Using Intuitive Test Statistics in a Randomization-Based Introductory Course

Todd Swanson[1], Jill VanderStoep[1], Nathan Tintle[2], Beth Chance[3], Allan Rossman[3], Soma Roy[3] and George Cobb[4] ([1]Hope College, [2]Dordt College, [3]California Polytechnic State University, [4]Mount Holyoke College)

Introduction

Randomization-based introductions to statistics are gaining momentum. We think this approach offers students a more intuitive, more direct path to understanding the logic and scope of inference. Recently published assessment data offer preliminary support for this hypothesis (Tintle, et al. 2011 and Tintle et al. 2012), although many questions remain.

One potential advantage of the randomization-based approach is its flexibility. The traditional approach is tied to test statistics and pivotal quantities like z, t, and F that rely on the Central Limit Theorem, and that require an understanding of variance. This approach also requires students learn about standardization, t versus z, and perhaps variance ratios while they are still learning the logic and scope of inference. Alternatively, the randomization-based approach allows students to work with more intuitive statistics such as a single proportion for one group, the difference in means or difference in medians for comparing two groups. For comparing three or more groups, they can work with the mean absolute pairwise differences of means or proportions and put off learning about more complicated statistics until later.

Throughout our course, we have students develop null distributions through coin-flipping and shuffling cards. Most of the time, however, we use computer applets to speed up this process. Finally, we show that theory-based distributions can be used to predict the shape, center, and variability of our null distributions when certain conditions are met.

Single Proportion: Do Dogs Understand Human Gestures?

In chapter 1, students explore results from an experiment done to determine if dogs can understand human gestures (pointing, bowing, looking towards one of two cups, etc.). In one set of trials a dog named Harley went to the cup that was bowed towards 9 out 10 times. We present the students with two possible explanations.

- He is merely guessing at random between the two objects and in these 10 trials happened to guess correctly in 9 of them.
- He is doing something other than merely guessing and perhaps understands the gesture of bowing towards an object.

Students are then asked to think about how we can conduct a simulation of the first explanation (null hypothesis). As a group, the students quickly come up with the idea of flipping a coin 10 times and counting the number of heads to represent one simulated trial. Each student flips a coin 10 times and records their results on the board to develop a null distribution like the one shown in Figure 1. At this point we are not even using a proportion as our statistic, but the number of successes. In the next section, they are introduced to the terminology surrounding a test of significance as well as the use of proportion as the statistic.

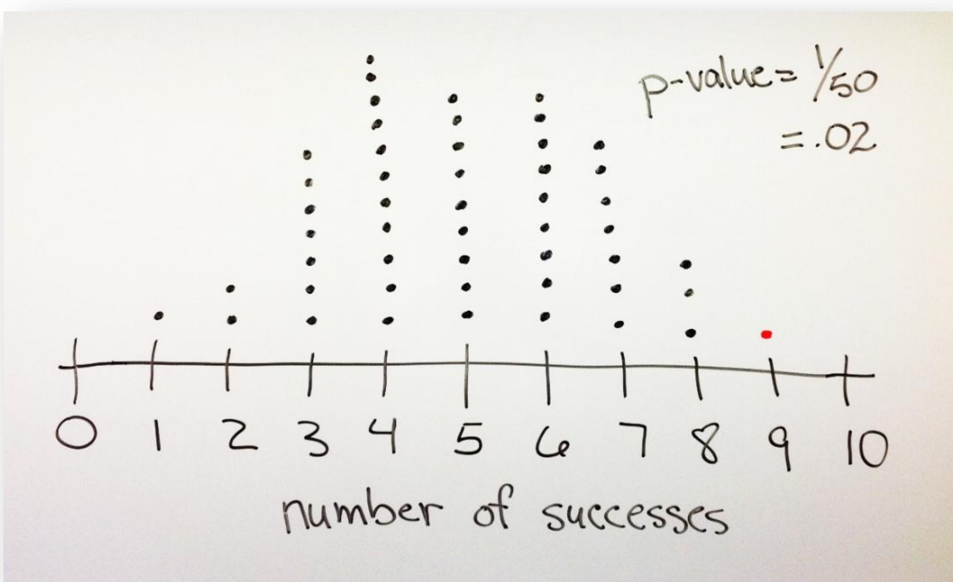


Figure 1: A null distribution of simulated statistics that represents the number of times Harley would choose the correct cup just by chance. From this, we can see it is highly unlikely Harley would make 9 or more correct choices if he was just guessing.

Matched Pairs: Rounding First Base

When stretching a single into a double in baseball, does the path you take to round first base make a difference? More specifically, does a runner make it to second base faster if they round first base with a narrow angle or with a wide angle? (See Figure 2.)

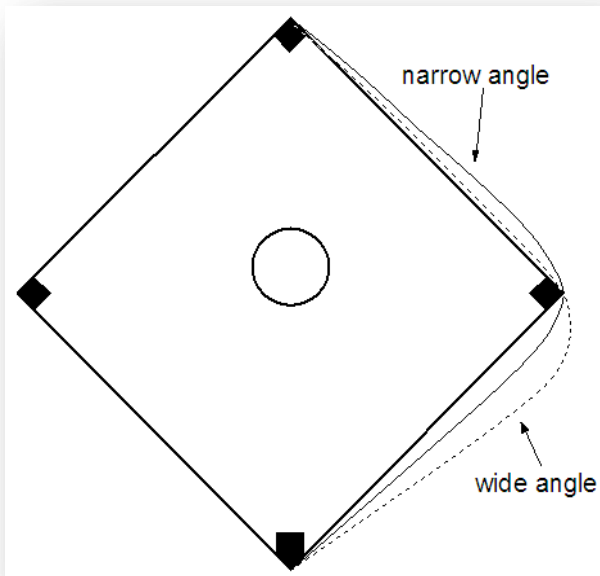


Figure 2: Two methods of rounding first base when a player plans to run to second base.

Hollander & Wolfe (1999) report on a Master's thesis by W. F. Woodward (1970) in which 22 runners were timed running each path. (They were actually timed from a spot 35 feet past home to a spot 15 feet before second.)   Table 1 shows the differences for the first 10 runners. The average of the differences in their times, 0.75, is our statistic.

Subject	1	2	3	4	5	6	7	8	9	10	
narrow angle	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	…
wide angle	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	…
difference	-0.05	-0.05	0.1	0.1	0.15	-0.05	0.05	0.15	0.15	0.10	…

Table 1: Running times for 10 of the runners (in seconds). The Last row is difference in times for each runner (narrow – wide).

The null hypothesis basically says the base running strategy doesn't make a difference in their time. To model this using simulation, we assume the same two times for each runner, but randomly decide (with a coin flip) which time goes with the narrow path and which with the wide path. The students do this, compute the new mean of the differences, and put their simulated statistics on the board to create a null distribution. From that, a p-value is determined. (See Figure 3.)

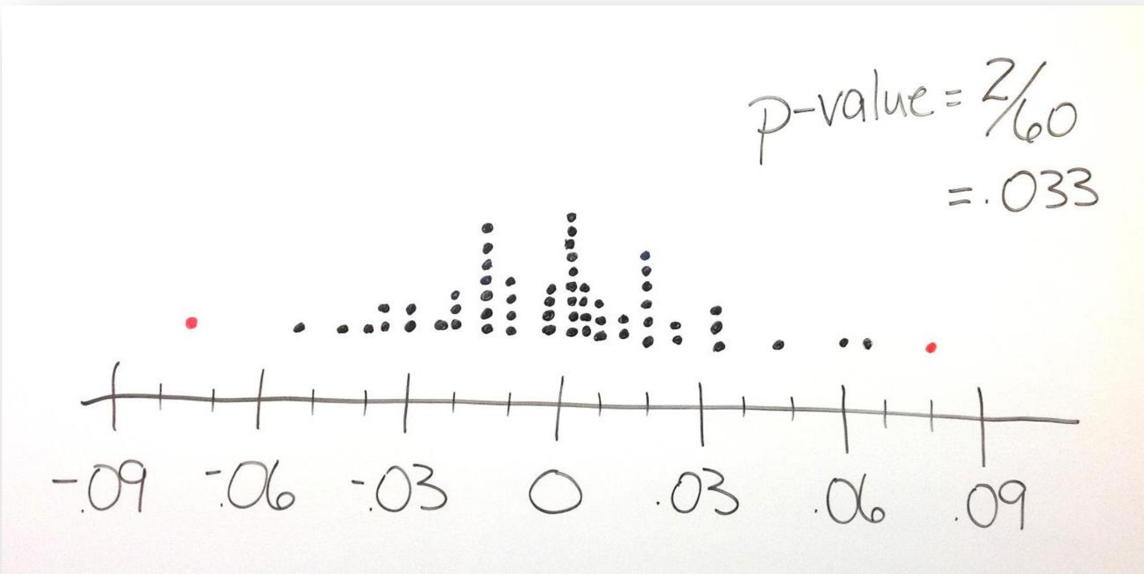


Figure 3: A null distribution of simulated means of the differences in base running times created by randomly assigning (through coin flips) the base running times to the paths for each pair.

The student-created null distribution is followed up using an applet that will simulate the same process quickly. This is followed up by using theory-based methods and a *t*-distribution to fit our simulated null distribution.

Comparing Multiple Proportions: Night Lights and Near-Sightedness

Recent studies have explored whether there is an association between development of near-sightedness and the use of nightlights with infants. One study (Quinn, et al, 1999) interviewed parents of 479 children who were seen in a university pediatric ophthalmology clinic. One of the questions asked was whether the child typically slept with the room light on, a night light on, or in darkness before age 2. Based on the child's most recent eye examination, the children were also separated into two groups: near-sighted or not near-sighted. They found a higher percentage of near-sighted children among those using a room light (54.7%) or with a night light (33.6%) compared to children who slept in darkness (10.5%). (See Table 2.)

	Room Light	Night Light	Darkness	Total
Near-sighted	41 (54.7%)	78 (33.6%)	18 (10.5%)	137
Not near-sighted	34	154	154	342
Total	75	232	172	479

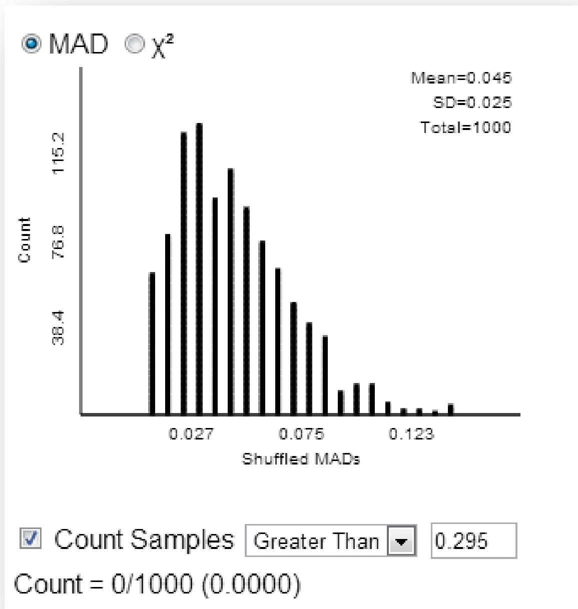
Table 2: We can see in this observational study that the more light that was in a bedroom at night, the more likely the child sleeping there would become near-sighted.

When comparing two proportions, it seems obvious that an appropriate statistic to use is the difference in the two sample proportions. What about comparing more than two proportions? What single number can be used to show how far apart our three proportions are from one another? Students will suggest some ideas, some of which will be viable and some not so much. With a small amount of prompting, an intuitive statistic they will quickly suggest is what we call the *MAD* statistic (or mean absolute differences). This mean of the absolute value of the differences in the conditional proportions provides a measure of how far apart these sample proportions are on average. For our night light and near-sightedness example, the *MAD* statistic is:

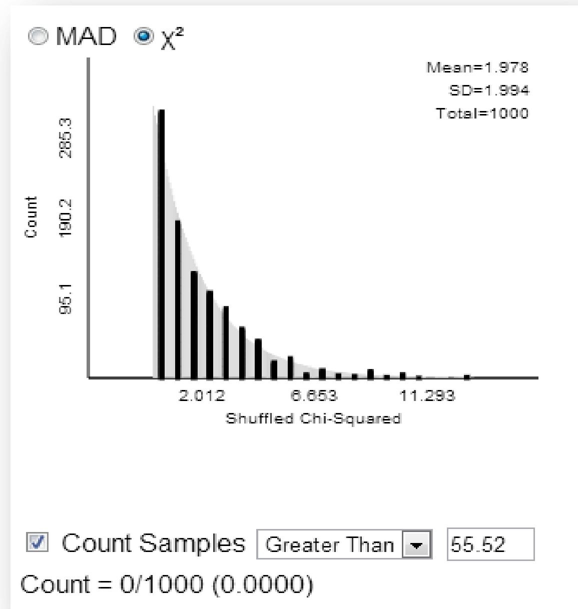
$$MAD = \frac{0.547 - 0.336 \; + \; 0.547 - 0.105 \; + \; 0.336 - 0.105}{3} = 0.295$$

It is also obvious to the students that the smallest value for the MAD statistic is 0 (when all the proportions are the same) and the greater the distance between proportions results in a larger *MAD* statistic.

Just as in comparing two means or two proportions, we can simulate the "no association" under the null by shuffling the values of the response variable. This models the situation where the explanatory variable (the amount of light) is not associated with eyesight. We used our applet to conduct 1000 shuffles of the data, which yielded 1000 simulated values of the statistic as shown in Figure 4(a).



(a)



(b)

Figure 4: **(a)** A null distribution of simulated *MAD* statistics resulting from the hypothesis of no association between the type of light in infants' bedrooms and whether or not they are near-sighted as children. **(b)** A null distribution of simulated chi-square statistics for the same scenario along with an overlay of the theory-based chi-square distribution.

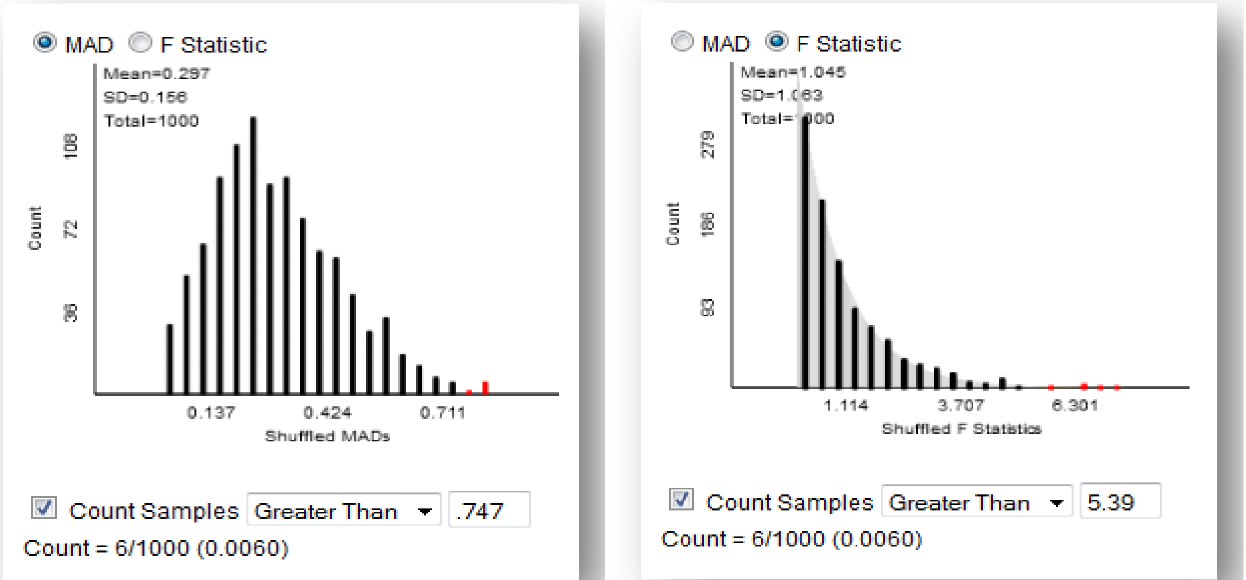
The applet allows us to use a chi-square statistic instead of the *MAD* statistic. (See Figure 4(b).) Students see that this statistic has many similarities to the *MAD* statistic, but also allows for more than two outcomes for the response variable as well as gives a smoother distribution that can be more easily modeled using a theory-based distribution. This helps students transition from an intuitive simulation-based test to the theory-based chi-square test.

Comparing Multiple Means: Which Diet Works Best?

Dieting always seems to be waxing and waning in pop culture. One study looked at 232 women who were randomly assigned to one of three different diets: Atkins, Ornish, or Zone. (Gardner, et al. 2007) The change in their body mass index (BMI) was measured. We can use the mean change in BMI for each diet to describe the effects of the diet, but we need a single summary statistic that compares the three groups. If we were comparing only two diet groups, the difference in means would be a logical summary statistic. Just as in comparing multiple proportions, when comparing three or more means we can again use the *MAD* statistic. In this study, the change in BMI for the subjects assigned to the Atkins diet was -1.65, for the Zone diet it was -0.53, and for the Ornish diet it was -0.77. This gives us a *MAD* statistic of:

$$MAD = \frac{-1.65 - (-0.53) \; + \; -1.65 - (-0.77) \; + \; -0.53 - (-0.77)}{3} = 0.747.$$

The null hypothesis is that there is no association between the type of diet and the change in BMI. To simulate this we shuffle all the BMI changes and randomly assign them to a diet, then recalculate the *MAD* statistic from the simulated data. This is the same basic process as is done when comparing two means, two proportions, or multiple proportions. We do this many times, say 1000, to get an idea of what values of the *MAD* statistic would be expected if indeed there were no association between diet type and change in BMI. From the applet we see that the observed *MAD* statistic from our study, or one more extreme, only occurred 6 out of 1000 simulations. (See Figure 5(a).)



(a) **(b)**

Figure 5: The simulated *MAD* statistics are used in (a) to develop a null distribution while the simulated *F*-statistics are used in (b). A theory-based *F*-distribution is also shown in (b) over the simulated distribution.

The applet also has the option of calculating the *F*-statistic in the simulation. This simulated null matches up nicely with the over-laid theory-based *F*-distribution. (See Figure 5(b).) Just as with comparing multiple proportions, this provides a nice transition from using the intuitive *MAD* statistic to the theory-based ANOVA test.

Literature Cited

- Gardner CD, Kiazand A, Alhassan S, et al. (2007), "Comparison of the Atkins, Zone, Ornish, and LEARN Diets for Change in Weight and Related Risk Factors Among Overweight Premenopausal Women: The A TO Z Weight Loss Study: A Randomized Trial," *JAMA.* 297(9):969-977.
- Hollander, M. and Wolfe D., (1999), *Nonparametric Statistical Methods, 2nd Edition*, Wiley.
- Quinn, G.E., Shin, C.H., Maguire, M.G., Stone, R.A., (1999), "Myopia and ambient lighting at night," Nature 399:113–114.
- Tintle, N., VanderStoep, J., Holmes, V-L, Quisenberry, B. & Swanson, T. (2011), "Development and assessment of a preliminary randomization-based introductory statistics curriculum," *Journal of Statistics Education*, 19(1).
- Tintle, N., VanderStoep, J., Holmes, V-L, Topliff, K. & Swanson, T. (2012) "Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum," *Statistics Education Research Journal*, 11(1):21-40.
- Udell M.A.R, Giglio R.F, Wynne C.D.L. (2008), "Domestic dogs (*Canis familiaris*) use human gestures but not nonhuman tokens to find hidden food." *Journal of Comparative Psychology*. 122(1):84-93.
- Woodward, W.F. (1970), *A Comparison of Base Running Methods in Baseball*, Master's Thesis, Florida State University.

Acknowledgments

Funding for this project has come from a Howard Hughes Medical Institute Undergraduate Science Education program grant to Hope College, the Great Lakes College Association Pathways to Learning Collegium, the Teagle Foundation, and the National Science Foundation (DUE 1140629).

Further information

This project will soon be published as *Introduction to Statistical Investigations* by John Wiley & Sons. For further information please contact Nathan Tintle at *nathan.tintle@dordt.edu or visit www.math.hope.edu/isi*.