

Excel Workbooks that Generate Sampling Distributions from Arbitrary Populations for Any Sample Statistic: An Overview and Invitation

Prof. John K. Kruschke
Dept. of Psychology, Indiana University
21 July 2005

To Instructors and Students of Introductory Statistics:

I have developed Excel workbooks that generate sampling distributions from arbitrary populations for any sample statistic, in a way that is easy for students to actively manipulate and observe, and easy for instructors to demonstrate and make assignments from. This document introduces you to the workbooks and invites you to incorporate them into your course.

What the Excel workbooks do:

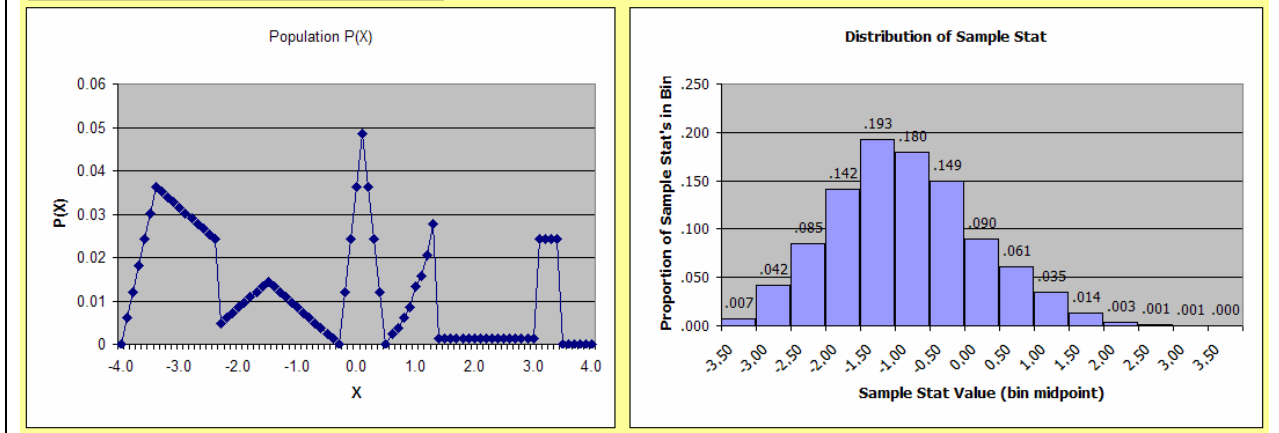
All the usual sampling distributions can be specified. Students can specify the usual null hypothesis populations and generate Monte Carlo sampling distributions of the mean, z , t (for one or two groups) and F (for three groups, or more with expansion of the workbook). From these sampling distributions, students can estimate critical values for hypothesis testing, and thereby get a visceral feeling for the meaning and origin of those critical values tabled in the back of their textbooks.

What makes the system more exciting is the ability to specify other populations and sample statistics, so that the students can

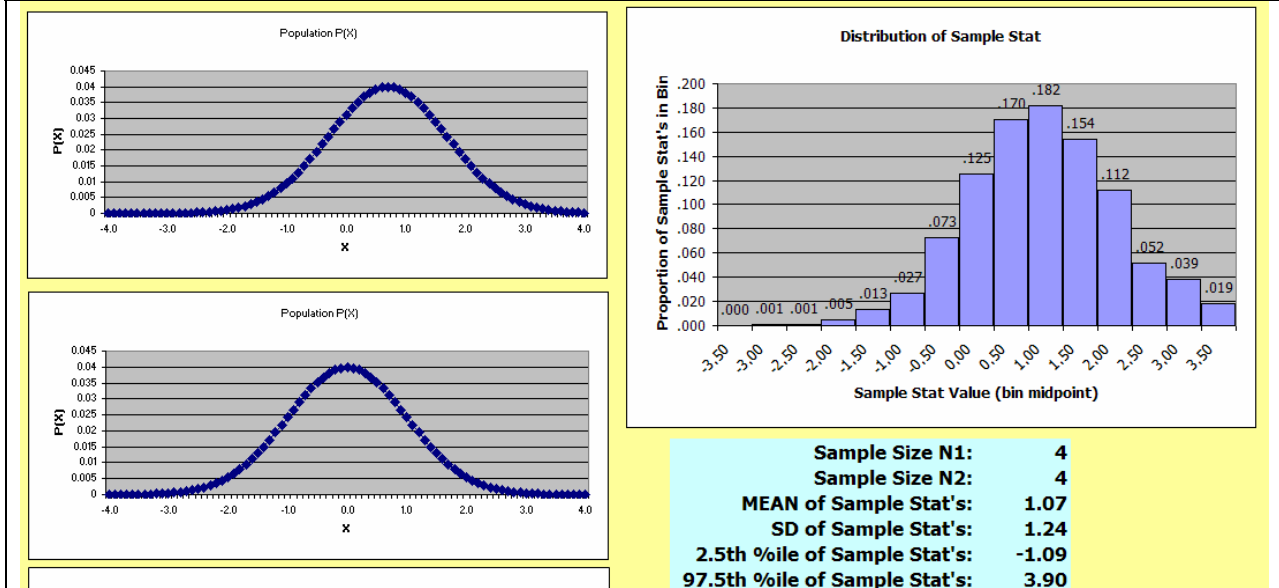
- generate sampling distributions for alternative hypotheses, and consequently estimate statistical power (so students get a hands-on understanding of what power is)
- determine critical values for sampling distributions from non-normal populations or from multi-group populations with non-homogeneous variances (so students actually understand why the heck they are told that the standard tests, using the tabled critical values, need to have data that are normal with homogeneous variances, but that t -tests are "robust against violations of normality")
- examine sampling distributions of other sample statistics, such as median or maximum.
- generate binomial sampling distributions drawn from two-valued populations.
- generate sampling distributions from arbitrary populations, and thereby explore resampling methods.

Some examples of these abilities are shown on subsequent pages.

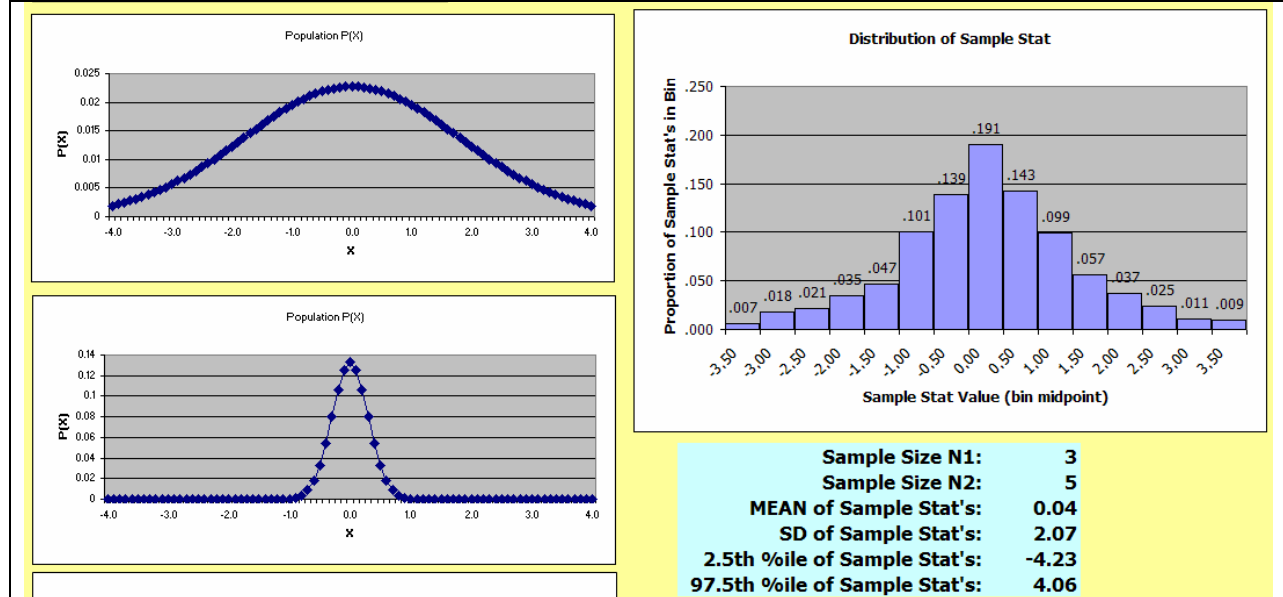
Example: An arbitrarily jagged population distribution can be seen in left panel of the screen shot below. The sampling distribution of the mean, for sample size $N=4$, is shown in the right panel.



Example: A two-group population with an effect size of 0.7 is seen on the left below, and on the right the sampling distribution of t for sample size $N=4$. Students can see that the t distribution is skewed, and they can estimate the statistical power of this alternative hypothesis. Analogous graphs are available for three-group F distributions.



Example: Below is a t distribution sampled from two groups with heterogeneous variances and unequal sample sizes. Students can see that the critical values in this sampling distribution are much larger than the critical values listed in the usual tables. Thus, it becomes clear why the usual t -tests come with otherwise mysterious admonishments about respecting the assumption of homogeneity of variance.



Why Excel?

You might be thinking, "So what? This could be done in SPSS or SAS or Matlab or Minitab or Visual Basic or C++ other programming languages. Why does Excel make it any better?"

Excel is a great tool for introductory statistics because

- So many students are already familiar with it before arriving in the statistics classroom. Excel does not have much "math anxiety" associated with it because it is so familiar. Many students *want* to learn more about Excel because they believe they will use it in other situations, not only in statistics.
- Excel is easier to learn than programming languages because of its intuitive spatial layout of arrays and numbers. There are no difficult abstractions like programming loops for arrays that can't be seen. No declaring of variable names and assignment statements; just click on the cells.
- Students can transfer their Excel skills between statistics and other topics (even personal finances), so there is no worry about teaching and learning idiosyncratic aspects of a specialized programming language that will likely never be used by the student again.
- Excel is cheap and ubiquitous. In fact, at I.U., it's free to all students and faculty, and can be found on every I.U. public computer. But it's also already on most students' personal computers.

Excel does have two disadvantages. First, it does not scale up well to doing real-world statistical analyses. But the introductory stats class does not involve large data sets and intricacies of actual data analysis. More advanced courses can include more specialized statistical packages. A second disadvantage is that Excel does not scale up well for generating really big Monte Carlo sampling distributions. But the moderate sizes included here are fine for pedagogical purposes.

How the workbooks are used.

Each workbook has just four basic worksheets in it, each with thorough instructions to the student for how to use it.

The left screenshot shows a text box with the following instructions:

General Samples for Our Group:
Excel workbooks by Dr. John Kruschke.

IMPORTANT: Because the workbooks will be generating a lot of random numbers, which takes time, automatic updating of the workbook should be turned off. To make sure it is off, do the following: From the menu, click **Tools > Options...**. Then click the **Calculation** tab. In the top section of that box, labeled **Calculation**, click the **Manual** button. (You may also check the box for recalculation before save.) Then, click **OK**. Remember, whenever you want the numbers in the workbooks to be recalculated, you must press F9.

The next worksheet, named **PopulationDistribution** on the tabs at the bottom of the Excel window, is where you specify the population distribution from which you will be sampling.

You specify the following:

- The Population X values.** Type X values for the population in cells A2 to A62. These are the only values that are actually sampled. It defaults to values -4.5, -2.5, -0.5, 1.5, 3.5, 5.5, 7.5, but you can type in any other values you want. In particular, for resampling, type in data values (conveniently as 100 data values).
- The Relative Probability of X.** Type nonnegative numbers for the relative probability of each value of X into cells B2 to B62. For a normal distribution (the default), type in cell B2 = `=NORM.DIST(A2,Pop_Mean*(2*Pop_SD^2))` and copy/paste that into all cells down to B62. For a uniform distribution from cell A2 to A, type in cell B2 to B62 = `1/(A62-A2)`, and in cell B63 to B62. For a two-valued population, e.g. to simulate heads/tails, type in cell B2 to B62 except type a 1 where X is 0 and a 1 where X is 1. Type in any nonnegative numbers to create an arbitrary population profile as desired and recalculate as you like.
- The Population Parameter Values.** Cells D2 and E2 are available for specifying parameter values, but these values are optional. Cell D2 is named **Pop_Mean** and cell E2 is named **Pop_SD**, but you can type in any values you want and use (do not use) those numbers in your specification of the relative probability you want you want. As an example, see how the normal distribution is specified in section 2 above.

Don't forget to press F9 after making changes, to have the worksheet recalculate all its values and update the graph of the population distribution.

The right screenshot shows a table with columns for Population X, Relative Probability of X, Parameter 1 (e.g., μ), Parameter 2 (e.g., σ), Actual Population Mean, and Actual Population SD. A normal distribution curve is plotted over the data.

Step 1: Specify the population distribution. Above are screen shots of the two worksheets where the population distribution is specified. The left screen shows the instructions, and the right screen shows where the population is actually entered. The user simply enters a column of possible population values and the relative probability of each value. The relative probabilities can be entered as raw values or as a formula (function) of the population value.

Step 2: Specify the sample size. At the right is a screen shot of the samples generated from the population. Each row is a sample, so the number of columns is the sample size. To change the sample size, the user simply copies or deletes columns. Full instructions are on a preceding screen that looks much like the instruction screen in the preceding step.

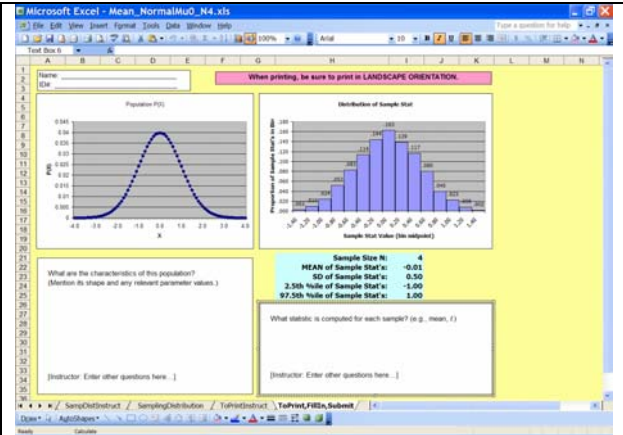
The screenshot shows a grid of random samples. The columns are labeled X1 through X16. The rows are numbered 1 through 33. Each cell contains a random value, such as 0.70, -0.40, 1.00, etc.

Step 3: Specify the sample statistic. Here the user specifies the statistic to be computed for each sample, and the bin values for the histogram. The screen shot at the right shows this worksheet. Like the others, it is preceded by an instruction screen with examples.

The screenshot shows a histogram titled "Distribution of Sample Stat". The x-axis is labeled "Sample Stat Value (bin midpoint)" and the y-axis is "Proportion of Sample Stat in Bin". A summary table below the histogram provides the following statistics:

Sample Size n:	4
MEAN of Sample Stat:	-0.01
SD of Sample Stat:	0.50
97.5th Ntile of Sample Stat:	1.00

Step 4: Summarize and answer questions from the instructor. The final worksheet collects the population and sampling graphs on one page. It also includes panels where questions from the instructor can be entered (and answered by the student). This page is intended to be printed out by the student and handed in as evidence of having completed the assignment. Assignments are therefore easy to generate and easy to grade.



Good payoff for small investment of time

The workbooks are almost self contained, and require relatively little investment of the instructor's time. The payoff is that students actually understand the foundational topic of hypothesis testing, which unfortunately remains a mystery to many students in an introductory statistics course. Realistically, the instructor will want to demonstrate the workbooks in class before assigning them, and will want to go over a few basic Excel skills such as entering a formula in a cell and copying formulas down a column. The only other thing the instructor needs to do is decide what additional questions s/he would like the students to address in the final worksheet; this allows the instructor to fit the workbooks to his or her own curriculum.

How to get the workbooks

The workbooks are available free from the author. Go to the web site
<http://www.indiana.edu/~jkkteach/ExcelSampler>

The author can be contacted via e-mail at kruschke@indiana.edu.

Acknowledgment

This project was supported in part by an Active Learning Grant from Instructional Support Services, Indiana University.