

Introducing Concepts of Statistical Inference

Allan J. Rossman, Cal Poly – San Luis Obispo

arossman@calpoly.edu

<http://statweb.calpoly.edu/arossman>

CensusAtSchool, 2nd International Workshop, UCLA, July 28-29, 2008

Sociology Study: Naughty or Nice?

We all recognize the difference between naughty and nice, right? What about children less than a year old- do they recognize the difference and show a preference for nice over naughty? In a study reported in the November 2007 issue of *Nature*, researchers 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 (Hamlin, Wynn, and Bloom, 2007). In one component of the study, sixteen 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"). The 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. The color and shape and order (left/right) of the toys were varied and balanced out among the 16 infants.

(a) In order for you to be reasonably convinced that infants in general are not just choosing blindly but genuinely choose the helper toy more often, how many of the 16 would have to choose the helper toy? Or would no outcome convince you? Explain your answer.

Suppose for the moment that the researchers' conjecture is wrong, and infants do not really have any preference for either type of toy. In other words, infants just blindly pick one toy or the other, without any regard for whether it was the helper toy or the hinderer. Put another way, the infants' selections are just like flipping a fair coin: choose the helper if the coin lands heads and the hinderer if the coin lands tails.

(b) If this is really the case (that infants show *no preference* between the helper and hinderer), what would be the *most likely* outcome (for number of infants choosing the helper toy) when this study is conducted on 16 infants?

(c) Still assuming that infants show *no preference* between the helper and hinderer, what kind of results (for number of infants choosing the helper toy) would you *not be surprised* to see when this study is conducted on 16 infants?

The researchers actually found that 14 of the 16 infants in the study selected the helper toy.

(d) Calculate the *proportion* of these infants who chose the helper toy. Is this more than half (a majority)?

(e) If it is really the case that infants show *no preference* between the helper and hinderer, which word do you believe best completes the sentence: It would be _____ for 14 out of 16 infants to choose the helper toy just by chance. (Circle your answer below.)

- 1) impossible
- 2) very surprising
- 3) somewhat surprising
- 4) not at all surprising

A key question is how to determine whether the observed result is surprising under the assumption that infants have no real preference. (We will call this assumption of no genuine preference the *null model*.) To answer this question, we will replicate the infants' selection process over and over, assuming that infants have no genuine preference and were essentially flipping a coin in making their choices (i.e., knowing the null model to be true). In other words, we'll *simulate* the process of 16 hypothetical infants making their selections, where we *know* those selections are by random chance (coin flip), and we'll see how many of them choose the helper toy. Then we'll do this again and again, over and over. Every time we'll see how many of the 16 infants choose the helper. Once we've repeated this process a large number of times, we'll have a pretty good sense for whether 14 of 16 is very surprising, or somewhat surprising, or not so surprising under the null model.

Hands-on simulation analysis:

Now the practical question is, how do we simulate this random selection (with no genuine preference)? One answer is to go back to the coin flipping analogy. Let's literally flip a coin for each of the 16 hypothetical infants: heads will mean to choose the helper, tails to choose the hinderer.

(f) Flip a coin 16 times. Count how many of your 16 hypothetical infants chose the helper toy (represented by heads):

(g) Combine your results with your classmates. Do this by producing a dotplot (of the number of infants who choose the helper toy) on the board, where you contribute a dot corresponding to your simulation result from (f). How many simulated experiments altogether are represented in the resulting plot (i.e., how many students put a dot on the board)?

(h) Granted, we have not conducted a large number of simulated experiments, but how is it going so far? Does it seem like the results actually obtained by these researchers would be surprising under the null model that infants do not have a genuine preference for either toy? Explain.

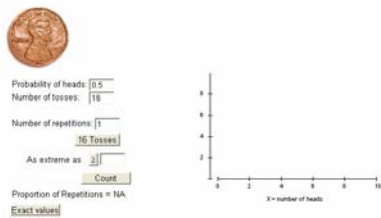
Let's make this question of surprising-ness more specific by quantifying how often the experimental result would occur.

(i) In how many of you and your classmates' simulated experiments did 14 or more infants choose the helper toy? In other words, in how many of the simulated experiments did randomness alone produce a result at least as extreme as the researchers found in the actual study? What proportion of the simulated experiments is this?

Computer simulation analysis:

We can use the computer to simulate this random process of 16 infants making this helper/hinderer choice, still assuming the null model that infants have no real preference and so are equally likely to choose either toy. The computer enables you to conduct many more repetitions much more efficiently.

(j) Open the applet at: <http://statweb.calpoly.edu/bchance/applets/BinomDist3/BinomDist.html>.
Simulating Coin Tossing



Make sure that the probability of heads is set to .5 (as our null model for the infants stipulates) and the number of tosses is 16 (corresponding to the number of infants in the study). Keep the number of repetitions at 1 for now, and click on “16 Tosses” to simulate 16 coin tosses, determining the number of heads and adding this result to the dotplot on the right. Then repeat this four more times (five total), noting how the number of heads in 16 tosses varies from repetition to repetition. Then change the number of repetitions to 995 (which will bring the total number of repetitions to 1000) and click on “16 Tosses” again.

Now enter 14 in the “as extreme as \geq ” box and press the “Count” button. Below the button, the applet will report the proportion of your 1000 repetitions that produced 14 or more infants choosing the helper toy. Record this number here.

(k) Based on this simulation analysis which assumes the null model that infants have no preference and so choose blindly, the actual result obtained by the researchers (14 of 16 infants choosing the helper toy) is _____ (circle one),

- 1) impossible
- 2) very surprising
- 3) somewhat surprising
- 4) not at all surprising

Terminology: The long-run proportion of times that an event happens when its random process is repeatedly indefinitely is called the **probability** of the event. We can approximate a probability empirically by simulating the random process a large number of times and determining the proportion of times that the event happens.

More specifically, the probability that randomness would produce data as (or more) extreme as an actual study, assuming the null model to be true, is called a **p-value**. Our analysis above approximated this p-value by simulating the random process a large number of times (under the null model) and finding how often we obtained results at least as extreme as the actual data. You can obtain better and better approximations of this p-value by using more and more repetitions in your simulation.

A small p-value indicates that the observed data would be surprising to occur by randomness alone, if the null model were true. Such a result is said to be **statistically significant**, providing evidence against the null model.

(l) Click on the applet's "Exact values" button, and report the exact (to three decimal places) p-value. Did your simulation analyses produce a close approximation to this exact p-value?

(m) Based on this simulation analysis, what conclusion should the researchers draw: does their experimental result provide strong evidence that infants in general are not just choosing blindly but genuinely choose the helper toy more often? Explain the reasoning process behind your conclusion, being sure to explain the role of the simulation analysis, as if to a classmate who is not in class today.

Conclusion:

Reasoning process:

(n) If the actual study had instead found that 9 of the 16 infants chose the helper toy, then what decision should the researchers make based on this result? Explain the reasoning process behind your answer.

Conclusion:

Reasoning process:

Medical Study: 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. 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 (15 subjects per group). Both groups engaged in the same amount of swimming and snorkeling each day, but one group (the animal care program) 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. For each subject, the researchers determined whether they showed "substantial improvement" in reducing their level of depression (Antonioli and Reveley, 2005).

As with the previous study, we will replicate the process of randomly assigning subjects to treatments over and over, but assuming that a subject is equally likely to experience improvement no matter which group he/she is assigned to (i.e., knowing the null model to be true). In other words, we'll *simulate* the random assignment process of these 30 subjects to two groups, where we *know* that 17 will improve and 13 will not regardless of which group they go to. Then we'll see how often we get a difference between the groups, favoring the dolphin therapy group, as large as in the actual study. Then we'll do this again and again, over and over. Every time we'll see how large the difference in success rates between the groups is. Once we've repeated this process a large number of times, we'll have a pretty good sense for whether a difference in success rates as large as in the actual study ($10/15 - 3/15 = 7/15 \approx .467$) is very surprising, or somewhat surprising, or not so surprising under the null model.

Hands-on simulation analysis:

Now the practical question is, how do we simulate this random selection (with no genuine preference)? One answer is to take 30 cards, designate 13 of them to represent improvers the other 17 as non-improvers, shuffle them up, and randomly deal out 15 to be in the dolphin therapy group and the other 15 in the control group.

(a) We will provide a set of 30 cards containing 13 face cards (jacks, queens, kings, aces) that represent improvers and 17 non-face cards that represent non-improvers. Shuffle these cards, and deal them into two groups of 15, representing the dolphin and control groups. (Each person should do this once.) Record the results in the following table:

	Dolphin therapy	Control group	Total
Showed substantial improvement (face cards)			13
No substantial improvement (non-face cards)			17
Total	15	15	30

Calculate the difference in success rates/proportions between the two groups. Be sure to take the dolphin success rate minus the control success rate. (If the success rate is higher in the control group, then your answer will be a negative number.)

Dolphin therapy success rate:

Control group success rate:

(b) Combine your results with your classmates. Do this by producing a dotplot of the number of successes in the dolphin therapy group on the board, where you contribute a dot corresponding to your simulation result from (a). How many simulated experiments altogether are represented in the resulting plot (i.e., how many students put a dot on the board)?

(c) Granted, we have not conducted a large number of simulated experiments, but how is it going so far? Does it seem like the results actually obtained by these researchers would be surprising under the null model that dolphin therapy is no more effective than swimming without dolphins? Explain.

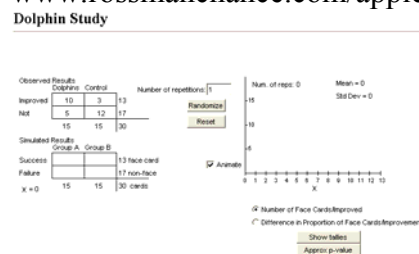
Let's make this question of surprising-ness more specific by quantifying how often the experimental result would occur.

(d) In how many of you and your classmates' simulated experiments was the difference in success rates positive and at least as large as .467? In other words, in how many of the simulated experiments did the random assignment process alone produce a result at least as extreme as the researchers found in the actual study? What proportion of the simulated experiments is this?

Computer simulation analysis:

We can use the computer to simulate this random assignment process of 30 subjects into two groups of 15, still assuming the null model that which group a subject is assigned to has no effect on whether or not he/she will improve. The computer enables you to conduct many more repetitions much more efficiently.

(e) Open Internet Explorer and open the “dolphin applet” from your course web page or at: www.rossmanchance.com/applets/Dolphins/Dolphins.html.



Click on “Randomize” and notice that the applet does what you have done: shuffle the cards (13 successes and 17 failures) and deal them out for the “dolphin therapy” group, separating face cards from non-face cards. The applet also determines the 2×2 table for the simulated results and calculates the number of improvers (face cards) randomly assigned to the “dolphin therapy” group and adds this to the dotplot on the right. Click the button below the dotplot for “difference in proportion of face card/improvement.”. Then click the Randomize button four more times (five total), noting how the difference in proportions varies from repetition to repetition

Now un-check the “Animate” button. Change the number of repetitions to 995 (which will bring the total number of repetitions to 1000) and click on “Randomize” again.

Now click on “Approx p-value.” Below the button, the applet will report the proportion of your 1000 repetitions that produced a difference in success rates at least as large as in the actual study (.467). Record this number here.

(f) The exact (to three decimal places) p-value can be found to be .013. Did your simulation analyses produce a close approximation to this exact p-value?

(g) Based on this simulation analysis, what conclusion should the researchers draw: does their experimental result provide strong evidence that dolphin therapy is more effective than swimming without dolphins? Explain the reasoning process behind your conclusion, being sure to explain the role of the simulation analysis, as if to a classmate who is not in class today.

Conclusion:

Reasoning process:

(h) Suppose that the actual study had instead produced the following results:

	Dolphin therapy	Control group	Total
Showed substantial improvement	7	6	13
Did not show substantial improvement	8	9	17
Total	15	15	30

What decision should the researchers make based on this result? Explain the reasoning process behind your answer. [Hint: Look again at the dotplot of your simulation results. Click on “Show tallies” to see how often each possible difference in success rates occurred.]

Conclusion:

Reasoning process:

Legal Study: 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 and Gelbach, 2005). Here are the data:

	Gilbert working on shift	Gilbert not working on shift	Total
Death occurred on shift	40	34	
Death did not occur on shift	217	1350	
Total			

Before we begin to analyze these data, let’s pause to note some similarities between this study and the previous one. First, we are again comparing two groups, investigating a conjecture that one group would “do better” on the response than the other. And the data can again be represented in a 2×2 table, because both the variables are categorical and binary. But there are two big differences in this study, which we will deal with when they arise.

(a) Were deaths more likely to occur on shifts that Gilbert was working than on shifts when she was not? Calculate and compare the relevant proportions from the two-way table above. [*Hint*: You will first need to fill in the totals of the table.]

Gilbert’s defense attorneys would not be pleased with the information revealed by your calculations. A death occurred on only 2.5% of the shifts that Gilbert did not work, but on a whopping 15.6% of the shifts on which Gilbert did work. The prosecution could correctly tell the jury that a death was more than 6 times as likely to occur on a Gilbert shift as on a non-Gilbert shift.

Could Gilbert’s attorneys consider a “random chance” argument for her defense? Is it possible that deaths are no more likely on her shifts and it was just the “luck of the draw” that resulted in such a higher percentage of deaths on her shifts? Possible, sure, but we have learned from the previous two examples how to use simulation to investigate how unlikely such a result is.

Now we come to one of ways in which this study differs from the dolphin study: we have very large group sizes here. We don’t just have 30 subjects this time. Our observational units are the 8-hour hospital shifts, and we have more than 1600 shifts in this analysis. This is not a complaint: presumably more data is better than less data, as long as it was collected well and recorded accurately. But this does mean that using playing cards for our simulation is out of the question. This time we will use technology from the start.

(b) Use an applet (www.rossmanchance.com/applets/Friendly1/friendly1.html) or other software to simulate 1000 repetitions of randomly assigning these shifts to the two groups. What values are not surprising (for the number of deaths on a Gilbert shift)? Did *any* of these repetitions produce a result as extreme as the actual data (40 or more deaths on a Gilbert shift)?

If any of your simulated repetitions produced such an extreme result, something probably went wrong with your analysis. Granted, it's not *impossible* to obtain such an extreme result, but the exact probability of this can be shown to be less than 1 in 100 trillion.

(c) What are we forced to conclude about the question of whether random chance is a plausible explanation for the observed discrepancy in death rates between the two groups? Can we (essentially) rule out random chance as the explanation for the larger percentage of deaths on Gilbert's shifts than on other shifts? Explain the reasoning process. [*Hint*: What is the null model in this study, and what did the simulation reveal about outcomes produced under the null model?]

If Gilbert's attorneys try to make this argument that "it's only random chance that produced such extreme values on Kristin's shifts," the prosecution will be able to counter that it's virtually impossible for random chance alone to produce such an extreme discrepancy in death rates between these groups. It's just not plausible for the defense attorneys to assert that random chance accounts for the large difference in the groups that you observed in (a).

(d) So while we can't use "random chance" as an explanation, does this analysis mean that there is overwhelming evidence that Gilbert is guilty of murder, or can you think of another plausible explanation to account for the large discrepancy in death rates? Explain.

Now we come to the second primary difference between this case and the dolphin study. This difference is subtle but crucial. It does not change our analysis strategy at all, but it does have a dramatic impact on how we interpret our findings, on the kind of conclusion that we're able to draw. To see what this second difference is, consider the following question:

(e) Did researchers randomly assign shifts to either be worked by Gilbert or not?

Presumably, Gilbert's work schedule was not determined at random. There must have been a schedule, some sort of system, by which shifts were assigned to nurses. It may not have been a rigid schedule or system, but it's certainly not the case that these 1641 shifts were all shuffled up and then 257 shifts were drawn at random to be worked by Gilbert.

(f) So, even after we rule out random chance as the explanation for the data, does that mean that Gilbert's guilt as a murderer is the only explanation that is left?

No, there are any number of other explanations that are possible and even plausible. Maybe Gilbert was assigned to work night shifts, and maybe more of these patients tend to die at night for some physiological reason. That would explain the large discrepancy in death rates between Gilbert shifts and other shifts.

But then, why aren't other explanations possible in the dolphin therapy study? Maybe people who like animals were assigned to the dolphin group and so were more likely to improve; is that a plausible explanation for why the dolphin group improved so much more often than the other group? No, it's not. The reason is *random assignment*. Subjects were randomly assigned to one group or the other, so that impersonal randomization should have balanced out all other characteristics between the two groups (for example, we should get a similar distribution of animal lovers in each group). The *only* difference between the groups should then be whether the subjects got dolphin therapy or a nature program. Because we (pretty much) ruled out random chance as the explanation for the difference in improvement rates, the only explanation left is that the dolphin therapy really helped.

A study such the case against Kristen Gilbert, where the observations units are observed as they occur naturally rather than randomly assigned by researchers, is called an *observational study*. We typically cannot draw cause-and-effect conclusions from observational studies, because the possibility of alternative explanations always exists. Randomized experiments, such as the dolphin therapy study, do allow for cause-and-effect conclusions when the observed experimental results are found to be very unlikely to occur under the null model of no difference between the groups.