2024 ■ Vol. 20 ■ no. 3

Teaching Quantitative Methods Vignettes

How can musicians be both more and less empathic than nonmusicians? Using R and Shiny to reveal how Simpson's Paradox emerges when comparing unbalanced groups

Floris van Vugt [a]

[a]Department of Psychology, University of Montreal

Abstract ■ Imagine we want to understand whether musicians are more empathic than non-musicians. We recruit two groups of individuals, musicians and nonmusicians, and administer empathy tests. To compare the empathy scores we could use a two-sample t-test, or, equivalently, a one-way ANOVA. However, before we do so, textbooks advise us we have to investigate whether the two groups are balanced when it comes to other variables, such as age and sex. Why is that important? Students may feel that this is merely a bureaucratic step, a formality of sorts. In this teaching vignette, we explore how group imbalances may affect the conclusions we draw from the dataset. We provide sample datasets as well as an online interactive interface that allows the students to explore when group imbalances are problematic and when they are not. We explore ways to test group imbalances and ways to deal with them, such as including the confounding variables into the analysis model. Overall, the learning activity aims to help sensitize students to the problem of unaccounted group imbalances.

Keywords ■ Put keywords here, in a comma separated list. **Tools** ■ R, Shiny.

floris.van.vugt@umontreal.ca

10.20982/tqmp.20.3.v004

Acting Editor ■ Sébastien Béland (Université de Montréal)

Reviewers
■ Denis Cousineau (Université d'Ottawa)
■ one anonymous reviewer

Concept to be presented

Sometimes, fairly straight-forward seeming questions, such as whether one group has higher scores than another, have surprisingly complicated answers. A famous example is the admittance records for UC Berkeley that seemed to show men were more likely to be admitted than women. However, when taking into account that women tended to apply more to more competitive departments, it was revealed men were actually less likely to be admitted (Bickel et al., 1975). This class of examples, where a group difference inverts when taking into account a third, confounding variable, are referred to as Simpson's paradox (Good & Mittal, 1987; Kievit et al., 2013). Here we use a variant of this paradox in a learning activity to show the importance of identifying and taking into account group imbalances. Statistics students may feel that checking for group balance is a mere formality, the stuff that overzealous analysts do for some obscure reason, but of no real consequence to the interpretation of the data. By contrast, here we aim to give practical examples how group imbalances can give rise to the wrong conclusion altogether.

Note that while most reporting guidelines indicate the importance of checking baseline balance between groups (Gerber et al., 2014), criticism has emerged over the use of statistical tests for that purpose (Senn, 1994; Mutz et al., 2019). Specifically, researchers often use chi-square tests on gender distributions between groups and use non-significance on this test to conclude groups are sufficiently balanced. This practice is questionable on multiple ac-

counts (Senn, 1994). Instead, covariates should be incorporated into the analysis model. This issue is shown in the present vignette.

Teaching strategy

We use simulated data which has been proposed to be an efficient teaching method (Revelle, 2020). It allows students to play with simulation parameters and directly observe the outcome, hence developing an intuition.

We focus on using visualizations (Wainer & Velleman, 2001) to help the student understand the relations between analyses and how they account for the data, especially when presented in an interactive setting (Forbes et al., 2014).

We promote active learning by guiding the student's own investigation (Dolinsky, 2001). Rather than presenting conclusions at the outset, we provide step-by-step instructions that will allow the student to discover the results on their own. To ensure accurate learning in this context, proper debriefing is essential to check that the student has indeed reached the intended conclusions instead of being led astray by mistakes along the road.

Finally, we will use paradox (seemingly conflicting accounts of the same dataset) as a teaching tool because of its potential to contribute to memorable teaching by challenging students to resolve it (Sowey, 1995; Székely et al., 1986).

Learning objectives

The overarching learning objective of this vignette is to understand how imbalances in variables of no interest may affect the conclusions drawn in group comparisons. Specifically, students will learn that unbalanced groups can yield wrong conclusions or reduce statistical power. We will cover important nuances such as that it is the imbalances between groups, rather than within groups, that are problematic. Finally, students will see how often applied tests for imbalance may not be appropriate to determine whether these imbalances are problematic.

Description of the activities

Activity #1: Conflicting analyses of a provided dataset

Rationale. Students are provided a simulated dataset and are asked to analyze it, with and without taking into account a confounding variable. The dataset has been cleverly constructed to yield opposite conclusions in these two cases: first the apparent conclusion is that musicians are less empathic than non-musicians, whereas second that musicians are more empathic (Figure 1). This paradox is resolved when the student understands that it is due to group imbalances on a confounding variable.

Instructions for the students. (Comments in parentheses are intended for the reader, not the students)

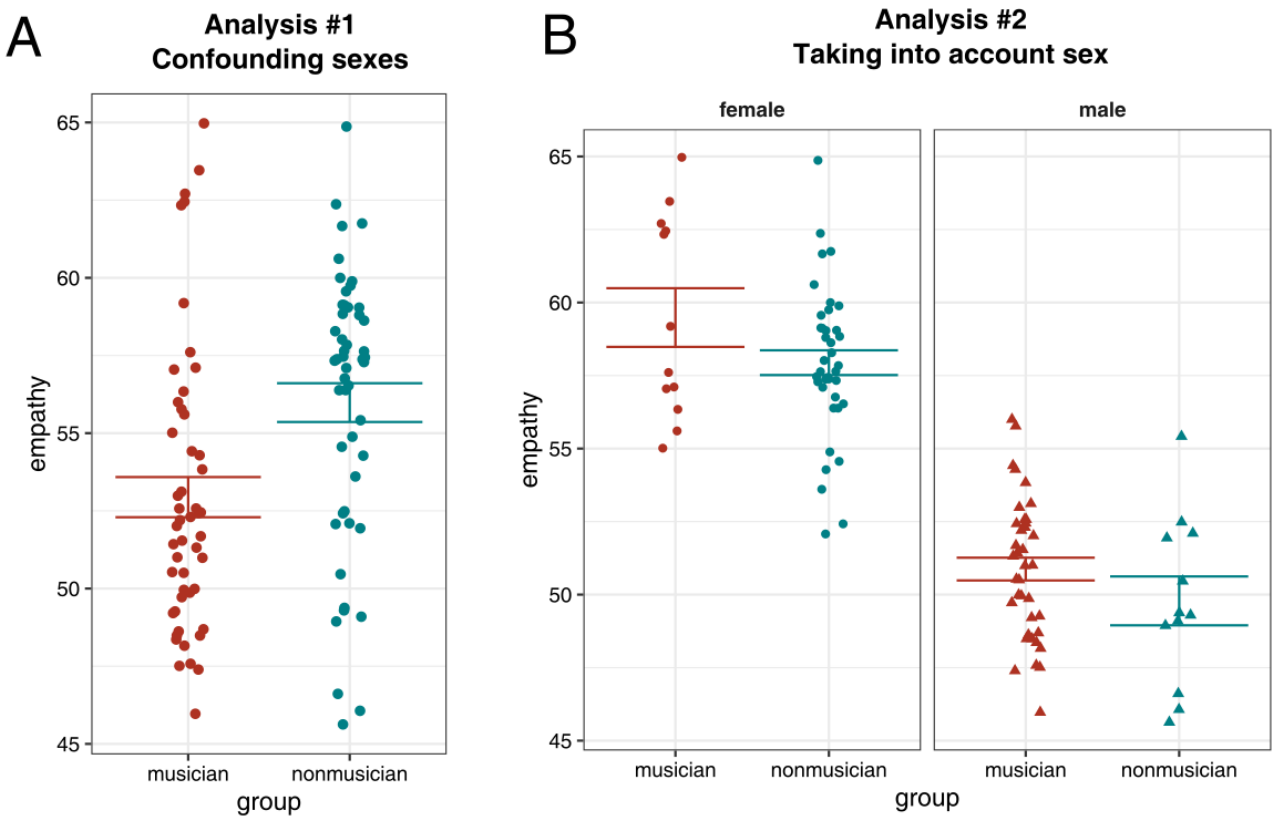
1) In this activity, we will investigate data from a hypothetical experiment testing whether musicians are more empathic than non-musicians. The simulated dataset has one participant per row, either musician or nonmusician, with the empathy score indicated numerically. Download the provided dataset and import it in R. Inspect the table to ensure you understand how the above information is represented.
2) Produce a visualization that indicates empathy scores of musicians and nonmusicians separately. Indicate standard errors using error bars. What do you observe?
3) Perform a one-way ANOVA to compare the empathy scores between musicians and non-musicians. What do you conclude? (We expect students at this point to conclude that musicians are less (!) empathic than musicians.)
4) The dataset also indicates the sex for each participant (male/female). What is the distribution of males and females in the two groups? Are the distributions comparable?
5) Produce a new visualization that splits the musician and nonmusician group according to sex. What do you observe? Run a two-way ANOVA with musicianship and sex as factors. Does empathy differ between males and females? Does empathy differ between musicians and non-musicians? Interpret the results. What do you conclude? (We expect students at this point to conclude that musicians are more empathic than nonmusicians)
6) Reflect on how it was possible to reach two different conclusions at points 3 and 5 above.

Debriefing. Discuss the conclusions drawn by the students in the entire class. The envisaged conclusion is that group imbalances can lead to erroneous conclusions, in this case that musicians are less empathic than non-musicians, whereas in reality the opposite is true. The reason that this occurred was that the groups were unbalanced according to sex (there were more females among the non-musicians than among the musicians), coupled with the effect that females are more empathic.

Activity #2 : Interactive simulations

Rationale. To understand when confounding variables are problematic, different configurations of group sizes, distributions, effect sizes and directions need to be explored. For this, an interactive simulation is proposed using the Shiny platform (Sievert, 2020); for another pedagogical application see Hoisington-Shaw and Pek (2021). This unique online platform allows the student to run preprogrammed simulations while using sliders to vary parameters such as group sizes and sex distributions, immediately

Figure 1 ■ The dataset that was purposefully created to suggest that nonmusicians are more empathic than musicians (panel A). Dots indicate simulated individuals and error bars indicate standard error. In reality, this difference is driven by a sex difference in empathy where females are more empathic than males, coupled with a greater proportion of females in the non-musician group. Accounting for this imbalance, musicians are in fact more empathic, which is true for both males and females (panel B).

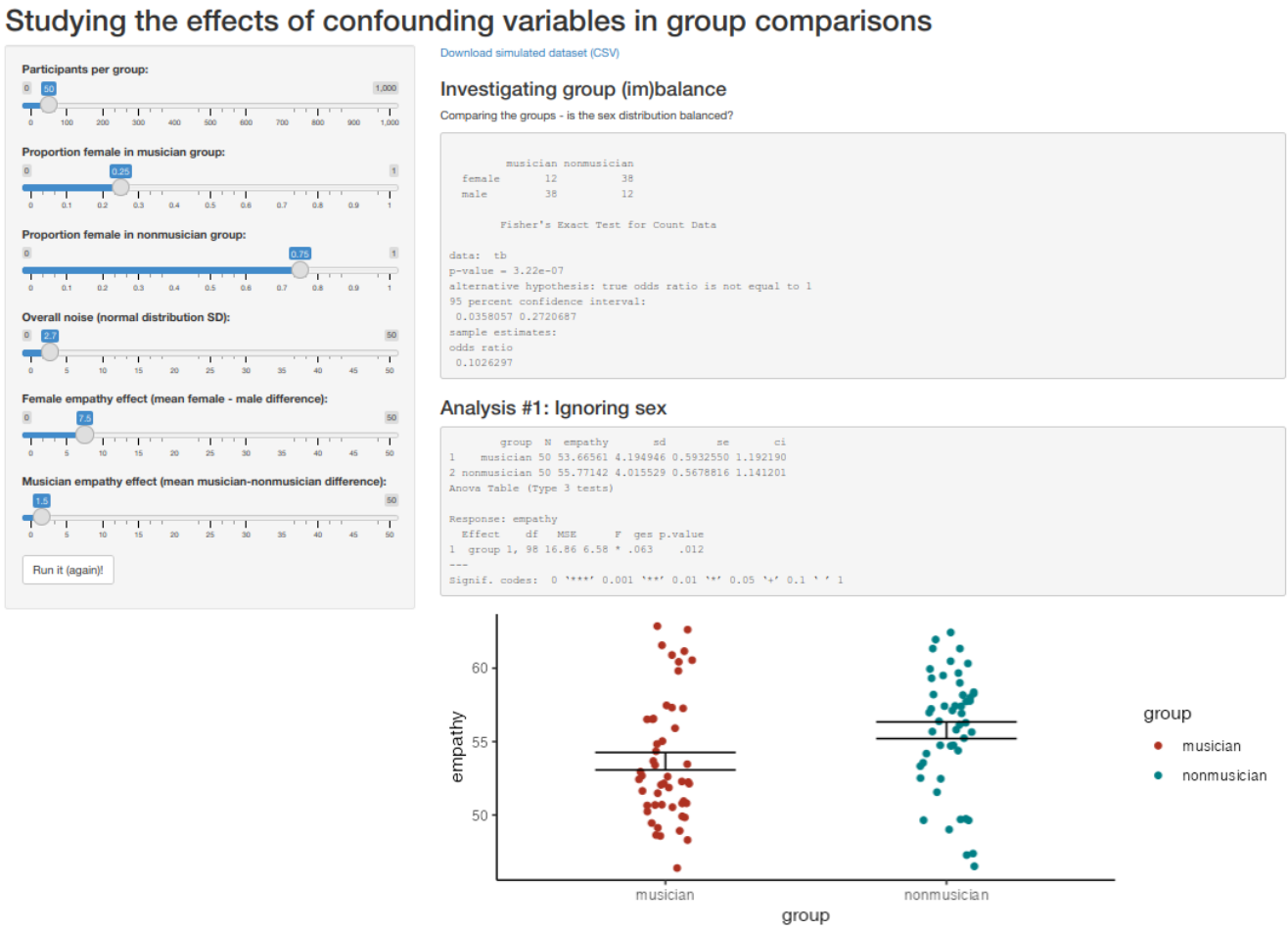


observing the results (Figure 2). This playful interactive activity will allow the students to develop and test intuitions about the effect of confounding variables.

Instructions for the students.

1) We will study in what situations confounding variables become problematic. We use the same hypothetical experiment in which we test whether musicians are more empathic than non-musicians. Open the address https://vanvugt.shinyapps.io/GroupImbalances/ in your web browser. On the left hand side, you can select the number of participants per group, the proportion of females in the musician and non-musician groups, separately, as well as other variables that we will get to later. For now, select 50 participants per group, set the proportion of females in the musician group to 0.25 and in the non-musician group to 0.75. Set the noise to 2.7. Set the female empathy effect to 7.5 and the musician empathy effect to 1.5. Click the "Run it (again)!" button. By clicking this button you can run the hypothetical experiment over and over, to yield different results using the same settings. Familiarize yourself with the analyses indicated on the right hand side of your screen, which should be similar to the previous assignment. What do you conclude in this particular simulation?
2) The one-way ANOVA not taking into account sex yields an erroneous conclusion. What error type is this (type I or II)? We could hypothesize that this is due to the sex imbalance between groups. To test this idea, set the proportion of females to 0.5 in both groups. What do you observe? (We expect the students to observe that the

Figure 2 ■ Interactive web-based interface that allows the student to change parameters of the simulation to directly observe the effects on the analyses. This was created using the Shiny platform.



one-way ANOVA now gives the correct result).

3) Gender imbalances may not always be so large that they cause the wrong conclusion. However, even for smaller imbalances, the analysis that incorporates sex (the two-way ANOVA) may still be preferable. Here we will explore why. Set the proportion female to 0.4 for musicians and 0.6 for nonmusicians. Leave the noise on 2.7. Reduce the female empathy effect to 4. Run the hypothetical analysis a few times and track the results of the one-way ANOVA and of the two-way ANOVA. What do you observe? Why is the two-way ANOVA preferable? (We expect the students to observe here that the musicianship effect tends to be non-significant for the one-way ANOVA but significant for the two-way ANOVA, indicating greater statistical power in the latter case).
4) How can we know when a confounding variable (sex in this case) is imbalanced enough between groups to be problematic? One typical practice is to use the Fisher Exact Test to test whether the sex distributions are different in the two groups, and consider it problematic when the test comes out significant. However, here we will see this test can reach significance without it really being a problem. Increase the group size to 1000. Set the proportion female to 0.45 for musicians and 0.55 for nonmusicians. Crank up the noise to 15 and reduce the female empathy effect of 0.6 and the musician empathy effect to 1.5. Run a few simulations and summarize the results: does the Fisher test tend to come out significant? Do the one-way or two-way ANOVA analyses tend to detect the musicianship effect? (We expect the students to notice that the Fisher test tends to be significant while both versions of the ANOVA tend to agree. Hence

the group imbalance is small enough, hence unproblematic). Advanced: can you find parameters for which the group imbalance is problematic but the Fisher test tends to come out non-significant? (We expect the students to observe that with small sample sizes and large imbalances, Fisher tests may not come out significant while the one-way ANOVA yields a wrong conclusion)

5) Advanced exercise: It is often thought that sexes need to be equally represented within each group. Here we study what happens if the sexes are unbalanced within both groups, but in exactly the same way in both groups. Use the sliders to create sex imbalances ensuring that the imbalance is the same in both groups, e.g. 20% females in both groups. What do you observe? Does the one-way ANOVA tend to yield the wrong conclusion? (We expect the student to pick up the subtlety that as long as the two groups show the same internal imbalance, the one-way ANOVA analysis tends to be correct)
6) Advanced exercise: What if there were no empathy difference between females and males? Set it to zero in the interface. In what cases do the one-way or two-way ANOVA risk drawing the wrong conclusion? (We expect the students to observe that in this case, group imbalances fail to be problematic. They could conclude that group imbalances are only problematic if the variables along which the groups are imbalanced show differences on the dependent variable).
7) Advanced open-ended exercise: Explore the simulation further to see if you can discover notable patterns beyond what we have discussed above. Articulate these patterns in the form of rules of thumb and discuss with your peers.

Debriefing. Use in-class discussion to check that students conclude that group imbalances can lead to erroneous conclusions. In less extreme, more common cases, they can lead to loss of statistical power. These problems occur only if the confounding variable (sex in this case) is associated with differences in the dependent variable (empathy). Typically, balance checks on the distribution of the confounding variable across groups (e.g. Fisher Exact test) are insufficient. Preferably confounding variables should be included in the model. We also introduce an important nuance that the problem shown here occurs with imbalances between groups, rather than within groups.

Strategies to assess the activity

All or part of the exercises in the instructions for the students sections above could be used as exam questions or for-grade assignments. The students' understanding can also be assessed by reverse problem construction as follows. The student can be asked to choose parameters in the interactive interface used in Activity #2 to illustrate a particular problem, such as cases in which group imbalances are not problematic.

Conclusion

The activity presented here allows the students to learn about the potential effects of imbalances on some variables between groups on the study of differences in other variables. The activity allows developing intuitions for patterns of results that can occur when three variables (an outcome variable, a grouping variable, and a confounding variable) conspire. This can provide the stepping stone for teaching advanced analyses such as mediation or path analysis.

Authors' note

This work was supported by NSERC Discovery Program RGPIN-2022-04362 and FRQNT Support for new academics #316243.

References

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, *187*(4175), 398–404. doi: 10.1126/science.187.4175.398.

Dolinsky, B. (2001). An active learning approach to teaching statistics. *Teaching of Psychology*, *28*(1), 55–56. https://psycnet.apa.org/record/2001-03050-015

Forbes, S., Chapman, J., Harraway, J., Stirling, D., & Wild, C. (2014). Use of data visualisation in the teaching of statistics: A new zealand perspective. *Statistics Education Research Journal*, *13*(2), 187–201. doi: 10.52041/serj.v13i2.290.

Gerber, A., Arceneaux, K., Boudreau, C., Dowling, C., Hillygus, S., Palfrey, T., ..., & Hendry, D. J. (2014). Reporting guidelines for experimental research: A report from the experimental research section standards committee. *Journal of Experimental Political Science*, *1*(1), 81–98. doi: 10.1017/xps.2014.11.

Good, I. J., & Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables. *The Annals of Statistics*, *15*, 694–711. https://www.jstor.org/stable/2241334

Hoisington-Shaw, K. J., & Pek, J. (2021). Using dynamic graphics to teach the sampling distribution with active learning. *The Quantitative Methods for Psychology*, *17*(2), v1–v9. doi: 10.20982/tqmp.17.2.v001.

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in psychology*, *4*, 513–513. doi: 10.3389/fpsyg.2013.00513.

Mutz, D. C., Pemantle, R., & Pham, P. (2019). The perils of balance testing in experimental design: Messy analyses of clean data. *The American Statistician*, *73*(1), 32–42. doi: 10.1080/00031305.2017.1322143.

Revelle, W. (2020). Teaching research methods using simulation. In W. Revelle (Ed.), *Teaching statistics and quantitative methods in the 21st century* (pp. 217–237). Routledge. doi: 10.4324/9780429442810-16.

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in medicine*, *13*(17), 1715–1726. doi: 10.1002/sim.4780131703.

Sievert, C. (2020). *Interactive web-based data visualization with r, plotly, and shiny*. CRC Press. doi: 10.1201/9780429447273.

Sowey, E. R. (1995). Teaching statistics: Making it memorable. *Journal of Statistics Education*, *3*(2), 1–9. doi: 10.1080/10691898.1995.11910487.

Székely, G. J., Alpár, M., & Unger, É. (1986). *Paradoxes in probability theory and mathematical statistics*. Dordrecht. https://archive.org/details/paradoxesinproba0000szek

Wainer, H., & Velleman, P. F. (2001). Statistical graphics: Mapping the pathways of science. *Annual Review of Psychology*, *52*(1), 305–335. doi: 10.1146/annurev.psych.52.1.305.

Appendix A: R listing for simulating the dataset

```
require('tidyverse')
set.seed(20230914)

n.pergroup    <- 50
p.f.mus       <- .25 # in musician group, probability of being female
p.f.nonmus    <- .75 # in nonmusician group, probability of being female
empathy.base  <- 50
sd.noise      <- 2.7 # amount of noise
f.empathy     <- 7.5 # boost in empathy for females relative to males
mus.empathy   <- 1.5 # boost in empathy for musicians relative to non-musicians

n.f.mus       <- round(n.pergroup*p.f.mus)
N.f.nonmus    <- round(n.pergroup*p.f.nonmus)

dat <- rbind(
    data.frame(
        group='musician',
        sex=c(rep('female', n.f.mus),rep('male', n.pergroup-n.f.mus))
    ),
    data.frame(
        group='nonmusician',
        sex=c(rep('female', n.f.nonmus),rep('male', n.pergroup-n.f.nonmus))
    )
)
dat$group <- factor(dat$group)
dat$sex   <- factor(dat$sex)

dat <- dat %>%
    mutate(
        empathy = empathy.base+
            sd.noise*rnorm(1:nrow(.))+
            f.empathy*(sex=='female')+
            mus.empathy*(group=='musician'),
        id=sprintf('s%d',1:nrow(.))
    )
```

```
write.csv(dat, file='empathymusic.csv')
```

Citation

van Vugt, F. (2024). How can musicians be both more and less empathic than nonmusicians? using R and Shiny to reveal how Simpson's paradox emerges when comparing unbalanced groups. *The Quantitative Methods for Psychology*, *20*(3), v4–v10. doi: 10.20982/tqmp.20.3.v004.

Received: 23/09/2023 ~ Accepted: 14/03/2024

Extended activity metadata

Concept illustrated	Group balance, confounding variables, statistical power, Type I/II error	*Type of activity*	Guided dataset analysis; interactive dataset generation
Prerequisite	N/A	*Types of data*	Simulated datasets
Co-requisite	(i) Basic descriptive statistics (mean, standard deviation); (ii) Between-subjects ANOVA; (iii) Fisher Exact Test; (iv) Basic knowledge of R: reading CSV data sets, producing basic visualizations (ggplot2), ANOVA; (v) For the advanced exercises: understanding statistical power.	*Computation by*	R, Shiny
Suitable class size	Small groups (5-15)	*Duration*	2 × 1-2 hours