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.

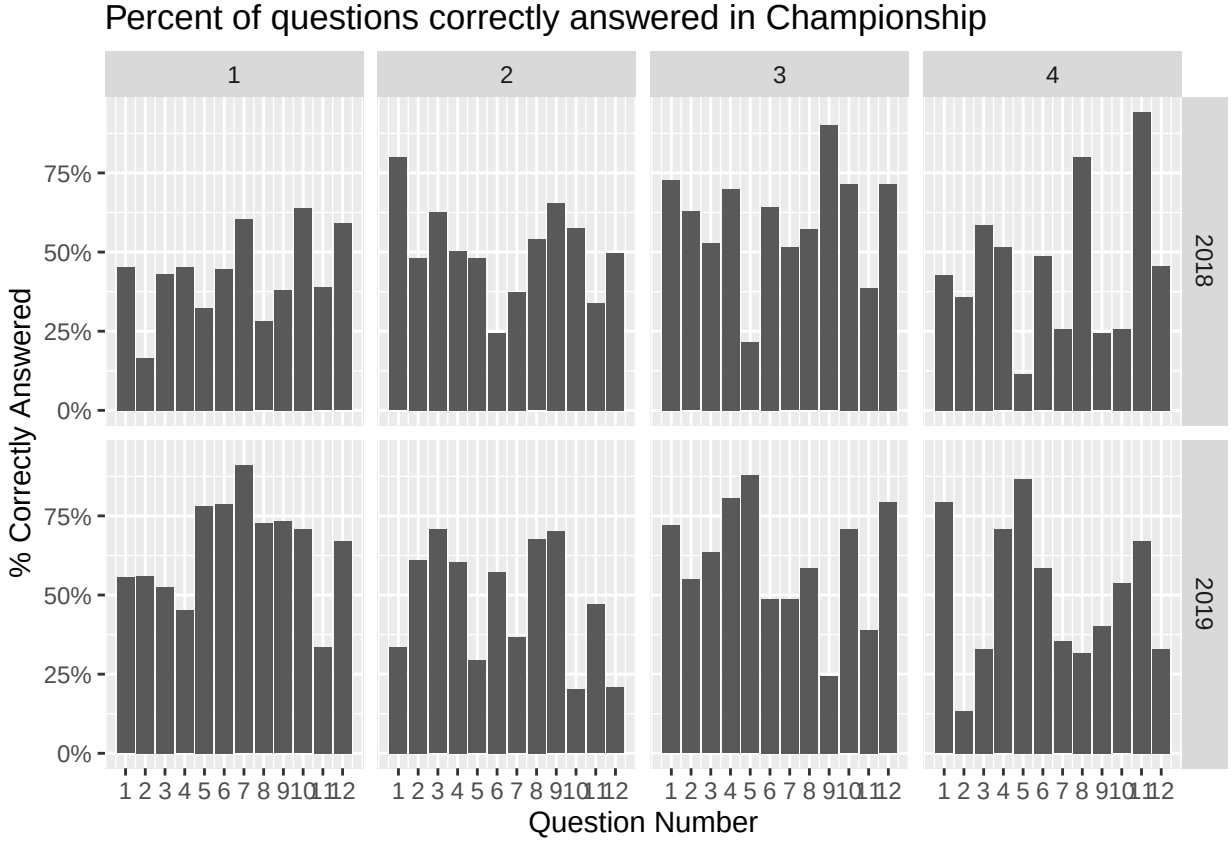
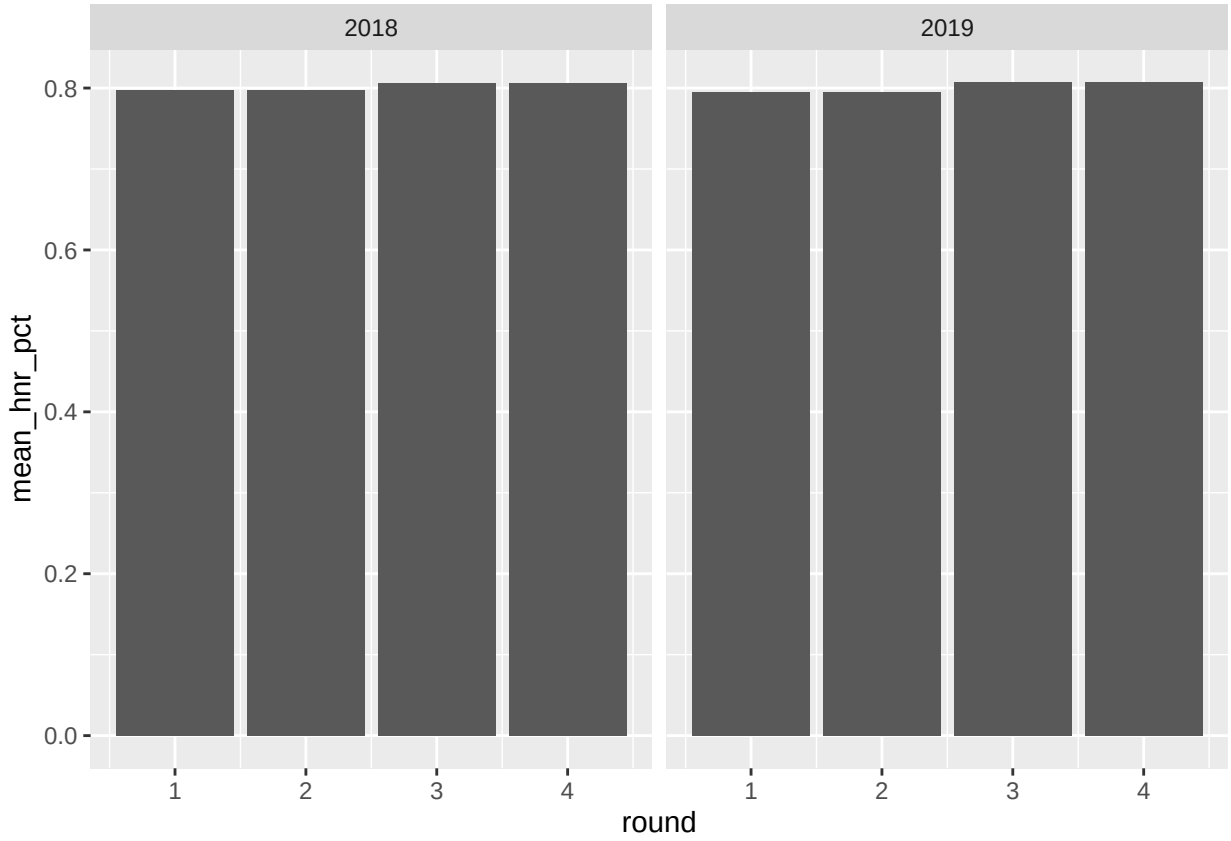


Table 1: 2018/2019 Championship Percentage Correct by Question

	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%

Table 2: 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)

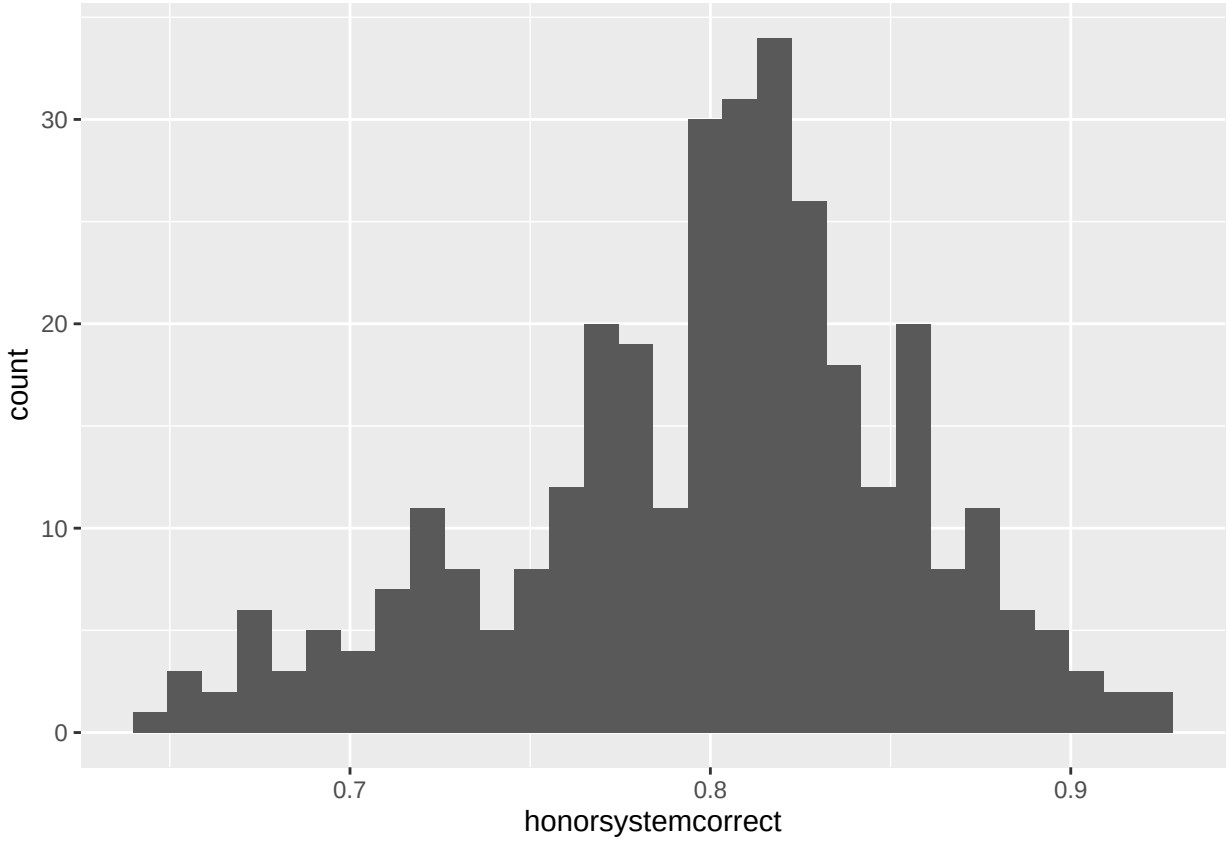


Let's focus on looking for cheating in the first two rounds of the Championship each year, before a bunch of people get eliminated. A critical piece in determining how much cheating there might be is to figure out how difficult the championship questions are relative to the regular season questions for people who play honestly during the regular season. Suggest at least two awesome strategies for coming up with such an estimate. Be extremely explicit about the assumptions that need to be true for each of your strategies to yield truthful estimates. Given the likely violation of your assumptions, and say whether your estimates are likely to overestimate or underestimate the true amounts of cheating.

```
# 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>
```