ECON 21300 Assignment 3

Yixin Hu

Question 1:

Without running any regressions, get to know the data. Describe interesting features you see in the data, especially with respect to the championship data and differences between the regular season and championship data.

Response:

Glimpsing at each of the data frames, we see that:

- qualifiers only contain ID of qualified individuals
- regular data contains season and matches to indicate the time the questions were taken, whereas championship data only contains rounds to indicate when the questions were taken
- in the regular data, individuals can take questions only belonging to a specific category, whereas the finals has no category whatsoever.
- id and user_id seem to indicate the same entity (identification of individual) in the championship and regular data respectively, but with different names.

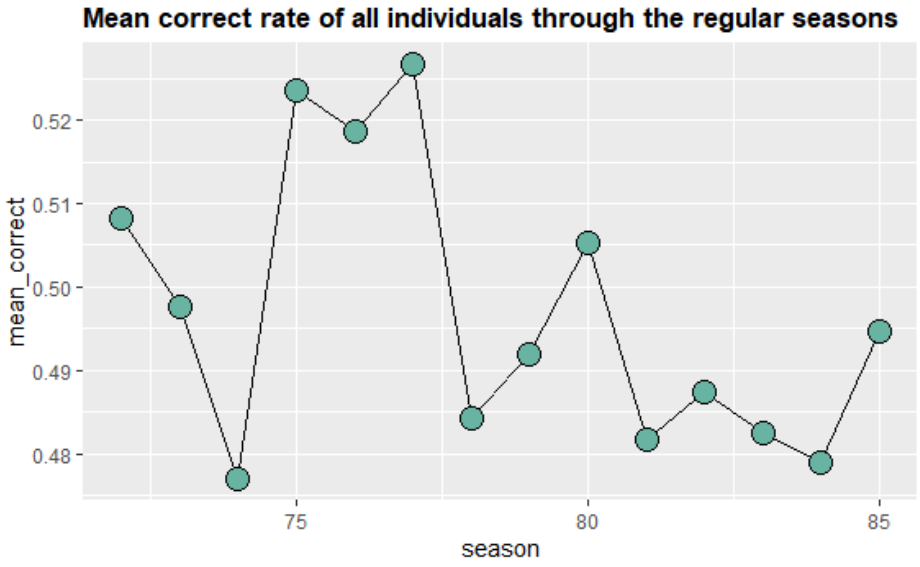
Furthermore, there are some interesting findings within each piece of data as well. For example, there are a total of 14 seasons in the regular season data. Since the data is from 2017-2020, we expect there to be 16 seasons. Therefore, one observation could be that (without looking at any other source) there is only data up to to mid-2020, and the last two seasons of 2020 was not observed, assuming that the data indeed starts on the first season of 2017. Therefore, I will treat the seasons' corresponding years accordingly.

From here, recall that we are most interested on the differential outcomes between regular and championship seasons to tease out whether or not individuals were 'cheating' in the regular season. This is

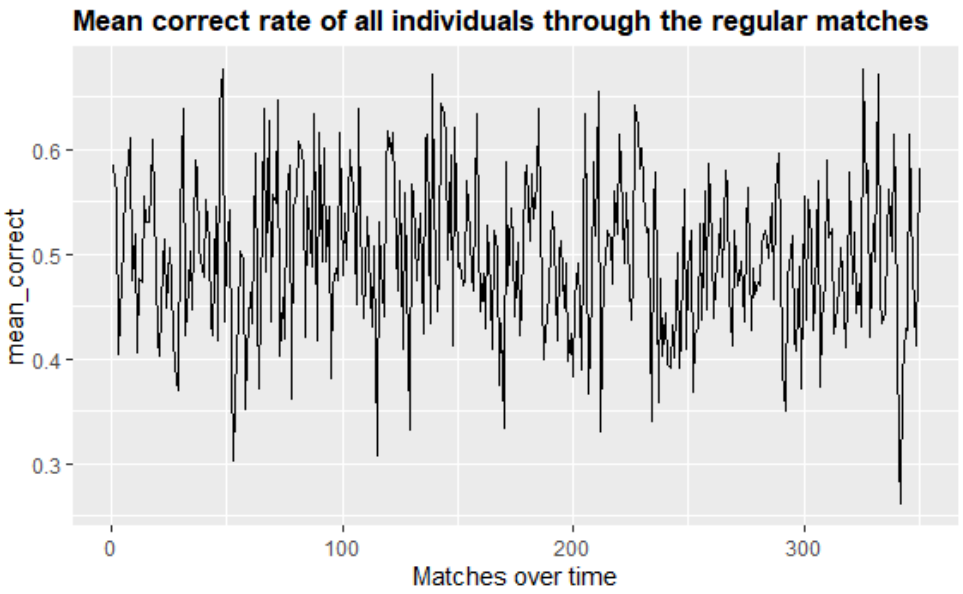
because the championships are live, and individuals cannot readily cheat in that context. Therefore, some basic visualizations will be used to explore the %correct between the two sets of data, as well as within each set of data (say, based on categories, seasons, or between matches). First, however, I will check the distributions of %correct based on categories, matches, and seasons to check if there is some persistence in the regular season data only. If there is some structural differences (i.e. for some reason, different categories have different distributions of mean correct rates), then it would be wise to take a closer look at these variables, and maybe control from them when building the models in the following questions. I will proceed in the following order:

1. take a look at the aggregate (not looking at different groups) data in the regular season.
2. take a look at the aggregate (not looking at different groups) data in the final season.
3. based on previous observations, address the main issue of finding the differences between the champions and regular season data.

First, I will show the mean correct rate based on different seasons to see if different years will have different outcomes for the entire group (i.e. If we see that some seasons have significantly different correct rates than any other season, it would be a good idea to consider this factor when predicting within the regular season data.)

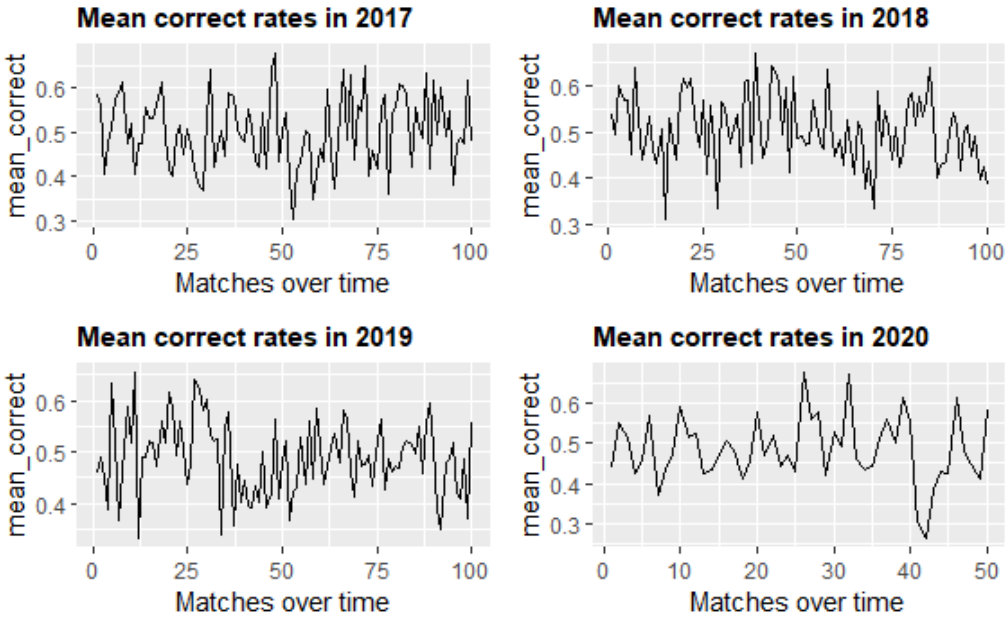


We see that on aggregate (not distinguishing between subgroups), the mean correct rate varies with seasons as well. In particular, we have large fluctuations in the first seasons which includes a spike up to around 52% correctness in seasons 75~77, followed by a sharp decrease back to around 48% in season 78, and generally staying around the 48%~49% range throughout the 5 most recent seasons. Therefore, if an individual participated in the seasons 75 instead of 74, one could potentially argue (of course this is very unsound) that on average, the individual who participated in 75 would be more likely to answer a question correct than someone who participated in season 74. Of course, there are lots of different confounding variables in this argument, where one could easily look at the distribution of individuals within the seasons and find that in season 75 there was an abundance of individuals who did very well (due to smarts, or even cheating) which led to this jump from the previous season, and thus does not represent any ordinary individual's ability to answer questions correctly. Nevertheless, because we see variation in the seasons, it would be good to consider this in further analysis. Now that we have checked a higher-level view of the seasons, I will now dive into the matches-level data to see trends or anything that pops out in the regular season on aggregate (without distinguishing subgroups).



From the graph above, we do not see any systematic trend over time, where these mean correct rates seems as though they were random draws from a normal distribution situated at round 50% (basically random noise), which is a very intuitive result. This indicates that (on average) there shouldn't be an underlying structural influence on the data due to different matches (i.e. it is not evident that there are some matches that are inherently rigged so that the mean correct rate is very high or low relative to every other match). This is already very granular data, and because we cannot see some matches being inherently different from other matches when only looking at the entire sample, there is no strong justification to be made about using this variable as a fixed-effect of any kind.

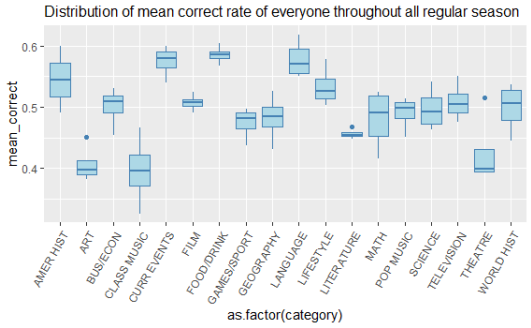
Of course, a critic can easily say that the different years will have different trends and this overall trend is misleading. To that I offer the following graphs, which separate each year's matches by the respective years.

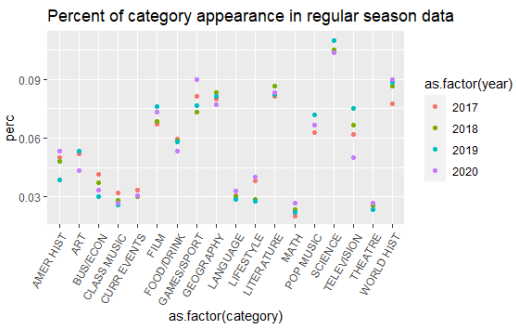


Here, even in the separate visual cases, it would be extremely hard to argue that there is some seasonality or similarly any inherent structural issues which will provide arguments to support including matches as an explanatory variable.

Now that we have seen that, on average, the regular season data is basically random noise, there is no reason to go even more granular and look at questions data. This is because the matches data, which is already quite granular compared to season data, already displays quasi-random noise behavior. As a result, the questions data, which are the components of the matches data, should not be introduced as an explanatory variable due to the random noise that it would bring into any model looking at the sample in general if directly added as some fixed effect.

Lastly because the regular data set also provided some category data, it would be interesting to see if the categories are all centered at a particular mean with similar variances throughout the years, or that they are clustered differently. In the graphs below, I have shown that each category has its own cluster of correct rates, and the rate of appearance of a question seems to relatively constant over time.

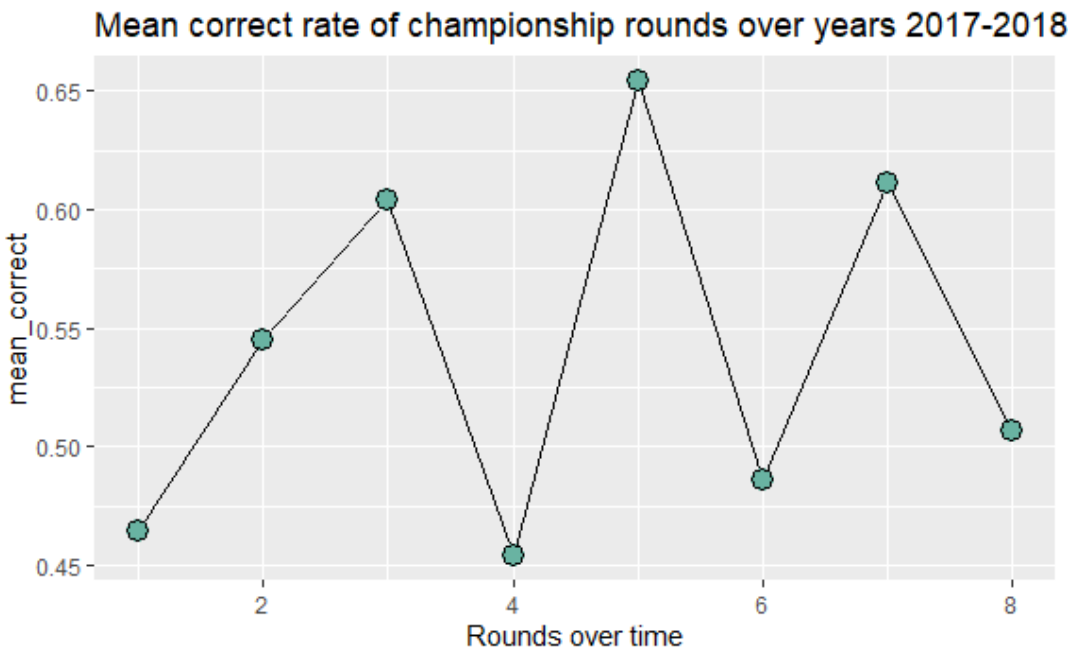




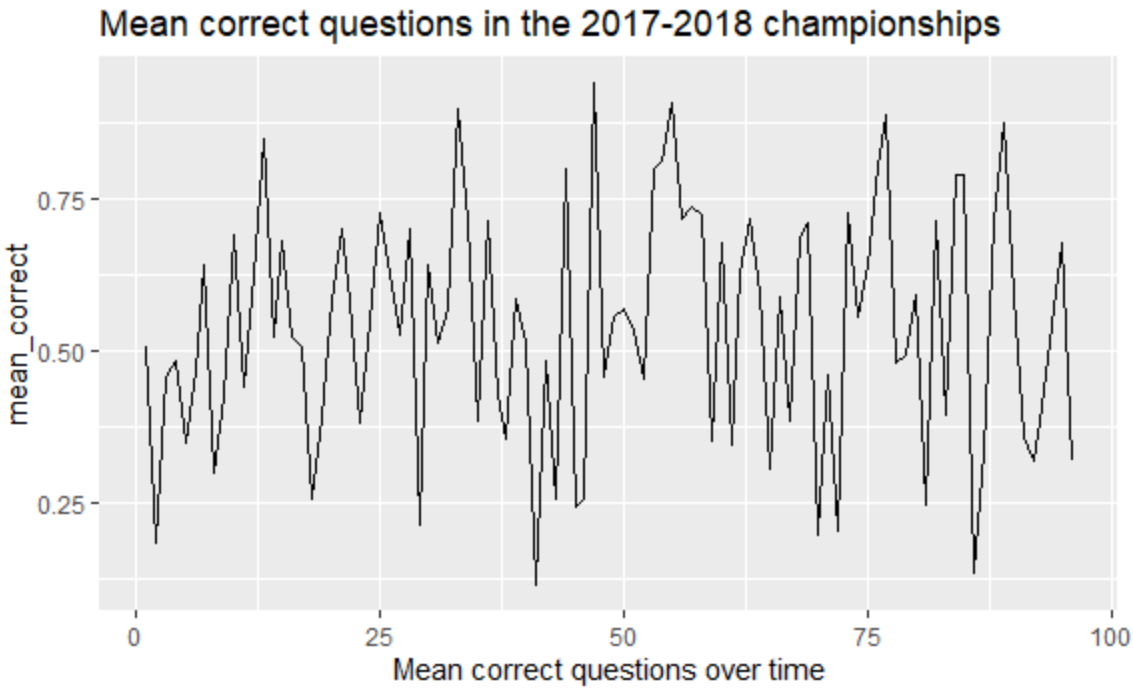
Notice that for each category, the distribution of the mean correct rate of everyone throughout the years have very different distributions. For example, throughout the years, people seem to do poorly relative to American history. This shows that each category has its own "**difficulty", where we define difficulty as an empirical difference between the mean correct rate of particular categories, matches, or questions** and **NOT NECESSARILY** how intrinsically "hard" or "complicated" a question is. Furthermore, it is worth noting that categories are more or less identically distributed with respect to time.

Now that we have seen how each of our variables of interest is presented in the sample in general within the regular seasons. I will now turn to the champions data and look at variables within that data frame accordingly.

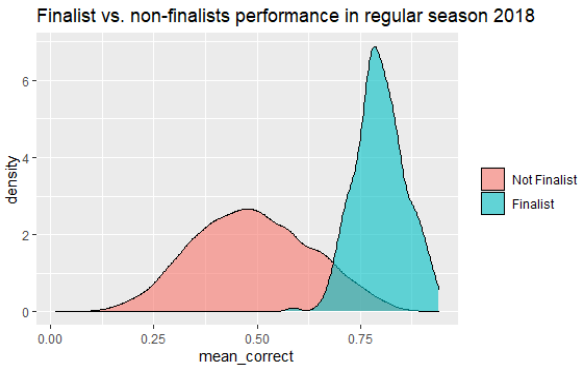
Glimpsing into the champions data frame we find that the round date is somewhat parallel to the matches data in the regular season. Therefore, I will take a look at that first and see how it might differ from the regular season's matches data. This graph below will hope to demonstrate the trends of the mean correct rate of final rounds over time, and see if there are any structural or inherent traits worth considering.

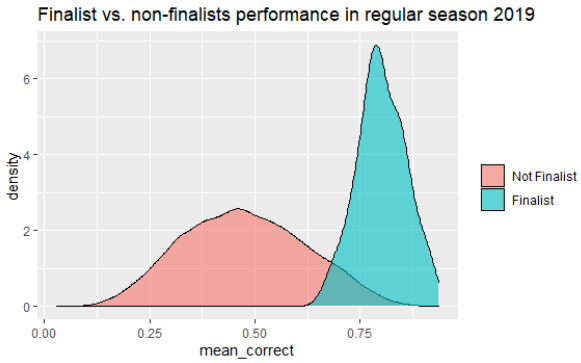


Notice that although there are not many data points (which warrants a slightly deeper dive into the question-level data), the mean correct rate of championship rounds do seem similar to a random draw from a normal distribution centered around 55%. This number is quite intuitive considering that the entire population in the regular season is centered around 50% (which includes these finalists), and this number is higher than the average regular season correct rate. Therefore, I tentatively claim that (at least looking very naively at the whole finalist sample), that the rounds played in the championships do not explain one's mean correct rate on average by itself as a fixed-effect variable. However, due to the small amount of data, I will now dive into the question level analysis over time.



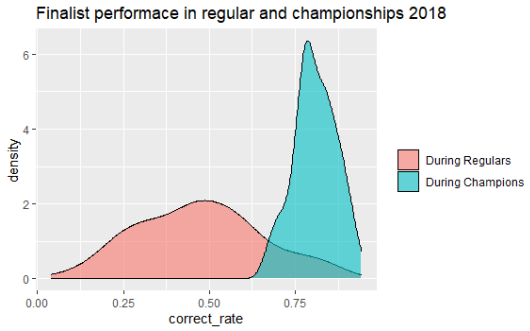
Now that we have taken a look at the regular and championships data separately, I will now address the important issue (which is what the question actually asked) of looking at the difference between the finalists and the non-finalists. First, I will note that according to Prof. Levitt, the final questions are harder than the regular season questions. Therefore, all else equal, we should expect the point estimate of the mean correctness to be much lower in the championships relative to the regular season. Of course, intuitively, the people who are in the championships are selected from the regular season to be the best around, and judging by the sheer difference between the size of the two data frames, it must be that the finalists must be extremely outstanding in terms of correct rates to be selected in. To confirm these two ideas, I will look at a very high-level analysis of individuals who made it to the championships and see 1) if the championship participants (finalists) are indeed much better than those who are not qualified, and 2) if the questions in the championships are indeed a lot harder. Where 'hard' here is simply defined as the lack of a concentrated distribution of mean correct rate at the higher range (for example, concentrated around 70%~90%).

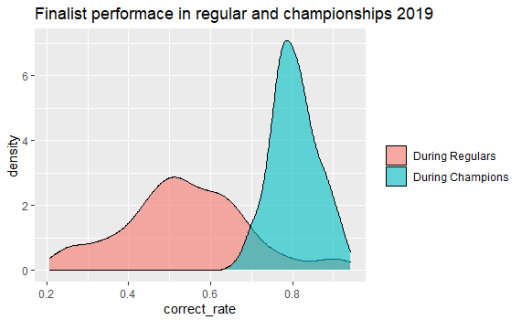




Here, we see a very intuitive outcome: Those who are in the finals have scored extremely well during the regular season. This shows that the intuition that the finalists are only those who are extremely outstanding in terms of correct percentage holds empirically.

Now looking into the second idea, we expect that the finals should have more difficult (where we have already defined difficult as an empirically different mean correct rate, so here "more difficult" means "lower mean correct rate") questions relative to the regular season. This should be identified if we see that peak that is the finalists performance during the regular season flatten out and return to the lower ranges of the mean correct rates. The following graph presents the performance of finalists during the regular season and the championships.

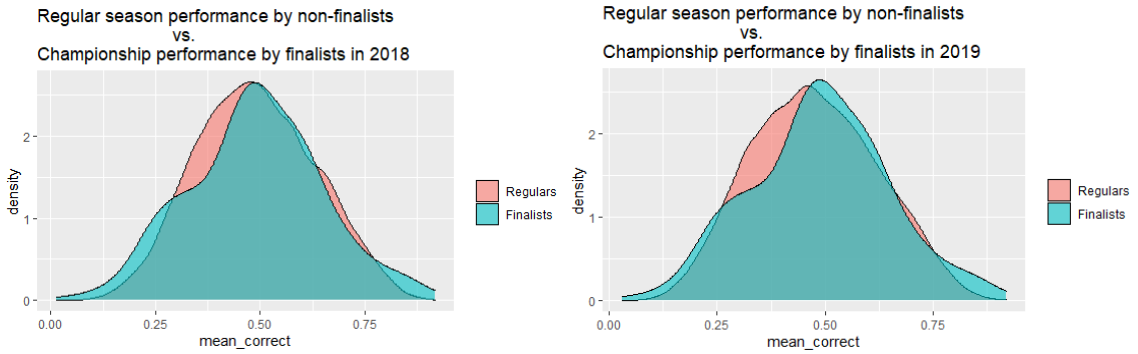




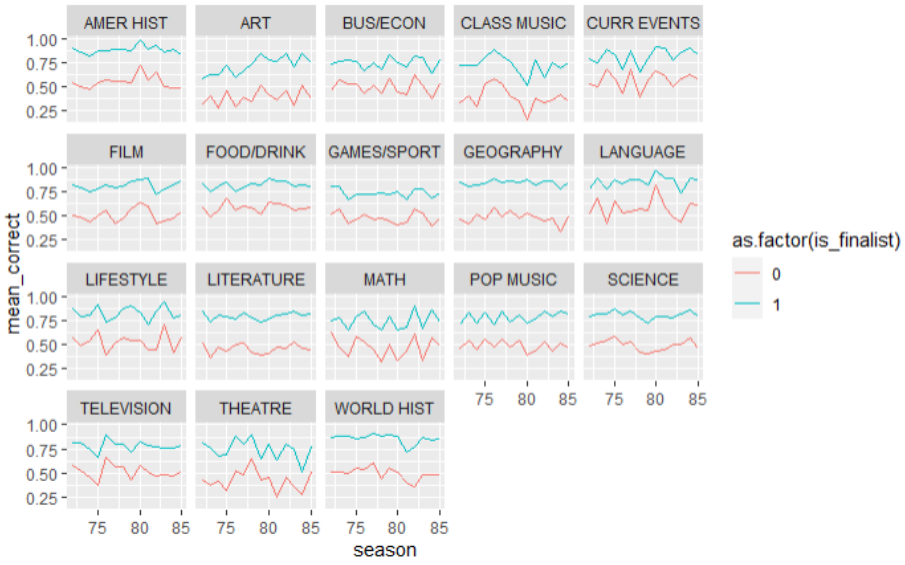
Notice that this graph is extremely similar to that above graph with different cohorts (finalists vs. non-finalists)! However, to be confused, this graph directly above shows the differential performance of the

finalists during the regular season (outlined blue), and the championships(outlined red). This empirically shows our second idea that the championships are much harder in terms of correct rates is true.

Furthermore, one could even argue that the finals questions have been designed so that it would exactly flatten the 'finalists' curve to have a similar distribution to that of the regular season participants. Let's see if this holds. Now that we have seen these two comparisons, lets check how the finalist do in the championship versus how non-finalists do in the regular rounds (basically overlaying the two red distributions from the previous two graphs)



Finally, checking the categories between finalists and non-finalists, we find the following graph which shows differential performance



In this large collection of graphs, the mean correct rate of individuals who got into the finals vs. the mean correct rate of individuals who did not get into the finals is shown. Notice that for all the graphs, those who got into the finals will have a strictly higher level scores than those who didn't (which makes sense, because that's how they got into the finals in the first place). However, what is more interesting is the trends over time. In particular most of the categories (such as Television, Science, Classical Music, Current events etc.).

Furthermore, we see that each category is not likely drawn from an uniform distribution. This, combined with the fact that each category has on average different mean correct rates, means that this will be quite important to take into consideration when modeling and predicting whether or not an individual will get a random question right.

To sum up, I have shown that looking at simple aggregate data on seasons and matches in the regular season data, as well as round and questions data in the championship data, all exhibit random behavior which does not help to predict the correctness of an individual's answer. Furthermore, I have shown that for different categories, the mean correct rate throughout the years are clustered differently. Furthermore, we see that finalists are strictly better than non finalists in each category in terms of mean correct rates. This might lead to an predictive variable we can use to predict an individual's ability to get a particular question right. As seen in the data this holds because, if the individual is a finalist instead of a non-finalists, and the question is about American history and not art, then it must be that we should expect her to do better than the case where she was a non-finalist doing art (check figures on categories for concrete justification). Finally, I have shown that finalists and non-finalists have very different mean correct rates. In particular, finalists are far better at answering regular questions than non-finalists (since that's how they are chosen as finalists anyways). However, finalists performance drop dramatically once the actual championship questions are presented, where they share a very similar distribution in mean correct rates as the non-finalists in the regular season.

Question 2:

Build a good model for predicting whether or a not a person will get a particular question right in the regular season. Describe why you made the modeling choices you made.

Response:

At the core of this prediction, this question boils down to two intuitive points that we need to ask ourselves before we respond to this individual A: how good at answering questions is individual A for answering a specific question type X? Motivated by these two points and connecting this with what we have discovered in question one, I believe a good model would include the following considerations:

To measure how "good" an individual is, I believe it is most intuitive to take do the following. Before anyone does the examination, each contestant is a blank slate with no information. Taking the questions, I will assume that this is a complete reflection of an individual's skill. Therefore, I retrieve the information "how good are you at answering a specific question?". Based on the observed information, I am effectively saying that "If the category is not hard for you (i.e. you got good scores in the past), then you shouldn't have a problem answering questions of the similar categories in the future, and vice versa". This will be the guiding principle of constructing my model.

However, there is another aspect to this. Just because someone is really good at economics in general, doesn't mean that he will get a good score if all the individual questions are super hard. Therefore, here is the second piece of the two-pronged approach: not only will I need to consider how good is an individual at answering a certain type of question, but I also need to consider how relatively difficult is the question. In order to do this, I will introduce a simple measurement of how difficult a question is, which is the mean correct rate of a question particular question for everyone in the given sample. One could argue that the two variables have interactions which will cause confounding, but I believe this is not the case. This is because we have effectively separated out the two variables by defining difficulty and individual skill like

so. In particular, if an individual's skill in a particular category has changed, that will surely affect a particular question's measure of difficulty. **HOWEVER**, this will not actually go back and re-influence the measure of individual's skill. **Therefore, these two separate variables of "individual's skill" and "question difficulty" will not be an issue where the loop back and feedback into one another. (this theme of "individual skill" and "question difficulty" will be the main idea for most of the assignment)**

In constructing the model, I will take this as the scenario: it is the case that we have both predictor and outcomes in the training sample (i.e., we have all the data before 2020), this is an in-sample prediction question, where we allocate regular season data in 2020 (id, question category, match number etc.) for testing purposes.

The model will make the assumptions/observations that:

- individuals scores fully reflect their inherent abilities (measure of skill)
- question type is not uniform in difficulty, some questions are more difficult than others (categorical differences)
- match, season, and year variation are uniform (no fixed effects)

Note that points 2 and 3 are proven empirically in question 1, where we have seen that different categories have different distributions, and that match, season, and year variables are like draws from a normal distribution, not worthy of fixed effects. The main lever that I will pull will be the individual's performance on specific categories of questions. i.e. if you were really good at economics/business before, I would have more faith on you to get the next economics/business question right than someone who really sucked at economics/business before.

Following that, we starting out with the extremely simple OLS model (check code for full model):

In this basic OLS regression, where we considered individual's performance as well as how difficult is this category among their peers. To restate, I have trained the model on the 2017~2019 subset and

tested on the rest of the regular data of 2020. Although this extremely parsimonious prediction does give a fairly decent predicted accuracy of around **71.5%,** we should still explain the model and the thought process involved.

To conclude, we have built an extremely parsimonious model where, by checking each individual's performance in each category type, we estimate how one's previous performance in the specific question type will influence the rate at which the same individual will get a question of the same category correct. Once we have completed the model, we get a point estimate of **0.94** on individuals ability to answer a particular category of question (id_cat_mean) with S.E. of **0.0005**, and **0.94** on the measure of difficulty on the question (q_difficulty) with S.E. of **0.0005**, which means that for every percent that an individual shows that they are good at a particular question type, we expect that they have a 0.94% more likely to get the question right next time, where as for every percent that the question is 1% more difficult, then the individual will be 0.94% more likely to get a question incorrect. Effectively, this is saying that if you are 1% better, but encounter a new question that we know is 1% harder, I suspect that your "ability" will negate the "difficulty" of the question. (This is extremely simple and an extremely trivial result. However, this is to be expected given the bare-bones data that we have.) Say if we had education level data, then I expect this estimate would be different) In the end, I have checked to see that our model has an correct rate of 71.5% by training on the 2017~2019 data and testing on the 2020 data.

Question 3:

Build a good model for predicting whether or not a person will get a particular question right in the first two rounds of the championship using ONLY information from the regular season. Describe why you made the modeling choices you made.

Response:

The wording of this question is not entirely clear as to who and what are we predicting. Therefore, I will interpret the question as follows: What is the overall likelihood that a someone gets a championship question X correct.

There are few pieces of information that we must consider before going into this question:

- the finals questions are empirically more difficult than the regular season data (see question 1)
- championships data do not provide any question categories
- only an extremely small subset of individuals actual participate in the finals (approximately 2.5% of all participants).
- we are limited to only running regressions on the regular season data, but looking at the finals data for general information (graphs) is allowed.
- the distribution of categories in the finals data will be same as the regular season's data
- we are only predicting the first two rounds of data, meaning that we won't be able to consider the differential between those who make it and those who don't just yet.

The main point of interest is with the first piece of information provided above, where we have already seen that the finals questions are very difficult relative to the regular questions (since the cluster of finalists which were skewed to the 80th percentile was pushed back to a normal distribution centered at around the 50th percentile). Therefore, the main type of questions that we must consider is the "more difficult questions", and we can no longer use the category variable directly here. I concede that my last

point is very unstable, where the we have no idea whether or not the distribution of categories are the same in the finals as in the regular season. I do not believe that my sampling of questions will be indicative of finals data because some categories are much more relatively difficult than others, and because of my sampling method (stated below), I will necessarily be unintentionally selecting a biased result. Since we have seen that the distribution of categories throughout the years in the regular season data is very similar (see question 1), I am making the tentative assumption that the unobserved distributions of categories in the final data will be the same as that seen in the four regular season's data.

Now that we are only left with even more bare-bones data (without), and also that we are stuck with the regular season data, then I think the most intuitive thing to do here is to look at how individuals do in the hardest questions in the regular season as a "proxy" for the questions in the championships.

Following our definition of difficulty, I must first choose what would be the cutoff for a "difficult question". While this sounds quite arbitrary, I will do my best to avoid introducing biases into the analysis. In particular, I will use a methodology that is extremely similar to the one from the previous question, where I simply substitute the "subject category" into the "difficult category.

Quick word on cheating: since we are using regular season data to predict finals data, it must be the case that I am assuming that those who potentially cheat also "want" to cheat during the finals. i.e. if they are allowed to cheat, then our predictions should be sound. However, because finals are cheat-free, it is necessary that our model will misfit those individuals who cheat.

The model will make the assumptions/observations that:

Once again, please note that points 2 and 3 are proven empirically in question 1:

1) a quick calculations looking at the number of IDs in the regular season vs the championships reveal that, throughout the years, there has only been 2.5% of all participants who were ever involved in the finals.

2) throughout time, the distribution for the finalists category moves from a sharp peak centered at around 80% correct rate in the regular season to a distribution highly resembling the distribution of mean correct rates for non-finalists during the regular season.

Furthermore, these are some assumptions which we are making:

- individuals who cheat will want to cheat in the finals, but can't (consistent action)
- only a small and selected subset of individuals actually participate in the finals (selectivity)
- difficulty in the questions raise dramatically during the finals (increased difficulty)
- match, season, and year variation are uniform (no fixed effects)
- consistent category distributions between regular and final data

Then, regression itself will take the highest 2.5% quantile (since about only 2.5% of individuals are admitted into the finals, and it has a good interpretation of 3 standard deviations above the mean), and check their performance on the most difficult question (so that the training data tries it's best to approximate the mean and standard deviation of the performance of finalists during the championships). The training will be conducted on the regular season data from 2017~2019, and the model will be tested on hardest questions from 2020 which approximate the finalist distribution.

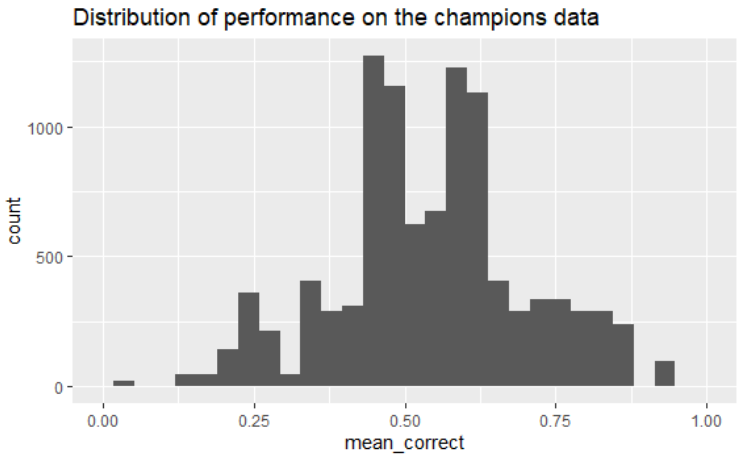
Please note: I am at no point constructing a model which uses championship data.

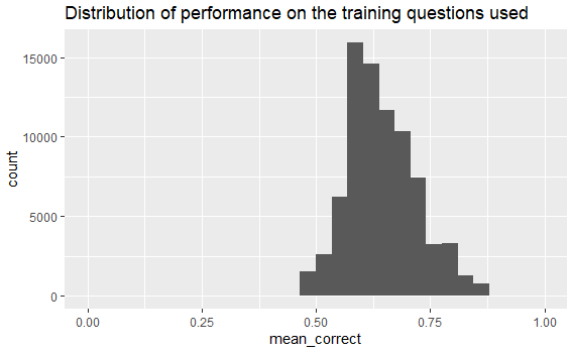
To conclude, we have created the most skilled individuals (top 2.5%) and considered their performance on the hardest questions (which approximates finals data the best to our ability, see below for graph). As a result, we have produced a OLS regression which states that, for every percent that a potential finalist is better at answering hard questions in the regular rounds (mean_correct), they are **0.99%** more likely to actually get a particular question during the finals correctly with an S.E. of **0.0214**. Furthermore, the second term that measures question difficulty (mean_correct_of_q), meaning that for each percent increase in difficulty here means a **1.43%** drop in the probability in getting this specific question correct, **0.0231**. This makes intuitive sense because we have consciously taken a biased sample

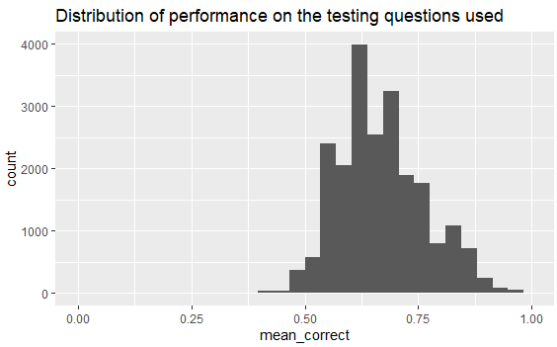
of difficult questions, and what this is saying is that "because the questions are so difficult, even if you are 1% better than you were before, I would still lose some faith in you (-0.44% worse) to get the question wrong, not because of your lack of ability, but because the question is so much more difficult". In effect, this actually explains how the finalists are swimming against the tide, where despite their talent, the difficult questions in the finals will decrease their overall correct rates!

Nevertheless, I have to mention one aspect of the regression, which is the absolutely abysmal R^2 value of **0.0707**. We know that a good R^2 value (or at least one that is not so close to zero) will be much more preferred, but I simply cannot think of a clear-cut variable that we can construct given the data that we are provided which will both make intuitive sense and also bolster the regression strength.

NOTE: Below is the distribution of the regular season questions' mean correct rate among finalists used to train the model, test the model and the actual distribution of questions in the championship data. Notice that we have somewhat approximated what we would have expected in the finals.







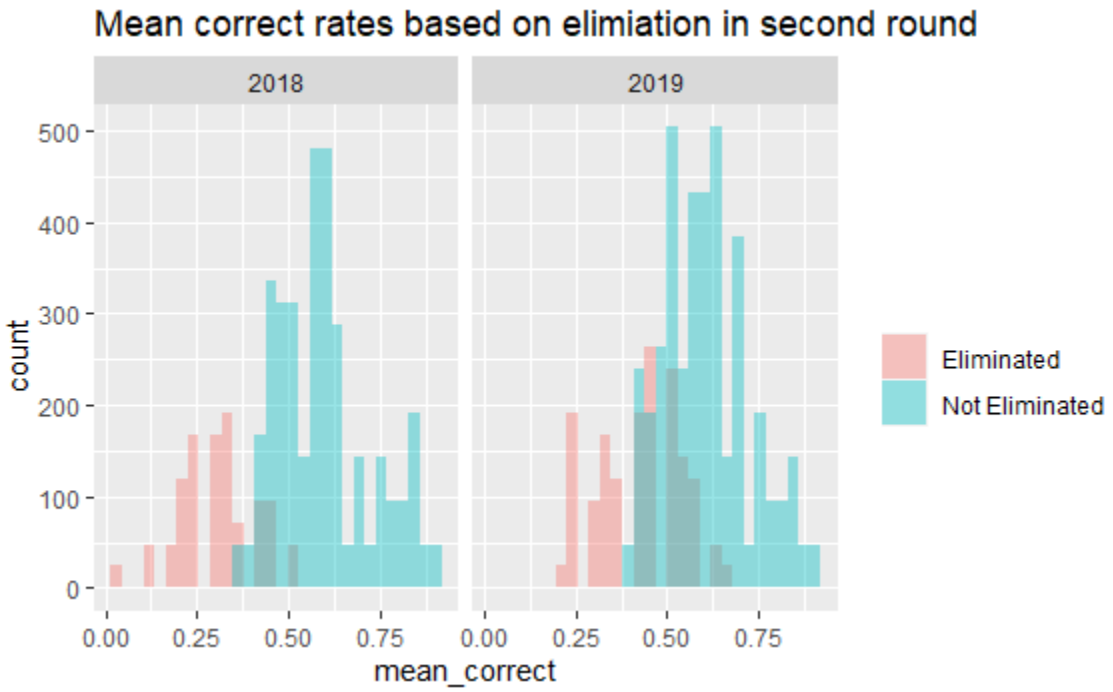
Question 4:

Using your model from question 3, do you find compelling evidence that some players were cheating in the regular season? What share of the players who qualified for the championship do you think cheated substantially in the regular season? What's your evidence?

Response:

For the following question, I will define potential cheaters as those who originally had a very high scores in the regular season, but did unexpectedly poorly during the finals. 'Unexpectedly Poorly' will be defined as falling two standard deviations lower than the average predicted scores from our model in question 3. Therefore, I will make the assumption that for this question, we are allowed to use championship data for testing purposes.

However, we should note that individuals are eliminated if they are not doing well in the finals at the second round. Checking the graphs shows that these two groups indeed have differential distributions:



That being said, I would like to note that because most cheaters should be eliminated in the first two rounds (by our definition of a cheater), then I will primarily take the group that was eliminated to do the analysis.

I will make the following notes before running the model:

- a potential cheater is ****NOT**** an actual cheater, it could have false positives where we claim someone who had a bad day was also a cheater. Therefore, I will frame this as just taging individuals which did unexpectedly poorly (as per our definition).
- I am currently assuming that our model in question 3 is the 'correct' model for estimation. In doing so, I am currently ignoring the possibility that our model is causing unwanted bias. (I am ****NOT**** saying that my model is good, rather the R^2 would tell me otherwise, but I am making this assumption for ceretus parabis analysis)
- I am currently only looking at the first two rounds to eliminate variation caused by the elimination. I am doing so because I suspect that cheaters, who cannot cheat in the finals, should not be able to make it past the elimination (since they are by definition those who do extremely poorly)
- Recall that the model for question 3, I have assumed that cheaters have “consistent action”, meaning that my model is predicting their performance "if they had kept on cheating". However, because finals are cheating-free, my model will be far off from what is actually the case. Therefore, this characteristic will be exploited to catch these potential cheaters

Now to estimate the model before on actual championship data, I assume that the modeling framework provided in question 3 can be used to with championship data, thus we are actually able to see who could be cheating and who is not. Therefore, I will take all the data from the 2018 and 2019, because these are the years where we have championship data. First off, however, I would like to distinguish between those individuals who have made it past the filter in the second round and those who don't. (check code for full model)

Here, we have found two individuals who satisfy the criteria for a potential cheater. In particular, we have seen that these two individuals' true finals performance is two standard deviations away from the predictions given by our model in question 3. This means that there are **2 individuals out of 227 finalists in 2018 and 2019 which satisfy our requirement**, amounting to around **1% of all finalists being potential cheaters**. If learned league truly claims to be such a prestigious competition, then this might be right.

Question 5:

Using whatever information we gave you, come up with your best prediction for the number of correct answers each player who qualified for the 2020 championship got correct in that championship. Describe your modeling choices. You will be graded based on the accuracy of your predictions.

Response:

I WILL BE ESTMIATING THE <u>FIRST TWO ROUNDS</u> ONLY, AS TO BE CONSISTENT WITH THE REST OF THE QUESTIONS

Since we are allowed to use every piece of information available, it would be ideal to introduce finalist data into our predictions. I believe that the best predictor of finals data will always be other finals data rather than regular season data since characteristics in skill and difficulty are relatively the same between finalist seasons, and quite different between regular seasons (see question 1).

Therefore, I will use the finals data whenever possible, and only resort to regular season data when needed (say for someone who has never participated in the finals before). Taking a quick look at the championship roster in 2020, we see that there are around 565 finalists in total for all the years, and 504 finalists in 2020, with 338 new finalists. While the previous finalists will be trained on the original finals data, the new finalists' predictions will be trained on the hardest questions that they have done in the regular season data (since new finalists by definition must be from the 2020 season).

For the new finalists, I have created a model with an accuracy rate of 69%, and the point estimate for individual skill (mean_correct) is **0.991** with an S.E. of **0.02** and the point estimate for the question difficulty is **0.977** with and S.E. of **0.033** (mean_correct_of_q). The interpretation is the same as for previous questions. (check code for full results)

Now for the previous finalists: I have constructed the expected questions correct in the previous finalists data as well, with model point estimates of **1** for both individual skill (mean_correct) and question difficulty (mean_correct_of_q) respectively. (check code for full results)

Now combining the two dataframes, the predictions for the first two rounds are given as **"Hu_Yixin_predictions_first_two_rounds.csv"** to indicate that I have only estimated the first two rounds, and submitted to Dropbox as required.