# Homework 3: Do Trivia Nerds Cheat?

*Neeraj Sharma*

*05/19/2020*

## Question 1: Clean data and report summary statistics of percent correct answers by year and round of the championship, as well as when these players are on the honor system. Provide hypotheses as to why these summary percentages might vary.

In order to clean this data set, the first thing I do is convert it from a wide data structure to a tidy structure. Several relational variables like name, year, round, and honorsystemcorrect remain fixed, but I pivot questions vertically to make question number a unique variable. The second cleaning action I perform is I convert the data type of the question number column I just created by removing the "Q" at the beginning of each entry and coercing the datatype. The third modification I perform to clean the data is I coerce all "-99" observations in the `honorsystemcorrect` column to be NA. Often times, -99 is code for missing values.

With the cleaned data, I was able to get insight into the aggregate performance of competitors in each year.

Table 1: Summary Statistics of Correct Answers by Year and Round of the Championship

| | 2018 | | | | | | 2019 | | | | | |
| | Questions Overall | | Individual Performance | | | | Questions Overall | | Individual Performance | | | |
| Round | N | % Correct | Min | Max | Mean | St Dev | N | % Correct | Min | Max | Mean | St Dev |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1668 | 43.0% | 0 | 12 | 5.16 | 2.81 | 1968 | 64.5% | 3 | 12 | 7.74 | 2.12 |
| 2 | 1668 | 51.0% | 0 | 11 | 6.12 | 2.75 | 1968 | 47.9% | 1 | 12 | 5.74 | 2.43 |
| 3 | 840 | 60.4% | 3 | 11 | 7.24 | 1.99 | 984 | 60.7% | 2 | 12 | 7.28 | 2.20 |
| 4 | 840 | 45.4% | 0 | 11 | 5.44 | 2.07 | 984 | 50.2% | 2 | 12 | 6.02 | 2.18 |

Table 2: Summary Statistics of Honor System Performance by Year and Round

| | 2018 | | | | 2019 | | | |
| Round | Mean | Median | Min | Max | Mean | Median | Min | Max |
|---|---|---|---|---|---|---|---|---|
| 1 | 79.7% | 79.8% | 65.3% | 92.7% | 79.5% | 79.6% | 64.8% | 92.7% |
| 2 | 79.7% | 79.8% | 65.3% | 92.7% | 79.5% | 79.6% | 64.8% | 92.7% |
| 3 | 80.7% | 80.9% | 65.3% | 92.7% | 80.7% | 80.1% | 64.8% | 92.7% |
| 4 | 80.7% | 80.9% | 65.3% | 92.7% | 80.7% | 80.1% | 64.8% | 92.7% |

In general, people who made it to later stages in the competition scored better in the regular season. This makes sense because you would expect the elimination championship format to only yield players who were actually good at trivia.

Across the board, however, the percentages in the championship were significantly lower than the percentages in the regular season. One reason this could be is that the questions were much harder, so people got them wrong more. Another reason could be that cheating was disallowed so people didn't have that added boost inflating their honor system scores. Furthermore, the questions asked in round could be poorly calibrated and skew in obscure directions because of the small sample size in the championship. This problem doesn't exist in the regular season because the number of questions is higher. Finally, the competitive, high

pressure environment of the championship could cause some people to crack and slump dramatically in the championship.

## Question 2: In the first two championship rounds, estimate how difficult the championship questions are relative to the regular season questions for people who play honestly during the regular season. Suggest at least two strategies for coming up with such an estimate. Be extremely explicit about the assumptions for each of your strategies to yield truthful estimates. Given the likely violation of your assumptions, say whether your estimates overestimate or underestimate the true amounts of cheating.

In office hours with Eric on 5/18, there was some confusion about the interpretation of this question. I interpret this question to ask us to find two distinct methods of identifying cheaters and then to come up with way to identify the increase in question difficulty based on the subset of fair players.

I think this this approach to this question makes a lot of sense. In order to quantify the impact that a change has on a sample group, one needs to only change that variable and keep everything else as stable as possible. In this case, cheaters have two dimensions of change: they go from cheating to not cheating, and "easy" questions to "hard" questions. For fair players, they only have one dimension of change: they go from easy questions to hard questions. Because fair players only experience one type of shift, the difference between their honor system performance and championship performance has only one motivating factor, while the cheaters have two factors driving differences in their performance.

Thus, identifying fair players is a vital and technical step to understanding how difficult these two phases of competition are. Once I have controlled for the participants' cheating tendencies, I can evaluate the magnitude and impact of the different types of questions in isolation.

I make two overarching assumptions that are reasonable, but I would like to formally state them just for completeness.

1. I assume that people try as hard as possible in the championship rounds and that their performance there is closely representative of their true potential.
2. I assume that people only cheat in ways that will improve their performance. This rules out Google giving people wrong answers more frequently than they would give wrong answers themselves.

### Step 1: Identify Cheaters (Three Strategies)

**Approach 1: Superior regular season performance is evidence of cheating**

One possible strategy to distinguish between cheaters and non-cheaters is by calculating each person's percentage of correct answers given in the first two rounds and comparing that number to their honor system answers. If we assume that their scores in live competition represent a firm upper limit on their ability, than any individual who performed better at home must have cheated.

The assumption that people who perform better at home than in competition cheat has several issues that cause this method to overestimate the amount of cheating. First, live competition is more stressful due to the stakes and pressure, so these external factors can impact performance. Second, this assumes that the sample selection in competition is equally suited for each person which is unlikely given the small sample size. In reality, any skew within the sample can cause the championship performance to drop slightly below the honor system performance. In summary, people could very reasonably under perform in the live round and not be cheaters. As a result, I believe this method will greatly overestimate the amount of cheating and greatly underestimate the difficulty of questions.

**Approach 2: High honor system scores but low scores on easy championship questions is evidence of cheating**

Honestly, if you are messing up on lots of easy questions, you are bad at trivia. People that are bad at trivia but appear to be good at trivia over a large sample size must regularly cheat. Thus, if I can identify people that fail to properly answer even the easiest questions in competition but have high success rates at home, they must have cheated at home.

One assumption I make is that there is a large enough sample of easy questions in rounds 1 and 2 of 2018 and 2019 that are perceived to be equally easy by all participants. Because all players have high accuracy on the honor system which spans questions of all difficulties, I expect that truthful players will answer easy questions correctly while cheating players will fail. I am unable to evaluate the cheating-ness of occasional cheaters because medium difficulty championship questions do not have the same filtering effect. Good players and bad players alike can fail on medium questions, but only bad players will fail on easy questions. This is why I only analyse failure rates on easy questions. Second, I assume that only those who under perform on easy questions cheat which rules out an entire class of people who are decent at trivia, but still cheat. This relates to the previous point, but has the effect that it means I remove potential cheaters from consideration.

Because I am unable to analyze "decent" players with this method (of which some percentage cheat), I will overestimate the amount of cheating with this method.

**Approach 3: Segment population into cheaters and fair players via T-Testing**

An optimal strategy is to identify people whose play differs (statistically) significantly in the tournament from their play on the honor system. Specifically, I segment people by performing a one-tailed one-sample t-test for every individual. This t-test is one tailed because I am only interested in people who perform significantly worse in the championship than at home. Superior performance in the championship compared to performance at home is legitimate due to the live broadcast. I perform a one-sample t-test because I assume that the number of questions they answer at home is sufficiently large to approximate a population mean for that individual person. This is realistic, as Professor Levitt indicated that every participant had answered hundreds to thousands of questions at home during the regular season. In summary, these t-tests will tell us is if an individual's performance in competition is in line with their performance at home, or if their performance in competition is statistically significantly worse than their performance at home.

I think this is the most accurate representation of the amount of cheating and difficulty of questions. However, I think this approach might slightly underestimate the amount of cheating because of how strong of a binary it creates between players. This assumes that players who play fairly play fairly 100% of the time and players who cheat also cheat 100% of the time. There are some work arounds I implement which I discuss later on to attempt to measure the intermediate amount of cheating, but as far as this approach goes mathematically, it only draws a concrete binary. The true distribution of cheaters is likely much more smooth than the one I construct, so I also then overestimate the true amount of cheating.

## Step 2: Difficulty between championship questions relative to the regular season questions (1 Strategy)

As discussed above, fair players only experience one shift in going from the regular season to championship. Thus, any difference in the mean values of these two is only due to different types of questions, not the type of player they are. What this yields is a percentage difference between the questions in the championship and at home.

# Question 3: Report your findings from the strategies in question 2.

Hint: the most sensible way to report your findings would be a predicted value for the percent of questions you would predict each player have gotten correct over the first two rounds if they were not cheating in the regular season.

For each strategy you used, answer the following questions:

- What is your estimate of the average percent of questions that are cheated on for the entire group during the regular season?
- What percent of the players do you think cheat on at least 3 percent of the regular season questions?
- How many individual players can you say cheat with a high degree of confidence?

## Approach 1

Graphing all individuals by their correctness in the first two rounds and their honor system percentage allows us to visualize this model.
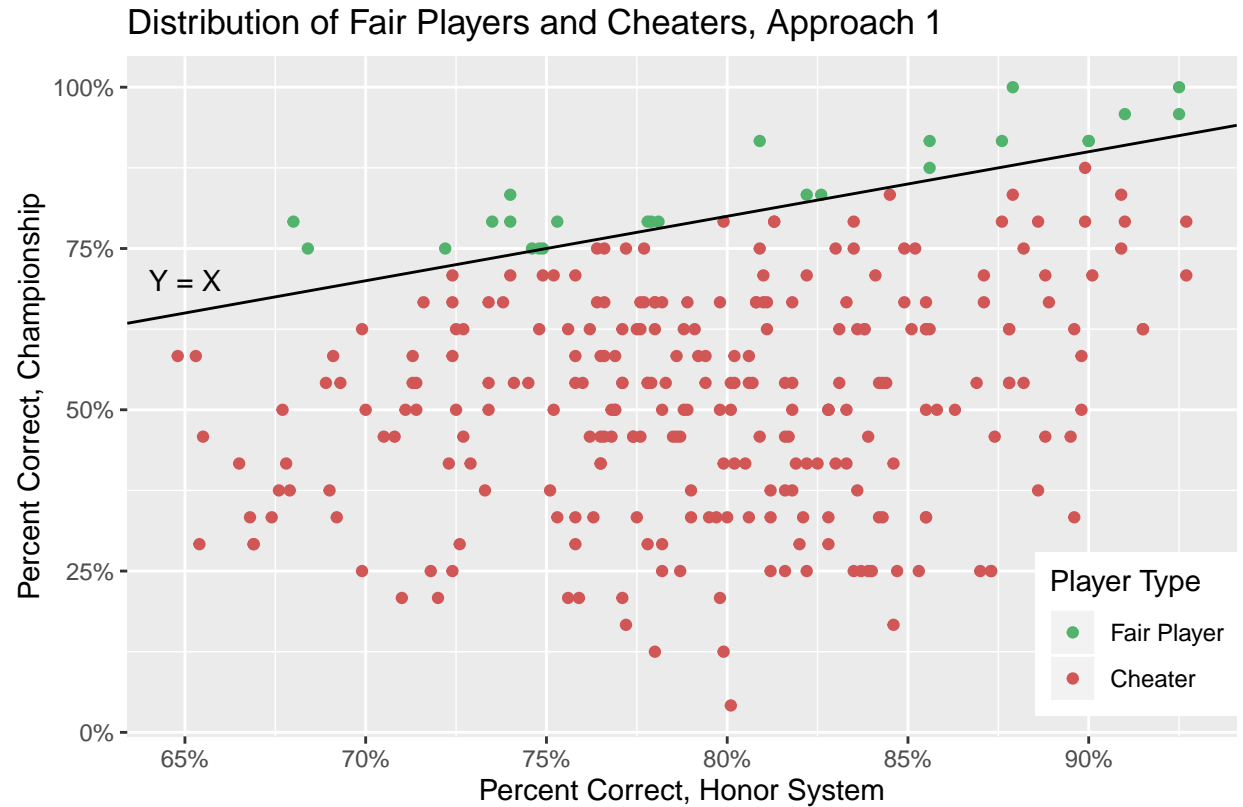
### Distribution of Fair Players and Cheaters, Approach 1



Source: LearnedLeague

Based on this partitioning of the data the mean percentage of correct answers given by honest players at home is 80.5% and the mean percentage of correct answers given by honest players in competition is 84.7%. This implies that the treatment effect of going from home questions to competition questions led to an increase in performance. Specifically, it implies that competition questions are **-4.2%** harder than home questions. That means they are actually easier.

- What is your estimate of the average percent of questions that are cheated on for the entire group during the regular season?
  - The `y = x` line provides an implicit upper limit on how one can score. If we accept the assumption that each person's championship score is representative of their true ability, then each individual's adjusted regular season score would be their championship score. The mean difference between the regular season and championship accuracy rate approximates the average percent of questions that are cheated on for the entire group during the regular season. That number is **28.6%**.
- What percent of the players do you think cheat on at least 3 percent of the regular season questions?
  - To find the percent of the players that cheat on at least 3% of the questions, I need to find the number of players who have a championship mean that is not within 3% of their honor system

mean. I count 10 individuals who meet this criteria. That means that **87.6%** percent of all players cheat on at least 3% of questions.

- How many individual players can you say cheat with a high degree of confidence?
  - Because of how limited this approach is, it's impossible to distinguish between players that certainly and possibly cheat. A core assumption is that no player can over perform in the regular season, so because of how cut and dry that assumption is players that over perform are automatically labeled cheaters. Thus, **258** players cheat according to this approach.

## Approach 2

I calculate a "difficulty coefficient" for each question in the first two rounds of 2018 and 2019 by calculating the percentage of correct answers given, and then taking \frac{1}{\text{% Correct}}. This means that difficult questions are weighted more heavily than easy questions. Summing up the "difficulty coefficients" of each question a person gets right gives them a score of how good they are at trivia based on the assumption that good players get lots of hard questions right.

The next step is to quantify which questions are uniquely difficult. I calculate the mean and standard deviation of the correct percentage for each round in each year. Those questions that are 1 standard deviation above or below I denote as easier or harder, respectively. Players that fail easy questions likely cheated.

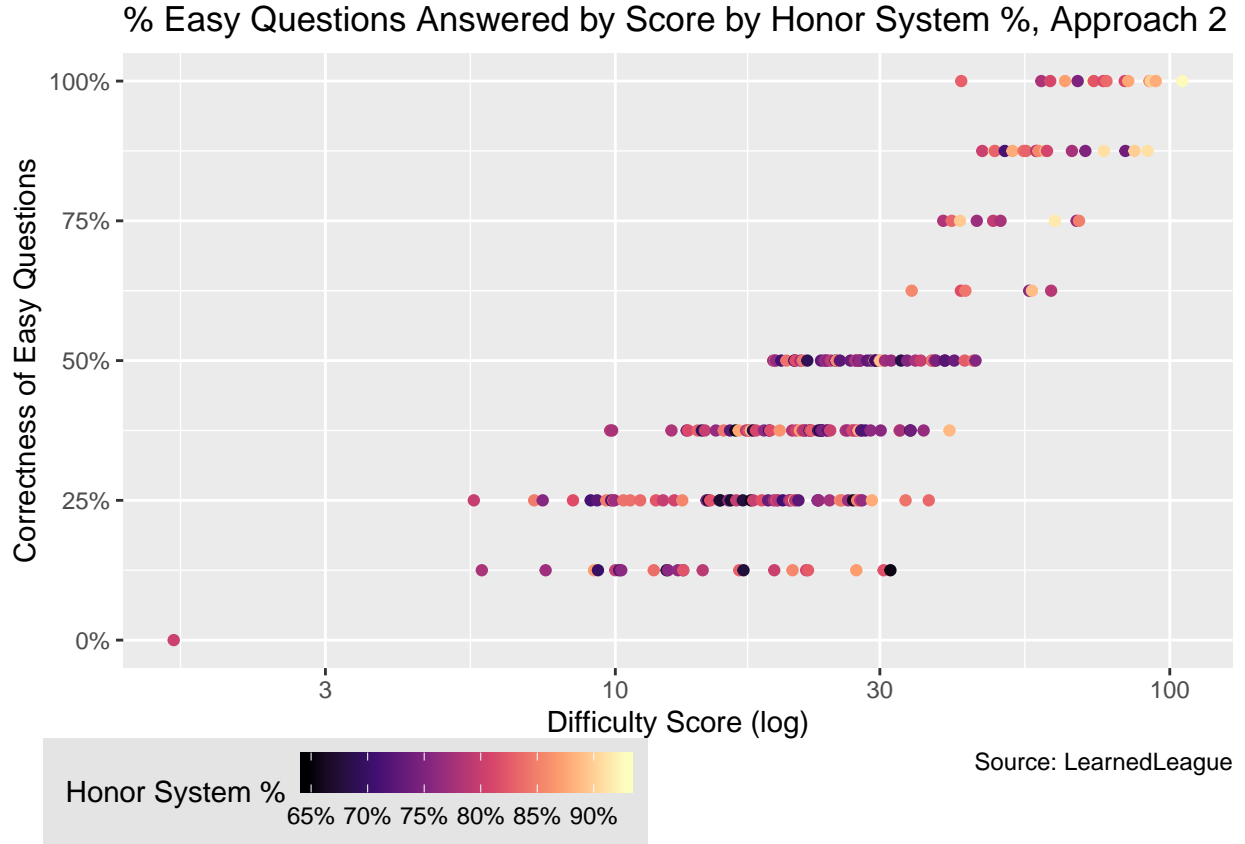### Difficulty of Questions by Round and Year, Approach 2
Bands are +/− 1 Standard Deviation from Mean
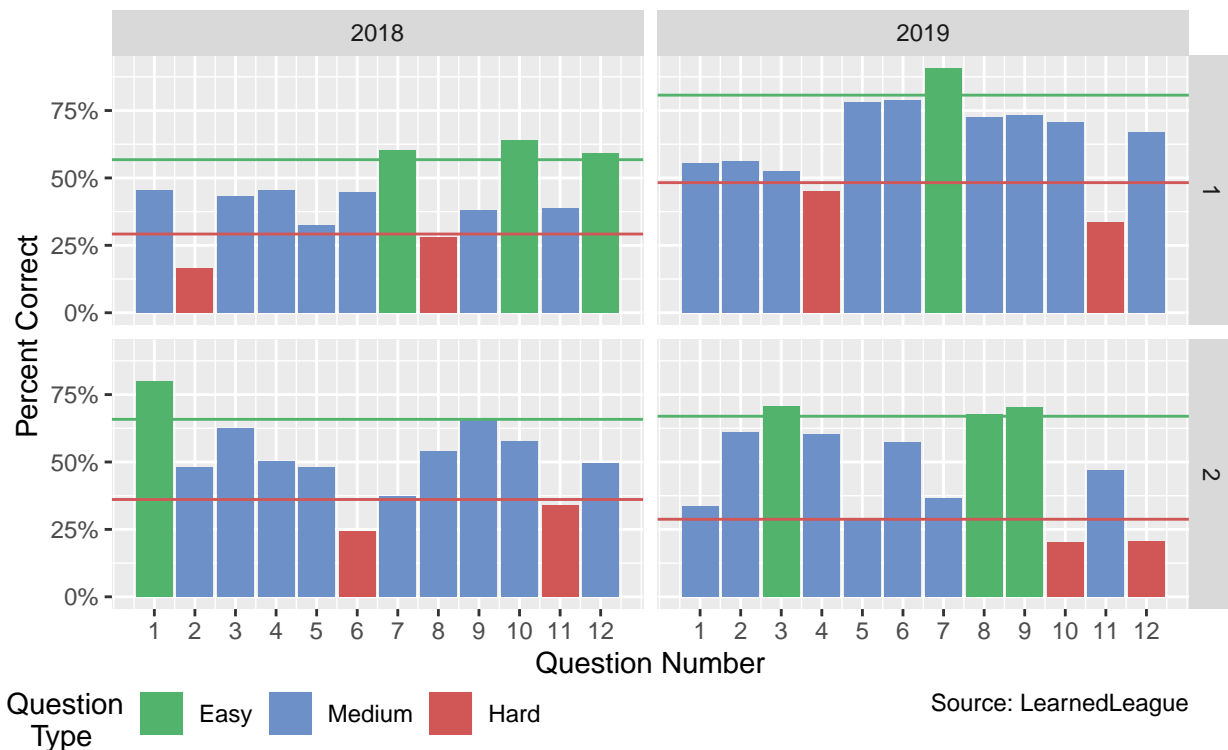


Having identified which questions are easy and which are hard, I can perform analysis to understand how individuals perform on the easy questions versus their difficulty scores of all questions the answered correctly versus their honor system percentage to make predictions about who is cheating.

Table 3: Worst 5 Performers by Difficulty Score

| Name | % Correct, Honor | % Correct, Easy Championship | Difficulty Score |
|---|---|---|---|
| VolkA | 80.1% | 0.0% | 1.598 |
| LutinM | 79.9% | 25.0% | 5.556 |
| di GiovanniA | 78.0% | 12.5% | 5.744 |
| VanceA | 84.6% | 25.0% | 7.139 |
| GirolamoA | 75.6% | 25.0% | 7.396 |

## % Easy Questions Answered by Score by Honor System %, Approach 2



Source: LearnedLeague

I identify cheaters by selecting those who perform poorly on easy questions and overall given their stated percentage on the honor system.
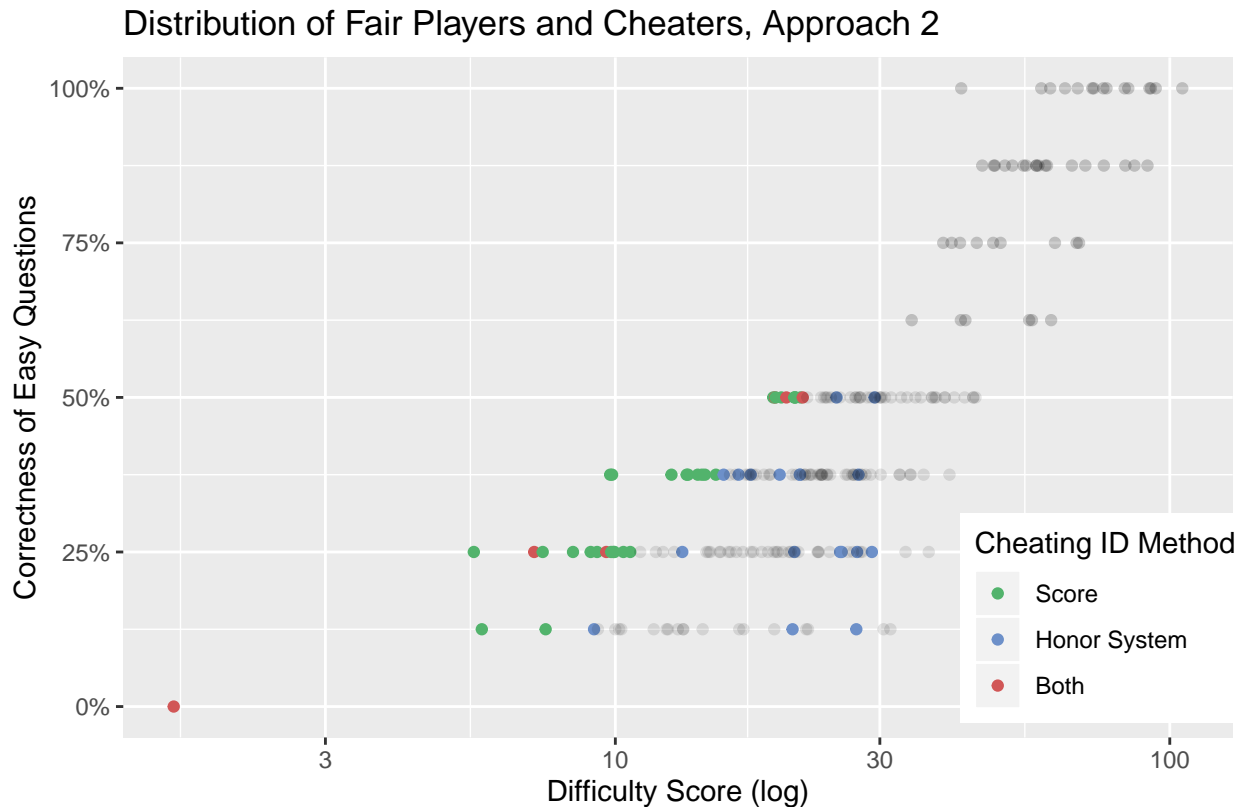
**Distribution of Fair Players and Cheaters, Approach 2**
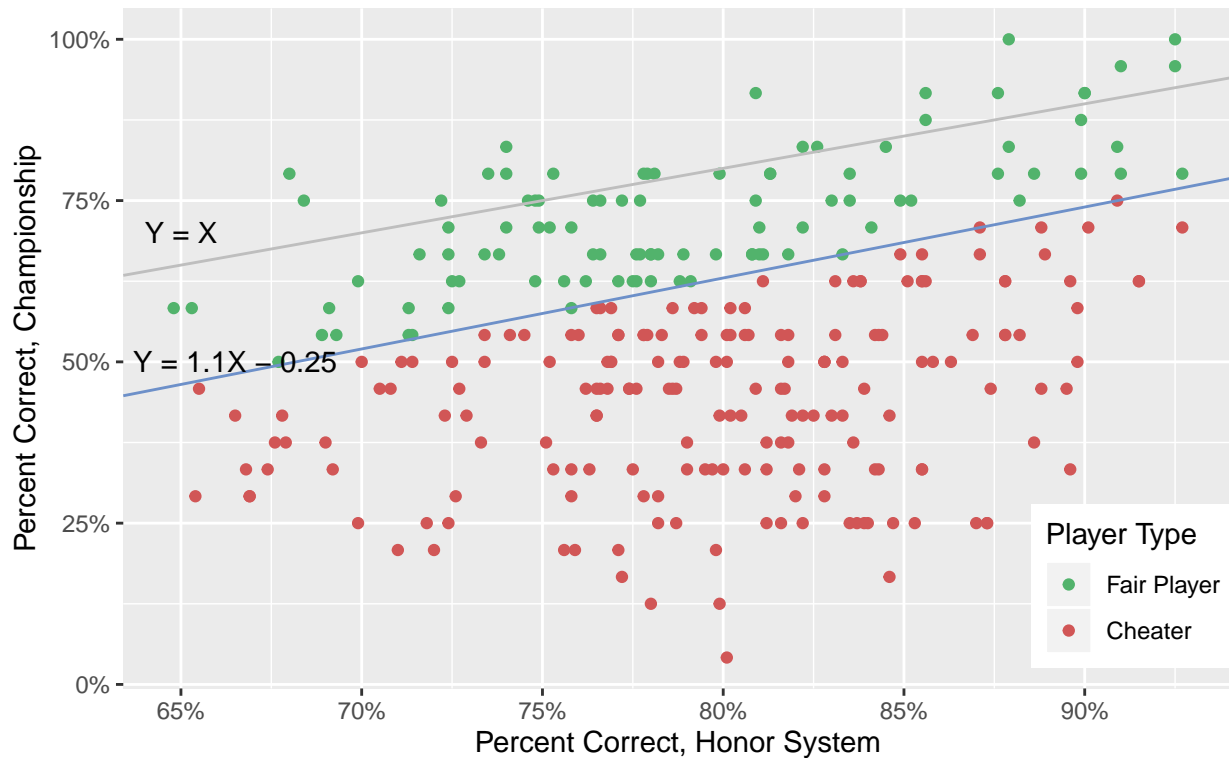
Source: LearnedLeague

Based on this partitioning of the data, the mean percentage of correct answers given by honest players at home is 79.2% and the mean percentage of correct answers given by honest players in competition is 57.7%. This implies that competition questions are **21.5%** harder than home questions.

- What is your estimate of the average percent of questions that are cheated on for the entire group during the regular season?
  - Based on this method, I project that 44.3% of questions were cheated on.
- What percent of the players do you think cheat on at least 3 percent of the regular season questions?
  - This approach grossly dichotomises players. It does a very poor job at evaluating how many questions people cheat on individually. I wouldn't recommend using this approach to answer this question, but for completeness's sake, I estimate that 52 players cheat on 3% or more questions.
- How many individual players can you say cheat with a high degree of confidence?
  - Similarly, this approach does not distinguish well with confidence intervals. I estimate that 52 cheated confidently.

## Approach 3

With this method, I run a t-test for every person to see if their performance in the championship could have come from their performance in the year. If their championship score was significantly worse, I identify them as a cheater. This is what that breakdown looks like with a dividing line plotted.

## Distribution of Fair Players and Cheaters, Approach 3



Source: LearnedLeague

Based on this categorization, the mean correct answers given by honest players at home is 78.9% and in competition is 72.6%. Thus, the treatment effect of going from home to competition questions is **6.3%**. Competition questions **6.3%** harder.

Showing the breakdown of accuracy by individual type shows that under this approach, fair players in the regular season actually fair much better in the postseason than cheaters. This make sense.

## % correctly answered in Championship by Player Type, Approach 3
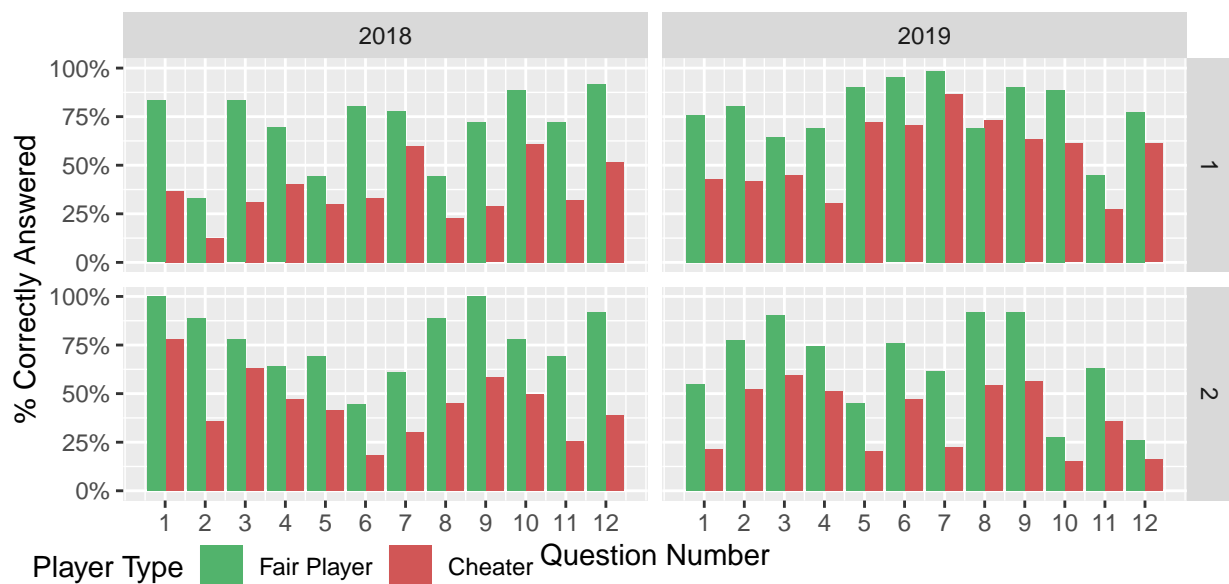


Source: LearnedLeague

- What is your estimate of the average percent of questions that are cheated on for the entire group during the regular season?
  - With this method, the `y = 1.1x - 0.25` line provides an implicit upper limit on how far away one's score can be in the regular season given their performance in the championship. If a person fails every question in the regular season but then gets a perfect score in the championship, this model predicts that the individual is actually more intelligent than their performance in the regular season would suggest. However, if a person gets 100% in the regular season then incorrectly answers every question in their championship, then they definitely cheated on 100% of questions. The mean difference between the adjusted regular season and championship accuracy rate approximates an upper limit on the percent of questions that are cheated on for the entire group during the regular season. That number is **17.3%**. See Appendix 1 for more.
- What percent of the players do you think cheat on at least 3 percent of the regular season questions?
  - To find the percent of the players that cheat on at least 3% of the questions, I need to find the number of players who have a championship mean that is not within 3% of the upper limit line. There are 17 players within 3% of the limit line and 98 that are above the limit line. Thus, I believe that **59.4%** of players cheat on at least 3% of all questions. Basically, most people cheat a little bit as evidenced by my answer to this question, but very rarely do people cheat a lot because that would drag my answer to the previous question up substantially.
- How many individual players can you say cheat with a high degree of confidence?
  - When I run this at a 95% confident interval, I get that **185** people cheated. If I increase this even further to 0.01 to get the highest of the high degree of confidence I get that **136** individuals definitely cheated.

## 4. Explain why it is easier or harder to make claims about the aggregate amount of cheating in a sample versus identifying individual cheaters.

Identifying individual cheaters was a relatively straight forward process. I propose one bad and two decent (in my opinion) strategies for separating cheaters from fair players in the data set. For individual cheaters, we are able to compare their personal scores to everyone else's scores. This allows us to do a sort of sample/population style statistical comparison. Being able to leverage the tendencies of the central limit theorem means that we are able to generate significantly more powerful predictions than without it.

When we are evaluating the amount of cheating in the population as a whole, we are unable to substantively compare it to anything else. One could employ some form of folded cross-validation technique, but that is inappropriate here because of the lack of extensive honor system data beyond the mean. I'm unsure of the variance within an individual sample, but can get a general sense of it in the population. This means that it is very difficult to estimate the aggregate amount of cheating in a sample of the population compared to evaluating if an individual cheats or not.

## 5. The players with -99 for honor code scores dropped out of the league after making one or both championships. Can you make any inferences about whether they cheated more or less than the players who have remained in the league, despite the fact you know nothing about their percent correct in the regular season?

I've identified the characteristics of what a cheater looks like so I can see what these people share in common with the cheaters. I constrain my answer to this question to only the the model I outline in Approach 3 because I believe that is the most robust approach. Using that model, I estimate that not a single person dropped out of the tournament was as a fair player. Not a single fair player had a similar percent of correct

answers given which makes me believe that their data points would fall in the cheaters category.
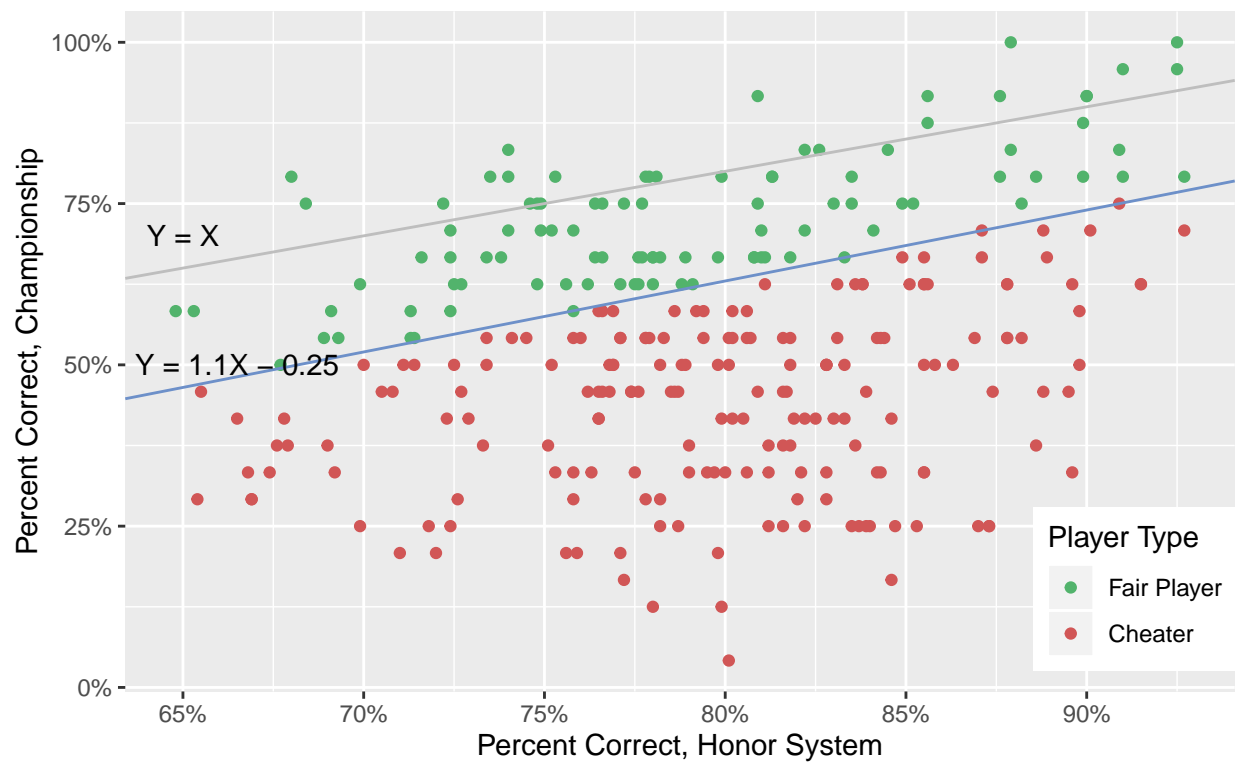


### Where do Competition Dropouts Fall? Compared to Approach 3.

Player Type  ● Fair Player  ● Cheater

Source: LearnedLeague
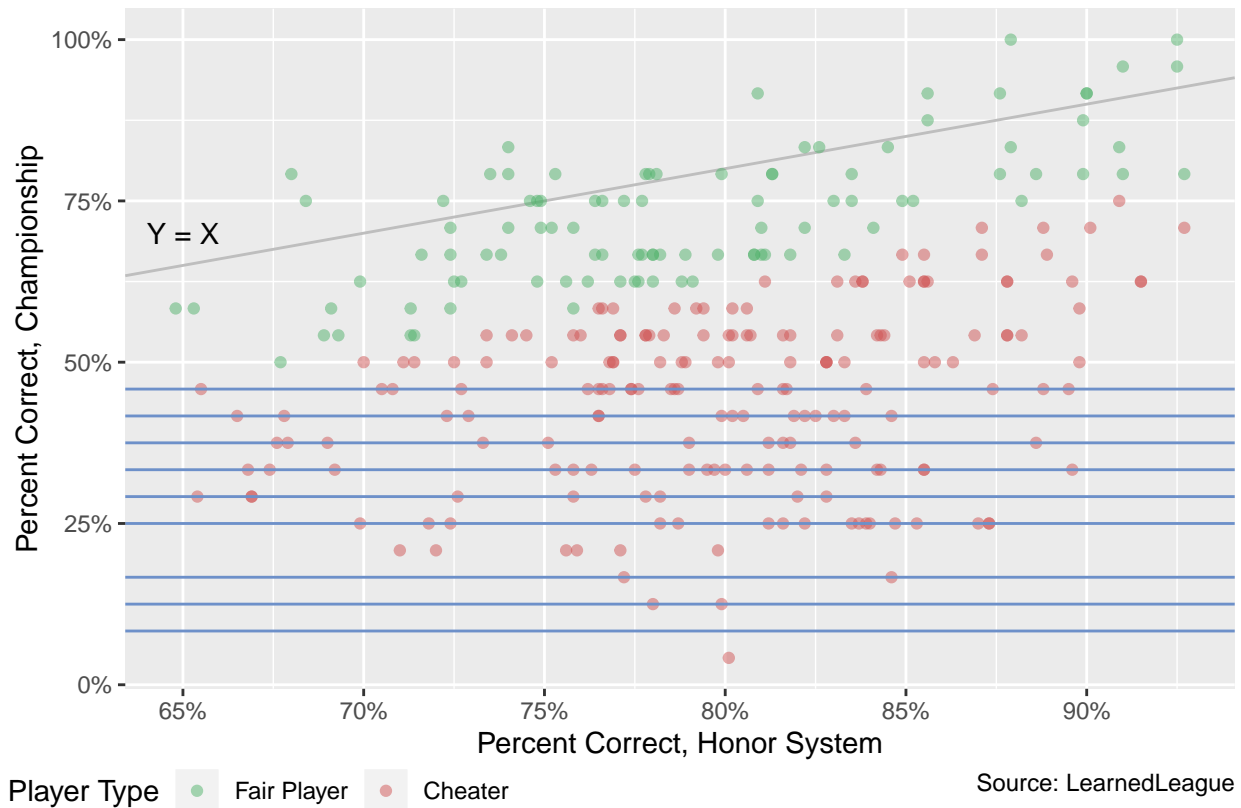
## 6. The person who runs this league is interested in learning your findings. Create one visual that you think best would summarize your insights showing the amount/non-existence of cheating in his league.

I think the most straightforward visualization to model this is the modeling scatter plot I produce in the third approach. I include that again here.
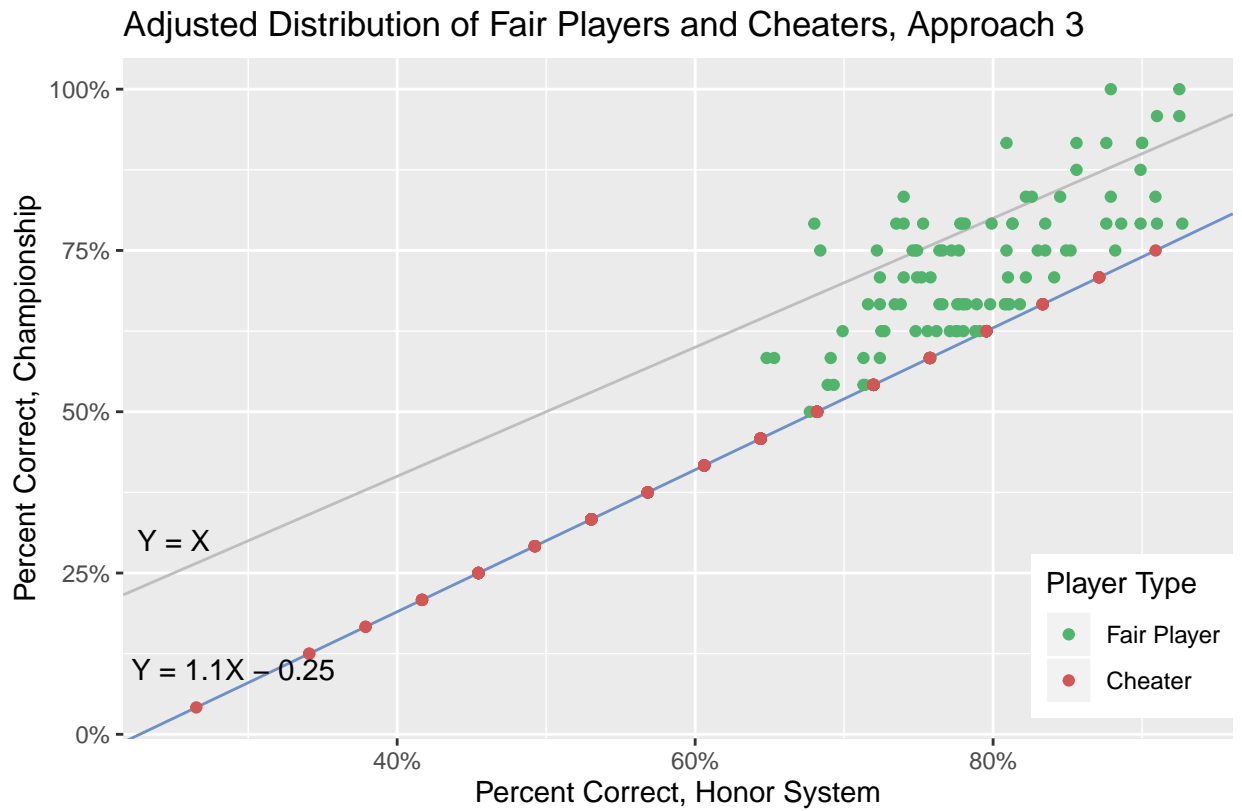
# Distribution of Fair Players and Cheaters, Approach 3



Y = X

Y = 1.1X − 0.25

Percent Correct, Championship

Percent Correct, Honor System

Player Type
- Fair Player
- Cheater

Source: LearnedLeague

# Appendix

1

## Adjusted Distribution of Fair Players and Cheaters, Approach 3



Source: LearnedLeague