

Resampling Stats Add-in for Excel User's Guide

Version 4

©statistics.com, LLC 2009

Preface

The presentation of resampling methods in this book owes a great debt to Julian Simon—resampling pioneer and creator of the original Resampling Stats software.

statistics.com, LLC
612 N. Jackson Street
Arlington, Virginia 22201
stats@resample.com
www.resample.com
703-522-2713

Contents

Preface	i
Contents	ii
List of Figures	v
List of Tables	x
1 Introduction	1
1.1 How to Use This book	1
1.2 Installation	2
1.3 About Resampling	3
1.4 The Resampling Stats Add-in (“RSXL”)	5
1.5 Probability by Resampling	6
1.6 Counting Results	13
1.7 The Frequency Function	18
2 Advanced Probability	21
2.1 Rates and Results	21
2.2 Simulation and Hard Problems	27
3 Confidence Intervals	37
3.1 Confidence Interval for Means	39
3.2 Confidence Interval for a Proportion	43
3.3 Confidence Intervals for Medians	45
3.4 Confidence Interval for Profit	47
3.5 Planning Inventory	49
4 Hypothesis Testing	55
4.1 Resampling and p-values	55

4.2	Testing for a Difference in Variability	63
4.3	Resampling in Complex Cases	65
4.4	Multiple Comparisons - Ad Clickthroughs	71
5	Contingency Tables	79
5.1	Chi-Squared Basics	79
5.2	Sir Ronald and the Tea Lady	79
5.3	Applying Resampling	82
6	Correlation and Regression	89
6.1	Applied Correlation: Baseball Salary vs. Rank	89
6.2	Regression Basics	92
6.3	Baseball Again: Running Regression from the Resampling Add-in	95
6.4	Multiple Linear Regression: Newspapers and Population	96
7	Analysis of Variance	101
7.1	Geyser Timing	101
7.2	Coagulation Time	103
7.3	Resampling and the F-test	105
8	Non-Parametric Statistics	111
8.1	Birthweight Revisited: A Signs Test	111
8.2	Birthweights a Third Time: A Paired Permutation Test	113
8.3	Rank Sum Test	115
8.4	Another Correlation Study	117
9	Stratified Resampling	119
9.1	Evaluating Corporate Mergers; “Shuffling Within Rows”	119
9.2	Mergers Revisited	122
9.3	Reading Methods: Use of Variable to Denote Strata	126
9.4	Darwin’s Plants: Strata in Separate Ranges	129
10	Formula Iteration	133
10.1	Iterative Solutions to Equations	133
10.2	Newton’s Method	135
10.3	The Logistic Equation	137
10.4	Predator-Prey Relationships	141
10.5	The Lorenz “Butterfly” Equations	144
	Summary	147
11	BCA Bootstrap	149

11.1 Process Temperature	150
11.2 Compare to Percentile Interval	153
Appendix to Chapter 11	155
12 Resampling Stats Operations	159
Resampling Stats Add-in Functions and Syntax	159
Auto-Range Selection	159
Auto-Reset	159
Auto-Numbering	160
Custom Functions	161
Escape	162
File Operations	162
Formulas (Resampling Formulas)	162
Histogram	163
Licensing Procedure	167
Macros	169
Maximum Number of Trials	169
Menu and Toolbar for the Resampling Stats Add-in	170
Multi-stage Resampling and Shuffling	175
Opening and Closing the Resampling Stats Add-in	175
Opening Files	175
Random Number Generator	175
Redo	179
Regression	179
Repeat and Score	181
Resample and Shuffle Options	183
Resample and Resampling (the Different Meanings of the Terms)	191
Reset	191
Saving and Opening Files and Storing Simulation Parameters	192
Score	194
Shuffle	194
Sort	194
Stratified Resampling and Shuffling	196
Toolbar and Excel Ribbons	198
Urn	199
Useful Excel Functions	202
Bibliography	209

List of Figures

1.1	The Resampling Toolbar	5
1.2	The Resampling Menu	5
1.3	Coin Flipping Setup	6
1.4	The Resampling Dialog Box	7
1.5	Counting the One's	8
1.6	Identify Score Cells	9
1.7	Worksheet Tabs	10
1.8	Results, Unsorted	10
1.9	Excel's Sort Buttons	10
1.10	Results, Sorted	11
1.11	Histogram Button	11
1.12	Histogram Dialog Box	12
1.13	Histogram, # of Heads in 8 Tosses of a Coin	12
1.14	Excel's Insert Function Button	13
1.15	Excel's Insert Function Dialog Box	14
1.16	COUNTIF Dialog Box	14
1.17	Rolling a Pair of Dice	16
1.18	Getting a 7 When Rolling Two Dice	17
1.19	Repeating the Repeat and Score	17
1.20	Results of 10000 Iterations of Tossing Two Dice	18
1.21	Frequency Distribution Worksheet Output	18
1.22	Names vs. Numbers	19
1.23	The FREQUENCY Dialog Box: (Control+Shift+Enter) should be pressed instead of OK	19
1.24	Analyzing 1000 Trials	20
2.1	Baseball Hits	22
2.2	BINOMDIST Dialog Box	23
2.3	Binomial Probability Table	24

LIST OF FIGURES

2.4	Choosing How to Create the Urn	25
2.5	The 12.4312.43 Urn Dialog Box	26
2.6	Totalling the Baskets	26
2.7	Zener Cards Used in ESP Experiment	27
2.8	Shuffle, the “Other” Resampling Function	28
2.9	The ESP Model	28
2.10	The IF Function	29
2.11	ESP Data	30
2.12	Autofill	31
2.13	Results of Autofill	32
2.14	Looking for Birthday Matches with MATCH	33
2.15	Counting Matches on the Results Sheet	34
2.16	The Secretary Problem	35
2.17	The Secretary Problem: 10 Letters and 1000 Trials	36
3.1	Finding the 2.5 th Percentile	40
3.2	C.I. Estimate, Drill Diameters	41
3.3	Histogram of Resampled Diameter Means	41
3.4	Rainfall in May	42
3.5	Results from a Small Poll	44
3.6	Larger Poll Results	44
3.7	Median Income	45
3.8	Median Income Revisited (Millionaire Version)	46
3.9	Price Elasticity Confidence Interval	48
3.10	Direct Mail Problem: 1000 Trials	50
3.11	Histogram of Results for Direct Mail Problem	50
3.12	A Sales-per-Week Model	52
3.13	A Sales-per-Week Model	52
3.14	Cell Formula for Inventory	54
3.15	Looking for Stockout	54
4.1	Fruit Flies and Gender	57
4.2	Fruit Flies and Gender: Results Sheet Sorted Output	58
4.3	A Histogram Table “On the Fly”	58
4.4	A Cure/Not-Cure Test	60
4.5	Checking Cancer Cure for Significance	61
4.6	A Classic Hypothesis Test	62
4.7	Too Close to Call	63
4.8	Instrument Design	64
4.9	Differences in Standard Deviation, Shuffled Pairs of Samples	65

4.10	Setup for Protein Analysis	67
4.11	Creating an Urn Full of Amino Acids	67
4.12	Resampling Results for DNA Protein Sequence	68
4.13	Employee Firing Test Statistic	70
4.14	Employee Firing Test Statistic: 1000 Trials	70
4.15	Employee Firing Test Statistic: 1000 Trials	71
4.16	Clickthrough Ads Multiple Urns	73
4.17	Difference in Resampled #3 and Other Ads	74
4.18	Unsorted Data	74
4.19	Resampling Stats Sort Dialog	75
4.20	Shuffled Sample Difference	75
4.21	Estimated p-value = $\frac{171}{1000}$	76
4.22	Observed Difference Compared to Shuffled Difference	77
5.1	Classic Probability and the “tea test”	80
5.2	Tea Testing: One of Many Possible Random Guesses	81
5.3	81
5.4	Numbers of Drivers Stopped and Not Stopped	82
5.5	Setting Up the Race Test	83
5.6	Driving While Black: Trial Results	83
5.7	Psychopharmacology	84
5.8	Results Sheet Showing Resampled Differences in Scores	85
5.9	Employee Drug Testing	86
6.1	Pay and Team Rank in Baseball	89
6.2	Correlation of Payroll and Rank in MLB	92
6.3	Regression via Resampling - a Simple Case	93
6.4	Analyzing Regression Output	94
6.5	X-Y Input for the Resampling Menu Regression Option	95
6.6	Slope	96
6.7	Another Section of the Same Regression Output	96
6.8	Population, Circulation, and Sales	97
6.9	Invoking the Regression Command in the Resampling Stats Add-in	98
6.10	Picking Resampled Regression Parameters for Repeat and Score (a Portion of the Resampled Regression Output is Shown)	98
6.11	Estimated 90% Confidence Intervals Using Resampling	99
7.1	Geyser Data	101
7.2	1000 Trials, Estimated p = .085	102

LIST OF FIGURES

7.3	Table of Coagulation Times	103
7.4	Dealing with asymmetric tables (data are shuffled together, then redistributed into a table of same structure as original table)	104
7.5	Low Probability Diet Outcome	105
7.6	Simple ANOVA Table: Heights of Beanstalks	105
7.7	Computing F from the Table	106
7.8	Simple Statistics for Table Analysis	107
7.9	p-values from Resampling	108
8.1	Shuffled by Rows	115
8.2	Shuffled by Rows	116
8.3	Ranksum Results	116
8.4	I.Q. and Athletic Ranks	117
8.5	Ranksum Results, Round Two	118
9.1	Merger data: revenue growth in year after merger	120
9.2	Merger Data, in Rankings	121
9.3	1.45 is Statistically Significant!	123
9.4	Original Merger Data	123
9.5	Shuffle Within Rows	125
9.6	Statistic of Interest	125
9.7	Histogram of Merged Data	126
9.8	Hypothetical Reading Data	127
9.9	Results of One Unstratified Shuffle	127
9.10	Results of One Stratified Shuffle	128
9.11	Stratified Shuffle Option	128
9.12	Resampling Stats “&n” Syntax	130
9.13	Region Selection - Resampling Stats “&n” Syntax	130
9.14	Shuffled Difference in Plant Heights	131
9.15	1000 Trials Shuffled Plant Height Differences	131
9.16	Resampled p-value =0.013 for 1000 Trials	132
10.1	$x = \cos(x)$	134
10.2	$f(x) = x^3 - x - 1$	136
10.3	Newton’s Method in Excel	137
10.4	The Logistic Equation in Excel	138
10.5	Logistic Equation: $r = 3$	139
10.6	Logistic Equation: $r = 3.1$	139
10.7	Logistic Equation: $r = 3.7$	139

10.8	Logistic Equation: Increment r	140
10.9	Logistic Chaos	141
10.10	Population Dynamics with Excel	142
10.11	Predator-Prey Data	143
10.12	Predator-Prey Parametric Plot	144
10.13	Lorenz Equations in Excel	145
10.14	Lorenz “Butterfly” Attractor	146
11.1	Process Temperature Readings	151
11.2	Bootstrap BCA from the Resampling Menu	151
11.3	Bootstrap BCA Dialog	152
11.4	Completed Bootstrap BCA Dialog	152
11.5	BCA Confidence Interval	153
11.6	Bootstrap Percentile CI	153
11.7	Bootstrap Percentile Method Interval	154
12.1	Auto-Range Select	160
12.2	Paste Special Dialog: Values	163
12.3	The Histogram Button	163
12.4	Selecting a Named Range	164
12.5	Entering a Range Name	165
12.6	The Histogram Dialog Box	166
12.7	Histogram: Rolling 2 Dice	166
12.8	Histogram Counts	166
12.9	Histogram Percent	167
12.10	Distribution Chart	167
12.11	Cumulative Frequency	167
12.12	First Run Licensing Dialog	168
12.13	Username and License Key Entry	168
12.14	Successful Username and License Key Entry	169
12.15	Resampling Menu	170
12.16	The Histogram Button	171
12.17	RSXL Random Number Distribution Functions	176
12.18	REDO Dialog Box	179
12.19	RSXL Regression Dialog Box	180
12.20	Regression Output	181
12.21	Repeat and Score Dialog Box	182
12.22	Resampled Data in Column B	184
12.23	Shuffled Data in Column B	184
12.24	Single Row/Column Resampling	185

12.25	Custom Resampled Output Range	186
12.26	Custom Shuffled Output Range	186
12.27	Matrix Resampling or Shuffling Dialog	187
12.28	Normal Matrix Shuffle	187
12.29	Shuffle Rows as Units	188
12.30	Shuffle Within Rows	189
12.31	Shuffle Columns as Units	189
12.32	Shuffle Within Columns	190
12.33	Shuffle a Single Column	190
12.34	Saving Simulation Parameters	193
12.35	Restoring Simulation Parameters	193
12.36	Sorting Resampled Data	195
12.37	Sort Dialog Box	195
12.38	Resampling Stats “&n” Syntax	197
12.39	Using the Shuffle Dialog with the “&n” Syntax	198
12.40	The Resampling Toolbar	198
12.41	The Excel Add-Ins Ribbon	199
12.42	Choosing the Urn Type	200
12.43	Creating an Urn with 48 “1’s” and 52 “2’s”	200
12.44	Specifying the Urn Contents on the Worksheet	201
12.45	Worksheet Urn dialog box	201
12.46	Autofill	202
12.47	Autofill Results	202
12.48	Countif Dialog	203
12.49	Frequency Dialog	205
12.50	Excel’s IF Function	205
12.51	Excel’s Insert Function “ f_x ”	206
12.52	Excel’s Percentile Function Dialog	207
12.53	Data Ribbon Sort Buttons	207

List of Tables

3.1	Price Elasticity	47
-----	----------------------------	----

3.2	Confidence Interval for Profit	48
3.3	Unit Sales for 15 Days	50
3.4	Deviation from the Average over 15 Days	51
4.1	Molecular Protein Sequences	66
4.2	Residues in COOH-term: Observed vs. Expected	66
4.3	Seniority of discharged workers (years)	68
4.4	Seniority of all workers	69
5.1	Prospective Employee Drug Related Claims	86
6.1	Major League Baseball - 1995-1997	90
8.1	Reading Scores	112
8.2	Birthweights	112
8.3	Hypothetical Athletic and I.Q. Scores for High School Boys	117
9.1	Revenue growth in year 1 following merger	120
9.2	Average Revenue Growth Rates	120
9.3	Ranked Within Matched Set: (1 = Worst, 3 = Best)	121
9.4	Average Revenue Growth Rates	122
11.1	Process Temperature Readings	150

Chapter 1

Introduction

1.1 How to Use This book

First, you should read this introductory chapter on the resampling method of solving problems in probability and statistics. Then you can decide whether to study additional illustrations (Chapters 2-11), or go directly to the section on syntax (entitled **Resampling Stats Operations**) to get an overview of all the Resampling Stats add-in functions.

- **Important note:** We assume that you are using Excel 2007 for Windows. Older versions of the Resampling Stats for Excel add-in are available for Excel 2003, Excel XP, and Excel 2000, but not all the functionality discussed in this user guide will be available. In addition, for your Resampling Stats add-in to work properly, make certain that the Analysis Toolpak and Analysis Toolpak VBA add-ins that come with Excel are activated. You may activate these add-ins by clicking on the “Excel Office Button” in the upper left corner of the Excel window and selecting the “Excel Options” button in the lower right corner of the dialog. In the pop-up dialog that appears, click on “Add-ins” (in the left pane), then “Go” on the “Manage: Excel Add-ins” button at the bottom of the dialog. You should then be able to select the required VBA add-ins (Analysis Toolpak and Analysis Toolpak-VBA) from the add-in manager dialog box.

1.2 Installation

The installation of the Resampling Add-in for Excel follows the standard Windows setup protocol. Double-click on the Resampling Stats for Excel installation file that you downloaded. This will install the add-in with the associated sample and help files. Once you have installed the RSXL add-in, you will need to enable the VBA macros in Excel. Usually when you launch the RSXL add-in, a security dialog will appear. Simply click on “Enable Macros” and RSXL should function properly. If RSXL does not work as expected, you may need to manually enable macros.

For security purposes, it is recommended that you allow Excel to prompt you, via the security dialog, to enable macros when you start the Resampling Stats add-in. Manually enabling macros will allow ALL VBA macros permission to run in Excel each time Excel is started.

To manually enable macros, select the “Excel Options” menu again from the Excel Office Button and select “Trust Center” from the left hand pane. Then in the right hand pane click on the “Trust Center Settings” button. Again, in the left hand pane select “Macro Settings” followed by “Enable All Macros” in the right pane. Click “OK” twice and the RSXL add-in should be completely installed. You may now run the Resampling Stats add-in from the Start menu. When you start Resampling Stats, it will automatically open Excel if Excel is not already open.

The first time you run the Resampling Stats add-in, you will need to enter licensing information. If you need help with the licensing procedure, please look in Chapter 12 (Resampling Stats Operations) under [Licensing](#).¹

¹Alternatively, you may start Resampling Stats as follows:

1. In your file manager, find and double-click the Resampling Stats .xla file that you downloaded and installed in your Program Files folder. This will open Excel, if it is not already open, and launch the Resampling Stats Add-In.
2. If you are already running Excel, you may select the “Excel Office Button” followed by “Open” and open the Resampling Stats .xla file as above.
3. There is also a third option. You can have Excel automatically load the Resampling Stats (RSXL) add-in every time you use Excel. As in the installation of the Analysis Toolpak Add-Ins, navigate to the Add-in Manager dialog and click the “Browse” button. Locate the add-in wherever you installed it and click “OK”. This will add it to your list of Excel add-ins; to make certain that RSXL loads automatically, be sure the box next to it is checked. Note that if you later install a different version of the add-in, you should delete or rename the old file – otherwise Excel will continue to try to load the old add-in. Likewise, if you later uninstall RSXL, Excel will continue

Note: Two versions of the Resampling Stats add-in may not be resident at the same time. This will generate an error box which informs the user of this problem. We'll assume that you are familiar with the basic features of Microsoft Excel, and show you the step-by-step solution of statistics problems using the Resampling Stats add-in for Excel. Some key Excel functions that are particularly useful for resampling are covered in the section entitled "Resampling Stats Operations."

1.3 About Resampling

The conventional analytic approach to inferential statistics requires that you understand complex formulas, and too often you can find yourself selecting the wrong formula. In contrast, resampling proceeds in stages that are easy to understand. Most problems can be tackled using the following 3-stage process:

1. Specify the population to sample from (random numbers, an observed data set, "0's" and "1's," etc.).
2. Specify the sampling procedure (number of samples, sizes of samples, sampling with or without replacement).
3. Specify the statistic you wish to monitor or score.

Resampling methods are typically used to address questions of statistical inference:

1. How much sampling error might there be in an estimate based on limited data (establishing confidence limits)?
2. How likely is it that chance sampling error might produce a sample result as extreme as the observed sample (hypothesis testing)?

to look for RSXL if you have selected this "autoload" option. For these reasons, unless you have a preference for having the add-in launch every time you use Excel, we recommend not including it in this add-in list and manually loading RSXL when needed.

1. INTRODUCTION

With resampling, you try to answer these questions by drawing simulated samples, or “resamples” from the data themselves, or from a reference distribution based on the data, and observing how the statistic of interest in these resamples behaves.

Early in the 20th century, when computers were unavailable to do the hard work of drawing all these samples, statisticians found they were able to represent the distributions of many sample statistics with calculated theoretical distributions of random variables.

William Gossett, the statistician better known by the pseudonym “Student” under which he published, repeatedly dealt out sets of randomly drawn cards with prisoners’ data written on them to see how the means of these resamples were distributed. He used this simulated data in deriving his now-famous t-distribution, which is used to approximate the sampling distribution of certain statistics, provided the samples are drawn from a distribution that is sufficiently normally-distributed (or large enough).

For example, suppose you have a data set of the heights of bean plants and would like to establish a confidence limit around the mean. In conventional statistics you generally proceed by assuming that your bean-heights are drawn from a huge, normally-distributed population of bean-heights.

Suitable theoretical approximations to sampling distributions were found for a variety of sample statistics, and were shown to be valid under a variety of circumstances. However, they are not available for all statistics in all circumstances. Approximations require assumptions about how data are distributed, and are generally good for large samples, but less accurate with small and imbalanced samples.

Resampling methods, including bootstrap and permutation methods, can be used with virtually any sample statistic and do not rely on assumptions about how the data are distributed, except for the assumption that the population’s data are distributed similarly to the sample. Permutation methods for significance testing have the added advantage that they produce “exact” p-values – guaranteed not to produce “significant” results more than 5% of the time when drawing from a null model (assuming you are testing at the 5% level of significance).

1.4 The Resampling Stats Add-in (“RSXL”)

The Resampling Stats add-in for Excel (“RSXL”) is a set of simple, intuitive commands that allow you to resample your worksheet data quickly and efficiently, with total understanding of the methods on your part. The RSXL installation file also contains a Worksheets directory which contains sample files for all the examples in this manual.

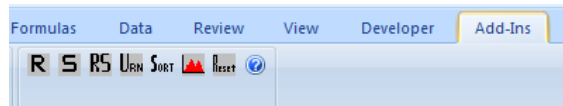


Figure 1.1: The Resampling Toolbar

When you start Resampling Stats, you’ll see the Resampling toolbar when you select the Add-ins menu in Excel (Figure 1.1). The same functions on the toolbar (plus additional ones) can also be found in the Resampling submenu in Excel’s Add-ins menu (Figure 1.2).

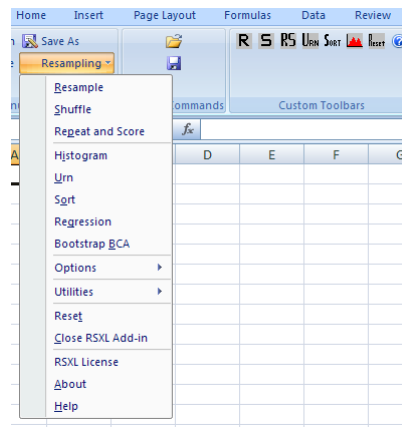


Figure 1.2: The Resampling Menu

To open Excel’s Add-ins menu, click on “Add-ins” in the top menu bar and look for the Resampling menu on the left side of the Excel window. Select “Resampling,” and you’ll see a short submenu of functions (many of which are on the toolbar).²

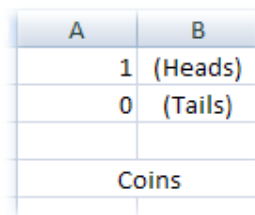
²If the Resampling Stats add-in is running, you can also bring up the resampling menu by right-clicking on a worksheet

For this chapter, all we'll need is Resample, the first choice on the menu (Figure 1.2) and the letter **R** on the Resampling toolbar. It's probably easiest to demonstrate what resampling means by showing a few problems in probability, so let's plunge right in. One feature of this approach may surprise you – resampling turns out to be not much harder to apply to challenging, complex problems than to simpler ones.

1.5 Probability by Resampling

Heads/Tails, Boys/Girls

Let's start where every introductory textbook starts, with coin flips. (The workbook Coins.xls contains the models discussed here.) A flipped coin can show heads or tails, so for a beginning modeling attempt we can just call these alternatives 1 or 0 (Figure 1.3).



A	B
1	(Heads)
0	(Tails)
Coins	

Figure 1.3: Coin Flipping Setup

To flip this coin 100 times, here's what we would do:

1. Select the range "A1:A2" (containing the 1 and 0) using a standard Excel click and drag.
2. Click the "**R**" on the Resampling toolbar. You'll see the dialog box in Figure 1.4; the input range (\$A\$1:\$A\$2) will be filled in. (Alternatively, from the Add-ins menu, select "Resampling" and then "Resample.")
 - Note that Excel automatically uses "absolute" cell references (\$A\$1:\$A\$2) that do not change when you copy them in formulas, as opposed to "relative" references that change when you copy formulas in order to maintain the same reference position relative to the new location of the formula.

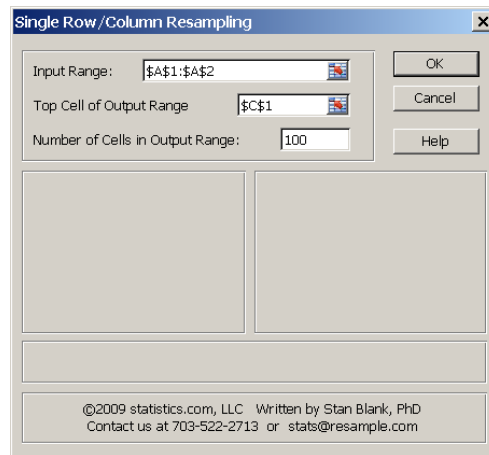


Figure 1.4: The Resampling Dialog Box

- If you did not correctly select the input range in step 1, you can type the correct range in the Input Range dialog.
3. Fill in your choice for the top cell of your output column of flips. An easy method for doing this is to click on the worksheet where you would like the top cell of your resampled output to go.
 4. Fill in the number of flips you'd like (i.e., the sample size) in the Number of Cells in Output Range box. The figure calls for 100, but you can try a larger number up to Excel's row limit.
 5. Click "OK".

This procedure gives you a column of randomly selected "1's" and "0's" in the range C1:C100. Let's plunge right in and apply this method to a real-world probability question:

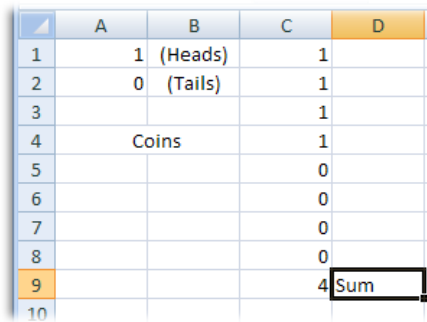
In families of eight children, how often should we expect to see three boys?

Assuming for simplicity that the probability of a boy vs. a girl is equal and independent of the previous birth (this is not strictly true), the results can be directly modeled by coin flips. The basic idea is: flip eight coins, count the heads, and repeat this procedure a large number of times. Here are the steps:

1. INTRODUCTION

1. Select as the input range the two-cell range containing the “1” (boy) and “0”(girl).
2. Specify C1 as the Top Cell of Output Range.
 - **IMPORTANT:** The output range should be separate from and not on top of the original data (the “1” and “0” here).
3. Specify “8” in the Number of Cells in Output Range box. The output will go into the range C1:C8. Click “OK”.
4. In cell C9, use the Excel formula =SUM(C1:C8) to add up the “1’s” (the boys, in this case.)
5. This produces a worksheet like the one shown in Figure 1.5.

Important: Do *not* put the resampled or shuffled output on top of the original data!



	A	B	C	D
1		1 (Heads)	1	
2		0 (Tails)	1	
3			1	
4		Coins	1	
5			0	
6			0	
7			0	
8			0	
9			4	Sum
10				

Figure 1.5: Counting the One's

Now we can call on a key feature of resampling, the Repeat and Score command. This command repeats your resampling operation many times, each time capturing the value in the cell(s) you designate. What we see in Figure 1.5 is a sample that represents a single family. To estimate the probability of 4 boys in 8 children, we should repeat the resampling of 8 “0’s” and “1’s” for a large number of samples, giving us more precise probability estimates as the number of samples grows. Here is how this is done:

1. Select the cell you want to score: C9.
2. Click the “**RS**” button on the Resampling toolbar. (Alternatively, select from the menu “Add-ins > Resampling > Repeat & Score.”) Figure 1.6 illustrates the Repeat and Score dialog.

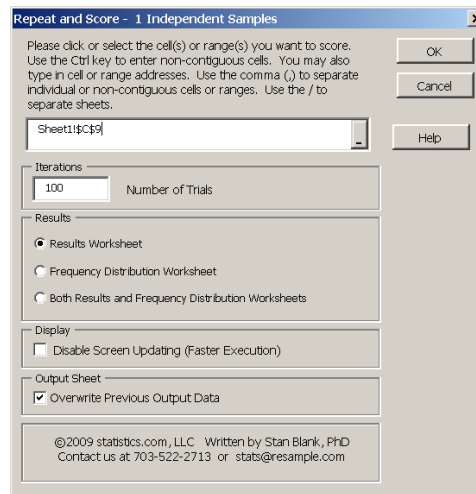


Figure 1.6: Identify Score Cells

3. Note that C9 is entered as the cell to score. You can also type it in. This score cell will automatically be highlighted in red for reference purposes.
4. Specify how many iterations (repetitions) you'd like to perform.³ For this simple demonstration, we'll try 100.

When we click “OK”, Resampling Stats then repeats the previous resampling instruction (resample eight selections from 0 and 1), which causes a new SUM value at each repetition, and writes these SUM values into a new worksheet, called the Results Sheet.

The output (Figure 1.8) for this case will be a list filling cells A1 to A100 on the Results sheet. Note how a new worksheet opened up to receive the results. You can navigate among the various worksheets via the tabs at the bottom of your screen as shown in Figure 1.7.

In this case, we get a distribution of numbers, ranging from 0 to 8 to reflect the number of boys (or the number of heads in eight coins flipped at once.)

Viewing the results is easier if the values are sorted. Click on the “A” at the top of the row, then select either the “Data Sort A to Z” or “Z to A” sort button on the main Excel toolbar. The Data Sort buttons are displayed in

³The iteration limit corresponds to the approximate worksheet row limit. For Excel 2003 worksheets, the limit is 65000. For Excel 2007, the limit is 1000000.

1. INTRODUCTION

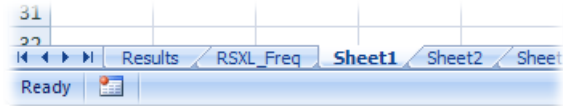


Figure 1.7: Worksheet Tabs

A screenshot of an Excel spreadsheet showing a column of data in column A, rows 1 through 18. The data values are: 3, 6, 4, 7, 5, 7, 3, 4, 5, 5, 5, 2, 3, 3, 3, 6, 3, 3.

	A
1	3
2	6
3	4
4	7
5	5
6	7
7	3
8	4
9	5
10	5
11	5
12	2
13	3
14	3
15	3
16	6
17	3
18	3

Figure 1.8: Results, Unsorted

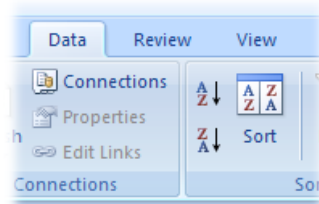
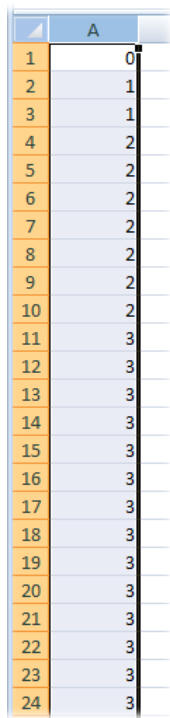


Figure 1.9: Excel's Sort Buttons



A screenshot of a spreadsheet showing a column of 24 rows. The first row is a header with the letter 'A'. The data values are: 0, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3. The rows are numbered 1 through 24 on the left side.

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

Figure 1.10: Results, Sorted



Figure 1.11: Histogram Button

Figure 1.9. The result of an “A to Z” (ascending) sort is displayed in Figure 1.10.

Next, while still on the Results sheet, use the Histogram feature of Resampling Stats to produce a frequency histogram of these results. Select the “Histogram” button (Figure 1.11) from the Resampling Stats toolbar, or “Histogram” from the Resampling menu.

Then, in the histogram dialog box (Figure 1.12), specify the input for the

1. INTRODUCTION

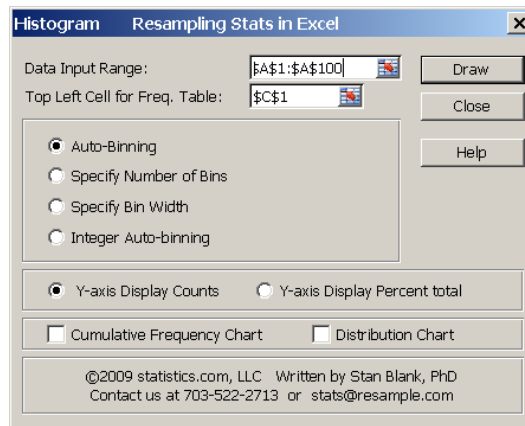


Figure 1.12: Histogram Dialog Box

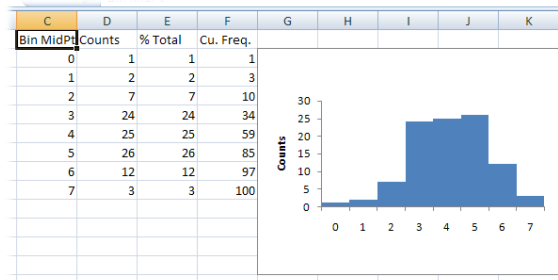


Figure 1.13: Histogram, # of Heads in 8 Tosses of a Coin

histogram, which is the results from the resampling experiments – A1:A100.⁴

For the Top Left Cell for Freq. Table specify the top left cell for any empty area – you can choose “C1” in this case. Change “Auto-Binning” to “Integer Auto-binning,” which works better with results that are exclusively integers. Click “Draw” to draw the histogram.

You should get something like Figure 1.13 (your results will be slightly different – remember that this is the result of 100 random trials).

You can move the graph around by clicking and holding somewhere inside

⁴An easy way to select the input for the histogram is to click on the top cell of the range for which you want to draw a histogram. Resampling Stats will proceed down until there is a gap in the data, and use that selection as the input. (The Data Input Range field in your histogram dialog must be active before you do this; click in it to make it active first.)

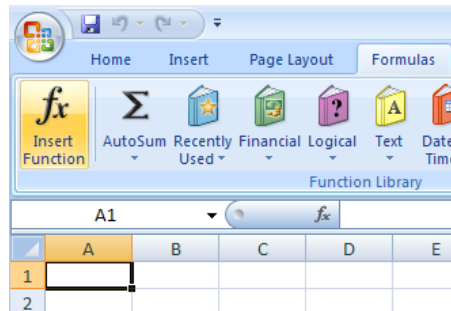


Figure 1.14: Excel's Insert Function Button

the graph area, and dragging. For additional details about histogram operations, see Histogram in the Resampling Stats Operations section at the end of this guide.

From the histogram output we can see that 25 of the 100 trials yielded exactly 4 “1’s.” So our estimate of the probability of having 4 boys in 8 children is 0.25.

To get a better estimate, we should now run the same procedure for a large number of trials, say, 1000 trials or more.

- **Note:** It’s always a good idea to try a Repeat and Score run of 100 trials first. If something wasn’t set up correctly, you’ll find that out in 1/100 of the time it would take for a 10,000-trial run.

1.6 Counting Results

One Excel function that you will be using regularly is the COUNTIF function, which you can reach through the Insert Function button in the Excel Formulas menu or by clicking the “*fx*” symbol immediately to the left of the formula bar as shown in Figure 1.14:

COUNTIF lets you count the number of values in a range meeting a specified criterion – “how many 1’s,” for example.

Still on the Results sheet, position the cursor on a blank cell where you want the count result to appear – say B1. Click on the “Insert Function” button, and the Insert Function dialog box comes up (Figure 1.15). You

1. INTRODUCTION

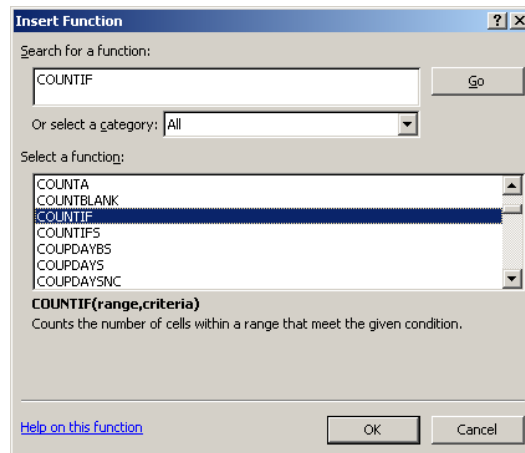


Figure 1.15: Excel's Insert Function Dialog Box

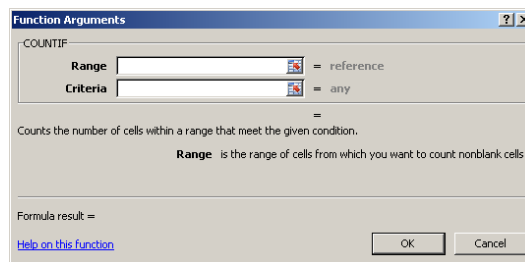


Figure 1.16: COUNTIF Dialog Box

may select the “COUNTIF” function by typing COUNTIF in the Search for a function edit box followed by clicking GO or you can select “All” as a category and scroll through the functions to find the one you want to use.

Select “COUNTIF,” click “OK,” and the COUNTIF dialog box will display as in Figure 1.16:

Enter a1:a100 in the Range field and =4 in the Criteria field; this will count the number of times the value 4 appears in the range a1:a100. The result, 25, is the same as we got reading directly off the frequency table.

Excel's Histogram function

Excel itself also has a histogram function, which you can reach from Data > Data Analysis > Histogram. (If you don't see the Data and Data Analysis menus, make certain the Analysis Toolpak is installed as described earlier.) Excel's histogram function does not do a good job of binning when there are small numbers of possible values, so we recommend the use of the Resampling Stats histogram function in most cases.

Frequency Distribution Worksheet

When the Repeat and Score (**RS**) dialog box is visible (Figure 1.6) it is possible to select the Results Worksheet (the default selection), the Frequency Distribution Worksheet (new in Version 4), or both. If either of the Frequency Distribution Worksheet options are selected, the RSXL.Freq worksheet (shown in Figure 1.7) will contain the frequencies of each of the possible score cell outcomes. These results are unsorted.

Auto-Reset

The Auto-Reset option is selected in the Resampling Options menu by default. This means that an automatic reset of all variables and ranges will be performed prior to each *new* simulation. A reset ensures that when it is time to Repeat and Score, only the resampling in the current problem gets repeated. If the Auto-Reset option is not selected, then each Repeat and Score will result in the scoring of ALL resampling and shuffling operations you have performed since the last reset. Of course, you may at any time click on the "Reset" button manually to clear all variables and ranges to start a new simulation.

Rolling the Dice

For another simple demonstration of resampling, let's look at a simple simulation: rolling a pair of dice (Figure 1.17) One die has six sides, with dots that represent the numbers one through six.

To simulate one roll of two dice:

1. INTRODUCTION

	A	B	C	D
1	1	<i>First</i>	6	
2	2	<i>Second</i>	3	
3	3			
4	4	DICE	9	SUM
5	5			
6	6			
7				

Figure 1.17: Rolling a Pair of Dice

1. List the numbers one through six in the range A1:A6 (see the workbook Dice.xls).
2. Click **“R”** (Resample) on the toolbar or select “Resample” from the Resampling menu to roll two dice at once, by specifying 2 as the Number of Cells in Output Range, then click “OK”.
3. Add these two numbers and put the sum into another cell with the SUM function.
4. Click **“RS”** on the toolbar, or select “Repeat and Score” from the Resampling menu, specify, say, 1000 trials, and again you’ll see the results appear on the Results worksheet.

There are several questions related to actual dice games you can answer from the basic Results sheet using a few Excel functions. One such question is “What is the probability of obtaining a 7?”

To answer this question, you want to count the number of “7’s” in the results range. With the cursor in a blank cell on the Results sheet (say, B1), use the Insert Function button on the Excel Toolbar and select the “COUNTIF” function. Enter a1:a1000 as the Range (recall that we did 1000 rolls of the dice), and =7 as the Criteria.

In the results in Figure 1.18, a 7 occurred in 168 out of the 1000 trials, for a proportion of 0.168.

Let’s run this simulation a second time. Click on the **“RS”** button (or select “Repeat and Score” from the “Resampling” menu). You should see a smaller dialog as shown in Figure 1.19. You have three options. You can use the previous resampled data and choose “New Score Cells.” You can “Redo the Previous Problem,” or you can redo the previous problem and

	A	B	C	D	E	F
1	5	2	7			
2	8	0.168				
3	6					
4	9					
5	4					
6	4					
7	6					
8	12					
9	10					
10	5					
11	9					

Figure 1.18: Getting a 7 When Rolling Two Dice

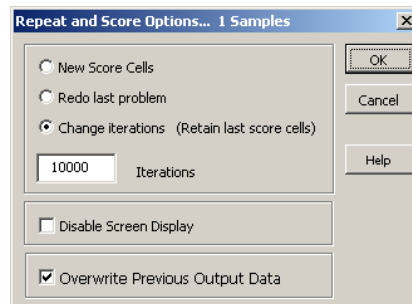


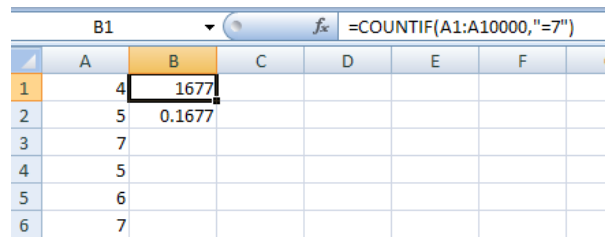
Figure 1.19: Repeating the Repeat and Score

“Change Iterations.” Select the “Change Iterations” option, which will retain the previous score cell, and type in the value 10000. This will toss the dice 10,000 times. Click “OK” and watch the iterations progress in the lower left corner of the Excel worksheet.

Figure 1.20 shows the results of tossing the two dice 10000 times. The =COUNTIF function has been modified to reflect the new a1:a10000 output range. If you selected both the “Results Worksheet” and the “Frequency Distribution Worksheet” options in the Repeat and Score dialog, selecting the “RSXL_Freq” worksheet will show output similar to Figure 1.21.

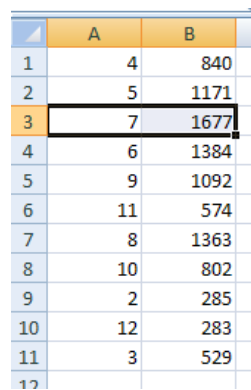
Notice that the frequency for 7 is 1677, which is identical to the COUNTIF output on the Results worksheet.

1. INTRODUCTION



	A	B	C	D	E	F	G
1	4	1677					
2	5	0.1677					
3	7						
4	5						
5	6						
6	7						

Figure 1.20: Results of 10000 Iterations of Tossing Two Dice



	A	B
1	4	840
2	5	1171
3	7	1677
4	6	1384
5	9	1092
6	11	574
7	8	1363
8	10	802
9	2	285
10	12	283
11	3	529
12		

Figure 1.21: Frequency Distribution Worksheet Output

Variations

In these simple examples, we used 1 and 0 to represent states (“heads” or “tails”, “seven” or “not seven”). In Excel, it’s easy to use more evocative labels, as shown in Figure 1.22. The Excel COUNTIF function (used in the worksheet Boys&Girls.xls) can tabulate resampled results of names in the same fashion as numbers, and the Resampling Stats add-in can resample or shuffle words as easily as numbers.

1.7 The Frequency Function

A Results sheet can also be analyzed with the Excel function called FREQUENCY (Figure 1.23). FREQUENCY is “live” – it updates itself every time you present it with new resampled output, while Histogram has to be called again every time the data changes. Here FREQUENCY is used (Fig-

1.7. The Frequency Function

	A	B	C	D
1	BOY		GIRL	
2	GIRL		GIRL	
3			BOY	COUNTIF Totals
4			BOY	# of Boys
5			BOY	4
6			BOY	
7			GIRL	
8				
9				
10				

Figure 1.22: Names vs. Numbers

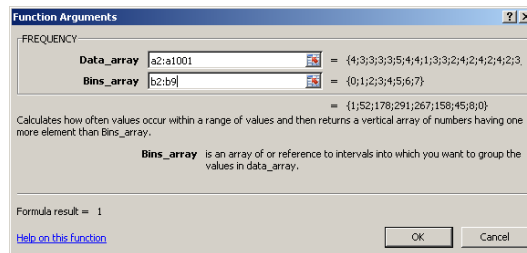


Figure 1.23: The FREQUENCY Dialog Box: (Control+Shift+Enter) should be pressed instead of OK

ure 1.24) to analyze 1000 trials of a “how many boys in a family of seven?” experiment.

We will be working with the Results sheet of this experiment. First, insert a row and add the label “output” at the top of column A. Next, enter the values 0-7 in the cells B2:B9. Then, position the cursor in cell C2 on the Results sheet and highlight the cells C2:C10. Select “FREQUENCY” via the Insert Function button. For the Data_array enter A2:A1001 and for the Bins_array enter B2:B9; see Figure and FREQUENCY in the Useful Excel Functions part of the syntax section (Resampling Stats Operations) for more detail.

- **Important:** Next, press **Control+Shift+Enter**. This is an “array function” (meaning that it works with arrays of numbers) and must be entered by holding down **Control+Shift** while you press “Enter”.

1. INTRODUCTION

	A	B	C
1	output	bins	frequencies
2	4	0	1
3	3	1	52
4	3	2	178
5	3	3	291
6	3	4	267
7	5	5	158
8	4	6	45
9	4	7	8
10	1		0
11	3		
12	3		
13	2		
14	4		

Figure 1.24: Analyzing 1000 Trials

Chapter 2

Advanced Probability

One advantage of resampling is that the same approach that produces answers in simple problems also produces answers in more complex problems, with little additional conceptual effort on your part. In simulating probability problems with resampling, the computer may have to work a little longer on harder problems, but *you* don't.

2.1 Rates and Results

Many probability questions ask you to decide how often a collection of events occurs, given a probability of the events happening one at a time (a base rate). The examples here are taken from sports, an area where journalists spill much ink on the matter of short-term departures from expected base rates.

Baseball

Here's a basic probability question that is just a bit beyond those in Chapter One. A baseball player has a .250 batting average. How often will he get 3 hits in five at-bats?

Take a look at Figure 2.1 for one way to model the situation (it's the Baseball.xls file in the Worksheets folder). First, the batting average information is represented in Column A by the numbers 0, 0, 0, and 1. If you pick from this set at random, you get a hit (a "1") one-fourth of the time. Now make 5 selections at a time from this using Resample (which selects with replacement,

2. ADVANCED PROBABILITY

so there is always a $1/4$ chance of a hit). Add up the hits in each set of 5 attempts using the Excel SUM function.

	A	B	C	D	E	F	G
1	BASEBALL: MODELING HITS FOR A 0.250 BATTING AVERAGE BINOMIAL TABLE (for comparison)						
2							
3						hits	prob
4	0		0	<-RESAMPLES		0	0.237
5	0		0			1	0.396
6	0		1			2	0.264
7	1		0			3	0.088
8			0			4	0.015
9			1	<-SUM		5	0.001
10							

Figure 2.1: Baseball Hits

Now select “Repeat and Score,” picking the sum cell as the cell to score, and run the selection as many times as you like. One way to get a quick approximate answer is to try 100 repetitions, view the Results worksheet and use Excel’s Sort tool (under Data on the menu) to sort Column A in descending order. If you do this, you can simply count the number of 3’s in the output, corresponding to 3 hits in 5 at-bats. Note that the theoretical probabilities for each number of hits are displayed on the right for comparison. We will show you how to compute the theoretical probabilities using the analytical approach later.

Once again, step-by-step:

1. Highlight the batting average range, A4:A7, and select “Resample” from the Resampling Stats toolbar (the R button) or the Resampling menu.
2. Enter C4 as the top cell of the output range, 5 as the Number of Cells in Output Range, and click “OK”.
3. In cell C9, enter the Excel formula =SUM(C4:C8) to sum up the number of hits in the resample.
4. Select “Repeat and Score” from the Resampling Stats toolbar or menu, and make sure C9 is the input cell; set the number of iterations to 100 and click “OK”.
5. On the Results sheet, use Excel’s sort button to sort the results in descending order.

6. Count the number of times 3 occurs (you can also use the COUNTIF function) and divide by 100 (number of iterations) to get the estimated probability of 3 hits in 5 at-bats.

In one set of 100 trials, a result of 3 was encountered 8 times, so an initial estimate of the probability of 3 hits in 5 at-bats would be 0.08.

Compare to the Analytical Approach

In the previous problem, the probabilities are easy to model analytically, using Excel's built-in function for binomial distribution (BINOMDIST, Figure 2.2). For the set of 5 trials (at bats) there are 5 possible outcomes. The probability of success is 0.250 in each at bat. Start by arraying in cells E3:E8 the possible outcomes of five at-bats, ranging from zero hits to five hits. To the right we will use BINOMDIST to calculate the theoretical probabilities of each outcome (using the binomial theorem to find the probability of x successes in n independent events with constant p probability of success in each event).

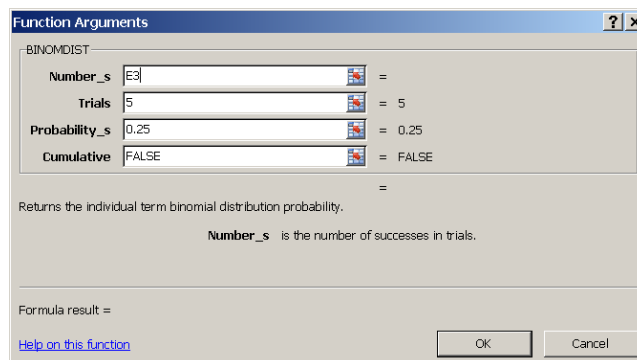


Figure 2.2: BINOMDIST Dialog Box

Starting in cell F4, use the Paste Function button to open the BINOMDIST dialog box (Figure 2.2), Number_s refers to the worksheet cell that contains the outcome whose probability you want to count (you start with cell E3, indicating zero hits). Trials in this case means the number of at-bats (5). Probability is .25 (the hitters .250 batting average). Cumulative is set to FALSE to indicate that you want an individual probability (the chance of getting the specified number of hits exactly) not a cumulative probability (the chance of getting up to the specified number of hits).

2. ADVANCED PROBABILITY

After entering the information for the dialog box, click “OK” and then copy the function to cells F4:F8. Note how the referenced outcomes changes. The resulting probabilities are shown in Figure 2.3.

E	F	G
BINOMIAL TABLE (for comparison)		
	hits	prob
	0	0.237
	1	0.396
	2	0.264
	3	0.088
	4	0.015
	5	0.001

Figure 2.3: Binomial Probability Table

With these theoretical probabilities in hand, you can estimate in advance what you’ll find in a simulation run. Resampling and Scoring 1000 trials, what do you find, compared to the 88 3’s predicted? You might want to do 1000 runs ten times or so, record the results, and inspect the distribution of outcomes. You might also try a single run of 10,000 or even 100,000 trials (if you are patient).¹

Basketball

Here is another example of attention being drawn to a short-term departure from long-run averages. On one occasion, former basketball great Larry Bird’s shots for a 3-day period were examined and compared to his expected accuracy of 48%. In 3 days, he made only 20 of 57 shots. What are the chances that, if his true accuracy rate has remained unchanged, he might do this badly in a series of 57 shots just by chance?

¹The Repeat and Score limits are 65000 trials for an Excel 2003 worksheet and 1000000 trials for an Excel 2007 worksheet. For trials above 65000, you *must* use an Excel 2007 workbook.

The Urn Function

For this example, perhaps the simplest modeling method is to use the Resampling Stats Urn function to create an “urn” with specified numbers of “baskets made” and “baskets missed” in it. You could also think of it as a hat or box containing slips of paper. The more morbid appellation “urn” has a long and distinguished usage in probability pedagogy.

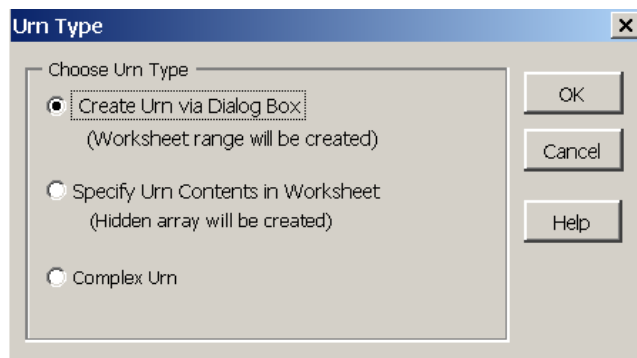


Figure 2.4: Choosing How to Create the Urn

You will be asked to choose between creating an urn via a dialog box, specifying the urn contents on the worksheet, or creating a complex urn. The first method is used in Figure 2.4. (See “Urn” in the “Resampling Stats Operations” section at the end of this guide.)

Creating an Urn Via a Dialog Box

The Urn dialog box (Figure 2.5) models a container with different numbers of possible outcome types. You could think of the model for Larry Bird as an urn with 48 black marbles (baskets made) and 52 white marbles (baskets missed). To simplify calculation, this example (Basket.xls) uses “1” for a basket and “0” for a miss. The Urn function then puts 100 entries, 48 “1’s” and 52 “0’s,” into Column A. Specify A1 as the Top Cell of Urn Output Range (you may type in A1 or you may click in the edit box and then click in cell A1 as was done in Figure 2.5).

To test the situation proposed in the problem, resample from these 100 cells, making 57 draws at a time. Use the SUM function to add up the number

2. ADVANCED PROBABILITY

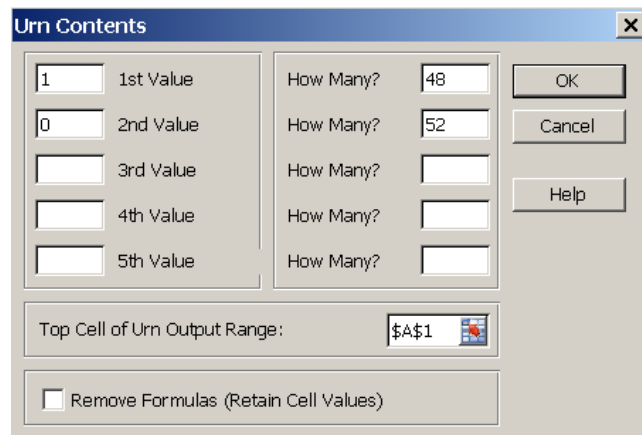


Figure 2.5: The Urn Dialog Box

of baskets in the simulated 57 shots, and Repeat and Score on the sum cell (C3 in Figure 2.6).

	A	B	C	D	E
1	1	0	<- 57 Selections from Column A		
2	1	0			
3	1	1	24		
4	1	0			
5	1	1			
6	1	0			
7	1	0			
8	1	0			

Figure 2.6: Totalling the Baskets

Sometimes, when using Repeat and Score, you will want to leave unchecked the Disable Screen Updating box in the Repeat and Score dialog. For a modest set of trials, say 1000 or so, it's worth taking the speed penalty to watch the numbers flicker past in the score cell. Try it a few hundred times, just to get a feeling for the probabilities. How often do you see a 20? a 19?

For a final estimate, do 2000 trials and use the "Histogram" function to summarize the results. How likely is it that Larry Bird would hit 20 out of 57, just by chance?

Since in his career Larry Bird had hundreds of series of 57 shots, it would

not have been unlikely for him to have done this badly just by chance a number of times.

2.2 Simulation and Hard Problems

The problems in the previous section were harder than Chapter One problems, but still fairly simple in that they could be solved by applying the binomial formula (BINOMDIST). To show some of resampling's power, let's look at three problems. Most introductory textbooks just state these problems with an answer, rather than explaining the calculation details.

ESP

The five symbols shown in Figure 2.7 are the central element in the most rigorous experiments yet performed to investigate extra-sensory perception. The experimenter deals out the five cards face down, and you try to guess which symbol is on each card.

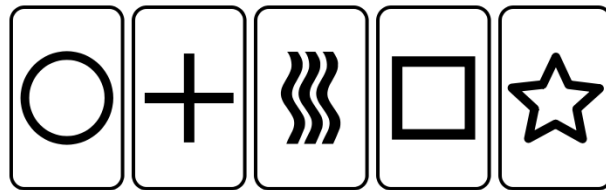


Figure 2.7: Zener Cards Used in ESP Experiment

How well might you do by just guessing?

There are actually two ways to do this experiment (File ESP.xls). It could involve just five cards, shuffled and dealt out over and over again. Or it could use a very large deck, shuffled once and dealt out in sets of five cards at a time. The difference here is that in the second case, the set of five facedown cards might contain, say, two stars, while in the first case each symbol appears only once.

For modeling the first case, we can use the Resampling function called Shuffle, which means exactly what it says. Like the dialog box for Resample,

2. ADVANCED PROBABILITY

the dialog box for Shuffle (Figure 2.8) asks for an input range, an output range, and number of output cells.

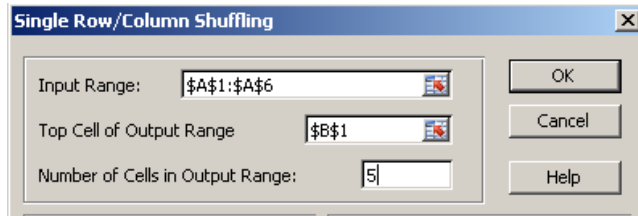
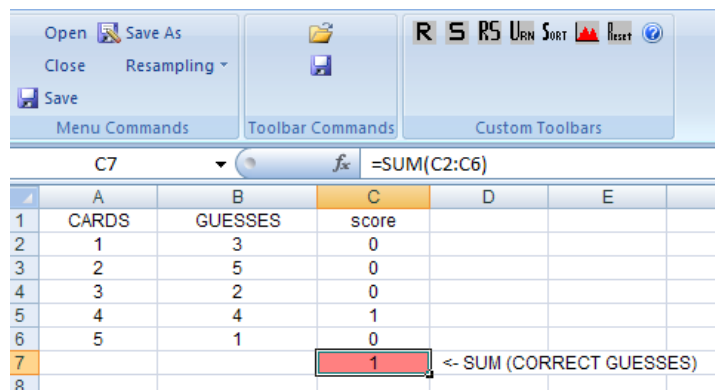


Figure 2.8: Shuffle, the “Other” Resampling Function

We proceed as follows:

1. In A2:A6 enter the numbers 1-5 to represent the 5 cards, and select this range.
2. From the Resampling Stats toolbar or menu, select “Shuffle” (“S” on the toolbar).
3. The input range should already be entered (it’s A2:A6), enter (or click on) B2 as the top cell of the output range and 5 as the Number of Cells in Output Range, click “OK”.



	A	B	C	D	E	F
1	CARDS	GUESSES	score			
2	1	3	0			
3	2	5	0			
4	3	2	0			
5	4	4	1			
6	5	1	0			
7			1			
8						

Figure 2.9: The ESP Model

We now have the “actual” cards in column A, and the “guesses” in column B as illustrated in Figure 2.9.

Now proceed as follows:

1. Put Excel's IF function in column C to see if there is a match between the values in columns A and B.

There are two ways to do this. You could select Excel's "Insert Function" button, choose "IF", and fill in the dialog box as shown in Figure 2.10:

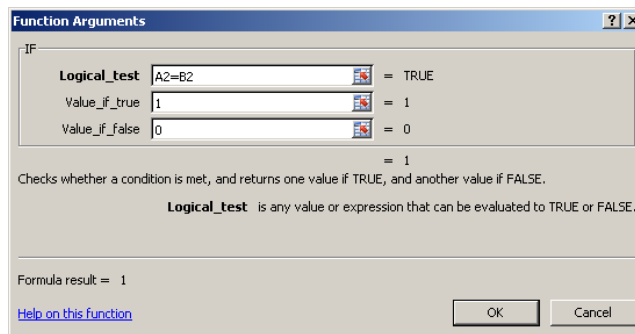


Figure 2.10: The IF Function

Or, you could enter the following formula directly in C2: `IF(A2=B2,1,0)`.

Either one translates to "if A2=B2 enter a '1' in C2, otherwise enter a '0' in C2."

2. Copy this formula down through C6.
3. In C7, SUM the values C2:C6. This is the number of matches by chance.
4. With C7 (our score cell) highlighted, select "Repeat and Score" from the Resampling Stats toolbar or menu (**RS** on the toolbar), and enter your desired number of iterations (repeats). On the Results sheet, we can use the Resampling Stats histogram function (Histogram on the menu, or the graph button on the toolbar) to produce a histogram of the results. Just select the top cell of the results column, select "Histogram" (or click the "Histogram" button), and Resampling Stats will select downward in the column until it encounters a gap in the data, and produce a histogram. In this case, we want to select the "Integer Auto-Binning" option. All our possible outcomes (the x-axis values) are integers, so we want to force the Histogram to have integers as the x-axis bin centers.

2. ADVANCED PROBABILITY

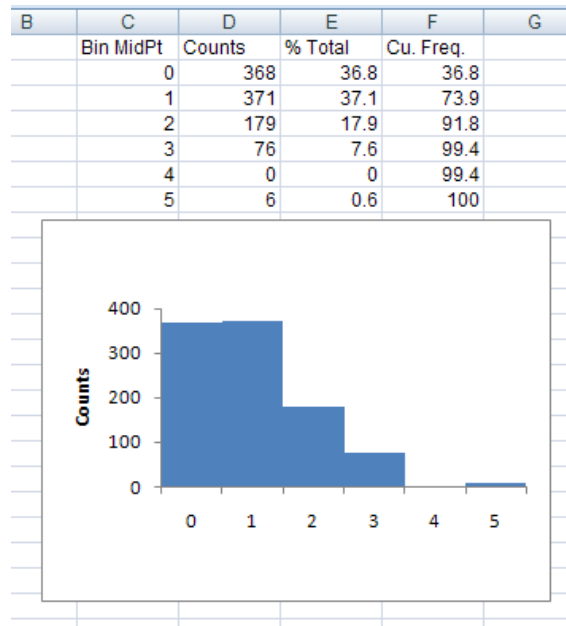


Figure 2.11: ESP Data

From the histogram in Figure 2.11, you can see that just over 35% of the time, there were zero matches, and that over 25% of the time there were two or more matches.

Sampling with Replacement

The second case involves a slight twist, to save the trouble of defining a very large deck and shuffling it. Since the deck is very large, with equal numbers of stars, circles, squares, etc., we can effectively regard the probability of drawing, say, a circle as unchanging from draw to draw. This we can model using Resample, instead of shuffle, since the “sampling with replacement” aspect of Resample guarantees unchanging probabilities from one draw to the next.

To model selection of five at a time from an infinitely large deck, Resample cells A2:A6 to B2:B6. The rest of the problem proceeds as before.

- Which case is more likely to show 5 correct guesses at a time?

- How would you modify your procedure to check for 25 guesses at a time (the classic original experiment)?

If you're curious about this problem, search the Web for the phrase "Zener cards," the name of the five symbol cards. There are continuous on-line experiments, and you can compare your Excel resampling output to the experimental results piling up.

The Birthday Problem

The classic birthday problem is this: "How many people have to be in a room for it to be likely (at least a 50% chance) that two of them have the same birthday?" Lets pose the question in a different manner, starting with the answer:

"If there are 23 people in a room, what is the probability that two or more of them share the same birthday?"

Let's proceed as follows:

1. Using Excel's Autofill function, we put the numbers 1 to 365 (all possible birthdays) in cells A1 to A365 respectively (we'll ignore leap years for simplicity).

Autofill

Excel's Autofill function lets you fill in adjacent cells simply by selecting several cells that establish the series pattern, then dragging down. Suppose you select the value "1" and "2" in cells A1:A2:

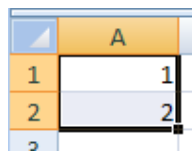
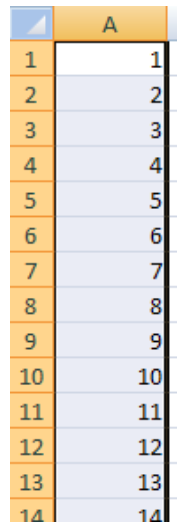


Figure 2.12: Autofill

Click precisely on the little square at the lower right of A2 (as shown in Figure 2.12), and drag down. The outcome is shown in Figure 2.13.



	A
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
14	14

Figure 2.13: Results of Autofill

Note that Excel detects the pattern and fills the rest of the range appropriately as you drag down. Had you selected simply the “2,” instead of the values “1” and “2,” Excel would have put “2’s” in the cells below as you dragged down.

2. Resample the range A1:A365 as input and set cell C1 as the top cell in the output range, followed by 23 as the number of output cells (for 23 people in the room).
3. Cells D1:D23 contain an Excel formula that will determine whether a match occurs. This formula:

$$=MATCH(C1,C1:C23,0)$$

in cell D1 (it’s already there in Birthday.xls) has been copied down to cells D2 through D23. The first term is a relative term (and will be adjusted accordingly during the copy procedure) and contains the reference to the cell we are examining – in this case, the cell immediately to the left of the formula. The second term ($\$C\$1:\$C\23) is the range against which we check for a MATCH with the first term, in this case range of the 23 randomly selected birthdays. The final term tells Excel to look for an exact match.

Essentially, the formula looks at the reference cell to the left and checks it against the entire range. If a match is found, the row number of the match is returned as the cell value (see Figure 2.14).

	A	B	C	D	E
1	1		101	1	11
2	2		162	2	
3	3		90	3	
4	4		131	4	
5	5		56	5	
6	6		116	6	
7	7		133	7	
8	8		12	8	
9	9		107	9	
10	10		238	10	
11	11		124	11	
12	12		153	12	
13	13		162	2	<- A MATCH!
14	14		191	14	
15	15		3	15	
16	16		302	16	
17	17		86	17	
18	18		130	18	
19	19		57	19	
20	20		258	20	
21	21		321	21	
22	22		297	22	
23	23		311	23	
24	24				

Figure 2.14: Looking for Birthday Matches with MATCH

If no matches occur, then the numbers 1 through 23 are returned in the range containing the MATCH formulas in D1:D23 (the only match is the cell to itself). If a match occurred, this will not be the case. Look at cells C2 and C13 (highlighted in Figure 2.14; both show day 162) and then look at cells D2 and D13. D2 indicates that C2 matches itself (position 2 in the C1:C23 reference range) but D13 also returns a 2, indicating that C13 encounters its first match in row 2.

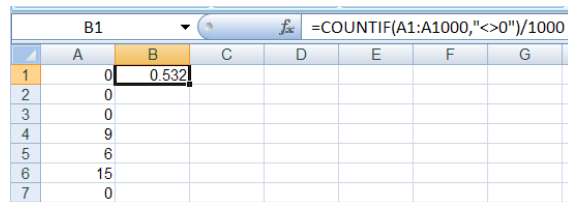
4. To determine if a match occurred (2 birthdays the same), compare the sum of the D1:D23 formula range with the sum of the numbers 1 through 23 (in A1:A23). If they are equal, no birthdays matched. If they are different, this is because a match occurred to some number other than itself, and we have a winner!

Cell E1 finds the difference in the range sums (A1:A23 minus D1:D23). In Figure 2.14, this difference is 11 indicating a match.

2. ADVANCED PROBABILITY

- Repeat and Score on cell E1. The Results sheet shows numbers that are either indicators of a match for that trial (result not equal to zero) or no birthday match (a zero).
- Using the COUNTIF function on the results column, count the number of times we found a birthday match (i.e., a non-zero value).

For 1000 trials, we typically get a value very close to 0.5, confirming that you only need 23 people in a room for the probability of a birthday match to reach 50%. To make this more informative, try this experiment yourself for 20 people and for 26 people. In Figure 2.15, we obtained a result of 0.532 for 1000 trials and 23 people.



	A	B	C	D	E	F	G
1	0	0.532					
2	0						
3	0						
4	9						
5	6						
6	15						
7	0						

Figure 2.15: Counting Matches on the Results Sheet

The Secretary Problem

In a classic textbook probability problem, a secretary mixes up fifty letters and envelopes at random. You're asked to determine the probability that at least one person receives the correct letter (that is, the one that matches the envelope address). To illustrate the method, let's start with a 10-letter case (the sample worksheet file is Secty.xls).

- Put the numbers 1 through 10 in a range (A3:A12) in column A. Select that range.
- Pick "Shuffle" from the Resampling Stats menu or toolbar.
- The input range is A3:A12, the top cell of the output range is B3, and the Number of Cells in Output Range is 10.
- To see if somebody got the right letter, select cell C3 next to the top two values (original and shuffled), and enter an IF function in C3.

If you do this via the Paste Function dialog, enter $A3=B3$ for Logical test. The “Value_if_true” line field has a “1” in it, and the “Value_if_false” line field has a “0” in it.

You can also type this function directly into the cell:

$=IF(A3=B3,1,0)$.

Either way, it translates to “If $A3=B3$, then place a ‘1’ in C3, otherwise place a ‘0’ in C3.”

Copy the IF formula into the ten cells C1:C10 in column C to test the match on all ten pairs. If you like, shuffle a few more times with this IF test in place, to watch how often the shuffled list makes a match.

5. Put $=SUM(C1:C10)$ into C14, or another cell near these columns.
6. Use this as the score cell for Repeat and Score (Figure 2.16).

	A	B	C
1	envelope	shuffled	match?
2		letter	
3	1	1	1
4	2	8	0
5	3	3	1
6	4	7	0
7	5	4	0
8	6	2	0
9	7	6	0
10	8	5	0
11	9	10	0
12	10	9	0
13			
14			2
15		# right	

Figure 2.16: The Secretary Problem

7. Try 1000 repetitions, then check the Results sheet.
8. Select the top cell of the results, then use Resampling Stats Histogram feature (again, select “Integer Auto-Binning”). The results of our simulation are shown in Figure 2.17.

2. ADVANCED PROBABILITY

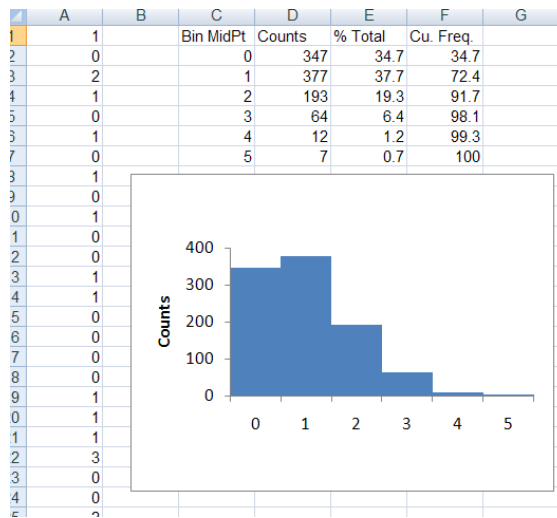


Figure 2.17: The Secretary Problem: 10 Letters and 1000 Trials

347 shuffles out of 1000 produce no matches, but 653 out of a hundred produce at least one match.

To treat the 50-letter case, just extend the list of numbers to 1 through 50, extend the shuffle to cover all 50, and copy the IF function down column C to test for matches. Now try a Repeat and Score for 1000 repetitions, and once again make a Histogram of the results. You will see something like $629/1000 = 0.629$ as the probability of at least one match. If you try a very long run, 100000 iterations or more, you'll get the exact first three digits of the result quoted in textbooks. Its probably safe, for this problem, to say you can find the answer faster with resampling than you could by trying to work out the analytic solution.

Chapter 3

Confidence Intervals

How accurate is an estimate based on a sample of data from a larger population? This depends on how variable different samples are when drawn from that same population. A confidence interval is an estimate of the range that would enclose most (say, 90% or 95%) of the sample estimates, given the sampling variability of the data. In the discussion below, we will consider the case of the sample mean.

The Resampling World

How do many different sample means turn out when the samples are drawn from the same population? If you've been following along in the spirit of resampling, you will quickly conclude that the best way to find out is to actually draw them.

Of course, we don't have available the entire population to draw samples from. If we did, we wouldn't be bothering to ask how accurate the sample was. To make an estimate of how samples drawn from that population behave, we need a proxy population that embodies everything we know about the real population, and which we can use to draw samples from.

One resampling technique is to replicate the sample data a huge number of times to create a proxy population based entirely on our sample. After all, the sample itself usually embodies everything we know about the population that spawned it, so it's often the best starting point for creating an artificial proxy population.

3. CONFIDENCE INTERVALS

Once the sample is replicated (say) millions of times, we can proceed to draw samples from this proxy population and record how they behave. In the case of a confidence interval for a sample mean, we can observe the distribution of sample means.

The Bootstrap

There is a shortcut that saves us the trouble of replicating the sample a huge number of times – simply sample with replacement from the original sample. By sampling with replacement, each sample observation has $\frac{1}{n}$ probability of being selected each time – just as if you were drawing without replacement from an infinitely large replicated population. This technique is called the bootstrap.

Drawing samples with replacement from the observed data, we record the means found in a large number of samples. Looking over this set of means, we can read the values that bound 90% or 95% of the entries. It's also a computationally simple matter, in resampling with Excel, to find confidence intervals for sample medians or other statistics.

For Comparison: The Classical Statistics World

In classical statistics, we still invoke the concept of the larger population. However, rather than creating a proxy population and actually drawing from it, classical statistics works from a mathematical description of this larger population, based on information provided by the sample.

Generally, this mathematical description of the larger population is built as follows:

1. We make the assumption that the real world data are normally-distributed (or invoke laws that state that nonnormally-distributed data can be used if the sample size is large enough);
2. We use the observed sample statistics (generally mean and standard deviation) to estimate these same parameters of the larger population.

Once the parameters of the larger, normally-distributed population have been estimated, we can look up in tables the distribution of sample means for samples of various sizes.

It is important to note that both the resampling and classical approaches start from the same point – the observed sample. They both use it to build a description of the larger population that we think spawned it. If the observed sample is way off base, both approaches are in trouble.

If the assumptions underlying our mathematical description are accurate, this mathematical description of the larger population will be a bit more complete than our bootstrap description, and perform a bit better. If the assumptions are not justified, the bootstrap may be a more appropriate procedure. Additionally, if the statistic you are interested in is not one whose distribution can be determined analytically, the bootstrap is the only way to proceed.

3.1 Confidence Interval for Means

Drills

Let's consider the problem of finding a confidence interval for drill-bit diameters (worksheet Drills.xls).

The question we wish to answer is “When estimating average drill bit diameter on the basis of a sample of 52 bits, how much might that estimate be in error?”

If we could, we would draw additional samples of 52 drill bits and see to what extent they differ from one another. Lacking the time and resources to gather more data (and this is almost always the case), we instead draw samples from a proxy population – the observed sample replicated over and over many times to create a huge artificial population of drill bit diameters.

Actually, we will use a shortcut – sampling with replacement from the observed data set.

The resampling procedure for estimating a 95% confidence interval (a bootstrap percentile confidence interval) for drill bit diameter follows these steps:

1. Draw 52 observations randomly and with replacement from the drill diameter data.
2. Take the mean of this resampled set.

3. CONFIDENCE INTERVALS

3. Repeat steps 1 and 2, say, 1000 times. Record the resampled mean each time.
4. Sort the results.
5. Find the 2.5th and 97.5th percentiles (these percentiles bound the 95% in the center).

Here's how to implement these steps using the Resampling Stats add-in:

1. Select the data (A3:A54) in the Drills.xls workbook and choose "Resampling" from the Resampling Stats menu or toolbar. Put the resampled data in an adjacent column by selecting cell B3 as the top cell of the output range and 52 as the number of output cells.
2. To find the mean of the resampled data, enter this formula in C3:
`=AVERAGE(B3:B54)`
3. Select "Repeat and Score" from the Resampling Stats menu or toolbar, and select C3 as the cell to score, with 1000 repetitions.
4. Sort the results (on the Results sheet) in ascending order.
5. Find the 2.5th and 97.5th percentiles using Excel's PERCENTILE function from the Insert Function button as shown in Figure 3.1:

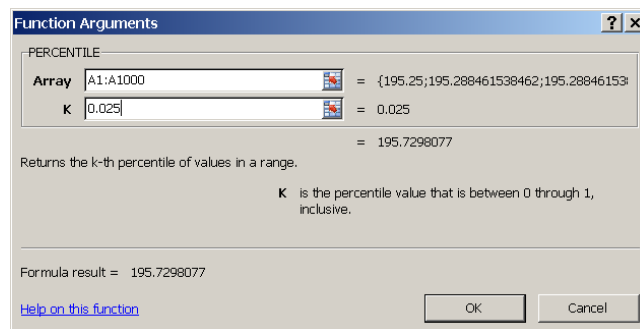


Figure 3.1: Finding the 2.5th Percentile

In the PERCENTILE function, the 2.5th percentile is specified as .025, and the 95th as .975.

3.1. Confidence Interval for Means

	A	B	C
1	195.25	2.5th Percentile:	195.7298
2	195.2885	97.5th Percentile:	197.5582
3	195.2885		
4	195.3077		
5	195.3269		
6	195.3269		
7	195.4808		
8	195.4808		

Figure 3.2: C.I. Estimate, Drill Diameters

With the cursor on a blank cell in the Results sheet (here, C1 in Figure 3.2), select “PERCENTILE” from the Insert Function button. The array that we want to find a percentile for is a1:a1000, and the percentile we want is .025 (this is the 2.5th percentile). Repeat the same procedure (with the cursor in a different cell, C2) to find the 97.5th percentile.

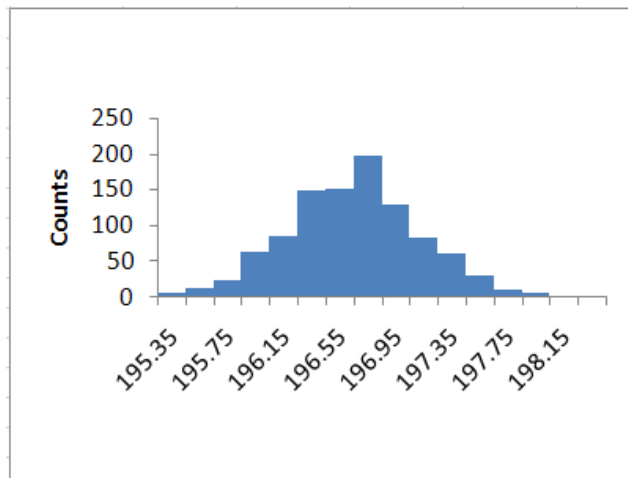


Figure 3.3: Histogram of Resampled Diameter Means

You can also call on Histogram in the Resampling Stats toolbar or menu to inspect the results of the procedure graphically (Figure 3.3).

3. CONFIDENCE INTERVALS

Weather

Cazadero, CA is a tiny town tucked in a redwood forest, with a unique and distinctly damp microclimate (about 90 inches of rain per year). There are more than 100 years of rainfall records, documenting the number of rainy days each month, since the days when the town was a lumber camp supplying San Francisco with building materials.

In May, it rains 15 days, on the average. In the early 1990's, a five-year period produced an average of 18 rainy days in May, and local newspapers began citing this as evidence of a weather shift related to global warming. To resolve this "Quality Control in the Sky" issue, consider the worksheet Rain.xls, which contains 100 years' worth of data on the number of rainy days in May. We resample the rainfall data, in sets of 5 resamples at a time, for 100 Repeat and Score trials (the selected cell is the average number of rainy days). This gives us an estimate of the range within which a 5-year average might be expected to fall (assuming each year is independent of the others). (Figure 3.4) In this case, with 100 cells in the output sheet, select the 5th and 95th cells as interval bounds. (This is not, strictly speaking, a confidence interval in that it does not involve making inference to a larger, unknown population.)

[Note that in Figure 3.4 we have added two rows for formatting, so the cells for interval bounds become A7 and A97.]

	A	B	C
1	Rain Output		
2			
3	8		
4	8.6		
5	9	5% =	9.8
6	9.6	mean =	15.062
7	9.8	95% =	19.2
8	10		
9	10		
10	10.8		
11	10.8		
12	11		

Figure 3.4: Rainfall in May

Figure 3.4 shows the interval values of 9.8 and 19.2 average rainy days in May, which enclose 90% of the results of this particular simulation.

The output suggests that the shift in the five-year running average is not evidence of a climate shift – an average of 18 days is well within the range of random fluctuation. Journalists are like everyone else in their tendency to conclude too readily that fluctuation in a process (rainfall, car theft, school graduation rate) represents a basic change in process parameters.

3.2 Confidence Interval for a Proportion

Finding a confidence interval for a proportion can be done in several ways – the proportion can be expressed as a fraction, as a percentage, or as raw data converted after resampling to either of these.

Here’s a test case. A political candidate has in hand a survey of a random sample of 100 New Hampshire voters. 55 of them favor the candidate, 45 her opponent. What are the bounds on the true percentage of the electorate that favors her? A straightforward resampling approach, following the logic in the drill program above, is this:

1. Use the Urn function to put 55 “1’s” and 45 “0’s” in column A (worksheet Votes.xls).
2. Resample 100 selections from the Urn.
3. Use COUNTIF to count the “1’s” in the resampled data (or SUM to sum the data).
4. Repeat and Score on the total cell for (say) 100 simulated trials.

This experiment will give you results like those in Figure 3.5 (the output has been formatted a bit). The 5th and 95th percentiles have been calculated. Right away, you should see the reason candidates take larger polling samples than 100 voters at a time, since a 55% favorable rating does not reasonably exclude losing!

How much more accurate would a larger poll be?

Repeat the experiment, taking resamples of size 1000 instead of 100, for 100 simulated trials, and you’ll get results resembling those in Figure 3.6. If you try 1000 trials, or better yet 10,000 trials, you’ll get a more nearly symmetrical confidence interval with 55% (550) at the interval center.

3. CONFIDENCE INTERVALS

	A	B	C
1	Poll Results		
2			
3	45	5% =	48
4	45	mean =	55.33
5	46	95% =	63
6	47		
7	48		
8	48		
9	48		

Figure 3.5: Results from a Small Poll

	A	B	C
1	Poll Results, N = 1000		
2			
3	519		
4	520		
5	520	5% =	522
6	521	mean =	550.24
7	522	95% =	581
8	524		
9	525		

Figure 3.6: Larger Poll Results

- Note:** Be sure to distinguish among sample size (the number of values in your original data set), resample size (the number of values you put in each resample or shuffle), and the number of iterations (also called simulations, or simulated trials). Confusion among these elements is perhaps the most common source of error in resampling. Generally, the resample size should match the sample size, and the number of iterations should be as large as practical, to achieve the most accurate result.

The problems above all have one feature in common: the techniques of classical statistics produce acceptable answers (as long as you remember the formula, for example, for the confidence interval of a proportion). Therefore, let's consider a few confidence interval problems where the conventional formula approach is not so straightforward.

3.3 Confidence Intervals for Medians

Every textbook has, of course, formulas for confidence intervals for proportions. Confidence intervals for the median, much less fancier constructs, can't be obtained by simple formulas. Therefore, a few examples using medians are useful for demonstrating resampling's extension into these areas.

Income data (in \$1000) for 100 sample families are tabulated in the worksheet `Income.xls`. The median income is \$25,500. We want to find a confidence interval for the median income of this survey group.

Again, if we had time and resources, we would go out and sample additional families to see how different one sample of 100 might be from another. Lacking time and resources, we will instead let our observed sample stand in as a proxy for the population that it came from and take bootstrap samples from it in the manner of the drill problem, above.

The procedure is simple:

1. Resample 100 values from the set of 100 incomes.
2. Use Excel's MEDIAN function [=MEDIAN(A1:A100)] to find the median of the resampled 100.
3. Use the median cell to Repeat and Score, say, 1000 simulated trials.
4. Sort the Results sheet data and find the 5th and 95th percentiles of the distribution (an estimated 90% confidence interval).

	A	B	C	D
1	20.7		C. I. For	
2	21.1		Median Income	
3	21.4			
4	21.4		5% =	23.75
5	21.5		mean =	25.58
6	21.6		95% =	27.3
7	21.65			
8	21.7			
9	21.7			
10	21.75			
11	21.8			

Figure 3.7: Median Income

3. CONFIDENCE INTERVALS

	A	B	C	D
1	21.3		C. I. For	
2	21.3		Median Income	
3	21.4			
4	21.4		5% =	23.6
5	21.4		mean =	25.54
6	21.5		95% =	27.2
7	21.5			
8	21.5			
9	21.5			

Figure 3.8: Median Income Revisited (Millionaire Version)

A typical result is shown in Figure 3.7. It doesn't seem particularly remarkable, but medians have a property that distinguishes them from analogous calculations with means: they resist outliers. If we change the highest income in the set of 100 from \$57,000 to \$1,000,000 and run the same calculation, we get approximately the same results (see Figure 3.8).

Confidence Interval for Median Price Elasticity

The responsiveness of demand to a price change (the “price elasticity”) has been measured for cigarette price changes in various states at various dates (computed from cigarette sales data preceding and following a tax change in a state) and is shown in table 3.1 (Lyon and Simon, 1958):

Price elasticity is defined as $\frac{\% \text{ Change in Demand}}{\% \text{ Change in Price}}$. The observed median is -0.511.

The curious point here is the presence of positive observations. This implies an increase in demand when the price rises, which runs against all theory. But the positive values might be considered to be the result simply of measurement errors, or of external factors, and treated as they stand. We can thus resample to estimate a confidence interval for the median price elasticity.

One procedure is:

1. Record the data in column A (worksheet Cigs.xls).
2. Resample 73 at a time to column B.
3. Take the median as the Repeat and Score cell.

3.4. Confidence Interval for Profit

1.725	-0.142	-0.377	-0.713	-1.066
1.139	-0.174	-0.383	-0.724	-1.118
0.957	-0.234	-0.385	-0.734	-1.145
0.863	-0.24	-0.393	-0.749	-1.146
0.802	-0.251	-0.444	-0.752	-1.157
0.517	-0.277	-0.482	-0.753	-1.282
0.407	-0.301	-0.511	-0.766	-1.139
0.304	-0.302	-0.538	-0.805	-1.42
0.204	-0.302	-0.541	-0.886	-1.443
0.125	-0.307	-0.549	-0.926	-1.478
0.122	-0.328	-0.554	-0.971	-2.041
0.106	-0.329	-0.6	-0.972	-2.092
0.031	-0.346	-0.613	-0.975	-7.1
-0.032	-0.357	-0.644	-1.018	
-0.100	-0.376	-0.692	-1.024	

Table 3.1: Price Elasticity

4. Try 1000 trials for a first study.
5. Sort the output, and find the 2.5th and 97.5th percentiles to determine the interval that bounds 95% of all the resampled medians.

The sample results (Figure 3.9) show the estimated resampling confidence interval for the median price elasticity, using Excel's PERCENTILE function to find the 2.5th and 97.5th percentiles.

3.4 Confidence Interval for Profit

This problem uses the bootstrap to derive a confidence interval around estimated profit.

A magazine company is planning a massive direct mail campaign to win customers and decides to test its offer out on a more limited mailing to 10,000 potential customers before mailing to millions.

The offer consists of a subscription to the magazine; as an enticement the firm is conducting a sweepstake drawing. The results of the mailing are shown in Table 3.2.

3. CONFIDENCE INTERVALS

	A	B	C	D
1	-0.753		C.I. For	
2	-0.752		Price Elasticity	
3	-0.752			
4	-0.749			
5	-0.749			
6	-0.734		2.5% =	-0.692
7	-0.734		mean =	-0.49685
8	-0.724		97.5% =	-0.357
9	-0.724			
10	-0.724			

Figure 3.9: Price Elasticity Confidence Interval

Action	n	proportion	profit rate	profit
No	5000	0.5	-0.41	-2050
Silent	4700	0.47	-0.4	-1880
Order/return	90	0.009	-8.5	-765
Order/bd	30	0.003	-9.5	-285
Order/pay	180	0.018	45	-1880
Profit				\$3120

Table 3.2: Confidence Interval for Profit

5000 people enter the sweepstakes, but decline the offer of the subscription (“no”). These cost \$0.41 apiece (mostly costs of the outbound mailing). 4700 people do not respond (“silent”), costing \$0.40 apiece (outbound mailing cost). 90 people order the magazine but then return it, costing \$8.50 apiece (shipping, processing). 30 people order and never pay for the subscription (bad debt), costing \$9.50 apiece (shipping, processing, collections). Finally, 180 people make it all worth while by ordering and paying, bringing in a value of \$45 apiece (the net present value of their current and expected future subscriptions).

The profit from the test run is \$3120, or \$0.312 per item mailed.

How reliable is this estimate? Put another way, how much might it differ in additional tests, if we could afford to run them? We can use the bootstrap to estimate a confidence interval around this estimated profit.

1. Put 10,000 slips of paper in an urn, 5000 marked “-\$0.41,” 4700 marked “-\$0.40,” 90 marked “-\$8.50,” 30 marked “-\$9.50,” and 180 marked “\$45.00.”
2. Draw 10,000 slips of paper, randomly and with replacement each time.
3. Record the sum of the values of the 10,000.
4. Repeat steps 2 and 3 many times.
5. Find the 5th and 95th percentiles, to estimate a 90% confidence interval.

With Resampling Stats in Excel, you may use the URN function (dialog box option) for step 1 to save time; (in the worksheet Sweeps.xls the 10,000 values are in A14:A10013 to allow room for more information at the top). Resample these data into the B column, then sum the B column into D1. This is the resampled profit, and is the Repeat and Score cell. 1,000 trials took approximately 39 seconds on an older laptop computer, so please be patient!

By using the PERCENTILE function with the results (which are in the Results sheet in cells A1:A1000), we can find the 5th percentile and the 95th percentile, which bound a 90% confidence interval. Remember that in the PERCENTILE function the 5th percentile is specified as .05, and the 95th as .95.

In Figure 3.10 we see that the score cell data recorded in the Results sheet has been sorted. This is not strictly necessary when using Excel’s PERCENTILE function, however in this case, notice that we are comparing Excel’s PERCENTILE function with the resampling percentile values in cells A50 and A950. Note that the two methods are in close agreement!

We therefore estimate that a 90% confidence interval for profit per 10,000 pieces mailed runs from about \$2075 to \$4164. Results of a histogram are displayed in Figure 3.11.

3.5 Planning Inventory

The following problem is not a strict confidence interval problem; rather it is a “what-if” simulation that uses the bootstrap as a key component of the model.

3. CONFIDENCE INTERVALS

	A	B	C
1	\$1,252.83		
2	\$1,329.56		
3	\$1,437.42	PERCENTILE Function	
4	\$1,453.82	5th percentile:	\$2,076.31
5	\$1,517.72	95th percentile:	\$4,164.29
6	\$1,568.64		
7	\$1,580.64		
8	\$1,623.94	Resampling Percentiles	
9	\$1,662.42	5th percentile:	\$2,075.38
10	\$1,694.07	95th percentile:	\$4,164.10
11	\$1,730.20		
12	\$1,762.52		

Figure 3.10: Direct Mail Problem: 1000 Trials

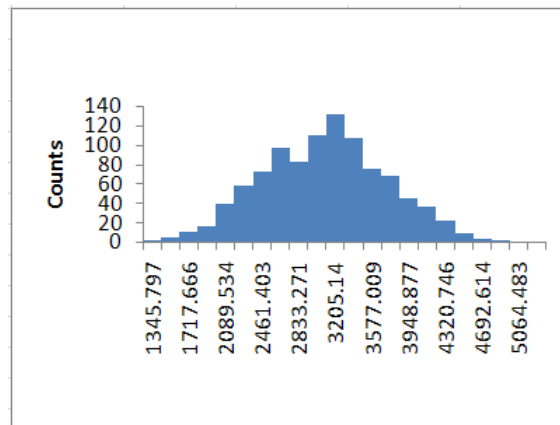


Figure 3.11: Histogram of Results for Direct Mail Problem

A company wants to determine an initial-stock and weekly reorder policy for a particular item. Table 3.3 shows daily unit sales of the item for a 15-day test period.

121	99	87	111	99	99	89	78	113
120	115	87	90	123	86			

Table 3.3: Unit Sales for 15 Days

How can you use this data to estimate future sales? Can you determine a

way to prevent a stockout, given past sales history? More specifically, can we estimate the probability of a stockout in any given week?

For a simple model, think of tomorrow's sales as having 2 components:

$$\text{Sales tomorrow} = \text{forecast level} + \text{random element}$$

The random component, of course, could be positive or negative. Let's use a very simple forecast model for tomorrow's sales – the average of sales in the recent past. (We could also use a regression slope, or we could adjust our estimate based on seasonality or other cyclicity. The estimate would then be more complex, but it would not affect the structure of the resampling procedure.)

How will we determine the random component? Our best guess is simply to look at the random components of the recent past. Unless we have some special knowledge of what luck will bring us tomorrow (and if we did, we wouldn't need to run this simulation), the best predictor of tomorrow's luck is yesterday's luck. Table 3.4 shows how the last 15 days' demand have departed from average (from the worksheet Inventory.xls).

Day	Sales	Average	Difference
1	121	101	20
2	99	101	-2
3	87	101	-14
4	111	101	10
5	99	101	-2
6	99	101	-2
7	89	101	-12
8	78	101	-23
9	113	101	12
10	120	101	19
11	115	101	14
12	87	101	-14
13	90	101	-11
14	123	101	22
15	86	101	-15

Table 3.4: Deviation from the Average over 15 Days

We will use the differences in the right hand column and resample them in groups of seven (we are interested in a week's worth of sales) (Figure 3.12).

3. CONFIDENCE INTERVALS

Then we will tack one of them on to our forecast demand for the week, representing the random component. In doing so, we are saying that the “randomness of the past is our best guess as to the randomness of the future.”

	A	B	C	D	E	F	G	H	I
1	Inventory								
2									
3	Sales	average	diff	(as values*)					
4	121	101.1	19.9	19.9					
5	99	101.1	-2.1	-2.1		A week's			
6	87	101.1	-14.1	-14.1		resampled			
7	111	101.1	9.9	9.9		differences:	-2.1333		
8	99	101.1	-2.1	-2.1			-11.133		
9	99	101.1	-2.1	-2.1			9.86667		
10	89	101.1	-12.1	-12.1			-15.133		
11	78	101.1	-23.1	-23.1			18.8667		
12	113	101.1	11.9	11.9			-12.133		
13	120	101.1	18.9	18.9			9.86667		
14	115	101.1	13.9	13.9		Week's sum	-1.9333	< repeat & score	
15	87	101.1	-14.1	-14.1		Forecast			
16	90	101.1	-11.1	-11.1		for week:	707.7		
17	123	101.1	21.9	21.9		TOTAL			
18	86	101.1	-15.1	-15.1		SALES	705.767		

Figure 3.12: A Sales-per-Week Model

Inventory on hand is a function of three things – how much you started with, how much you add through restocking, and how much is subtracted by sales (demand). Company choices (policies) determine the first two; the last is what we will model via random selections from the recorded sales fluctuations, added to the forecast level. First let’s make up a run of 100 simulated “sales weeks” by using Repeat and Score on the cell that sums the total sales, including the forecast component. The output from this procedure (the sequence is important, so we don’t sort it!) is shown graphically in Figure 3.13 (created using Excel’s Chart Wizard).

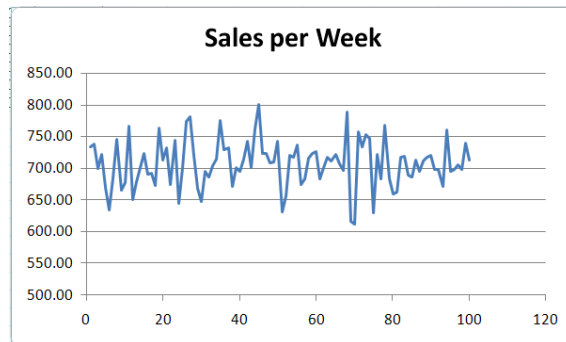


Figure 3.13: A Sales-per-Week Model

Having estimated weekly sales, including both a forecast and random com-

ponent, our task now is to model the change in inventory over time:

1. Define a starting weekly inventory cell.
2. Define “first day of the week” reorder cell (i.e. the amount to be ordered on the first of the week).
3. Subtract sales-per-week from these two cells.
4. Carry this difference forward as the starting inventory of the next week.
5. Repeat this procedure down the column to record 100 weeks’ starting inventory.

This simple setup is shown in Figure 3.14. (The value at the bottom of the visible part of column D – 310.33 – is arrived at by adding the reorder quantity (700) to the previous end of week inventory (308.10) then subtracting this week’s sales (697.77).) The point, of course, is to decide how to juggle starting inventory and weekly reorder quantity so that you just avoid running out of stock. In Excel, this means that you try different values for starting inventory and for reorder, and scan down the column (column D in the Results sheet of Inventory.xls) looking for negative values (see Figure 3.15 – they really stand out if you format the numbers in the column so that negatives are red). The proportion of negative values is an estimate of the probability of a stockout.

So far, we have only estimated the probability of a stockout. A full business analysis would incorporate the costs of a stockout, and balance these costs against the costs of carrying inventory.

3. CONFIDENCE INTERVALS

fx =D2-C3+\$E\$2			
B	C	D	E
Average	Sales per	start	reorder
	week	400	700
707.7	787.77	312.23	
707.7	688.77	323.47	
707.7	699.77	323.70	
707.7	704.77	318.93	
707.7	707.77	311.17	
707.7	717.77	293.40	
707.7	675.77	317.63	
707.7	728.77	288.87	
707.7	680.77	308.10	
707.7	697.77	310.33	

Figure 3.14: Cell Formula for Inventory

C	D	E
Sales per	start	reorder
week	400	700
787.77	312.23	
688.77	323.47	
699.77	323.70	
704.77	318.93	
707.77	311.17	
717.77	293.40	
730.7667	52.23	
690.7667	61.47	
729.7667	31.70	
741.7667	(10.07)	<-OUT!
704.7667	(14.83)	<-OUT!
755.7667	(70.60)	<-OUT!

Figure 3.15: Looking for Stockout

Chapter 4

Hypothesis Testing

Hypothesis testing is one of the main applications of statistics in actual practice. Besides freedom from assumptions about the data being examined, and the ability to produce the sampling distribution of virtually any statistic, resampling has the additional advantage of extreme conceptual simplicity. As you survey the examples in this section, you will almost certainly notice that resampling arrives at answers using the same straightforward procedure in every case.

4.1 Resampling and p-values

A few examples should show the simplicity and consistency of resampling procedures in hypothesis testing.

Zapping Fruitflies

In a biology experiment, fruitflies are irradiated to test whether or not the irradiation increases the ratio of males to females among the offspring. Of 20 offspring, 14 are male and 6 are female. Is this a statistically significant departure from 50/50?

Statistic of Interest

It is important to identify that statistic which measures what you are interested in. In this case it is the number of “1’s,” where “1” is arbitrarily set to represent a male.

Putting the test in familiar textbook terms, the “null hypothesis” is that irradiation has no effect. We ask how often a result as extreme as the observed result might occur just by chance if the null hypothesis is true.

We can test this directly by creating a hypothetical population that embodies the null hypothesis (here, a population comprised of 50% males and 50% females) and repeatedly drawing samples of 20 fruitflies from it. The steps are as follows:

1. Generate 20 “0’s” and “1’s” randomly
2. Record the number of “1’s” (males)
3. Repeat steps 1 and 2 many times
4. Determine how often you get 14 or more “1’s”

Figure 4.1 shows this example as resampling in Excel. If we let “1” stand for male and “0” stand for female, we can resample “0’s” and “1’s” repeatedly to produce 20 “offspring.” Totaling the number of “1’s” in the column of 20 resampled values gives the number of males in that set of 20. Then, to estimate the probability of 14 or more males, simply Repeat and Score this experiment a large number of times. We do a one-sided test here counting resamples with 14 or more males, but not 14 or more females because we are investigating only whether the treatment increases the male to female ratio.

In Resampling Stats:

1. Enter a “1” (male) in cell A5, a “0” (female) in cell A6, then highlight this range (A5:A6).
2. Select “R” for resample, note that A5:A6 is already entered as the input range.
3. Enter B4 as the top left cell of the output range (or just click on cell B4).

	A	B	C	D
1	Fruit Fly Irradiation			
2				
3	<i>choices</i>	<i>20 selections</i>		
4	1	0		
5	0	0		
6		0		
7		0		
8		1		
9		1		
10		1		
11		1		
12		1		
13		0		
14		0		
15		0		
16		0		
17		1		
18		1		
19		1		
20		0		
21		1		
22		0		
23		1	10	number of males

Figure 4.1: Fruit Flies and Gender

4. Enter “20” as the number of cells in the output range, click “OK”.
5. Using the Excel SUM formula, sum these resampled values in (say) C23.
6. With the cursor in this cell (C23), select “Repeat and Score.”
7. Set the number of iterations to (say) 1000, leave other settings at their defaults, and click “OK”.

The Repeat and Score command, as usual, reports its results to the Results sheet. You could simply sort the output (use “Sort” in the “Data” menu, and click the button for descending, “Z to A”) and locate the bottom of the “14’s,” which is at row 63 in Figure 4.2.

From this you can see directly that 63 out of 1000 trials were as extreme as (or more extreme than) the observed result. This is the estimated “probability value,” or “p-value.” You can also use COUNTIF to tally the output sheet, without sorting it. In this sample 1000-trial run, we estimate the number of

4. HYPOTHESIS TESTING

55	14
56	14
57	14
58	14
59	14
60	14
61	14
62	14
63	14
64	13
65	13

Figure 4.2: Fruit Flies and Gender: Results Sheet Sorted Output

times 14 or more males happen in a set of 20 flies by applying in an empty cell `=COUNTIF(A1:A1000, ">=14")`. You can enter this formula directly, or construct it by selecting “COUNTIF” from Excel’s Insert Function menu and filling in the blanks.

You can also use the FREQUENCY function on the output sheet to make a frequency distribution (Figure 4.3) that changes with each new round of output; see FREQUENCY in the section on Resampling Stats Operations (“Useful Excel Functions”). Figure 4.3 illustrates a second simulation of 1000 trials.

	A	B	C	D
1	6	Bins	Frequency	>=14 Total
2	8	20	0	60
3	10	19	0	
4	11	18	0	
5	7	17	1	
6	15	16	6	
7	7	15	16	
8	9	14	37	
9	6	13	79	
10	11	12	119	
11	5	11	163	
12	10	10	164	
13	9	9	143	

Figure 4.3: A Histogram Table “On the Fly”

Conclusion

A result as extreme as the observed result (14 males), or a result more extreme, occurred 6.0% of the time in Figure 4.3 (a p-value of .060). While unusual, this does not quite attain the traditional 5% benchmark required in order to be called statistically significant.

Clinical Trial: Cancer (a Permutation Test)

As another example of the power of simple “0 and 1” binomial models, here’s a hypothetical drug experiment. Note that the set-up is a bit more sophisticated than the Flies example, which could have been analyzed with Excel’s BINOMDIST function as well as resampling.

A new medicine produced 5 cures in 6 patients, while a placebo group showed cures in only 2 in 6.

Statistic of Interest

The statistic of interest here could be the difference in the number cured – 3. (There are other possibilities – the difference in cure rates, for example.)

What is the probability that, if the medicine is ineffective, a difference this big or bigger might occur by chance? We will test by combining all results together (7 cures and 5 no-cures; this embodies the null hypothesis of no difference between treatment and placebo), shuffling, and drawing out two resamples of size six each. If this rarely produces a difference between the first resample and second resample as big as the observed difference, we can say that the observed difference is not likely due to chance.

To obtain the answer, record the data as two columns of six elements each, with five “1’s” in the first and two “1’s” in the second as shown in Figure 4.4 (from worksheet Cancer.xls).

Then,

1. Shuffle (Normal Shuffle) the contents of the columns into two new columns, representing outcomes drawn at random from this small set of possibilities (note: although shuffling preserves the 2-column format for the shuffled output, in the Normal Shuffle all data are combined for shuffling purposes and are not confined to their respective columns).

4. HYPOTHESIS TESTING

	A	B	C	D	E
1	Cancer Drug Test				
2					
3	drug	placebo	<i>[randomized results]</i>		
4	1	1	1	1	
5	1	1	0	1	
6	1	0	0	0	
7	1	0	1	1	
8	1	0	1	0	
9	0	0	0	1	
10			3	4	
11					
12			difference	-1	

Figure 4.4: A Cure/Not-Cure Test

2. Sum the “1’s” (cures) in each column.
 3. Take the difference between these column sums (C – D).
 4. Repeat and Score on the difference cell.
- **Note:** For a discussion of resampling with replacement (bootstrap-style) vs. without replacement (permutation style), see the end of section 7-3.

Results for a small first test (100 trials) are shown on the sorted Results sheet in Figure 4.5. How common is it to find the observed cure rate difference in a sample this small? In this test a difference of 3 or more “excess” cures in the column C group compared to the column D group happened 18 times in 100 trials, so the first indication is that this result would not be considered significant at the usual clinical significance value of $p < 0.05$. In practice, you would want to repeat this experiment for 1,000 and for 10,000 trials (or more).

This process of shuffling the data and, for each shuffle, recalculating the test statistic, is known as a permutation test.

Clinical Trial: Birthweight of Babies

The format of the previous example can be taken as a departure point for any sort of two-sample comparison.

	A	B
1	Drug Test Output	
2		5
3		3
4		3
5		3
6		3
7		3
8		3
9		3
10		3
11		3
12		3
13		3
14		3
15		3
16		3
17		3
18		3
19		3
20		1
21		1
22		1
23		1
24		1

Figure 4.5: Checking Cancer Cure for Significance

An experimental diet is developed for expectant mothers (hypothetical example drawn from Rosner, p. 257). In the test (results shown on the left side of Figure 4.6) the babies born to mothers placed on the experimental diet show a higher average weight than babies born to mothers in the control group. The question is: for this sample size, is the weight gain significant? The null hypothesis is that there's no difference, and significance means “p-value less than 0.05.”

Statistic of Interest

Here the statistic of interest is the average weight difference between the two groups 0.62 pounds (7.01-6.39) shown in Figure 4.6.

The average weight difference would be significant if it turns out that a

4. HYPOTHESIS TESTING

	A	B	C	D	E	F
1	Diet and Birthweight					
2						
3	Diet	Control		<i>[randomized results]</i>		
4	6.9	6.4		7.6	6.8	
5	7.6	6.7		8	6.9	
6	7.3	5.4		6.7	6.9	
7	7.6	8.2		5.4	6.6	
8	6.8	5.3		4.2	6.6	
9	7.2	6.6		5.3	7.1	
10	8	5.8		6.8	6.2	
11	5.5	5.7		5.8	8.6	
12	5.8	6.2		6.8	8.2	
13	7.3	7.1		5.7	7.3	
14	8.2	8		5.5	4.7	
15	6.8	6.9		7.2	7.6	
16	6.8	6.6		7.3	6.4	
17	4.7	4.2		8.2	5.8	
18	8.6	6.8		6.8	8	
19	7.01	6.39	0.62	6.49	6.91	-0.43
20	average	average	diff			diff

Figure 4.6: A Classic Hypothesis Test

difference this large is rarely seen when the results are randomly shuffled into two groups. So the values in the experimental results columns in Figure 4.6 are shuffled together and re-distributed into two columns (resampling without replacement). Then Excel is used to compute the average weight difference between the two shuffled sets (workbook Babies.xls).

Step by step:

1. Record the diet treatment birthweights in column A, the control group in column B.
2. Highlight these data (A4:B18) and select “Shuffle” from the Resampling Stats menu or toolbar.
3. The input range should already be filled in, enter D4 as the “Top Left Cell of Output Range” and make sure that Normal Shuffle is selected, click “OK”.
4. Enter or copy the formulas required to calculate the mean of each shuffled column, and the difference between those means.

5. Select the difference cell, and select “Repeat and Score” from the Re-sampling Stats menu or toolbar. Set the iterations to 1000 and click “OK”.

Again, there are several ways to inspect the Results sheet, but the simplest is probably just to use COUNTIF on the output column (Figure 4.7). If you were doing this diet research yourself, first you’d do 10,000 trials, and then apply for funding to do a larger study – the results here (56 out of 1000) are perched right on the edge of the $p = 0.05$ significance borderline.

	A	B	C	D
1	-0.14667			
2	-0.37333		Count if result >= 0.62	
3	-0.05333			
4	-0.29333			56
5	-0.37333			
6	-0.69333			
7	0.253333			
8	0.026667			
9	0.04			

Figure 4.7: Too Close to Call

4.2 Testing for a Difference in Variability

In scientific instrument design, there’s constant effort to design new measuring devices that reduce measurement variability.

In the worksheet Measure.xls (data adapted from Hirsch, *ASQC Stat. Div. Newsletter*, Spr. 91), measurements of a concentration of a particular chemical in a bath are listed in column A. The measurements are made over a short period of time in which the chemical concentration being measured remains unchanged. At the same time, measurements are also being made by a new prototype device, whose output is listed in column B. We measure each instrument’s variability by the standard deviation of the measurements it produces. Is the measurement variability the same? Our null hypothesis is that the measurements produced by systems A and B belong to the same population of measurements, and that the difference between A and B is due to chance. We test this by combining the A and B measurements together (this is our

4. HYPOTHESIS TESTING

best guess of what the null population would be), then drawing two resamples from this common “population.”

	A	B	C	D	E
1	System System				
2	A	B		<i>[shuffled columns]</i>	
3	137.9	141.6		138.6	143.2
4	143	138.9		141.3	141.4
5	143.2	140		141.5	139.5
6	140	141.9		140.1	140
7	140.2	140.5		140.5	143
8	139.3	138.6		140.7	142
9	141.4	141.5		142	138.9
10	140.1	141.5		137.2	141.5
11	142	140.7		141.6	141.5
12	137.2	141.5		137.9	140.4
13	139.5	140.4		142.7	140
14	142.7	142		141.9	141
15	141.3	141		140.2	139.3
16	1.8828	1.0856		1.6694	1.3626
17	<i>(StDev A)</i>	<i>(StDev B)</i>			
18	diff =	0.7972		diff =	0.3068
19					

Figure 4.8: Instrument Design

Statistic of Interest

The statistic of interest, as shown in Figure 4.8, is the difference in standard deviations between A and B.

1. Shuffle the data into two columns (when you select the range containing the two columns, the Resampling Stats add-in will automatically use the Matrix Shuffle option to distribute values into two columns also; you should select “Normal Shuffle” in which a value being shuffled might end up in either column, any row).
2. Find the difference between the standard deviations for the shuffled columns (use STDEV to calculate them) and select the difference cell as the Repeat and Score cell.
3. For a first look, try 1000 trials.

The results in Figure 4.9 show that a difference as large as the observed difference in Figure 4.8 happens rarely (27 counts per 1000 trials in this simulation) in chance draws. Therefore, we conclude that the difference between the two devices is significant.

	A	B	C	D
1	0.813243	Countif result >=0.7972		
2	-0.03636	27 <- Counts per 1000		
3	0.400271			
4	0.155052			
5	-0.68388			
6	-0.30613			
7	-0.30052			
8	-0.11981			

Figure 4.9: Differences in Standard Deviation, Shuffled Pairs of Samples

4.3 Resampling in Complex Cases

The cases above (and most examples you might find in standard textbook) can be studied by resampling or shuffling small arrays of “1’s” and “0’s,” or shuffling small arrays of data. The next cases need just slightly fancier modeling, in terms of setting up the initial array to resample or shuffle.

Molecular Biology

Here’s a research example of hypothesis testing, drawn from Karlin & Brendel, 1992.

Molecular protein sequences are comprised of charged amino-acid residues of three types: basic, acidic, or mixed. A question that arises in protein analysis is whether the composition of a sub-sequence differs from its “master” with respect to the composition of charged residues.

For example, the protein called GCN4 (a yeast transcriptional activator) has 36 basic residues, 46 acidic residues, and 201 mixed residues.

One particular 46-residue section of GCN4, called the COOH-terminal DNA-binding domain, contains 15 basic and 7 acidic residues, and 24 mixed.

4. HYPOTHESIS TESTING

From the above data (shown in Table 4.1) on the amino-acid composition of GCN4, a randomly selected sequence of 46 GCN4 residues would be expected to have 5.9 basic, 7.5 acidic, and 32.7 mixed residues. Is the departure from expectation more than might readily occur by chance?

	GCN4	COOH-term
Basic	36	15
Acidic	46	7
Mixed	201	24
Total	283	46

Table 4.1: Molecular Protein Sequences

Statistic of Interest

We will measure departure by summing the absolute values of the observed values less the expected values, where “expected” is what we would get if the 46 were drawn perfectly proportionally from the 283.

COOH-term	Observed	Expected	Absolute Difference
Basic	15.0	5.9	9.1
Acidic	7.0	7.5	0.5
Mixed	24.0	32.7	8.7
Total	46.0	46.0	18.3

Table 4.2: Residues in COOH-term: Observed vs. Expected

The observed value of this statistic is 18.3, as shown in Table 4.2 and in Figure 4.10; the Excel summary of these tables is in workbook Protein.xls.

We can simulate random composition of the COOH-terminal segment of the protein, drawing randomly from the amino acid set of GCN4, in these steps:

1. Constitute an urn with 36 “1’s,” 46 “2’s,” and 201 “3’s” representing 283 amino-acids, of three types (use the Create Urn via Dialog Box tool in Resampling Stats, see Figure 4.11). In the worksheet Protein.xls, this Urn function is invoked from cell H4, so the 283 “1’s,” “2’s,” and “3’s” proceed down from there.

4.3. Resampling in Complex Cases

	A	B	C	D	E	F	G
1	Protein Sequence Analysis						
2							
3		GCN4	COOH-		observed	found	shuffled
4		total	terminal	expected	difference *	in shuffle	difference **
5	Basic	36	15	5.85	9.15	12	6.15
6	Acidic	46	7	7.48	0.48	3	4.48
7	Mixed	201	24	32.67	8.67	31	1.67
8	Total	283	46				
9							
10				sum	18.30	score	12.30
11							
12	*Absolute difference between the observed count (column C) and the expected						
13	count (column D)						
14	**Absolute difference between the shuffled count (column F) and the expected						
15	count (column D)						
16							

Figure 4.10: Setup for Protein Analysis

2. Shuffle and draw 46 (Using the Shuffle tool, specify 46 cells in the output range)
3. Count the number of “1’s,” “2’s,” and “3’s” in the 46 values (3 separate COUNTIF statements) (F5:F7).
4. Record the sum of absolute deviations between these counts and the expected numbers of “1’s,” “2’s,” and “3’s” (5.9, 7.5 and 32.7) (G10).
5. Repeat steps 2-4 many times.
6. Count how often the sum of absolute deviations is greater than 18.3.

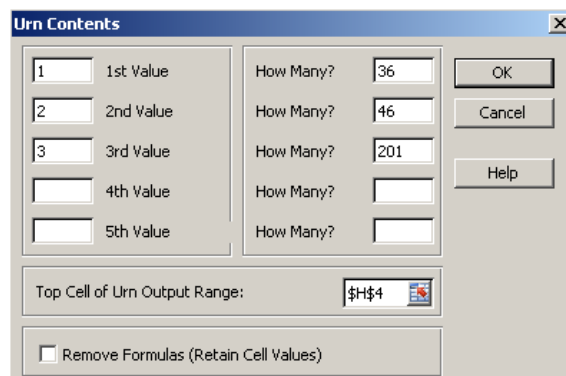


Figure 4.11: Creating an Urn Full of Amino Acids

4. HYPOTHESIS TESTING

A quick survey run of only 100 trials, sorted by the size of the difference, shows one value at 22, and all the rest less than 18.3 (Figure 4.12).

	A	B
1	17.34276	Sorted
2	14.65724	Difference
3	13.34276	Scores
4	12.65724	
5	12.29682	
6	12.29682	
7	11.04594	
8	11.04594	
9	10.65724	
10	10.29682	

Figure 4.12: Resampling Results for DNA Protein Sequence

That’s an indication that the amino-acid composition of the COOH-terminal segment is very non-random, and that runs of 1000 and 10,000 trials are an obvious next step for estimating the p-value more precisely.

Age Discrimination in Employment

This problem illustrates a case in which a special “home-grown” statistic is needed (one for which no tabulated distribution has been established).

XYZ company has been accused of firing workers (it has 50) when they get close to the level of seniority at which their pension would be vested (25 years). The union notes that the levels of seniority of 7 fired workers in the last 12 months were unusually close to 25 years. Four of the seven were within 6 years of vesting, and three within 2 years of vesting. (The worker with 25 years seniority has already been vested.) Table 4.3 displays the ages of the workers who were fired.

23	19	24	23	25	2	5
----	----	----	----	----	---	---

Table 4.3: Seniority of discharged workers (years)

This age data also exists in column B, and in bold italic in column A of the workbook Firing.xls. The seniority of all workers is displayed in Table 4.4.

11	8	24	36	20	19	11	9	10
9	5	4	21	9	21	16	17	11
1	1	23	19	24	40	28	5	7
1	34	20	16	31	23	50	4	1
8	8	14	12	32	1	15	12	25
19	5	24	2					

Table 4.4: Seniority of all workers

The company counters that operational considerations were the only factors in each of the firings and that the proximity of the firing dates to pension vesting dates was purely coincidental, the result of random chance.

To evaluate the union’s claim, we need a measure of the degree to which firing dates cluster just below 25 years seniority. Let’s subtract from 25 the tenure of each fired, unvested employee then sum those values. The lower this sum, the more evidence there is that the firings cluster around 25.

But what about zero and negative values? These result when a fired worker has 25 or more years seniority – they all signify vested pensions. Whatever the reason for these firings, it was not to avoid pension responsibility. The more such nonpositive values there are, the greater the evidence against the union’s proposition that the firings are pension-related.

Hence, we need to incorporate non-positive values in an appropriate way. One reasonable approach is to say that all workers fired after vesting count at equal and maximum weight against the union’s contention. A new worker fired at maximum time before (i.e. least connection with) vesting gets a 25, so we will also recode all non-positive values as 25’s.

Statistic of Interest

To calculate the statistic of interest, subtract the workers’ seniorities from 25, recode non-positive values as 25, then sum.

The formula used in column D (Figure 4.13) helps calculate this statistic by recoding non-positive values as 25:

$$=IF (C11<=0, 25, C11)$$

This means “if the value in C11 is ≤ 0 , enter 25, otherwise enter the value

4. HYPOTHESIS TESTING

in C11.”

	A	B	C	D	E
1	Employee firings				
2					
3	seniority:	seniority:			
4	all	fired		recode	
5	employees	employees	(25 -)	<=0 as 25	
6	50	23	2	2	
7	40	19	6	6	
8	36	24	1	1	
9	34	23	2	2	
10	32	25	0	25	
11	31	2	23	23	
12	28	5	20	20	
13	25			79	< SUM

Figure 4.13: Employee Firing Test Statistic

For the observed data, the value of the test statistic is 79. Now we select at random 7 workers from the total 50, whose years of tenure are listed in column A. We make this same calculation for the random sample – subtract the values from 25, and recode all negative values (“pension vested”) as “25.” Then observe whether the sum is more or less than that actually observed. We repeat this procedure and find what proportion of our 1000 trials produces sums equal to or less than that observed.

The results for 1000 trials (Figure 4.14) indicate that the firing pattern is concentrated on near-seniority employees, but not at levels that are thought of as unambiguously statistically significant – the probability value is about 0.11, well above the usual cutoff of $p = 0.05$.

	A	B	C	D
1	133	Employee firings		
2	116	Count if result <=79		
3	83	111 <- counts per 1000		
4	81			
5	96			
6	108			
7	71			
8	104			

Figure 4.14: Employee Firing Test Statistic: 1000 Trials

4.4 Multiple Comparisons - Ad Clickthroughs

In this problem we will address the issue of making comparisons when more than two groups are being tested (Clickthroughs.xls).

A web site ad developer is testing four different web ads representing four different marketing concepts. The ads are placed on a web site and rotated through the same pages on the web site. They have different exposure levels because they were completed and uploaded at different times.

Data are collected to see how many times web viewers click on the advertisements, and which ad has the best clickthrough rate. The results over a week are shown in Figure 4.15. A “click” (also called a clickthrough) represents a viewing of the ad (an impression, in web terminology) in which the ad is clicked on by the user. “No click” represents a viewing in which the ad is not clicked on. The click rate is the percent of impressions resulting in a click. The “vs. others” is the difference between an ad’s click rate, and the average click rate on the other ads. (This average is a weighted average; in other words, total clicks for the other ads together, divided by total clicks and non-clicks for the other ads together.)

	A	B	C	D	E	F
1	OBSERVED					
2		Ad #1	Ad #2	Ad #3	Ad #4	Overall
3	Clicks	3	8	7	6	24
4	No-clicks	234	760	291	569	1854
5	Click-rate	1.27%	1.04%	2.35%	1.04%	1.28%
6	Vs. others	-0.03%	-0.42%	1.26%	-0.36%	
7						

Figure 4.15: Employee Firing Test Statistic: 1000 Trials

Ad #3 appears to do the best, with about double the clickthrough rates of the other ads. Should the agency proceed with the concept behind #3 and implement it more widely?

Considerable effort has gone into the development of the other concepts, and the firm wants to be sure the evidence in favor of #3 is conclusive before it discards the other concepts. It will also be easier to tell the designers of ads 1, 2 and 4 that their concepts will not be pursued if the evidence in favor of #3 is strong.

The firm would therefore like to be reasonably certain that the favorable results of #3 are not due to chance.

The Problem of Multiple Comparisons

The problem of multiple comparisons can be stated simply: the more experiments you do, or the more you undertake separate examinations of the data looking for different things, the more likely it is that you will find something by chance.

If you are told that Mr. Smith can toss a coin 10 times and get at least 9 heads, and he then tosses a coin 10 times and gets 9 heads, you find the result sufficiently surprising to provide good evidence of his claimed ability.

If, on the other hand, you are told that Mr. Smith was one of 30 people in a room, each of whom tossed a coin 10 times, and that he was singled out as the most successful, his success is less surprising. In fact, the chances are better than 1 in 4 that at least one person among the 30 will do that well just by chance.

The question at issue is therefore “If all four ads are equally likely to be clicked on, what is the chance that one of them will do this much better than average just by chance?”

Statistic of Interest

One can imagine several statistics of interest; one could be the difference between the best clickthrough rate, and the clickthrough rate for the other ads.

The null model is therefore that the 24 observed clicks were randomly distributed among the ads, and that ad #3 had a better rate just by chance. We can test whether this is reasonable to believe as follows:

1. Create an urn with 24 clicks and 1854 no-clicks.
2. Shuffle the urn and take four samples without replacement of 237, 768, 298 and 575.
3. Count the number of clicks for each sample, determine the click rates, and find the difference between each sample’s rate and the rate for the rest of the samples.
4. Sort those differences, and record the largest difference.
5. Repeat 2-4 many times.

4.4. Multiple Comparisons - Ad Clickthroughs

- Find out how often this randomly-produced “largest difference” equals or exceeds the observed largest difference of 1.26%.

In Clickthroughs.xls multiple urns have been created, one for each ad as shown in Figure 4.16.

	A	B	C	D	E	F
1	OBSERVED					
2		Ad #1	Ad #2	Ad #3	Ad #4	Overall
3	Clicks	3	8	7	6	24
4	No-clicks	234	760	291	569	1854
5	Click-rate	1.27%	1.04%	2.35%	1.04%	1.28%
6	Vs. others	-0.03%	-0.42%	1.26%	-0.36%	
7						
8						
9	Urns:	1	1	1	1	
10		1	1	1	1	
11		1	1	1	1	
12		0	1	1	1	
13		0	1	1	1	
14		0	1	1	1	
15		0	1	1	0	
16		0	1	0	0	
17		0	0	0	0	
18		0	0	0	0	
19		0	0	0	0	
20		0	0	0	0	
21		0	0	0	0	
22		0	0	0	0	
23		0	0	0	0	
24		0	0	0	0	
25		0	0	0	0	

Figure 4.16: Clickthrough Ads Multiple Urns

Taken together (the range B9:E776) they constitute the single urn of step 1 above.

The next step is to shuffle the urns to an adjacent range (H9:K776). If we choose “Normal Shuffle” then the blank cells are ignored (i.e. considered fixed and not part of the shuffling). Non-blank cell contents are shuffled across all ads then “re-dealt” in the same configuration (i.e. the same sample sizes), which is step 2 above. The same formulas used to calculate clicks and click rates for the observed data can be copied to calculate the click rates for the shuffled data (step 3 above).

In Figure 4.17, the difference between sample #3 (which is the resampled counterpart to ad #3) and the other samples is highlighted. But this is NOT the statistic we will be tracking. Recall that the statistic of interest is not the difference between sample #3 and the others, but rather the MAXIMUM difference.

4. HYPOTHESIS TESTING

H	I	J	K	L
SHUFFLED				
Ad #1	Ad #2	Ad #3	Ad #4	Overall
2	10	3	9	24
235	758	295	566	1854
0.008439	0.013021	0.010067	0.015652	1.28%
-0.51%	0.02%	-0.34%	0.40%	
0	0	0	0	
0	0	0	1	
0	0	0	0	
0	0	0	0	
0	0	0	0	

Figure 4.17: Difference in Resampled #3 and Other Ads

Using the Sort Feature in Resampling Stats

To find the maximum difference, we can use the Resampling Stats SORT feature, which can sort columns and rows and iterate the sort as part of each resample. (If you use Excel’s sort function, the sort will *not* be repeated for each resampling trial.)

First, create a column with the shuffled differences (N1:N4) as shown in Figure 4.18.

M	N
Unsorted:	-0.51%
	0.02%
	-0.34%
	0.40%

Figure 4.18: Unsorted Data

Then with that range highlighted, select “SORT” on the resampling toolbar or menu. The Resampling Stats Sort dialog shown in Figure 4.19 is displayed.

Select “Sort Columns Independently,” then select “Sort Selection to New Range” and click on the top cell of the new range you want to sort the values to (P1 in this case). Click on “Desc” for descending (Asc means ascending)

4.4. Multiple Comparisons - Ad Clickthroughs

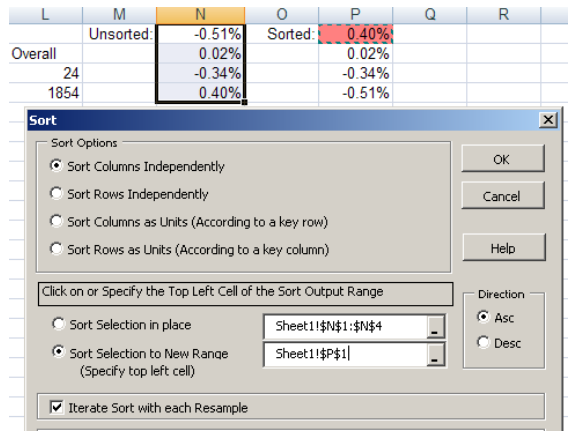


Figure 4.19: Resampling Stats Sort Dialog

and make sure “Iterate Sort with each Resample” is selected.

Next, select “Repeat & Score;” the cell to score is the top cell in the sorted range – this is the maximum difference.

For comparison purposes, Clickthroughs.xls also scores the difference between shuffled sample #3 (which is the resampled counterpart to ad #3) and the other samples. This cell is the one highlighted in Figure 4.20.

H	I	J	K	L
SHUFFLED				
Ad #1	Ad #2	Ad #3	Ad #4	Overall
2	10	3	9	24
235	758	295	566	1854
0.008439	0.013021	0.010067	0.015652	1.28%
-0.51%	0.02%	-0.34%	0.40%	
0	0	0	0	
0	0	0	1	
0	0	0	0	
0	0	0	0	

Figure 4.20: Shuffled Sample Difference

We are thus tracking the p-values for two problems simultaneously:

1. Comparing the observed difference (for ad #3) to the maximum difference; and

4. HYPOTHESIS TESTING

2. Comparing the observed difference (for ad #3) to the difference for the shuffled counterpart to ad #3.

Comparing the observed difference (for ad #3) to the maximum difference, we see in Figure 4.21 that a maximum difference (shuffled) as big as the observed difference occurred in just over 17% of the 1000 trials for an estimated p-value of .17.

165	0.012613
166	0.012613
167	0.012613
168	0.012613
169	0.012613
170	0.012613
171	0.012613
172	0.011574
173	0.011574
174	0.011574

Figure 4.21: Estimated p-value = $\frac{171}{1000}$

Conclusion

A p-value of .17 is not very conclusive evidence (by itself) that ad #3 is better. The firm would probably be better off letting more time elapse and additional evidence accumulate before making a decision.

Consider how different the conclusion would be if we instead compare the observed difference (for ad #3) to the difference for the shuffled counterpart to ad #3. A difference in shuffled sample #3 (the counterpart to ad #3) as big as the observed difference occurred in only 7.8% of the 1000 trials for an estimated p-value of .078 as shown in Figure 4.22.

The firm might well conclude that the observed result is so unlikely to happen by chance that the other ad concepts can be abandoned now, without awaiting further data.

72	0.014226	0.012613
73	0.014226	0.012613
74	0.014226	0.012613
75	0.014226	0.012613
76	0.014226	0.012613
77	0.014226	0.012613
78	0.014226	0.012613
79	0.014226	0.008611
80	0.014226	0.008611
81	0.014226	0.008611
82	0.014226	0.008611

Figure 4.22: Observed Difference Compared to Shuffled Difference

Discussion

Why is it appropriate to make our comparison to the maximum difference (which makes the observed difference seem not so unusual), rather than to the difference for shuffled sample #3 (which makes the observed difference seem more rare)?

The resampling comparison needs to mirror what happened in the real world. In the real world, we looked at the results from four ads, and selected the best one. To make a judgement about whether chance variation might be responsible, in the resampling world we must also look at the shuffled results for four ads, and select the best one.

Chapter 5

Contingency Tables

5.1 Chi-Squared Basics

In classical analysis of contingency tables, values in cells of the tables are compared to “expected” values, and a “chi-squared” (χ^2) statistic is computed by summing the squared differences between observed and expected values and dividing this sum by the expected values. Then a value for p (the probability that a departure from expected as extreme as the observed departure could occur under the null hypothesis) is found from a table, using the χ^2 value and the degrees of freedom in the problem. This classical method is built into Excel as CHITEST.

Because resampling is a general technique, it approaches calculations of probability values from two-way and multi-way tables by designing a simulation and calculating probabilities directly. This gives it an advantage in handling problems with low cell counts, where the traditional method breaks down. Resampling (in this case, “exact” or “permutation” procedures) reports correct probabilities in situations where a χ^2 test is unreliable.

5.2 Sir Ronald and the Tea Lady

Here’s a rather charming, and almost certainly true, story about the origins of exact methods.

In Britain, it’s common to take tea with milk, and customary (called

5. CONTINGENCY TABLES

“mother’s way”) to put milk in the teacup first. One afternoon, after a lady friend of the great statistician Ronald Fisher remarked that she could always tell a tea-first from a milk-first cup, Fisher decided to make a little test (see workbook Tea.xls).

Fisher set before the tea-taster 4 “tea-first” cups and 4 “milkfirst” cups, all arranged in random order, and she correctly identified six of the eight. But might she have done this well by chance?

For eight cups of tea, as Fisher proposed, it’s possible to enumerate all possible ways one could guess randomly and calculate how many of those ways would yield 6 or more correct guesses. Fisher’s Exact Test does just that for $R \times C$ (row by column) tables, with only the practical limitation that as R and C (and n) get larger, great ingenuity is required (consult www.cytel.com for the software details) to perform the actual outcomes-counting.

Fisher’s exact result for this case was $p = 0.243$, quite different from the one-sided (right-tailed, actually) χ^2 result of 0.07865 (Figure 5.1). Of course, the small number of counts-per-cell suggests immediately that a standard χ^2 calculation isn’t appropriate.

	A	B	C	D	E	F	G	H
1								
2								
3		the guess	poured first		row total			
4			milk	tea				
5		milk	3	1	4		2	2
6		tea	1	3	4		2	2
7		column total	4	4	8			
8							CHITEST	0.15730
9							one-sided	0.07865

Figure 5.1: Classic Probability and the “tea test”

The resampling worksheet for this table shows an equivalent way of modeling the experiment.

1. Shuffle the eight cups (4 milk first and 4 tea first).
2. Use an IF statement to find how many matches to the correct choice there are in each set of eight (Figure 5.2). Note that we can shuffle and use IF logic with alpha values, just as we can with numeric values.
3. For 1000 trials, use the cell with the number of matches as the Repeat and Score cell. A sample sorted output is shown in Figure 5.3.

[actual]	[guesses]		
tea	milk	0	
tea	milk	0	
tea	tea	1	
tea	milk	0	
milk	tea	0	
milk	tea	0	
milk	tea	0	
milk	milk	1	
		2	<-correct

Figure 5.2: Tea Testing: One of Many Possible Random Guesses

	A
226	6
227	6
228	6
229	6
230	6
231	6
232	6
233	6
234	6
235	6
236	4
237	4
238	4

Figure 5.3:

These results show that, in this particular sampling, random choice produced six or more correct guesses 235 times out of 1000. If you're willing to wait through 10,000 trials, you'll remain close to Fisher's exact result of $p = 0.243$. For example, on the author's laptop, 10,000 trials took 17 seconds and resulted in 6 or more correct guesses 2436 times out of 10,000.

This example also illustrates a bit of the practical flexibility of resampling. On an issue of the grave importance of "tea before milk?" one hundred trials would be plenty for proving a point. Also, the usual statistical standard for significance of $p < 0.05$ might not make for graceful conversation in the circumstances of the actual experiment and in practice Sir Ronald could hardly require the lady to taste 200 cups to settle things once and for all.

5.3 Applying Resampling

Resampling provides a simple way to analyze contingency tables, with essentially the same procedure for every table. It also gives reliable answers even for tables with cell counts that are too low to be analyzed reliably with classical methods.

Driving While Black

As a real-data example, centered on a social issue, we will now look at a contingency table drawn from an ACLU Web site on freeway stops and searches in a state in the Eastern U.S. The data sample shown in Figure 5.4 (from worksheet Black1.xls) represents a small subset of these data, in fact the records of the two most race-neutral officers studied over a small time sampling of traffic. Clearly, blacks have a greater probability of being stopped, but is this difference statistically significant?

	A	B	C	D
1	"Driving While Black?"			
2				
3		STOPPED	NOT	<i>row</i>
4	WHITE	11	72	83
5	BLACK	7	10	17
6		18	82	100
7				
8		CHITEST	0.0063302	
9				

Figure 5.4: Numbers of Drivers Stopped and Not Stopped

Statistic of Interest

There are several ways we might measure the tendency to stop black drivers more than white. Here we will simply ask how probable it is that as many as 7 of the 18 drivers stopped would be black, given that blacks comprise only 17 out of the 100 drivers passing the two officers. So the statistic of interest is “number of blacks stopped.”

To model this in resampling for comparison, try the following steps:

1. Use Urn (the dialog box option) to make a column of 100 numbers, with 17 “1’s” representing black drivers and 83 “0’s” representing white drivers (Figure 5.5).

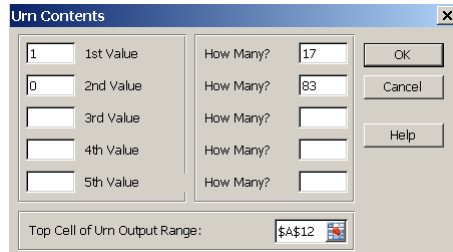


Figure 5.5: Setting Up the Race Test

2. Choose “Shuffle;” make sure the urn of 100 numbers is the input; specify 18 as the Number of Cells in Output Range, representing the drivers stopped (in the worksheet Blacks1.xls, the shuffled output starts in B12).
3. Use SUM to count the number of black drivers in the 18.
4. Repeat and Score 1000 times on the SUM cell.
5. Sort the Results sheet, and see how often 7 or more of those stopped are black.

	A	B	C
1	9		
2	9		
3	8		
4	8		
5	7		
6	7		
7	7	Output:	
8	7	Driving While Black	
9	7		
10	7		
11	7		
12	7		
13	6		
14	6		

Figure 5.6: Driving While Black: Trial Results

A typical run is shown in Figure 5.6. Only 12 of the 1000 trials yielded 7 or more “1’s,” for an estimated p-value of .012 (12/1000). We conclude that

chance is not likely to be the explanation for the larger proportion of blacks stopped.

When all the permutations are systematically counted, this test is known as Fisher's Exact Test.

Drug Response

Consider the table in Figure 5.7. Your first reaction might be “get more data.” That's entirely correct, but in this case the investigator (studying methods for dealing with obsessive-compulsive disorder) wanted to see if the earliest small results indicated promise for a larger, more expensive study.

	A	B	C	D	E
1	Evaluating therapies				
2					
3		<i>same</i>	<i>better</i>	<i>cured</i>	
4	hypnosis	2	1	0	3
5	drug	0	0	3	3
6		2	1	3	6
7					

Figure 5.7: Psychopharmacology

Statistic of Interest

The researcher decides to evaluate the outcome in terms of “scores” for each treatment, with a cure scored as 2, improvement scored as 1, and no improvement scored as zero. The comparative scores are then: drug 6, hypnosis 1 (see Hypnosis.xls).

Our null hypothesis is that both groups share the same distribution of scores and that the difference between them (6 vs. 1) is attributable to chance. Thus, the null model is a single urn with two “sames” (0's), one “better” (1), and three “cures” (2's), from which we draw (randomly and without replacement) two samples of three numbers each. Then we calculate the difference in scores, Repeat and Score on that cell, and determine how often we get a difference of 5 or more.

In Resampling Stats:

1. With the cursor at the top of an empty range (it's in A11 in Hypnosis.xls), select "Urn" from the Resampling Stats menu or toolbar, using the Create Urn Via Dialog Box option. "0" is the first value, and you want two of those. Then one "1" and three "2's."
2. Note that two "0's," one "1" and three "2's" have been entered in the range below where your cursor was (A11:A16 in Hypnosis.xls). Select this range, then select "Shuffle" from the Resampling Stats menu or toolbar. Enter the top of the output range (the range where you want the shuffled values to go; D11 in Hypnosis.xls), and 6 as Number of Cells in Output Range, then click "OK."
3. Consider the first three values in the shuffled output as the first sample (the resampling counterpart to the hypnosis group).
4. Consider the remaining three values in the shuffled output as the second sample (the resampling counterpart to the drug group) and use SUM to sum the scores (in cell E16 in Hypnosis.xls).
5. Find the difference in scores (cell E18 in Hypnosis.xls), highlight this cell, select "Repeat and Score" from the Resampling Stats toolbar or menu (choosing 1000 iterations), and click "OK".

	A	B	C	D	E	F
1		-1	Count If Results >= 5			
2		1	52			
3		-1				
4		-5				
5		3				
6		1				
7		-1				
8		-1				

Figure 5.8: Results Sheet Showing Resampled Differences in Scores

How often, in a thousand random shuffles, do we see a score difference as large as the one measured? A run of 1000 trials produced a score difference as big as the observed difference only 52 times in 1000 trials (Figure 5.8), suggesting that something significant might be producing the difference. As a curious note, the actual experiment reported here was not only repeated as an experiment on 240 human subjects, but also led to the discovery that the same drug greatly reduces odd "compulsive" behaviors in dogs!

Drug Testing

A government agency administers drug tests to 4000 prospective employees, and proceeds to hire 816 employees without regard to the drug test result. Later, the agency determines that, of 75 employees who tested positive and who had insurance coverage with a certain carrier, 4 (5.3%) made claims for drug-related health problems. By contrast, of 741 who tested negative, only 12 (1.6%) made drug related claims. Is submission of a drug related claim associated with a positive drug test result in initial screening? These data are summarized in Table 5.1:

	No Claim or Non-Drug Claim	Drug Claim
-test	729	12
+test	71	4

Table 5.1: Prospective Employee Drug Related Claims

Statistic of Interest

The statistic of interest is the number of drug claims in the group of 75 (the positive-testers).

We want to know, if there are 16 drug claims, what the probability is that 4 or more of them would fall in the “+ test” group. The basic setup for this problem (see the worksheet Drug.xls) is shown in Figure 5.9.

	A	B	C
1	total list	75 random	count
2	[816 entries]	selections from A	1's in B
3	0	0	0
4	0	0	
5	0	0	
6	0	0	
7	0	0	
8	0	0	
9	0	0	
10	0	0	
11	0	0	
12	0	0	
13	0	0	

Figure 5.9: Employee Drug Testing

1. Use the Urn function to set up column A with 800 “0’s” and 16 “1’s” to embody the null hypothesis that the negative testers and the positive testers belong to the same universe with respect to later drug-based health claims.
2. Select the urn data in column A, select “Shuffle” from the Resampling Stats menu or toolbar, enter B3 as the top of the output range, and enter 75 as the Number of Cells in Output Range.

We want to see how often, when shuffling the 800 “0’s” and 16 “1’s” and drawing a group of 75, we find 4 or more of the “1’s” (drug-related claims) in the set of 75.

3. Use COUNTIF in cell C3 to count the number of “1’s.”
4. With the count cell (C3) selected, choose “Repeat and Score” from the Resampling Stats menu or toolbar and enter (say) 100 as the number of iterations.

The Results sheet for a small run (100 trials) will show you that there’s a low probability that four or more drug-related claims would be found in 75 employees selected at random from a universe of 16 claims and 800 no-claims.

Chapter 6

Correlation and Regression

6.1 Applied Correlation: Baseball Salary vs. Rank

Is baseball payroll (1995-97 total) correlated with team rank (won-lost record over same period)? The observed Pearson correlation coefficient (see Table 6.1) and the statistic of interest is -0.71 , meaning that larger payrolls tend to be associated with lower rank numbers (i.e. better performance). Is this statistically significant?¹

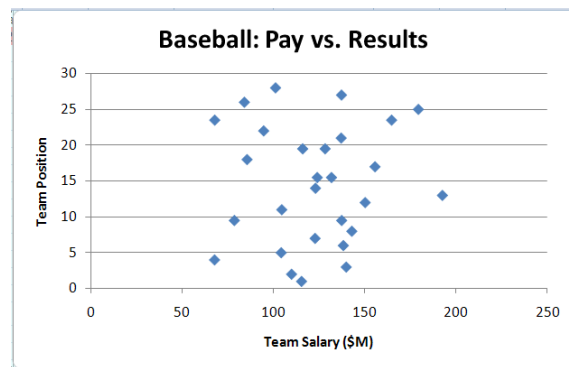


Figure 6.1: Pay and Team Rank in Baseball

¹The rank is determined by the games won and lost over the 3 year period. Data are from the *Washington Post*, March 27, 1998. Statistics compiled by the *Post* according to the formula of the Player Relations Council.

6. CORRELATION AND REGRESSION

	Total Payroll (\$ Million)	Rank
NY Yankees	192.7	3
Baltimore	179.5	4
Atlanta	164.8	1
Cleveland	155.7	2
Chicago WS	150.3	14
Cincinnati	143	9.5
Texas	139.9	11
Colorado	138.3	8
Toronto	137.4	25
St. Louis	137.3	19.5
Seattle	137.1	6
Boston	131.8	7
Los Angeles	128.3	5
San Francisco	124	18
Chicago Cubs	123	21
Florida	122.8	12
Anaheim	116	15.5
Houston	115.4	9.5
Philadelphia	109.9	26
San Diego	104.5	13
NY Mets	104.2	17
Kansas City	101.1	22
Minnesota	94.6	27
Oakland	85.5	23.5
Detroit	84	28
Milwaukee	78.5	19.5
Pittsburgh	67.7	23.5
Montreal	67.6	15.5

Table 6.1: Major League Baseball - 1995-1997

Figure 6.1 displays a scatterplot of team rank vs. payroll for the major league clubs represented by the data.

Statistic of Interest

The statistic of interest is the Pearson correlation coefficient, which measures linear correlation on a scale from -1 (perfect negative correlation) to +1 (perfect positive correlation) and is calculated as follows:

$$r = \frac{1}{(n - 1)} \sum \frac{(x_i - \bar{x})}{s_x} \cdot \frac{(y_i - \bar{y})}{s_y}$$

X_i and y_i represent the x and y values, respectively, for the i th element. \bar{x} and \bar{y} are the averages of the x values and the y values. s_x and s_y are the standard deviations for the x and y values. Excel's CORREL from the Paste Function button will calculate this for you.

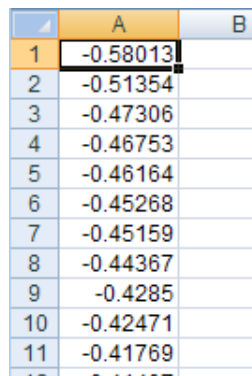
The null hypothesis is that there is no association between payroll and rank, and that the negative value of the correlation coefficient simply arose by a chance alignment of independent variables. Our resampling procedure, then, is to ask how this observed correlation of -.71 compares to correlation coefficients obtained after repeated shuffling of one of the variables relative to the other. This shuffling embodies the null model of no association – after we shuffle one variable we can be sure that any measured correlation between it and the other is simply the product of chance. In Excel we will shuffle the performance data and recalculate the correlation coefficient, then Repeat & Score on the cell that contains that value. Here are the steps spelled out in detail:

1. With the cursor in a blank cell on the data worksheet in Baseball-c.xls, select “CORREL” from the “Insert Function” button.
2. Use B3:B30 as the first array in CORREL input, and C3:C30 as the second.
3. You'll see -0.71 as the function result.
4. Copy the array C3:C30 to the cells H3:H30 (a temporary “parking place”).

6. CORRELATION AND REGRESSION

5. Use the Shuffle function on the Resampling Stats menu or toolbar to shuffle H3:H30 into cells C3:C30. You'll see the correlation value change.
6. Select the "CORREL" cell for Repeat and Score.

How likely is a correlation value of -0.71 by chance? In this set of trials (Figure 6.2), the most negative value was -0.58013, suggesting that the observed value of -.71 is extremely unlikely to have happened by chance.



	A	B
1	-0.58013	
2	-0.51354	
3	-0.47306	
4	-0.46753	
5	-0.46164	
6	-0.45268	
7	-0.45159	
8	-0.44367	
9	-0.4285	
10	-0.42471	
11	-0.41769	

Figure 6.2: Correlation of Payroll and Rank in MLB

6.2 Regression Basics

A Simple Case Using Excel's SLOPE and INTERCEPT Functions

Microsoft Excel comes with functions for calculating regression in data sets, as well as a Regression routine in the Data Analysis set that reports traditional confidence interval values for the regression parameters, for example slope and intercept in a simple x-y case. But as is the case with all other traditional parametric procedures, these confidence interval calculations are based on the assumption that the variables involved are normally distributed.

A resampling approach, in contrast, makes the assumption that the x-y pairs available for study were drawn from a much larger population of possible x-y pairs that is well-represented by the sample at hand. To simulate this population and use it to estimate confidence intervals, we draw randomly and

with replacement from the set of x-y pairs – we bootstrap the cases. This has the effect of “weighting” the data points differently in different rounds of selection, which produces a range of regression parameter estimates in the various simulations. The same procedure, resampling x-y pairs, is easily extended to [x1, x2, x3....,y] sets for regression as well.

	A	B	C	D	E	F
1	Resampling and Regression: The Basics					
2						
3	x1	y	<i>(resampled rows)</i>			
4	2.1	3.7	2.6	15	1.3264	<- slope
5	2	3	2	3	1.2542	<- intercept
6	2.6	15	8	12.2		
7	3.9	4.1	3.9	4.1		
8	4.3	2	3.9	4.1		
9	6	7.5	8	12.2		
10	7.3	9	9.7	16		
11	8	12.2	3.9	4.1		
12	9.7	16	4.3	2		
13						

Figure 6.3: Regression via Resampling - a Simple Case

Let’s take a simple case in which we are interested in the reliability of the estimated y-intercept. Figure 6.3 shows a small collection of x-y pairs. The procedure is:

1. Select the data in both columns (see the workbook Basic.xls) with a standard click-drag.
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 C4 as the top left output cell.
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(D4:D12, C4:C12) into cell E4 and enter =INTERCEPT(D4:D12, C4:C12) into cell E5, as shown in Figure 6.3. These commands mean “find the slope and intercept of the regression line fitted to the data in the referenced range.”
4. Repeat and Score 100 trials on the cell for INTERCEPT.

C	D	E
<i>Bin</i>	<i>Frequency</i>	<i>Cumulative %</i>
-6.4871461	1	1.00%
-4.6497006	1	2.00%
-2.8122551	2	4.00%
-0.9748096	18	22.00%
0.86263585	8	30.00%
2.70008133	21	51.00%
4.53752682	26	77.00%
6.3749723	11	88.00%
8.21241778	6	94.00%
10.0498633	4	98.00%
More	2	100.00%

Figure 6.4: Analyzing Regression Output

If you analyze the Results sheet for the 100 trials with the Histogram function and check the Cumulative output check box, you'll see something like the output in Figure 6.4. It certainly indicates that the computed intercept (in the "Bin" column) wanders all over the y-axis for the resampled data sets. You can compare these results to the results of standard regression analysis with Excel's Regression routine (under Tools/Data Analysis), although it would be well to invest a few minutes in 10,000 trials for the resampling procedure.

In the above approach, we used the functions SLOPE and INTERCEPT. There are two other options for resampling regressions (and you'll want to use one of these when dealing with multivariate problems):

1. The Regression command in the resampling menu calls a Visual Basic for Applications program that uses Excel's own Regression macro.
2. LINEST, Excel's built-in all-purpose regression function. This is a bit tricky to use, since LINEST's output is array formula output, requiring the user to define the exact output range in advance, which will be different for problems with different numbers of variables. We mention this because, if you are an experienced Excel user, you can set up LINEST for your regression problem, Repeat and Score on appropriate cells in LINEST's output range, and pick up a noticeable speed advantage over the Resampling menu's Regression command which must run the Excel macro.

Important Note: Resampling macros (such as Excel's own regression macro)

in the data analysis menu does not work. The Repeat & Score function will not cause these macros to be re-executed. To resample Excel's regression macro, you must execute it from the Resampling menu.

6.3 Baseball Again: Running Regression from the Resampling Add-in

Here we're studying again a simple x-y pairs example, using the above baseball data (workbook Baseball-r.xls). Now the question is:

Can we estimate a linear relationship between payroll and performance, and use resampling to determine how reliable that estimate is?

In posing the question "How reliable is the estimate?" we are in effect asking how the relationship might change were we to pick a different set of data points. In this case, this would not be a different sample from the years, since we have exhausted all data for the given years. Instead, we might consider this to be a sample from an ongoing process that will continue. (Of course, this is not strictly the same thing as a random sample. Things might change in our ongoing process, although baseball is a relatively stable process compared to other things in life. To the extent that things do change, our calculations, whether via resampling or conventional procedures, to determine a confidence interval for our estimated relationship will tend to underestimate the width required for the confidence interval.)

If we call the Regression function from the resampling menu, we'll be asked to identify (Figure 6.5) the x-range, the y-range, and the beginning output cell and a confidence interval.

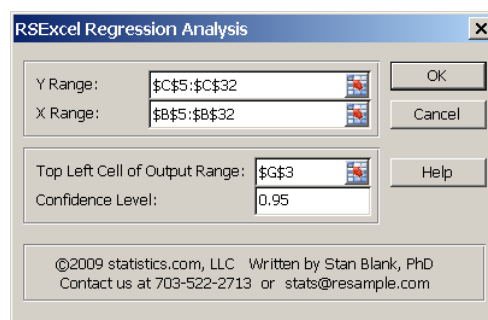


Figure 6.5: X-Y Input for the Resampling Menu Regression Option

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.

In the regression routine output, we could select slope (the cell X Variable 1) and intercept (see Figure 6.6) as Repeat and Score cells, along with the parameter Multiple R (Figure 6.7) to develop a confidence interval for the correlation. If you like, you can perform 1,000 trials in the workbook Basic.xls, both to get a estimate of the time involved in these calculations and to check the agreement with the classical result in the full regression output.

	<i>Coefficients</i>	<i>Standard Error</i>
Intercept	161.9827123	8.729909918
X Variable 1	-2.710876712	0.526023646

Figure 6.6: Slope

Resampled regression output	
SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.662520192
R Square	0.438933004
Adjusted R Sq	0.417353505
Standard Error	5.886124787
Observations	28

Figure 6.7: Another Section of the Same Regression Output

6.4 Multiple Linear Regression: Newspapers and Population

Now let's use resampling in regression to analyze a multivariate problem. In a model taken from Daniel Terrell's *Business Statistics* (1975, Houghton Mifflin), daily newspaper circulation (in thousands) is predicted on the basis

6.4. Multiple Linear Regression: Newspapers and Population

of cities' total retail sales (\$ million) and population per sq. mile (part of the data set is shown in Figure 6.8). We'll use resampling to establish confidence intervals for the parameters of this equation.

	A	B	C
1	Circulation	Sales	Population
2	3	22	48
3	3.3	24	51
4	4.7	37	77
5	3.9	29	66
6	3.2	23	52
7	4.1	32	65
8	3.6	26	57
9	4.3	32	67
10	4.7	35	76
11	3.5	25	53
12	4	31	67
13	3.5	26	56
14	4	30	66

Figure 6.8: Population, Circulation, and Sales

Multiple linear regression yields the following relationship for 25 cities (workbook News.xls):

$$circ = 0.057(sales) + 0.030(pop) + 0.345$$

But how reliable are these estimated coefficients? To answer this question, we'll repeatedly resample the data, calculating the regression coefficients for each resample.

1. Select all the data, choose "Resample," and select "Resample Rows as Units." You can designate D2 as the Top Left Cell of Output Range.
2. From the Resampling Stats menu, apply regression to the RESAMPLED set of 25 cities (Figure 6.9 – the resampled y-range is now in d2:d26, the two x-variables are in e2:f26, the Top Left Cell of Output Range could be H2).
3. Select the regression parameters as Repeat and Score cells.

The regression parameters, in this case, will be the cells labeled Intercept, X Variable 1, and X Variable 2, as shown in Figure 6.10 (for the resampled data).

6. CORRELATION AND REGRESSION

4. Run 100 trials for a first look at the output.

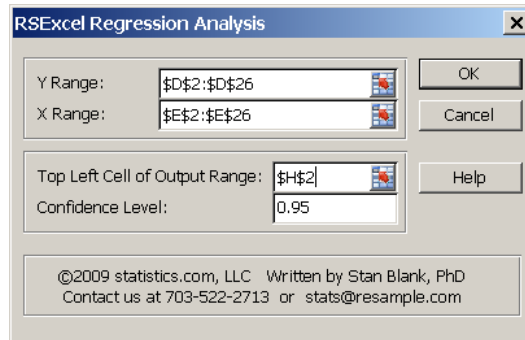


Figure 6.9: Invoking the Regression Command in the Resampling Stats Add-in

ANOVA		
	<i>df</i>	<i>SS</i>
Regression	2	0.3645029
Residual	22	5.7330971
Total	24	6.0976
	<i>Coefficients</i>	<i>Standard Error</i>
Intercept	3.3176664	0.718250743
X Variable 1	-0.09329394	0.120668473
X Variable 2	0.05543181	0.060015

Figure 6.10: Picking Resampled Regression Parameters for Repeat and Score (a Portion of the Resampled Regression Output is Shown)

The estimated 90% confidence limits by Resampling are calculated on the Results sheet by using sorted data to find the 5th and 95th percentiles for the output variables (Figure 6.11).

6.4. Multiple Linear Regression: Newspapers and Population

	A	B	C	D	E	F	G	H	I
1	News data: sample run of 100 trials,sorted								
2	2.039171			-0.37734			-0.1569		
3	2.114388			-0.3748			-0.1565		
4	2.14235	Intercept		-0.3144	sales		-0.12583	pop	
5	2.240847			-0.31011			-0.10804		
6	2.278585	2.278585	at 5 of 100	-0.27981	-0.27981	at 5 of 100	-0.10191	-0.10191	at 5 of 100
7	2.339872	3.931348	mean	-0.27673	-0.0139	mean	-0.09901	0.006169	mean
8	2.381414	5.181095	at 95 of 100	-0.27601	0.200799	at 95 of 100	-0.09133	0.132646	at 95 of 100
9	2.519951			-0.21521			-0.09083		
10	2.653211			-0.21514			-0.08932		

Figure 6.11: Estimated 90% Confidence Intervals Using Resampling

Chapter 7

Analysis of Variance

Analysis of variance is a straightforward extension of the reasoning in hypothesis testing.

7.1 Geyser Timing

For decades, park rangers at Yellowstone, perhaps for lack of other statistical excitement, have recorded the time in minutes between the eruptions of the geyser Old Faithful. Monthly data for an assortment of years is shown in Figure 7.1 (workbook Faithful.xls). As an ANOVA example, in this case the year for the data is the “treatment,” and we want to determine if geological underpinnings are really changing from year to year.

y1951	y1985	y1995	y1996	
60	54	88	42	
74	68	79	86	
62	62	89	89	
62	86	85	79	
68	88	62	85	
85	86	91	74	
52	65	57	56	
104	94	83	82	
60	85	66	58	
79	65	94	95	
60	90	91	86	
74	62	89	84	
70.00	75.42	81.17	76.33	75.73
Sum of absolute group mean deviations from grand mean:				12.08

Figure 7.1: Geyser Data

7. ANALYSIS OF VARIANCE

The null hypothesis is that all the yearly data comes from the same underlying process, and that the variations from one year to the next are just due to chance variation.

Statistic of Interest

We will measure the degree of variation from one year to another by taking the sum of absolute deviations between the yearly means and the overall (grand) mean of 75.73 minutes.

For the observed data, this statistic is 12.08.

1. Record each years data in a separate column, calculate the column means and the grand mean (see Figure 7.1 and columns A to D in Faithful.xls).
2. Find the sum of the absolute deviations between the yearly means and the grand mean (E18).
3. Shuffle the observed data to an adjacent range (in Faithful.xls the top left cell of the shuffled range is F4), and copy the formulas to calculate the means, as well as the statistic of interest – the sum of absolute deviations between the yearly means and the grand mean (J18).
4. With this cell (J18) selected, choose “Repeat and Score” from the Resampling Stats menu or Toolbar and enter (say) 1000 iterations, click “OK”.

	A	B	C	D	E	F
1	24.54167		Old Faithful			
2	19.75					
3	19.54167		How many of the 1000 trials produced			
4	18.79167		a value >= the observed value of 12.08?			
5	18.54167					
6	18.41667		85			
7	18.08333					
8	17.95833		Estimated p-value =		0.085	
9	17.95833					
10	17.95833					
11	17.79167					
12	17.54167					
13	17.125					
14	17.125					

Figure 7.2: 1000 Trials, Estimated $p = .085$

A run of 1000 trials gives the result in Figure 7.2. This indicates that the years are not significantly different (at the $p = 0.05$ level).

7.2 Coagulation Time

The following example shows how the above procedures can be used easily with asymmetric tables. In the example of Diet.xls, the starting point is the table shown in Figure 7.3. Here 24 animals (Box, Hunter, Hunter; *Statistics for Experimenters*, p. 166) are randomly allocated to 4 different diets, but the numbers allocated to different diets are not the same. The coagulation time for blood is measured for each animal. Are the diet-based differences significant?

	A	B	C	D	E	F
1		Diet A	Diet B	Diet C	Diet D	
2	grand mean:	62	63	68	56	
3	64	60	67	66	62	
4		63	71	71	60	
5		59	64	67	61	
6			65	68	63	
7			66	68	64	
8					63	
9					59	
10	Group mean:	61	66	68	61	
11	abs devs	3	2	4	3	
12		Sum of absolute differenc			12	
13						

Figure 7.3: Table of Coagulation Times

Statistic of Interest

The statistic of interest is the sum of the absolute deviations between the group means and the overall mean.

The null hypothesis is that the diet makes no difference in coagulation time, and that the differences among the groups can be accounted for by chance variation.

We can test this null hypothesis by combining all observations together, shuffling them, then dealing them out into groups of 4, 6, 6 and 8 and recalculating the statistic of interest. If the sum of absolute differences for the

7. ANALYSIS OF VARIANCE

shuffled data frequently exceeds 12, we can conclude that chance variation might be to blame.

Although the sample sizes for the four groups are different, the analysis proceeds much the same as with the previous problem.

1. Record the result of each diet in a separate column, calculate group means, the grand mean, and the sum of absolute deviations of the group means from the grand mean (Figure 7.3, columns B-E in Diet.xls).
2. Select these data, and Shuffle them to a nearby range (B17 is the top left cell for the shuffled range in Diet.xls). Choose “Normal Shuffle” and leave the “Shuffle Blank Cells” box unchecked. Figure 7.4 illustrates the shuffled data.

	A	B	C	D	E
15					
16		SHUFFLED DATA			
17	grand mean:	68	63	63	67
18	64	59	63	71	67
19		66	59	68	60
20		68	65	64	64
21			61	63	56
22			62	62	66
23					60
24					71
25	Group mean:	65.25	62.17	65.17	63.88
26	abs devs	1.25	1.83	1.17	0.13
27		Sum of absolute difference			4.38
28					

Figure 7.4: Dealing with asymmetric tables (data are shuffled together, then redistributed into a table of same structure as original table)

3. Find the group means and sum of absolute deviations from the grand mean for the shuffled data.
4. Repeat and Score on the sum of absolute deviations for the shuffled data, doing (say) 1000 trials.
5. Then, on the Results sheet, use COUNTIF on the results to determine how often the value for the test statistic was greater than or equal to 12.

For these data, greater than or equal to 12 did not occur in 1000 trials and only twice in 10,000 trials (Figure 7.5), indicating that the diets' difference for coagulation time is certainly significant (the classical ANOVA result agrees on significance, reporting an extremely low p-value). The main point of this particular example is that, even if there were seven diets and seven different sample sizes, it would be the same easy process to set up this test statistic. That's not the case with the F statistic – relatively few professional statisticians can set up ANOVA for large asymmetric tables without consulting a reference or trusting commercial ANOVA software.

	A	B	C	D	E	F
1	12.41667					
2	12.08333	Diet Output 10000 Trials				
3	10.875					
4	10.83333		Only 2 Resampled trials showed a			
5	10.75		sum of absolute deviations >= 12			
6	10.75		(the observed value)			
7	10.66667					
8	10.625		Est. p-value = .0002			
9	10.625					
10	10.58333					
11	10.54167					
12	10.5					
13	10.45833					

Figure 7.5: Low Probability Diet Outcome

7.3 Resampling and the F-test

The simple data set shown in Figure 7.6 is a slightly modified (the table entries aren't just integers) version of an example (workbook F.xls) that's been used in textbooks continuously over the last fifty years. For purposes of illustration, we could say that the three treatments are different fertilizers and the numbers are heights of beanstalks. The question is whether treatments a, b, and c produce significantly different heights.

<i>treatment a</i>	48.20	49.30	50.70	49.50
<i>treatment b</i>	47.44	46.23	48.84	48.16
<i>treatment c</i>	49.56	51.36	50.05	50.17

Figure 7.6: Simple ANOVA Table: Heights of Beanstalks

7. ANALYSIS OF VARIANCE

The standard way to answer this question is to use single-factor ANOVA, first computing the “F” statistic from the data table and then using tables to look up a probability p that corresponds to that F value (7.42), for the number of degrees of freedom in the problem. This calculation is shown in Figure 7.7.

grand mean =	49.13
total var =	22.87
var within	8.63
var between	14.24
MS within	0.96
MS between	7.12
F(2,9)	7.42

Figure 7.7: Computing F from the Table

We can also obtain this p -value via resampling.

The null hypothesis is that there is no difference among the fertilizers – they all result in equal growth of the bean plants. Under this null hypothesis, we shuffle together all the bean plant heights and draw 3 resamples (here we will draw without replacement) to see whether these 3 resamples differ among themselves as much as the 3 observed treatment samples differed.

As noted above, the standard method uses the F statistic. The F statistic measures between-group variation in relation to the total variation, which allows you to compare the F statistic from any table against a single tabulated F distribution.

Possible Statistics of Interest

From the point of view of resampling, F is just another possible useful statistic. It’s a simple matter to resample the original data back into the table structure, recompute F each time, and save the F results in the output sheet. Then you compare the observed F statistic to the list of F values in the resampling output sheet to see how often you get a resampled F as extreme as the observed value.

With resampling, other statistics may be used. In this case, the sum of absolute differences of the group means from the overall table mean would be a natural choice. In this example, we use the means of each row, and sum the three absolute differences of these means from the overall mean. Then we compare it to the same statistic calculated after shuffling the table.

This statistic has only positive values, and its observed value will only be very large (comparatively) when the variation between group means is large. In Figure 7.8, the cell corresponding to this statistic is called *stat*. For an additional check, this particular example also uses a statistic called *stat2*, the sum of squares of the elements in *stat*. Note that:

1. Analyzing the distribution of either statistic would be a daunting task without a computer.
2. Nonetheless, it's quite obvious how to formulate this statistic in a great variety of different situations. Either row means or column means may be measured, and asymmetrical tables pose no problem.

<i>row mean</i>	<i>difference</i>
49.43	0.30
47.67	1.46
50.29	1.16
stat	2.92
stat2	3.56
stat = sum of absolute differences from grand mean	
stat2 = sum of abs. diff. elements squared	

Figure 7.8: Simple Statistics for Table Analysis

We can select all three of these statistics (“F,” “stat” and “stat2”) as Repeat and Score cells for resampling, and use COUNTIF to find the probability of finding a value for each statistic as large as or larger than the original-data value. The F results give a probability (Figure 7.9) of approximately 93 counts in 10,000 greater than or equal to the observed value of 7.42, or p of about 0.0093. Stat (sum of absolute deviations) yielded a p-value of 0.0096, and stat2 a p-value of 0.0057.

7. ANALYSIS OF VARIANCE

	A	B	C	D	E	F	G	H
1	stat	stat2	F					
2	3.23666667	4.734954	21.6731638					
3	3.23666667	4.734954	21.6731638				COUNTIF	P-VALUE
4	3.23666667	4.453817	15.8529373				result	
5	3.23666667	4.453817	15.8529373					
6	3.23666667	4.423517	15.3765624			stat	96	0.0096
7	3.23666667	4.394117	14.9351801			stat2	57	0.0057
8	3.23666667	4.394117	14.9351801			F	93	0.0093
9	3.23666667	4.365617	14.525627					
10	3.23666667	4.365617	14.525627					
11	3.23666667	4.365617	14.525627					

Figure 7.9: p-values from Resampling

Technical Note: Sampling with replacement versus sampling without replacement

After creating our null universe, should we be sampling with or without replacement? In technical terms, should we be resampling bootstrap style (with replacement), or permutation style (without replacement, also called “shuffling”)?

Permutation tests date from the 1930’s work by Fisher and Pitman, and are classic tests regarded as the “gold standard” in the sense that they yield “exact” p-values. A test is exact if, when testing multiple samples from the null population, it yields erroneous “significant” results 5% of the time or less (when you conduct the test at $\alpha = .05$). In other words, an exact test preserves Type I error at or below the level of the test.

Fisher and Pitman worked with “exhaustive” permutation tests, in which all possible permutations of the combined data into two or more samples were enumerated. The principle is the same for the tests we have used above in which the data are shuffled and then randomly permuted or divided into two or more resamples. The latter are an unbiased estimate of the former.

Bootstrap style resampling (that is, with replacement) from a null model cannot make this guarantee. However, Westfall and Young (1992) point out the results of simulations that show that bootstrap style tests for a difference in binomial proportions preserve Type I error pretty effectively, and yield more power than permutation style tests. Permutation tests in such cases are conservative.

Because permutation (shuffling) tests are classically regarded as standard treatments, we will use them in most of the examples where we have two or

more samples that can be combined for this type of test.

Chapter 8

Non-Parametric Statistics

Most nonparametric tests were developed because dealing with data converted to ranks or signs permits analysis that is computationally easy and does not depend on assumptions about the distribution of the data.

8.1 Birthweight Revisited: A Signs Test

Consider again the birthweight problem introduced in Chapter 4. Let us suppose now that this is a matched-pair study, instead of a study with two independently selected groups of women. Each woman on treatment is matched with a control subject of similar physical and socio-economic characteristics.

Why do this? The purpose is to allow us, in our significance test, to factor out the variation from one subject to another that might otherwise obscure the true effect. As a hypothetical example, consider (Table 8.1) just a few reading scores on the same subjects, where the subjects take a short test after reading a passage without background music and, a week later, after reading a similar passage with music.

While there does seem to be an effect from the music, it is very small compared to the differences among the individuals. If we did a standard resampling (permutation) test in which all the scores get tossed in the same hat before shuffling, our pairs of resamples will show big differences between them just due to the random selection of subjects. Such a test will not do a good job of identifying the music effect. In statistical terms, it will lack power.

When we have paired subjects, we would like to perform a hypothesis test

8. NON-PARAMETRIC STATISTICS

Subject #	Without Music	With Music
1	24	27
2	79	80
3	17	18
4	50	50
5	98	99

Table 8.1: Reading Scores

that, when we resample, preserves the association of each treatment subject with her paired control.

Table 8.2 shows the birthweight data table again, showing the birthweights of babies born to women on treatment to prevent low birthweights, and on placebo (workbook Birth.xls). In this nonparametric version, each row in the table is shown with a score: “1” if the treatment baby in that row had a higher birthweight, and “0” if the weight was not higher.

Treatment	Control	Score
6.9	6.4	1
7.6	6.7	1
7.3	5.4	1
7.6	8.2	0
6.8	5.3	1
7.2	6.6	1
8.0	5.8	1
5.5	5.7	0
5.8	6.2	0
7.3	7.1	1
8.2	8.0	1
6.8	6.9	0
6.8	6.6	1
4.7	4.2	1
8.6	6.8	1
7.01	6.39	←—Mean

Table 8.2: Birthweights

Statistic of Interest

The statistic of interest is the number of times treatment does better (i.e., the number of “1’s”).

Our null hypothesis is, as before, that there is no benefit from the treatment – that each pair is as likely to get a “0” as a “1.” Under the null hypothesis, we attribute the fact that the treatment group got 11 “1’s” to chance. Therefore, we can test the null hypothesis as follows:

1. Flip a coin 15 times and record the number of heads (“treatment” wins).
2. Repeat step 1 many times.
3. How often did we get 11 or more heads?

Note that each treatment/control pair is now treated as a self-contained unit.

In Excel:

1. Put a “0” in A1 and a “1” in A2.
2. Take a resample of size 15 and put it in B1:B15.
3. In B17, SUM B1:B15 to get the number of “1’s.”
4. This is our statistic of interest, to be compared to the observed value of 11, so B17 is what we Repeat & Score. Do, say, 1000 simulations and examine the Results sheet.

Use the COUNTIF function to count how often we get an 11 or greater in the range A1:A1000 in the Results sheet. This is the estimated p-value, which should be fairly close to .06.

8.2 Birthweights a Third Time: A Paired Permutation Test

Perhaps you are struck by the fact that the paired sign test, above, yielded a p-value very similar to the unpaired test presented in chapter 4. Two factors are at work here:

1. We gain power (lowers p-values) by pairing.
2. We lose power (raises p-values) by converting measured data to 0/1 data.

Also note that the amount of gain we get by pairing depends on the relative size of the treatment effect when compared to the variation among subjects.

In the “reading with music” example above, the treatment effect was very small compared to the variation among subjects, hence easily obscured by the latter. In the birthweight problem, the treatment effect is not as small, relative to the variation among subjects.

So let us try a third approach – a “paired permutation test.”

Statistic of Interest

As was the case the first time we did this problem (Chapter 4), the statistic of interest is the difference in average birthweights.

Our null hypothesis is still “no treatment benefit,” and it implies that the birthweights for each matched pair could belong to either member of the pair. We test by randomly shuffling the pairs of birthweights, then recalculating the mean birthweight for each column. If the difference in means is rarely as large as the observed difference in means, we conclude that the observed difference in means is probably not due to chance.

In Excel (see worksheet Birthweight-pairs.xls)

1. Click and drag to select the two columns of data values.
2. Using Shuffle on the Resampling Stats menu or toolbar, select “Shuffle Within Rows.”
3. Take the mean (average) of the weight differences in the shuffled data, and find the difference in means (F19).
4. Use this difference cell for Repeat and Score.

Figure 8.1 shows this; the scores used in our previous procedure can now be ignored. The procedure we now use is like the one we used in Chapter 4, except that we choose “Shuffle Within Rows.”

	TREATMENT	CONTROL		Shuffled	
	6.9	6.4		6.4	6.9
	7.6	6.7		7.6	6.7
	7.3	5.4		5.4	7.3
	7.6	8.2		7.6	8.2
	6.8	5.3		5.3	6.8
	7.2	6.6		6.6	7.2
	8.0	5.8		5.8	8.0
	5.5	5.7		5.5	5.7
	5.8	6.2		6.2	5.8
	7.3	7.1		7.3	7.1
	8.2	8.0		8.2	8.0
	6.8	6.9		6.8	6.9
	6.8	6.6		6.8	6.6
	4.7	4.2		4.7	4.2
	8.6	6.8		8.6	6.8
Mean	7.01	6.39	Mean	6.59	6.81
	Difference	0.61		Difference	-0.23

Figure 8.1: Shuffled by Rows

The result for 1000 trials is a p-value of 0.006. As we expect, it is smaller than those produced by either of the previous two procedures. This reflects the fact that we gain the increased power from pairing without losing power by converting measured data to 0/1 data.

8.3 Rank Sum Test

As another nonparametric test example, look at this textbook problem (Mario F. Triola, *Elementary Statistics*, 8th ed., Addison-Wesley) that investigates stress and pay levels in different occupations (the worksheet Pay.xls). Figure 8.2 shows the jobs and their stress/pay rankings. We've already accomplished something just by sorting – the sorted data make it appear that there's a very strong correlation.

Statistic of Interest

To evaluate correlation, we find the sum of ranks for the first five stress entries. Generally speaking, the lower this score, the greater the correlation. The sum for these ranks is 16, as shown. The minimum possible score is 15.

One possible resampling test is to shuffle the set of stress ranks and take the sum of the top half (first five positions) as the Repeat and Score cell. The

8. NON-PARAMETRIC STATISTICS

	A	B	C	D	E	F	G	H
1	Pay vs. Stress				Pay vs. Stress			
2					<i>sorted by pay rank</i>			
3		pay	stress			pay	stress	
4	stockbroker	2	2		pilot	1	1	
5	zoologist	6	7		stockbroker	2	2	
6	EE	3	6		EE	3	6	sum of first
7	principal	5	4		psychologist	4	3	five stress ranks=
8	hotel mgr.	7	5		principal	5	4	16
9	bank officer	10	8		zoologist	6	7	
10	safety	9	9		hotel mgr.	7	5	
11	home ec	8	10		home econ	8	10	
12	psy	4	3		safety	9	9	
13	pilot	1	1		bank officer	10	8	

Figure 8.2: Shuffled by Rows

observed sum of the top half is 16 – is this a lower number than might be expected in a random ordering of ranks?

Figure 8.3 shows that this ranksum or a lower value occurred only once in 100 resample trials. To sharpen this probability estimate further, we could try 1000 trials, but even this quick test suggests pay and stress are correlated at a statistically significant level.

	A	B
1	16	
2	18	
3	19	Pay/Stress
4	19	Output
5	19	[100 trials]
6	20	
7	20	
8	21	
9	21	
10	21	
11	21	
12	22	
13	22	
14	22	

Figure 8.3: Ranksum Results

8.4 Another Correlation Study

Hypothetical physical-education scores of a group of ten high-school boys are shown in Table 8.3 ordered from high to low, along with the I.Q. score for each boy. The ranks for each student's athletic and I.Q. scores are also shown.

Athletic Score	I.Q. Score	Athletic Rank	I.Q. Rank
97	114	1	3
94	120	2	1
93	107	3	7
90	113	4	4
87	118	5	2
86	101	6	8
86	109	7	6
85	110	8	5
81	100	9	9
76	99	10	10

Table 8.3: Hypothetical Athletic and I.Q. Scores for High School Boys

A little inspection shows that this example is exactly the same situation as the pay/stress example above, once the IQ scores are converted to ranks. We have one set of ranks 1 through 10 linked to another set (Figure 8.4, from workbook Ath.xls).

	A	B	C	D	E	F
1	Nonparametric correlation					
2						
3	Athletic	IQ	Athletic	IQ rank		
4	score		rank			
5	97	114	1	3		
6	94	120	2	1		
7	93	107	3	7		
8	90	113	4	4		
9	87	118	5	2	17	ranksum
10	86	101	6	8		
11	86	109	7	6		
12	85	110	8	5		
13	81	100	9	9		
14	76	99	10	10		

Figure 8.4: I.Q. and Athletic Ranks

8. NON-PARAMETRIC STATISTICS

So, the steps are:

1. List the two data columns together, with first column (Athletic score) in one-through-ten order and the second column listing the linked (IQ) ranks. Note the “first five” ranksum in this order.
2. Shuffle the second column, and use the cell with the “first five” ranksum as the Repeat and Score cell for 1000 trials.
3. Sort the results, and see where the original data ranksum value occurs in the list (Figure 8.5).

	A	B	C	D	E
1	15		1000 Trials		
2	15				
3	15		17 trials had a rank sum with		
4	15		as low a value as the observed		
5	15		ranksum (17)		
6	16				
7	16		Estimated p-value = .017		
8	16				
9	17				
10	17				
11	17				
12	17				
13	17				
14	17				
15	17				
16	17				
17	17				
18	17				
19	17				
20	17				
21	17				
22	18				
23	18				

Figure 8.5: Ranksum Results, Round Two

Chapter 9

Stratified Resampling

In several problems, we have seen Resampling Stats' ability to confine the resampling or shuffling to columns or rows, in effect putting up “walls” between rows or columns to keep shuffled or resampled values from crossing those barriers.

Let's now examine several problems that take advantage of these and other features useful for stratified resampling.

9.1 Evaluating Corporate Mergers; “Shuffling Within Rows”

In a study by Simon, Mokhtari, and Simon (1996), 33 advertising agencies that merged over a period of years were each compared to advertising agencies that did not merge. Specifically, each pair of firms that merged was compared against:

- a) a pair of non-merging firms that were roughly the same size as the merging firms before the merger, and
- b) a single firm that was roughly the same size as the merged entity *after* the merger.

The first entry in the data table (see the worksheet mergers9-1.xls) is shown in Table 9.1. (provided by the authors).

9. STRATIFIED RESAMPLING

Set #	Merged	Match1	Match2
1	-0.20000	0.02564	0.000000

Table 9.1: Revenue growth in year 1 following merger

Comparisons were made in several years before and after the mergers to see whether the merged entities did better or worse than the non-merging entities they were matched with by the researchers, but for simplicity we may focus on just one of the more important years in which they were compared – say, the revenue growth rates in the year after the merger. Figure 9.1 displays the top rows of data in the mergers9-1.xls workbook.

	A	B	C	D
2	Merger study			
3	SET #	MERGED	MATCH1	MATCH2
4	1	-0.2	0.02564	0
5	2	-0.34831	-0.125	0.08046
6	3	0.07514	0.06322	-0.023121
7	4	0.12613	-0.04199	0.164671
8	5	-0.10169	0.08	0.277778
9	6	0.03784	0.14907	0.430168
10	7	0.11616	0.15183	0.142857
11	8	-0.09836	0.03774	0.04

Figure 9.1: Merger data: revenue growth in year after merger

Here are those average revenue growth rates for all 33 entries for the three groups in Table 9.2:

Entity Type	Year's Revenue Growth
Merged	-0.021300
Match 1	0.092085
Match 2	0.095931

Table 9.2: Average Revenue Growth Rates

Is the poorer performance of the merged firms statistically significant?

Our null hypothesis is that there is no difference among the three groups with respect to revenue growth. In light of the fact that we have matched groups, we must consider carefully how to implement a resampling test of this

9.1. Evaluating Corporate Mergers; “Shuffling Within Rows”

Set #	Merged	Match1	Match2
1	1	3	2

Table 9.3: Ranked Within Matched Set: (1 = Worst, 3 = Best)

null hypothesis. (See the birthweight problem in the previous chapter for an analogy.)

The authors felt that it was inappropriate to average together growth rates for firms of widely different sizes. Moreover, any resampling scheme must maintain the segregation of each set from the others.

To meet both these problems, we can use a simple resampling procedure that maintains the separation between matched sets by converting each observation into a rank (1, 2 or 3) within the matched set (Figure 9.2), then shuffling within that set. Here’s an example of the ranking for the first data entry in Table 9.3:

Note the assignment of ranks – “1” to the worst, “3” to the best.

That is, for the first line in the table, the two separate firms (Match 1) did best, the merged firm did the worst, and the single larger firm landed in the middle (for revenue growth).

set	mrg	mt1	mt2
1	1	3	2
2	1	2	3
3	3	2	1
4	2	1	3
5	1	2	3
6	1	3	2
7	1	3	2
8	1	2	3
9	1	2	3

Figure 9.2: Merger Data, in Rankings

The average ranks for the three groups are displayed in Table 9.4:

Entity Type	Year's Revenue Growth
Merged	1.45
Match 1	2.18
Match 2	2.36

Table 9.4: Average Revenue Growth Rates

Statistic of Interest

The statistic of interest is the average rank for the merged group.

The null hypothesis is that the ranks within each set were assigned randomly, and that “merged” came out so poorly just by chance. We are specifically interested in whether the “*merged*” group could come out this poorly; we are not interested in investigating whether any group could come out this poorly. The following procedure simulates random assignment of ranks to the “merged” group:

1. Put the numbers 1 through 3 in a column.
2. Select a number at random 33 times.
3. Average the 33 numbers, and Repeat and Score on the AVERAGE cell.
4. Try 1000 trials, sort the output and see how often the average is as low as 1.45.

A sample run, shown in Figure 9.3, shows that the poor performance of the merged firms is statistically significant (1.55 is the lowest value!) and not a random happening. You can easily confirm this result yourself – try a run of 10000 or 20000 trials and find where the value 1.45 appears in the sorted output.

9.2 Mergers Revisited

Another approach to the mergers problem would keep the data in its original continuous form, rather than converting it to ranks as shown in Figure 9.4:

Let's outline a procedure that uses these data as they are.

	A	B	C
1	1.55		
2	1.58		
3	1.58	[1000 cells,	
4	1.61	merger ranking]	
5	1.61		
6	1.64		
7	1.64		
8	1.64		
9	1.64		
10	1.64		
11	1.67		
12	1.67		
13	1.70		
14	1.70		
15	1.70		
16	1.70		
17	1.70		
18	1.70		
19	1.70		
20	1.70		
21	1.70		

Figure 9.3: 1.45 is Statistically Significant!

	A	B	C	D
2	Merger study			
3	SET #	MERGED	MATCH1	MATCH2
4	1	-0.2	0.02564	0
5	2	-0.34831	-0.125	0.08046
6	3	0.07514	0.06322	-0.023121
7	4	0.12613	-0.04199	0.164671
8	5	-0.10169	0.08	0.277778
9	6	0.03784	0.14907	0.430168
10	7	0.11616	0.15183	0.142857
11	8	-0.09836	0.03774	0.04

Figure 9.4: Original Merger Data

Refer again to Table 9.4, the average revenue growth rates for all 33 groups.

We are interested in the difference between the merged firms and and their two matches, so we might choose as our test statistic the difference between the mean of all the merged firms and the mean of both sets of matches.

Statistic of Interest

The statistic of interest is the mean of the merged firms' revenue growth minus the average of both sets of matches.

The observed value of the test statistic is $-.1153$.

The null hypothesis is that the results within each triplet (merged firm, single firm match, two-firm match) are indistinguishable from one another – that each result could just as well have happened to any of the three. The alternative hypothesis is that the merged firms did more poorly than the un-merged matches, to a greater extent than chance would predict.

As with the birthweight problem, we will confine the random shuffling within each matched set, reflecting the fact that each matched set has characteristics that are shared by that set, but not necessarily by other matched sets. This ensures that variation from one matched set to another does not obscure the variation we are interested in – the difference between the merged firms and their un-merged matches.

The following procedure simulates the null model's random assignment of results:

1. Array the data in a matrix where column 1 is the merged firm, column 2 is match 1 and column 3 is match 2, and each row is a set of entities with approximately the same level of business.
2. Shuffle the values within each row.
3. Find the means of each column, and the average of the means of columns 2 and 3.
4. Subtract the average of the means of columns 2 and 3 from the mean of column 1 and record.
5. Repeat steps 2-4 (say) 1000 times.
6. Observe how often the shuffled test statistic is less than or equal to the observed value of $-.1153$.

In Resampling Stats (file mergers9-2.xls):

1. Select the data and click on "Shuffle."

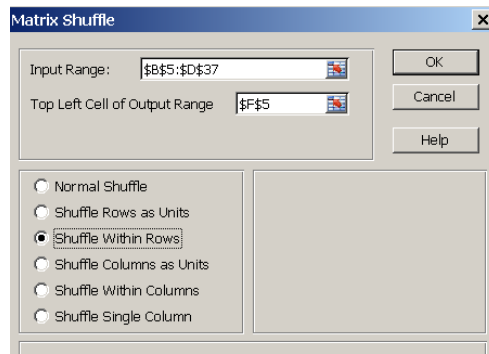


Figure 9.5: Shuffle Within Rows

Mergers - Example 9.2						
Merger Study						
SET #	MERGED	MATCH1	MATCH2	MERGED	MATCH1	MATCH2
1	-0.2	0.02564	0	-0.2	0.02564	0
2	-0.34831	-0.125	0.08046	-0.125	-0.34831	0.08046
30	-0.00676	0.25494	0.237047	0.25494	-0.00676	0.237047
31	-0.16298	0.01124	0.190476	-0.16298	0.190476	0.01124
32	0.19182	0.15048	0.151994	0.151994	0.19182	0.15048
33	0.06116	0.17045	0.093525	0.06116	0.093525	0.17045
	-0.02130	0.09208	0.09593	0.05282	0.04358	0.07032
	Average M1, M2		0.09401	Average M1, M2		0.05695
	Avg. - merged	0.11531		Avg. - merged	0.00413	

Figure 9.6: Statistic of Interest

2. Choose “Shuffle Within Rows” as shown in Figure 9.5.
3. At the bottom of the data in mergers9-2.xls, formulas have been entered to calculate the means of the columns, as well as the statistic of interest, for both the observed and the shuffled data. The cell we want to score is the statistic of interest for the shuffled data, highlighted in Figure 9.6:
4. Do 1000 trials and you will see that a shuffled value of the test statistic as low as the observed value of -.1153 is extremely rare.

Figure 9.7 indicates that the observed inferiority of the merged firms is not easily explainable by chance variation.

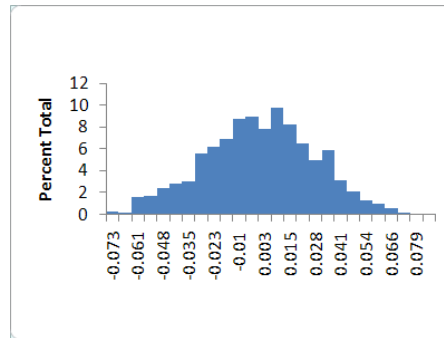


Figure 9.7: Histogram of Merged Data

9.3 Reading Methods: Use of Variable to Denote Strata

Your data may be in a row/column format in which the strata are indicated by values in one or more stratum variable columns. Resampling Stats lets you specify up to two stratification variables (columns), in which case resampling or shuffling will be confined within each value (stratum) of this variable. The purpose of stratified resampling is to control for the effect of variables not of interest (“nuisance parameters”) when testing variables of interest. For example, if you want to test the performance of two reading methods, you could stratify by class so that differences between classes do not obscure the effect of the reading method. Consider the following hypothetical data in Figure 9.8 from the workbook reading.xls:

Is method B’s superiority statistically significant? Ignoring, for the moment, the class variable, the standard permutation test would have us repeatedly shuffle all the scores together and reallocate them to the two columns randomly, then determine how often we got a randomly generated “improvement” as great as the observed value of 1.34. Here is the result of one such shuffling in Figure 9.9 (again, the first column, denoting the student’s class, is ignored):

When the data are shuffled in this fashion, note that the values associated with class 1 (which are lower) tend to get shuffled together with those for class 2 (which are higher). The difference in the scale of the values, even though they show much the same effect (superiority of B) when taken by class, will mean that a lot of noise gets introduced into the randomized distribution of

9.3. Reading Methods: Use of Variable to Denote Strata

DATA (hypothetical)			
Class	Method A	Method B	Ratio
1	25	39	1.56
1	19	33	1.74
1	21	19	0.90
1	21	24	1.14
2	55	61	1.11
2	36	52	1.44
2	43	65	1.51
2	39	59	1.51
2	30	41	1.37
2	40	46	1.15
avg. improvement ratio			1.34

Figure 9.8: Hypothetical Reading Data

SHUFFLE 1			
unstratified			
1	40	33	0.825
1	65	30	0.461538
1	25	46	1.84
1	19	55	2.894737
2	41	24	0.585366
2	39	36	0.923077
2	61	43	0.704918
2	59	39	0.661017
2	21	19	0.904762
2	21	52	2.47619
score cell 1			1.227661

Figure 9.9: Results of One Unstratified Shuffle

the ratio when they are all shuffled together. This means the observed result will not seem that extreme when compared to the randomization distribution (it will have a higher p-value). The solution is to confine the shuffling to within each class (stratify) as shown in Figure 9.10:

The results confirm that the stratified test yields a lower p-value than the unstratified test. Only 1.4% of the randomly shuffled stratified resamples showed a ratio as large (or larger than) the observed value, in contrast to 20.7% of the unstratified results (see the Results tab in Reading.xls).

9. STRATIFIED RESAMPLING

SHUFFLE 2			
stratified by column 1			
1	25	33	1.32
1	19	21	1.105263
1	21	19	0.904762
1	39	24	0.615385
2	43	41	0.953488
2	55	59	1.072727
2	36	46	1.277778
2	40	65	1.625
2	30	52	1.733333
2	39	61	1.564103
score cell 2			1.217184

Figure 9.10: Results of One Stratified Shuffle

Stratification is an option that appears once you invoke a shuffle or resample dialog. Simply check the Stratified Sample box and specify the column to be used for stratification. If more than one is specified, the second will be nested in the first. Resampling Stats will automatically confine resampling or shuffling within the defined strata. You can click on the column to define it (while the cursor is in the box where you specify the column), you can type in column letters (A, BA, etc.), or column numbers (e.g. 2, signifying the second column from the left in the selected range). Figure 9.11 illustrates the Stratified Shuffle dialog option.

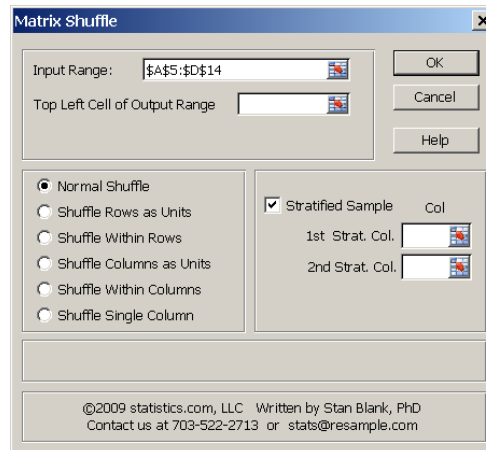


Figure 9.11: Stratified Shuffle Option

9.4 Darwin's Plants: Strata in Separate Ranges

In this example, the data are structured with separate ranges where each of the strata are located. Resampling Stats lets you specify where each stratum ends using a marker (&n, where "n" is the stratum number).

Charles Darwin (1900) tested the growth of plants produced according to two methods of fertilization. Some were fertilized with the pollen of other plants and an equal number were fertilized with their own pollen. He grew the plants in four pots, each pot containing both cross-fertilized and self-fertilized plants, then measured the plant heights after several weeks.

The crossed plants averaged 20.198 inches, which was 2.623 inches higher than the self-fertilized plants. Is this difference significant?

Statistic of Interest

The statistic of interest is the difference in means between the cross-fertilized group and the self-fertilized group.

Our null model is that there is no inherent difference between the growth propensities of cross- and self-fertilized plants, and that the difference between the two groups arose merely through the random assignment process.

We test this null hypothesis by repeatedly shuffling the self- and cross-fertilized heights together, and finding whether the difference in shuffled means is often as great as the observed difference.

Stratified Resampling - Resampling Stats "&n" syntax

To control for the variation introduced by using different pots, we shuffle within pots. Using Resampling Stats' &n syntax to indicate the end of strata (in this case, pots) makes this easy to do. Place "&1" at the top left of the first stratum (pot) to be shuffled, "&2" at the top left of the second stratum, and so on, placing && at the bottom left of the last stratum. This procedure is illustrated in Figure 9.12 from workbook Darwin.xls.

Note that you should enter a parallel set of &1, &2, etc. (but no &&) in the region where you plan to place the shuffled output.

You can then select the entire data set, select "Shuffle" (or "Resample"), and Resampling Stats will automatically confine the whatever shuffling (or

9. STRATIFIED RESAMPLING

	A	B	C	D	E	F
5	Observed				Shuffled (within pots)	
6		Crossed	Selfed		Crossed	Selfed
7		&1			&1	
8	Pot I	23.5	17.4			
9		12	20.4			
10		21	20			
11		&2			&2	
12	Pot II	22	20			
13		19.2	18.4			
14		21.5	18.6			
15		&3			&3	
16	Pot III	22.2	18.6			
17		20.4	15.2			
18		18.3	16.5			
19		21.6	18			
20		23.2	16.2			
21		&4			&4	
22	Pot IV	21	18			
23		22.1	12.8			
24		23	15.5			
25		12	18			
26		&&				

Figure 9.12: Resampling Stats “&n” Syntax

resampling) operation you select within the bounds of each stratum.

Important Note: When you use the &n syntax for stratified resampling or shuffling, for the output range you cannot select merely the top left cell. You must select the entire destination range that contains the &1, &2, etc. (i.e. the range where you intend to place the shuffled or resampled data) as shown in Figure 9.13.

The screenshot shows the 'Matrix Shuffle' dialog box with the following settings:

- Input Range: \$B\$7:\$C\$26
- Top Left Cell of Output Range: \$F\$7:\$E\$26
- Selected Shuffle Method: Normal Shuffle
- Other options: Shuffle Rows as Units, Shuffle Within Rows, Shuffle Columns as Units, Shuffle Within Columns, Shuffle Single Column (all unselected).
- Stratified Sample: (unchecked)
- Shuffle blank cells in data: (unchecked)

Figure 9.13: Region Selection - Resampling Stats “&n” Syntax

9.4. Darwin's Plants: Strata in Separate Ranges

Figure 9.14 shows the results for one shuffled difference in heights: -1.68 inches:

5	Observed		Shuffled (within pots)			
6	Crossed	Selfed	Crossed	Selfed		
22	Pot IV	21	18	18	12	
23		22.1	12.8	18	23	
24		23	15.5	21	22.1	
25		12	18	12.8	15.5	
26	&&					
27	MEAN	20.2	17.57333	MEAN	18.04667	19.72667
28						
29	DIFFERENCE	2.626667		DIFFERENCE	-1.68	
30						

Figure 9.14: Shuffled Difference in Plant Heights

Repeat and Score for 1000 trials; Figure 9.15 shows the histogram of the results.

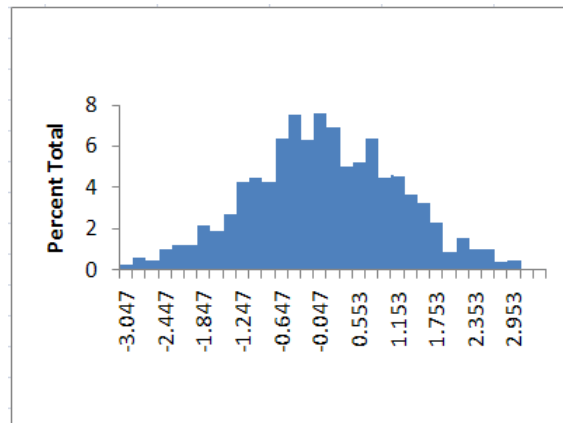


Figure 9.15: 1000 Trials Shuffled Plant Height Differences

If you use the COUNTIF function to find out how many of the resampled differences in means equalled or exceeded the observed value of 2.626667 you will find that it is a rare occurrence. In one set of 1000 trials, only 13 trials yielded a difference (in favor of the resampling counterpart to the “selfed” plants) this big (Figure 9.16).

9. STRATIFIED RESAMPLING

	A	B	C
1	3.32		13
2	3.173333	p-value:	0.013
3	3.026667		
4	2.933333		
5	2.92		
6	2.88		
7	2.866667		

Figure 9.16: Resampled p-value =0.013 for 1000 Trials

Chapter 10

Formula Iteration

The Resampling Stats in Excel add-in allows the user the option to iterate formulas. This feature provides powerful and interesting capabilities for Excel that are not normally available. Several of these capabilities are discussed in the sections that follow. Note: This chapter of the user's guide is not intended to serve as a rigorous mathematical treatment. The intention is to illustrate some potentially useful (and interesting) features of both Excel and the Resampling Stats in Excel add-in.

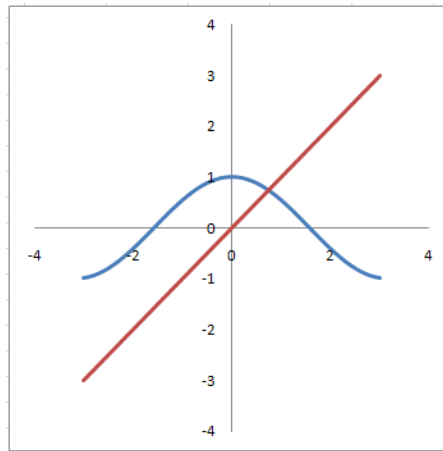
10.1 Iterative Solutions to Equations

The add-in may be used to iteratively solve equations through numerical means. An example of this is the equation $x = \cos x$. An attempt to solve this equation for x algebraically can be frustrating. One can see that the equation does have a solution by graphing the equations $y = x$ and $y = \cos x$ and noting that there is indeed an intersection (as plotted by the Resampling Stats in Excel add-in and the Excel chart feature in Figure 10.1)¹, but what is the value of the solution?

Turn off Auto-Range Select

In Resampling Stats, Auto-Range Select is on by default. This means that when you select "Resample" or "Shuffle" from either the Resampling Stats

¹See the *Note* at the end of this Chapter for detailed plotting instructions.

Figure 10.1: $x = \cos(x)$

toolbar or menu, the current cell and all of its contiguous neighbors will be selected as the input region automatically. In some instances, this is not the behavior you want, particularly when dealing with formula iteration problems. For the problems in this chapter, it is best to uncheck the Auto-Range Select option to disable this feature, and select the input range manually with the Resample or Shuffle dialog.

To find the solution to the $x = \cos x$ problem using the Resampling Stats in Excel add-in, simply do the following:

1. In cell B1 enter the formula: =COS(A1)
2. In cell A1 enter a “seed” or estimate of the root. 0.5 is a good value.
3. Select cell “B1” containing the formula and click on “R” on the Resampling Toolbar or select “Resample” from the Resampling menu.
4. Verify that cell B1 is the input range and select cell “A1” as the Top Left Cell of the output range.
5. Enter 1 for the number of output cells.
6. Click “OK.”
7. The Resampling Stats in Excel add-in will notify you that you are attempting to resample/iterate a formula. Since this is exactly what we want to do, click “OK” on this message box.

8. Select “RS,” or “Repeat and Score”.
9. Enter or select cell “A1” for the score cell.
10. Click “OK” (Note: 100 iterations is fine for this problem).

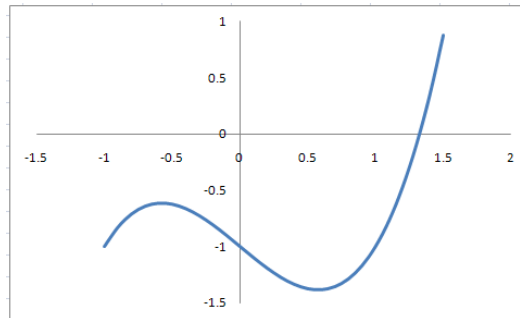
You'll notice that the number in cell A1 (and B1, for that matter) rapidly converges to the value 0.739085. This value is the root for the equation $x = \cos x$ to 6 decimal places (more places can be displayed by viewing the formula bar or making column A wider). View the Results sheet output to see that the solution is found on the 33rd iteration.

Now, without selecting or changing any menu items, we will use the current stored resampling procedure to solve the equation $x = \sin x + G$. In cell B1, enter the following: `=SIN(A1) + 0.25` (this equation overwrites the previous equation in cell B1). We can use the value in cell A1, 0.739085 (the root of the previous equation), as a first guess. Simply click “RS” or select “Repeat and Score” and “Redo” the model. Notice that the value converges to 1.17123 quickly. This value is the approximate root of our new equation.

Since we are iterating a formula, the Resampling Stats in Excel add-in will use the stored resampling procedure to iterate any formula we type in cell B1, with the condition that the initial guess of the root is resident in cell A1. We don't have to select “R” or “Resample” again unless we have clicked on “Reset” or we want to change the number of cells or the cell references containing our equations. This “reusability” feature is very useful when solving many problems of the same form.

10.2 Newton's Method

The iterative feature of the Resampling Stats in Excel add-in, when coupled with a tool from Calculus, can help provide real roots or solutions to nearly any function (assuming real roots exist). Newton's method uses the first derivative of a function and iteration to converge to a root. As an example, let $f(x) = x^3 - x - 1$. A graph of this function (as drawn by the Resampling Stats in Excel add-in and Excel's chart feature) is shown in Figure 10.2. It's readily apparent that only one real root exists since the function crosses the x axis at only one point.

Figure 10.2: $f(x) = x^3 - x - 1$

Newton's Method is represented by the equation:

$$x_{n+1} = x_n \frac{f(x_n)}{f'(x_n)}$$

where $f(x)$ represents the function (in this case, $f(x) = x^3 - x - 1$) and $f'(x)$ represents the derivative of the function (in this case, $f'(x) = 3x^2 - 1$). x_n is the initial guess or “seed” for the root of the function. The term $x_n + 1$ is the result of the first iteration, which then is used as the next value of x_n . After several iterations, the method converges to a real local root if one exists.

The Resampling Stats in Excel add-in can implement Newton's Method in a fashion similar to the equations in Section 10.1. To solve the example problem above, try the following procedure:

1. In cell B1 enter the formula: $=A1-(A1^3-A1-1)/(3*A1^2-1)$
2. In cell A1 enter a “seed” or estimate of the root. 1 is a good value.
3. Select cell “B1” containing the formula and click on “R” on the Resampling Toolbar or select “Resample” from the Resampling menu.
4. Verify that cell B1 is the input range and select cell “A1” as the “Top Left Cell” of the output range.
5. Enter 1 for the number of output cells.
6. Click “OK.”
7. Select “RS,” or “Repeat and Score” from the Resampling menu.

8. Enter or select cell “A1” for the score cell and click “OK”.

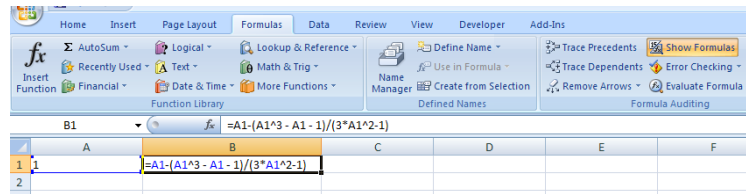


Figure 10.3: Newton’s Method in Excel

Figure 10.3 illustrates how the spreadsheet should look prior to step 3 above. Note that the Formulas|Show Formulas option has been selected so that you can see the correct formula in cell B1. Note also the “1” in cell A1 as the initial seed for the equation in B1.

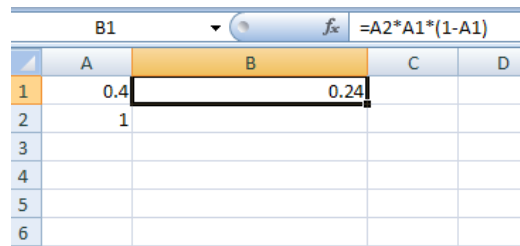
When the model is iterated by Repeat and Score, you should see the values in cells A1 and B1 converge rapidly to 1.324717957, which is the approximate real root of the function. It should be noted that Newton’s Method finds only one root at a time. If a function has more than one real root, the user should utilize a graph and choose initial seeds in the neighborhood of each root to find all real roots. For a more thorough treatment of Newton’s Method and its limitations, the user is referred to any basic Calculus text.

10.3 The Logistic Equation

The logistic equation, $x_{n+1} = rx_n(1 - x_n)$, has been used to model population dynamics and is discussed popularly in Gleick (1987) and Stewart (1992) and more formally by May (1976). In this model, x represents the population, expressed as a proportion of the maximum carrying capacity of the habitat; x_n represents the current population and x_{n+1} the population in the next period. The parameter r represents all factors that affect the population we are studying (food, competition, climate, etc.). The subscripts n and $n + 1$ indicate that the output, x_{n+1} , will be used as the input, x_n , during the next iteration.

Let’s place this model on a worksheet (displayed in Figure 10.4). Note the formula in cell B1: $=A2*A1*(1-A1)$. The initial value for r is in cell A2 and the initial value for x is in cell A1. At this point, follow the same procedure as in previous sections:

10. FORMULA ITERATION



	A	B	C	D
1	0.4	0.24		
2	1			
3				
4				
5				
6				

Figure 10.4: The Logistic Equation in Excel

1. Select cell B1.
2. Choose “R” from the toolbar, or “Resample” from the menu.
3. Choose cell A1 as the output cell.
4. Enter “1” for the number of output cells.
5. Click “OK.”
6. Choose “RS,” or “Repeat and Score,” and this time use 1000 iterations.
7. Select cell A1 for the score cell.
8. Click “OK” and run the model.

You will notice on the Results sheet that the values seem to be heading toward zero, but there is no convergence at this point. Is there a minimum set of conditions (the parameter r) for a population to sustain itself? Now enter 2 in cell A2 as a new value for r . Enter 0.4 in cell A1 and Repeat and Score again (redo the model). What happens? Interesting! We have a stable population of 0.5 of the maximum carrying capacity. Now enter 3 in cell A2 and repeat and score. The population appears to be fluctuating between two approximate values and these values seem to be converging slowly as shown in the graph (Figure 10.5).

Now change cell A2 to 3.1, repeat and score and view the output. The chart in Figure 10.6 clearly shows that the equation, instead of converging, is fluctuating between two values after an initial diverging pattern.

Finally, enter 3.7 into cell A2 and repeat and score. Notice the values fluctuate wildly! Figure 10.7 charts the behavior of the output values.

10.3. The Logistic Equation

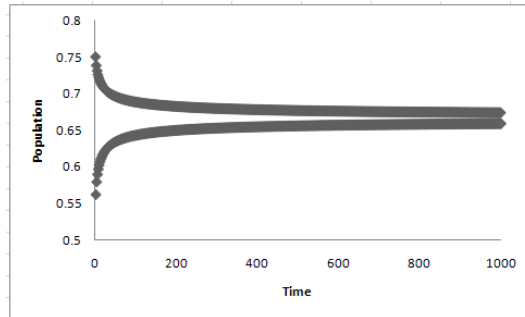


Figure 10.5: Logistic Equation: $r = 3$

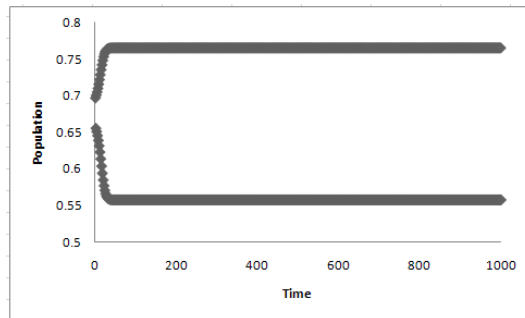


Figure 10.6: Logistic Equation: $r = 3.1$

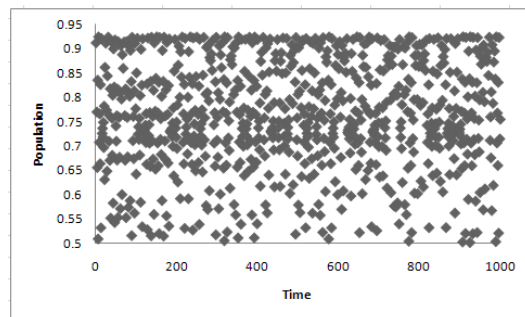


Figure 10.7: Logistic Equation: $r = 3.7$

10. FORMULA ITERATION

Users who are familiar with non-linear dynamics and chaos will recognize Figure 10.7 as a chaotic pattern of data. To have even more fun with our model, return to the original worksheet and click on “Reset” and add the formula $=A2+0.0005$ in cell B2 (shown in the Formula Bar) as demonstrated in Figure 10.8.

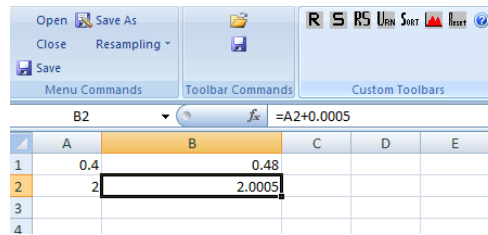


Figure 10.8: Logistic Equation: Increment r

If you haven't already clicked “Reset”, do so now in order for this model to execute correctly. Then, follow the steps below:

1. Select cells B1:B2.
2. Click on “R,” or “Resample.”
3. Select cell A1 for the output cell.
4. Enter “2” for the number of output cells (to update both cells A1 and A2).
5. Select “RS,” or “Repeat and Score.”
6. Select cell A1 as the score cell.
7. Enter “4000” for the number of iterations.
8. Click “OK.”

The values on the output sheet show a gradual rise for a period of time and then begin to demonstrate chaotic behavior. A chart showing the result of this model is in Figure 10.9. Notice the unusual appearance of the data. Again, those users familiar with chaos theory will immediately recognize the bifurcation diagram (although somewhat crudely drawn). Note: When using a system of two or more equations which are dependent on one another, you

should “stack” the system as presented in Figure 10.8. The add-in must be able to resample/iterate all equations during the same operation and “stacking” the equations in a contiguous range facilitates this operation. We will demonstrate this further in the next two sections.

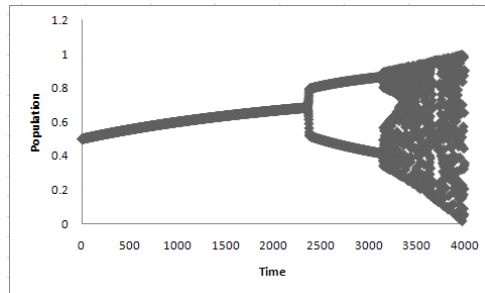


Figure 10.9: Logistic Chaos

10.4 Predator-Prey Relationships

In keeping with the previous section, we can expand the population model to include populations which are dependent on one another, such as a predator-prey relationship. Volterra’s predator-prey model was presented as a system of two differential equations (Braun, 1993) as follows:

$$\frac{dy}{dt} = ax - bxy \quad \text{and} \quad \frac{dx}{dt} = -cy + dxy$$

In this specific model, the populations of predators y and prey x change as a function of time. The prey population would ordinarily grow according to a Malthusian law of growth ax , but contacts with predators and prey subtract from this growth $-bxy$. The predator population would normally, in the absence of food (prey), be expected to decrease $-cy$. Predator-prey contact results in an increase in predator population growth $+dxy$. The parameters a , b , c , and d represent various factors, such as the environment, that can affect the two populations.

In order to model these equations using Excel and the Resampling Stats in Excel add-in, we must realize that these equations represent the change in the populations with respect to time. If we want the population numbers, a bit of “fudging” must be done. Excel is not able to handle the infinitesimal

10. FORMULA ITERATION

values represented by the differentials dx , dy , and dt , so we will consider them to be finitely small values. We will now rewrite the above equations a bit differently:

$$dx = (ax - bxy)dt \quad \text{and} \quad dy = (-cy + dxy)dt$$

The current population of each species is represented by y for predators and x for prey. The population for predators at the end of the next time increment is equal to the current population plus the change in the population (the differential) over that time interval. The same formula applies to the number of prey. These formulae are represented by the following equations:

$$x_{n+1} = x_n + dx \quad \text{and} \quad y_{n+1} = y_n + dy$$

Finally, substituting for dx and dy :

$$x_{n+1} = x_n + (ax_n - bx_ny_n)dt \quad \text{and} \quad y_{n+1} = y_n + (-cy_n + dx_ny_n)dt$$

Where, as before, the n and the $n + 1$ terms represent the population at the beginning and the end of the time increments respectively. Now, let's use Excel and the Resampling Stats in Excel add-in to simulate this system of equations.

	A	B	C	D	E	F	G	H
1				a	b	c	d	dt
2	x	5	=B2+(D2*B2-E2*B2*B3)*H2	0.7	0.5	0.3	0.2	0.02
3	y	1	=B3+(-F2*B3+G2*B2*B3)*H2					
4								
5								
6								

Figure 10.10: Population Dynamics with Excel

Using Excel's Formula View feature, Figure 10.10 shows the formulas typed in cells C2 and C3 for x and y respectively. For clarity, they are repeated below:

In cell C2: $=B2+(D2*B2-E2*B2*B3)*H2$

In cell C3: $=B3+(-F2*B3+G2*B2*B3)*H2$

Type in boundary values for the populations of prey and predator (5 and 1) and values for the constants and dt as shown. To run the model:

1. Select cells C2:C3
2. Select “R,” or “Resample.”
3. Enter or select cell B2 as the top cell of the output region.
4. Enter “2” as the number of output cells (to update both x and y).
5. Repeat and Score with cells B2 and B3 as score cells.
6. Enter 4000 iterations.
7. Click “OK.”

Figure 10.11 shows a very nice representation of this mythical predator-prey relationship. Notice the cyclic nature of both populations. While this model is certainly not indicative of any real-world system, it is nevertheless interesting and capable of demonstrating the rudiments of elementary population dynamics. More knowledgeable users may adapt this for their own use and develop far more complex models.

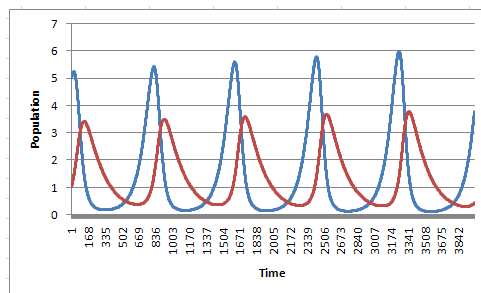


Figure 10.11: Predator-Prey Data

Figure 10.12 is a different view of the same Predator-Prey data using an X-Y (Scatter) Chart with the smooth line option. Essentially, this plot is equivalent to a parametric plot with respect to time and definitely demonstrates the cyclic nature of the two populations.

Feel free to experiment with different chart types, various boundary values of x and y , and different values for the parameters a , b , c , d , and dt . In the next section, we are going to take a look at a very interesting system of 3 relatively famous differential equations.

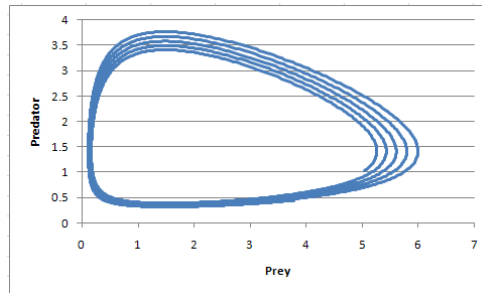


Figure 10.12: Predator-Prey Parametric Plot

10.5 The Lorenz “Butterfly” Equations

Ed Lorenz has been credited by many with starting the Chaos revolution/fad (Gleick, 1987) (Stewart, 1992). According to these popular accounts, Lorenz had created a very simplistic model of the weather based on 3 differential equations (Lorenz, 1999):

$$\frac{dx}{dt} = -10x + 10y$$

$$\frac{dy}{dt} = 28x - y - xz$$

$$\frac{dz}{dt} = -\frac{8}{3}z + xy$$

According to Gleick’s account, Lorenz had programmed these equations into his ancient Royal Bee computer. An interesting run of data caught his attention, so he stopped program execution and started that particular sequence again. Much to his surprise (after returning from a coffee break) he found an entirely different data set (representing atmospheric conditions) as output. At first, it was thought that the computer was at fault (vacuum tubes were somewhat prone to failure), but it was later realized that instead of entering the full precision of numbers as starting values, they had been truncated to 3 decimal places. It was not believed that such a small input error would result in such a wide variance in output. From this, the phrase “sensitive dependence on initial conditions” was born. In certain coupled or feedback systems, small (even tiny!) variances in initial conditions can quickly lead to completely unpredictable outcomes a short time later. The end result

was an explanation of why it is impossible to accurately predict the weather more than a few days in advance.

To model Lorenz’s system of equations, we’ll rely on the same method that we employed in the previous section and rewrite the equations as follows:

$$x_{n+1} = x_n + (-10x_n + 10y_n)dt$$

$$y_{n+1} = y_n + (28x_n - y_n - x_nz_n)dt$$

$$z_{n+1} = z_n + (-\frac{8}{3}x_n + x_ny_n)dt$$

Figure 10.13 illustrates the Lorenz system entered into an Excel worksheet:

	A	B	C	D
1				dt
2	x	8	=10*(B3-B2)*D2+B2	0.02
3	y	8	=(28*B2-B3-B2*B4)*D2+B3	
4	z	24	=(-(8/3)*B4+B2*B3)*D2+B4	
5				
6				

Figure 10.13: Lorenz Equations in Excel

As we did with the worksheet view in the last few sections, we are showing the formulas here to illustrate the correct entry.

The formulas in column C are as follows:

$$=10*(B3-B2)*D2+B2$$

$$=(28*B2-B3-B2*B4)*D2+B3$$

$$=(-(8/3)*B4+B2*B3)*D2+B4$$

The time increment, dt , is placed in cell D2. The initial values for x , y , and z are found in cells B2:B4. Also, note the “stacked” nature of the equations in contiguous cells in the same column. This facilitates the execution of models of this type. The parameters 10, 28, and $8/3$ were in Lorenz’s original equations (Stewart, 1992, p. 136).

To execute this model, take the following steps:

1. Select cells C2:C4 containing the Lorenz equations.

10. FORMULA ITERATION

2. Click “R,” or select “Resample” from the Resampling Menu.
3. Select or enter B2 as the top cell of the output region.
4. Enter 3 for the number of output cells (so we can include all three variables).
5. Click “OK.”
6. Click “RS,” or select “Repeat and Score” from the Resampling Menu.
7. Enter or select cells B2:B4 for the score cells.
8. Enter 4000 for the number of iterations.
9. Click “OK” and watch the numbers fly!

There are three columns of data on the Results sheet. The Lorenz system represents a 3D phase space, but selecting any two columns and charting those columns (using the X-Y (Scatter) option and the smooth line graph) will result in a planar projection of the data. Three different views of the data are possible, one of which is represented in Figure 10.14.

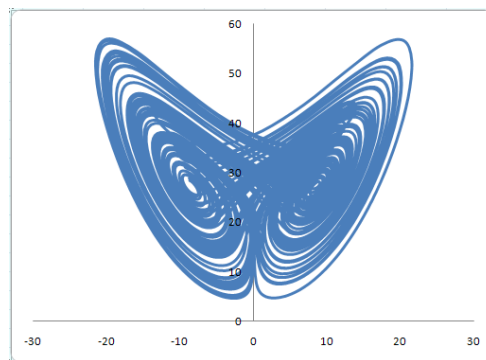


Figure 10.14: Lorenz “Butterfly” Attractor

The “Butterfly Effect” was coined to represent sensitive dependence on initial conditions. The flapping of the wings of a butterfly in Brazil could cause (or prevent) a tornado in Kansas the next week. How serendipitous that one projection of Lorenz’s equations actually resembles a butterfly – this is an amazing result from 3 relatively simple differential equations! One can see that while the model’s “weather” is constrained (it isn’t going to be 160

degrees Fahrenheit tomorrow!), the chaotic and unpredictable nature of the data is evident.

The user is urged to explore various parameters of these equations—perhaps even restructuring the worksheet so that easy entry of other parameters is possible. There are certain constant values in these equations that lead to stable conditions (such as the summer doldrums?). Can you find them? Also, experimenting with various charts and views of the data can be interesting.

Summary

The purpose of this chapter was to introduce the user to some of the more esoteric features of Excel and the Resampling Stats in Excel add-in. As stated previously, this chapter was not intended to be a mathematics text. It is hoped that many users who need to use these features will either possess the requisite mathematical knowledge or refer to appropriate textbooks.

This chapter hopefully has served to stir the imagination of the interested user. While not the focus of the chapter, Figures 10.1 and 10.2 represent an application of the Resampling Stats in Excel add-in for graphing simple functions. Many other applications, including Euler’s and the Runge-Kutta methods of solving differential equations through numerical means, are possible. Also, using conditional formatting, cellular automata might be explored using Excel and the Resampling Stats in Excel add-in. The possibilities are endless and limited only by the expertise and creativity of the user.

Note: The graphs in this chapter were created using the Chart Wizard and XY scatter plots, selecting the smooth lines sub-type. In this sub-type, the Series menu tab allows the user to pick the X and Y data series. If a plot such as the Lorenz equations (Figure 10.14) is chosen, then the data series will be whichever 2 of the 3 columns you select. Charts such as the Lorenz chart and the Parametric Population Chart (the concentric elliptical structures) are created in this fashion. You already have two columns of data from Repeat and Score and can choose whichever column you want for X and Y. For the simple graphs as in Figures 10.1 and 10.2, an additional column of X values was added to the Results sheet. Essentially, you need two columns of data to do an XY plot, so in order to do this, try the following steps:

1. Place a “seed” value (say “0”) in cell A2.

10. FORMULA ITERATION

2. In cell B1, put the formula: =COS(A2)
3. In cell B2, type: =A2 + 0.001
4. Select “Resample” via the “R” button or via the Resampling menu.
5. Select cells B1 and B2 as the input region.
6. Select cell A1 as the “Top Cell of the Output Region.”
7. Type 2 as the number of cells in the output region.
8. Click “OK.”
9. Repeat and Score on cells A1 and A2.

If you select A2 and A1 independently, choosing A2 first, then the X values will be in the first column of the output sheet. Call the Chart Wizard and XY Scatter Plots (smooth lines) and choose the correct columns for X and Y under the Series tab. This should create the chart.

Chapter 11

BCA Bootstrap

Confidence Intervals With Improved Coverage

An important question in evaluating confidence intervals is how well they perform - i.e., does a 95% confidence interval actually capture the parameter value 95% of the time? Of course, in any real world application, you won't know - you can only know in simulated situations where you draw a sample from a known population, construct a confidence interval from the sample, then record whether or not it includes the actual known population parameter.

The bootstrap confidence intervals that have been discussed to this point are termed “percentile intervals.” They perform well in many circumstances, but more complex bootstrap intervals have been developed that have, on balance, superior coverage properties. Their superior performance must be offset against their greater complexity and opaqueness (they lack one of resampling's great strengths - transparency and ease of understanding).

Perhaps the most widely-used such method is the “bias-corrected and accelerated” (“BCA”) interval. The algorithm is somewhat involved (see the [Appendix](#)), but the idea is to use the original sample and the bootstrap samples to estimate two quantities:

- a) Bias (the difference between the true value of the population parameter being estimated and the median of the sampling distribution of that parameter), and
- b) Acceleration (the degree to which the variance increases or decreases as the value of the population parameter increases).

These estimates are then used to derive an adjusted “bias-corrected and accelerated” (“BCA”) sampling distribution. For details, please see the Appendix to this chapter. For the original exposition of the BCA process, see *An Introduction to the Bootstrap*, Bradley Efron and Robert Tibshirani, CRC Press, 1993. The BCA method has been shown to have superior properties to both standard procedures, and the simpler percentile method. See *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley, Cambridge University Press, 1997.

Let’s see how to use Resampling Stats for BCA intervals through an example.

11.1 Process Temperature

Temperature readings are recorded for an industrial process and listed in Table 11.1:

431	450	431	453
481	449	441	476
460	482	472	465
421	452	451	430
458	446	466	476

Table 11.1: Process Temperature Readings

The average reading was 454.55. How much might this be in error, simply based on the random variation in the selection of the sample? Let’s apply the BCA method in Resampling Stats. Figure 11.1 shows the data in an Excel worksheet. The steps are as follows:

1. Select “Bootstrap BCA” from the Resampling Menu (Figure 11.2).
2. In the Resampling Stats Bootstrap BCA dialog, select the temperature data range (A2:A21) as the input range (Figure 11.3).
3. Select the cell containing the statistic of interest (in this case, the mean temperature) and then select the “BCA CI Output Cell” (Figure 11.4). The default values for replications and confidence level are sufficient for this example. Click “OK.”

11.1. Process Temperature

	A	B
1	data orig.	
2	431	
3	450	
4	431	
5	453	
6	481	
7	449	
8	441	
9	476	
10	460	
11	482	
12	472	
13	465	
14	421	
15	452	
16	451	
17	430	
18	458	
19	446	
20	466	
21	476	
22	454.55	
23		

Figure 11.1: Process Temperature Readings

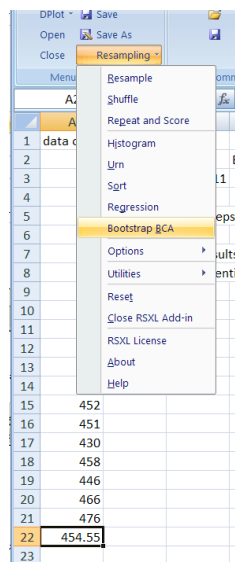


Figure 11.2: Bootstrap BCA from the Resampling Menu

11. BCA BOOTSTRAP

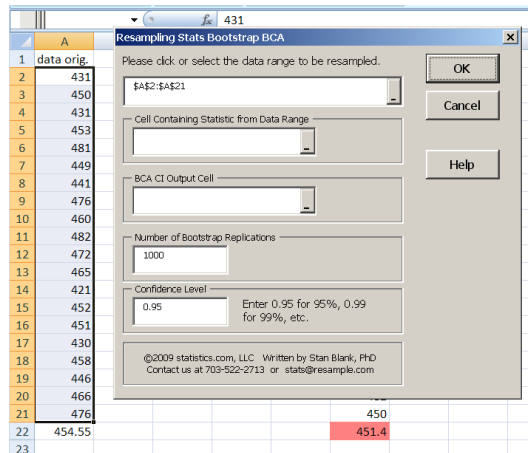


Figure 11.3: Bootstrap BCA Dialog

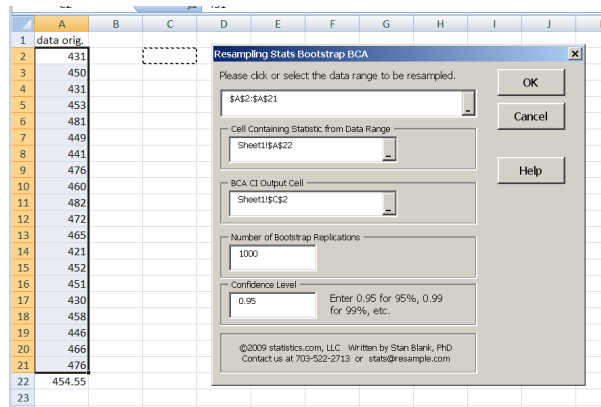


Figure 11.4: Completed Bootstrap BCA Dialog

- Figure 11.5 displays the result of the Bootstrap BCA procedure. The 95% CI is between 446.78 (BCA LCL) and 461.85 (BCA UCL).¹

Note: The Resampling Bootstrap BCA contains its own Repeat and Score procedure. You will not use the standard Repeat and Score methods for BCA Bootstrap problems.

¹LCL = Lower Confidence Level, UCL = Upper Confidence Level

11.2. Compare to Percentile Interval

	A	B	C	D	E
1	data orig.				
2	431		BCA LCL	BCA UCL	
3	450		446.7811	461.85	
4	431				
5	453		10,000 reps		
6	481				

Figure 11.5: BCA Confidence Interval

11.2 Compare to Percentile Interval

How does the bootstrap BCA procedure compare with the bootstrap percentile CI calculations we outlined in Chapter 3? Let's see:

1. Resample with replacement the temperatures into a new column and find the resample mean (Figure 11.6).

	A	B	C	D
1	data orig.		resampled	
2	431		472	
3	450		460	
4	431		421	
5	453		430	
6	481		476	
7	449		452	
8	441		482	
9	476		452	
10	460		482	
11	482		431	
12	472		431	
13	465		421	
14	421		441	
15	452		431	
16	451		458	
17	430		450	
18	458		430	
19	446		476	
20	466		482	
21	476		450	
22	454.55		451.4	
23				

Figure 11.6: Bootstrap Percentile CI

2. Repeat and Score the resampled mean temperature (cell E16 in Figure 11.6). Use 1000 iterations.

11. BCA BOOTSTRAP

The Results sheet from this simulation is shown in Figure 11.7.

	A	B	C	D	E	F
1			bootstrap percentile interval			
2	455.85		446.90	lower		
3	454.7		462.20	upper		
4	461.75					
5	452.6			10,000 reps		
6	458.15					
7	457.05		BCA LCL	BCA UCL		
8	452.35		446.781068	461.85		
9	462.85					
10	450.1					
11	457.1					
12	451.95					
13	453.25					
14	451.45					

Figure 11.7: Bootstrap Percentile Method Interval

Both the Bootstrap BCA Bootstrap and the Percentile method yield similar results; the BCA interval is slightly wider.

Appendix to Chapter 11

If you are interested in the details of how the BCA interval is computed, a good step-by-step algorithm is described in *Data Analysis by Resampling*, by Cliff Lunneborg, Duxbury Press (Brooks/Cole), 2000, p. 164:

1. Compute the plug-in estimate, t , from the sample distribution, x .
2. Compute the n omitted-case estimates, $t_{(-i)}$.
3. Compute the average of the omitted-case estimates, $t_{(\cdot)}$.
4. Compute the n influence statistics,

$$\frac{U_i}{n} = \left(\frac{n-1}{n} \right) (t_{(\cdot)} - t_{(-i)})$$

5. Compute the acceleration estimate,

$$\hat{a} = \frac{\sum_{i=1}^n \left(\frac{U_i}{n} \right)^3}{6 \left[\sum_{i=1}^n \left(\frac{U_i}{n} \right)^2 \right]^{\frac{3}{2}}}$$

6. Form \hat{X} from one or more copies of x .
7. Set $\hat{\theta}$ equal to the t of step 1.
8. Draw a bootstrap sample, x_b^* , from \hat{X} .
9. Compute the estimate, t_b^* , from the bootstrap sample.
10. Repeat steps 8 and 9 a total of B times, forming the bootstrap sampling distribution of t_b^* .
11. Compute $\hat{\pi}$, the proportion of the bootstrap sampling distribution smaller than $\hat{\theta}$.
12. Select α , the confidence level for the $(1 - 2\alpha)100\%$ CI.
13. Use the tabled distribution of the standard normal distribution to determine $z_{[\hat{\theta}]}$, $z_{[\alpha]}$, and $z_{[1-\alpha]}$. These are the z -scores that cut off the lower $\hat{\pi}$, α , and $(1 - \alpha)$ proportions of the standard normal distribution.

11. BCA BOOTSTRAP

14. Use the acceleration estimate from step 5 and the z -scores from step 13 to compute

$$\hat{z}_{lo} = z_{[\hat{\pi}]} - \frac{(z_{[1-\alpha]} - z_{[\hat{\pi}]})}{1 + \hat{\alpha}(z_{[1-\alpha]} - z_{[\hat{\pi}]})}$$

and

$$\hat{z}_{up} = z_{[\hat{\pi}]} + \frac{(z_{[\hat{\pi}]} - z_{[\alpha]})}{1 + \hat{\alpha}(z_{[\hat{\pi}]} - z_{[\alpha]})}$$

15. Use the tabled distribution of the standard normal distribution to find q_{lo} and q_{up} , the proportions of the distribution falling below the two z -scores computed at step 14, \hat{z}_{lo} and \hat{z}_{up} .

16. Use the two step 15 proportions and the B of step 10 to compute

$$lo = int[q_{lo} \times (B + 1)]$$

and

$$up = (B + 1) - int[(1 - q_{up}) \times (B + 1)]$$

17. Sort the bootstrap sampling distribution of step 10 in order from the smallest element, $t_{[1]}^*$, to the largest element, $t_{[B]}^*$.

18. Use the integers computed in step 16 to select $t_{[lo]}^*$ and $t_{[up]}^*$ from the sorted bootstrap sampling distribution. The selected $t_{[lo]}^*$ and $t_{[up]}^*$ are the q_{lo} and q_{up} quantiles of the bootstrap sampling distribution, just as \hat{z}_{lo} and \hat{z}_{up} were the q_{lo} and q_{up} quantiles of the standard normal distribution.

19. Steps 15 through 18 describe the translation of the CI limits from the standard normal to the bootstrap sampling distribution. The lower and upper limits of our $(1 - 2\alpha)100\%$ BCA confidence interval are given by $t_{[lo]}^*$ and $t_{[up]}^*$.

The phases of the algorithm are summarized here:

1. Compute the acceleration estimate from a set of jackknifed² estimates (steps 2-5)

²In a jackknife procedure, the first observation in the sample is removed, and the estimate is recalculated. This procedure is then repeated with the second observation removed, the third observation, etc. The set of n jackknife estimates are then available to work with.

2. Produce a bootstrap sampling distribution from the original sample (steps 6-10)
3. Compute the bias estimate (step 11)
4. Incorporate z-scores from the normal distribution, along with the results of the prior steps, to derive the BCA confidence interval.

Chapter 12

Resampling Stats Operations

Resampling Stats Add-in Functions and Syntax

For a quick-start introduction to the add-in, we suggest you read through [Chapter 1](#), then review the [Resample](#), [Shuffle](#) and [Repeat and Score](#) items below. These are the heart of all resampling operations.

Auto-Range Selection

When you want to select a range for resampling or shuffling, simply place your cursor in any cell in the range and select “Resample” or “Shuffle” – Resampling Stats will automatically select the entire contiguous range. There are times when this is not suitable – you can simply reselect the range manually, or turn off auto-range selection in “Resampling>Options” as shown in [Figure 12.1](#).

Auto-Reset

Normally, Resampling Stats resets when you select a new “Resample” or “Shuffle” routine after a Repeat and Score operation is completed. Otherwise you are likely to accumulate an overhead of irrelevant resampling operations if you forget to reset as you progress through a work session. (Typically, these accumulated resampling operations do not affect the accuracy of your current procedure, they just slow things down.) If you want to use the output of a Re-

12. RESAMPLING STATS OPERATIONS

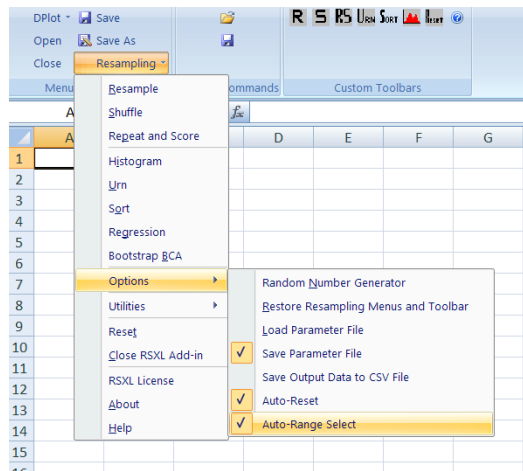


Figure 12.1: Auto-Range Select

peat & Score operation as an input to a second Resample/Shuffle and Repeat & Score problem, make sure Auto-Reset is NOT checked in the Resampling > Options menu (see Figure 12.1).

Auto-Numbering

Resampling Stats provides an option to insert a column that sequentially numbers the rows of a data set (useful if you want to verify where data rows go during resampling). To use this option:

1. Go to the new Resampling > Utilities menu and click on the “Add Sequential Numbering” option (no check mark will appear; it simply sets an internal variable to “True”).
2. Click on a cell in your data set and select “Resample” or “Shuffle”
3. Click on “Yes” to the message box that appears. “No” will cancel sequential numbering and the add-in will proceed normally.

You must click on the “Add Sequential Numbering” every time you want to use this option. It only works once per click.

Also, the sequential numbering is useful only for “Rows as Units” resampling/shuffling. Otherwise, the numbers will be resampled/shuffled with the data.

Custom Functions

Resampling Stats can iterate your custom written procedures, provided they can be expressed as a function in Excel. You can write custom functions in Visual Basic using Excel’s Visual Basic Editor. Some Excel users post libraries of custom functions on the web - search for “Excel UDF” (user defined functions).

Example

In Excel 2007, select “Developer > Visual Basic”¹ and then use “Insert > Module” to start with a new blank VBA code sheet. Here is a simple example that calculates the n^{th} root of x .

In the new code module sheet, type the following:

```
Function findroot(x as double, n as double)
    Application.Volatile
    If n = 0 Then
        msgbox n MUST be unequal to zero!
        Exit Function
    End If
    findroot = x^(1/n)
End Function
```

Excel will automatically add the “End Function” when you complete the first line of the function. Excel will also capitalize and provide coloration for words it knows (like Function, Application.Volatile, If, etc.).

¹If the Developer Menu is not visible, you must enable it from the Office Button > Excel Options > Popular > Show Developer Tab in Ribbon check box.

Important:

You **MUST** have the lines “Application.Volatile” and “findroot = x^(1/n)” in order for the function to work properly. The Application.Volatile allows Repeat and Score to work with the function, while the “findroot=” returns the value of the function (the name of the function **MUST** be used here in order to return a value).

Now, select “Debug” > “Compile” then “File” > “Close and Return to Microsoft Excel” to display the Excel worksheet. You can now use your new custom function as you would use any Excel function by:

- Typing “=findroot(7,5)” (to calculate the 5th root of 7) into a worksheet cell (you can use any pair of numbers, of course, as long as the function has a defined root). Use Excel’s Insert Function button and then select “ALL” functions. Your Findroot function should be visible in the right hand pane.

Escape

Pressing the Esc key will halt a resampling simulation, allowing you to terminate it or have it continue.

File Operations

See [Saving and Opening Files and Storing Simulation Parameters](#)

Formulas (Resampling Formulas)

The add-in will warn you if a formula is being resampled or shuffled.

If you are dealing with statistical data that just happens to contain formulas, and wish to resample or shuffle the data for statistical purposes (the main topic of this user guide), then you should convert those values to pure data first. Copy and “Paste Special” the data to a new range. When the Paste Special dialog comes up (Figure 12.2), click the “Values” button:

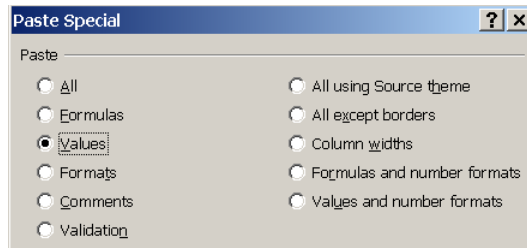


Figure 12.2: Paste Special Dialog: Values

It is possible to use the add-in to solve equations (such as the Lorenz “butterfly” equations) by resampling the equation system, selecting the appropriate output range (which serves as input for the equations), clicking “OK” to the dialog box, and to the subsequent message box, and then using Repeat and Score.

If you would like to repeatedly iterate a formula (formula iteration), you can do so by using Repeat and Score with an appropriate score cell without doing any resampling or shuffling operation. See [Chapter 10](#) on formula iteration.

Histogram

The Histogram feature produces a frequency histogram of a specified range. In resampling operations, you would typically produce a histogram from the output or Results sheet data. Select the “Histogram” button (Figure 12.3) from the Resampling Stats toolbar (or, from the menu, “Add-ins” > “Resampling” > “Histogram”):



Figure 12.3: The Histogram Button

Next, in the histogram dialog box, specify the input for the histogram, which is the output from the resampling experiments.

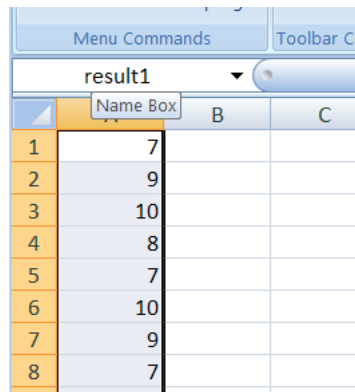


Figure 12.4: Selecting a Named Range

Hint:

An easy way to select the input for the histogram is to click on the top cell of the data range from you want to use to create a histogram. Resampling Stats will select all data in that column until it encounters an empty cell in the data. The histogram feature will then use that selection as the input. (The Data input range field in your histogram dialog must be active before you do this; click in it to make it active.)

Another method for entering the input range into the histogram dialog box is to use the Named Range feature. The Resampling Stats add-in always gives names to the score cell output in the Results sheet. The output of the first score cell is named “result1,” the output of the second score cell “result2,” etc. To work with these range names, simply select the “Name Box” (Figure 12.4) and select the named range. Selecting “result1” will also select the output from the first score cell automatically.

Another method for entering the Data Input Range in the histogram feature is to simply type the name of the data range you wish to use in creating the histogram. This method is illustrated in Figure 12.5 using the data range named “result1”, which corresponds to cells A1:A1000.

For the Top Left Cell for Freq. Table specify the top left cell in any empty area.

You have several options for how Resampling Stats will determine the bins to be used in drawing the histogram. (Histograms have bars whose height on

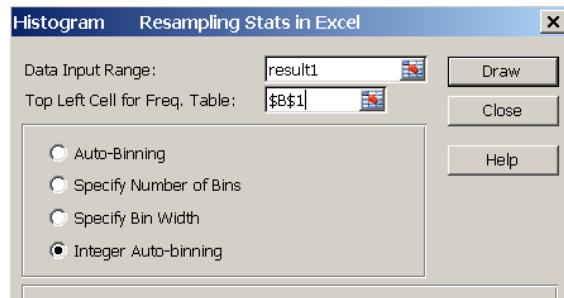


Figure 12.5: Entering a Range Name

the y-axis depends on the number of values that fall in each bin, or range on the x-axis. All bins have the same width – for example 0-4, 5-9, 10-14, etc.).

- **Auto-Binning:** Resampling Stats will determine the number of bins and where they are located.
- **Specify Number of Bins:** You tell Resampling Stats how many bins to use.
- **Specify Bin Width:** You specify the width for the bins.
- **Integer Auto-Binning:** You force the bin centers to be integers (used typically when you have integer-only data).

The histogram dialog box is shown in Figure 12.6.

The result will look something like Figure 12.7.

You can move the graph around by clicking and holding somewhere inside the graph area, and dragging.

You also have the option of displaying counts or percents on the y-axis. With 1000 repetitions, here is how the histogram of rolling 2 dice looks choosing “Counts” (Figure 12.8) and then choosing “Percent” (Figure 12.9):

The “Distribution Chart” option produces a line graph by connecting the mid-points of the bars on the Histogram (Percent) chart (Figure 12.10).

The “Cumulative Frequency” option produces a line graph that is a cumulative version of the “Distribution Chart” graph (Figure 12.11).

12. RESAMPLING STATS OPERATIONS

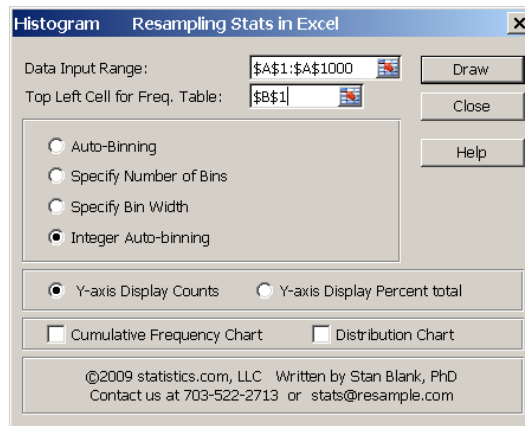


Figure 12.6: The Histogram Dialog Box

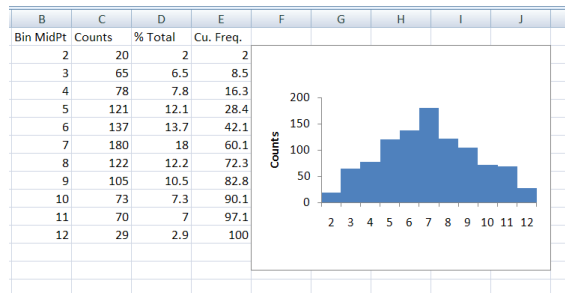


Figure 12.7: Histogram: Rolling 2 Dice

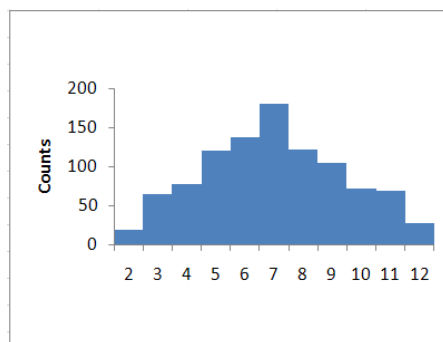


Figure 12.8: Histogram Counts

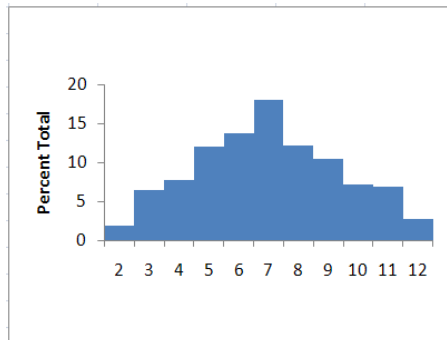


Figure 12.9: Histogram Percent

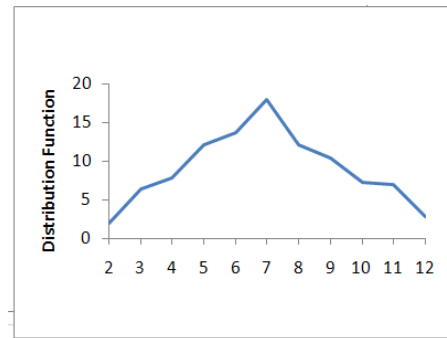


Figure 12.10: Distribution Chart

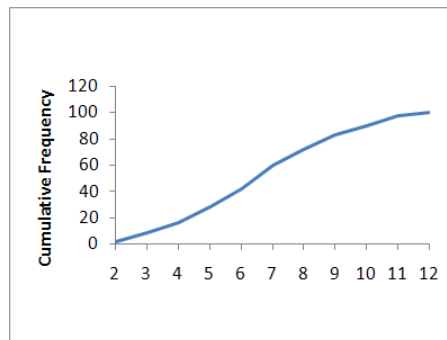


Figure 12.11: Cumulative Frequency

Licensing Procedure

The first time you start the Resampling Stats Add-in, you will see the dialog in Figure 12.12.

12. RESAMPLING STATS OPERATIONS

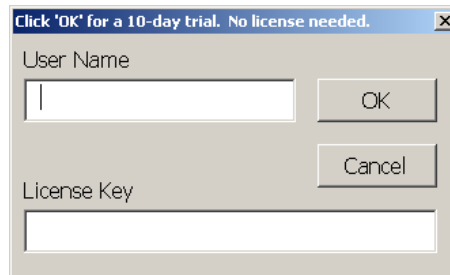


Figure 12.12: First Run Licensing Dialog

If you click “OK”, you will enable a 10-day trial of the Resampling Stats add-in. If you purchased a license, you will have been sent a username and license key that looks something like the following:

```
ima user  
281E-04157D9-9D5B-555
```

This license key is time sensitive and should be entered within a few days of receipt. You would enter this username and license key as shown in Figure 12.13.

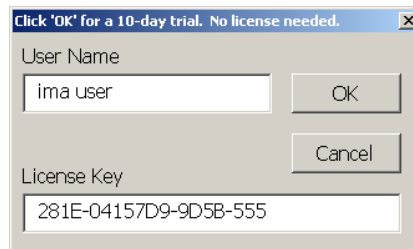


Figure 12.13: Username and License Key Entry

If the username and license key are valid, you will see something similar to Figure 12.14, which includes the expiration date of the software.

At any time you may check your license status by choosing the “Add-ins > Resampling > RSXL License” menu.

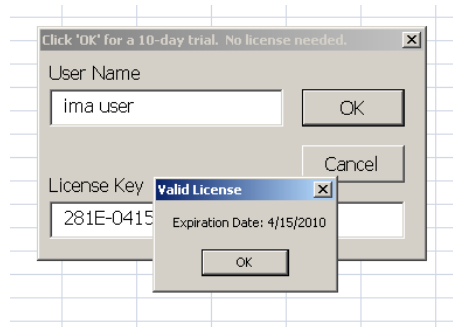


Figure 12.14: Successful Username and License Key Entry

Macros

Any user-created macro that automatically updates each time the worksheet is recalculated can be iterated with Repeat and Score. To ensure that the macro will update on recalculation, it should be declared “Application.Volatile.”

Maximum Number of Trials and Score Cell Limits

The maximum number of trials is determined by the maximum number of rows in an Excel worksheet. The maximum number of score cells is determined by the number of columns in an Excel worksheet. For Excel 2003 and below, the row limit is 65536 and the column limit is 256. The Resampling Stats add-in limits Excel 2003 (and versions below 2003) workbooks to 65000 trials with a maximum of 256 score cells.

If you are using an Excel 2007 workbook, the worksheet has a row limit of 1048576 and a column limit of 16384. The Resampling Stats add-in limits the maximum number of iterations in Excel 2007 workbooks to 1000000 trials. You can also have up to 3000 score cells in Excel 2007 with the caveat that the product of the score cells and iterations can not exceed 100 million. This is a memory limitation in Excel 2007.

Menu and Toolbar for the Resampling Stats Add-in

You can reach the Resampling Stats add-in menu via the Add-Ins ribbon on the main Excel menu, once you have opened the add-in. The key functions are duplicated on a floating toolbar (see [Toolbar and Excel Ribbons](#)). Figure 12.15 shows the Resampling Menu.

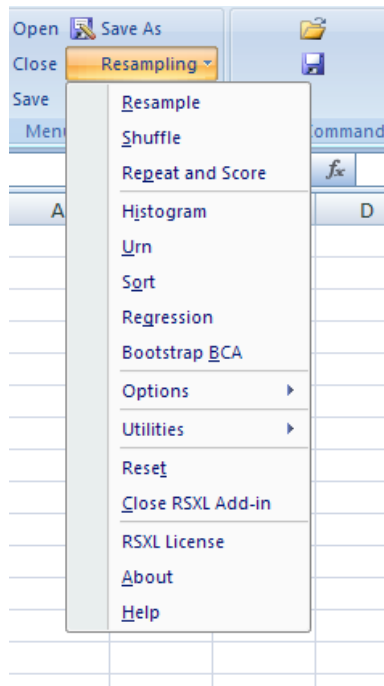


Figure 12.15: Resampling Menu

You can also bring up the menu by right-clicking on a worksheet (provided the add-in has been opened).

Resample (“R” on the Resampling Stats Toolbar)

Takes a random sample with replacement from a selected range and places the resample wherever you specify. For details, see [Resample and Shuffle Options](#).

Shuffle (“S” on the Resampling Stats Toolbar)

Takes a random sample without replacement from a selected range and places the resample wherever you specify. For details see [Resample and Shuffle Options](#).

Repeat and Score (“RS” on the Resampling Stats Toolbar)

Repeats for a specified number of times those resampling and shuffling operations that have been conducted on the worksheet since the last reset, and, for each iteration (repeat), records on the Results sheet the value(s) in specified score cell(s). See the more detailed [Repeat and Score](#) section, below.

Histogram (On the Resampling Stats Toolbar)

The Histogram feature is accessed from the Resampling Stats menu or toolbar. The toolbar icon is displayed in [Figure 12.16](#).



Figure 12.16: The Histogram Button

The Histogram feature produces a frequency histogram and table of the specified range. In resampling operations this is typically used with the results range. See the more detailed [Histogram](#) section above.

Urn (“Urn” on the Resampling Stats Toolbar)

Provides a facility for easily creating a range with specified quantities of values – say, 200 “0’s” and 15 “1’s.” See the more detailed [Urn](#) section, below.

Sort (“Sort” on the Resampling Stats Toolbar)

Sorts a specified range as part of a resampling operation. (Excel’s SORT function can sort a range, but will not repeat the sort automatically as part of

a repeated resampling operation.) See the more detailed [Sort](#) section, below.

Regression

The Regression feature performs a multiple linear regression within the Resampling Stats add-in. (Excel's Data Analysis Toolpak does multiple linear regression, but will not repeat the regression automatically as part of a repeated resampling operation.) See the [Correlation and Regression](#) chapter for more details on using the Resampling Stats Regression feature as well as the [Regression](#) section below.

Bootstrap BCA

The Bootstrap BCA feature implements the “bias-corrected and accelerated” (BCA) method for determining a confidence interval based on a data sample. The Bootstrap BCA procedure relies on both resampling and analytic methods. See the [Bootstrap BCA](#) chapter for further details.

Options

The Options menu is visible in [Figure 12.1](#) and contains the following items:

- **Random Number Generator**

Allows the user to either use the default random number seed generator or to select a specific random number seed if desired. Choosing a specific random number seed allows the user the ability to replicate simulations for the purposes of testing simulation validity.

- **Restore Resampling Menus and Toolbar**

If for any reason the Resampling Toolbar or right-click menu disappears, this selection will attempt to restore them.

- **Load Parameters File**

If a parameter file has been saved for a particular simulation, you may load the file from this menu item. Loading a parameter file allows for an “instant” replay of a stored simulation. See “Saving and Opening Files and Keeping the Same Simulation” below.

- **Save Parameters File**

This option is enabled by default. When using the “Save” or “Save As” feature in Excel while a Resampling operation is currently open (a simulation has been run, but Reset has NOT been clicked), a parameter file with the same name as the open workbook and an “.rxl” extension is saved. See [Saving and Opening Files and Storing Simulation Parameters](#) below.

- **Save Output Data to a CSV File**

If you wish to save a column or columns of data to a CSV (comma-separated values) file, use this option. The method is simple: select the *top* cell of all columns of data that you wish to save and then click this menu item. You will be prompted for a filename. Type in a filename and the CSV file will be saved in your current working directory.

- **Auto-Reset**

This option is enabled by default. After a Repeat and Score has been executed, any new Resample or Shuffle operation will generate a Reset. The Resampling Stats add-in can “remember” up to 100 resampling or shuffling simulations. The purpose of Auto-Reset is to keep current resampling and/or shuffling operations from being hindered by previous simulations. To disable this feature, click on the “Auto-Reset” menu item.

- **Auto-Range Select**

This option is enabled by default. Auto-Range Select allows you to select a single cell within a contiguous range of cells and the Resample

or Shuffle dialog will automatically select the entire region as input. You may disable this feature by clicking the “Auto-Range Select” menu item once.

Utilities

The Utilities menu contains a single sub-menu at this point.

- **Add Sequential Numbering**

See [Auto-Numbering](#)

- **Reset (“Reset” on the Resampling Stats Toolbar)**

Erases the add-in’s memory of prior resampling operations on the worksheet. For more details see the [Reset](#) section, below.

- **Close RSXL Add-in**

This menu selection allows the user to gracefully exit the Resampling Stats for Excel add-in. This option will NOT save a parameter file. Your current workbook will still be intact, however.

- **RSXL License**

Displays the current license status and expiration date. You also have the option to enter a new username and license key if desired. See [Licensing](#).

- **About**

This item displays the current version, copyright, and contact information.

- **Help**

Displays the current User Guide.

Multi-stage Resampling and Shuffling

See [Resample and Shuffle Options](#).

Opening and Closing the Resampling Stats Add-in

Resamplings Stats can be opened from the Start Menu in Windows, whether or not Excel is running. You can also open the add-in just as if you were opening a file in Excel. With Excel running go to the Office Button > Open button and select the add-in. It will be located in the “Recent Files” section or in the installation directory (the default path is C:\Program Files\Resampling Stats for Excel 2007). The name of the file will be similar to “Resample-for-Excel-2007.xla.”

You can close the add-in (without closing Excel) from the Resampling Stats menu. You can also close the add-in by closing all open workbooks and then clicking “Close” one more time. The add-in will inform you that there is no workbook open and ask you if you want to close the add-in. Clicking “Yes” will close the Resampling Stats add-in.

If the add-in has been loaded via the Office Button > Excel Options > Add-ins > Manage Add-ins dialog (not recommended), then unchecking the box associated with the add-in will unload it. Only one version of the Resampling Stats add-in may be open at a time. Attempting to start a second instance of the Resampling Stats add-in will open a dialog allowing the user to exit gracefully from the conflicting situation.

Opening Files

See [Saving and Opening Files and Storing Simulation Parameters](#).

Random Number Generator

The Resampling Stats add-in is equipped with its own random number generator. You can set the seed to the random number generator by selecting Options on the Resampling Stats menu. Otherwise, the seed is set from the

computer's clock. The random number generator in Resampling Stats utilizes a linear congruential algorithm (see <http://www.resample.com>) for more information).

You can generate random numbers from a variety of distributions, as listed below.

These functions can be entered from Excel's "Insert Function" menu; the easiest way to locate the RSXL functions is to select "All" and then scroll down to the functions that begin with "RSXL." All of these functions can be iterated with Repeat & Score – it is not necessary to resample them. Figure 12.17 displays some of the RSXL random number functions.

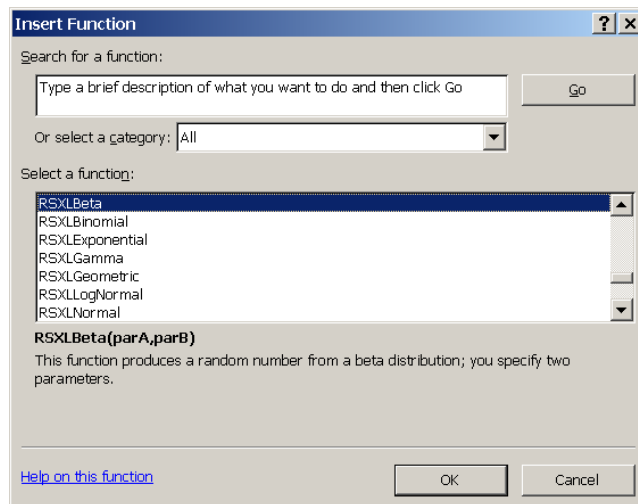


Figure 12.17: RSXL Random Number Distribution Functions

RSXLBeta

This function produces a random number from a beta distribution. The random numbers can take on values between 0 and 1; the shape of the distribution depends on the two parameters you specify.

RSXLBinomial

This function produces a random integer from a binomial distribution that can take on values between 0 and n . The random variate, x , is the number of

successes in an experiment with n Bernoulli (yes/no) trials. You specify the probability of a success (p) and the number of trials (n).

RSXLExponential

This function produces a random number from an exponential distribution, which can take on values between 0 and ∞ . One application of the exponential distribution is to model the distribution of the “time to next event” when an event has a constant probability of happening in each successive (tiny) increment of time. You specify the mean.

RSXLGamma

This function produces a random number from a Gamma distribution, where you specify an integer shape parameter. (This distribution is also called an Erlang gamma distribution, to distinguish it from more generalized gamma distributions where the scale as well as the shape can be specified.)

RSXLGeometric

This function produces a random integer from a Geometric distribution, which models the number of Bernoulli (yes/no) trials that are conducted to get to the first success (yes). You specify one parameter – the probability of a success (constant on all trials).

RSXLLognormal

This function produces a random number from a Lognormal distribution, for which you specify the mean and standard deviation. The lognormal distribution is skewed right (has a long tail to the right) and is bounded on the left by zero. If you transform the data by taking their logs, the resulting transformed distribution will have an approximately normal distribution.

RSXLNormal

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

RSXLPareto

This function produces a random number from a Pareto distribution, where you specify a location parameter (a) and a shape parameter (c). A typical application of the Pareto distribution is to model percentages of complaints accounted for by percent of customers (e.g. 80% of the complaints come from 20% of the customers).

RSXLPoisson

This function produces a random number from a Poisson distribution, where you specify the mean. A typical use of the Poisson distribution is to describe the number of events happening in a time period (say, incoming phone calls at a call center) where the average rate of the event remains constant.

RSXLRand

This function produces a random number from a uniform distribution between 0 and 1, inclusive. You do not need to provide any information (arguments) for this function. It is analogous to Excel's RAND function, except that it uses the Resampling Stats random number generator.

RSXLRandbetween

This function produces a random integer from a uniform distribution between the high and low values you specify. It is analogous to Excel's Randbetween function, except that it uses the Resampling Stats random number generator.

RSXLUniform

This function produces a random number from a uniform distribution between the high and low values you specify. It differs from RSXLRandbetween in that this number need not be an integer.

RSXLWeibull

This function produces a random number from a Weibull distribution. You specify two parameters: scale and shape. A Weibull distribution is typically used to model survival times, or time to failure.

Redo

To redo a simulation while the worksheet is open, click on “Repeat and Score” after you have already run the simulation. The dialog in Figure 12.18 will appear.

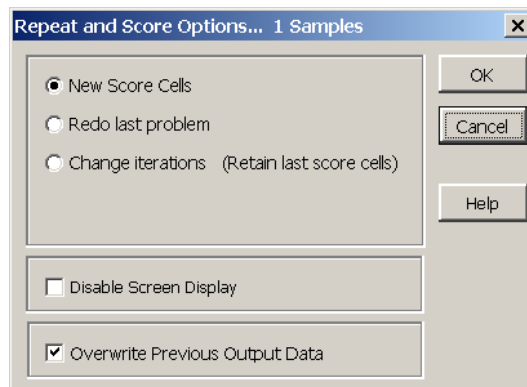


Figure 12.18: REDO Dialog Box

Select “Redo Last Problem” if you want to run the problem with no changes. Select “Change Iterations” if you want to alter the number of repeats, but keep everything else the same. Select “New Score Cells” to re-run the same simulation, but scoring different cells. See the [Repeat and Score](#) section below for information on the Disable Screen Display and Overwrite Previous Output Data options.

Regression

If you try to resample Excel’s regression routine (or any of the Excel tools reached via the Data > Data Analysis menu), you will find that it does not work. The regression is not iterated (repeated) for each resample.

12. RESAMPLING STATS OPERATIONS

In order to repeat the regression analysis for each resample, you need to use the regression option reached via the Resampling Stats menu (Add-ins > Resampling > Regression). (The Resampling Regression feature utilizes Excel’s built-in regression function.) The dialog in Figure 12.19 will appear:

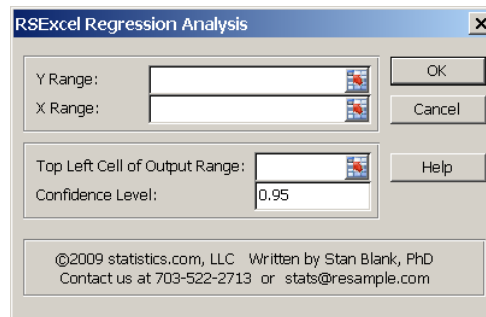


Figure 12.19: RSXL Regression Dialog Box

You will need to enter the Y Range (this is the resampled Y range), the X Range (again, the resampled X range), and the top left cell of the location where you want the output to appear. The routine also asks for a Confidence Level. This is for a conventional (non-resampling) calculation of a confidence limit for the data – as long as some value is in there, this can be ignored. You’ll be determining confidence limits via resampling, and will not likely have any use for individual formula-based confidence limits for each resample.

Only one regression model may be iterated. (Experienced Excel users can use Excel’s LINEST to iterate more than one regression model in the same resampling procedure.)

After you select “OK,” the regression output will appear in the designated location, looking like Figure 12.20.

You will typically be interested in the values in the “Coefficients” column (“Intercept,” “X Variable” and “X Variable” above), and also perhaps the “R Square” value (which estimates the extent of variance explained by the regression). These would be the cells to Repeat and Score.

Please see the chapter on **Correlation and Regression** for more details.

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.90379488							
R Square	0.81684519							
Adjusted R Square	0.76451524							
Standard Error	10.6617881							
Observations	10							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	2	3548.78392	1774.39196	15.609517	0.00262945			
Residual	7	795.716082	113.673726					
Total	9	4344.5						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 0.95%</i>	<i>Upper 0.95%</i>
Intercept	-10.589319	13.1711069	-0.8039809	0.44784614	-41.734038	20.5553998	-10.751828	-10.42681
X Variable 1	0.83588407	0.56441632	1.48097077	0.18216354	-0.4987484	2.17051658	0.82892014	0.842848
X Variable 2	1.39543586	0.70572636	1.97730445	0.08853289	-0.2733418	3.06421352	1.3867284	1.40414332

Figure 12.20: Regression Output

Repeat and Score

Use this feature after you have done your resample or shuffle operation(s) and calculated a statistic or other estimate based on your resample. For example, say that you have resampled your data from the A column into the B column, and calculated some statistic from the B column and entered the formula for calculating that statistic in C9. Select “C9,” and then select “Repeat And Score” from either the Resampling Stats menu or toolbar. C9 will be identified as the cell to be scored; you also need to enter the number of repetitions (iterations) you want to perform. The Repeat and Score Dialog Box is represented in Figure 12.21.

When you click “OK”, Resampling Stats will repeatedly perform the resampling or shuffling operation, each time recalculating the statistic in C9 and placing each successive value in column A on the Results sheet.

Multiple Score Cells

Within this one dialog box, you can select for scoring multiple cells in the same or different worksheets in the same file. You can score up to 256 cells in Excel 2003 (and versions below) and up to 3000 cells in Excel 2007.² If you select more than one cell to score, the second cell selected will have its results

²Subject to the limitation that the product of the score cells and iterations can not exceed 100 million due to memory limitations in Excel.

12. RESAMPLING STATS OPERATIONS

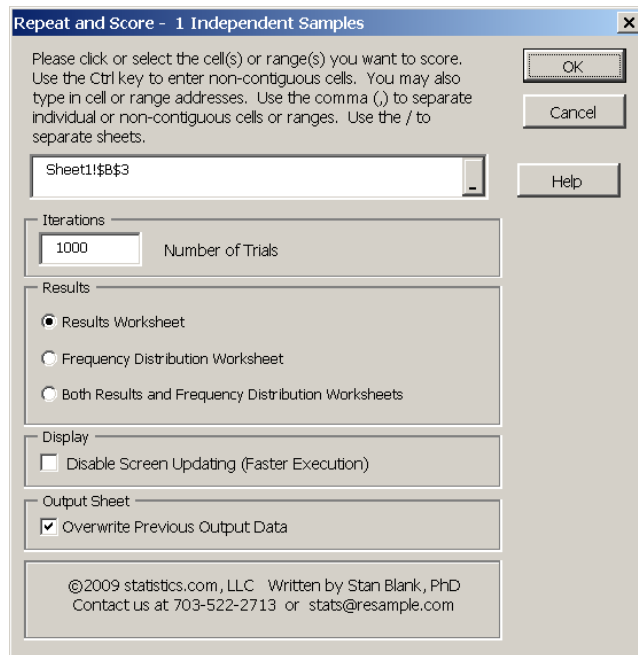


Figure 12.21: Repeat and Score Dialog Box

scored to column B in the Results sheet, the third cell to column C in the Results sheet, and so on.

Results Worksheet

The Results Worksheet option produces Repeat and Score output to the Results sheet only.

Frequency Distribution Worksheet

Choosing this option produces Repeat and Score output on the RSXL_Freq worksheet only. The output is in the form of frequencies of each possible score cell value. The output is unsorted.

Both Results and Frequency Distribution Worksheets

Choosing this option produces Repeat and Score output in both the Results worksheets and the Frequency Distribution worksheet.

Disable Screen Updating

Selecting “Disable Screen Updating” will marginally increase the simulation speed at the expense of not being able to view the resample or shuffle operations during the Repeat and Score. On the author’s computer, a simple dice simulation with 1 score cell and 100,000 trials took 78 seconds with screen updating enabled and 70 seconds with screen updating disabled.

Overwrite Previous Output Data

If this option is selected, every new Repeat and Score operation will write data starting with column A on the Results sheet (assuming one of the Results Worksheet options is selected). This will overwrite any existing data in those columns. If this option is not selected “old” Repeat and Score output will be retained and new output will begin in the first empty column to the right of the existing data. This option works for both the Results worksheet and the RSXL_Freq worksheet.

Resample and Shuffle Options

Resample takes a random sample with replacement from a specified range and puts it where you specify. In other words, after each element is selected randomly and recorded for the resample, it is replaced in the original data range so that it might be selected again. This is continued until the specified size for the resample is reached.

Here’s an example of the numbers 1-10, resampled (Figure 12.22). Notice that 2, 8, and 9 are duplicates and 3, 4, and 6 do not appear. This is normal in resampling *with replacement*.

Shuffle takes a random sample *without replacement* from a specified range and puts it where you specify. In other words, after each element is selected randomly for the resample, it is not replaced in the original data range and

12. RESAMPLING STATS OPERATIONS

	A	B	C
1	Original	Resampled	
2		1	2
3		2	7
4		3	5
5		4	2
6		5	9
7		6	8
8		7	1
9		8	10
10		9	8
11		10	9
12			

Figure 12.22: Resampled Data in Column B

therefore is unavailable to be selected again. This is continued until the specified size for the resample is reached (which must be less than or equal to the size of the original sample). If the resample size is equal to the size of the original data then SHUFFLE amounts to simply rearranging (shuffling) the original data.

Here's an example of the numbers 1-10, shuffled (Figure 12.23). Each number appears exactly once in the shuffled output.

	A	B	C
1	Original	Shuffled	
2		1	10
3		2	2
4		3	1
5		4	6
6		5	7
7		6	4
8		7	8
9		8	3
10		9	5
11		10	9
12			
13			

Figure 12.23: Shuffled Data in Column B

Single Row or Column Resampling or Shuffling

If you select a single column or row and then select “Resample” or “Shuffle,” a dialog box like Figure 12.24 pops up:

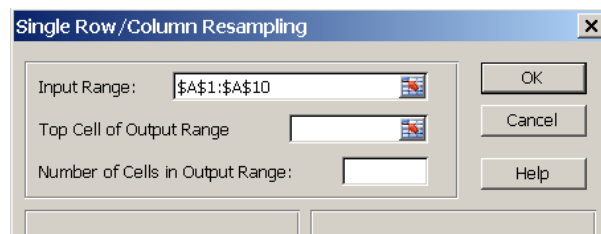


Figure 12.24: Single Row/Column Resampling

In the dialog box you specify the range you want to resample (or shuffle, if that is what you selected), where you want to place the resample (you specify the top cell only), and the number of cells in the output range (i.e. the size of the resample).

An efficient way to work is as follows:

1. On the worksheet, select the range you want to resample or shuffle.
2. Select “Resample” or “Shuffle” – the range you selected in step 1 will be entered as the input range.
3. On the worksheet, click on the top cell of the range where you want the output to go – this cell will be entered in the “Output Range” cell in the dialog box.
4. Type in the value you want for “Number of Cells in Output Range” (this is generally the original sample size).

Of course, you can also type the desired ranges, instead of selecting them in the worksheet.

Note that the output range need not contain the same number of values as the input range. Figure 12.25 displays the numbers 1-10 resampled, with a resample size of 15 (i.e. 15 cells in the output range).

And here are the same numbers shuffled (Figure 12.26), with only 6 cells in the output range:

12. RESAMPLING STATS OPERATIONS

	A	B
1	1	3
2	2	8
3	3	9
4	4	2
5	5	6
6	6	8
7	7	9
8	8	6
9	9	6
10	10	7
11		1
12		9
13		3
14		3
15		9
16		

Figure 12.25: Custom Resampled Output Range

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

Figure 12.26: Custom Shuffled Output Range

Of course, the output of a shuffle cannot be more than the number of elements in the original input range. Shuffling is the same thing as sampling randomly *without replacement*, and once the shuffled output reaches the same size as the original sample you will have run out of data to shuffle.

Matrix Resampling or Shuffling

If you select a matrix – more than one row or column – several choices present themselves as shown in Figure 12.27.

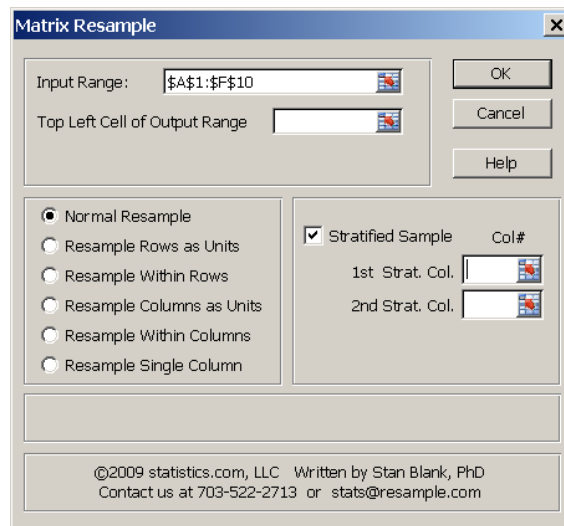


Figure 12.27: Matrix Resampling or Shuffling Dialog

Normal Matrix Resampling or Shuffling

This option takes all the data in the matrix and resamples or shuffles it to a matrix of similar configuration. For example, Figure 12.28 shows the result of a normal shuffle of the data in A1:C6, with the resample placed in the range E1:G6.

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

Figure 12.28: Normal Matrix Shuffle

Note in Figure 12.28 how the first column in the original data has 6 values, the second column 4, and the last column 3. The shuffled resample has the same structure the blank cells in the matrix stay in the same relative location in the resample. A value from anywhere in the original data might end up anywhere in the resample.

If there are blank cells in your data and you want to include the blank cells in the resampled or shuffled output, check the “Include Blank Cells in Data” checkbox.³ This would result in the blank cells being interspersed throughout the resampled or shuffled output as if they were normal data cells.

Resample or Shuffle Rows as Units

This option treats rows as units, so that the values in the row remain together in the same order in any resampling or shuffling. Here in Figure 12.29 is a shuffling of the same data as above with the Rows as Units option checked.

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

Figure 12.29: Shuffle Rows as Units

By checking the Custom Sample Size box, you are given the option of changing the sample size (i.e. you would end up with a different number of rows in the resample than in the original data). Of course, if you are using the Shuffle option, the number of rows in the resample cannot exceed the number of rows in the original data set.

Resample or Shuffle Within Rows

This option conducts shuffling or resampling by row, as if there were a barrier between rows that values could not cross. Here in Figure 12.30 is a shuffling of the same data with the Within Rows option checked.

³This option appears in the Resampling Dialog Box only when there are blank cells in your data input range.

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

Figure 12.30: Shuffle Within Rows

Resample or Shuffle Columns as Units

This option treats columns as units, so that the values in the column remain together in the same order in any resampling or shuffling. Here in Figure 12.31 is a shuffling of the same data as above with the Columns as Units option checked.

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

Figure 12.31: Shuffle Columns as Units

By checking the Custom Sample Size box, you are given the option of changing the resample size (i.e. you would end up with a different number of columns in the resample than in the original data). Of course, if you are using the Shuffle option, the number of columns in the resample cannot exceed the number of columns in the original data set.

Resample or Shuffle Within Columns

This option conducts shuffling or resampling by column, as if there were a barrier between columns that values could not cross. Figure 12.32 shows a shuffling of the same data with the Within Columns option checked.

12. RESAMPLING STATS OPERATIONS

	A	B	C	D	E	F	G
1	1	2	3		1	22	33
2	1	2	3		1	2	3
3	1	22	33		11	2	3
4	11	22			1	22	
5	11				11		
6	11				11		
7							

Figure 12.32: Shuffle Within Columns

Resample or Shuffle a Single Column

There may be instances where you would need to resample or shuffle a single column in a matrix of data, leaving all other columns fixed or unaltered. This option allows you to do that. When you select “Resample (or “Shuffle”) Single Column,” a range reference box will appear in the right pane of the dialog box. Select this “Column” reference box and then click in either the top cell of the column of data you wish to resample/shuffle or choose the entire column by clicking on the “A”, “B”, “C”, etc. at the top of the worksheet. The output of this operation looks something like Figure 12.33. You can see that the second column of data (highlighted in the output range) has been shuffled, but columns 1 and 3 have remained the same.

	A	B	C	D	E	F	G	I
1	1	6	9		1	2		9
2	2	5	8		2	6		8
3	3	4	7		3	5		7
4	4	3	6		4	4		6
5	5	2	5		5	1		5
6	6	1	4		6	3		4
7								
8								

Figure 12.33: Shuffle a Single Column

Multistage Resampling and Shuffling

You can include multiple resampling and shuffling operations in the same worksheet, and you can select the output of a Resample or Shuffle operation as input for a new Resample or Shuffle, and do this to multiple levels. Resampling and Shuffling can be done across multiple worksheets in the same file.

Stratified Resampling and Shuffling

See the section on Stratified Resampling and Shuffling for additional options.

Resample and Resampling (the Different Meanings of the Terms)

The terms “resampling” and “resample” are used with slightly differing meanings in different contexts.

1. The menu and toolbar item “Resample” means to sample with replacement (also called “bootstrap” sampling). See [Resample and Shuffle Options](#).
2. More generally, the term resampling is used to mean the process of repeated simulated sampling (with or without replacement).
3. The term “resample” used as a noun means the simulated sample drawn during a simulation.

Reset

Normally, Resampling Stats remembers all resampling operations done on your worksheet up to the time you decide to Repeat and Score, then repeats those resampling operations for each Repeat and Score. Thus, if you resample A1:A10 to B1:B10 then discover that you meant to resample A1:B11 to A1:B11 and do it over again, the RSXL add-in will actually do BOTH resamplings during each iteration.

Resampling Stats will erase its memory of resampling and shuffling operations when:

1. You click on “Reset”.
2. You open a previously saved workbook.
3. An error occurs in the operation of the add-in.

4. You have done a prior Repeat and Score operation and begin a new Resample or Shuffle operation (IF the “Auto-Reset” option is checked in the “Add-Ins > Resampling > Options” menu).

Opening a new workbook via the “Office Button > New” method will not reset the add-in. Also, adding a new worksheet to an existing workbook that has a “live” resampling operation will not reset the add-in.

Reset should be used whenever a model is finished and before starting a new, unrelated model. This saves system resources and speeds things up considerably. If you want to try a new model on an existing worksheet but don’t want to lose the old simulation, you should save the worksheet and parameters (make sure that the “Save Parameter File” box in the Resampling Options Menu is checked) before you use reset. See ‘Saving and Opening Files and Storing Simulation Parameters’ for further information.

Saving and Opening Files and Storing Simulation Parameters

You can save the parameters of a simulation by saving the worksheet after running the resampling operation; this allows you to reopen the file and run the simulation again without going through all the steps of defining the problem again. Resampling Stats will remind you that the parameters have been saved (Figure 12.34), and give you the name of the file that contains these parameters. In order for the simulation parameters to be saved, you should use the “Save” or “Save As” buttons in the Add-ins ribbon (these should be visible next to the Resampling menu). These “Save” and “Save As” buttons were created by the Resampling Stats add-in and will insure that a parameter file is created.

You may also use the traditional Office Button “Save” or “Save As” menus, although this is *not* recommended. You can *not* use the Ctrl-S keyboard shortcut or the save file shortcut icon to the right of the Office Button to save simulation parameters.

The name of the simulation parameters file will be the same as the main worksheet, except ending in .rxl.

If you want to re-open a workbook that has an associated saved parameters file, you should use the “Open” menu next to the “Resampling” menu. This

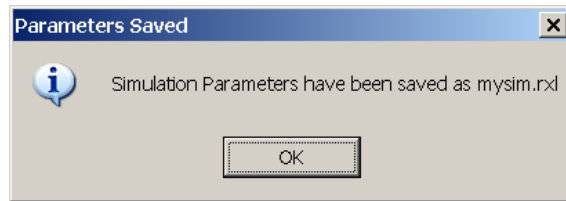


Figure 12.34: Saving Simulation Parameters

“Open” menu button was created by the Resampling Stats add-in and insures that Excel will search in the correct directory for the parameter file. If the parameter file is found and loaded successfully, you’ll get a message like Figure 12.35.

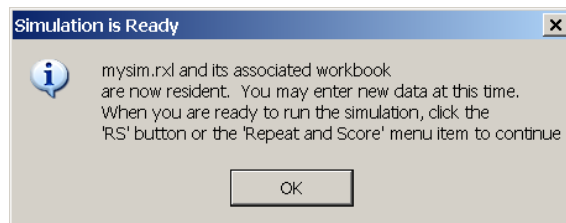


Figure 12.35: Restoring Simulation Parameters

Note: The most common error in loading the simulation parameter file is a failure to utilize the “Open” button next to the Resampling menu.

If you know that a simulation parameter file exists, but was not loaded when you re-opened a saved workbook, you may attempt to load the parameter file from the “Add-Ins > Resample > Options > Load Parameter” menu item. If the parameter file is found, it will then load. If not, then you will be informed that the parameter file does not exist. This may be due to the fact that Excel is looking in the wrong directory. To remedy this situation, *close* the *current* workbook. Then, using the “Add-Ins” menu, click on the “Open” button near the Resampling menu. This action will utilize the Resampling Stats “Open” function and insure that both the workbook and its associated parameter file will be loaded properly.

You can turn off the “Save Parameter File” option by unchecking this option in the Resampling Stats menu (under Options).

Adding New Data

You can add new rows of data below the current data, as long as you don't add columns. The Add-In will automatically incorporate it into the current model and extend the resampling that you have already defined to cover the new data.

Running the Simulation Again (With or Without Modifications)

Once the parameter file is loaded, you may click on "RS" or "Repeat and Score." The Redo dialog box should appear . You now have 3 options:

1. Use "Change Score Cell" to change the designation of the cell(s) you want to score to the output sheet.
2. Use "Redo" to run the simulation with no change in the score cells, or the number of iterations (repeats).
3. Use "Change Iterations" to increase or decrease the number of repeats.

See also: [Redo](#)

Score

See [Repeat and Score](#)

Shuffle

See [Resample and Shuffle Options](#)

Sort

Sort can be reached via the Resampling Stats menu or toolbar. Sort lets you sort a range in such a way that the sort operation is repeated with each

iteration of the simulation. If you use Excel's Sort facility, the sort operation will not be repeated for each resample.

In order to demonstrate the Resampling Stats add-in Sort feature, let's use an example. In Figure 12.36 the original data are in column A. They have been resampled into column B. Then they have been sorted into column C. This sort will be redone each time the data are resampled.

	A	B	C
1	1	6	1
2	2	7	2
3	3	2	2
4	4	6	5
5	5	2	5
6	6	5	6
7	7	1	6
8	8	5	7
9			

Figure 12.36: Sorting Resampled Data

You have several options with Sort, as shown in Figure 12.37.

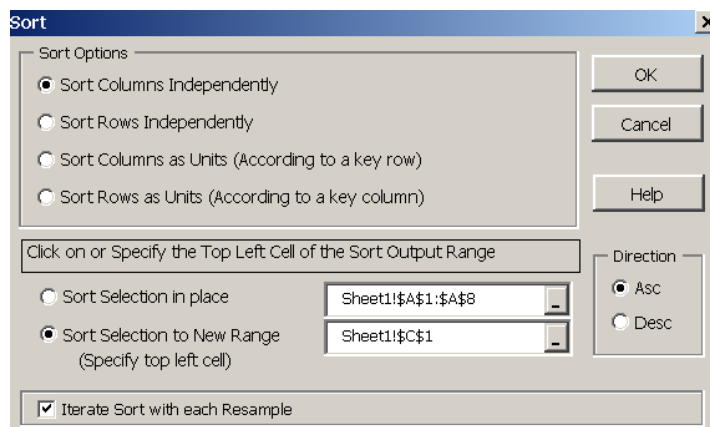


Figure 12.37: Sort Dialog Box

If you select a matrix of multiple columns and rows, the sort operation could proceed in several ways.

- **Sort Columns Independently** will sort by column, treating each column as a separate unit.
- **Sort Rows Independently** will sort by row, treating each row as a separate unit.
- **Sort Columns as Units** will keep each column together as a unit and all columns will be sorted according to the order of a single “key” row that you specify (you will be given a prompt for the key row if you specify this option).
- **Sort Rows as Units** will keep each row together as a unit and all rows will be sorted according to the order of a single “key” column that you specify (you will be given a prompt for the key column if you specify this option).
- **Sort Selection in Place** will cause the sorted data to overwrite the original data that were selected for the sort.
- **Sort Selection to New Range** will place the sorted data in a new range; you will be given a field in which you can specify this new location.
- **Iterate Sort with each Resample** will cause the sort to be repeated with each resample. This option is on by default.

Stratified Resampling and Shuffling

Often you will want to confine the resampling or shuffling operation to strata (clusters or subgroups) within the overall sample. Doing so eliminates the “noise” caused by the variability that occurs from stratum to stratum. Resampling Stats has several tools that let you do this.

Resampling or Shuffling Within Rows (or Columns)

This option creates barriers between each row (or column) which resampled or shuffled values cannot cross. Thus, when shuffling within rows, for example, the values in a given row get shuffled around in that row and cannot end up in other rows. See [Resample and Shuffle Options](#) in this section.

&n Syntax

When it is not convenient to record the values in a stratum to a single row or column, you can define a matrix as a stratum using the &n syntax.

Place “&1” at the top left of the first stratum to be shuffled, “&2” at the top left of the second stratum, and so on, placing && at the bottom left of the last stratum as shown in Figure 12.38.

	A	B	C	D	E	F
5	Observed					Shuffled (within pots)
6		Crossed	Selfed		Crossed	Selfed
7		&1			&1	
8	Pot I	23.5	17.4			
9		12	20.4			
10		21	20			
11		&2			&2	
12	Pot II	22	20			
13		19.2	18.4			
14		21.5	18.6			
15		&3			&3	
16	Pot III	22.2	18.6			
17		20.4	15.2			
18		18.3	16.5			
19		21.6	18			
20		23.2	16.2			
21		&4			&4	
22	Pot IV	21	18			
23		22.1	12.8			
24		23	15.5			
25		12	18			
26		&&				

Figure 12.38: Resampling Stats “&n” Syntax

Note that you should enter a parallel set of &1, &2, etc. (but no &&) in the region where you plan to place the shuffled or resampled output.

You can then select the entire data set, select “Shuffle” (or “Resample”), and Resampling Stats will automatically confine the shuffling (or resampling) operation you select within the bounds of each stratum.

Important Note: When you use the &n syntax for stratified resampling or shuffling, for the output range you cannot select merely the top left cell. You must select the *entire* destination range that contains the &1, &2, etc. (i.e. the range where you intend to place the shuffled or resampled data). The destination range **MUST** be identical in size to the input range.

Figure 12.39 illustrates the highlighted input range and the outlined output range for the Matrix Shuffle dialog. Note that both ranges are identical in size.

12. RESAMPLING STATS OPERATIONS

The output range contains the &1, &2, etc. cells in the first column.

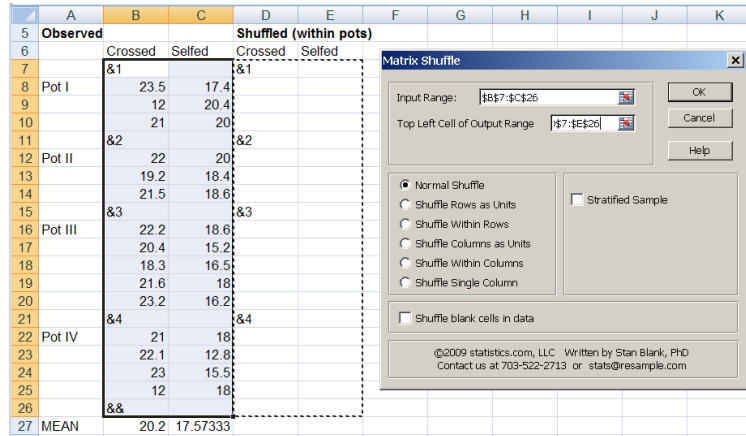


Figure 12.39: Using the Shuffle Dialog with the “&n” Syntax

Non-contiguous Strata

Multiple input ranges, whether contiguous or not, can be specified in the dialog box for Resample or Shuffle. They need to be separated by commas, and there needs to be a separate Top Left Cell for Output Range for each input range.

Toolbar and Excel Ribbons

The main resampling functions can be accessed from the Resampling Stats toolbar; see the section on [Menu and Toolbar for the Resampling Stats Add-in](#) for complete explanations of these functions. The Resampling Stats toolbar is displayed in Figure 12.40.

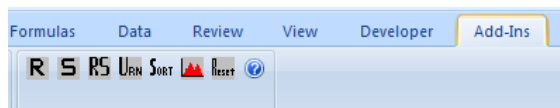


Figure 12.40: The Resampling Toolbar

Excel itself comes with several Ribbons⁴, one or more of which can be displayed by default directly below the Excel menu. Figure 12.41 shows the Ribbon associated with Add-ins and displays the Resampling Menu and its associated “Open”, “Save”, “Save As”, and “Close” buttons.

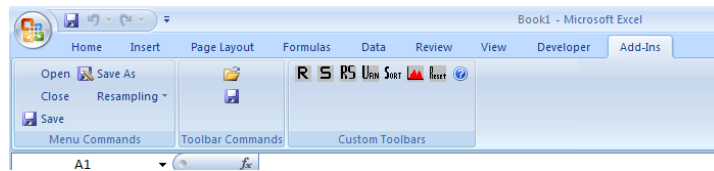


Figure 12.41: The Excel Add-Ins Ribbon

Note: Remember to use the file buttons in the Add-ins Ribbon next to the Resampling menu to properly save and restore parameter files.

Showing the Resampling Stats Toolbar

You can recover a “disappeared” Resampling Stats toolbar from the Resampling Stats menu. Select “Add-Ins” > “Resampling” > ”Options” and click the “Restore Resampling Menus and Toolbar” item.

Urn

Urn facilitates the entry of categorical data. It is the computer equivalent of filling an urn (or a box or hat) with slips of paper so that you can draw samples from the urn. The slips of paper might be marked white and black, “1,” “2,” and “3,” or in some other fashion that you specify.

You can create an urn in two ways – via a dialog box, or by specifying its contents on the worksheet.

- **Dialog Box Option**

If you want to use the Urn dialog box option, click the “Urn” button on the Resampling Stats toolbar or select the “Urn” option in the Resampling menu. The Urn Type dialog will appear as in Figure 12.42.

⁴Ribbons are new in Excel 2007. You can think of them as analogous to toolbars.

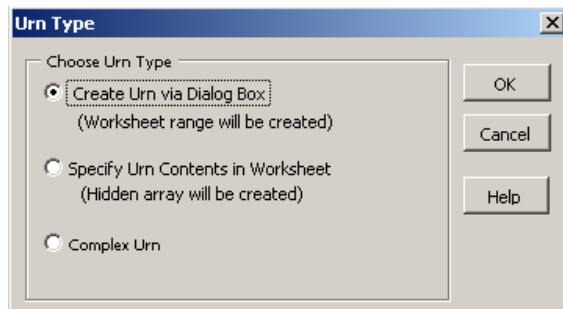


Figure 12.42: Choosing the Urn Type

Select “Create Urn Via Dialog Box” and click “OK.” With the dialog box option, you specify up to five values (alpha or numeric) and how many you want of each (Figure 12.43).

To specify the output range, you can click in “Top Cell of Output Range” field then click on the worksheet in the top cell of your desired output range. Or you may simply type in the desired range.

You can be creative and use formulas to enter values. By selecting “Remove Formulas (Retain Cell Values),” the values will be kept and the urn will contain no formulas.

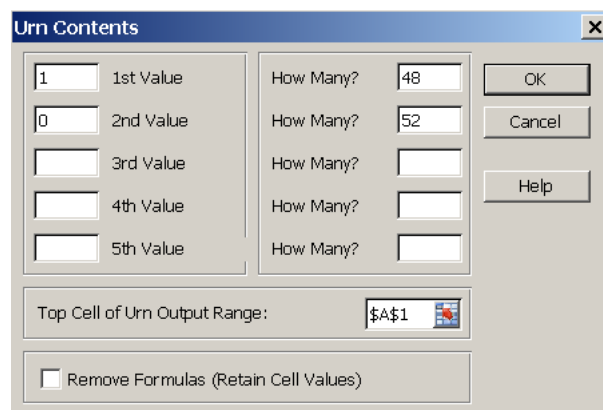


Figure 12.43: Creating an Urn with 48 “1’s” and 52 “2’s”

- **Worksheet Urn**

If you have more than five values represented in the urn, you need to specify the urn contents on the worksheet itself. (Another way to look at this is that you are creating an empirical probability distribution.) Let's say you have the following information in cells A1:B6 to indicate that you will want an urn with 1 red, 5 blacks, 21 greens, etc. (Figure 12.44)

	A	B	C
1	1	red	
2	5	black	
3	21	green	
4	4	white	
5	5	gray	
6	5	yellow	
7			

Figure 12.44: Specifying the Urn Contents on the Worksheet

Note that the quantity needed precedes the value itself.

Next, specify this range in the Worksheet Urn dialog (Figure 12.45).

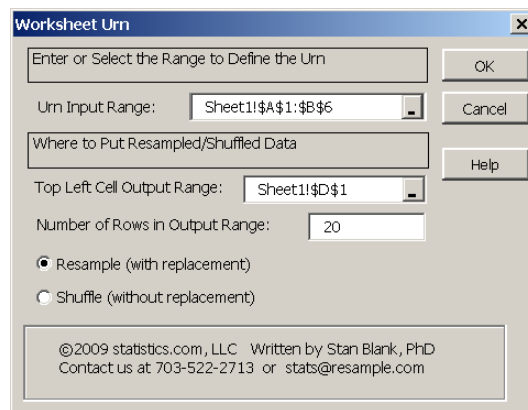


Figure 12.45: Worksheet Urn dialog box

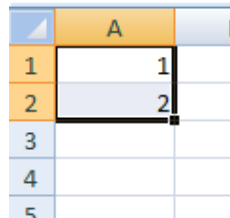
Note that when you create an urn based on the worksheet, the contents are resampled or shuffled as part of the same operation. The original

urn exists only in memory as a hidden array – only the resampled or shuffled urn appears on the worksheet. This allows the resampling of tiny probabilities, such as 99999 reds, and 1 black.

Useful Excel Functions

Autofill

Excel's Autofill function lets you fill in adjacent cells simply by selecting several cells that establish the series pattern, then dragging down. Suppose you select the values "1" and "2" in cells A1:A2 as in Figure 12.46.

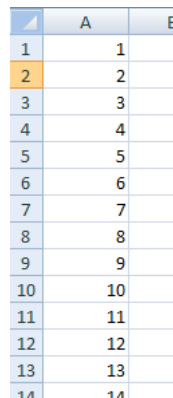


The image shows a portion of an Excel spreadsheet. The columns are labeled A and B, and the rows are numbered 1 through 5. Cells A1 and A2 are selected, containing the values 1 and 2 respectively. A small black square (the fill handle) is visible at the bottom right corner of cell A2.

	A	B
1	1	
2	2	
3		
4		
5		

Figure 12.46: Autofill

Click precisely on the little square at the lower right of A2, and drag down. The result of this operation is shown in Figure 12.47.



The image shows the same Excel spreadsheet as in Figure 12.46, but now the values 1 and 2 from cells A1 and A2 have been autofilled down to cell A14. The fill handle is no longer visible.

	A	B
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	
8	8	
9	9	
10	10	
11	11	
12	12	
13	13	
14	14	

Figure 12.47: Autofill Results

Note that Excel detects the pattern and fills the rest of the range appropriately as you drag down. Had you selected simply the “2,” instead of the values “1” and “2,” Excel would have put 2’s in the cells below as you dragged down.

Countif

Countif lets you count the number of values in a range meeting a specified criterion (“= 3”, “>= 11”, etc.). You can enter the Countif arguments directly, or access the function through the Insert Function menu (Figure 12.48).

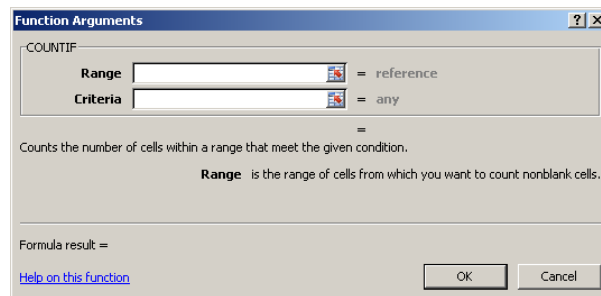


Figure 12.48: Countif Dialog

You enter the range you want to examine in the “Range” field, and a criterion like “>= 11” in the criteria field.

A typical use of Countif is to examine the results of your resampling operation, and determine how many of the resampled results are as extreme (e.g. >=) as the observed value.

Using a variable in the criteria field

Suppose you do not want to ask “how many values are >= 11”, but rather “how many values are greater than or equal to the value in cell A5?”

The proper syntax to use in the dialog box’s Criteria field is as follows:

“>=”&A5

Frequency

This Excel function produces output similar to the frequency table produced by Histogram, but with the advantage that it is “live” – it updates itself each time you redo the problem and produce new output.

Before you use Frequency, your worksheet (and typically you would be using the Results worksheet) must have on it not only the resampled output data, but a range of bin values as well. These bin values are numbers you choose to represent the upper bounds of the bins into which your data will be grouped.

For example, suppose you want your table to have 9 bins containing, respectively, the values ≤ 0 , between 0 and 1 (including 1), between 1 and 2 (including 2), and so on up to the top bin which would be values > 7 . Your bin values would be:

- 0** [contains values ≤ 0]
- 1** [contains values between 0 and 1, including 1]
- 2** [contains values between 1 and 2, including 2]
- 3** etc.
- 4**
- 5**
- 6**
- 7** [contains values between 6 and 7, including 7]
- 8** [contains values > 7]

If your output is in cells A2:A1001 and the bin range is in cells B2:B9, you would select “Frequency” from the Insert Function menu and fill in the fields accordingly (Figure 12.49):

Important: Next, press Control+Shift+Enter.

Frequency is an “array function,” meaning that it works with arrays of numbers, and must be entered by using the Ctrl+Shift+Enter key combination.

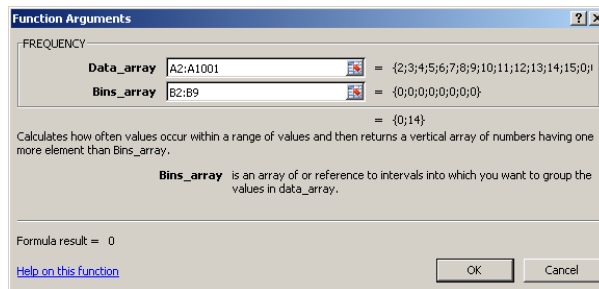


Figure 12.49: Frequency Dialog

Functions

See [Insert Function](#).

IF

Excel’s IF function determines whether a referenced cell meets a specified criterion, and returns one of two values, depending on whether the condition is met. Here is an example in Figure 12.50 that returns a “1” if the number in B7 equals the number in A7 (otherwise it returns a “0”).

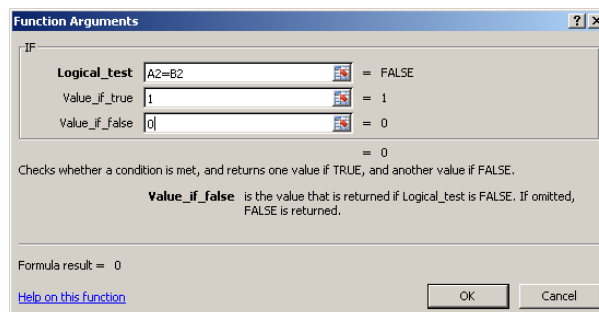


Figure 12.50: Excel’s IF Function

For Logical Test, “A2=B2” means “find out whether A2 equals B2.” The “Value_if_true” line field has a “1” in it, meaning “this formula yields a ‘1’ if A2 = B2.” The “Value_if_false” line field has a “0” in it, meaning “this formula yields a ‘0’ if A2 does not equal B2.” You can also type this function directly into the cell:

IF(A7=B7,1,0)

Insert Function

Many of Excel's statistical and other functions can be entered through the Insert Function button (Figure 12.51):

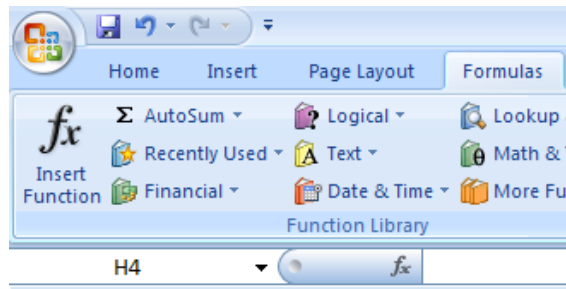


Figure 12.51: Excel's Insert Function “ f_x ”

Note: You can select either “ f_x ” icon; the large one on the left in the Ribbon or the smaller version next to the formula bar.

When you select either “ f_x ,” a menu system opens, from which you can select the function you are interested in using. When you select a function, a dialog box will open in which you can enter the required information for the function.

Percentile

In resampling operations you are often interested in finding some percentile of the results. You can use Excel's Percentile function (from the Insert Function) menu. The Percentile Function dialog is shown in Figure 12.52.

Let's say you have done 1000 trials, and want to find the 2.5th and 97.5th percentiles.

Working from the Results sheet and the cursor on a blank cell, the array that you want to find a percentile for is A1:A1000, and the percentile you want is 0.025 (this is the 2.5th percentile). Repeat the same procedure (with the cursor in a different cell) to find the 97.5th percentile which is entered as 0.975.

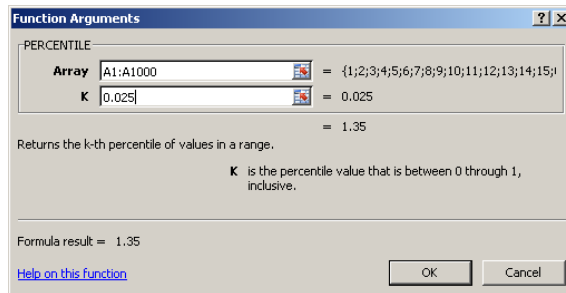


Figure 12.52: Excel's Percentile Function Dialog

Sort (Excel's Sort Capability)

Excel can sort a range of data in ascending or descending order. Simply select the range to be sorted and click the “A to Z” or “Z to A” button on the Excel Data Ribbon, depending on whether you want an ascending or descending sort. Figure 12.53 displays the Sort buttons in the Data Ribbon control.

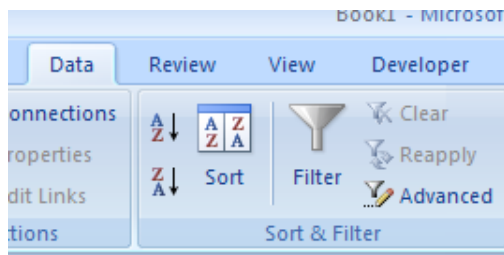


Figure 12.53: Data Ribbon Sort Buttons

Important: Excel's Sort function will *not* be repeated inside a resampling loop. Use Resampling Stats' Sort feature instead (which also offers greater flexibility and functionality).

Bibliography

- [1] Atkinson, D. T. (1975), *A Comparison of the Teaching of Statistical Inference by Monte Carlo and Analytical Methods*, Ph.D. Thesis, University of Illinois.
- [2] Braun, M. (1993), *Differential Equations and Their Applications*, Springer-Verlag, 443-444.
- [3] Chernick, M. (1999), *Bootstrap Methods: A Practitioner's Guide*, New York: Wiley.
- [4] Chung, J.H., and Fraser, D.A.S. (1958), *Randomization Tests for a Two-Sample Problem*, Journal of the American Statistical Association, 53, 729-735.
- [5] Davison, A.C. and Hinkley, D.V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, 1997.
- [6] Dwass, M. (1957), *Modified Randomization Tests for Nonparametric Hypotheses*, Annals of Mathematical Statistics, 29, 181-187.
- [7] Edgington, E. (1987), *Randomization Tests*, 2nd ed., New York: Dekker.
- [8] Efron, B. (1983), *Bootstrap Methods; Another Look at the Jackknife*, The Annals of Statistics, 7, 1-26.
- [9] Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: SIAM.
- [10] Efron, B., and Tibshirani, R. (1993), *Introduction to the Bootstrap*, New York: Chapman & Hall.
- [11] Fisher, R.A. (1935), *The Design of Experiments*, London: Oliver and Boyd.

BIBLIOGRAPHY

- [12] Gleick, J. (1987), *Chaos: Making a New Science*, Penguin Books.
- [13] Good, P. (1994), *Permutation Tests – A Practical Guide to Resampling Methods for Testing Hypotheses*, New York: Springer Verlag.
- [14] Karlin & Brendel, *Chance and Statistical Significance in Protein and DNA Sequence Analysis*, Science, v257, July 1992, p. 39.
- [15] Lyon, H.L. and Simon, J.L. *Price Elasticity of the Demand for Cigarettes in the United States*, American Journal of Agricultural Economics, v50, No. 4, Nov. 1958, p. 891.
- [16] Lunneborg, Cliff (2000), *Data Analysis by Resampling*, Duxbury Press (Brooks/Cole) p. 164-166
- [17] Manly, B. (1992), *Randomization and Monte Carlo Methods in Biology*, New York: Chapman & Hall.
- [18] May, R. M. (1976), *Simple mathematical models with very complicated dynamics*, Nature, Vol. 261, 459-467.
- [19] Noreen, E. (1989), *Computer-Intensive Methods for Testing Hypotheses*, New York: Wiley.
- [20] Peterson, I. (1991), *Pick a Sample*, Science News, July 27, 56-58.
- [21] Pitman, E.J.G. (1937), *Significance Tests Which May Be Applied to Samples from Any Population*, Royal Statistical Society Supplement 4, 119-130.
- [22] Pitman, E.J.G. (1938), *Significance Tests Which May Be Applied to Samples from Any Population: III. The Analysis of Variance Test*, Biometrika, 29, 322-335.
- [23] Rosner, Bernard (1982) *Fundamentals of Biostatistics*, Boston: Duxbury.
- [24] Shevokas, C. (1974), *Using a Computer-Oriented Monte Carlo Approach to Teach Probability and Statistics in a Community College General Mathematics Course*, Ph.D. Thesis, University of Illinois.
- [25] Simon, J. L. (1969), *Basic Research Methods in Social Science*, 362-432, New York: Random House; 2nd ed., 1978; 3rd ed., with Paul Burstein, 1985.

- [26] Simon, J. L., Atkinson, D. T., and Shevokas, C. (1976), *Probability and Statistics: Experimental Results of a Radically Different Teaching Method*, The American Mathematical Monthly, 83, November, 733-739.
- [27] Simon, J. L., and Bruce, P. (1991), *Resampling: A Tool for Everyday Statistical Work*, Chance, 4(1), 22-32.
- [28] Simon, J. L., and Holmes, A. (1969), *A Really New Way to Teach Probability and Statistics*, The Mathematics Teacher, Vol. LXII, April, 283-288.
- [29] Simon, J.L., Mokhtari, M., and Simon, D.H. (1966), *Are Mergers Beneficial or Detrimental? Evidence from Advertising Agencies*, International Journal of the Economics of Business, v. 3, n. 1, 69-82.
- [30] Simon, J. L., and Weidenfeld, D. (1974), *SIMPLE: Computer Program for Monte Carlo Statistics Teaching*, American Statistician, November, (letter).
- [31] Stewart, I. (1992), *Does God Play Dice? The Mathematics of Chaos*, Blackwell Publishing, 136.
- [32] Westfall, P., and Young, S. (1992), *Resampling-Based Multiple Testing*, New York: Wiley.

Index

- About Resampling, 3
- Acceleration, 149
- Add-Ins menu, 5
- Adding New Data, 194
- Advanced Probability, 21
- advertising agencies, 119
- Age Discrimination in Employment, 68
- Analysis of Variance, 101
- Analysis Toolpak, 1
- Analytical Approach, 23
- Another Correlation Study, 117
- ANOVA, 101, 105, 106
- array function, 19
- asymmetric tables, 105
- Auto Binning, 12, 165
- Auto-Range Selection, 159
- Autofill, 202

- Babies.xls, 62
- Baseball, 21
- baseball payroll, 89
- baseball salary vs. rank, 91
- Baseball-c.xls, 91
- Baseball.xls, 21
- Basic.xls, 93
- Basket.xls, 25
- Basketball, 24
- BCA, 150
- BCA Bootstrap, 150

- beta distribution, 176
- Bias, 149
- bias-corrected and accelerated, 149
- BINOMDIST, 23
- binomial distribution, 23, 176
- binomial models, 59
- bins, 164, 204
- Birth.xls, 112
- birthday problem, 31
- Birthday.xls, 32
- birthweight of babies, 60
- Birthweight Revisited: A Signs Test, 111
- Birthweights a Third Time, 113
- Black1.xls, 82
- bootstrap, 38, 149
- Boys&Girls.xls, 18

- Chart Wizard, 52
- chi-squared, 79
- CHITEST, 79
- classical statistics, 38
- Clickthroughs.xls, 71
- coagulation time, 103
- Coins.xls, 6
- confidence interval, 37
- confidence interval for a proportion, 43
- confidence interval for the median income, 45

- confidence interval specification in regression, 96
- confidence intervals for the median, 45
- confidences interval for means, 39
- contingency tables, 79
- correlation, 89
- correlation and regression, 89
- correlation coefficient, 89, 91
- COUNTIF, 13, 203
- Cumulative Frequency, 165
- Custom Functions, 161
- dice, 15
- Diet.xls, 104
- difference in variability, 63
- direct mail, 47
- Disable Screen Updating, 26
- Drills.xls, 39
- Driving While Blank, 82
- Drug Response, 84
- Drug Testing, 86
- Drug.xls, 86
- ESP, 27
- ESP.xls, 27
- exact p-values, 4
- Excel's Sort function, 207
- exponential distribution, 177
- F statistic, 105, 106
- F.xls, 105
- Faithful.xls, 101
- File Operations, 162
- Firing.xls, 68
- Fisher's Exact Test, 80, 84
- formula iteration, 133, 163
- formulas, 162
- FREQUENCY, 18, 204
- frequency distribution, 58
- fruitflies, 55
- Gamma distribution, 177
- Geometric distribution, 177
- Geyser Timing, 101
- Heads/Tails, Boys/Girls, 6
- histogram, 11, 15, 163, 165
- Hypothesis testing, 55
- IF function, 35, 205
- Income.xls, 45
- Insert Function, 13, 206
- installation, 2
- INTERCEPT, 93
- introduction, 1
- inventory, 49
- Inventory.xls, 51
- Iterative Solutions to Equations, 133
- jackknife, 156
- Larry Bird, 24
- LINEST, 94
- Lognormal distribution, 177
- macros, 94
- MATCH, 32
- matched groups, 120
- matched-pair study, 111
- Matrix Resampling or Shuffling, 187
- maximum number of trials, 169
- Measure.xls, 63
- menu, 170
- mergers, 119
- mergers9-1.xls, 119
- Molecular Biology, 65
- Multiple Comparisons, 71
- Multiple Linear Regression, 96
- Multiple Score Cells, 181
- Multistage Resampling and Shuffling, 175, 190
- News.xls, 97

-
- Newspapers and Population, 96
 - normal distribution, 177
 - Normal Matrix Resampling or Shuffling, 187
 - null hypothesis, 56, 63
 - Old Faithful, 101
 - one-sided test, 56
 - Opening Files, 175
 - Options, 172
 - outliers, 46
 - p-value, 57
 - paired permutation test, 113
 - Pareto distribution, 178
 - Paste Special, 162
 - Pay.xls, 115
 - Pearson correlation coefficient, 91
 - percentile, 206
 - percentile confidence interval, 39
 - Permutation methods, 4
 - permutation procedures, 79
 - Poisson distribution, 178
 - price elasticity, 46
 - probability by resampling, 6
 - Protein.xls, 66
 - proxy population, 39
 - Rain.xls, 42
 - random number generator, 175
 - Rank Sum Test, 115
 - Redo, 179
 - Regression, 92, 172, 179
 - Regression Basics, 92
 - Repeat and Score, 8, 181
 - Resample, 6, 183, 191
 - Resample and Shuffle Options, 183
 - Resample or Shuffle a Single Column, 190
 - Resample or Shuffle Columns as Units, 189
 - Resample or Shuffle Rows as Units, 188
 - Resample or Shuffle Within Columns, 189
 - Resample or Shuffle Within Rows, 188
 - resampling, 191
 - Resampling and p-values, 55
 - Resampling in Complex Cases, 65
 - Resampling Stats Add-in, 5
 - Resampling Stats Operations, 159
 - Resampling toolbar, 5
 - resampling without replacement, 62
 - Reset, 174, 191
 - Results Sheet, 9
 - RSXLBeta, 176
 - RSXLBinomial, 176
 - RSXLExponential, 177
 - RSXLGamma, 177
 - RSXLGeometric, 177
 - RSXLLognormal, 177
 - RSXLNormal, 177
 - RSXLPareto, 178
 - RSXLPoisson, 178
 - RSXLRand, 178
 - RSXLRandbetween, 178
 - RSXLUniform, 178
 - RSXLWeibull, 179
 - running the simulation again, 138
 - sample size, 7, 188
 - Sampling with Replacement, 30
 - versus sampling without replacement, 108
 - Saving and Opening Files, 192
 - Score, 194
 - Secretary Problem, 34
 - Showing the Resampling Toolbar, 199
 - Shuffle, 27, 171, 194

INDEX

Shuffle Within Rows, 114
Signs Test, 111
Single Row or Column Resampling
 or Shuffling, 185
SLOPE and INTERCEPT, 93
Sort, 74, 171, 194, 207
STDEV, 64
stratified resampling, 119, 129, 196
Sweeps.xls, 49

tabulate resampled results, 18
Tea Lady, 79
tea-taster, 80
Tea.xls, 80
toolbar, 170, 198

uniform distribution, 178
Urn, 25, 171, 199
Urn dialog box option, 199
Useful Excel Functions, 202

Weather, 42
Weibull distribution, 179
Worksheet Urn, 201