REASONING ABOUT INFORMAL STATISTICAL INFERENCE: ONE STATISTICIAN'S VIEW

ALLAN J. ROSSMAN
California Polytechnic State University – San Luis Obispo
arossman@calpoly.edu

ABSTRACT

This paper identifies key concepts and issues associated with the reasoning of informal statistical inference. I focus on key ideas of inference that I think all students should learn, including at secondary level as well as tertiary. I argue that a fundamental component of inference is to go beyond the data at hand, and I propose that statistical inference requires basing the inference on a probability model. I present several examples using randomization tests for connecting the randomness used in collecting data to the inference to be drawn. I also mention some related points from psychology and indicate some points of contention among statisticians, which I hope will clarify rather than obscure issues.

Keywords: *Statistical reasoning; Statistical significance; Randomization tests*

1. PRELIMINARY DEFINITIONS

In preparing these comments I began by consulting an online dictionary (dictionary.com), where I found two definitions of "infer" that seem especially relevant to statistical inference:

1. to derive by reasoning; conclude or judge by premises or evidence;
2. to draw a conclusion, as by reasoning.

Similarly, the following two definitions of "informal" struck me as appropriate for this discussion:

1. without formality or ceremony, casual;
2. not according to the prescribed, official, or customary way or manner; irregular; unofficial.

I also consulted a statistics textbook, *The Statistical Sleuth* (Ramsey & Schafer, 2002), in which I read the following definitions:

1. An *inference* is a conclusion that patterns in the data are present in some broader context.
2. A *statistical inference* is an inference justified by a probability model linking the data to a broader context.

Informed by these definitions, I suggest that inference requires going beyond the data at hand, either by generalizing the observed results to a larger group (i.e., population) or by drawing a more profound conclusion about the relationship between the variables (e.g., that the explanatory variable causes a change in the response).

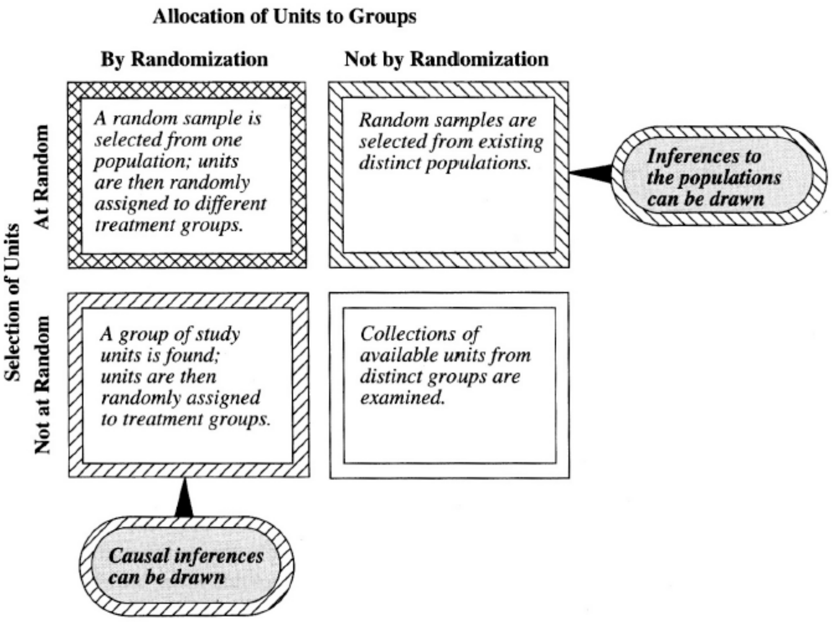
Statistical inference has traditionally been the focus of introductory courses at the tertiary level, and this topic has become more prevalent in the K-12 curriculum. For example, the K-12 GAISE (Guidelines for Assessment and Instruction on Statistics Education) report endorsed by the American Statistical Association (Franklin et al., 2005) argues that by the end of their secondary schooling, students should learn to “look beyond the data.” This GAISE report also emphasizes that these students should understand the nature of “chance variability.”

This notion of *chance* variability is fundamental to drawing statistical inferences. Statisticians deliberately introduce randomness into the process of collecting data, in large part to enable inferences to be made on a probabilistic justification. This randomness takes one of two forms (or both), depending on the research question being addressed:

1. Random *sampling* from a population enables results about the sample to be generalized to the larger population.
2. Random *assignment* of units to treatment groups allows for cause-and-effect conclusions to be drawn about the relationship of the explanatory and response variables.

Figure 1, taken from *The Statistical Sleuth,* summarizes these points.

Display 1.5 Statistical inferences permitted by study designs



Allocation of Units to Groups

Selection of Units	By Randomization	Not by Randomization
At Random	*A random sample is selected from one population; units are then randomly assigned to different treatment groups.*	*Random samples are selected from existing distinct populations.*
Not at Random	*A group of study units is found; units are then randomly assigned to treatment groups.*	*Collections of available units from distinct groups are examined.*

Inferences to the populations can be drawn

Causal inferences can be drawn

Figure 1. Statistical inferences permitted by study designs from Ramsey and Shafer (2002)

2. AN EXAMPLE OF INTUITIVE INFERENTIAL REASONING

Example 0: Funny Dice Beth Chance, inspired by Jeff Witmer, introduced me to the following activity for introducing students to the reasoning of statistical significance. Take to class a pair of dice that appear to be fair and ordinary but are actually not: One of the dice contains only fives on its faces and the other has half twos and half sixes. The dice will therefore produce sums of only seven and eleven. (An internet search for “7 11 dice” will reveal many places to purchase such dice.) Roll the dice, or better yet ask a student to roll them, and call out the sum. Do this repeatedly, and observe the reactions of students in the class as the sevens and elevens accumulate.

Students generally have no reactions to the first two or three rolls. By the time they see a seven or eleven for the fourth or fifth time, some start to snicker or otherwise indicate that the results seem suspicious. By the sixth or seventh roll, many in the class openly voice conviction that the dice are not fair. After the tenth roll, almost all students in the class are convinced (without looking at the faces of the dice) that the dice are not fair. Most students provide a good account of their reasoning process, explaining that it would be extremely unlikely to observe so many results of seven or eleven if the dice were truly fair.

This reasoning process, which seems to come very naturally for students, is a classic example of Fisherian inductive reasoning. Students are assessing the strength of evidence against a claim. They do this by determining how unlikely the observed result would be, if in fact the claim being tested were true. Of course, all of this happens not just informally but intuitively. If I impose more structure here, I assert that students' intuitive reasoning process in this example involves:

- Starting with an unspoken belief that the dice are fair (we could call this the null model or null hypothesis);
- Evaluating that the observed data (nothing but sevens and elevens) would have been very unlikely if that belief (null model) were true (intuitively calculating a p-value);
- Rejecting the initial belief (null model) based on the very small p-value, rather than believe that a very rare event has occurred by chance alone.

3. AN EXAMPLE OF INFORMAL INFERENTIAL REASONING

Example 1: Toy Preference A recent study (Hamlin, Wynn, & Bloom, 2007) investigated whether infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction. In one component of the study, 10-month-old infants were shown a "climber" character (a piece of wood with "google" eyes glued onto it) that could not make it up a hill in two tries. Then they were shown two scenarios for the climber's next try, one where the climber was pushed to the top of the hill by another character ("helper") and one where the climber was pushed back down the hill by another character ("hinderer"). Each infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer) and asked to pick one to play with. Preferences were recorded for a sample of 16 infants, with 14 choosing the helper toy.

Clearly more than half of these infants chose the helper toy, but the inferential question is whether this result provides evidence of a *genuine* preference, either among a larger population of infants or among these same 16 infants if they were to be tested repeatedly. When asked for their initial impressions, students give widely varying reactions. Some are willing to conclude a genuine preference merely because more than half chose the helper; others argue that they would remain unconvinced about a genuine preference even if all 16 chose the helper toy because of what they perceive as a prohibitively small sample size.

Making a statistical inference requires a probability model. Fortunately, a simple model presents itself, one that is both familiar and understandable at the school level. Under the null model that infants have no genuine preference, we can model their selections as flips of a fair coin. In this manner we can simulate the selections by 16 infants over and over again, in order to assess how surprising it would be to obtain 14 or more of them choosing the helper in a sample of 16 infants if there were, in fact, no genuine preference. Asking students in a class to conduct 16 coin flips and count the number of heads, we quickly find that it is quite unusual to obtain 14 or more heads. Repeating this process 1000 times produces results like those in Figure 2.

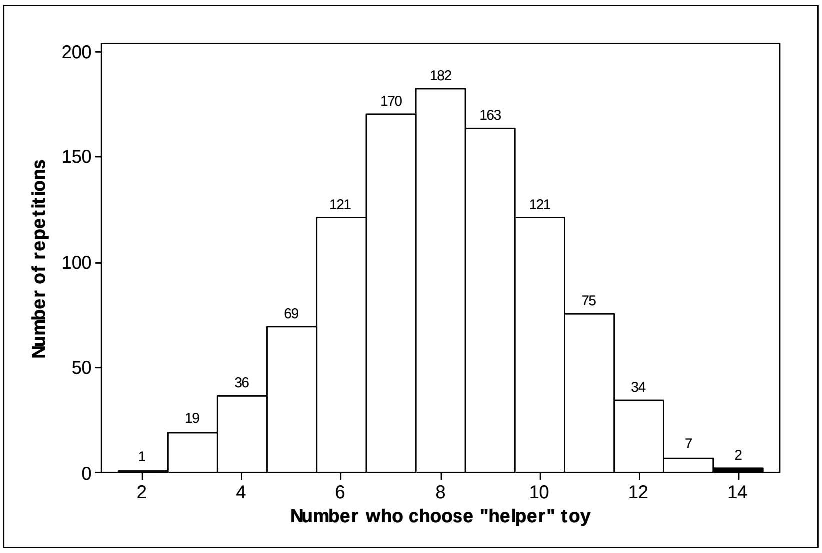


Figure 2. Simulation results for helper toy study

Notice that 14 chose the helper toy in 2 of these 1000 simulated repetitions. This graph therefore reveals that it is not impossible to find 14 or more choosing the helper toy even when there is no genuine preference, but such a result is very unlikely. Accordingly, we have strong evidence that infants genuinely do tend to prefer the helper toy over the hinderer.

This reasoning process is identical to that with the 7/11 dice, but it does not come as naturally to students. With the dice, intuition correctly tells us that it is very unlikely to get a string of exclusively sevens and elevens with fair dice. But our intuition is much less reliable for knowing the distribution of coin flip results. Simulation enables us to estimate the probability distribution and the p-value from this study. This activity and analysis are amenable to use with schoolchildren as well as college students.

But is this inferential analysis informal? I contend that it is. We are not doing a formal calculation of an exact p-value, which we could do using the binomial distribution. We are also not calculating a test statistic or approximate p-value based on a normal distribution. But we are using a reasonable process, based on a probability model, to draw an inference beyond the data at hand.

What other reasoning would I like students to think about, and begin to learn about in this context? Three issues, in order of increasing conceptual difficulty:

1. Students should recognize the key role that the sample proportion of successes plays in this inferential reasoning process. For example, if only 10 of the 16 infants had chosen the helper toy, this would provide much weaker support for concluding that infants have a genuine preference for the helper. Why? Because a result as extreme as 10 or more successes would not be at all surprising under the null model of no genuine preference (as the above histogram shows).
2. Students should come to appreciate the important role played by sample size. Ask about a different (hypothetical) study in which 100% choose the helper toy: Does that provide strong evidence of a genuine preference? Well, not if the study only

involved two infants. What if 60% choose the helper- is that statistically significant? Not if the study involved 10 infants, but yes if the study involved 100 infants.

3. Students should eventually learn to use the same reasoning process to investigate claims beyond a 50/50 null model. For example, do the study results provide evidence that infants actually prefer the helper by more than a 2-to-1 ratio over the hinderer? The reasoning process is the same, but the simulation needs to use a 2/3 probability of each infant's choosing the helper.

4. INFORMAL INFERENCE WITH RANDOMIZATION TESTS

Example 2: Dolphin Therapy Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility, researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression (Antonioli & Reveley, 2005). Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group did so in the presence of bottlenose dolphins and the other group (outdoor nature program) did not. At the end of two weeks, each subjects' level of depression was evaluated, as it had been at the beginning of the study. The results are summarized in Table 1 and Figure 3.

Table 1. Results of dolphin therapy experiment

	Dolphin therapy	Control group	Total
Showed substantial improvement	10	3	13
Did not show substantial improvement	5	12	17
Total	15	15	30

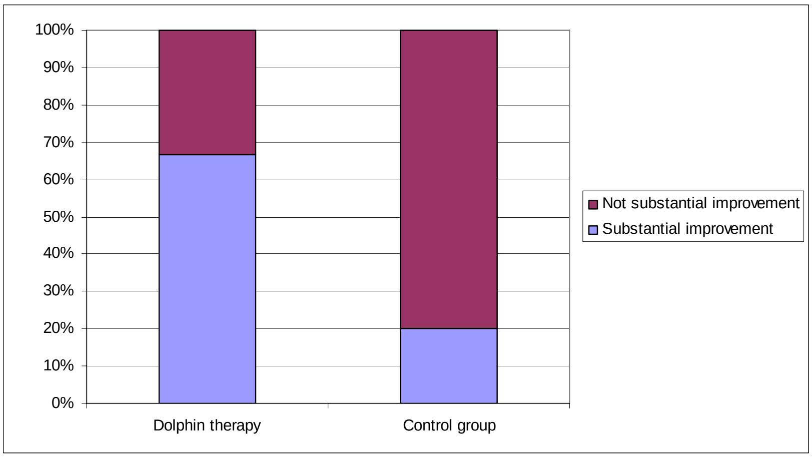


Figure 3. Results of dolphin therapy experiment

Clearly the dolphin therapy group had a larger success rate than the control group (66.7% vs. 20.0%). Can we reasonably *infer* that the dolphin therapy really is more effective than the control? To address this key question we must consider the role of chance variability.

The randomness in this study arises from researchers randomly assigning the 30 subjects to one of the treatment groups. Is it possible that this randomization process alone, even if dolphin therapy were no more effective than the control, would have produced results as extreme as the researchers found? Sure, it's possible. But is that possibility so unlikely that it discredits that explanation?

We could proceed directly to a probability calculation to assess this unlikeliness. But with introductory students I recommend investigating this with simulation. Students can simulate this randomization process by taking 30 playing cards, marking 13 to represent those who showed substantial improvement and the other 17 to represent those who did not improve substantially. Then shuffle the cards and randomly deal out 15 to be in the dolphin therapy group with the other 15 in the control group. Note that this shuffling/dealing process simulates the random assignment process actually used by the researchers to put subjects in treatment groups. Also note that this simulation process assumes that there is really no benefit of the dolphin therapy, because it assumes that the 13 subjects who improved were going to improve regardless of which group they were assigned to. Then observe the results of the simulated random assignment, either by calculating the difference in success proportions between the two groups or simply by noting the number of "successes" in the dolphin therapy group. Figure 4 shows the results of 1000 simulated random assignments.

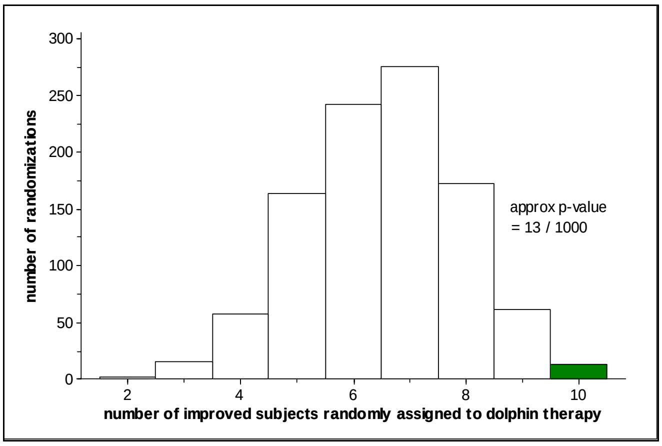


Figure 4. Simulation results from dolphin therapy experiment

In only 13 of the 1000 random assignments did the simulated result turn out to be as extreme as the actual experimental result (10 or more successes in the dolphin therapy group). So, it is indeed possible to have obtained such an extreme result by chance alone, even if the dolphin therapy had no effect, but this simulation reveals that this possibility is fairly unlikely. We therefore conclude that the experimental data provide fairly strong evidence that dolphin therapy really is more effective than the control.

The exact probability can be calculated to be 0.0127, to four decimal places. This procedure is known as Fisher's Exact Test. This is an example of a type of inference procedure called a randomization test. One advantage of this procedure for introducing introductory students to the reasoning process of statistical inference is that it makes clear the connection between the random assignment in the design of the study and the inference procedure. It also helps to emphasize the interpretation of a p-value as the long-term proportion of times that a result at least as extreme as in the actual data would have occurred by chance alone under the null model. For an overview of randomization tests, see Ernst (2004).

Notice that the reasoning process here is again the same as with the 7-11 dice, and also as with the helper/hinderer toy. But students seem to struggle more to understand it in this context. There are two possibilities (dolphin therapy is more effective than control, or it is not). The experimental data would be very unlikely to occur if dolphin therapy was not more effective, so we have strong evidence to reject that explanation and conclude that dolphin therapy really is more effective than the control. But of course this conclusion is not definitive, and there remains a possibility that dolphin therapy is not more effective and the researchers just happened to witness a rare event. Many students seem to be troubled by this lack of certainty in their conclusion, more so than with the 7-11 dice.

What else do I want students to learn about the reasoning of statistical inference in such settings? Similar to my list following Example 1, I want students to appreciate and understand the importance of how different the success rates are between the two groups, and also the importance of the numbers of subjects in the two groups.

Example 3: Murderous Nurse For several years in the 1990s, Kristen Gilbert worked as a nurse in the intensive care unit (ICU) of the Veteran's Administration hospital in Northampton, Massachusetts. Over the course of her time there, other nurses came to suspect that she was killing patients by injecting them with the heart stimulant epinephrine. Part of the evidence against Gilbert was a statistical analysis of more than one thousand 8-hour shifts during the time Gilbert worked in the ICU (Cobb & Gelbach, 2005). Table 2 and Figure 5 display the data.

Table 2. Data from Kristen Gilbert trial

	Gilbert working on shift	Gilbert not working on shift	Total
Death occurred on shift	40	34	74
Death did not occur on shift	217	1350	1567
Total	257	1384	1641

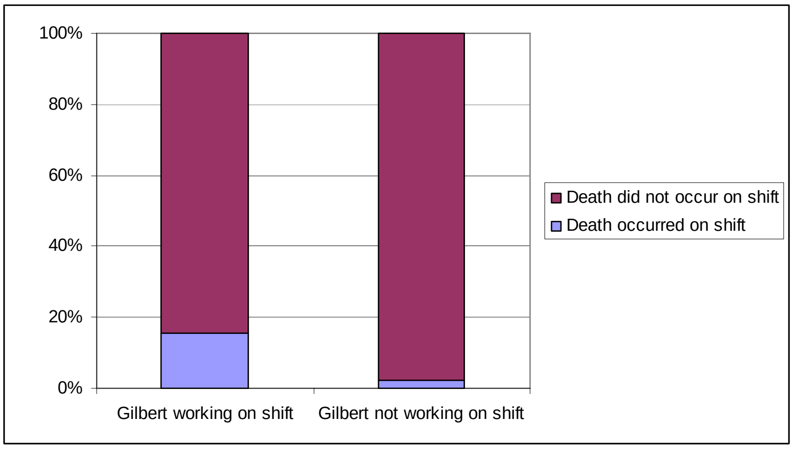


Figure 5. Results from Kristen Gilbert trial

As with the previous two examples, this study involves comparing two groups, with data presented in a 2×2 table, and investigating a conjecture that one group would "do better" on the response than the other. But a big difference is that, unlike the dolphin therapy experiment, this is an observational study in which no random assignment occurred. With this lack of randomness in the data production phase, some statisticians would argue that no randomization test should be performed here. But others would contend that we can still conduct a randomization test to assess whether the observed difference in death proportions (0.156 on a Gilbert shift vs. 0.025 otherwise) is large enough to infer that random variation is not a reasonable explanation. The p-value turns out to be less than 1 in a trillion, which effectively rules out "luck of the draw" as an argument for Gilbert's defense.

Because this is an observational study, however, we cannot conclude that Gilbert's presence on the shift is the *cause* of the higher death rate. The significance test does not rule out the possibility that other (confounding) variables may have differed between Gilbert shifts and no-Gilbert shifts. For example, without examining more detailed data, it is possible that Gilbert might have worked shifts during a particular time of day when deaths were more likely to occur.

This distinction in scope of conclusions that can be drawn from randomized experiments as opposed to observational studies is a key aspect of inference that I expect all students to learn. The GAISE guidelines for the K-12 curriculum include this distinction for secondary students.

5. SOME CONSIDERATIONS FROM PSYCHOLOGY

Why is this reasoning process difficult for many people (including, of course, students)? Part of the answer surely rests in all of the research that has shown how difficult probabilistic reasoning is for people (Nickerson, 2004). I particularly like how Keith Stanovich phrases this issue in his book *How to Think Straight About Psychology* (2007). Stanovich refers to probabilistic reasoning as the "Achilles heel of human cognition." One example is that the concept of a statistical *tendency* is much more difficult for people to grasp than a deterministic relationship. Also, human beings are generally not comfortable with "luck of the draw" as an explanation; we tend to ascribe deterministic explanations to chance phenomena and tend not to consider variability in general, and chance variation in particular.

Notice also that the reasoning process of Fisherian inductive inference is related to a modus tollens argument in logic, but with a probabilistic aspect thrown in for good measure. Rethinking the "7/11 dice" example in these terms, we reason as follows:

1. If the dice were fair, it would be extremely surprising to observe a long string of sevens and elevens.
2. We observe a long string of sevens and elevens.
3. Therefore, we have extremely strong evidence that the dice are not fair.

In this particular context students seem to apply the reasoning process effectively and intuitively. But they often struggle to apply the reasoning process in less familiar contexts, and they often struggle mightily to understand it in the abstract. This bears many similarities to the well-known logic problem known as the Wason selection task (Wason & Johnson Laird, 1972).

The abstract version of the Wason task presents subjects with four cards that have a letter on one side and a number on the other. Subjects are then told the following rule: Every card with a vowel on one side has an even number on the other side. The four cards shown reveal an A, a B, a six, and a seven. Subjects are then asked which cards should be turned over in order to detect whether the rule has been violated. Studies show that a small percentage of people choose the correct answer, which is to turn over the A and the seven cards (Wason & Johnson Laird, 1972). Most people select the six rather than the seven card, failing to realize that the rule will be violated if the seven reveals a vowel on the other side (by a modus tollens argument). But when this same task is presented in concrete terms that are familiar to the subject, most people do indeed answer correctly. For example, suppose the rule is that all people who drink alcohol must be at least 21 years old. Consider four people: a 30-year-old, an 18-year-old, a beer drinker, and a soda drinker. Most subjects have no difficulty in realizing that they should check on how old the beer drinker is and on what the 18-year-old is drinking. The logical structure of this problem is identical to that with the cards, but the concreteness and familiarity make it much easier to solve correctly.

One of my points in mentioning the Wason task is that a modus tollens argument is hard for people to make intuitively, so the reasoning of statistical significance, which invokes a modus tollens-like argument with uncertainty thrown in, is all the more daunting. But my larger point is that students can apply this reasoning process for themselves if we start with concrete examples in a familiar setting.

6. COBB'S 3Rs

George Cobb (2007) refers to the randomization test approach to statistical inference as the 3Rs:

- Randomize data production.
- Repeat by simulation to see what's typical (and what's not).
- Reject any model that puts your data in its tail.

Cobb argues that introductory statistics students have a better chance of understanding the core logic of inference if it is presented in this manner as opposed to a more conventional approach based on calculations from normal-based probability distributions. Cobb writes: "Our curriculum is needlessly complicated because we put the normal distribution, as an approximate sampling distribution for the mean, at the center of our curriculum, instead of putting the *core logic of inference* at the center." While Cobb is referring to the introductory curriculum at the tertiary level, Scheaffer and Tabor (2008) advocate teaching statistical inference at the secondary level through this process of simulating randomization tests. Chance and Rossman (2006) adopt this approach in a tertiary course for mathematically inclined students.

The three examples discussed above have all involved categorical variables, but this 3Rs approach to statistical inference can also be applied with a quantitative response variable. Scheaffer and Tabor (2008) include such an example, as do Ernst (2004) and Cobb (2007). I prefer giving students experience with a categorical response variable first, because the quantitative response involves several complicating factors. One is that summarizing the difference between the groups is less straightforward; for example, you could use the difference in group means, or the difference in group medians, or some other statistic. Another complication is summarized by Wild (2006), who writes: "Assessment of 'significance' balances three factors—effect size, variability and sample size—in a very complicated way." By starting with categorical variables, we eliminate the variability issue because effect size and sample size are the only relevant factors.

7. POINTS OF CONTENTION, ALTERNATIVE APPROACHES

I should admit that the examples above involve some thorny issues on which statisticians disagree. In Example 2 about dolphin therapy, one question is why keep both margins (not only the number of subjects in each group but also the number who improved and did not improve) fixed when conducting the simulation. Lehmann (1993) points out that, although Fisher favored this analysis, others have criticized this procedure for being too conservative and having low power. Another tricky question is why to calculate the p-value by considering results *more extreme* than the actual results when calculating the p-value. The common answer is that with a large sample size, any one particular outcome is bound to have a small probability. (Imagine flipping a fair coin 10,000 times; obtaining 5000 heads is the most likely result but has probability only 0.008.) Bayesian statisticians do not accept this practice, however; a Bayesian analysis (discussed below) conditions on the data observed, not on data that were not observed.

The examples above all illustrate the Fisherian approach to statistical inference (1925, 1935a, 1935b). Fisher's approach emphasizes the strength of evidence provided by observed data against a null model. This strength of evidence is captured in the p-value, which measures the probability of having obtained such an extreme result (or more extreme) if the null model were true.

An alternative approach associated with Neyman (1935, 1955) adopts a more mathematical viewpoint. This view regards statistical inference as principally concerned with making a decision between competing hypotheses. The decision procedure is chosen optimally by specifying some condition on the two error probabilities. Typically this condition is to set the desired probability of type I error (rejecting a true null hypothesis), universally denoted by α.

Table 3 summarizes the different perspectives and emphases of the Fisher and Neyman approaches to statistical inference. In the last row of this table I suggest that the Fisherian approach is closer to informal inference and Neyman's to formal inference. As my earlier examples attest, I favor introducing students to the Fisherian approach.

Table 3. Comparing Fisher's and Neyman's perspectives on statistical testing

Fisher	Neyman
Significance testing	Hypothesis testing
Null model	Competing hypotheses
Strength of evidence	Error probabilities
Inductive inference	Inductive decision
p-value	α-level
Data-based	Mathematics-based
Informal inference?	Formal inference?

Are there implications of this contention for introductory teaching and learning? Hubbard and Bayarri (2003) contend that many teachers, authors, and researchers do not recognize and appreciate the differences between these approaches. They write: "Because statistics textbooks tend to anonymously cobble together elements from both schools of thought, however, confusion over the reporting and interpretation of statistical tests is inevitable." In his discussion of this article, Carlton (2003) argues that students "can handle both approaches" and suggests that α levels be introduced as prespecified thresholds for determining whether a p-value is small enough to constitute convincing evidence against the null model. Lehmann (1993) offers that the Fisher and Neyman perspectives are more compatible than others have realized. See Salsburg (2001) and Lehmann (2008) for readable accounts of the Fisher-Neyman dispute.

A third perspective takes a very different approach to statistical inference. All of the examples and discussion above, including both the Fisher and Neyman perspectives, adopt the classical (sometimes called frequentist) approach to statistical inference. Adherents of the Bayesian viewpoint adopt a subjectivist view of probability as measuring personal degree of belief in the proposition being considered. In the 7/11 dice example, a Bayesian would start with a prior probability that the dice are fair, before they are even rolled. Then the Bayesian updates this probability as the dice rolls are observed, using Bayes' Theorem as the mechanism. The result then is a conditional probability, given the observed data, that the dice are fair.

For example, suppose that you start by believing very strongly that the dice are fair, let's say with a probability of 0.99 (and with a very small 0.01 prior probability that the dice produce only sevens and elevens). Then after five consecutive rolls resulting in seven or eleven, Bayes' Theorem calculates that the updated (conditional) probability that the dice are fair becomes 0.051. In other words, those five rolls serve to reduce your belief that the dice are fair from a prior probability of 0.99 to a new probability of 0.051. After eight rolls of seven or eleven, your probability that the dice are fair drops to 0.0006. These are based on starting with a very high prior probability (0.99) that the dice are fair. But if instead you start with only a 0.5 probability that the dice are fair (so prior to seeing any rolls you think it's equally likely that the dice are fair or not), then the updated probability that the dice are fair drops to 0.0005 after five rolls of seven or eleven, and this probability falls all the way to 0.0000006 after eight such rolls.

Bayesians contend that talking of the probability that the dice are fair is very natural and interesting, yet this probability is nonsensical to a classical statistician: The dice are either fair or not; there's nothing random about that, so the classical statistician cannot assign a probability to that proposition. The only probability that can be determined by a classical statistician is the probability of obtaining such extreme results if a pair of fair dice are rolled repeatedly.

Proponents of the Bayesian approach cite many advantages for it. One is that it seems to correspond with how people actually reason. Another is that it results in probability statements about the null model being tested and for the parameter being estimated. Instructors of introductory statistics cringe when students interpret a p-value as a probability that the null model is true, or interpret a confidence interval by saying that there is a 95% chance that the parameter is within the interval. These statements are not only wrong but nonsensical from a classical approach, but they are quite appropriate and accurate from a Bayesian perspective.

A third advantage is that in many cases there truly is prior information that is relevant to making an inference. For example, suppose I tell you that I observed 8 members of a profession and saw that 4 were men and 4 were women. What would you infer about the overall proportion of women in that profession, if I tell you nothing else? But then what if I tell you that I am talking about mechanical engineers—would this information, and your prior knowledge about the relatively small proportion of engineers who are women, affect your inference? Or what if you learn that the occupation in question is pilots, or flight attendants? I suspect that you would draw quite different inferences from the same data in these situations, and quite appropriately so, based on your prior knowledge, or at least impressions, of the proportion of women in those various professions.

Another point of contention has emerged in the past decade, with a growing movement arguing that significance testing has been overused and misused, often serving as a substitute for thoughtful analysis, particularly in the social sciences. Harlow, Mulaik, and Steiger (1997) edited a collection of essays with the provocative title *What if There Were No Significance Tests?* A new book by economists Ziliak and McCloskey (2008) has the even more provocative title *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Two of the principal complaints lodged against significance testing are that with a large enough sample size, nearly all null models are rejected, and statistical significance does not necessarily imply practical significance. These critics often recommend estimating effect sizes as a replacement for assessing statistical significance.

Even proponents of significance testing admit the importance of the other major concept of statistical inference: estimating with confidence. The next section addresses this aspect of statistical inference.

8. INFORMAL REASONING ABOUT INTERVAL ESTIMATION

Example 4: Kissing couples Most people are right-handed and even the right eye is dominant for most people. Molecular biologists have suggested that late-stage human embryos tend to turn their heads to the right. German bio-psychologist Onur Güntürkün (2003) conjectured that this tendency to turn to the right manifests itself in other ways as well, so he studied kissing couples to see if they tended to lean their heads to the right while kissing. He and his researchers observed couples in public places such as airports, train stations, beaches, and parks. They were careful not to include couples who were holding objects such as luggage that might have affected which direction they turned. For each couple observed, the researchers noted whether the couple leaned their heads to the right or to the left. Of the 124 couples observed, 80 leaned to the right. Does this sample provide evidence that more than half of all kissing couples lean to the right? In light of the sample data, what proportion of the population might lean to the right?

We'll treat this sample as if it were a random one from the population of all kissing couples. As with the helper/hinderer study, we'll simulate 1000 repetitions of 124 couples that are equally likely to lean right or left. Figure 6 displays the results of one such simulation.

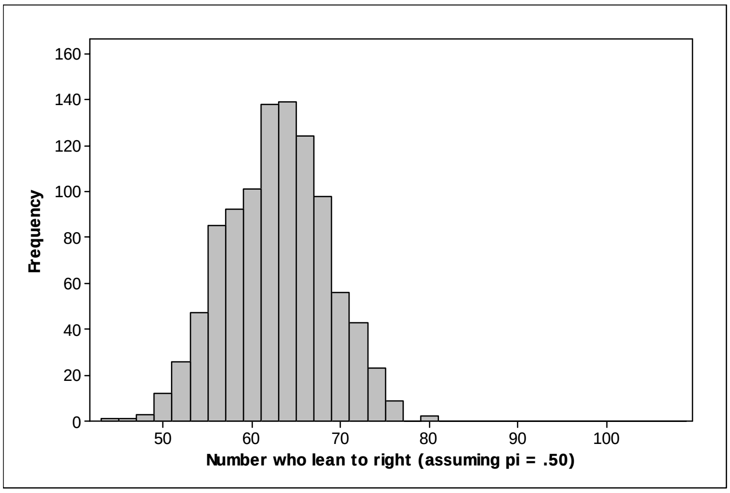


Figure 6. Simulation results for the kissing study

Notice that the observed result (80 couples who lean to the right) is way out in the tail of this empirical sampling distribution. The sample data therefore provide very strong evidence to reject that couples are equally likely to lean to the right or left. We have very strong evidence that kissing couples do indeed tend to lean to the right.

But the natural follow-up question is: How much more than half lean to the right? In other words, what proportion of the population of all kissing couples leans to the right? We can investigate the plausibility of values other than 0.5 by repeating the simulation analysis with those values. Our strategy remains the same: Reject any value of the population proportion that puts the observed data in the tail of its sampling distribution.

Figure 7 displays the results from six different simulations of 1000 repetitions each, changing the value of the population proportion each time.

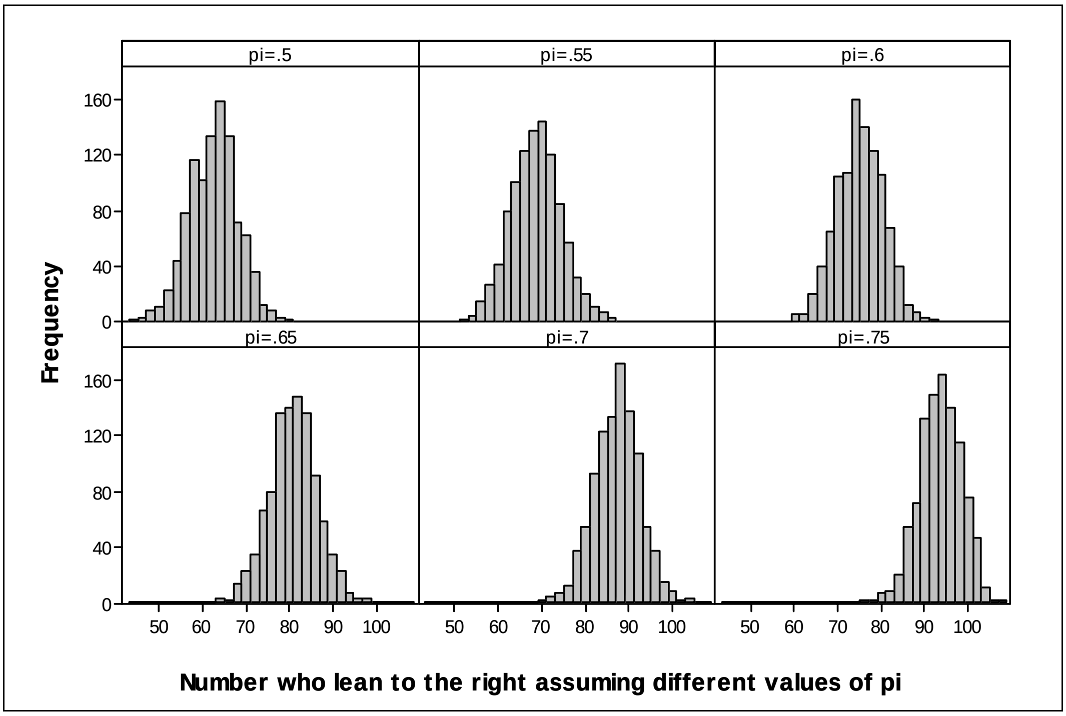


Figure 7. Comparing multiple models via simulation for the kissing study

Notice that the observed result (80 of 124 couples leaning to the right) is surprising (in the tail of the distribution) when the population proportion equals 0.5, 0.55, and 0.75, but not surprising when it equals 0.6, 0.65, and 0.7. Therefore, we include the values 0.6, 0.65, and 0.7 as plausible values of the population proportion who lean to the right. A more thorough analysis reveals an interval of plausible values (using a 5% level of significance) to be from about 0.56 to 0.73. This Fisherian approach to interval estimation is very different from calculating a confidence interval with a formula based on the normal distribution, such as: $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$. This simulation approach strikes me as a more informal method that is likely to involve and increase students' reasoning abilities.

9. CONCLUSION

I suggest that simulation of randomization tests provides an informal and effective way to introduce students to the logic of statistical inference. One advantage of this strategy is that it emphasizes the key role played by chance variation in statistical inference. During the SRTL-5 conference, Tim Erickson observed that asking this "what could have happened if the experiment/sampling had been repeated?" question is paramount in statistical inference. Harradine (2008) provided similar ideas and activities for introducing students to this issue. As the GAISE report suggests, understanding this reasoning process should be attainable by students at the secondary as well as tertiary levels.

I have emphasized categorical variables in the examples, in part because I think such variables provide a simpler context in which students can focus on key ideas of inference. I also suggest that the secondary curriculum often underutilizes categorical variables. I also worry that the secondary curriculum pays far more attention to sampling contexts rather than experimental studies, and I propose that more should be done with experimental studies and activities at this level.

ACKNOWLEDGEMENTS

I thank Joan Garfield and Dani Ben-Zvi for their vision in organizing the SRTL forums and for inviting me to speak to the group. Thanks also to Janet Ainley and David Pratt for providing such a hospitable environment for a very thought-provoking conference, as well as for editing this special issue. More thanks to all of the SRTL participants for their insightful presentations and discussions. Special thanks to Beth Chance for extensive and very helpful suggestions on early drafts of my remarks.

Many of the ideas presented here have emerged from a curriculum development grant from the National Science Foundation (NSF-DUE-CCLI #0633349), through very productive collaborations with Beth Chance, George Cobb, John Holcomb, and Matt Carlton.

REFERENCES

Antonioli, C., & Reveley, M. (2005). Randomized controlled trial of animal facilitated therapy with dolphins in the treatment of depression. *British Medical Journal, 331*(7527), 1231-1234.

Carlton, M. (2003). Comment on "Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing." *The American Statistician*, *57*, 179-181.

Chance, B., & Rossman, A. (2006). *Investigating statistical concepts, applications, and methods*. Belmont, CA: Cengage.

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1).
[Online: http://repositories.cdlib.org/uclastat/cts/tise/]

Cobb, G., & Gelbach, S. (2005). Statistics in the courtroom. In R. Peck, et al. (Eds.), *Statistics: A guide to the unknown*. Belmont, CA: Thomson.

Ernst, M. (2004). Permutation methods: A basis for exact inference. *Statistical Science, 19*, 676-685.

Fisher, R. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R. (1935a). The logic of inductive inference. *Journal of the Royal Statistical Society, 98*, 39-54.

Fisher, R. (1935b). Statistical tests. *Nature*, *136*, 474.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., et al. (2005). *Guidelines for assessment and instruction in statistics education (GAISE) report: A Pre-K-12 curriculum framework.* Alexandria, VA: American Statistical Association.
[Online: http://www.amstat.org/education/gaise/]

Güntürkün, O. (2003). Adult persistence of head-turning asymmetry. *Nature, 421*, 711.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*, 557-560.

Harlow, L., Mulaik, S., & Steiger, J. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Harradine, A. (2008, July). *Birthing big ideas in the minds of babes*. Paper presented at the IASE/ICMI Roundtable Conference, Monterrey, Mexico.

Hubbard, R., & Bayarri, M. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *The American Statistician, 57*, 171-178.

Lehmann, E. (1994). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88*, 1242-1249.

Lehmann, E. (2008). *Reminiscences of a statistician: The company I kept*. New York: Springer.

Neyman, J. (1935). Discussion of "Logic of inductive inference." *Journal of the Royal Statistical Society, 98*, 74-75.

Neyman, J. (1955). The problem of inductive inference. *Communications in Pure and Applied Mathematics, 8*, 13-46.

Nickerson, R. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ramsey, F., & Schafer, D. (2002). *The statistical sleuth: A course in methods of data analysis* (2nd ed.). Belmont, CA: Duxbury Press.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: W. H. Freeman.

Scheaffer, R., & Tabor, J. (2008). Statistics in the high school mathematics curriculum: Building sound reasoning under uncertainty. *Mathematics Teacher, 102*(1), 56.

Stanovich, K. (2007). *How to think straight about psychology* (8th ed.). Upper Saddle River, NJ: Allyn & Bacon.

Wason, P., & Johnson Laird, P. (1972). *Psychology of reasoning: Structure and content*. London: Batsford.

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal, 5*(2), 10-26.
[Online: www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Wild.pdf]

Ziliak, S., & McCloskey, D. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

ALLAN J. ROSSMAN
Department of Statistics
Cal Poly
San Luis Obispo, CA 93407