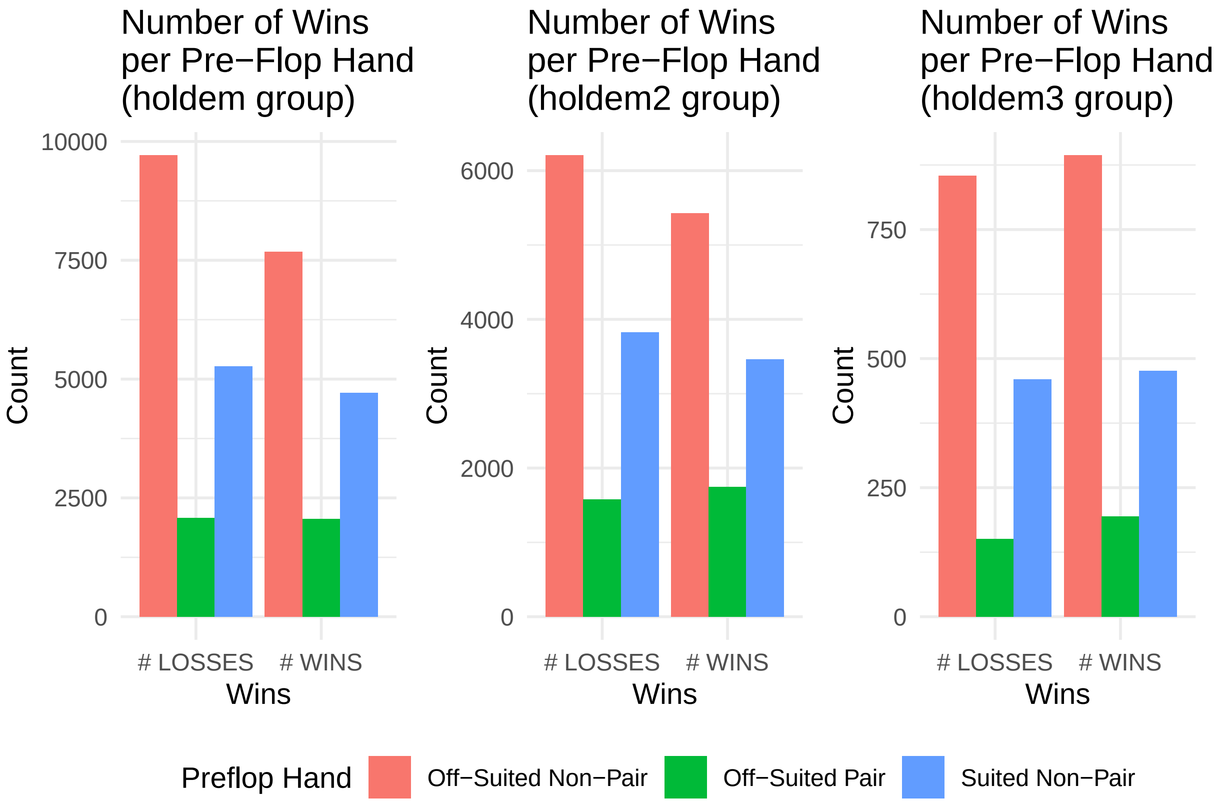
```r
library(tidyverse)
library(ggplot2)

filenames <- list.files("data/holdem")
full_paths <- paste("data/holdem", filenames, sep="/")
all_data <- lapply(full_paths, function(x) {
  read.table(x, header = FALSE, fill = TRUE, stringsAsFactors = FALSE)
})
max_cols <- max(sapply(all_data, ncol))
all_data_char <- lapply(all_data, function(df) {
  for(i in 1:ncol(df)) {
    df[,i] <- as.character(df[,i])
  }
  return(df)
})

games1 <- bind_rows(all_data_char)

games_with_preflop_1 <- games1 %>%
  filter(V8 != "-") %>%
  filter(!is.na(V12), !is.null(V12), V12 != "") %>%
  mutate(rank12 = substr(V12, 1, 1)) %>%
  mutate(rank13 = substr(V13, 1, 1)) %>%
  mutate(suit12 = substr(V12, 2, 2)) %>%
  mutate(suit13 = substr(V13, 2, 2)) %>%
  mutate(preflop = case_when(
    rank12 == rank13 & suit12 != suit13 ~ "Off-Suited Pair",
    rank12 != rank13 & suit12 == suit13 ~ "Suited Non-Pair",
    rank12 != rank13 & suit12 != suit13 ~ "Off-Suited Non-Pair"
  )) %>%
  mutate(wins = case_when(
    V11 > 0 ~ "# WINS",
    TRUE ~ "# LOSSES"
  )) %>%
  select(wins, preflop)
```

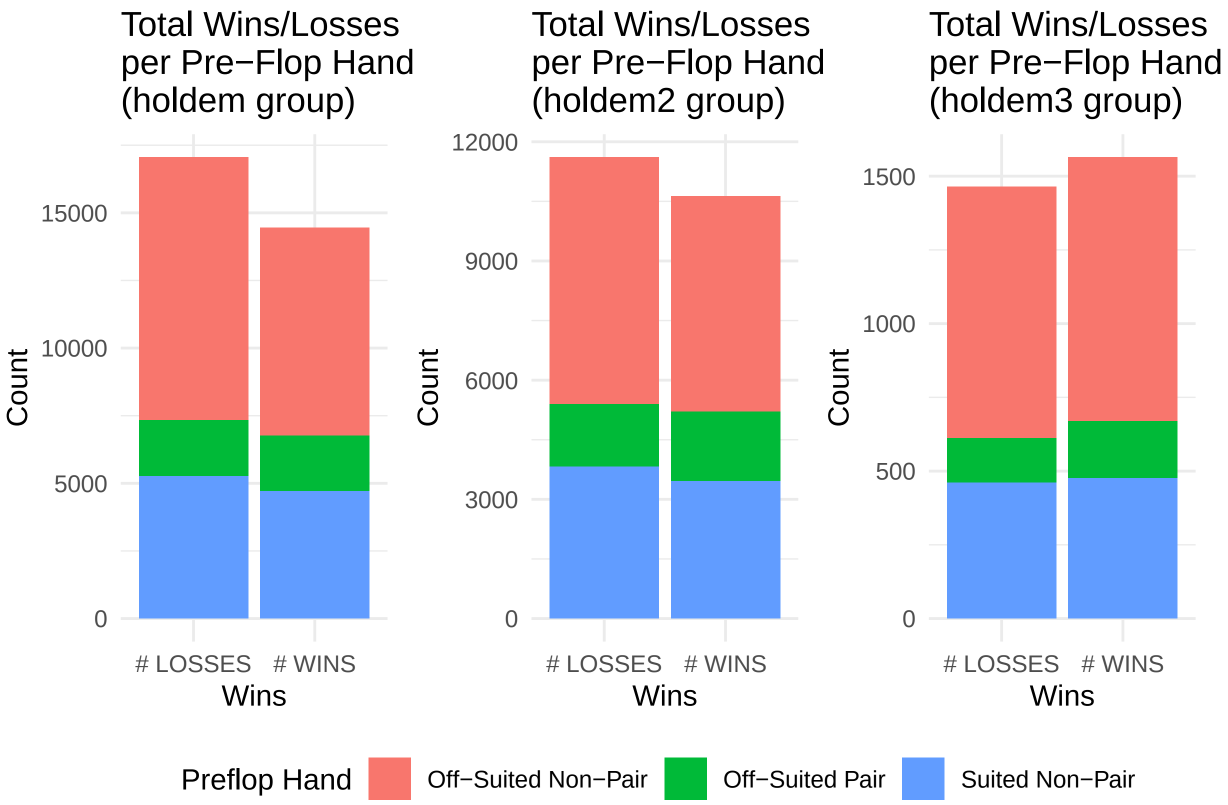
For the sake of space, code to import data from holdem2 and holdem3 have been omitted from this compiled report. However, the code for those is nearly identical.

Pre-Flop to Showdown Analysis

I first examined the chance of winning a showdown given the two hands a player is dealt at the pre-flop stage. This is to see if luck of the draw plays a significant part in win rates. In our research, we determined that different hands can be "bucketed" into categories to reduce the total number of hand combinations and make analysis more efficient. For a starting hand, the buckets are **"Suited Non-Pair"** (same suit, different rank), **"Off-Suited Pair"** (different suit, same rank), and **"Off-Suited Non-Pair"** (different suit, different rank).

```
## Warning: package 'patchwork' was built under R version 4.4.2
```





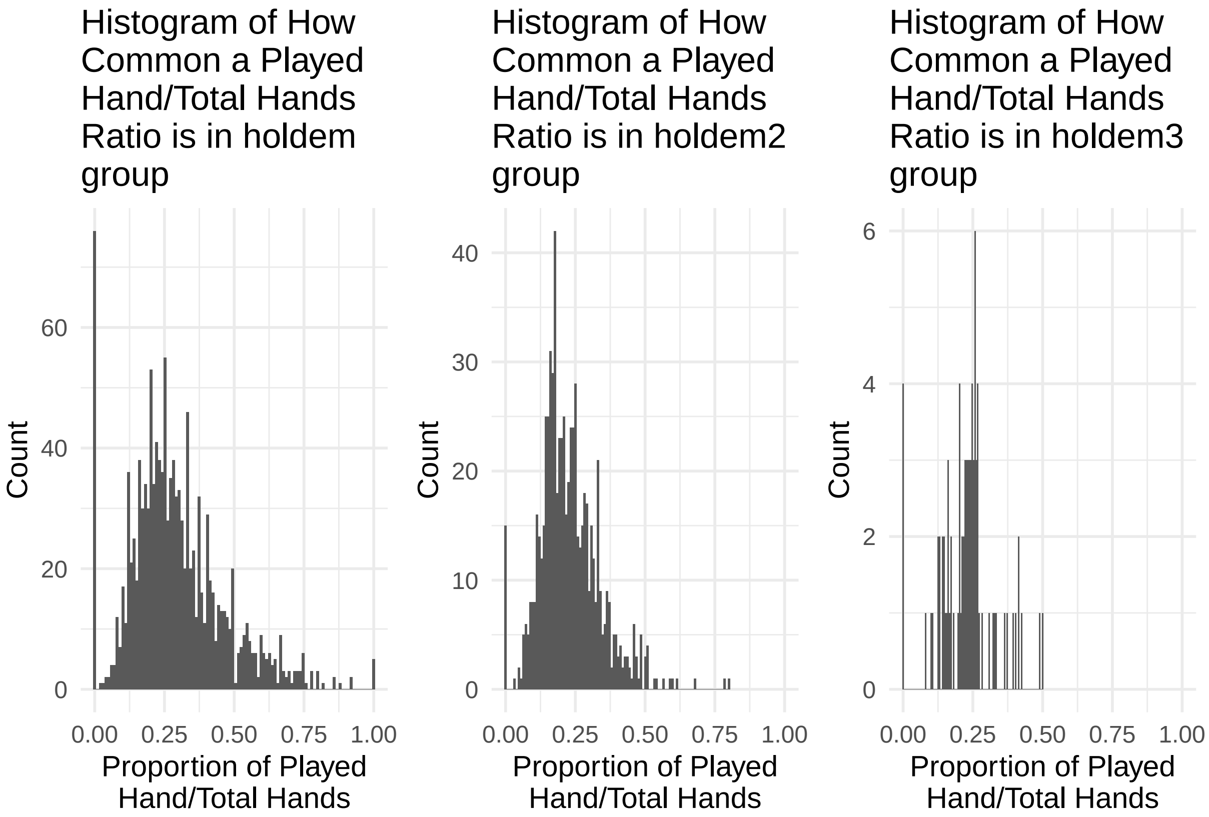
In the least experienced groups, players seem to have more reliance on the dealt hand to win because wins/losses are least consistent in the worst dealt hand, the off-suited non-pair. Meanwhile, the most experienced group seems to know how to handle the off-suited non-pair hands, whether it's by bluffing or finding a good final hand with a straight. This could also imply that more experienced players know when to fold to minimize their losses.

The rarest starting hand, the pair, seems to generally allow players to win more than they lose, which means that the starting hand *can* help a player win a showdown, but as seen in the difference between newer players and experienced players, good poker players will also be able to make the most out of a bad starting hand. Do note that this data analysis does not include any folds that occurred, and only analyzes showdowns (where both players show their hand, with the stronger one being the winner). Furthermore, we don't have the numbers of the bets nor the pools, which means that some of these losses could just be to improve the credibility of their bluffs. There could be further analysis to be performed on how often good players fold versus how often newer players fold, and whether that has an effect on the win rate. We could also analyze how often good players bluff, what kind of bluffs they perform, and how effective these bluffs are for improving win rate.

For now, let's examine the frequency of play (folding vs calling) among the different groups to see if it contributes to win rate.

Frequency of Play Analysis

```
freq1 <- ggplot(games_played_hands_1, aes(x=proportion_played))+
  geom_histogram(bins=100)+
  coord_cartesian(xlim = c(0, 1))+
  labs(
    title = str_wrap("Histogram of How Common a Played Hand/Total
                     Hands Ratio is in holdem group", width = 20),
    x = str_wrap("Proportion of Played Hand/Total Hands", width = 20),
    y = "Count"
  ) +
  theme_minimal()
freq2 <- ggplot(games_played_hands_2, aes(x=proportion_played))+
  geom_histogram(bins=100)+
  coord_cartesian(xlim = c(0, 1))+
  labs(
    title = str_wrap("Histogram of How Common a Played Hand/Total
                     Hands Ratio is in holdem2 group", width = 20),
    x = str_wrap("Proportion of Played Hand/Total Hands", width = 20),
    y = "Count"
  ) +
  theme_minimal()
freq3 <- ggplot(games_played_hands_3, aes(x=proportion_played))+
  geom_histogram(bins=100)+
  coord_cartesian(xlim = c(0, 1))+
  labs(
    title = str_wrap("Histogram of How Common a Played Hand/Total
                     Hands Ratio is in holdem3 group", width = 20),
    x = str_wrap("Proportion of Played Hand/Total Hands", width = 20),
    y = "Count"
  ) +
  theme_minimal()
combined_plot <- freq1 + freq2 + freq3
combined_plot
```



As experience increases, the desire to play hands tend to decrease, and the charts between less experienced players and the most experienced players have clear differences. It seems then, that we can assume the decision-making has a big part in how good of a poker player you are.

From our preliminary data analysis, we can determine with a decent amount of certainty that the decisions made between experienced players and new players have significant difference. While Texas Hold'em poker has an element of luck, good players seem to find more consistent winnings by being smart with their playing strategy. When to fold, when to raise, how to bluff, etc., are considerations that a good player should be making when playing the game. Therefore, if a model is to be theoretically considered for this game, we should find a way to include all of these parameters that should influence our decision.

To do this, our research has concluded that one way to do this is by using Hidden Markov Models (HMMs), which is a time series analysis that helps us model opponents' strategies over time and infer their hand strength based on their aggressiveness in betting compared to their hand strength, their tendency to fold, and other such patterns. However, we found that we can also perform such analyses through Bayesian Networks, so we will primarily focus on those. The variables and charts that we already examined are examples of possible variables that can be included in a Bayes net. With many variables and lots of observations of an opponent, we should be able to model their behaviour and improve our poker strategy.