**Teaching the p-value alone is not enough: A discussion of p-hacking and other pitfalls of using p-values in research**

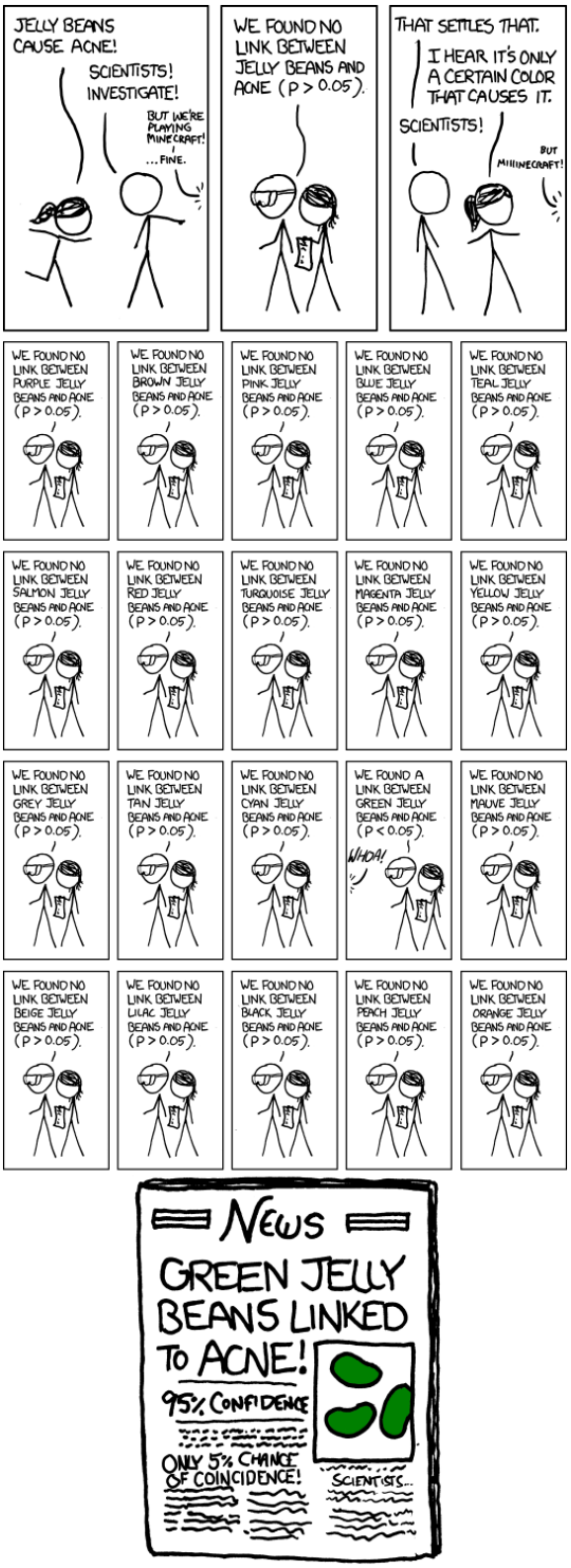
Billy W. Esra II
Math Teacher at Thomasville High School Scholars Academy
Adjunct Professor at Thomas University

According to an article that has shown up in my Facebook feed several times over the last few weeks, "smelling farts may be good for your health" (DeMaria, 2014). After reading the headline, I was skeptical and decided to take the bait and read the linked article which eventually explained that the study referenced actually had nothing to do with farts and long lasting health, but instead with a correlation that was found between the delivery of small amounts of a hydrogen sulfide compound to mitochondria and damage reparation of those cells. To spin the findings into something that people are more likely to click on and read, the news media made the connection between the hydrogen sulfide compound that was used in the study with the hydrogen sulfide compound that is expelled when the human body emits a fart. This article points to several potential pitfalls in scientific research and science reporting. AP Statistics teachers discuss many of these issues when discussing experimental design. We ask students to be critical of reported science and to ask questions of the research process used by the scientists and any assumptions made by non-science periodicals.

Unfortunately, researchers sometimes appear to be using appropriate scientific and statistical methods, while, in fact, they have p-hacked the results to show statistical significance. In order to show statistical significance when performing a significance test, researchers need a p-value that is below the acceptable significance level. A p-value is essentially the probability that a researcher gets the sample outcome that they got or something more extreme, assuming that the null hypothesis is true. If the p-value is less than the significance level (sometimes called the alpha level), then the researcher has statistically significant evidence of a difference from the null hypothesis. For example, my local grocery store is selling five pound bags of peaches. If I gathered a sample of bags of peaches and weighed each bag, the resulting average weight is very likely not going to be exactly five pounds. There is inherent variability in the weight of individual peaches, and the bags cannot be filled to exactly five pounds. We would expect some bags to be a bit heavier and some to be a bit lighter than the grocer claims. If we suspect that the grocer is under filling the bags, we can perform a one sample t-test (a type of significance test) to check this claim. If we get a p-value of 0.013, assuming conditions for inference are met, then if the bags of peaches do have an average weight of five pounds, the probability that we would get the sample average weight that we got or something more extreme is 1.3%, which is not very likely. We might conclude that the bags might not actually be filled with advertised five pounds of peaches.

The significance level is a sort of "line in the sand" where the researcher feels that the results, based on the p-value, are unlikely to have been obtained solely by the inherent randomness of

sampling. Traditionally, many fields of science and social science use 0.05 as an accepted significance level, though some fields and publications may require a much more conservative significance level of 0.01 or something smaller. One of the problems with using a significance level of 0.05 is that if the null hypothesis is indeed true, 5% of the samples obtained, or one in 20, will cause the researcher to reject the null hypothesis in error (if the procedure were repeated with the same sample size from the same population). A fantastic webcomic by xkcd.com (Munroe) displays this concept beautifully. After testing 20 different jelly bean colors for a link between jelly beans and acne, the researcher found a statistically significant link between only green jelly beans and acne. Ignoring the fact that no other color led to a low enough p-value, the newspaper headline warned of a link between green jelly beans and acne. Obviously, the results of the green jelly bean study had a few issues.




AP Statistics teachers discuss this type of error with our students with the moniker “Type 1 Error,” where the null hypothesis is incorrectly rejected. Unfortunately, this type of error is sometimes exploited by researchers. Often in the world of academia, research findings will not be published unless

there is a statistically significant result. According to the old adage, "publish or perish," advancement in degree programs and careers can be stymied without publishable results. Scientific research can be extremely challenging, then, because, often the results are not statistically significant. To find statistically significant results, some researchers have begun to throw as many variables as they can at a research question in order to find at least one variable that shows statistical significance, and is, hence, publishable. This practice of hacking the p-value, or p-hacking, is fairly deplorable from a statistical standpoint, but it often achieves results. For an example of how easy p-hacking can be, check out fivethirthyeight.com's article "Science Isn't Broken" (Aschwanden, 2015). An app available in the article that uses real data can show, with statistical significance, that Republicans in office are bad for the economy or that Republicans in office are good for the economy or that Democrats in office are bad for the economy or that Democrats in office are good for the economy. By controlling which variable(s) are used to measure politicians in office (Presidents, Governors, Senators, and/or Representatives) and which variable(s) are used to measure economic performance (Employment, Inflation, GDP, and/or Stock Prices) you can prove any relationship that you'd like. The practice of p-hacking detracts public confidence from the large quantity of scientists and statisticians that are using inference procedures appropriately. Even, comedian John Oliver, in his show "Last Week Tonight with John Oliver," brought attention to this misuse of statistics in a show on scientific studies (Oliver, 2016). (Warning – John Oliver often uses vulgar language as a part of his comedic news reporting. A PG Version (Oliver, 2016) of the episode was edited by a statistics teacher for use in the classroom.)

P-hacking and other misuses of the p-value led the American Statistical Association (ASA) to release "six principles" of p-values (McCook, 2016) in order to outline the appropriate uses and interpretations of p-values. The actual article written for the ASA can be found on the AMSTAT website (Wasserstein & Lazar, 2016). The ASA feels that people are not purposefully misusing the p-value, but rather that data analysis is not being performed by statisticians who are specifically trained on statistical analysis. Essentially, the p-value should not be the end-all decision maker for whether science is publishable, or important. If statistical significance is found, the researcher should be led to even more questions. Can the results be reproduced, or replicated, in additional studies? (Obtaining funding for replication studies can be difficult.) What is the effect size? (While effect size is not explicitly part of the AP Statistics curriculum, it would be worth discussing with students.) What is the practical importance of the findings? (Just because a study finds statistically significant results, there are no assurances that any practical implications of the findings will be evident.) Ultimately, a low p-value, by itself, should not be used in isolation to display a studies merit. As AP Statistic teachers, we coach our students on how to make an inference decision of rejecting the null hypothesis or failing to reject the null hypothesis based on the p-value. We may do our students, and the future of scientific discovery, a service by spending a little bit of time on how the p-value should be used in research. We should also caution students about practices, like p-hacking, that lower confidence in scientific discoveries and statistical analysis. My hope is that one day, after enough AP Statistics students pass through our classrooms, my social media newsfeed will no longer be filled with studies that link smelling farts to living longer!

References:

Aschwanden, C. (2015), “Science Isn’t Broken: It’s just a hell of a lot harder than we give it credit for,” FiveThirtyEight, available at http://fivethirtyeight.com/features/science-isnt-broken/#part1

DeMaria, M. (2014), “Study: Smelling farts may be good for your health,” The Week, available at http://theweek.com/speedreads/450160/study-smelling-farts-may-good-health

McCook, A. (2016), “We’re using a common statistical test all wrong. Statisticians want to fix that,” Retraction Watch, available at http://retractionwatch.com/2016/03/07/were-using-a-common-statistical-test-all-wrong-statisticians-want-to-fix-that/

Munroe, R. “Significant,” xkcd, available at https://xkcd.com/882/

Oliver, J. (2016), “Last Week Tonight with John Oliver: Scientific Studies (HBO),” YouTube Video available at https://www.youtube.com/watch?v=0Rnq1NpHdmw

Oliver, J. (2016), “Oliver_Scientific Studies.mp4,” an edited for content version of “Last Week Tonight with John Oliver: Scientific Studies (HBO),” available at https://drive.google.com/file/d/0B_TxGBATbtQtQUdSOUxCRnR5UG8/view?pli=1

Wasserstein, R. L., and Lazar, N. A. (2016), “The ASA’s statement on p-values: context, process, and purpose,” The American Statistician, DOI: 10.1080/00031305.2016.1154108, available at http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108