1/17/2023
STAT 27850
Real Data Analysis Critique
Soren Dunn, Aim Wonghirundacha, Tytus Wilam, Simone Zhang

Summary

The study's main question is whether "the effect of peers on adolescents' decision making is mitigated by the presence of a slightly older adult". They begin by citing previous literature that shows adolescents are more risk-taking in the presence of their peers. Therefore, the authors seek to learn whether having an adult present within the peer group would mitigate the increased risk-taking. In other words, would adolescents have a higher appetite for risk in the presence of peers than on their own or in the presence of an older individual? The researchers conjectured that increased propensity to take risks might result from adolescents' changing preference for immediate rewards: when the adolescents prefer immediate rewards, as measured by higher discount rates, they would take more risks. The authors therefore asked two questions, one about the effect of adult presence on risk taking, and one about the relationship between risk taking and a preference for immediate rewards. To answer these questions, the researchers designed two tests: the first was a game in which targets could engage in risky behavior and the second was a task designed to reveal their discount rate. The game asked the targets to drive a vehicle through a busy intersection where risk taking is encouraged by offering monetary incentives for those who complete the game in a shorter time. Based on the game, the authors constructed a risk index, defined as the proportion of intersections at which the brakes were not applied. From the delay discounting task they calculated the indifference points and discount rates which are two highly correlated measures of target's preference toward immediate rewards. They conducted each task in 3 different social contexts: target takes the tests on his or her own, target takes the tests in the presence of peers, and target takes the test in the presence of two peers and one adult. In the analysis of the data, they run linear regressions to determine whether different social contexts affect the risk index, indifference point, and discount rate. They report the difference in measured risk index between targets in each of the settings.

Summary of reported results

For the risk index, they reported that the solo condition has a risk index of 0.10 lower than the risk index for the peer-group condition with a p-value of less than 0.001, thereby confirming the fact that being with peers increases risk-taking behavior. Then, they reported the difference between the solo-group and the adult-present group as well as the difference between the peer group and the adult-present group. By calculating the pairwise difference between groups, they argued that the adult-present condition mitigates the effect of the risk-taking, answering their first question. To answer their second question, they attempted to argue that the indifference point and discount rate show a similar pattern, which is that the peer group is significantly different from the solo and adult-present condition. They followed the same format of reporting pairwise differences between groups and the corresponding p-value for each comparison. However, they were unable to conclude that the peer group and adult-present group had significantly different indifference points and discount rates. Therefore, they were unable to

decisively conclude that the preference for immediate rewards leads to risk-taking. At the end of the paper, they concluded that those in the peer group took significantly more risks than solo and the adult-present group. They also concluded that those in the peer group are more drawn to immediate reward but could not argue that the adult-present condition had the same mitigating effect as it did in the risk-taking task.

Multiple Testing Issues

We think that the authors report the wrong significance levels because of a multiple comparisons problem. When reporting results of the three conditions on the measured risk propensities of adolescents, the authors emphasize the fact that they have significant findings in all three comparisons they test: risk in the peer condition is larger than in both the solo and adult-present conditions and there isn't a significant difference between the solo and adult-present conditions. It makes sense that the authors care about all three results because their theory predicts all three. If even one of the findings was not there, that would put in doubt the validity of the conclusions (for the same reason the results for the indifference points and discount rates are weaker because only some of the expected findings are found).

The authors, however, do not look at the family-wise possibility of a false positive in the three tests reported when those are taken as a group. They should report their confidence in the result that all three comparisons are simultaneously significant. They report three different comparisons and a p-value for each of them. What they use later on, however, is the fact that all three p-values show what their theory predicted (two rejections of the null and one failure to reject). Their findings would be much weaker if any of those hypotheses were not rejected:

1) H_a: ß_solo is significantly different from ß_adult
2) H_b: ß_solo is **not** significantly smaller than ß_peer
3) H_c: ß_adult is **not** significantly smaller than ß_peer

Therefore, they should either compute the family-wise error rate for this family of hypotheses or use a correction. If using the Bonferroni correction, which is acceptable in this scenario because it allows for dependence, it would mean adjusting the alpha = 0.5 to alpha/n = 0.5/3 = 0.0166 . In the case of their tests for risk, the p-values would withstand this more stringent condition. But in the case of the indifference point, their p-value of 0.012 (p. 327) is very close to the Bonferroni adjusted critical value.

We constructed simulations to check the importance of some of these multiple testing and degrees of freedom issues. We include the code we used in the Appendix. First, we used a simulation to estimate the probability of getting any or all p values for the risk analysis below a 0.05 significance threshold given that all individuals' risk indexes were independently drawn from a normal distribution with mean 0.25 and variance distributed normally between 0.0001 and 0.1. The parameters for these distributions were chosen somewhat arbitrarily as being plausibly similar to the actual data collected. Over one-thousand simulations there were no instances of all three differences reaching the significance threshold, but there were 125 simulations where at least one of the differences reached significance. Although the results observed from the risk-index analysis were unlikely to have occurred if the experimental conditions had no effect on the individuals' risk index, since the paper did not account for

multiple comparisons the probability of the authors getting any significant results was around 12% type I error (significantly higher than the commonly applied 5% significance threshold).

This issue is compounded if one considers that there is no guarantee that the authors always intended the article to focus primarily on the risk analysis. If either the discount rate or indifference point analysis had gotten more significant results, it is quite possible the authors may have focused on those instead of the risk analysis. This concern suggests that the probability of the experimenters getting significant results even if their experimental conditions had no effect might be even higher than suggested above. To test the extent of this issue, a second simulation was run another thousand times to estimate the overall probability of getting significant results over all three analyses performed in the study.

Risk indices were generated as in the first simulation, log discount rates were generated as normally distributed with mean negative six and standard deviation one and the actual monetary rewards chosen by the subjects as well as their delay interval were drawn from their own normal distributions. Additionally, indifference points were calculated for each subject based on a combination of their log discount rate, chosen monetary rewards, and chosen delay interval as described in the paper. Although almost no simulations got five or more significantly different group means as in the study (three from the risk analysis and one each from the indifference point and discount rate analyses), a full 284 of them got significant results for at least one of the comparisons checked. If the authors did not choose their primary analysis beforehand, then given their procedure the simulation estimated that they would get significant results almost a third of the time even if there was no difference between the different conditions.

Overall, the strength of the results presented in the paper were unlikely to occur if the experimental conditions had no effect on the subjects. However, the authors' failure to adjust for multiple comparisons and the number of degrees of freedom present in the study greatly inflated their chances of getting any significant results even if the experimental conditions had no effect on the subjects (without accounting for any other potential issues in the study).

Other Issues

In terms of statistical issues not related to multiple comparisons, there might be omitted variable problems in the study. For example, they chose not to include any demographic covariates, claiming that "the three conditions (solo, peer-group, adult-present) did not differ on any demographic variables for the target subjects." However, we didn't know how they reached this conclusion and whether any tests were performed. From the table that's provided in the paper we can see that there's a relatively large difference between the percentage of white among targets in the peer-group condition (54%) and the adult-person condition (69%) which comes from the fact that the study was not fully randomized but the volunteers were recruited into the control and test conditions at different times. It's not impossible that the risk taking behavior is related to race which may be due to culture, neighborhood lifestyle, or a number of other reasons. Since we see a difference in these demographic variables, it would be reasonable for the researchers to control for them in their regressions. On the other hand, if the researchers did include all the demographic variables and then decided not to include them because they appeared to be not

significant or even if they made a decision not to include them after seeing the data, this would be an **implicit multiple testing** problem. If dropping the demographic variable was a data-dependent choice it would cause a degree of freedom problem.

Another noticeable problem in the research is that the researchers never mentioned that they checked the linear model assumptions held before they discussed the result generated by the fitted linear models. We know that for a linear model, the constant variance and normality assumption should be satisfied, otherwise the confidence interval computed for the ß is not accurate. Moreover, since they didn't provide the actual regression model they implemented, we interpreted a linear regression model based on the information in the paper and noticed that there might be perfect collinearity in one of the models they use. According to the paper, the researchers claim that "Terms for the interactions between each of the 11 confederate dummy variables and the adult-present dummy variable were also included in the model" and the 11 confederate dummy variables were coded 0 for the solo and peer group. Therefore, for the Risk score comparison model for example, the regression should be:

$$Y_i = \beta_0 + \beta_1 \mathbb{1}_{adult} + \beta_2 \mathbb{1}_{solo} + \sum_{i=1}^{11} \beta_i \mathbb{1}_{confederate\ i} + \sum_{j=1}^{11} \beta_j \mathbb{1}_{confederate\ j,\ adult}$$

However, the last two terms are perfectly collinear because they are the exact same thing. The indicator $\mathbb{1}_{confederate\ i}$ and $\mathbb{1}_{confederate\ j,\ adult}$ have equal values all the time because $\mathbb{1}_{confederate\ i}$ can only equal 1 when it's under the adult-present condition. If our interpreted model is correct, this perfect collinearity violates the linear model assumption seriously and we shouldn't trust the β value they reported at all. This is because the linear model wouldn't be able to determine the power of the last two terms in terms of explaining the response variable and would distribute the power randomly. In that case, the p-value and confidence intervals wouldn't have much statistical value because they're not reliable.

We also notice that there are removed observations in this study that appeared to be suspicious. As claimed by the author, "ten target subjects were excluded from the analyses because their data were incomplete." Among the removed ten targets, five are in the solo condition and five are in the adult-present condition. We find it suspicious because, first the research was not in the form of a questionnaire but in person, and second it's such a coincidence that the number of incomplete data points happen to be the same for the two condition groups. It is possible that for example, to get more significant result, the researchers removed the five targets who took most risk in the solo and adult-present group; or that the researched found there was an X-Y plot where the trend is nonlinear or there was serious heteroskedasticity, but after removing the ten data points, the plot seemed more linear or the heteroskedasticity was mitigated. Removing these data points would be a poor choice because it leads us to use a linear model that actually is a poor fit to a non-negligible proportion of the population that the data is drawn from.

There is also a problem with the representativeness of the population and the ability to generalize. As the authors note, their sample included only male college-educated subjects which may not generalize to the population of people in the military. While they acknowledge that the results may not generalize to females, they attempt to justify their use of college-educated subjects. They argue that individuals who make important decisions in teams are likely to be the most highly educated within the team. However, they do not provide data or cite other research

to support this explanation of the limitation. Additionally, if the most likely decision-maker in the team is the most highly educated, then their experimental design does not account for this team structure very well. This is because the most highly educated is the graduate student or "adult" in the team, but they are requiring the younger and less educated subject to make the decisions. Given knowledge of the likely team structure, a more realistic experiment would be to have the adult confederate be someone who is not a graduate student and rather someone who is older but less educated than the subject.

The authors note that they did not vary the extent to which participants know each other. Since the study was not conducted with subjects who all know each other well, the authors acknowledge that they do not know whether their results would be suitable in situations where the team are already well acquainted. Therefore, this is an issue of generalizability since teams may be already well acquainted in the military. However, they also claim that there is no difference between the behavior of targets when the peers are acquaintances and when they are strangers. They claim to have verified this but we are not told whether this was explicitly tested. If they compared means or ran a statistical test, then there may be a multiple testing issue.

The evidence they do provide is that in both the peer-only and adult-present groups, there is a similar percentage of people who had at least one friend in their group. While there is a similar percentage of people who had at least one friend, we don't know whether the similar percentage of people includes the confederates. If it does not include the confederates, then the adult confederate is always a stranger to the rest of the group. Therefore, if they only checked for prior relationships between the subjects and not the confederates, then the adult-present group always contains more strangers than the peer-only group. If this is the case, the number of strangers could have been a potential confounder and therefore should be included in the model. This may also be a plausible explanation of the result since a group with more strangers might encourage you to act in a less risky way.

Construct validity describes how well an indicator represents a concept that is not directly measurable. We think that the constructs presented in the paper are not valid on at least two counts. The first concept the authors claim to be measuring is the effect of including an adult in a peer group. Do they succeed in creating an adequate construct for this purpose? Do they measure what they want to measure? We think they fail and that the most plausible interpretation of the experimental situation they create is different from intended. The adult that is included in the peer group in the experimental condition has special characteristics that increase his or her status:

1. He or she is a co-organizer of the experiment, which could induce an authority effect similar to those demonstrated in Stanley Milgram experiments.
2. He or she has seniority by academic level compared to the test subjects. Note that the test subjects are recruited from an undergraduate psychology seminar, which might make them more receptive to the status of a graduate psychology student.

One might think that the researchers would make an effort to conceal the fact that the confederate is part of the experiment, but they don't. This means that the effect that is reported could well be a status effect and not an age effect. This is not harmful in the primary intended destination of the study, as in the military formal status and age go hand in hand. But in other contexts where adolescents make risky choices the age could be uncorrelated with status or be

negatively correlated with status, so the findings should not be reported as an universal effect of adult presence on adolescent behavior.

The second concept the authors would like to have measured is the type of risk that is relevant for the sponsors of the study, the military, and for other "organizations that place adolescents in situations in which risky and myopic decision making is problematic" (p. 322). It is unclear that they succeed at capturing this risk and their construct faces validity issues.
The study situation was a game and we do not know its full specification, but the authors write that "at each intersection, the subjects could (a) stop, (b) cross successfully, or (c) crash (as a result of either failure to brake or taking too long to brake after the light turned red)" (p. 325). This suggests that the probability of a crash is a function of both risk taking and skill of the test participant. The authors do not look at actual performance of the test subjects; they categorize their behavior as reckless regardless of the payoff achieved. It is possible that the skilled players could have increased their expected payoff in the game without increasing their overall risk by much (crossing on a red light without a crash) but that behavior, regardless of the outcome, would still be classified as risky behavior by the experimenters. Some of the participants might have taken *legitimate* risks which led to significant gains in performance -- which isn't the type of "risky or myopic behavior" that is problematic in organizations like the military. The authors claim to be measuring the propensity to make reckless decisions, but in fact they are also measuring a skill and the ability to take legitimate risks.

Data Critique Simulations

2023-01-17

```
set.seed(1)
total_people_tested = 290
number_simulations = 1000

regression_sim = function(mean, variance_upper_bound, additional) {
  variances = runif(total_people_tested, 0.0001,variance_upper_bound)
  response = rnorm(total_people_tested, mean, variances)
  condition = factor(c(rep("solo",95), rep("peer",100), rep("adult",95)))
  df = data.frame(response = response, condition = condition)

  lm1 = lm(response ~ condition, data = df)
  peer_adult_p_value = t(summary(lm1)$coefficients[,4])[2]
  solo_adult_p_value = t(summary(lm1)$coefficients[,4])[3]

  # Now rerun the regression with peer as the reference level
  condition = relevel(condition, ref = "peer")
  reordered_df = data.frame(response = response, condition = condition)
  reordered_lm = lm(response ~ condition, data = reordered_df)
  solo_peer_p_value = t(summary(reordered_lm)$coefficients[,4])[3]

  # Only run this part of the code if you want to run the additional indifference
  # point analysis based on the main regression and return all 6 p values
  if (additional) {
    actual = rnorm(total_people_tested, 1000, 200)
    D = rnorm(total_people_tested, 292, 20)
    df$indifference_point = actual / (1 + exp(df$response) * D)

    lm1 = lm(indifference_point ~ condition, data = df)
    pa2 = t(summary(lm1)$coefficients[,4])[2]
    sa2 = t(summary(lm1)$coefficients[,4])[3]

    condition = relevel(condition, ref = "peer")
    reordered_df = data.frame(indifference_point = df$indifference_point, condition = condition)
    reordered_lm = lm(indifference_point ~ condition, data = reordered_df)
    sp2 = t(summary(reordered_lm)$coefficients[,4])[3]

    return(c(peer_adult_p_value, solo_adult_p_value, solo_peer_p_value,pa2,sa2,sp2))

  }

  return(c(peer_adult_p_value, solo_adult_p_value, solo_peer_p_value))
}
```

```
extreme_results = 0
any_results = 0
```

```
simes_any_result = 0

for (i in 1:number_simulations) {

  comparison_values <- regression_sim(0.25, 0.1,FALSE)

  if (sum(comparison_values < 0.05) == 3) {
    extreme_results = extreme_results + 1
  }

  if (sum(comparison_values < 0.05) != 0) {
    any_results = any_results + 1
  }

}

extreme_results/number_simulations
```

```
## [1] 0
```

```
any_results/number_simulations
```

```
## [1] 0.125
```

```
extreme_results = 0
any_results = 0

for (i in 1:number_simulations) {

  risk_ps = regression_sim(0.25, 0.1,FALSE)
  log_k_indifference_ps = regression_sim(-6, 1,TRUE)
  all_ps = c(risk_ps, log_k_indifference_ps)

  if (sum(all_ps < 0.05) >= 5) {
    extreme_results = extreme_results + 1
  }

  if (sum(all_ps < 0.05) != 0) {
    any_results = any_results + 1
  }
}

extreme_results/number_simulations
```

```
## [1] 0.001
```

```
any_results/number_simulations
```

```
## [1] 0.29
```

Stat 27850/30850: real data analysis critique

This is a group assignment—please work in groups of size 2, 3, or 4.
Only one student from each group should submit the assignment to Gradescope (then add the other students as group members).

Select one of the following papers to critique:

1. "Optimism is associated with exceptional longevity in 2 epidemiologic cohorts of men and women", Lee et al, PNAS 2019
   `https://www.pnas.org/content/116/37/18357`
2. "Lung cancer incidence decreases with elevation: evidence for oxygen as an inhaled carcinogen", Simeonov & Himmelstein, PeerJ 2015
   `https://peerj.com/articles/705/`
3. "Adolescents in Peer Groups Make More Prudent Decisions When a Slightly Older Adult Is Present", Silva et al, Psychological Science 2015
   `http://pss.sagepub.com/content/early/2016/01/14/0956797615620379.full.pdf+html`

Please note that to access these articles when not on campus, you may need to log in through UChicago Library.

These papers are chosen as recent research across a range of health & psychology disciplines; they are chosen to represent typical studies, not necessarily as examples of good or bad statistical methodology. Your job is to critique the statistical analyses performed in the paper you choose from the point of view of multiple testing, replicability, and valid inference. You should also assess the validity of the conclusions from a statistician's point of view: think about modeling assumptions, possible confounding variables, whether the data collected is representative of the claim being made, etc.

Some questions you might think about:

- Are there assumptions being made (implicitly or explicitly), e.g., by using certain models, where the validity of the conclusions may depend on these assumptions? Were the assumptions checked in any way?
- Are there explicit issues of multiple comparisons—e.g. looking for effects of multiple variables on the response? If so, were these accounted for in the analysis?
- Are there implicit issues of multiple comparisons, where the analyses that were run, were chosen by examining the data? Can you think of a hypothesis which was not tested, which is equally plausible to the one that was tested?
- Are there many "degrees of freedom" in the choices made for the analysis, such as which subjects to exclude, how to partition variables into categories, which variables to control for, etc? Were these chosen in advance or as part of the analysis?
- Statistical issues not related to multiple comparisons, e.g. possible confounding variables, representativeness of the population, issues of study design, dropping some individuals or observations from the data set, etc.
- And any other relevant issues that come to mind.

What you should hand in: a short report (typically 3–8 pages, but this is flexible) that summarizes the study (including methods, who was in the study, main questions asked, etc), then discusses any relevant points or concerns regarding multiple comparisons or validity of the statistical analyses in the paper. You are encouraged to consider designing a simulation to illustrate any possible issues that you identify.