Confidence in Sampling: Why Every Lawyer Needs to Know the Number 384

By John G. McCabe, M.A. and Justin C. Mary

Both John (john.mccabe.555@gmail.com) and Justin (justin.mary@cgu.edu.) are in Ph.D. programs at Claremont Graduate University. John is freelance trial consultant and a Ph.D. Candidate who is currently writing his dissertation and hopes to defend it in December. Justin, a researcher in statistics education, has just finished collecting the data for his Masters Thesis. Both are students in the Applied Cognitive Psychology program of CGU's School of Behavioral and Organizational Sciences.

Before revealing and explaining the importance of the number 384, a little background. Our hope is that by explaining the significance of the number 384 and some related concepts, we can help the reader cope with e-discovery, not be fooled by people who, let's say, *oversell* the use of statistics, and put the application of research findings to your case in some perspective.

Let's begin with a hypothetical scenario in which you are a contestant on a reality TV show. You are in a large warehouse and the floor is entirely covered with pennies. The producers have purposely put a specific number of the pennies heads and tails up. You are told that there are 250,000 pennies and your job is to determine what percentage of the pennies are heads up and tails up. How many pennies should you examine and record (i.e., sample) at random to be confident that you have a good estimate of the true number of pennies that are heads and tails in the entire warehouse? Let's define *confident* in this rather obtuse way: if we recreated this exercise 100 times, 95 of those 100 times (95%), the percentage of your sample that are heads and tails, plus or minus (or ±) 5%, would capture the true percentage of heads and tails for all of the pennies in the warehouse. This is obviously a complex question, but most of us can agree that the more you sample the better. In fact, one eager observer might just argue that we should look at all quarter-million pennies! The problem is that we live in a world where we have limited

time and resources, so we are forced to strike a balance in determining how large a sample we should take to achieve an acceptably accurate estimate while minimizing the investment of time and money. With very large populations, the magical number that fits these criteria is, as you may have guessed, the number 384. With a random sample of 384 we can be "95%" confident that the percentage of heads and tails in our sample, ± 5%, captures the true proportion of heads and tails in the entire warehouse. What may be surprising, though, is that even when there are more than 250,000 in the population of pennies, you still only need 384. Essentially, the curve goes up and flattens out around 384. Consequently, 384 is the biggest bang for your buck. Fewer than 384 will reduce your confidence or enlarge the margin of error, while more than 384 results in a quickly diminishing return. For example, the level of confidence could be increased from 95% to 97.5%, but that would take a quadrupling of the sample size to 1536. Barring special circumstances, 95% confidence and a margin of error of ± 5% are generally sufficient, and that's what you get with 384.

So why is this knowledge useful to attorneys? A few examples: first, suppose you have e-discovery and your opponent sends over hard drives containing a million pages of documents. How can you estimate what percentage of pages are relevant to your case? You can't possibly spend the time and money to have associates go through all of them. The relatively paltry sum of 384 pages is actually all you need to have an estimate that is very likely to be within 5% of the percentage on the hard drives. Like a coin flip can only be either heads or tails, whether a page is case-relevant can be seen as binary. It either is or it isn't. So if you choose 384 pages at random and assigned to them a label of either case-relevant or not, you would know with 95% confidence the percentage of pages in

the data dump that are relevant to your case, within 5% either way. Armed with this knowledge, you can better plan for how you will sift what you need from what you don't.

Taking the example a step further, let's say you receive that e-discovery of one million pages of documents. The cost for preliminarily reviewing these documents can be roughly $3,000,000, or $3 per page. Assume that by randomly selecting 384 pages and assigning them dichotomous labels of case-relevant or not you find that 25%, ± 5%, of the documents are case-relevant. At this point, you can develop various filters using key words and phrases, the date the document was created, by whom, etc., and re-sample based on those filters. All of this is with the goal of increasing the richness (the percentage of case-relevant pages) of the documents to be reviewed. You could also then create a scale of case-relevance that is more sensitive than the dichotomous labels. At some point, though, you will have to decide where the lines are drawn between what is and is not to be reviewed. Still, consider the benefits if, for example, you sample 384 pages five times, but can increase the richness of the documents by just 5 %. The investment of some $6,000 to review sampled pages could save your client $150,000 (5% of 1 million pages or 50,000 pages that did not need to be reviewed at $3 per page). And, obviously, the same advantage holds for those producing discovery, but you would want to create a filter for privilege. For those who would say, "Yes, but you could mistakenly eliminate a key document," fallible humans reviewing all one million documents could do the same. And, given how inexpensive sampling is relative to reviewing most or all documents, additional quality control measures to help ensure documents don't slip through the cracks are affordable. Ultimately, though, it is still a question of balancing the maximization of accuracy and the minimization of the cost.

For a second example, you have a case and hire a trial consultant to run focus groups to give you a sense of what the outcome of your trial will be given your strategy. Assume the outcome is essentially a coin toss, 50-50, and is being tried in Colorado Springs with a population of jury-eligible citizens of around 250,000. Assume also that you could create a condensed, but still accurate, version of the future trial to show your participants. How many people would you have to randomly select and run through your focus group so that you could say with high confidence that the proportion of votes from these mock jurors for or against your client is representative of the greater population, ± 5%? The answer is 384. So, if a trial consultant comes to you and says, "Based on the 40 (60, or even 80) participants in this focus group, we think your chances at trial are good (bad)," you have our permission to take that information with a grain of salt. Just so you know, assuming you randomly selected 80 participants out of a population of 250,000, you can be confident that the proportion of votes for or against you are within ± 11%, using the most conservative assumption of a 50-50 split in the population. So a finding that 44% favor your client's side could be as high as 55% in the actual population - more than half, or as low as 33% - less than a third. That's what we would have guessed (around 50-50) and we've never even read your case. We do not mean to imply that small group research for a lawsuit is a bad idea, only that caution must be taken when extrapolating outcomes from small samples to large populations. There is considerable knowledge that can be gained from small group research other than highly suspect predictions of trial outcomes. Still, only by substantially increasing your sample size can you the decrease the margin of error and/or increase your confidence and come up with more predictive information.

The last example concerns research for jury selection and is intentionally oversimplified. Let's suppose that you have research done on a very large sample (around 384 for a population of 250,000 or more) that shows that women who ride bikes favor your case by a 2 to 1 margin. In other words, 67% (± 5%) of women sampled who ride bikes favored your case, whereas only 40% (± 5%) of women who don't ride bikes favored your case. You have the opportunity to strike a non-bike riding female potential juror who will be replaced by a bike riding female juror. Absent any other information, what you should do is obvious - strike the non-bike riding women and try to get the female bike rider on the panel. What may not be as clear is the strength of the rationale for doing this.

To explain, consider two six-sided dice. As the reader may know, the most likely total of two rolled dice is 7. This is because of the 36 possible outcomes of rolling two dice, there more combinations that total 7 than any other number (six combinations; 1-6, 2-5, 3-4, 4-3, 5-2, 6-1). In contrast, there is only one combination that totals 12 (6-6 or *boxcars*). So there is a 17% chance the two dice will total 7, but only a 3% chance of rolling boxcars. However, probabilities in dice predict what will happen over many rolls, over the very long term, not what will happen on any individual roll. Any individual roll of the dice could still turn up 12. In fact, there is nothing about the probabilities that says 12 could not come up repeatedly on successive rolls; probabilities only dictate that over the very long term 7 will come about six times more often than 12.

Getting back to our female, bike riding juror, when we say that the research has identified 67% (± 5%) of female bike riders favor the case, this refers to the entire population. Our confidence is significantly reduced when we hope to predict the attitude

and behavior of a small group or individual. Like rolling boxcars on successive rolls, there is nothing to say that even if we gathered a dozen female biker riders together that they would necessarily include someone who favors your case. In the aggregate, all tolled, there should be 62% to 72% who favor your case, but there is no guarantee at the individual level. This is why probability theory is based on the *Law of Large Numbers* [emphasis added]. The difference is the size of the sample and the confidence we can derive from that sample. In the end, given the absence of other information to motivate another choice, it is reasonable to challenge the non-bike rider in favor of the bike rider. As a wise old lawyer used to paraphrase the Bible and say, "The race does not always go to the swift, nor the battle to the strong…but if you're going to bet, that's where you should put your money." Given the information available, using a preemptory challenge is the wise choice. Still, it is good to keep in mind that this choice may or may not have much impact in predicting an individual's attitudes or behavior.

Many attorneys, particularly those who work in the personal injury field, will recognize this as a nomothetic vs. idiographic problem. *Nomothetic* means regarding a rule, such as a law in science, while *idiographic* means pertaining to the individual. Let's say that we know that, as a rule, exposure to Chemical X causes an increase in a specific type of cancer over base rate in 25% of people exposed to the chemical. Assume for simplicity's sake that the base rate is 0 or that in none of the cases making up the base rate had the person been exposed to Chemical X. Larry Jones has this specific type of cancer and has been exposed to Chemical X. Twenty-five percent of people who are in Larry's position can rightfully claim that Chemical X caused their cancer, but can Larry? In a lawsuit, a jury might look at Larry's case and decide, because there is only a one-in-

four chance Larry can rightfully claim that Chemical X caused his cancer, that there is a greater probability that Chemical X *did not* cause it and so he should get nothing. If every jury used this same logic, none of the 25% who could rightfully claim injury would ever win their suit. Change the percentage to 75% of people exposed to the chemical, and using the same logic, these plaintiffs would never lose. Obviously, neither is correct nor the ideal. This problem becomes even more complicated when we posit a base rate of greater than 0 because we then would have to wrestle with the question of whether Larry Jones would have gotten the cancer regardless of his exposure to Chemical X. It is a stubborn problem in jury research and the legal system as a whole: it is one thing to know as a rule something in the aggregate; it is another to apply that rule to the individual case.

In this article we have discussed the importance of adequate sampling and provided a few examples of its usefulness and limitations. When determining a course of action, we need to be as confident as possible in the quality of our inferences while also being efficient with time and money. With the e-discovery example, most people are surprised at *how few* data points, just 384, are needed to describe with confidence a population in the hundreds of thousands or even millions. Equally as surprising is *how many* people must be surveyed to confidently describe a population's opinion, 384. And when trying your case, it is more than reasonable to use all of research findings you have despite the possibility that they may not affect the outcome, because of course they still may! Finally, if you value your bottom line, as well as being right more often than wrong, you should remember the number 384.