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

In both years, the round to round percentage of questions answered correctly fluctuated

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%

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.

I organize my strategies in two distinct steps as each step addresses a different part of this question.

The first step is to identify which people are cheaters and which people play fairly. This is important because I need to isolate a group of people that only experience one dimension of change to quantify the impact that change has. For cheaters, they go from cheating to not cheating, and "easy" questions to "hard" questions. That is a total of two changes. For fair players, they only go from easy questions to hard questions. That is only one change. Because fair players only experience one type of shift, the difference in their performance between honor system and live game play is only due to the increased difficulty of the questions. Thus, identifying these players and calculating the rate at which they miss questions in the regular season versus championship is key to understanding how difficult these two phases of competition are.

The second step is to actually evaluate the difficulty of championship questions relative to regular season questions. Once I have controlled for the participants' cheating tendencies, I can evaluate the magnitude and impact of the different types of questions in isolation.

TWO STRATEGIES FOR PICKING CHEATERS, 1 STRATEGY FOR PICKING DIFFICULTY.

Step 1: Identify Cheaters (Two Strategies)

Approach 1: Assuming uniform cheating

One approach to estimate the increase in difficulty of championship questions compared to regular season questions begins with the assumption that all individuals play (un)fairly at the same rate. What this would mean is that every player cheats on questions as much as any other given player. It doesn't mean that all players are cheaters 100% of the time or all players live and die by the rule book, but whatever behavior one individual exhibits is shared by all others. Essentially, all players are uniform and benefit from cheating proportionally.

Because this assumption of uniformity flattens out the skilled aspects of trivia, this approach has limited potential to yield an accurate estimate. All individuals are boosted equally by cheating so the net delta between mean correct answer percentage of championship and regular season questions remains fixed. However, parsing out the portion of the net delta that is due to the boost provided by cheating and the drag created by harder questions is difficult to calculate. If questions are equally difficult in both rounds, than the increase in the mean of the honor system metric is solely due to cheating. In order to calculate this

Approach 1 assumes that all players express the same propensity to cheat uniformly across the population. Obviously, this is a very flawed approach as it assumes all individuals are uniform which is unrealistic. Given that in the real world, some players cheat and others play fairly at variable rates, the total number of cheaters is lower in reality than this model would project. As a result, this method will likely overestimate the true amount of cheating.

Approach 2: Assume superior regular season performance is evidence of cheating

A more nuanced strategy is to distinguish between cheaters and non-cheaters 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.

This assumes highly optimized behavior because it requires players to achieve and exhibit their full potential in a competitive environment.

This assumes that every player is playing at their full potential in the first two rounds of the championship. Secondly, it assumes that

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.

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

Approach 1: Difference between regular season and championships

Approach 1 assumes that all players express the same propensity to cheat uniformly across the population. Obviously, this is a very flawed approach as it assumes all individuals are uniform which is unrealistic. In essence, it assumes all players cheat at exactly the same rate. However, because some players cheat and others play fairly at variable rates, the total number of cheaters is lower in reality than this model would project. As a result, this method will likely overestimate the true amount of cheating.

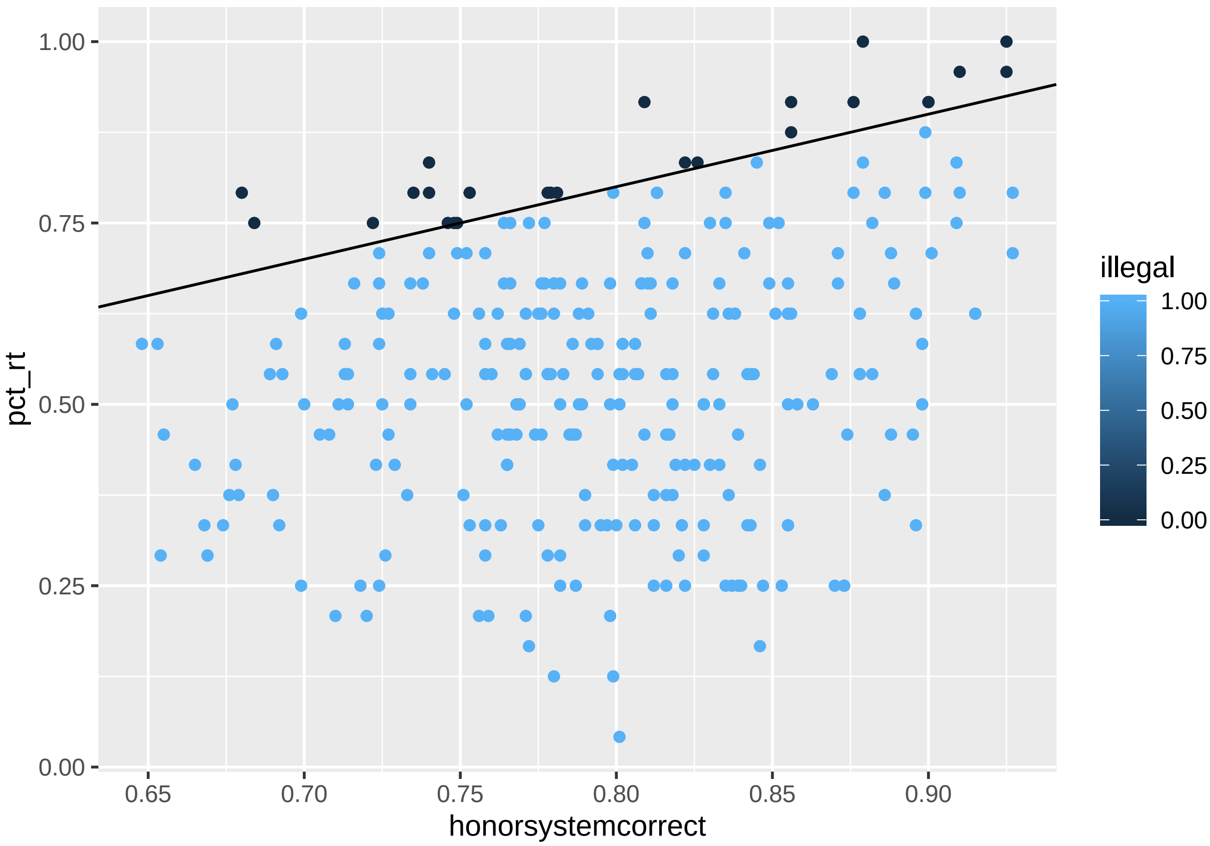
Question 3: Report your findings from the strategies in from question 2. (Hint: the most sensible way to report your findings would be a predicted value for the percent of questions you would predict each player have gotten correct over the first two rounds if they were not cheating in the regular season.). 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: Difference between regular season and championships

By this approach, 80.2% of regular season questions were answered correctly and 52.7% of championship questions were answered correctly. Assuming uniformity, that means that the championship questions were 27.5% harder. Given questions are 27.5% harder, A way to project is the minimum between the percent the answer correctly adjusted for difficulty (% correct + difficulty) and the number of questions actually answered correctly.

Approach 2

```
edited %>%
  filter(round %in% c(1, 2)) %>%
  group_by(year, name, honorsystemcorrect) %>%
  summarize(pct_rt = mean(ans, na.rm = TRUE)) %>%
  drop_na() %>%
  mutate(illegal = if_else(honorsystemcorrect >= pct_rt, 1, 0)) %>%
  ggplot(aes(x = honorsystemcorrect, y = pct_rt, color = illegal)) +
  geom_point() +
  geom_abline(slope=1)
```

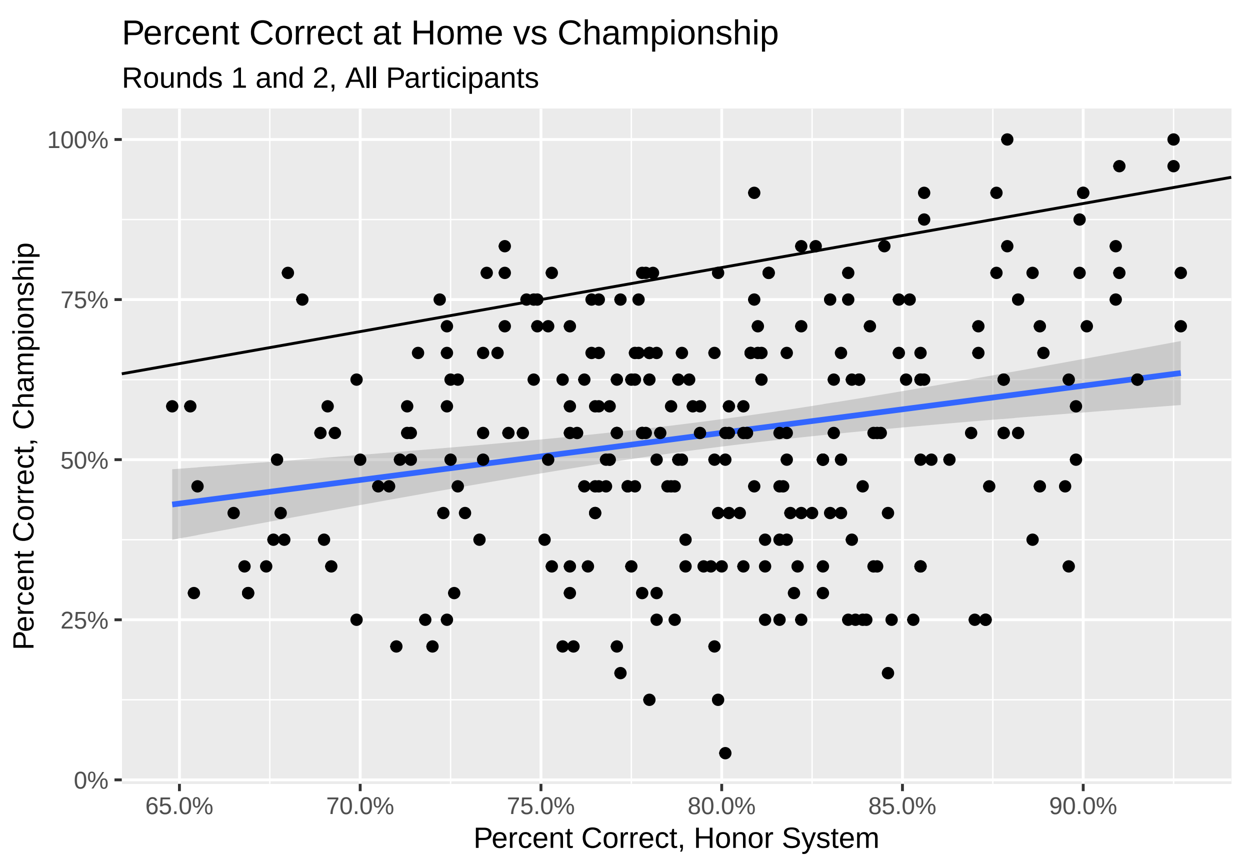



```
edited %>%
  filter(round %in% c(1, 2)) %>%
  group_by(year, name, honorsystemcorrect) %>%
  summarize(pct_rt = mean(ans, na.rm = TRUE)) %>%
  drop_na() %>%
  mutate(illegal = if_else(honorsystemcorrect >= pct_rt, 1, 0))
```

```
## # A tibble: 283 x 5
## # Groups:   year, name [283]
##      year name            honorsystemcorrect pct_rt illegal
##     <dbl> <chr>                        <dbl>  <dbl>   <dbl>
##  1  2018 Abou-SayedT                  0.853   0.25        1
```

```
##   2  2018 BahnamanS                   0.899  0.792       1
##   3  2018 BawdonG                     0.828  0.292       1
##   4  2018 BerrettJ                    0.776  0.458       1
##   5  2018 BlishS                      0.835  0.75        1
##   6  2018 ButschekAHeyHey             0.816  0.375       1
##   7  2018 BuxtonK                     0.786  0.458       1
##   8  2018 CalcagnoR                   0.759  0.208       1
##   9  2018 CannonS                     0.802  0.542       1
## 10  2018 CareyC                      0.769  0.5         1
## # ... with 273 more rows
```

```r
pp <- edited %>%
  filter(round %in% c(1, 2)) %>%
  group_by(year, name, honorsystemcorrect) %>%
  summarize(pct_rt = mean(ans)) %>%
  drop_na() %>%
  ggplot(aes(honorsystemcorrect, pct_rt)) +
  labs(subtitle = "Rounds 1 and 2, All Participants",
       x = "Percent Correct, Honor System",
       y= "Percent Correct, Championship") +
  geom_abline(slope=1) +
  geom_smooth(method = "lm") +
  scale_x_continuous(labels = scales::label_percent()) +
  scale_y_continuous(labels = scales::label_percent())
pp +
  geom_point() +
  labs(title = "Percent Correct at Home vs Championship")
```



Approach 3

A better strategy is to run a one-sample t-test for every person's performance at home.

Using this approach, I identify a subset of people who I believe cheated in at home.

NOTE THAT THIS DATA ONLY INCLUDES PEOPLE WHO HAD PLAYED AT HOME AND PLAYED IN COMPETITION. THOSE WITHOUT VALUES FOR HONORSYSTEMCORRECT AND ANS WERE DISCOUNTED.

A one-sample t-test is used to compare the mean value of a sample with a constant value denoted mu0. The test has the null hypothesis that the population mean is equal to mu0 and the alternative hypothesis that it is not equal to mu0. http://www.instantr.com/2012/12/29/performing-a-one-sample-t-test-in-r/

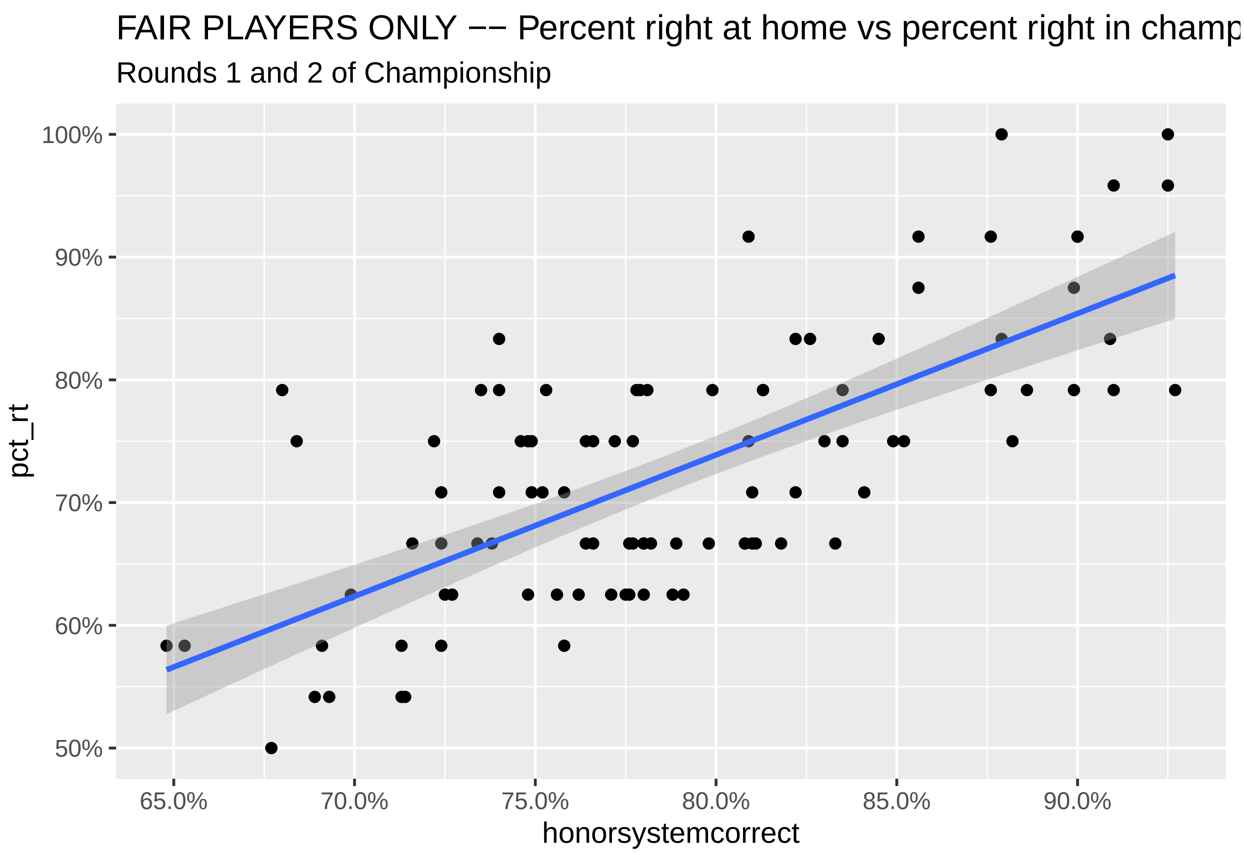
```r
corr_r_1and2 <- edited %>%
  filter(round %in% c(1, 2)) %>%
  drop_na(honorsystemcorrect, ans) %>%
  group_by(year, name, honorsystemcorrect) %>%
  summarize(pct_rt = mean(ans))

ttester <- function(current_selection, year_sel) {
  obs <- edited %>%
    filter(round %in% c(1, 2),
           year == year_sel,
           name == current_selection)
  # Case where people get a perfect score in competition returns no std dev so
  # do not do a ttest in those situations
  if (sum(obs$ans) == 24) {
    return(1)
  }
  honsyscorr_num <- obs %>%
    magrittr::extract2(1,6)
  pval <- t.test(obs$ans, mu = honsyscorr_num, alternative = "less") %>%
    broom::tidy() %>%
    magrittr::extract2(1, 3)
}

pvals <- edited %>%
  drop_na(honorsystemcorrect, ans) %>%
  distinct(name, year) %>%
  mutate(pvals = map2(name, year, ttester),
         statsig = if_else(pvals < 0.05, 1, 0))

allplayers <- left_join(corr_r_1and2, pvals)
```

```
## Joining, by = c("year", "name")
```

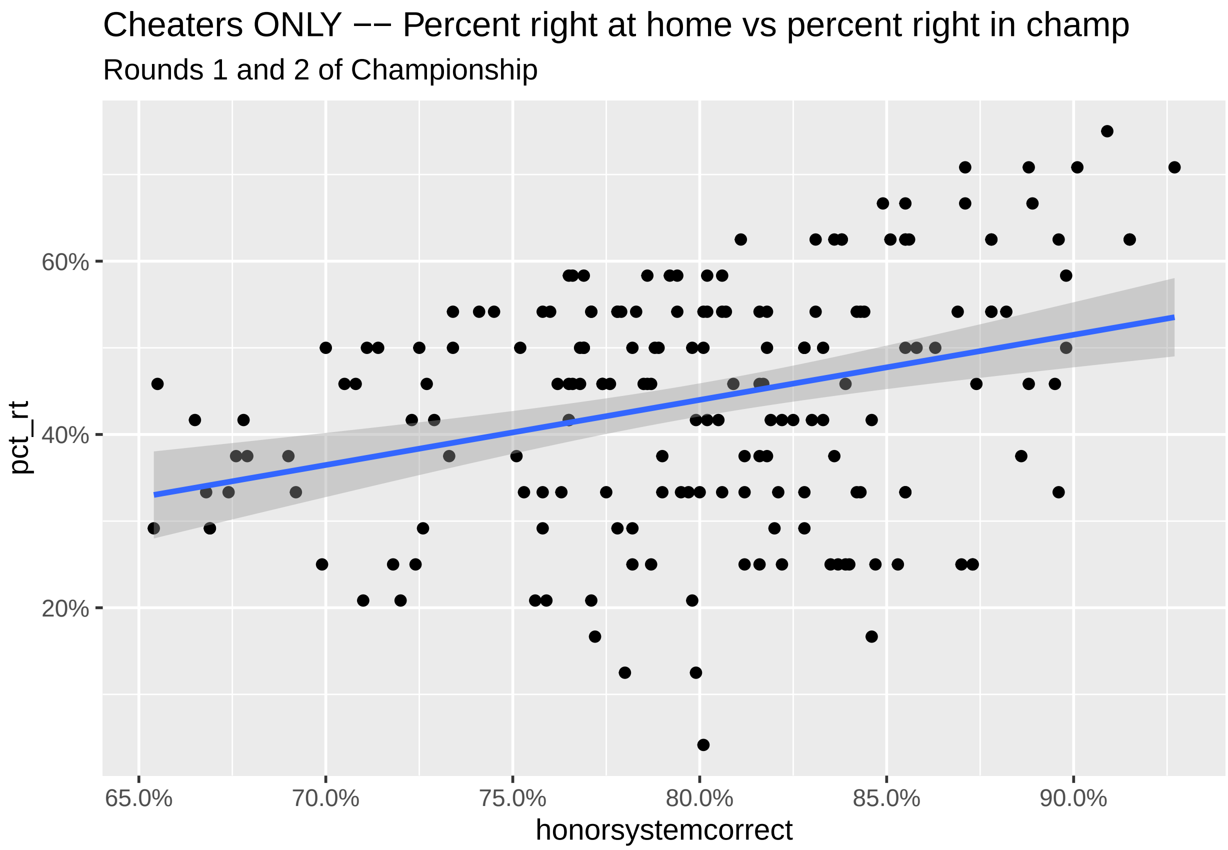
```r
# List of people that cheated
cheaters <- allplayers %>%
  filter(statsig == 1) %>%
  select(-statsig) %>%
  mutate(type = "cheater")

# List of people that didn't cheat
fairplayers <- allplayers %>%
  filter(statsig == 0) %>%
  select(-statsig) %>%
```

```r
  mutate(type = "fair")
# %>%
 # mutate(honorsystemcorrect = percent(honorsystemcorrect, accuracy = 0.1),
 #         pct_rt = percent(pct_rt, accuracy = 0.1))
```

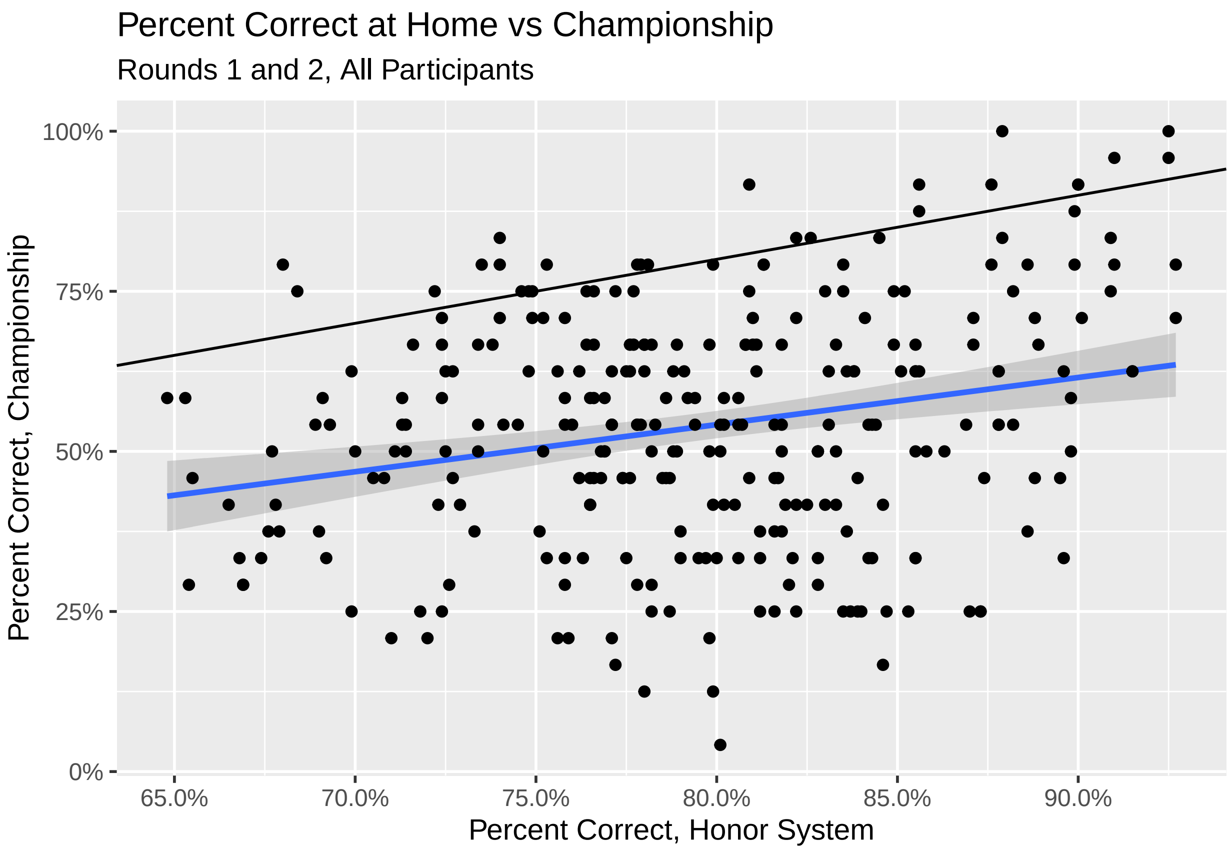
```r
fairplayers %>%
  ggplot(aes(honorsystemcorrect, pct_rt)) +
  geom_point() +
  labs(title = "FAIR PLAYERS ONLY -- Percent right at home vs percent right in champ",
       subtitle = "Rounds 1 and 2 of Championship") +
  scale_x_continuous(labels = scales::label_percent()) +
  scale_y_continuous(labels = scales::label_percent()) +
  geom_smooth(method = "lm")
```



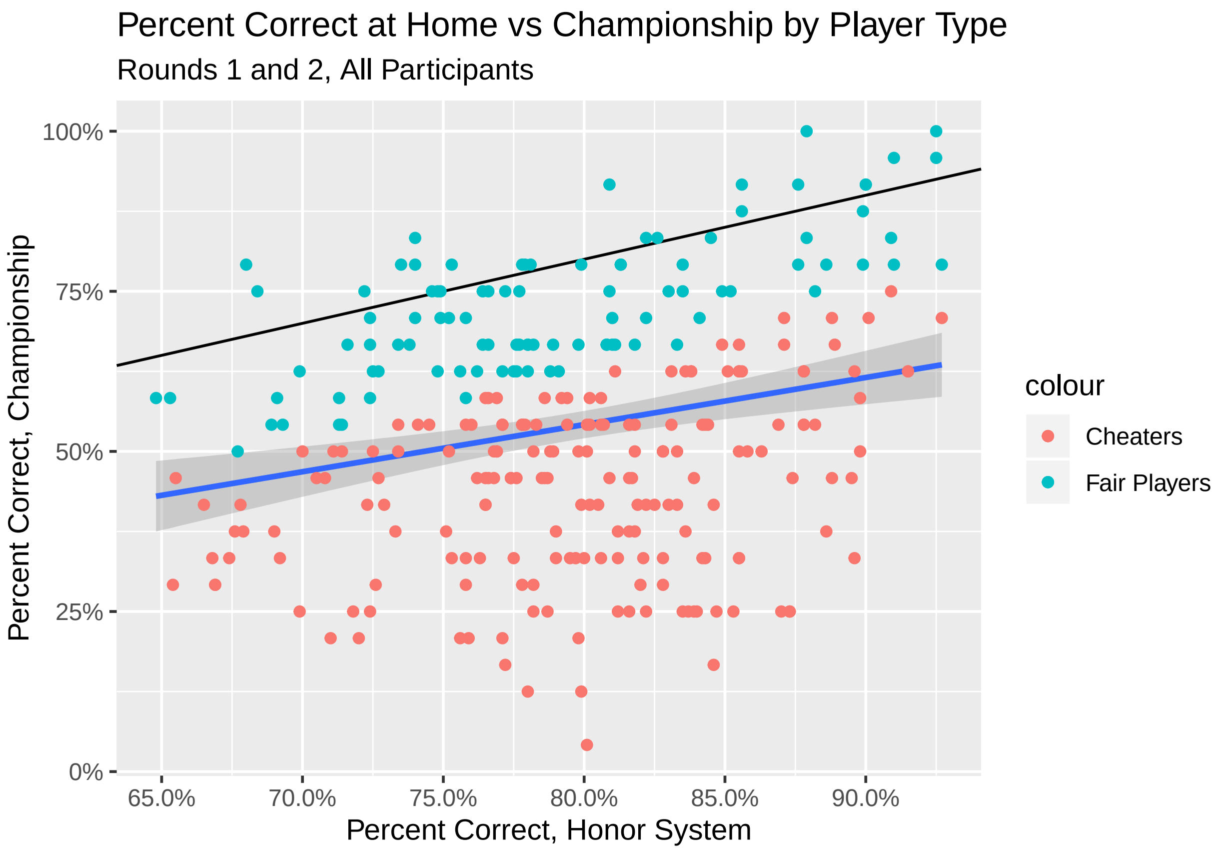

```r
cheaters %>%
  ggplot(aes(honorsystemcorrect, pct_rt)) +
  geom_point() +
  labs(title = "Cheaters ONLY -- Percent right at home vs percent right in champ",
       subtitle = "Rounds 1 and 2 of Championship") +
  scale_x_continuous(labels = scales::label_percent()) +
  scale_y_continuous(labels = scales::label_percent()) +
  geom_smooth(method = "lm")
```




```
pp +
  geom_point() +
  labs(title = "Percent Correct at Home vs Championship")
```




```r
pp +
  geom_point(cheaters, mapping = aes(honorsystemcorrect, pct_rt, color = "Cheaters")) +
  geom_point(fairplayers, mapping = aes(honorsystemcorrect, pct_rt, color = "Fair Players")) +
  labs(title = "Percent Correct at Home vs Championship by Player Type")# +
```




```
#   theme(legend.position = c(0.93, 0.1))

mean(fairplayers$honorsystemcorrect)
```

```
## [1] 0.7890204
```

```
mean(fairplayers$pct_rt)
```

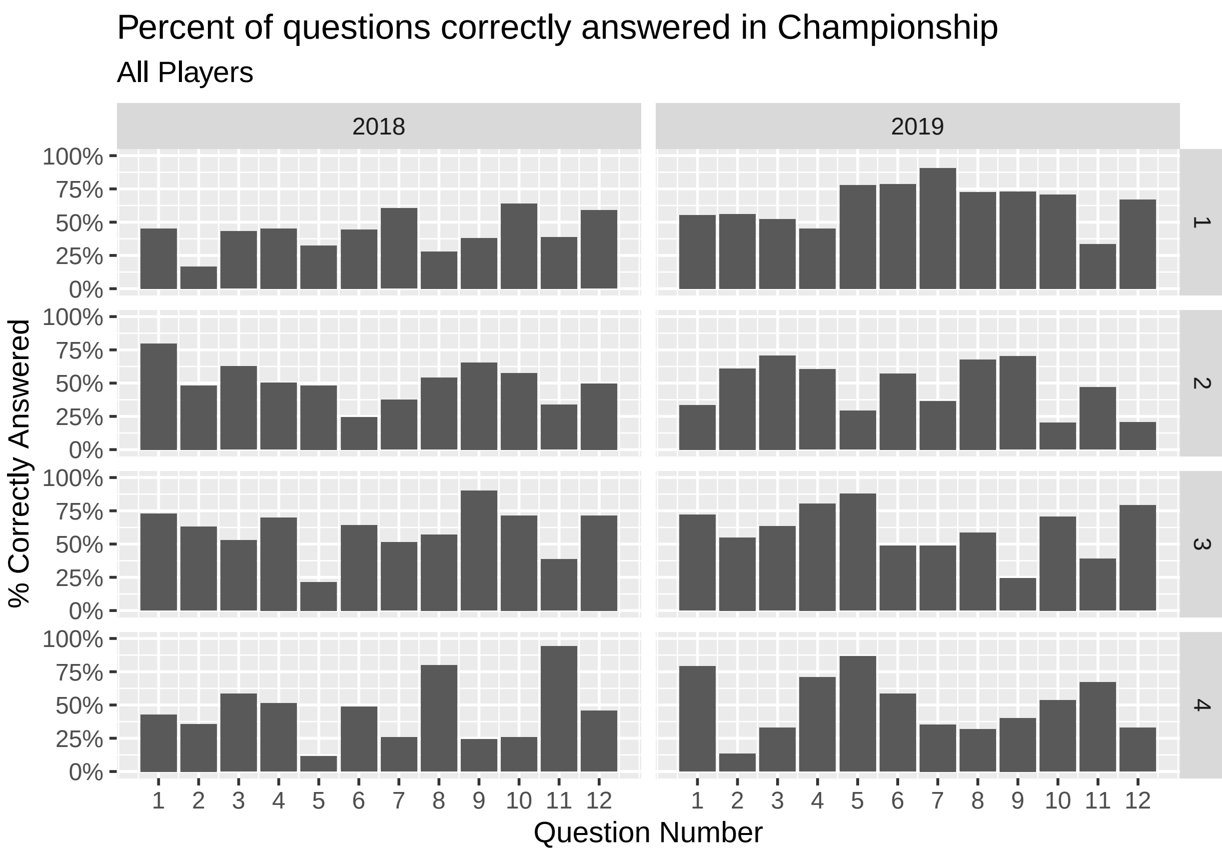
```
## [1] 0.7261905
```

```
edited %>%
  filter(round %in% c(3, 4)) %>%
  drop_na(honorsystemcorrect, ans) %>%
  group_by(year, name, honorsystemcorrect) %>%
  summarize(pct_rt = mean(ans))
```

```
## # A tibble: 151 x 4
## # Groups:   year, name [151]
##     year name      honorsystemcorrect pct_rt
##    <dbl> <chr>                  <dbl>  <dbl>
##  1  2018 BahnamanS              0.899  0.833
##  2  2018 BerrettJ               0.776  0.458
##  3  2018 BlishS                 0.835  0.792
##  4  2018 CannonS                0.802  0.333
##  5  2018 CareyC                 0.769  0.417
##  6  2018 ChiltonC2              0.758  0.542
##  7  2018 CohenD2                0.831  0.625
##  8  2018 CraneN                 0.713  0.333
##  9  2018 CurtinR                0.788  0.542
```

```
## 10  2018 DhuwaliaR             0.879  0.75
## # ... with 141 more rows
```

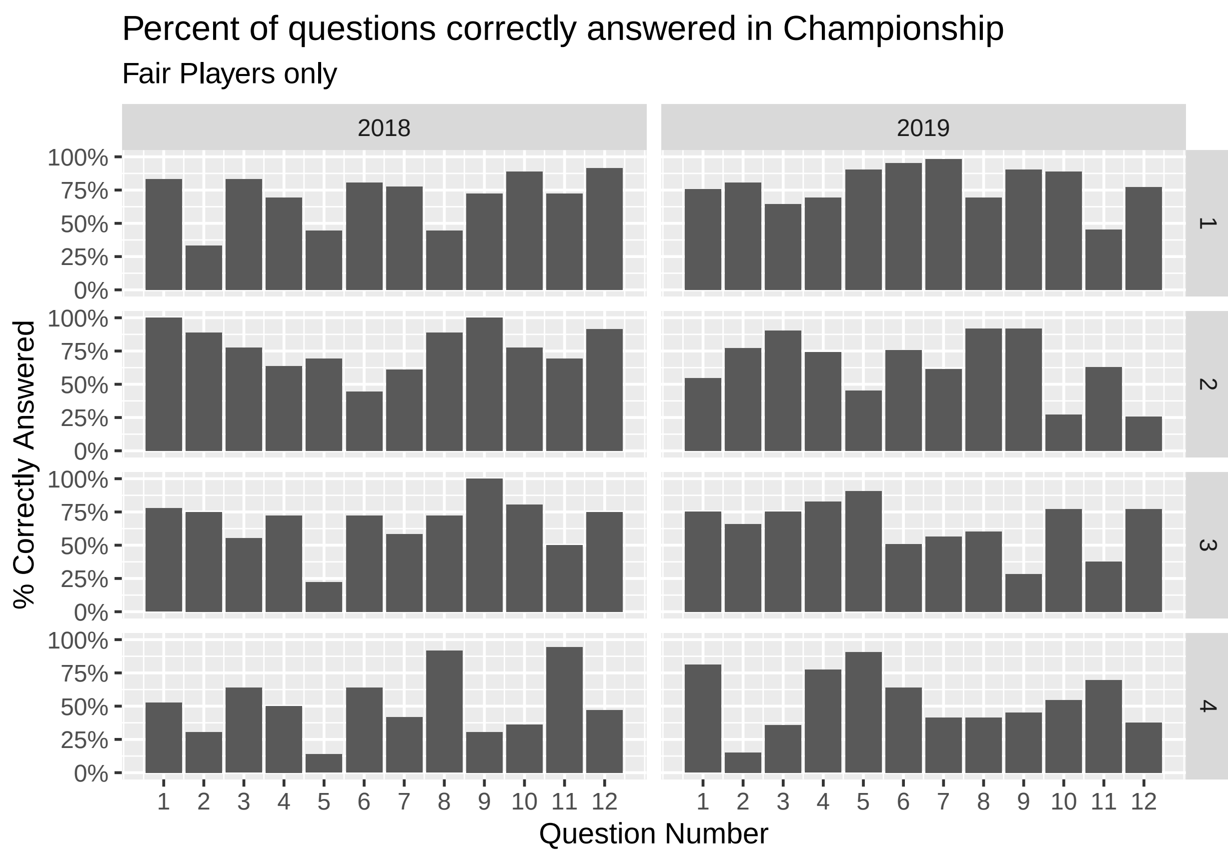
```
edited %>%
  filter(merge == "merged") %>%
  group_by(year, round, qno) %>%
  summarize(correct_ans = sum(ans), times_asked = n()) %>%
  ggplot(mapping = aes(x = qno, y = correct_ans/times_asked)) +
  geom_col() +
  facet_grid(round ~ year) +
  scale_x_continuous(breaks = seq(1, 12)) +
  scale_y_continuous(limits = c(0, 1), labels = label_percent()) +
  labs(title = "Percent of questions correctly answered in Championship",
       subtitle = "All Players",
       x = "Question Number",
       y = "% Correctly Answered")
```




```
inner_join(fairplayers, edited) %>%
  group_by(year, round, qno) %>%
  summarize(correct_ans = sum(ans), times_asked = n()) %>%
  ggplot(mapping = aes(x = qno, y = correct_ans/times_asked)) +
  geom_col() +
  facet_grid(round ~ year) +
  scale_x_continuous(breaks = seq(1, 12)) +
  scale_y_continuous(limits = c(0, 1), labels = label_percent()) +
  labs(title = "Percent of questions correctly answered in Championship",
       subtitle = "Fair Players only",
       x = "Question Number",
```

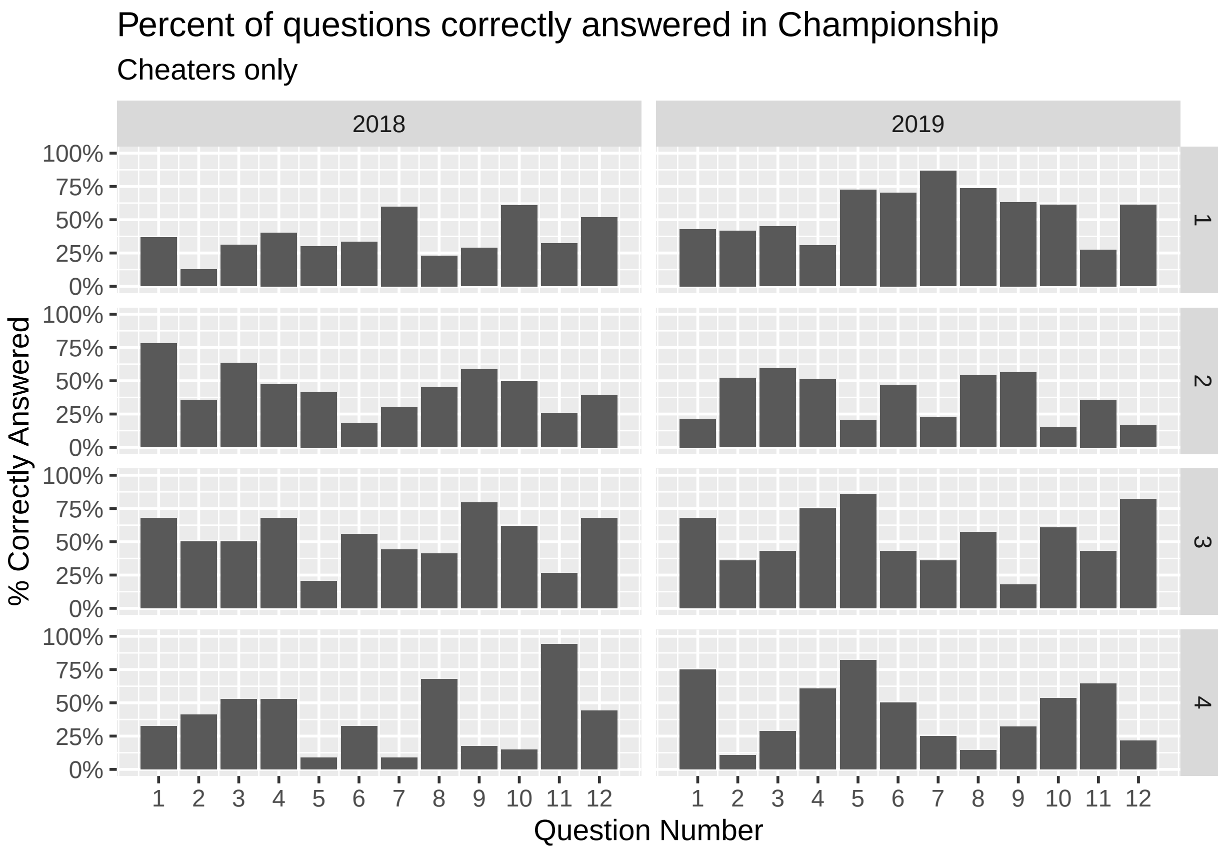
```
       y = "% Correctly Answered")
```

```
## Joining, by = c("year", "name", "honorsystemcorrect")
```



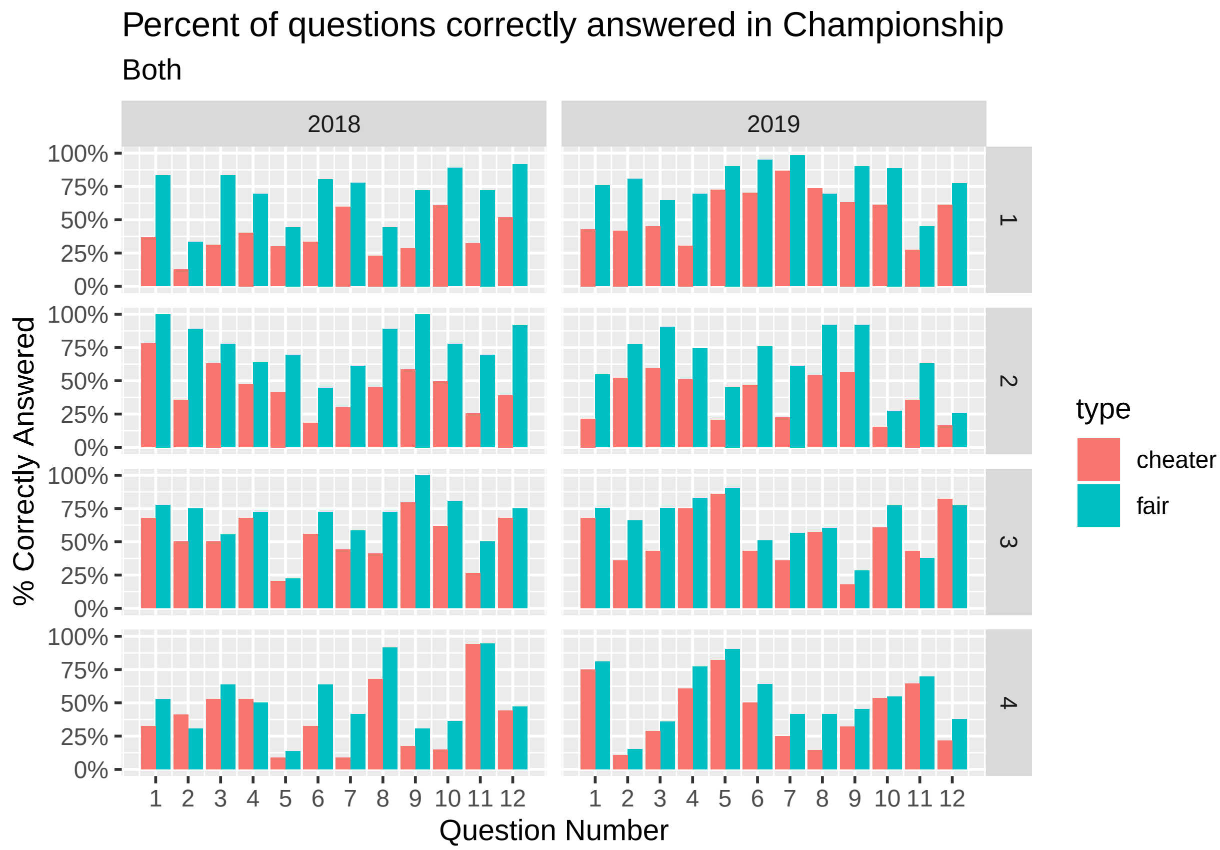

```
inner_join(cheaters, edited) %>%
  group_by(year, round, qno) %>%
  summarize(correct_ans = sum(ans), times_asked = n()) %>%
  ggplot(mapping = aes(x = qno, y = correct_ans/times_asked)) +
  geom_col() +
  facet_grid(round ~ year) +
  scale_x_continuous(breaks = seq(1, 12)) +
  scale_y_continuous(limits = c(0, 1), labels = label_percent()) +
  labs(title = "Percent of questions correctly answered in Championship",
       subtitle = "Cheaters only",
       x = "Question Number",
       y = "% Correctly Answered")
```

```
## Joining, by = c("year", "name", "honorsystemcorrect")
```




```r
bind_rows(cheaters, fairplayers) %>%
  inner_join(edited) %>%
  group_by(year, round, qno, type) %>%
  summarize(correct_ans = sum(ans), times_asked = n()) %>%
  ggplot(aes(x = qno, y = correct_ans/times_asked, group = type, fill = type)) +
  geom_col(position = "dodge") +
  facet_grid(round ~ year) +
  scale_x_continuous(breaks = seq(1, 12)) +
  scale_y_continuous(limits = c(0, 1), labels = label_percent()) +
  labs(title = "Percent of questions correctly answered in Championship",
       subtitle = "Both",
       x = "Question Number",
       y = "% Correctly Answered")
```

```
## Joining, by = c("year", "name", "honorsystemcorrect")
```




```r
cheaters %>%
  mutate(honorsystemcorrect = percent(honorsystemcorrect, accuracy = 0.1),
         pct_rt = percent(pct_rt, accuracy = 0.1)) %>%
  slice(1:10) #%>%
```

```
## # A tibble: 185 x 6
## # Groups:   year, name [185]
##     year name             honorsystemcorrect pct_rt pvals     type
##    <dbl> <chr>            <chr>              <chr>  <list>    <chr>
##  1  2018 Abou-SayedT      85.3%              25.0%  <dbl [1]> cheater
##  2  2018 BawdonG          82.8%              29.2%  <dbl [1]> cheater
##  3  2018 BerrettJ         77.6%              45.8%  <dbl [1]> cheater
##  4  2018 ButschekAHeyHey  81.6%              37.5%  <dbl [1]> cheater
##  5  2018 BuxtonK          78.6%              45.8%  <dbl [1]> cheater
##  6  2018 CalcagnoR        75.9%              20.8%  <dbl [1]> cheater
##  7  2018 CannonS          80.2%              54.2%  <dbl [1]> cheater
##  8  2018 CareyC           76.9%              50.0%  <dbl [1]> cheater
##  9  2018 ChiltonC2        75.8%              54.2%  <dbl [1]> cheater
## 10  2018 ChinG            80.0%              33.3%  <dbl [1]> cheater
## # ... with 175 more rows
```

```r
#  kable(col.names = c("Year", "Name", "Percent Correct, Honor System", "Percent Correct, Competition",
```

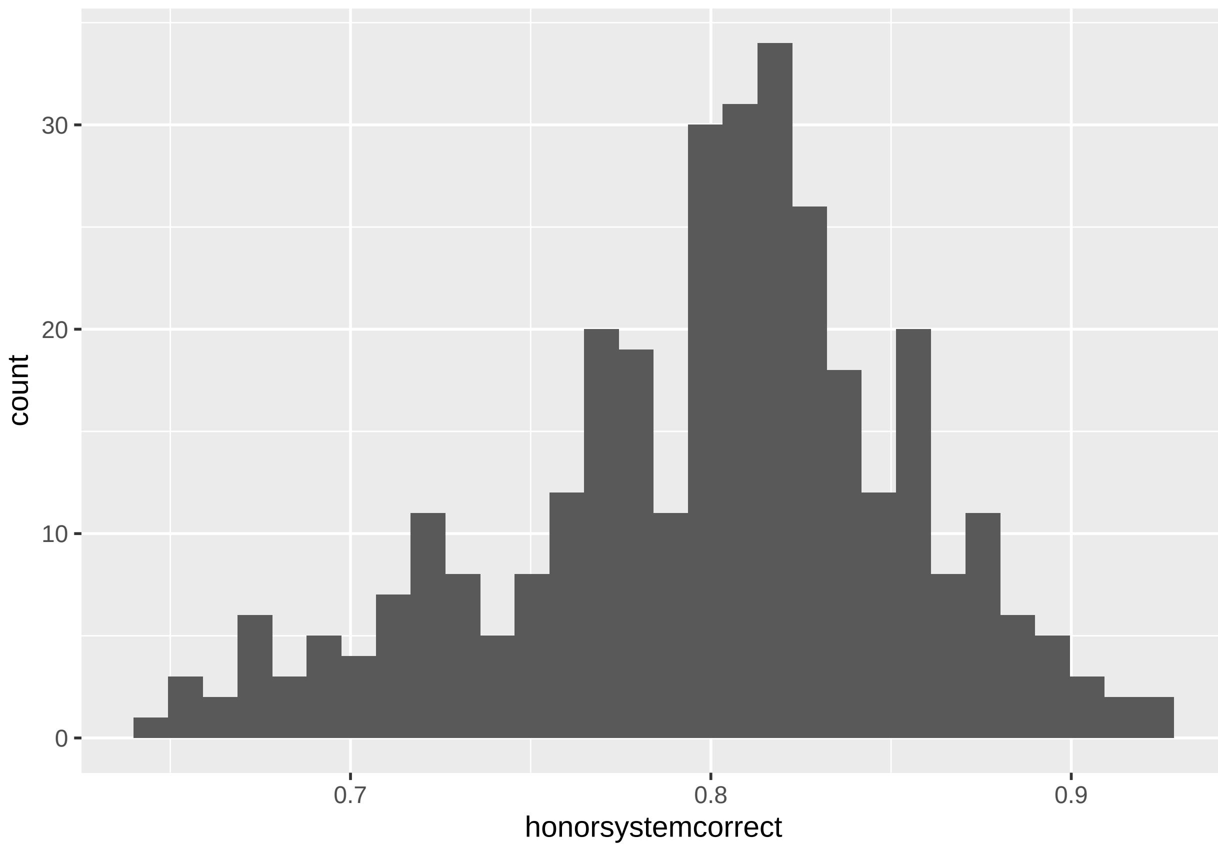
Step 2: how difficult the championship questions are relative to the regular season questions for the above group of people

One fairly awful strategy is to calculate mean correctness percentage of all eligible participants at home, and then compare that to the mean correctness percentage of all eligible participants in competition. By eligible, I mean participants that pass the tests I define in step 1.

One better strategy is to

```
# Lets find people who I think played honestly during the regular season.
edited %>%
  distinct(name, honorsystemcorrect) %>%
  ggplot(aes(honorsystemcorrect)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 19 rows containing non-finite values (stat_bin).
```




```
edited %>%
  group_by(name, year, round) %>%
  mutate(stdev = sd(ans), competcorrect = mean(ans)) %>%
  mutate(tstatistic = (competcorrect - honorsystemcorrect)/(stdev/sqrt(12)))
```

```
## # A tibble: 12,132 x 11
## # Groups:   name, year, round [1,011]
##    name   year round merge numbercorrect honorsystemcorr~   qno   ans stdev
##    <chr> <dbl> <dbl> <chr>         <dbl>            <dbl> <dbl> <dbl> <dbl>
##  1 Abou~  2018     2 merg~             2            0.853     1     0 0.389
##  2 Abou~  2018     2 merg~             2            0.853     2     0 0.389
##  3 Abou~  2018     2 merg~             2            0.853     3     0 0.389
##  4 Abou~  2018     2 merg~             2            0.853     4     1 0.389
##  5 Abou~  2018     2 merg~             2            0.853     5     0 0.389
##  6 Abou~  2018     2 merg~             2            0.853     6     0 0.389
```

```
## 7 Abou~  2018     2 merg~              2            0.853     7     1 0.389
##  8 Abou~  2018     2 merg~              2            0.853     8     0 0.389
##  9 Abou~  2018     2 merg~              2            0.853     9     0 0.389
## 10 Abou~  2018     2 merg~              2            0.853    10     0 0.389
## # ... with 12,122 more rows, and 2 more variables: competcorrect <dbl>,
## #   tstatistic <dbl>
```

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

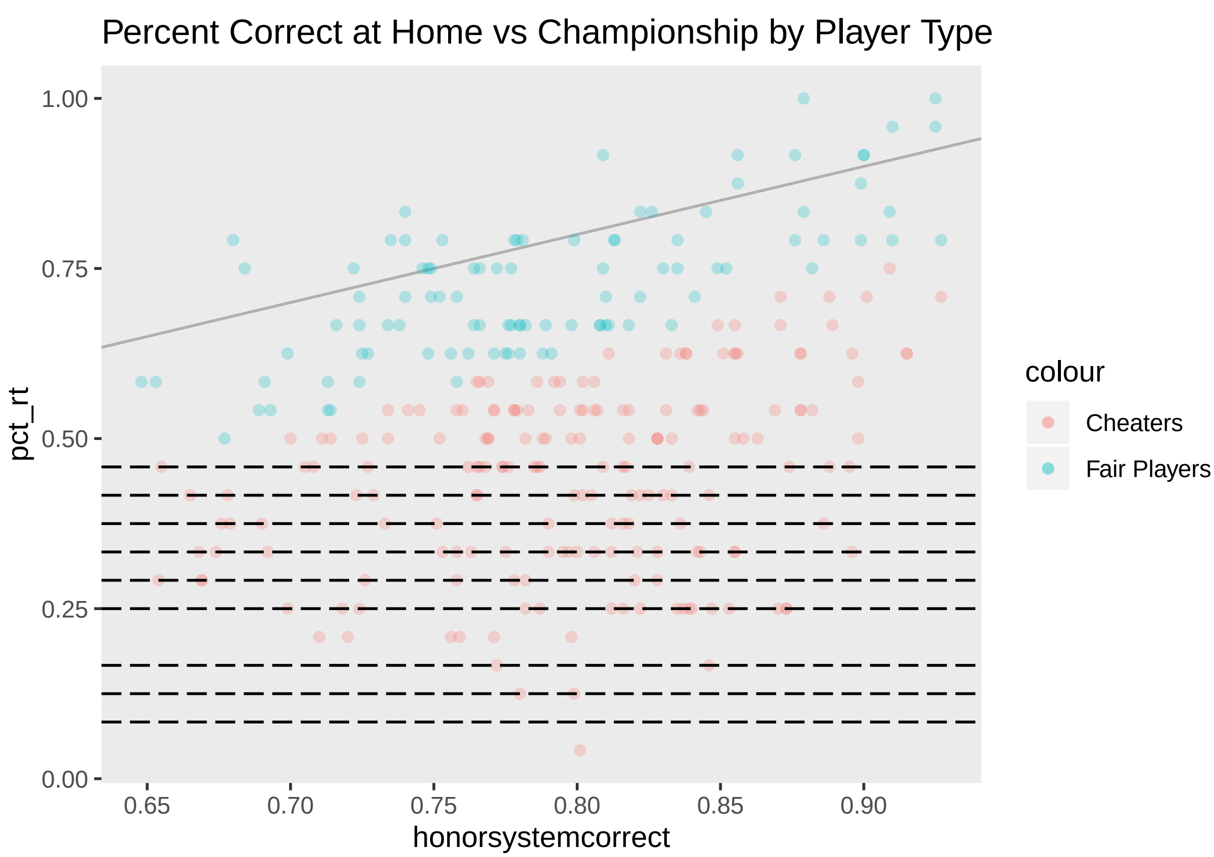
5. The players with -99 for honor code scores dropped out of the league after making one or both championships. Can you make any inferences about whether they cheated more or less than the players who have remained in the league, despite the fact you know nothing about their percent correct in the regular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

I've identified the characteristics of what a cheater looks like, and I think all these people fit the bill based on one axis. Not a single person who stayed who had the same championship scores as the -99ers was categoriesed as a fair player. Thus, I think they are cheaters.

```
dropouts <- edited %>%
  filter(is.na(honorsystemcorrect)) %>%
  distinct(name, year) %>%
  pull(name)

x <- edited %>%
  filter(name %in% dropouts) %>%
  group_by(name, year) %>%
  summarize(pct_corr = mean(ans)) %>%
  ungroup() %>%
  count(pct_corr)

ggplot() +
  geom_abline(slope = 1, alpha = 0.25) +
  geom_point(cheaters, mapping = aes(honorsystemcorrect, pct_rt, color = "Cheaters"), alpha = 0.25) +
  geom_point(fairplayers, mapping = aes(honorsystemcorrect, pct_rt, color = "Fair Players"), alpha = 0.
  geom_hline(x, mapping = aes(yintercept = pct_corr), linetype = "longdash") +
  labs(title = "Percent Correct at Home vs Championship by Player Type") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```

Table 3: 2018/2019 Championship Percentage Correct by Question

	2018				2019			
Round Number	1	2	3	4	1	2	3	4
Sample Size (n =)	1668	1668	840	840	1968	1968	984	984
Question 1	45.32%	79.86%	72.9%	42.9%	55.49%	33.54%	72.0%	79.3%
Question 2	16.55%	48.20%	62.9%	35.7%	56.10%	60.98%	54.9%	13.4%
Question 3	43.17%	62.59%	52.9%	58.6%	52.44%	70.73%	63.4%	32.9%
Question 4	45.32%	50.36%	70.0%	51.4%	45.12%	60.37%	80.5%	70.7%
Question 5	32.37%	48.20%	21.4%	11.4%	78.05%	29.27%	87.8%	86.6%
Question 6	44.60%	24.46%	64.3%	48.6%	78.66%	57.32%	48.8%	58.5%
Question 7	60.43%	37.41%	51.4%	25.7%	90.85%	36.59%	48.8%	35.4%
Question 8	28.06%	53.96%	57.1%	80.0%	72.56%	67.68%	58.5%	31.7%
Question 9	38.13%	65.47%	90.0%	24.3%	73.17%	70.12%	24.4%	40.2%
Question 10	64.03%	57.55%	71.4%	25.7%	70.73%	20.12%	70.7%	53.7%
Question 11	38.85%	33.81%	38.6%	94.3%	33.54%	46.95%	39.0%	67.1%
Question 12	58.99%	49.64%	71.4%	45.7%	67.07%	20.73%	79.3%	32.9%



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.

Table 4: Percentages of Honor System Success by Round and Year

Year	1	2	3	4
2018	79.7% (n = 1668)	79.7% (n = 1668)	80.7% (n = 840)	80.7% (n = 840)
2019	79.5% (n = 1968)	79.5% (n = 1968)	80.7% (n = 984)	80.7% (n = 984)

Appendices