INFORMAL INFERENCE: A SCIENCE CONNECTION

Tim ERICKSON
Epistemological Engineering
Oakland, California, USA

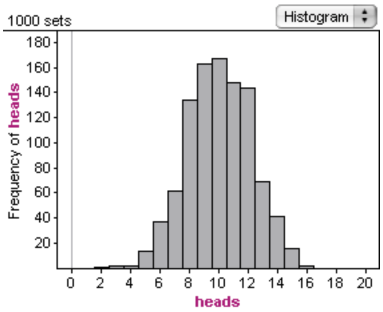
ABSTRACT

We compare and contrast statistical inference with scientific inference, looking for perspectives and insights to help us improve instruction in this difficult topic.

Let's begin with a thought-experiment that involves informal inference in statistics:

You flip a coin 20 times. Suppose all 20 flips are heads. For most people, this is reason enough to doubt the fairness of the coin. If you had gotten 10 heads—ignoring order—most people would say that the coin appears to be fair. Even if you had gotten 11 or 12 heads, most people would say the same. After all, you won't always get exactly 10. But somewhere between 12 and 20 heads, you go from acceptance to doubt.

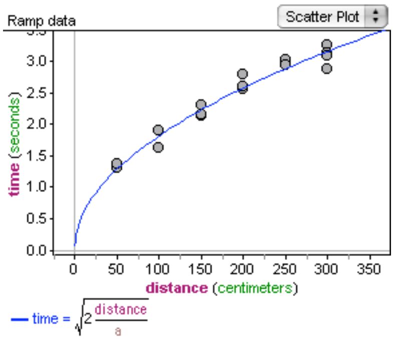
Now suppose you get 16 heads in 20 flips. Is the coin fair? In making your decision, you *infer* about the coin based on observations of the coin flips. We might go so far as to simulate flipping a fair coin 20 times, repeatedly, and see that 16 heads is unusual:



Let's compare this to inferential reasoning in science. Suppose we have this conjecture:

Objects rolling down a ramp accelerate uniformly.

In a typical school science lab to test this statement, students roll a ball down a ramp, timing it from various heights. The resulting data should show a quadratic relationship between distance and time—and they generally do:



From the point of view of a statistics teacher, these are two problems from different parts of the course: the coin problem is from binomial probability and the introduction to hypothesis testing, while the rolling ball problem is from the unit on fitting curves. But we are not interested here in finding the best value for the parameter (in this case, the acceleration *a* under the square root). Instead, we want to assess whether the conjecture that gave rise to the graph is correct. Do rolling objects accelerate uniformly? What do the data tell us about that?

This follows the hypothetico-deductive model for how science works, as described by Popper (1968). In broad strokes: Scientists observe phenomena and generate hypotheses. They imagine the consequences of a hypothesis, and design experiments to see whether the consequences do in fact occur. If they do not, this will (eventually, allowing for error and other possibilities) *falsify* the hypothesis, and force us to discard it.

Applying this to the coin situation makes sense: we have a hypothesis—the null hypothesis—that the coin is fair. We design an experiment (flip it 20 times) to test that hypothesis. We predict what will happen in our experiment if the hypothesis is true. And finally, when our result does not match with prediction, we reject that null hypothesis.

How are these two situations the same, and how are they different? We will begin with logical similarities, but let us foreshadow one of the big differences between statistics and science: In statistics we usually try to show that the relevant hypothesis—the null hypothesis—is *false*. In science, we usually try to show that it is *true*.

UNDERLYING LOGIC

Not only do science and statistics share some language (e.g., *hypotheses*), but they share the same underlying logic as well. This derives from the classical *modus tollens*: If I know that **P** implies **Q**, and I observe that **Q** is false (not-**Q**, written ¬**Q** here), I can conclude that **P** is false as well. For example, if I accept the conditional (if-then) statement,

If it is raining, then the streets are wet.

and I observe that the streets are dry, I can conclude that it is not raining.

Symbolically,

$$\mathbf{P} \rightarrow \mathbf{Q}$$
$$\neg\mathbf{Q}$$
$$\therefore \neg\mathbf{P}$$

This maps directly onto our statistics example. The logic of the hypothesis test rests on a conditional statement such as:

If the coin is fair, only results between 5 and 15 heads are plausible.

When we see 16 heads, we conclude that the coin is not fair. To be sure, there is the chance that we are wrong (and commit a Type I error); variation pierces the watertight logic of *modus tollens*—but the pattern of reasoning still holds. Mapping this logic into a science context works as well, but is more subtle:

If rolling objects accelerate uniformly, and if I make good measurements, then my distance data will be proportional to the square of the time.

If my results don't show that quadratic pattern, there are two possibilities: either the ball does not accelerate uniformly, or the measurements are not good. In either case, the "conjunction" is false.

Let us move on to another similarity.

PROVING HYPOTHESES AND ACCEPTING THE NULL

In both fields, we warn students that certain tempting statements are incorrect. In statistics, we do not "accept the null hypothesis." Instead, we insist that students *fail to reject* the null. Similarly, in science, we don't want students to say that they proved their hypothesis, but rather that the data are *consistent* with it.

The correct language helps students avoid a fallacy related to the underlying logic:

If I observe that the streets are wet (**Q**), I cannot conclude that it is raining (**P**); that would be the fallacy of *affirming the consequent*. In the statistics case, getting 10 or 11 heads out of 20 (i.e., **Q**) doesn't show that the coin is fair. In science, the fact that the ball data was quadratic doesn't prove that the data follow the same pattern between measurements.

Does such a result have any use? Yes. Common sense (or a Bayesian perspective, or Occam's Razor) comes into play: if the streets are wet, the idea that it might be raining gains more weight. In addition, the scientific community values alternative hypotheses. We imagine their consequences and test them. For example, you could suggest a hypothesis involving a water-balloon fight to explain the wet streets. A failed search for balloon fragments would support the rain theory. In science, this is often described as the evidence in favor of a conjecture or hypothesis accumulating until the hypothesis is generally accepted.

VARIABILITY, SAMPLE SIZE, AND REPETITION: LESSONS FROM SCIENCE

In the coin simulation, even a fair coin will give different results because of the inherent random nature of the process. More broadly, when we're talking about informal inference in statistics, we want students to be thinking about whether chance (and the null hypothesis) could have brought about the data we see, and assess the null's plausibility that way—even if, because this is *informal* inference, we have been loose about defining the null.

We think about variability differently in science. We ask whether inevitable measurement errors can explain any deviations of our data from the theory. In practical terms, we make error bars for our points and see if the curve goes through—or close enough to—the bars.

How big do we make the bars, though? We could use an *a priori* value: timing a ball with a stopwatch, we might assign an error of ±0.2 seconds. It might be better, however, to calculate the size of the error bars. Scientists often do this by taking repeated measurements and calculating some number of standard errors (e.g., 2), essentially making a confidence interval (in this case, about 95%), and plotting that range as the bar. The more points you have, the smaller the bars will be. Thus repeated measurements in science behave like increased sample size in statistics: beneficial, but subject to diminishing returns.

Yet there's a pedagogical sinkhole: calculating standard error, and distinguishing it from standard deviation, is a black box to most science students. If we're trying to be informal, can we dispense with the calculation?

Yes. If you look back at the graph of rolling-ball data, you can see the vertical stacks of points that serve as visual error bars. Erickson and Cooley (2005) even suggested that in informal data analysis, we should dispense with error bars—and the black-box calculation of that standard error—in favor of eyeballing the curve through the stacks of points. Even a few points give an impression of how much variability there is and give you a sense both of the range of possible parameter values and of whether the form of the function is plausible.

The same seems true in any curve-fitting situation: an eyeball fit, especially with residuals, does a good job at estimating the parameter, and longer or shorter error bars won't make a lot of difference (Erickson 2008). But it may not be true that an eyeball estimation of the mean, say, from a large sample will be any *narrower* than that from a small sample. The way sampling distributions get narrower with sample size may still be non-intuitive.

At this point, it will be good to address two common misconceptions.

First: Repeated measurements help. One flip of the coin tells us nothing about its fairness. So we flip it 20 or 200 times. But what about that histogram that showed how unusual 16 heads were? Is that another example of repeated measurements?

Yes and no. In that graph, one "unit" is 20 flips. The graph shows 1000 such units, so in a way it's like doing our whole experiment 1000 times. But be careful: the graph no longer shows 20 flips of *our* coin but 20000 flips of the null-hypothesis *fair* coin. It is true that the graph shows the size of our logical *modus tollens* leak (about 1%), and it is a tool for helping us understand our data, but it does not show our data at all.

Second: We often say in science that we value *reproducible* experiments. But this practice of repeating measurements is *not* about reproducibility. Reproducible experiments ensure that your results are *generalizable*: they must occur in other people's labs as well as your own.

This has parallels in statistics: If you are concerned that the particular way you flip a coin might skew the results, you should get other people to flip the coin, assigning flippers at random to different flips. Put another way, reproducibility is to science what randomization (and random sampling) is to statistics: both support generalizability. But that's *not* what repetition is for.

LANGUAGE, LOGIC, AND THE SUBJUNCTIVE MOOD

Consider this statement:

> *If I were to flip a fair coin 20 times, the results would range roughly from 5 to 15 heads.*

This informal statement—with its ill-defined "roughly"—is exactly the kind that leads well to understanding formal statistical procedures.

In an informal setting, even though we might not ask for probabilities, we would like students to make a statement like the one above, and use it to reason about their coin. In fact, of all the elements that contribute to informal approaches, this one seems particularly useful—and correspondingly (and ironically) unlikely to arise by chance in the mind of the naïve student.

Why is that? One problem is that the question is about a *hypothetical* fair coin—*not* about the actual data. From a linguistic point of view, it is in *subjunctive mood*. It is contrary to fact. I have not flipped a fair coin 20 times, nor will I (in a traditional stats curriculum). Our pedagogical question becomes: What prompts a student with a real coin to compare it to a fictitious one?

Let's look at a parallel statement in the science context:

> *If rolling objects accelerate uniformly, and if I were to roll a ball down the track for various distances and record the times, then the distances should be proportional to the squares of the times.*

This conditional statement maps onto the logic we saw earlier; it even has the two-part antecedent. And it is easier to imagine students coming up with it. Perhaps this is because it is something the students expect to *do*: unlike flipping the fair coin, they will in fact roll the ball down the ramp from various distances.

Even so, if we ask science students to produce a statement giving an overview of the purpose of their upcoming lab experience, they are more likely to say this:

> *If I were to roll a ball down the track for various distances and record the times, then the distances should be proportional to the squares of the times.*

That is, they omit the "mechanism" part of the statement, the actual theory they need to test. They accept as fact that the rolling objects accelerate uniformly (it's in the book, after all) and interpret the school activity as testing not the theory of rolling objects but their ability to measure accurately. Put more bluntly, they miss the point.

CAUSALITY, MECHANISMS, FACTS, AND RELATIONSHIPS

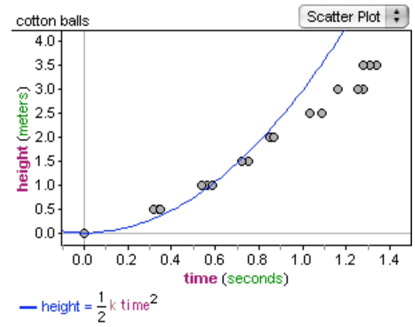
This problem of omitting the mechanism seems to be one where science students have a problem that statistics students do not. For in general, scientific inference is all about evaluating mechanism and causality (does gravity work the way we think?), whereas statistical inference is about assessing facts (is the coin fair?) and, in fact, often *avoiding* causality as dangerous.

Let's explore this idea of "mechanism" briefly, using an example from Erickson and Ayars (2005). Students read an obviously fake physics paper, "On the Descent of Cotton Balls," which said that the distance *s* a cotton ball falls in time *t* is proportional to the square of the time:

$$s = \frac{1}{2}kt^2$$

where k is the acceleration of the ball. The paper explained that air resistance affects the cotton ball, so k is less than g, the acceleration of gravity. The assignment challenged students to design and perform an experiment to test the hypothesis in the paper. (The paper is *wrong*. *See* Note 1.)

Here are some data with a curve showing $k = 6$ m/s^2:



A typical response from a class of state university students—in the second semester of a physics sequence for engineering majors—was to do repeated drops of the cotton ball from one height, average the results, compute k, and report it as confirmation of the theory. In a more sophisticated (but still troubling) response, students dropped the ball from various heights, computed k for each height, and reported the various values of the "constant" k—again, as confirmation of the theory.

Evidently the students did not recognize that the *form* of the function in the paper, its particular symbolic representation, is part of the hypothesis they were supposed to test. It also seems likely that the students were so used to doing labs designed to confirm a formula in the book that the notion of *falsifying* a conjecture—which to Popper is central to scientific work—did not occur to them.

So two things may make this hard for students: first, the conjecture is about a mechanism whose prediction is a relationship rather than a single quantity; and second, students' school experience in science is to verify as true what they have been given. This is a cautionary tale for those of us interested in creating richer activities for students in statistics.

SUMMARY AND SUGGESTIONS

Looking back over this discussion, the problem of getting the underlying logic right—and all of its ramifications, linguistic and otherwise—seems to be a persistent problem that science and statistics share. Basically, students need to know what they can conclude from their observations, and how to describe their results correctly.

Even informally, students need to develop a sense of what's plausible. And even if they use what Edwards et al. (1963) called the "interocular" effect to make a determination (it hits you right between the eyes), they need to compare their data to something—the thing that will eventually turn into the null hypothesis. That comparison arises from the antecedent (the if-clause) in an underlying conditional statement: *if I were to flip a fair coin 20 times…*

And that is the problem: getting students to recognize, when flipping their real coin, that they need to imagine the nonexistent fair one. This problem really has two parts: recognizing the need for a conditional statement at all, and finding the right antecedent. The antecedent problem arises because it's so hypothetical. Is there a way to make the null hypothesis real?

Yes: through simulation and resampling. Simon (1993) wrote an early book about it; Erickson (2006) commented that simulation can help students make the null hypothesis real; and Cobb (2007) essentially called for the introductory statistics course to be remade with randomization-based inference at its core. The statement will cease to be subjunctive and will become, *when I flip a fair coin 20 times, repeatedly….* This is much more like science.

As to the problem of making the conditional at all, science educators address it head-on. Eugenia Etkina and her colleagues (2002) have devised an instructional scheme (*ISLE* for Investigative Science Learning Environment) in which, as part of the process of designing their own experiments, students explicitly learn to construct conditional statements like the ones we used above. They have their undergraduate students essentially fill in the blanks in a formulaic template: "If my idea _____ about this phenomenon is correct, and I do _____, then ______ will occur," and then, "but _____ happened, therefore I cannot reject the idea yet, (or) but _____ did not happen therefore I either need to consider my assumptions or reject the idea."

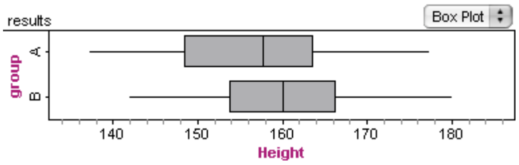
Being that explicit is worth considering in statistics. If we take Cobb's suggestion and focus on randomization, students will no longer have to memorize the confidence-interval mantra; perhaps the time could be productively spent constructing conditional statements like Etkina's.

Even so, (as suggested early in this paper) the science situation is different because scientists generally hope that their hypotheses are true while statisticians hope that they are false. (We are not supposed to bring our desires into either field, but I claim that it makes a difference; you can see how Etkina's focus on falsification is designed to help students break free of that.)

One obvious suggestion, then, is that we give science students more experience with false hypotheses and statistics students more experience with true ones. The "fake paper" activity described above (Erickson and Ayars 2005) is an example of the former, and the frightening student performance on it is a good indicator of its importance.

What would the corresponding activity look like in a statistics classroom? What irrejectable conjectures can we have them explore? Several ideas come to mind, all of which occur in some classrooms already.

- Plenty of exercises in existing texts result in a failure to reject—but these most often require formal inference methods. It would be interesting to have more that are informal. For example, the box plot below shows heights of two groups of 50 young students in centimeters. Maria tells you that the two groups are from the same school, chosen the same way. Do you think Maria's method of choosing people was completely random?



- Students should collect data from apparently fair dice and coins (as they do already in order to learn about probability distributions); but then they should look carefully at how dramatically these diverge from the expected patterns, and come to accept that amount of variation as plausibly due to chance.
- They should (like the scientists) challenge themselves to find alternative explanations for what appear to be null results. For example, if you hear that Gustavo had 10 heads in 20 flips, how could the coin be unfair? What could you ask or observe to convince you? (For example, suppose Gustavo's coin alternated heads and tails.) This will help them both in science and later in statistics, for example, when they need to develop sampling strategies.

An even greater challenge would be to figure out how to incorporate mechanism and causality naturally into more statistics activities. After all, the most interesting data are interesting precisely because we have suspicions about their causes.

So let's address causality head-on, taking a page from science. We can present engaging datasets (such as U.S. income data by sex) and ask: What competing hypotheses explain the data? What additional data could we get to distinguish among the competing explanations? Even if our formal tools may not yet be able to determine causes, the informal insights we might get—and the increased interest—could be worth the trouble.

REFERENCES

Cobb, G. 2007. “The Introductory Statistics Course: A Ptolemaic Curriculum.” *Technology Innovations in Statistics Education.* **1**, 1.

Edwards, W, Lindman, H & Savage, L J. 1963. “Bayesian statistical inference for psychological research.” *Psychological Review* **70** : 193-242.

Erickson, T. 2008. “A Pretty Good Fit.” *The Mathematics Teacher*. In press.

Erickson, T. 2006. “Using Simulation to Learn about Inference” in A Rossman and B Chance, (eds.) *Proceedings of the Seventh International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute. (Available online at http://www.ime.usp.br/~abe/ICOTS7/Proceedings/index.html)

Erickson, T and Ayars, E. 2005. “Fake Papers as Investigation Prompts.” *Physics Education* **40**(6) 550–555.

Erickson, T and Cooley, B. 2005. *A Den of Inquiry*. Oakland, CA: eeps media.

Etkina E, Van Heuvelen A, Brookes D T and Mills D. 2002 “Role of experiments in physics instruction—a process approach.” *Physics Teacher* **40** 351

Popper, K. 1968. *The Logic of Scientific Discovery*. New York: Harper.

Simon, J L. 1993. *Resampling: The new statistics*. Belmont, CA: Duxbury. (This book also comes with the software, *Resampling Stats*, available in several forms through http://www.resample.com.)

NOTES

1: About the physics behind the falling cotton balls: Air resistance does not simply reduce the acceleration, because the force of air resistance depends on speed. As a consequence, the distance-time relationship is not quadratic. The curve starts out quadratic but approaches a straight line when t is large; the slope of that line is the terminal velocity.

ACKNOWLEDGEMENTS

The author is grateful to KCP Technologies for support in attending ICME; and to the U.S. National Science Foundation, award DMI-0216656, for support developing some of the ideas behind this work.