August 7, 2014 3:52, Raphael Silberzahn wrote

Dear all,

We are very grateful for your contributions to the crowdsourcing research project. After receiving all outstanding submissions, we are happy to inform you about the results. In total we received results from 29 teams, involving 61 researchers worldwide.

A wide variety of analytical approaches have been used and have resulted in varying conclusions by researchers. After submitting their final results, ofthe 29 teams, the leaders of 16 teams found it likely or very likely thatsoccer referees tended to give more red cards to dark skinned players. 7 team leaders found this to be unlikely and 6 team leaders found it neither likely nor unlikely.

For research questions 2a and 2b, results were more homogeneous. Believes in an association between skintone preferences in referees' countries of originand number of red cards for dark skinned players was indicated for implicit preferences by 3 team leaders and for explicit preferences by 1 team leader. All other teams found such association of implicit/explicit country scores unlikely (16/18) and 12/10 team leaders found it neither likely nor unlikely.

All the reports and a summary file can be found here:

https://osf.io/vae2d/files/

The diversity of analytic approaches and conclusions is a key "finding" for this crowdsourcing project. Despite the same research question and dataset, there was substantial diversity in the analytic approach to answering the question and the conclusion reached. This alone is a notable finding and,as far as we are aware, this has not been demonstrated in the literature. Illustrating this diversity, and the evolution of the analysis strategiesand conclusions from the two rounds of analysis will be the first theme ofthe paper.

There is, however, another important question to resolve: Does player skin-tone predict likelihood of receiving a red card? Indeed, the present results do not offer a definitive answer. To resolve this question, we now transition to open discussion among the teams.

There are multiple ways to resolve the question including, but not limited to:

1. Use the central tendency of all analysis strategies

2. Use a weighted central tendency with weightings based on features of each analysis, such as our collective evaluation of their appropriateness giventhe data

3. Selection of a single analysis (or a small subset) that represents the most defensible methodology for evaluating the research question

We think that the paper will be instructive to be inclusive of presentingall of these approaches and noting their convergence or discussing the reasons for their divergent conclusions. For example, it could be that one family of analytic techniques tended to

elicit a positive result where as others tended to elicit a negative result. And, there may be clear or debatable reasons to prefer one approach to the other.

Besides the general approach to answering the research question, there are some particulars in individual analysis strategies that are worth discussing as a group. These may inform updates to individual analyses, or suggest a final collectively-determined analysis strategy that considers the variety

of issues identified by the various teams. Some illustrations:

1. Team 17 first found a positive association but after removing 7 outliers (Player IDs 544,1804,1136,1343,1158,18,418) the effect disappeared.

2. Team 12 was the only team that reported an interaction between skin color and implicit bias in the referee's country of origin on red cards. Why didother teams not find such an interaction?

In sum, now is the time for us to collectively discuss the results, thevariation in strategies, and their implications for answering the original research question. Your comments in reply-all are welcome and encouraged.

Soon, we will also make available a google doc with a starting draft of the manuscript so that we can transition this open-ended discussion into the concrete discussion in the paper.

Once again thank you very much for your participation in this exciting project. Have a look at other teams' reports, they are very interesting!

All the best,Raphael, Eric, Dan and Brian

From: <bahniks@seznam.cz>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 11:50:11 AM GMT+02:00

Hi all, couple of thoughts.

1) From a cursory view, it seems to me that teams using some form of multilevel modeling approach were more likely to observe the effect of dark skin tone in the first question. While there are exceptions, this might explain a large part of the difference in results.

2) I think that there might be a problem with understanding the summary questions. At least I had a problem with their interpretation. For instance, the first question "How likely do you think it is that soccer referees tend to give more red cards to dark skinned players?" may be intepreted in couple of different ways. It might be viewed as: 1) how likely do you think that the effect follows from the result; 2) how likely it follows from the dataset; 3) how likely it applies in general; or 4)

how likely it applies in general if we were able to take into account all possible covariates. The first interpretation is related only to the results of a given analysis and might take into account just some uncertainty from possible mistakes in the analysis. The second intepretation includes uncertainty from the chosen analysis method. The third adds uncertainty from the dataset. The last interpretation would be about a general opinion about the existence of bias. Pre-existent opinions about the existence of bias should play a different role in the four interpretations. While they should play no role in the first interpretation, they should influence more the other intepretations. If we want to use the answers to these questions in describing the results, it might be good to clarify what is meant by them. My preferred interpretation was 1) or 2), but it seems that some people interpreted the questions in a different way since they answered that they don't think that an effect is likely even though they observed the effect in their results.

3) Would it be possible to release the summary data of the observed effect sizes? It would be easier to interpret the results then.

Best, Štěpán

From: Tim Heaton <t.heaton@sheffield.ac.uk>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 2:23:19 PM GMT+02:00

Also I also think it would be preferable to have some kind of online forum where we can discuss this rather than it all take place on email. In my view it'd be much easier for me to view thread progression rather than trying to track lots of emails responding to potentially different issues. It'd also mean that those who reply to messages doesn't get lots of out of office autoreplies.

Tim

From: Erikson Kaszubowski <erikson84@yahoo.com.br>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 8:24:07 PM GMT+02:00

Dear all,

Disaggregating the data was done by many teams that opted for a Bernoulli likelihood. We and some other teams didn't disaggregate the data but used a binomial likelihood instead. The binomial regression estimates should be exactly the same of a logistic regression with the disaggregated data, but without the burden of actually having to reshape the dataset.

Given the dataset and our research questions, I think that negative binomial and Poisson are possible options, but hardly justifiable given what we are interested in: even using the games as an exposure variable, we would have to truncate the distribution beyond one because it's not possible to receive more than one red card per game. As far as I understant it, using and exposure variable allows us to model the rate of events (cards/game) instead of raw counts, and our rate is bounded between 0 and 1. Red cards are rare events and truncating might not influence the coefficient estimation much, though (but would give a lot work to implement!).

In my opinion, the 4th issue pointed by Lammertjan is our main problem: I guess no model is going to be able to differentiate between referee bias and player agressiveness

when estimating skin tone effect. I thought that a mixed effect models with skin tone coefficient varying by player and referee could do it, but I couldn't run the model because it is too big for my computer.

Garret's points are also interesting. The conditions he stated for the research question ('given the dataset and inability to identify causal effects') certainly influenced the model we developed. There are tons of covariates that could be present, but, given what we have, some covariates help us estimate skin tone effect better *under a predictive interpretation*. While it is easy to slip to causal interpretation, it should be clear, as stated by the project leaders, that causal inference wouldn't be possible give the available dataset.

I think that, for the final article, we should go with something along the lines of option (2) (weighted central tendency). While I still do not know the best way to do this, I agree with Alicia that the results should reflect the variaty of approaches. We could go crazy and do some kind of bayesian model averaging with weights based on our priors for which model best suits the problem and dataset, but this would take a small cluster to compute given the number of models, parameters and data points.

Great discussion! I hope we don't drive the project coordinators crazy, though!

Best regards, Erikson

From: Seth Spain <smspain@gmail.com>**Subject:** **Re: [Spam:******] Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 29, 2014 12:29:05 PM GMT+02:00**To**

Hi all,

To follow up on Lammertjan's point. Our original analysis was a type of Poisson (we started with the intention of controlling for a variety of game-sum factors, like goals, but nots games themselves). The mid-round feedback convinced us that a binomial logistic model made the most sense, modeling red cards ~ binomial(theta_i, games) where theta_i is the linear predictor. Regardless, the point I wish to emphasize is that the internal review and feedback process *strongly* affected our approach.

In the end, we fitted the model using glmer. We tried using both JAGS and Stan to fully Bayesian analyses, which worked fine for relatively small subsamples, but we couldn't get either to run in reasonable time for the full data--I'm curious if the groups that did Bayesian analysis used more computationally efficient programming (esp. in Stan, which I had expect to run quite fast for this kind of a problem), or if they are just more patient than we were.

Best, Seth

From: "Morey, R.D." <r.d.morey@rug.nl>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 13, 2014 9:28:59 AM GMT+02:00**To:**

Hi Erikson, all,

Yes, it was averaged over dyads. Two things: I had plotted it both ways, but it doesn't change the outlier plot that much to take a sum (actually the sum is in the code but commented out). These are extreme points. I ended up settling on that plot due to the deadline (I would have done the outlier selection slightly differently with more time, but the model takes a day to run). The model fits, however, do not rely on the averaging choice, yet tell the same story.

Second, I think we should focus on the medians look in your box plots. If the effect were robust, wouldn't you expect to see regularity there, as well as in the mean? The means are a questionable descriptive here, due to the high skew and, as you point out, large numbers of zeros. If the medians are not regular, the effect in the means must be driven by a minority of higher-card players. It is possible that our random effects model yields enough shrinkage to moderate those effects if they are of moderate size, but simply cannot do so with the extreme outliers. I am not sure, but that is my hunch.

At any rate, I'll rerun the analysis when I get back home next week. It would be strange if the outliers only had an effect on *our* analysis, so we should get to the bottom of it.

Best, Richard

From: Frederik Aust <frederik.aust@uni-koeln.de>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 11, 2014 8:14:54 PM GMT+02:00

Hi everyone,

thanks for plotting the estimates, Dan. This is a nice overview of the results.

I have another small correction, though: Our analysis (Team 28) estimated the OR to be 1.382 [1.020, 1.705]. The plot shows an estimate of 1.205 with a very narrow confidence interval.

Best regards, Frederik

From: Dan Martin <dpmartin42@gmail.com>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 11, 2014 8:24:49 PM GMT+02:00

Sorry for the repeat email, the last figure had an error. The OR column and the CI plot should be flipped. Thanks to those who pointed out the error to me. I uploaded the new file to our osf project page to avoid more correction emails (https://osf.io/68vpr/)

If you think there might be another error (either in effect size, analytic approach, or otherwise), feel free to send me an email and I'll make the correction (just reply to me, no need to involve the whole group for minor corrections)

Thanks! Dan

From: "Morey, R.D." <r.d.morey@rug.nl>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 13, 2014 9:53:35 AM GMT+02:00

One other thing; I think we were the only group to use all the data, including those without skin tone data. I can't see how this would change the effect of the outliers, but a very large proportion of the data was kissing, so it is conceivable that this could account for some of the difference in the outlier effect.

I will rerun our analysis throwing out players with no skin tone data, to see what happens.

Best, Richard

From: Erikson Kaszubowski <erikson84@yahoo.com.br>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 7, 2014 6:48:21 PM GMT+02:00

Dear Raphael, Eric, Dan, Brian and all other teams!First of all, congratulations on the exciting project! I am quite thrilled with the results, and I have yet to read all the reports.How should we discuss the results? Is it possible to setup a forum in the OpenScience Framework to organize it? Or should we keep it in the e-mails? Anyway, I checked the two issues you mentioned and have an opinion on them.About the effect disappearing when removing seven outliers:
The dataset we worked with in our model already excluded most outliers detected by Team 17, because they were missing information on player position and/or skin tone rating. In fact, only one 'outlier' remained with 0.055 proportion of games with red cards – hardly far from the distribution.The main effect for player skin tone remained statiscally significant (95% CI excluded zero) in our final model.How can we explain such a difference when both Team 17 model and ours use the same likelihood function? Three outliers didn't have information on skin tone rating, so they should not impact so much the coefficient estimation.

The Team could provide more details on the model specification. We are specially interested in why the team used a categorical distribution to model the skin tone ratings instead of using the values directly.

About the interactions with the rating variable:

Our final model with skin tone rating coefficients varying by referee country included coefficients for implicit and explicit country bias in the second level of the hierarchical model. They can also be interpreted as an interaction between the rating and both country bias variables. Our estimates had high uncertainty, so we did not discuss them much.The difference might be due to the model specification: Team 12 used a zero-inflated poisson regression. We thought about the possibility of using a Poisson regression, but we decided that it was not adequated to model the data because a player can receive only one red card per game. Well, that's it for now. I will read the remaining reports to make some suggestions on how to aggregate the results.

Regards, Erikson

From: Raphael Silberzahn <rts27@cam.ac.uk>
Subject: **Crowdstorming Project: It's time to discuss the results!****Date:** August 7, 2014 8:51:27 AM GMT+02:00
Dear all,

We are very grateful for your contributions to the crowdsourcing research project. After receiving all outstanding submissions, we are happy to inform you about the results. In total we received results from 29 teams, involving 61 researchers worldwide.

A wide variety of analytical approaches have been used and have resulted in varying conclusions by researchers. After submitting their final results, of the 29 teams, the leaders of 16 teams found it likely or very likely that soccer referees tended to give more red cards to dark skinned players. 7 team leaders found this to be unlikely and 6 team leaders found it neither likely nor unlikely.

For research questions 2a and 2b, results were more homogeneous. Believes in an association between skintone preferences in referees' countries of origin and number of red cards for dark skinned players was indicated for implicit preferences by 3 team leaders and for explicit preferences by 1 team leader. All other teams found such association of implicit/explicit country scores unlikely (16/18) and 12/10 team leaders found it neither likely nor unlikely.

All the reports and a summary file can be found here: https://osf.io/vae2d/files/

The diversity of analytic approaches and conclusions is a key "finding" for this crowdsourcing project. Despite the same research question and dataset, there was substantial diversity in the analytic approach to answering the question and the conclusion reached. This alone is a notable finding and, as far as we are aware, this has not been demonstrated in the literature. Illustrating this diversity, and the evolution of the analysis strategies and conclusions from the two rounds of analysis will be the first theme of the paper.

There is, however, another important question to resolve: Does player skin-tone predict likelihood of receiving a red card? Indeed, the present results do not offer a definitive answer. To resolve this question, **we now transition to open discussion among the teams.**

There are multiple ways to resolve the question including, but not limited to:

1. Use the central tendency of all analysis strategies

2. Use a weighted central tendency with weightings based on features of each analysis, such as our collective evaluation of their appropriateness given the data

3. Selection of a single analysis (or a small subset) that represents the most defensible methodology for evaluating the research question

We think that the paper will be instructive to be inclusive of presenting all of these approaches and noting their convergence or discussing the reasons for their divergent conclusions. For example, it could be that one family of analytic techniques tended to elicit a positive result where as others tended to elicit a negative result. And, there may be clear or debatable reasons to prefer one approach to the other.

Besides the general approach to answering the research question, there are some particulars in individual analysis strategies that are worth discussing as a group. These

may inform updates to individual analyses, or suggest a final collectively-determined analysis strategy that considers the variety of issues identified by the various teams. Some illustrations:

1. Team 17 first found a positive association but after removing 7 outliers (Player IDs 544,1804,1136,1343,1158,18,418) the effect disappeared.2. Team 12 was the only team that reported an interaction between skin color and implicit bias in the referee's country of origin on red cards. Why did other teams not find such an interaction?

In sum, now is the time for us to collectively discuss the results, the variation in strategies, and their implications for answering the original research question. Your comments in reply-all are welcome and encouraged.

Soon, we will also make available a google doc with a starting draft of the manuscript so that we can transition this open-ended discussion into the concrete discussion in the paper.

Once again thank you very much for your participation in this exciting project. Have a look at other teams' reports, they are very interesting!

All the best,

Raphael, Eric, Dan and Brian

On 08/08/14 10:12, Dam, L. wrote:

Dear All,

I would be very much in favor for option 3:"Selection of a single analysis (or a small subset) that represents the most defensible methodology for evaluating the research question".

I don't mean to be blunt/arrogant/offending people, but was invited to be critical:

1. Using count data techniques (Poisson, Negative Binomial) is simply wrong. It assumes the same number of games played for each player, and controlling for number of games does not solve this problem. ANOVA is throwing away information.

2. I actually believe that the way the data is structured is not ideal and I think this is the reason for why people use so many different approaches. If the data was on game level, I guess almost all of us would have used a binary outcome model (Logit, Probit, LPM, etc).

3. There are also many possible omitted variable biases: for example, we lack a time dimension. If referees have become more strict over time, and soccer teams have become more international over time, this would create spurious correlation.

4. It is impossible with this data to disentangle whether dark skinned players simply make more fouls, or referees are somewhat racist. I know the research question is simply whether they get more red cards, but the second research question points more

towards the second explanation of the findings. Even if this is the case, it may even be that referees are easily influenced by the spectators, and the spectators are actuallyracist, not the referee.

I know that the last 3 comments are quite general in nature and not easily solved by some methodology, but the point is that I think that we should be very careful in how we phrase the findings.

Regards, Lammertjan Dam

From: Raphael Silberzahn <rts27@cam.ac.uk>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 12:52:49 AM GMT+02:00

Dear Erikson, Dear all,

Thanks for your interesting thoughts! We would encourage the discussion via e-mail (rather than on the OSF) to enable quick replies and have everyone included.

Thanks Silvia for pointing out that the report from Team 27 was not in the zip file. It is now. All other pdf reports have also been uploaded directly to the OSFfor convenient online viewing.

https://osf.io/vae2d/files/

This discussion of results is important for shaping the contribution of our research. This is a first for everyone and I believe we can achieve a great result by having an open discussion. Some have written me with concerns. You may feel hesitant to be critical of other approaches. It is important however, that all participants in this project are scientists who are interested in finding out what is closer to the truth. We encourage a culture in which we can openly discuss about the validity of different approaches without misperceiving critique of a work to be personal. It is normal that we don't know the ins and outs of every method. Some of us have applied a method they have previously been unfamiliar with and have explicitly stated that they seek feedback to help them learn.

Further, I believe the format of this project gives us a great space to openly discuss, learn and update our knowledge with the help of each other and prior to engaging the wider (scientific) public. As none of the results have been published, there is no need to take a defensive stance and the more open we discuss, the more WE as a joint research are prepared to defend our final conclusions!

In case that this process leads you to change or update your individual finding or report, it is important to note that you will be given a small space to state how if your view of your results have changed during the discussion. This phrase is then presented prominently alongside your one-sentence summary and you can include a link to further analyses if you wish so.

It will be great to hear your opinion and impressions! Best,Raphael

---

From: Silvia Liverani <liveranis@gmail.com>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 1:45:35 PM GMT+02:00

Dear all,I completely agree with Lammertjan that using count data techniques is wrong because it assumes the same number of games played for each player.Lammertjan, could expand on why you think it is wrong also when controlling for number of games?I am against the first of Raphael's suggestions (1. Use the central tendency of all analysis strategies) as, like Lammertjan, I believe there are some major flaws in a few of the methods used.

However, aside from methods which have been misused in this context, I believe that a large number of the methods used are appropriate, but based on different sets of assumptions and model choices. Therefore I am very much in favour of the second of Rapahael's suggestions (2. Use a weighted central tendency with weightings based on features of each analysis, such as our collective evaluation of their appropriateness given the data).

Going for the third suggestion (3. Selection of a single analysis (or a small subset)) might be tricky, but I would find that sensible as well.

At this point I think we should all evaluate all methods used, giving a value on a scale on whether we think the method is appropriate. Or should we use a small pool of "experts" on the methods used to judge which ones are most suited to answer the research question? The experts could be the project leaders (ie Raphael, Dan, etc) if they felt they could do the job well. Alternatively we can do a mix of the two suggestions above: Raphael, Dan etc make the final call on the methods to be included in the final conclusion, based on our recommendations of methods which should be excluded.

Thanks to everyone who has partecipated: it has been a really interesting project to be part of so far!

Best wishes, Silvia

From: Tim Heaton <t.heaton@sheffield.ac.uk>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 2:14:02 PM GMT+02:00

Hi,

My preference would be either option 2 or 3. If we are going to go for Option 3 (selection of a single analysis) then it would be good for someone (the project co-ordinators?) to first narrow down the contenders fairly significantly. Trying to get all of us to write comments and critically assess all 29 projects and end up with a single one may be practically difficult. If however we only had to critique and select between 4-5 I would find that much easier.

Having cut the contenders we could then all pose more questions to those project teams about questions we might have before making our decision, Tim

---

From: Alicia Hofelich Mohr <hofelich@umn.edu>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 3:19:19 PM GMT+02:00

Hi all, I would be interested in hearing more about the count analyses - I was under the impression that the offset in a Poisson model is meant to deal with different levels of exposure, which in this dataset would be number of games. I agree it may not be sufficient simply to add number of games as a variable in the model. But wouldn't the offset address your concerns, Silvia? Best, Alicia

---

From: Tim Heaton <t.heaton@sheffield.ac.uk>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 3:42:12 PM GMT+02:00

Dear Alicia and Silvia,

My view on this is as follows (I may have misunderstood your question though). You can use an offset in a Poisson model to account for different levels of exposure but in a different situation than modelled here. Fundamentally, for any player I don't see the number of red cards they receive given n games played as a Poisson random variable (you can't get 4 red cards in 3 matches).

Potentially if the data here were in terms of minutes on the pitch and you were interested in the red card rate then a Poisson model is likely to be appropriate with an offset for exposure (in terms of the minutes played). However that is quite a different model and we don't have the data,

Tim

From: Alicia Hofelich Mohr <hofelich@umn.edu>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 4:39:22 PM GMT+02:00

Hi Tim,

Is the issue you bring up the fact that using games as exposure places an upper bound on the poisson distributed counts? I agree it would be much better to use minutes of play as the exposure if we had it. But since not, each game within a dyad is essentially treated as a Bernoulli trial (which we also assume are independent, but that could be arguable) where a red card is treated as a "success". The Poisson with the offset is used to approximate a binomial distribution when the probability of success low and the number of observations are relatively high - which would result in an upper bound based on the number of trials observed.

Our team chose this route (using log(games) as the offset), because it seemed the most logical strategy without massive remodeling of the dataset (to get each row to represent a single game - which, to be honest, I started doing in my first attempt, but kept crashing the loop to reformat...). Thinking more about your comments, I think the first assertion of the approximation (low probability of success) holds more strongly than the second assertion (high number of observations), when at the game level, rather than with the number of minutes.

I do think if one reshapes the data, a logit/probit model would be the best option, as Lammertjan mentioned before. I also think that in terms of the article, even if we do decide on a single approach for analysis, it is important to capture the fact people were/are influenced by the way the data is organized. Whether this occurs because we are primed to think of certain models/analysis strategies when we see certain data structures, or whether it is because of actual limits faced in computational power or knowledge to reshape, I think we would lose something if this is not reflected in the results/discussion (and as a psychologist, it's such an interesting point!).

Best, Alicia

---

9 aug 2014 kl. 20:24 skrev Eric-Jan Wagenmakers <ej.wagenmakers@gmail.com>:

Dear all,

This has been an insightful project. I think what it has illustratedis that statistics is more than some kind of universal, objectivetool; it is a scientific discipline like any other. Consequently,different teams have come in with different assumptions, and reached different conclusions.

This diversity in outcomes is something to emphasize, not to hide. Therefore I strongly recommend against selecting a single "best" analysis (unless this analysis is the one Richard and I conducted, of course :)).

What *could* be done is to analyse the differences and similarities between the approaches and identify a few key features that may be responsible for the difference in conclusions.

And then a more specific comment related to the outliers. For our analysis (team 17) excluding the outliers made a world of difference. The details of our analysis are here: https://osf.io/wfvpc/ It wouldbe interesting and important to figure out why we reach different conclusions from Erikson's team. I have not have time to look into this carefully, but one key issue might be the choice of predictors. For instance, we had predictors for player aggression and referee strictness.

In general, figuring out why different analyses lead to different conclusions could require a rather time-intensive effort by itself. Nonetheless, one could proceed by looking at what likelihoods were used, and what predictors were included. Perhaps there is some hidden order here.

Finally, I think this project also underscores the advantage for research teams to preregister their analysis. If we were given the assignment "analyze the data in a sensible way and reveal an effect", we could have done so. And if we were given the assignment "analyze the data in a sensible way and show there is no effect", we could have done so as well.

In statistics, one usually acknowledges variability due to noise (sampling variability, or uncertainty). But most of these efforts assume the model is fixed. The current projects highlights that the uncertainty is much larger than we might think if we only look at the statistics for a single model that happens to be favored by the research team reporting

the results.

Cheers, E.J.

On 11/08/14 09:00, Rickard Carlsson wrote: Dear all,

Regardless of how we proceed next I think that this project has already been a huge success. This is one of the most intriguing, and fun, projects I have taken part in. _____ Regarding the alternatives on how we should present the data, I feel thata summarizing graph with all the odds ratios and 95 % CI along with some brief summary of the approach would be very illustrative. I think this isvery important because it will move the focus of whether we found all found statistically significant findings to whether our findings are similar and consistent in terms of a) effect size and b) precision.

I think this type of summary and discussion on it is important on its own right, but then we should be able to, based on some discussion, narrow it down to a few approaches.

As already pointed out, I think that the data structure is partlyresponsible for the different kinds of approaches used. In my initial approach I tried to disaggregate the data, but could not make it work. Ialso tried some analyses (multi-level modeling) that ended up crashing my computer. Initially, I settled for an analysis that I was not satisfied with(I noted this in the submission). In my final analysis I used a binomiallogistic regression that (as others have also pointed out) does the jobsimilar to disaggregating the data. It further did not crash my computer :).I do feel that this aspect of data structure and challenges along with it is important to take into the discussion. Some of us are more likely used to analyzing our own data, where we have perfect understanding (and control) over the data structure, whereas others are more familiar with the task of analyzing a given set of data. I find it interesting to hear that othershave also had some problems with computers, software etc. Personally, I do not have a version of STATA that is advanced enough to handle this type of large dataset and this has guided me in my choices. Had this been a project I had worked in, I would probably have upgraded my software or switched to something else. What I am trying to say is thus that if the data had been structured in a different manner and we all had software that is able to handle this type of dataset properly, then there may have been higher agreement among approaches.

Another aspect I find really important is how similar the approaches thatseveral of us feel are "wrong" actually perform. Further, it is quitedifferent if the approaches are invalid or if they are just unreliable. Forexample, analyses that throw away some data (as suggest that ANOVA would do) - what are the consequences for the analyses? Are they simply more blunt, or could the be biased in favor of a certain interpretation? Have some of use

chosen the wrong specialized tool from the tool box, or have some of us simply used a too blunt all-purpose tool? This is quite a different scenario in my view.

Finally, I found Team 17 discussion on outliers intriguing. My own approach to check for the influence of outliers was to base the analysis on censored data where it was only possible to get one single red card, reasoning that players that receive *several* red cards are probably quite unusual players that may influence the findings too much. This

did not change the conclusions I drew. Of course, having a single red card in a few games is more extreme than having a single red card in a great number of games, but I did not find these extreme enough to exclude. In either case, if removal of .3% of the data is enough to eliminate the effect, then the effect is indeed not very robust, which is not surprising since it is small.

Kind regards,

Rickard Carlsson

From: "Dam, L." <l.dam@rug.nl>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 11, 2014 10:20:48 AM GMT+02:00

Hi all,All very interesting thoughts. Maybe I am too cynical as well, but I too a large extent I can relate to Garret's remark: "I have trouble caring about complicated statistical details when there seems to be such a fundamental flaw."

Anyway, I still want to follow up on some of the questions/remarks: 1) A negative binomial distribution models the number of Bernoulli successes for a given number of trials. In this case the number of trials (games) is not the same for each observation. Even if you control for the number of games played, the model would still be able to predict 5 red cards for a player that has only played 3 games, for example if other covariates have a strong positive effect.

2) You do not necessarily have to disaggregate the data yourself. What we did was using "frequency weights", so that the software knows how many underlying observations each unit of analysis represents.

(By the way, by no means I think we have the "best" model. And I also do not know whether any of my concerns would lead to some sort of bias, but I thought at least I could mention them.)

3) I do agree that one has to make certain assumptions for any analysis, and so different teams will make different assumptions leading to different methods. But when the assumptions made are clearly violated, the researchers should be able to justify the approach by specifying whether or not it would lead to a bias. (e.g. Often data is not normally distributed, but OLS is fairly robust and can usually be applied). If they cannot, we should perhaps discard those analyses.

Lammertjan

Begin forwarded message:

From: "Dam, L." <l.dam@rug.nl>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 13, 2014 9:35:53 AM GMT+02:00**To:**

Hi Erikson,

I like Wagenmakers approach (we also used red-card proportion) but I agree with you that the red card proportion should be per number of games, and not a red card proportion per player. That is why we used frequency weights.

I redid our own analysis, throwing out outliers, and then the coefficient was just significant at 10%. But I doubt how robust this is to other specifications. Lammertjan

---

Od: Ismael Flores Cervantes <ismaelflorescervantes@westat.com>Komu: Dam, L. <l.dam@rug.nl>, Erikson Kaszubowski <erikson84@yahoo.com.br> Datum: 13. 8. 2014 18:34:31 Předmět: Re: Crowdstorming Project: It's time to discuss the results!

Hello,

My 2 cents. My initial impression of this project was to examine (and if possible to measure) the level of agreement among researchers when they deal with the same problem. In a way, this is an indirect way to study the level of agreement in more complex problems, for example, climate change which has been studied by many researchers using different methodologies (of course, most of them agree that it is real). If there is no agreement for simpler problems studied using the same data file, then there is no much hope for agreement in more complex problems. The fact that there are different results (product of different methodologies) is an interesting finding. The fact that a consensus seems to be forming (based on the results in plot) is encouraging (just a visual count shows 78% the teams found that odds of dark skin toned players are likely to receive red flags is statistically higher that those light skin toned players. In other words, these odds are not result of random chance). Based on my quick review of the team's reports, it seems also that a consensus seem to be forming in that the referee's implicit /explicit bias do not explain these differences. Of course this was not a controlled experiment and we had feedback among teams so we don't know if there was consensus before the feedback, but this exercise mimics the way scientific studies develop since there are not isolated researchers. This is good news.

It would be interesting to sort out the methodologies (seems like same approaches may have different names), what software was used, or special programming or canned subroutines were used and what assumptions support these (these are proxies for available resources for researchers, and I would hope that we can arrive to the same conclusions using specialized software and open source software). What are the researchers backgrounds (i.e., mathematicians, psychologist, etc.)? Notice that I'm not advocating for a democratic science (i.e., we vote for what the result is) but if different (and sensible) approaches find this phenomenon, then this phenomena is likely to exits in real life so corrective actions can be taken.

IMHO, trying to determine which method produces the "most" correct answer may be a futile effort. Of course, some models may be better than others, but we are dealing here is with mathematical constructs based on assumptions . Observed phenomenon does not necessarily follows these models and assumptions. These models are approximations to reality and they are good as long they can explain it. When they fail to do so, then a new explanation is needed.

For example, one can challenge the study's results based on the large amount of missing data that was removed. If I say that most of the missing data included dark skin toned players with lots of red cards, we can be more confident of this form of prejudice actually exits. If I say that most of these players were light skin toned with lots of red cards, the observed effect is more likely to be a random occurrence. If the missing data is just similar to those with data, the results hold. Is there an easy way to test this? One is to reconstruct the data filling for missing values. That's doesn't seem possible. Other is to impute but the number of auxiliary variables does not seem to be large enough to do a good job. Furthermore, the imputation can also been challenged the imputed also depends on models and (implicit/explicit) assumptions. A possibility is to do a sensitivity analysis assuming that all missing player are all darker skin tones with different number of red cards or assuming that they are all light skin toned. Would the conclusions hold? Is the analysis robust when there are departures or violations in the assumptions? This also takes time. As some teams footnoted (including myself) that the results were conditioned on the available data in a prominent footnote.

Other challenges to these results can come from the measurement of errors in the implicit/explicit bias (measurement errors can affect the values of beta coefficients) The skin tone assessment is subjective and also has measurement error (the two raters do not agree 100%). Is this linear? Bayesian methods can better reflect the measurement error (categorical misclassification for skin tone and continuous values for implicit /explicit bias). There are not canned subroutines on the frequentist side that account for these measurement errors that I was able to find for correlated data (my type of analysis). This would require additional programing that would take time to implement (and advanced the methodology). If we ignore the measurement errors, would the inference hold? Some teams determined that there is a high agreement in skin tone for the two raters. Is this high agreement enough to conclude that odds ratio will not include 1 (i.e., not significant) Is there a way to test this assumption? My CI was (1.10, 1.75). Can accounting for measurement errors push the CI to include 1? This is a weak spot on my analysis that I would be interested knowing how other non-Bayesian approaches deal with it or if they just assume that this would not make a difference (one team used Bayesian approach and reflected some of these).

Just a couple words about the outliers. But dropping them seems a way to force my model to the data. What about if they really exist. Should I consider a distribution with longer tails? Are these the result of mixed distributions? What about if they are indicator of something else (i.e., error in the data)? Suppose they are real and these people are receiving more red cards than everybody else. Isn't this an indicator that there is prejudice for these players? FYI, I reran my programs dropping these players and I still got the same result.

Just some thoughts Ismael Flores Cervantes

From: Garret S Christensen <gchrist1@swarthmore.edu>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 8, 2014 6:52:15 PM GMT+02:00

All,□I agree with all of Štěpán's thoughts. RE: 1 and 3, perhaps the organizers could alter the summary spreadsheet to include estimated effect sizes from the Qualtrics survey? That would save us the time of reading through all the pdfs to do it by hand.

RE: Point 2, I agree that interpretation of the question is key. If you interpret "How likely do you think it is that soccer referees tendto give more red cards to dark skinned players?" as literally "Do darker skin toned players get more redcards?" or "Are darker skin toned players more likely to get a redcard" then it seems like you shouldn't really control for any covariates--just run a t-test or super simple bivariate model and you're done. But I don't think any of us did that, because the more interesting question is "Do darker skin toned players get more redcards because they have darker skin?" But it seems obvious to me that we've got all sorts of potential omitted variable problems. My pet favorite is league of play: players with darker skin tone could be more likely to play in one league (say, in a country with looser immigration laws or better international recruiting) and that country happens to have stricter referees. (I don't know if that's at all realistic in European soccer, but comparing NHL hockey to Olympic hockey, or the NBA to FIBA, it seems plausible.) The organizers said that the league variable that many of us used in our initial analysis didn't actually track players over time, which I found disappointing.

My pet favorite omitted variable aside, there are numerous potential others. So I'm skeptical that any method can be used with the given dataset to get anything that's well identified causally. Maybe my training has made me focus too much on causal identification, but I have trouble caring about complicated statistical details when there seems to be such a fundamental flaw. If it's this obvious that you have OVB, just compare the average number of redcards for dark players and for light players and be done with it. Maybe that's too nihilistic or cynical, or maybe if we defined the research question more precisely, we could come to a better consensus on an appropriate statistical method. Maybe something like "Given the available data, and the inability to identify a causal effect, what is the best way to model the correlation between skin tone and receiving a redcard?" Maybe that's what some of us assumed to be the question to begin with?

Sincerely,Garret Christensen

On 11/08/14 09:00, Rickard Carlsson wrote:

> Dear all,

> Regardless of how we proceed next I think that this project has already

been a huge success. This is one of the most intriguing, and fun, projects I

have taken part in.

Regarding the alternatives on how we should present the data, I feel that

a summarizing graph with all the odds ratios and 95 % CI along with some

brief summary of the approach would be very illustrative. I think this is

very important because it will move the focus of whether we found all found

statistically significant findings to whether our findings are similar and

consistent in terms of a) effect size and b) precision.

I think this type of summary and discussion on it is important on its own right, but then we should be able to, based on some discussion, narrow it down to a few approaches.

As already pointed out, I think that the data structure is partly responsible for the different kinds of approaches used. In my initial approach I tried to disaggregate the data, but could not make it work. I also tried some analyses (multi-level modeling) that ended up crashing my computer. Initially, I settled for an analysis that I was not satisfied with (I noted this in the submission). In my final analysis I used a binomial logistic regression that (as others have also pointed out) does the job similar to disaggregating the data. It further did not crash my computer :).

I do feel that this aspect of data structure and challenges along with it is important to take into the discussion. Some of us are more likely used to analyzing our own data, where we have perfect understanding (and control) over the data structure, whereas others are more familiar with the task of analyzing a given set of data. I find it interesting to hear that others have also had some problems with computers, software etc. Personally, I do not have a version of STATA that is advanced enough to handle this type of large dataset and this has guided me in my choices. Had this been a project I had worked in, I would probably have upgraded my software or switched to something else. What I am trying to say is thus that if the data had been structured in a different manner and we all had software that is able to handle this type of dataset properly, then there may have been higher agreement among approaches.

Another aspect I find really important is how similar the approaches that several of us feel are "wrong" actually perform. Further, it is quite different if the approaches are invalid or if they are just unreliable. For example, analyses that throw away some data (as suggest that ANOVA would do) - what are the consequences for the analyses? Are they simply more blunt, or could the be biased in favor of a certain interpretation? Have some of use chosen the wrong specialized tool from the tool box, or have some of us simply used a too blunt all-purpose tool? This is quite a different scenario in my view.

Finally, I found Team 17 discussion on outliers intriguing. My own approach to check for the influence of outliers was to base the analysis on censored data where it was only possible to get one single red card, reasoning that players that receive *several* red cards are probably quite unusual players that may influence the findings too much. This did not change the conclusions I drew. Of course, having a single red card in a few games is more extreme than having a single red card in a great number of games, but I did not find these extreme enough to exclude. In either case, if removal of .3% of the data is enough to eliminate the effect, then the effect is indeed not very robust, which is not surprising since it is small.

Kind regards,

Rickard Carlsson

From: Tim Heaton <t.heaton@sheffield.ac.uk>> Reply: Tim Heaton <t.heaton@sheffield.ac.uk>>> Date: August 11, 2014 at 12:53:44

Hi again,

>

> In relation to outliers (team 17) before deciding making any comment about this can I ask a question. From memory this was a Bayesian

> analysis. Can you confirm how you checked convergence? Did you run

> multiple chains from over-dispersed starting values? What was the BGR

> statistic for these chains?

>

> Also I personally don't support an argument that you can just wrap up

> all the other factors into a single player random effect - I don't think

> that you can just ignore the different leagues (or the different

> positions) and say that these are accounted for in a single player

> random effect term. It's not if there is a different mean for each

> league/position that needs to be shared amongst all individuals in that

> unit (league/group) and (based on the AIC) the position/league does make

> unit (league/group) and (based on the AIC) the position/league does make > a difference. What did you get for the BIC when you included

> league/position?

>

> Would the logical extension of an argument that you can wrap up all

> the other effects into a single random effect term and forget about them > not be that you can do the same for the skin tone.

>

> Finally do we know why some individual had missing values? A previous > email suggested that most of the outliers considered by team 17 also had > missing data. It would seem that if most of the other groups removed

> these outliers (as they had missing data) and still found skin tone to

> be significant that actually the results are considerably more robust

> than team 17 suggests. It would however be useful to actually see how

> robust the other groups analyses were to single individuals who may not > satisfy the requirements of a normally distributed random effect.

>

> All of the above is from memory of reading the submissions last week

> so please forgive me if I have mis-remembered,

>

> Tim

From: Tim Heaton <t.heaton@sheffield.ac.uk>Reply: Tim Heaton <t.heaton@sheffield.ac.ukDate: August 11, 2014 at 12:53:44

Hi again,

In relation to outliers (team 17) before deciding making any comment about this can I ask a question. From memory this was a Bayesian analysis. Can you confirm how you checked convergence? Did you run multiple chains from over-dispersed starting values? What was the BGR statistic for these chains?

Also I personally don't support an argument that you can just wrap upall the other factors into a single player random effect - I don't thinkthat you can just ignore the different leagues (or the differentpositions) and say that these are accounted for in a single playerrandom effect term. It's not if there is a different mean for each league/position that needs to be shared amongst all individuals in thatunit (league/group) and (based on the AIC) the position/league does make a difference. What did you get for the BIC when you included league/position?

Would the logical extension of an argument that you can wrap up all the other effects into a single random effect term and forget about them not be that you can do the same for the skin tone.

Finally do we know why some individual had missing values? A previous email suggested that most of the outliers considered by team 17 also had missing data. It would seem that if most of the other groups removed these outliers (as they had missing data) and still found skin tone to

be significant that actually the results are considerably more robustthan team 17 suggests. It would however be useful to actually see how robust the other groups analyses were to single individuals who may not satisfy the requirements of a normally distributed random effect.

All of the above is from memory of reading the submissions last week so please forgive me if I have mis-remembered,

Tim

On 12 August 2014 11:25, Richard D. Morey <r.d.morey@rug.nl>:

Hi all,

There have been several questions about our Team's (17) analysis, so I thought I'd post and clear things up.

On the subject of outliers and missing data: Due to the conversion to integer player values possibly being different depending on how you clean the data, I thought I'd post the names of the outliers we eliminated to clear up any ambiguity. Here they are, with skin tone ratings:

	player	rater1	rater2
154	Leyti N'Diaye	5	5
841	Cyril Jeunechamp	1	1P:
2850	Maxime Poundje	4	4
3967	Larry Azouni	NA	NA
4427	Abel Khaled	NA	NA
6904	Sidy Koné	4	4
30848	Dylan Tombides	NA	NA

As you can see, three of the 7 actually don't have skin tone ratings. We did not eliminate players solely on the basis that they did not have skin tone ratings, because we wanted to use their data to help estimate the referee random intercept parameters. But as you can see here, most of the outliers we eliminated do, in fact, have skin tone ratings and therefore could be having an inordinate influence on the estimate of the relationship between skin tone and red cards. If you look at scatter of the average red card probability against skin tone (as can be seen in our analysis files on OSF), it is obvious who these players are. They have *way* too high a red card percentage to be accounted for even by adding a random effect of participant, and so our logistic regression model places some of the effect into the slope between skin tone and red card percentage. This is not surprising, really.

Regarding convergence: I do not have access to all my analysis files, so I cannot give you the convergence statistics, but I can tell you that it is my sense that cannot be accounted for by lack of convergence. I assessed the analysis in several ways:

a. Visual inspection of samples of the parameters (there's too many parameters to check them all)b. Agreement with classical LME (using lme4): results agreed, when the lme4 analysis converged (which wasn't always...)c. Multiple runs with different seeds/starting values: all led to verysimilar inferencesd. Single chain convergence statistics (e.g. Geweke's diagnostic)

These affirmed the well-behaved nature of these chains. Logistic regression models of this complexity (i.e., medium complexity, not too many covariates) are typically well-behaved. I would have been shocked had the chains *not* been of good quality.

Regarding how we dealt with missing skin tone data, there appears to be somewhat of a misunderstanding. We did not treat skin tone categorically, except for the purpose of imputing the missing data. We first computed a single skin-tine rating (averaging the z scores of the two raters). The missing scores were assumed to be drawn from the same distribution as the existing scores, so in a sense they were treated as discrete. In the regression model, however, the skin-tone entered in as a continuous covariate, as one would expect.

Best, Richard

From: Dan Martin [mailto:dpmartin42@gmail.com] **Sent:** Monday, August 11, 2014 1:36 PM

Hey everyone, Great discussion so far. A few of you mentioned wanting to see a figure summarizing the results. Here it is!

I'm still working with two groups to get a finalized estimate with confidence intervals, but this should give you all a good idea of the overall results for the first research question. For simplicity, all effect sizes and CIs were converted to odds ratios units (from either incidental risk ratios, standardized betas, or Cohen's d estimates). It also includes the corresponding team number along with the analysis description that the team submitted on qualtrics.

Best, Dan

From: Tim Heaton <t.heaton@sheffield.ac.uk>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 11, 2014 1:52:40 PM GMT+02:00

Hi again,

In relation to outliers (team 17) before deciding making any comment about this can I ask a question. From memory this was a Bayesian analysis. Can you confirm how you checked convergence? Did you run multiple chains from over-dispersed starting values? What was the BGR statistic for these chains?

Also I personally don't support an argument that you can just wrap up all the other factors into a single player random effect - I don't think that you can just ignore the different leagues (or the different positions) and say that these are accounted for in a single player random effect term. It's not if there is a different mean for each league/position that needs to be shared amongst all individuals in that unit (league/group) and (based on the AIC) the position/league does make a difference. What did you get for the BIC when you included league/position?

Would the logical extension of an argument that you can wrap up all the other effects into a single random effect term and forget about them not be that you can do the same for the skin tone.

Finally do we know why some individual had missing values? A previous email suggested that most of the outliers considered by team 17 also had missing data. It would seem that if most of the other groups removed these outliers (as they had missing data) and still found skin tone to be significant that actually the results are considerably more robust than team 17 suggests. It would however be useful to actually see how robust the other groups analyses were to single individuals who may not satisfy the requirements of a normally distributed random effect.

All of the above is from memory of reading the submissions last week so please forgive me if I have mis-remembered, Tim

---

From: Ismael Flores Cervantes <ismaelflorescervantes@westat.com>**Subject:** **RE: Crowdstorming Project: It's time to discuss the results!****Date:** August 11, 2014 7:43:26 PM GMT+02:00

Hi, A small correction on the plot, for team 32, the OR is 1.39 (1.10, 1.75) . The plot incorrectly shows 1.178 with 1 inside the CI. Thanks Ismael

---

From: "Richard D. Morey" <r.d.morey@rug.nl>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 12, 2014 4:25:08 PM GMT+02:00

Hi all,There have been several questions about our Team's (17) analysis, so I thought I'd post and clear things up.

On the subject of outliers and missing data: Due to the conversion to integer player values possibly being different depending on how you clean the data, I thought I'd post the names of the outliers we eliminated to clear up any ambiguity. Here they are, with skin tone ratings:

	player	rater1	rater2
154	Leyti N'Diaye	5	5
841	Cyril Jeunechamp	1	1P:
2850	Maxime Poundje	4	4
3967	Larry Azouni	NA	NA
4427	Abel Khaled	NA	NA
6904	Sidy Koné	4	4
30848	Dylan Tombides	NA	NA

As you can see, three of the 7 actually don't have skin tone ratings. We did not eliminate players solely on the basis that they did not have skin tone ratings, because we wanted to use their data to help estimate the referee random intercept parameters. But as you can see here, most of the outliers we eliminated do, in fact, have skin tone ratings and therefore could be having an inordinate influence on the estimate of the relationship between skin tone and red cards. If you look at scatter of the average red card probability against skin tone (as can be seen in our analysis files on OSF), it is obvious who these players are. They have *way* too high a red card percentage to be accounted for even by adding a random effect of participant, and so our logistic regression model places some of the effect into the slope between skin tone and red card percentage. This is not surprising, really.

Regarding convergence: I do not have access to all my analysis files, so I cannot give you the convergence statistics, but I can tell you that it is my sense that cannot be accounted for by lack of convergence. I assessed the analysis in several ways:

a. Visual inspection of samples of the parameters (there's too many parameters to check them all)b. Agreement with classical LME (using lme4): results agreed, when the lme4 analysis converged (which wasn't always...) c. Multiple runs with different seeds/starting values: all led to very similar inferencesd. Single chain convergence statistics (e.g. Geweke's diagnostic)

These affirmed the well-behaved nature of these chains. Logistic regression models of this complexity (i.e., medium complexity, not too many covariates) are typically well-behaved. I would have been shocked had the chains *not* been of good quality.

Regarding how we dealt with missing skin tone data, there appears to be somewhat of a misunderstanding. We did not treat skin tone categorically, except for the purpose of imputing the missing data. We first computed a single skin-tine rating (averaging the z scores of the two raters). The missing scores were assumed to be drawn from the same distribution as the existing scores, so in a sense they were treated as discrete. In the regression model, however, the skin-tone entered in as a continuous covariate, as one would expect.

Best, Richard

On Tue, Aug 12, 2014 at 9:54 PM, Erikson Kaszubowski <erikson84@yahoo.com.br> wrote:

Dear all,

Thanks to Richard Morey for giving more details on his team's approach. The use of a categorical distribution for missing data imputation on skin tone rating sounds like a good idea!

But I still can't believe that those outliers have such an impact on theskin tone coefficient. The first version of our model didn't assume linear effect for skin tone rating. Treating the variable as categorical, we were surprised to notice an almost linear effect, with lighter toned players below the mean probability and darker toned players above the mean probability. Those estimates were based on the data with only one outlier, Jeunechamp, who has a skin tone rating of 1 (0 on the new scale).

I ran a simplified version of Team 17's model (without the interactions and country bias variables) in Stan, excluding the seven outliers, and the coefficient for the skin tone rating was statistically significant and closeto what the other teams found (1.36 [1.12, 1.72]) (convergence was OK). I think that the interaction terms might be introducing some kind of collinearity problem in the model; but that doesn't explain why the first version worked fine.

The results summary looks very interesting! Even though there are some differences, fourteen teams found statiscally significant result with very close estimates and similar CI, even when the model distribution was different (Poisson, Negative-Binomial and Binomial). This shows how some assumptions made huge differences and others were not so meaningful. I thought that the likelihood distribution would influence the estimations a lot more, but now it seems that the models that somehow weighted the outcome using the number of games reached similar results.

Best, Erikson

On Aug 12, 2014, at 10:12 PM, Eric-Jan Wagenmakers <ej.wagenmakers@gmail.com> wrote:

I still have a hard time believing the "significant" result, not just because our model does

not produce it, but also because it simply does not appear to be there if you look at the data (outliers taken out).

Consider the descriptives shown in Figure 6 of the attached file and tell me there is a positive relationship between skin tone rating and red card proportion. I can't see it. In light of these descriptives,it seems to me that any model that finds for a positive relationship has some explaining to do.

@Erikson: Thanks for running the model in Stan, that's cool.

Cheers, E.J.

From: Erikson Kaszubowski <erikson84@yahoo.com.br>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 13, 2014 8:46:31 AM GMT+02:00

Dear all,

I will try show that the exploratory analysis of the data does support the effect, at least to some extent. Thanks to Dr. Eric-Jan Wagenmakers for proposing the challenge!

First, I think that the plot might be misleading, given the great number of points. But the main problem is, in my opinion, what is really plotted in it. According to the Team 17's code, the card.prob variable is the mean proportion of red cards per game for each player. But the observed proportion of matches in which a player received a red card shouldn't average over all dyads, but simply divide the sum of red cards by the sum of games.

For example: player John Utaka received 2 red cards in 431 games. The observed proportion of red cards per game for him is simply 2/431=0.00464. Team 17's code computes the mean of the proportion for every referee that is paired with the player; in this case, the result is 0.00733. The comment in the code seems to assume it is computing the first value (card.prob=card.sum/card.N, that is, for each player, sum(redCards)/sum(games)), but it computes something else (for each player, mean(redCards/games)).

There is a problem with both calculations: for every player that did not receive any red card the result will be 0, regardless of the number of games. I still didn't figure out a way around it, but here are some suggestions that support some models point estimate for skin tone rating.

1) We can calculate the proportions of red cards per player-game for each level in the skin tone scale (I am working with a dataset that excludes the cases with missing skin tone rating).

> tapply(data$redCards, data$rater1, sum)/tapply(data$games, data$rater1, sum) 0 0.25 0.5 0.75 1

0.003813149 0.004422772 0.004543566 0.004671890 0.005191931

> tapply(data$redCards, data$rater2, sum)/tapply(data$games, data$rater2, sum) 0 0.25 0.5 0.75 1

0.003746957 0.004288672 0.004945711 0.004436464 0.004943899

The results do not differ much when we exclude the outliers.

> tapply(dataOut$redCards, dataOut$rater1, sum)/tapply(dataOut$games, dataOut$rater1, sum) 0 0.25 0.5 0.75 1

0.003735998 0.004422772 0.004543566 0.004534005 0.005077875

> tapply(dataOut$redCards, dataOut$rater2, sum)/tapply(dataOut$games, dataOut$rater2, sum) 0 0.25 0.5 0.75 1

0.003637475 0.004288672 0.004945711 0.004301075 0.004849422

It's interesting that the odds ratio between levels 0 and 1 ranges from 1.32 to 1.36.

2) We can also compute the proportion of red cards per game for each player (sum(redCards)/sum(games)) and then compute the mean for each skin tone level. The attached image is a graphical representation of the results (outliers present but not shown).

The dotted line represents the overall proportion mean (0.00455) and each diamond represents the group mean:

> tapply(meanCards, rater1.player, mean) 0 0.25 0.5 0.75 1

0.003917657 0.004610303 0.004293529 0.006271244 0.006238067

> tapply(meanCards, rater2.player, mean) 0 0.25 0.5 0.75 1

0.004142242 0.004246992 0.004676824 0.005911500 0.005978804

Here the odds ratio are higher: 1.6 and 1.46. Even tough it might be a bit of a stretch to say that the effect is linear, players with darker skin tone do have higher mean red car ds per game.

Given this, I could accept a wider interval with a lower point estimate when excluding the outliers, but I think it is not possible that it would make the effect disappear.

Cheers, Erikson

P.S.: (For Dr. Eric-Jan Wagenmakers) Your model was interesting to implement in Stan! If you want, I can send you the source file so you can extend it and run the full model.

From: Rickard Carlsson <rickard.carlsson@lnu.se>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 13, 2014 9:54:24 AM GMT+02:00

Dear all,It seems to me that Erikson and I have worked on similar things regarding these outliers.

My approach was a binomial logistic regression with clustered standard errors (on the level of player and referee). Thus, the primary unit of analysis is the observations and it was on this level I screened for outliers. I thus simply calculated redCards / games and plotted this to check for outliers. (STATA: generate red_prob = redCards / games). I found no outliers in my analysis using this approach.

However, Team17's approach made it very clear to me that is important to screen for outliers on the aggregated level of players as well. Otherwise it would be possible for outlier players to be masked by several observations of probabilities that are plausible by themselves, but not when combined. Since skin tone is measured on this level, it may bias the result. I believe this is a *very* important point raised by Team 17. So I next collapsed [(collapse red_prob, by(player)] the probabilities across players and then plotted the same histogram that Team 17 used to detect outliers: histogram red_prob, frequency. I have attached it.

Indeed, now I am also able to see the very same seven outliers. I am not entirely convinced about the cut-off at .05. It seems to me that there is simply a long tail and the values that do stand out are the four players above .08. However, this is a small detail so I decided on the same cut-off and remove these seven outliers and then re-ran my analysis.

The results were virtually identical. Hence, it seems to me that these seven outliers do not explain the difference between my findings and Team 17's quite different approach.

Kind regards,

Rickard Carlsson

From: <bahniks@seznam.cz>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 13, 2014 7:05:48 PM GMT+02:00**To:**

Hi,

I also tried to do the analysis (multilevel logistic regression) without the outliers and the results did not change (in fact only one of the seven had been included in my model before). I also looked at estimated random effects for players and found that there are three outliers, but removing them did not change the results too.

Best, Štěpán

---

13 de Agosto de 2014 14:08, "bahniks@seznam.cz" <bahniks@seznam.cz> escreveu:

Hi,

I also tried to do the analysis (multilevel logistic regression) without the outliers and the results did not change (in fact only one of the seven had been included in my model before). I also looked at estimated random effects for players and found that there are

three outliers, but removing them did not change the results too.

Best, Štěpán

From: Erikson Kaszubowski <erikson84@yahoo.com.br>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 15, 2014 10:18:00 PM GMT+02:00

Dear all,

@Richard,

I agree with you that the mean isn't a good representative value of the distribution of red cards. On the other hand, the median by itself is also problematic because of the great number of zeros. I also agree that a more robust effect would be more evident and it would show in the median values.

But I still believe that this exploratory analysis show some evidence of skin tone effect on the number of red cards, even though it's hard to pinpoint it to one summary statistic.The attached graph shows the density of the proportion of red cards per game at the player level, in three categories: 1 (skin tone rating <=0.25), 2 (>0.25 & <0.75) and 3 (>=0.75). The density at 0 decreases as the skin tone rating increases. The density for category 3 is higher than the others at higher values. The density of the first category is lower for higher values.

The mean (sd) for each category is (excluding the values above 0.05): 0.0042 (0.006); 0.0045 (0.006); 0.0050 (0.0067). Obviously, the number of zeros certainly helps lower the mean for category 1, but I also think it tells us something about how the proportion of red cards per game is affected by player skin tone.

@Magnus,

Assuming a consistent skin tone effect between red cards and yellow cards does make sense, but I don't think it's necessary. The fact that the skin tone coefficient has a negative sign when predicting yellow cards per game could be explained by referee bias, too: we could say that referees are more strict with dark toned players and more condescending to lighter tone players.

Our data shows that the rate of red cards per yellow card also increases with skin tone rating! This can be modeled using a Poisson likelihood and the yellow cards as the exposure variable, aggregating the data at player level (I should have thought about this earlier!):

red/yellow ~ Pois(exp(a_position + B * raterMean))

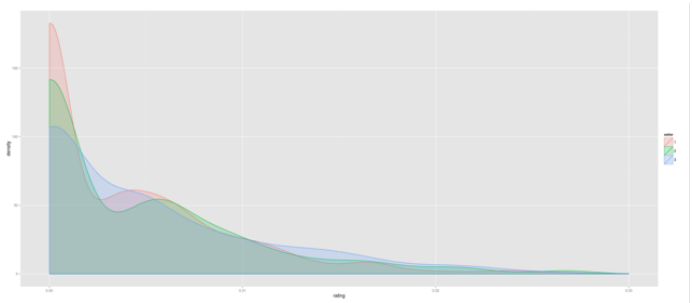
The estimate for the skin tone effect is similar to what the other models have found: 0.34 (s.e. 0.09). It's not surprising, giving that we expect higher count of red cards for dark toned players and we also know that the number of yellow cards do not differ much between each level of the skin tone rating in our data. If we include the rater variable as a factor, the coefficient increases from one level to the next.

@Ismael,

I followed your suggestion and fitted a model with random effects for player and referee with the missing skin tone rating as 0 or 1. The results are interesting. If we fit the model with all missing skin tone data as 0, the coefficient for skin tone decreases to 0.26 (s.e. 0.1). If we fit the same model with all missing skin tone data as 1, the coefficient decreases to 0.21 (s.e. 0.08). The coefficient decreases in both cases, but it's still positive, not very far from the original estimation, and more than two s.e. from 0.

I really think the best way to deal with this is to use a measurement error model and then do some kind of imputation like Team 17 did so we can correctly deal with the uncertainty in the data. I will try this later, as this will take some time.

All the best, Erikson



From: Felix Schönbrodt <felix@nicebread.de>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 20, 2014 9:27:11 AM GMT+02:00

Hi all,I just returned from my holidays, and worked through your very inspiring discussion! **First, three general remarks:**

A)

I agree with previous commenters that averaging the results (approach 1) is not appropriate. Even if all scholars agree that the world is flat, this consensus can be far away from the truth.

So what could be another “gold standard”? If this had been set up as a predictive challenge, we could simply test the models against a held back test data set. But this was no predictive challenge (in which case we would probably have seen a lot of neural networks and SVMs), and explanatory and predictive models are not exchangeable (Shmueli, 2010).

Hence, from my point of view, only *good arguments* can give a hint which approach is "closer to the truth” (... always keeping in mind that essentially all models are wrong). A single good argument can outweigh a previous consensus of many.At the end, of course, reasonable people can still disagree; but we could and should make these different

points of view explicit in the paper.

B)

Regarding the RQs - Assumed we had a sponsor that wants to base a policy on our scientific conclusions. What should we recommend?Personally, I am not so interested in the specific research question (in particular in the light of the possible confounders and omitted variables many of you pointed out). But the process *how* we come to a joint recommendation could be a blueprint (or at least a thought-provoking impulse) for similar situations in the future. I liked Ismaels link to the Panel on Climate Change. They have a protocol for reaching decisions:

"In taking decisions, and approving, adopting and accepting reports, the Panel [...] shall use all best endeavours to reach consensus. If consensus is judged by the relevant body not possible: (a) for decisions on procedural issues, these shall be decided according to the General Regulations of the WMO; (b) for approval, adoption and acceptance of reports, differing views shall be explained and, upon request, recorded. Differing views on matters of a scientific, technical or socio-economic nature shall, as appropriate in the context, be represented in the scientific, technical or socio-economic document concerned." (http://www.ipcc.ch/pdf/ipcc-principles/ipcc-principles.pdf)

C)

One insight for me is quite similar to Eric-Jan's: The variability in modeling approaches.We should keep in mind that we are discussing on a very high level. When I think of the typical psychology departments I know, only very few people outside the methods sections would be able to set up a crossed random effects GLMM. Nobody ever used zero-inflated hurdle models (nor heard of it), and when it comes to hand-carved Bayesian models, well ... these are completely uncharted waters.

To put it bluntly: If a group of highly skilled statistic professors etc. gets to such diverging conclusions about a complex research question - what happens when "normal" researchers (i.e., non-statisticians) fiddle around with latent longitudinal growth models, dyadic diary data, fMRI measures, and so on?

Is one possible consequence to have (external?) professional statisticians available for complex RQs, as is mandatory for some clinical trials ("**data monitoring committee**", http://en.wikipedia.org/wiki/Clinical_trial)?Or better go for **open data**, which allows a even larger group of interested researchers to run their own models on the data (post-publication-style) and to reproduce the conclusions? This, however, would only work well when we also have a central hub where alternative analytical approaches for a data set can be published and discussed. (I am thinking of OSF.io, or https://curatescience.org/).

Some specific comments:

Erikson wrote:

The binomial regression estimates should be exactly the same of a logistic regression with the disaggregated data, but without the burden of actually having to reshape the

dataset.

It is. We did both approaches.

I thought that a mixed effect models with skin tone coefficient varying by player and referee could do it, but I couldn't run the model because it is too big for my computer.

I think “varying by player” does not work, as skin tone does not vary within player. This would an interesting approach if we had “aggressiveness" of each player in each game. But IMO, the skintone effect can only vary within referees.

Eric-Jan wrote:

This diversity in outcomes is something to emphasize, not to hide.

Absolutely agree!

Rickard wrote:

In either case, if removal of .3% of the data is enough to eliminate the effect, then the effect is indeed not very robust, which is not surprising since it is small.

I had a similar impression: if the effect really is driven by a few outliers (which, however, does not seem to be true for all models), we should be very skeptical.

Finally, I agree with Tim that an **online forum**, such as google groups, could facilitate discussions. At the moment we discuss different topics in a single thread; for me it would work better to have separate topics in separate threads.

Now I am going to run our model without the outliers ...

Cheers, Felix

References:

Shmueli, G. (2010). To explain or to predict? Statistical Science, 25(3), 289–310. doi:10.1214/10-STS330

From: Felix Schönbrodt <felix@nicebread.de>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 20, 2014 11:47:38 AM GMT+02:00

***Science* Establishes Dedicated Statistical Review Panel**
http://magazine.amstat.org/blog/2014/08/01/science-review-panel/

"*Science* magazine editor-in-chief, Marcia McNutt, recently announced the July 1 appointment of a new statistical board of reviewing editors (SBoRE)—composed of prominent members of the statistical community— that will help address reproducibility issues and increase confidence in the papers published in the magazine.“So why is *Science* taking this additional step? Readers must have confidence in the conclusions published in our journal, and that we have taken reasonable measures to verify the accuracy of those results,”

said McNutt in a recent editorial. “We believe that establishing the SBoRE will help avoid honest mistakes and raise the standards for data analysis, particularly when sophisticated approaches are needed.”

The statistical experts appointed to five-year terms on the newly created SBoRE are:

Ron Brookmeyer, professor of biostatistics at the University of California at Los Angeles’s School of Public Health**Alison Motsinger-Reif**, associate professor of statistics at North Carolina State University**Giovanni Parmigiani**, professor of biostatistics at the Harvard School of Public Health and chair of the Dana-Farber Cancer Institute’s Department of Biostatistics and Computational Biology **Richard L. Smith**, professor of statistics and biostatistics at The University of North Carolina and director of the Statistical and Applied Mathematical Sciences Institute**Jane-Ling Wang**, professor of statistics at the University of California at Davis**Chris Wikle**, professor of statistics at the University of Missouri-Columbia**Ian Wilson**, professor of structural biology and chair of the department of integrative structural and computational biology at The Scripps Research Institute"

Maybe we could just send our approaches to them, and they determine which one is right and which one is wrong ;-) F

From: Brian Nosek <nosek@virginia.edu>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 20, 2014 7:53:39 PM GMT+02:00

Hi all --

This has been an excellent discussion and the developing consensus about how to best present this research process is very consistent with my own view. We have most of a first draft completed. When that is circulated, that should help to organize this discussion.

My sense is that there we will eventually arrive at a highly engaging summary report (that reflects on the project at a relatively high level), and a rich set of supplementary materials that delves into the process and many of the issues that have been raised already.

Once the draft has started to circulate (Google Docs for collaborative editing), then we can decide if discussion is effectively contained within that and the email list, or if an online forum would be a helpful complement.

Regards, Brian

From: Johannes Ullrich <j.ullrich@psychologie.uzh.ch>**Subject:** **Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 22, 2014 4:50:49 PM GMT+02:00

Dear E.J. (and all)

I wish the role of the outliers could be made clearer. These 7 players seem to make a big difference in your analysis. Others have reported that their conclusions don't change when these players are excluded. Likewise, we have done our analyses again without the cases that you identified, but obtained virtually identical results (i.e., small, but non-zero bias). It's puzzling to me why such a tiny fraction of the sample should completely change your conclusions.

In terms of simple descriptive analyses, I agree that the scatterplot in your report does not reveal the effect. The attached figure simply compares the observed vs. expected number of red cards across levels of skin- tone. Even without the 7 critical players, the effect is easy to see. All skin-tone categories equal to or greater than 2 have more red cards associated with them than would be expected if the frequency of red cards was proportional to the number of players with a given skin-tone.

So I'm wondering at what point in the process of model building and statistical inference does the effect go away? This would be a great learning experience for me. CheersJohannes

---

From: "Dam, L." <l.dam@rug.nl>
Date: August 29, 2014 9:20:07 AM GMT+02:00

Hi All,I do not have much more to add, but I would like to say that by now I am convinced myself that some sort of logit/logistic analysis would be appropriate (which our team did not do btw). Regards,Lammertjan

---

From: Seth Spain [smspain@gmail.com] **Sent:** Friday, August 29, 2014 6:29 AM **To:**

Subject: Re: Crowdstorming Project: It's time to discuss the results!

Hi all,

To follow up on Lammertjan's point. Our original analysis was a type of Poisson (we started with the intention of controlling for a variety of game-sum factors, like goals, but nots games themselves). The mid-round feedback convinced us that a binomial logistic model made the most sense, modeling red cards ~ binomial(theta_i, games) where theta_i is the linear predictor. Regardless, the point I wish to emphasize is that the internal review and feedback process *strongly* affected our approach.

In the end, we fitted the model using glmer. We tried using both JAGS and Stan to fully Bayesian analyses, which worked fine for relatively small subsamples, but we couldn't get either to run in reasonable time for the full data--I'm curious if the groups that did Bayesian analysis used more computationally efficient programming (esp. in Stan, which I had expect to run quite fast for this kind of a problem), or if they are just more patient than we were.

Best, Seth

From: Russ Clay <Russ.Clay@csi.cuny.edu>**Subject:** **RE: [Spam:******] Re: Crowdstorming Project: It's time to discuss the results!****Date:** August 29, 2014 3:11:07 PM GMT+02:00

Hi all,

That is interesting. My original approach was a binomial logistic model, and the interim feedback persuaded me to switch to a Poisson regression analysis with games as an offset. I would say that I was strongly influenced by the feedback as well, but it pushed me in the opposite direction.

Russ