■ RICK HESSE, Feature Editor, Graziadio Graduate School of Business, Pepperdine University

Resampling Calculations in a Spreadsheet

W. J. Hurley, Department of Business Administration, The Royal Military College of Canada

With the advent of high-speed personal computers, resampling procedures are now a realistic alternative to the traditional parametric approach to problems in statistics such as confidence intervals and hypothesis testing. The interested reader is referred to Efron and Tibshirani (Effron & Tibshiravi, 1993) or Simon (Simon, 1997) for an introduction to resampling techniques. There are a variety of software packages available to execute resampling procedures including SAS, S, S-PLUS, and Resampling Stats. However, to my knowledge, nobody has demonstrated the usefulness of spreadsheets for resampling. In this column article I show how resampling calculations can be done within an Excel spreadsheet.

Resampling

We begin with an example from Efron and Tibshirani (Effron & Tibshiravi, 1993):

> Table 1 shows the results of a small experiment, in which 7 out of 16 mice were randomly selected to receive a new medical treatment, while the remaining 9 were assigned to the non-treatment (control) group. The treatment was intended to prolong survival after a test surgery. The table shows the survival time following surgery, in days, for all 16 mice. Did the treatment prolong survival?

On the face of it, the experimental data suggests the treatment might be effective since the treatment group survived an average of 86.68 days and the control group an average of 56.22 days—a difference of 30.63 days. However, there is significant variation in the data and it may be that the difference is due to chance. It would be worthwhile determining if the results might change if this experiment were run again. What is needed is a way to redo the experiment without actually having to spend the time (and mice). Therefore, resampling allows us to randomly choose the results from the experiment N times and calculate a p-value from these N experiments. In this case, the p-value is the probability that we would observe a difference of 30.63 or higher between the two sample means under the null hypothesis that the treatments are the same. If this probability is sufficiently low, we would reject the null.

Group	Data			Sample Size	Mean	Estimated Standard Error
Treatment	94	197	16			
	38	99	141			
	23			(7)	86.86	25.24
Control	52	104	146			
	10	50	31			
	40	27	46	(9)	56.22	14.14
				Difference:	30.63	28.93

Table 1: The mouse data.

Bill Hurley
is a professor in the Department of Business Administration at the Royal Military College. His research interests include decision analysis, game theory, and transportation modeling.
hurley-w@rmc.ca.

To estimate this p-value, we could employ the following resampling experiment:

1. Generate randomly, with replacement, seven observations from the Treatment dataset. One way to do this would be to put seven pieces of paper in a hat, one for each observation in the Treatment dataset, and draw seven out sequentially making sure to replace each after it had been drawn and recorded. These seven observations are termed the Treatment Resample.
2. Generate randomly, with replacement, nine observations from the Control dataset. These nine observations are termed the Control Resample.
3. Compute the mean of the Treatment Resample and the mean of the Control Resample.
4. Repeat steps 1, 2, and 3 N times and count the number of times the mean of the Treatment Resample exceeded the mean of the Control Resample by at least 30.63. Our estimate of the p-value is this number divided by N.

Resampling in Excel

To execute this experiment in a spreadsheet we need to be able to draw out of a hat with replacement. Here is how it can be done. Excel has a function SMALL(array,k) which specifies the kth smallest element from a specified array. For instance, if the array elements are {2,9,7,4,13,5} specified in the cells C3:C8, the function SMALL(C3:C8,3) returns 5, the third smallest element of the array.

Suppose now we have a sample of size n specified in a range called data_range. Then the following function call generates one of the sample points randomly:

=SMALL(data_range,INT(COUNT (data_range)*RAND())+1).

Note that INT(COUNT(data_range)* RAND())+1) generates one of the integers 1,2,3,:.,n at random. The SMALL function then selects the sample element corresponding to the random integer generated. The function LARGE(array,k) could be used just as easily.

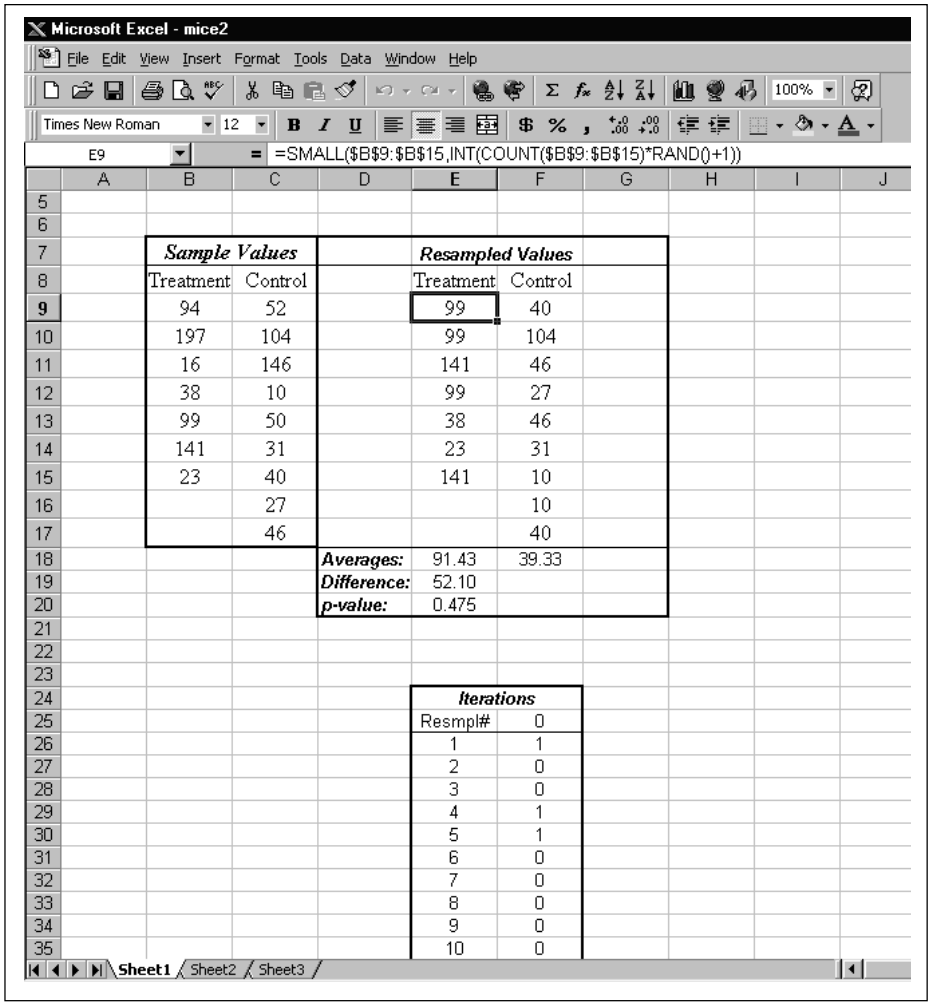


Figure 1: EXCEL output for the Efron/Tibshirani example.

In Figure 1 we show a snapshot of an Excel sheet set up to calculate the p-value for the Efron/Tibshirani example. There are three areas of interest in this sheet. In the range B7:C17, the original data is input under the heading "Sample Values." In the box adjacent to it (D7:G20), under the heading "Resampled Values," a single resampling is executed. Note the distinguishing feature of resampling: in each resample, some of the elements of the original sample appear more than once. For instance in the "Treatment," the element 99 occurs three times. Note also the formula for the active cell, E9. This formula is just the SMALL call discussed above. For the particular resamples shown, the mean of the "Treatment"' resample is 91.43, the mean of the "Control" resample is 39.33, giving a difference of 52.10.

Having set up the resampling experiment, we would now like to repeat it a number of times, each time recording whether or not the difference exceeded 30.63. In the end we would simply count the number of times we observed a difference exceeding 30.63 and then divide by the number of repetitions in order to estimate the p-value. One way to do this would be to hit F9 a number of times making sure to record results manually after each recalculation. By far the easiest way to do this is through the DATA/TABLE command. We have done this in the range E24:F35, labeled "Iterations." The key cell is F26, which contains the formula

=IF(E19>=30.63,1,0).

This IF statement checks whether the resampled difference in sample means (cell

E19) equals or exceeds 30.63. If it does, the formula returns a 1, otherwise, it returns a 0. Once this formula is input, the DATA/TABLE command can be executed for any number of iterations. In this sheet we have done it for 200 iterations. Only the first 10 are shown in FIGURE 1. The column input cell can be any cell not used in the calculations, and it is safest to use cell E25.

To complete the calculation, we simply take an average of the range F26:F225. This is done in cell E20. Therefore our estimate of the p-value is 0.475. This value is not small enough to warrant rejection of the null hypothesis. Hence we conclude there is not enough evidence to say that the difference in the two treatments is 30.63 days or more.

Acknowledgment: *I would like to thank Rick Hesse for pointing out how to automate the resampling using the DATA/TABLE command.*

References

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Simon, J. L. (1997). *Resampling: The new statistics*. Arlington VA: Resampling Stats. ■

Rick Hesse
Graziadio Graduate School of Business
Pepperdine University
Malibu, CA 90265
email: rickhesse@aol.com

The SPECIALIST WITH A UNIVERSAL MIND

■ **ANDREW VAZSONYI, Feature Editor, McLaren School of Business, University of San Francisco**

Get Ready! Aim! Tinker in Cyberspace!

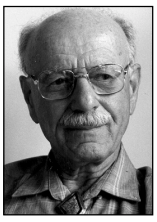
Andrew Vazsonyi, Feature Editor

Have you forgotten the games you played as a child? Remember the old saying, "All work and no play makes Jack a dull boy"?

Kindergarten teachers have known for over a hundred years that the best way to teach children is through play. War games have been around for hundreds of years. Zoologists have studied with great care the playing habits of animals. Psychologist know that people live their lives by playing out games. Von Neuman revolutionized economic theory by the discovery of game theory. S. M. Ulam changed research in the physical sciences by introducing Monte Carlo simulation. Herman Kahn projected the future of the world by playing with scenarios. Herbert Simon changed sciences by artificial intelligence and simulation. Creativity gurus advocate playing with games to discover novel solutions. Microsoft Excel is making simulation available for millions as a fundamental tool of decision making.

The guiding principle of finding solutions is heuristic search; the principle of making choices is bounded rationality.

The essence of simulation is hidden mathematics. ■



Andrew Vazsonyi
is an internationally recognized author, researcher and educator. He is the author of over 70 technical articles, and seven textbooks, in English, German, Spanish, French, Russian, Japanese and Hungarian. Dr. Vazsonyi received a Ph.D. from the University of Budapest. He is currently an emeritus professor at San Francisco University and has 20 years of teaching experience. Prior to becoming an educator, he served for 25 years in industrial positions. These days he focuses on books and articles that apply Microsoft Excel and VBA to production and operations management.

Dr. Andrew Vazsonyi
156 Oak Island Dr.
Santa Rosa, CA 95409
(707) 539-0272
fax: (707) 537-1833
compuserve: 102113,1352
email: avazsonyi@compuserve.com