USING SIMULATION/RANDOMIZATION TO INTRODUCE P-VALUE IN WEEK 1

Soma Roy[1], Allan Rossman[1], Beth Chance[1], George Cobb[2],
Jill VanderStoep[3], Nathan Tintle[4], and Todd Swanson[3]
[1]Department of Statistics, California Polytechnic State University, San Luis Obispo, CA, USA
[2]Mount Holyoke College, South Hadley, MA, USA
[3]Department of Mathematics, Hope College, Holland, MI, USA
[4]Department of Mathematics, Computer Science and Statistics,
Dordt College, Sioux Center, IA, USA
soroy@calpoly.edu

In traditional introductory Statistics courses, statistical inference is often not introduced until the last third of the course, leaving little time for students to develop a strong understanding of the meaning of a p-value. We use simulation and randomization methods to introduce statistical inference early in our introductory Statistics courses, which allows us to discuss concepts of statistical investigations, significance, and p-value in week 1. We start with one-proportion examples to build on students' intuition about "Is the observed result surprising, if both outcomes are equally likely?" Having established the core concept of the logic of inference, we repeat the cycle for situations involving one mean, two proportions, and so on, to contexts involving several proportions, several means, and two quantitative variables. Here we describe the implementation of this approach by showing examples of our student activities, and demonstrating our use of applets to bolster student understanding and learning.

BACKGROUND

In the past two decades, there has been much discussion about "Stat 101," the algebra-based introductory statistics course for non-majors. As a result this course has seen much change in terms of course content, pedagogy, and use of technology. [See, for example, Cobb (1992), Cobb (1994), Moore (1997), Garfield (2000), and Garfield (2002).] As part of this statistics education reform, the question of what topics we teach in Stat 101 and how we teach these has been in the limelight for a while now. Many statistics educators now implement the GAISE [Garfield et al. (2005)] guidelines in their classrooms by always using real data and active learning. In that sense, the present day Stat 101 tends to be much more "modern" than it was twenty years ago.

One aspect, though, that is still quite "traditional" in most introductory statistics courses is the sequencing of topics, in the sense that the first third of the course typically covers descriptive statistics and data collection, and the second third some probability, sampling distributions, and the normal distribution; thus, saving statistical inference until the last third of the course. This sequencing leaves little time for students to develop a strong conceptual understanding of the meaning of p-values, and confidence intervals, and the reasoning process of statistical significance and confidence. Though there has been discussion of the order in which topics should be introduced in Stat 101 in the past, only recently have attempts at changes picked up momentum. For example, Chance and Rossman (2005) introduces statistical inference in week 1 or 2 of a 10-week quarter in a calculus-based introductory statistics course. Malone et al. (2010) discuss reordering of topics such that inference methods for one categorical variable are introduced in week 3 of a 15-week semester, in Stat 101 type courses.

PHILOSOPHY AND APPROACH

In our Stat 101 course, we introduce statistical inference in week 1, using simulation and randomization methods to do so. This is driven by Cobb (2007)'s challenge to statistics educators to make the logic of inference rather than specific techniques based on the normal model as the focus of Stat 101. We use a spiral approach to present the statistical investigation process that begins by asking the research question, moves on to designing the study and collecting data, analyzing the data, and ends with formulating conclusions and suggesting follow-up research questions. We start in the context of simple one-proportion scenarios in week 1, and then repeat the process in new scenarios throughout the term. The philosophy behind this is to expose students to the logic of statistical inference early, and give them time to develop and strengthen their

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2014 ISI/IASE]

understanding as they repeatedly revisit the logic of inference, specifically the concepts of p-value and confidence interval, in new scenarios.

The key to introducing concepts of statistical inference early and often is to adopt a simulation/randomization-based approach. This approach makes use of modern computing power and puts the logic of statistical inference at the center of the curriculum, as advocated by Cobb (2007). Also, this approach is easier to motivate as it does not rely on theoretical distributions and formal discussion of probability. Instead this approach makes use of tactile simulations using everyday objects like coins, dice, and cards, and applet-based simulations to generate sampling/randomization distributions. For every scenario that students encounter in this course, they first learn how to make inferences using simulations of chance models. Later we introduce students to theory-based procedures for statistical inference, as an alternative approximation to the randomization-based methods. For example, when students are introduced to the analysis of quantitative data with the goal of comparing two groups, we first simulate a randomization test to assess the strength of evidence provided by the observed data that the groups differ, and then we proceed to introduce and use a two-sample *t*-test for making that inference (when the necessary conditions for using such a test are satisfied).

Implementing a randomization-based approach requires effective use of technology. After students have carried out a few trials of a tactile simulation, they move on to use technology to implement a large number of repetitions. Rather than ask students to learn to use a statistical software package, we have designed easy-to-use web-applets that conduct all of the simulations and perform all of the analyses presented in this course. A few of the applets come in versions that are tailored to particular studies, while other applets come with "default" data that students can practice with, or they can enter and analyze their own data. The applets have been developed to be self-explanatory, and show students visual representations of what the simulation does; also we have been deliberate about keeping the look and feel of the applets consistent across the various applets, so that the transition from applet to applet is not bumpy. Another nice feature of our applets is that they allow the user to run one trial of simulation/randomization at a time, before jumping to a large number of trials, say 1000. With the advancement that technology has seen since the 1980s, it is only appropriate to use technology to enhance student learning.

COURSE MATERIALS

One of the challenges of teaching this course was that there was no textbook available that teaches the topics in the sequence that we wanted, and alternated between simulation/randomization-based methods and theory-based methods to present statistical inference. So, we started working on developing our own materials with the following key features:

- Use the spiral approach to the statistical investigation process
- Use simulation/randomization-based methods to introduce statistical inference
- Focus on the logic and scope of inference
- Integrate exposition, examples, and active explorations
- Integrate easy-to-use web-based applets for carrying out analyses
- Always use real data from genuine studies that matter.

The sequence in which statistical inference is introduced and revisited in different contexts is follows:

- Single binary variable (inference for a process probability and then a population proportion)
- Single quantitative variable
- Comparing binary variables between two groups (inference for 2×2 table)
- Comparing quantitative variables between two groups
- Analyzing paired data
- Comparing categorical variables across multiple groups
- Comparing quantitative variables across multiple groups
- Association between two quantitative variables

In each context, students first see a simulation/randomization-based approach to analyzing the data, followed by the equivalent theory-based test. We start with simple one-proportion examples to build on students' intuition about "*Is the observed result surprising to happen by* *chance alone, if both outcomes are equally likely?*" Then having established the core concept of the logic of inference, we move on to repeat the cycle in situations involving one mean, two proportions, and so on. Descriptive statistics are introduced as and when the need arises. For example, segmented bar charts are introduced when we start comparing two groups on a categorical response; side-by-side boxplots are introduced when we start comparing two or more groups on a quantitative response. The importance of random sampling is introduced near the beginning, and the distinction between observational studies and randomized experiments is made when we start comparing data from two groups.

SAMPLE EXAMPLES/EXPLORATIONS

One feature of our materials is that every concept is introduced via a complete worked-out example, as well as an exploration – an activity that an instructor could use in class to help students "discover" and learn statistical concepts. This gives the instructor a choice of using an expository-based or activity-based approach to introducing topic in class. Below we present three sample examples/explorations to give you an idea of how we use simulation/randomization-based methods to motivate the evaluation of strength of evidence. The first two concern inference about one proportion, while the last addresses inference for comparing two means.

Sample Example/Exploration 1: Introduction to Chance Models

This is one of the first examples/explorations students in our classes may see. We use this study to motivate the logic of inference, and introduce students to *chance models.* In 1978, researchers Premack and Woodruff published a study in *Science* magazine, reporting an experiment where an adult chimpanzee named Sarah was shown videotapes of eight different scenarios of a human being faced with a problem. After each videotape showing, she was presented with two photographs, one of which depicted a possible solution to the problem. Sarah could pick the photograph with the correct solution for seven of the eight problems.

We begin by asking students what two possible explanations might be for Sarah getting 7 correct answers out of 8. With some guidance students are able to arrive at the two possible explanations being: (1) Sarah was just guessing and got lucky, and (2) Sarah can do better than just guessing. The next question students have to answer is, "Which explanation do you think is a better explanation for Sarah's performance on the given task?" Almost always students pick explanation (2), which leads to a follow-up question such as, "I (the instructor) think explanation (1) is better. How will you convince me that (1) is not a better explanation?" This indirectly emphasizes that the assumption being made is that the null hypothesis is true. Students then discuss ways to refute explanation (1), and suggest looking at what Sarah's results could have been if she just guesses. Almost always students can figure out that tossing a coin is the way to model "just guessing." We talk about why guessing between two choices is equivalent to flipping a coin, and we define "heads" as getting the answer correct, and "tails" as getting the answer incorrect. Of course, the choice of "heads" as correct is arbitrary. Next, we talk about what the expected number of correct answers ("heads") would be if Sarah were just guessing ("flipping a coin") on all 8 questions, and how even though we expect this value to be 4 (half of 8), we know not every set of 8 coin tosses results in 4 heads. Thus, we need to repeat the set of 8 tosses several times to generate the pattern for correct answers ("heads" out of 8 tosses) that can happen in the long-run.

Thus, having established that we need to mimic what happened in the actual study, *but assuming Sarah is just guessing*, we conduct a tactile simulation with each student tossing a coin 8 times to generate possible values for the number of correct answers Sarah could have got (out of 8) just by guessing. Students then combine their findings by making a dotplot of their individual simulation results on the chalkboard. This leads into a discussion of what a dotplot displays, what is the label on the *x*-axis (in this case, the number of correct answers or the number of heads). We also make note of where the graph centers (in this case, 4) and discuss why that should not surprise us. We observe that even though we were tossing (presumably fair) coins, not every set of 8 tosses resulted in 4 heads; there is always chance variability. Next, we make note of where Sarah's actual result (7 correct out of 8) is on this graph. This result is seen to be very rare, and in right tail of the distribution of results that could have been, had Sarah been guessing. Thus, this indicates that Sarah's result is very unlikely (or surprising) to have happened if she were just guessing. Then, we are fairly convinced that Sarah is doing better than just guessing. We also discuss that had Sarah got 5 out of 8 correct, we would not be as convinced of Sarah's ability to do better than guess, because 5 out 8 correct ("heads") is not surprising to happen by chance alone.

Depending on the class size the number of repetitions of the tactile simulation can be fairly low, and this can be used as motivation to use the One Proportion applet, to crank up the number of repetitions to 1000 or even higher to generate the long-run pattern. This applet is designed so that when the chance model involves two equally likely outcomes, the visual is that of coin tosses, and "heads" is defined as a "success." We think doing the tactile simulation first helps students get a better understanding of what the applet is doing. Once the applet has completed the simulation, students use the applet output to address the same kind of questions as before: *Is the observed result of 7 correct out of 8 unlikely to have happened by chance alone? Explain. Consequently, is "just guessing" a plausible explanation for Sarah's results? Explain.*

To wrap the discussion on this example/exploration we point out that to evaluate whether an observed result could have happened by chance alone, we use a chance model (such as coin tossing) to simulate results that could have happened by chance, and compare the observed result to the simulated results. If the observed result is surprising to happen by chance alone, we find ourselves convinced that observed results did not happen by chance alone, and something other than chance is at play.

We are deliberate about starting with a context that involves a relatively simple50-50 scenario for the null model, and one in which the observed result is quite clearly in the tail of the null distribution. The discussion for this first example/exploration is kept straightforward and informal with the aim being to introduce the core logic behind statistical inference.

As follow-up questions, we ask: "What if Sarah had got 14 out of 16 answers correct? Would you be more convinced or less convinced than before that she does better than just guess?" "What if Sarah got 3 out of 8 correct?" "Sarah had been raised in captivity, and had received extensive training using photos and symbols. Based on the results of this study, would it be okay to say that all chimpanzees do better than just guess when presented with problems like the ones presented to Sarah? Explain why or why not."

Sample Example/Exploration 2: Measuring the Strength of Evidence

This example/exploration is representative of what is used to introduce the concept of a p-value. Also, this example moves away from the one-proportion scenario of two equally likely outcomes, to a non 50-50 scenario. Statistician Jessica Utts has conducted extensive analyses of studies that have investigated psychic functioning. One type of study involves having one person (called the "sender") concentrate on an image while a person in another room (the "receiver") tries to determine which image is being "sent." The receiver is given four images to choose from, one of which is the actual image that the sender is concentrating on. Utts (1995) cites research from Ben and Honorton (1994) that analyzed studies using a technique called Ganzfeld. These researchers analyzed a total of 329 sessions, and found that 106 of these sessions produced a "hit" (correct identification of the image being "sent"). The research question then is, *Do these data provide convincing evidence that the "hit" rate in such studies is higher than what would be expected by chance alone?*

Now the coin tossing model doesn't work anymore, because under the chance model the probability of a "hit" is 0.25, and not 0.5. We use this to motivate students to think about other possible ways to model chance, such as, spinners, cards, and dice. Students use the One Proportion applet to generate the null distribution of results (number of "hits" in 329 sessions) that could have happened by chance alone, and then compare the observed number of hits (106) to the null distribution. Students use the applet output to address the same kind of questions as before: *Is the observed result of 106 hits in 329 sessions unlikely (that is, surprising) to have happened by chance alone? Explain. Consequently, is "random chance" a plausible explanation for the results? Explain.*

At this point we introduce the concept of p-value as a measure of "how unlikely (that is, surprising) the observed results are to have happened by chance alone." One way to measure the surprising-ness of the observed results is to find out how often a result as or more surprising than the observed result happened by chance in the simulation. This number expressed as a proportion (out of the number of repetitions) is the *p-value*. The smaller the p-value is, the less often a result as or more surprising than the observed results happened by chance – thus providing strong evidence against the observed results happening due to chance alone.

A natural follow-up to this is the discussion of standardized scores (or *z*-scores*)* as an alternative to finding a p-value. Students find the standardized score corresponding to the observed proportion of successes using the mean and standard deviation of the simulated null distribution that are produced by the One Proportion applet.

Another nice follow-up for this example/exploration is that it is also great to introduce the theory-based (normal distribution) approach to find the p-value, because the sample size is large enough that the distribution of proportion of successes is bell-shaped and "filled-in."

A few things we are deliberate about: for the second example/exploration, we include some formal discussion of the null and alternative hypotheses, and start to move away from 50-50 scenarios. We still stay with one-sided tests, so that introduction to p-values is simple in that it only looks at one tail of the null distribution. The One Proportion applet is designed so that students have the choice of which region of the null distribution (left tail or right tail) they look at to determine the p-value. We believe that making this choice forces students to think about what the p-value measures. Another nice feature of the One Proportion applet is that the visual changes from being a coin to a spinner when the chance model is not a 50-50 model.

Sample Example/Exploration 3: Comparing Two Groups

This is an example of how we introduce inference for comparing two groups on a quantitative response. In a study published in *Nature Neuroscience*, researchers Stickgold, James, and Hobson (2000) investigated the effects of sleep deprivation. Twenty-one volunteers, aged 18 to 25 years, were first trained on a visual discrimination task that involved watching stimuli appear on a computer screen and reporting what was seen. Performance was recorded as the minimum time (in milliseconds) between the appearance of stimuli and an accurate response. Then one group (of 11 participants) was randomly assigned to be deprived of sleep for 30 hours, followed by two full nights of unrestricted sleep, whereas the other group (of 10 participants) was allowed to get unrestricted sleep on all three nights. Following this, both groups were retested on the task. Researchers recorded the *improvement score* as being the decrease in time required at retest compared to training. The unrestricted sleep group's average score was higher than that of the deprived group by 15.92 ms.

To generate the null distribution for the difference in average scores that could have been by random chance alone, students first carry out a tactile simulation. They write down each of the 21 improvement scores on index cards – one score per card, and shuffle and redistribute the scores into two groups – 10 for unrestricted and 11 for deprived, and find the difference in average scores for the shuffled data. This shuffling and redistribution of scores simulates the random assignment of the subjects to treatment groups, under the assumption that treatment has no effect on improvement scores (that is, the responses are fixed). Then, the students combine their findings to create a dotplot of the null distribution, and as before compare the observed results to the null distribution to answer the questions: *Is the observed result of difference in averages of 15.92 unlikely (that is, surprising) to have happened by chance alone? Explain. Consequently, is "random chance" a plausible explanation for the results? Explain.*

Through this example/exploration students see that again the core logic of inference stays the same, only the chance model is formulated differently based on what the context of the study is.

Note: The applets we use can be found at http://rossmanchance.com/ISIapplets.html

ADVANTAGES OF THIS APPROACH

Some advantages of this approach are: does not rely on a formal discussion of probability, and hence can be used to introduce statistical inference as early as week 1; has a lot of opportunity for activity/exploration-based learning; interpreting and evaluating the p-value becomes much easier when students experience it as a long-run probability through simulation; allows one to revisit the entire statistical investigation process over and over again in new contexts. Where in the past we have found students struggle to interpret p-values, with the simulation/randomization approach we have noticed that students find it easier to interpret p-values because all they have to do is describe what the simulation does in order to calculate it.

Another advantage of using the randomization-based approach to introduce statistical inference is that it allows the use of any statistic one may be interested in. In our materials, we have students think of different statistics that can be used to compare groups, and then run randomization-based tests for these statistics. Examples of such statistics are: difference in medians, relative risk, ratio of standard deviations; in the context of comparing multiple groups, examples of statistics that students find more intuitive than the χ^2 statistic or the F-ratio are mean absolute deviations, or sum of pairwise differences. These are statistics for which it would be hard to run theory-based tests, but for which the randomization-based tests are easily extendable.

An additional benefit of this approach, in our opinion, is that this is a much more enjoyable sequence and way in which to teach the Stat 101 topics for many instructors.

CONCLUSION

We believe that it is important to introduce students to the logic of inference early in the term, and revisit the concept often and in new contexts to help them understand that the logic stays the same regardless of context. To accomplish this we use simulation/randomization-based methods, and incorporate tactile simulations as well as simulations using web applets that have been developed to support student learning. The applets were designed keeping in mind that we would like students to not be distracted by or be left confused by a “black box” applet, but see exactly how the applet creates the null distribution and how we can find the p-value from it. Thus, most applets have visuals such as coin flips, spinners, and card shuffles. We use examples and explorations to guide student learning, samples of which are presented above. We hope that with this approach we will be able to help students get a better understanding of statistical inference. It is with this hope that we have developed our course materials.

REFERENCES

Chance, B., & Rossman, A. (2005). *Investigating statistical concepts, applications, and methods*. Duxbury Press.

Cobb, G. (1992). Teaching statistics. In L. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action*. Washington, D. C.: Mathematical Association of America, Notes #22, 3-43.

Cobb, G. (1995). Statistics education: A National Science Foundation conference. *Journal of Statistics Education*, *1*(1).

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum. *Technology Innovations in Statistics Education, 1*(1).

Garfield, J. (2000). *An evaluation of the impact of statistics reform: Year 1 report*. National Science Foundation, REC-9732404.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10*(3).

GAISE College Report (2005). http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf

Malone, C., Gabrosek, J., Curtiss, P., & Race, M. (2010). Resequencing topics in an introductory applied statistics course. *The American Statistician, 64*(1).

Moore, D. (1997). New pedagogy and new content: The case of statistics (with discussion). *International Statistical Review*, *65*, 123-165.

Premack, D., & Woodruff, G. (1978). Chimpanzee problem-solving – Test for comprehension. *Science, 202* (4367).

Stickgold, R., James, L., & Hobson, J. (2000). Visual discrimination learning requires sleep after training. *Nature Neuroscience*, *3*, 1237-1238.

Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1).

Utts, J. (1995). An assessment of the evidence for psychic functioning. *Journal of Parapsychology*, *59*, 289-320. Online at: http://www.ics.uci.edu/~jutts/air.pdf