Software Supplement

To Accompany "Introductory Statistics and Analytics: A Resampling Perspective"

Stan Blank, Ph.D.

Table of Contents

1 Designing and Carrying Out a Statistical Study

Software Notes:

Excel

Although it is not as robust as standard statistical software, Microsoft Excel is widely used for routine statistical analysis, particularly in business. We assume some familiarity with Excel and there are many Excel tutorials available online if you need a refresher. It is difficult to perform resampling or bootstrapping using Excel alone, so we make available two software packages (Box Sampler and Resampling Stats) as Excel add-ins for this purpose.

Box Sampler

Box Sampler is freeware and runs on Windows using Microsoft Excel. Box Sampler works with Excel versions 2003, 2007, and 2010. It is recommended that you start or load Box Sampler from the Start>Programs menu or from the desktop icon (if the icon is installed). We do NOT recommend loading Box Sampler from the Excel Add-ins dialog, using found in the Excel Tools or Options menu. If you do not have Box Sampler downloaded, you can use the following link to download the software:

http://www.resample.com/box-sampler/

Installation is simple, but you do need to enable macros in Excel for Box Sampler to work properly. If you don't know how to enable macros, see the Resampling Stats instructions on the Resampling Stats download page (the link is below) for instructions specific to your version of Excel.

Box Sampler is somewhat slow in execution and is limited to a Sample Size of 200, but the program provides an excellent visual representation of resampling procedures. If you are using Excel 2003, Box Sampler will have its own menu entry at the top of the Excel worksheet. If you are using Excel 2007 or Excel 2010, the Box Sampler menu is found in the Excel Add-ins ribbon menu at the top of the worksheet. Box Sampler can be found in the left panel in the Add-ins menu. A Box Sampler toolbar will appear in all versions of Excel.

Resampling Stats

Resampling Stats is a commercial software package. We will provide you with a 180-day license at no additional charge in order for you to complete your homework assignments. You can download Resampling Stats from the links provided in the software section from Lesson 1 in your course. Resampling Stats 2003 works with Excel 2003 (and earlier versions of Excel) and Resampling Stats 2007 works with Excel 2007, Excel 2010 and Excel 2013.

If you have not already downloaded Resampling, you can use the link below this paragraph to access the download page.

http://www.resample.com/

Please follow all the installation instructions in Steps 1 and 2 on the linked page above for your version of Excel. You must enable both Analysis Toolpak add-ins and enable macros in order for the Resampling Stats add-in to function properly.

If you have questions or receive an error message, please visit the following page:

http://www.resample.com/troubleshooting.shtml

If you have followed the directions on both pages carefully and you still receive an error message, please contact: stats@resample.com for help.

The Resampling menu is found in the Tools menu in Excel 2003 or in the Add-ins ribbon menu in Excel 2007, Excel 2010 or Excel 2013. The Resampling menu will be in the left panel of the Add-ins menu. Resampling Stats will display a toolbar in all versions of Excel. If you are using Excel 2013, all references to Excel 2010 should work properly for you.

NOTE: We do NOT recommend starting either Box Sampler or Resampling Stats using the Excel Add-ins menu. Please use the desktop icons (if installed) or the Start>Programs menu to load the desired program. You may also start Excel and load either Box Sampler or Resampling Stats from the File menu. It may be necessary to navigate to the proper folder where the programs are installed. Usually the Box Sampler and Resampling Stats folders are found in the C:\Program Files directory.

StatCrunch

StatCrunch is a very powerful statistics package that can solve many resampling or bootstrapping problems. In addition StatCrunch has many formula based statistical functions. You must visit the StatCrunch website, http://www.statcrunch.com/ and purchase a license. The cost is very reasonable. One major advantage of StatCrunch is that the program is Java-based and will work on the Mac, Windows, and Linux platforms. However, it is not quite as flexible as Resampling Stats when using resampling for more complex problems.

Note: There are now two versions of StatCrunch, the classic version and the new version. The supplement illustrates the NEW version of StatCrunch. The new version has some differences in the dialogs and menus when compared with the classic version, but the commands are identical in both versions and the appropriate dialog buttons and edit boxes, while having different styles and captions in some instances, are analogous. The author prefers the classic version of StatCrunch and has had both versions running successfully on the Mac (OSX 10.8.x) and Windows (XP, 7, 8) operating systems using Safari, FireFox and Chrome. It is important to update your java platform to the latest version and to allow java plug-ins to run in your browser. Please read the online help available on the StatCrunch website for further information.

Using the Supplement

This textbook supplement is aligned closely with the textbook. Where possible, step-by-step procedures for all three software packages are used to work or solve the examples found in the text. You should read both the textbook and supplement carefully to get the greatest benefit from the course… and, of course, actually use your software along with the supplement! In addition, the Resources course has many tutorials available for Excel, Box Sampler, Resampling Stats, and StatCrunch. As is true in any endeavor, the more you practice, the more proficient you will become.

1.2 Is Chance Responsible? The Foundation of Hypothesis Testing.

Prior to working with specific software examples we should explore one of the most important Excel menu items.

Insert Function (Function Wizard)

The Excel Insert Function feature (also called the function wizard) allows you to quickly search for and insert native Excel functions. The function wizard is opened by selecting a worksheet cell clicking the ***fx*** button to the left of the formula bar.

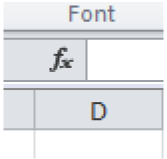
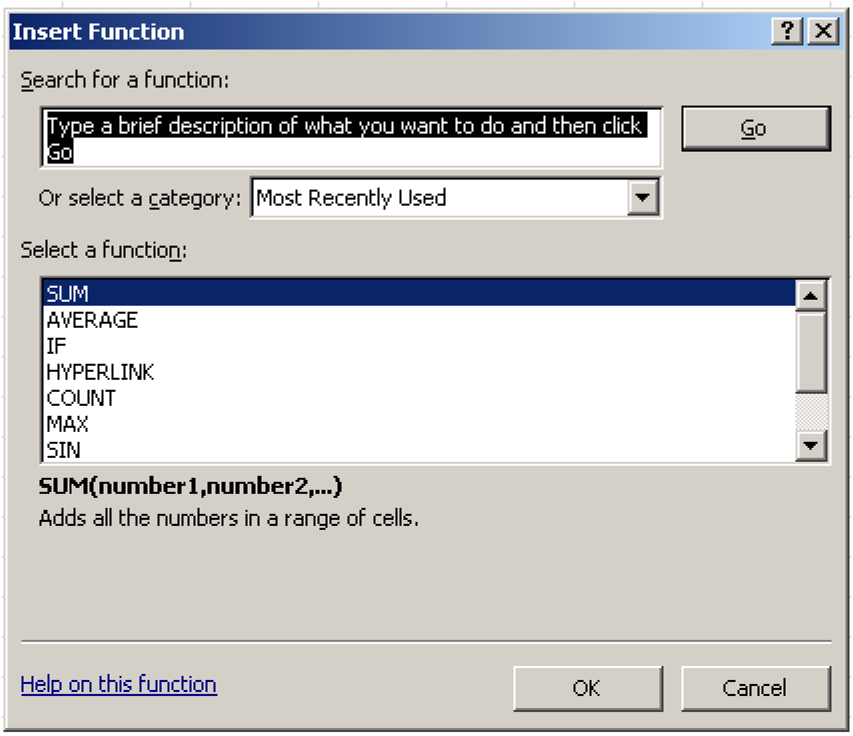


Figure 1-1: Function wizard ***fx***

Click on the ***fx*** button to display the following dialog:



Search for a function: Type a brief description of what you want to do and then click Go

Or select a category: Most Recently Used

SUM
AVERAGE
IF
HYPERLINK
COUNT
MAX
SIN

SUM(number1,number2,...)

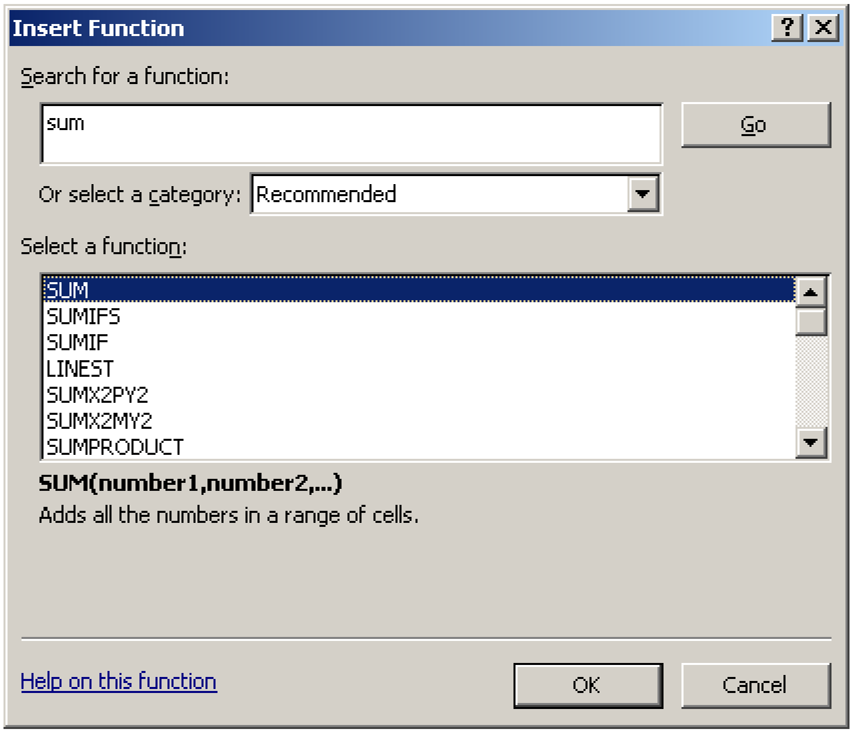
Adds all the numbers in a range of cells.

Figure 1-2: Function wizard dialog (Insert Function)

Let's try an example. Enter the values 1 through 20 in cells A1:A20 inclusive. You can enter the numbers individually or you can try the following shortcut:

1) Enter 1 in cell A1
2) Enter 2 in cell A2
3) Select both cells (click in cell A1 and drag down to cell A2)
4) Click on the small box in the lower right corner of cell A2 (you should see a + instead of a mouse cursor if you are in the correct position). This is the "fill handle" and allows you to copy the contents of cells in any direction. If there is a sequence of values in a range, Excel will attempt to extend the sequence.
5) With the left mouse button pressed, drag down to cell A20. You should have the values 1-20 in cells A1:A20. Note that Excel assumed that you wanted to continue with the sequence 1, 2 by adding 3, 4, 5 … 19, 20 (later, try other sequences such as odd numbers, even numbers, 1, 4, 7…. , etc. to see how intelligent Excel can be in copying/extending sequences).

Find the sum of these values by selecting cell A21 and then clicking the Insert Function (function wizard) button. The SUM function is most likely the top entry in the "Select a Function:" box, but go ahead and type "sum" in the "Search for a function:" edit box and click the "Go" button (figure 1-3):



Search for a function: sum

Or select a category: Recommended

Select a function:

SUM
SUMIFS
SUMIF
LINEST
SUMX2PY2
SUMX2MY2
SUMPRODUCT

SUM(number1,number2,...)

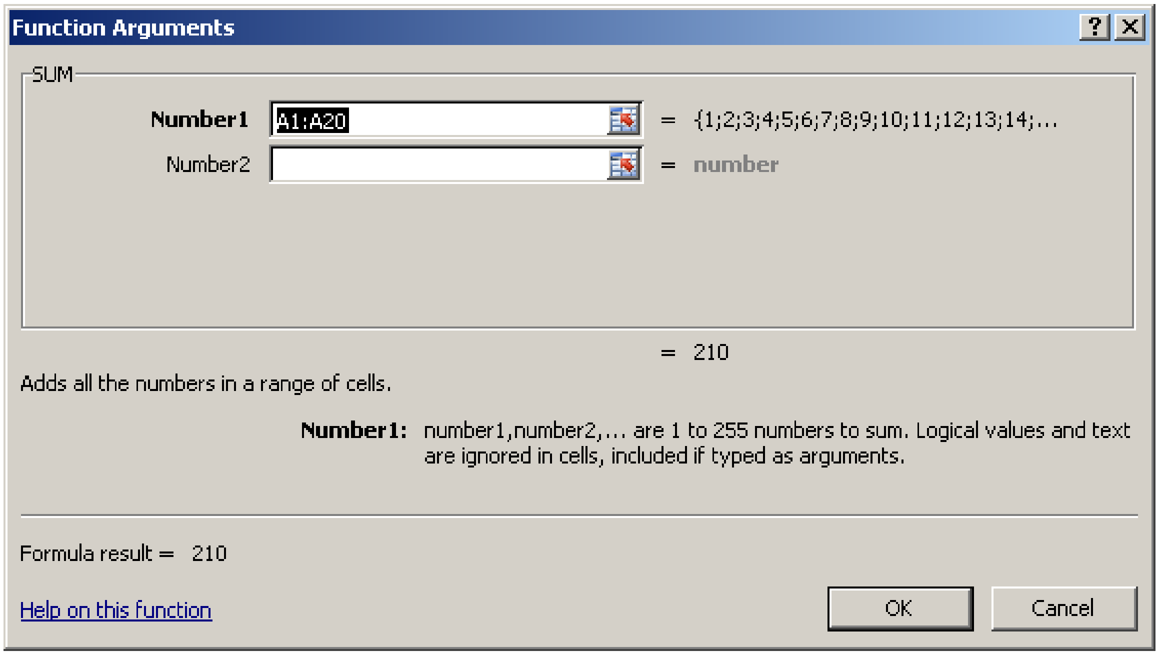
Adds all the numbers in a range of cells.

Figure 1-3: Excel Insert Function wizard

Select the SUM function and click OK. The function wizard will display a dialog that allows you to enter the proper function arguments.

> ***Definition:* Argument**
> An argument is a data value supplied as an input to a function, algorithm or computer procedure.

In this example, Excel correctly assumes you want to find the sum of the range A1:A20 (figure 1-4). Click OK.



Number1 A1:A20 = {1;2;3;4;5;6;7;8;9;10;11;12;13;14;...

Number2 = number

= 210

Adds all the numbers in a range of cells.

Number1: number1,number2,... are 1 to 255 numbers to sum. Logical values and text are ignored in cells, included if typed as arguments.

Formula result = 210

Figure 1-4: Enter function arguments

The function wizard completes the proper function syntax and enters the formula in cell A21. The sum 210 is displayed immediately. Note the formula bar and the function syntax: =SUM(A1:A20)

Each Excel function MUST begin with the "=" sign. If you want to multiply the cells A1 and A2 and store the value in cell C1, you would use the following function entered in cell C1: =A1*A2

If you want to find the square root of the product of cell A1 and A2, in cell C1 you can enter: =SQRT(A1*A2)

The possibilities are endless and there are additional Excel tutorials in the Resources.

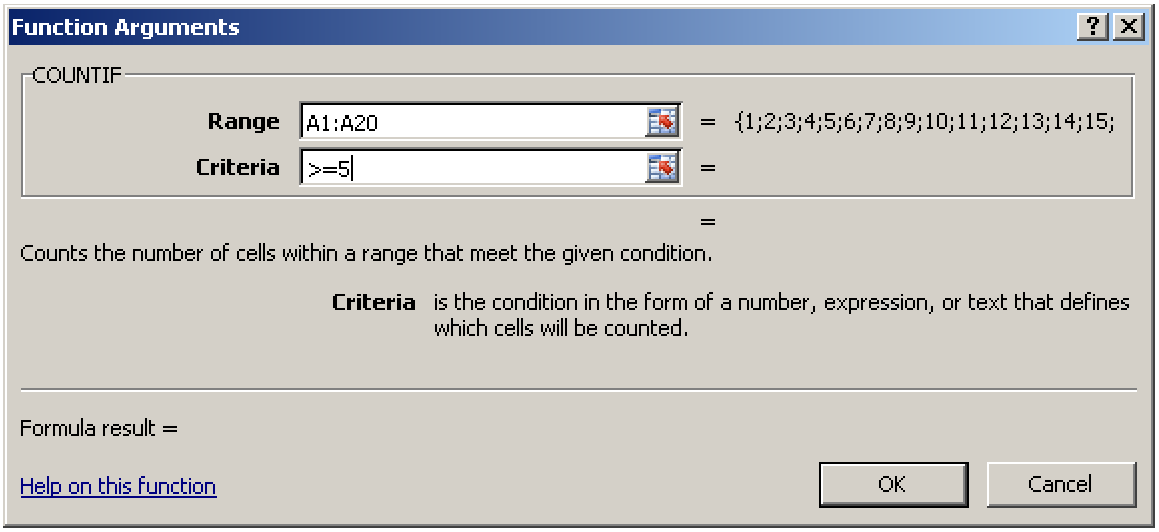
Let's continue…

Now find the number of values in cells A1:A20 that are >=5 (greater than or equal to 5).

1) Select a different cell, say B1, to store the output. Click the ***fx*** button
2) Search for COUNT
3) Select the COUNTIF function in the dialog
4) Click OK

5) Enter the range A1:A20 in the Range edit box. You can also click in the edit box and select the range using the mouse (click in cell A1 and drag down to cell A20)

6) Enter >=5 in the “Criteria” edit box (figure 1-5):



Range A1:A20 = {1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;

Criteria >=5 =

=

Counts the number of cells within a range that meet the given condition.

Criteria is the condition in the form of a number, expression, or text that defines which cells will be counted.

Formula result =

Figure 1-5: =COUNTIF function

7) Click OK

8) As expected, there are 16 values >=5

9) Note the function syntax in cell B1: =COUNTIF(A1:A20, ”>=5”)

Finally, select a new empty worksheet cell, click the ***fx*** button again and click the “Select a category:” dropdown list box (figure 1-6). All native Excel functions are grouped by category. Selecting a category will populate the “Select a function” box with all functions in that category. Click on “Statistical” and observe the result.

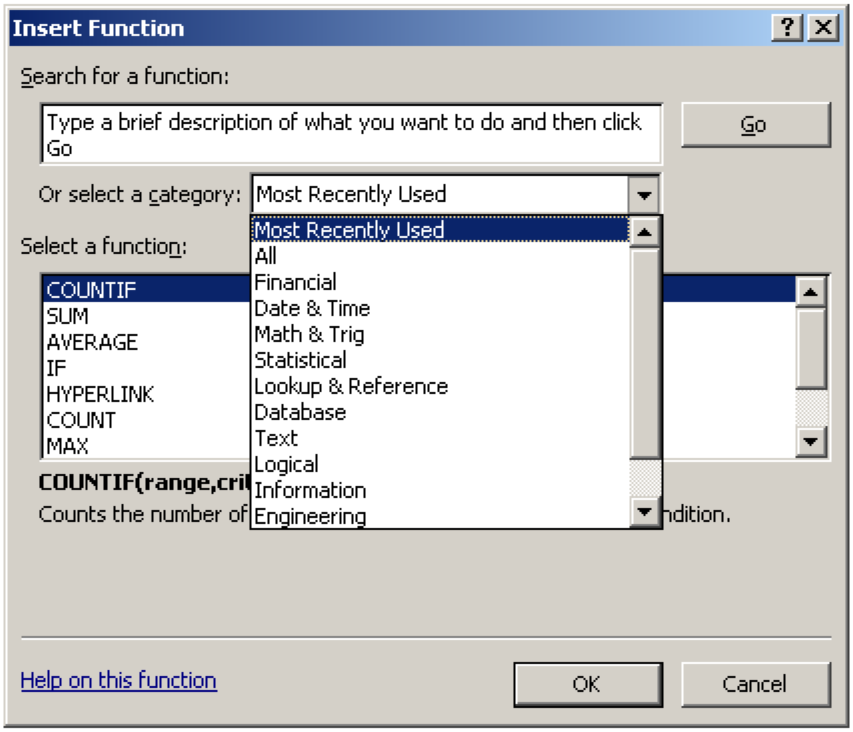


Figure 1-6: Select a category

All native Excel functions can be searched for and entered by using the Insert Function wizard.

Try It Yourself

Toss a coin ten times and record the number of heads and the number of tails. We will call the ten tosses one trial. Then repeat that trial eleven more times for a total of twelve trials and 120 tosses. To try this exercise on your computer, click <here> to download a macro-enabled Excel workbook containing a Box Sampler model. To view the procedure in Resampling Stats, reference section 1.2 in the Text Supplement.

Did you ever get seven (or more) heads in a trial of ten tosses?

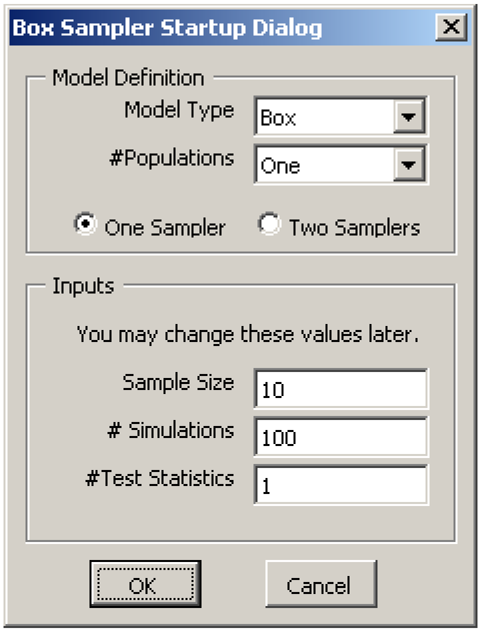
Box Sampler Coin Toss

The Try It Yourself exercise (above) has a link that provides a downloadable macro-enabled workbook demonstrating a Box Sampler solution for the coin toss scenario. In the course Resources, there is a video <here> that illustrates how to create a coin toss experiment using Box Sampler. However, it may be helpful to view a step-by-step procedure of how the Box Sampler workbook was created for the Try It Yourself exercise.

1) Open Box Sampler in Excel by loading the Box Sampler add-in as you would an Excel workbook:
    a. You can use the File>Open menu in Excel 2003 or Excel 2010/2013 or the Office Button menu in Excel 2007. Search for the BoxSampler.xla add-in and open the file. Most likely you will find the file in the C:\Program Files\BoxSampler\ or C:\Program Files (x86)\BoxSampler\ folder.
    b. You can also open/load the Box Sampler add-in by using the Start menu (or its equivalent in Windows 8) and double-clicking the Box Sampler shortcut.
    c. Another method for opening Box Sampler is to find the Box Sampler installation folder (as indicated in step a. above) and drag the Box Sampler add-in icon into the Excel worksheet.
    d. You can facilitate opening Box Sampler in the future by copying the Box Sampler shortcut (in the Box Sampler menu or in the Box Sampler folder) and pasting the shortcut to your desktop. Opening Box Sampler then becomes a simple process of double-clicking the shortcut icon.

2) After the Box Sampler add-in has been opened in Excel 2007, 2010 or 2013, you will find the Box Sampler menu in the Add-Ins menu on the main Excel worksheet page. In Excel 2003, Box Sampler will have its own menu at the top of the Excel worksheet.

3) Select the Add-ins>Box Sampler>New Model menu or in Excel 2003, the Box Sampler>New Model menu. You should see the following dialog (figure 1-7):



Model Type: Box

#Populations: One

● One Sampler ○ Two Samplers

You may change these values later.

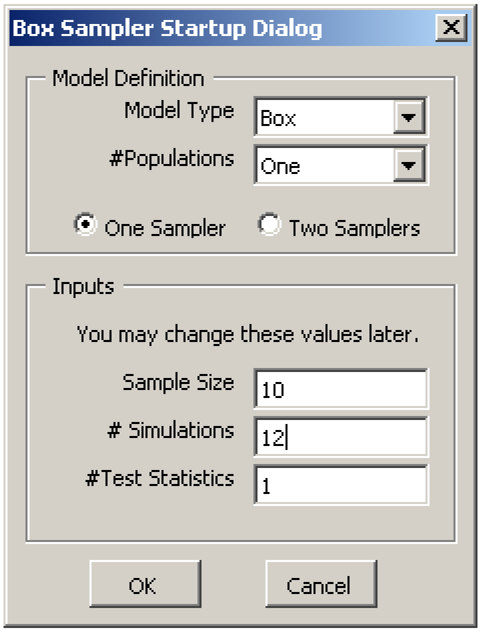
Sample Size: 10

Simulations: 100

#Test Statistics: 1

Figure 1-7: Box Sampler Startup Dialog

4) Change the # Simulations to 12. All other entries are OK as they are. You should now see the following dialog (figure 1-8):



Model Definition

Model Type: Box

#Populations: One

● One Sampler ○ Two Samplers

Inputs

You may change these values later.

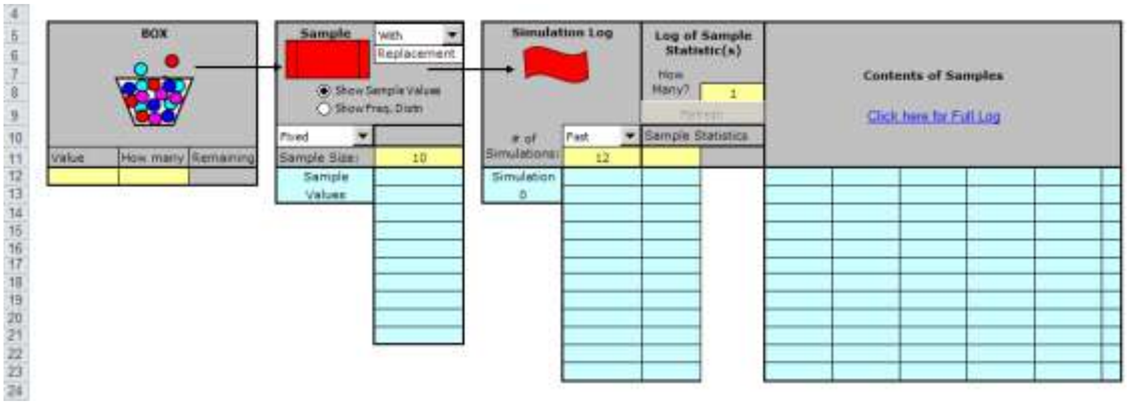
Sample Size: 10

Simulations: 12

#Test Statistics: 1

Figure 1-8: Enter 12 for the # Simulations

5) Click OK and the following Box Sampler setup will be generated (figure 1-9):



Sample Size: 10

# of Simulations: 12

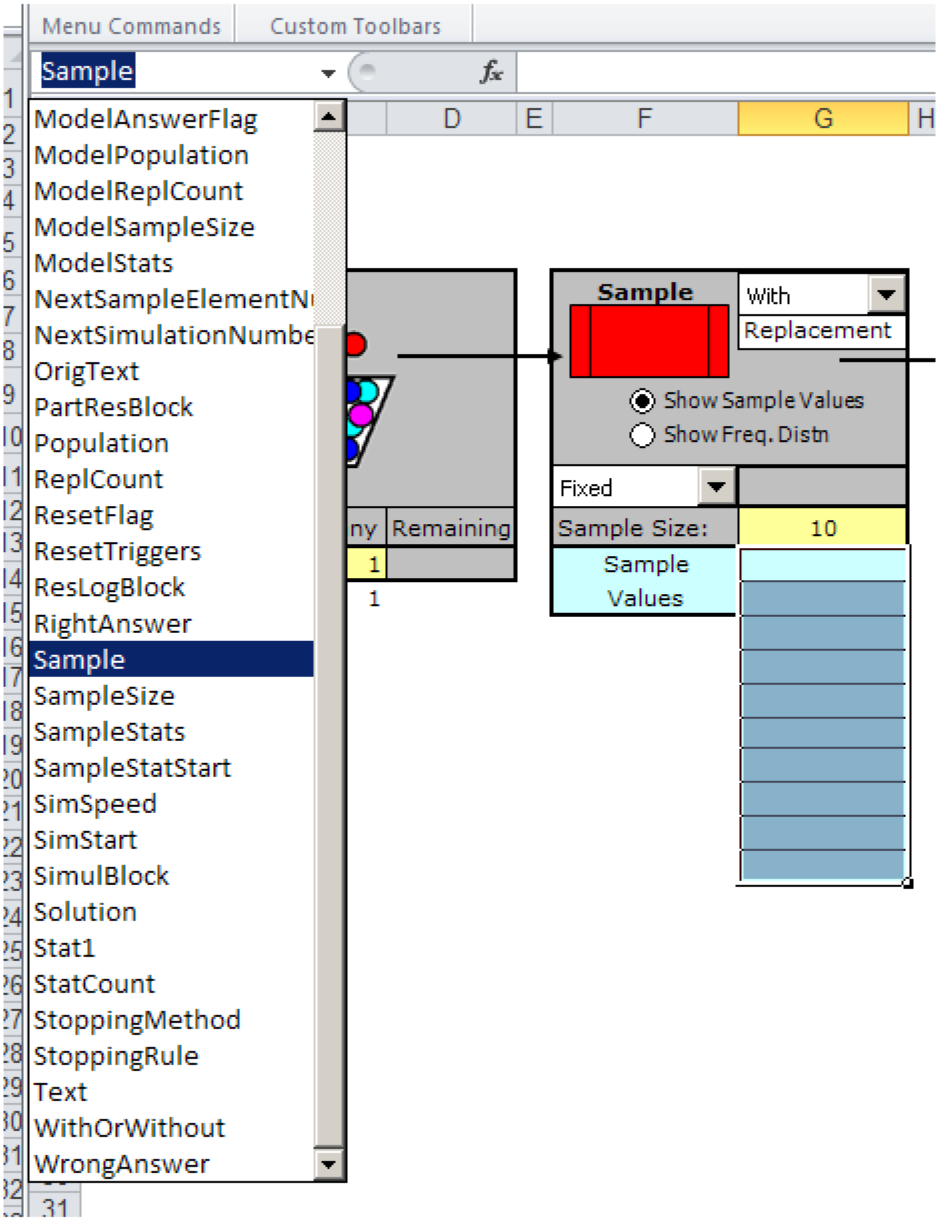
How Many? 1

Figure 1-9: Initial Box Sampler model

6) Under the Value label in the Box (cell B12), enter heads. In cell B13 enter tails. We want one instance of each outcome, so we enter a 1 (under the How many label) in cells C12 and C13. Do not worry if the tails entry is not highlighted in yellow at this point. The highlighting will occur automatically when you run the simulation. In cell K11 (under Sample Statistics) enter the following formula:

=COUNTIF(Sample,"=heads")

The Excel function =COUNTIF() is very useful. This particular instance checks the Sample range (cells G12:G21 in this model) and counts the number of times that heads appears in the Sample range. Box Sampler always names several areas of the worksheet with specific names. To see these names and the areas they encompass, you can check the Range dropdown menu in your Excel worksheet as shown in figure 1-10. Note that the Sample range name refers to cells G12:G21 as indicated above.



With Replacement

● Show Sample Values
○ Show Freq. Distn

Fixed

Sample Size: 10

Figure 1-10: Named Ranges in Box Sampler

7) Click the Simulate button (the >> button as shown in figure 1-11).

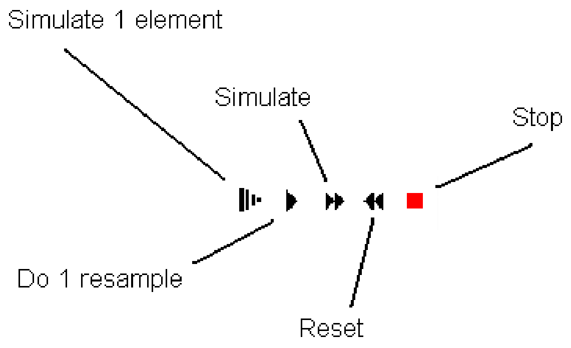
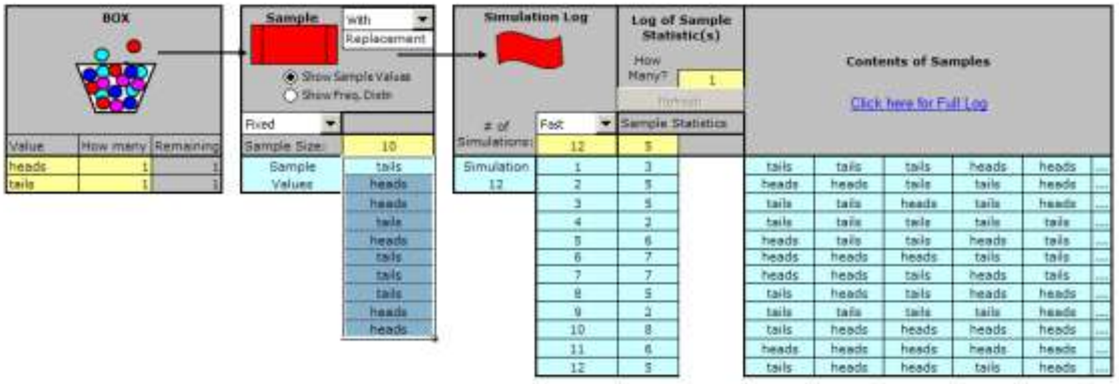


Figure 1-11: The Box Sampler control menu

8) After clicking on the Simulate button, you should see something similar to the following (figure 1-12):



Value	How many	Remaining
heads	1	1
tails	1	1

Fixed	
Sample Size:	10
Sample Values	tails
	heads
	heads
	tails
	heads
	tails
	tails
	tails
	heads
	heads

How Many? 1

# of Simulations:	Fast	Sample Statistics
	12	5
Simulation 12	1	3
	2	5
	3	5
	4	2
	5	6
	6	7
	7	7
	8	5
	9	2
	10	8
	11	6
	12	5

tails	tails	tails	heads	heads	...
heads	heads	tails	tails	heads	...
tails	tails	heads	tails	heads	...
tails	tails	tails	tails	tails	...
heads	tails	tails	heads	tails	...
heads	heads	heads	tails	tails	...
heads	heads	tails	heads	tails	...
tails	heads	tails	heads	heads	...
tails	tails	tails	tails	heads	...
tails	heads	heads	heads	heads	...
heads	heads	heads	heads	heads	...
tails	heads	heads	tails	heads	...

Figure 1-12: Results of one experiment with 12 simulations (trials) of 10 coin tosses

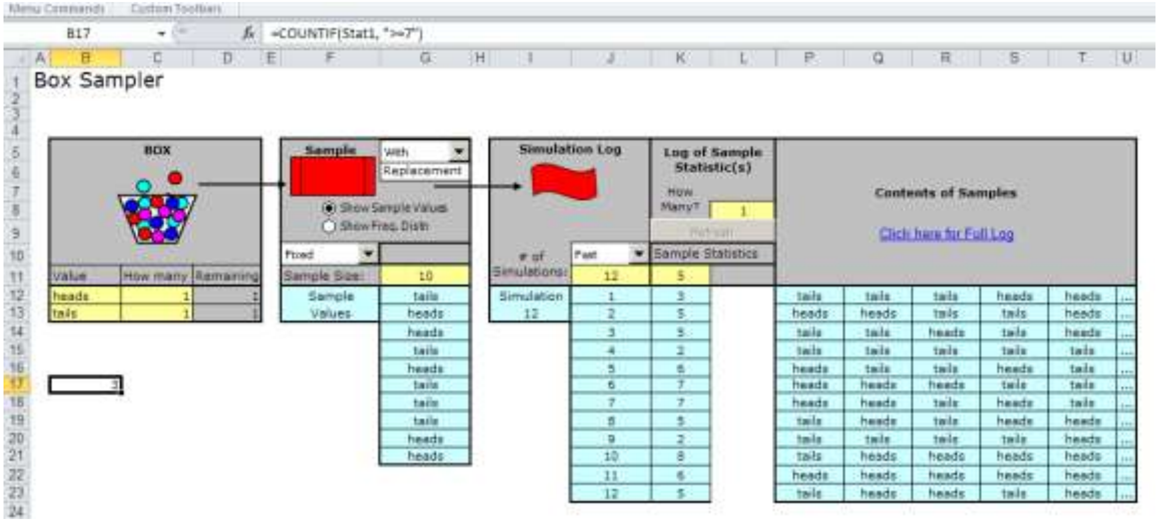
9) With such a small number of simulations (12) we can easily count the number of times we have 7 or more heads. In the above example, there are 3 instances that meet our criteria. However, if we changed the number of simulations to 100 or 1000, the counting task becomes a bit more difficult. We can easily remedy this situation by using =COUNTIF() once again. In the Try It Yourself download example, the macro-enabled workbook has the =COUNTIF() formula below the Sample Statistic output range (cells K12:K23… and by the way, no matter how large or small this range, this output range is named Stat1 by Box Sampler… very handy as we shall see!). Let's place this formula in another location so that it will not hinder us if we choose to increase the number simulations. Cell B17 is a good location, so in cell B17, we could enter:

=COUNTIF(K12:K23, ">=7")

But if we increase the number of simulations, we would need to change the K12:K23 range to reflect the new output range. Instead, we'll use the following formula in cell B17:

=COUNTIF(Stat1, “>=7”)

Now run the simulation again. You should see something similar to the following output in figure 1-13.



B17 =COUNTIF(Stat1, ">=7")

B17: 3

Value	How many	Remaining
heads	1	1
tails	1	1

Sample Size: 10

Sample Values: tails, heads, heads, tails, heads, tails, tails, tails, heads, heads

How Many?: 1

# of Simulations: 12

Simulation	5
1	3
2	5
3	5
4	2
5	6
6	7
7	7
8	5
9	2
10	8
11	6
12	5

P	Q	R	S	T	U
tails	tails	tails	heads	heads	...
heads	heads	tails	tails	heads	...
tails	tails	heads	tails	heads	...
tails	tails	tails	tails	tails	...
heads	tails	tails	heads	tails	...
heads	heads	heads	tails	tails	...
heads	heads	tails	heads	tails	...
tails	heads	tails	heads	heads	...
tails	tails	tails	tails	heads	...
tails	heads	heads	heads	heads	...
heads	heads	heads	heads	heads	...
tails	heads	heads	tails	heads	...

Figure 1-13: =COUNTIF(Stat1, “>=7”) in cell B17 to count the instances of heads >= 7

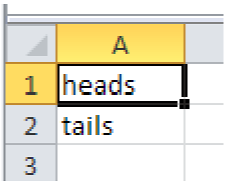
10) Run the simulation again. Now change the # Simulations from 12 to 100 and try again. If time permits, try 1000 simulations. You can also experiment with the number of tosses (Sample Size). Feel free to experiment!

Hopefully the above steps, along with the tutorials and videos in the Resources, will give you a good start in using Box Sampler. There are additional step-by-step procedures for Box Sampler later in this supplement.

Resampling Stats for Excel (RSXL)

Note: Resampling Stats for Excel refers to the original (real) sample as "sample" and to simulated/resampled samples as "resamples."

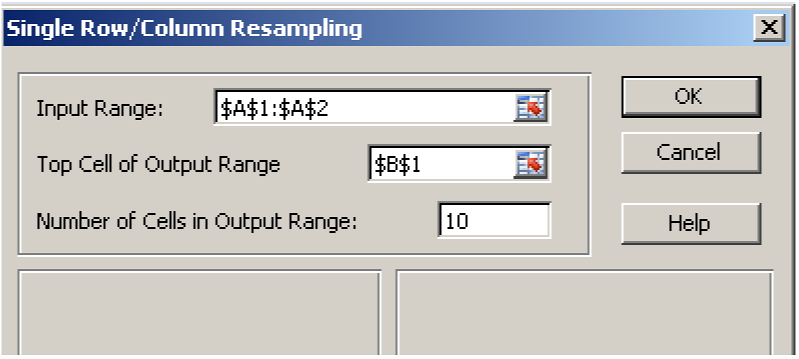
1) Open RSXL and enter “heads” in cell A1 and “tails” in cell A2.



	A
1	heads
2	tails
3	

Figure 1-14: RSXL coin toss worksheet

2) Select cell A1 and choose “Resample” from the Resampling menu (or “R” from the Resampling Toolbar). **Note** that the Resample option resamples the data WITH replacement. The Shuffle option (“S” from the Resampling Toolbar) resamples the data WITHOUT replacement and can’t be used in this particular example (Why? Think about it… we have only 2 data entries, heads and tails. If we don’t replace the data, we run out of data after 2 selections!). Continuing, select cell B1 as the Top Left Cell of Output Range and enter 10 as the Number of Cells in the Output Range. Click OK.



Input Range: $A$1:$A$2

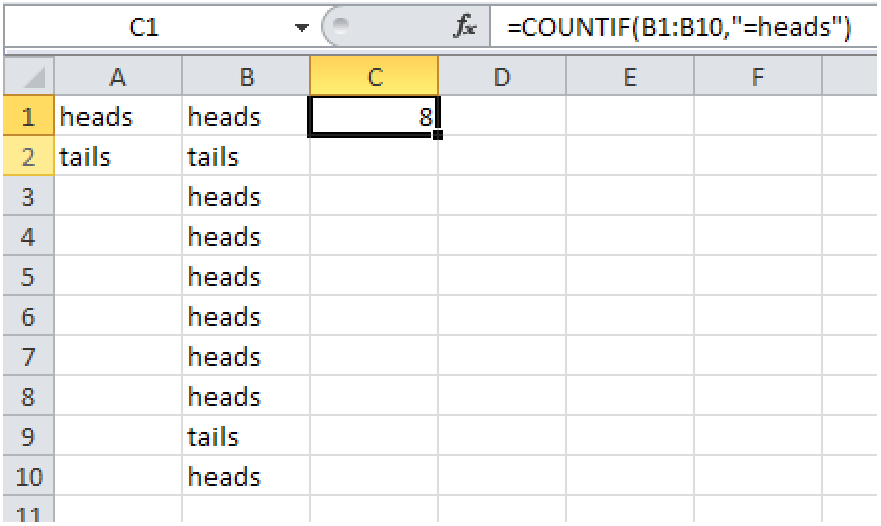
Top Cell of Output Range $B$1

Number of Cells in Output Range: 10

Figure 1-15: Resample dialog

3) In cell C1, enter =COUNTIF(B1:B10,"=heads")

<u>Note</u>: =COUNTIF() is one of the more useful native Excel functions. It allows you to search a range of cells (i.e. cells B1:B10 above) in an Excel worksheet for a particular value or condition (“=heads” in this case). The result is the count of the number of cells in the range that meet the condition. In this example the value returned is the number cells in range B1:B10 that contain “heads.” In Figure 1-16, there are 8 instances of “heads” in the range B1:B10.

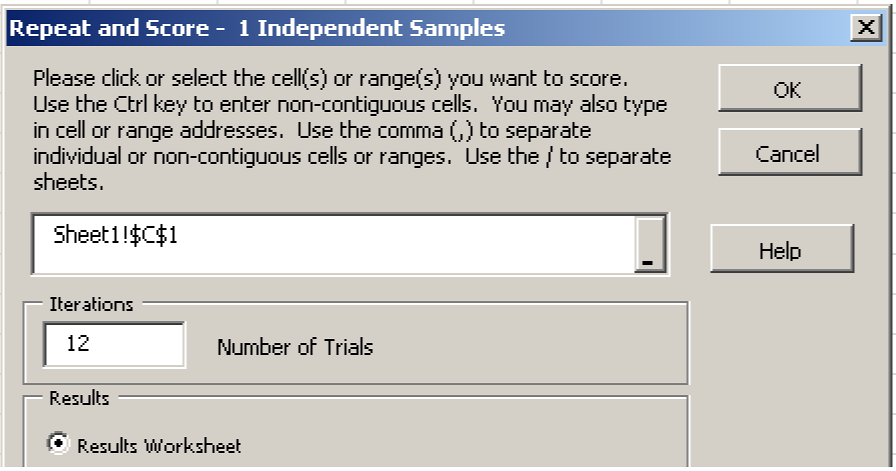


C1 =COUNTIF(B1:B10,"=heads")

	A	B	C	D	E	F
1	heads	heads	8			
2	tails	tails				
3		heads				
4		heads				
5		heads				
6		heads				
7		heads				
8		heads				
9		tails				
10		heads				

Figure 1-16: Results of initial resample

4) As indicated in figure 1-16 above, there are 8 heads in this particular trial, so that could count as 1 success. Select cell C1 and choose Repeat and Score from the Resampling menu (or RS from the Resampling toolbar). We want a total of 12 trials and we already have one successful trial. However, for the purposes of illustration, let's enter 12 as the number of trials and start from the beginning. Click OK and view the Results sheet.



Repeat and Score - 1 Independent Samples

Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

Sheet1!$C$1

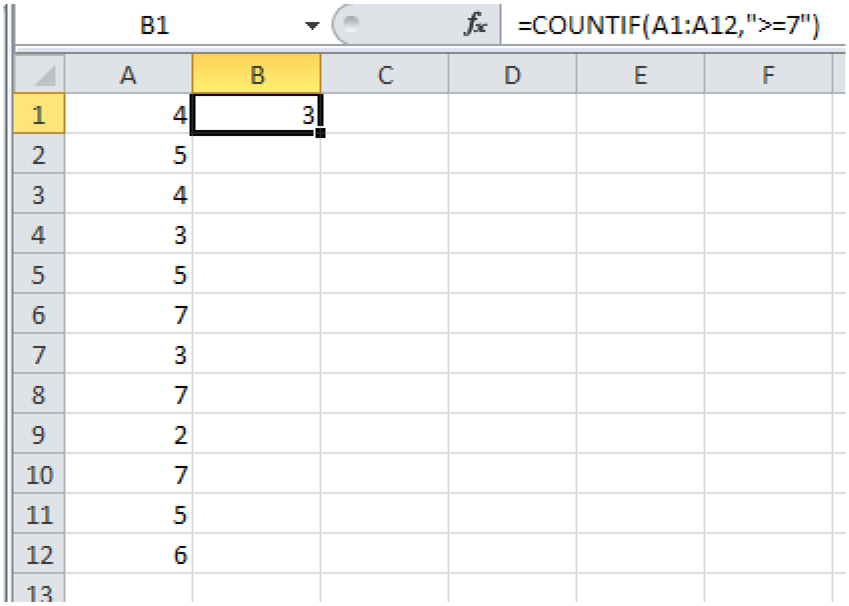
Iterations

12 Number of Trials

Results

● Results Worksheet

Figure 1-17: Repeat and Score dialog



B1 =COUNTIF(A1:A12,">=7")

	A	B	C	D	E	F
1	4	3				
2	5					
3	4					
4	3					
5	5					
6	7					
7	3					
8	7					
9	2					
10	7					
11	5					
12	6					
13						

Figure 1-18: Results worksheet

5) In figure 1-18, the values in cells A1:A12 represent the number of heads in each trial of 10 tosses. For example, in cell A1 (which represents the first trial of 10 tosses), there were 4 heads. In the second trial of 10 tosses, represented by cell A2, there were 5 heads. Trials 6, 8, and 10, represented by cells A6, A8, and A10, had a total of 7 heads each. Those 3 cells containing 7 heads represent the successes indicated by the =COUNTIF(A1:A12,">=7") function in cell B1.

   Note: We are interested in the number of instances of 7 or more heads, which means that we would accept trials of 7, 8, 9, or 10 as successes. The =COUNTIF(A1:A12, ">=7") means that we want Excel to count the number times 7 or more (">=7") appears in the output range A1:A10. In Figure 1-18, there are 3 instances of 7 or more heads.

6) The number of tails in each trial of 10 tosses is not shown. However, this is easy to calculate. For example, in cell A12, the total number of heads is 6, so the number of tails is 10 – 6 = 4.

7) As mentioned in step 5, the =COUNTIF(A1:A12,">=7") in cell B1 tells us that in this simulation, we had 3 successes or instances where there were 7 or more heads.

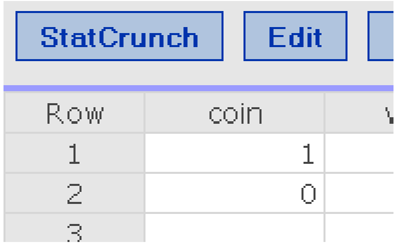
   Note: We could have also used "1" for heads and "0" for tails. We'll use this notation for StatCrunch (although StatCrunch can also handle string or word notation).

StatCrunch

A special thanks to Webster West, the creator of StatCrunch, for providing ideas and hints on how to use StatCrunch to solve many of the exercises in the text.

Note: StatCrunch uses the term "sample" to refer both to the original (real) sample, and to the simulated/resampled samples.

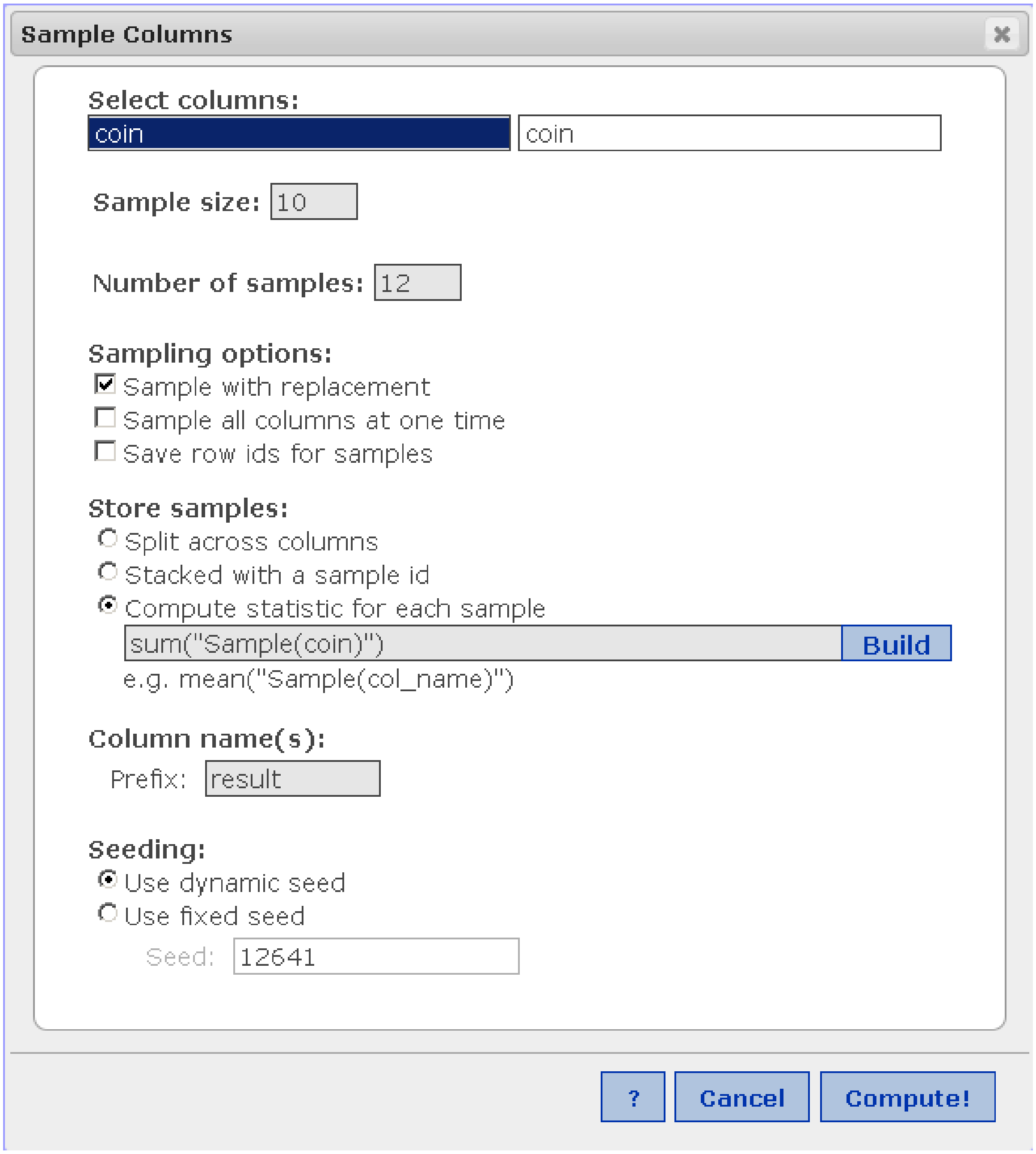
1) Open StatCrunch and re-label the first data column by clicking on the column name (to highlight the label box) and using the Backspace or Delete key to erase the existing label. Type in the new column label (coin) and enter the following (1 = heads, 0 = tails) in rows 1 and 2 as shown in figure 1-19:



Row	coin
1	1
2	0
3	

Figure 1-19: StatCrunch initial worksheet

2) Select Data>Sample and choose the coin column. Enter a Sample size of 10, check Sample with Replacement (we need to re-use the coin!), and enter 12 as the Number of samples. In the Store Samples section, select Compute statistic for each sample. The Statistic is: sum("Sample(coin)") as shown below. With 0/1 data, finding the sum is the same as counting the number of 1's, so the sum function effectively tallies the number of 1's (heads) in each trial. The number of tails will not be calculated or displayed in the results. Enter "result" for the Column name and click Compute!



Select columns:
coin
coin

Sample size: 10

Number of samples: 12

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

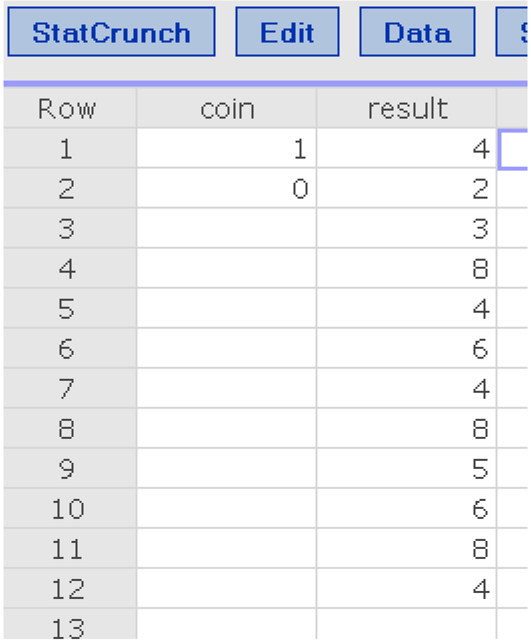
Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
sum("Sample(coin)")
e.g. mean("Sample(col_name)")

Column name(s):
Prefix: result

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 1-20: Sample Columns dialog

3) The result of one set of 12 trials is shown below. Note that there are three instances of 8 heads in Rows 4, 8 and 11 (and by subtraction, 10 – 8 = 2 tails in each of those rows).

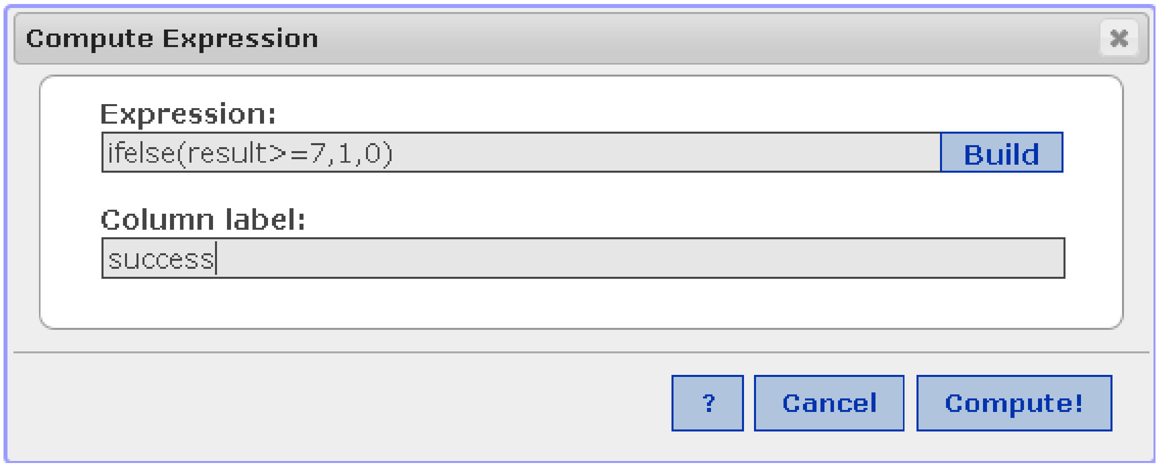


Row	coin	result
1	1	4
2	0	2
3		3
4		8
5		4
6		6
7		4
8		8
9		5
10		6
11		8
12		4
13		

Figure 1-21: Results of 12 trials of 10 coin tosses

4) As in the Resampling Stats procedure above, each Row represents one trial of 10 tosses. Each value in the result column of figure 1-21 represents the total number of heads in 10 tosses of the trial represented by that row. Row 1, for example, represents the first trial of 10 tosses which had a total of 4 heads. There are 12 trials of 10 tosses each, so we have 12 Rows of results.

5) Although the example has only 12 trials and we can easily see the three trials where 8 heads occurred, with a large number of trials we might want to have StatCrunch count how often we got >=7 heads. Select Data>Compute Expression and enter the expression: ifelse(result>=7,1,0) in the Expression area. This assumes the column name of the results of the trials is “result”. The interpretation of the expression is: If the value in each result row is greater than or equal to 7, then enter a 1, otherwise (else) enter a 0. Enter a New column name of “success” and click Compute!

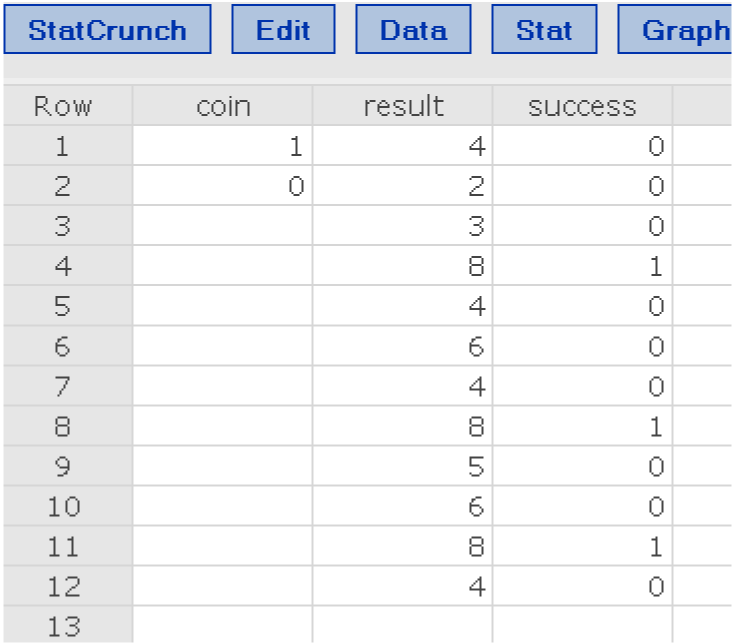


Expression: ifelse(result>=7,1,0)

Column label: success

Figure 1-22: Compute expression dialog

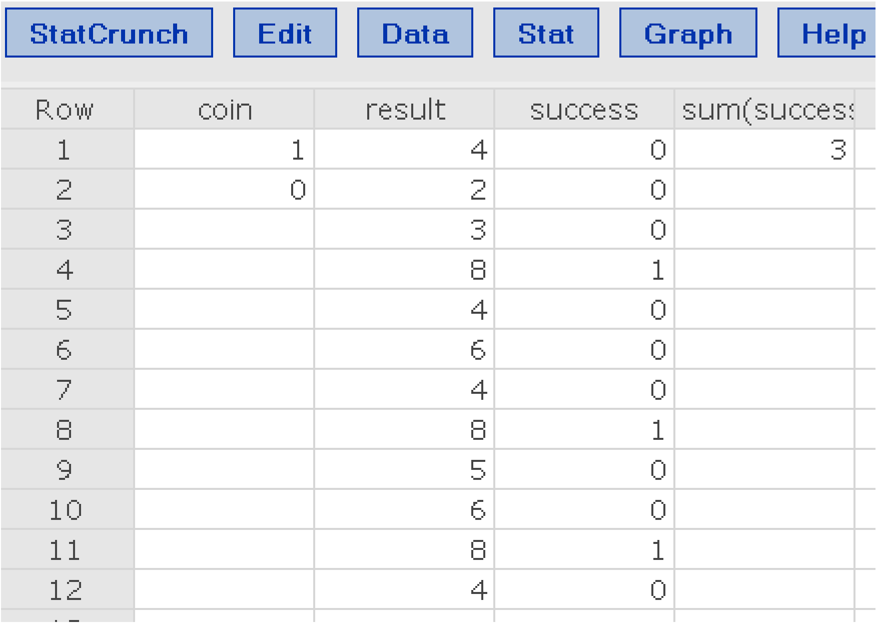
6) The result based on our simulation is:



Row	coin	result	success
1	1	4	0
2	0	2	0
3		3	0
4		8	1
5		4	0
6		6	0
7		4	0
8		8	1
9		5	0
10		6	0
11		8	1
12		4	0
13			

Figure 1-23: Marking successes (heads >= 7)

7) Note that the 1s in the success column are next to the 8s in the result column. To tally the number of successes (the number of times we get 7 or more heads in 10 tosses), sum the success column. Choose Data>Compute Expression and enter: sum(success) in the Expression box. Click Compute! and you should see something similar to the figure below:



Row	coin	result	success	sum(success
1	1	4	0	3
2	0	2	0	
3		3	0	
4		8	1	
5		4	0	
6		6	0	
7		4	0	
8		8	1	
9		5	0	
10		6	0	
11		8	1	
12		4	0	

Figure 1-24: Tallying the number of successes

8) Your results will most likely be slightly different due to the nature of resampling.

Try It Yourself

Let's double the sample size and imagine that the study had revealed 14 errors in one year and 6 the following (instead of 7 and 3). Now regroup your twelve simulations of ten tosses each into six trials of 20 tosses each. Combine the first and second sets, the third and fourth, etc. Then do six more trials of 20 tosses each for a total of 120 additional tosses. You should now have twelve sets of 20 tosses. If you want to try a computer simulation, click <here> to download a Box Sampler macro-enabled Excel workbook.

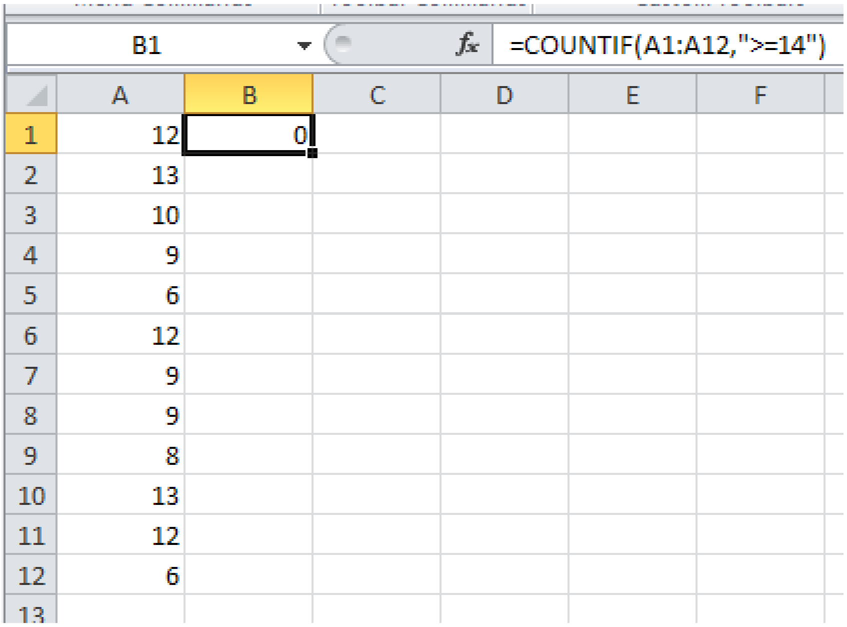
Section 1.2 of the Textbook Supplement contains a Resampling Stats procedure for this problem.

Did you ever get fourteen (or more) heads in a trial of twenty tosses?

Resampling Stats for Excel

1) Recreate the Resampling Stats procedure from the previous exercise, except use 20 as the number of cells in the output range. In cell C1 enter: =COUNTIF(B1:B20,"=heads")

2) Repeat and score on cell C1 and use 12 trials as before.



B1 =COUNTIF(A1:A12,">=14")

	A	B	C	D	E	F
1	12	0				
2	13					
3	10					
4	9					
5	6					
6	12					
7	9					
8	9					
9	8					
10	13					
11	12					
12	6					
13						

Figure 1-25: 12 trials of 20 tosses

3) In this simulation, there were no instances of 14 or more heads in 12 trials.

1.6 What to Measure – Variability

Variance and standard deviation for a sample

Try It Yourself

(optional) In your resampling software, randomly generate a "population" of 1000 values. It doesn't matter what population you generate - let's say a population of randomly selected numbers between 0 and 9. In Excel you can do this with the RANDBETWEEN function. Next, find the variance of this population using the "population" variance formula. Then repeatedly take resamples of size 10 and calculate the variance for each resample according to the same "population" formula. How does the mean of the resample variances compare to the "population" variance?

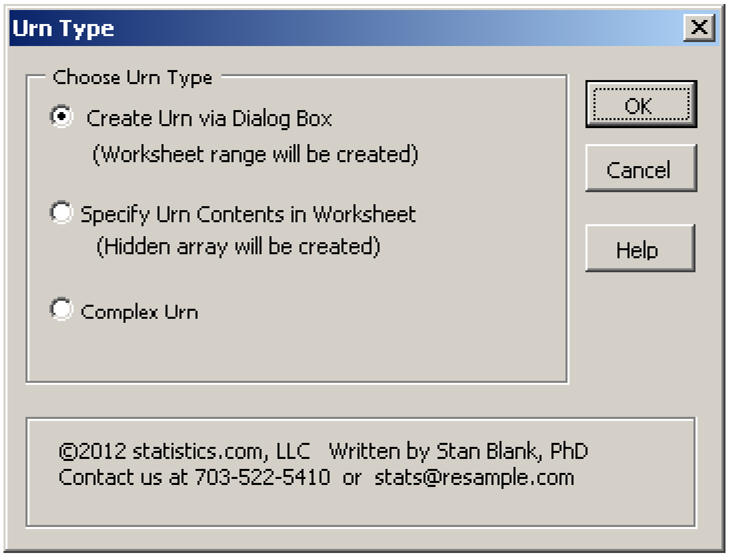
Tutorials for this exercise using Resampling Stats for Excel and StatCrunch can be found in Section 1.6 of the Textbook Supplement.

For a Box Sampler resampling tutorial based on this exercise, click <here>.

Resampling Stats for Excel

1) We will need 1000 random values between 0 and 9 inclusive. One way to do this is to enter the formula =RANDBETWEEN(0, 9) in cell A1 and then copying this formula down the column by clicking on the AutoFill box/button in the lower right-hand corner of cell A1 and dragging downward to cell A1000. However, there is a much easier method using the RSXL Urn feature. The Urn feature is specifically designed to reduce the tedium of entering multiple values of the same kind in a worksheet.

2) Open RSXL and select the Urn feature from the Add-ins>Resampling menu (in Excel 2003 use Tools>Resampling or the floating toolbar) or from the Resampling toolbar in the Add-ins menu.



Choose Urn Type

● Create Urn via Dialog Box
(Worksheet range will be created)

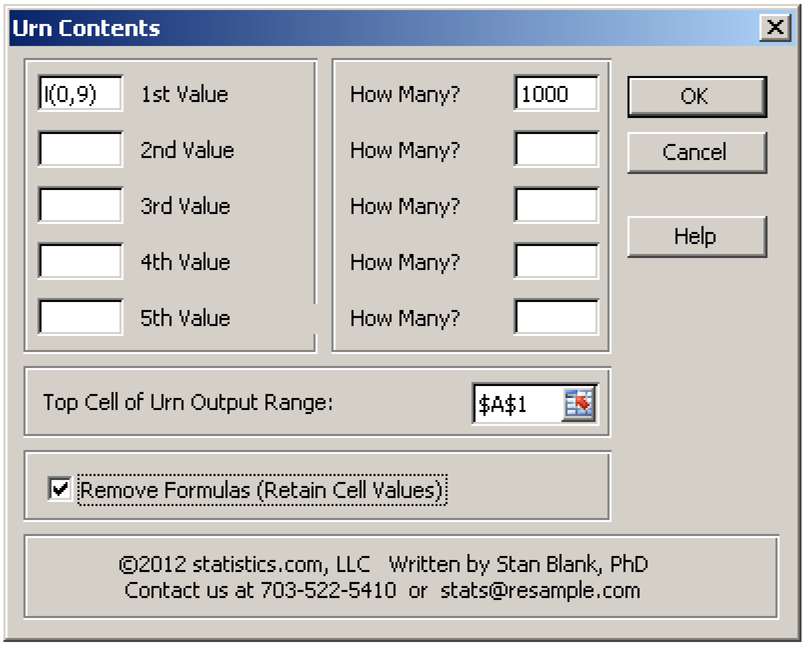
○ Specify Urn Contents in Worksheet
(Hidden array will be created)

○ Complex Urn

©2012 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 1-26: Initial Urn dialog

3) Use the top Urn option and click OK. In the Urn Contents dialog, enter the formula =RANDBETWEEN(0,9) in the 1st Value box and 1000 in the corresponding "How Many?" box. Click in the "Top Left Cell of Urn Output Range" edit box and select cell $A$1. Check the "Remove Formulas (Retain Cell Values)" box as shown below and click OK.



I(0,9) 1st Value — How Many? 1000

2nd Value — How Many?

3rd Value — How Many?

4th Value — How Many?

5th Value — How Many?

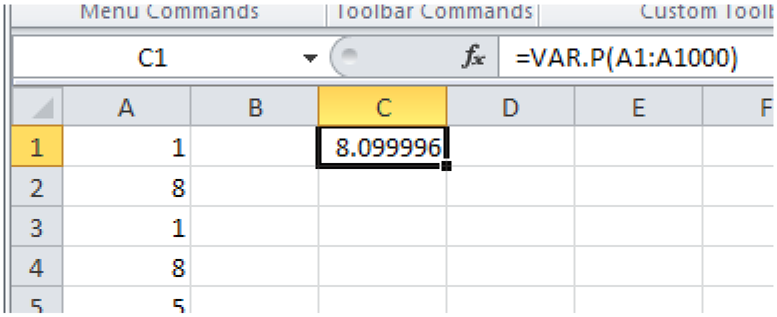
Top Cell of Urn Output Range: $A$1

☑ Remove Formulas (Retain Cell Values)

©2012 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 1-27: Filling the Urn

4) You should see column A fill with 1000 integers from 0 to 9. Take the population variance of this data by entering the formula =VAR.P(A1:A1000) in cell C1. If you are using Excel 2007 or Excel 2003, use =VARP(A1:A1000). The worksheet should look like the following:

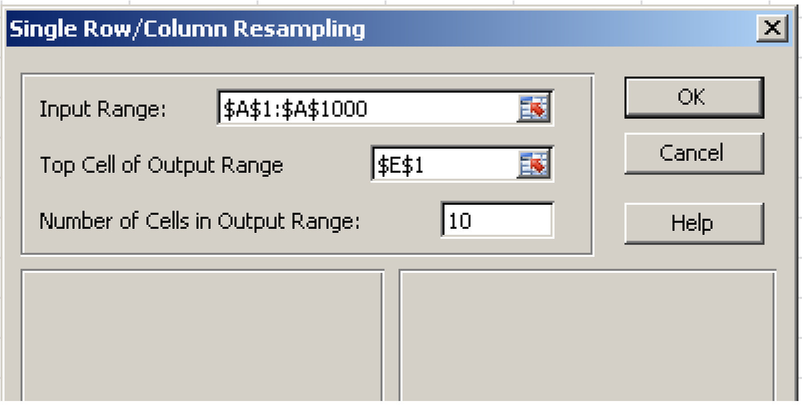


C1 =VAR.P(A1:A1000)

	A	B	C	D	E	F
1	1		8.099996			
2	8					
3	1					
4	8					
5	5					

Figure 1-28: Population variance: Native Excel function

5) Select cell A1 and choose the Resample menu from the Add-ins>Resampling menu or the Add-ins>Resampling floating toolbar. Choose cell E1 as the Top Left cell of the Output Range and 10 as the Number of Trials.


Single Row/Column Resampling

Input Range: $A$1:$A$1000

Top Cell of Output Range $E$1

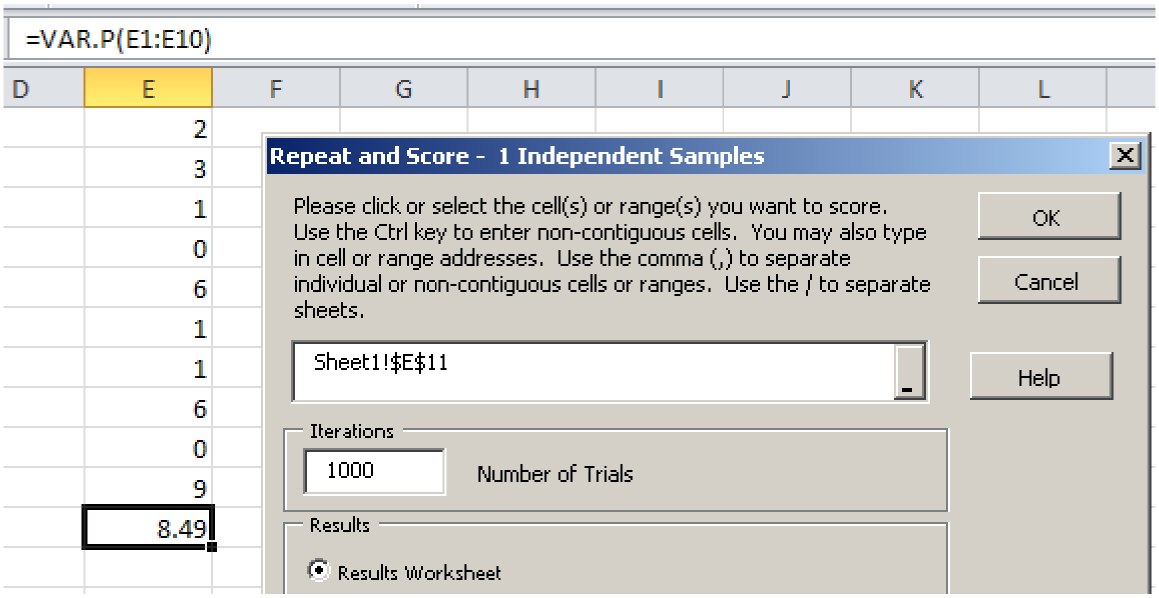
Number of Cells in Output Range: 10

Figure 1-29: Resampling dialog

6) Click OK. In cell E11, enter the formula =VAR.P(E1:E10) to calculate the sample variance (for Excel 2007 and 2003, use =VARP(E1:E10).

7) You should use the VARP (population variance) formula you used previously for the sample variance in step 5 - the point of this exercise is to show that simply using the population variance formula for a sample underestimates the population variance.

8) Select the Repeat and Score feature from the Resampling menu. Choose cell E11 as the Score cell and 1000 as the number of iterations. Click OK.


=VAR.P(E1:E10)

D	E
	2
	3
	1
	0
	6
	1
	1
	6
	0
	9
	8.49

Repeat and Score - 1 Independent Samples

Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

Sheet1!$E$11

Iterations

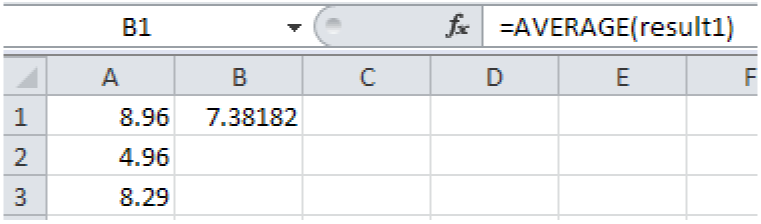
1000 Number of Trials

Results

● Results Worksheet

Figure 1-30: Repeat and Score on sample variance

9) Find the average of the values on the Results sheet by using the formula =AVERAGE(result1)



B1 =AVERAGE(result1)

	A	B	C	D	E	F
1	8.96	7.38182				
2	4.96					
3	8.29					

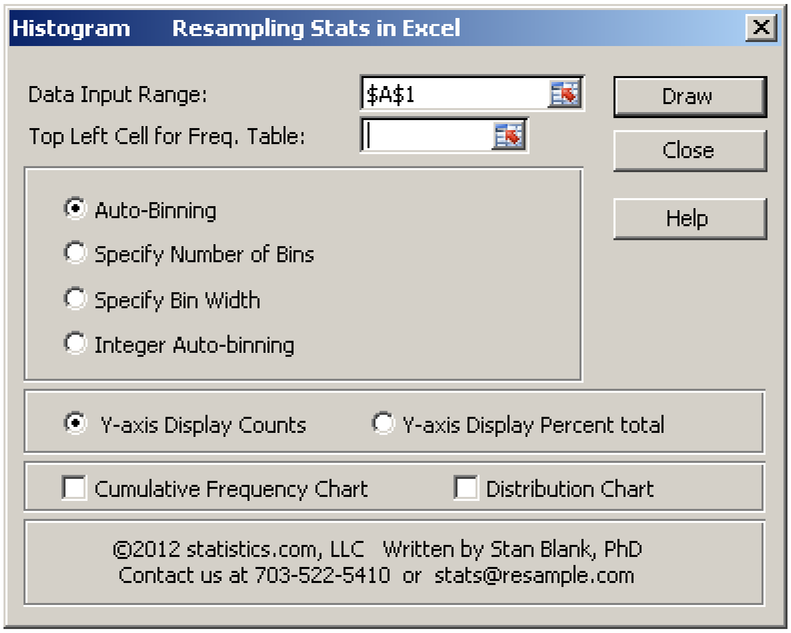
Figure 1-31: Calculating the average of 1000 sample variances

10) The average of 1000 sample variances (using the Excel population variance function VAR.P) underestimates the population variance by 8.100 – 7.380 or about 0.72.

Don't erase the worksheet yet… let's make some histograms first!

Create a Histogram with RSXL

Histograms are very important statistical tools. RSXL has the capability of creating histograms with a minimum of effort. Using the Results worksheet from the previous example, erase the contents of cell B1 while keeping the values in cells A1:A1000. Select cell A1 and choose the histogram feature from the Resampling menu.



Histogram Resampling Stats in Excel

Data Input Range: $A$1

Top Left Cell for Freq. Table:

● Auto-Binning
○ Specify Number of Bins
○ Specify Bin Width
○ Integer Auto-binning

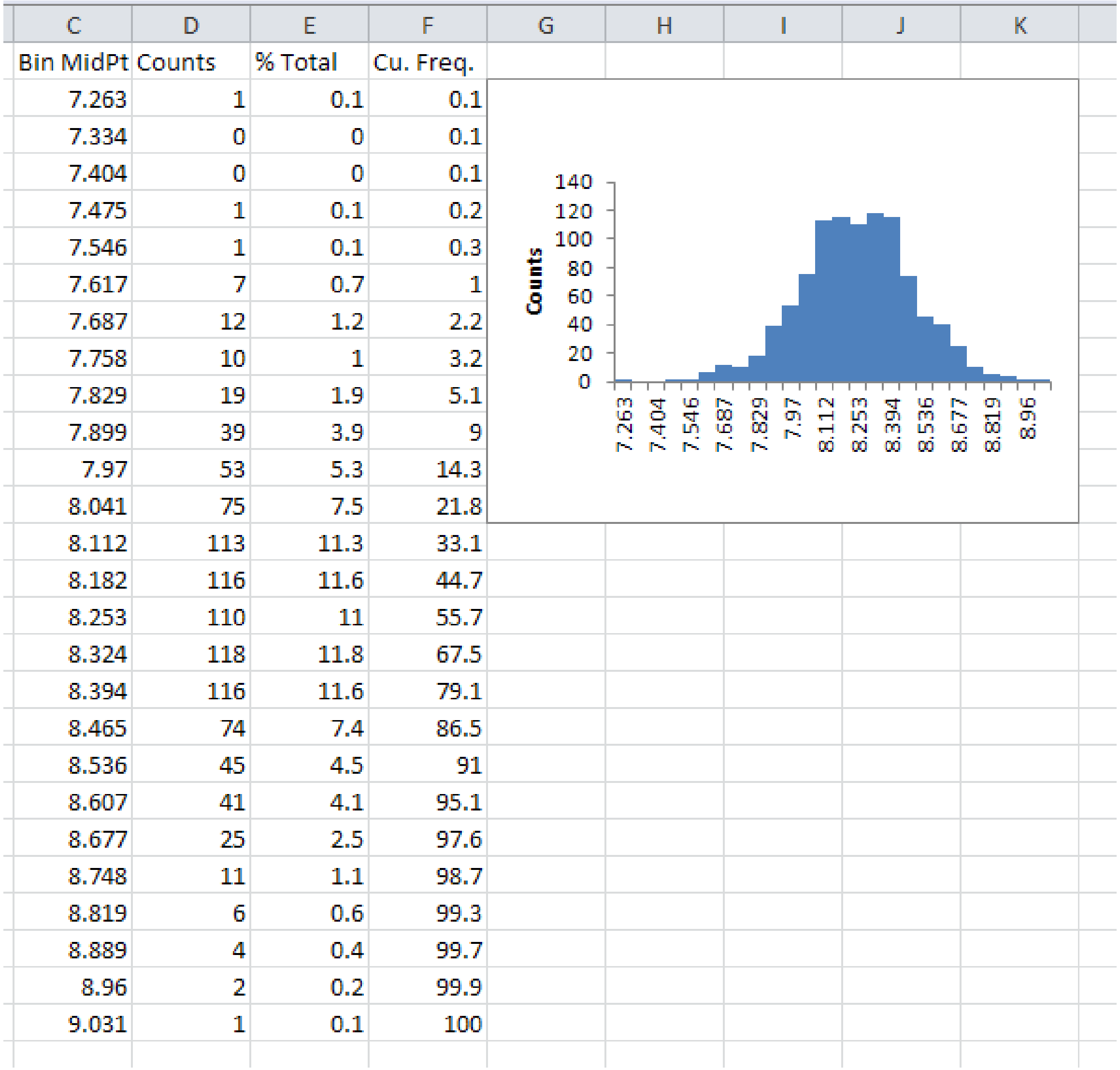
● Y-axis Display Counts ○ Y-axis Display Percent total

☐ Cumulative Frequency Chart ☐ Distribution Chart

©2012 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 1-32: Histogram dialog

Click in the "Top Left Cell…" edit box and choose cell C1. Keep all other settings the at their default values and click "Draw." A sample result is in figure 1-33.



Bin MidPt	Counts	% Total	Cu. Freq.
7.263	1	0.1	0.1
7.334	0	0	0.1
7.404	0	0	0.1
7.475	1	0.1	0.2
7.546	1	0.1	0.3
7.617	7	0.7	1
7.687	12	1.2	2.2
7.758	10	1	3.2
7.829	19	1.9	5.1
7.899	39	3.9	9
7.97	53	5.3	14.3
8.041	75	7.5	21.8
8.112	113	11.3	33.1
8.182	116	11.6	44.7
8.253	110	11	55.7
8.324	118	11.8	67.5
8.394	116	11.6	79.1
8.465	74	7.4	86.5
8.536	45	4.5	91
8.607	41	4.1	95.1
8.677	25	2.5	97.6
8.748	11	1.1	98.7
8.819	6	0.6	99.3
8.889	4	0.4	99.7
8.96	2	0.2	99.9
9.031	1	0.1	100

Figure 1-33: Histogram results

RSXL will make a valiant attempt to find suitable bin ranges in order to plot a proper histogram. Note the columns that represent bin midpoints, the frequency counts in each bin, the percentage of the total in each bin, and the cumulative frequency.

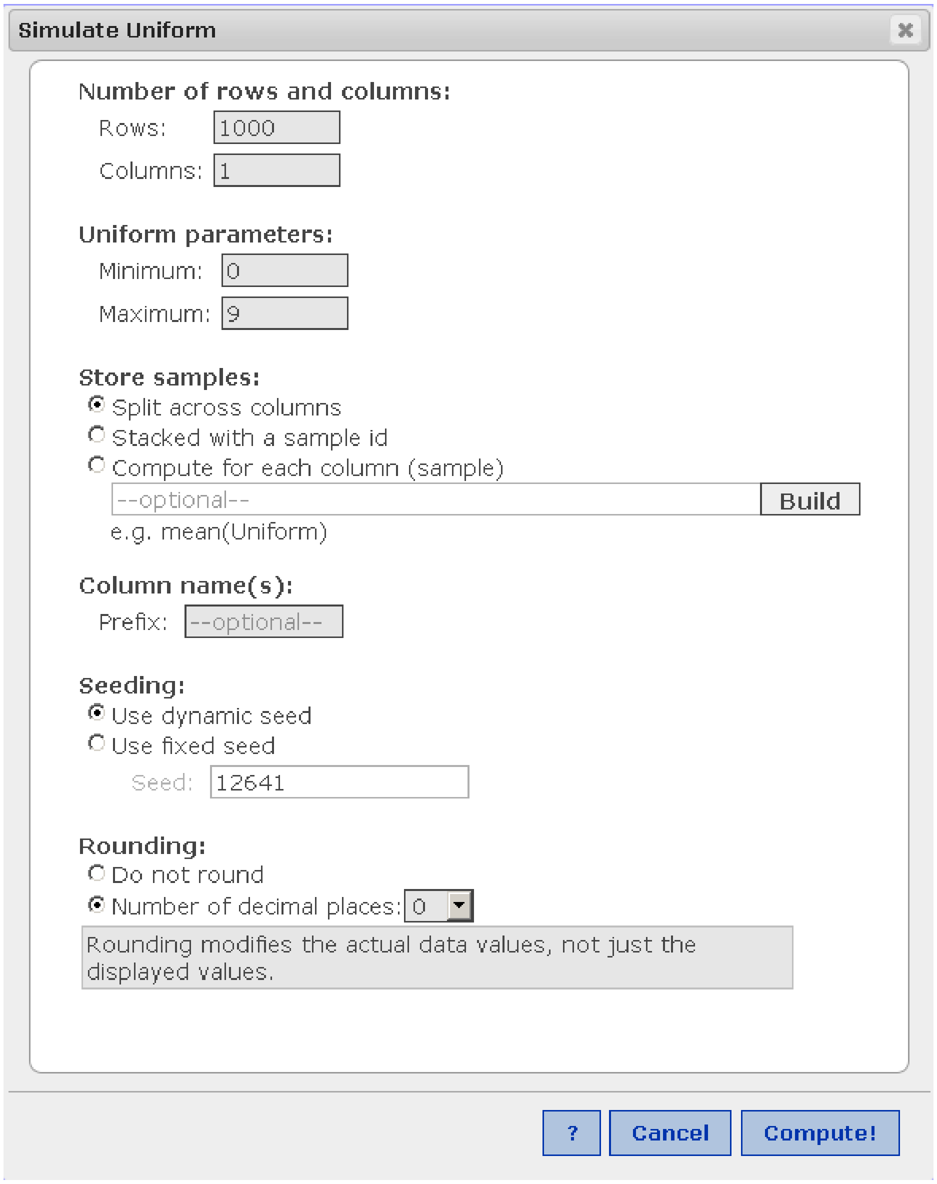
Creating a histogram is both art and science. The RSXL histogram feature allows you to adjust the bin widths, the number of bins, and you can select integer binning for integral values if desired.

There are more examples of using the histogram feature later in the supplement. Also, the Box Sampler histogram feature is identical to RSXL.

StatCrunch

Likewise, this procedure needs to use the StatCrunch statistic "unadj variance" throughout the exercise for the same reason as noted above.

1) Open StatCrunch and select Data>Simulate>Uniform.



Simulate Uniform

Number of rows and columns:
Rows: 1000
Columns: 1

Uniform parameters:
Minimum: 0
Maximum: 9

Store samples:
● Split across columns
○ Stacked with a sample id
○ Compute for each column (sample)
--optional--
e.g. mean(Uniform)

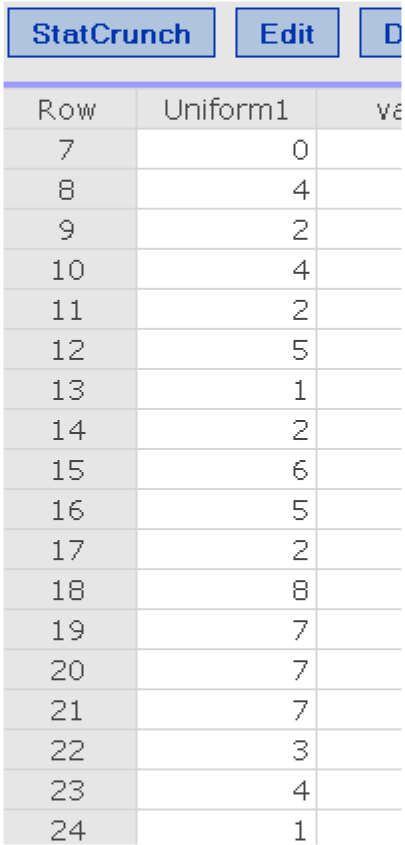
Column name(s):
Prefix: --optional--

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Rounding:
○ Do not round
● Number of decimal places: 0
Rounding modifies the actual data values, not just the displayed values.

Figure 1-34: StatCrunch Uniform samples

2) Enter 1000 for the number of rows and 1 for the number of columns. The Uniform Parameters will be from a minimum of 0 to a maximum of 9. Under Rounding, set the number of decimal places to 0. Click Compute! and a worksheet similar to the following should appear:

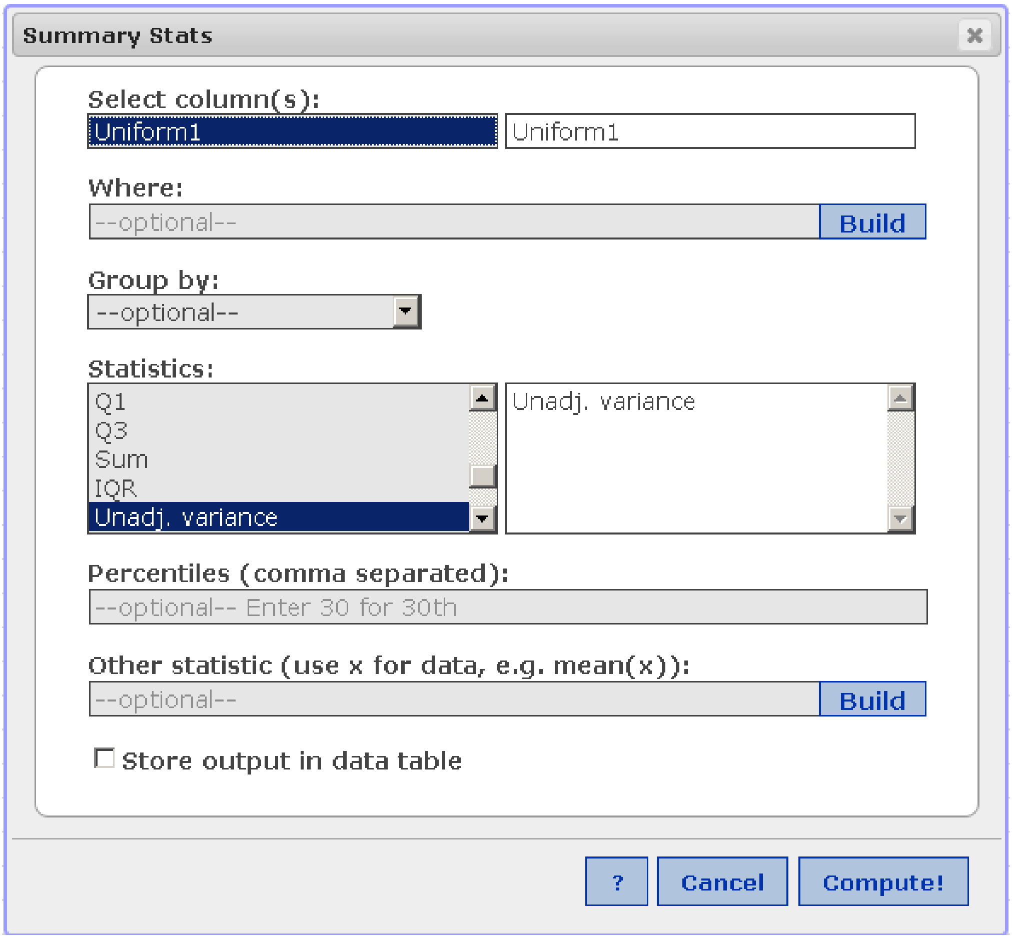


Row	Uniform1	va
7	0	
8	4	
9	2	
10	4	
11	2	
12	5	
13	1	
14	2	
15	6	
16	5	
17	2	
18	8	
19	7	
20	7	
21	7	
22	3	
23	4	
24	1	

Figure 1-35: Uniform distribution from 0 to 9 inclusive

3) We have generated a random list of 1000 numbers between 0 and 9 inclusive as shown in figure 1-35.

4) To find the population variance, select Stat>Summary Stats>Columns and choose the Uniform1 column. Select Unadj. Variance (as in figure 1-36). Click Compute!



Select column(s): Uniform1 | Uniform1

Where: --optional--

Group by: --optional--

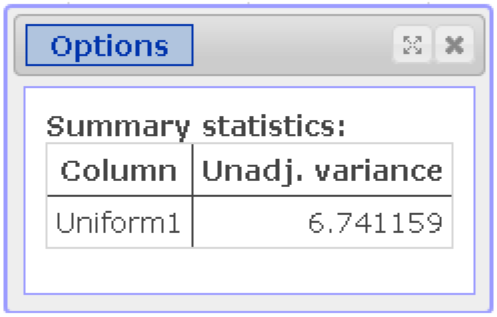
Statistics: Q1, Q3, Sum, IQR, Unadj. variance | Unadj. variance

Percentiles (comma separated): --optional-- Enter 30 for 30th

Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 1-36: Unadj. Variance (population variance)



Summary statistics:

Column	Unadj. variance
Uniform1	6.741159

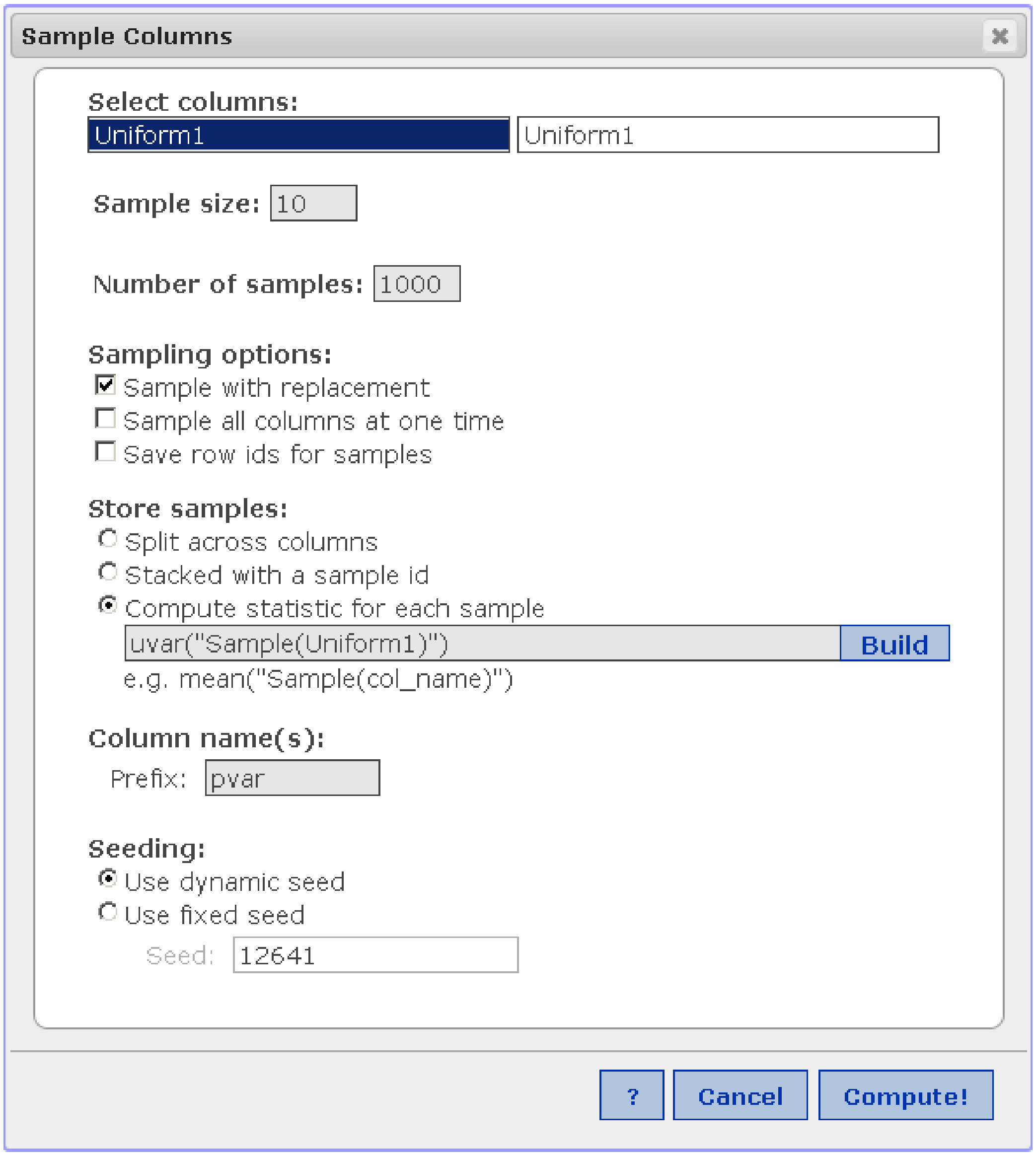
Figure 1-37: Population variance

5) The population variance for the Uniform1 column is 6.741159.

6) To calculate the mean of 1000 sample variances (sample size = 10) using the population variance formula, choose Data>Sample and select the Uniform1 column. Enter a Sample Size of 10, enter 1000 for the number of samples and check the Sample with replacement box. The following “Compute statistic…” expression calculates the population variance (uvar = unadjusted variance) of each sample (resample) taken from Uniform1:

   ```
   uvar("Sample(Uniform1)")
   ```

   Change the name of the new column to “pvar” or something similar and click Compute! You should see a new column of 1000 sample variances.



Select columns: Uniform1

Sample size: 10

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

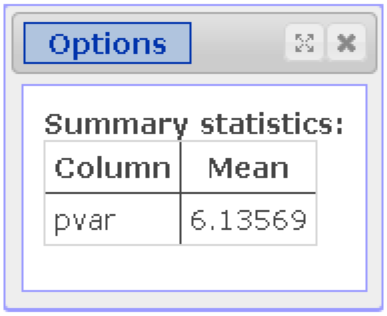
Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
uvar("Sample(Uniform1)")
e.g. mean("Sample(col_name)")

Column name(s):
Prefix: pvar

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 1-38: Calculating variance of each of 1000 samples

7) Find the mean or average of the "pvar" column by selecting Stat>Summary Stats>Columns, choose the pvar column and select the mean. Click Compute!



Summary statistics:

Column	Mean
pvar	6.13569

Figure 1-39: Mean of sample variances

8) The mean of the unadjusted variances of the samples is 6.13569 for an underestimate of 0.605469 compared to the population variance of the original data set.

1.7 What to Measure - Distance (nearness)

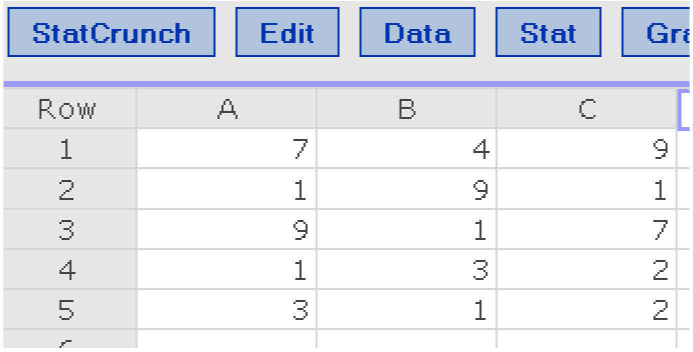
The question exercise posed in section 1.7 of the text involves calculating the Euclidean distance between two vectors. The answer to the question was an Excel workbook with the formulas and calculations on display. Let’s follow with the StatCrunch solution to the problem.

Question: *Let's say that you own a music store, A, B and C are all customers of yours, and that A and B have both just made purchases. You want to recommend one of these purchases for C. Which one would you recommend?*

ANSWER

StatCrunch

1) Open StatCrunch and enter the vectors for customers A, B and C (in StatCrunch, the vectors will be in columns instead of rows) as shown in figure 1-40.

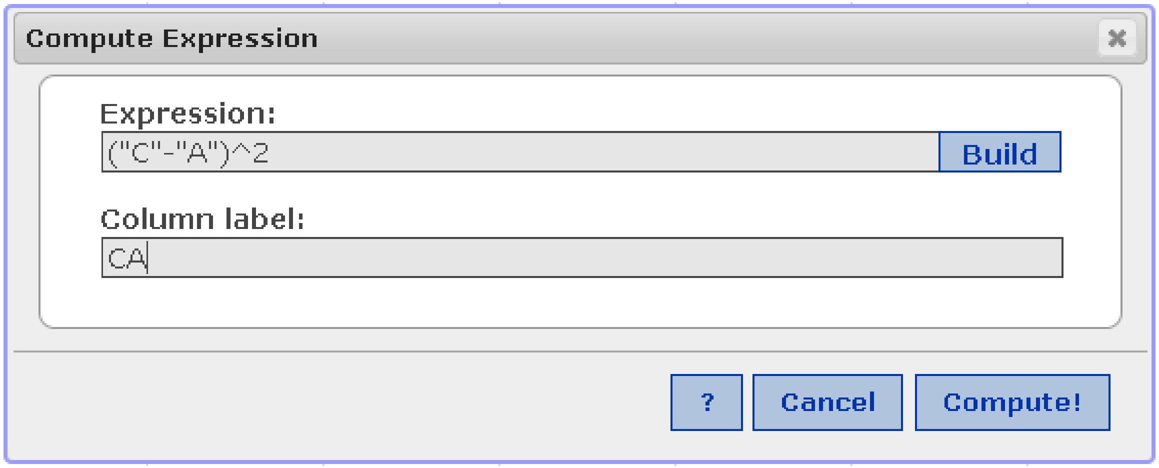


Row	A	B	C
1	7	4	9
2	1	9	1
3	9	1	7
4	1	3	2
5	3	1	2

Figure 1-40: Vectors for customers A, B and C

2) Select Data>Compute Expression and enter (“C”-“A”)^2 for the expression. This formula subtracts each element in vector A from the corresponding element in vector C and squares each individual result. The New column name should be CA. Click Compute! (figure 1-41).

Note: Quotes are needed for column/vector labels unless the label is enclosed by another command. The expression: mean(A) is OK and calculates the mean of column A. The expression A-B will NOT work because StatCrunch would attempt to interpret both A and B as built-in commands. To avoid an error, individual column labels must be entered as “A” - “B”.

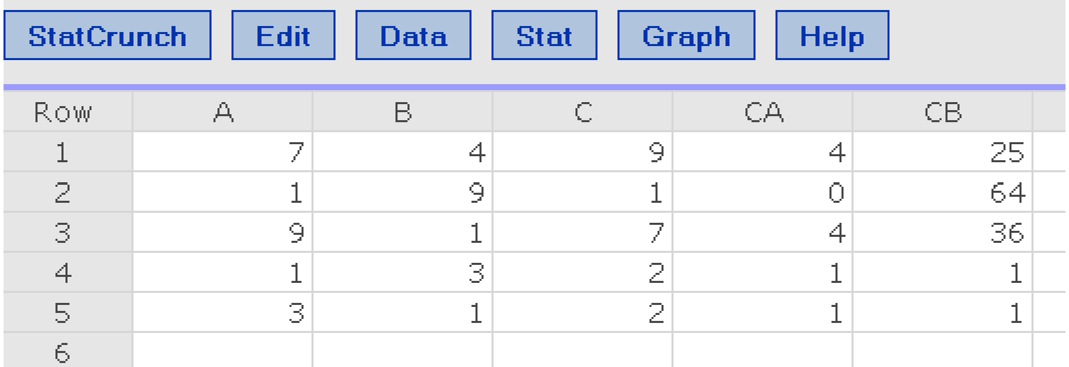


Expression: ("C"-"A")^2

Column label: CA

Figure 1-41: Squared differences of vector C – vector A

3) Repeat step 2, except use the expression (“C”-“B”)^2 and the New column name should be CB. Click Compute! Your StatCrunch worksheet should look like figure 1-42.

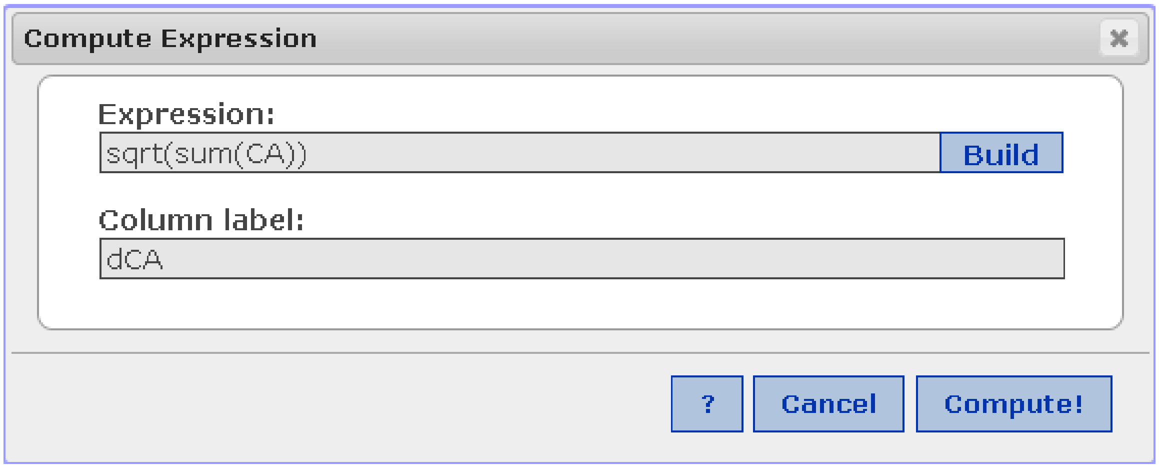


Row	A	B	C	CA	CB
1	7	4	9	4	25
2	1	9	1	0	64
3	9	1	7	4	36
4	1	3	2	1	1
5	3	1	2	1	1
6					

Figure 1-42: Squared differences

4) To find the square root of the sum of the squared differences (the Euclidean distance) of the CA and CB columns, again choose Data>Compute Expression. Enter the expression: sqrt(sum(CA)) to calculate the square root of the sum of

column/vector CA. The new column name should be dCA (figure 1-43). Click Compute.



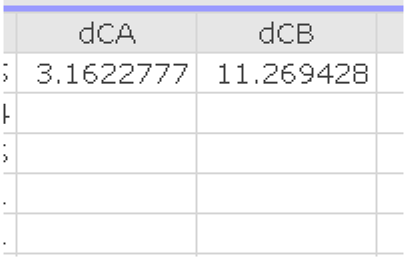
Expression: sqrt(sum(CA))

Column label: dCA

Figure 1-43: Calculate the square root of the squared difference in column/vector CA

5) Repeat step 4, using sqrt(sum(CB)) to calculate the square root of the squared differences in column/vector CB. The new column name should be dCB. Click Compute.

6) The final result is shown in figure 1-44. We should recommend customer A's music choices to customer C. The Euclidean distance between C and A is 3.16, while the distance between C and B is 11.27.



dCA	dCB
3.1622777	11.269428

Figure 1-44: Results: Recommend customer A's choices to customer C

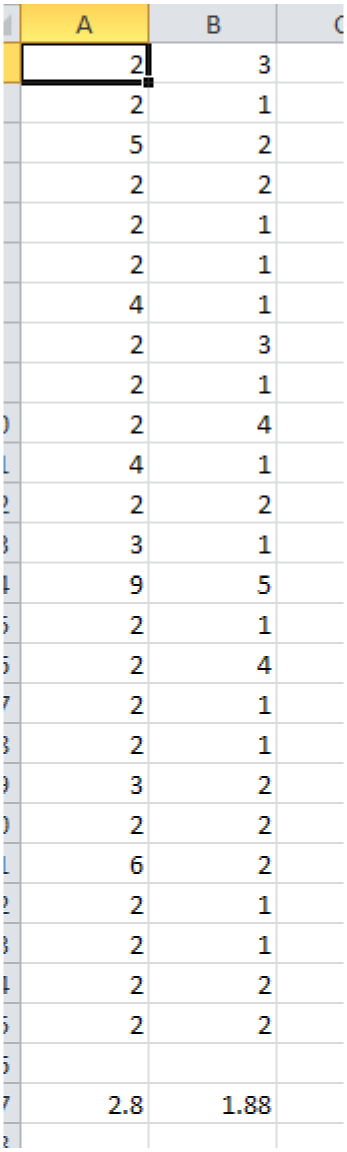
2 Statistical Inference

2.1 Repeating the Experiment

Resampling Stats

We can try the medical errors example ourselves using Resampling Stats for Excel. You can enter the data from table 1-10 manually or you can download the workbook <here>.

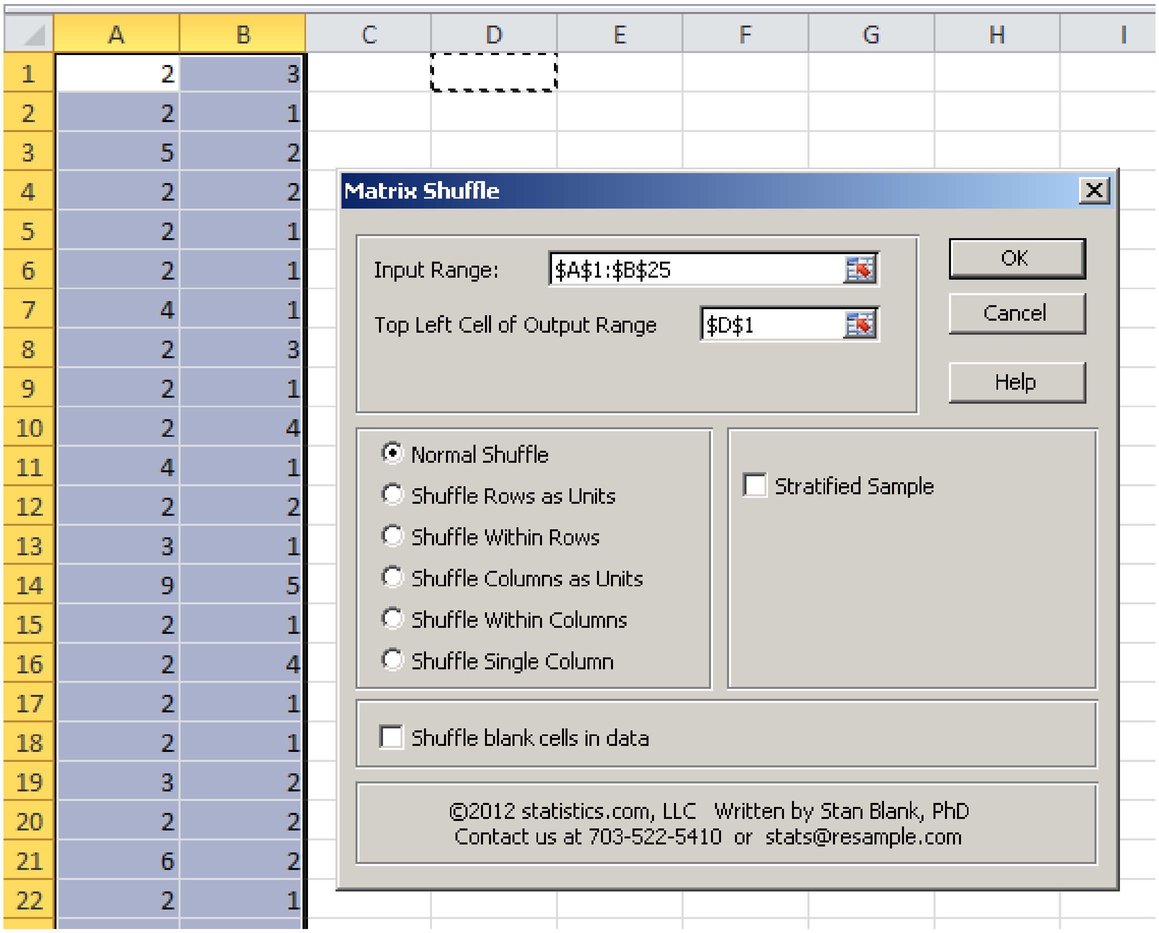
1) Figure 2-1 illustrates the medical error data



A	B
2	3
2	1
5	2
2	2
2	1
2	1
4	1
2	3
2	1
2	4
4	1
2	2
3	1
9	5
2	1
2	4
2	1
2	1
3	2
2	2
6	2
2	1
2	1
2	2
2	2
2.8	1.88

Figure 2-1: Medical errors data from Table 1-10

2) Select cell A1 and choose Shuffle from the Resampling menu; Resampling Stats auto-completes the rest of the range to be shuffled. Note that the Shuffle option resamples the data WITHOUT replacement.



	A	B
1	2	3
2	2	1
3	5	2
4	2	2
5	2	1
6	2	1
7	4	1
8	2	3
9	2	1
10	2	4
11	4	1
12	2	2
13	3	1
14	9	5
15	2	1
16	2	4
17	2	1
18	2	1
19	3	2
20	2	2
21	6	2
22	2	1

Matrix Shuffle

Input Range: $A$1:$B$25

Top Left Cell of Output Range $D$1

● Normal Shuffle
○ Shuffle Rows as Units
○ Shuffle Within Rows
○ Shuffle Columns as Units
○ Shuffle Within Columns
○ Shuffle Single Column

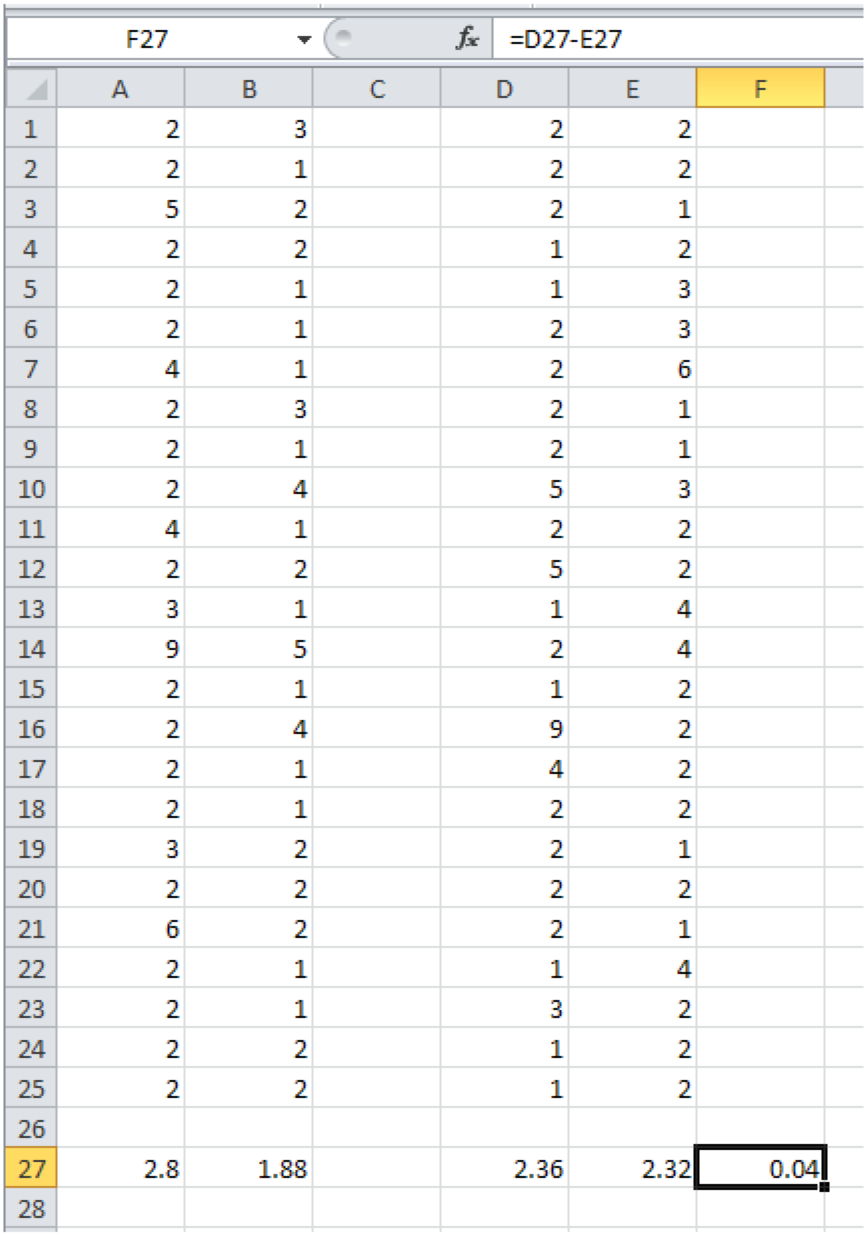
☐ Stratified Sample

☐ Shuffle blank cells in data

©2012 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 2-2: Matrix shuffle

3) Select cell D1 as the Output cell and choose a Normal Shuffle (which takes all the values in the range, shuffles them, then places the shuffled values in a new range with the same number of columns and rows). Click OK. Find the average of each new column and calculate the difference in the means as shown in Figure 2-3.

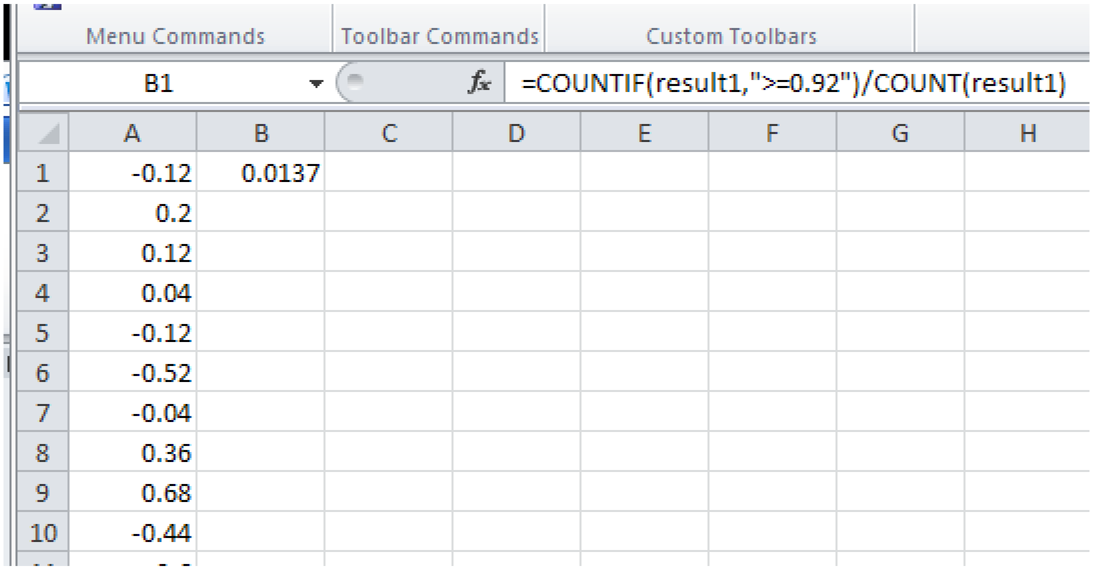


F27 =D27-E27

	A	B	C	D	E	F
1	2	3		2	2	
2	2	1		2	2	
3	5	2		2	1	
4	2	2		1	2	
5	2	1		1	3	
6	2	1		2	3	
7	4	1		2	6	
8	2	3		2	1	
9	2	1		2	1	
10	2	4		5	3	
11	4	1		2	2	
12	2	2		5	2	
13	3	1		1	4	
14	9	5		2	4	
15	2	1		1	2	
16	2	4		9	2	
17	2	1		4	2	
18	2	1		2	2	
19	3	2		2	1	
20	2	2		2	2	
21	6	2		2	1	
22	2	1		1	4	
23	2	1		3	2	
24	2	2		1	2	
25	2	2		1	2	
26						
27	2.8	1.88		2.36	2.32	0.04
28						

Figure 2-3: Difference in means of shuffled medical errors

4) Select the difference in means (cell F27 in Figure 2-3). Choose Repeat and Score and try 20000 trials. After the simulation finishes (40 seconds on an older laptop), go to the Results window, and use Excel's Function wizard and the same COUNTIF function to count how many of the results were >= 0.92, then divide by 20,000. Alternatively, you can enter the formula directly: enter =COUNTIF(result1,”>=0.92”)/COUNT(result1) to calculate an estimated p-value (Figure 2-4)



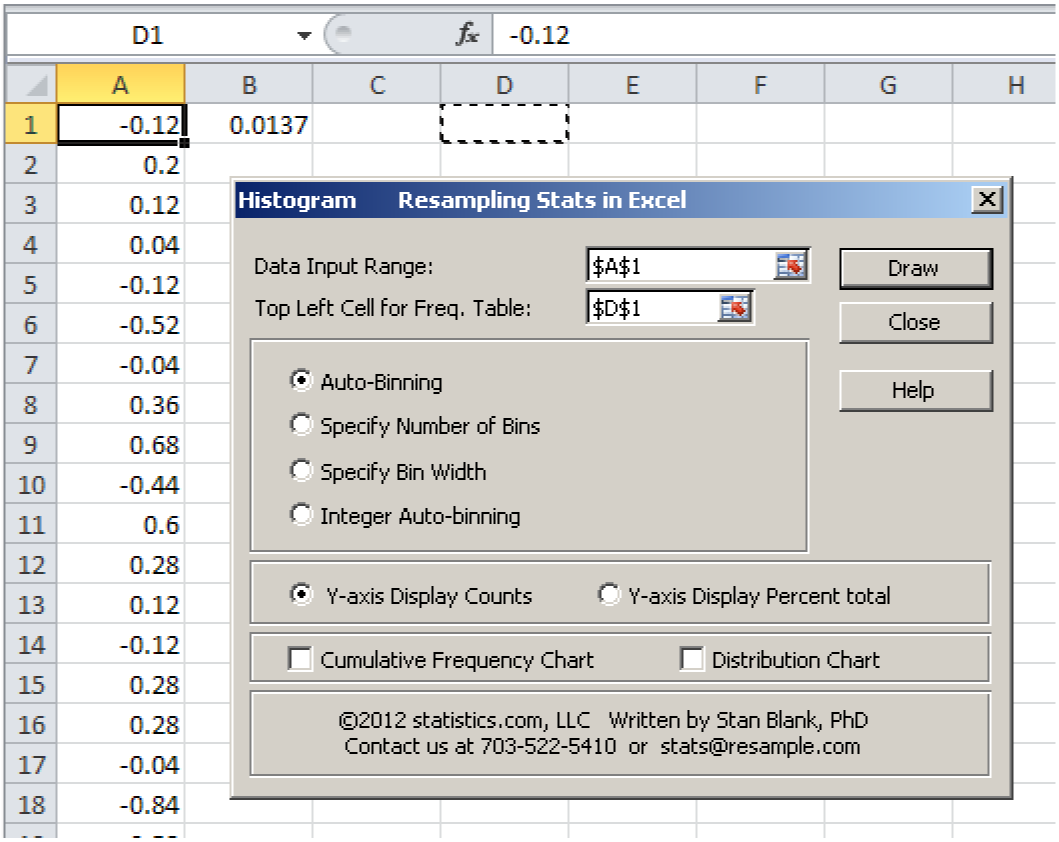
B1 =COUNTIF(result1,">=0.92")/COUNT(result1)

	A	B	C	D	E	F	G	H
1	-0.12	0.0137						
2	0.2							
3	0.12							
4	0.04							
5	-0.12							
6	-0.52							
7	-0.04							
8	0.36							
9	0.68							
10	-0.44							

Figure 2-4: Estimated p-value from 20000 trials

5) The estimated p-value of 0.0137 supports the conclusion that the difference in the reduction of hospital errors was most likely not due to chance.

6) Create a histogram of the results by selecting cell A1 and choosing the Histogram feature in Resampling Stats as shown in figure 2.5.



D1 f_x -0.12

	A	B	C	D	E	F	G	H
1	-0.12	0.0137						
2	0.2							
3	0.12							
4	0.04							
5	-0.12							
6	-0.52							
7	-0.04							
8	0.36							
9	0.68							
10	-0.44							
11	0.6							
12	0.28							
13	0.12							
14	-0.12							
15	0.28							
16	0.28							
17	-0.04							
18	-0.84							

Histogram Resampling Stats in Excel

Data Input Range: $A$1

Top Left Cell for Freq. Table: $D$1

● Auto-Binning

○ Specify Number of Bins

○ Specify Bin Width

○ Integer Auto-binning

● Y-axis Display Counts ○ Y-axis Display Percent total

☐ Cumulative Frequency Chart ☐ Distribution Chart

©2012 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 2-5: Histogram dialog

7) Select cell D1 as the Top Left Cell for Freq. Table and click Draw. The resulting histogram from this simulation is shown in figure 2-6.

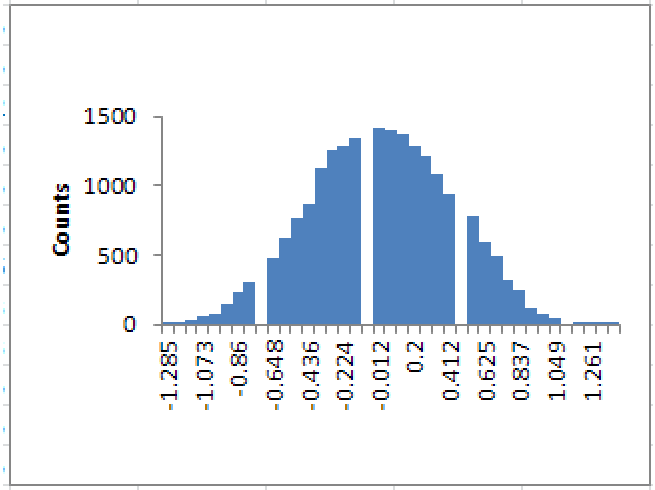
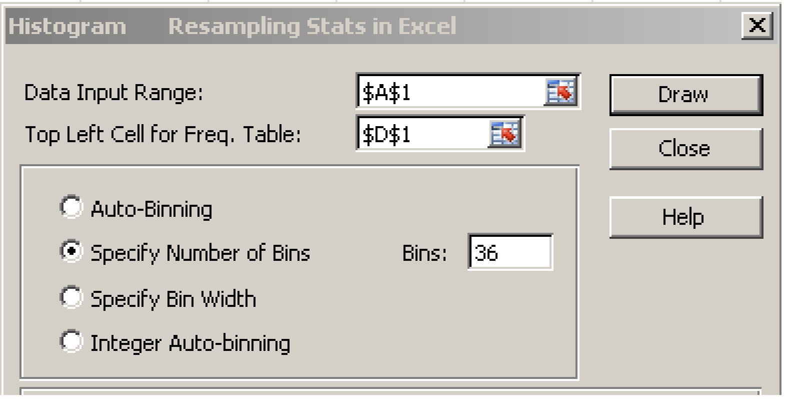


Figure 2-6: Histogram

8) The gaps are due to the fact that there is a limited number of feasible differences in average errors, simply as a result of the given data. Three of the histogram bins

chosen by the automatic binning algorithm just happened to contain no feasible values for the difference in means under *any* shuffling of the data. We can eliminate the gaps by reducing the number bins (by trial) until the gaps disappear.



Data Input Range: $A$1
Top Left Cell for Freq. Table: $D$1
● Specify Number of Bins   Bins: 36

Figure 2-7: Reducing the number of bins

9) The original number of bins was 39 and if we choose 36 bins as shown in figure 2-7, we should see a histogram without gaps. The result is shown in figure 2-8.

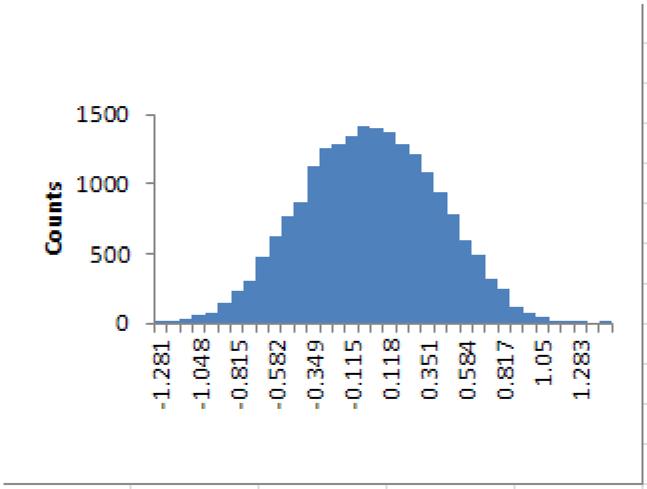


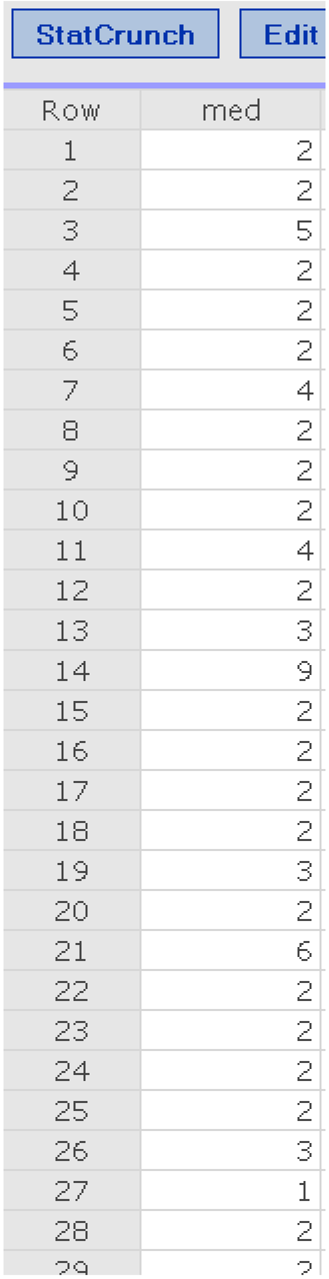
Figure 2-8: Modified histogram with 36 bins

10) Note the original medical error mean difference of 0.92 is on the far right of the histogram, indicating that such a value is not likely due to chance. Why do you think the histogram resembles a Normal curve?

StatCrunch

We have to be somewhat creative to model the medical errors problem in StatCrunch. StatCrunch does not behave like Excel!

1) Download the Excel medical errors workbook from the Resampling Stats section above. Open StatCrunch and copy/paste column A of the medical error data from the Excel worksheet into the first column of the StatCrunch worksheet filling cells 1-25. Copy/paste the data from column B from the Excel worksheet into the same StatCrunch column filling cells 26-50. Name the column med.



Row	med
1	2
2	2
3	5
4	2
5	2
6	2
7	4
8	2
9	2
10	2
11	4
12	2
13	3
14	9
15	2
16	2
17	2
18	2
19	3
20	2
21	6
22	2
23	2
24	2
25	2
26	3
27	1
28	2
29	2

Figure 2-9: Medical error data in single stacked column

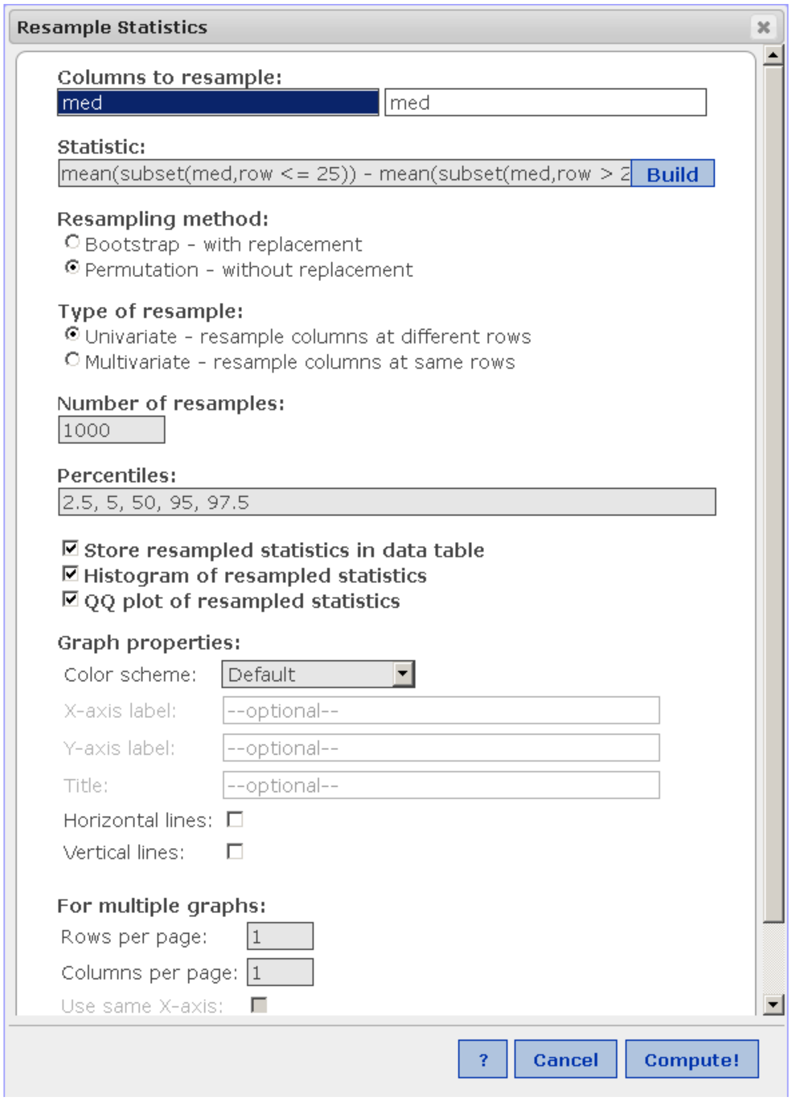
2) Select Stat>Resample>Statistic

3) Choose the med column. For the Statistic, enter (or copy/paste):

mean(subset(med,row <= 25)) - mean(subset(med,row > 25))

This statement uses the subset function to find the mean of the first 25 rows/cells of the med column and subtracts from that value the mean of rows/cells 25-50. In other words, this statement finds the difference in means between the two medical error data sets. See figure 2-10.

4) Choose the Permutation test (without replacement) and enter 1000 as the number of resamples. Check the Store resampled statistics in data table and click Compute! See figure 2-10. This is a lengthy calculation so you may have to click Continue if the script warning dialog appears. Rename the new column “diff”.



Resample Statistics

Columns to resample: med | med

Statistic: mean(subset(med,row <= 25)) - mean(subset(med,row > 2

Resampling method:
○ Bootstrap - with replacement
● Permutation - without replacement

Type of resample:
● Univariate - resample columns at different rows
○ Multivariate - resample columns at same rows

Number of resamples: 1000

Percentiles: 2.5, 5, 50, 95, 97.5

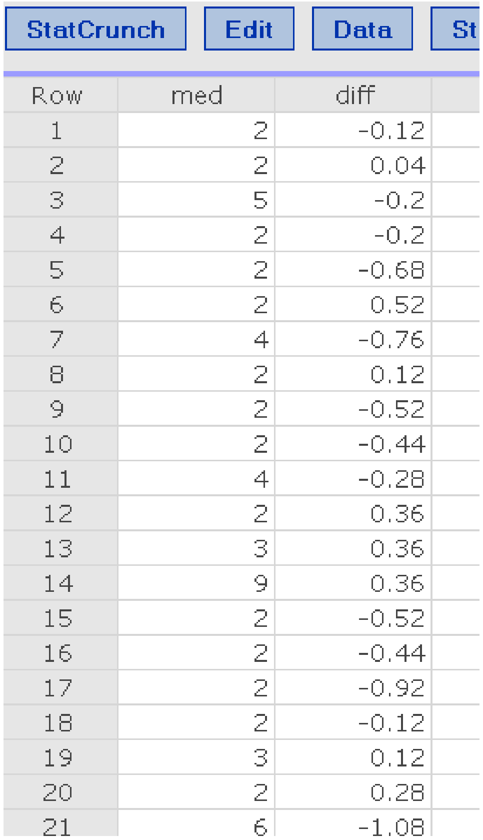
☑ Store resampled statistics in data table
☑ Histogram of resampled statistics
☑ QQ plot of resampled statistics

Graph properties:
Color scheme: Default
X-axis label: --optional--
Y-axis label: --optional--
Title: --optional--
Horizontal lines: ☐
Vertical lines: ☐

For multiple graphs:
Rows per page: 1
Columns per page: 1
Use same X-axis: ☐

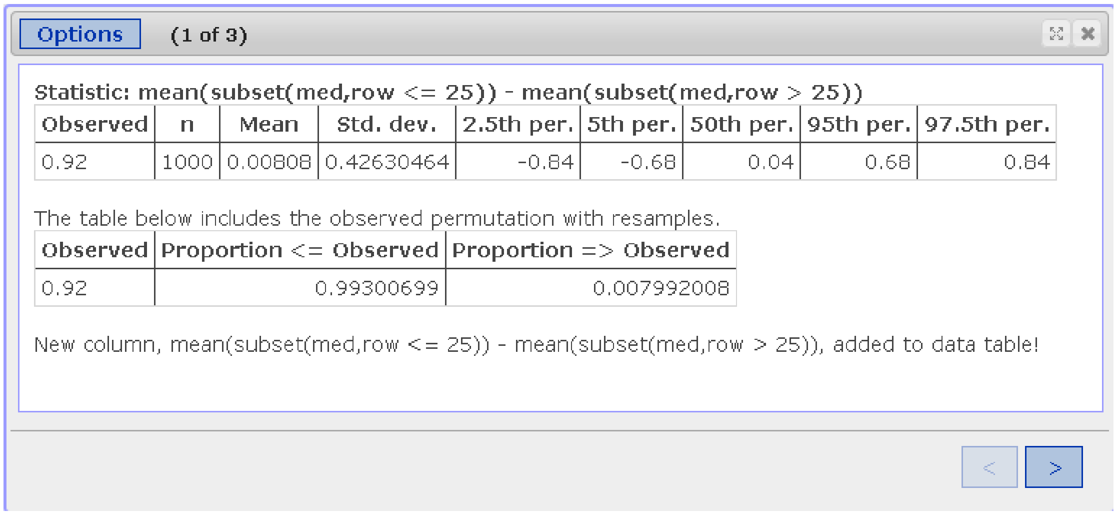
Figure 2-10: Resample Statistics

5) Here is some sample output:



Row	med	diff
1	2	-0.12
2	2	0.04
3	5	-0.2
4	2	-0.2
5	2	-0.68
6	2	0.52
7	4	-0.76
8	2	0.12
9	2	-0.52
10	2	-0.44
11	4	-0.28
12	2	0.36
13	3	0.36
14	9	0.36
15	2	-0.52
16	2	-0.44
17	2	-0.92
18	2	-0.12
19	3	0.12
20	2	0.28
21	6	-1.08

Figure 2-11: Sample output for difference in means (medical errors)



Statistic: mean(subset(med,row <= 25)) - mean(subset(med,row > 25))

Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
0.92	1000	0.00808	0.42630464	-0.84	-0.68	0.04	0.68	0.84

The table below includes the observed permutation with resamples.

Observed	Proportion <= Observed	Proportion => Observed
0.92	0.99300699	0.007992008

New column, mean(subset(med,row <= 25)) - mean(subset(med,row > 25)), added to data table!

Figure 2-12: Summary results

6) Note the difference in means of the original data (observed) is 0.92. The proportion of resampled differences => the observed is 0.007992008. Click on the > button in the lower right corner of the results to view a histogram.

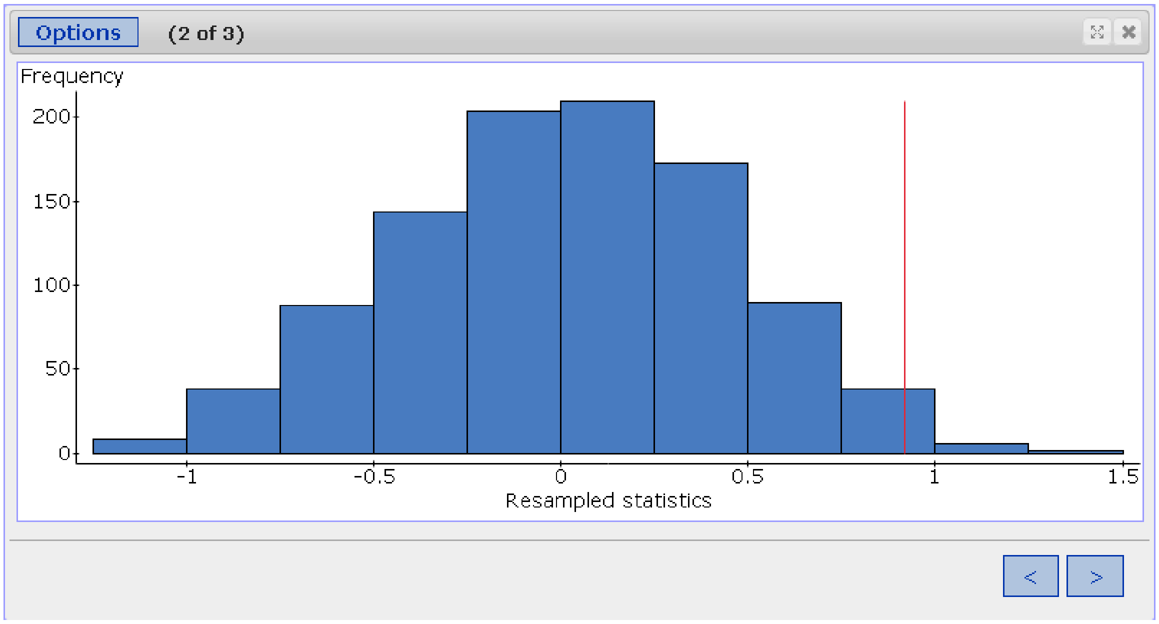
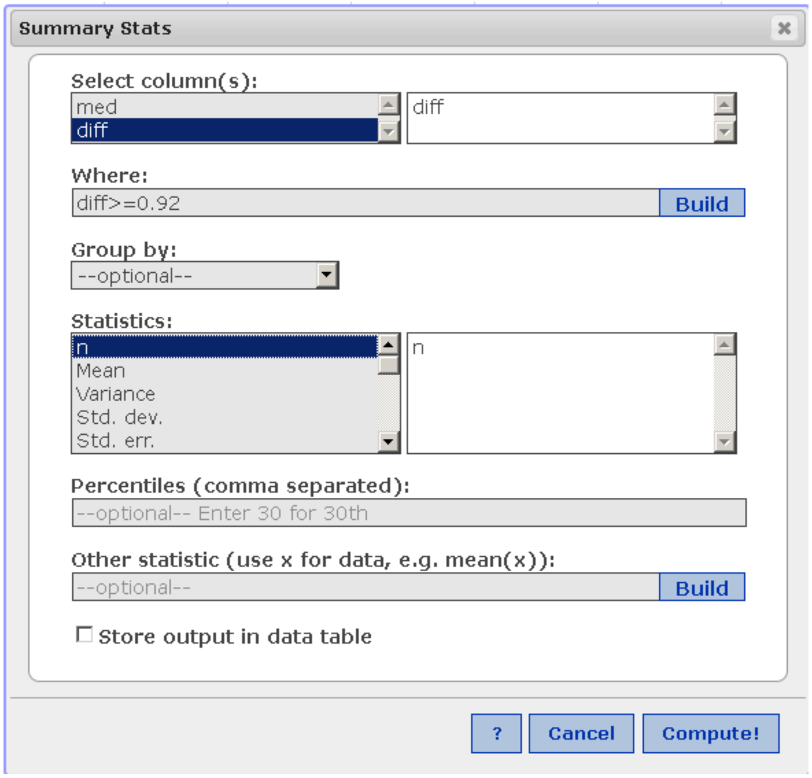


Figure 2-13: Histogram

7) We can verify the results by actually counting how many times the resampled data was >= (or =>) 0.92 by selecting Stats>Summary Stats>Column, selecting the diff column, entering: diff>=0.92 in the "Where" box and clicking n in the Statistics box (see figure 2-14). Click Compute! Figure 2-15 displays the output from one complete experiment.

8) Note: Remember that the results of most of the experiments in the supplement and the results from your own computer are based on pseudo-random numbers will most likely not agree. However, if enough trials are done, the results should be reasonably close.



Select column(s): diff

Where: diff>=0.92

Group by: --optional--

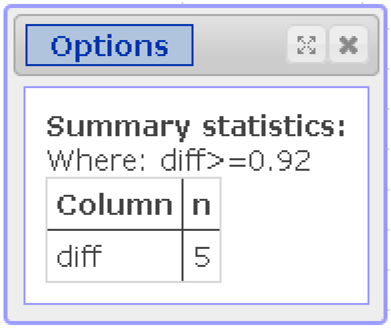
Statistics: n

Percentiles (comma separated): --optional-- Enter 30 for 30th

Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 2-14: Summary Stats dialog



Summary statistics:
Where: diff>=0.92

Column	n
diff	5

Figure 2-15: Number of instances (7) of difference in means >= 0.92

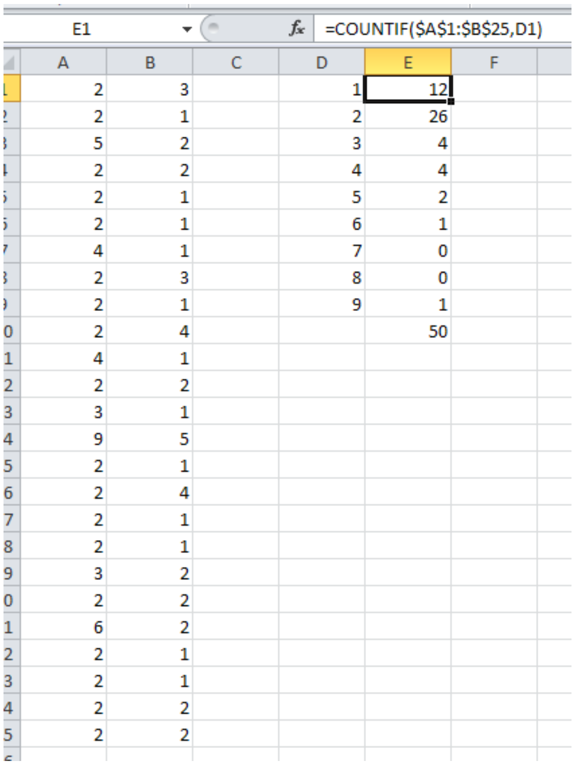
Box Sampler

The Box Sampler model for the medical errors simulation is somewhat more involved than the coin toss simulation. If you would rather download and run a macro-enabled workbook, you can click this <link>. I would still recommend that you read through the step-by-step procedures to see how the Box Sampler model was created.

Here are the step-by-step procedures:

1) Download the mederrors.xls file using this <link>.

2) In order to correctly populate the Box, we need to know the frequency of each number of errors. The mederrors.xls workbook calculates the frequency for us. The data is placed in cells A1:B25 and it is evident that the values in the data range from 1 to 9. We entered those values in cells D1:D9. The frequencies (which could also be counted manually) are calculated by the Excel =COUNTIF($A$1:$B$25,D1) functions in cells E1:E9. The worksheet is displayed in Figure 2-16:



E1 =COUNTIF($A$1:$B$25,D1)

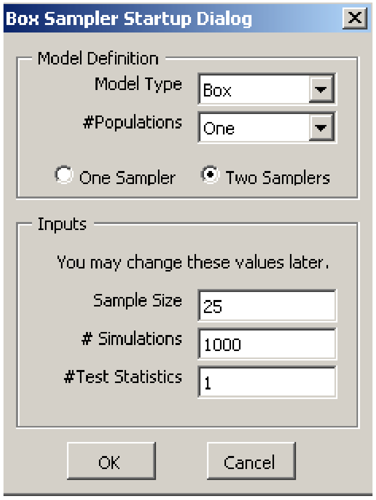
A	B	C	D	E	F
2	3		1	12	
2	1		2	26	
5	2		3	4	
2	2		4	4	
2	1		5	2	
2	1		6	1	
4	1		7	0	
2	3		8	0	
2	1		9	1	
2	4			50	
4	1				
2	2				
3	1				
9	5				
2	1				
2	4				
2	1				
2	1				
3	2				
2	2				
6	2				
2	1				
2	1				
2	2				
2	2				

Figure 2-16: Frequency of medical errors

Look closely at the =COUNTIF functions in cells E1:E9. Note that we are using absolute references to the medical error data range in cells $A$1:$B$25. Absolute references are denoted by the $ sign and do not change when the formula is copied to another location. Also note that we use the cell reference D1 to refer to the value 1 in cell D1. This value is our comparison value or the value we are looking to count. The D1 cell reference is a relative reference, which means that if we copy and paste the formula to another cell location, the D1 reference will change to reflect the new location. So, the formula in cell E2 will reference the comparison value in cell D2, E3 references D3, etc.

If we enter the function or formula as shown in cell E1 and then copy this formula down to cell E9 (click on the box in the lower right corner of cell E1 and drag downward to cell E9), the data input will remain constant (absolute references), but the comparison cell will change (relative references) to reflect the relative positions of the new cell location. The result is a frequency table, which is exactly what we want. The value (50) in cell E10 is a checksum to make certain we have included all the data. Please make certain you look at the formulas in cells E1:E9 so you can understand how the frequency table was created.

3) Start Box Sampler and instead of using the default settings, choose Two Samplers (for the two columns of data), a Sample Size of 25 (for the number of rows of data) and 1000 Simulations as shown in Figure 2-17:



Box Sampler Startup Dialog

Model Definition

Model Type: Box

#Populations: One

○ One Sampler ● Two Samplers

Inputs

You may change these values later.

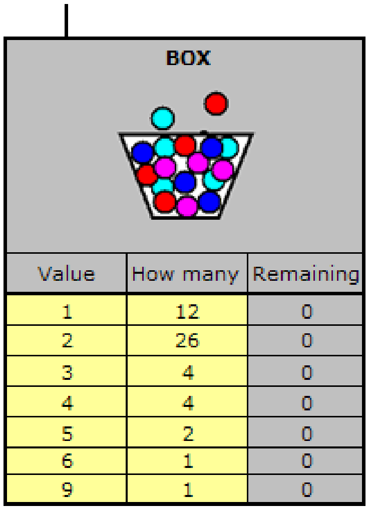
Sample Size: 25

Simulations: 1000

#Test Statistics: 1

Figure 2-17: Box Sampler setup

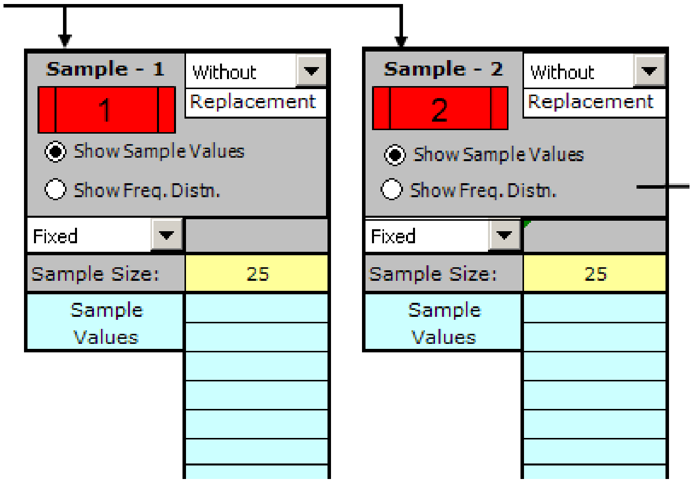
4) Click OK. Enter the frequency data calculated from step 2 in the Box as shown in Figure 2-18. Note that there is no need to enter 7 or 8 (why?).



Value	How many	Remaining
1	12	0
2	26	0
3	4	0
4	4	0
5	2	0
6	1	0
9	1	0

Figure 2-18: Box contents

5) In both Sample-1 and Sample-2 output boxes, change the With Replacement setting to Without Replacement setting. The explanation for With vs. Without replacement will be explained later in the course. For now, follow along and change the setting (Figure 2-19):



Sample - 1: Without Replacement; ● Show Sample Values; ○ Show Freq. Distn.; Fixed; Sample Size: 25

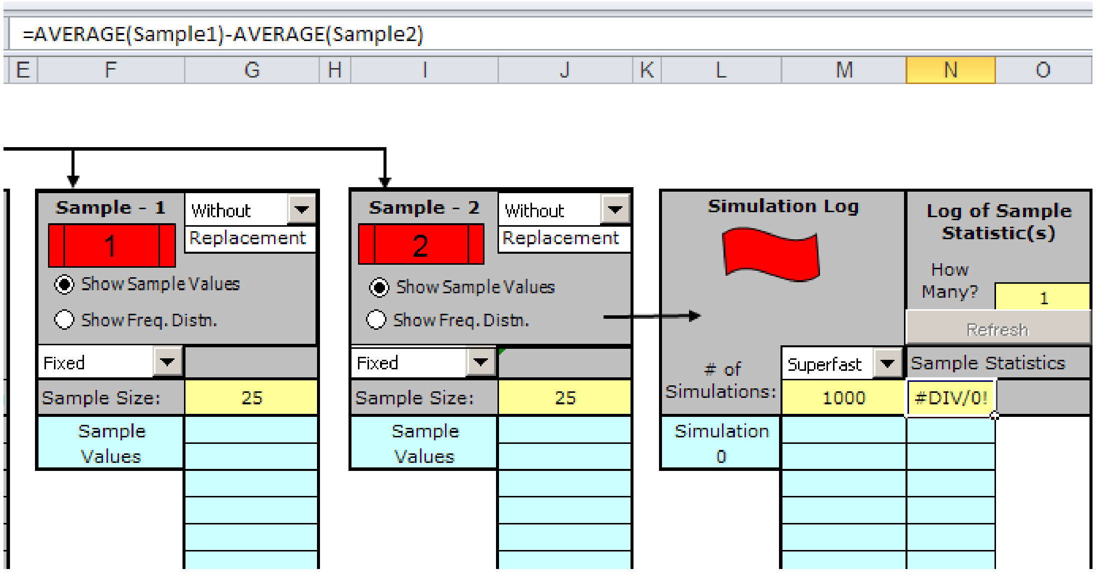
Sample - 2: Without Replacement; ● Show Sample Values; ○ Show Freq. Distn.; Fixed; Sample Size: 25

Figure 2-19: Without replacement

6) Change the simulation speed to Superfast and verify that the number of simulations = 1000. In the Sample Statistics cell (cell N11), enter the formula

=AVERAGE(Sample1) - AVERAGE(Sample2)

as shown in figure 2-20. Don't worry about the #DIV/0? error. That will go away after some data is 'fed' to the formula by running the simulation.



=AVERAGE(Sample1)-AVERAGE(Sample2)

Sample - 1: Without Replacement; ● Show Sample Values; ○ Show Freq. Distn.; Fixed; Sample Size: 25

Sample - 2: Without Replacement; ● Show Sample Values; ○ Show Freq. Distn.; Fixed; Sample Size: 25

How Many? 1

# of Simulations: Superfast; 1000

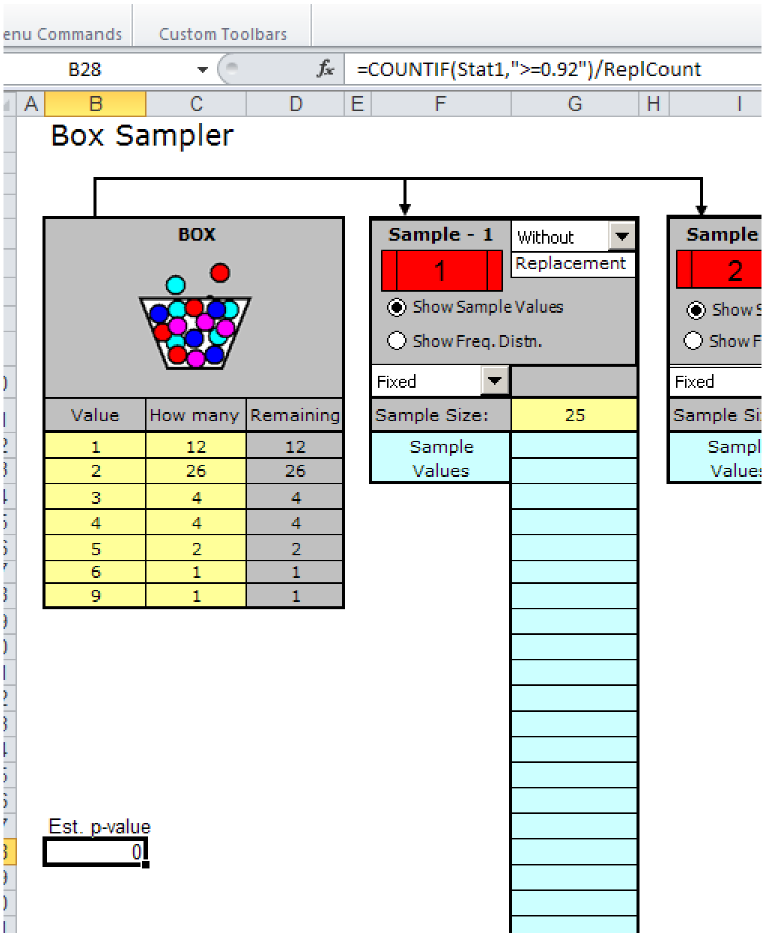
Sample Statistics: #DIV/0!

Figure 2-20: Enter the Sample Statistics formula in cell N11

7) Finally, in an open cell (I used cell B28), enter the formula:

=COUNTIF(Stat1, ">=0.92")/ReplCount

so we can calculate a p-value estimate. See figure 2-21. ReplCount is the named range for the number of simulations, which in this case is 1000. Stat1 is the named range for the output of the Sample Statistics in cell M12 and downward.



B28 =COUNTIF(Stat1,">=0.92")/ReplCount

Value	How many	Remaining
1	12	12
2	26	26
3	4	4
4	4	4
5	2	2
6	1	1
9	1	1

Without Replacement

● Show Sample Values
○ Show Freq. Distn.

Fixed

Sample Size: 25

Sample Values

Fixed

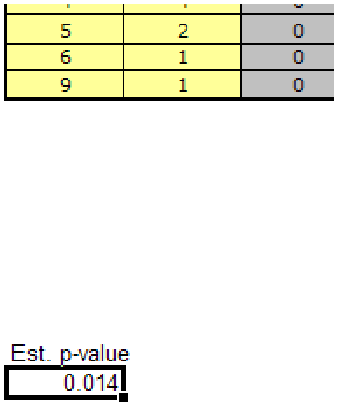
Sample Si

Est. p-value 0

Figure 2-21: Estimated p-value

8) Click the Simulate button to run the model. Be patient… this model is SLOW. The Superfast setting will not output as often, so it may appear that nothing is happening. Again, be patient. The p-value estimate will update as the model begins to create output.

9) Figure 2-22 illustrates the output from one simulation of 1000 trials. The result is consistent with the Resampling Stats estimated p-value. Again, you can download the macro-enabled workbook at the beginning of this section if you are having problems creating the model.



5	2	0
6	1	0
9	1	0

Est. p-value

0.014

Figure 2-22: Output from one simulation of the medical errors example

3 Displaying and Exploring Data

No specific resampling procedures in this chapter. You might take this opportunity to explore the charting features of Excel and/or the software package you are using.

4 Probability

No specific resampling procedures in this chapter. As an exercise, you might try tossing virtual coins and/or dice using your software package to see how closely the experimental results agree with the theoretical results.

5 Relationship Between Two Categorical Variables

5.1 Two-Way Tables

Could Chance be Responsible?

The following is the admission rate resampling model:

1. Put 584 chips in a hat to represent the 584 applicants. Of these, 147 are marked Admitted and 437 are marked Rejected.

2. Shuffle the hat

3. Draw 393 chips from the hat (that's the size of the female group), count the number of Admits, and calculate this as a proportion.

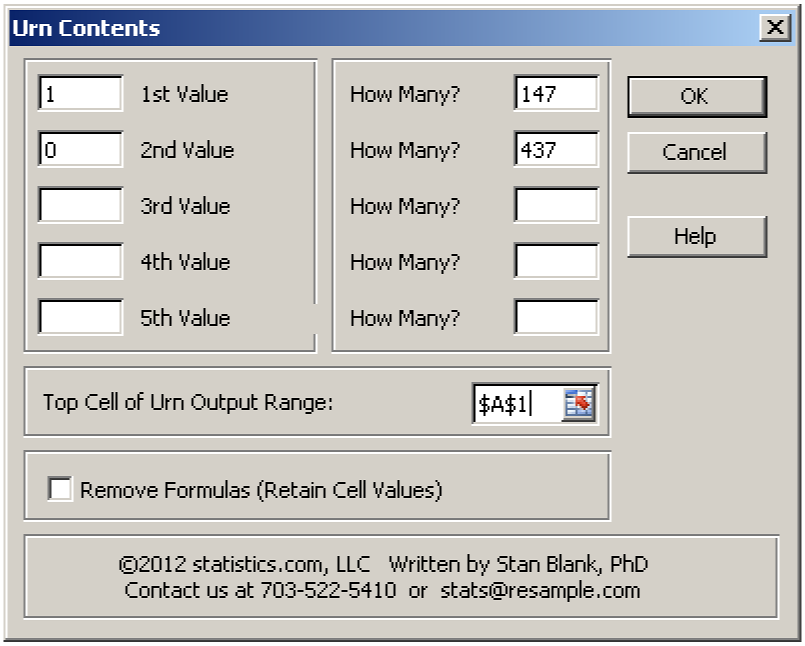
4. Draw the remaining 191 chips from the hat (the size of the male group), count the number of Admits, and calculate this as a proportion.

5. Record the criterion (statistic) of interest, which is the difference in the acceptance rates (women minus men). The actual difference was 23.92 - 27.75 = -3.83 percentage points.

6. Repeat the trial many (say 1000) times, and find out how often we get a difference as extreme as -3.83 percentage points.

Resampling Stats

1) Open Resampling Stats and select the Urn feature. Let 1 = Admitted and 0 = Rejected. Select A1 as the Top Left cell and click OK.

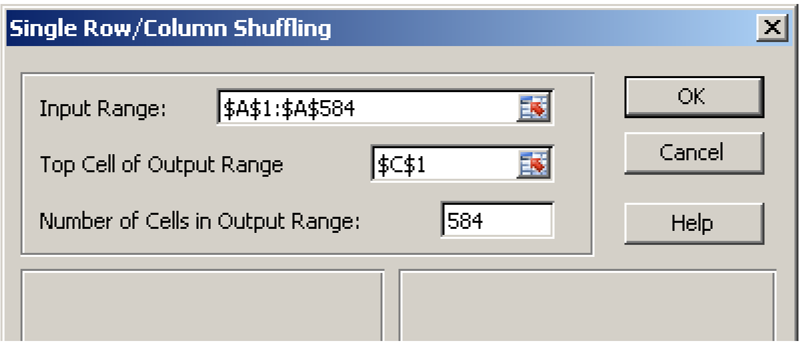


1 1st Value — How Many? 147
0 2nd Value — How Many? 437
Top Cell of Urn Output Range: $A$1
☐ Remove Formulas (Retain Cell Values)

©2012 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 5-1: Populate the Urn

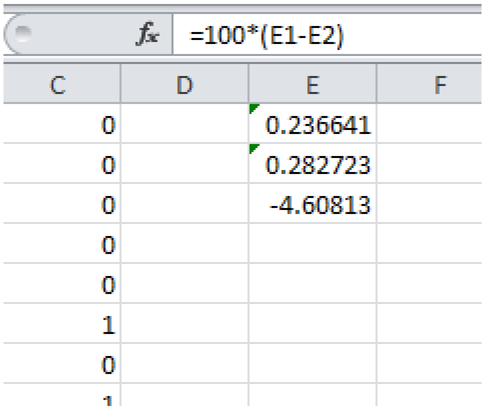
2) Select cell A1 and choose Shuffle (Shuffle resamples without replacement) from the Resampling menu. Select cell C1 as the Top Left cell. Keep the default number of cells in the output range (584). Click OK.



Input Range: $A$1:$A$584
Top Cell of Output Range $C$1
Number of Cells in Output Range: 584

Figure 5-2: Shuffle dialog

3) In cell E1, enter the formula =SUM(C1:C393)/393. This formula calculates the proportion of women who were admitted to the program. In cell E2, enter the formula =SUM(C394:C584)/191. This formula calculates the proportion of men who were admitted to the program. In cell E3, enter the formula =100*(E1-E2) to find the difference in the two proportions (women – men) expressed as a percentage.

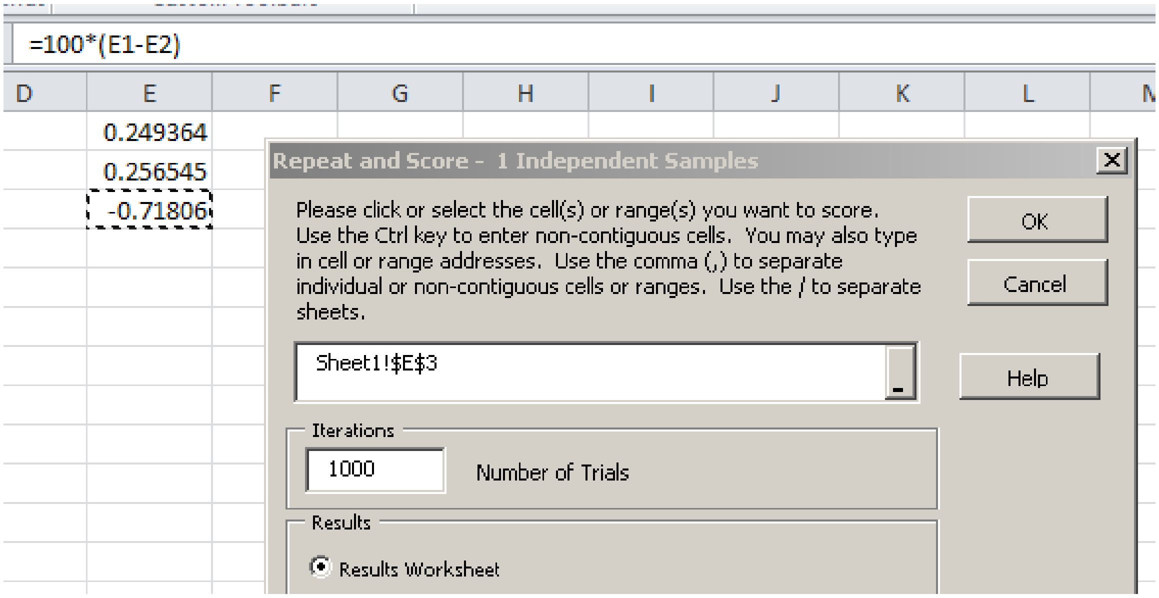


=100*(E1-E2)

C	D	E	F
0		0.236641	
0		0.282723	
0		-4.60813	
0			
0			
1			
0			

Figure 5-3: Calculate the difference in proportions expressed as a percentage (cell E3)

4) Select cell E3 and choose Repeat and Score from the Resampling menu. Click OK.



=100*(E1-E2)

D	E	F	G	H	I	J	K	L
	0.249364							
	0.256545							
	-0.71806							

Repeat and Score - 1 Independent Samples

Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

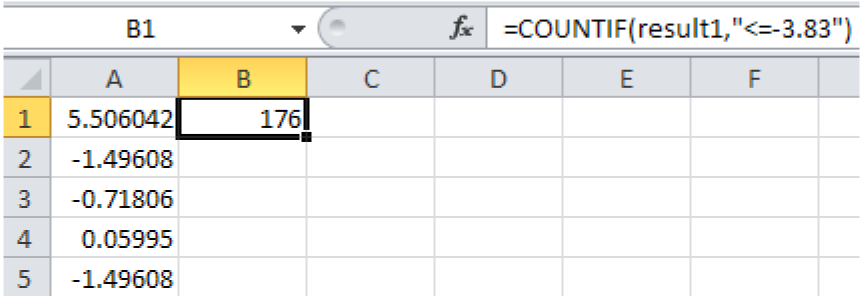
Sheet1!$E$3

Iterations: 1000 Number of Trials

Results: ● Results Worksheet

Figure 5-4: Repeat and Score on the difference in proportions

5) The partial output of this simulation is shown in figure 5-5. The Excel formula =COUNTIF(result1,"<=-3.83") in cell B1 finds the number of instances in the output that equal or are further from zero (more extreme) than the -3.83 percentage in the original data. NOTE: Column A… the output range is automatically named "result1" by Resampling Stats. We can use this range name to refer to the entire output data for both Excel formulas and in creating histograms.



B1 =COUNTIF(result1,"<=-3.83")

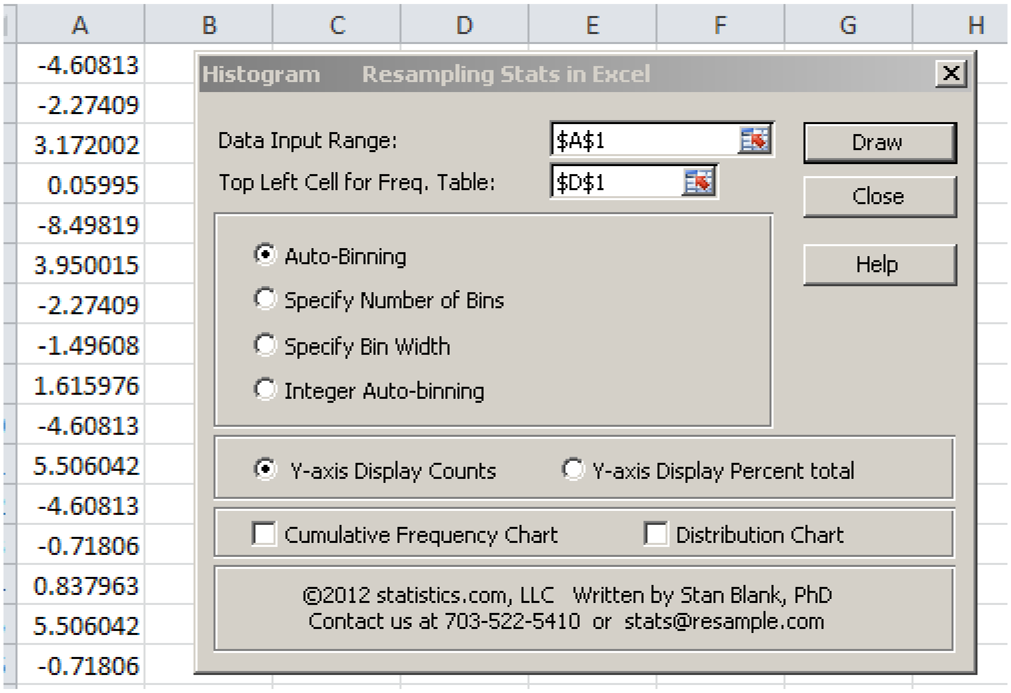
	A	B	C	D	E	F
1	5.506042	176				
2	-1.49608					
3	-0.71806					
4	0.05995					
5	-1.49608					

Figure 5-5: Estimated p-value = 0.176

6) The estimated p-value from our simulation can be calculated by dividing the 176 values as extreme (or more extreme) as the original difference of -3.83 by 1000 or 176/1000, which equals 0.176. This p-value is not too unusual, so the difference in admission rates could be due to chance.

7) Create a histogram by selecting the cell A1 on the Results sheet and choosing the histogram feature from the Resampling menu. The histogram dialog should appear with cell A1 in the range input box. The histogram function will automatically select the entire column for you. Choose cell D1 as the top cell of the output range and click Draw.

Note: You can select the top cell of the data column (A1 in most instances), enter the named data region (result1, result2, etc.), or enter the actual data range (A1:A1000) for the Data Input Range in the histogram dialog.



A
-4.60813
-2.27409
3.172002
0.05995
-8.49819
3.950015
-2.27409
-1.49608
1.615976
-4.60813
5.506042
-4.60813
-0.71806
0.837963
5.506042
-0.71806

Histogram Resampling Stats in Excel

Data Input Range: $A$1

Top Left Cell for Freq. Table: $D$1

● Auto-Binning
○ Specify Number of Bins
○ Specify Bin Width
○ Integer Auto-binning

● Y-axis Display Counts ○ Y-axis Display Percent total

☐ Cumulative Frequency Chart ☐ Distribution Chart

©2012 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 5-6: Histogram dialog

8) Here is the histogram for 1000 trials (figure 5-7)

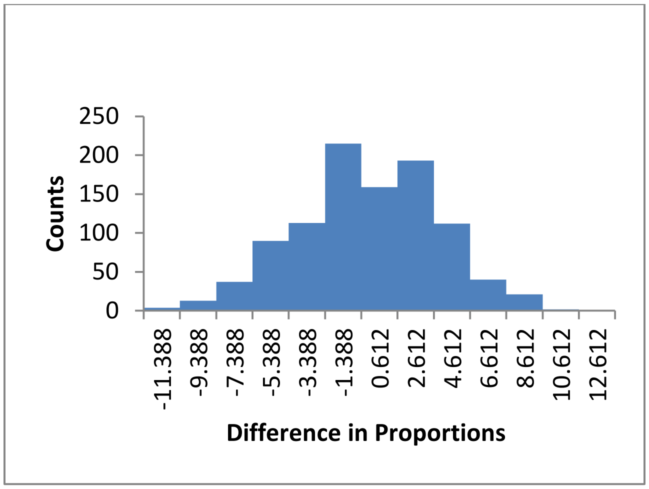
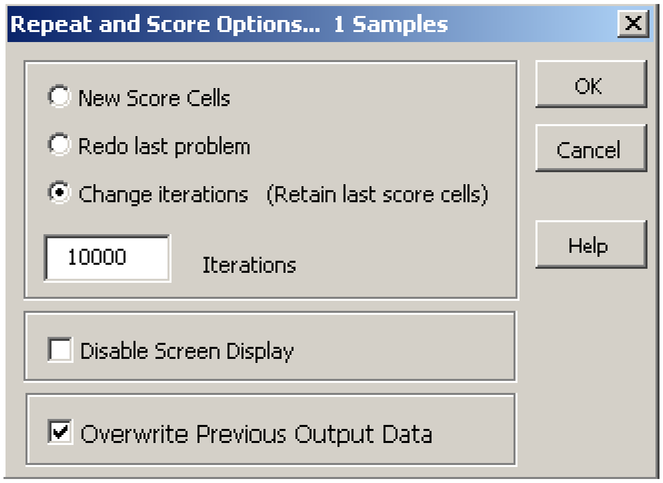


Figure 5-7: Histogram: 1000 trials

9) Run the model again by selecting Repeat and Score and then change the iterations to 10000.



○ New Score Cells
○ Redo last problem
● Change iterations (Retain last score cells)
10000 Iterations
☐ Disable Screen Display
☑ Overwrite Previous Output Data

Figure 5-8: Repeat and Score again

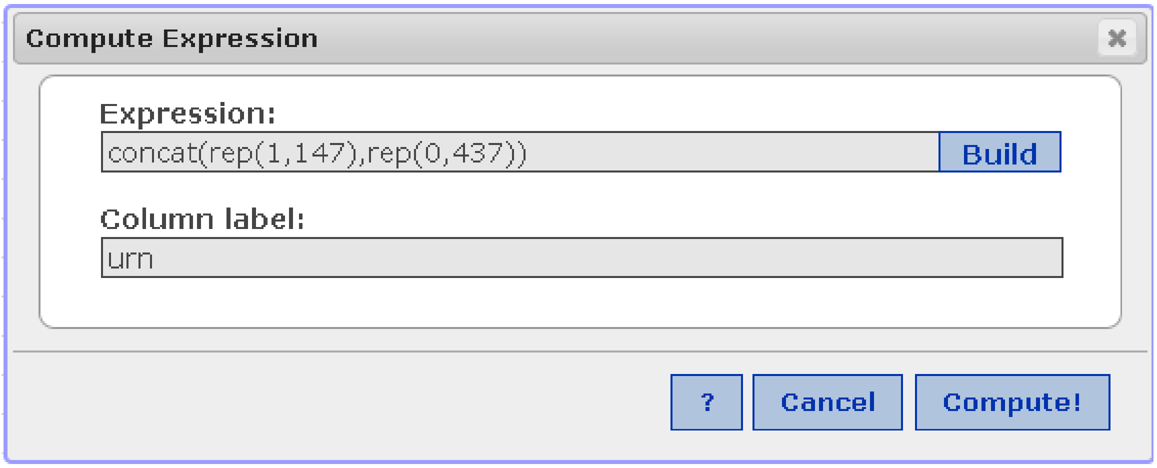
10) Create a histogram from the new model and compare it to figure 5-7.

StatCrunch

1) Open StatCrunch and populate the Urn as follows: Select Data>Compute Expression. In the Compute Expression dialog, enter:

concat(rep(1,147),rep(0,437))

The word "concat" is an abbreviation of concatenate, which means (roughly) "to add to," so this statement creates a list of 147 repetitions of the number 1 added to another list of 437 repetitions of the number 0 and places the continuous list of 1's and 0's in the first column. Change the column name to "urn" as shown in Figure 5-9. Click Compute!



Expression: concat(rep(1,147),rep(0,437))

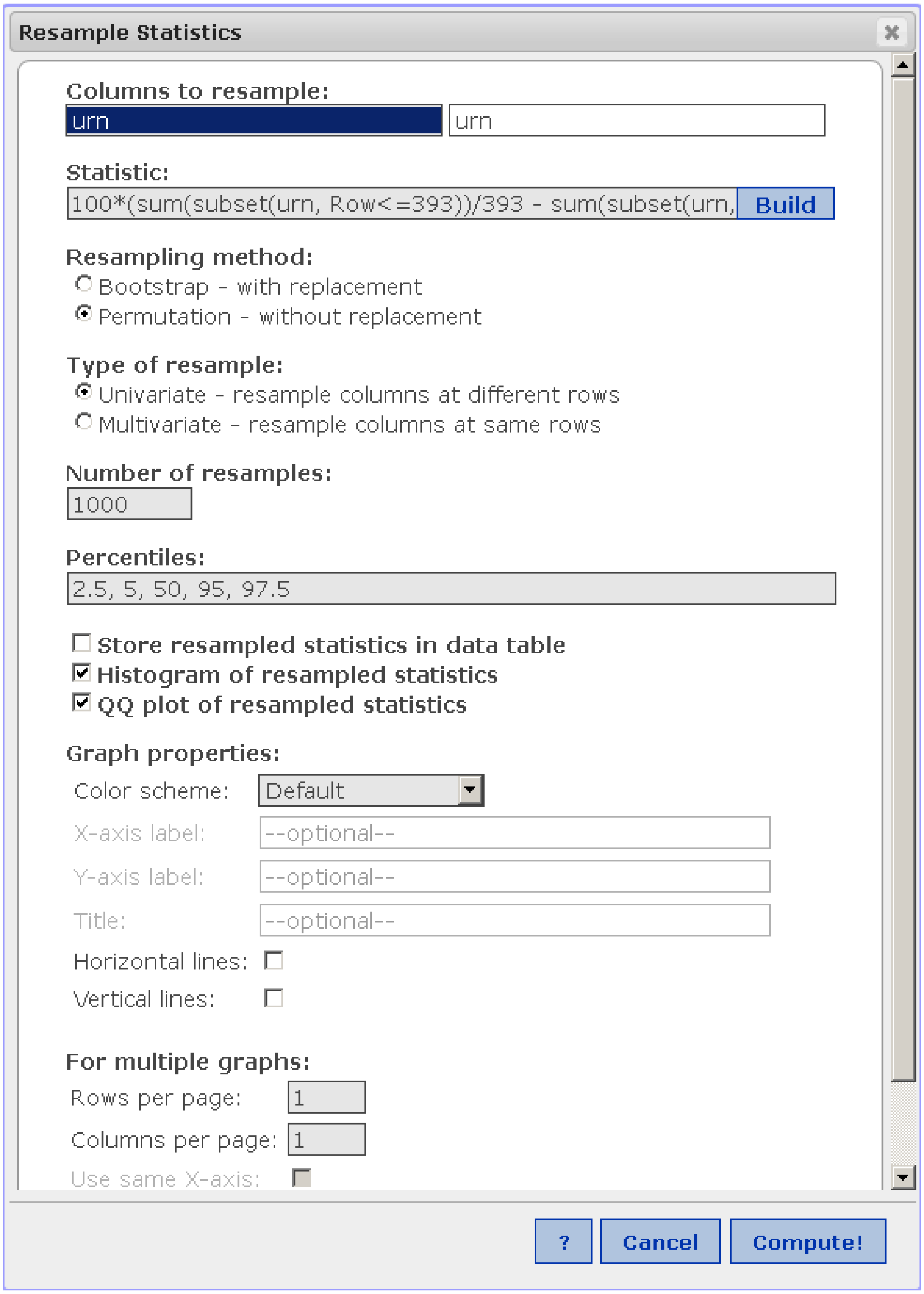
Column label: urn

Figure 5-9: Populate urn

2) Choose Stat>Resample>Statistic and choose the urn column. The statistic expression is somewhat complex and uses the powerful "subset" command which allows us to designate a range of rows as input for a StatCrunch operation. Subset is particularly useful in permutation operations where we resample without replacement. Here is the statistic expression (you can copy/paste):

   100*(sum(subset(urn, Row<=393))/393 - sum(subset(urn, Row>393))/191)

   Select the Permutation – without replacement option and click Compute! (see figure 5-10). You may receive an unresponsive javascript message due to the number of calculations. If this happens, click Continue to continue the execution of the script.



Columns to resample: urn | urn

Statistic: 100*(sum(subset(urn, Row<=393))/393 - sum(subset(urn,

Resampling method:
○ Bootstrap - with replacement
● Permutation - without replacement

Type of resample:
● Univariate - resample columns at different rows
○ Multivariate - resample columns at same rows

Number of resamples: 1000

Percentiles: 2.5, 5, 50, 95, 97.5

☐ Store resampled statistics in data table
☑ Histogram of resampled statistics
☑ QQ plot of resampled statistics

Graph properties:
Color scheme: Default
X-axis label: --optional--
Y-axis label: --optional--
Title: --optional--
Horizontal lines: ☐
Vertical lines: ☐

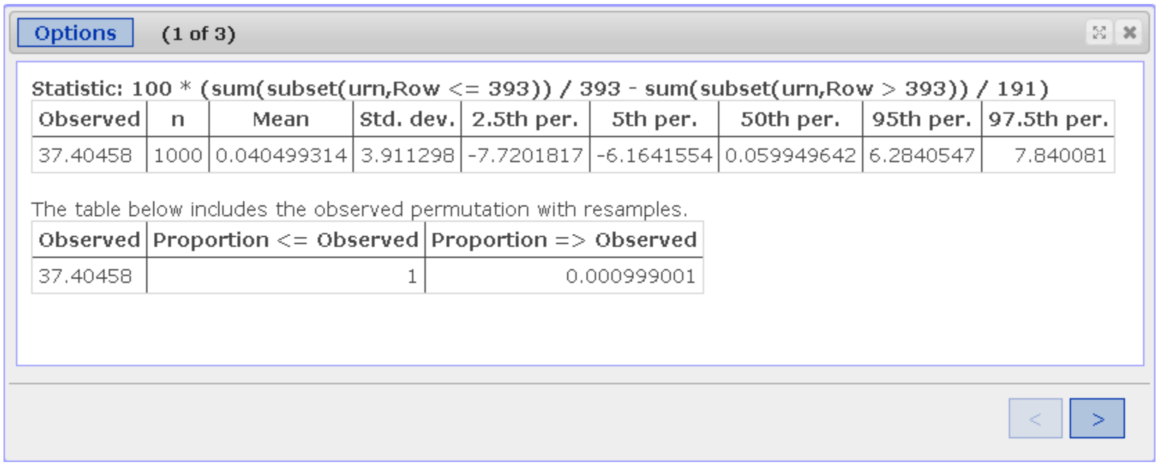
For multiple graphs:
Rows per page: 1
Columns per page: 1
Use same X-axis: ☐

Figure 5-10: Resample Statistic dialog

3) The Statistic expresson in step 2 (and figure 5-10) takes the sum of a subset of the urn column, specifically the sum of the rows from 1 to 393 (the female admissions) and divides this sum by 393 to obtain the proportion of female admissions. Then, the statistic expression subtracts a second subset of the urn column (the male admissions) that consists of the sum of the rows from 394 to

584 divided by 191 (to calculate the male admissions proportion) . This difference in proportions is then multiplied by 100 to obtain a percentage.

4) The results of one model run can be seen in figure 5-11 (you may need to resize the result dialog).



Statistic: 100 * (sum(subset(urn,Row <= 393)) / 393 - sum(subset(urn,Row > 393)) / 191)

Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
37.40458	1000	0.040499314	3.911298	-7.7201817	-6.1641554	0.059949642	6.2840547	7.840081

The table below includes the observed permutation with resamples.

Observed	Proportion <= Observed	Proportion => Observed
37.40458	1	0.000999001

Figure 5-11: StatCrunch female/male admissions results

5) The original -3.83 difference in proportions is well within the 90% confidence interval of -6.16 to 6.28. This means that the -3.83 difference in admissions between females and males is most likely due to chance.

6) Click the > box in the lower right of the dialog to view a histogram (figure 5-12).

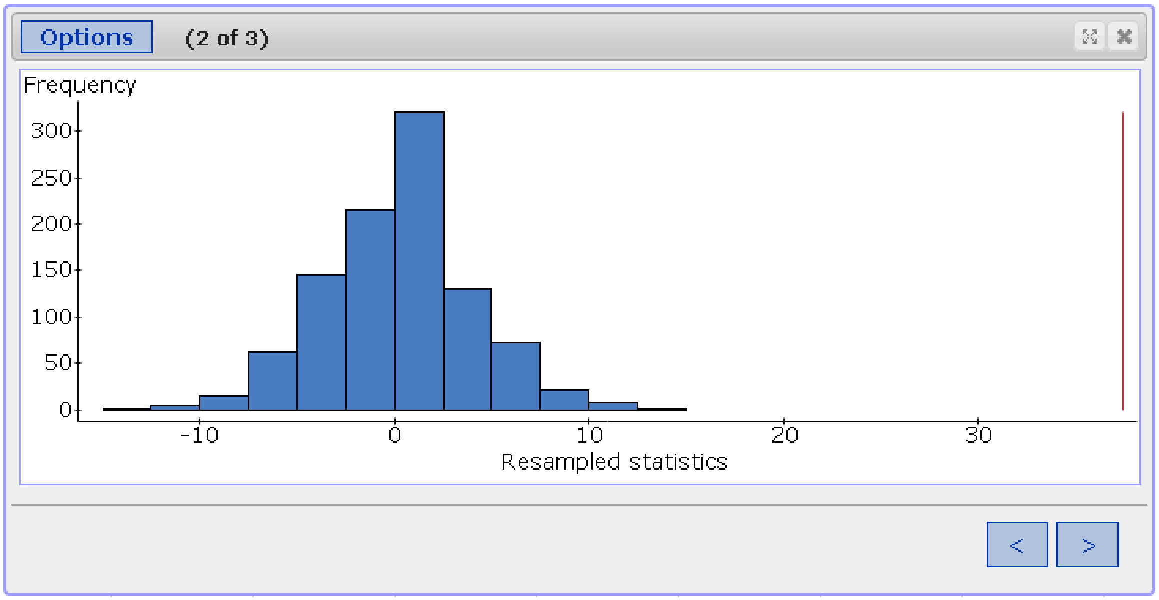


Figure 5-12: Histogram of difference in admissions

6 Surveys and Sampling

6.2 Margin of Error: Sampling Distribution for a Proportion

Solving the Voter Survey Problem with Resampling Stats for Excel

1) Use 1 for positive and 0 for not positive.

2) Start RSXL and select the Urn feature. Accept the default “Create Urn via Dialog Box” (Figure 6-1)

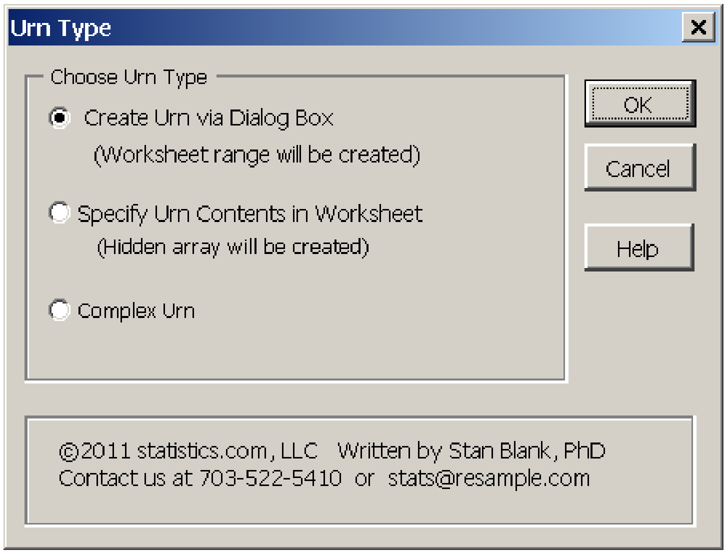

Urn Type
Choose Urn Type
● Create Urn via Dialog Box
(Worksheet range will be created)
○ Specify Urn Contents in Worksheet
(Hidden array will be created)
○ Complex Urn
OK
Cancel
Help
©2011 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

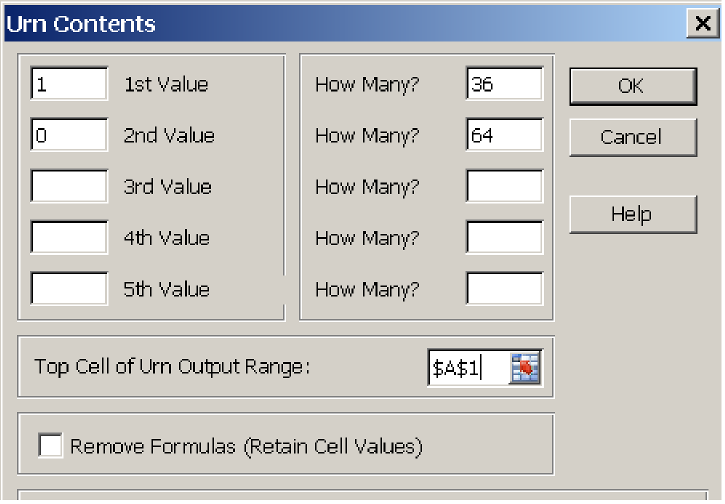
Figure 6-1: Initial Urn dialog

3) Use 1 for the 1st value and 36 for “How Many?”

4) Use 0 for the 2nd value and 64 for “How Many?”

   Note: You could also use 72 for “How Many?” in step 3 and 128 for “How Many?” in step 4. As stated in the text, the important concept is to keep the percentage of positive and negative ratings equivalent to 36% and 64% respectively. Please read the corresponding section of the textbook for further information.

5) Click in the field that says “ Top cell of urn output range, then click the cell on the spreadsheet where you want the top of the urn to be (A1 will do) then click “OK” (Figure 6-2)



Urn Contents

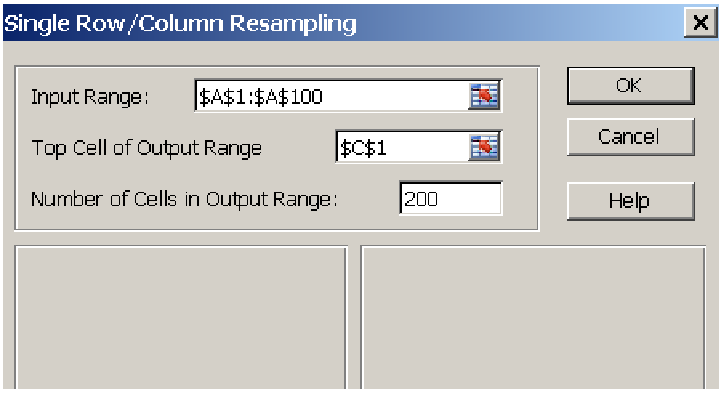
1	1st Value	How Many?	36
0	2nd Value	How Many?	64
	3rd Value	How Many?	
	4th Value	How Many?	
	5th Value	How Many?	

Top Cell of Urn Output Range: $A$1

☐ Remove Formulas (Retain Cell Values)

Figure 6-2: Filling Urn contents

6) Select cell A1 and choose the Resample dialog (this resamples with replacement). Cells A1:A100 should automatically be selected as the input range, choose C1 as the Top Cell of the Output Range, and 200 as the Number of Cells in the Output Range (Figure 6-3)



Single Row/Column Resampling

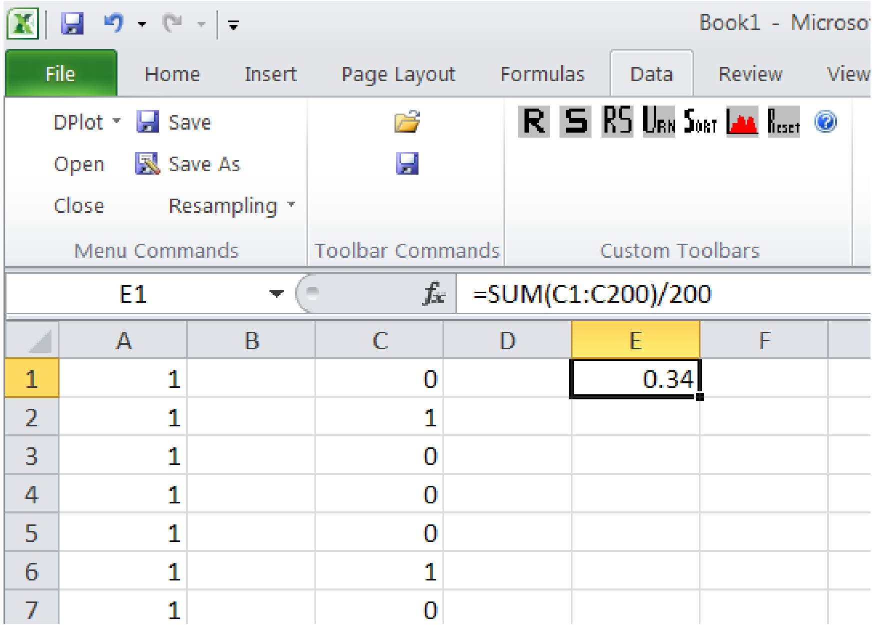
Input Range: $A$1:$A$100

Top Cell of Output Range $C$1

Number of Cells in Output Range: 200

Figure 6-3: Resampling dialog

7) In cell E1 (or any empty cell) place the formula "=SUM(C1:C200)/200" to calculate the proportion of 1's, representing "positive" handling of economy (Figure 6-4).



E1 =SUM(C1:C200)/200

	A	B	C	D	E	F
1	1		0		0.34	
2	1		1			
3	1		0			
4	1		0			
5	1		0			
6	1		1			
7	1		0			

Figure 6-4: Proportion of positive responses in one trial

8) Highlight cell E1 and select “Repeat and Score”. Make certain cell E1 is in the top edit box. E1 is the score cell. Use 5000 iterations.

9) Create a histogram of the data in cells A1:A5000 on the Results sheet using the histogram feature of RSXL (Figure 6-5).

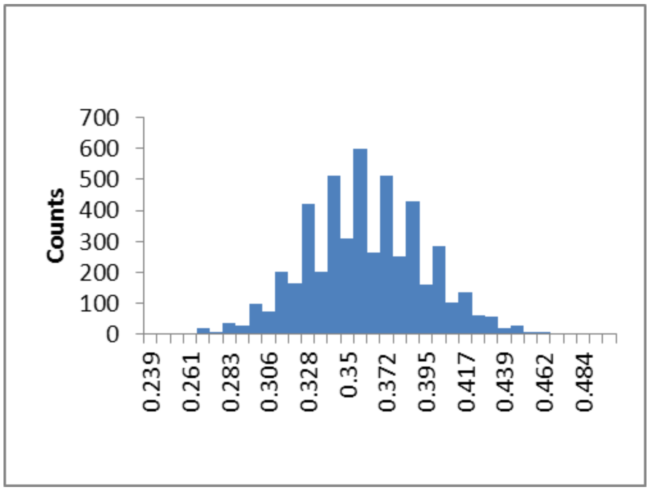


Figure 6-5: Histogram

Roughly speaking we can see that the vast majority of the results run from about 30% positive to 42% positive.

Solving the Voter Survey Problem with StatCrunch

Note: The StatCrunch procedure below requires that the resample size be the same size as the "box" that we are drawing from; this is simply because the program lacks the facility to enter a separate value for the resample size. This is not a problem - as we saw in the text, the "box" can be a variety of different sizes. What matters is the resample size. So here we need to recognize that the resample size must be N=200 (to match the original sample), and make the "box" (from the "concat" statement) the same size.

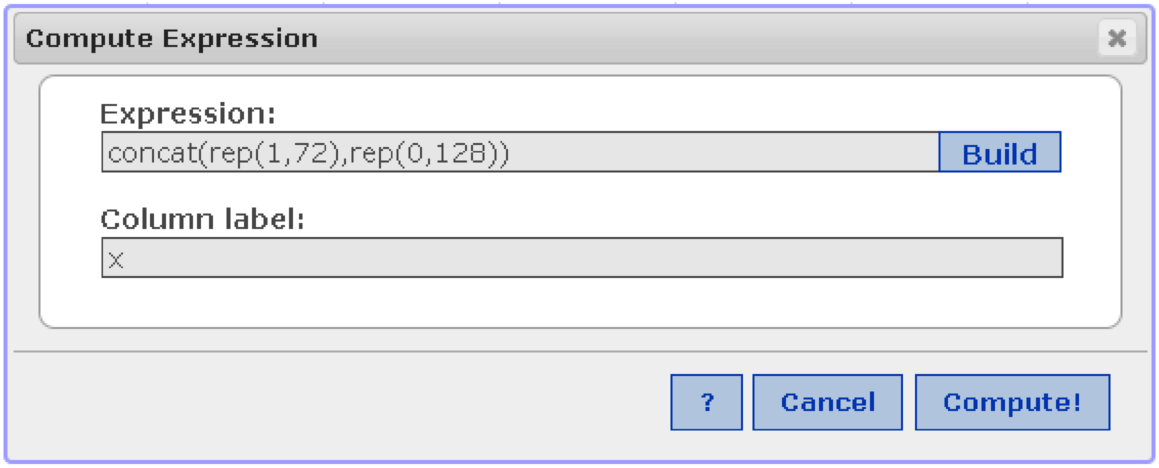
1) After opening StatCrunch, select Data -> Compute Expression

2) In the Expression box, type:

   concat(rep(1,72),rep(0,128))

   to fill the urn (box). This expression physically joins or concatenates (concat) a list of 72 repetitions of the number 1 (rep(1,72)) to a list of 128 repetitions of the number 0 (rep(0,128). The list will be generated and placed in StatCrunch automatically. Note: This list must be the same size as the resample we draw later, because the "Resample" function used in step 4 automatically assumes that the resample size is the same as the sample size. So we make the concat statement produce a list of 200 1's and 0's, 36% or 72 values = 1.

3) Type x for the new column name at the bottom of the dialog and click Compute! (Figure 6-6)



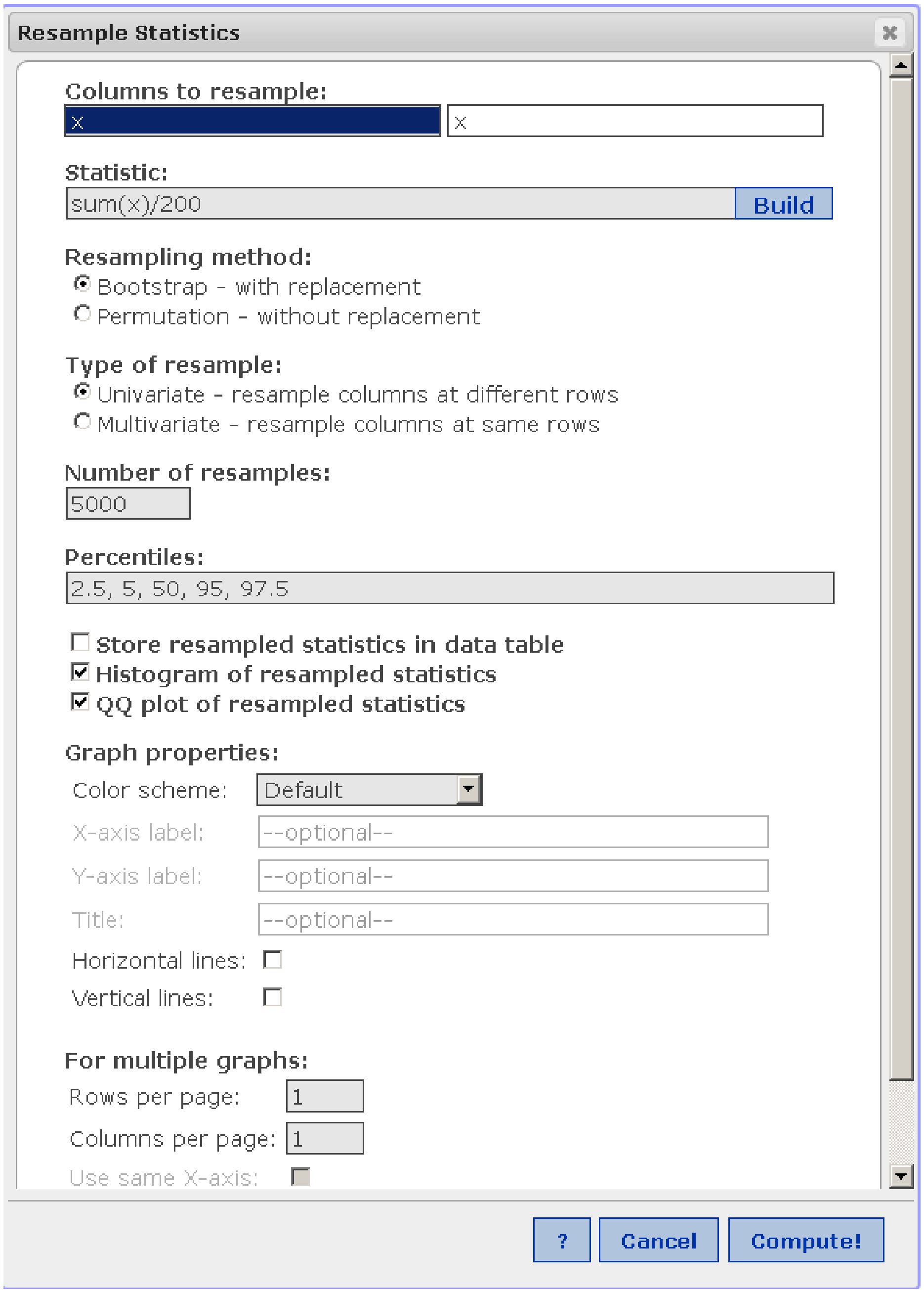
Expression: concat(rep(1,72),rep(0,128))

Column label: x

Figure 6-6: StatCrunch Compute Expression dialog (define Urn/Box)

4) Select Stat -> Resample -> Statistic and click on the x in the Columns to Resample.

5) Type: sum(x)/200 in the Statistic: edit box and 5000 for the Number of Resamples. Click Compute! (Figure 6-7)



Resample Statistics

Columns to resample: x | x

Statistic: sum(x)/200

Resampling method:
● Bootstrap - with replacement
○ Permutation - without replacement

Type of resample:
● Univariate - resample columns at different rows
○ Multivariate - resample columns at same rows

Number of resamples: 5000

Percentiles: 2.5, 5, 50, 95, 97.5

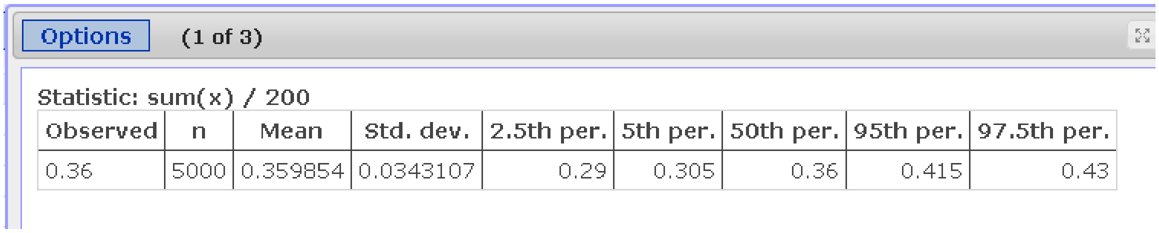
☐ Store resampled statistics in data table
☑ Histogram of resampled statistics
☑ QQ plot of resampled statistics

Graph properties:
Color scheme: Default
X-axis label: --optional--
Y-axis label: --optional--
Title: --optional--
Horizontal lines: ☐
Vertical lines: ☐

For multiple graphs:
Rows per page: 1
Columns per page: 1
Use same X-axis: ☐

Figure 6-7: StatCrunch Resample Statistic

6) Here is the result (Figure 6-8). Again, you may need to resize the dialog.



Statistic: sum(x) / 200

Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
0.36	5000	0.359854	0.0343107	0.29	0.305	0.36	0.415	0.43

Figure 6-8: StatCrunch resample results

7) Click the > button (lower right) to see the StatCrunch histogram (Figure 6-9)

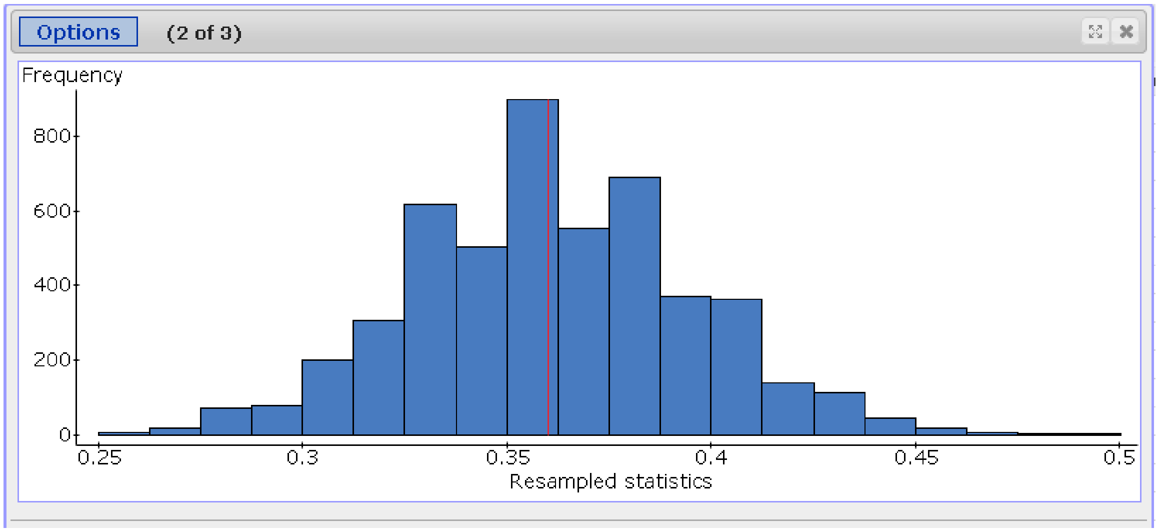


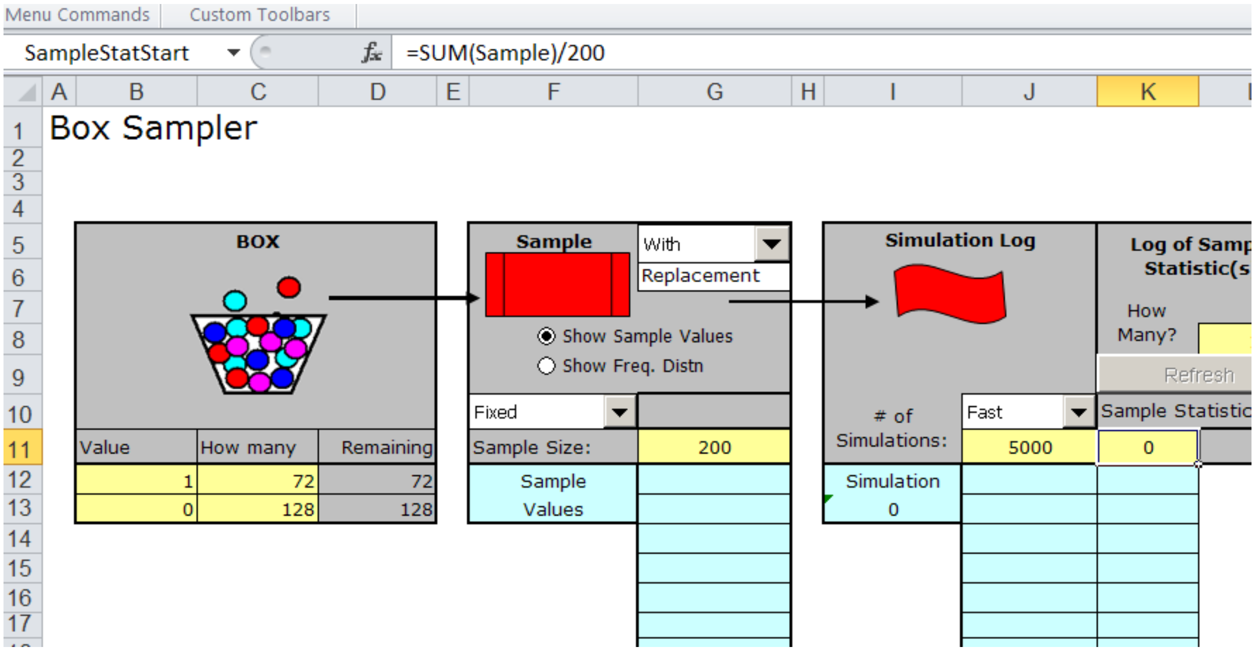
Figure 6-9: StatCrunch histogram

Roughly speaking, we can see from the percentiles in Fig. 5-8 that the vast majority of the results are between 30% and 42%.

Solving the Voter Survey Problem Using Box Sampler

1) Start Box Sampler and click OK on the opening dialog

2) Edit Box Sampler so it is identical to Figure 6-10.



=SUM(Sample)/200

Value	How many	Remaining
1	72	72
0	128	128

With Replacement

● Show Sample Values

○ Show Freq. Distn

Fixed

Sample Size: 200

Fast

# of Simulations: 5000

Sample Statistic: 0

Figure 6-10: Box Sampler setup for the voter survey problem

3) The Sample Statistic cell (K11 in Figure 6-10) should have the formula “=SUM(Sample)/200”

4) Click the “Simulation: Simulate” button (Figure 6-11) or select “Simulate” from the Box Sampler Menu.

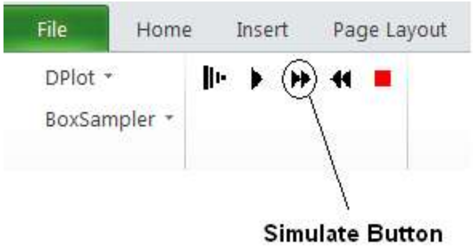


Figure 6-11: Simulate button

5) When the simulation is complete (be patient!), create a histogram by selecting the “BoxSampler -> StatCharts -> Histogram” menu. Use “Stat1” for the Data Input Range. A histogram is shown in Figure 6-12.

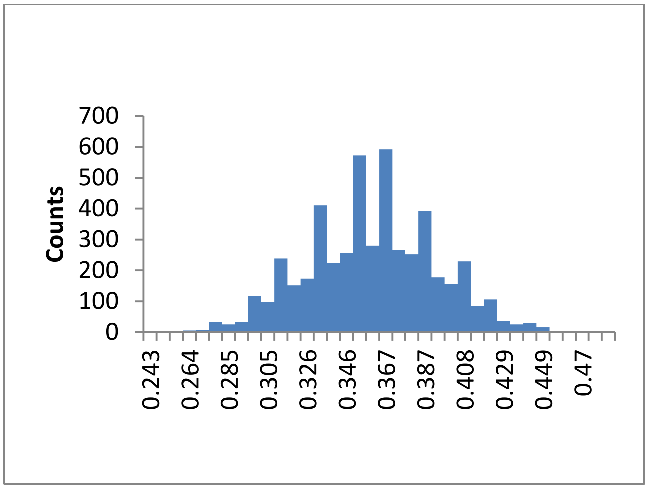


Figure 6-12: Box Sampler histogram

Roughly speaking, we can see that the vast majority of the results are between 30% and 42%.

7 Confidence intervals

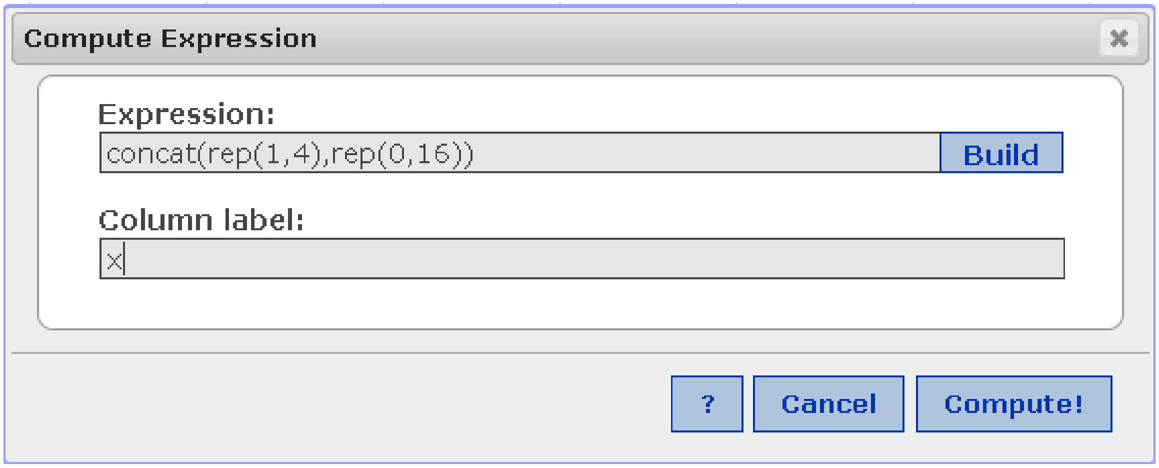
7.3 Confidence Intervals for a Mean

There are video tutorials for RSXL, StatCrunch, and Box Sampler for the Toyota price data. These videos can be found at www.introductorystatistics.com in the Book menu under the Videos sub-menu.

7.6 Confidence Intervals for a Single Proportion

StatCrunch Procedure

1) Open StatCrunch. Choose Data-> Compute Expression
2) Enter: concat(rep(1,4),rep(0,16)) in the Expression edit box
3) New column name x
4) Click Compute! (figure 7-1)

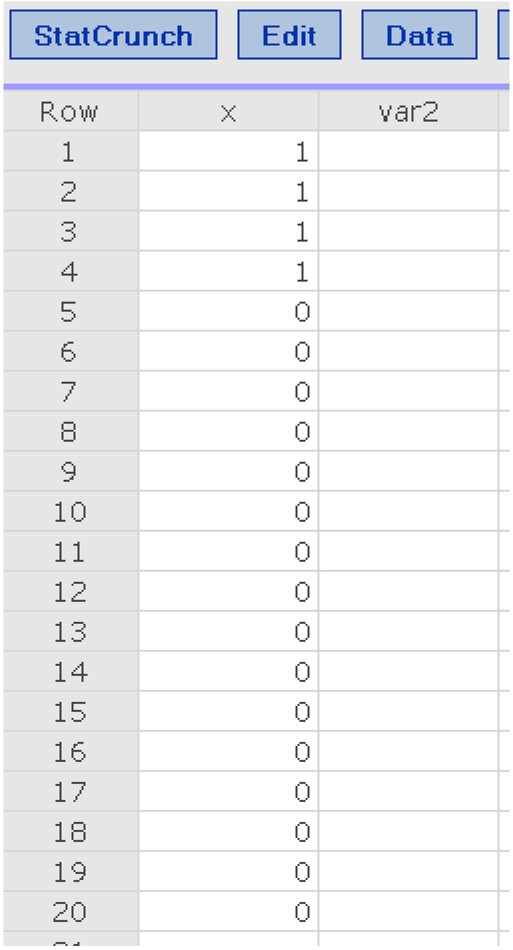


Expression: concat(rep(1,4),rep(0,16))

Column label: x

Figure 7-1: StatCrunch Compute expression dialog

5) The returned data is found in column x (figure 7-2)



Row	x	var2
1	1	
2	1	
3	1	
4	1	
5	0	
6	0	
7	0	
8	0	
9	0	
10	0	
11	0	
12	0	
13	0	
14	0	
15	0	
16	0	
17	0	
18	0	
19	0	
20	0	

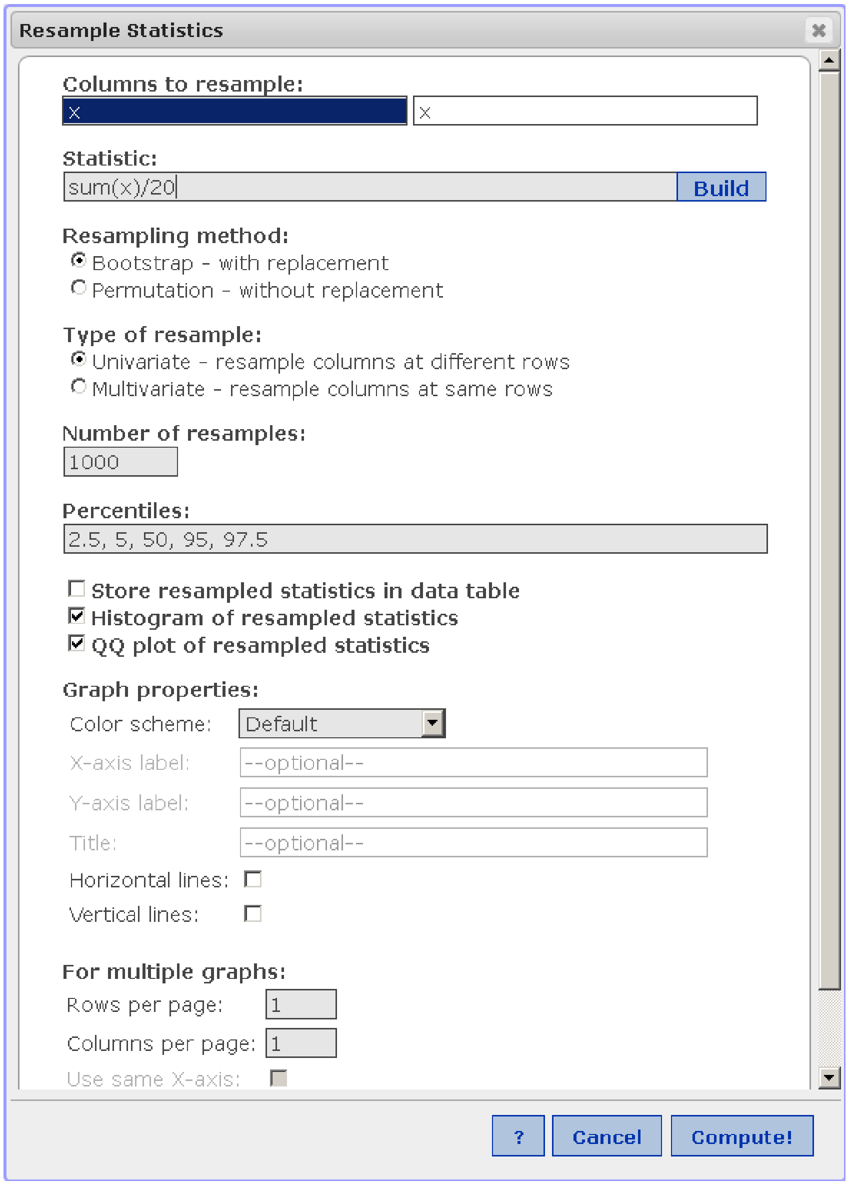
Figure 7-2: Data in column x

6) Choose Stat-> Resample-> Statistic

7) Select column x as the Columns to resample

8) Enter: sum(x)/20 as the Statistic. Dividing the sum by 20 calculates the decimal proportion of returns.

9) Click Compute! (figure 7-3) (StatCrunch automatically sets the size of the resample at 20, the same size as the original box (column x)



Columns to resample: x | x

Statistic: sum(x)/20

Resampling method:
● Bootstrap - with replacement
○ Permutation - without replacement

Type of resample:
● Univariate - resample columns at different rows
○ Multivariate - resample columns at same rows

Number of resamples: 1000

Percentiles: 2.5, 5, 50, 95, 97.5

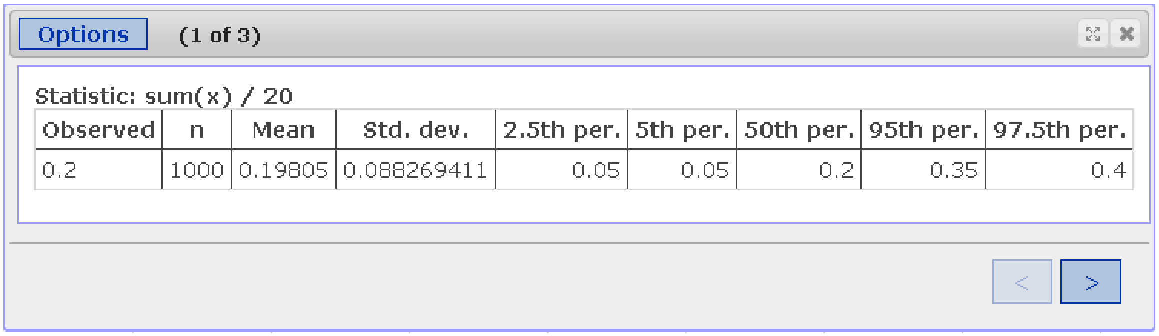
☐ Store resampled statistics in data table
☑ Histogram of resampled statistics
☑ QQ plot of resampled statistics

Graph properties:
Color scheme: Default
X-axis label: --optional--
Y-axis label: --optional--
Title: --optional--
Horizontal lines: ☐
Vertical lines: ☐

For multiple graphs:
Rows per page: 1
Columns per page: 1
Use same X-axis: ☐

Figure 7-3: Resample Statistic dialog

10) The 90% confidence interval goes from 0.05 to 0.35: the 5th percentile = 0.05 and the 95th percentile = 0.35 (figure 7-4).



Statistic: sum(x) / 20

Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
0.2	1000	0.19805	0.088269411	0.05	0.05	0.2	0.35	0.4

Figure 7-4: 90% confidence interval (between 5th and 95th percentiles)

11) View histogram by clicking the > button (figure 7-5)

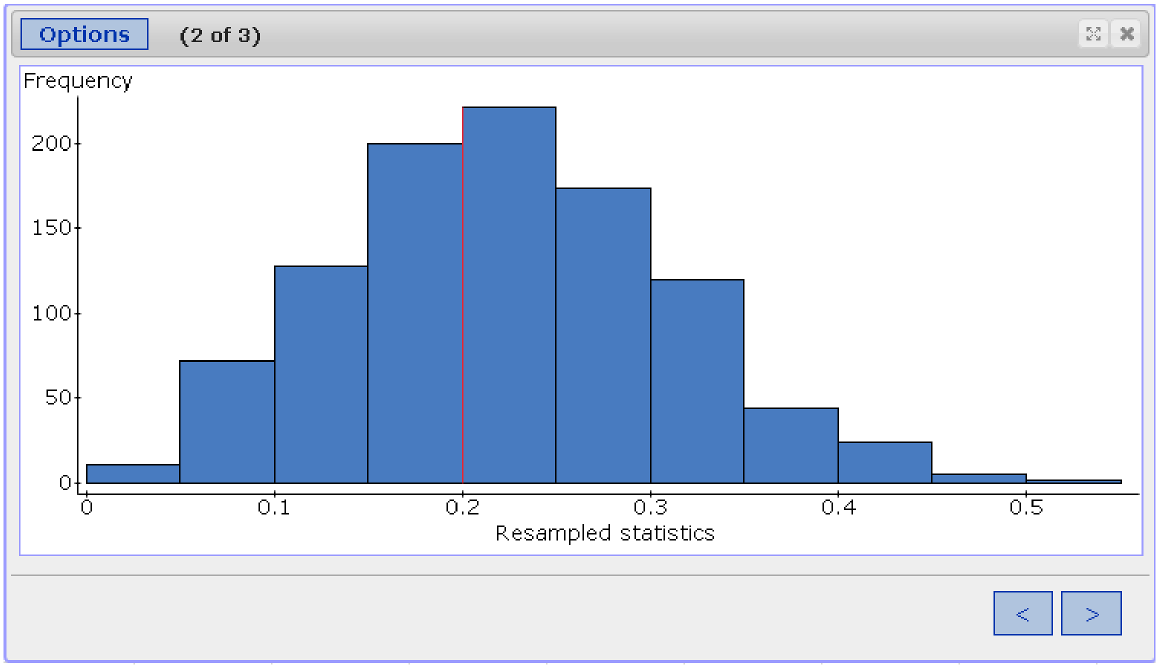


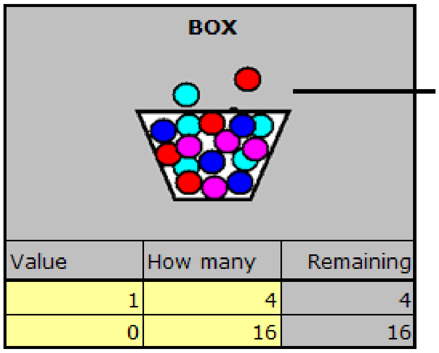
Figure 7-5: StatCrunch histogram

Box Sampler Procedure

1) Start Box Sampler

2) Accept default settings in the opening dialog

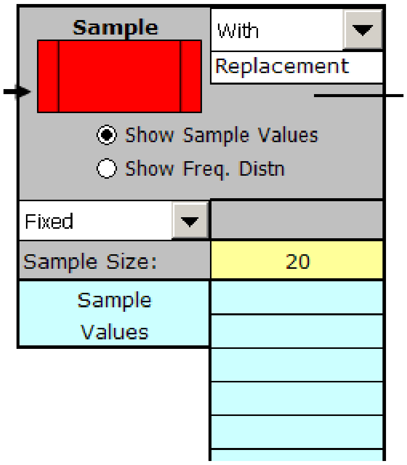
3) Create the Box (figure 7-6)



Value	How many	Remaining
1	4	4
0	16	16

Figure 7-6: Box setup

4) Enter a sample size of 20 (figure 7-7)



With Replacement

● Show Sample Values

○ Show Freq. Distn

Fixed

Sample Size: 20

Figure 7-7: Sample size

5) Enter 1000 for the # of Simulations (figure 7-8)

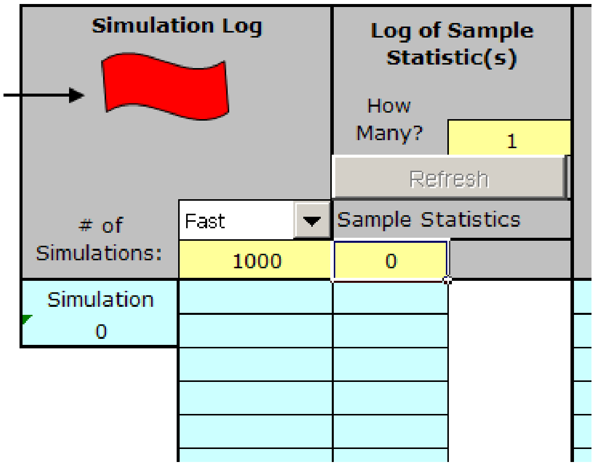
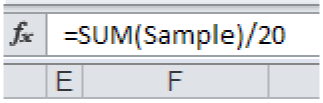


Figure 7-8: # of simulations

6) The Sample Statistics formula (Cell K11) is “=SUM(Sample)/20” (figure 7-9). Recall that “Sample” is the named range for the Sample Values. Dividing by 20 calculates the proportion of returns.



=SUM(Sample)/20

Figure 7-9: Sample statistic formula in cell K11

7) Click the “Simulation: One Element” button several times to see the first few elements of the resample.

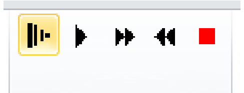


Figure 7-10: Simulation: One Element button

8) Click the “Simulation: One Resample” button to fill in the remainder of the first resample. You can click this button a few more times to watch the “Stat1” named range (output range below Sample Statistics formula) begin to fill.

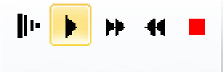


Figure 7-11: Simulation: One Resample

9) Click the “Simulation: Simulate” button to complete the entire simulation.

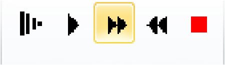
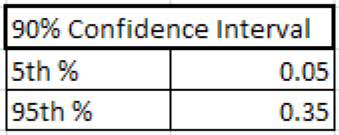


Figure 7-12: Simulation: Simulate

10) Calculate the 90% Confidence Interval by selecting an empty cell and entering the Excel function for the 5th percentile as follows: =PERCENTILE(Stat1, 0.05)

11) In the cell below the 5th percentile formula, enter: =PERCENTILE(Stat1, 0.95) to calculate the 95th percentile.

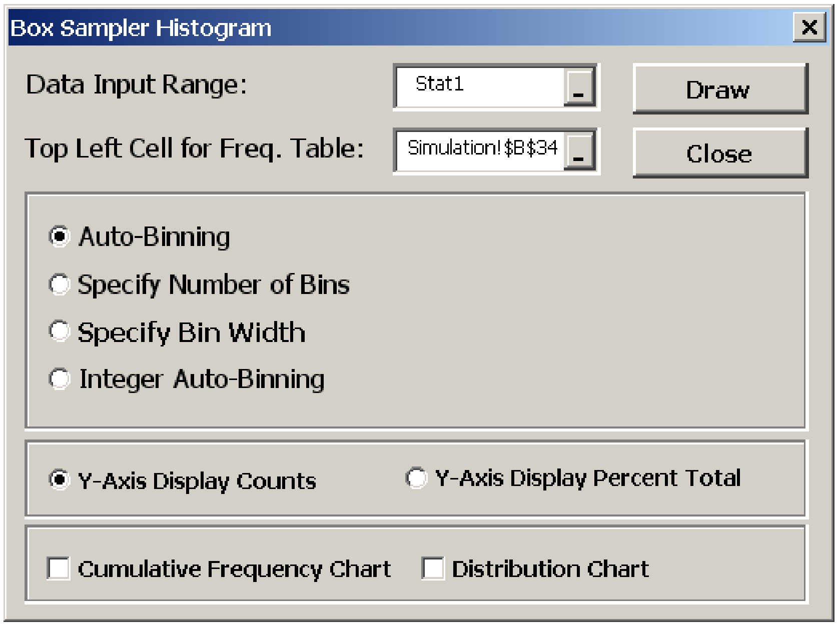
12) The 90% confidence interval is from 0.05 to 0.35 (identical to the StatCrunch simulation)



90% Confidence Interval	
5th %	0.05
95th %	0.35

Figure 7-13: 90% confidence interval

13) Create a histogram by selecting "BoxSampler-> StatCharts-> Histogram" and enter "Stat1" for the Data Input Range and an empty cell for the "Top Left Cell for Freq. Table". Click Draw (figure 7-14)



Box Sampler Histogram

Data Input Range: Stat1

Top Left Cell for Freq. Table: Simulation!$B$34

● Auto-Binning
○ Specify Number of Bins
○ Specify Bin Width
○ Integer Auto-Binning

● Y-Axis Display Counts ○ Y-Axis Display Percent Total

☐ Cumulative Frequency Chart ☐ Distribution Chart

Figure 7-14: Box Sampler Histogram dialog

14) Box Sampler Histogram (figure 7-15)

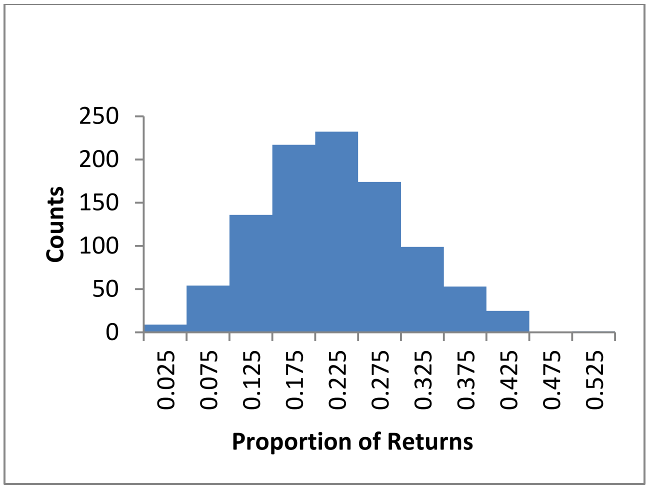
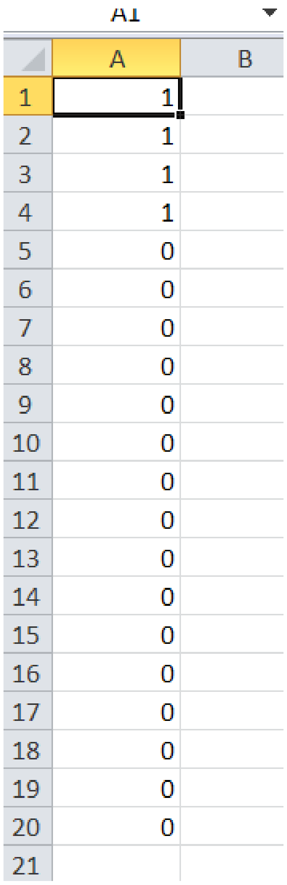


Figure 7-15: Box Sampler Histogram output

Resampling Stats Procedure

1) After loading the Resampling Stats add-in, use the Urn feature or manually enter 4 “1’s” and 16 “0’s” in cells A1:A20 (figure 7-16).

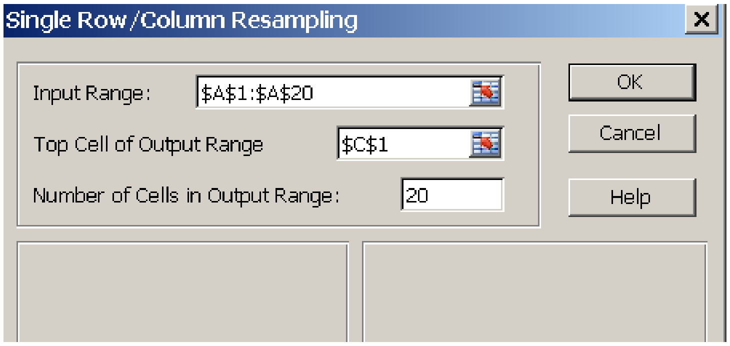


	A	B
1	1	
2	1	
3	1	
4	1	
5	0	
6	0	
7	0	
8	0	
9	0	
10	0	
11	0	
12	0	
13	0	
14	0	
15	0	
16	0	
17	0	
18	0	
19	0	
20	0	
21		

Figure 7-16: Product purchase and returns data

2) With cell A1 selected, choose “Resample” (with replacement) from the Resampling menu.

3) Select cell C1 for the Top Cell of the Output Range and use 20 for the Number of Cells in the Output Range. Click “OK” (figure 7-17).



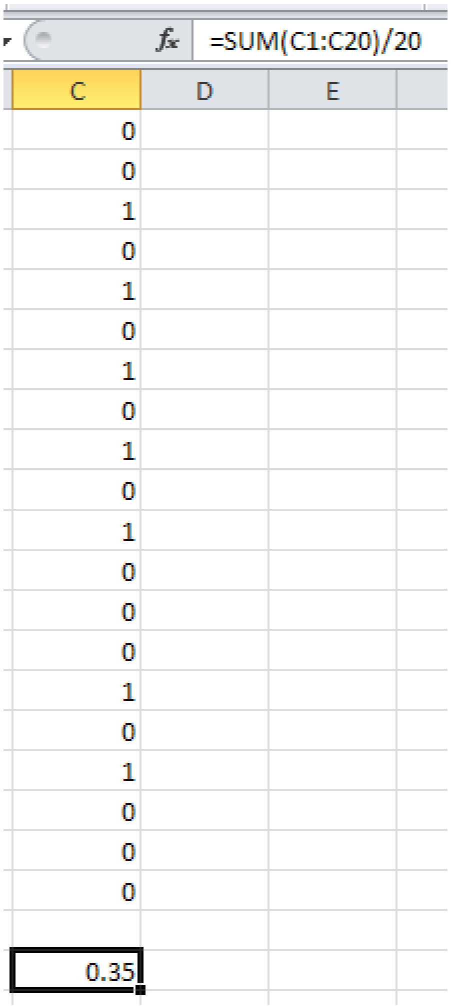
Input Range: $A$1:$A$20

Top Cell of Output Range $C$1

Number of Cells in Output Range: 20

Figure 7-17: Resample dialog

4) Use the Excel function “=SUM(C1:C20)/20” to find the proportion of returns. This formula can be placed in any empty cell near the resample output range (for example, cell C22).



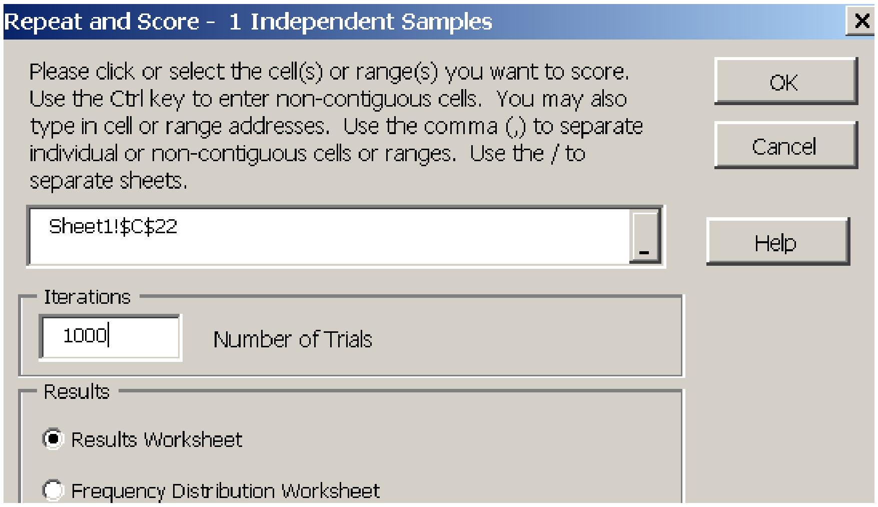
=SUM(C1:C20)/20

C	D	E
0		
0		
1		
0		
1		
0		
1		
0		
1		
0		
1		
0		
0		
0		
1		
0		
1		
0		
0		
0		
0.35		

Figure 7-18: Find the proportion of returns using the formula =SUM(C1:C20)/20

5) With the “=SUM(C1:C20)/20” cell selected, choose “Repeat and Score” from the Resampling menu.

6) Use 1000 for the Number of Trials. Click “OK” (figure 7-19).



Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

Sheet1!$C$22

Iterations

1000 Number of Trials

Results

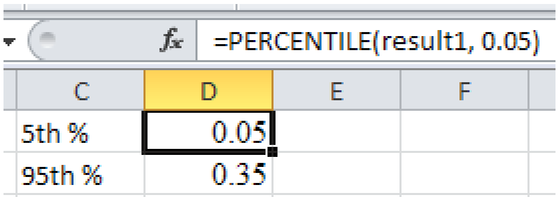
● Results Worksheet

○ Frequency Distribution Worksheet

Figure 7-19: Repeat and Score dialog

7) When the simulation has finished, order the data (sort) on the Results sheet in column A from low to high.

8) The range from A1 to A1000, which contains the sums of all the resamples, is named "result1". Find the 5th and 95th percentiles as we did in Box Sampler. In an empty cell (D1 for example), enter the native Excel formula: =PERCENTILE(result1, 0.05) for the 5th percentile and in cell D2, enter =PERCENTILE(result1, 0.95) for the 95th percentile. Again we see a 90% confidence interval from 0.05 to 0.35 (figure 7-20).

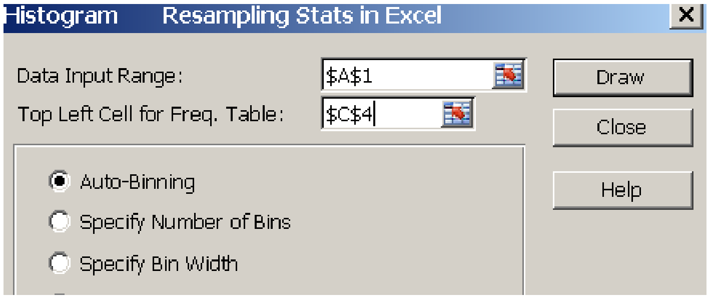


=PERCENTILE(result1, 0.05)

C	D	E	F
5th %	0.05		
95th %	0.35		

Figure 7-20: 90% confidence interval using Excel's =PERCENTILE() function

9) Create a histogram by choosing the histogram feature from the Resampling menu. Select cell A1 for the Data Input Range and any convenient empty cell for the Top Left Cell for Freq. Table. Click "OK" (figure 7-21).



Histogram Resampling Stats in Excel

Data Input Range: $A$1

Top Left Cell for Freq. Table: $C$4

● Auto-Binning

○ Specify Number of Bins

○ Specify Bin Width

Figure 7-21: Resampling Stats Histogram dialog

10) The histogram is similar to Box Sampler.

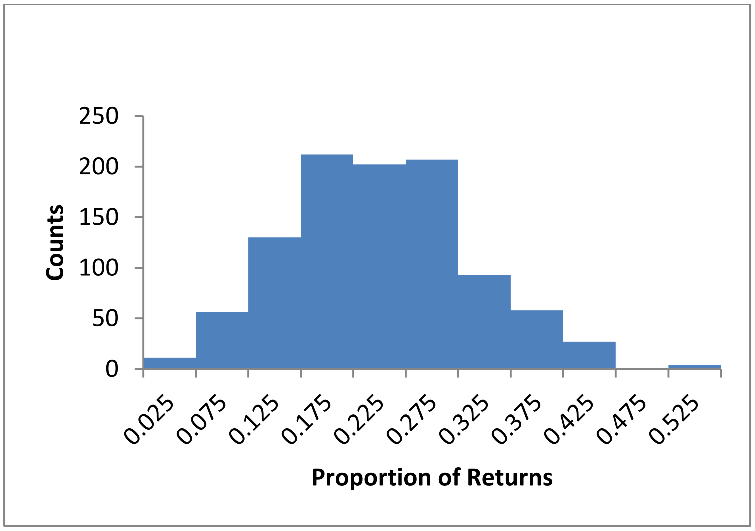


Figure 7-22: Resampling Stats histogram

Note: In the examples above, we divided the sum of returns in each resample by 20 in order to calculate the proportion of returns. This is somewhat inefficient and results in an unnecessary division for each trial. If we use 1000 trials, then we have 1000 divisions! A more efficient procedure would be to resample the sums of the returns and calculate a 90% confidence interval using the sums only. Then divide the 5th and 95th percentile sums by 20 to produce the proportions. To illustrate, in each of the above examples the 5th percentile sum = 1 and the 95th percentile sum = 7. Dividing both 1 and 7 by 20 (1/20 and 7/20) returns the correct percentile values for the proportions, 0.05 and 0.35 respectively. However, computer time is inexpensive and there is nothing wrong with doing a few extra divisions!

Binomial Formula

A baseball player has a 0.3 probability of getting a hit in each at bat (a 0.300 hitter). What is the probability that he will get exactly 3 hits in 5 at-bats?

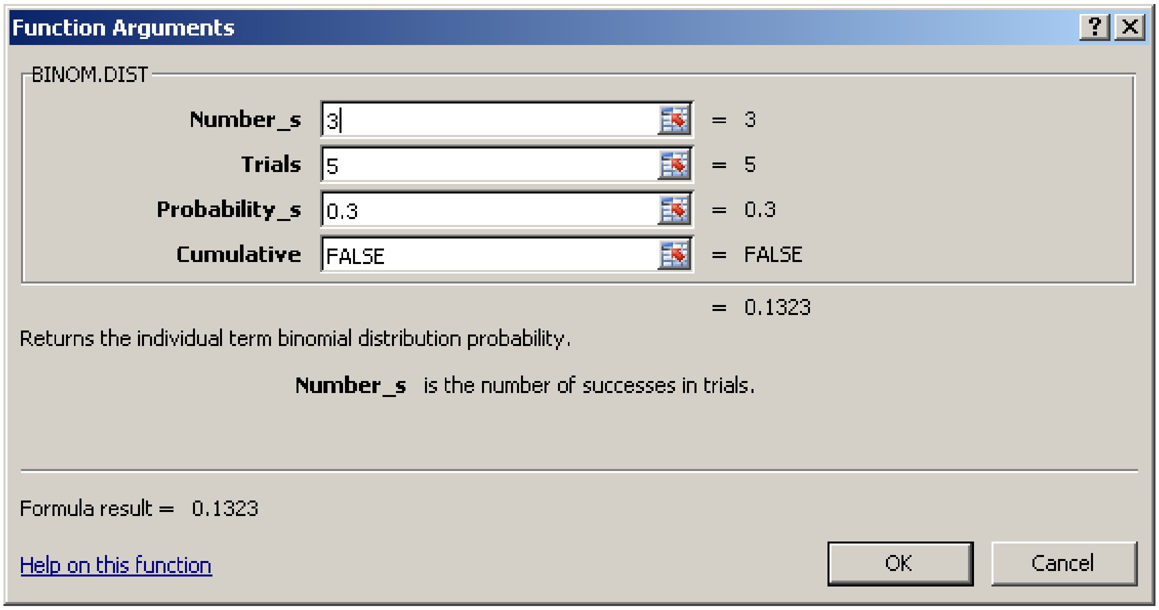
Excel

We can use Excel to provide a formula approach as follows:

1) Start Excel and select a cell to hold the needed function

2) Click on the ***fx*** button to start the Insert Function dialog

3) Enter BINOM.DIST in the Search for a function and click OK



Function Arguments

BINOM.DIST

Number_s 3 = 3

Trials 5 = 5

Probability_s 0.3 = 0.3

Cumulative FALSE = FALSE

= 0.1323

Returns the individual term binomial distribution probability.

Number_s is the number of successes in trials.

Formula result = 0.1323

Figure 7-23: =BINOM.DIST arguments

4) The =BINOM.DIST function takes 4 arguments. The first argument is Number_s, the number of successes in trials. In this problem, this argument = 3, which is the number of hits.

5) The second argument is the number of trials. There are five at-bats, so the number of trials is 5.

6) The third argument is the probability, 0.3.

7) The final argument is either TRUE or FALSE depending on whether we want a cumulative (TRUE) or exact (FALSE) value. We want the probability of getting <u>exactly</u> 3 hits, so the argument is FALSE.

8) **Note**: In step 7, if we were to choose TRUE, we would be determining the probability of getting 3 or fewer hits in 5 at-bats. This results in a much higher probability since we would count 3, 2, 1 or 0 hits as successes.

9) You can see the value of =BINOM.DIST(3, 5, 0.3, FALSE) in the Function Arguments dialog (= 0.1323), but click OK to complete the function wizard and place the formula in the worksheet.

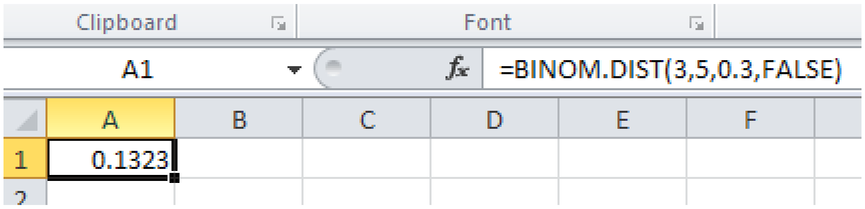
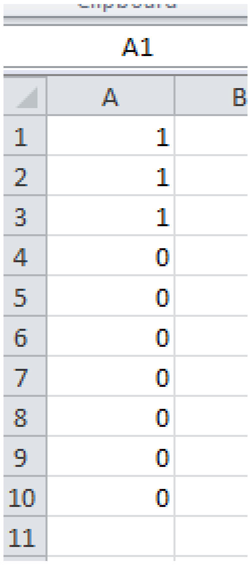


Figure 7-24: Probability of getting exactly 3 hits

Resampling Stats Procedure

1) Load Resampling Stats into Excel and create the following data range in cells A1:A10 to simulate the 3 chances in 10 of getting a hit (0.300 batting average):

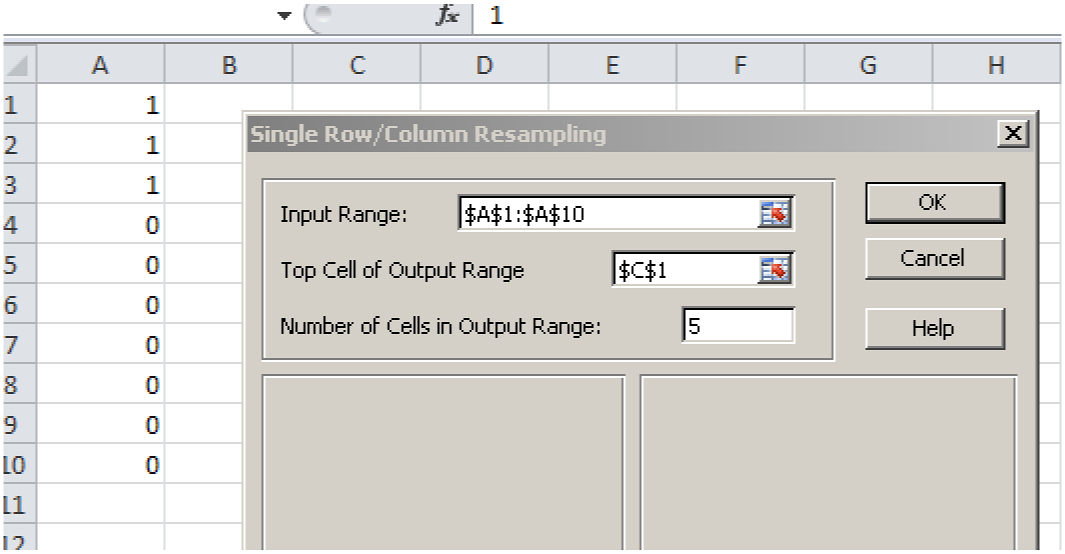


	A
1	1
2	1
3	1
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	

Figure 7-25: Resampling Stats model for hits in 5 at-bats

2) Select cell A1 (the top cell in the data range) and choose Resample from the Resampling menu (or click the R button in the Resampling toolbar).

3) Select cell C1 as the top cell of the output range and enter 5 as the number of output cells:



fx 1

	A
1	1
2	1
3	1
4	0
5	0
6	0
7	0
8	0
9	0
10	0

Input Range: $A$1:$A$10

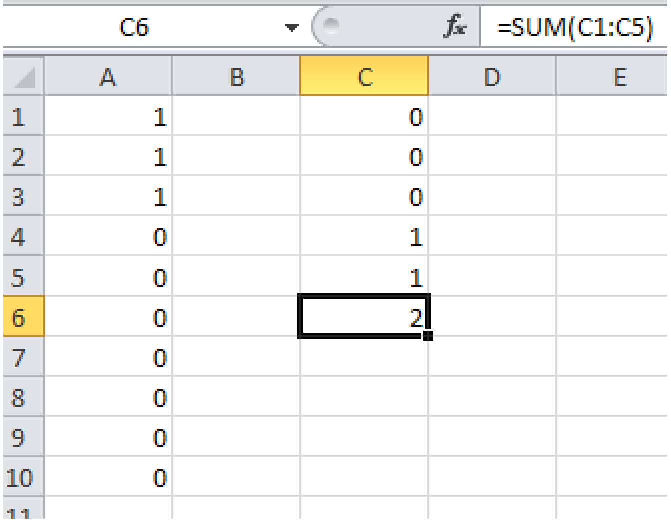
Top Cell of Output Range $C$1

Number of Cells in Output Range: 5

Figure 7-26: Resample dialog for hits example

4) Click OK

5) In cell C6, enter the formula =SUM(C1:C5)

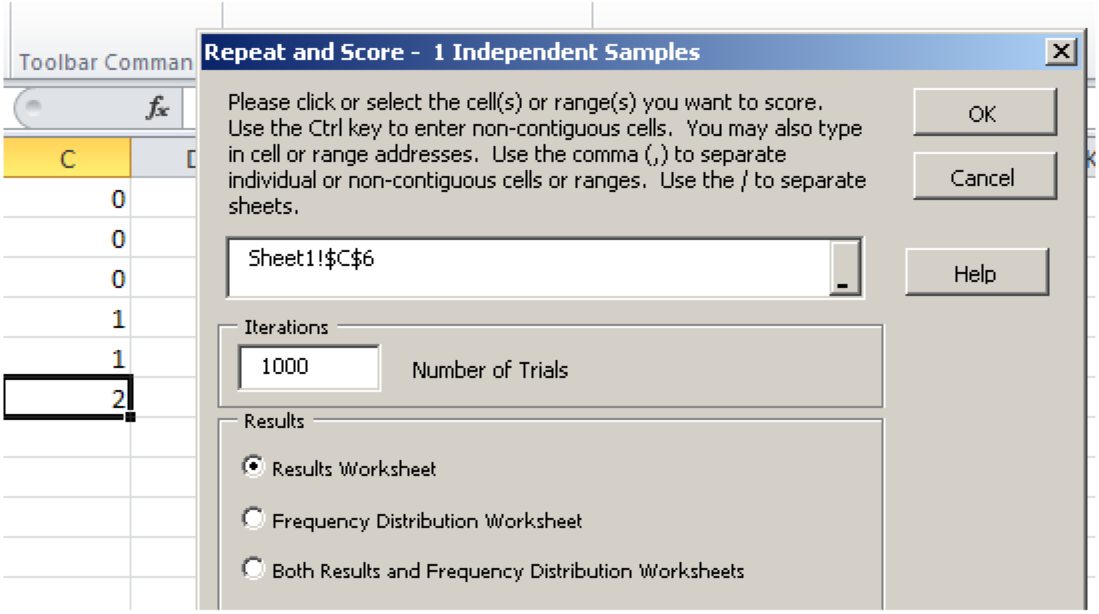


C6 fx =SUM(C1:C5)

	A	B	C	D	E
1	1		0		
2	1		0		
3	1		0		
4	0		1		
5	0		1		
6	0		2		
7	0				
8	0				
9	0				
10	0				

Figure 7-27: Calculate the number of hits

6) Select cell C6 (or whichever cell contains the =SUM() formula) and choose Repeat and Score from the Resampling Menu:



Repeat and Score - 1 Independent Samples

Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

Sheet1!$C$6

Iterations

1000 Number of Trials

Results

● Results Worksheet

○ Frequency Distribution Worksheet

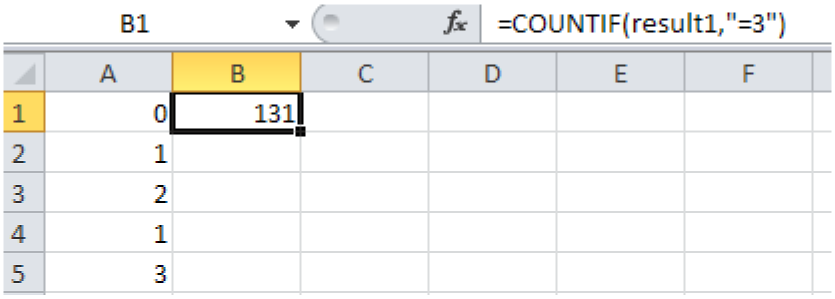
○ Both Results and Frequency Distribution Worksheets

C
0
0
0
1
1
2

Figure 7-28: Repeat and score the number of hits

7) Enter 1000 trials and click OK

8) In the Results sheet, enter =COUNTIF(result1, “=3”) in cell B1



B1 =COUNTIF(result1,"=3")

	A	B	C	D	E	F
1	0	131				
2	1					
3	2					
4	1					
5	3					

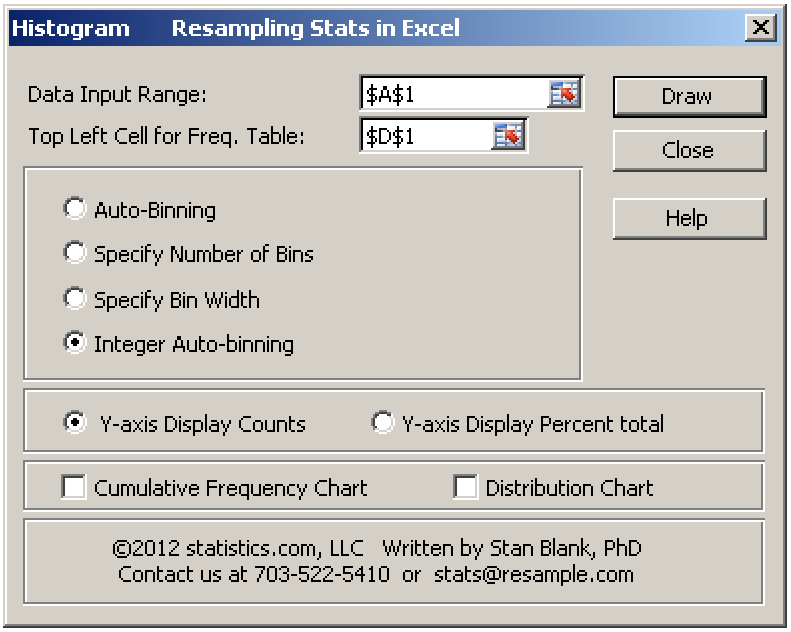
Figure 7-29: Instances in 1000 trials of exactly 3 hits in 5 at-bats

9) In this example, there were 131 instances of 3 hits in 5 at-bats. This is a probability of 131/1000 or 0.131.

10) To create a histogram, select the top cell of the output range in the Results worksheet (cell A1).

11) Choose Histogram from the Resampling menu or toolbar.

12) Select cell D1 as the Top left cell for the frequency table and select the Integer Auto-binning option.



Histogram Resampling Stats in Excel

Data Input Range: $A$1

Top Left Cell for Freq. Table: $D$1

○ Auto-Binning
○ Specify Number of Bins
○ Specify Bin Width
● Integer Auto-binning

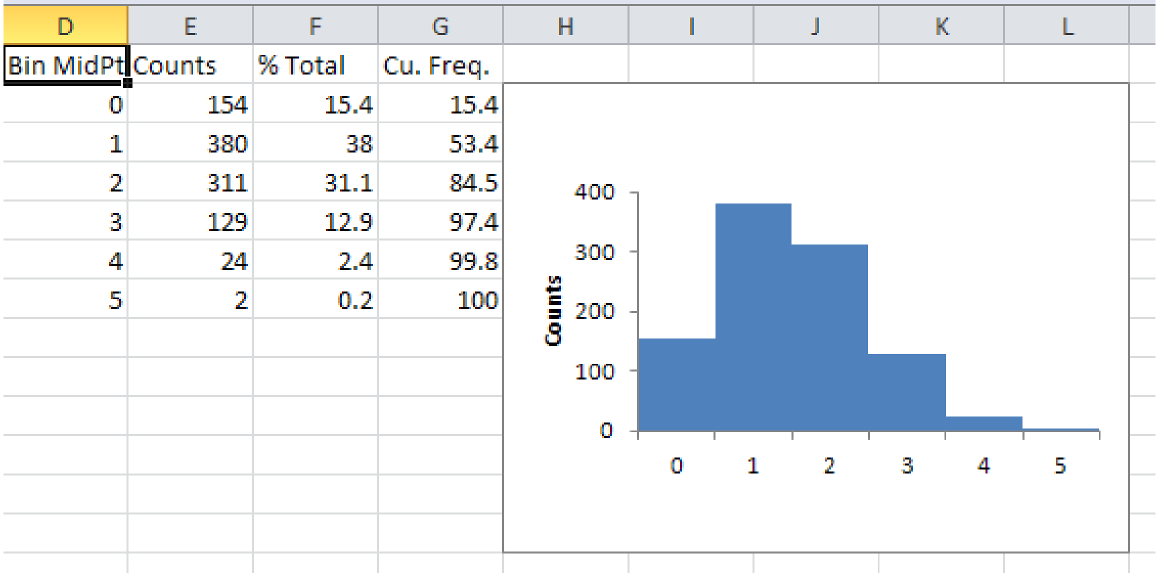
● Y-axis Display Counts ○ Y-axis Display Percent total

☐ Cumulative Frequency Chart ☐ Distribution Chart

©2012 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 7-30: Histogram dialog for hits example

13) Click Draw



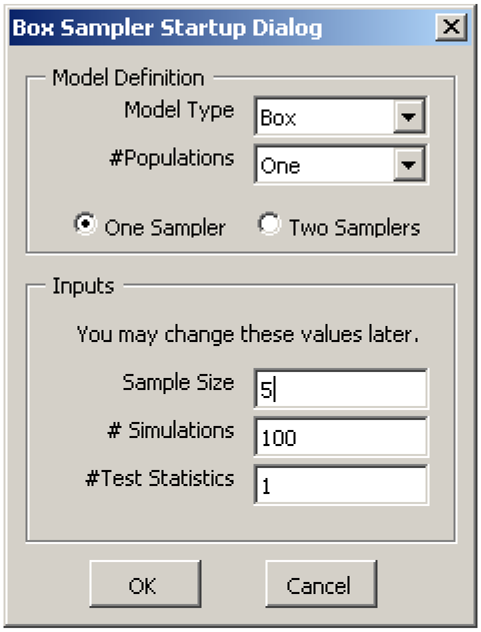
Bin MidPt	Counts	% Total	Cu. Freq.
0	154	15.4	15.4
1	380	38	53.4
2	311	31.1	84.5
3	129	12.9	97.4
4	24	2.4	99.8
5	2	0.2	100

Figure 7-31: Histogram for hits example using Resampling Stats

14) The histogram in Figure 7-31 is similar to figure 7-5 in the text. Note the Counts in each bin. The figure above was created using a new run of 1000 trials. In this new example, there were 129 instances of exactly 3 hits in 5 at-bats. How do the counts (divided by 1000) compare with the table preceding figure 7-5 in the text? See if you can figure out the meaning of the %Total and Cu. Freq. (Cumulative Frequency) columns.

Box Sampler

1) Load Box Sampler in Excel and choose the Box Sampler>New Model menu. Enter 5 for the Sample Size in the opening dialog and click OK.



Box Sampler Startup Dialog

Model Definition

Model Type: Box

#Populations: One

● One Sampler ○ Two Samplers

Inputs

You may change these values later.

Sample Size: 5

Simulations: 100

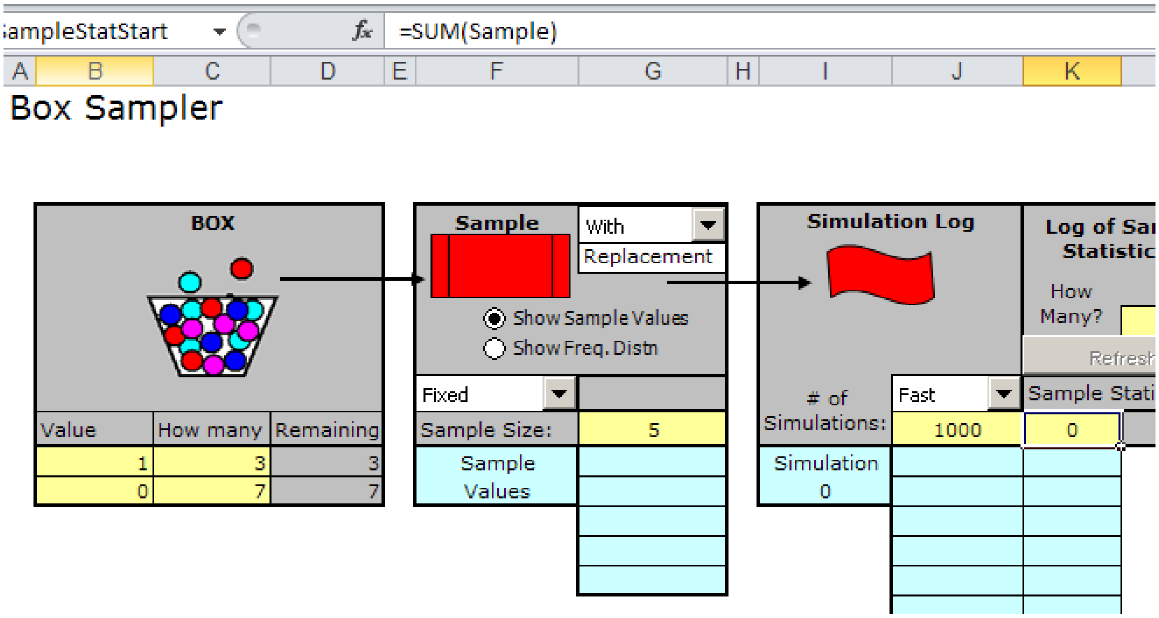
#Test Statistics: 1

Figure 7-32: Box Sampler opening dialog

2) In the Box, enter 1 as the first value and 3 for How Many. Enter 0 for the second value and 7 for How Many.

3) Enter 1000 for the # of Simulations and in cell K11 (the Sample Statistic cell), enter the formula: =SUM(Sample)

   (Recall that Sample is the name of the range holding the sample values)



ampleStatStart =SUM(Sample)

Value	How many	Remaining
1	3	3
0	7	7

With Replacement

● Show Sample Values
○ Show Freq. Distn

Fixed

Sample Size: 5

Fast

# of Simulations: 1000

Sample Stati 0

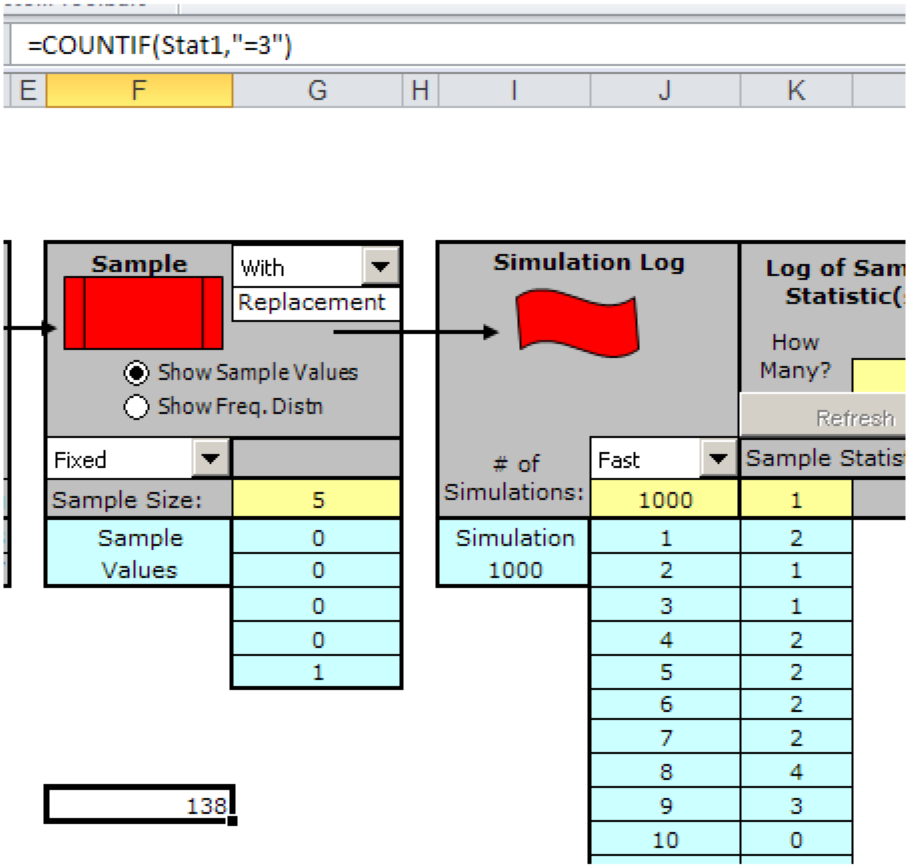
Figure 7-33: Box Sampler setup for hits example

4) In the Box Sampler toolbar or menu, click “Simulate”

5) The sum (hits) from each sample will be placed in the Stat1 range below the Sample Statistic cell.

6) When the simulation has completed, choose an empty cell, say cell F20, and enter the following formula: =COUNTIF(Stat1, “=3”)

   (Recall that Stat1 is the name of the range of output values below the Sample Statistic)



=COUNTIF(Stat1,"=3")

With Replacement

● Show Sample Values
○ Show Freq. Distn

Fixed

Sample Size: 5

Sample Values: 0, 0, 0, 0, 1

138

How Many?

Fast

Sample Statis

# of Simulations: 1000

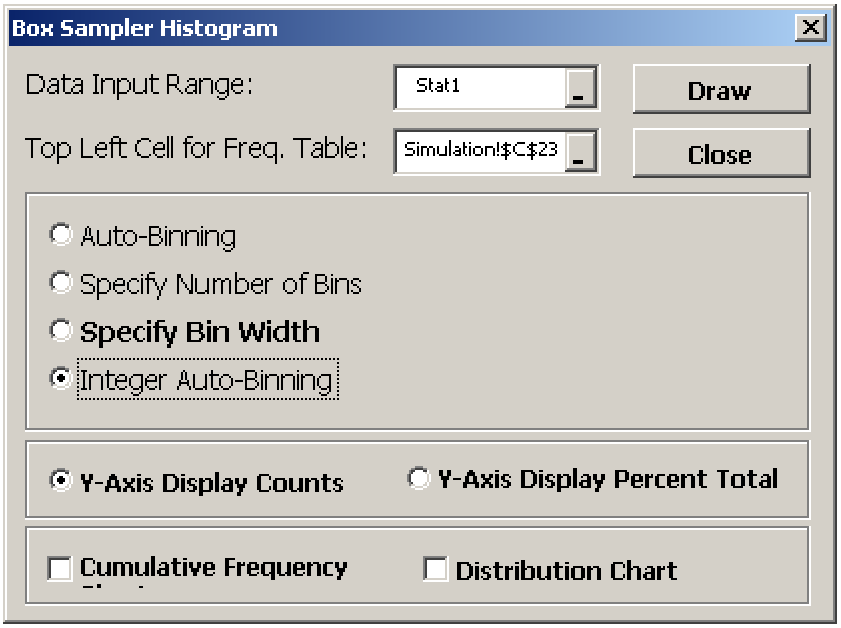
Simulation 1000	1000	1
	1	2
	2	1
	3	1
	4	2
	5	2
	6	2
	7	2
	8	4
	9	3
	10	0

Figure 7-34: 138 instances in 1000 trials of exactly 3 hits in 5 at-bats

7) In this example, there were 138 instances of exactly 3 hits in 5 at-bats. This gives us a probability of 0.138

8) You can create a histogram by selecting the Box Sampler>StatCharts>Histogram menu item.

9) Enter Stat1 in the Input cell and select an empty cell (say C23) for the Top left cell of the frequency table. Choose the Integer Auto-binning option.



Data Input Range: Stat1

Top Left Cell for Freq. Table: Simulation!$C$23

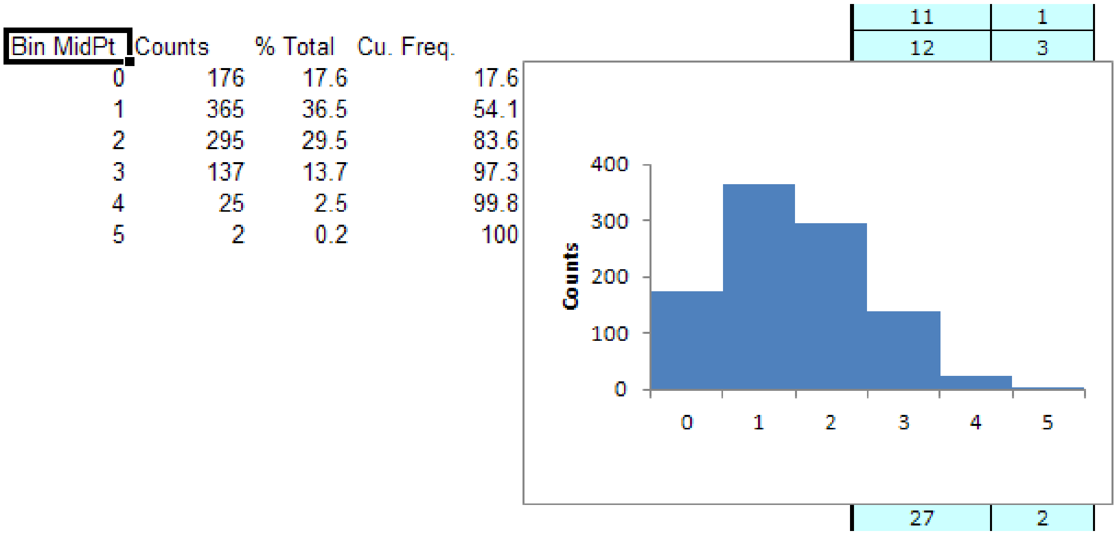
○ Auto-Binning
○ Specify Number of Bins
○ Specify Bin Width
● Integer Auto-Binning

● Y-Axis Display Counts
○ Y-Axis Display Percent Total

☐ Cumulative Frequency
☐ Distribution Chart

Figure 7-35: Box Sampler histogram for hits example

10) Click Draw



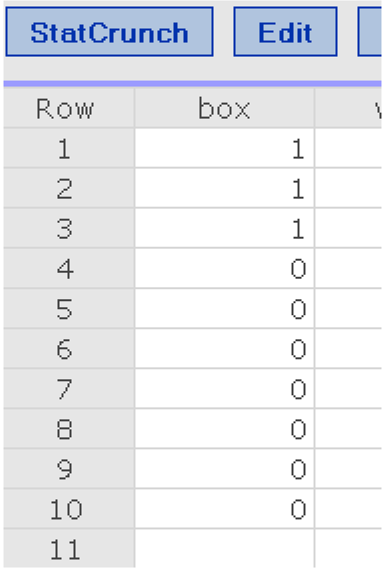
Bin MidPt	Counts	% Total	Cu. Freq.
0	176	17.6	17.6
1	365	36.5	54.1
2	295	29.5	83.6
3	137	13.7	97.3
4	25	2.5	99.8
5	2	0.2	100

Figure 7-36: Box Sampler histogram output

11) The histogram in figure 7-36 was based on a new run of 1000 trials. Note that it is similar to the histogram in figure 7-5 in the text. In this new run, there were 137 instances in 1000 trials of exactly 3 hits in 5 at-bats. Compare the counts above (divided by 1000) with the table preceding figure 7-5 in the text. **Note**: I had to make column E wider in order to view the values in the % Total column. What do you think these %Total values represent? What about the values in the Cu. Freq. column?

StatCrunch

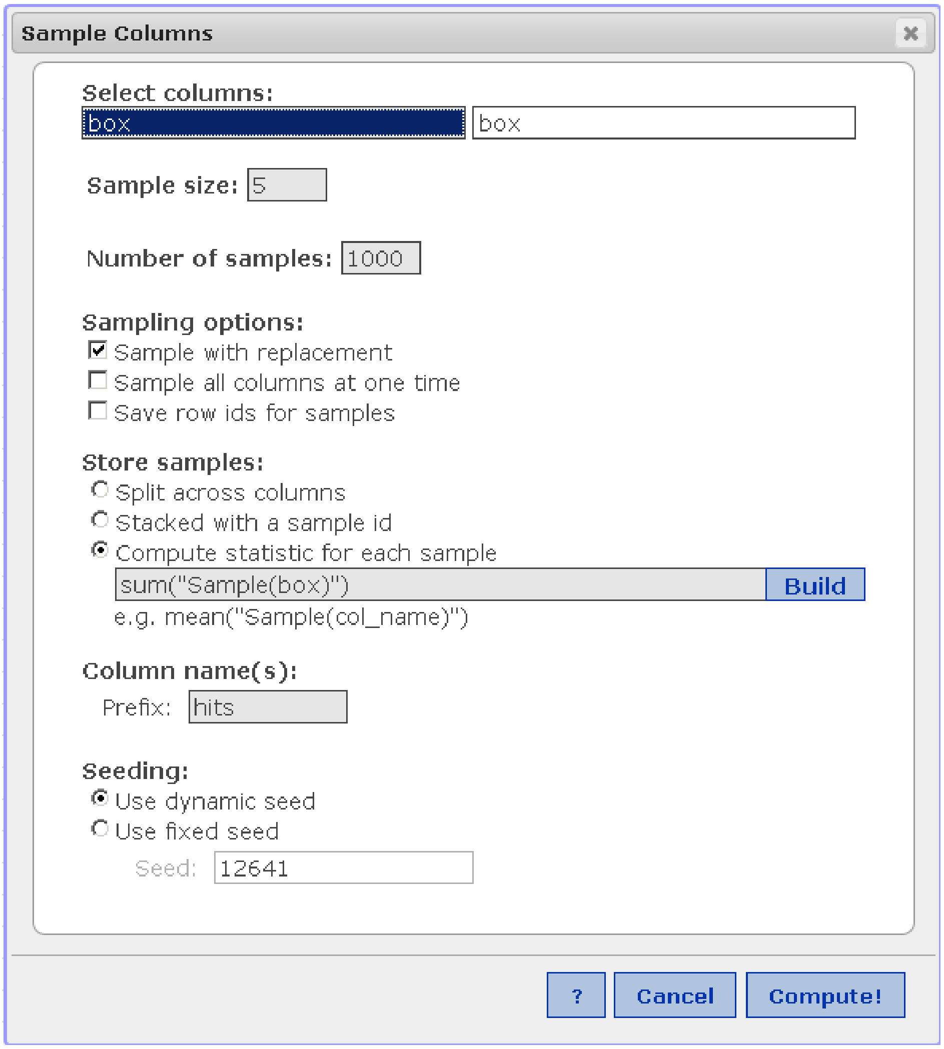
1) Start StatCrunch and enter 3 1s and 7 0s. Change the column name to “box” as shown below:



Row	box
1	1
2	1
3	1
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	

Figure 7-37: StatCrunch model for hits example

2) Select Data>Sample and choose box for the column. Enter 5 for the Sample Size, 1000 for the Number of Samples and select Sample with Replacement. Choose Compute Statistic and enter: sum("Sample(box)") as shown in the next figure.



Select columns: box | box

Sample size: 5

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
sum("Sample(box)")
e.g. mean("Sample(col_name)")

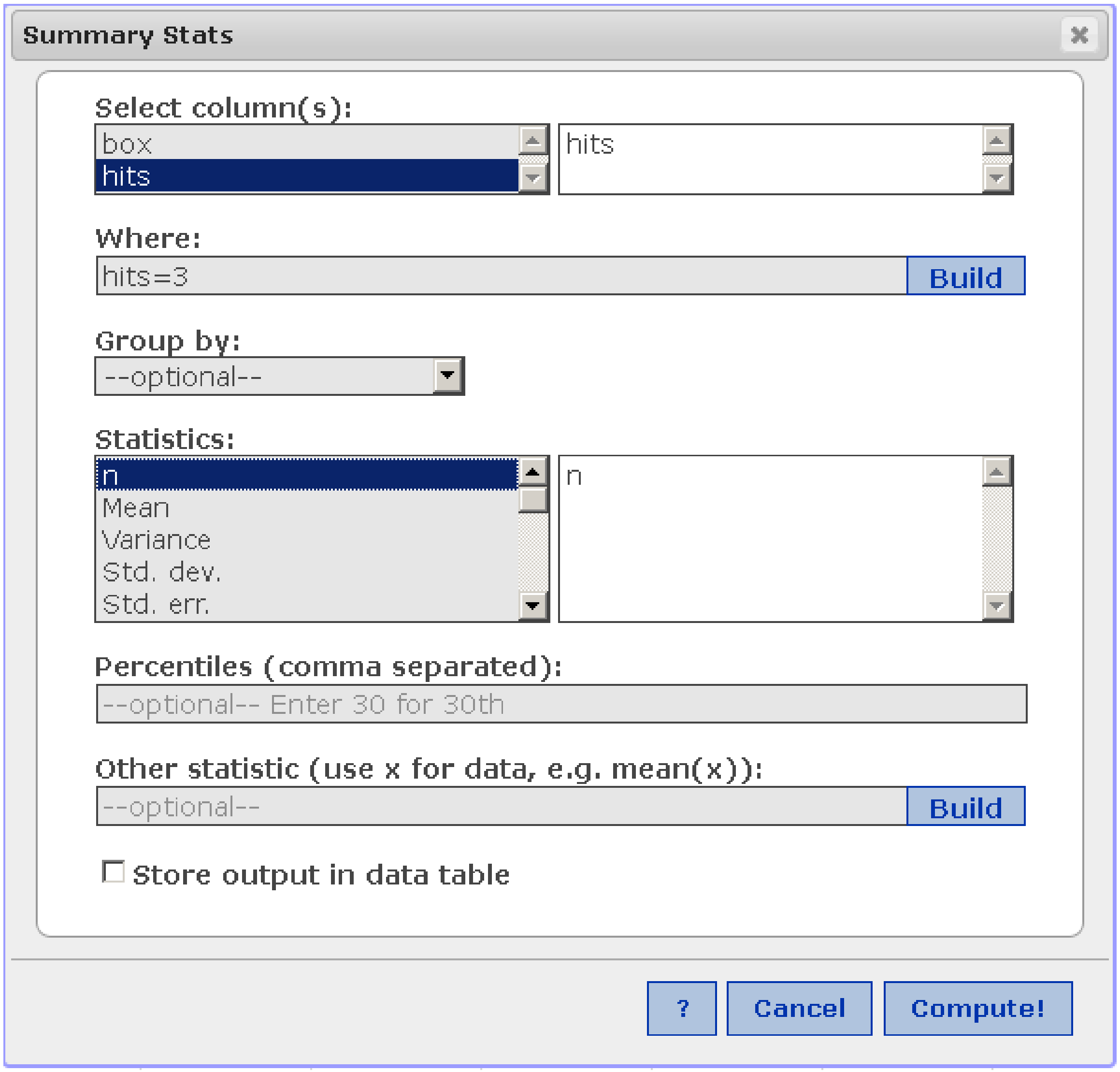
Column name(s):
Prefix: hits

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 7-38: Sample columns setup

3) Enter hits for the Column name and click Compute!

4) We now need to count the number of occurrences of exactly 3 hits. Select Stat>Summary Stats>Columns and choose the hits column. Enter “hits=3” in the Where: editbox (as follows):



Summary Stats

Select column(s): hits

Where: hits=3

Group by: --optional--

Statistics: n

Percentiles (comma separated): --optional-- Enter 30 for 30th

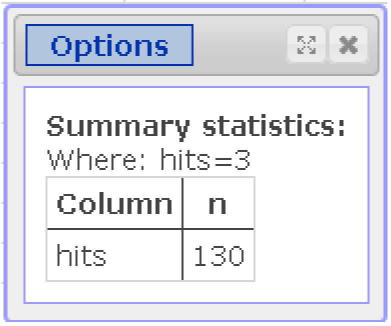
Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 7-39: Summary statistics

5) Click n in the Statistics box. The n entry provides a count of the number of 3’s in the hits column.

6) Click Compute!



Summary statistics:
Where: hits=3

Column	n
hits	130

Figure 7-40: 130 instances in 1000 trials of exactly 3 hits in 5 at-bats

7) There were 130 instances in 1000 trials of exactly 3 hits in 5 at-bats.

8) Making a histogram is simple. Click Graph>Histogram and choose the hits column. Click Compute!

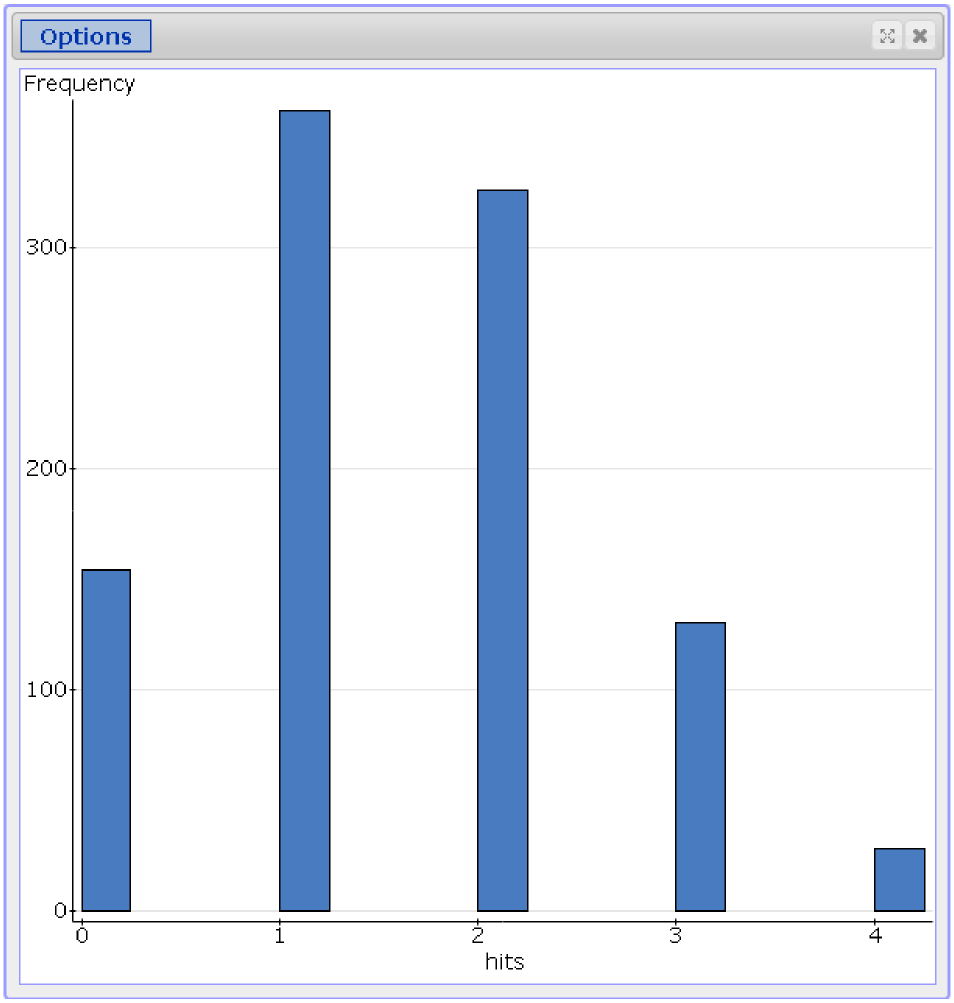


Figure 7-41: StatCrunch histogram for hits example

7.7 Confidence Interval for a Difference in Means

Resampling Stats Procedure

1) If you haven't done so already, copy (or download) the data from Vendor A in column A and the data for Vendor B in column B. If you have been working in RSXL, press reset in the RSXL menu to clear the model.

2) Highlight the entire data range, A1:B12 and select Resample from the RSXL menu.

3) Choose D1 as the Top Left Cell of the Output Range and the Resample Within Columns option. This option creates a two-box model by keeping the vendor columns separate and independent of each other. Click OK.

4) In cell F1, enter the formula "=AVERAGE(D1:D12)" and in cell F2, enter the formula "=AVERAGE(E1:E10)".
5) In cell F3, enter the formula "=F1-F2" to calculate the difference in means between the independent column (two-box) resamples.

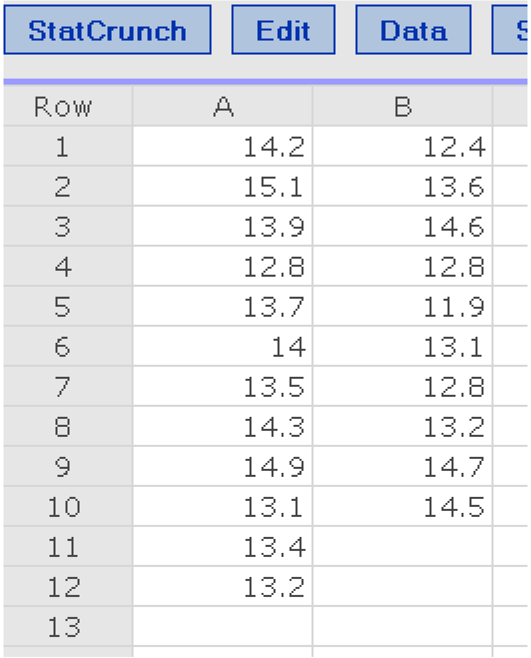
6) Highlight cell F3 as the score cell and select Repeat and Score. Try 1000 trials. Click OK.

7) Create a 90% confidence interval by entering the formula "=PERCENTILE(result1,0.05)" in cell C1 and "=PERCENTILE(result1,0.95)" in cell C2.

StatCrunch

1) We will use two columns for the two box model. Open StatCrunch and copy/paste the Vendor data from the Excel workbook (download <here>) or load the Excel workbook directly into StatCrunch using "Data> Load Data> from file" and select "My computer". Browse and select the Excel workbook downloaded above, making certain to uncheck the "Use first line as column names" (the first line contains data!). Click "Load File" at the bottom of the dialog.

2) Rename the first column "A" and the second column "B" (figure 7-42).



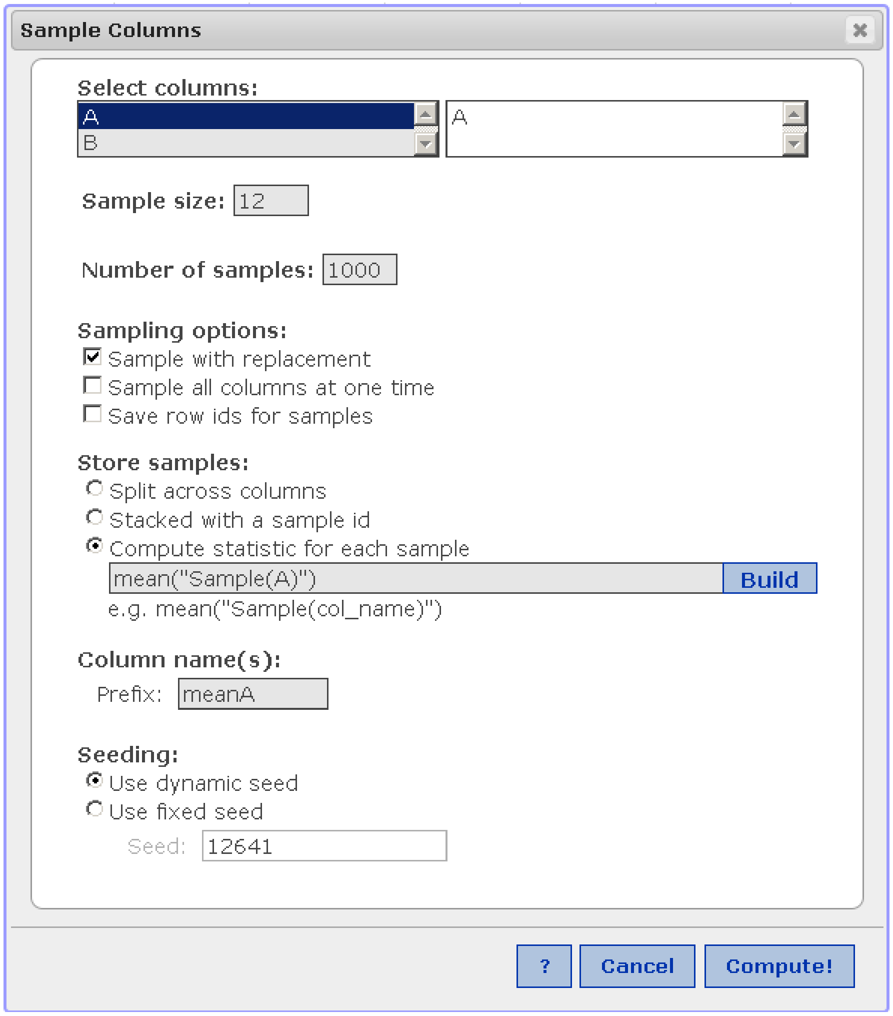
Row	A	B
1	14.2	12.4
2	15.1	13.6
3	13.9	14.6
4	12.8	12.8
5	13.7	11.9
6	14	13.1
7	13.5	12.8
8	14.3	13.2
9	14.9	14.7
10	13.1	14.5
11	13.4	
12	13.2	
13		

Figure 7-42: Two box model

3) For vendor A, select Data> Sample. Select column A, enter 12 as the sample size, check "Resample with Replacement", and enter 1000 as the number of trials. The Statistic expression is:

   mean("Sample(A)")

   Name the new column “meanA” and click Compute! (figure 7-43).



Select columns: A, B (selected: A)

Sample size: 12

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
mean("Sample(A)")
e.g. mean("Sample(col_name)")

Column name(s):
Prefix: meanA

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 7-43: Sample columns dialog two-box model (Vendor A)

4) For vendor B, repeat the above, selecting column B, 10 as the sample size, and Statistic expression:

   mean("Sample(B)")

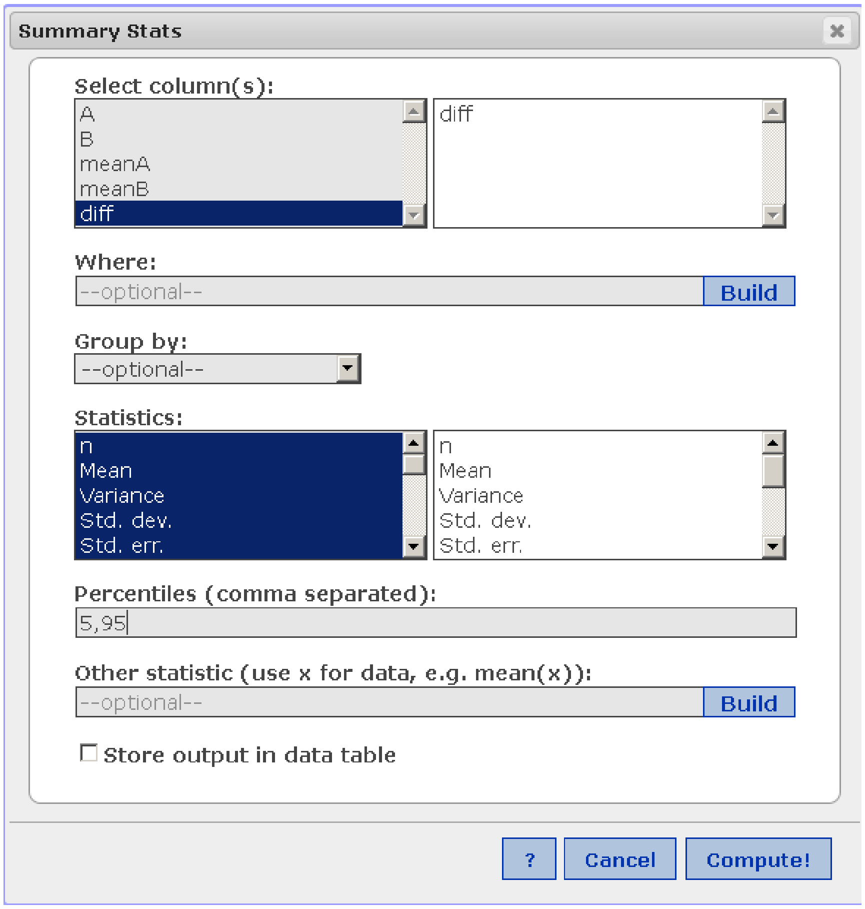
   Rename the new column "meanB".

5) Find the difference in the means by selecting "Data>Compute Expression" and entering:

   meanA - meanB

   as the expression and "diff" as the new column name. Click Compute!

6) Find the 90% confidence interval by selecting "Stat>Summary Stats>Columns". Select the "diff" column and enter 5, 95 in the Percentiles box (figure 7-44).

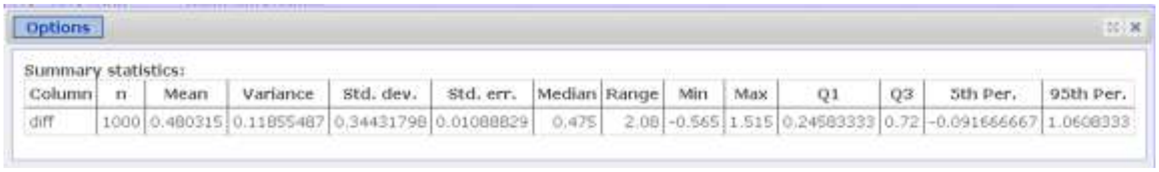


Percentiles (comma separated): 5,95

☐ Store output in data table

Figure 7-44: Calculate the 90% confidence interval by specifying percentile range

7) The output of one complete 1000 trial simulation is shown below in figure 7-45. Note that the 90% confidence interval contains 0, which means that we can't rule out random factors or chance as a cause for the difference in means.



Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3	5th Per.	95th Per.
diff	1000	0.480315	0.11855487	0.34431798	0.01088829	0.475	2.08	-0.565	1.515	0.24583333	0.72	-0.091666667	1.0608333

Figure 7-45: Summary statistics with 90% confidence interval

Box Sampler

We will use two-boxes and enter the data as shown in the table. To create a two-box Box Sampler workbook.

1) In the startup dialog, select "#Populations: Two"

2) Enter the data from the table into the two boxes. Vendor A in box 1 and Vendor B in box 2.

3) In cell R11 (Sample Statistics cell) enter the formula =AVERAGE(Sample1)-AVERAGE(Sample2) to calculate the difference in means

4) In cell P1, enter the formula =PERCENTILE(Stat1, 0.05)

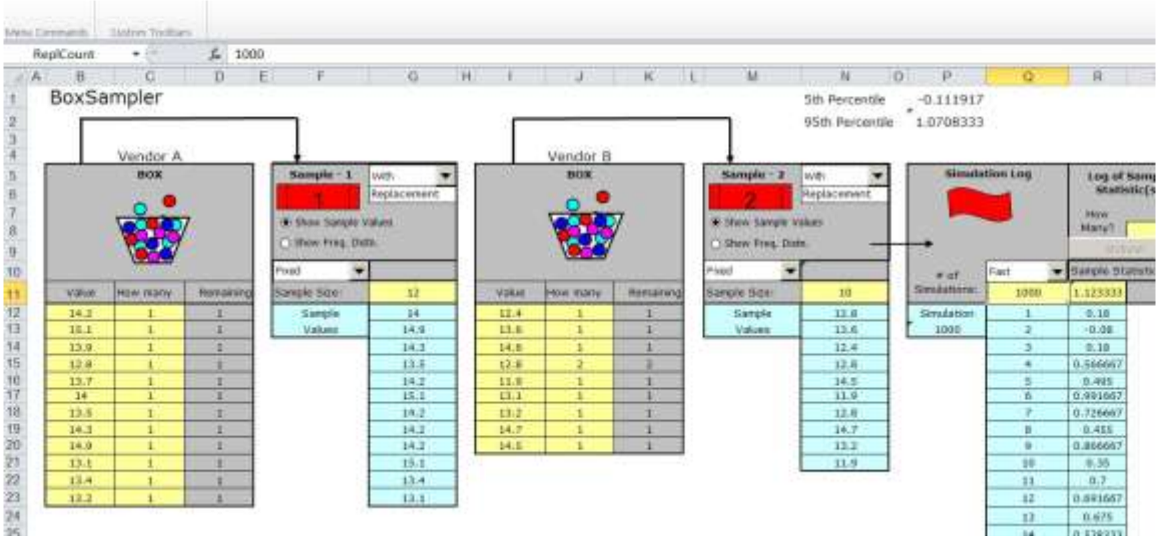
5) In cell P2, enter the formula =PERCENTILE(Stat1, 0.95)

6) Run the model and note the confidence interval expressed by the percentile functions in cells P1 and P2.

   A macro-enabled workbook illustrating this Two-Box simulation can be downloaded <here>

   Note: The number of simulations in the macro-enabled workbook is 1000. If you have a slower computer, you may want to use the "Superfast" setting (above the value for the number of simulations) or enter a smaller number of simulations.

   An example workbook is shown below in figure 7-46:



RepICount 1000

BoxSampler

5th Percentile -0.111917

95th Percentile 1.0708333

Sample Size: 12

Sample Size: 10

# of Simulations: 1000

Sample Statistic: 1.123333

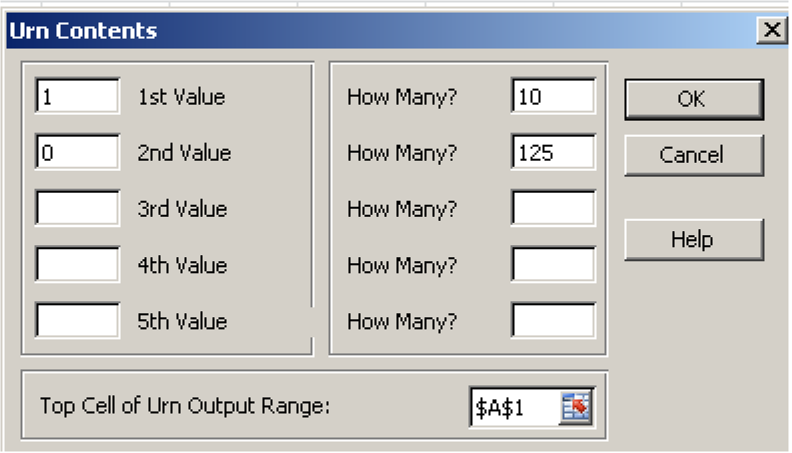
Figure 7-46: Box Sampler output

Notice the two boxes, one for Vendor A and one for Vendor B. The 90% confidence interval is calculated at the top of the worksheet in cells P1 and P2. The difference in sample means of 0.48 is well within the 90% confidence interval calculated in this example simulation.

7.8 Confidence Interval for a Difference in Proportions

Resampling Procedure for Cholesterol and MI

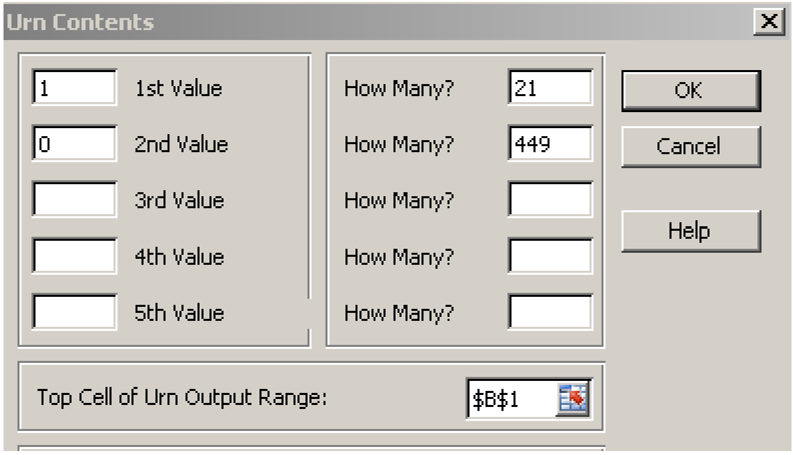
1) Use column A for the “high” box and column B for the “low” box. Select the Urn feature and enter the following for the high box (figure 7-47). Click OK.



1st Value: 1 — How Many? 10
2nd Value: 0 — How Many? 125
Top Cell of Urn Output Range: $A$1

Figure 7-47: High box urn

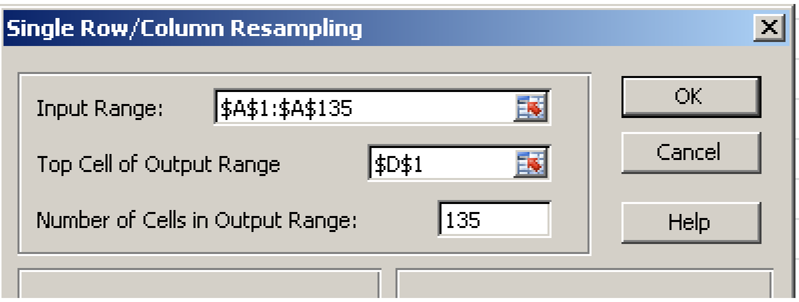
2) Select the Urn feature again to populate the low box as shown in figure 7-48. Click OK.



1st Value: 1 — How Many? 21
2nd Value: 0 — How Many? 449
Top Cell of Urn Output Range: $B$1

Figure 7-48: Low box urn

3) Select Resample and fill in the dialog for the high cholesterol resample as shown in figure 7-49. Click OK.



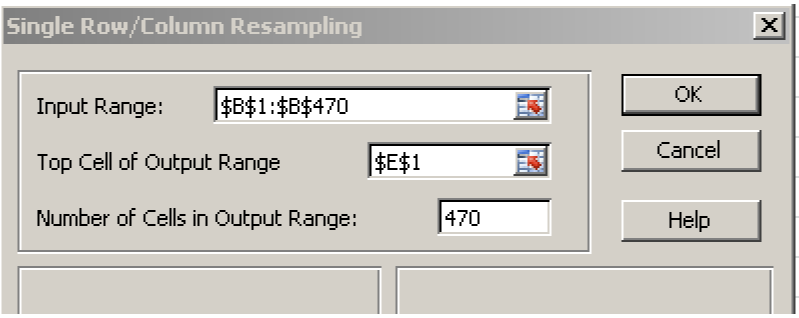
Input Range: $A$1:$A$135

Top Cell of Output Range $D$1

Number of Cells in Output Range: 135

Figure 7-49: Resample from the high box/urn

4) Repeat step 3 for the low box/urn. Fill in the Resample dialog as shown in figure 7-50. Click OK.



Input Range: $B$1:$B$470

Top Cell of Output Range $E$1

Number of Cells in Output Range: 470

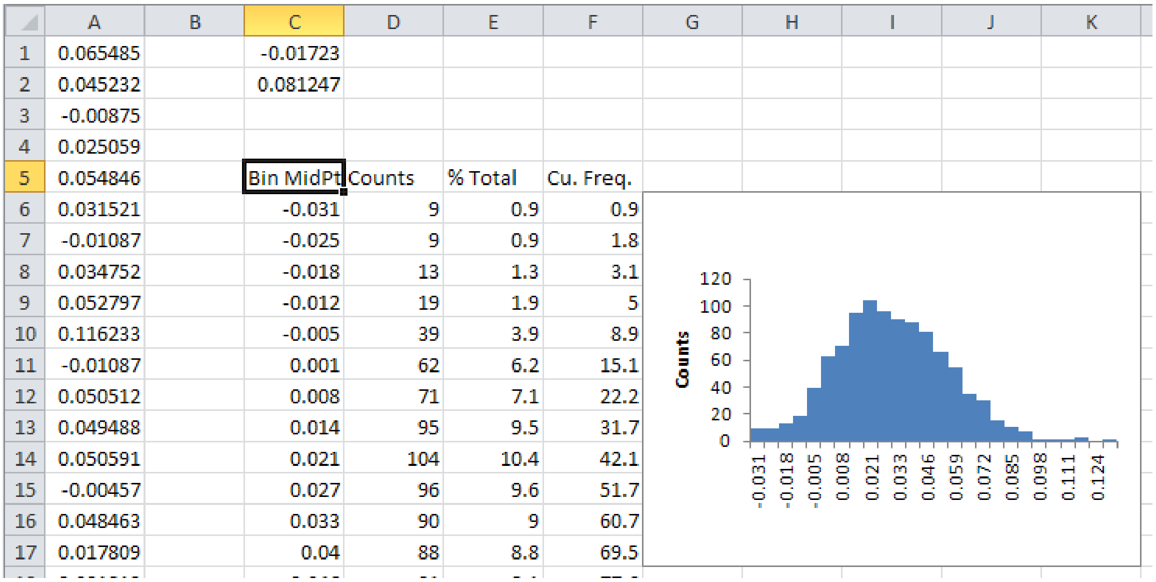
Figure 7-50: Resample from the low box/urn

5) In cell G1 enter the formula =SUM(D1:D135)/135 and in cell G2, enter the formula =SUM(E1:E470)/470. These formulas calculate the proportions of 1's in each box/urn. In cell G3, enter =G1-G2 to calculate the difference in the proportions.

6) Select cell G3 and choose Repeat and Score. Cell G3, the difference in proportions, is the statistic of interest in this model and is the score cell. Enter 1000 trials and click OK.

7) On the Results worksheet, enter =PERCENTILE(result1,0.025) in cell C1 and =PERCENTILE(result1, 0.975) in cell C2 to calculate the 95% confidence interval.

8) Create a histogram by selecting the top cell of the results (A1) and choosing the Resampling Stats histogram feature. Select cell C5 as the Top Left Cell for the Freq. Table. Click Draw. The resulting histogram for one simulation is shown in figure 7-51.



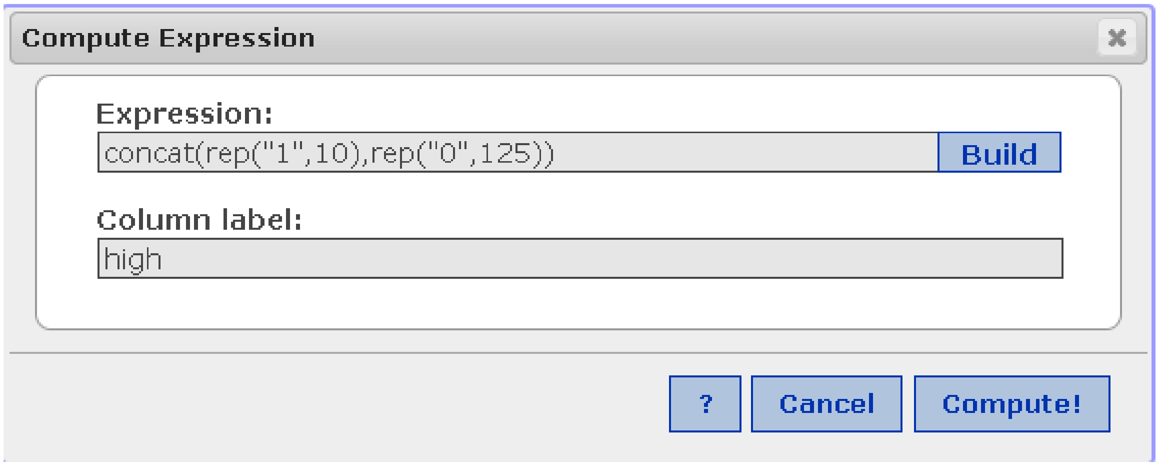
	A	B	C	D	E	F	G	H	I	J	K
1	0.065485		-0.01723								
2	0.045232		0.081247								
3	-0.00875										
4	0.025059										
5	0.054846		Bin MidPt	Counts	% Total	Cu. Freq.					
6	0.031521		-0.031	9	0.9	0.9					
7	-0.01087		-0.025	9	0.9	1.8					
8	0.034752		-0.018	13	1.3	3.1					
9	0.052797		-0.012	19	1.9	5					
10	0.116233		-0.005	39	3.9	8.9					
11	-0.01087		0.001	62	6.2	15.1					
12	0.050512		0.008	71	7.1	22.2					
13	0.049488		0.014	95	9.5	31.7					
14	0.050591		0.021	104	10.4	42.1					
15	-0.00457		0.027	96	9.6	51.7					
16	0.048463		0.033	90	9	60.7					
17	0.017809		0.04	88	8.8	69.5					

Figure 7-51: 95% confidence interval (cells C1 and C2) and histogram

StatCrunch Procedure

1) Open a StatCrunch worksheet and select "Data>Compute Expression"

2) In the "Expression" box, enter: concat(rep("1",10),rep("0",125)) and enter "high" for the New column name as shown in figure 7-52.



Expression: concat(rep("1",10),rep("0",125))

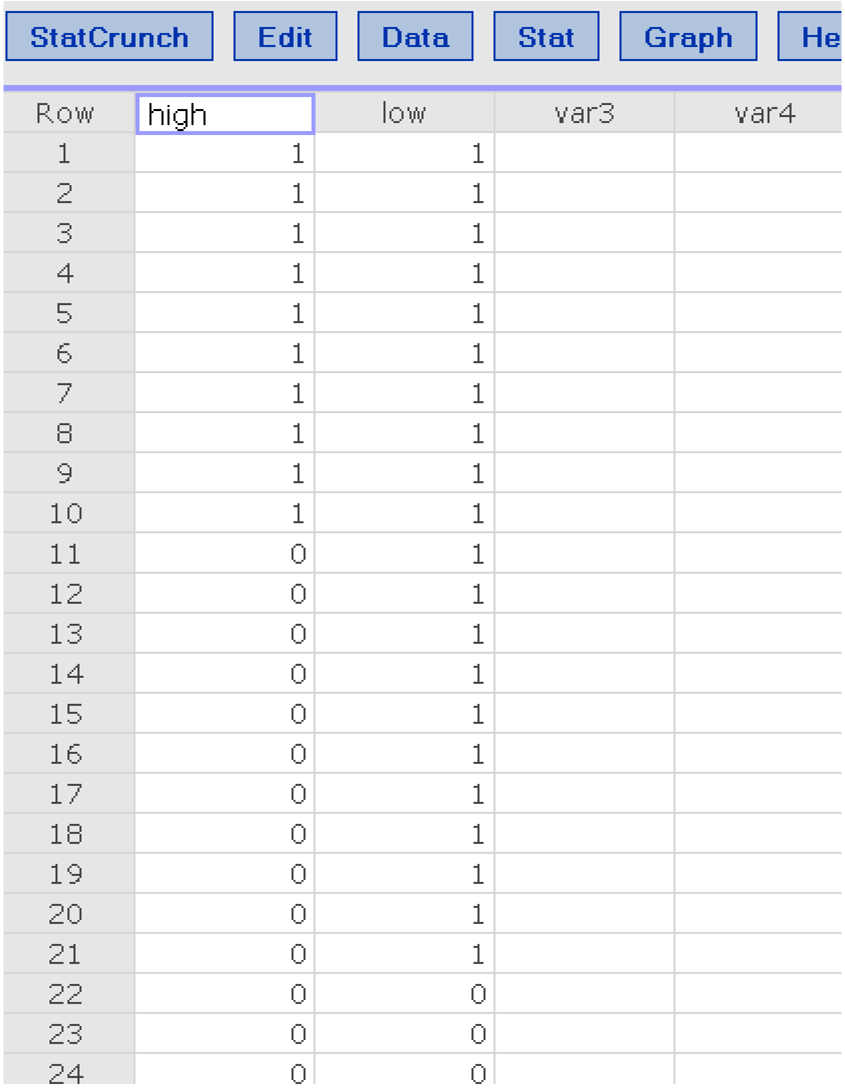
Column label: high

Figure 7-52: Compute expression

3) Click Compute!

4) Repeat the procedure above using the expression: concat(rep("1",21),rep("0",449)) and "low" for the New column name. Click Compute!

5) You should see the following data in StatCrunch



Row	high	low	var3	var4
1	1	1		
2	1	1		
3	1	1		
4	1	1		
5	1	1		
6	1	1		
7	1	1		
8	1	1		
9	1	1		
10	1	1		
11	0	1		
12	0	1		
13	0	1		
14	0	1		
15	0	1		
16	0	1		
17	0	1		
18	0	1		
19	0	1		
20	0	1		
21	0	1		
22	0	0		
23	0	0		
24	0	0		

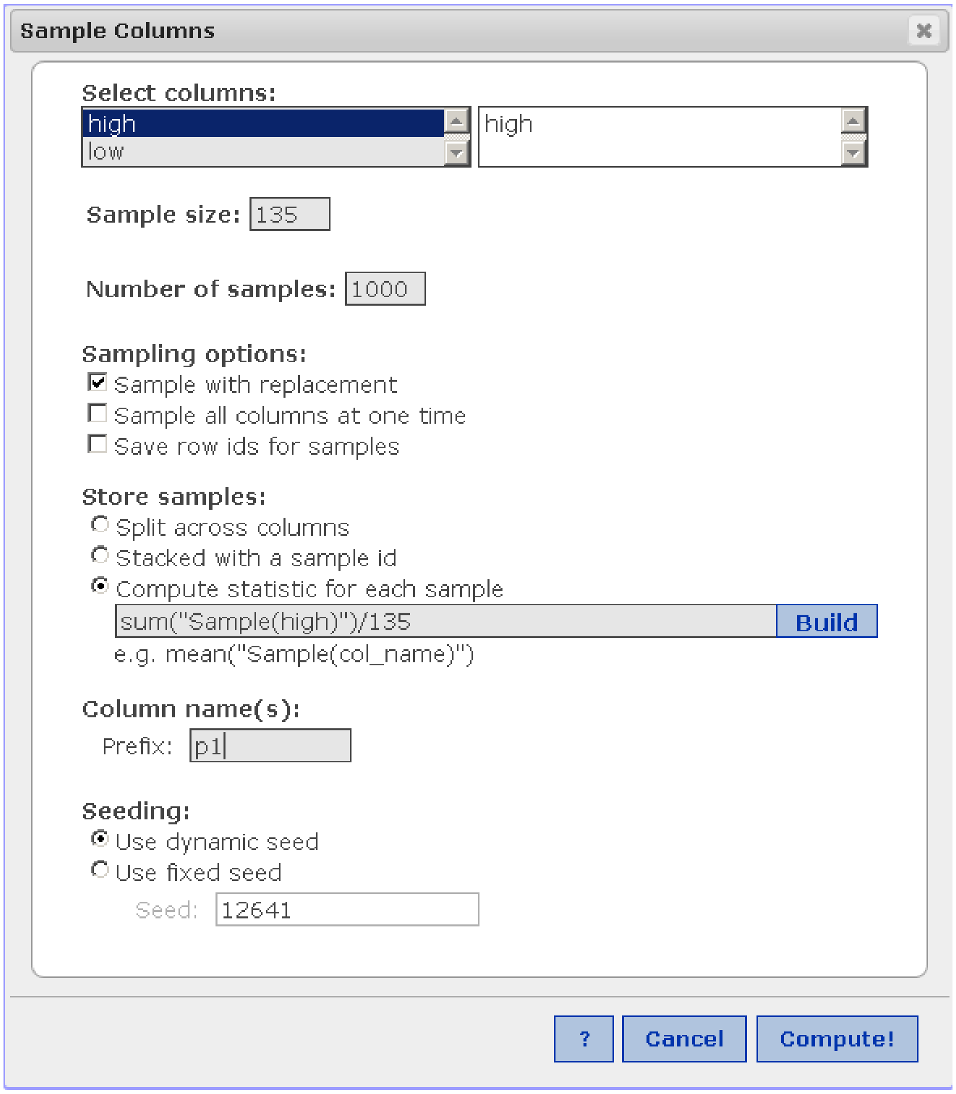
Figure 7-53: High/low box setup

6) Select Data>Sample

7) Select high as the column to sample. Enter 135 as the Sample size, 1000 as the number of samples, and check the Sample with replacement option.

8) Select Compute statistic below for each sample and enter: sum("Sample(high)")/135 to calculate the proportion of 1s for each resample.

9) Name the new column "p1" and click Compute! (figure 7-54).



Select columns: high (selected), low; high

Sample size: 135

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
sum("Sample(high)")/135
e.g. mean("Sample(col_name)")

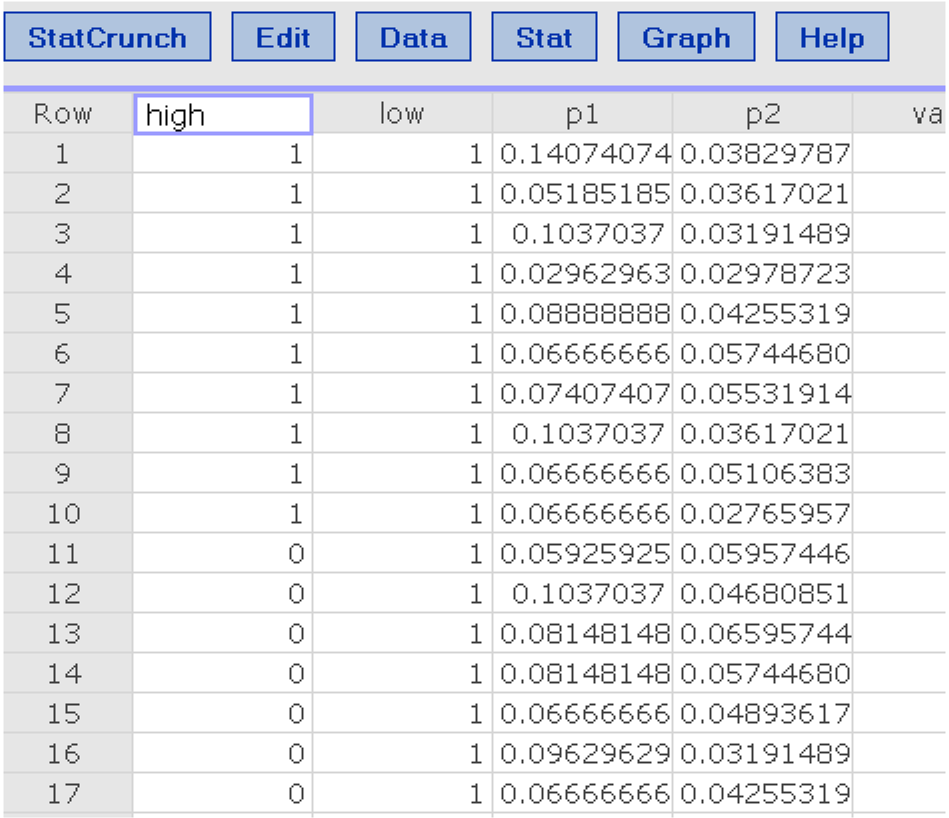
Column name(s):
Prefix: p1

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 7-54: Sample columns

10) Repeat the above procedure and this time select low as the column and 470 for the Sample Size. Check Sample with replacement and the Statistic formula is: sum("Sample(low)")/470

11) Name the new column "p2" and click Compute! (see sample result in figure 7-55):



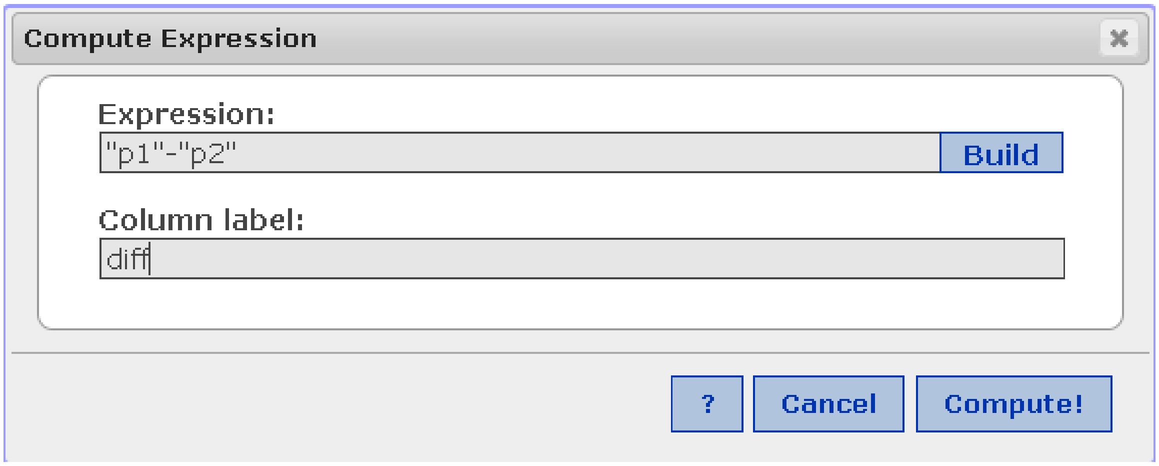
Row	high	low	p1	p2	va
1	1	1	0.14074074	0.03829787	
2	1	1	0.05185185	0.03617021	
3	1	1	0.1037037	0.03191489	
4	1	1	0.02962963	0.02978723	
5	1	1	0.08888888	0.04255319	
6	1	1	0.06666666	0.05744680	
7	1	1	0.07407407	0.05531914	
8	1	1	0.1037037	0.03617021	
9	1	1	0.06666666	0.05106383	
10	1	1	0.06666666	0.02765957	
11	0	1	0.05925925	0.05957446	
12	0	1	0.1037037	0.04680851	
13	0	1	0.08148148	0.06595744	
14	0	1	0.08148148	0.05744680	
15	0	1	0.06666666	0.04893617	
16	0	1	0.09629629	0.03191489	
17	0	1	0.06666666	0.04255319	

Figure 7-55: Sample proportions output

12) To calculate the difference between the resampled proportions, click "Data>Compute Expression"

13) Enter "p1"-"p2" in the Expression box (you MUST include the quotes!).

14) Enter "diff" as the New column name (see below) and click Compute! Your output should be similar to figure 7-56.

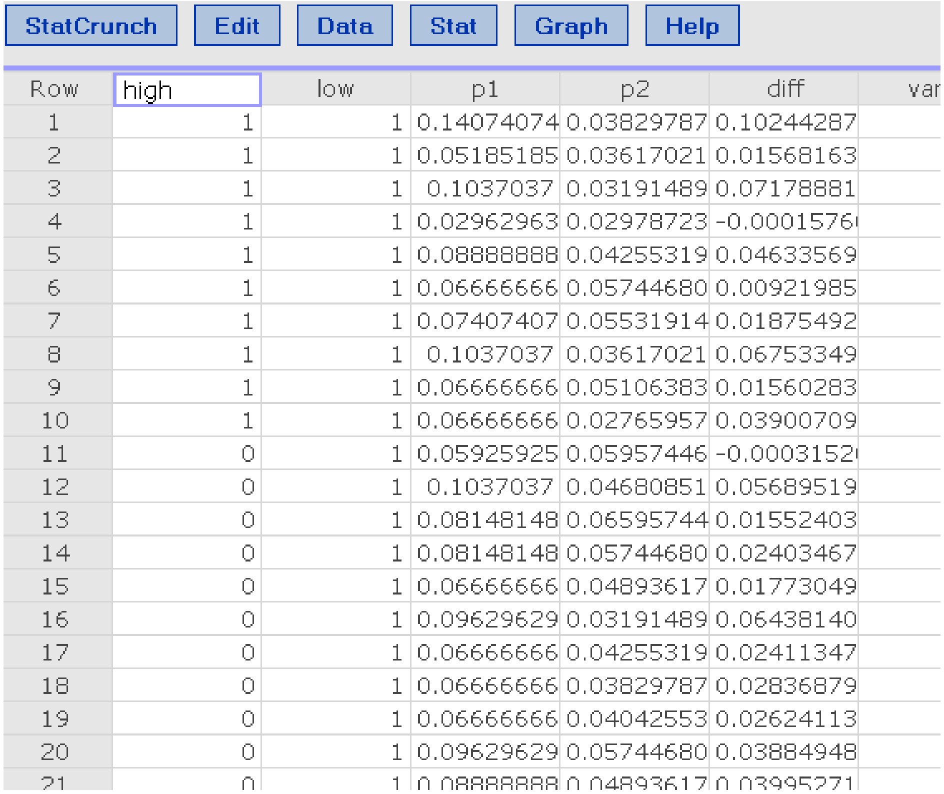


Expression: "p1"-"p2"

Column label: diff

Figure 7-56: Calculating the difference in proportions in columns p1 and p2

15) The results will be similar to figure 7-57:

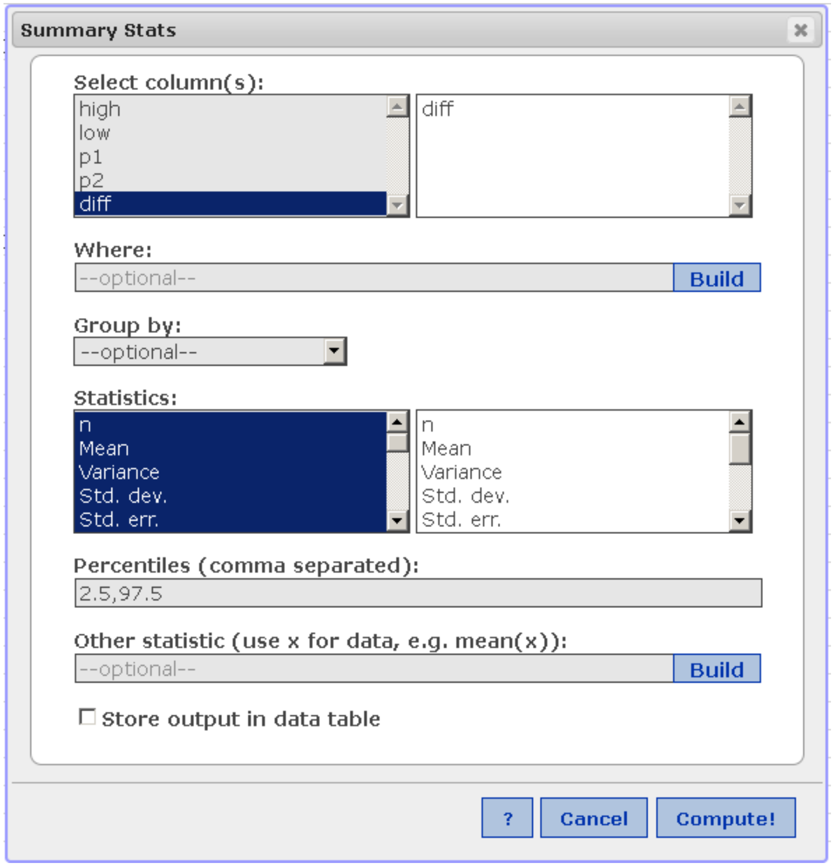


Row	high	low	p1	p2	diff	var
1	1	1	0.14074074	0.03829787	0.10244287	
2	1	1	0.05185185	0.03617021	0.01568163	
3	1	1	0.1037037	0.03191489	0.07178881	
4	1	1	0.02962963	0.02978723	-0.0001576	
5	1	1	0.08888888	0.04255319	0.04633569	
6	1	1	0.06666666	0.05744680	0.00921985	
7	1	1	0.07407407	0.05531914	0.01875492	
8	1	1	0.1037037	0.03617021	0.06753349	
9	1	1	0.06666666	0.05106383	0.01560283	
10	1	1	0.06666666	0.02765957	0.03900709	
11	0	1	0.05925925	0.05957446	-0.0003152	
12	0	1	0.1037037	0.04680851	0.05689519	
13	0	1	0.08148148	0.06595744	0.01552403	
14	0	1	0.08148148	0.05744680	0.02403467	
15	0	1	0.06666666	0.04893617	0.01773049	
16	0	1	0.09629629	0.03191489	0.06438140	
17	0	1	0.06666666	0.04255319	0.02411347	
18	0	1	0.06666666	0.03829787	0.02836879	
19	0	1	0.06666666	0.04042553	0.02624113	
20	0	1	0.09629629	0.05744680	0.03884948	
21	0	1	0.08888888	0.04893617	0.03995271	

Figure 7-57: Difference in proportions in diff column

16) To calculate the 95% confidence interval, click "Stat>Summary Stats>Columns"

17) Select the diff column, enter 2.5, 97.5 in the Percentiles edit box (as shown below) and click Compute!



Select column(s): diff

Where: --optional--

Group by: --optional--

Statistics: n, Mean, Variance, Std. dev., Std. err.

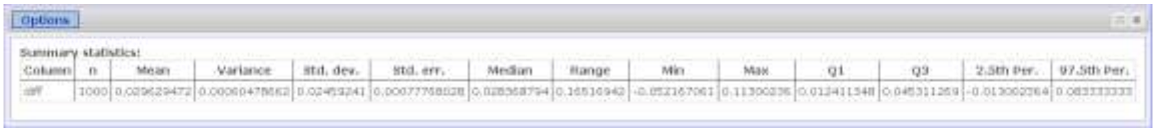
Percentiles (comma separated): 2.5,97.5

Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 7-58: Summary column statistics including 95% confidence interval (using the 2.5 and 97.5 percentiles)

18) The results are shown in figure 7-59. The 95% confidence interval is from (2.5th Percentile) -0.013002364 to (97.5th Percentile) 0.083333333.



Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3	2.5th Per.	97.5th Per.
diff	1000	0.029629472	0.00060478662	0.02459241	0.00077768028	0.028368794	0.16516942	-0.052167061	0.11300236	0.012411348	0.045311269	-0.013002364	0.083333333

Figure 7-59: Summary statistics

7.9 Appendix A – The Parametric Bootstrap (OPTIONAL)

Resampling Stats

Using the Toyota data set as an example, here are the steps for Resampling Stats:

1) Find the mean and standard deviation (SD) of the data. Cell A23 in figure 7-60 contains the function =AVERAGE(A2:A21) and cell A24 contains the function

=STDEV.S(A2:A21). In this case, the mean of the Toyota data is 17685 and the sample (‘n-1’ version) SD is 3507 as shown in figure 7-60.

2) In cell D2 enter =RSXLNormal(17685, 3507) and copy this formula down the worksheet to cell D21. This Resampling Stats function will generate a pseudo-random number (using its RNG) from a Normal distribution having a mean of 17685 and a SD of 3507. See figure 7-60.

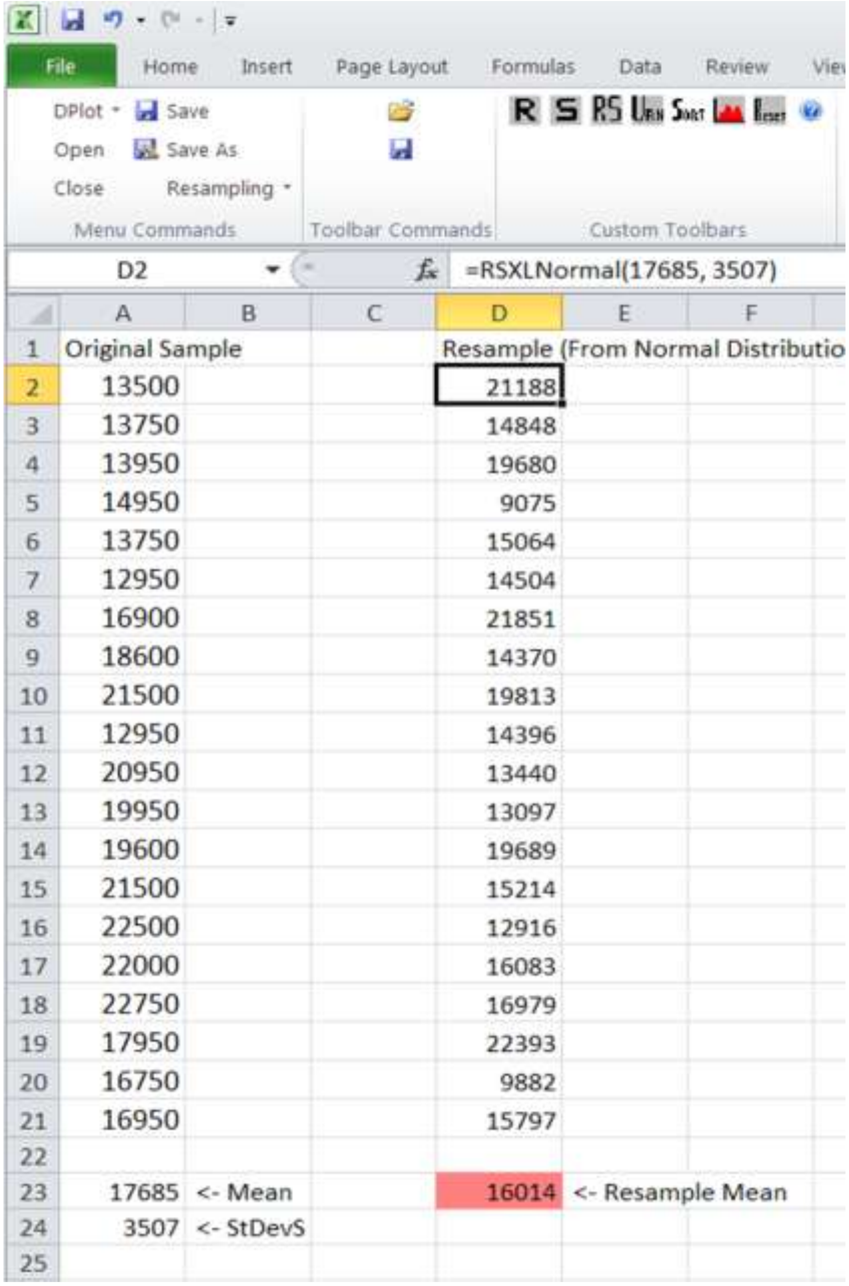
3) Enter =AVERAGE(D2:D21) in cell D23.

4) Select cell D23 and immediately choose Repeat and Score. When generating random numbers, you do NOT use Resample prior to Repeat and Score. Each time the worksheet updates, the random numbers will update, so the correct procedure is to choose Repeat and Score as the first step.

5) Enter 1000 for the number of trials and click OK. Sample results are in figure 7-62. Note that all values have been rounded to integers.

6) On the Results sheet we can find the 5th and 95th percentiles by using the =PERCENTILE function. In cell E1, enter =PERCENTILE(result1, 0.05) and in cell E2, enter =PERCENTILE(result1, 0.95). These functions will calculate the 5th and 95th percentiles respectively. In this example, the 90% confidence interval is from 16443 to 19010. See figure 7-61.

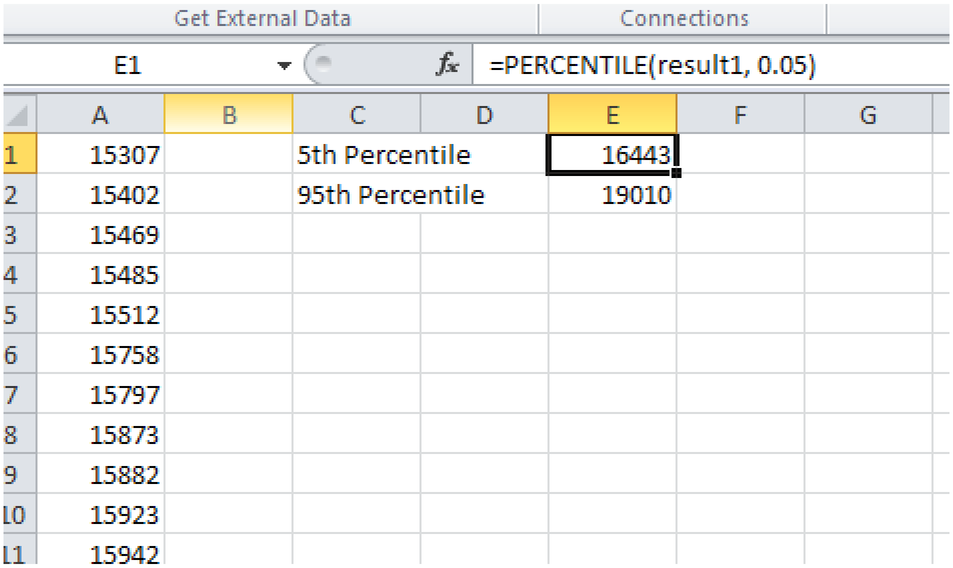
7) Select cell A1 on the Results sheet and create a histogram (figure 7-62).



D2 =RSXLNormal(17685, 3507)

	A	B	C	D	E	F
1	Original Sample			Resample (From Normal Distributio		
2	13500			21188		
3	13750			14848		
4	13950			19680		
5	14950			9075		
6	13750			15064		
7	12950			14504		
8	16900			21851		
9	18600			14370		
10	21500			19813		
11	12950			14396		
12	20950			13440		
13	19950			13097		
14	19600			19689		
15	21500			15214		
16	22500			12916		
17	22000			16083		
18	22750			16979		
19	17950			22393		
20	16750			9882		
21	16950			15797		
22						
23	17685	<- Mean		16014	<- Resample Mean	
24	3507	<- StDevS				
25						

Figure 7-60: Parametric bootstrap using Resampling Stats



E1 =PERCENTILE(result1, 0.05)

	A	B	C	D	E	F	G
1	15307		5th Percentile		16443		
2	15402		95th Percentile		19010		
3	15469						
4	15485						
5	15512						
6	15758						
7	15797						
8	15873						
9	15882						
10	15923						
11	15942						

Figure 7-61: Parametric bootstrap results

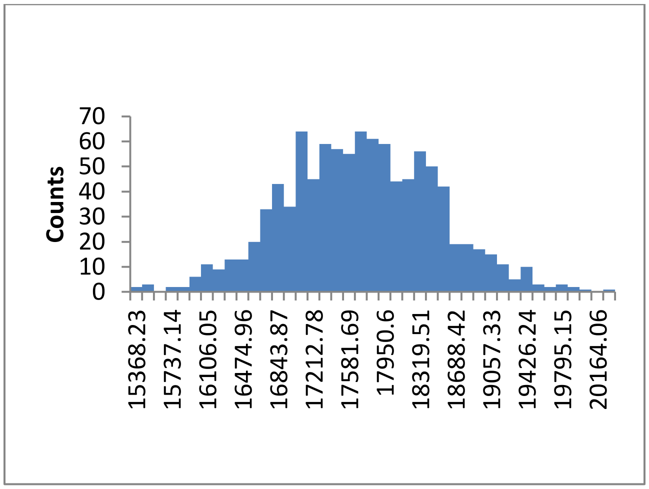


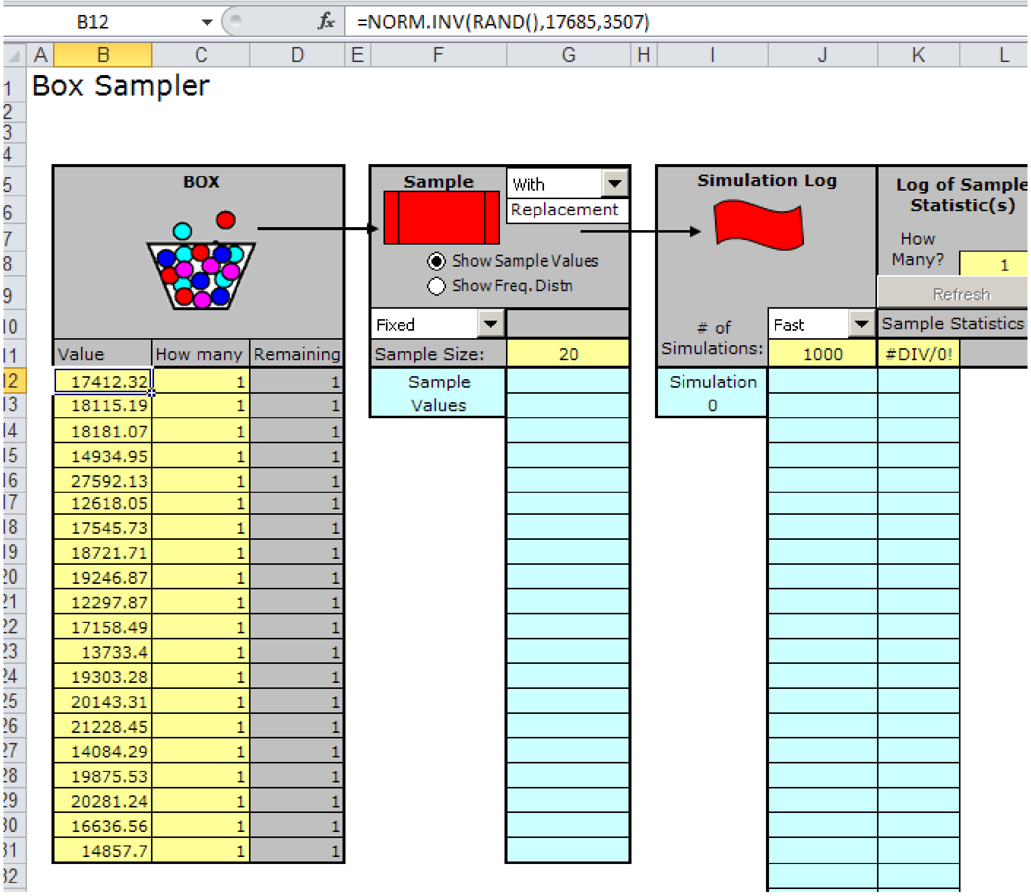
Figure 7-62: Histogram of sample means

Box Sampler

We will use the mean and SD of the Toyota data set (mean = 17685, SD = 3507)

1) Open Box Sampler and enter the following formula in the Box (cell B12): =NORM.INV(RAND(),17685,3507)

2) Copy this formula down to cell B31 for a total of 20 cells. Make certain that each corresponding “How many” box has a value of 1. Enter 20 for the Sample Size and 1000 for the # of Simulations (figure 7-63).



B12 =NORM.INV(RAND(),17685,3507)

Box Sampler

Value	How many	Remaining
17412.32	1	1
18115.19	1	1
18181.07	1	1
14934.95	1	1
27592.13	1	1
12618.05	1	1
17545.73	1	1
18721.71	1	1
19246.87	1	1
12297.87	1	1
17158.49	1	1
13733.4	1	1
19303.28	1	1
20143.31	1	1
21228.45	1	1
14084.29	1	1
19875.53	1	1
20281.24	1	1
16636.56	1	1
14857.7	1	1

Sample: With Replacement

● Show Sample Values

○ Show Freq. Distn

Fixed

Sample Size: 20

# of Simulations: 1000

Fast

Log of Sample Statistic(s) How Many? 1

Sample Statistics: #DIV/0!

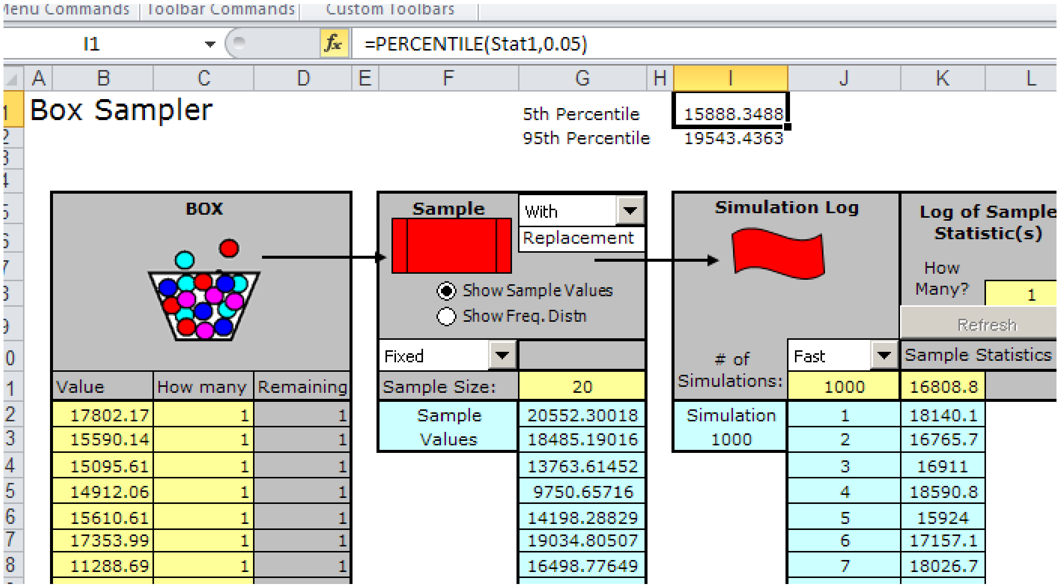
Simulation 0

Figure 7-63: Box Sampler setup for parametric bootstrap

3) In cell K11 (the Sample statistics cell) enter: =AVERAGE(Sample)

4) Run the simulation.

5) When the simulation finishes (be patient!), enter =PERCENTILE(Stat1, 0.05) in cell I1 and =PERCENTILE(Stat1, 0.95) in cell I2 to calculate the 90% confidence interval (figure 7-64). In this example, the 90% confidence interval is from 15888 to 19543 (rounded to integers).



I1 =PERCENTILE(Stat1,0.05)

Box Sampler

5th Percentile 15888.3488
95th Percentile 19543.4363

Sample: With Replacement
● Show Sample Values
○ Show Freq. Distn
Fixed
Sample Size: 20

Log of Sample Statistic(s)
How Many? 1

Fast
# of Simulations: 1000

Value	How many	Remaining
17802.17	1	1
15590.14	1	1
15095.61	1	1
14912.06	1	1
15610.61	1	1
17353.99	1	1
11288.69	1	1

Sample Values
20552.30018
18485.19016
13763.61452
9750.65716
14198.28829
19034.80507
16498.77649

Simulation 1000	Sample Statistics
1000	16808.8
1	18140.1
2	16765.7
3	16911
4	18590.8
5	15924
6	17157.1
7	18026.7

Figure 7-64: 90% CI for the parametric bootstrap

StatCrunch

Using the Toyota data mean and SD (mean = 17685, SD = 3507) we will have StatCrunch model the parametric bootstrap.

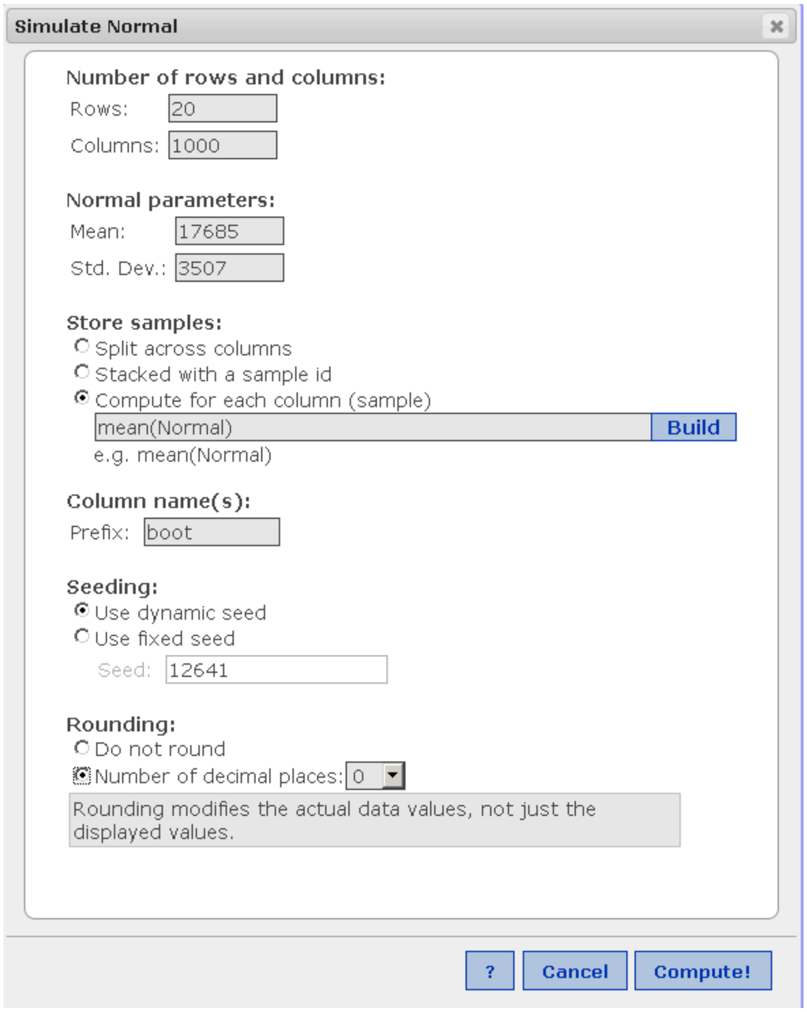
1) Open StatCrunch

2) Select Data>Simulate>Normal

3) As illustrated in figure 7-65, enter 20 for the number of rows (this is the sample size) and 1000 for the number of columns (this is the number of trials). Enter the mean (17685) and SD (3507) in the appropriate boxes. In the Store samples section, enter: mean(Normal) in order to calculate the mean of each of the 1000 samples. Use 'boot' for the column label or Prefix and use Rounding to round the values to 0 decimal places. Click Compute!

4) Calculate the 90% confidence interval by selecting Stat>Summary Stats>Columns and selecting the boot column, choose a statistic (i.e. mean) and enter 5, 95 in the Percentiles box (figure 7-66). Click Compute!

5) The results are displayed in figure 7-67.



Number of rows and columns:

Rows: 20

Columns: 1000

Normal parameters:

Mean: 17685

Std. Dev.: 3507

Store samples:

○ Split across columns

○ Stacked with a sample id

● Compute for each column (sample)

mean(Normal)

e.g. mean(Normal)

Column name(s):

Prefix: boot

Seeding:

● Use dynamic seed

○ Use fixed seed

Seed: 12641

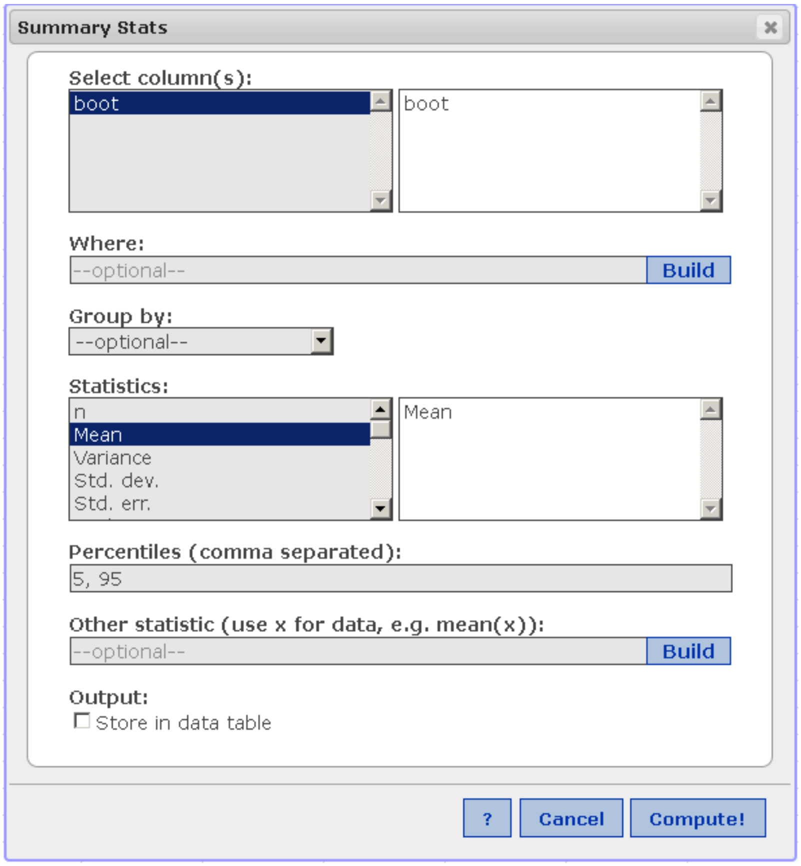
Rounding:

○ Do not round

● Number of decimal places: 0

Rounding modifies the actual data values, not just the displayed values.

Figure 7-65: Simulate a Normal distribution, mean = 17685, SD = 3507



Summary Stats

Select column(s): boot

boot

Where: --optional--

Group by: --optional--

Statistics: n, **Mean**, Variance, Std. dev., Std. err.

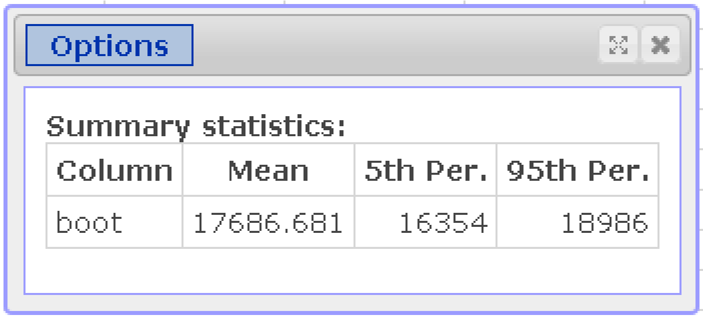
Mean

Percentiles (comma separated): 5, 95

Other statistic (use x for data, e.g. mean(x)): --optional--

Output: ☐ Store in data table

Figure 7-66: Find the 90% confidence interval



Summary statistics:

Column	Mean	5th Per.	95th Per.
boot	17686.681	16354	18986

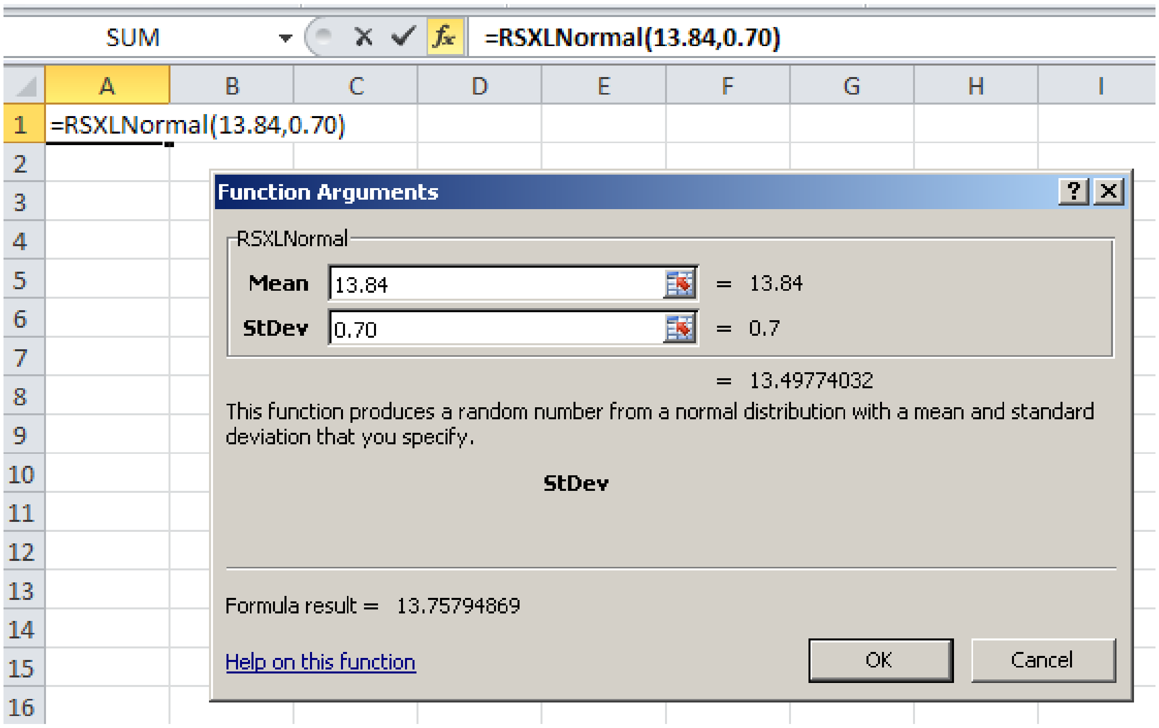
Figure 7-67: Parametric bootstrap results 90% CI

7.10 Appendix B (Optional)

Resampling Stats

1) To create a box for Vendor A, enter the Resampling Stats function =RSXLNormal(13.84, 0.70) in cell A1. This function generates a pseudo-

random number from the distribution having the specified mean and standard deviation. Copy this cell down to cell A12 (figures 7-68 and 7-69).



=RSXLNormal(13.84,0.70)

Function Arguments

RSXLNormal

Mean	13.84	= 13.84
StDev	0.70	= 0.7

= 13.49774032

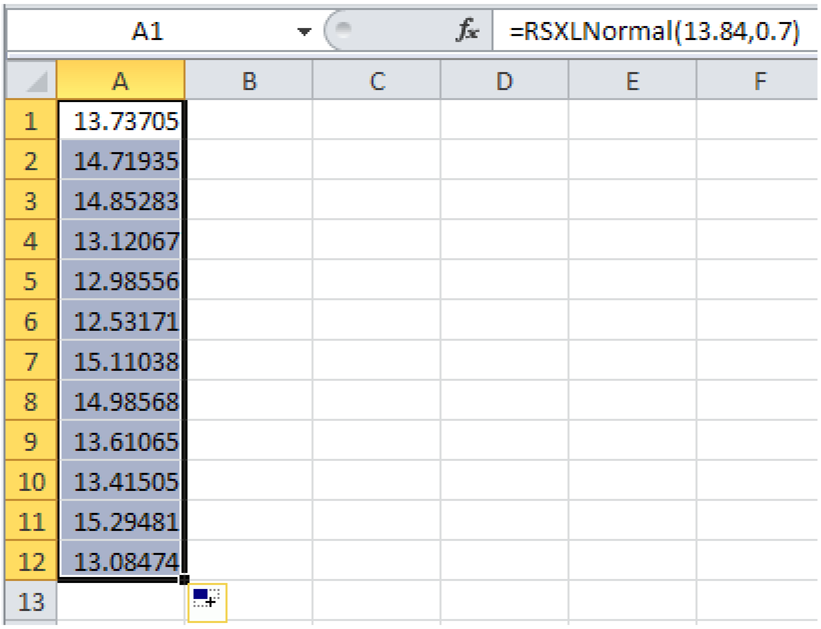
This function produces a random number from a normal distribution with a mean and standard deviation that you specify.

StDev

Formula result = 13.75794869

Help on this function

Figure 7-68: Resampling stats RSXLNormal function



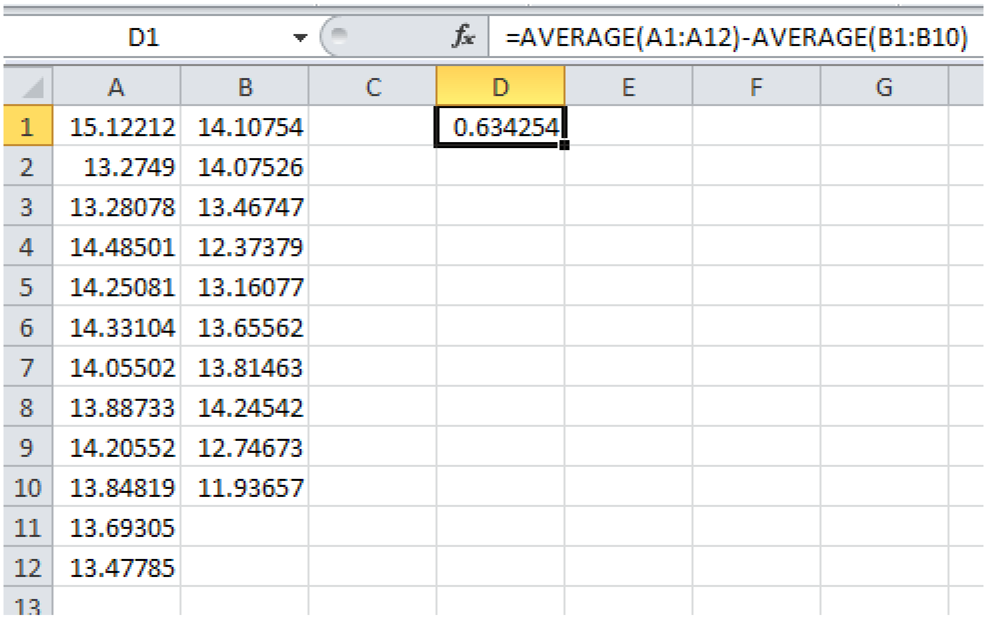
=RSXLNormal(13.84,0.7)

	A
1	13.73705
2	14.71935
3	14.85283
4	13.12067
5	12.98556
6	12.53171
7	15.11038
8	14.98568
9	13.61065
10	13.41505
11	15.29481
12	13.08474
13	

Figure 7-69: Copy cell A1 downward to cell A12

2) Repeat the process for the Vendor B box, using a mean of 13.36 and a standard deviation of 0.97. Enter =RSXLNormal(13.36, 0.97) in cell B1 and copy downward to include cell B10.

3) We do not have to use the Resample dialog prior to Repeat and Score. The =RSXLNormal() function will automatically generated a different random number each time the worksheet is updated. This function behavior is standard for both native Excel functions and user-created function procedures such as the =RSXL random number generators. All we need to do now is to find the difference in means between the two boxes. In cell D1, enter =AVERAGE(A1:A12) – AVERAGE(B1:B10) as in figure 7-70.



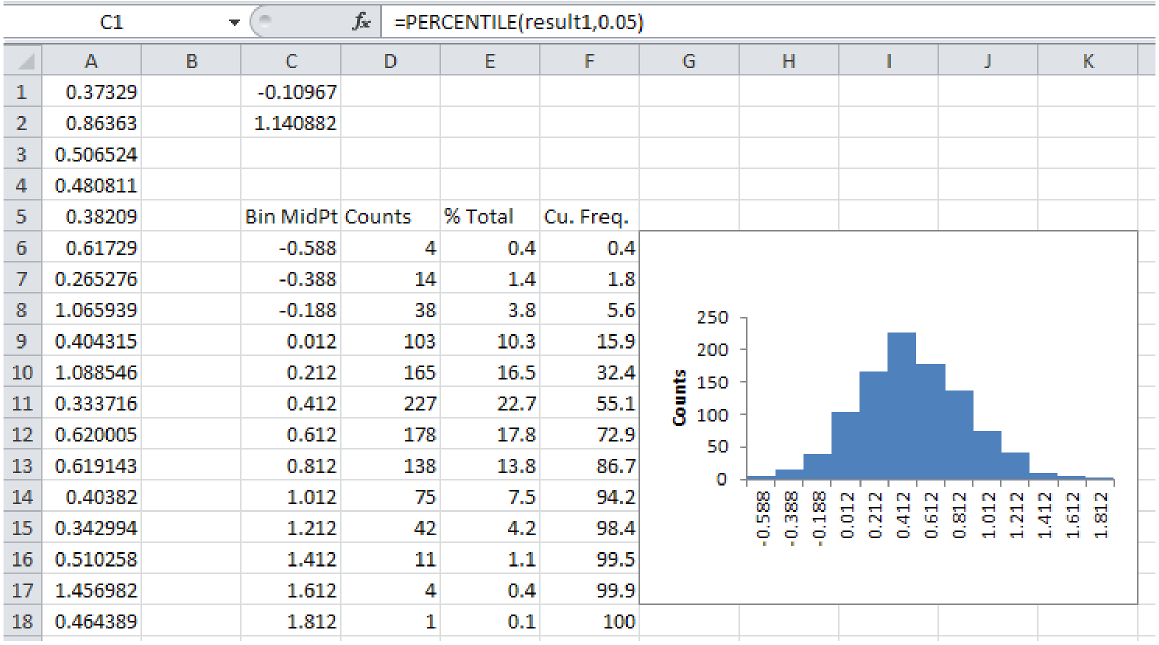
D1 =AVERAGE(A1:A12)-AVERAGE(B1:B10)

	A	B	C	D	E	F	G
1	15.12212	14.10754		0.634254			
2	13.2749	14.07526					
3	13.28078	13.46747					
4	14.48501	12.37379					
5	14.25081	13.16077					
6	14.33104	13.65562					
7	14.05502	13.81463					
8	13.88733	14.24542					
9	14.20552	12.74673					
10	13.84819	11.93657					
11	13.69305						
12	13.47785						
13							

Figure 7-70: Difference in resample means: Vendor A – Vendor B

4) Cell D1 immediately becomes our statistic of interest or score cell. Select cell D1 and choose Repeat and Score. Enter 1000 for the number of trials and click OK. Note that the values in both boxes update with each iteration.

5) Find the 90% confidence interval and create a histogram using the data in the Results sheet. A sample output is shown in figure 7-71.



C1 =PERCENTILE(result1,0.05)

	A	B	C	D	E	F	G	H	I	J	K
1	0.37329		-0.10967								
2	0.86363		1.140882								
3	0.506524										
4	0.480811										
5	0.38209		Bin MidPt	Counts	% Total	Cu. Freq.					
6	0.61729		-0.588	4	0.4	0.4					
7	0.265276		-0.388	14	1.4	1.8					
8	1.065939		-0.188	38	3.8	5.6					
9	0.404315		0.012	103	10.3	15.9					
10	1.088546		0.212	165	16.5	32.4					
11	0.333716		0.412	227	22.7	55.1					
12	0.620005		0.612	178	17.8	72.9					
13	0.619143		0.812	138	13.8	86.7					
14	0.40382		1.012	75	7.5	94.2					
15	0.342994		1.212	42	4.2	98.4					
16	0.510258		1.412	11	1.1	99.5					
17	1.456982		1.612	4	0.4	99.9					
18	0.464389		1.812	1	0.1	100					

Figure 7-71: 90% confidence interval and histogram

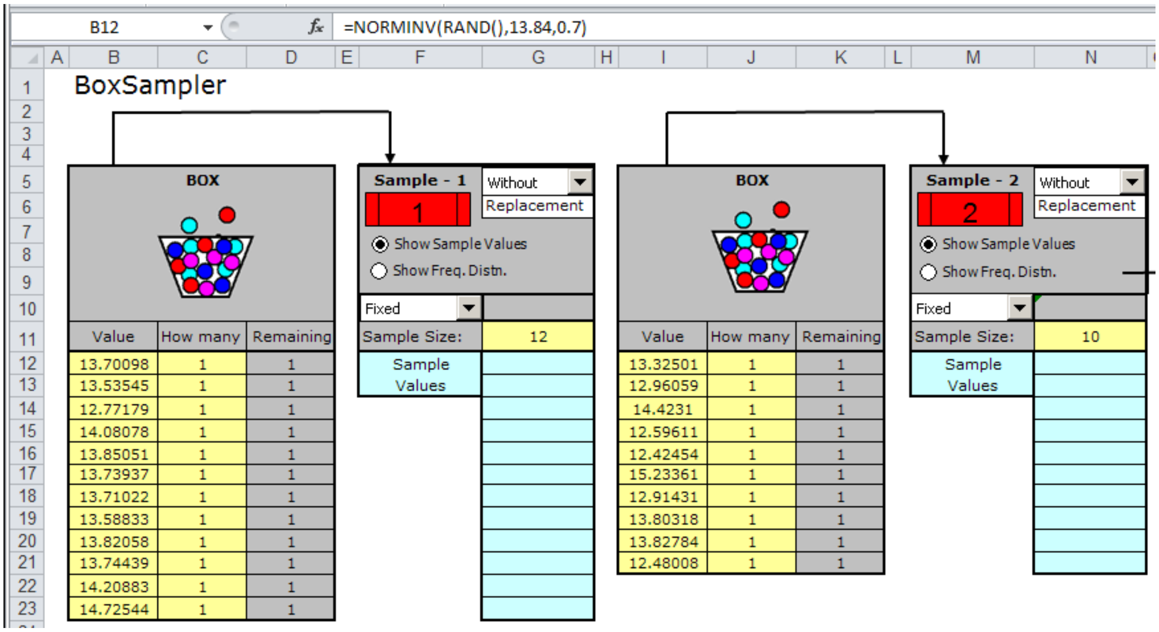
6) Cell C1 and C2 contain the =PERCENTILE(result1, 0.05) and =PERCENTILE(result1, 0.95) functions respectively and display the 90% confidence interval. The histogram was created by using the Resampling Stats histogram feature as we have done in previous procedures.

Box Sampler

1) We need a Two-Box model for Box Sampler. For the first Box, enter the native Excel formula =NORMINV(RAND(), 13.84, 0.70) in the Value cell (cell B12) and for Box 2, enter the formula =NORMINV(RAND(), 13.36, 0.97) in Value cell I12. Both of these formulas generate specific pseudo-random values from Normal distributions having means of 13.84 and 13.36 respectively and standard deviations of 0.7 and 0.97 respectively. We need sample sizes of 12 for Box 1 and 10 for Box 2, so copy cell B12 downward to cell B23 and copy cell I12 downward to cell I21. Enter 12 for the Sample Size of Box 1 and 10 for the Sample Size of Box 2. In order to replicate drawing 12 and 10 numbers from the box, without the values in the box changing as we draw each value, we will draw each new sample WITHOUT replacement. Figure 7-72 illustrates the initial setup.

<u>Note</u>: The Box Sampler setup is a bit different than the Resampling Stats setup. In Resampling Stats, all pseudo-random numbers are generated or drawn from the box before the means are calculated. In Box Sampler, as each value is drawn from the box and placed in the Sample area, a worksheet update is executed, which changes the values in the boxes prior to the calculations of the means. We do not want to erase and rewrite new numbers on the slips we've already drawn! So, we need to

resample WITHOUT replacement in Box Sampler using this model. This may seem like it should not make a difference in the final confidence interval. Try the simulation both with and without replacement to see what happens to the upper and lower bounds of the 90% confidence interval.



B12 =NORMINV(RAND(),13.84,0.7)

Sample - 1: Without Replacement; ● Show Sample Values; ○ Show Freq. Distn.; Fixed; Sample Size: 12

Sample - 2: Without Replacement; ● Show Sample Values; ○ Show Freq. Distn.; Fixed; Sample Size: 10

Value	How many	Remaining
13.70098	1	1
13.53545	1	1
12.77179	1	1
14.08078	1	1
13.85051	1	1
13.73937	1	1
13.71022	1	1
13.58833	1	1
13.82058	1	1
13.74439	1	1
14.20883	1	1
14.72544	1	1

Value	How many	Remaining
13.32501	1	1
12.96059	1	1
14.4231	1	1
12.59611	1	1
12.42454	1	1
15.23361	1	1
12.91431	1	1
13.80318	1	1
13.82784	1	1
12.48008	1	1

Figure 7-72: Two-Box setup without replacement

2) Enter 1000 in cell Q11 (# of simulations) and in cell R11, enter =AVERAGE(Sample1) - AVERAGE(Sample2) for the statistic of interest (the difference in the means Vendor A – Vendor B). Click Simulate.

3) When the simulation has ended, find the 90% confidence interval by using the native Excel =PERCENTILE as you did in the Resampling Stats procedure. For your convenience, Box Sampler names the statistic output range “Stat1”. You may also create a histogram using the BoxSampler>StatCharts>Histogram feature.

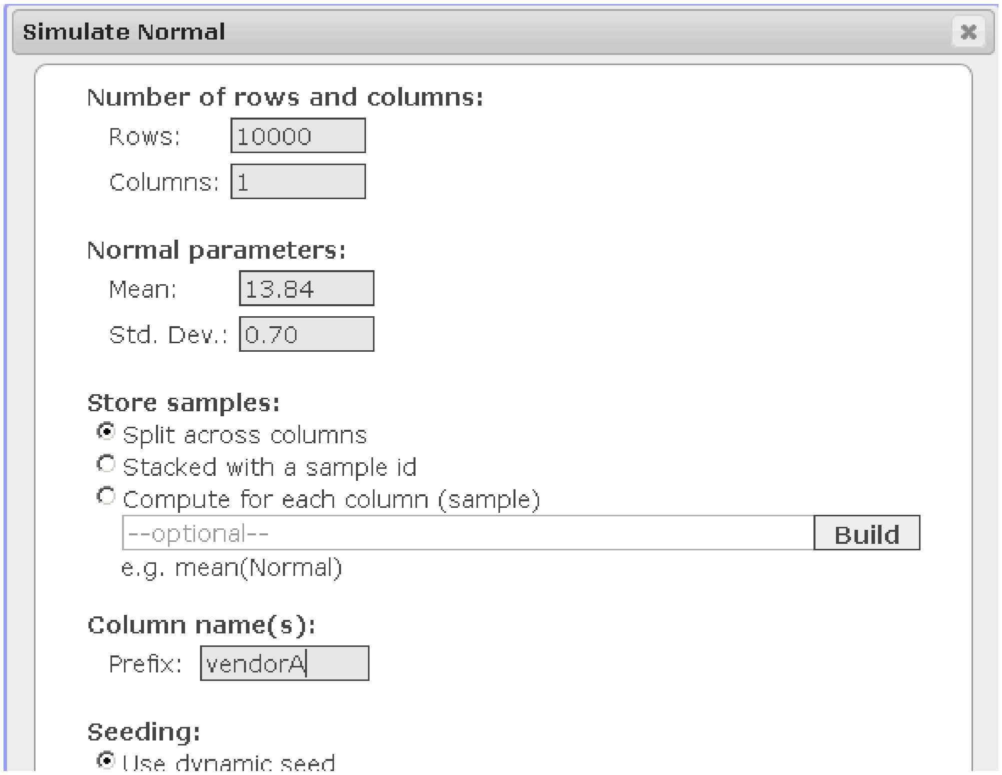
4) You can download a macro-enabled Box Sampler workbook for this simulation <here>.

StatCrunch

1) The procedure for StatCrunch will be a bit different than the procedures for Resampling Stats and Box Sampler. We will not generate pseudo-random numbers “on the fly”, but instead, we’ll generate two lists (boxes) of 10000 random numbers each. The first list (Vendor A) of 10000 numbers will be generated from a Normal distribution having a mean of 13.84 and standard

deviation of 0.70. The second list (Vendor B) will be generated from a Normal distribution having a mean of 13.36 and a standard deviation of 0.97.

2) Open StatCrunch and select Data>Simulate>Normal. Enter the values as shown in figure 7-73, keeping the default settings for all other entries. Name the column vendorA or something similar. Click Compute!



Simulate Normal

Number of rows and columns:
Rows: 10000
Columns: 1

Normal parameters:
Mean: 13.84
Std. Dev.: 0.70

Store samples:
● Split across columns
○ Stacked with a sample id
○ Compute for each column (sample)
--optional--
e.g. mean(Normal)

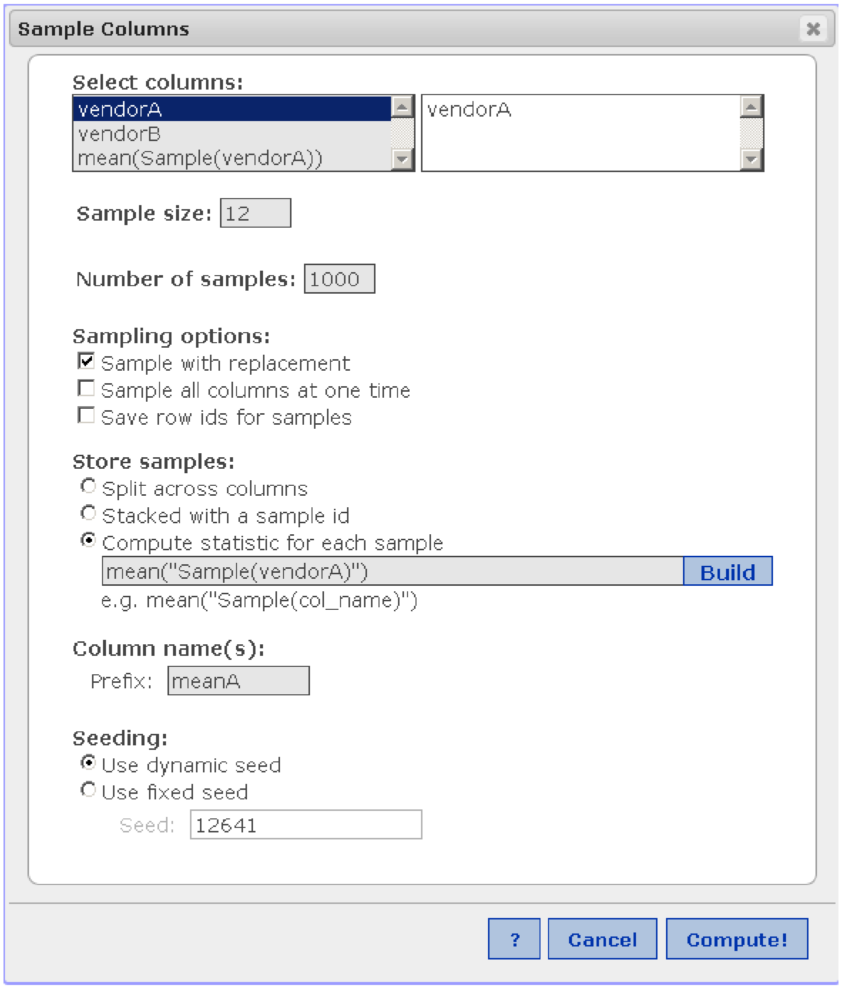
Column name(s):
Prefix: vendorA

Seeding:
● Use dynamic seed

Figure 7-73: Simulate Vendor A using Normal distribution, mean = 13.84, SD = 0.70

3) Repeat step 2 for Vendor B, using a mean of 13.36 and SD of 0.97. Rename the column vendorB or something similar.

4) Select Data>Sample, choose the vendorA column, enter a sample size of 12, check Sample with replacement, and enter 1000 as the number of samples. In the Compute for each column (sample) enter: mean(vendorA). Name the column meanA or something similar. Click Compute! (see figure 7-74).



Sample Columns

Select columns:
vendorA
vendorB
mean(Sample(vendorA))

vendorA

Sample size: 12

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

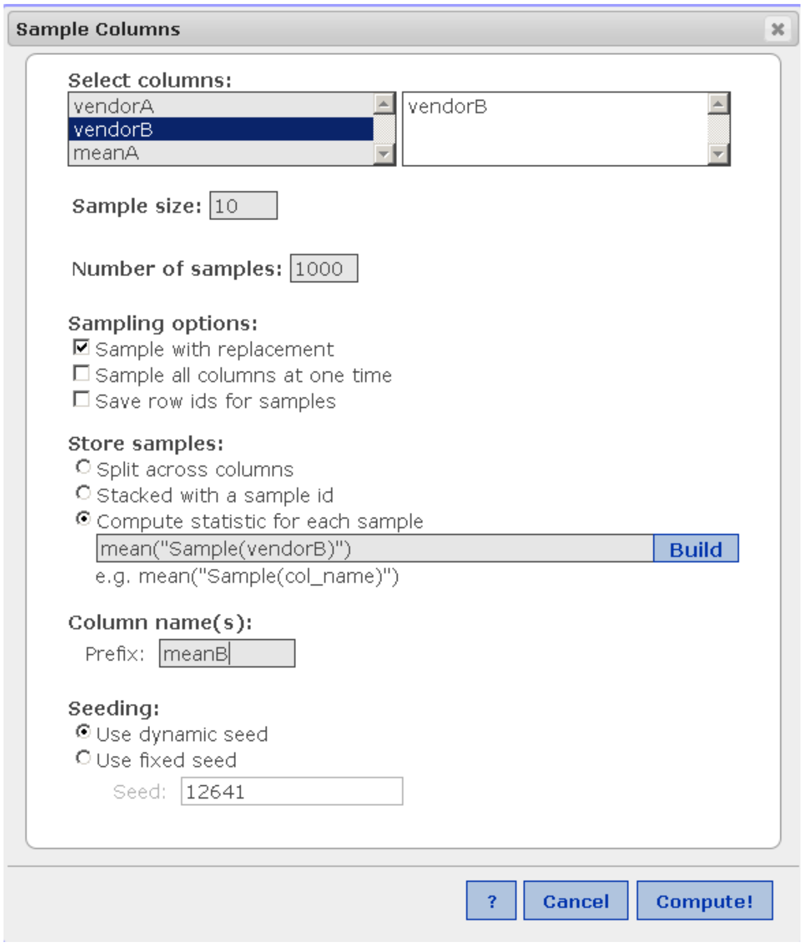
Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
mean("Sample(vendorA)")
e.g. mean("Sample(col_name)")

Column name(s):
Prefix: meanA

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 7-74: Resample for Vendor A

5) Repeat the process for Vendor B as shown in figure 7-75.



Select columns: vendorA, vendorB (selected), meanA → vendorB

Sample size: 10

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

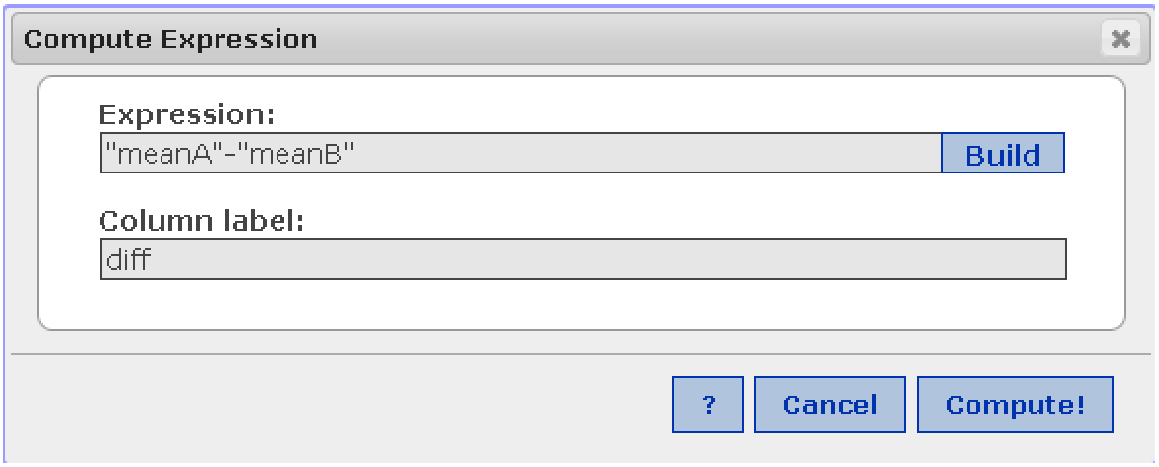
Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
mean("Sample(vendorB)")
e.g. mean("Sample(col_name)")

Column name(s):
Prefix: meanB

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 7-75: Vendor example (vendorB)

6) Select Data>Compute Expression. Enter the expression "meanA"-"meanB" (you MUST use quotes!) and a column label of diff (figure 7-76). Click Compute!

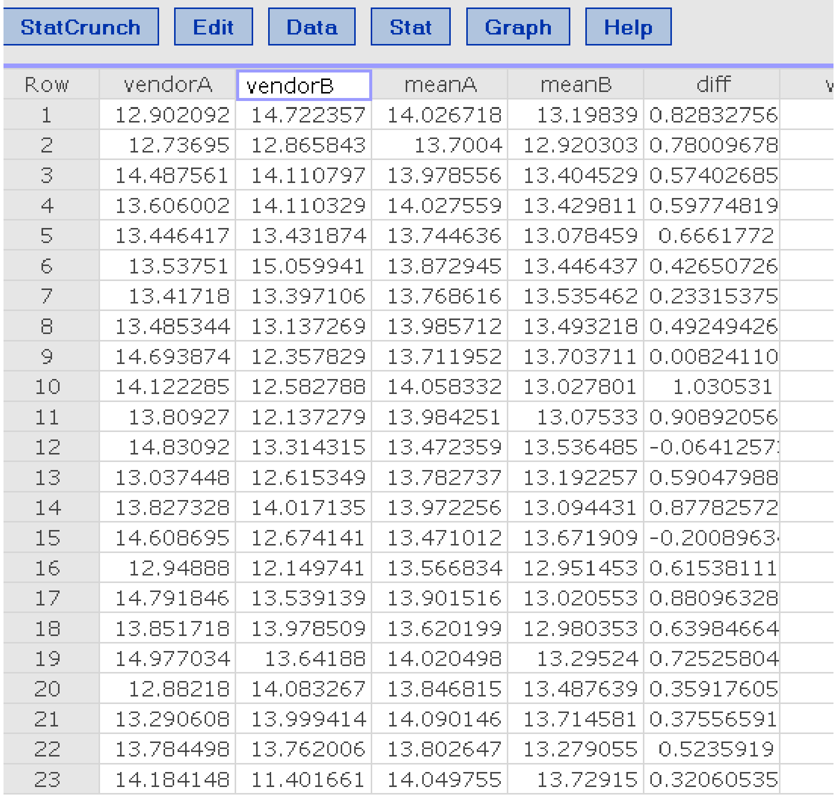


Expression: "meanA"-"meanB"

Column label: diff

Figure 7-76: Calculating the difference in resample means

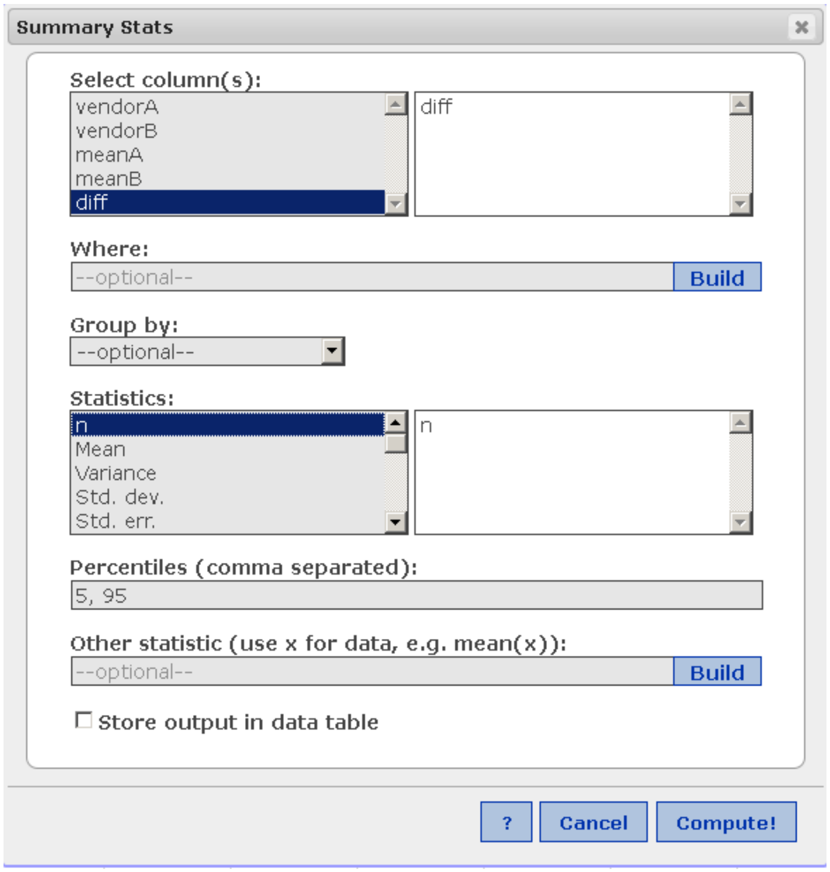
7) Your worksheet should look something like figure 7-77.



Row	vendorA	vendorB	meanA	meanB	diff
1	12.902092	14.722357	14.026718	13.19839	0.82832756
2	12.73695	12.865843	13.7004	12.920303	0.78009678
3	14.487561	14.110797	13.978556	13.404529	0.57402685
4	13.606002	14.110329	14.027559	13.429811	0.59774819
5	13.446417	13.431874	13.744636	13.078459	0.6661772
6	13.53751	15.059941	13.872945	13.446437	0.42650726
7	13.41718	13.397106	13.768616	13.535462	0.23315375
8	13.485344	13.137269	13.985712	13.493218	0.49249426
9	14.693874	12.357829	13.711952	13.703711	0.00824110
10	14.122285	12.582788	14.058332	13.027801	1.030531
11	13.80927	12.137279	13.984251	13.07533	0.90892056
12	14.83092	13.314315	13.472359	13.536485	-0.0641257
13	13.037448	12.615349	13.782737	13.192257	0.59047988
14	13.827328	14.017135	13.972256	13.094431	0.87782572
15	14.608695	12.674141	13.471012	13.671909	-0.2008963
16	12.94888	12.149741	13.566834	12.951453	0.61538111
17	14.791846	13.539139	13.901516	13.020553	0.88096328
18	13.851718	13.978509	13.620199	12.980353	0.63984664
19	14.977034	13.64188	14.020498	13.29524	0.72525804
20	12.88218	14.083267	13.846815	13.487639	0.35917605
21	13.290608	13.999414	14.090146	13.714581	0.37556591
22	13.784498	13.762006	13.802647	13.279055	0.5235919
23	14.184148	11.401661	14.049755	13.72915	0.32060535

Figure 7-77: StatCrunch worksheet after calculating difference in means

8) Find the 90% confidence interval for the diff column by selecting Stat>Summary Stats>Columns and selecting the diff column. Click n (you must select at least one statistic) enter 5, 95 in the Percentiles box as shown in figure 7-78. Click Compute!



Select column(s): diff

Where: --optional--

Group by: --optional--

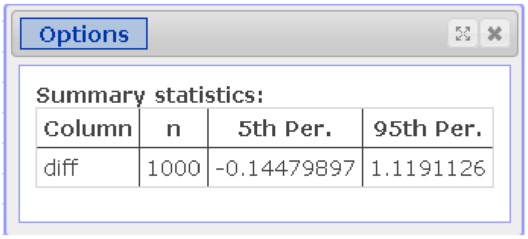
Statistics: n

Percentiles (comma separated): 5, 95

Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 7-78: Column Statistics for diff (n and Percentiles only)



Summary statistics:

Column	n	5th Per.	95th Per.
diff	1000	-0.14479897	1.1191126

Figure 7-79: 90% confidence interval

9) The 90% confidence interval for the example simulation is shown in figure 7-79.

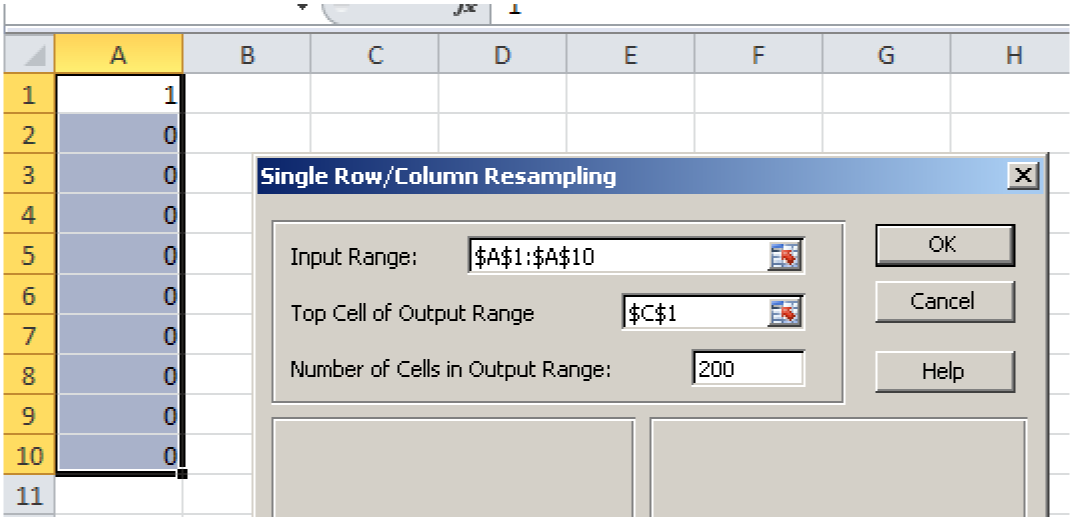
8 Hypothesis Tests

8.1 Review of Terminology

Example
An online merchant has historically experienced a 10% return rate in the "kitchen gadget" category. In an effort to reduce returns, it does a pilot in which it invests in additional explanatory information and pictures about several products. Out of the next 200 purchases, 16 are returned. Is the pilot effective?

Resampling Stats

1) Use the Urn or manually enter 9 0's and a 1 in cells A1:A10. Select Resample and complete the dialog as shown in figure 8-1.



Single Row/Column Resampling

Input Range: $A$1:$A$10

Top Cell of Output Range $C$1

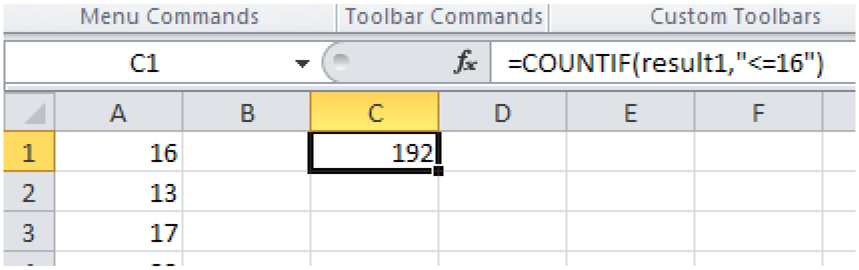
Number of Cells in Output Range: 200

Figure 8-1: Return rate

2) In cell E1, enter =SUM(C1:C200)

3) Select cell E1 (the score cell) and choose Repeat and Score. Enter 1000 trials and click OK.

4) On the Results sheet, enter =COUNTIF(result1,"<=16") in cell C1. Figure 8-2 shows the results of an experiment:



C1 =COUNTIF(result1,"<=16")

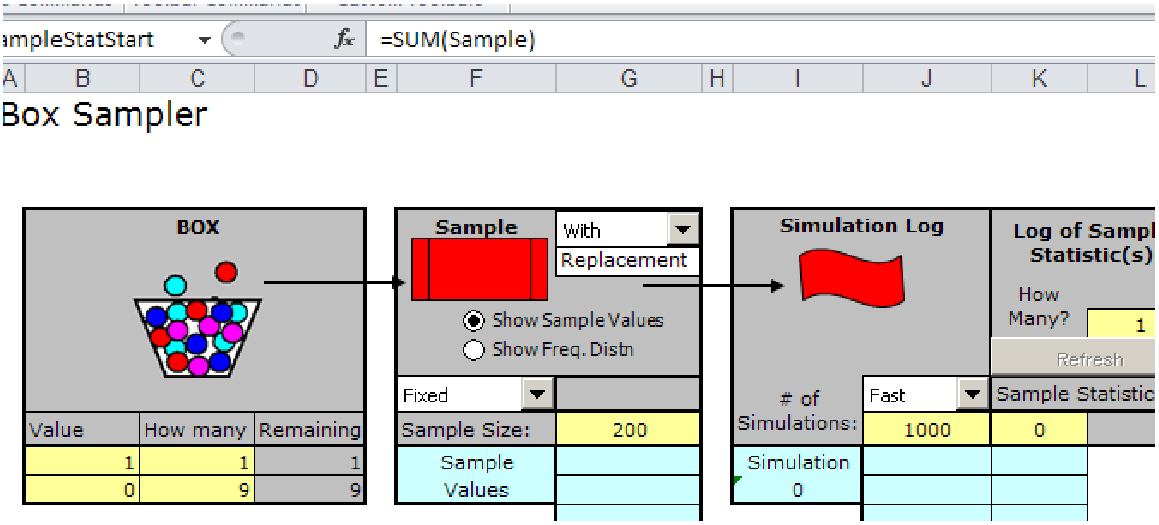
	A	B	C	D	E	F
1	16		192			
2	13					
3	17					

Figure 8-2: 192 values: an estimated p-value of 0.192

5) The estimated p-value is 192/1000 or 0.192. Create a histogram and compare it to figure 8-1 in the text.

Box Sampler

1) Open Box Sampler, choose a one box model, and enter values as shown in figure 8-3.



ampleStatStart =SUM(Sample)

Value	How many	Remaining
1	1	1
0	9	9

Sample: With Replacement

● Show Sample Values

○ Show Freq. Distn

Fixed

Sample Size: 200

Sample Values

Fast

# of Simulations: 1000

Simulation 0

How Many? 1

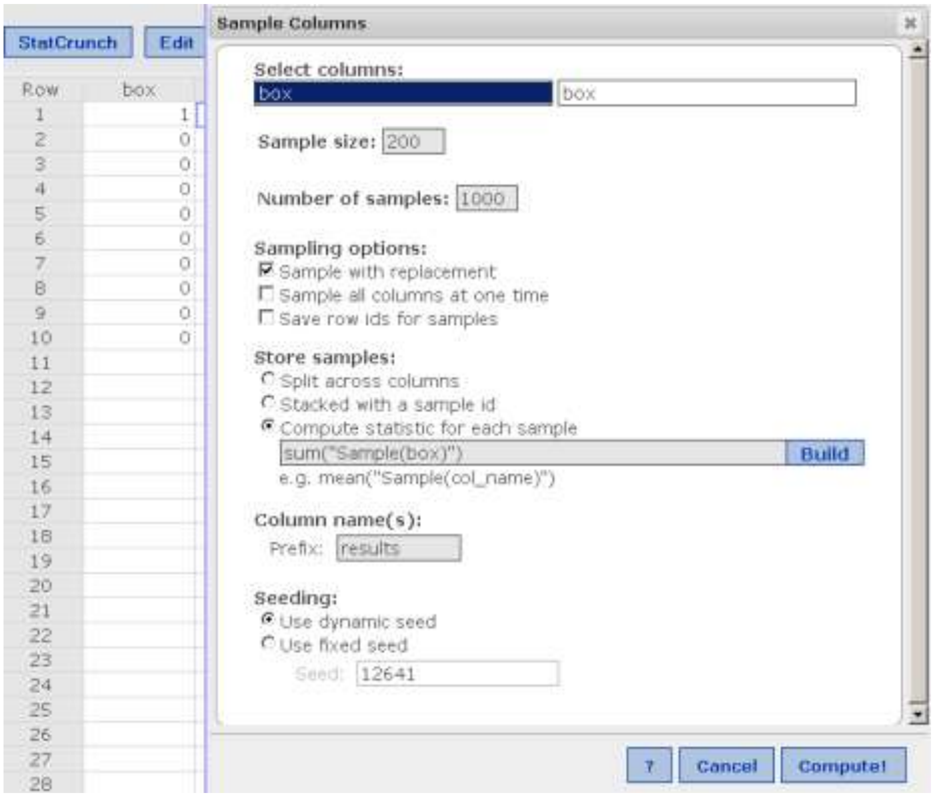
Sample Statistic 0

Figure 8-3: Box Sampler setup

2) Click Simulate and have patience! When the simulation has ended, enter =COUNTIF(Stat1,"<=16") in an empty cell such as J1. One such result ended with 189 instances of 16 or fewer returns for an estimated p-value of 0.189.

StatCrunch

1) Open StatCrunch and enter the values representing the box (1-1 and 9-0's). Select Data>Sample and enter the information in the dialog as shown in figure 8-4 (the new column name should be "results"). Click Compute!



Row	box
1	1
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0

Sample Columns

Select columns: box

Sample size: 200

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

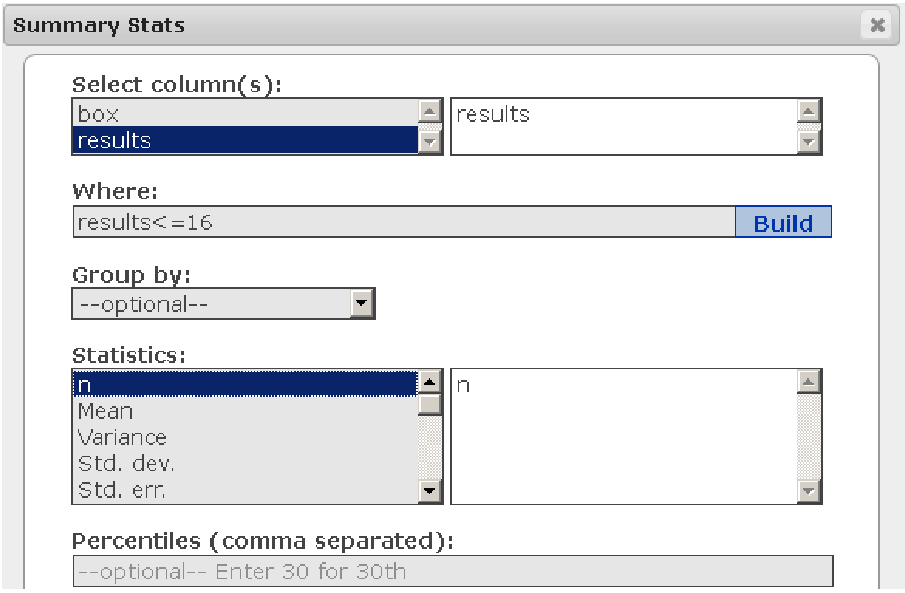
Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
sum("Sample(box)")
e.g. mean("Sample(col_name)")

Column name(s):
Prefix: results

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 8-4: Sample Columns

2) Select Stat>Summary Stats>Columns and enter the data shown in figure 8-5. Click Compute!



Summary Stats

Select column(s): box, results → results

Where: results<=16

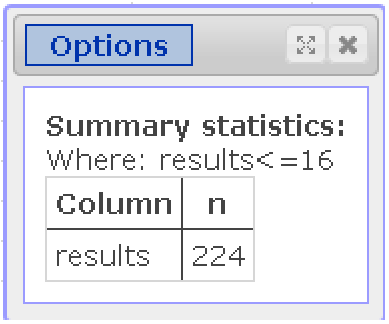
Group by: --optional--

Statistics: n, Mean, Variance, Std. dev., Std. err. → n

Percentiles (comma separated): --optional-- Enter 30 for 30th

Figure 8-5: Column statistics using the Where: function

3) The result of one experiment is shown in figure 8-6.



Summary statistics:
Where: results<=16

Column	n
results	224

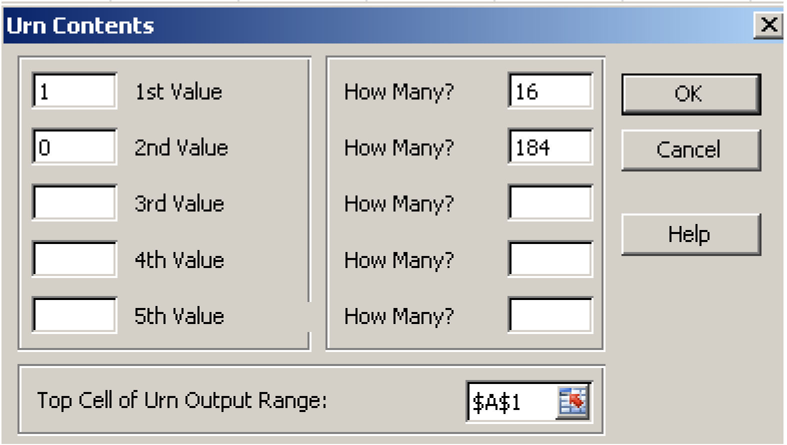
Figure 8-6: p-value estimate = 0.224

4) There were 224 “successes” which is a p-value estimate of 224/1000 or 0.224.

   Example
   The next example uses the same sample results, but a different scenario, to illustrate the different question that the confidence interval seeks to address. An online merchant wants to get an accurate estimate of its return rate. Out of the next 200 purchases, 16 are returned. This is an 8% return rate, but how accurate is this estimate?

Resampling Stats

1) Create a box/urn using the Urn feature (figure 8-7).

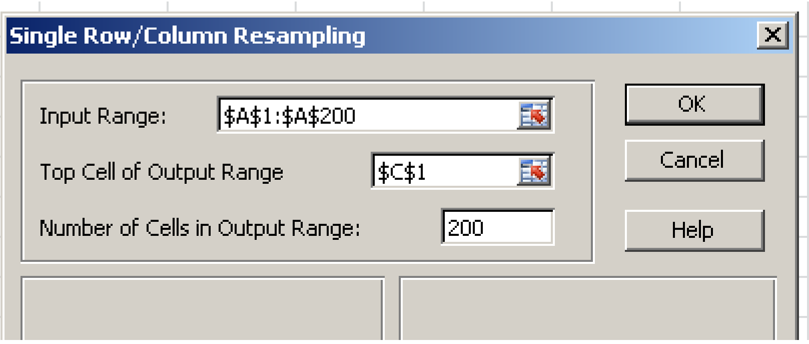


1	1st Value	How Many?	16
0	2nd Value	How Many?	184
	3rd Value	How Many?	
	4th Value	How Many?	
	5th Value	How Many?	

Top Cell of Urn Output Range: $A$1

Figure 8-7: Resampling Stats Urn feature

2) Select Resample and fill in the dialog as follows (figure 8-8).



Single Row/Column Resampling

Input Range: $A$1:$A$200

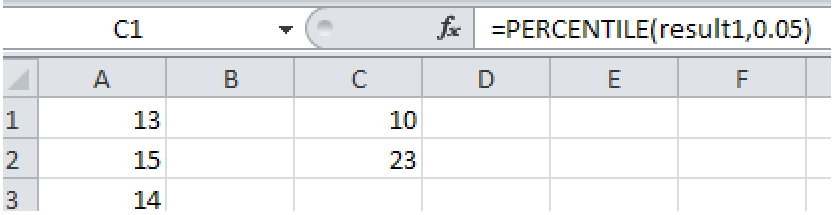
Top Cell of Output Range $C$1

Number of Cells in Output Range: 200

Figure 8-8: Resample dialog

3) Enter =SUM(C1:C200) in cell E1. Select cell E1 and choose Repeat and Score. Enter 1000 trials and click OK.

4) Find the 90% confidence interval by using the =PERCENTILE(result1, 0.05) and =PERCENTILE(result1, 0.95) functions in cells C1 and C2 (figure 8-9).

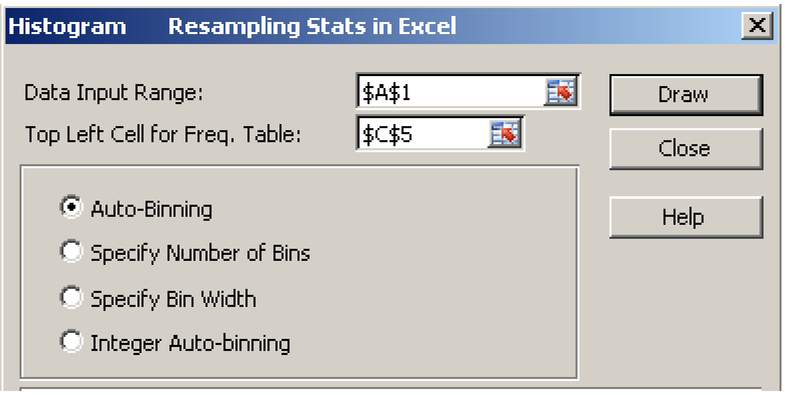


C1 =PERCENTILE(result1,0.05)

	A	B	C	D	E	F
1	13		10			
2	15		23			
3	14					

Figure 8-9: 90% confidence interval using the =PERCENTILE() functions

5) The 90% confidence interval for this experiment was 10 to 23. Create a histogram by selecting cell A1(for automatic data range selection by the histogram function), choosing the histogram feature and filling in the dialog as follows in figure 8-10. Note: you could also enter result1 or A1:A1000 for the Data Input Range. Click Draw.



Histogram Resampling Stats in Excel

Data Input Range: $A$1

Top Left Cell for Freq. Table: $C$5

● Auto-Binning

○ Specify Number of Bins

○ Specify Bin Width

○ Integer Auto-binning

Figure 8-10: Histogram dialog

6) The results of the histogram are as follows:

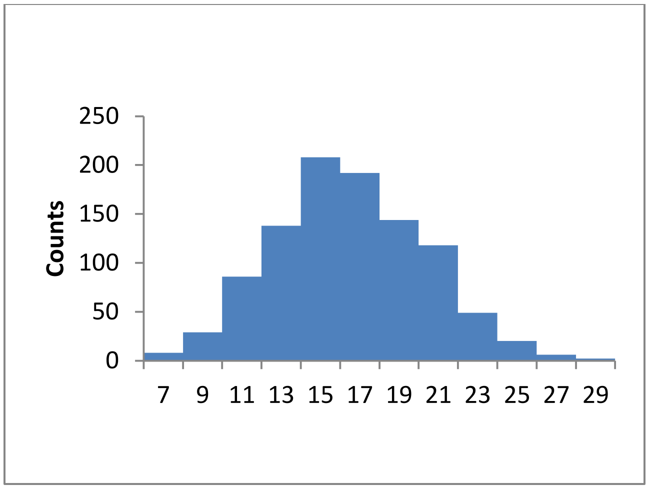
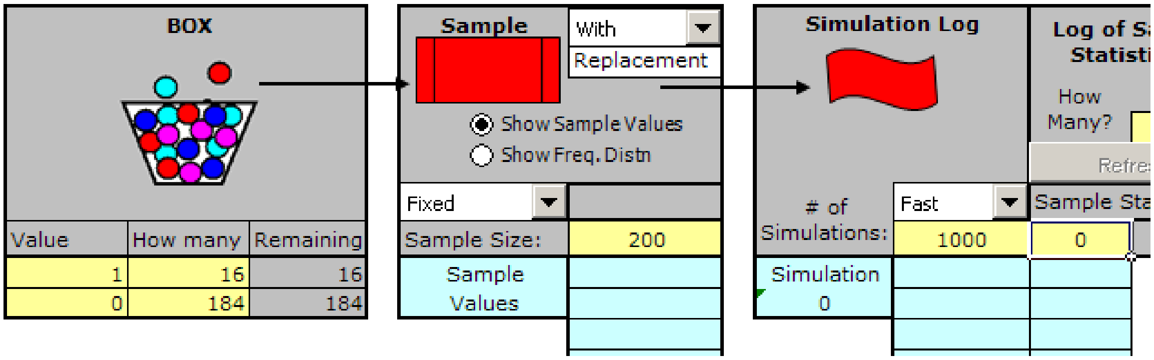


Figure 8-11: RSXL histogram

Box Sampler

1) Use a one Box model and enter the following values (figure 8-12). The Sample Statistic is =SUM(Sample) as in the previous model.



Value	How many	Remaining
1	16	16
0	184	184

With Replacement

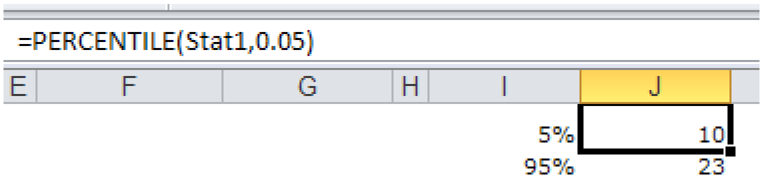
● Show Sample Values
○ Show Freq. Distn

Fixed
Sample Size: 200
Sample Values

Fast
# of Simulations: 1000
0
Simulation 0

Figure 8-12: Box Sampler setup

2) Click Simulate and have patience! The results of one experiment are shown in figure 8-13.



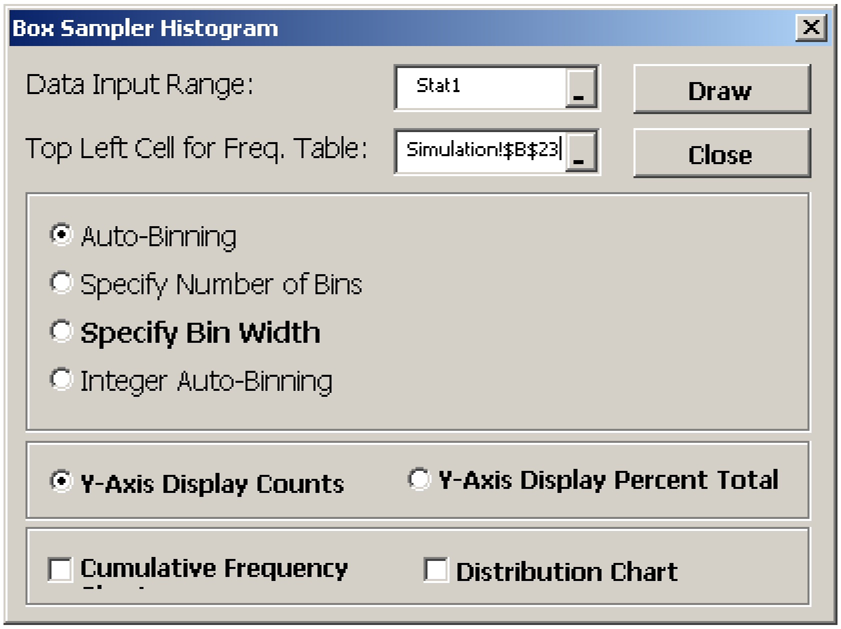
=PERCENTILE(Stat1,0.05)

E	F	G	H	I	J
				5%	10
				95%	23

Figure 8-13: 90% confidence interval

3) The 90% confidence interval is from 10 to 23. You can create a histogram of the results in the Stat1 range. Select the BoxSampler>StatCharts>Histogram menu

item and fill in the Histogram dialog with Stat1 as the Data Input and click on cell B23 for the Top Left cell (Figure 8-14). Click Draw. The resulting histogram is displayed in figure 8-15.



Data Input Range: Stat1

Top Left Cell for Freq. Table: Simulation!$B$23

● Auto-Binning
○ Specify Number of Bins
○ Specify Bin Width
○ Integer Auto-Binning

● Y-Axis Display Counts ○ Y-Axis Display Percent Total

☐ Cumulative Frequency Chart ☐ Distribution Chart

Figure 8-14: Histogram dialog

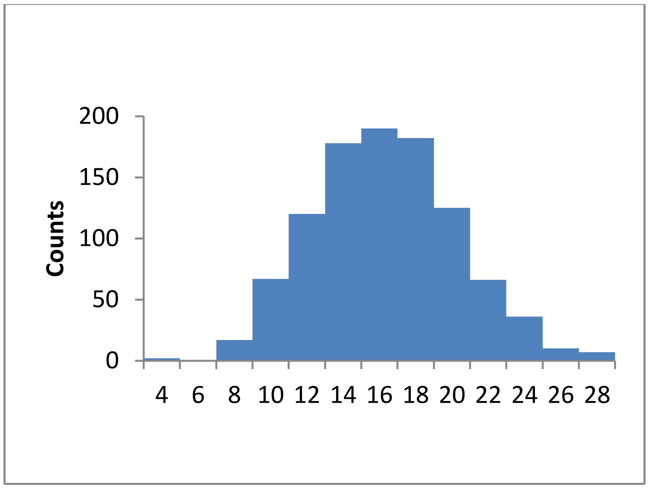


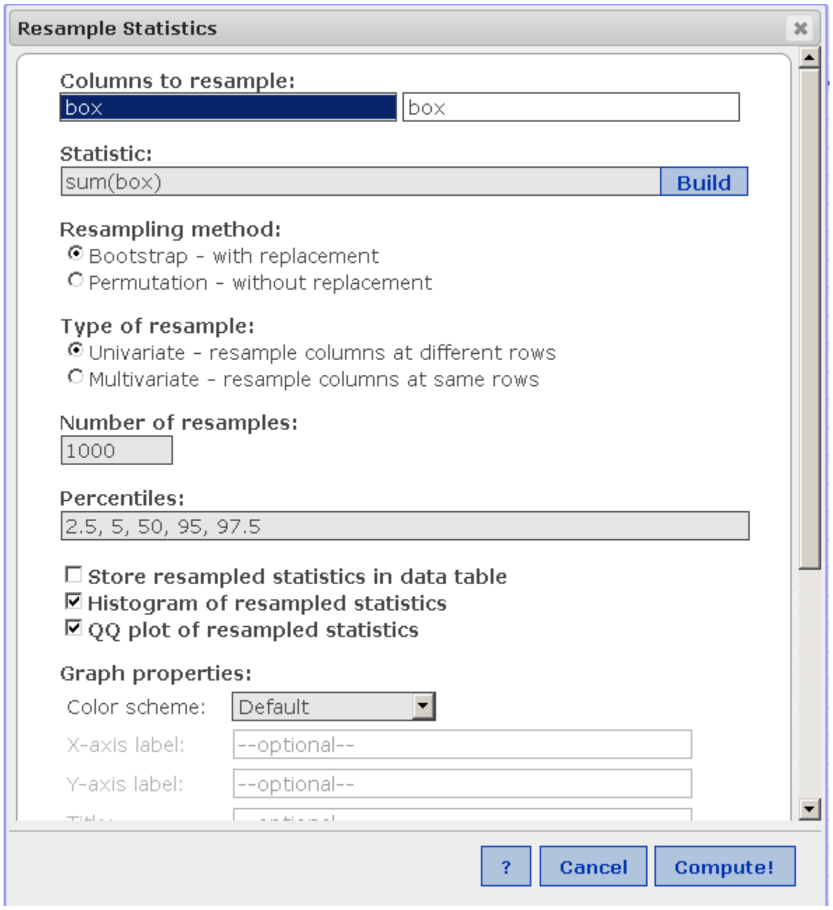
Figure 8-15: Histogram

StatCrunch

1) Select Data>Compute Expression and enter: concat(rep(1,16),rep(0,184)) to create the box/urn containing 16 1's and 184 0's.

2) Enter box for the Column label and click Compute!

3) Select Stat>Resample>Statistic and select the box column. The Statistic expression is: sum(box) (figure 8-16).



Columns to resample: box | box

Statistic: sum(box)

Resampling method:

● Bootstrap - with replacement

○ Permutation - without replacement

Type of resample:

● Univariate - resample columns at different rows

○ Multivariate - resample columns at same rows

Number of resamples: 1000

Percentiles: 2.5, 5, 50, 95, 97.5

☐ Store resampled statistics in data table

☑ Histogram of resampled statistics

☑ QQ plot of resampled statistics

Graph properties:

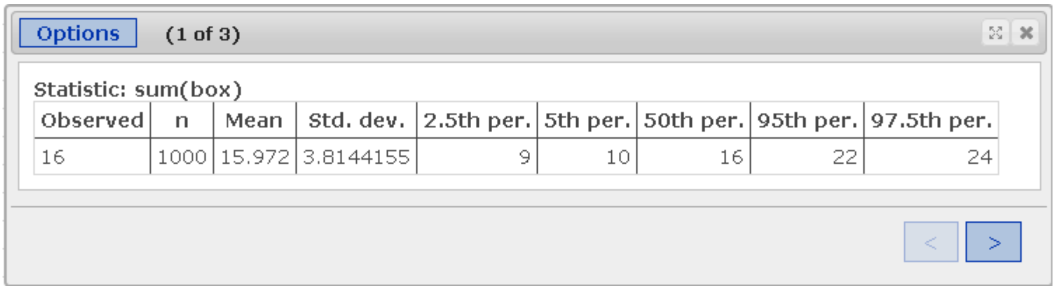
Color scheme: Default

X-axis label: --optional--

Y-axis label: --optional--

Figure 8-16: Resample Statistic

4) Click Compute! to view the results (figure 8-17).



Statistic: sum(box)

Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
16	1000	15.972	3.8144155	9	10	16	22	24

Figure 8-17: Results with confidence intervals

5) The 90% confidence interval is from 10 to 22. Click the > button to view a histogram of the data. You may need to resize the output dialog box.

8.3 Comparing Two Means

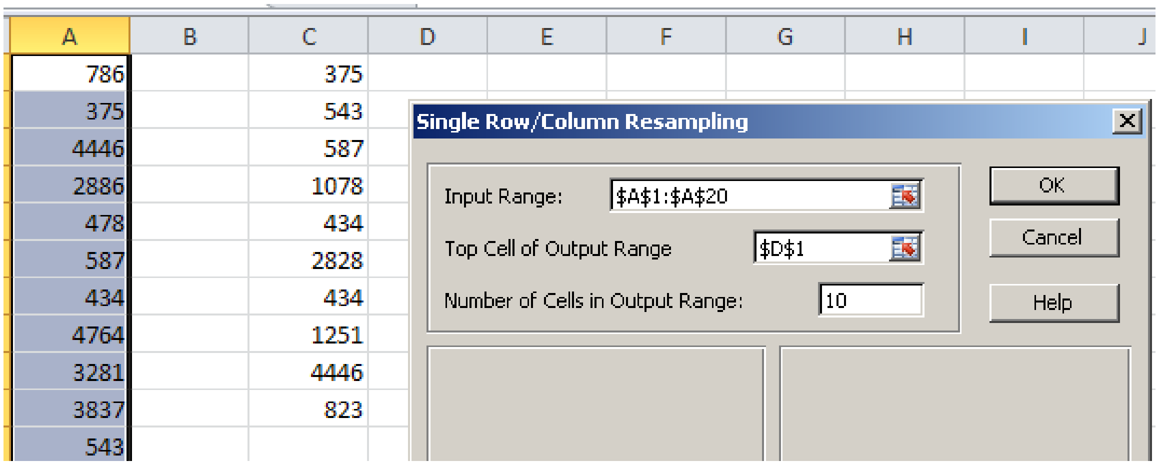
Procedures for the Blood Loss in Pigs example.

Resampling Stats

1) Enter all 20 pig blood loss values in cells A1:A20.

2) Select cell A1 and choose Resample. Choose cell C1 as the Top Cell of the Output Range and 10 as the Number of Cells. Click OK.

3) Repeat the procedure in step 2, using cell D1 as the Top Cell (figure 8-18).



A	B	C
786		375
375		543
4446		587
2886		1078
478		434
587		2828
434		434
4764		1251
3281		4446
3837		823
543		

Single Row/Column Resampling

Input Range: $A$1:$A$20

Top Cell of Output Range $D$1

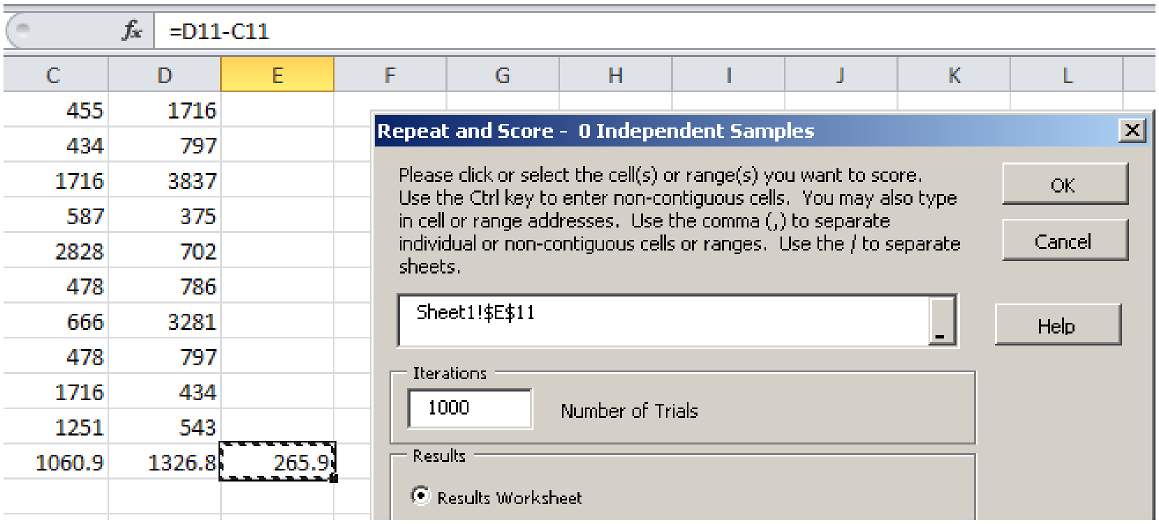
Number of Cells in Output Range: 10

Figure 8-18: Second resample from the Pig Blood box

4) Find the average for each resample by entering =AVERAGE(C1:C10) in cell C11 and =AVERAGE(D1:D10) in cell D11.

5) Subtract the mean of the first resample from the mean of the second resample by entering =D11-C11 in cell E11.

6) Select cell E11 (the score cell or statistic of interest) and choose Repeat and Score. Enter 1000 for the number iterations or trials (figure 8-19). Click OK.



=D11-C11

C	D	E
455	1716	
434	797	
1716	3837	
587	375	
2828	702	
478	786	
666	3281	
478	797	
1716	434	
1251	543	
1060.9	1326.8	265.9

Repeat and Score - 0 Independent Samples

Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

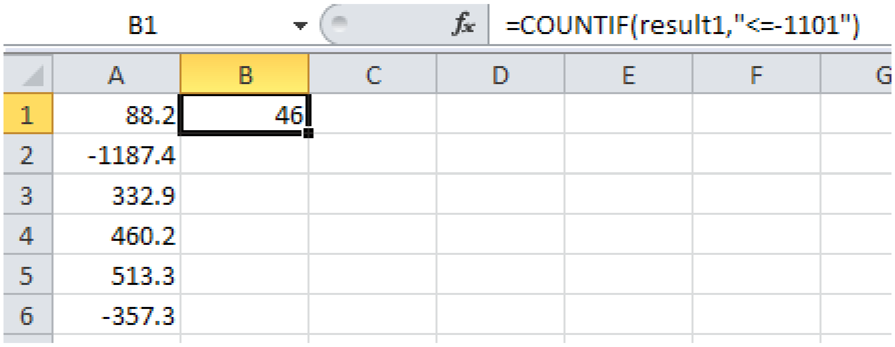
Sheet1!$E$11

Iterations: 1000 Number of Trials

Results: ● Results Worksheet

Figure 8-19: Repeat and Score

7) On the Results worksheet, enter =COUNTIF(result1,"<=-1101") in cell B1 (figure 8-20).



B1 =COUNTIF(result1,"<=-1101")

	A	B	C	D	E	F	G
1	88.2	46					
2	-1187.4						
3	332.9						
4	460.2						
5	513.3						
6	-357.3						

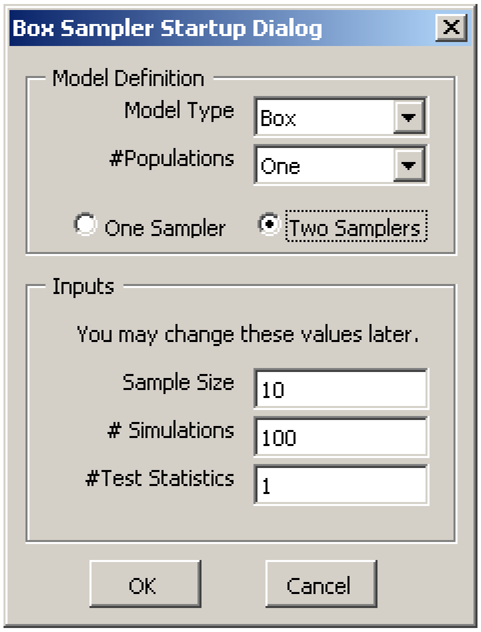
Figure 8-20: Results: estimated p-value = 0.046

8) There were 46 instances of values <=-1101, for an estimated p-value = 0.046.

9) Create a histogram of your results and compare it to figure 8-2 in the text.

Box Sampler

1) This experiment requires taking two samples from a single box, so in the initial Box Sampler dialog we need to check the "Two Samplers" option button (figure 8-21).



Box Sampler Startup Dialog

Model Definition

Model Type: Box

#Populations: One

○ One Sampler ● Two Samplers

Inputs

You may change these values later.

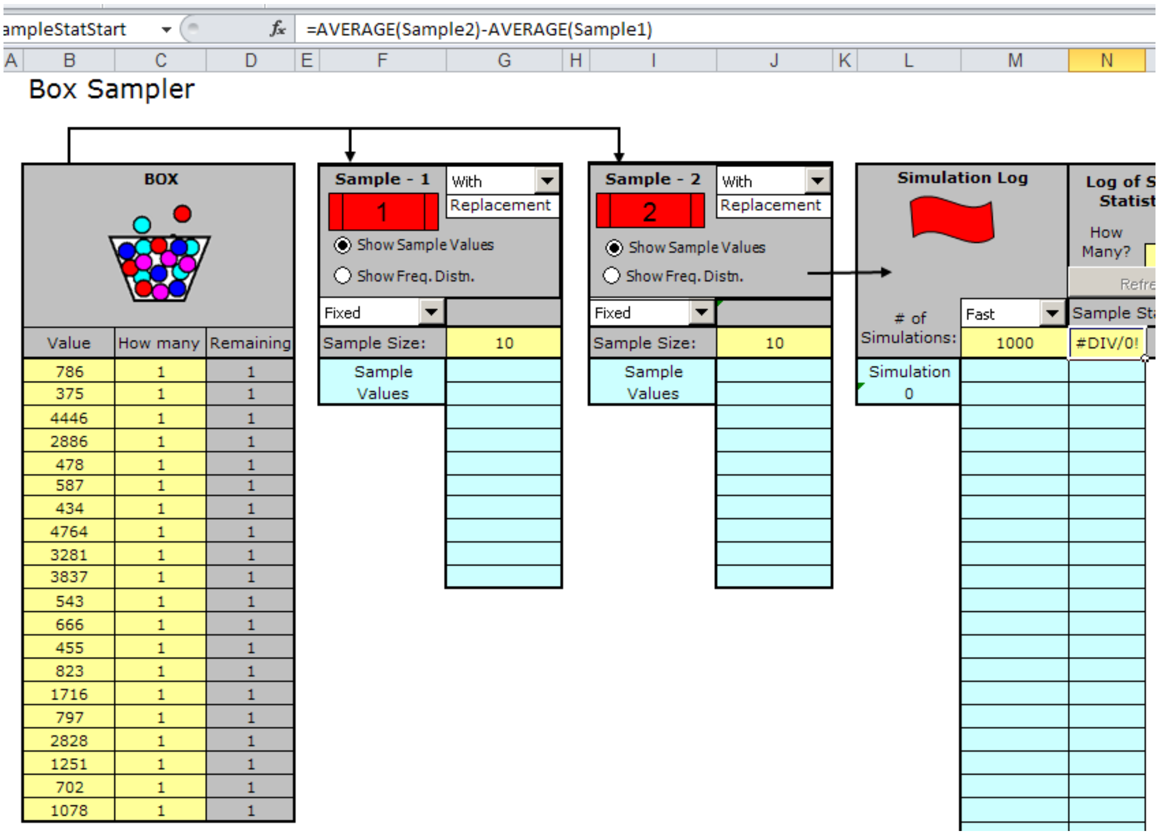
Sample Size: 10

Simulations: 100

#Test Statistics: 1

Figure 8-21: Two Samplers in initial Box Sampler dialog

2) Enter the pig blood loss data and in cell N11 (the Sample Statistics cell), the function =AVERAGE(Sample2) – AVERAGE(Sample1) where Sample2 and Sample1 are the names for the range of sample values in Sample-2 and Sample-1 respectively. Enter 1000 for the # of Simulations (figure 8-22). Click the Simulate button.



ampleStatStart

=AVERAGE(Sample2)-AVERAGE(Sample1)

Box Sampler

Value	How many	Remaining
786	1	1
375	1	1
4446	1	1
2886	1	1
478	1	1
587	1	1
434	1	1
4764	1	1
3281	1	1
3837	1	1
543	1	1
666	1	1
455	1	1
823	1	1
1716	1	1
797	1	1
2828	1	1
1251	1	1
702	1	1
1078	1	1

Sample - 1: With Replacement; ● Show Sample Values ○ Show Freq. Distn.; Fixed; Sample Size: 10

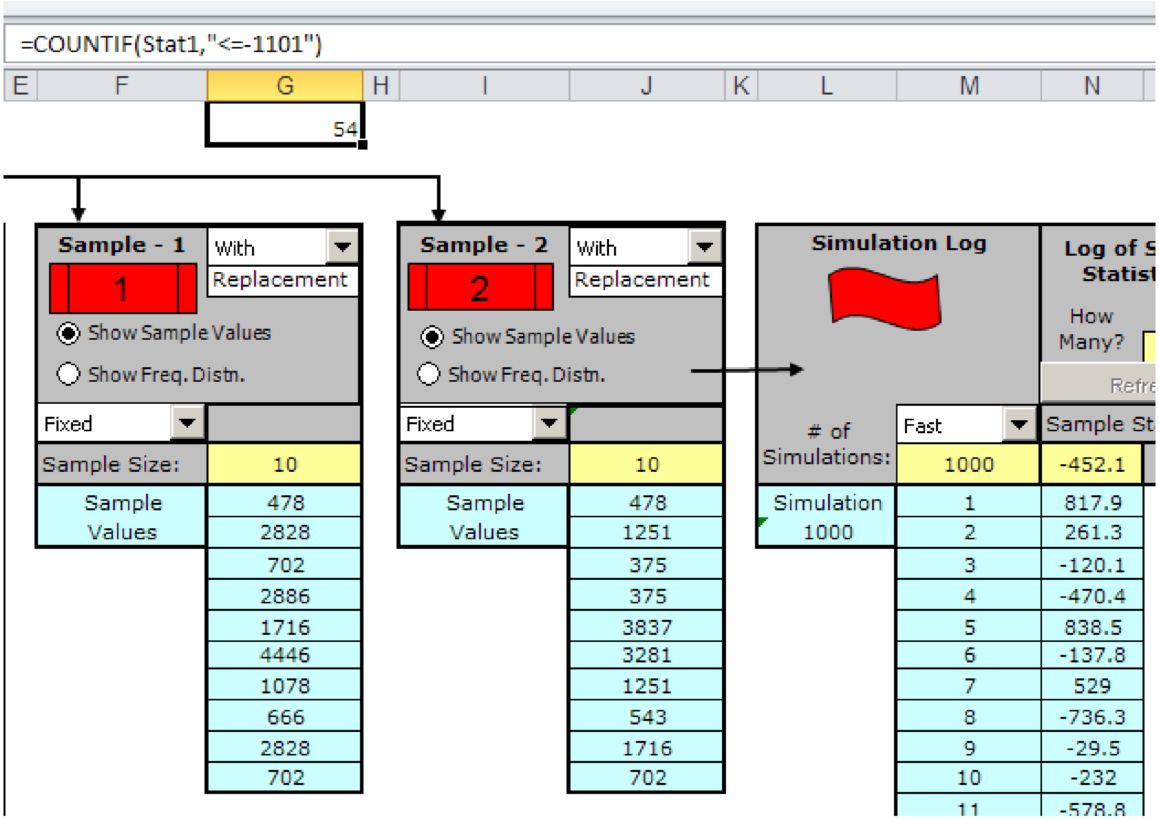
Sample - 2: With Replacement; ● Show Sample Values ○ Show Freq. Distn.; Fixed; Sample Size: 10

Simulation Log: # of Simulations: 1000; Fast; Simulation 0

Sample St: #DIV/0!

Figure 8-22: Initial Box Sampler setup

3) When the simulation has finished, enter =COUNTIF(Stat1,”<=-1101”) in an empty cell such as G1 (figure 8-23).



=COUNTIF(Stat1,"<=-1101")

G: 54

Sample - 1: 1 — With Replacement — ● Show Sample Values ○ Show Freq. Distn. — Fixed — Sample Size: 10

Sample - 2: 2 — With Replacement — ● Show Sample Values ○ Show Freq. Distn. — Fixed — Sample Size: 10

Simulation Log: # of Simulations: Fast 1000 — Sample St -452.1

Sample Values (Sample - 1)	Sample Values (Sample - 2)
478	478
2828	1251
702	375
2886	375
1716	3837
4446	3281
1078	1251
666	543
2828	1716
702	702

Simulation 1000	Sample St
1	817.9
2	261.3
3	-120.1
4	-470.4
5	838.5
6	-137.8
7	529
8	-736.3
9	-29.5
10	-232
11	-578.8

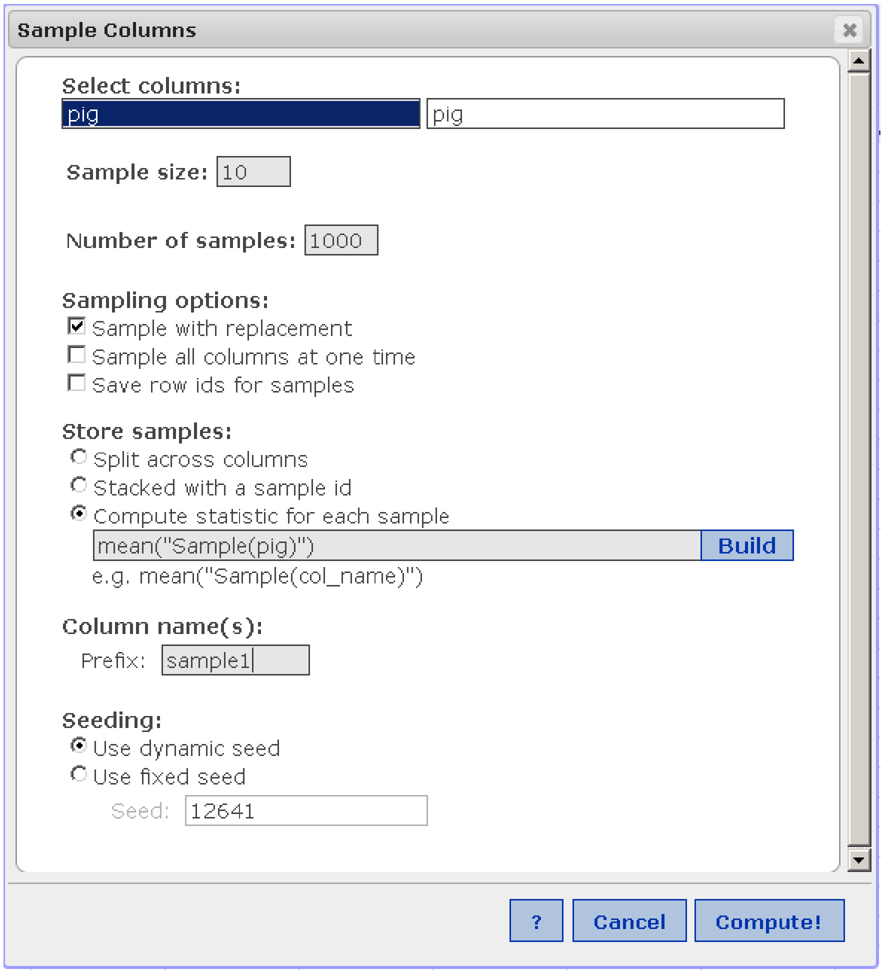
Figure 8-23: Estimated p-value = 0.054

4) The estimated p-value from this experiment was 0.054.

StatCrunch

1) Enter the pig blood values in the first column. Rename the column "pig".

2) Select Data>Sample, select the pig column, enter a Sample size of 10, check Sample with replacement, and enter 1000 Samples. The Statistic expression is mean("Sample(pig)") and will find the mean of each sample. Enter sample1 for the name of the column and click Compute! (figure 8-24).



Select columns: pig | pig

Sample size: 10

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

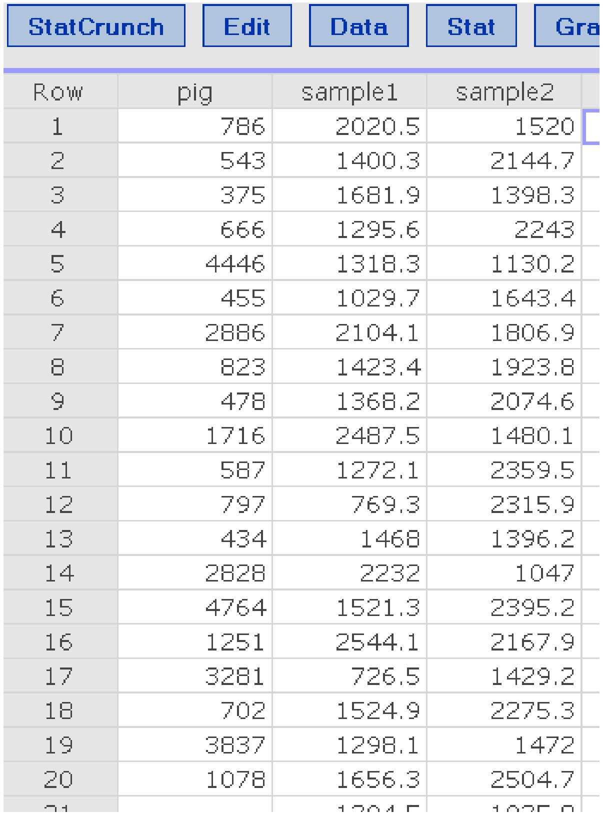
Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
mean("Sample(pig)")
e.g. mean("Sample(col_name)")

Column name(s):
Prefix: sample1

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 8-24: Sample columns

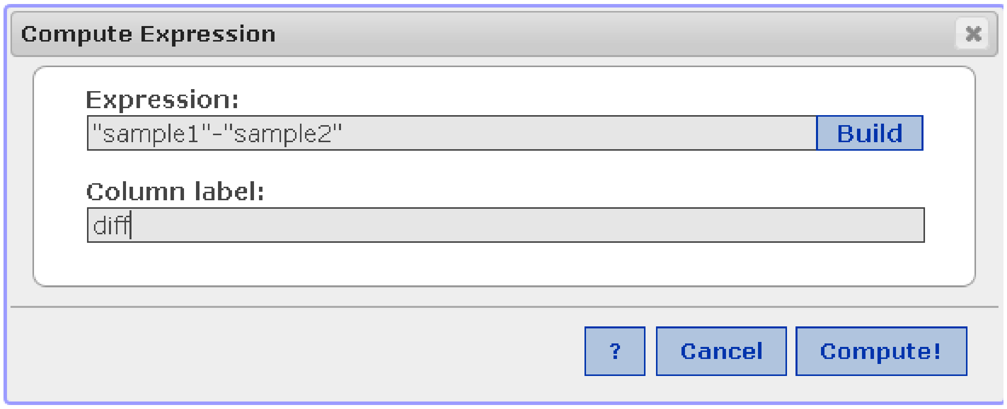
3) Repeat step 2 for the second sample. Name the new column sample2 and click Compute! See figure 8-25 for an example of the output.



Row	pig	sample1	sample2
1	786	2020.5	1520
2	543	1400.3	2144.7
3	375	1681.9	1398.3
4	666	1295.6	2243
5	4446	1318.3	1130.2
6	455	1029.7	1643.4
7	2886	2104.1	1806.9
8	823	1423.4	1923.8
9	478	1368.2	2074.6
10	1716	2487.5	1480.1
11	587	1272.1	2359.5
12	797	769.3	2315.9
13	434	1468	1396.2
14	2828	2232	1047
15	4764	1521.3	2395.2
16	1251	2544.1	2167.9
17	3281	726.5	1429.2
18	702	1524.9	2275.3
19	3837	1298.1	1472
20	1078	1656.3	2504.7

Figure 8-25: StatCrunch worksheet after steps 1-3

4) Choose Data>Compute Expression and enter "sample2"-"sample1" as the expression and diff as the new column name (figure 8-26).



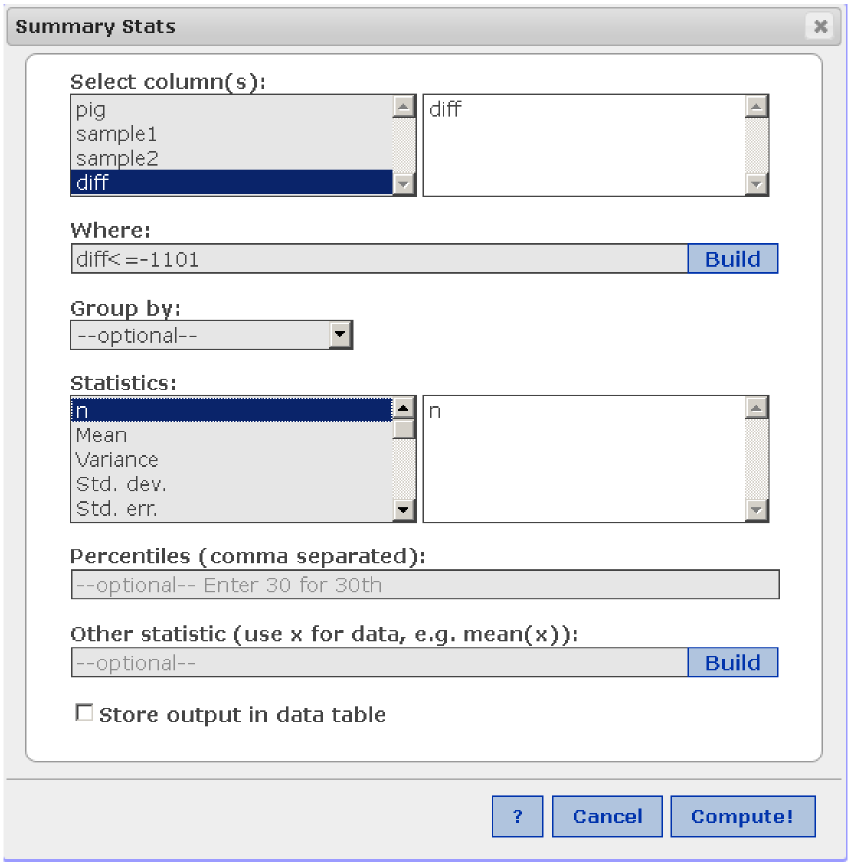
Expression: "sample1"-"sample2"

Column label: diff

Figure 8-26: Compute expression: finding the difference in sample means

5) Select Stat>Column Stats and choose the diff column. For the Where: expression, enter diff<=-1101. We are interested in the value for n, which would be the number of instances in which the condition of the difference in means (diff<=-1101) is met. We could click Compute! at any point after entering the Where

expression and n would be included in the results. However, in this example we only click n to simplify the output (figure 8-27). Click Compute!



Summary Stats

Select column(s): diff

Where: diff<=-1101

Group by: --optional--

Statistics: n

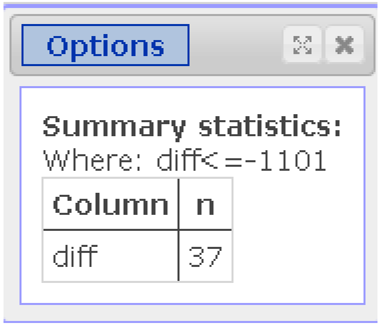
Percentiles (comma separated): --optional-- Enter 30 for 30th

Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 8-27: Column Statistics: Where: diff<=-1101

6) The results are shown in figure 8-28.



Summary statistics:
Where: diff<=-1101

Column	n
diff	37

Figure 8-28: n = 37

7) The estimated p-value is 0.037.

8) Create a histogram of the diff column. Try the Normal option in the Overlay listbox and a color option in the Color scheme listbox. Add axis labels and a title. See an example in figure 8-29.

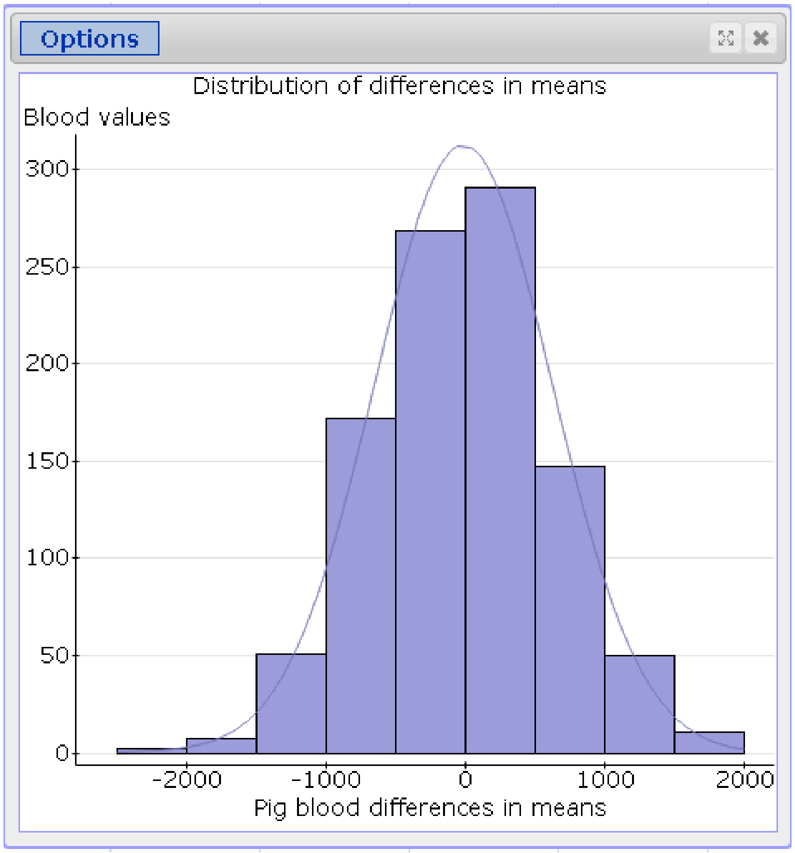
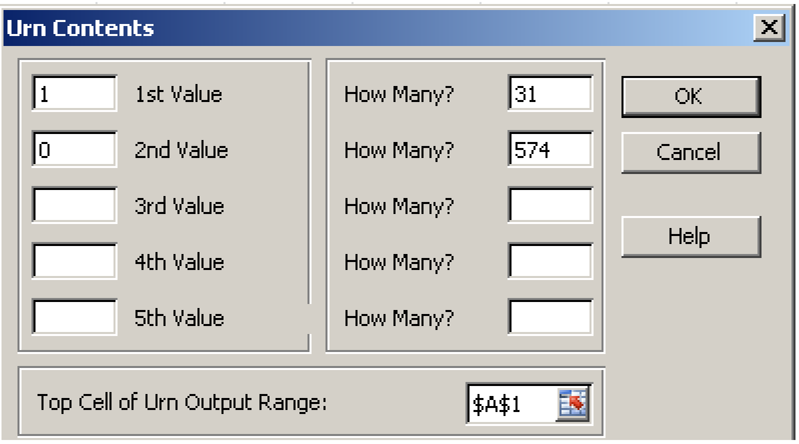

Options
Distribution of differences in means
Blood values
300
250
200
150
100
50
0
-2000
-1000
0
1000
2000
Pig blood differences in means

Figure 8-29: Histogram of pig blood differences

8.4 Comparing Two Proportions

Resampling Stats (Cholesterol and MI)

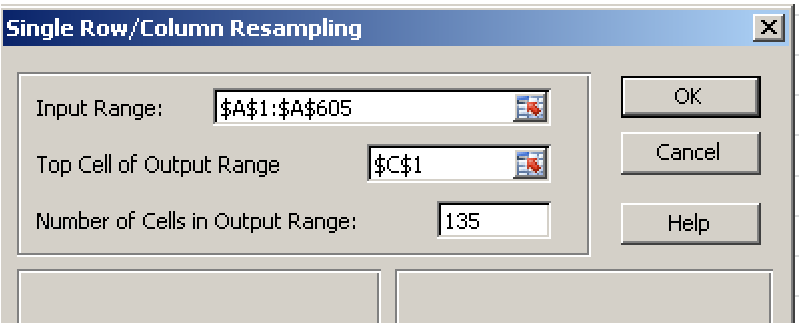
1) Use the Urn feature to populate the box with 31 1's and 574 0's (figure 8-30). Click OK.



1st Value: 1 — How Many? 31
2nd Value: 0 — How Many? 574
3rd Value: — How Many?
4th Value: — How Many?
5th Value: — How Many?
Top Cell of Urn Output Range: $A$1

Figure 8-30: Urn contents for Cholesterol and MI

2) Select cell A1 and choose Resample. Click cell C1 for the Top Cell and enter 135 for the Number of Cells (figure 8-31). Click OK.

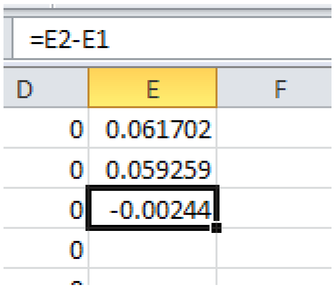


Input Range: $A$1:$A$605
Top Cell of Output Range $C$1
Number of Cells in Output Range: 135

Figure 8-31: First resample

3) Repeat step 2 using cell D1 for the Top Cell and 470 for the Number of Cells. Click OK.

4) To find the difference in proportions (proportion 2 – proportion 1), in cell E1 enter =SUM(D1:D470)/470 and in cell E2, enter =SUM(C1:C135)/135. In cell E3, enter =E2-E1 to subtract the first proportion from the second proportion (figure 8-32).

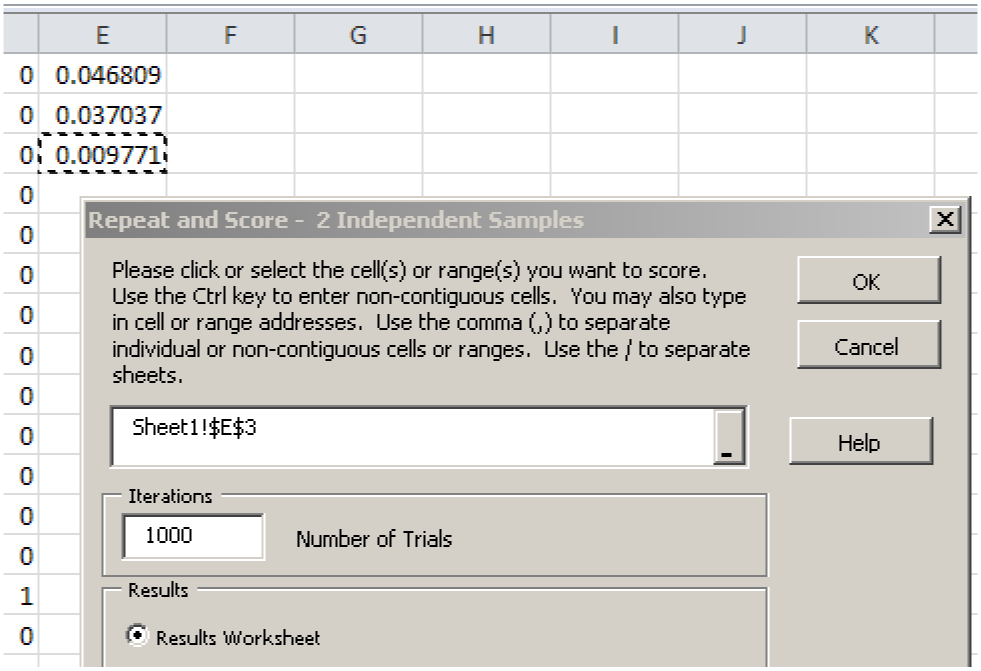


=E2-E1

D	E	F
0	0.061702	
0	0.059259	
0	-0.00244	
0		

Figure 8-32: Finding the difference in proportions

5) Select cell E3 and choose Repeat and Score. Enter 1000 for the number of iterations/trials and click OK (figure 8-33).



	E	F	G	H	I	J	K
0	0.046809						
0	0.037037						
0	0.009771						
0							
0							
0							
0							
0							
0							
0							
0							
0							
0							
1							
0							

Repeat and Score - 2 Independent Samples

Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

Sheet1!$E$3

Iterations

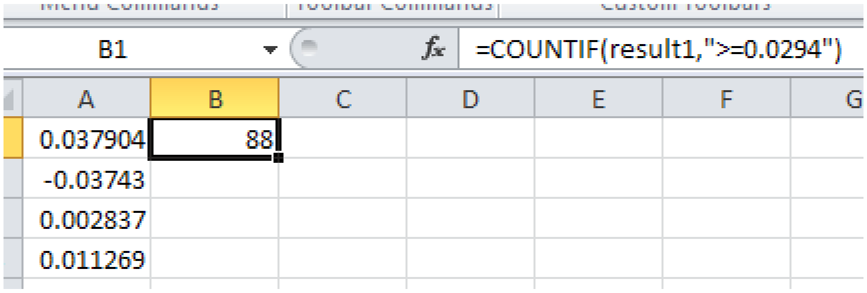
1000 Number of Trials

Results

● Results Worksheet

Figure 8-33: Repeat and Score with two independent samples

6) On the Results sheet, enter =COUNTIF(result1,"”>=0.0294”) in cell B1 (figure 8-34).



B1 =COUNTIF(result1,">=0.0294")

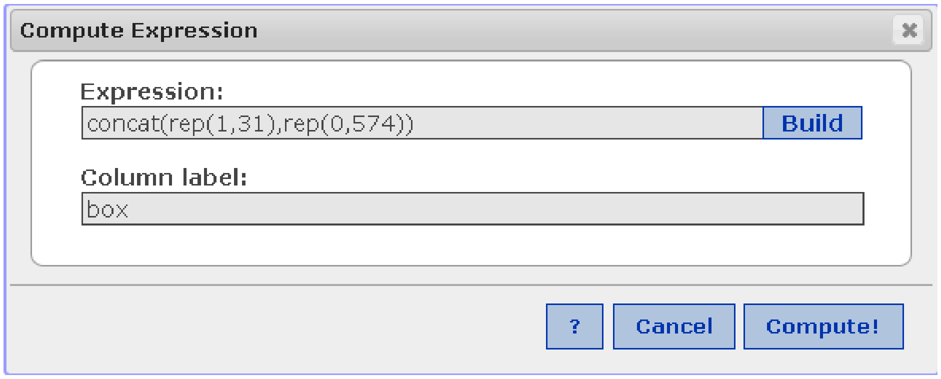
A	B	C	D	E	F	G
0.037904	88					
-0.03743						
0.002837						
0.011269						

Figure 8-34: Estimated p-value = 0.088

7) There were 88 instances where the difference in proportions was >=0.0294. The estimated p-value is 88/1000 or 0.088.

StatCrunch

1) Select Data>Compute Expression and enter the expression: concat(rep(1,31),rep(0,574)) to populate the box with 31 1s and 574 0s (figure 8-35).

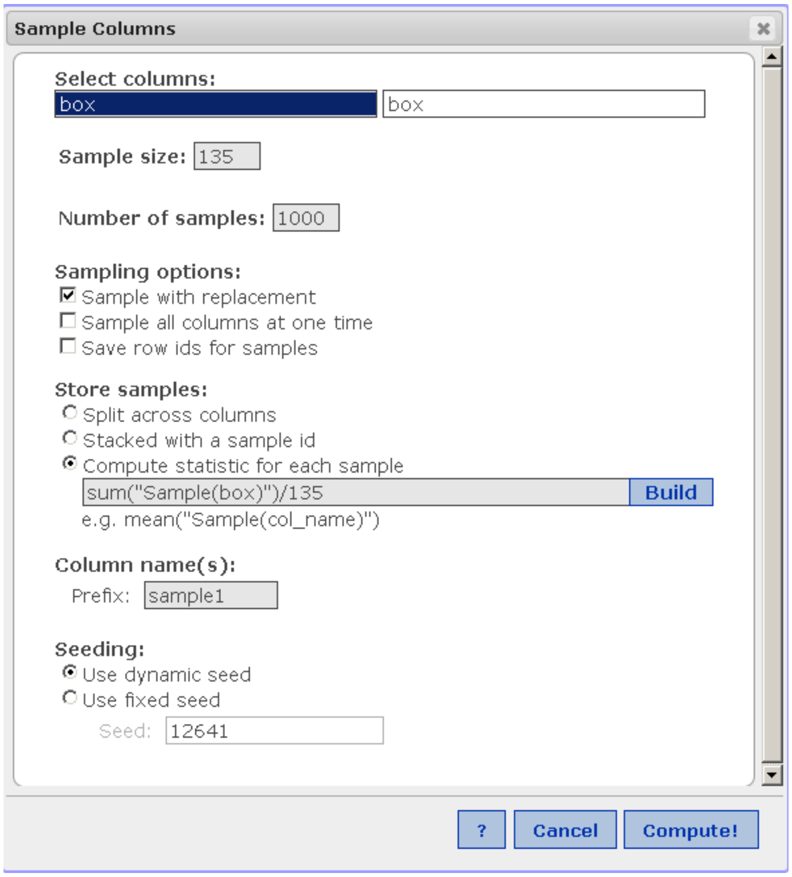


Expression: concat(rep(1,31),rep(0,574))

Column label: box

Figure 8-35: Compute expression to populate box

2) Choose Data>Sample and select the box column, enter a Sample Size of 135, check Sample with replacement, and enter 1000 for the number of samples. The Statistic expression to calculate the proportion for the first sample is: sum("Sample(box)")/135 and the column name is sample1 (figure 8-36). Click Compute!



Select columns: box | box

Sample size: 135

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
sum("Sample(box)")/135
e.g. mean("Sample(col_name)")

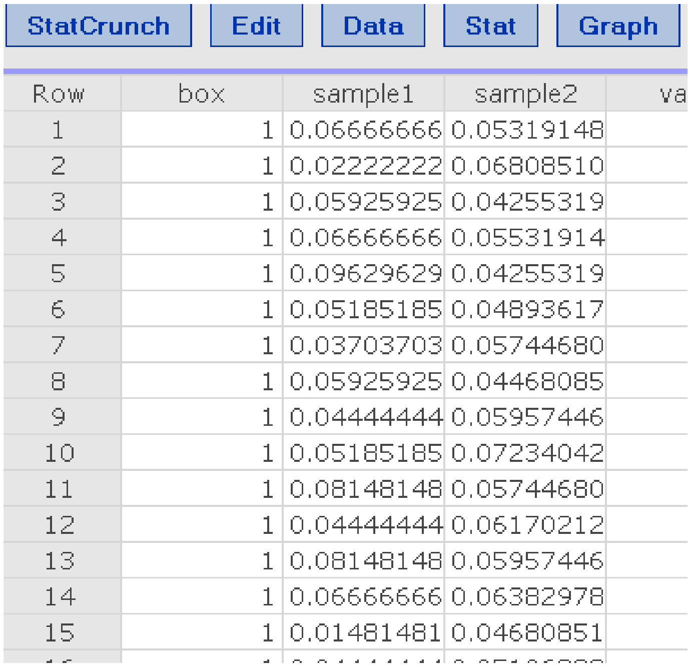
Column name(s):
Prefix: sample1

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 8-36: Sample columns dialog for sample 1

3) Repeat step 2 for the second sample. Select the box column, enter 470 for the Sample size, check Sample with replacement, use 1000 for the Number of samples, and enter the expression: sum("Sample(box)")/470 for the Statistic expression. Name the new column sample2 and click Compute!

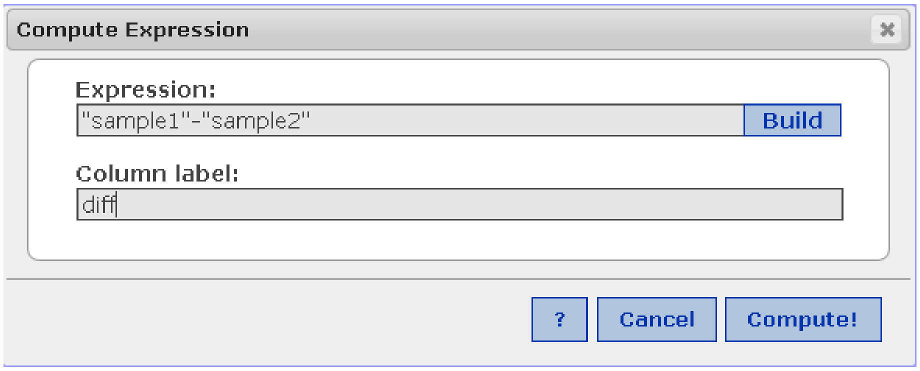
4) Your worksheet should look something like figure 8-37.



Row	box	sample1	sample2
1	1	0.06666666	0.05319148
2	1	0.02222222	0.06808510
3	1	0.05925925	0.04255319
4	1	0.06666666	0.05531914
5	1	0.09629629	0.04255319
6	1	0.05185185	0.04893617
7	1	0.03703703	0.05744680
8	1	0.05925925	0.04468085
9	1	0.04444444	0.05957446
10	1	0.05185185	0.07234042
11	1	0.08148148	0.05744680
12	1	0.04444444	0.06170212
13	1	0.08148148	0.05957446
14	1	0.06666666	0.06382978
15	1	0.01481481	0.04680851

Figure 8-37: Sample proportions

5) To calculate the differences in the proportions, select Data>Compute Expression and enter the expression: "sample2"-"sample1" and enter diff as the column label (figure 8-38).



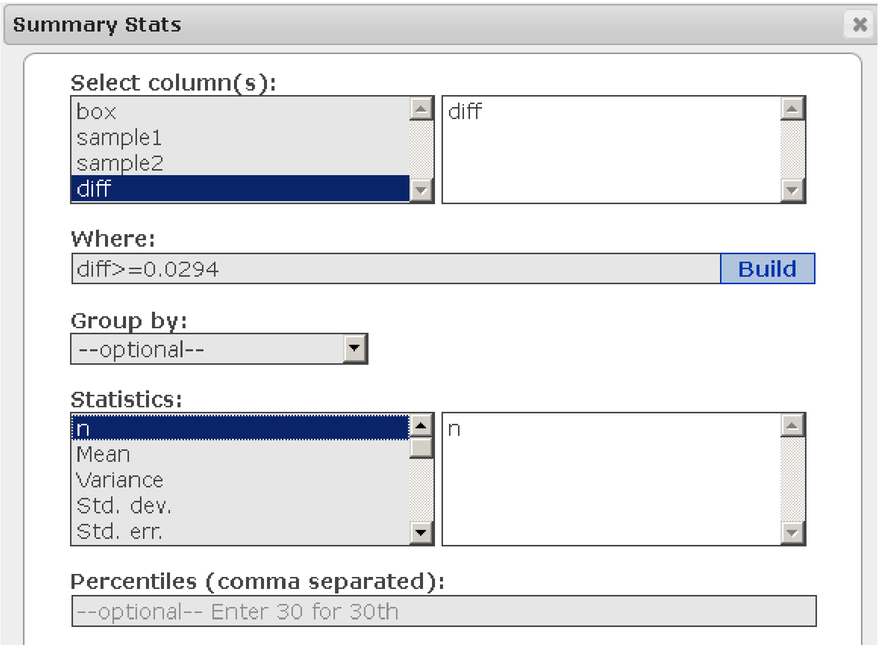
Expression: "sample1"-"sample2"

Column label: diff

Figure 8-38: Compute expression for the difference in proportions

6) To calculate the estimated p-value from the difference in proportions, choose Stat>Summary Stats>Columns and select the diff column. In the Where box,

enter: diff>=0.0294 to calculate the number of instances where the difference in proportions was >=0.0294. Select n for the statistic and click Compute! (figure 8-39).



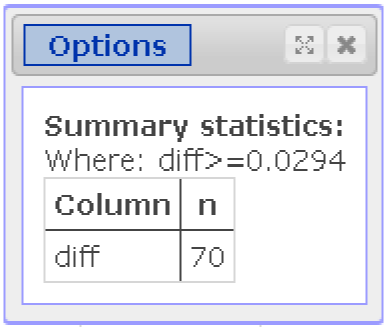
Select column(s): diff

Where: diff>=0.0294

Statistics: n

Figure 8-39: Finding n

7) In this experiment, there were 70 instances where the difference in proportions was >=0.0294. The estimated p-value is 70/1000 or 0.070 (figure 8-40).



Summary statistics:
Where: diff>=0.0294

Column	n
diff	70

Figure 8-40: Estimated p-value = 0.070

8.7 Paired Comparisons

The resampling procedures for the first paired comparisons test are similar to the resampling procedures in chapter 7.

Resampling Stats (Reading Scores example)

1) Enter all 22 reading score values in cells A1:A22.

2) Select cell A1, and choose Shuffle (without replacement). Enter cell C1 as the Top Cell and 11 as the number of Output cells. Click OK.

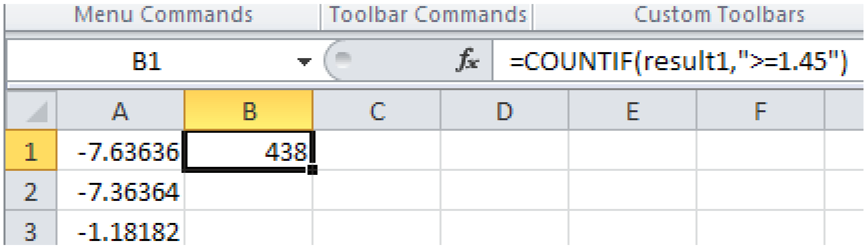
3) Repeat step 2, using cell D1 as the Output cell. Click OK.

4) In cell C12, enter =AVERAGE(C1:C11) and in cell D12, enter =AVERAGE(D1:D11).

5) In cell E12, enter =D12 – C12 to find the difference in means.

6) Select cell E12 (the score cell) and choose Repeat and Score. Enter 1000 iterations and click OK.

7) On the Results sheet, enter =COUNTIF(result1,”>=1.45”) in cell B1.

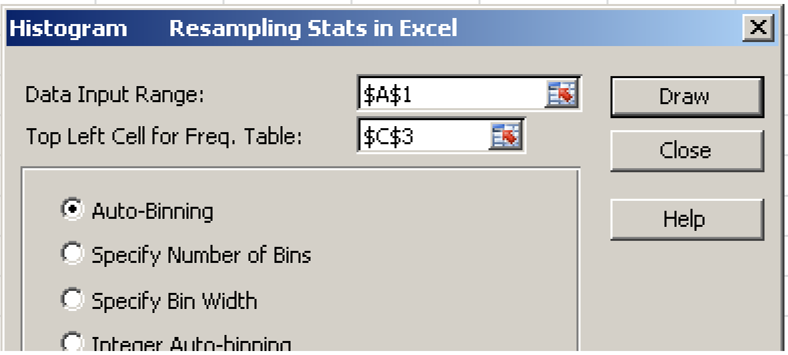


B1 =COUNTIF(result1,">=1.45")

	A	B	C	D	E	F
1	-7.63636	438				
2	-7.36364					
3	-1.18182					

Figure 8-40: Estimated p-value = 0.438

8) Create a histogram by selecting cell A1 (on the Results sheet) and choosing the Histogram feature (figure 8-41).



Data Input Range: $A$1

Top Left Cell for Freq. Table: $C$3

● Auto-Binning

○ Specify Number of Bins

○ Specify Bin Width

○ Integer Auto-binning

Figure 8-41: Histogram dialog

9) Click Draw. How does the resulting histogram compare to figure 8-8 in the textbook?

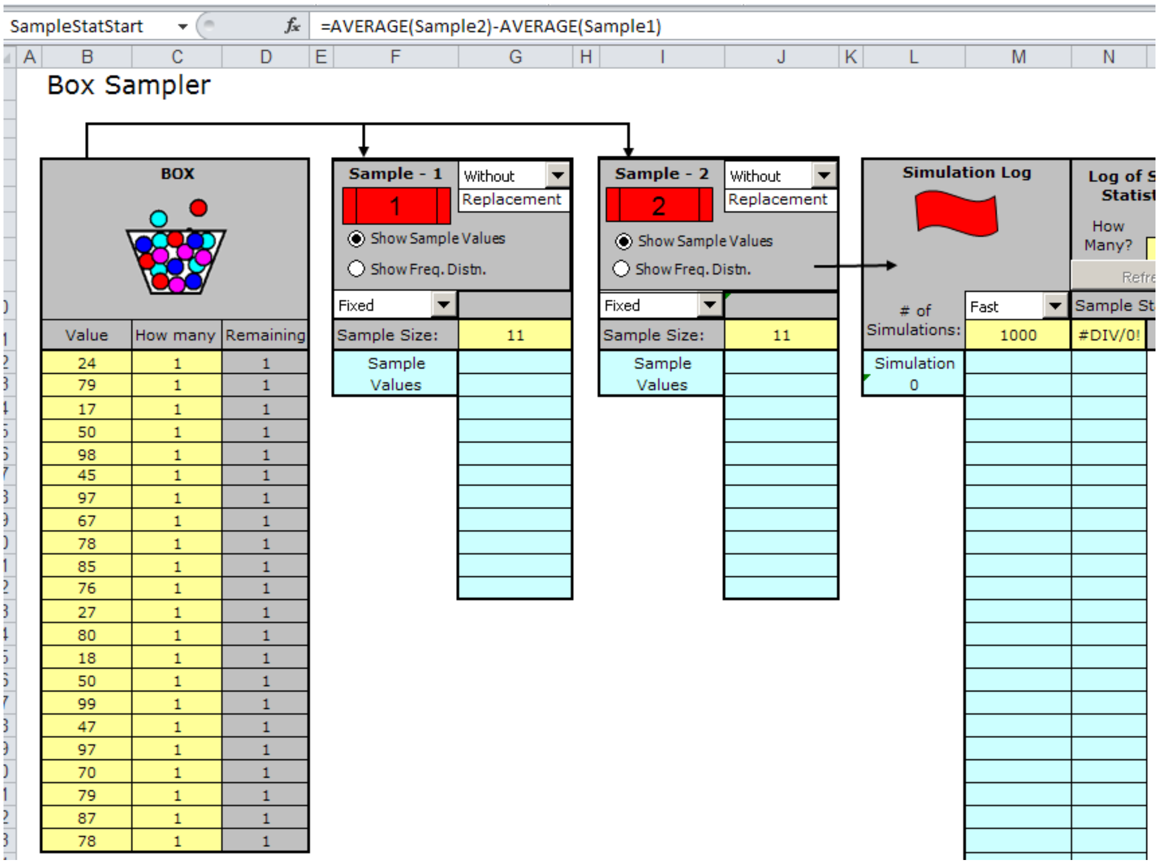
Box Sampler

1) As in section 8.3, we will need to select Two Samplers from the initial Box Sampler dialog, keeping the other settings at the default.

2) Enter the 22 reading score values, choose 11 for the Sample Size in both Samples, select Without Replacement, use 1000 for the # of Simulations, and enter:

   =AVERAGE(Sample2)-AVERAGE(Sample1)

   in cell N11 (the Statistic cell). See figure 8-42.



SampleStatStart

=AVERAGE(Sample2)-AVERAGE(Sample1)

Box Sampler

BOX

Value	How many	Remaining
24	1	1
79	1	1
17	1	1
50	1	1
98	1	1
45	1	1
97	1	1
67	1	1
78	1	1
85	1	1
76	1	1
27	1	1
80	1	1
18	1	1
50	1	1
99	1	1
47	1	1
97	1	1
70	1	1
79	1	1
87	1	1
78	1	1

Sample - 1: Without Replacement
● Show Sample Values
○ Show Freq. Distn.
Fixed
Sample Size: 11

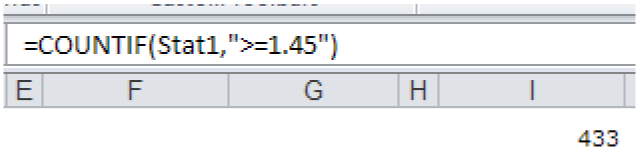
Sample - 2: Without Replacement
● Show Sample Values
○ Show Freq. Distn.
Fixed
Sample Size: 11

Simulation Log
Fast
# of Simulations: 1000
Simulation 0

Log of S… Statist…
How Many?
Sample St…
#DIV/0!

Figure 8-42: Box Sampler setup

3) In an empty cell, enter =COUNTIF(Stat1,”>=1.45”) to see how many times the difference in mean reading scores was >=1.45 (figure 8-43). You may also create a histogram to compare to figure 8-8 in the text.



=COUNTIF(Stat1,">=1.45")

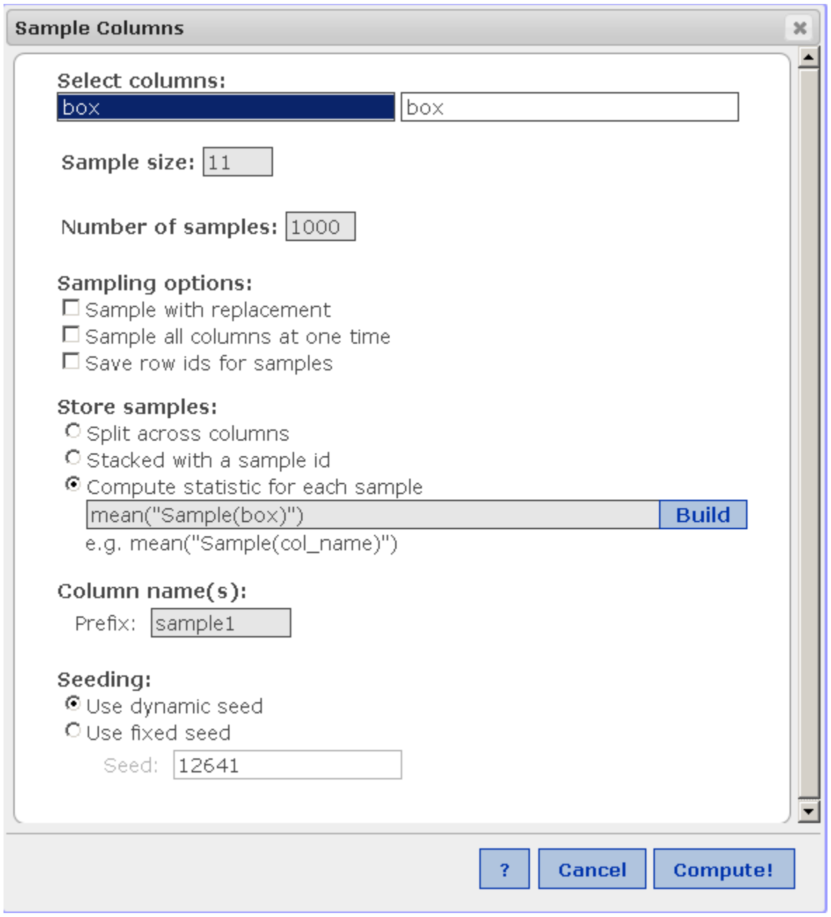
433

Figure 8-43: Estimated p-value = 0.433

StatCrunch

1) Enter or load the reading scores in the first column and rename the column “box” or something similar.

2) Select Data>Sample and complete the dialog as shown in figure 8-4.



Sample Columns

Select columns: box

Sample size: 11

Number of samples: 1000

Sampling options:

☐ Sample with replacement

☐ Sample all columns at one time

☐ Save row ids for samples

Store samples:

○ Split across columns

○ Stacked with a sample id

● Compute statistic for each sample

mean("Sample(box)")

e.g. mean("Sample(col_name)")

Column name(s):

Prefix: sample1

Seeding:

● Use dynamic seed

○ Use fixed seed

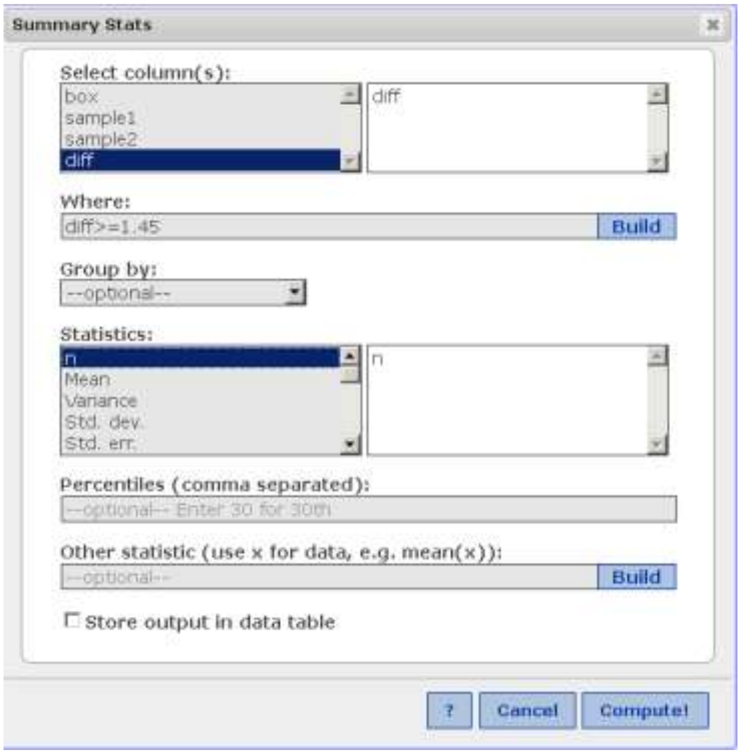
Seed: 12641

Figure 8-44: Reading scores Sample Columns

3) Repeat step 2 for the second sample and name that column sample2. Click Compute!

4) Select Data>Compute Expression and enter the expression “sample2” – “sample1” again, remember to include the quotes! The new column label should be diff or something similar. Click Compute!

5) Choose Stat>Summary Stats>Columns, select the diff column, and in the Where box, enter diff>=1.45 as shown in figure 8-45.



Select column(s): diff

Where: diff>=1.45

Group by: --optional--

Statistics: n

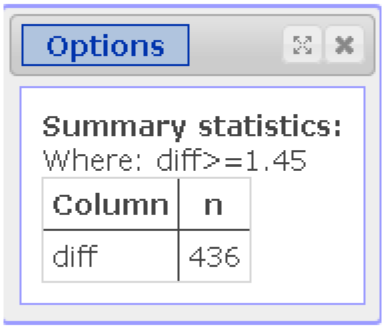
Percentiles (comma separated): --optional-- Enter 30 for 30th

Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 8-45: Summary stats

6) Click Compute! to find the result



Summary statistics:
Where: diff>=1.45

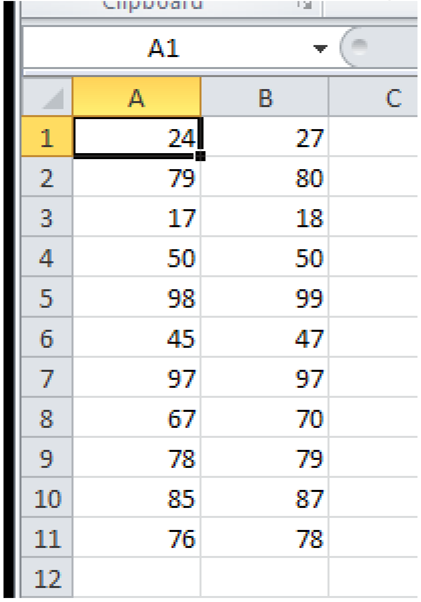
Column	n
diff	436

Figure 8-46: Estimated p-value = 0.436

7) The estimated p-value is 436/1000 or 0.436. This level of difference is not rare.

Paired Differences – Resampling Stats

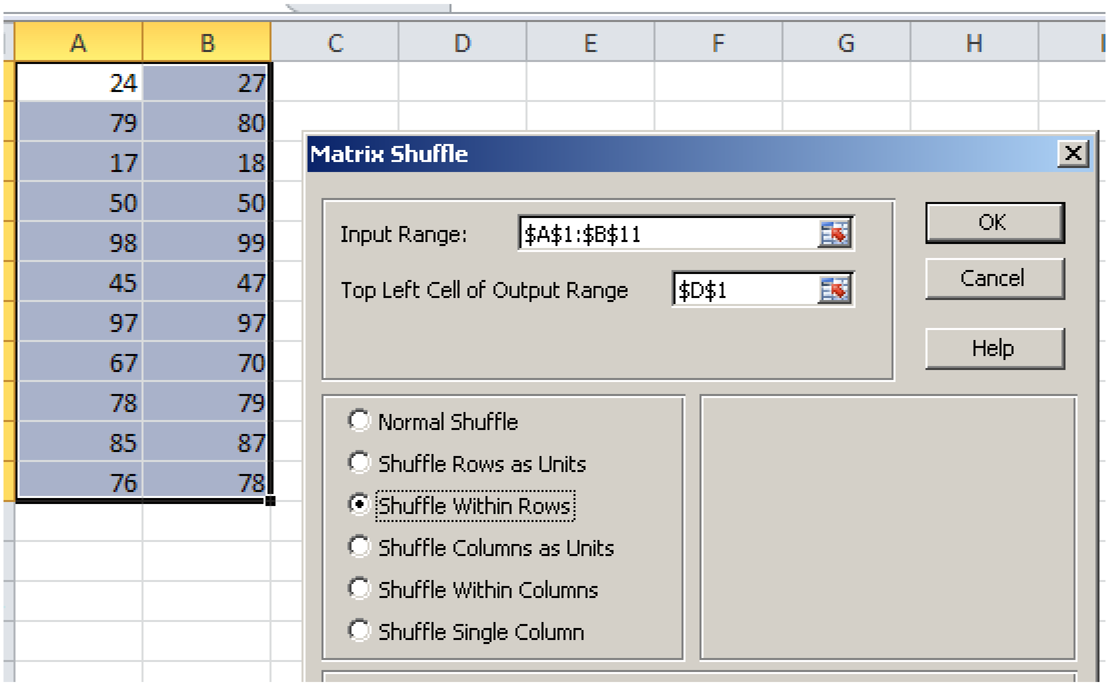
1) Enter the paired data in the worksheet as in table figure 8-47.



	A	B	C
1	24	27	
2	79	80	
3	17	18	
4	50	50	
5	98	99	
6	45	47	
7	97	97	
8	67	70	
9	78	79	
10	85	87	
11	76	78	
12			

Figure 8-47: Paired data

2) Select cell A1 and choose Shuffle (without replacement). We do NOT want to resample with replacement because we need to maintain the integrity of both scores for each subject. Notice that the entire data set is highlighted and more options are available. Choose cell D1 as the top left cell and select the Shuffle Within Rows option (figure 8-48).



A	B
24	27
79	80
17	18
50	50
98	99
45	47
97	97
67	70
78	79
85	87
76	78

Input Range: $A$1:$B$11

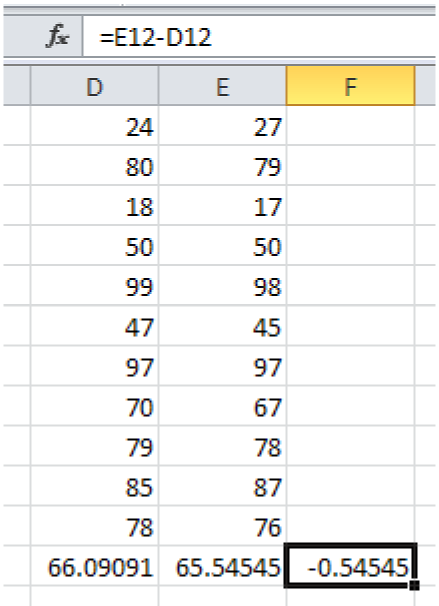
Top Left Cell of Output Range $D$1

○ Normal Shuffle
○ Shuffle Rows as Units
● Shuffle Within Rows
○ Shuffle Columns as Units
○ Shuffle Within Columns
○ Shuffle Single Column

Figure 8-48: Shuffle Within Rows

3) The Shuffle Within Rows option will keep each subject's set of scores intact and randomly shuffle the two scores. Click OK.

4) In cell D12, enter the formula =AVERAGE(D1:D11) and in cell E12, enter the formula =AVERAGE(E1:E12). Enter the formula =E12-D12 in cell F12.



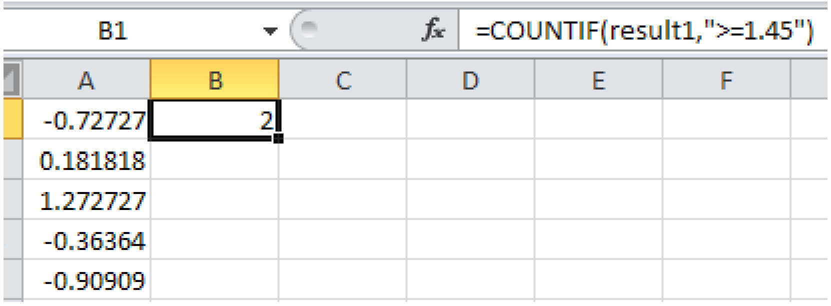
fx =E12-D12

D	E	F
24	27	
80	79	
18	17	
50	50	
99	98	
47	45	
97	97	
70	67	
79	78	
85	87	
78	76	
66.09091	65.54545	-0.54545

Figure 8-49: Within Rows resampling: difference in means

5) Select cell F12 and choose Repeat and Score. Enter 1000 iterations and click OK.

6) On the Results sheet, enter the formula =COUNTIF(result1,”>=1.45”) in cell B1. The outcome of one experiment yielded 2 instances where the difference in means was >=1.45, which is an estimated p-value =0.002 (figure 8-50).



B1

=COUNTIF(result1,">=1.45")

A	B	C	D	E	F
-0.72727	2				
0.181818					
1.272727					
-0.36364					
-0.90909					

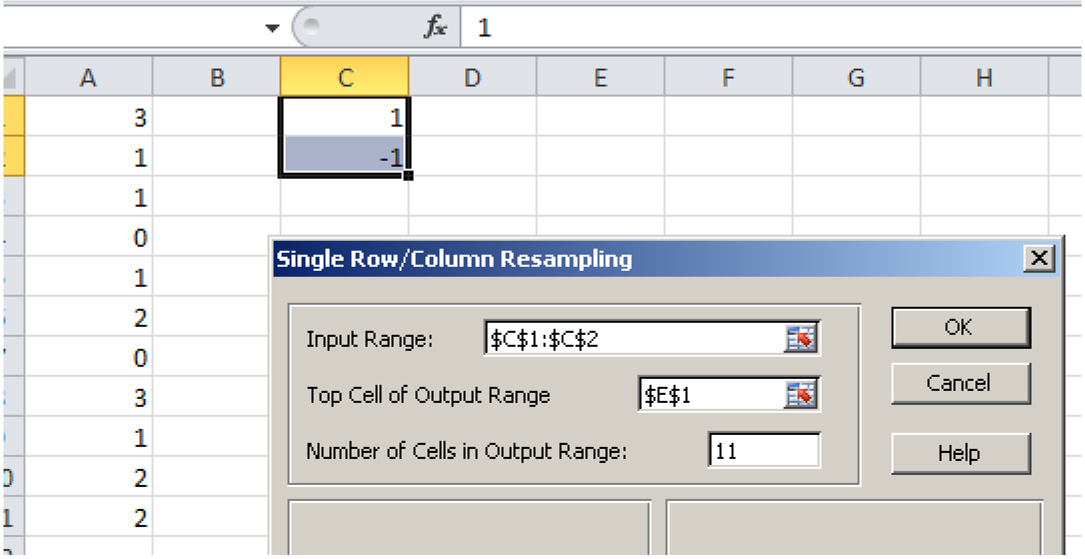
Figure 8-50: Estimated p-value = 0.002

7) Create a histogram of the results and compare your histogram to figure 8-8 in the text.

Resampling Stats – Alternate Method

1) A second method for analyzing the data would involve looking at the differences in the scores and resampling those differences using random alternating signs. From table 8-4 in the text, enter the differences in reading scores in cells A1:A11. In cell C1 and C2, enter 1 and -1 respectively.

2) Select cell C1 and choose Resample. Use cell E1 for the Top Left cell and enter 11 for the number of cells (figure 8-51). Click OK.



fx 1

A	B	C
3		1
1		-1
1		
0		
1		
2		
0		
3		
1		
2		
2		

Single Row/Column Resampling

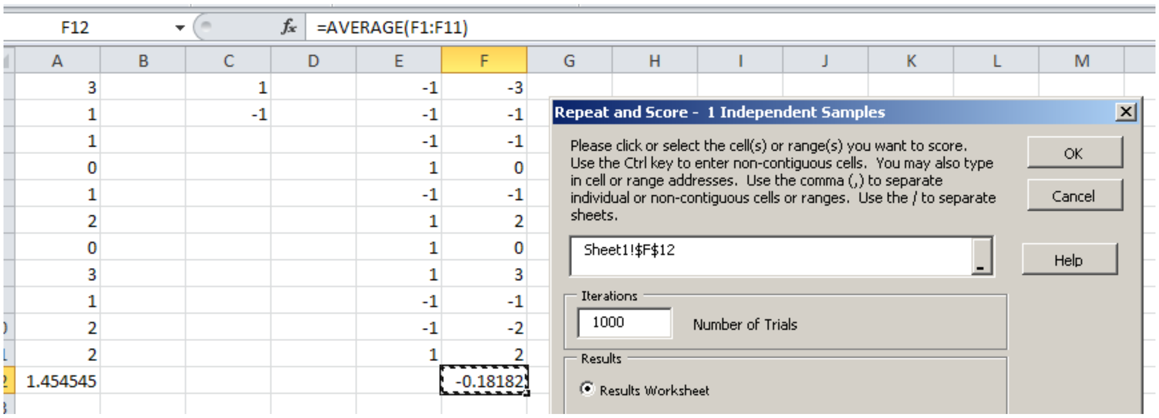
Input Range: $C$1:$C$2

Top Cell of Output Range $E$1

Number of Cells in Output Range: 11

Figure 8-51: Paired comparisons alternate method

3) In cells E1:E11 you should see a random set of 1’s and -1’s. We are going to use this 1/-1 resample to change the signs of the differences in reading scores in cells A1:A11. This operation will have the same effect as the first method when we Shuffled within rows.
4) We need to multiply each value in A1:A11 by the corresponding 1 or -1 in cells E1:E11 and then take the average of these values. We can do this by entering the formula =A1*E1 in cell F1 and copying this formula down the column. In cell F12, take the average of the column by entering the formula =AVERAGE(F1:F11). Cell F12 is our score cell, so select F12 and choose Repeat and Score. Enter 1000 iterations and click OK (figure 8-52).



F12 fx =AVERAGE(F1:F11)

A	B	C	D	E	F
3		1		-1	-3
1		-1		-1	-1
1				-1	-1
0				1	0
1				-1	-1
2				1	2
0				1	0
3				1	3
1				-1	-1
2				-1	-2
2				1	2
1.454545					-0.18182

Repeat and Score - 1 Independent Samples

Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

Sheet1!$F$12

Iterations: 1000 Number of Trials

Results: ● Results Worksheet

Figure 8-52: Repeat and Score paired comparisons alternate method

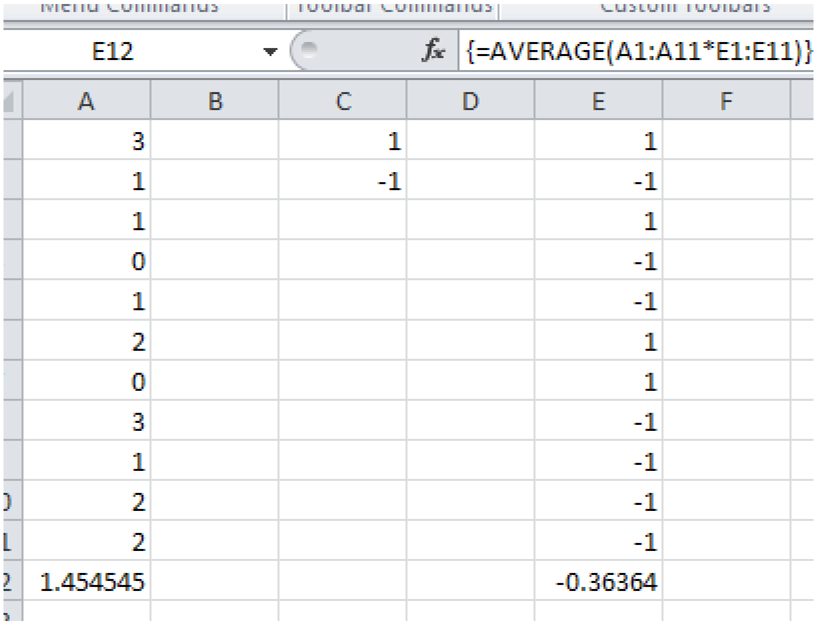
5) On the Results worksheet, use =COUNTIF(result1,”>=1.45”) to find the number of instances where the mean of the differences is >=1.45. Note the mean of the original differences in reading scores in cell A12 = 1.45. Create a histogram if you wish.

Resampling Stats – Formula Array

1) Recreate steps 1-3 in the previous method.

2) In cell E12, enter the following formula BUT DON'T PRESS ENTER: =AVERAGE(A1:A11*E1:E11)

3) When you have typed in the formula, press SHIFT+CTRL+ENTER… all three keys simultaneously, and release. You should see a value appear in cell E12 and the formula in the formula bar will look something like {=AVERAGE(A1:A11*E1:E11} (figure 8-53).



E12 — *fx* {=AVERAGE(A1:A11*E1:E11)}

A	B	C	D	E	F
3		1		1	
1		-1		-1	
1				1	
0				-1	
1				-1	
2				1	
0				1	
3				-1	
1				-1	
2				-1	
2				-1	
1.454545				-0.36364	

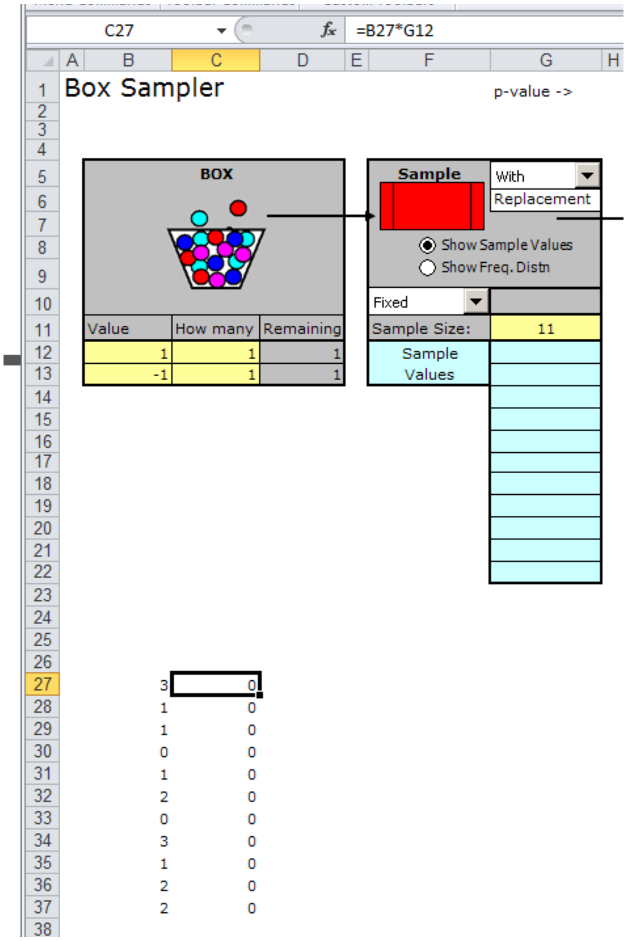
Figure 8-53: Formula array

4) This bracketed function is called a formula array and it serves to multiply the entire range of A1:A11 by the corresponding values in E1:E11. The average of this array multiplication is then calculated. Formula arrays are a very powerful (and advanced) feature of Excel. Cell E12 is the score cell, so select it and choose Repeat and Score. Again, enter 1000 iterations and click OK. If the =COUNTIF(result1,">=1.45") function is not in cell B1, enter this formula to estimate the p-value.

Box Sampler

1) Open Box Sampler and use 1 and -1 for the box contents (1 each). Enter the differences in means in cells B27:B37. In cells C27 enter the formula =B27*G12

and copy this cell down to cell C37. The Excel formulas should automatically adjust so that cell C28 has the formula =B28*G13, cell C27, cell C29 has the formula =B29*G14, etc. until cell C37, which should have the formula =B37*G22. See figure 8-54.



C27 =B27*G12

Box Sampler

p-value ->

BOX

Value	How many	Remaining
1	1	1
-1	1	1

Sample

With Replacement

● Show Sample Values
○ Show Freq. Distn

Fixed

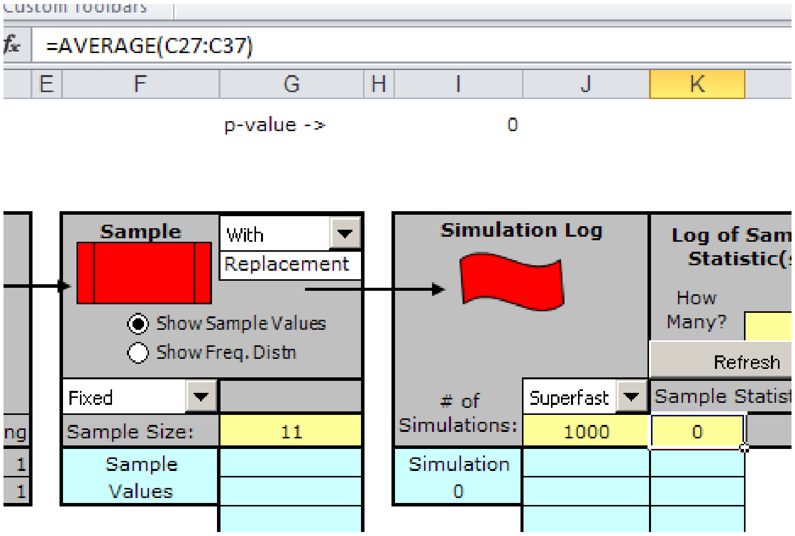
Sample Size: 11

Sample Values

Row	B	C
27	3	0
28	1	0
29	1	0
30	0	0
31	1	0
32	2	0
33	0	0
34	3	0
35	1	0
36	2	0
37	2	0

Figure 8-54: Paired differences

2) As shown in figure 8-54, use 11 for the Sample Size and 1000 for the # of Simulations (you probably should use the Superfast setting). In cell K12 (the Sample Statistics cell) enter the formula =AVERAGE(C27:C37) as shown in figure 8-55.



=AVERAGE(C27:C37)

p-value -> 0

With Replacement

● Show Sample Values
○ Show Freq. Distn

Fixed

Sample Size: 11

Sample Values

How Many?

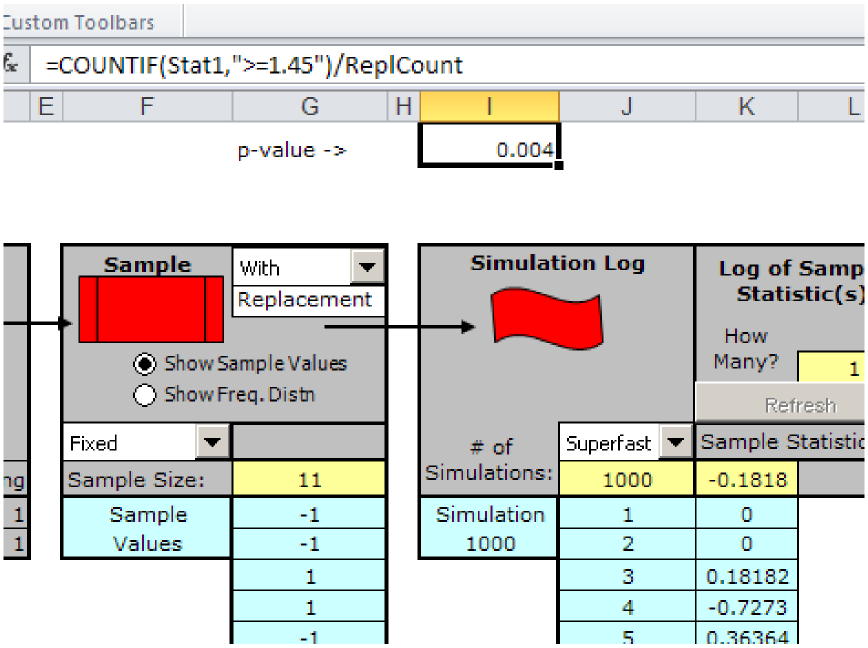
# of Simulations: Superfast 1000

Sample Statist 0

Simulation 0

Figure 8-55: Sample statistics cell formula

3) To calculate the p-value, enter =COUNTIF(Stat1, ”>=1.45”)/ReplCount in cell I1 (see figure 8-56). ReplCount is the name for the # of Simulations cell and equals 1000 in this model. Run the simulation to calculate the p-value. Figure 8-56 shows the results from one experiment.



=COUNTIF(Stat1,">=1.45")/ReplCount

p-value -> 0.004

With Replacement

● Show Sample Values
○ Show Freq. Distn

Fixed

Sample Size: 11

Sample Values: -1, -1, 1, 1, -1

How Many? 1

# of Simulations: Superfast 1000

Sample Statistic -0.1818

Simulation 1000

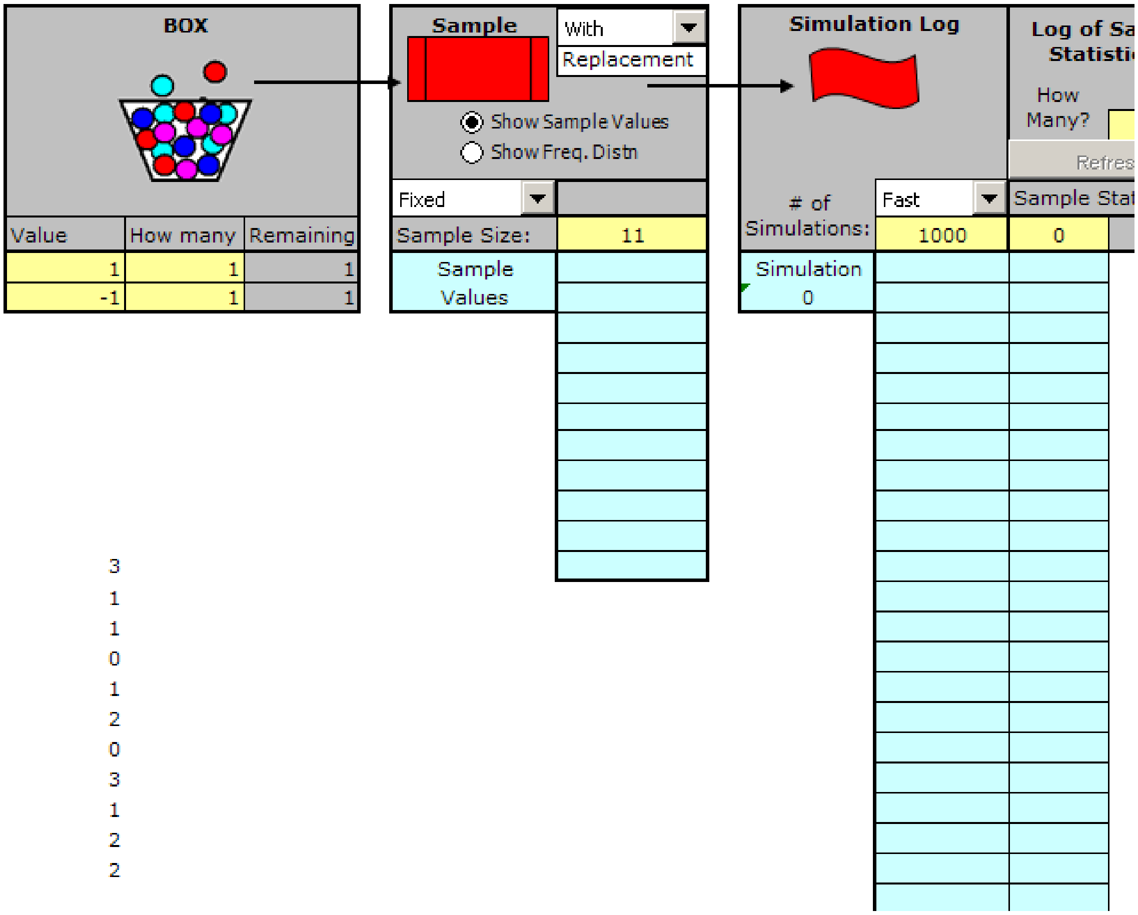
1	0
2	0
3	0.18182
4	-0.7273
5	0.36364

Figure 8-56: Results of one experiment, p=0.004

Box Sampler – Alternate Method

1) We can use the idea behind the alternate methods for Resampling Stats to solve this exercise in Box Sampler. The setup is a bit different than usual. After

loading Box Sampler and selecting the default setup, enter the differences in reading scores from Table 8-4 in an area below the Box Sampler box. In figure 8-57, we use cells B22:B32. The two values 1 and -1 are placed in the box, the Sample Size is 11, and the Sample Statistic in cell K11 is the formula array: {=AVERAGE(B22:B32*Sample)} Remember: Do NOT type the braces. Enter the formula and press SHIFT+CTRL+ENTER to create the formula array.



BOX

Value	How many	Remaining
1	1	1
-1	1	1

Sample

With Replacement

● Show Sample Values

○ Show Freq. Distn

Fixed

Sample Size: 11

Sample Values

Simulation Log

Fast

# of Simulations: 1000

Sample Stat: 0

Simulation 0

3
1
1
0
1
2
0
3
1
2
2

Figure 8-57: Box Sampler setup paired comparisons

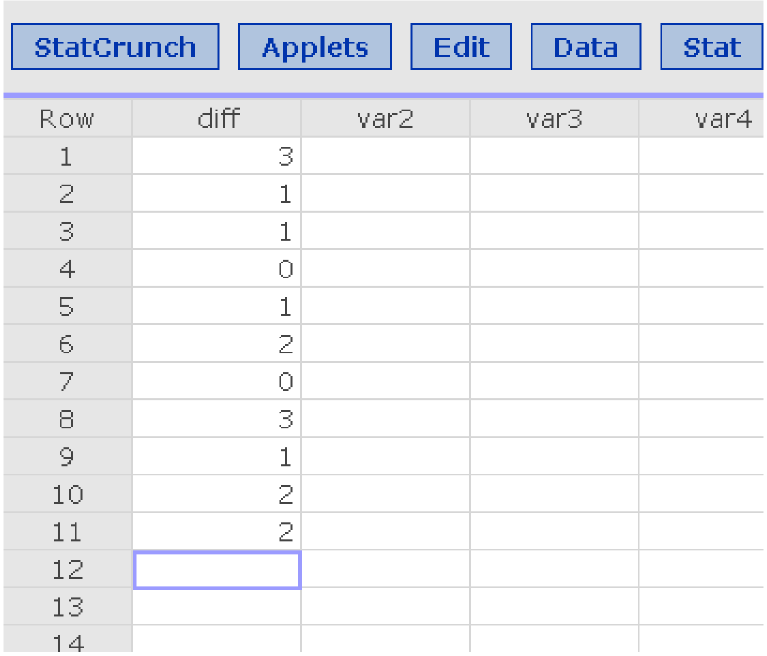
2) Click Simulate. When the simulation is finished, enter =COUNTIF(Stat1,”>=1.45”) in an empty cell to estimate the p-value.

StatCrunch

Note: Thanks to Dr. Webster West, the author of StatCrunch, for the expression used in this StatCrunch procedure.

1) Similar to Box Sampler, we will use the idea from the Resampling Stats alternative methods to solve the exercise in StatCrunch. Enter the differences

(absolute values) in reading scores in the first column and rename the column "diff" as shown in figure 8-58.



Row	diff	var2	var3	var4
1	3			
2	1			
3	1			
4	0			
5	1			
6	2			
7	0			
8	3			
9	1			
10	2			
11	2			
12				
13				
14				

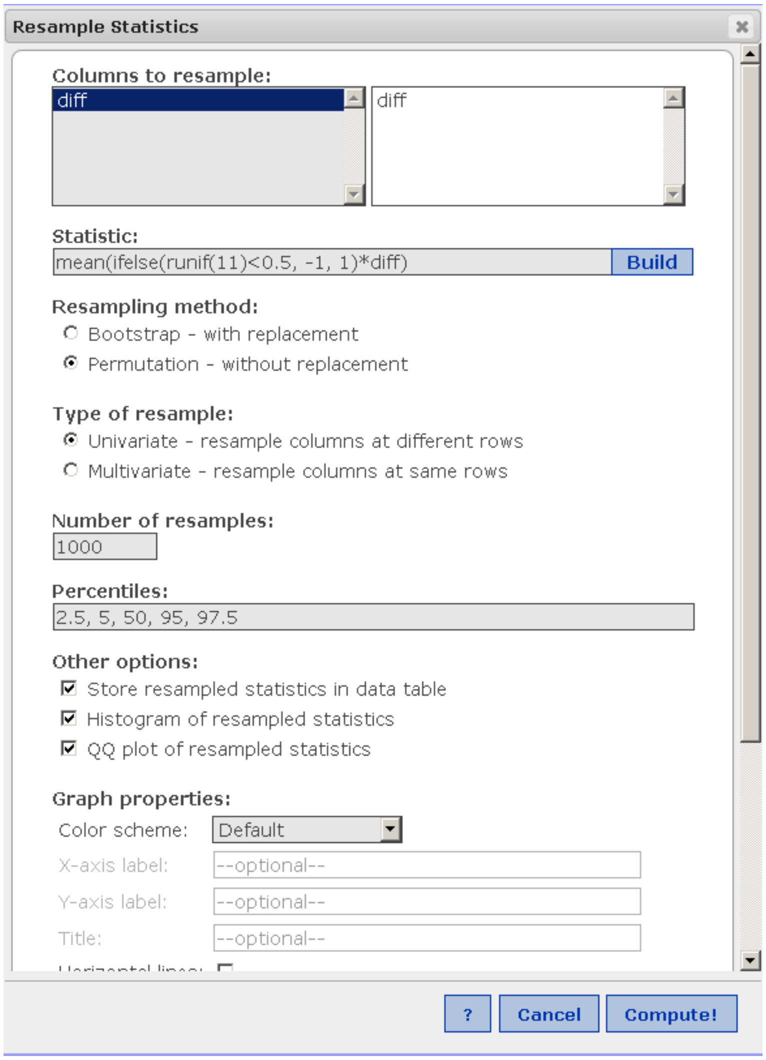
Figure 8-58: StatCrunch setup for reading score differences

2) Select Stat>Resample>Statistic and choose the diff column. The somewhat complex Statistic expression is (again, thanks to Dr. Webster West!):

   mean(ifelse(runif(11) < 0.5, -1, 1)*diff)

   Before we go further, let's look at this expression from the inside out. The runif(11) function produces 11 random numbers from the uniform distribution (uniformly distributed random real numbers between 0 and 1). The next layer, the ifelse conditional statement, checks to see which (if any) of the random numbers are < 0.5. If a number is < 0.5, then the ifelse conditional returns a -1. If not (or the 'else' case), the ifelse returns a 1. This is exactly equivalent to a coin toss and the result is a virtual (hidden) column of random 1s and -1s, which is exactly what we need! Then, the virtual column of 11 random 1s and -1s are multiplied by the diff column, which produces another virtual column of paired differences that are randomly positive or negative. Finally, the mean of the 11 paired differences is found using the mean function. **Note:** Don't worry if you find this function a bit too complicated… it IS complicated. At the end of this section, I'll show you how to dissect the function to make it a bit more understandable.

3) Select Permutation - without replacement, keep the default number of resamples and check the Store resampled statistics in data table in order to calculate the p-value. When everything matches figure 8-59, click Compute!



Columns to resample: diff | diff

Statistic: mean(ifelse(runif(11)<0.5, -1, 1)*diff)

Resampling method:
○ Bootstrap - with replacement
● Permutation - without replacement

Type of resample:
● Univariate - resample columns at different rows
○ Multivariate - resample columns at same rows

Number of resamples: 1000

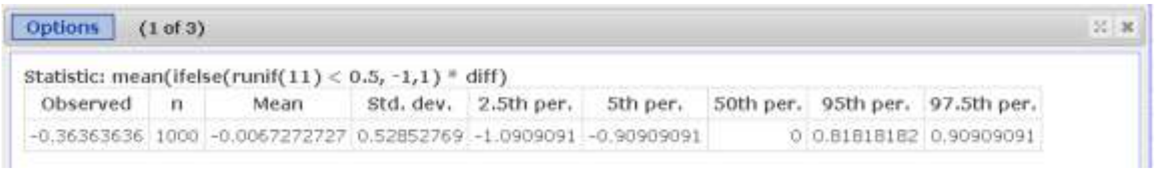
Percentiles: 2.5, 5, 50, 95, 97.5

Other options:
☑ Store resampled statistics in data table
☑ Histogram of resampled statistics
☑ QQ plot of resampled statistics

Graph properties:
Color scheme: Default
X-axis label: --optional--
Y-axis label: --optional--
Title: --optional--
Horizontal lines: ☐

Figure 8-59: Resample Statistic paired comparison

4) The confidence intervals are shown in figure 8-60.



Statistic: mean(ifelse(runif(11) < 0.5, -1,1) * diff)

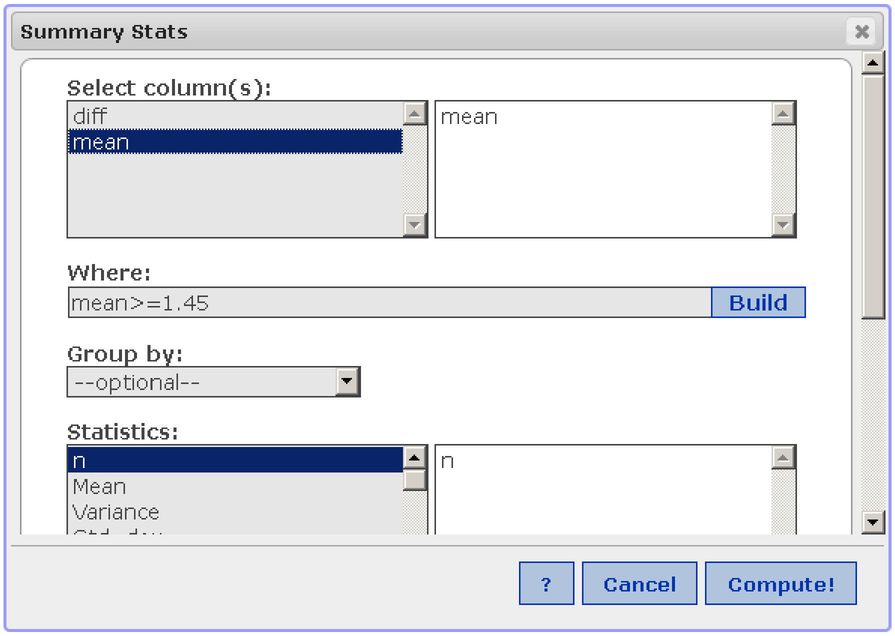
Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
-0.36363636	1000	-0.0067272727	0.52852769	-1.0909091	-0.90909091	0	0.81818182	0.90909091

Figure 8-60: Results and percentiles

5) The mean difference in scores of 1.45 in the original data is extreme. It is well outside the 97.5th percentile.

6) To calculate an estimated p-value, rename the new column ‘mean’ or something similar (for convenience). Select Stat>Summary Stats>Columns and choose the

new column (mean in this example). Enter the expression: mean>=1.45 in the Where edit box and choose n in the Statistics box. See figure 8-61. Click Compute!



Summary Stats

Select column(s): mean

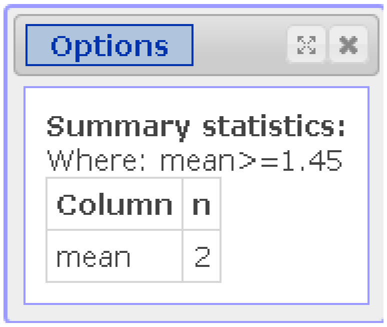
Where: mean>=1.45

Group by: --optional--

Statistics: n

Figure 8-61: Summary Stats setup to find p-value

7) In this example, there were only 2 instances in 1000 trials where the difference in means was >= the original mean of 1.45 (figure 8-62). This gives us an estimated p-value of 0.002.



Summary statistics:
Where: mean>=1.45

Column	n
mean	2

Figure 8-62: Estimated p-value = 0.002

Using Compute Expression to Understand a StatCrunch Function

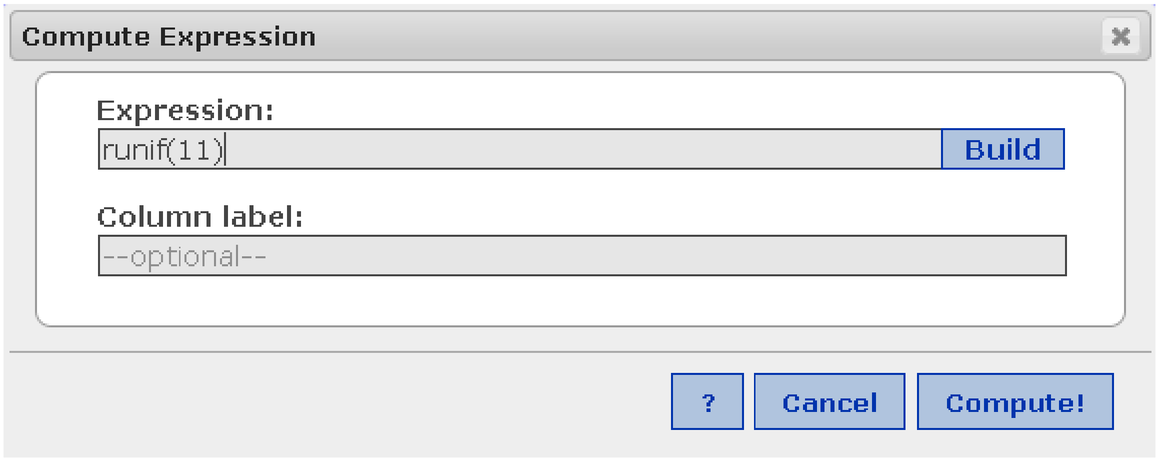
The StatCrunch function:

mean(ifelse(runif(11) < 0.5, -1, 1)*diff) may be more understandable if we can examine the expression more closely using the Compute Expression dialog.

Here are the steps. Remember that the following calculations are for illustrative purposes and the goal is to dissect and understand a complicated StatCrunch expression.

1) Enter the paired differences in the first column of a new worksheet and use diff as the column label.

2) Choose Data>Compute Expression and enter runif(11) in the Expression box. Click Compute! (figure 8-63). Notice the results consist of 11 random real values between 0 and 1. These values are from a uniform distribution where every value in this range has an equal chance of being chosen.



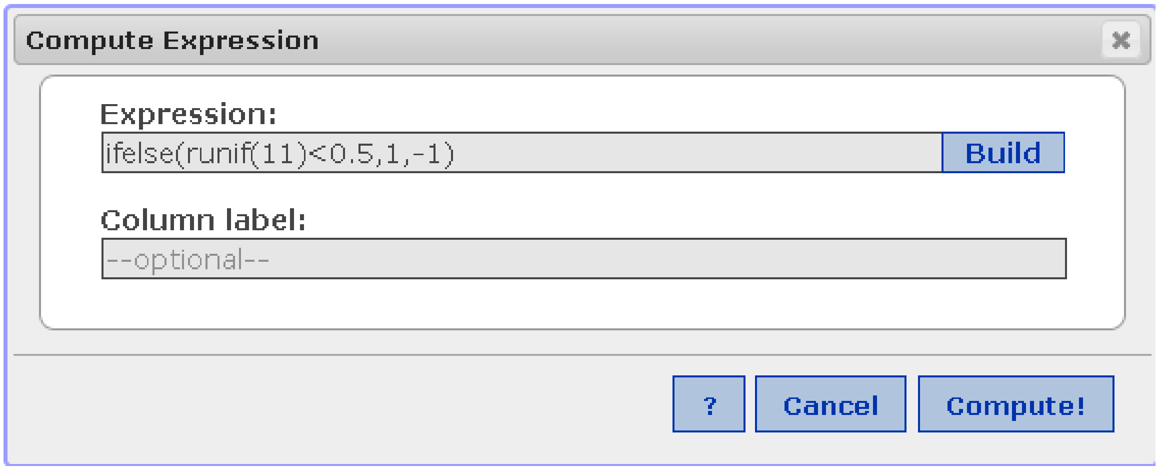
Compute Expression

Expression: runif(11)

Column label: --optional--

Figure 8-63: Create 11 random values from the uniform distribution

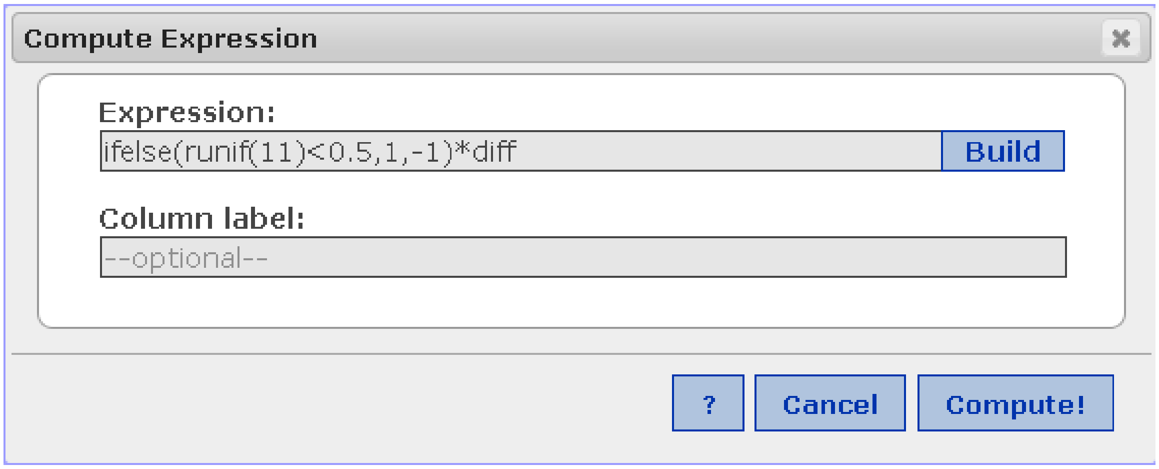
3) Now let’s try this again, except this time we’ll add the ifelse() function. Choose Data>Compute Expression and enter ifelse(runif(11) <0.5, -1, 1) in the Expression box and click Compute! (figure 8-64). You should see a new column consisting of random 1s and -1s. This new column is completely independent of the previous calculation, so don’t worry if the 1s and -1s don’t align with the column of random values in the second column.



ifelse(runif(11)<0.5,1,-1)

Figure 8-64: Adding the ifelse statement

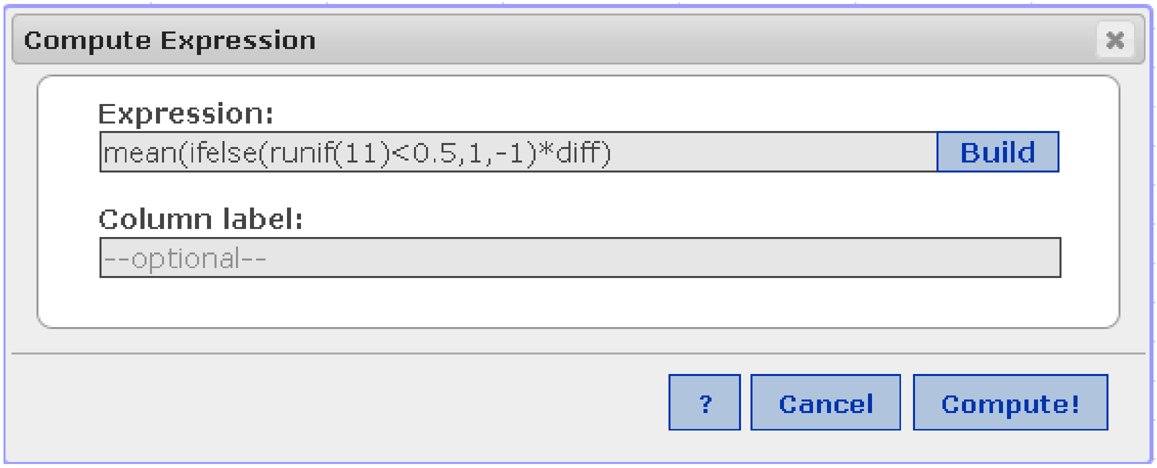
4) Now, we'll add the *diff operation. Once again, select the Data>Compute Expression dialog and enter ifelse(runif(11) <0.5, -1, 1)*diff in the Expression box and click Compute! (figure 8-65). You should see a new column with values that match the diff column with some random differences in sign. This value represents the mean of one single trial of the resampled paired differences.



ifelse(runif(11)<0.5,1,-1)*diff

Figure 8-65: Adding the *diff operation

5) Finally, we'll wrap everything in the mean() function. Select Data>Compute Expression and enter mean(ifelse(runif(11) < 0.5, 1, -1)*diff) in the Expression box and click Compute! (figure 8-66). You should see a single value in a new column. This value represents the mean difference in the paired comparisons in one trial. Remember that in the experiment in the previous section, we had StatCrunch perform 1000 trials!

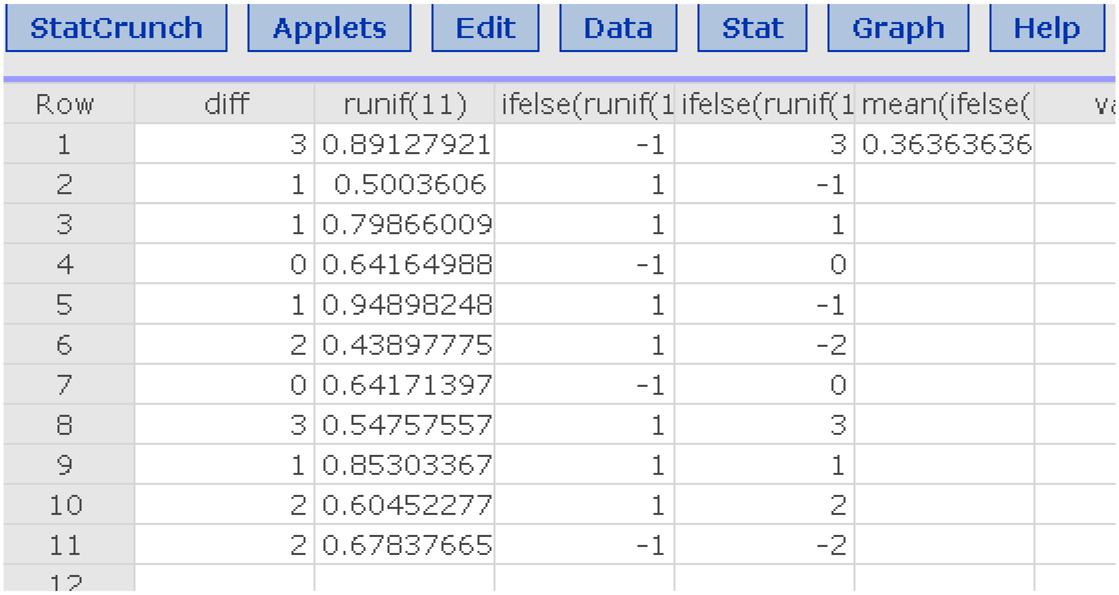


Expression: mean(ifelse(runif(11)<0.5,1,-1)*diff)

Column label: --optional--

Figure 8-66: Adding the mean function

6) Figure 8-67 illustrates the results of all of our expression building. Remember that the calculated columns are independent of each other and were used to illustrate how to dissect a complicated StatCrunch expression.



Row	diff	runif(11)	ifelse(runif(1	ifelse(runif(1	mean(ifelse(	va
1	3	0.89127921	-1	3	0.36363636	
2	1	0.5003606	1	-1		
3	1	0.79866009	1	1		
4	0	0.64164988	-1	0		
5	1	0.94898248	1	-1		
6	2	0.43897775	1	-2		
7	0	0.64171397	-1	0		
8	3	0.54757557	1	3		
9	1	0.85303367	1	1		
10	2	0.60452277	1	2		
11	2	0.67837665	-1	-2		
12						

Figure 8-67: Results of a systematic expression dissection in StatCrunch

9 Hypothesis Testing - 2

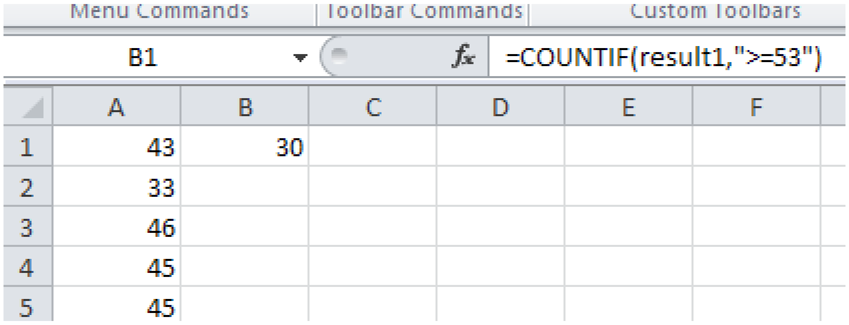
9.1 A Single Proportion

Resampling Stats

1) Enter 3 0's in cells A1:A3. Enter a 1 in cell A4.

2) Select cell A1 and choose Resample. Let cell C1 be the Top Cell of the output range and enter 165 for the number of output cells. Click OK.

3) Enter =SUM(C1:C165) in cell D1. Select cell D1 and choose Repeat and Score. Enter 1000 for the number of iterations (trials). Click OK.



=COUNTIF(result1,">=53")

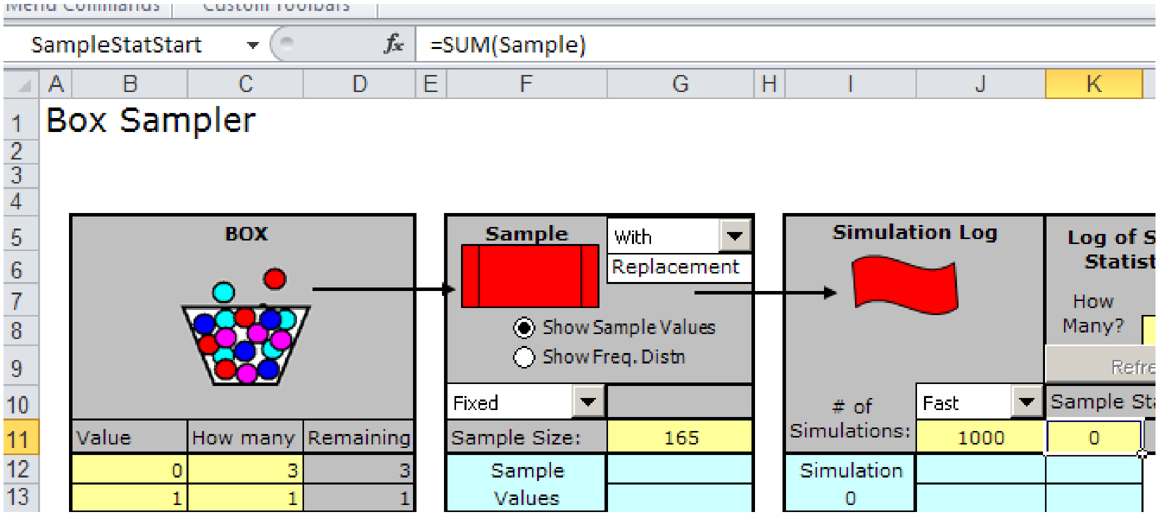
	A	B	C	D	E	F
1	43	30				
2	33					
3	46					
4	45					
5	45					

Figure 9-1: Gym full membership rate

4) The p-value estimate in figure 9-1 is 0.03. Create a histogram and compare it to figure 9-1 in the text.

Box Sampler

1) Select New Model from the Box Sampler menu and accept the default settings. Enter the values shown in Figure 9-2 and enter the formula =SUM(Sample) in cell K11 (the Sample Statistics cell). Note the "How Many" cells used to specify 3 0's and 1 instance of 1 in the Box.



Value	How many	Remaining
0	3	3
1	1	1

Figure 9-2: Box Sampler setup

2) Click Simulate and when the simulation has ended (it will take some time, so be patient and/or change the simulation speed to SuperFast), enter the formula =COUNTIF(Stat1,”>=53”) in an empty cell (say, cell G1). See figure 9-3 for the results of one experiment.

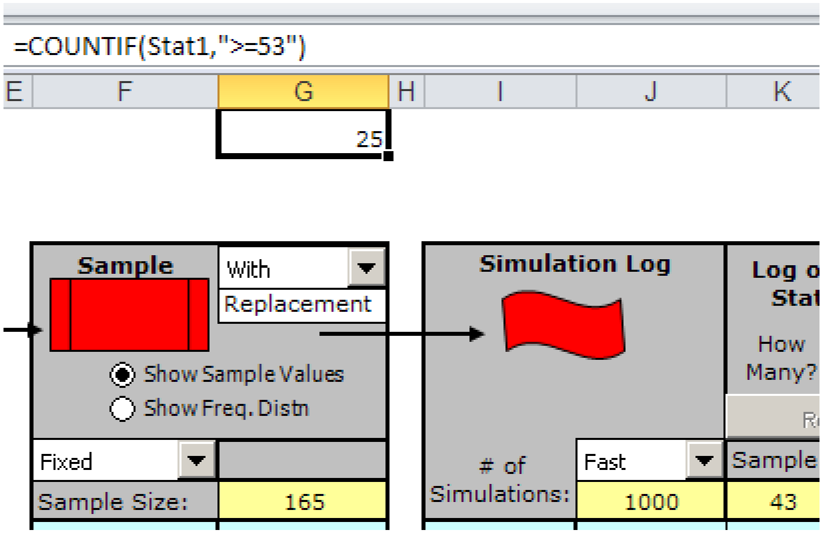


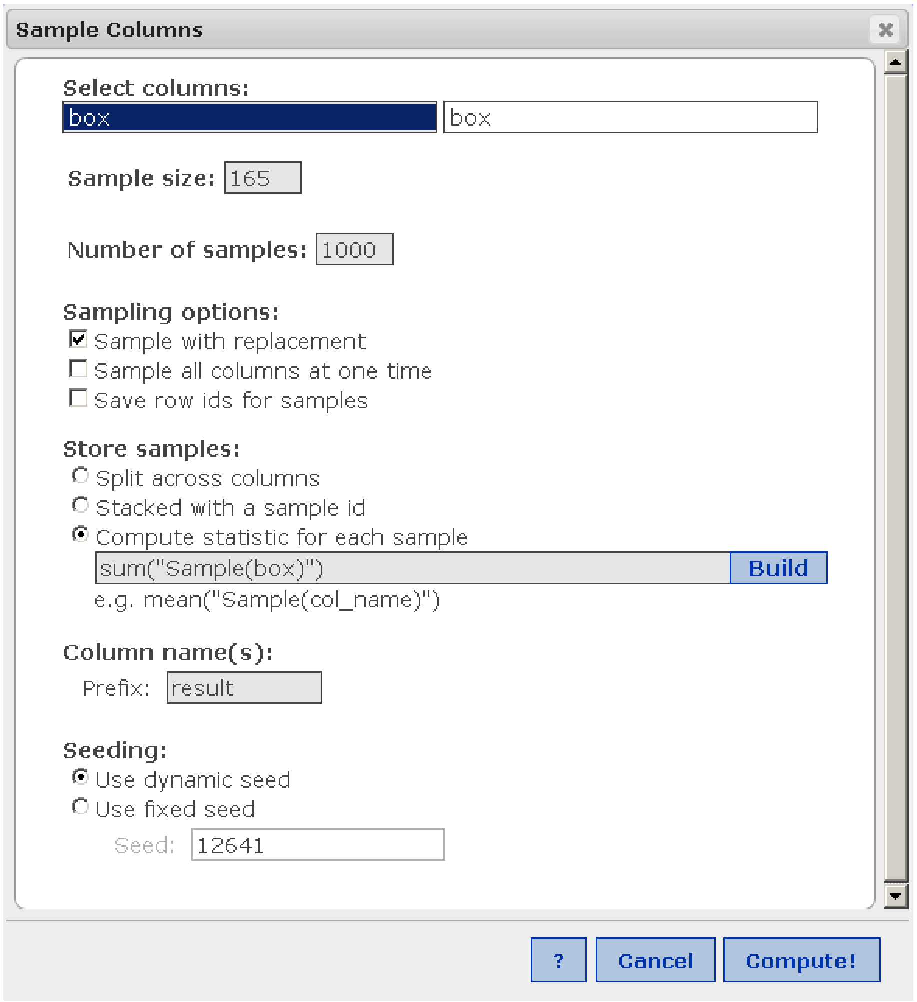
Figure 9-3: Estimated p-value = 25

StatCrunch

1) Enter 3 0s and a 1 in the first column. Rename this column “box”.

2) Select Data>Sample and choose the box column. Enter a Sample size of 165, check Sample with replacement, and enter 1000 for the Number of samples. The

Statistic expression is: sum("Sample(box)"). Label the new column "result" and click Compute!



Select columns: box | box

Sample size: 165

Number of samples: 1000

Sampling options:

☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

Store samples:

○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample

sum("Sample(box)")

Column name(s):

Prefix: result

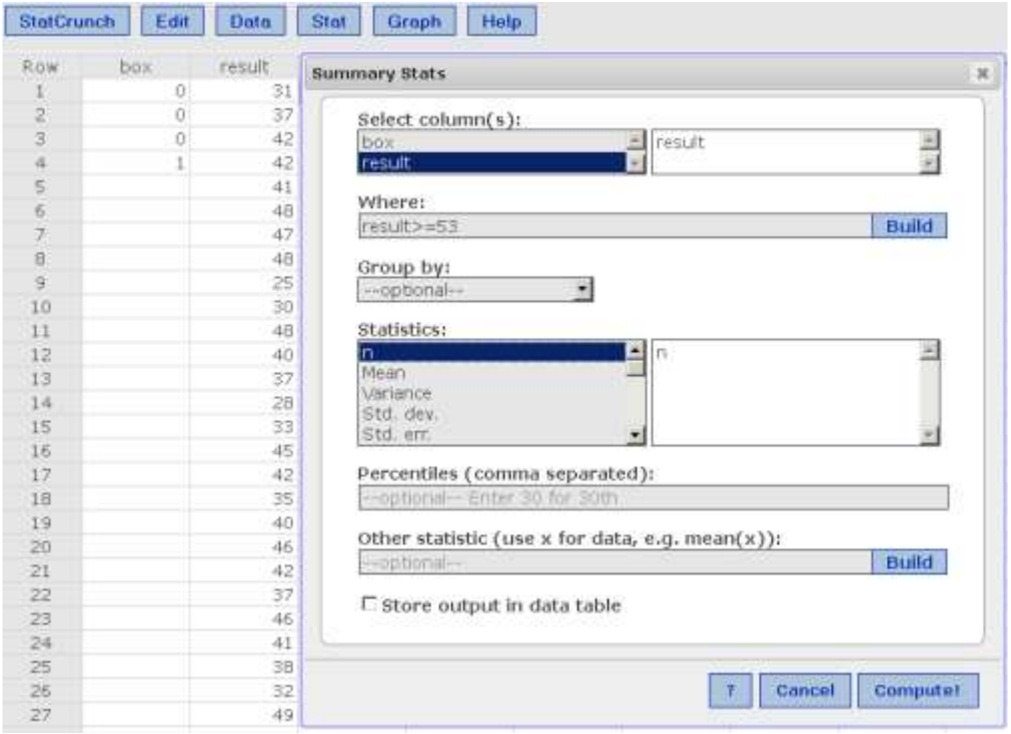
Seeding:

● Use dynamic seed
○ Use fixed seed

Seed: 12641

Figure 9-4: Sample columns dialog

3) Select Stat>Summary Stats>Columns and select the result column. In the Where box, enter: result>=53 and select n in the Statistics box (figure 9-5). Click Compute!



Row	box	result
1	0	31
2	0	37
3	0	42
4	1	42
5		41
6		48
7		47
8		48
9		25
10		30
11		48
12		40
13		37
14		28
15		33
16		45
17		42
18		35
19		40
20		46
21		42
22		37
23		46
24		41
25		38
26		32
27		49

Select column(s): box, result — result

Where: result>=53

Group by: --optional--

Statistics: n, Mean, Variance, Std. dev., Std. err. — n

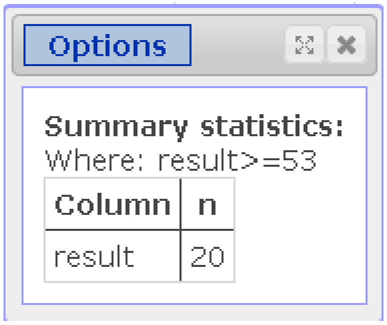
Percentiles (comma separated): --optional-- Enter 30 for 30th

Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 9-5: Column Statistics

4) The result of one experiment is shown in figure 9-6



Summary statistics:
Where: result>=53

Column	n
result	20

Figure 9-6: StatCrunch gym membership result

5) The estimated p-value = 0.020 is similar to Resampling Stats and Box Sampler.

9.2 A Single Mean

Resampling Stats Procedure for the Confidence Interval

1) Enter the moisture output values in cells A1:A12

2) With cell A1 highlighted, select Resample. Choose cell C1 for the top cell of the output range and 12 as the number of cells in the output range.

3) In cell D1, use the Excel formula =AVERAGE(C1:C12) to calculate the mean of each resample.

4) With cell D1 selected, choose Repeat and Score and use 1000 trials.

5) Sort the data in column A of the Results sheet. The 90% confidence can be calculated from the sorted data (the values in cells 50 and 950) or by using the =PERCENTILE() function. The 90% interval from one experiment was from 13.52 to 14.17.

6) The histogram for the output is shown in figure 9-7.

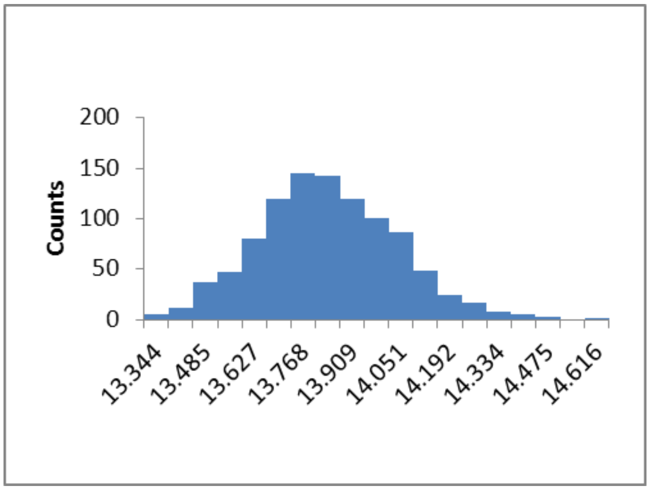
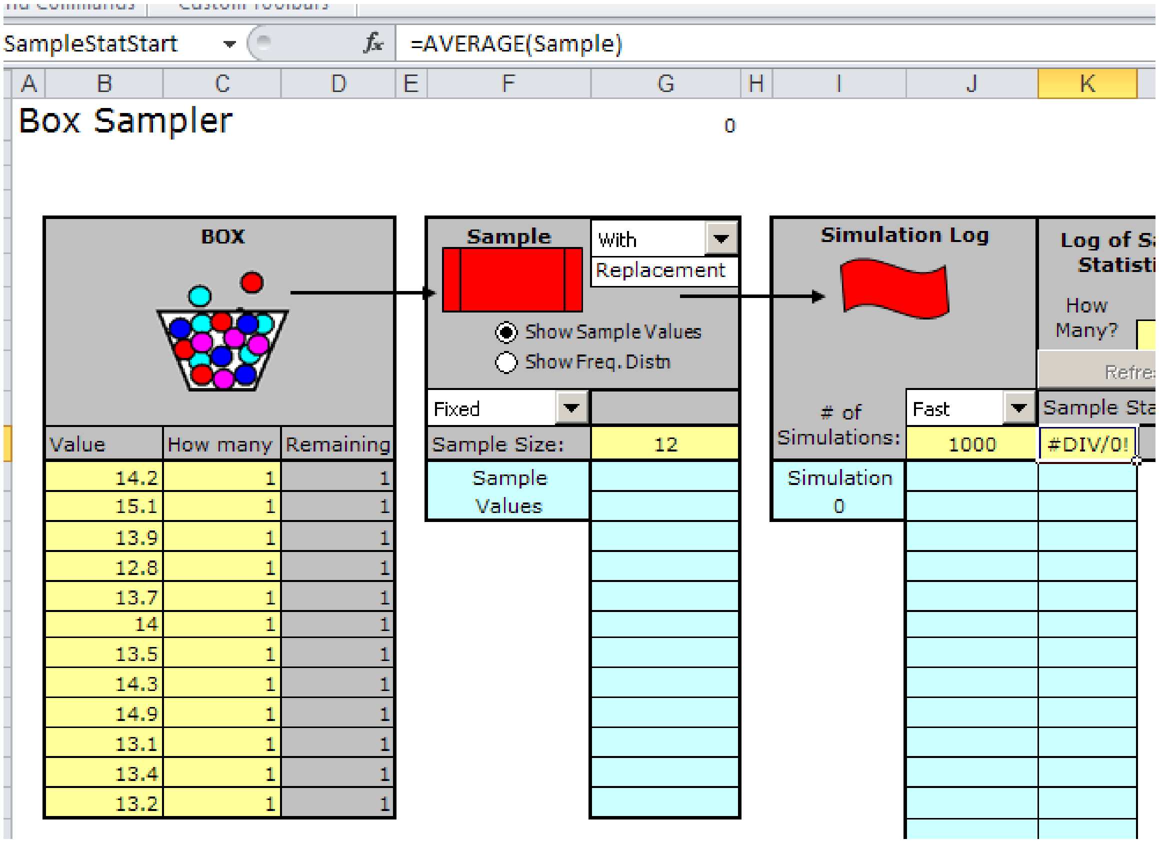


Figure 9-7: Histogram for resampling procedure: 14 is not rare

Box Sampler

1) Start Box Sampler and use the default options in the initial dialog. Figure 9-8 illustrates the setup for the initial Box Sampler setup. Enter the moisture output values in the Box, use 12 for the Sample Size, 1000 for the # of Simulations, and =AVERAGE(Sample) for the Sample Statistic. Click Simulate.



SampleStatStart

=AVERAGE(Sample)

Box Sampler

Value	How many	Remaining
14.2	1	1
15.1	1	1
13.9	1	1
12.8	1	1
13.7	1	1
14	1	1
13.5	1	1
14.3	1	1
14.9	1	1
13.1	1	1
13.4	1	1
13.2	1	1

Sample: With Replacement

● Show Sample Values

○ Show Freq. Distn

Fixed

Sample Size: 12

# of Simulations: 1000

Fast

#DIV/0!

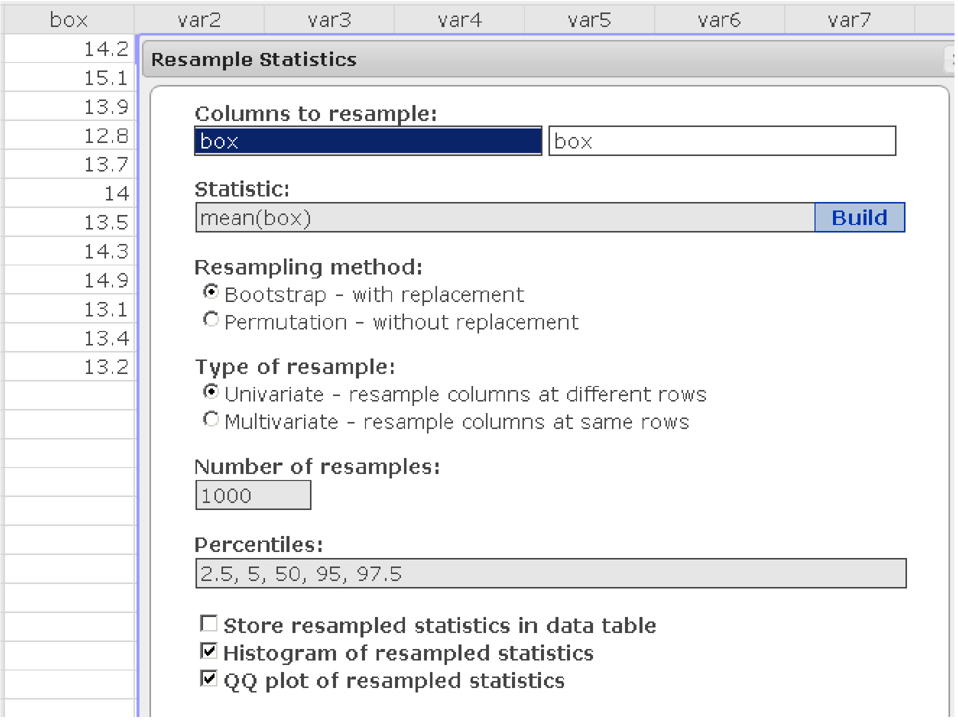
Figure 9-8: Box Sampler setup

2) Use the =PERCENTILE(Stat1, 0.05) and =PERCENTILE(Stat1, 0.95) functions to find the 90% confidence interval. One experiment with Box Sampler resulted in a 90% confidence interval of 13.53 to 14.15.

StatCrunch

1) Enter the moisture output data in the first column and rename the column “box”.

2) The resample and sample sizes are the same, so we can use the Stat>Resample>Statistic feature. Select box and enter: mean(box) for the Statistic. Keep the other default options and click Resample Statistic.



box
14.2
15.1
13.9
12.8
13.7
14
13.5
14.3
14.9
13.1
13.4
13.2

Columns to resample: box

Statistic: mean(box)

Resampling method:
● Bootstrap - with replacement
○ Permutation - without replacement

Type of resample:
● Univariate - resample columns at different rows
○ Multivariate - resample columns at same rows

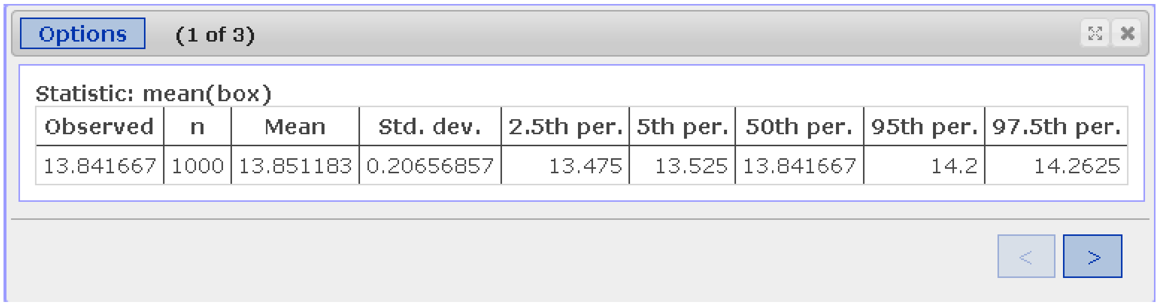
Number of resamples: 1000

Percentiles: 2.5, 5, 50, 95, 97.5

☐ Store resampled statistics in data table
☑ Histogram of resampled statistics
☑ QQ plot of resampled statistics

Figure 9-9: Resample Statistic

3) The results are shown in figure 9-10.



Statistic: mean(box)

Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
13.841667	1000	13.851183	0.20656857	13.475	13.525	13.841667	14.2	14.2625

Figure 9-10: StatCrunch 90% confidence interval: 13.525 – 14.2

4) Click > button to view the StatCrunch Histogram, resizing if needed (figure 9-11). The graphic clearly shows 14 near the middle of the histogram.

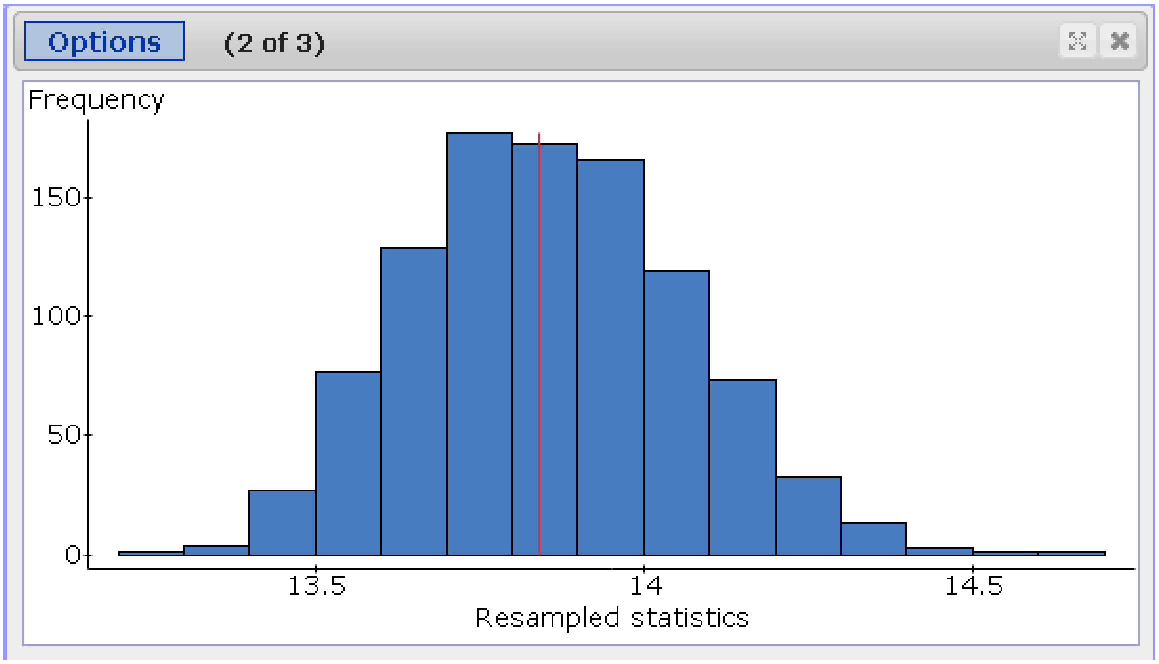
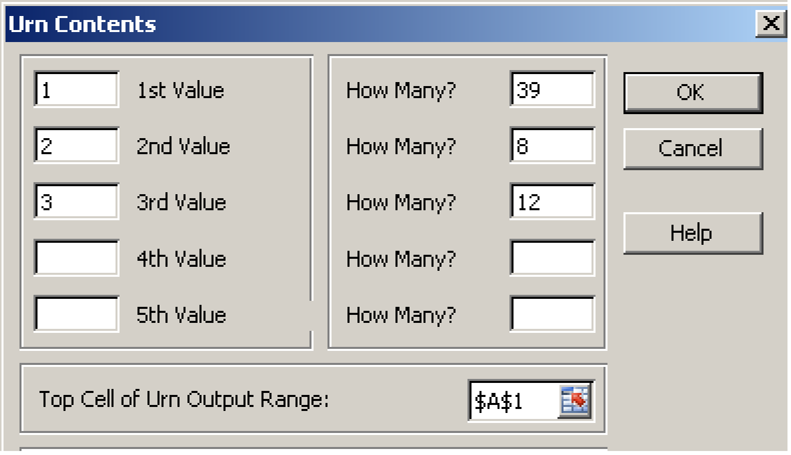


Figure 9-11: StatCrunch histogram

9.3 More than Two Categories or Samples

Resampling Stats

1) Use the Urn feature to create a box/urn containing 39 one's (for HM), 8 two's (for DM) and 12 three's (for D) (figure 9-12). Note: You could enter HM, DM, and D for the urn contents instead of numbers. The =COUNTIF statements in steps 3-5 below would then be modified as follows: =COUNTIF(C1:C29,"=HM"), etc.



1	1st Value	How Many?	39
2	2nd Value	How Many?	8
3	3rd Value	How Many?	12
	4th Value	How Many?	
	5th Value	How Many?	

Top Cell of Urn Output Range: $A$1

Figure 9-12: Resampling Stats Urn dialog

2) Select cell A1 and choose Shuffle from the Resampling Stats menu. Choose cell C1 as the Top Left Cell and keep the default of 59 for the Number of Cells. Click OK.

3) In cell E1, enter =COUNTIF(C1:C29,"=1") and in cell E2, enter =COUNTIF(C30:C59,"=1")

4) In cell F1, enter =COUNTIF(C1:C29,"=2") and in cell F2, enter =COUNTIF(C30:C59,"=2")

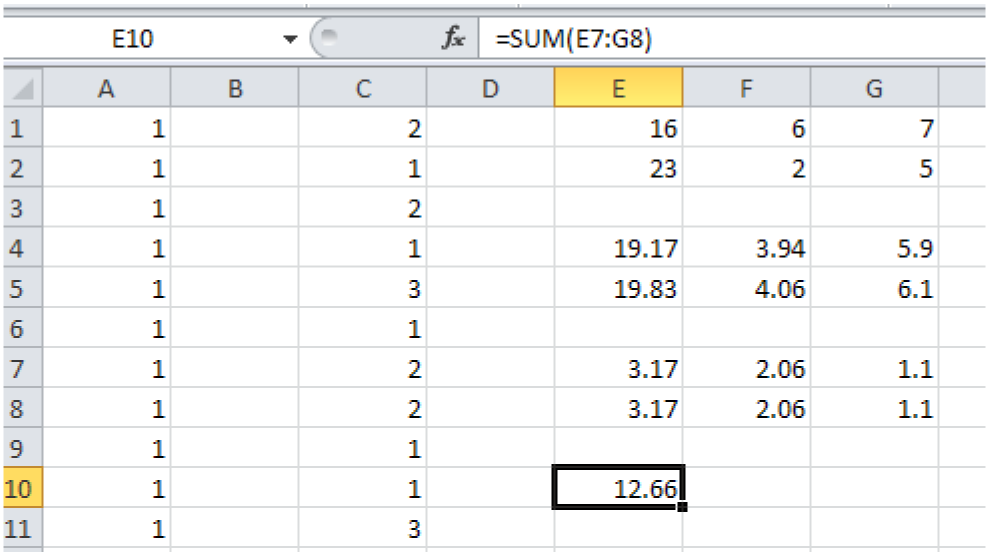
5) In cell G1, enter =COUNTIF(C1:C29,"=3") and in cell G2, enter =COUNTIF(C30:C59,"=3")

6) Enter the "expected table" in cells E4:G5.

7) Cells E7:G8 contain formulas that find the absolute difference between corresponding values in the two tables. In cell E7, enter the formula =ABS(E1-E4). You can then copy this formula down and to the right to fill the entire E7:G8 matrix. The relative cell addresses will automatically adjust during the copy procedure.

8) In cell E10, enter the formula: =SUM(E7:G8)

9) Figure 9-13 shows a sample worksheet at this point.



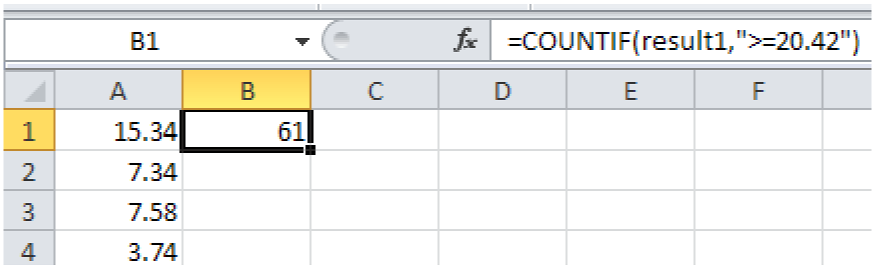
=SUM(E7:G8)

	A	B	C	D	E	F	G
1	1		2		16	6	7
2	1		1		23	2	5
3	1		2				
4	1		1		19.17	3.94	5.9
5	1		3		19.83	4.06	6.1
6	1		1				
7	1		2		3.17	2.06	1.1
8	1		2		3.17	2.06	1.1
9	1		1				
10	1		1		12.66		
11	1		3				

Figure 9-13: Sum of the absolute differences

10) Select cell E10 and choose Repeat and Score. Enter 10000 iterations/trials.

11) On the Results sheet, enter =COUNTIF(result1,">=20.42") in cell B1 (figure 9-14).


B1 =COUNTIF(result1,">=20.42")

	A	B	C	D	E	F
1	15.34	61				
2	7.34					
3	7.58					
4	3.74					

Figure 9-14: Estimated p-value = 61/10000 or 0.0061.

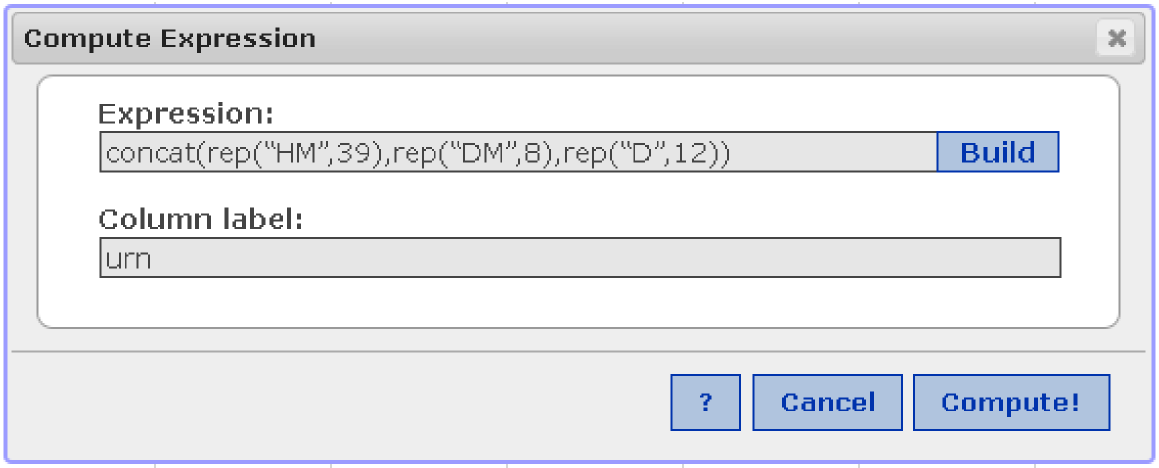
12) Create a histogram.

Statcrunch

1) Open StatCrunch and select “Data>Compute Expression” and enter (copy/paste) the following formula to populate the box or urn:

   concat(rep(“HM”,39),rep(“DM”,8),rep(“D”,12))

   This expression combines 39 repetitions of happily married (HM) couples with 8 distressed marriage (DM) couples and 12 divorced (D) couples in a single box or urn. Name the new column urn and click Compute! (see Figure 9-15).


Compute Expression

Expression: concat(rep(“HM”,39),rep(“DM”,8),rep(“D”,12))

Column label: urn

Figure 9-15: Compute expression dialog to populate the marriage urn

2) The second step involves using a rather complex formula that sums the absolute values of the differences between the groups. Select Data>Sample and choose urn as the column, enter 59 for the sample size, sample without replacement (do NOT check the sample with replacement box), and choose 10000 for the number of trials. Enter the following formula for the statistic. It is recommended that you copy/paste the formula (see Figure 9-16):

```
abs(sum(subset("Sample(urn)"=HM,Row<=29))-19.17) + abs(sum(subset("Sample(urn)"=DM,Row<=29))-
3.94) + abs(sum(subset("Sample(urn)"=D,Row<=29))-5.90) + abs(sum(subset("Sample(urn)"=HM,Row
>=30))-19.83) + abs(sum(subset("Sample(urn)"=DM,Row>=30))-4.06) +
abs(sum(subset("Sample(urn)"=D,Row>=30))-6.10)
```

The formula is long and complex, but it can be translated into English as follows (where # = “number of” or “sum of”):

In rows 0 – 29, do the following calculation:

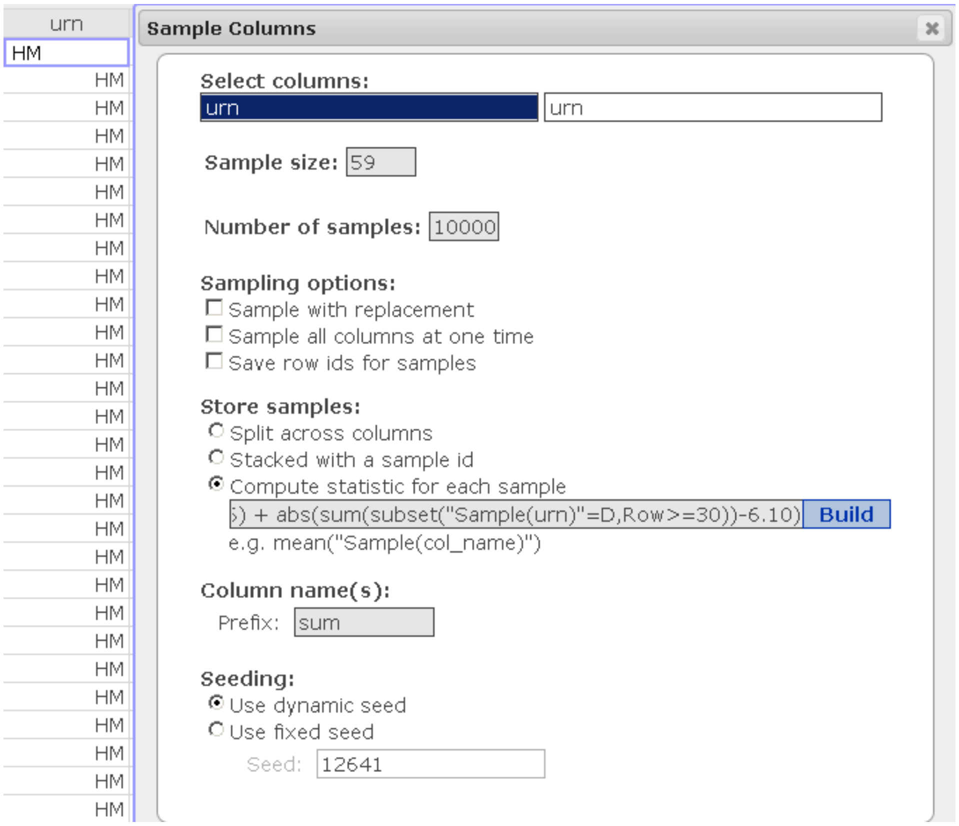
|#HM – 19.17| + |#DM – 3.94| + |#D – 5.90|

(this formula equals the sum of the deviations between the counts in the resample, and the expected counts for the sample size of 29, the same size as the original behaviorial group).

Then, in rows 30-59 (30 rows -- the size of the insight group), do the following:

|#HM – 19.83| + |#DM – 4.06| + |#D – 6.10|

Finally, add the sum from the second calculation to the sum from the first calculation to obtain the total sum of the absolute deviations.



urn
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM
HM

Sample Columns

Select columns: urn urn

Sample size: 59

Number of samples: 10000

Sampling options:
☐ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
) + abs(sum(subset("Sample(urn)"=D,Row>=30))-6.10) Build
e.g. mean("Sample(col_name)")

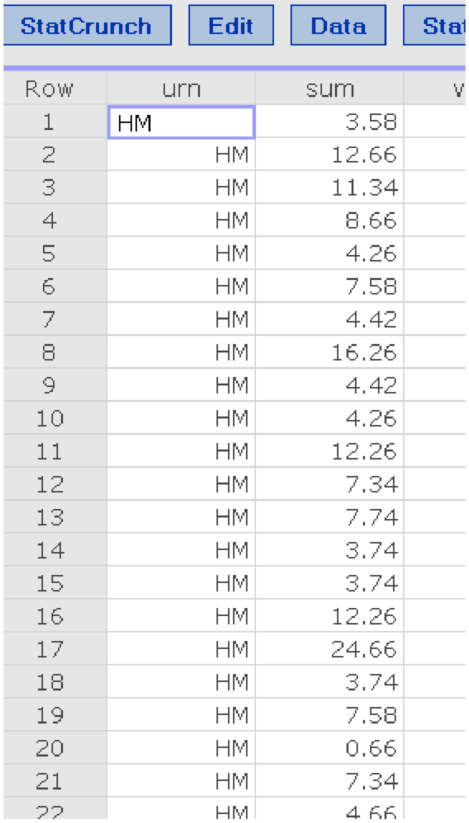
Column name(s):
Prefix: sum

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 9-16: Sample columns dialog for the marriage therapy problem

<u>Note</u>: The seeding can be ignored at this point. Computers generate pseudo-random numbers based on algorithms and these algorithms need to "start somewhere." Random number seeds are generally based on the system clock or some other changing value so that new pseudo-random numbers sets are different from those previously generated. This randomness explains how two runs of the same experiment can result in slightly different results. However, if you are attempting to debug a program or statistical model, then it may be advantageous to use the identical number set multiple times. In order to do this, you can provide a fixed seed. Also, different programs may utilize different algorithms to generate pseudo-random numbers. For example, it is possible (and indeed, quite likely) that programs such as Resampling Stats and StatCrunch will differ in the final output. However, the difference is usually minor and within the error constraints expected from random numbers.

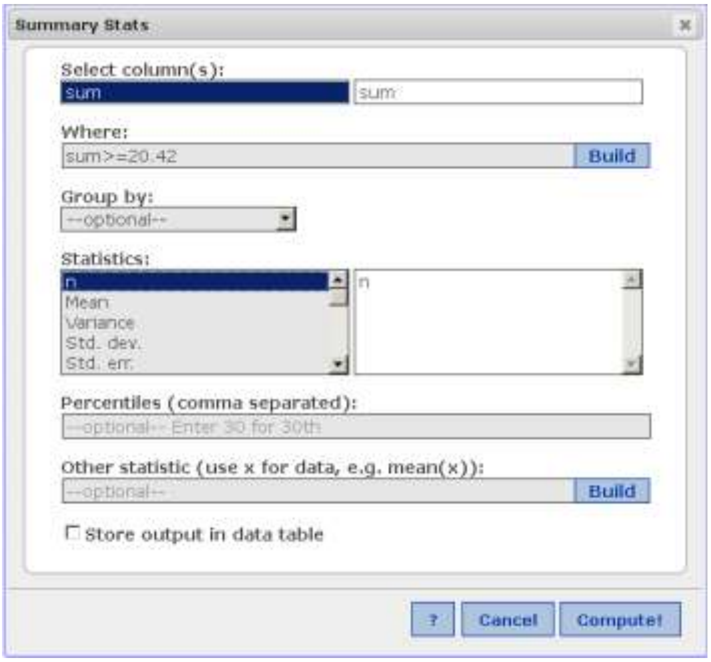
3) Enter a column label (such as sum) and click Compute! This is a LONG calculation and may produce one or more javascript warnings. Let the script continue and be patient. After a few minutes (more or less) the new column will contain the sum of the absolute deviations in the marriage therapy groups. The following figure (figure 9-17) is an example of how the StatCrunch worksheet might appear at this point.



Row	urn	sum	v
1	HM	3.58	
2	HM	12.66	
3	HM	11.34	
4	HM	8.66	
5	HM	4.26	
6	HM	7.58	
7	HM	4.42	
8	HM	16.26	
9	HM	4.42	
10	HM	4.26	
11	HM	12.26	
12	HM	7.34	
13	HM	7.74	
14	HM	3.74	
15	HM	3.74	
16	HM	12.26	
17	HM	24.66	
18	HM	3.74	
19	HM	7.58	
20	HM	0.66	
21	HM	7.34	
22	HM	4.66	

Figure 9-17: Example of a StatCrunch worksheet for the marriage problem

4) The final step involves finding out how many times the sum of the absolute deviations is greater than 20.42 in the 10000 trials. We can use this information to approximate a p-value. Select “Stat>Summary Stats>Columns” and select sum as the column. In the Where box, enter sum>=20.42, click n in the Statistics box and click Compute! (Figure 9-18).



Summary Stats

Select column(s): sum

Where: sum>=20.42

Group by: --optional--

Statistics: n (Mean, Variance, Std. dev., Std. err.)

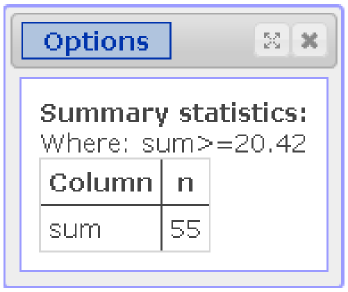
Percentiles (comma separated): --optional-- Enter 30 for 30th

Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 9-18: Column Statistics for the sum of deviations column

The output of this particular simulation is shown below (figure 9-19). The important statistic is n. The other values are computed automatically and are not important in this problem.



Summary statistics:
Where: sum>=20.42

Column	n
sum	55

Figure 9-19: Output of Column Statistics for an example simulation

5) In this trial, there were 55 instances in which the sum of the deviations from both groups was greater than or equal to 20.42. In order to estimate a p-value from this number, we should divide n/10000. This produces a p-value estimate of 55/10000 or 0.0055, which is consistent with the Resampling Stats result and constitutes a

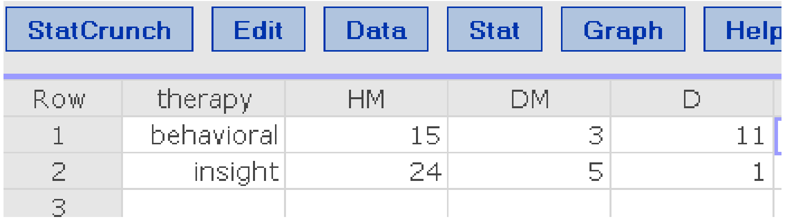
very rare event. The interpretation is that the differences among the three treatments seem to be real, and not the product of chance.

StatCrunch Chi-Square Formula Example

Let's try the marriage therapy problem again using a formula approach, instead of resampling. We will use StatCrunch and a contingency table.

1) Open StatCrunch

2) Enter the marriage therapy data as shown below in figure 9-20

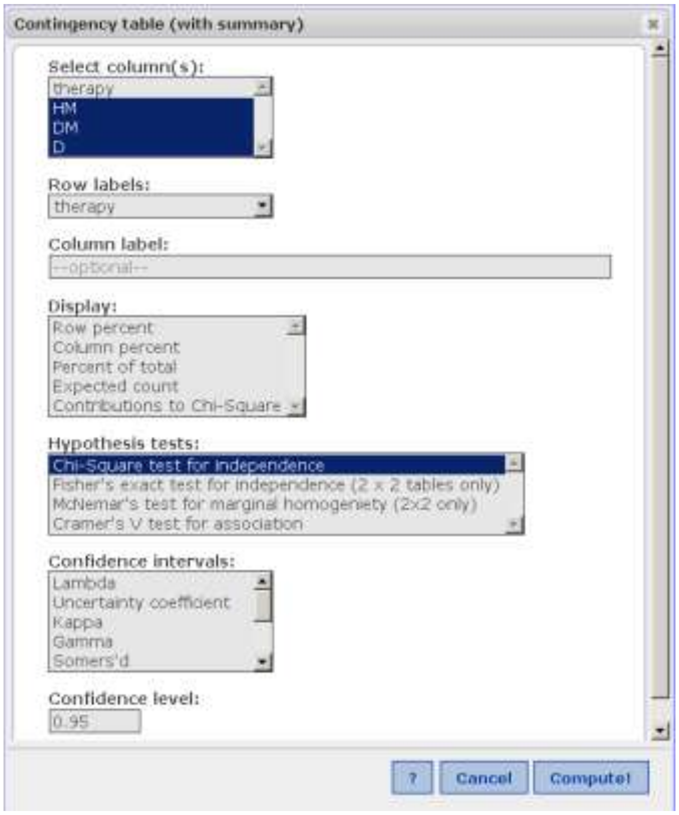


Row	therapy	HM	DM	D
1	behavioral	15	3	11
2	insight	24	5	1
3				

Figure 9-20: StatCrunch with marriage data

3) Select Stat-> Tables-> Contingency-> With Summary

4) Select HM, DM, and D as the columns for the table and "Therapy" for the row labels as shown in figure 9-21



Select column(s):
therapy
HM
DM
D

Row labels:
therapy

Column label:
--optional--

Display:
Row percent
Column percent
Percent of total
Expected count
Contributions to Chi-Square

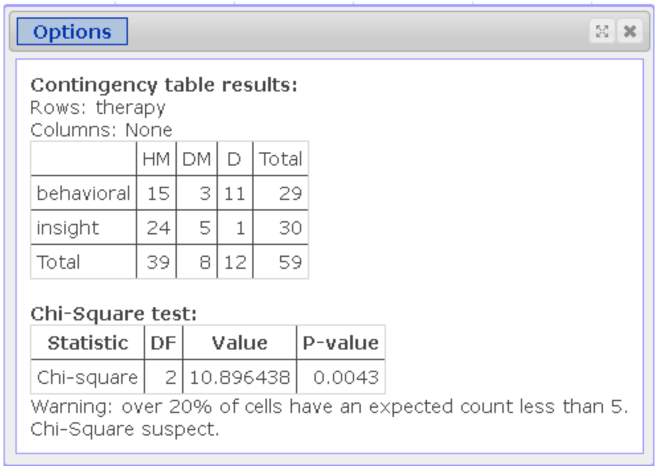
Hypothesis tests:
Chi-Square test for independence
Fisher's exact test for independence (2 x 2 tables only)
McNemar's test for marginal homogeniety (2x2 only)
Cramer's V test for association

Confidence intervals:
Lambda
Uncertainty coefficient
Kappa
Gamma
Somers'd

Confidence level:
0.95

Figure 9-21: Selecting row and column variables

5) Click Compute! to view the contingency table and the Chi-Square results (figure 9-22).



Contingency table results:
Rows: therapy
Columns: None

	HM	DM	D	Total
behavioral	15	3	11	29
insight	24	5	1	30
Total	39	8	12	59

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	10.896438	0.0043

Warning: over 20% of cells have an expected count less than 5.
Chi-Square suspect.

Figure 9-22: StatCrunch Chi-Square results

6) The p-value of 0.0043 indicates the results are probably not due to chance.

9.5 Goodness-of-Fit

Resampling Stats (RSXL) Solution

1) Open Resampling Stats and enter the digits 0 through 9 in cells A1:A10

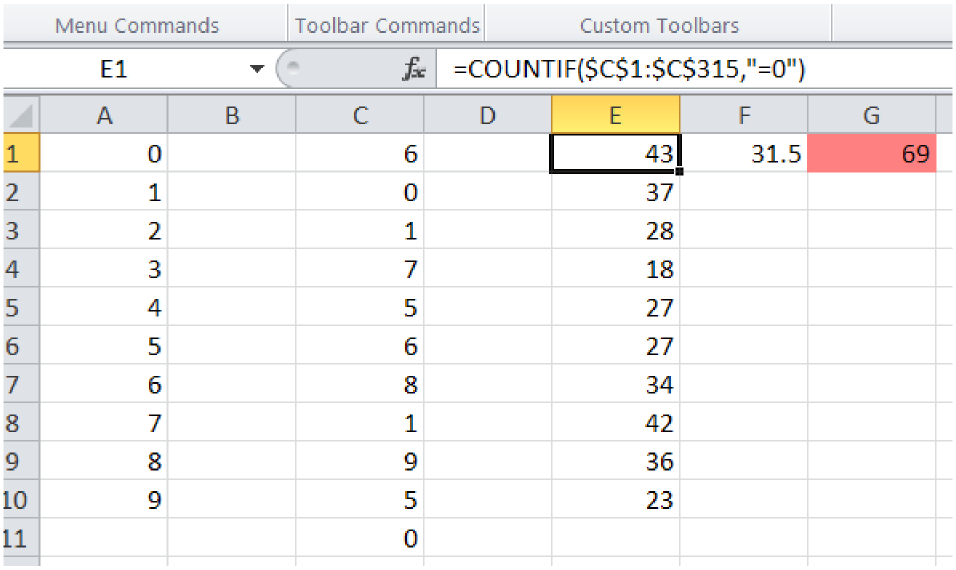
2) Select “Resample” from the RSXL menu and choose C1 as the top cell of the output range and 315 as the number of cells in the output range.

3) To simplify calculations, in cells E1:E10, enter the formula shown in the figure below, adjusting the formula for each value from 0 to 9.

4) Enter 31.5 in cell F1.

5) In cell G1, enter the following formula (all on one line) to calculate the sum of the absolute difference between 31.5 (in cell F1) and the totals of each of the digits from 0-9:

   =ABS(F1-E1)+ABS(F1-E2)+ABS(F1-E3)+ABS(F1-E4)+ABS(F1-E5)+ABS(F1-E6)+ABS(F1-E7)+ABS(F1-E8)+ABS(F1-E9)+ABS(F1-E10)



E1 =COUNTIF($C$1:$C$315,"=0")

	A	B	C	D	E	F	G
1	0		6		43	31.5	69
2	1		0		37		
3	2		1		28		
4	3		7		18		
5	4		5		27		
6	5		6		27		
7	6		8		34		
8	7		1		42		
9	8		9		36		
10	9		5		23		
11			0				

6) Select cell G1 as the score cell and try 10000 trials.

7) On the Results sheet, enter: =COUNTIF(A1:A10000,">=216") in cell B1.

8) Divide this result by the number of trials to get the p-value.

9) If you have patience, a fast computer, and Excel 2007 or Excel 2010, try 100000 trials and see what happens. Try 1000000 trials (if you have patience!). You will have to adjust the formula in cell B1 of the Results sheet to reflect the new number of trials.

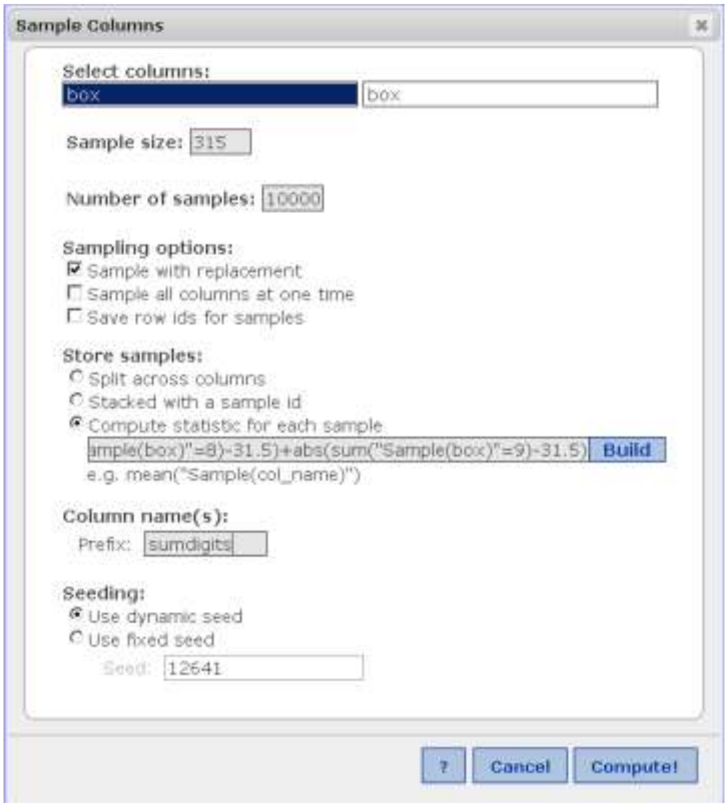
StatCrunch

1) Enter the digits 0-9 in the first column. Rename the column “box” or something similar.

2) Select Data>Sample and select the box column. Enter 315 for the sample size, check Sample with Replacement, and enter 10000 for the number of samples. The Statistic expression is complex. Look at it carefully and either type or copy/paste into the expression box (see figure 9-23). Label the new column sumdigits. Click Compute!

   ```
   abs(sum("Sample(box)"=0)-31.5)+abs(sum("Sample(box)"=1)-
   31.5)+abs(sum("Sample(box)"=2)-31.5)+abs(sum("Sample(box)"=3)-
   31.5)+abs(sum("Sample(box)"=4)-31.5)+abs(sum("Sample(box)"=5)-
   31.5)+abs(sum("Sample(box)"=6)-31.5)+abs(sum("Sample(box)"=7)-
   31.5)+abs(sum("Sample(box)"=8)-31.5)+abs(sum("Sample(box)"=9)-31.5)
   ```

3) Each component of the above expression sums the raw count of each digit in turn, and takes the absolute value of the difference in that count and 31.5. The sum of the entire expression is our statistic of interest. The results (after a short wait for 10000 trials… the javascript in the new SC isn’t quite as fast as java in the classic SC!) are shown in figure 9-24.



Sample Columns

Select columns: box

Sample size: 315

Number of samples: 10000

Sampling options:

☑ Sample with replacement

☐ Sample all columns at one time

☐ Save row ids for samples

Store samples:

○ Split across columns

○ Stacked with a sample id

● Compute statistic for each sample

ample(box)"=8)-31.5)+abs(sum("Sample(box)"=9)-31.5)

e.g. mean("Sample(col_name)")

Column name(s):

Prefix: sumdigits

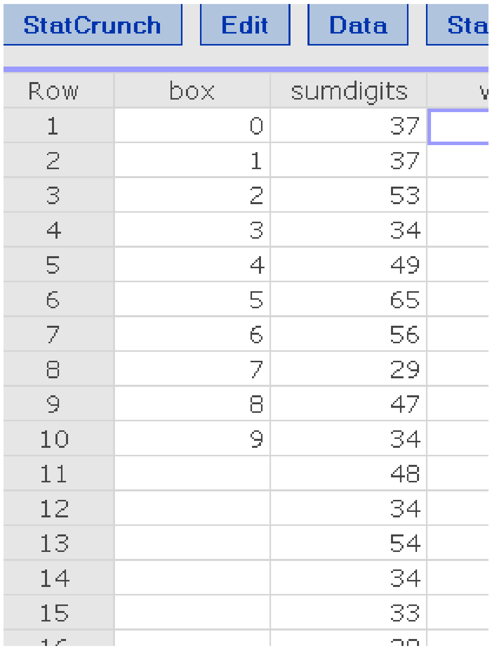
Seeding:

● Use dynamic seed

○ Use fixed seed

Seed: 12641

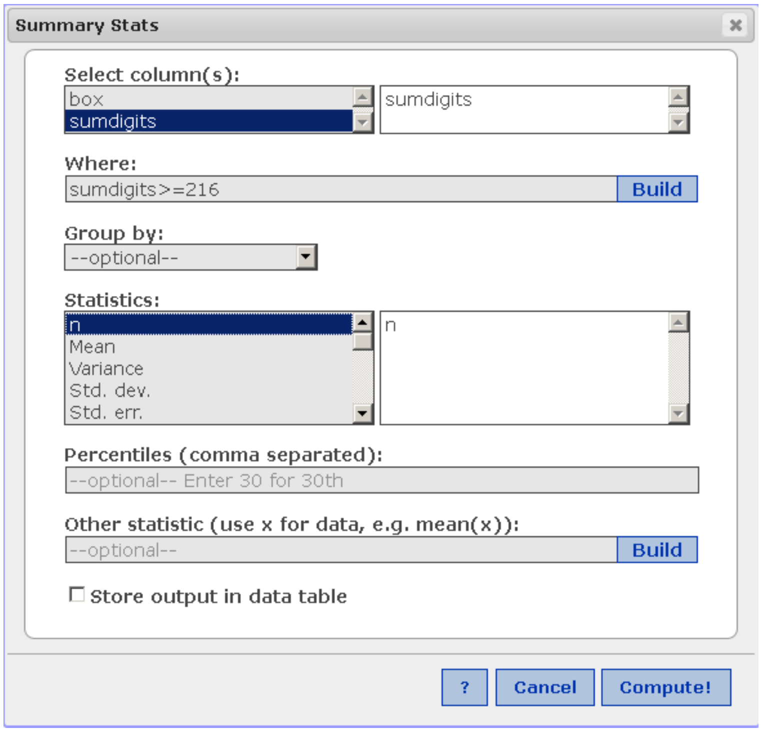
Figure 9-23: Sample columns setup



Row	box	sumdigits
1	0	37
2	1	37
3	2	53
4	3	34
5	4	49
6	5	65
7	6	56
8	7	29
9	8	47
10	9	34
11		48
12		34
13		54
14		34
15		33

Figure 9-24: Sum of difference in individual digit counts and 31.5

4) At this point, we need to find out how many times in 10000 trials the sum of the digit count differences >=216. Select Stat>Summary Stats>Columns and fill in the dialog as follows in figure 9-25. Click Compute!


Summary Stats
box
Group by:
--optional--
Mean
Variance
Std. dev.
Std. err.
Percentiles (comma separated):
--optional-- Enter 30 for 30th
Other statistic (use x for data, e.g. mean(x)):
--optional--
Build
Build
☐ Store output in data table
?
Cancel
Compute!

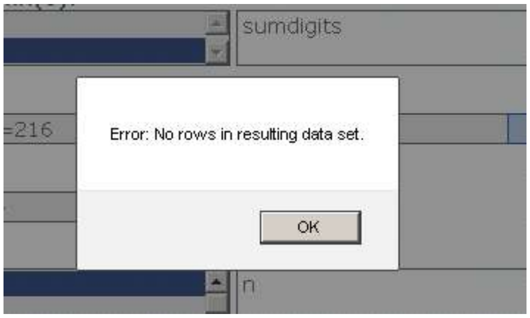
Select column(s): sumdigits

Where: sumdigits>=216

Statistics: n

Figure 9-25: Summary Statistics: sumdigits>=216

5) The result was an error message indicating that there were no rows meeting this condition (figure 9-26).



Error: No rows in resulting data set.

Figure 9-26: Estimated p-value = 0.0000

6) There were no instances >=216 for an estimated p-value = 0.0000. In fact, for this experiment, the maximum value in 10000 trials was 90 (using Data>Compute Expression and max(sumdigits) for the expression).

10 Correlation

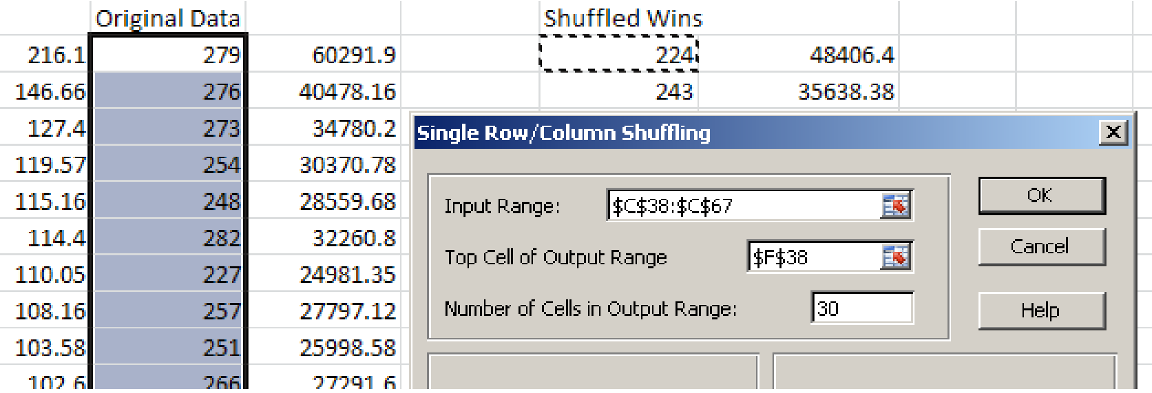
10.3 The Vector Product and Sum Test

Download the Baseball_Payroll.xlsx file <here>

Resampling Stats (Vector Product)

1) Download the Baseball_Payroll.xlsx file. There is a copy of the original data starting in cell B38. The workbook is generally ready for resampling. Column G, starting in cell G38, contains the vector product formulas (you could use a formula array instead of the column G formulas if you wish). The score cell is highlighted and is cell G68.

2) Select cells C38:C67 and choose Shuffle (without replacement) from the Resampling Stats menu. Select cell F38 as the Top Left Cell, keep the default number of output cells at 30. Click OK (figure 10-1).



	Original Data		Shuffled Wins	
216.1	279	60291.9	224	48406.4
146.66	276	40478.16	243	35638.38
127.4	273	34780.2		
119.57	254	30370.78		
115.16	248	28559.68		
114.4	282	32260.8		
110.05	227	24981.35		
108.16	257	27797.12		
103.58	251	25998.58		
102.6	266	27291.6		

Single Row/Column Shuffling

Input Range: $C$38:$C$67

Top Cell of Output Range: $F$38

Number of Cells in Output Range: 30

Figure 10-1: Shuffle the wins

3) Select cell G68 as the score cell and choose Repeat and Score. Try 1000 iterations/trials.

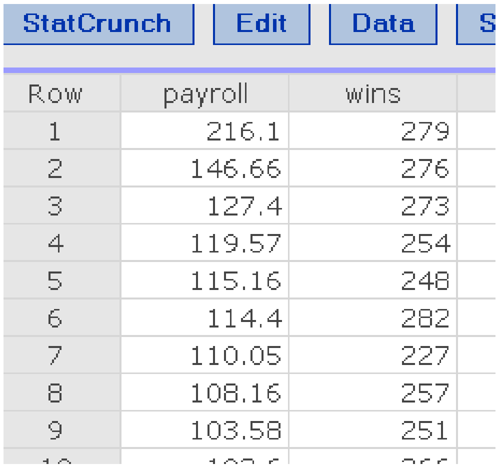
4) On the Results worksheet, enter =COUNTIF(result1,”>=668620”)

5) The results of one experiment found no values >=668620, which corresponds to the histogram (figure 10-4) in the text.

StatCrunch Resampling Method

1) Start StatCrunch and copy/paste the baseball payroll data (in column B, starting with cell B2) into the first column of StatCrunch. Label the payroll column

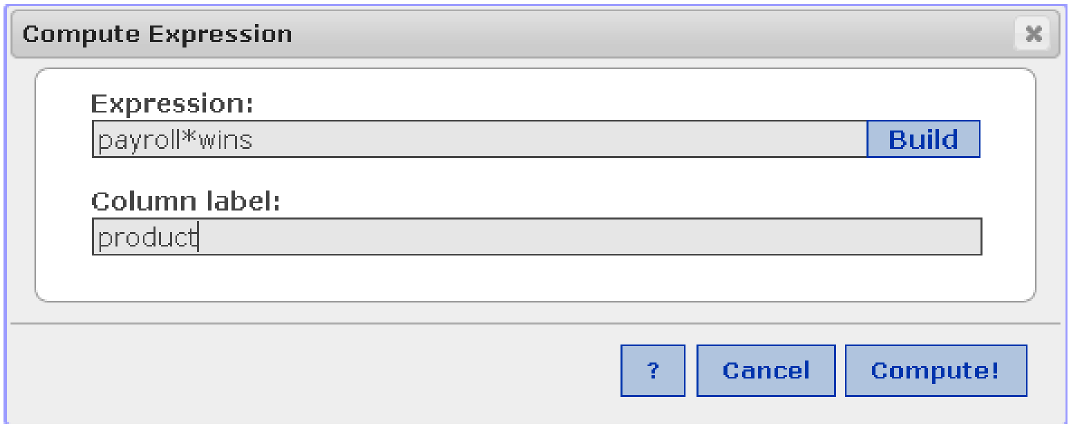
"payroll" or something similar. Copy/paste the Total Wins data (in column C, starting with cell C2) to the second column in the StatCrunch worksheet. Label this new column "wins" as shown in the following figure:



Row	payroll	wins
1	216.1	279
2	146.66	276
3	127.4	273
4	119.57	254
5	115.16	248
6	114.4	282
7	110.05	227
8	108.16	257
9	103.58	251

Figure 10-2: Baseball payroll example

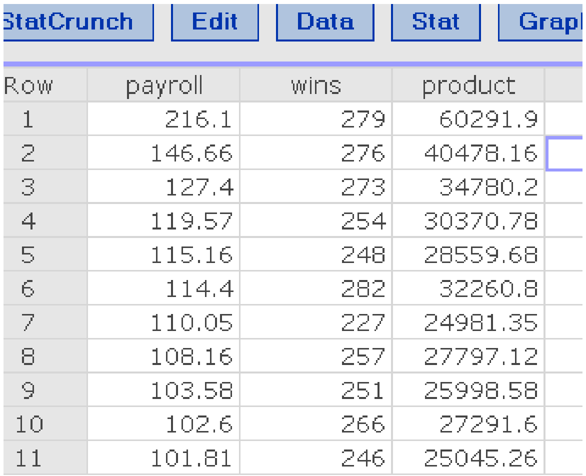
2) Column 3 in StatCrunch will hold the vector product of payroll*wins. Select Data>Compute Expression and enter payroll*wins in the Expression box. Name the new column product as shown in figure 10-3. Click Compute! Figure 10-4 illustrates the product column.



Expression: payroll*wins

Column label: product

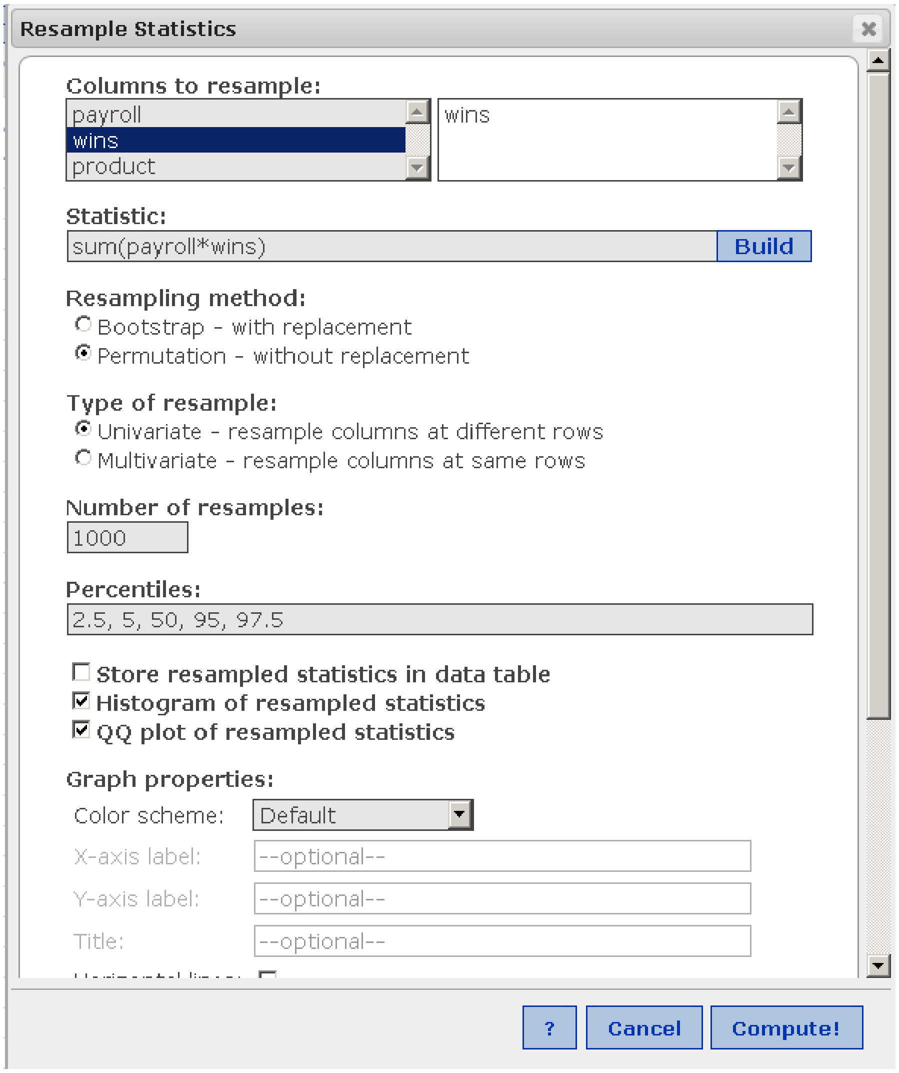
Figure 10-3: Compute expression payroll*wins



Row	payroll	wins	product
1	216.1	279	60291.9
2	146.66	276	40478.16
3	127.4	273	34780.2
4	119.57	254	30370.78
5	115.16	248	28559.68
6	114.4	282	32260.8
7	110.05	227	24981.35
8	108.16	257	27797.12
9	103.58	251	25998.58
10	102.6	266	27291.6
11	101.81	246	25045.26

Figure 10-4: Vector product payroll*wins

3) Perform the resampling procedure by selecting "Stat>Resample>Statistic" and choosing the "wins" column to resample. Select Permutation-without replacement and enter sum(payroll*wins) as the Statistic expression. Click the Resample Statistic button (figure 10-5).



Columns to resample:
payroll
wins
product

wins

Statistic: sum(payroll*wins)

Resampling method:
○ Bootstrap - with replacement
● Permutation - without replacement

Type of resample:
● Univariate - resample columns at different rows
○ Multivariate - resample columns at same rows

Number of resamples: 1000

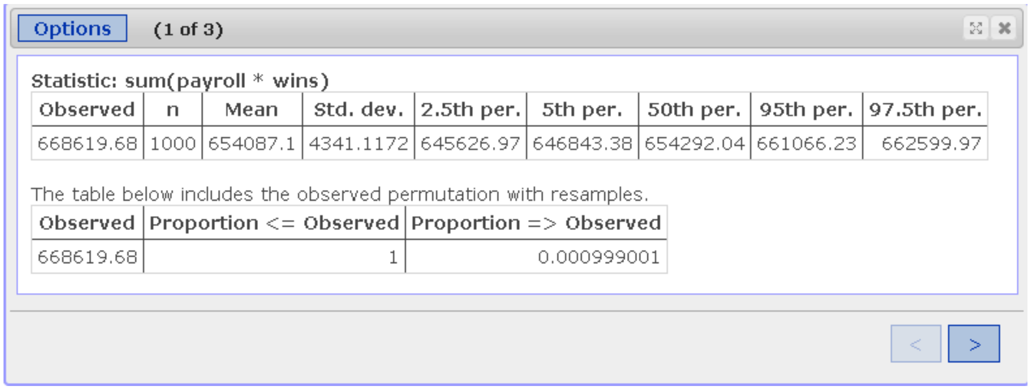
Percentiles: 2.5, 5, 50, 95, 97.5

☐ Store resampled statistics in data table
☑ Histogram of resampled statistics
☑ QQ plot of resampled statistics

Graph properties:
Color scheme: Default
X-axis label: --optional--
Y-axis label: --optional--
Title: --optional--

Figure 10-5: Resample wins column (without replacement)

4) The “wins” column is the only column that will be resampled. The “payroll” column will remain static. The sum of the vector product between the static “payroll” column and the resampled “wins” column will be calculated 1000 times. The results are shown in figure 10-6. The sum of the vector product in the original data was 668620. This value is outside the 97.5th percentile of the results of this particular experiment, making it highly unlikely that the results were due to chance. Figure 10-7 is a histogram of the results. The red line on the right represents the vector product of the original data.



Statistic: sum(payroll * wins)

Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
668619.68	1000	654087.1	4341.1172	645626.97	646843.38	654292.04	661066.23	662599.97

The table below includes the observed permutation with resamples.

Observed	Proportion <= Observed	Proportion => Observed
668619.68	1	0.000999001

Figure 10-6: Baseball payroll results

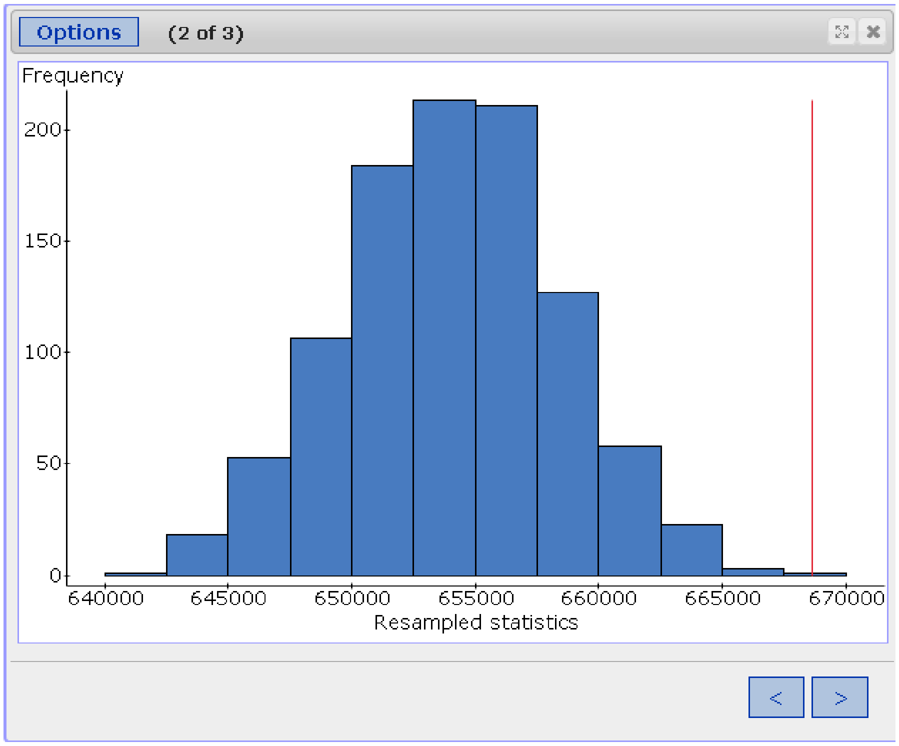
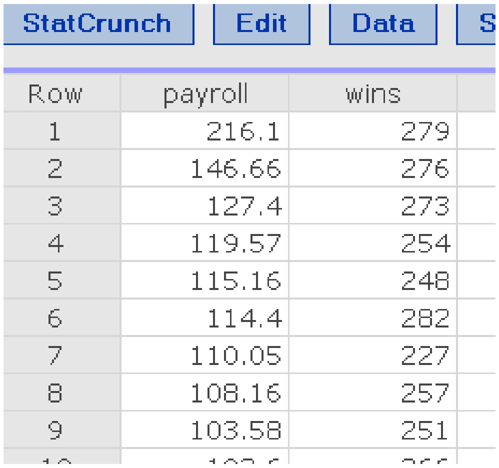


Figure 10-7: Histogram of baseball payroll vector products

StatCrunch Formula Method

Let’s use StatCrunch and its formula method to verify the Resampling Stats results.

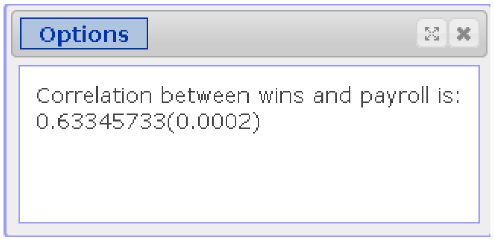
1) Copy the payroll and wins data from the Excel Baseball_Payroll.xlsx workbook and paste into StatCrunch (figure 10-8). Rename the columns “payroll” and “wins”.



Row	payroll	wins
1	216.1	279
2	146.66	276
3	127.4	273
4	119.57	254
5	115.16	248
6	114.4	282
7	110.05	227
8	108.16	257
9	103.58	251

Figure 10-8: Initial StatCrunch worksheet

2) Select Stat>Summary Stats>Correlation and choose both the payroll and wins columns. Check the two-sided P-value in the Display in Correlation Matix section. Click Compute! Figure 10-9 shows the results.



Correlation between wins and payroll is:
0.63345733(0.0002)

Figure 10-9: Correlation coefficient r=0.633

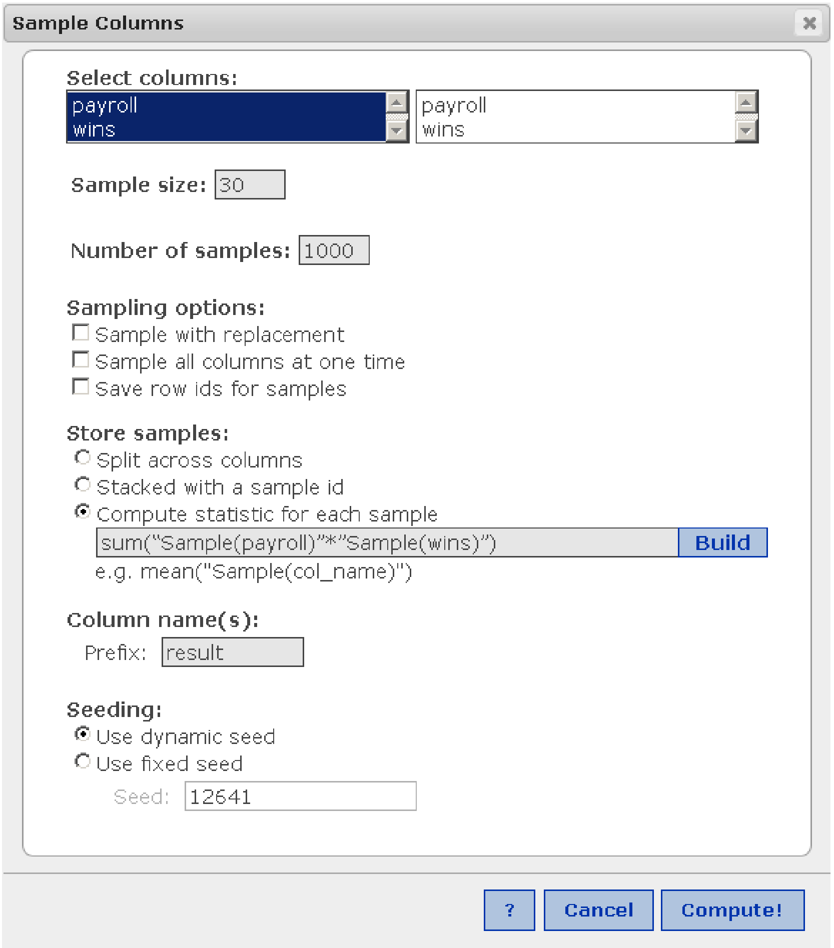
3) The correlation coefficient indicates a strong relationship and is statistically significant (two-sided p=0.0002).

StatCrunch Resampling (Vector Product) Revisited

The following procedure uses a slightly different method in StatCrunch to solve the baseball payroll example.

1) Use the original StatCrunch data above or recopy/paste from the Excel Baseball_Payroll.xlsx workbook. If you re-use the original data, choose Edit>Columns>Delete and select the product column for deletion (if the column still exists).

2) Select Data>Sample and choose both columns (figure 10-10). It doesn't matter whether we randomize the order of one column or both columns. The result will be a random pairing of payroll and wins and the final outcome will be the same.



Select columns: payroll, wins

Sample size: 30

Number of samples: 1000

Sampling options:
☐ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
sum("Sample(payroll)"*"Sample(wins)")
e.g. mean("Sample(col_name)")

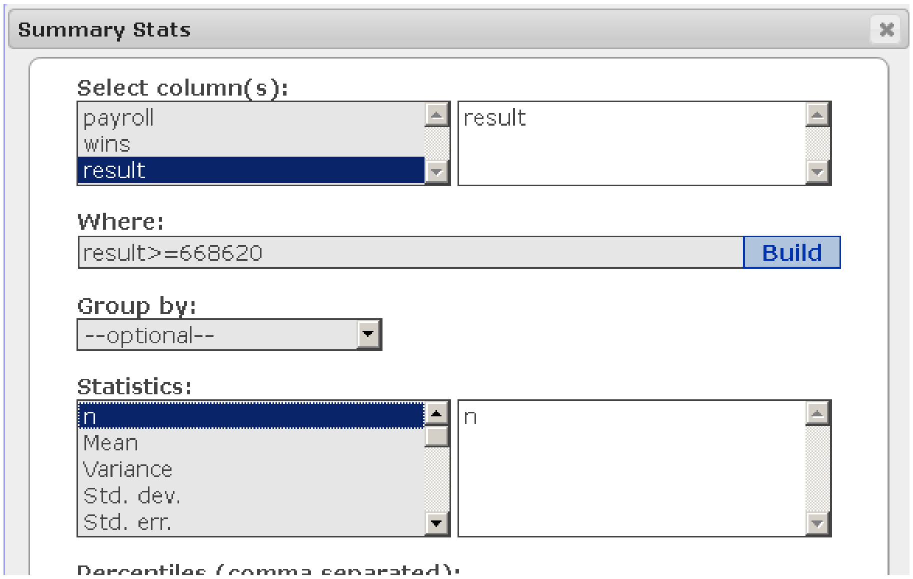
Column name(s):
Prefix: result

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 10-10: Sample columns with statistic expression

3) As shown in figure 10-10, enter 30 for the Sample size, 1000 for the Number of samples, and sum("Sample(payroll)"*"Sample(wins)") for the statistic expression. Name the new column result or something similar and click Compute!

4) Select Stat>Summary Stats>Columns, choose the result column, enter the expression: result>=668620 and click n for the statistic. Click Compute! (figure 10-11).



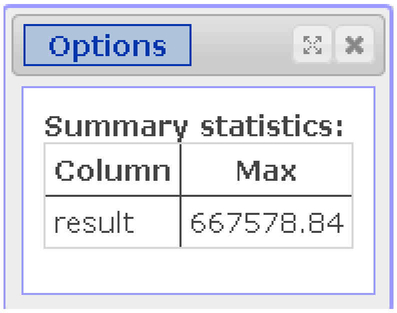
Select column(s): result

Where: result>=668620

Statistics: n

Figure 10-11: Summary statistic Where: expression

5) There were no examples that matched the Where expression in the above experiment. Trying the summary stats again, this time without the Where expression and with the Max function selected in the Statistics box, the maximum value was displayed and is shown in figure 10-12.



Summary statistics:

Column	Max
result	667578.84

Figure 10-12: Max value

6) The max value is close to the criteria in the original Where statement. If we would try more samples, perhaps we could find a success? Retry the exercise and use 10000 samples.

Resampling Stats Online Dating (Try It Yourself)

Try It Yourself

(optional - requires Resampling Stats or StatCrunch)

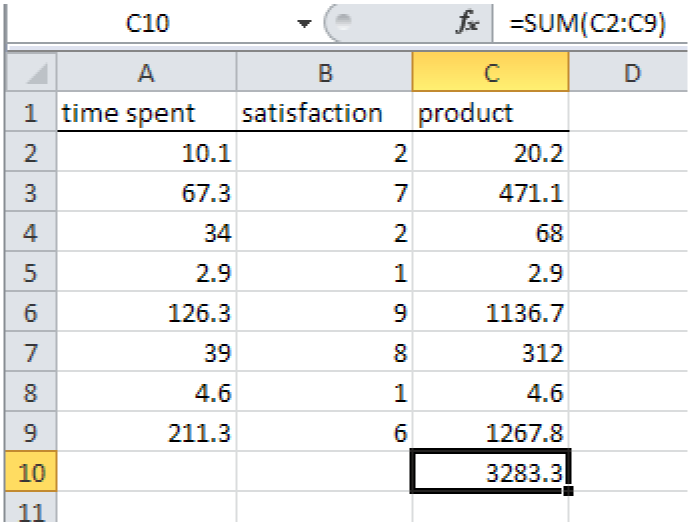
An online dating site seeks to learn more about what it can do to encourage successful outcomes to the relationships formed by its customers. It collects various data, including satisfaction surveys 6 months after a customer first signs up, and also how much time the customer has spent on its dating site. Satisfaction data is recorded as an integer between 1 and 10, and time-on-site is recorded in minutes. Here are hypothetical results:

time spent	satisfaction
10.1	2
67.3	7
34	2
2.9	1
126.3	9
39	8
4.6	1
211.3	6

Calculate the vector product sum and use a resampling procedure to test whether there is a correlation between time spent and satisfaction.

This problem is nearly identical to the baseball payroll example. Here are the steps using Resampling Stats:

1) Start Excel and load the Resampling Stats add-in. Enter the “time spent” and “satisfaction” data in adjacent columns in Excel. Create a “product” column by multiplying the “time spent” and “satisfaction” columns as you did for the baseball payroll example (figure 10-13).



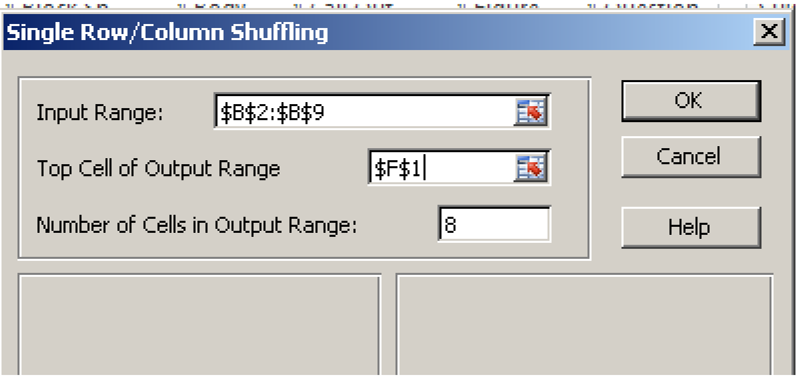
C10 =SUM(C2:C9)

	A	B	C	D
1	time spent	satisfaction	product	
2	10.1	2	20.2	
3	67.3	7	471.1	
4	34	2	68	
5	2.9	1	2.9	
6	126.3	9	1136.7	
7	39	8	312	
8	4.6	1	4.6	
9	211.3	6	1267.8	
10			3283.3	
11				

Figure 10-13: Vector product online dating data

2) Copy the dating data in cells A2:A9 and paste this data in cell E1.

3) Select Shuffle from the Resampling menu. Choose cells B2:B9 as the input range and cell F1 as the top cell of the output range (figure 10-14).



Single Row/Column Shuffling

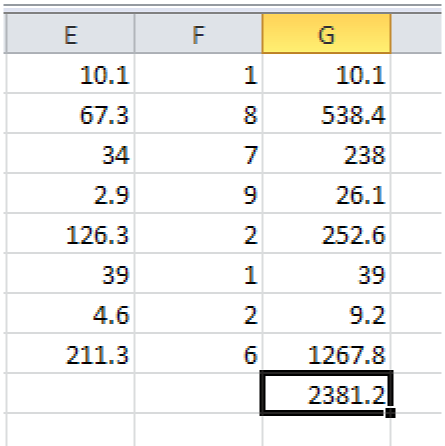
Input Range: $B$2:$B$9

Top Cell of Output Range $F$1

Number of Cells in Output Range: 8

Figure 10-14: Shuffle dialog online dating

4) Click OK and create a new product column using the data in cells E1:E8 and the shuffled data (permutation resampling… without replacement) in cells F1:F8. Calculate the sum of the new product column in cell G9 (figure 10-15).



E	F	G
10.1	1	10.1
67.3	8	538.4
34	7	238
2.9	9	26.1
126.3	2	252.6
39	1	39
4.6	2	9.2
211.3	6	1267.8
		2381.2

Figure 10-15: Onine dating vector product sum after resampling

5) Select Repeat and Score, choose the vector product sum (cell G1 in figure 10-15) and try 1000 iterations/trials.

6) In the Results worksheet, enter the Excel function =COUNTIF(result1, “>=3283.3”) in cell C1 to find the number of instances where the resampled vector product sum meets or exceeds the vector product sum of the original online dating data.

7) As figure 10-16 illustrates, there were 49 instances of a vector product sum exceeding the original value of 3283.3. This results in a p-value of 49/1000 or .049. Based on this experiment (your results may differ slightly), the original results are most likely not due to chance and there is a correlation between time spent online and satisfaction with the dating service. A histogram of the results is shown in figure 10-16.

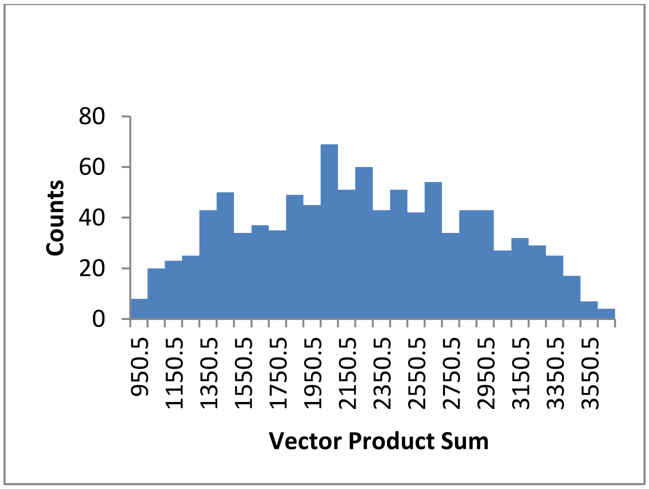


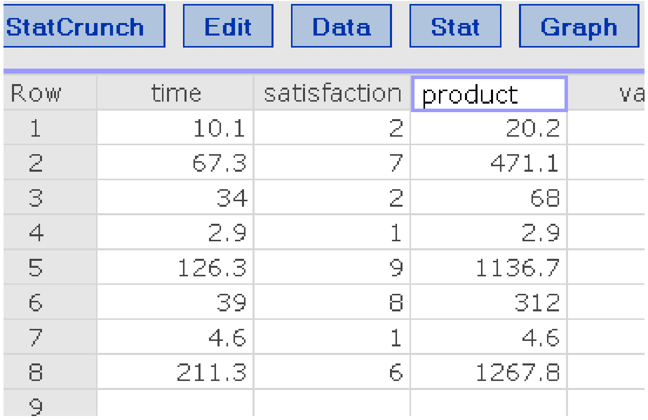
Figure 10-16: Vector product sum online dating

<u>Note</u>: This is an interesting result in that the p-value is very close to 0.05, a common p-value established in many behavioral or educational studies as the “cut-off” or criterion for statistical significance. It might be worthwhile to try this experiment multiple times or with more trials in order to establish whether or not the $p<0.05$ has actually been met. See the StatCrunch results and the subsequent note for further information.

StatCrunch Online Dating Resampling Procedure (Try It Yourself)

The steps for simulating the online dating satisfaction example in StatCrunch are nearly identical to the procedure used for solving the baseball payroll example. You can use any of the StatCrunch methods illustrated in the baseball payroll example, however this example will follow the first resampling method which uses the “Stat>Resample>Statistic” procedure.

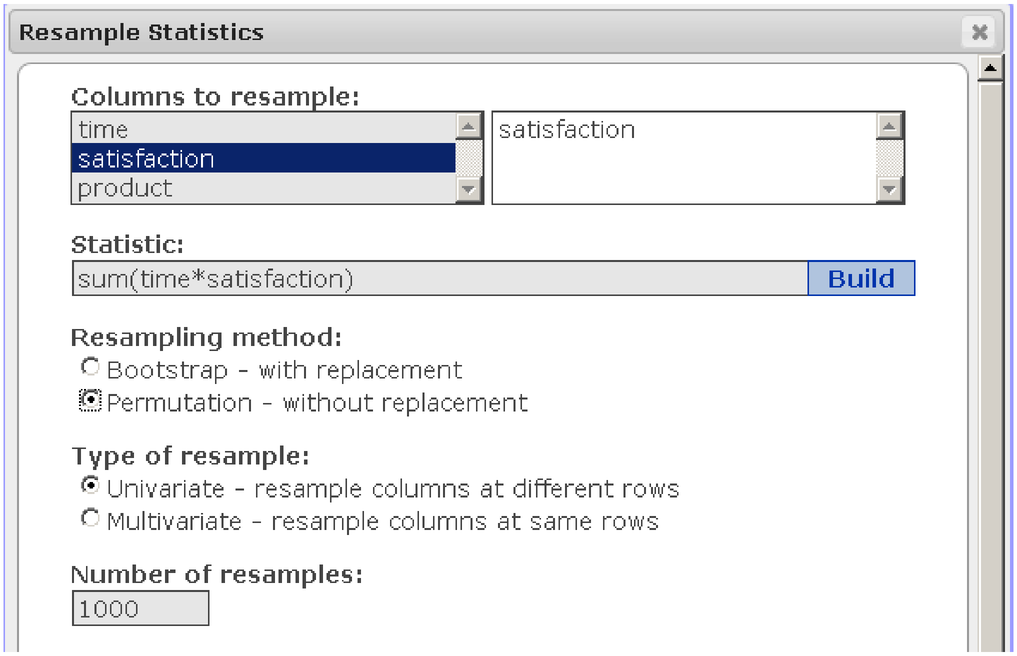
1) Start StatCrunch and load the online dating data. Label the columns appropriately and create a product column as shown in figure 10-17.



Row	time	satisfaction	product	va
1	10.1	2	20.2	
2	67.3	7	471.1	
3	34	2	68	
4	2.9	1	2.9	
5	126.3	9	1136.7	
6	39	8	312	
7	4.6	1	4.6	
8	211.3	6	1267.8	
9				

Figure 10-17: StatCrunch worksheet online dating satisfaction example

2) Select “Stat>Resample>Statistic” and complete the dialog as shown in figure 10-18.



Resample Statistics

Columns to resample:
- time
- satisfaction
- product

Selected: satisfaction

Statistic: sum(time*satisfaction)

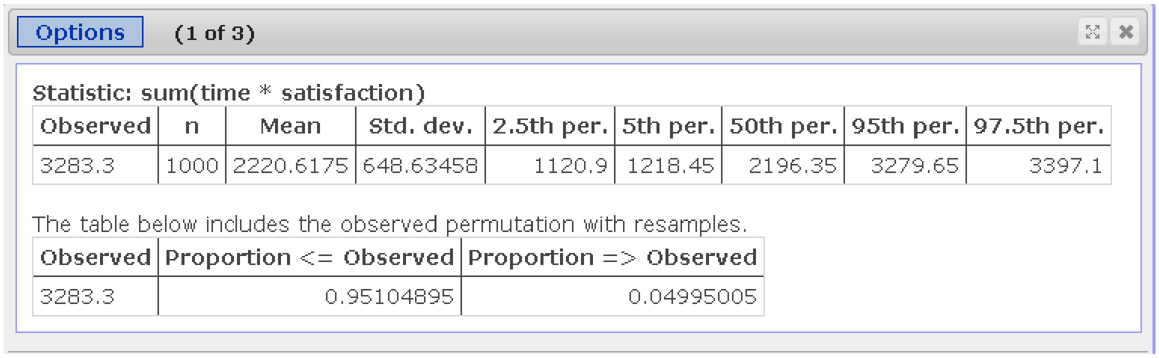
Resampling method:
- ○ Bootstrap - with replacement
- ● Permutation - without replacement

Type of resample:
- ● Univariate - resample columns at different rows
- ○ Multivariate - resample columns at same rows

Number of resamples: 1000

Figure 10-18: Resample Statistic dialog online dating example

3) Click Compute! and view the results (figure 10-19):



Statistic: sum(time * satisfaction)

Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
3283.3	1000	2220.6175	648.63458	1120.9	1218.45	2196.35	3279.65	3397.1

The table below includes the observed permutation with resamples.

Observed	Proportion <= Observed	Proportion => Observed
3283.3	0.95104895	0.04995005

Figure 10-19: Results online dating satisfaction

4) The results of the StatCrunch simulation indicated that the p-value was 0.04995005 which can be rounded to 0.050.

Note: Why is this result slightly different than the 0.049 p-value obtained in the Resampling Stats experiment? Remember that resampling methods employ random selection of data, either with or without replacement. It is the nature of randomness that we should not expect the same result each time we run an experiment or simulation. However, averaging repeated sets of experiments or using a large number of trials (10000+) we would expect the results of resampling methods to approach a correct value. Following the results of the StatCrunch experiment, the author repeated the Resampling Stats example using the same data, but with 10000 trials. The p-value obtained in the new experiment was 0.051.

10.4 Correlation Coefficient

Resampling Stats (Baseball Payroll)

1. Place the team payroll data in cells B2:B31. The corresponding team wins data should go in cells C2:C31.

2. With the cursor in a blank cell, type the "=CORREL" function (cell E1 in Figure 10-20). Choose cells B2:B31 as the first array and C2:C31 as the second array.

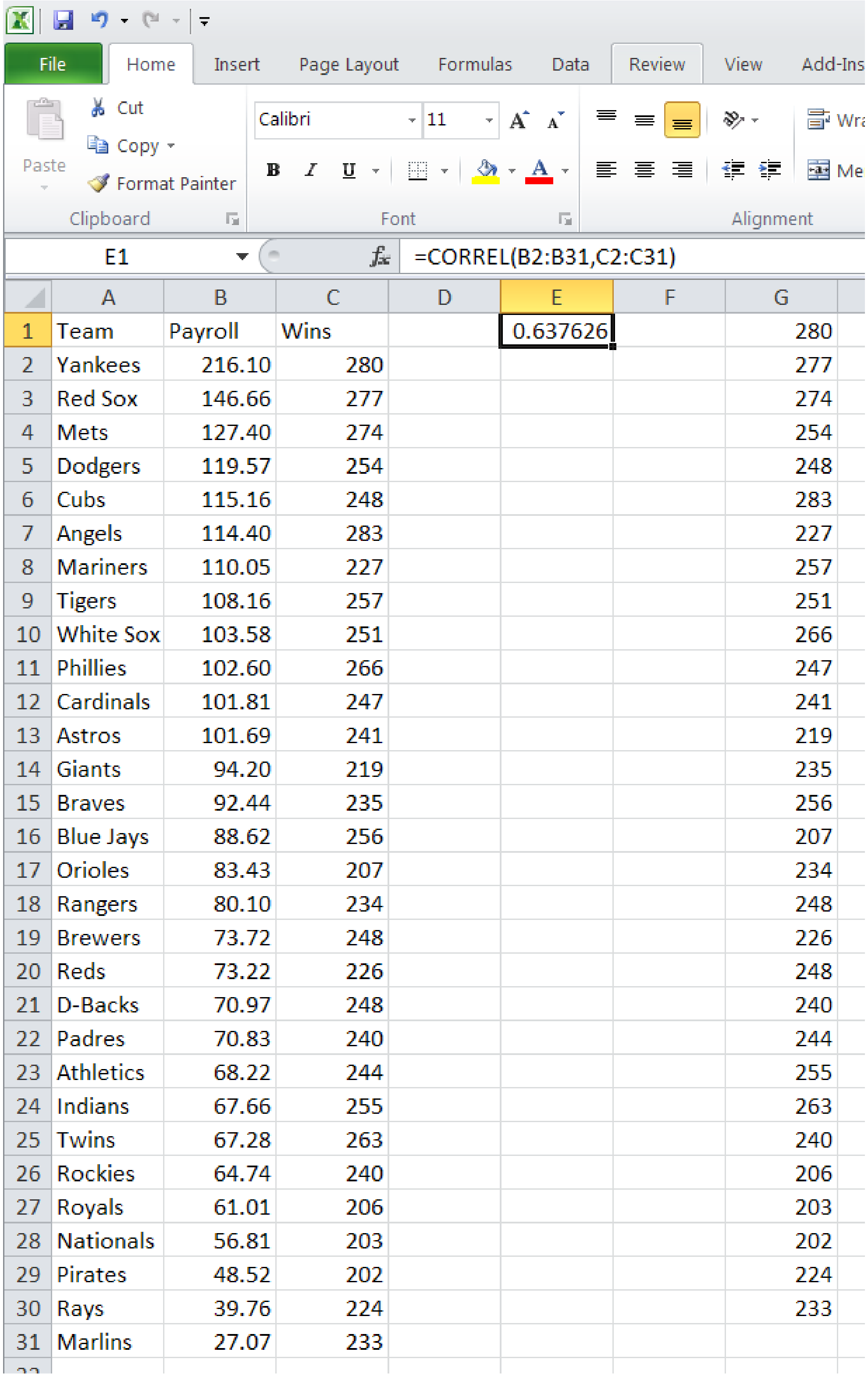
3. The function result is .64 (rounded), indicating a strong relationship between payroll and wins.

4. Copy the array C2:C31 to a new location (a temporary "parking" place) such as cells G1:G30 (see Figure 10-20).

5. Use the Shuffle function on the Resampling Stats menu to shuffle the G1:G30 range back into C2:C31 by using cell C2 as the Top Cell of the output range. Note the change in the correlation coefficient.

6. Select the "=CORREL" cell for Repeat and Score and try 1000 iterations.

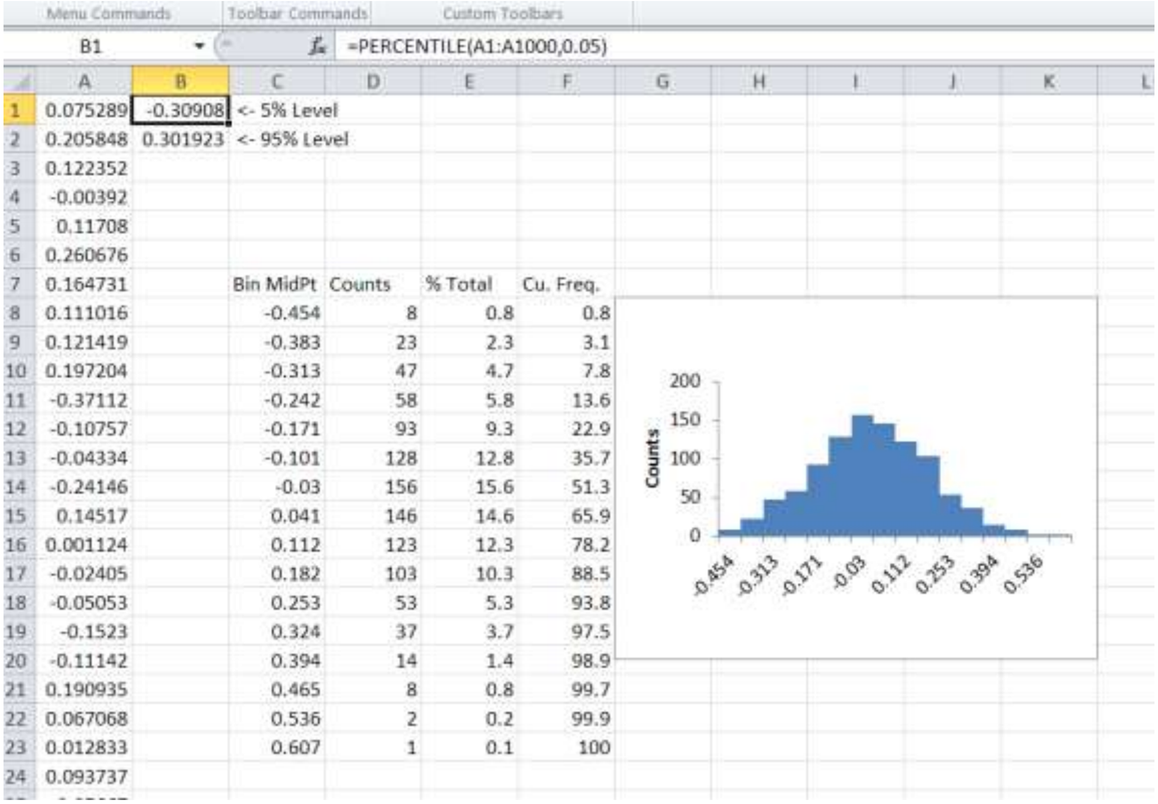
7. On the "Results" worksheet, find a 90% confidence interval using the "=PERCENTILE" function. Display a histogram of the Results (Figure 10-21).



E1 =CORREL(B2:B31,C2:C31)

	A	B	C	D	E	F	G
1	Team	Payroll	Wins		0.637626		280
2	Yankees	216.10	280				277
3	Red Sox	146.66	277				274
4	Mets	127.40	274				254
5	Dodgers	119.57	254				248
6	Cubs	115.16	248				283
7	Angels	114.40	283				227
8	Mariners	110.05	227				257
9	Tigers	108.16	257				251
10	White Sox	103.58	251				266
11	Phillies	102.60	266				247
12	Cardinals	101.81	247				241
13	Astros	101.69	241				219
14	Giants	94.20	219				235
15	Braves	92.44	235				256
16	Blue Jays	88.62	256				207
17	Orioles	83.43	207				234
18	Rangers	80.10	234				248
19	Brewers	73.72	248				226
20	Reds	73.22	226				248
21	D-Backs	70.97	248				240
22	Padres	70.83	240				244
23	Athletics	68.22	244				255
24	Indians	67.66	255				263
25	Twins	67.28	263				240
26	Rockies	64.74	240				206
27	Royals	61.01	206				203
28	Nationals	56.81	203				202
29	Pirates	48.52	202				224
30	Rays	39.76	224				233
31	Marlins	27.07	233				

Figure 10-20: Correlation between payroll and wins



=PERCENTILE(A1:A1000,0.05)

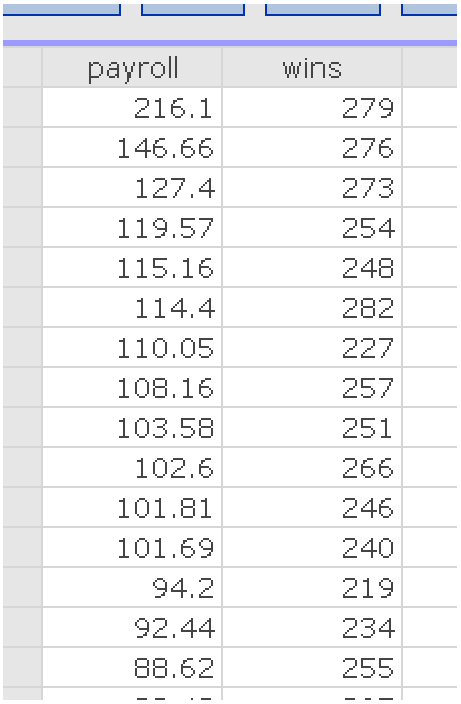
	A	B	C	D	E	F
1	0.075289	-0.30908	<- 5% Level			
2	0.205848	0.301923	<- 95% Level			
3	0.122352					
4	-0.00392					
5	0.11708					
6	0.260676					
7	0.164731		Bin MidPt	Counts	% Total	Cu. Freq.
8	0.111016		-0.454	8	0.8	0.8
9	0.121419		-0.383	23	2.3	3.1
10	0.197204		-0.313	47	4.7	7.8
11	-0.37112		-0.242	58	5.8	13.6
12	-0.10757		-0.171	93	9.3	22.9
13	-0.04334		-0.101	128	12.8	35.7
14	-0.24146		-0.03	156	15.6	51.3
15	0.14517		0.041	146	14.6	65.9
16	0.001124		0.112	123	12.3	78.2
17	-0.02405		0.182	103	10.3	88.5
18	-0.05053		0.253	53	5.3	93.8
19	-0.1523		0.324	37	3.7	97.5
20	-0.11142		0.394	14	1.4	98.9
21	0.190935		0.465	8	0.8	99.7
22	0.067068		0.536	2	0.2	99.9
23	0.012833		0.607	1	0.1	100
24	0.093737					

Figure 10-21: Resampling distribution of correlation coefficient for baseball, under the null hypothesis

StatCrunch Resampling (Baseball Payroll)

We can generally replicate the Resampling Stats procedure for the resampling distribution of the correlation coefficient using StatCrunch. The StatCrunch expression cor(x,y) is similar to Excel's =CORREL(Range1,Range2) function. Where the Excel function =CORREL(Range1,Range2) calculates the correlation coefficient between the data in Range1 and Range2, the StatCrunch expression cor(x,y) calculates the correlation coefficient between data in columns x and y. Here is the procedure:

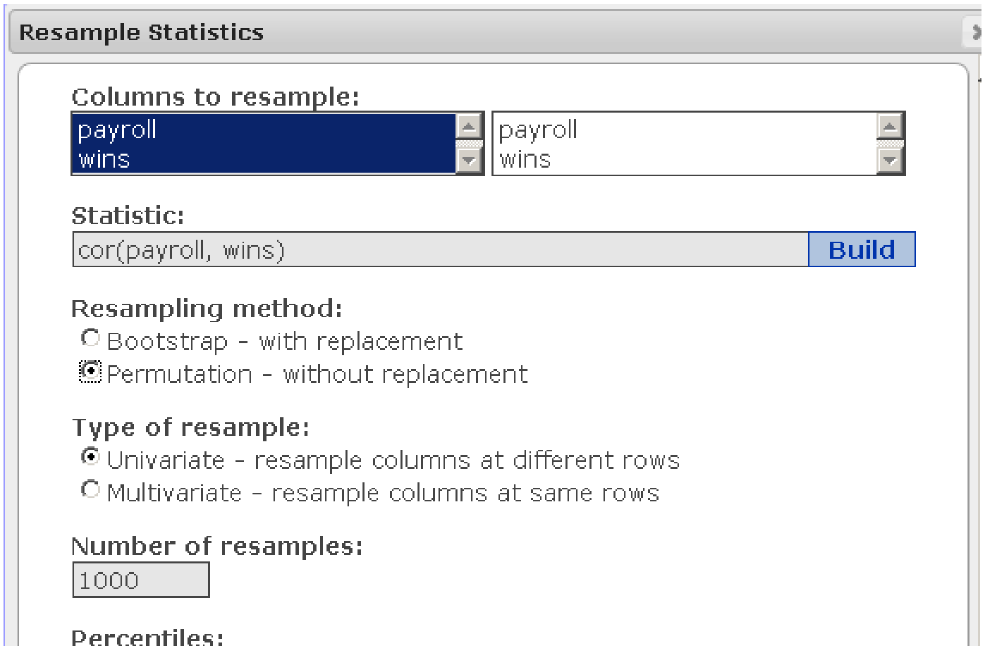
1) Start StatCrunch and load or paste the baseball payroll data into the worksheet (figure 10-22).



payroll	wins
216.1	279
146.66	276
127.4	273
119.57	254
115.16	248
114.4	282
110.05	227
108.16	257
103.58	251
102.6	266
101.81	246
101.69	240
94.2	219
92.44	234
88.62	255

Figure 10-22: Baseball payroll data

2) Select Stat>Resample>Statistic. Choose both the payroll and wins columns as Columns to resample. In the Statistic edit box, enter the expression cor(payroll,wins). Choose the Permutation - without replacement option. Keep all other default settings and click Compute! (figure 10-23).



Columns to resample: payroll, wins

Statistic: cor(payroll, wins)

Resampling method:
○ Bootstrap - with replacement
● Permutation - without replacement

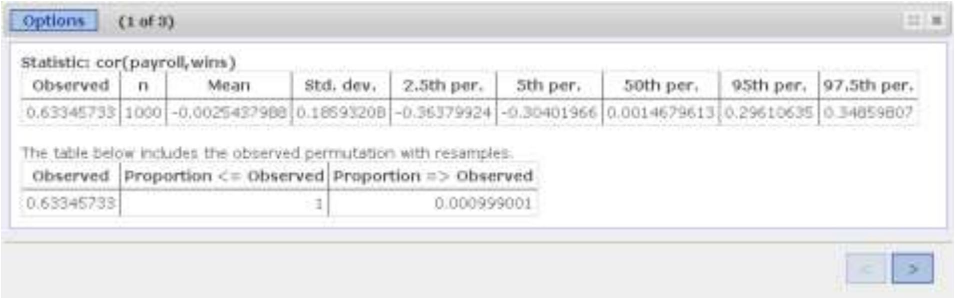
Type of resample:
● Univariate - resample columns at different rows
○ Multivariate - resample columns at same rows

Number of resamples: 1000

Percentiles:

Figure 10-23: Resample Statistic dialog

3) The result of one experiment of 1000 resamples is shown in figure 10-24.



Statistic: cor(payroll,wins)

Observed	n	Mean	Std. dev.	2.5th per.	5th per.	50th per.	95th per.	97.5th per.
0.63345733	1000	-0.0025437988	0.18593208	-0.36379924	-0.30401966	0.0014679613	0.29610635	0.34859807

The table below includes the observed permutation with resamples.

Observed	Proportion <= Observed	Proportion => Observed
0.63345733	1	0.000999001

Figure 10-24: Result of one experiment

4) This tells you that the observed correlation was 0.633, and that, in 1000 resamples (permutations), the resampled correlation was => the observed value less than 0.099% of the time (0.00099 is the estimated p-value). So the observed value is not likely something that happened by chance, and is statistically significant.

5) In addition, the 90% confidence interval is from the 5th percentile to the 95th percentile and is consistent with the Resampling Stats result. Click the > button to view the histogram (figure 10-25).

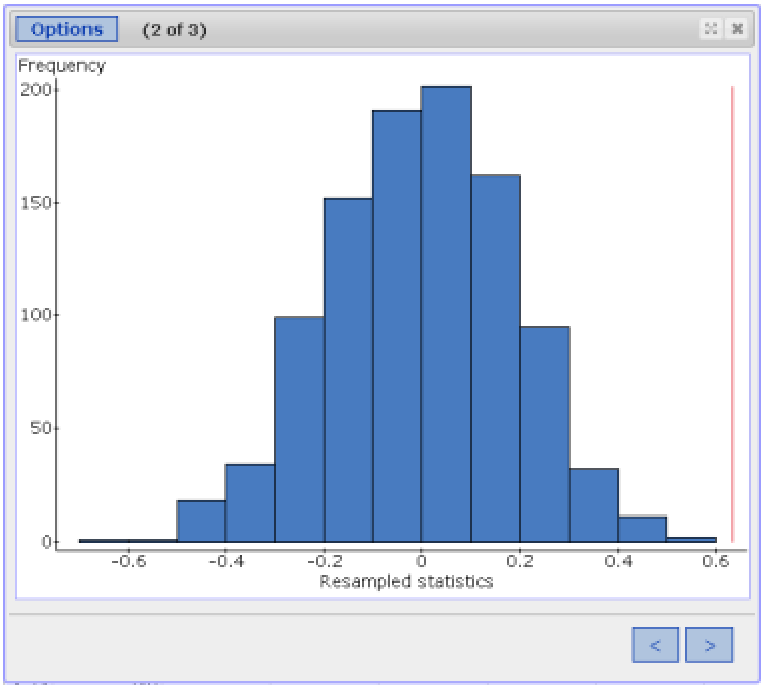


Figure 10-25: Histogram of the distribution of the confidence intervals (resampling)

6) Note that the calculated correlation coefficient of 0.633 (red vertical line) is outside the boundary of this histogram.

11 Regression

11.4 Inference for Regression

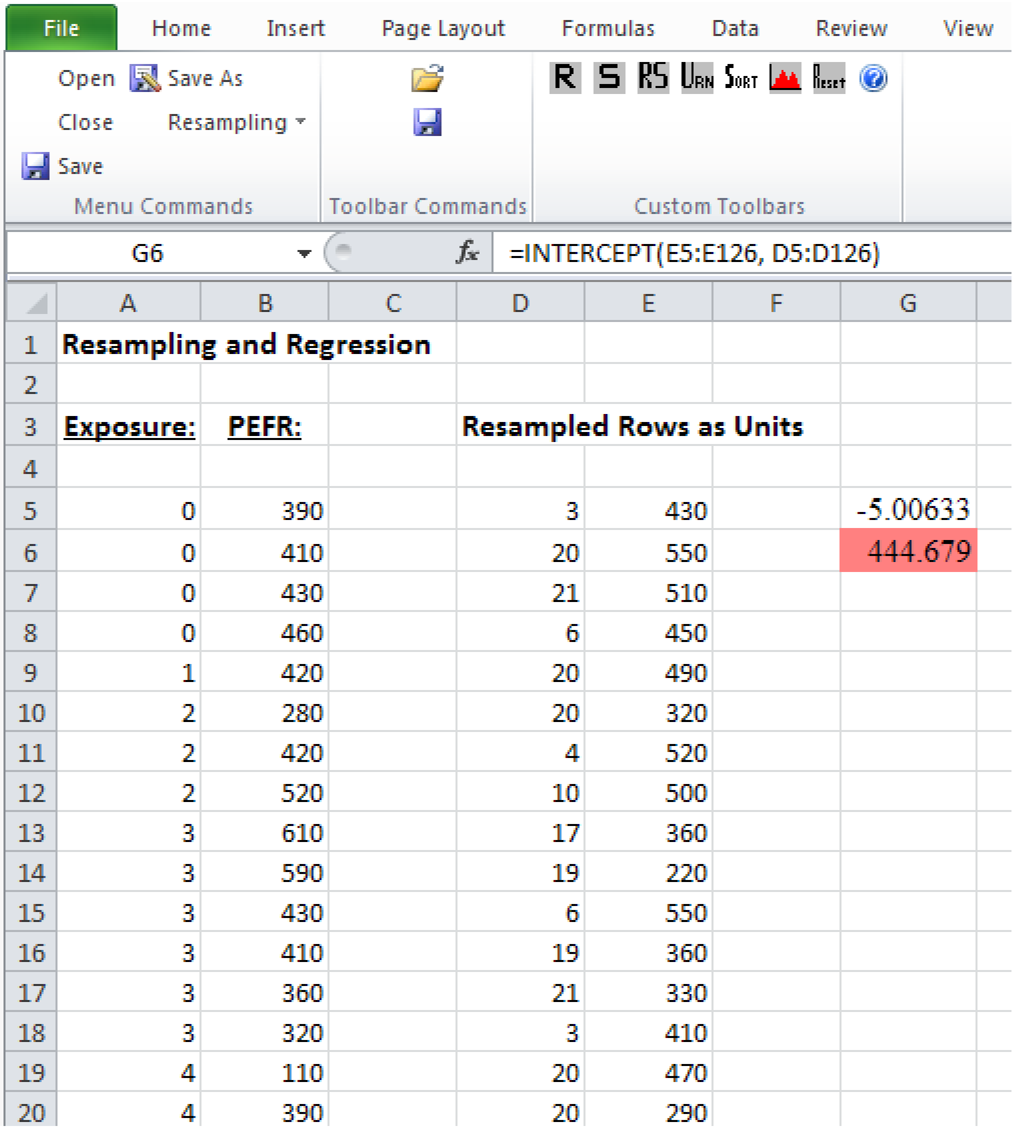
Resampling Stats

1) Click <here> to download the LungDisease.xlsx workbook.

2) Select cell A5 and choose “Resample” from the Resampling Stats menu or toolbar, and select the “Resample Rows as Units” option. (This option causes resampling to proceed on the basis of rows – when a selection is made for the resample, it is of a whole row as a unit, not individual elements separately). The input range should be both columns of Exposure and PEFR data (cells $A$5:$B$126 in figure 11-1). The Top Left cell of the output range should be cell D5 in this particular worksheet example.

3) To see how a line-fitting analysis works in Excel, select empty cells to place the functions SLOPE and INTERCEPT. That is, enter =SLOPE(E5:E126, D5:D126) into cell G5 and enter =INTERCEPT(E5:E126, D5:D126) into cell G6, as shown in Figure 11-1. These commands mean “find the slope and intercept of the regression line fitted to the data in the referenced range.”

4) Repeat and Score 1000 trials on the cell for INTERCEPT.

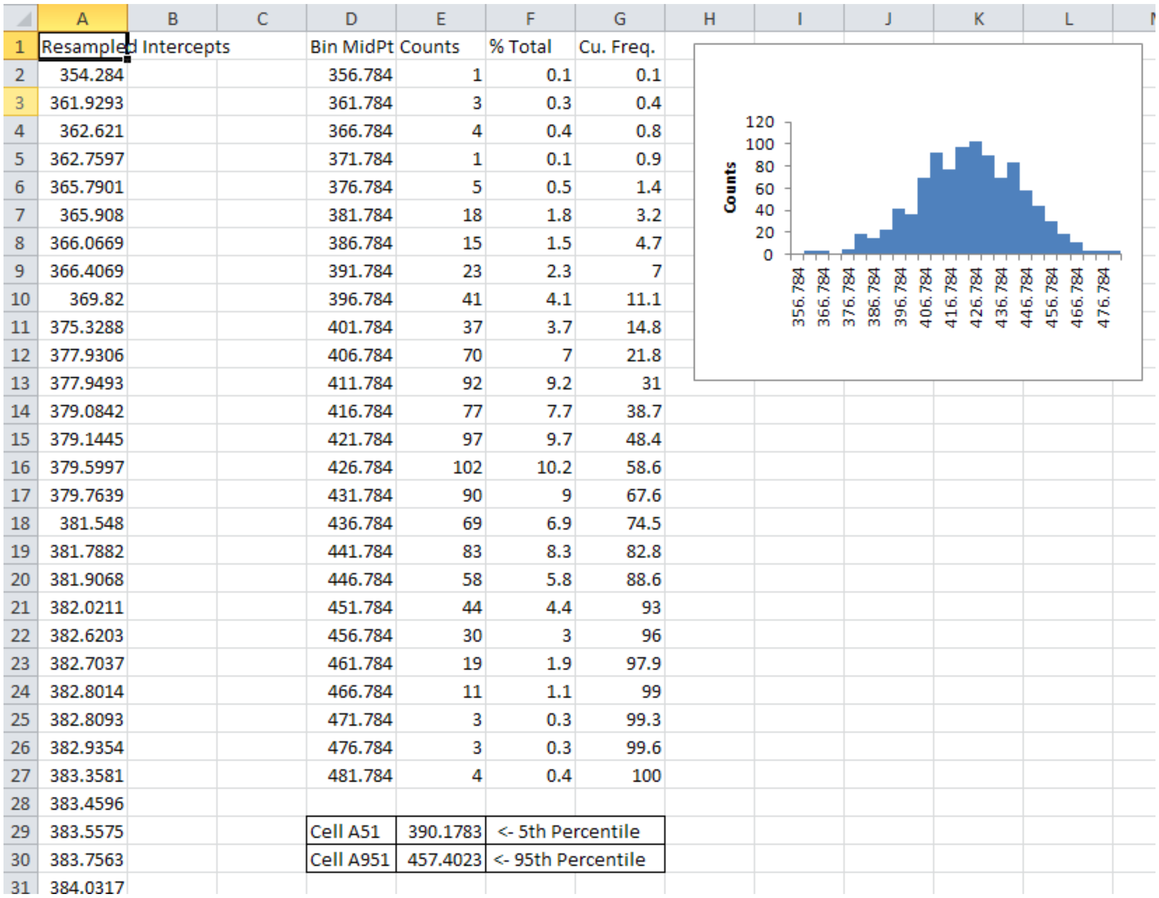


G6 =INTERCEPT(E5:E126, D5:D126)

	A	B	C	D	E	F	G
1	**Resampling and Regression**						
2							
3	**Exposure:**	**PEFR:**		**Resampled Rows as Units**			
4							
5	0	390		3	430		-5.00633
6	0	410		20	550		444.679
7	0	430		21	510		
8	0	460		6	450		
9	1	420		20	490		
10	2	280		20	320		
11	2	420		4	520		
12	2	520		10	500		
13	3	610		17	360		
14	3	590		19	220		
15	3	430		6	550		
16	3	410		19	360		
17	3	360		21	330		
18	3	320		3	410		
19	4	110		20	470		
20	4	390		20	290		

Figure 11-1: Regression via resampling – revisit the PEFR data

5) If you analyze the Results sheet (with sorted output values) for the 1000 trials with the Histogram function and check the Cumulative Frequency Chart check box, you'll see something like the output in Figure 11-2. It certainly indicates that the computed intercept (in the "Bin MidPt" column) has a wide range of values over the y-axis (PEFR) for the resampled data sets. Also notice the 90% confidence interval (5th and 95th Percentile values). You can compare these results to the results of standard regression analysis with Excel's Regression routine (under Tools/Data Analysis), although it also would be interesting to invest a few minutes in 10,000 or 100,000 trials for the resampling procedure. You should try the simulation again using the SLOPE value as a statistic of interest.



	A	B	C	D	E	F	G
1	Resampled Intercepts			Bin MidPt	Counts	% Total	Cu. Freq.
2	354.284			356.784	1	0.1	0.1
3	361.9293			361.784	3	0.3	0.4
4	362.621			366.784	4	0.4	0.8
5	362.7597			371.784	1	0.1	0.9
6	365.7901			376.784	5	0.5	1.4
7	365.908			381.784	18	1.8	3.2
8	366.0669			386.784	15	1.5	4.7
9	366.4069			391.784	23	2.3	7
10	369.82			396.784	41	4.1	11.1
11	375.3288			401.784	37	3.7	14.8
12	377.9306			406.784	70	7	21.8
13	377.9493			411.784	92	9.2	31
14	379.0842			416.784	77	7.7	38.7
15	379.1445			421.784	97	9.7	48.4
16	379.5997			426.784	102	10.2	58.6
17	379.7639			431.784	90	9	67.6
18	381.548			436.784	69	6.9	74.5
19	381.7882			441.784	83	8.3	82.8
20	381.9068			446.784	58	5.8	88.6
21	382.0211			451.784	44	4.4	93
22	382.6203			456.784	30	3	96
23	382.7037			461.784	19	1.9	97.9
24	382.8014			466.784	11	1.1	99
25	382.8093			471.784	3	0.3	99.3
26	382.9354			476.784	3	0.3	99.6
27	383.3581			481.784	4	0.4	100
28	383.4596						
29	383.5575			Cell A51	390.1783	<- 5th Percentile	
30	383.7563			Cell A951	457.4023	<- 95th Percentile	
31	384.0317						

Figure 11-2: Analyzing PEFR regression intercept output

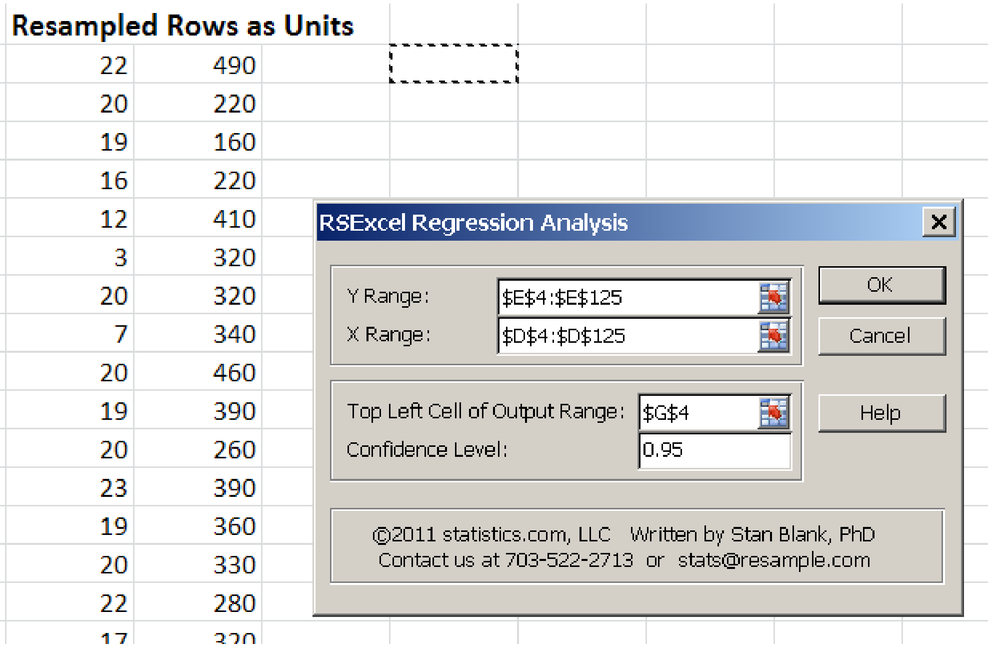
PEFR Again: Running Regression from the Resampling Add-in

1) Select the data in both columns with a standard click-drag (click <here> to download LungDisease.xlsx).

2) Choose "Resample" from the Resampling Stats menu or toolbar, and select the "Resample Rows as Units" option. (This option causes resampling to proceed on the basis of rows – when a selection is made for the resample, it is of a whole row as a unit, not individual elements separately.) In this case, use D4 as the top left output cell.

3) From the Resampling menu (the Add-ins menu or right-click a cell for the cell menu) choose "Regression." If we call the Regression function from the Resampling menu, we'll be asked to identify the x-range, the y-range, and the beginning output cell and a confidence interval (Figure 11-3).

4) Click "OK" and the output in Figure 11-4 will display in the Excel worksheet.



Resampled Rows as Units

22	490
20	220
19	160
16	220
12	410
3	320
20	320
7	340
20	460
19	390
20	260
23	390
19	360
20	330
22	280
17	320

Y Range: $E$4:$E$125
X Range: $D$4:$D$125

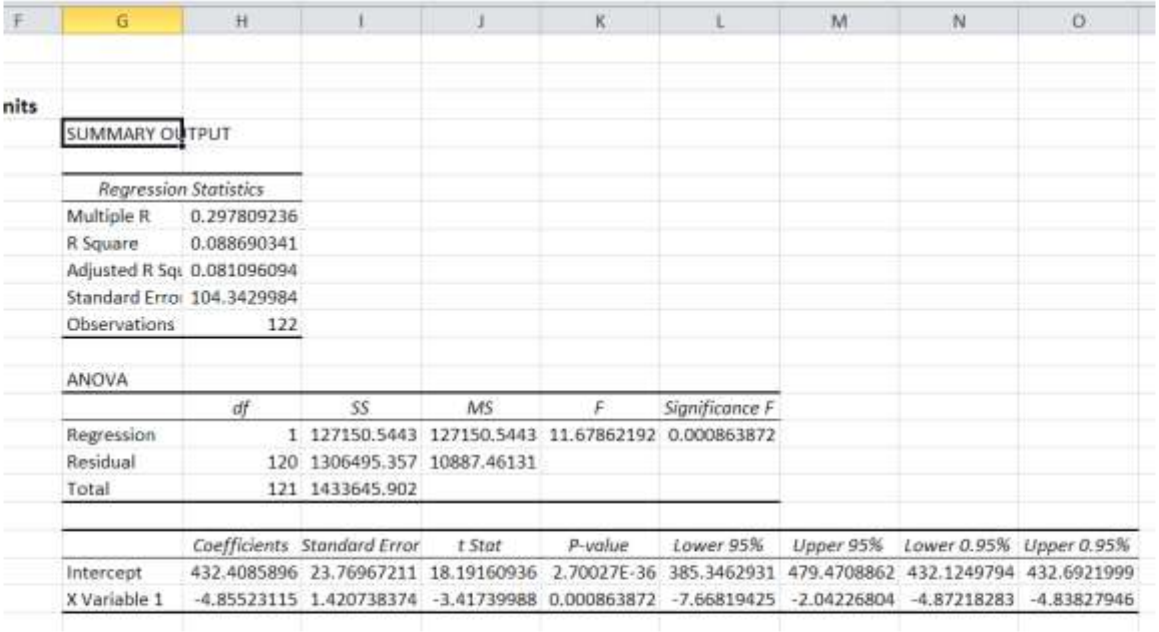
Top Left Cell of Output Range: $G$4
Confidence Level: 0.95

©2011 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-2713 or stats@resample.com

Figure 11-3: X-Y input for the Resampling menu regression option

Note: This confidence interval specification in regression is a conventional (non-resampling) confidence interval needed as an input to Excel's regression routine; it is NOT related to the confidence interval we will be developing through Repeat & Score. Think of it as a meaningless number you must fill in for the regression routine to work.

5) The output from one experiment is shown in figure 11-4.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.297809236
R Square	0.088690341
Adjusted R Squ	0.081096094
Standard Error	104.3429984
Observations	122

ANOVA

	df	SS	MS	F	Significance F
Regression	1	127150.5443	127150.5443	11.67862192	0.000863872
Residual	120	1306495.357	10887.46131		
Total	121	1433645.902			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 0.95%	Upper 0.95%
Intercept	432.4085896	23.76967211	18.19160936	2.70027E-36	385.3462931	479.4708862	432.1249794	432.6921999
X Variable 1	-4.85523115	1.420738374	-3.41739988	0.000863872	-7.66819425	-2.04226804	-4.87218283	-4.83827946

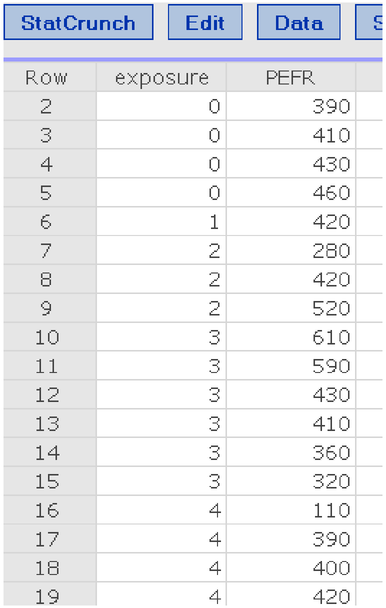
Figure 11-4: Output of Excel regression (in the Data Analysis Toolpak)

Note: The regression equation for the above experiment can be formed from the Intercept and X Variable 1 Coefficients at the bottom of the output. In this example, the regression equation (rounded) is:

y = -4.855 x + 432.41

StatCrunch

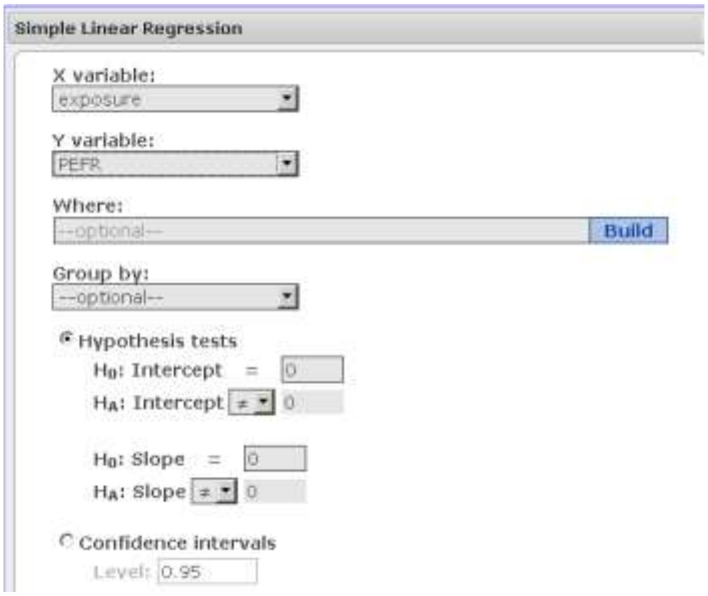
This section illustrates calculating a regression equation using StatCrunch. Load the PEFR data into StatCrunch (copy/paste) and label the columns as shown in figure 11-5:



Row	exposure	PEFR
2	0	390
3	0	410
4	0	430
5	0	460
6	1	420
7	2	280
8	2	420
9	2	520
10	3	610
11	3	590
12	3	430
13	3	410
14	3	360
15	3	320
16	4	110
17	4	390
18	4	400
19	4	420

Figure 11-5: PEFR data

1) Select Stat>Regression>Simple Linear. Choose the exposure column for the X variable and PEFR for the Y variable (figure 11-6). Click Compute!


Simple Linear Regression

X variable: exposure

Y variable: PEFR

Where: --optional--

Group by: --optional--

● Hypothesis tests

H_0: Intercept = 0

H_A: Intercept ≠ 0

H_0: Slope = 0

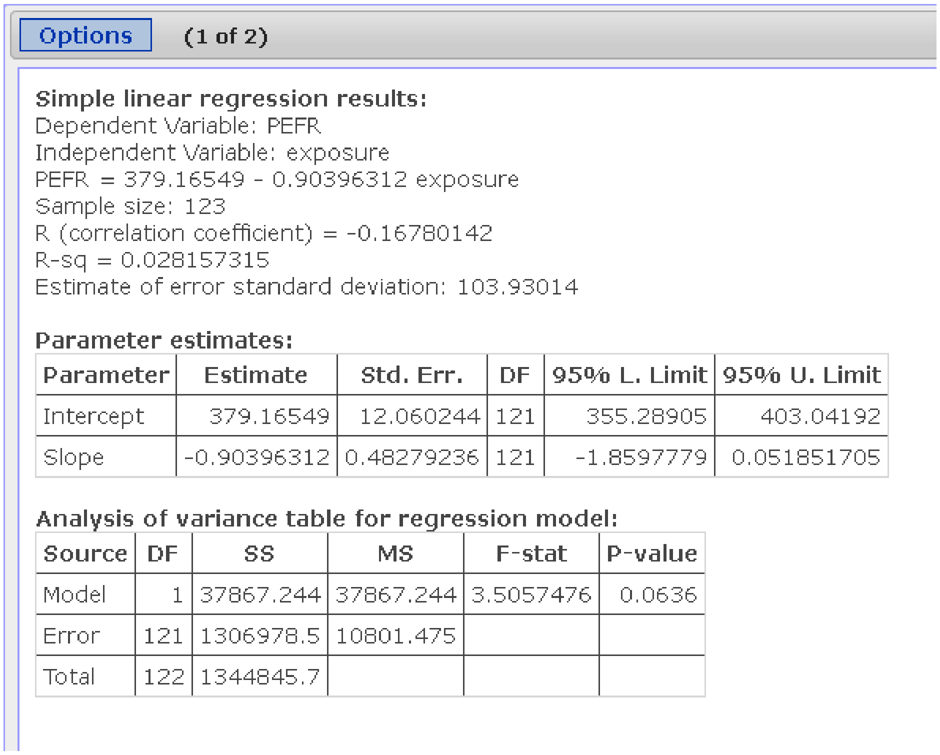
H_A: Slope ≠ 0

○ Confidence intervals

Level: 0.95

Figure 11-6: StatCrunch simple linear regression

2) The results are shown in figure 11-7



Simple linear regression results:
Dependent Variable: PEFR
Independent Variable: exposure
PEFR = 379.16549 - 0.90396312 exposure
Sample size: 123
R (correlation coefficient) = -0.16780142
R-sq = 0.028157315
Estimate of error standard deviation: 103.93014

Parameter estimates:

Parameter	Estimate	Std. Err.	DF	95% L. Limit	95% U. Limit
Intercept	379.16549	12.060244	121	355.28905	403.04192
Slope	-0.90396312	0.48279236	121	-1.8597779	0.051851705

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	37867.244	37867.244	3.5057476	0.0636
Error	121	1306978.5	10801.475		
Total	122	1344845.7			

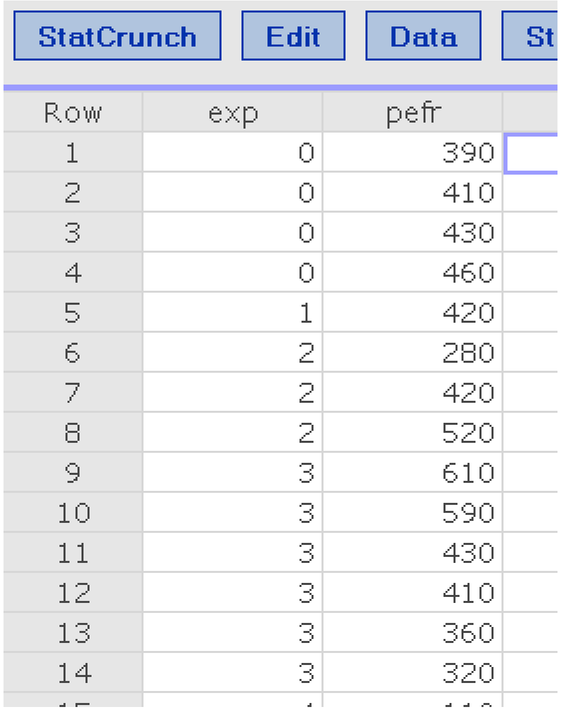
Figure 11-7: PEFR = 379.16549 – 0.90396312*exposure

3) The regression equation from StatCrunch is different from the Resampling Stats equation. Why? If you ran the Resampling Stats experiment again, would you get the same regression equation? How about the StatCrunch procedure? Finally, compare the confidence intervals in both Resampling Stats and StatCrunch. Are the CIs reasonably close?

StatCrunch (PEFR Lung Disease Linear Regression – Resampling Solution)

The StatCrunch resampling solution was prepared after some kind suggestions from Webster West, the author of StatCrunch.

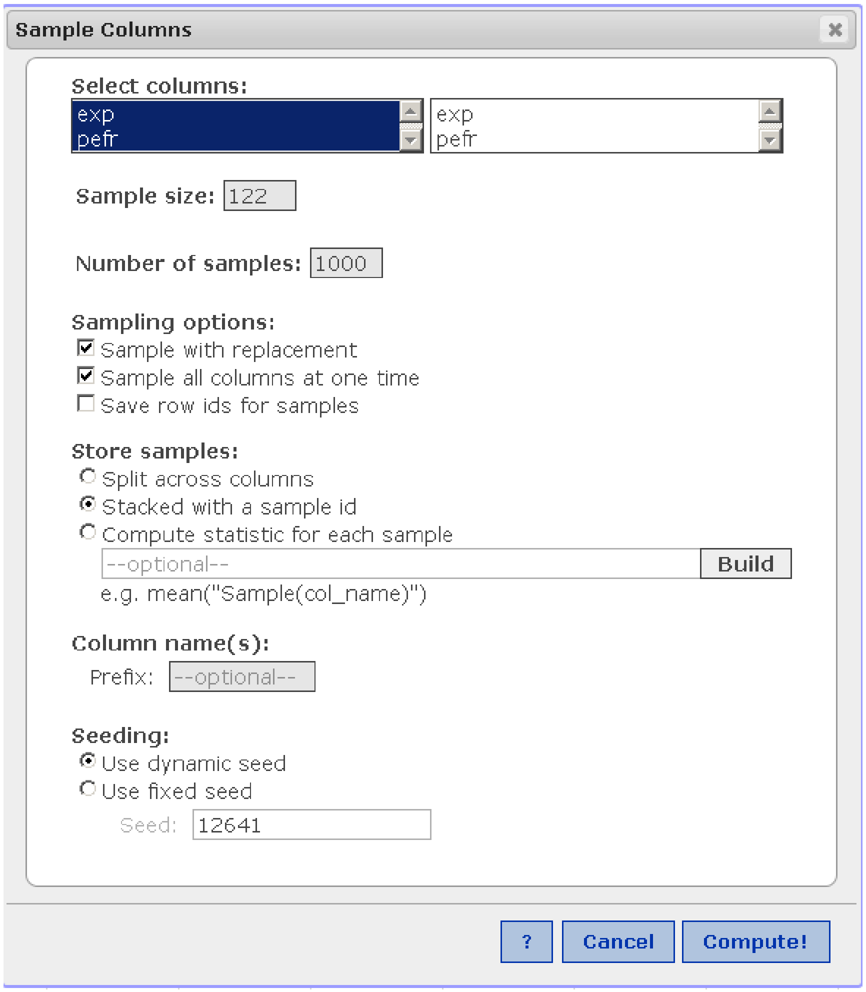
1) Enter the lung disease data into StatCrunch. The easiest method would probably be a copy/paste from the LungDisease.xlsx workbook, which you can download from the Resampling Stats subsection in this chapter. Figure 11-8 illustrates the data in StatCrunch.



Row	exp	pefr
1	0	390
2	0	410
3	0	430
4	0	460
5	1	420
6	2	280
7	2	420
8	2	520
9	3	610
10	3	590
11	3	430
12	3	410
13	3	360
14	3	320

Figure 11-8: PEFR data in StatCrunch

2) Select Data > Sample. Select both columns, enter a Sample Size of 122 and 1000 for the Number of samples. Check Sample with replacement and also check Sample all columns at one time. In the Store Samples section, select Stacked with a sample id. Figure 11-9 illustrates the dialog with options selected. Click Compute!



Sample Columns

Select columns:
exp
pefr

Sample size: 122

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☑ Sample all columns at one time
☐ Save row ids for samples

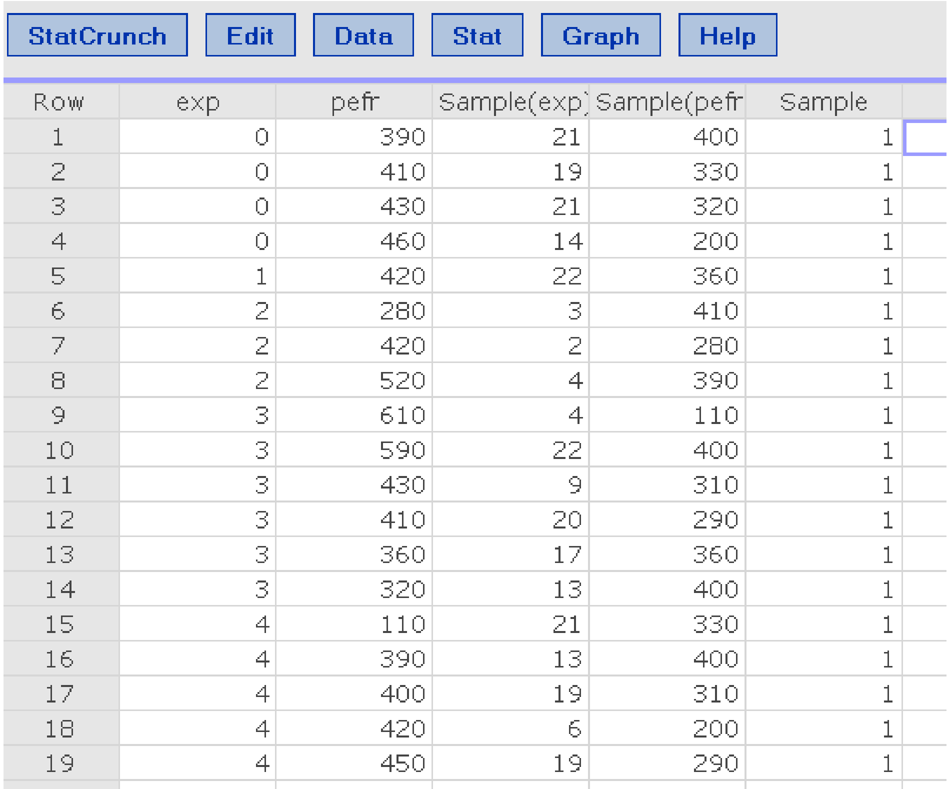
Store samples:
○ Split across columns
● Stacked with a sample id
○ Compute statistic for each sample
--optional-- Build
e.g. mean("Sample(col_name)")

Column name(s):
Prefix: --optional--

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 11-9: Sample columns dialog

3) This method of resampling will add 122*1000 (122,000) rows and 3 new columns to the StatCrunch worksheet. If you scroll downward, you can see that the Sample column contains sample IDs for each of the 1000 resamples. The IDs range from 1 to 1000, with 122 row entries associated with each ID. Figure 11-10 shows the top few rows of the revised worksheet. All of the 1000 resamples have been stacked in the worksheet, with each resample of 122 rows assigned a numerical ID from 1 to 1000 in order.

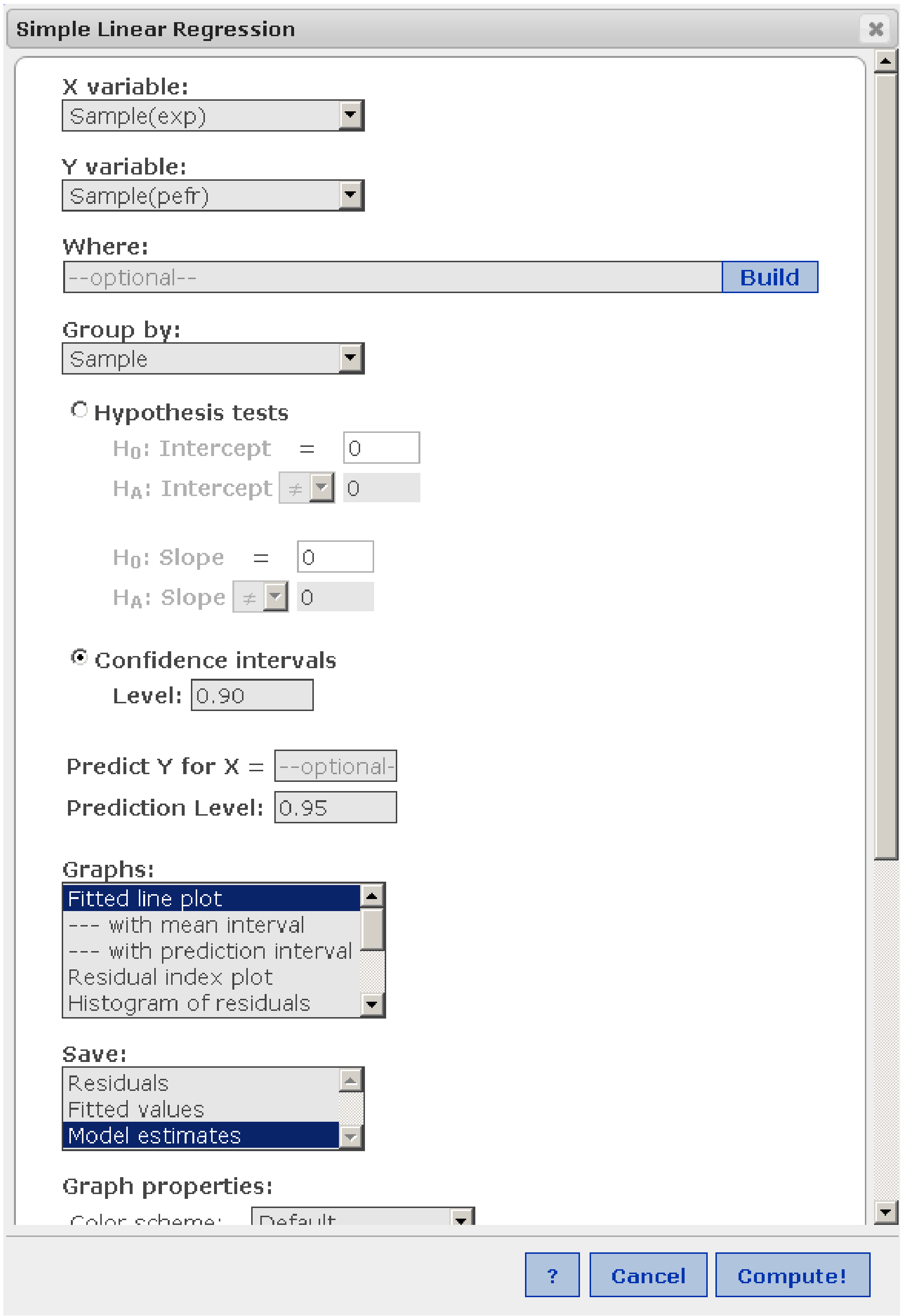


Row	exp	pefr	Sample(exp)	Sample(pefr	Sample
1	0	390	21	400	1
2	0	410	19	330	1
3	0	430	21	320	1
4	0	460	14	200	1
5	1	420	22	360	1
6	2	280	3	410	1
7	2	420	2	280	1
8	2	520	4	390	1
9	3	610	4	110	1
10	3	590	22	400	1
11	3	430	9	310	1
12	3	410	20	290	1
13	3	360	17	360	1
14	3	320	13	400	1
15	4	110	21	330	1
16	4	390	13	400	1
17	4	400	19	310	1
18	4	420	6	200	1
19	4	450	19	290	1

Figure 11-10: New columns and rows added

4) Now for the regression! Select Stat > Regression > Simple Linear. Choose the Sample(exp) column for the X-Variable and Sample(pefr) for the Y-variable. Group by the Sample column and choose the Confidence intervals option (enter 0.90). Under Save, choose Model estimates (figure 11-11). Click Compute!

5) You will receive a warning message such as “Whoa… lot’s of values… do you want to bin them?” (or something similar). You don’t want to bin the values, so click Cancel. You should, after a short wait, see something like figure 11-12.



Simple Linear Regression

X variable: Sample(exp)

Y variable: Sample(pefr)

Where: --optional--

Group by: Sample

○ Hypothesis tests

H_0: Intercept = 0

H_A: Intercept ≠ 0

H_0: Slope = 0

H_A: Slope ≠ 0

● Confidence intervals

Level: 0.90

Predict Y for X = --optional-

Prediction Level: 0.95

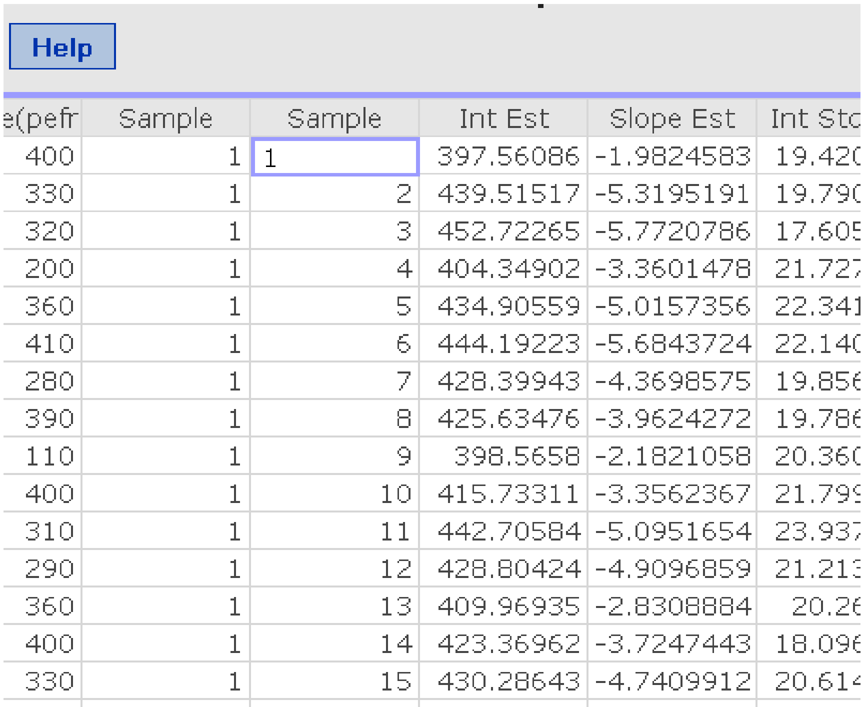
Graphs: Fitted line plot

Save: Model estimates

Graph properties:

Color scheme: Default

Figure 11-11: Regression dialog for PEFR

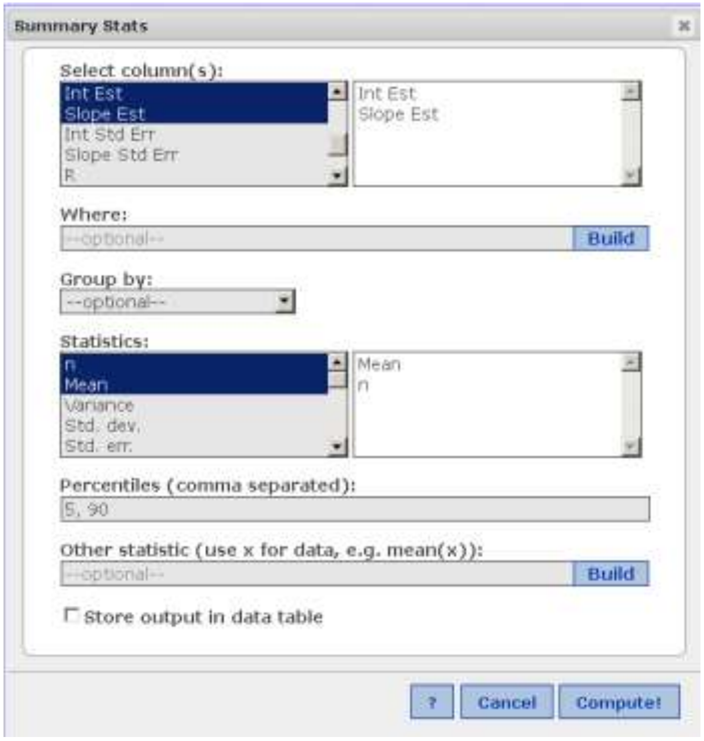


e(pefr	Sample	Sample	Int Est	Slope Est	Int St
400	1	1	397.56086	-1.9824583	19.42
330	1	2	439.51517	-5.3195191	19.79
320	1	3	452.72265	-5.7720786	17.60
200	1	4	404.34902	-3.3601478	21.72
360	1	5	434.90559	-5.0157356	22.341
410	1	6	444.19223	-5.6843724	22.14
280	1	7	428.39943	-4.3698575	19.85
390	1	8	425.63476	-3.9624272	19.78
110	1	9	398.5658	-2.1821058	20.36
400	1	10	415.73311	-3.3562367	21.79
310	1	11	442.70584	-5.0951654	23.93
290	1	12	428.80424	-4.9096859	21.21
360	1	13	409.96935	-2.8308884	20.2
400	1	14	423.36962	-3.7247443	18.09
330	1	15	430.28643	-4.7409912	20.61

Figure 11-12: Output from linear regression

6) The Int Est and Slope Est columns contain 1000 resampled intercepts and slopes. Use Stat > Summary Stats > Column and select the Int Est and Slope Est columns. Click n and mean in the Statistics section and enter 5, 95 for the percentiles (see figure 11-13). Click Compute!

7) The results are displayed in figure 11-14.



Select column(s): Int Est, Slope Est

Where: --optional--

Group by: --optional--

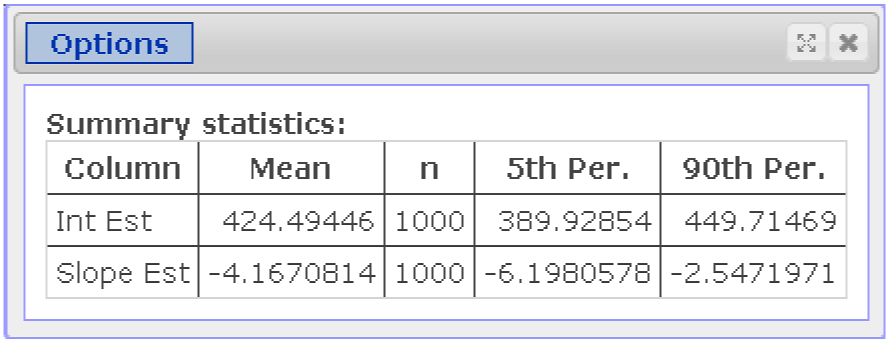
Statistics: Mean, n

Percentiles (comma separated): 5, 90

Other statistic (use x for data, e.g. mean(x)): --optional--

☐ Store output in data table

Figure 11-13: Summary stats including 90% CI



Summary statistics:

Column	Mean	n	5th Per.	90th Per.
Int Est	424.49446	1000	389.92854	449.71469
Slope Est	-4.1670814	1000	-6.1980578	-2.5471971

Figure 11-14: Regression by resampling results

8) The results are comparable to Resampling Stats

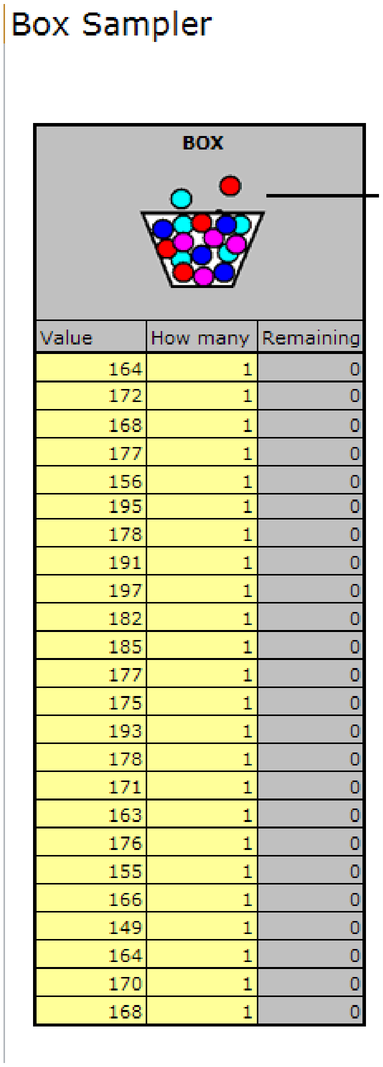
12 Analysis of Variance - ANOVA

12.3 A Single Test

Box Sampler

Box Sampler can be used to solve the donut fat problem.

1) Enter all 24 fat values in the box as shown in figure 12-1.



Box Sampler

BOX

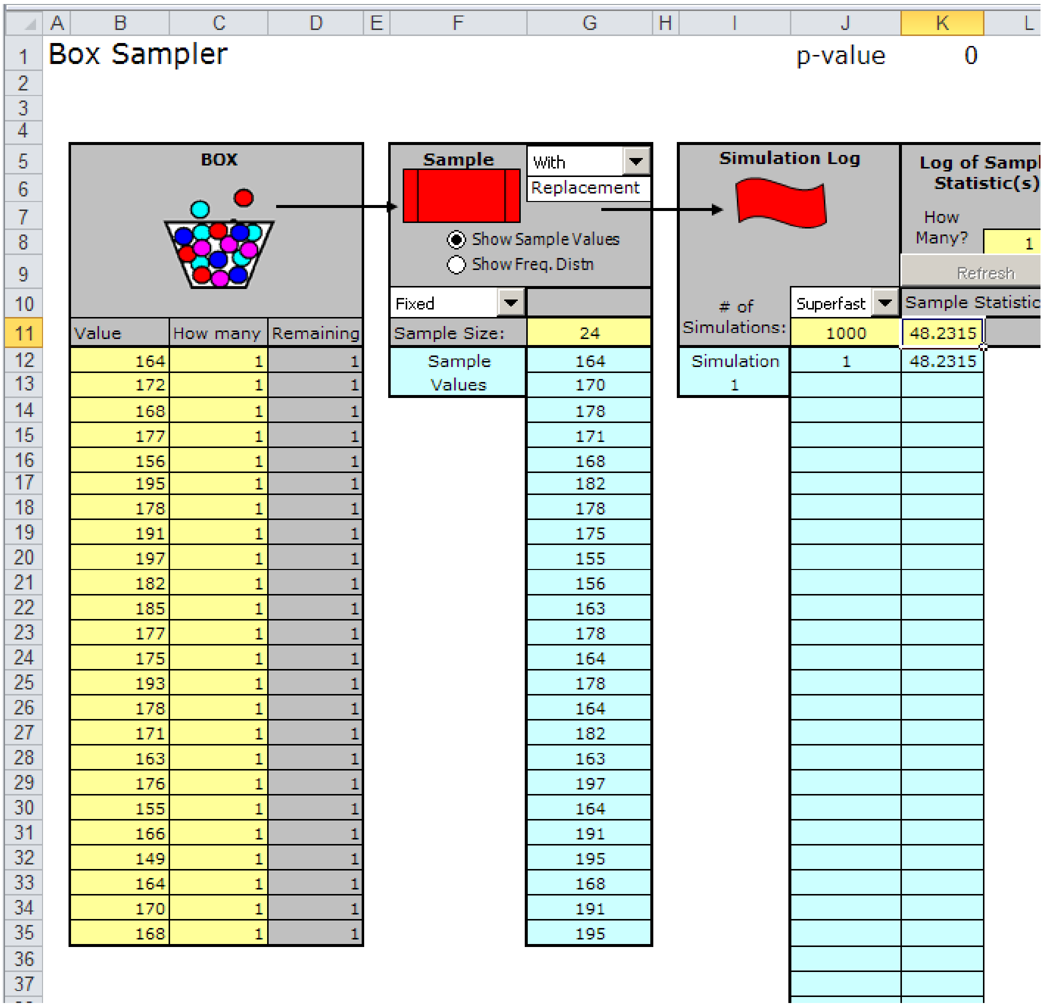
Value	How many	Remaining
164	1	0
172	1	0
168	1	0
177	1	0
156	1	0
195	1	0
178	1	0
191	1	0
197	1	0
182	1	0
185	1	0
177	1	0
175	1	0
193	1	0
178	1	0
171	1	0
163	1	0
176	1	0
155	1	0
166	1	0
149	1	0
164	1	0
170	1	0
168	1	0

Figure 12-1: Box Sampler setup for the donut fat example

2) The Sample Size is 24 and use 1000 for the # of Simulations. Enter the following function should be entered in the Sample Statistics cell:

=VAR.S(AVERAGE(G12:G17),AVERAGE(G18:G23),AVERAGE(G24:G29),AVERAGE(G30:G35))

Figure 12-2 illustrates the Box Sampler worksheet with one trial.



Box Sampler

p-value 0

Sample: With Replacement

● Show Sample Values
○ Show Freq. Distn

Fixed

Sample Size: 24

# of Simulations: 1000

Superfast

How Many? 1

Sample Statistic: 48.2315

Row	Value	How many	Remaining	Sample Values
12	164	1	1	164
13	172	1	1	170
14	168	1	1	178
15	177	1	1	171
16	156	1	1	168
17	195	1	1	182
18	178	1	1	178
19	191	1	1	175
20	197	1	1	155
21	182	1	1	156
22	185	1	1	163
23	177	1	1	178
24	175	1	1	164
25	193	1	1	178
26	178	1	1	164
27	171	1	1	182
28	163	1	1	163
29	176	1	1	197
30	155	1	1	164
31	166	1	1	191
32	149	1	1	195
33	164	1	1	168
34	170	1	1	191
35	168	1	1	195

Simulation		Sample Statistic
Simulation 1	1	48.2315

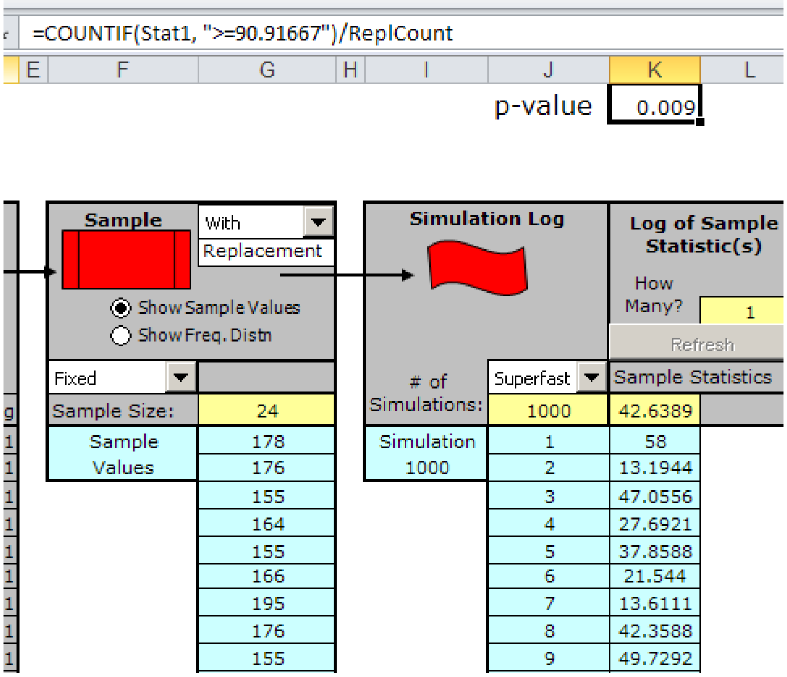
Figure 12-2: Donut fat example after 1 simulation

3) To calculate the p-value, enter:

=COUNTIF(Stat1, ">=90.91667")/ReplCount

in cell K1.

Figure 12-3 shows the results of one 1000 simulation experiment.



```
=COUNTIF(Stat1, ">=90.91667")/ReplCount
```

p-value 0.009

Sample: With Replacement

● Show Sample Values
○ Show Freq. Distn

Fixed

Sample Size: 24

Simulation Log

# of Simulations: Superfast 1000

Log of Sample Statistic(s)

How Many? 1

Sample Statistics: 42.6389

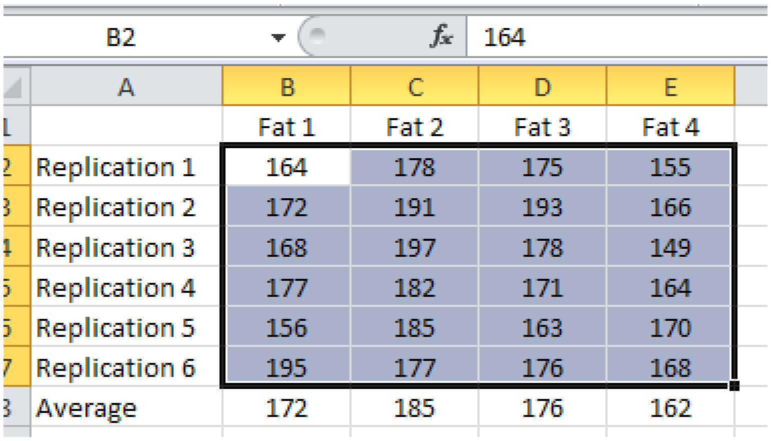
Sample Values	Simulation 1000	Sample Statistics
178	1	58
176	2	13.1944
155	3	47.0556
164	4	27.6921
155	5	37.8588
166	6	21.544
195	7	13.6111
176	8	42.3588
155	9	49.7292

Figure 12-3: Box Sampler donut fat results for 1000 simulations

Resampling Procedure

Download the donut_fat.xlsx file <here>. This workbook contains the table form of the donutHT.xls data.

1) Select the data area, cells B2:E7 (figure 12-4).

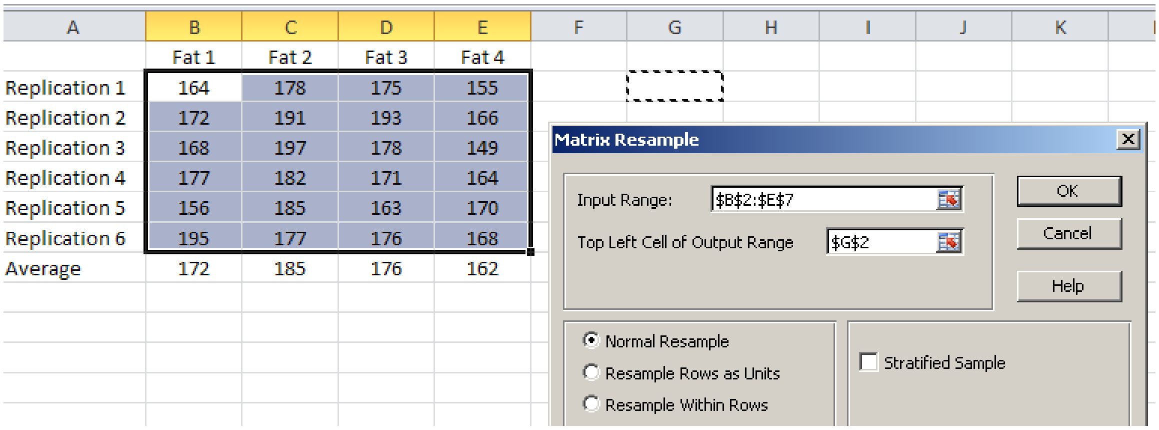


B2 = 164

	Fat 1	Fat 2	Fat 3	Fat 4
Replication 1	164	178	175	155
Replication 2	172	191	193	166
Replication 3	168	197	178	149
Replication 4	177	182	171	164
Replication 5	156	185	163	170
Replication 6	195	177	176	168
Average	172	185	176	162

Figure 12-4: Select the 24 fat absorption values

2) Choose Resample from the Resampling Stats menu. Select cell G2 as the Top Left Cell and choose Normal Resample (figure 12-5). Click OK.



A	B	C	D	E
	Fat 1	Fat 2	Fat 3	Fat 4
Replication 1	164	178	175	155
Replication 2	172	191	193	166
Replication 3	168	197	178	149
Replication 4	177	182	171	164
Replication 5	156	185	163	170
Replication 6	195	177	176	168
Average	172	185	176	162

Matrix Resample

Input Range: $B$2:$E$7

Top Left Cell of Output Range $G$2

● Normal Resample

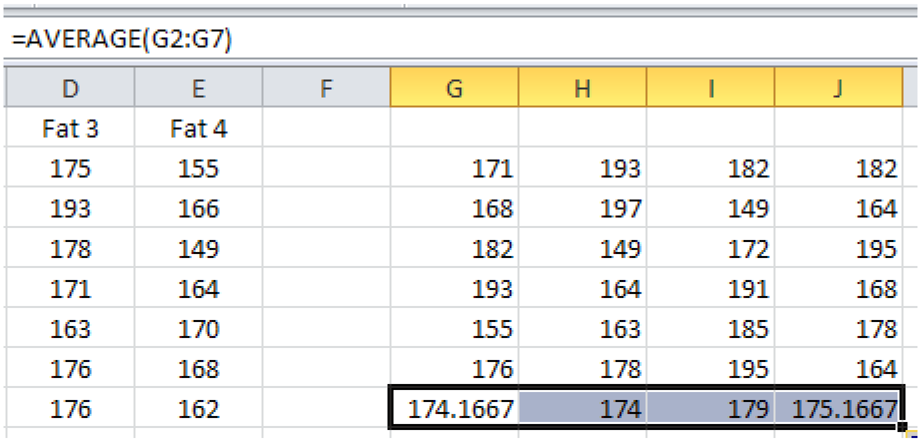
○ Resample Rows as Units

○ Resample Within Rows

☐ Stratified Sample

Figure 12-5: Matrix resample (with replacement)

3) Enter the formula =AVERAGE(G2:G7) in cell G8 and copy this formula to the right, stopping at cell J8 (figure 12-6).



=AVERAGE(G2:G7)

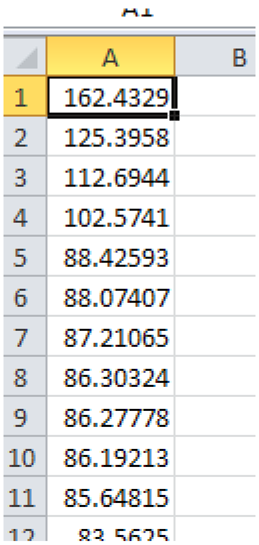
D	E	F	G	H	I	J
Fat 3	Fat 4					
175	155		171	193	182	182
193	166		168	197	149	164
178	149		182	149	172	195
171	164		193	164	191	168
163	170		155	163	185	178
176	168		176	178	195	164
176	162		174.1667	174	179	175.1667

Figure 12-6: Find the mean of each resample group

4) In cell K8, enter the formula =VAR.S(G8:J8) or =VAR(G8:J8) if you are using Excel 2003. These formulas calculate the sample variance of the means.

5) Select cell K8 and choose Repeat and Score. Enter 1000 iterations/trials and click OK.

6) Sort the data in the Results sheet from highest to lowest (figure 12-7).



	A	B
1	162.4329	
2	125.3958	
3	112.6944	
4	102.5741	
5	88.42593	
6	88.07407	
7	87.21065	
8	86.30324	
9	86.27778	
10	86.19213	
11	85.64815	
12	83.5625	

Figure 12-7: Four variances >= 90.92

7) There were 4 instances of the variance of the means being >=90.92. This is an estimated p-value = 0.004.

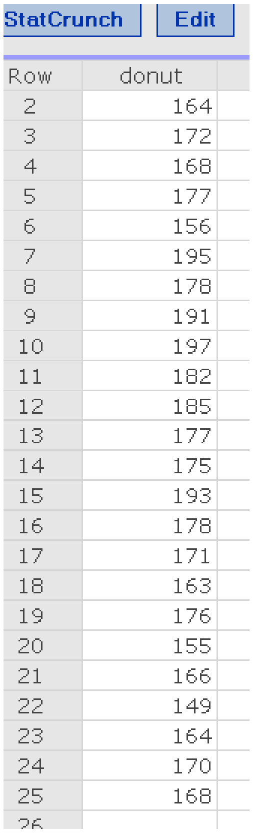
8) Create a histogram and compare it to figure 12-4 in the text.

NOTE: Sometimes it is the case that the various groups (e.g. Fat1, Fat2, etc. above) have different sizes. This does not matter in Resampling Stats - Simply select the range that contains the entire set of data, and Resampling Stats will preserve the sample sizes in the shuffled or resampled range. The rest of the procedure is the same.

StatCrunch Procedure

Download the donutHT.xls file <here>. This workbook contains the original donut weight data in column format. You can also enter the data manually from figure 12-8.

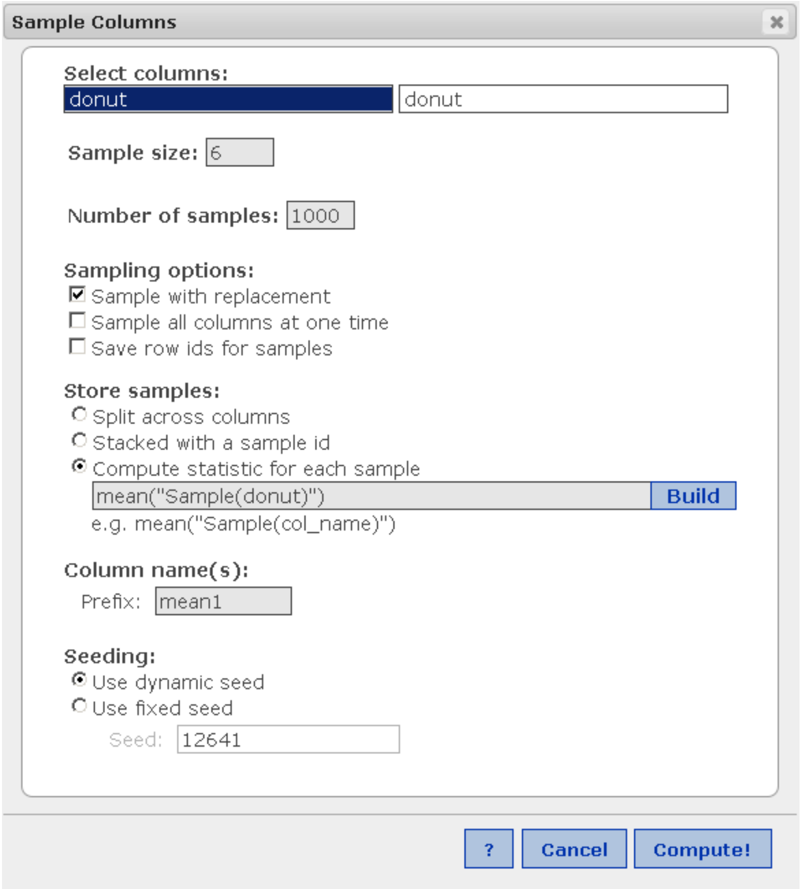
1) Open and copy/paste the original donut weight data (found in cells C3:C26) into the first column of the StatCrunch worksheet. Rename this column donut.



Row	donut
2	164
3	172
4	168
5	177
6	156
7	195
8	178
9	191
10	197
11	182
12	185
13	177
14	175
15	193
16	178
17	171
18	163
19	176
20	155
21	166
22	149
23	164
24	170
25	168
26	

Figure 12-8: Original donut weight data

2) Choose Data>Sample and select the donut column. Enter 6 for the Sample size and select Resample with replacement. The Statistic expression is: mean("Sample(donut)") . See figure 12-9 for details.



Sample Columns

Select columns: donut | donut

Sample size: 6

Number of samples: 1000

Sampling options:
☑ Sample with replacement
☐ Sample all columns at one time
☐ Save row ids for samples

Store samples:
○ Split across columns
○ Stacked with a sample id
● Compute statistic for each sample
mean("Sample(donut)")
e.g. mean("Sample(col_name)")

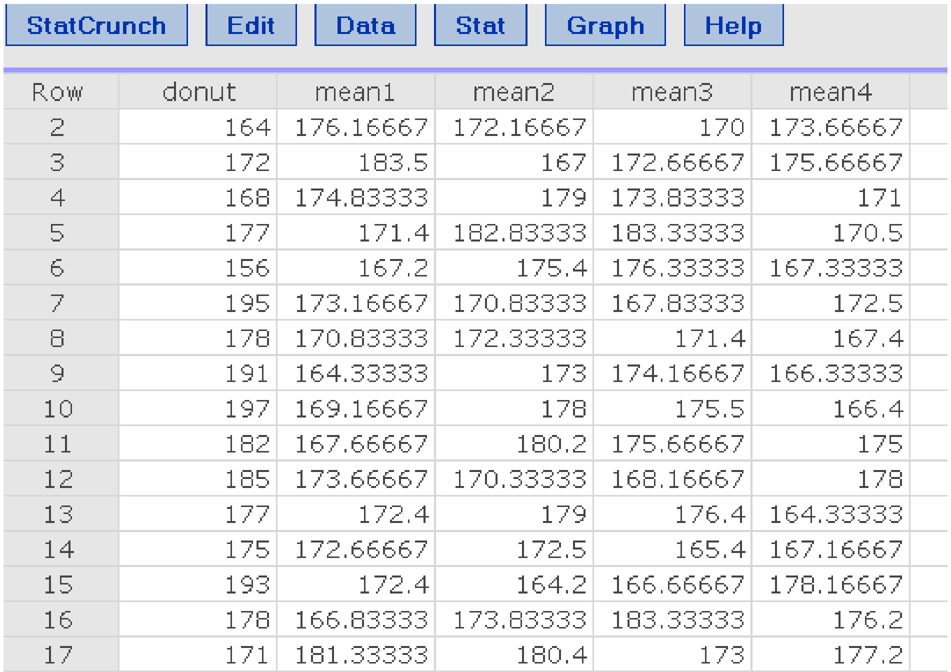
Column name(s):
Prefix: mean1

Seeding:
● Use dynamic seed
○ Use fixed seed
Seed: 12641

Figure 12-9: Sample columns

3) Name the new column mean1 and click Compute!

4) Repeat step 2 three more times, selecting the donut column for resampling and renaming the new columns mean2, mean3, and mean4 respectively.



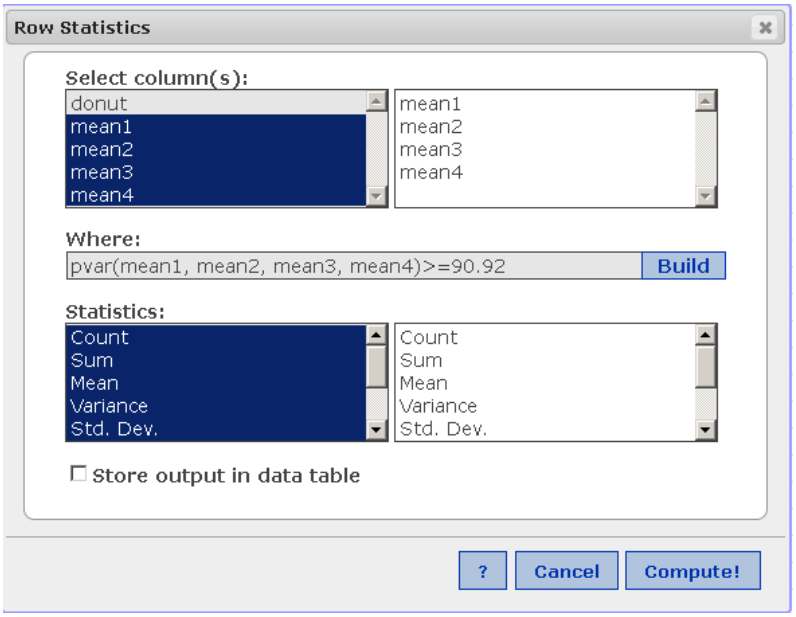
Row	donut	mean1	mean2	mean3	mean4
2	164	176.16667	172.16667	170	173.66667
3	172	183.5	167	172.66667	175.66667
4	168	174.83333	179	173.83333	171
5	177	171.4	182.83333	183.33333	170.5
6	156	167.2	175.4	176.33333	167.33333
7	195	173.16667	170.83333	167.83333	172.5
8	178	170.83333	172.33333	171.4	167.4
9	191	164.33333	173	174.16667	166.33333
10	197	169.16667	178	175.5	166.4
11	182	167.66667	180.2	175.66667	175
12	185	173.66667	170.33333	168.16667	178
13	177	172.4	179	176.4	164.33333
14	175	172.66667	172.5	165.4	167.16667
15	193	172.4	164.2	166.66667	178.16667
16	178	166.83333	173.83333	183.33333	176.2
17	171	181.33333	180.4	173	177.2

Figure 12-10: Output from the Sample columns procedures

5) To find the respective variances of the resampled means, select Stat>Summary Stats>Rows and choose the 4 columns: mean1, mean2, mean3, mean4. In the Where: edit box, enter:

   pvar(mean1, mean2, mean3, mean4)>=90.92

   This statement selects only the variances in the means greater than or equal to 90.92 (the variance in the means of the original data). The StatCrunch variance command pvar( ) operates on rows, rather than columns. Click Compute! (see figure 12-11).



Select column(s):

donut
mean1
mean2
mean3
mean4

mean1
mean2
mean3
mean4

Where: pvar(mean1, mean2, mean3, mean4)>=90.92

Statistics:

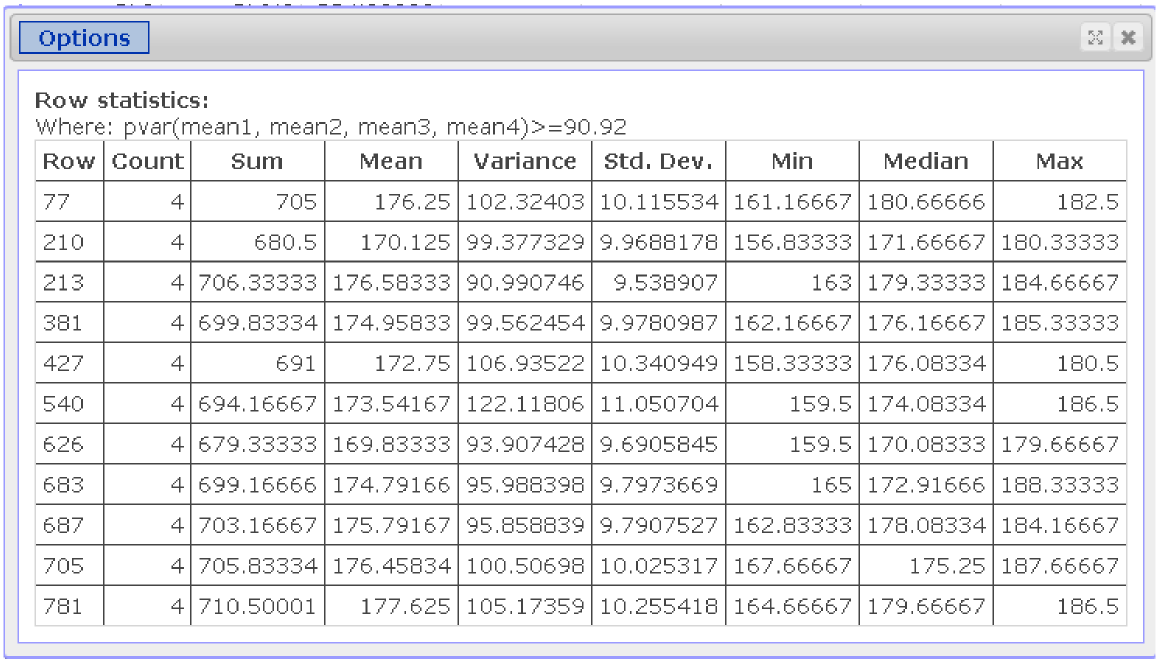
Count
Sum
Mean
Variance
Std. Dev.

Count
Sum
Mean
Variance
Std. Dev.

☐ Store output in data table

Figure 12-11: Row Statistics using pvar()

6) The figure below is the result of one simulation.



Row statistics:
Where: pvar(mean1, mean2, mean3, mean4)>=90.92

Row	Count	Sum	Mean	Variance	Std. Dev.	Min	Median	Max
77	4	705	176.25	102.32403	10.115534	161.16667	180.66666	182.5
210	4	680.5	170.125	99.377329	9.9688178	156.83333	171.66667	180.33333
213	4	706.33333	176.58333	90.990746	9.538907	163	179.33333	184.66667
381	4	699.83334	174.95833	99.562454	9.9780987	162.16667	176.16667	185.33333
427	4	691	172.75	106.93522	10.340949	158.33333	176.08334	180.5
540	4	694.16667	173.54167	122.11806	11.050704	159.5	174.08334	186.5
626	4	679.33333	169.83333	93.907428	9.6905845	159.5	170.08333	179.66667
683	4	699.16666	174.79166	95.988398	9.7973669	165	172.91666	188.33333
687	4	703.16667	175.79167	95.858839	9.7907527	162.83333	178.08334	184.16667
705	4	705.83334	176.45834	100.50698	10.025317	167.66667	175.25	187.66667
781	4	710.50001	177.625	105.17359	10.255418	164.66667	179.66667	186.5

Figure 12-12: Results of Row Statistics

7) Note that 11 instances out of 1000 trials resulted in a variance greater than 90.92. This corresponds to an estimated p-value of 11/1000 or 0.011. Remember that resampling procedures are based on random sampling from the data, so your results may differ from the results found in the text or text supplement.

NOTE: Sometimes it is the case that the various groups (e.g. Fat1, Fat2, etc. above) have different sizes. In SC, you would simply specify the appropriate resample sizes in step 2 above. The rest of the procedure is the same.

For example, if the original group sizes were 6, 8, 8 and 10, in step 2 you would enter “6” for the Sample size of the first Sample columns procedure, “8” for the Sample size of the second Sample columns procedure, “8” for the Sample size of the third Sample columns procedure, and “10” for the Sample size of the final Sample columns procedure.

12.5 Two-Way ANOVA

Interaction

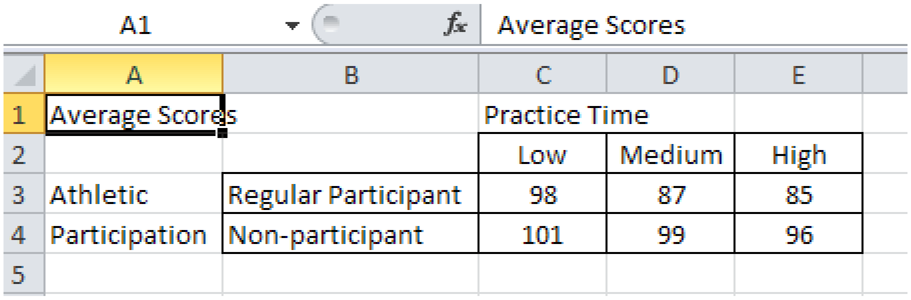
Creating an Interaction Plot using Excel

A group of beginning golfers takes 3 lessons during a 3 week period. In between lessons, the golf students are told to practice on their own. After the third lesson they play a round of golf (18 holes) and record their scores. The pro then asks them whether they regularly play another sport that involves hand-eye coordination (basketball, football, baseball, racquet sport), and divides the scores into "regular participants in other sport" and “non-participants in other sports”. In addition, the pro divides the student golfers into 3 levels of practice, based on the time students spent practicing on their own. Average scores were recorded for each student for the round they played. The average scores were divided into 6 categories:

- regular participants in other sports, low practice
- regular participants in other sports, medium practice
- regular participants in other sports, high practice
- non participants in other sports, low practice
- non participants in other sports, medium practice
- non participants in other sports, high practice

Instructions

1) Create your table of variables and means. We’ll use the example shown in the figure below. The data represent the average scores of student golfers on the final round after a 3 week period of 3 lessons and interim practice time. For example, the “98” in cell C3 represents the average score of student golfers who regularly participate in other sports requiring hand-eye coordination and also had minimal (low) practice time between lessons.

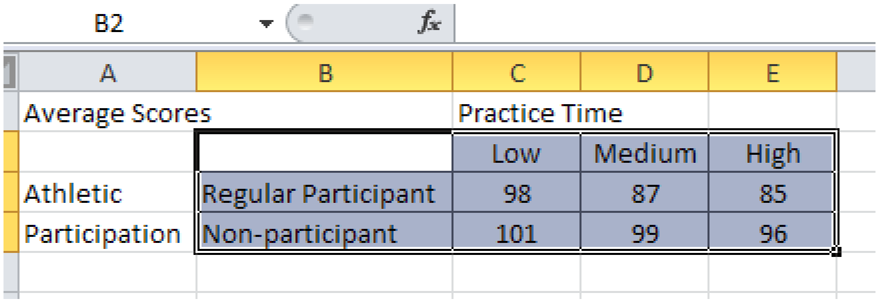


A1 Average Scores

	A	B	C	D	E
1	Average Scores		Practice Time		
2			Low	Medium	High
3	Athletic	Regular Participant	98	87	85
4	Participation	Non-participant	101	99	96
5					

Figure 12-13: Table of variables and mean scores

2) Select the data cells and their respective labels (see below):



B2

A	B	C	D	E
Average Scores		Practice Time		
		Low	Medium	High
Athletic	Regular Participant	98	87	85
Participation	Non-participant	101	99	96

Figure 12-14: Select data table

3) Select the Insert menu and choose the 2D Line graph ribbon menu. Choose the upper left 2-D Line chart as shown below:

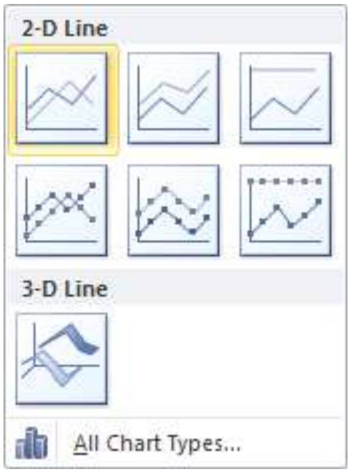
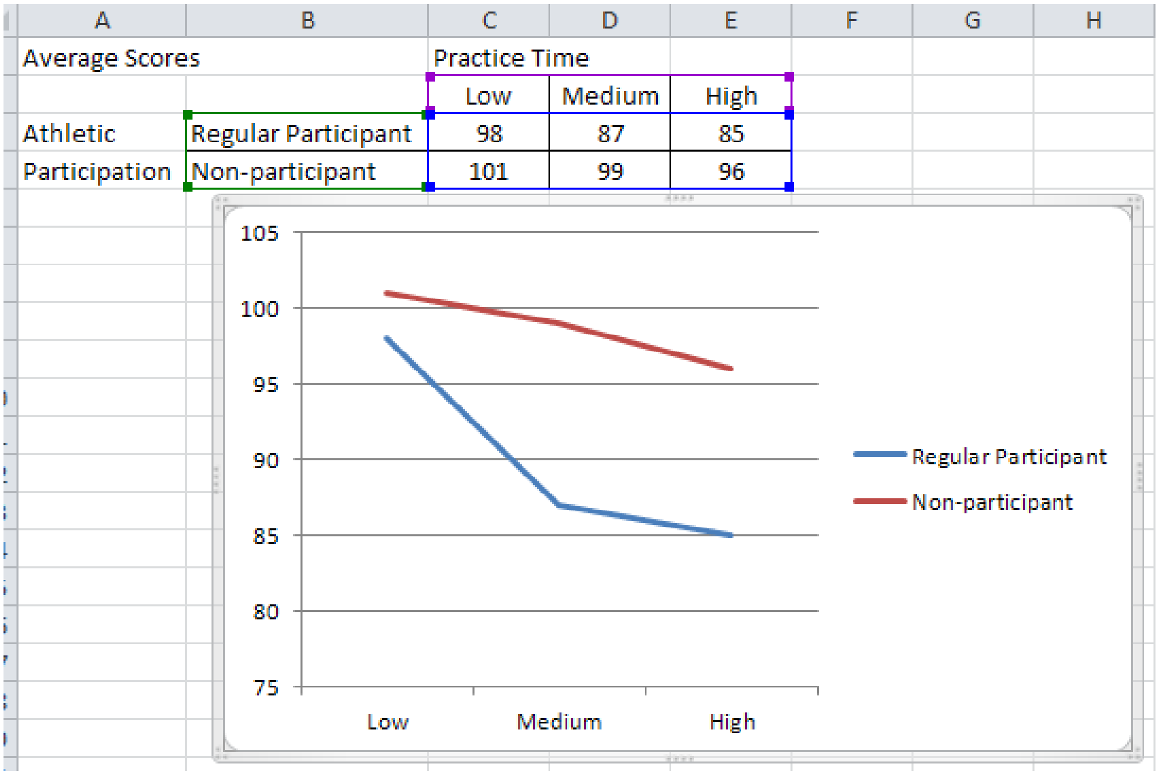


Figure 12-15: 2-D line graph menu (Insert menu)

4) Excel will automatically create the interaction plot:



A	B	C	D	E	F	G	H
Average Scores		Practice Time					
		Low	Medium	High			
Athletic	Regular Participant	98	87	85			
Participation	Non-participant	101	99	96			

Figure 12-16: Interaction chart

How can we interpret this plot? It is apparent that practice improves all golfers, but it yields a bigger and more immediate improvement for those who are skilled in other hand-eye coordination sports.

13 Multiple Regression

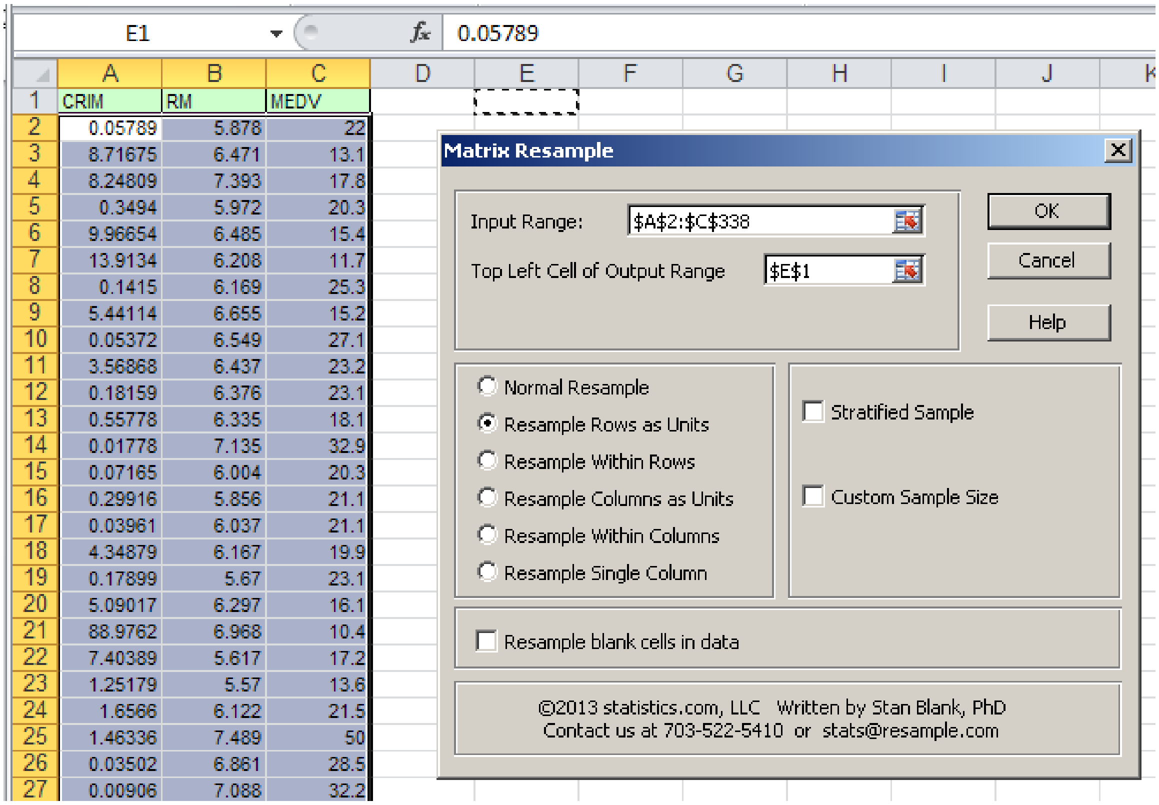
13.4 Model Assessment and Inference

The resampling model in this chapter is more complex. We've provided a downloadable workbook for Resampling Stats <here>. The workbook has the results of a previous simulation and will open in the Results worksheet where the output is immediately available, including the 90% confidence interval in cells E1:H3. Let's do another simulation from the beginning.

1) If you haven't already, download the BostonHousing_CRIM_RM_RSXL.xls workbook using the link above.

2) Load Resampling Stats and open the BostonHousing_CRIM_RM_RSXL.xls workbook.

3) Select the Data worksheet tab. Select cells A2:C338 and choose Resample (figure 13-1).



E1 0.05789

	A	B	C
1	CRIM	RM	MEDV
2	0.05789	5.878	22
3	8.71675	6.471	13.1
4	8.24809	7.393	17.8
5	0.3494	5.972	20.3
6	9.96654	6.485	15.4
7	13.9134	6.208	11.7
8	0.1415	6.169	25.3
9	5.44114	6.655	15.2
10	0.05372	6.549	27.1
11	3.56868	6.437	23.2
12	0.18159	6.376	23.1
13	0.55778	6.335	18.1
14	0.01778	7.135	32.9
15	0.07165	6.004	20.3
16	0.29916	5.856	21.1
17	0.03961	6.037	21.1
18	4.34879	6.167	19.9
19	0.17899	5.67	23.1
20	5.09017	6.297	16.1
21	88.9762	6.968	10.4
22	7.40389	5.617	17.2
23	1.25179	5.57	13.6
24	1.6566	6.122	21.5
25	1.46336	7.489	50
26	0.03502	6.861	28.5
27	0.00906	7.088	32.2

Matrix Resample

Input Range: $A$2:$C$338

Top Left Cell of Output Range $E$1

○ Normal Resample
● Resample Rows as Units
○ Resample Within Rows
○ Resample Columns as Units
○ Resample Within Columns
○ Resample Single Column

☐ Stratified Sample

☐ Custom Sample Size

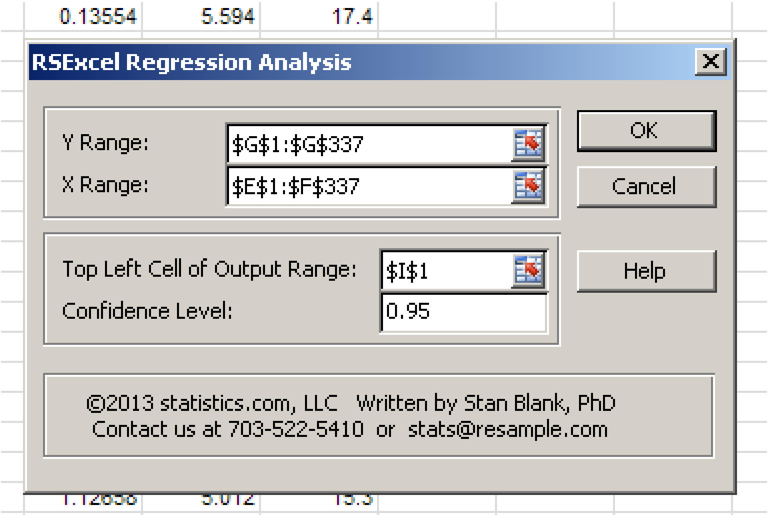
☐ Resample blank cells in data

©2013 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 13-1: Resample Rows as Units

4) Select cell E1 as the Top Left Cell and choose Resample Rows as Units (to keep the individual housing data together on the same "slip" of paper). Click OK.

5) Select Regression from the Resampling menu. For the Y Range (MEDV) select cells G1:G337 and for the X Range, select cells E1:F337. The Top Left Cell should be I1. Since we are bootstrapping, the confidence interval is meaningless in this context so leave it at the default of 0.95 (figure 13-2). Click OK.



Y Range: $G$1:$G$337

X Range: $E$1:$F$337

Top Left Cell of Output Range: $I$1

Confidence Level: 0.95

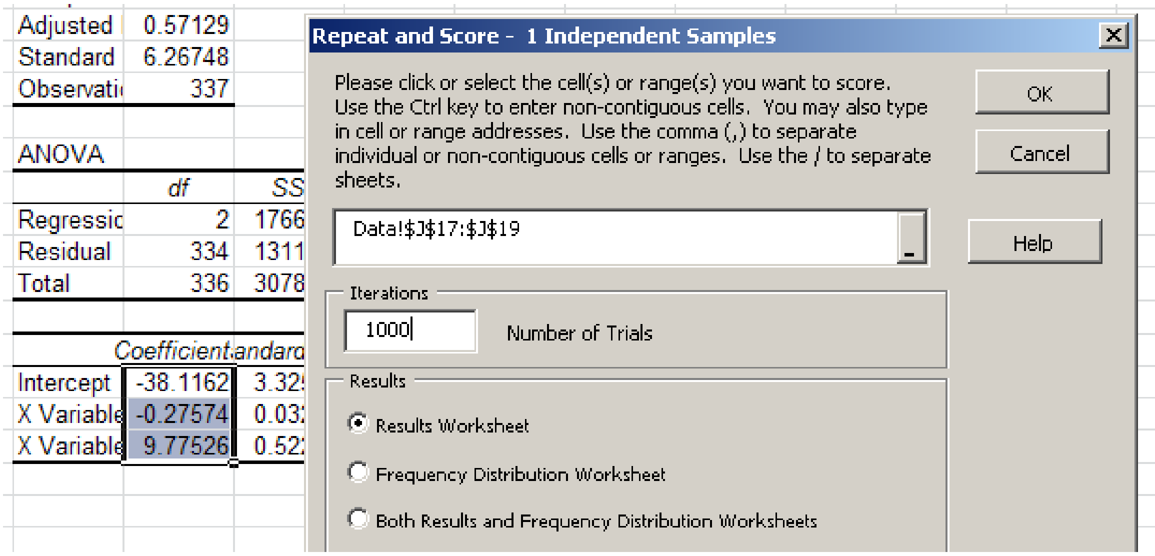
©2013 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 13-2: Resampling Stats Regression dialog

6) A dialog will ask if you want to overwrite existing data, click OK.

7) The score cells will be the coefficients (in order) in cells J17:J19. Select all 3 of these cells and click Repeat and Score (figure 13-3).

   Note: the coefficients correspond to the independent variables as follows: X Variable 1 = CRIM and X Variable 2 = RM.



Adjusted	0.57129	
Standard	6.26748	
Observati	337	
ANOVA		
	df	SS
Regressic	2	1766
Residual	334	1311
Total	336	3078
	Coefficient	andard
Intercept	-38.1162	3.32
X Variable	-0.27574	0.03
X Variable	9.77526	0.52

Repeat and Score - 1 Independent Samples

Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

Data!$J$17:$J$19

Iterations

1000 Number of Trials

Results

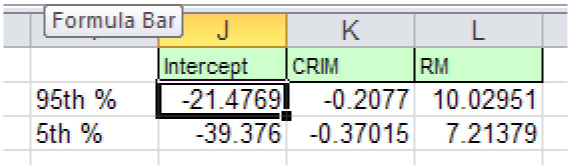
● Results Worksheet

○ Frequency Distribution Worksheet

○ Both Results and Frequency Distribution Worksheets

Figure 13-3: Coefficients as score cells for Repeat and Score

8) Click OK. When the simulation ends, go to the Results worksheet and view the 90% confidence interval for each of coefficients (figure 13-4).



	J	K	L
	Intercept	CRIM	RM
95th %	-21.4769	-0.2077	10.02951
5th %	-39.376	-0.37015	7.21379

Figure 13-4: 90% confidence intervals for one experiment

Addendum:

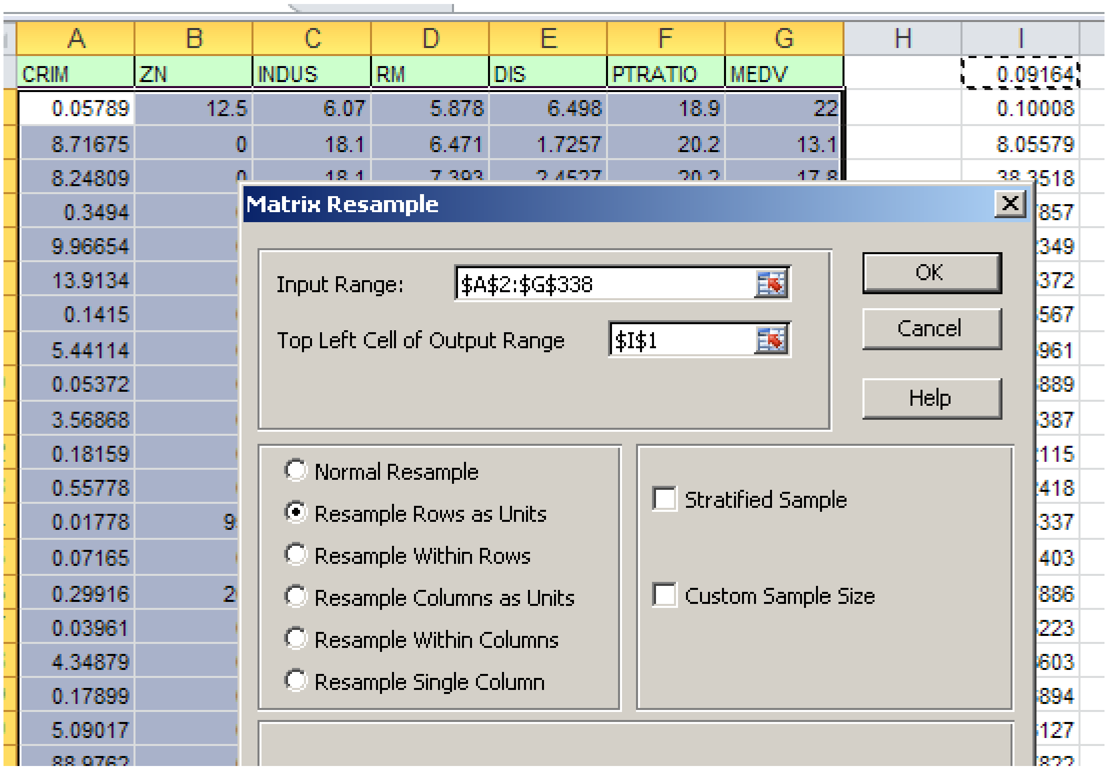
We can also perform a more complex multiple regression using the full Boston Housing data set. The full data set contains many more independent variables than the CRIM and RM example we've just completed.

The resampling model in this chapter is complex. We've provided a downloadable workbook for Resampling Stats <here>. The workbook has the results of a previous simulation and will open in the Results worksheet where the output is immediately available, including the 90% confidence interval in cells I1:P3. Let's do another simulation from the beginning.

1) If you haven't already, download the BostonHousingMain_RSXL.xls workbook.

2) Load Resampling Stats and open the BostonHousingMain_RSXL.xls workbook.

3) Select the Data worksheet tab. Select cells A2:G338 and choose Resample (figure 13-1).



Matrix Resample

Input Range: $A$2:$G$338

Top Left Cell of Output Range $I$1

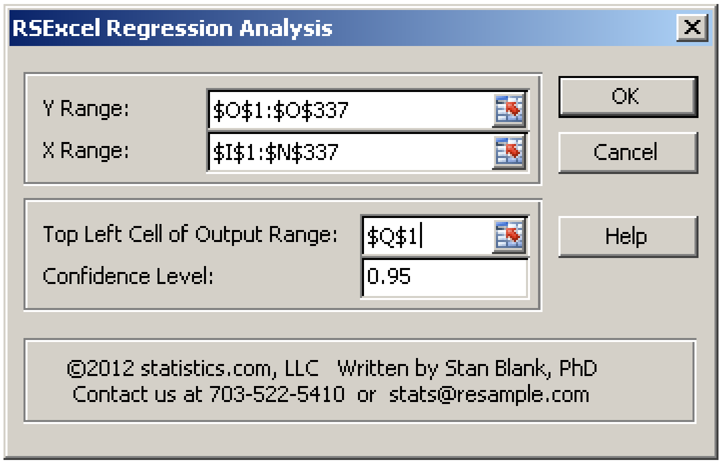
○ Normal Resample
● Resample Rows as Units
○ Resample Within Rows
○ Resample Columns as Units
○ Resample Within Columns
○ Resample Single Column

☐ Stratified Sample
☐ Custom Sample Size

Figure 13-5: Resample Rows as Units

4) Select cell I1 as the Top Left Cell and choose Resample Rows as Units (to keep the individual housing data together on the same "slip" of paper). Click OK.

5) Select Regression from the Resampling menu. For the Y Range (MEDV) select cells O1:O337 and for the X Range, select cells I1:N337. The Top Left Cell should be Q1. Since we are bootstrapping, the confidence interval is meaningless in this context so leave it at the default of 0.95 (figure 13-2). Click OK.



RSExcel Regression Analysis

Y Range: $O$1:$O$337

X Range: $I$1:$N$337

Top Left Cell of Output Range: $Q$1

Confidence Level: 0.95

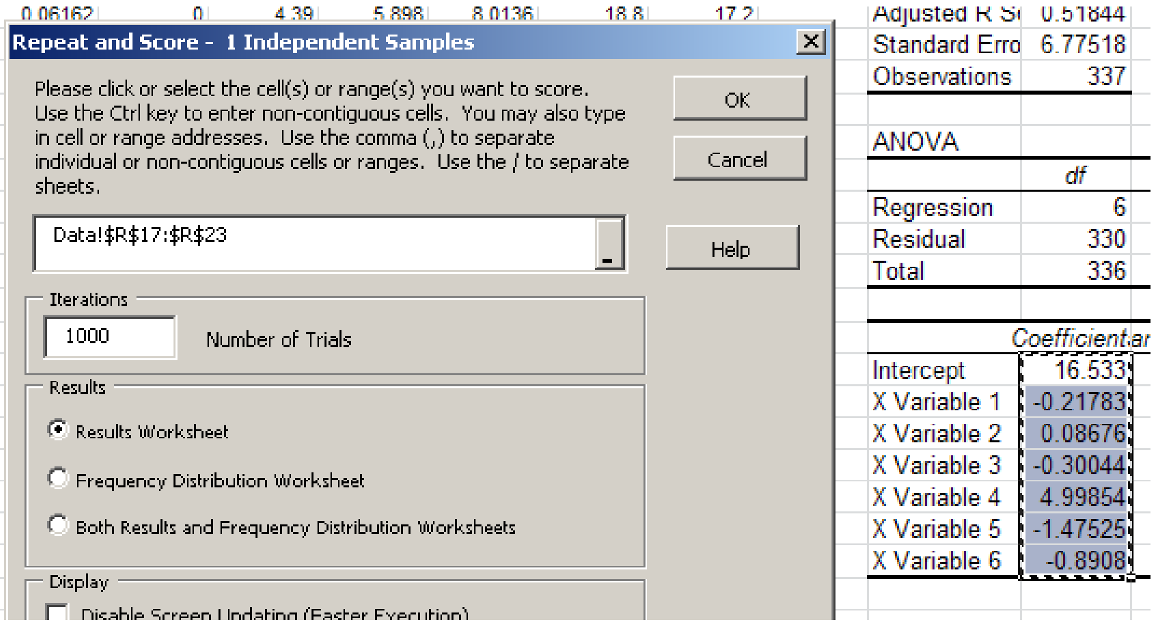
©2012 statistics.com, LLC Written by Stan Blank, PhD
Contact us at 703-522-5410 or stats@resample.com

Figure 13-6: Resampling Stats Regression dialog

6) A dialog will ask if you want to overwrite existing data, click OK.

7) The score cells will be the coefficients (in order) in cells R17:R23. Select all 7 of these cells and click Repeat and Score (figure 13-3).

   Note: the coefficients correspond to the independent variables as follows: CRIM-X Variable 1, ZN-X Variable 2, INDUS-X Variable 3, RM-X Variable 4, DIS-X Variable 5, and PTRATIO-X Variable 6.



0.06162 | 0 | 4.39 | 5.898 | 8.0136 | 18.8 | 17.2

Repeat and Score - 1 Independent Samples

Please click or select the cell(s) or range(s) you want to score. Use the Ctrl key to enter non-contiguous cells. You may also type in cell or range addresses. Use the comma (,) to separate individual or non-contiguous cells or ranges. Use the / to separate sheets.

Data!$R$17:$R$23

Iterations

1000 Number of Trials

Results

● Results Worksheet

○ Frequency Distribution Worksheet

○ Both Results and Frequency Distribution Worksheets

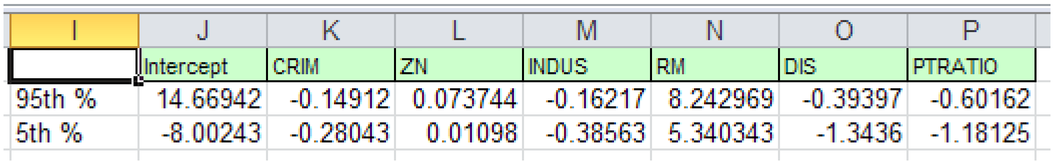
Display

☐ Disable Screen Updating (Faster Execution)

Adjusted R S	0.51844
Standard Erro	6.77518
Observations	337
ANOVA	
	df
Regression	6
Residual	330
Total	336
	Coefficient
Intercept	16.533
X Variable 1	-0.21783
X Variable 2	0.08676
X Variable 3	-0.30044
X Variable 4	4.99854
X Variable 5	-1.47525
X Variable 6	-0.8908

Figure 13-7: Coefficients as score cells for Repeat and Score

8) Click OK. When the simulation ends, go to the Results worksheet and view the 90% confidence interval for each of coefficients (figure 13-4).



I	J	K	L	M	N	O	P
	Intercept	CRIM	ZN	INDUS	RM	DIS	PTRATIO
95th %	14.66942	-0.14912	0.073744	-0.16217	8.242969	-0.39397	-0.60162
5th %	-8.00243	-0.28043	0.01098	-0.38563	5.340343	-1.3436	-1.18125

Figure 13-8: 90% confidence intervals for one experiment