Fathom™

Dynamic Statistics™ Software

Learning Guide

Key Curriculum Press
Key College Publishing

Fathom Dynamic Statistics™ Software February 2001

Authors: William Finzer, Tim Erickson, and Jill Binker

Lead Programmer: Kirk Swenson

Development Team: William Finzer, Kirk Swenson, Nick Jackiw, Eugene Chen, Gavin Peters, Tony Thrall, Tim Erickson, Jill Binker, Matthew Litwin, Zach Teitler, Denise Howald, Caroline Wales, and Vadim Keylis

Statistical Consultant: Tony Thrall

Production Editors: Deborah Cogan, Kristin Ferraioli

Copy Editors: Joan D. Saunders, Erin Milnes

Production and Manufacturing Manager: Diana Jean Parks

Production Coordinator: Ann Rothenbuhler

Art and Design Coordinator: Caroline Ayres

Cover Designer: Jenny Somerville

Cover Illustration: Rick Helf

Cover Photo Credit: Gary Buss/FPG ©1995

Prepress and Printer: Data Reproductions

Executive Editor: William Finzer

Publisher: Steven Rasmussen

Portions of this material are based upon work supported by the National Science Foundation under award number III–9400091. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Key Curriculum Press
1150 65th Street
Emeryville, CA 94608

techsupport@keypress.com
http://www.keypress.com

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1 04 03 02 01 ISBN 1–55953–515–6

Acknowledgments

Many people over many years helped us in many ways.

On the original NSF advisory board, Gail Burrill focused us on improving the lot of practicing teachers; Chris Hancock taught us the Tabletop metaphor and how to be concrete; Cliff Konold started us on the fruitful path of resampling; Andee Rubin not only warned us off dice but kept us thinking about learning; Mike Shaughnessy showed us the relationship between offline and online simulation; and Frank Sheehan brought us back time and time again to the statistical and probabilistic heart of the matter.

Of the hundreds of high school teachers who gave us feedback, deserving special thanks are those members of the statistics study group who gave of their Wednesday evenings once a month over two years trying out activities and software: Marianne Nolte for her insights into students' learning; Elisabeth Klein for her appetite for new ideas; Bev Austin for her brave willingness to try new things; Lew Douglas for his ideas on teaching statistics; Carlos Cabana for both his attention to detail and his ability to see the big picture; Judy Quan for her generous hospitality at her school; Don De Pasquale for his reality-tempered enthusiasm; and Carolyn Kirkpatrick for the student work she brought for us all to look at.

Many teachers served as reviewers, consultants, and field testers. We mention Gretchen Davis, who did not fear to tell us the hard truths; Chris Olsen, who told us what he really wanted; Leona Mirza, who had faith in Fathom from the beginning; Howard Amerlan, who funneled both his own and his students' ideas to us; Paul Myers, who helped us see the power of Fathom; Susann Mathews, who helped us think about uses for the network; Dudley Brooks, who gave countless good suggestions; Chris True, who made the first Fathom poster; Sue Gordon, who encouraged us with stories of student understanding; Sharon Shrader, who took Fathom home from the Teaching Statistics with Technology Institute and even surveyed her students about it; Mark Vosskamp, who advocated a balance between discovery and structure; Jim Fensom, who holds the record for bug submissions; David Hoppe, who gave us insightful criticism; and Joyce Smart, who was available for on-the-phone debugging; Len Fellman who has stuck with us through thick and thin, and Robin Lock for invaluable critiques, insights, and suggestions.

Ron Avitzur provided not only an excellent formula engine and editor, but also valuable insights into how his engine should and could link to Fathom.

Many people at Key Curriculum Press contributed to Fathom beyond their normal call. Dan Bennett critiqued just about everything; James Browne found last-minute bugs; Stefanie Liebman tested and told us she enjoyed it; Jennifer Strada tested tours and instructions; and Candice Miller helped us with naming Fathom.

Steve Rasmussen believed in us, supported us, critiqued everything from ad copy to user interface, and did more than anyone else outside the team to make Fathom real.

Nick Jackiw not only contributed in major ways to Fathom's design, he continually brought us back to the fundamental vision of educational software we are pursuing.

No project this size comes to fruition without the help of many family members and friends. Among these are Brigid McCaw, Lauren Finzer, Devin Finzer, Ramona Swenson, Anne Erickson, and Meg Holmberg.

Contents

Introduction

How will you learn to use this program?

Different people learn to use programs in different ways. Some read all the manuals before they ever start a program. Others never read manuals and jump right in. Here we describe the tools available to you for learning Fathom—starting with this learning guide and including other materials that come with Fathom as well. Let's go over what you have to help you on this journey.

The Structure of Fathom: This short article, included in this introduction, is a description, from the designers' point of view, of how the program works. It is just to read; there's nothing you have to do.

Guided Tours: There are nine guided tours in this book. If you do them all in order, you will learn a great deal (but not all) of what Fathom has to offer. It's best to do the tours step by step sitting in front of a computer. Each tour is discussed in greater detail shortly.

The Quick Reference: This is a six-page summary of Fathom. It has pictures and descriptions of the various gadgets that you'll find on the screen. You might want to keep this handy as you use Fathom, especially at first.

The *Reference Manual*: This is a comprehensive description of the entire program. Here you will find descriptions of every command and function. The *Reference Manual* is best used for looking things up, rather than for reading from front to back. If you have a student edition of Fathom, you do not have the *Reference Manual* on paper, though there is a copy on the CD.

On-Line Help: This is an on-line manual, complete with index, that covers a lot of the same material as the *Reference Manual*. When you're running Fathom, choose **Fathom Help** from the **Help** menu. This launches your browser, opened to Fathom help.

Data in Depth: This is a book of curriculum materials that comes with the instructor's editions only. Many of these activities teach you to use the program while they help you learn about mathematics and data analysis. If you're a teacher with students who do not own their own copies of Fathom's documentation, the students will probably learn to use the program by doing these activities.

Internet Resource: The Fathom Web site, found at www.keypress.com/fathom, has upgrades, a User's FAQ with answers to frequently asked questions, sample documents, links to data, and other resources.

Workshops for Teachers: Key Curriculum Press offers one-day workshops during the school year and three-day workshops and a week-long institute during the summer.

The Guided Tours

The guided tours make up most of this book. They're written for you to do in order and step by step. We expect you to have a computer in front of you, running Fathom. You may, of course, use the tours differently—jumping from topic to topic, for example—as long as you realize that you may come across instructions in the wrong order. That is, in later tours we expect you to know basics covered in the first tours.

Unless you're very comfortable with computers and with statistical concepts, you probably don't want to do the tours in one run. Consider doing a couple of tours, then using Fathom for a while to get used to the basics, before doing later tours.

From time to time we ask you to open a file. These files are in the Learning Guide Starters folder, in the Sample Documents folder, in the Fathom directory.

Tour 1: An Introduction to Exploring Data—Residents of Beverly Hills, California

This is an introduction to Fathom and all of the basics. You'll open files, make graphs and tables, and learn to manage Fathom's screen.

Tour 2: Data, Formulas, and Prediction—Wrist Versus Height

In this tour, you'll enter data, fit lines to data in a scatter plot, use a residual plot to improve your fit, and learn about the squares in a least-squares line. You will also see how you can change data in Fathom simply by dragging it—which is one reason we call Fathom "Dynamic Statistics" software.

Tour 3: Analysis of Groups

This tour features techniques for comparing groups. You'll compare groups by splitting plots according to group membership, and by computing summary statistics for the entire data set and for each group separately. You will also filter data—that is, look at part of the data set. Some of these tasks require that you use Fathom's formula editor. This will be your first encounter with that prodigious tool.

Tour 4: Formulas, Functions, and Data—The Planets

You'll use the formula editor even more in this tour. You'll learn to define new attributes with formulas and to plot functions on a scatter plot. You'll also meet another essential ingredient of dynamic statistics: the slider. A slider is a variable parameter you will use to control the function you're fitting to the data. As you drag the slider, its value changes and the function moves.

Tour 5: Generating Mathematics—Triangular Numbers, Fibonacci, and Chaos

We visit pure mathematics in this tour, learning how to define functions recursively in Fathom. Our collection of cases will be a sequence, with each case linked to previous cases by a formula. You can use "regular" functions to compare the recursive model to its closed-form counterpart, to discover the golden ratio in the Fibonacci numbers, and to see how the parameter in the logistic function affects its values.

Tour 6: Simulation—Polling Voters

In this tour, you will construct a simulation of a sample of voters using a random number and a slider. Then you'll see how well you can predict the outcome of an election. This requires understanding a new Fathom concept: the measure. A measure is an attribute of an entire data set. You'll collect measures (poll results) repeatedly to build up a distribution of possible polls.

Tour 7: Estimating Confidence Using the Bootstrap

By how much do the readings on two thermometers differ? We can compute the difference, but what is our margin of error? In this tour, you'll use sampling to find a confidence interval for a parameter. This technique, called the bootstrap, is relatively new in statistics—and you can do it in Fathom. You'll coordinate your knowledge of sampling and measures to build up this inferential procedure.

Tour 8: Testing a Hypothesis

Darwin did an experiment with plants to test whether cross-pollinated plants grew taller. But he had a small sample. What should he do? This time, instead of constructing a statistical procedure from scratch, we'll use a traditional procedure built in to Fathom: Student's *t*-test for the mean. Along the way, we'll still use simulation, this time to assess whether the original data are consistent with the normal distribution.

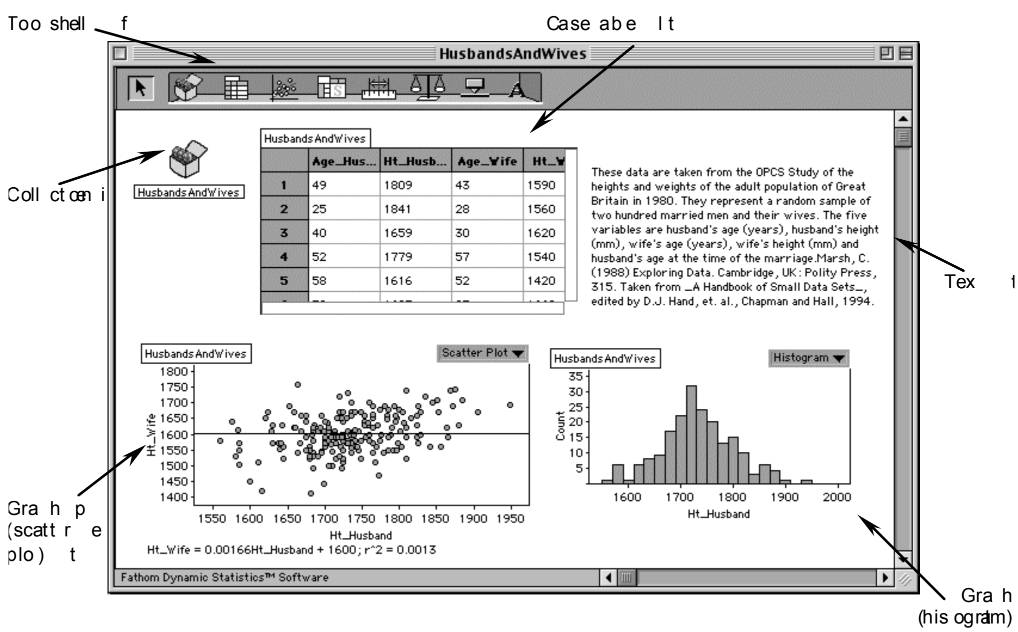
Tour 9: Pets and Sports—Testing for Independence

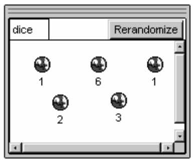
In this tour, you will study poll results in which schoolchildren were asked whether they prefer dogs or cats. Can we see sex differences in the preferences? That is, do girls like cats more? This last tour synthesizes traditional and computer-intensive methods. You will simulate the null hypothesis to generate a distribution, and you will also just plug the data into a chi-square test for independence.

The Structure of Fathom

Some people want to know what the designers think about the structure of a program. If that's you, this section is what you want. Others like to figure out a program by using it. That's a good way, too. But even if you seldom read manuals, you may want to read this section for background.

The Fathom Window





A collection, opened up

A collection, made small (iconified)

A Fathom document lives in a window and contains objects. There are several different kinds of objects, led by the collection (which holds the data). The collection looks like a rectangle with gold balls in it, or, if you make it small, a box of gold balls.

Graphs, tables, and statistical tests are objects, too. But they do not contain data. They just provide ways of looking at or altering data. (Deleting a graph or table leaves your data intact; deleting the collection *deletes your data.*)

Sliders are also objects; they control variable parameters, such as the coefficient of a function or a probability in a simulation.

Text resides in objects, too. Text objects allow you to add explanatory notes.

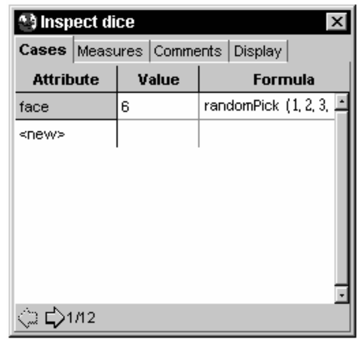
When you first use Fathom, collections are boring and may even appear unnecessary. Why can't the data just live in a table, like in a spreadsheet? The answer is deep; one way to look at it is that Fathom lets you distinguish between the way things *are* and the way they *look*. A table is just one representation of the data, and we didn't want to give it the privileged position of being the data's "native" appearance. So we invented another representation, the gold ball. Another reason for the existence of collections is that it makes some more advanced data manipulation, such as sampling, easier. For example, you sample from among the gold balls in the collection—and the balls go into another collection.

Cases and Attributes

Each gold balls is a case. A case has attributes (also known as variables, but we had to pick a word and stick with it) and attributes have values. If the values are all numeric, then it's a continuous (measurement) attribute; if the values are text, it's a categorical attribute. If your cases are people, height and age are continuous; sex and race are categorical.

Whether an attribute is continuous or categorical determines many things—including the kinds of graphs you can make to represent it.

The Inspector

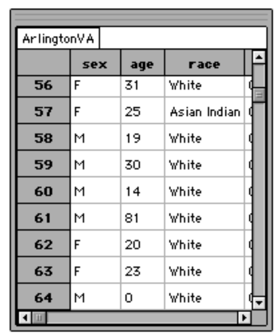


Inspector (Windows platform)

When you double-click a collection, its inspector appears. It will show you the names and values of the attributes, one case at a time. You may edit the values here. You can also add measures (more on this later) and comments to the collection, change how the cases look, and control advanced functions, such as sampling, using the inspector.

If you're typing a lot of data into Fathom, the inspector is *not* the most convenient place for it; you'll want to use a case table instead.

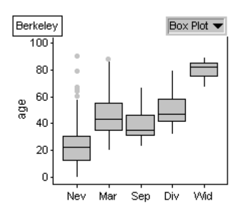
Case Tables



The case table is the familiar tabular view of data. Each case appears as a row; each attribute as a column. You can enter and edit data quickly in a case table and add new attributes.

You can also rename attributes, move columns by dragging them, resize columns by dragging their borders, sort the data, and even hide attributes to make your table less crowded. But though you can change data through a case table, it is only one view of the data. If you delete the case table, the data are still present—in the collection.

Graphs



Fathom also lets you view data graphically. The graph object supports various kinds of charts and plots. You tell Fathom what to graph by dragging attribute names to an axis of the graph. You specify the graph type by using a menu in the corner of the graph object.

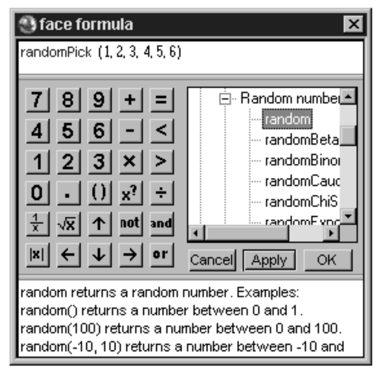
The configuration of attributes determines what kinds of graphs are available. For example, if you put age on an axis and leave the other axis blank, you can get a box plot or a histogram—because age is a continuous attribute. But if you put sex on that same axis, you can't get a histogram—because sex is categorical. You get a bar chart instead.

You can also add things to graphs, such as functions. Functions can be anything you can express with a formula (see the next section) as well as the more familiar lines such as the least-squares linear regression line or the median-median line.

In most kinds of graphs, you can change the data by dragging unless the data are locked. So, for example, you can drag points in a scatter plot and see how they affect a least-squares line.

Ubiquitous Formulas

Formulas are everywhere in Fathom. You can use formulas as filters to tell which cases you want to see and you can use them to compute values (including random values) for attributes. You can use formulas to devise statistical measures and to plot formulas as functions in graphs.



(Windows platform)

You write formulas in the formula editor. The formula editor appears when you double-click the place where a formula belongs, or when you select certain menu items. For example, to make an attribute that is the logarithm of income, make a new attribute, then select its name. Choose **Edit Formula** from the **Edit** menu. In the formula editor, enter log(income) and click **OK.** The new attribute will fill with the logs of the incomes. Later, any change in the incomes will instantly be reflected in their logs.

The formula editor supports a wide array of functions, from trig functions to statistical distributions to conditionals such as if() and switch(). They are all listed in the function and attribute list, a pane at the right in the editor. Double-click an item in the list to enter it into the formula pane above. You can also enter it by typing. The formula editor supports syntax coloring, that is, it displays the names of functions and attributes in special colors. If you type an attribute name or function and it remains black, Fathom does not recognize it (you probably have a typo).

Using the Formula Editor

The peculiar nature of mathematical formulas makes editing them different from editing ordinary text. First, certain characters—parentheses, quotes, and absolute-value signs—must appear in pairs. Fathom always makes the second of the pair when you enter the first. Second, grouping and order of operations require special editing capabilities. New users of Fathom can use their intuition most of the time, but there is one new principle they must learn: The effect of some keys depends on what, if anything, is selected when you press them. For example, if a + b—without parentheses—is selected, and you press (, you get (a + b)—with parentheses. If you press * (for multiplication), you get (a + b)*.

The Effect of Selection

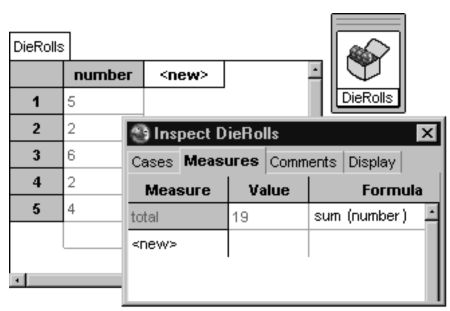
Before	*After*
$a + b$	$a + \frac{b}{4}$
$a + b$	$\frac{a + b}{4}$
$(a + b)$	$\frac{(a + b)}{4}$

In each of the examples at left, the user types /4.

Finally, you need to click **OK** to close the formula editor. But sometimes you want to see the result of your formula—to see if it's right—before closing the editor. In that case, click **Apply** on the screen or press **Enter** on your keyboard; this applies your changes, leaving the editor open. However, you must close the formula editor before you can go on to do other things.

The formula editor is *not* case sensitive; that is, it doesn't care whether you type capital or lowercase letters.

Measures

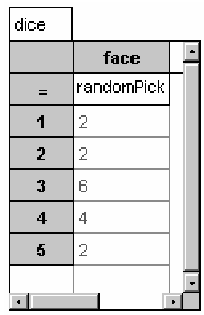


A regular attribute is a case attribute: It has a separate value for each case in the collection. But there are also measures—attributes with one value for the entire collection. For example, in a collection of five dice, each might have an attribute called number. But the collection itself could have a measure, or total, that applies to the entire collection. It would have a formula: sum(number). So, you can think of measures as statistics that summarize your collection, though they can serve other functions as well.

You create measures by using the inspector. Double-click a collection to make the inspector appear, then go to the Measures pane.

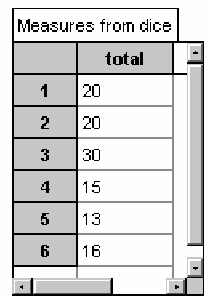
Derived Collections

Some of Fathom's collections fill with data automatically, according to rules you specify. These are called derived collections. There are several kinds. The two most important are samples and measures collections. A sample is just what it sounds like: a collection that is a sample of some other collection, called its source. Select the source and choose **Sample Cases** from the **Analyze** menu. The sample collection appears. Control how many cases it samples and whether it samples with or without replacement in its inspector.



dice

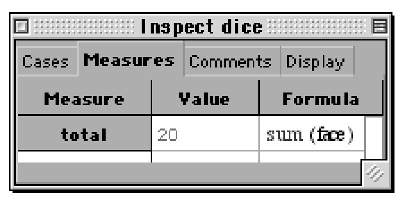
	face
=	randomPick
1	2
2	2
3	6
4	4
5	2



Measures from dice

	total
1	20
2	20
3	30
4	15
5	13
6	16

The measures collection is the key to simulation and analysis. It converts measures into case attributes—so you can record statistics (measures) about your collections. (It's a bit tricky and is best understood after you've tried it.) Suppose you have a collection with five (randomly generated) dice in it and a measure, or total, that contains their sum. If you connect a measures collection to that source collection, the measures collection can record the total as a case. If you tell Fathom to collect 100 measures, it will instruct the dice to re-roll 100 times, it will compute each total, and it will make a collection of those 100 values. You have built up the distribution of that statistic.



Inspect dice

Cases | Measures | Comments | Display

Measure	Value	Formula
total	20	sum (face)

To make a measures collection, select the source collection and choose **Collect Measures** from the **Analyze** menu. Again, control how many measures it collects in its inspector.

Getting Started

Here's what you need to get started.

System Requirements

Windows

- Pentium-based system or equivalent
- Windows 95, Windows 98, Windows 2000, or Windows NT 4
- 16 megabytes of memory
- CD-ROM drive

Macintosh

- OS 8 or better
- 16 megabytes of RAM
- CD-ROM drive

Install Fathom

Windows

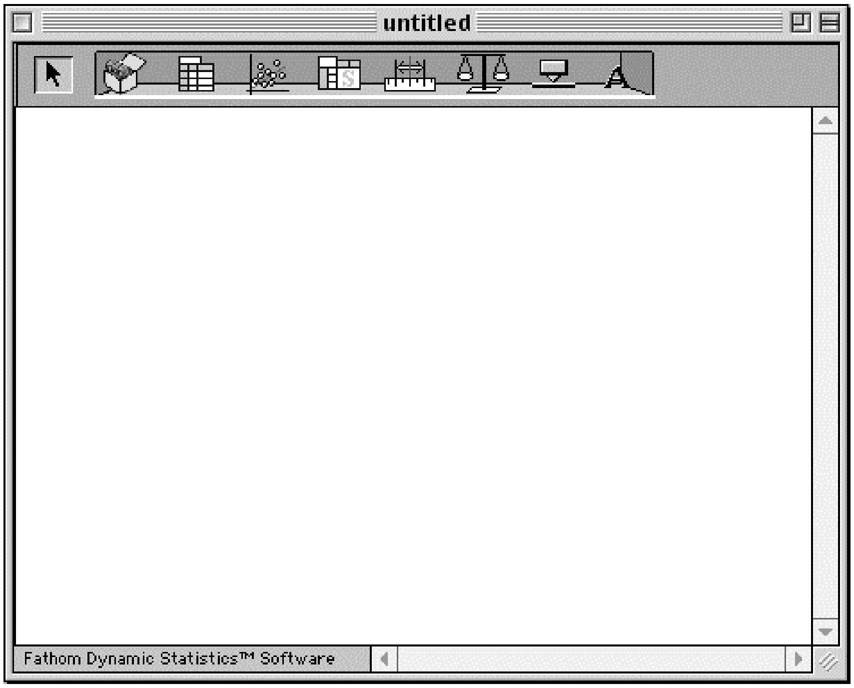
1) Insert the Fathom CD into your CD drive.
2) Double-click the **My Computer** icon on your desktop.
3) Double-click the CD icon. (It may be labeled either **D:** or **Fathom**.)
4) Double-click the **Setup** icon and follow the directions on the screen.

Macintosh

1) Insert the Fathom CD into your CD drive.
2) Double-click the CD icon.
3) Double-click the **Installer** icon and follow the directions on the screen.

Starting Fathom

The installer places a shortcut to Fathom on your desktop. (It looks like a little gold ball.) Double-click it. You should get an empty window similar to the one on the next page. From there, use commands in the **File** menu to open an existing Fathom document, import data from a text file or URL, or make a case table to type data into.



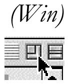
(Macintosh platform)

Using the Help System

- The on-line help system has most of what you need to know. Choose **Fathom Help** from the **Help** menu. Fathom launches your browser software allowing you to access help information.

About Screen Space and Objects

(Win)



(Mac)

- Fathom uses a lot of screen space. You may need to maximize your window by clicking the maximize box in the upper right of your screen.
- Only one *object* (collection, table, graph, text, statistical test, or slider) in the window is active at a time. The active object has a border around it.
- To make an object active, click on it.

- You can move and resize objects by clicking and dragging on their frames when they are active. Use the top of the frame to move an object. Drag sides or corners to resize it.
- Keep your objects small. (If you make them *very* small, they become little icons.) Use the **View Object in Window** command in the **Display** menu when you want a big object.
- The graph, summary table, and analysis objects each have their own menus, which appear only when the object is active.

Getting Data

- One way to get data is to open a Fathom document. Choose **Open** from the **File** menu and look in the **Sample Documents** folder. Many of these documents have notes in them explaining what the data set is and where it came from. Some have notes about things you can do with the data.
- You can also import data from a text file. Tab-delimited text works well, but Fathom can figure out the format of some files that are more complicated than that. Use the **Import From File** command on the **File** menu, or drag a text file from the desktop into an empty area of a Fathom document. Fathom creates a new collection.
- The Internet is an amazing source of data. Fathom comes with some URLs that point to data in the Data and Story Library (DASL). If your computer is connected to the Internet, simply drag one of these URLs into a Fathom document and you will get the data. Documentation for the data will be in the collection comments. You can access many other data sources on the Internet just by dragging their URLs from your Web browser into a Fathom document. However, some will *not* import correctly or will require tweaking to format properly.
- You can also copy data from other programs (such as a word processing table, spreadsheet, or Web page) and paste it into an empty collection. If importing doesn't work well, copy/paste might work better. However, you need to copy and paste comments separately.

About Undo

- The **Undo** command in the **Edit** menu will help you a lot. Choose it to undo the last thing you did. You can keep choosing it to keep undoing previous actions.

Tour 1: An Introduction to Exploring Data—Residents of Beverly Hills, California

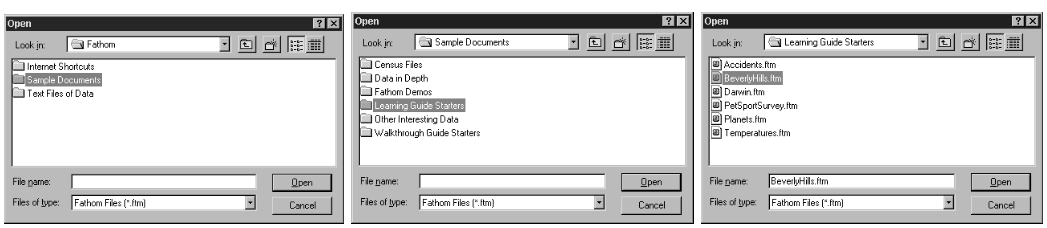
This tour shows you how you can begin to explore data using graphs and tables. We'll look at a sample of residents of Beverly Hills, California. The data were obtained from the 1990 U.S. Census "microdata," that is, data on individuals.

You Will Learn How to …

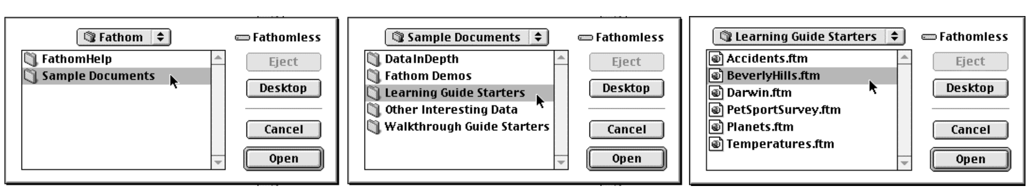
- open a previously saved Fathom document.
- call up an inspector.
- create, change, and move objects (collections, tables, and graphs).
- see with graphs and tables how a single attribute, such as age, varies in the sample; and how combinations of attributes, such as age and income, vary.

Opening the File

1) If Fathom is not already running, start it by double-clicking its icon.
2) From the **File** menu choose **Open**.
3) In the resulting dialog box, open the **Sample Documents** folder, then the **Learning Guide Starters** folder, and then the **BeverlyHills** file.



The sequence of folders as seen on a Windows computer



The sequence of folders as seen on a Macintosh

The documents that come with Fathom are locked so that each person going through a tour (or using a file for a *Data in Depth* activity) on a given machine will encounter the same document. Before going further, you should save this file.

4) Choose **Save As** from the **File** menu.

5) In the resulting dialog box, navigate up two levels to the Fathom folder.

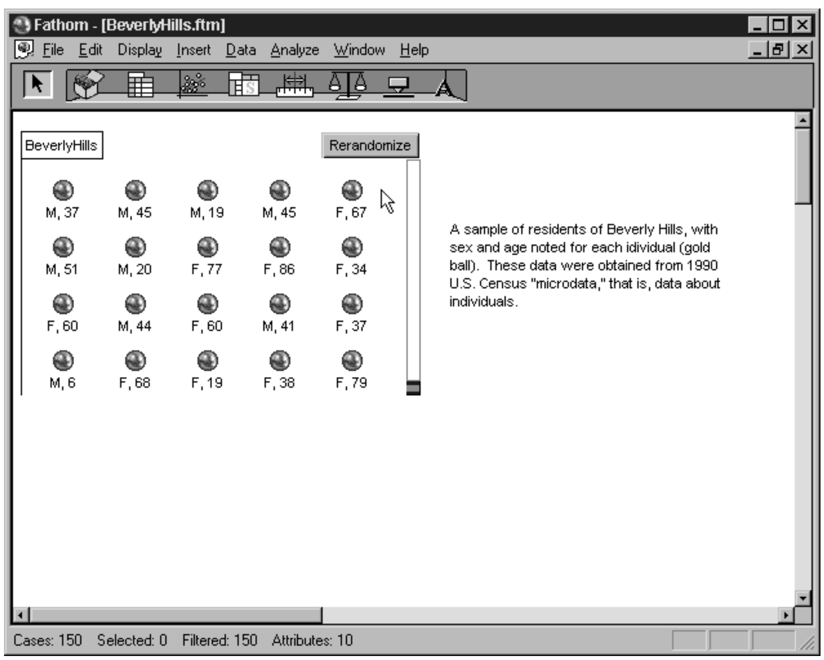
(Win)

(Mac)

6) Name the document **Tour1** and click **Save**.

7) Maximize the window (if it isn't already maximized) by clicking the appropriate button in the upper right corner of your screen.

   Your window should look similar to the one shown below.

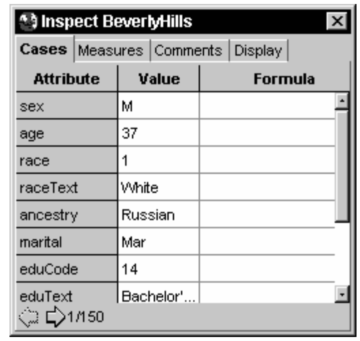


Each gold ball in the collection represents one case in the collection, that is, one person. When the cursor is over the collection (or case table, if there were one), the status bar—bottom left of the Fathom window—gives information about the data: This collection has 150 cases, and each case has 10 attributes. Let's see what the attributes are.

Inspecting the Data

You can also bring up the inspector by selecting the collection, then pressing **Ctrl-I** *(Win)* **⌘-I** *(Mac).*

8) Double-click the collection.



The inspector appears. If you double-click on a case, the inspector shows that case; otherwise, it defaults to the first case in the collection. The inspector shows 8 of the 10 attribute names. To see the remaining attributes, use the scroll bar to scroll down the list, or you can resize the inspector by dragging on the window border.

9) Click on the inspector's **Comments** tab.

You will see a brief description of this data set. The comments pane of the collection is a good place to keep documentation and notes about the data. It is also where notes and attribute definitions appear in collections you download from the Internet.

10) Click on the **Cases** tab.

11) Close the inspector by clicking its close box.

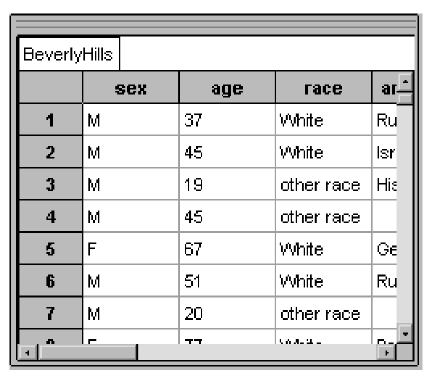
Making a Case Table

Often you want to see the data in a table. Here's one way to make a case table in Fathom.

The keyboard shortcut for inserting a case table is **Ctrl-T** *(Win)* **⌘-T** *(Mac).*

12) Click once on the collection to select it. (When it is selected, it has a border around it.)

13) Choose **Case Table** from the **Insert** menu.

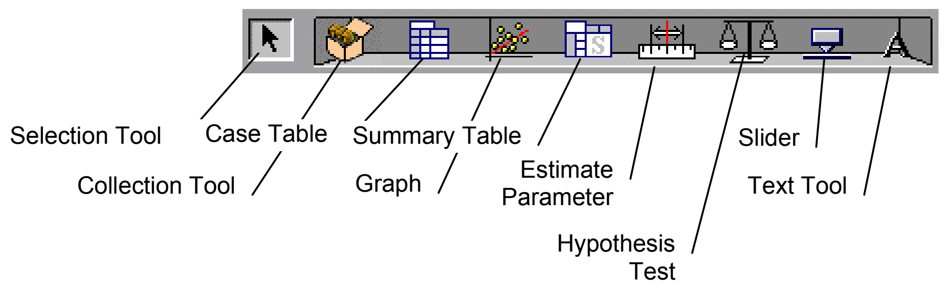


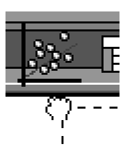
Each row in a case table shows attribute values for one case. Each column corresponds to one attribute. You can use a case table to change values, to enter new data, to add attributes, and to define formulas for attributes.

14) Drag a bottom corner of the case table to enlarge the table to see more cases and attributes.

Making a Graph

Let's look at the range of ages among the people in our sample by creating a graph of age. Across the top of the Fathom window is a shelf with tools for making frequently used objects.





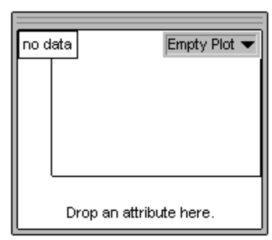
15) Drag a graph from the shelf into your document.

When you click, your mouse pointer becomes a closed fist. As you move the mouse over the document, you see a frame that shows where the graph will be placed when you release the mouse.

16) Find a spot in the document where you want to place the graph and release.

The result should be an empty graph, as shown at right.

You can also create a graph by selecting **Graph** from the **Insert** menu or pressing **Ctrl-G** (Win) **⌘-G** (Mac).



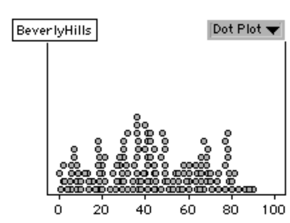
The graph is empty because you haven't told Fathom which attribute to graph. You do this by dragging an attribute from the case table to an axis of the graph.

If your case table takes up too much of the screen, shrink it by dragging on its border. Also shrink the collection.

17) Make sure you can see the age attribute in the case table.

18) Drag the age attribute by its name to the horizontal axis of the graph over the spot labeled **Drop an attribute here**. (As the mouse moves over the appropriate area of the graph, a black border will appear.)

The result should be a dot plot of the ages of the people in the sample, as shown at right.

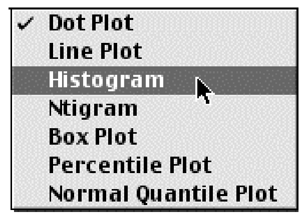


Another way to force cases not to overlap is to choose **Stack Dots** *from the* **Graph** *menu.*

We see that the ages of the people in our sample range from zero to almost 100. But at this resolution, the dot plot does not show which ages are more common than others. You can change the type of graph from a dot plot to a histogram.

19) Click on the tab in the upper right corner of the graph labeled Dot Plot.

A popup menu, as shown at right, appears.



20) Choose **Histogram** from the popup menu.

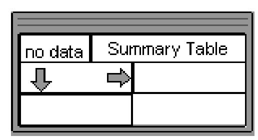
By default this histogram groups people into bins according to their ages. The height of each bin represents the number of people in the age range indicated on the *x*-axis. If you hold the cursor over a bin, the number of people the bin represents appears in the status bar (lower left of the Fathom window). Notice that in 1990 there was quite a peak of people in the range from thirty to forty. These are the baby boomers.

Making a Summary Table

Now that we've seen the distribution of ages among the people in our sample, let's look at how many females and males we have. There are several ways to do this; one is to make a summary table.

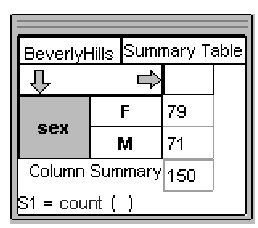
The keyboard shortcut for inserting a summary table is **Ctrl-U** (Win) **⌘-U** (Mac). You can also drag one off the shelf (the table with the S on it).

21) Choose **Summary Table** from the **Insert** menu. Fathom puts an empty table in your document, as shown at right.



22) Drag the sex attribute to the summary table, dropping it on the down-pointing arrow. (Again, a black outline appears when you are over a drop area.) The result should be similar to the illustration.

From this table we see that there are 79 females and 71 males.



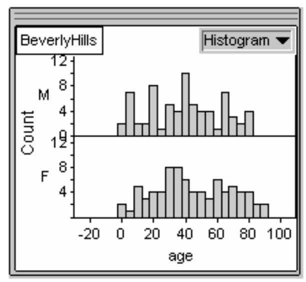
By default, when showing a categorical attribute, the summary table displays the number of cases in each group, but you can display other things in addition to, or instead of, the count by editing the formula. (See page 38.)

Splitting a Graph

We've now seen how age and sex vary among the people in our sample, but we have not checked to see whether males and females have roughly the same distribution of ages. (Do you have any reason to expect a difference?)

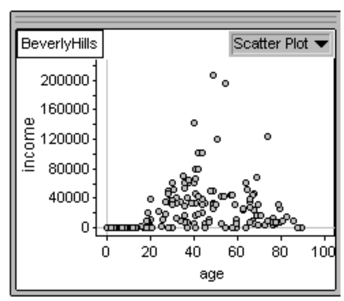
23) Drag the sex attribute to the *vertical* axis of the age histogram. (As before, a thick black outline surrounds the drop area.)

The result should be similar to that shown here. The sex attribute has split the histogram into an upper histogram of males and a lower histogram of females. We say that we *split* the histogram of the numeric attribute, age, by the categorical attribute, sex. The two histograms seem similar, as we might expect from a random sample of residents. (But are there perhaps more older women?)



Graphing Two Numeric (Continuous) Attributes

Now we'll look at relationships between two numeric attributes, age and income.



24) Drag income to the vertical axis of the age histogram, thereby replacing the sex attribute already there.

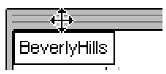
You should get a scatter plot like the one shown here. The type of graph has automatically changed from a split histogram to a scatter plot because you can't split a histogram by a continuous attribute.

25) We see that very young people earn practically nothing, as expected, and that a few middle-aged residents in our sample have incomes around $200,000 per year.

Moving, Resizing, and Deleting Objects

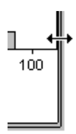
You may need to move, resize, and even delete graphs, tables, and collections in your document. (But be careful about deleting a collection, since that will delete the data.)

Move



Drag the bar at the top of whatever object you want to move. (You'll have to select it first to see the bar, but if you know where the bar is, you can just drag.) You can also move objects by holding down the **Alt** key (Win) **⌘-Option** keys (Mac), and dragging the object. Notice the kind of cursor you get when you can move an object.

Resize



The left, right, and bottom edges of the frame around an object are used for resizing. You can also drag any corner to resize in two directions. If you shrink an object enough, you *iconify* it, turning it into a little icon which you can move by dragging its body.

Delete

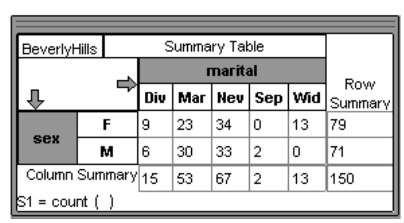
To delete an object, first click on it to select it, then press **Delete** (Win) **⌘-D** (Mac), or choose **Delete [Object]** from the **Edit** menu.

Two Attributes in the Summary Table

When we are interested in two categorical attributes, such as sex and marital status, summary tables can tell us the number of people (frequencies) with each combination of values.

26) Drag the marital attribute to the other (right-pointing) arrow of the summary table. (Again, you get a black outline when you're in the right area.)

The result should be similar to that shown here. Consider the cell with 23 in it. It tells us that there are 23 married females, because it is in the column for married people and in the row for females.



		marital					
		Div	Mar	Nev	Sep	Wid	Row Summary
sex	F	9	23	34	0	13	79
	M	6	30	33	2	0	71
Column Summary		15	53	67	2	13	150

We see that all 13 of the people who are widowed are women. Marital status, unlike age, seems to be distributed differently among the women and men in this sample from Beverly Hills.

Graph and Summary Table Tips and Techniques

You now know the basics. Exploration will tell you much of what you need to know. Fathom comes with many collections of data. Take some time to explore this and other data sets. Here are some tips about graphs and summary tables that you may find useful. When you are through with this file, close it.

F, 42 (42, 100000)

This 42-year-old female has an income of $100,000.

- As you move the mouse over data displayed in a graph, you will see information about the data in the status bar in the lower left corner of the window (see left).
- You can have as many graphs as you like, displaying data in different ways. You can select cases (or whole bars in histograms and bar charts) on the graphs by clicking on them. Since selecting in one graph selects in all the others, this is a great way to see relationships among many attributes.
- In a dot plot or scatter plot, double-click a point to bring up an inspector for that case.

Don't forget the help system!

- Rescale an axis by dragging its numbers with the mouse pointer (which turns into a hand). There are three drag areas: with the hand pointed up you shift the axis; with the hand pointed left or right, you can push numbers off of, or pull numbers onto, the axis.
- When a graph's axes have been changed, you can quickly set them so that all the data is visible by choosing **Rescale Graph Axes** from the **Graph** menu.
- You can make the bars of histograms narrower or wider by dragging their edges.
- Look in the **Graph** menu to see what else you can add to the graph. Depending on the graph, you can add lines of best fit, add a movable line, or you can plot functions and values on it.
- When you drag an attribute onto an already occupied axis, the new attribute replaces the attribute that was there.
- There are commands in the **Graph** menu to remove attributes from a graph.
- Try putting other attributes on the summary table. Dropping an attribute on an existing attribute replaces it. Dropping the same kind of attribute (categorical or numeric) on the arrow pointing at an attribute adds the new one, leaving the first in place.

- Just as adding a categorical attribute to a summary table gives you the *count* by default, adding a numeric attribute gives you the *mean* by default.
- When a summary table is selected, a **Summary** menu appears. Use it to add formulas to the summary table.
- Formulas using rowProportion, columnProportion, rowTotal, and columnTotal are particularly useful when your summary table has a categorical attribute on each axis.

Tour 2: Data, Formulas, and Prediction—Wrist Versus Height

Here we look at entering data with an eye toward using the values of one attribute to predict another.

You Will Learn How to …

- enter data into a new collection.
- make a scatter plot and drag points to change the data.
- fit a line through data in a scatter plot in two different ways.
- change the captions of the cases (gold balls) in a collection.
- show the influence of outliers.

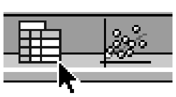
An "outlier" is a data point far from the main cluster of points.

Preparing to Enter Data

Do taller people have bigger wrists? We expect a correlation between the two quantities height and wrist circumference, but we don't expect it to be perfect. How well can we predict height if all we know is wrist size?

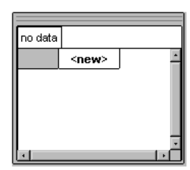
1) If Fathom is not already running, start it by double-clicking its icon. (If you don't already have an empty document, make one with the **New** command in the **File** menu.)

2) Maximize the window.



The shortcut for inserting a case table is **Ctrl-T** *(Win)* **⌘-T** *(Mac).*

3) Drag a case table from the shelf into the document, or choose the **Case Table** command in the **Insert** menu. You should get an empty table similar to the one shown at right.

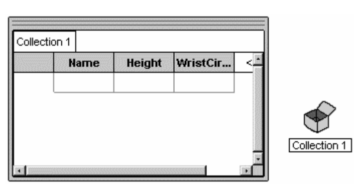


4) Click once on <new>.

5) Type Name for the first attribute and press **Enter**.

6) Make two more attributes: Height and WristCircum.

You now have a table and an iconified collection as shown below.



7) Make the WristCircum column wider by dragging its right edge.
8) Double-click the label Collection 1 in the title of either the table or the collection. This should bring up a dialog box for renaming the collection.
9) Type People and click **OK**.

Entering Data

You're ready to enter the names and numbers for our five people. (All measurements are in inches.)

10) Click in the cell below Name.
11) Type Bill and press **Tab**.
12) Enter the rest of the data shown in the table at right. Use **Tab** to move to the next cell. (At the end of the first case, **Tab** will take you to the first cell in the next case.)
13) Save this file to preserve the data.

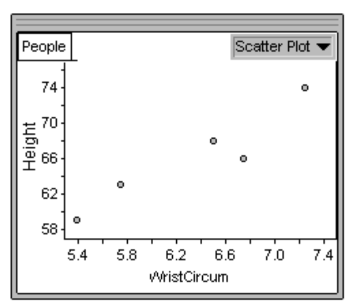
Name	Height	WristCircum
Bill	66	6.75
Jill	68	6.5
Denise	63	5.75
Tim	74	7.25
Lauren	59	5.4

Graphing Height Versus Wrist Circumference

Now let's look at the relationship between the two continuous attributes, Height and WristCircum.

The shortcut for **New Graph** *is* **Ctrl-G** *(Win)* **⌘-G** *(Mac).*

14) Choose **Graph** from the **Insert** menu (or drag the graph icon from the shelf to an empty area in the document).
15) Drag the column header for Height and drop it on the vertical axis of the graph.
16) Similarly, drag the column header for WristCircum and drop it on the horizontal axis.

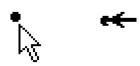


Your graph should look similar to the one shown above. Resize it to fill as much of the empty space as possible.

Dragging Data

You can drag data in a graph to change it, making it easy to observe the effect of changes in the data on graphs and analyses.

17) Move the mouse so that the tip of the arrow is on top of one of the data points.



Notice that the arrow's orientation changes from northwest to west. This is your clue that you're positioned to drag the point.

Why do we call Fathom "Dynamic Statistics" software? Whatever you drag in Fathom, everything else in your document changes during the drag. These dynamic changes allow you to see how things are connected and how they work.

Notice also that the status bar at the bottom left of the Fathom window shows the coordinates of the case.

18) Drag the data point.

As you drag, you should see the values for the two attributes changing in the case table.

19) Choose **Undo Drag** from the **Edit** menu.

The point returns to its original position.

Fathom has *unlimited undo*, allowing you to undo backward and redo forward through the changes you have made.

You can prevent data from being accidentally changed by selecting the collection and choosing **Lock Collection** from the **Data** menu. (If you try this, unlock it to continue the tour.)

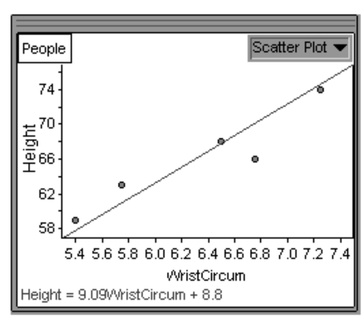
Fitting a Movable Line

Recalling that our goal is to be able to predict height from wrist circumference, and noticing that the points in the graph appear reasonably linear, we place a line in the graph.

You can also right-click (Win) or **Ctrl-** *click (Mac) on the graph to invoke the* **Graph** *menu.*

20) Select the graph by clicking on it once.

21) Choose **Movable Line** from the **Graph** menu.



The line that appears in the graph is *not* a fitted line. You can change its slope and intercept by dragging it. Dragging on one end of the line causes it to rotate around the other end. Dragging on the middle of the line moves it parallel to itself. The cursor changes shape to suggest what will happen when you drag. Notice that the equation of the line shown below the graph updates as you drag.

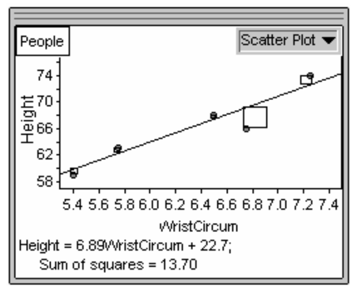
You might suppose that a line through this data should have zero for a *y*-intercept. You can try out this idea by choosing **Lock Intercept at Zero** from the **Graph** menu. You can uncheck it by choosing it again.

22) Experiment with dragging the line. Position the line so that it appears to give a good fit to the data.

While eyeballing a fit through data points is sometimes sufficient, we often need some criterion for a *best* fit. One commonly used criterion is a least-squares fit. You can see how least-squares works as follows.

23) Choose **Show Squares** from the **Graph** menu.

The graph now shows a square constructed from each point to the line, a square whose side length is equal to the difference between the actual and predicted value for the point—the *residual*.



24) Experiment with dragging data points.

Notice that the squares change as you drag, but the line does not move. (Be sure to use **Undo** to put the points back where they started before going on.)

25) Experiment with dragging the line. Notice that the squares change and that the sum of squares reported below the graph changes.

26) Adjust the line so that the sum of squares of the residuals is as small as possible.

The line that satisfies this criterion is called the *least-squares regression line*. Fathom can compute this line.

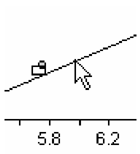
27) With the graph selected, choose **Least-Squares Line** from the **Graph** menu.

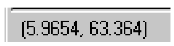
How closely did you manage to adjust the movable line to match the least-squares regression line?

Making a Prediction—Looking at Residuals

28) Choose **Movable Line** from the **Graph** menu (to uncheck it).

29) Measure your own wrist circumference in inches.

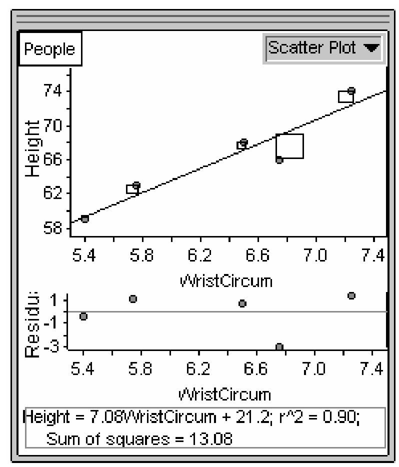




30) Move the mouse pointer along the least-squares line, noting that the coordinates of the tip of the arrow are reported in the status bar in the lower left corner of the window. When the *x*-coordinate of the mouse pointer is at your wrist circumference, you can read off a prediction for your height.

Probably this prediction was a bit off. Let's see how much off we might *expect* it to be.

31) With the graph selected, choose **Make Residual Plot** from the **Graph** menu.



A plot of the residuals appears below the main scatter plot. Corresponding to each point in the original graph is another point below whose *y*-value is the difference between the predicted value and the actual value (its vertical distance from the line).

32) Slowly drag one of the points in the top graph.

Notice how its residual changes in the residual plot. Notice also that, since the least-squares line is changing in response to dragging the point, the other residual points are changing as well.

The shortcuts are **Ctrl-Z** *(Win)* **⌘-Z** *(Mac) for* **Undo**, *and* **Ctrl-R** *(Win)* **⌘-R** *(Mac) for* **Redo**.

33) Use **Undo** to return the data to its original values.

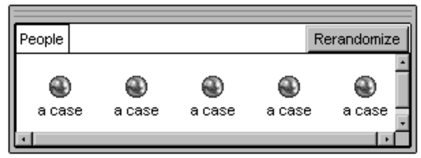
The residuals appear to lie roughly in a band from –3 to +1. How far off was the prediction of your height? Did it lie within that band?

34) Add yourself as a data point in the case table.

How much does this change the slope and intercept of the least-squares line?

Changing Cases' Captions

From the scatter plot alone, you can't tell which person is which, but Fathom lets you control how cases are *captioned*.



35) Make the collection larger until you see all of the cases, each of which is a gold ball labeled “a case,” the default.

36) Select **Name** in the case table.

37) Choose **Use as Caption** from the **Display** menu.

The cases are now labeled with each person’s name. When you hold the cursor over a data point, the caption appears in the status bar, bottom left of Fathom’s window.

38) Drag the right bottom corner of the collection up and to the left to iconify it again.

The Influence of an Outlier

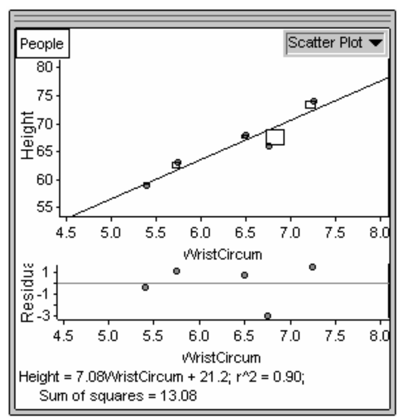
Let’s play around a bit with the idea that one data point may or may not have a significant influence on the regression line.

First, we need a little room to create an outlier.

The shortcut for zooming in is **Ctrl-click** *(Win)* **Option-click** *(Mac). Add the* **Shift** *key to zoom out.*

39) Manipulate the axes by dragging the upper and lower ends of each axis toward the middle.

You should have something similar to the graph shown at right.



40) Drag the rightmost data point away from the line.

Notice how wildly the slope and intercept of the regression line can change in response to changes in one of the points. A least-squares regression line is quite susceptible to outliers, making it especially important that you look at your data graphically before reporting a least-squares slope or intercept.

41) When you are finished with this file, close it.

Ideas for Further Exploration

- Where does the outlier have to be to make the least-squares line a negative slope? That is, how far out (and in what direction) does an outlier have to be to really mess you up?
- Which data point is most influential on the least-squares line?
- Add a median-median line to the graph. Compare how much it is influenced by changing data to how much the least-squares line is influenced by changing data.

To learn how to make attribute values depend on a formula, see page 42.

- Fathom has built-in functions for computing the slope and intercept of a least-squares regression line. You can use these to create a new attribute that computes the WristCircum values from the Height values. Your formula will look something like this:

```
linRegrSlope(WristCircum, Height) * WristCircum + linRegrIntercept(WristCircum, Height)
```

- Add another formula-driven attribute to compute the residuals. Find the standard deviation of the residuals using a summary table (see page 38 for how to use summary tables).

Tour 3: Analysis of Groups

Determining the differences and similarities between groups is a frequent theme running through data analysis and statistics. In this tour, we'll look at 735 highway accidents that took place in 1996 in Connecticut. Each accident resulted in at least one fatality. We will look at two questions: (1) Are people who wear seat belts more likely to escape injury in an accident than those who do not? (2) How does age vary among different groups?

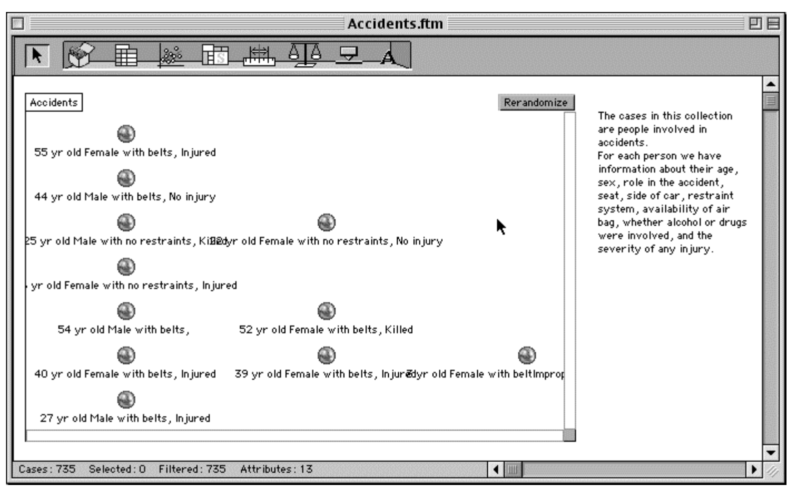
You Will Learn How to …

- filter a collection.
- make bar charts and ribbon charts.
- split bar charts and ribbon charts.
- make a summary table.
- compute quantities for each group in a summary table.

Looking at the Accident Data

Our data consist of the first 300 fatal crashes in Connecticut in 1996, compiled by the U.S. government's Fatal Accident Reporting System (FARS).

1) Open the **Accidents** file in the **Learning Guide Starters** folder. Save, move, and rename this file.



As usual, each gold ball represents one case—for this data, a person involved in an accident. The captions below the gold balls describe the person. The cases have been arranged so that each row represents one accident. If you haven't yet, maximize Fathom's window by clicking the control in the upper right corner of the window.

We need some screen space.

2) With the collection selected, choose **View in Window** from the **Display** menu to make the collection fill a separate window. Scroll around and look at some of the cases. When you're through, click its close box.

3) Drag the lower right corner of the collection frame toward the upper left corner until the collection shrinks into an icon. You can easily move it by dragging it from anywhere *except* its name or its edge.

Looking at Attributes

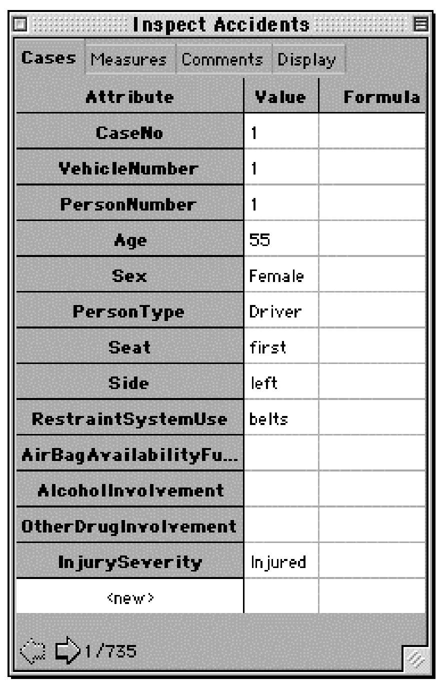
Let's see what attributes are in the collection and how we might use them to answer the two questions posed at the beginning of this tour.

4) Double-click the collection anywhere except its name (or press **Ctrl-I** (Win) **⌘-I** (Mac)).

   This brings up the inspector. In it you can see the list of attributes for cases in this collection.

5) Resize the inspector or scroll down until you find attributes called RestraintSy... and InjurySeverity. (You can resize the columns.)

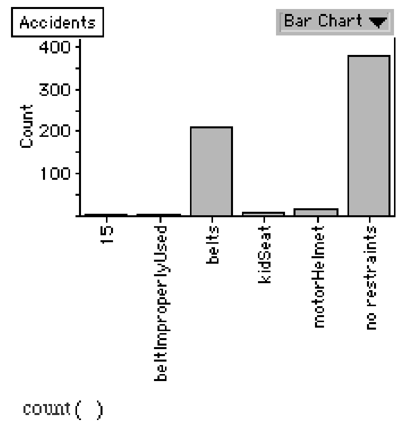
These two attributes, along with age, are the ones we'll use in this tour.



Making a Bar Chart

6) Insert a graph.

7) Drag the RestraintSystemUse attribute from the inspector onto the graph's horizontal axis and release.



In practice, we would look more carefully at the other categories before discarding them.

Your graph should look like the one at right. Notice that most of the cases are in two categories: belts and no restraints. The other four categories get in the way. We're going to *filter* out cases belonging to those four categories.

Filtering Data

Filters are expressions whose values are either true or false.

In Fathom, you can add filters to collections, tables, graphs, summary tables, estimates, and tests. We'll add one to the collection so that anything we make from that collection will only show the cases that pass through the filter.

8) Select the collection.

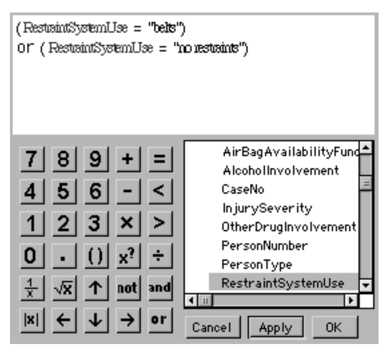
9) From the **Data** menu, choose **Add Filter**.

The keyboard shortcut for adding or removing a filter is **Ctrl-F** *(Win)* **⌘-F** *(Mac).*

The collection will expand, a line labeled <new filter> will appear underneath it, and the formula editor opens for you to put in an expression that will serve as a filter. (To edit an existing filter, double-click it.)

10) Type an open parenthesis (or use the formula editor's keypad).

Notice that you get both open and close parentheses, with the cursor between them.



11) Click the open/close control for Attributes in the formula list.

Notice that attribute names become colored in the expression.

You should see a list of attribute names from the collection.

12) Scroll down to RestraintSystemUse and double-click it.

This enters the attribute name in the formula editor without your having to type it.

13) Press =.

14) Type "belts" (with the quotes).

15) Press the closing quote and then the right parenthesis key to move the cursor to the right of the close parenthesis (or click there with the mouse).

16) Click the **or** button on the keypad in the formula editor (or type it).

17) Create the second half of the expression as you did the first, substituting "no restraints" for "belts".

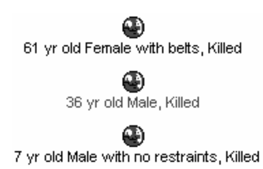
```
(RestraintSystemUse = "belts")
or (RestraintSystemUse = "no restraints")
```

18) Click the **OK** button in the formula editor.

Cases: 735, Selected: 0, Filtered: 593, Attributes: 13

When the cursor is over the collection, the information in the status bar in the lower left of the Fathom screen says that now only 593 out of the original 735 cases passed the filter. If you scroll through the collection, you will see that some of the cases have grayed-out captions. These are the ones that don't pass the filter.

19) Make the collection small again by dragging the lower right corner of the frame toward the upper left.

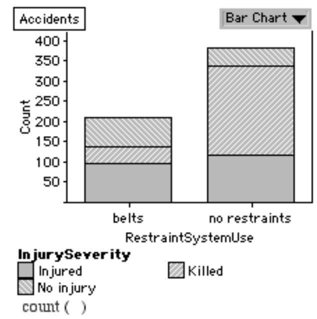


The bar chart should now have two bars, one for each of the categories we care about. Now we're ready to look at injuries in relation to restraints.

Splitting a Bar Chart

20) From the inspector, drag InjurySeverity to the *middle* of the graph (drop it when a big black box outlines the plot area).

When you release the mouse, you should see a *stacked* bar chart. Each bar has been split into three different regions, one for each of the three injury categories. The legend underneath the graph provides a key to the regions.



Notice that the portion of the no restraints bar that corresponds to killed is large, especially compared with that portion in the belts bar.

Making and Splitting a Ribbon Chart

Bar charts are good for comparing counts, but they are hard to use when what we care about is *proportions.* A ribbon chart works better.

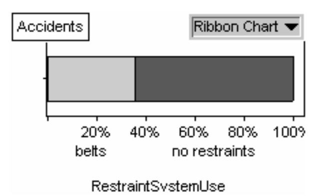
Leaving the old graph on the screen makes it easy to compare the two. You could, instead, change the existing graph, rather than making a new one.

21) Make a new empty graph.

22) Drag RestraintSystemUse to the horizontal axis of the new graph.

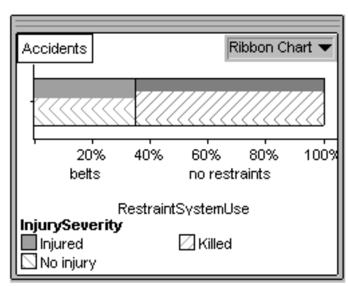
23) From the popup menu in the upper right corner of the graph, choose **Ribbon Chart**.

Notice that you can read the proportion of cases for belts right off the axis scale.



24) Drag InjurySeverity to the *middle* of the ribbon chart.

The ribbon chart splits so that each of the two main parts of the bar has three smaller regions, each corresponding to an injury category, as shown at right and described in the legend that appears below the graph.



The height of each small region corresponds to the proportion of its bar. So we can estimate that about half of those with no restraints were killed, while fewer than one-fifth of those using seat belts were killed.

This answers our first question. We conclude that, for this collection of accidents, your chances of getting killed were about two times greater if you did not wear seat belts than if you did.

Making a Summary Table

What about the second question: How does age vary among the groups? Neither bar charts nor ribbon charts will help us here. We'll use a summary table so that we can compute the median age for each group.

25) Choose **Summary Table** from the **Insert** menu or use **Ctrl-U** (Win) **⌘-U** (Mac).

26) Drag RestraintSystemUse to the right arrow and InjurySeverity to the down arrow.

Accidents	Summary Table			
		RestraintSystemUse		Row Summary
		belts	no restraints	
InjurySeverity	Injured	94	116	210
	Killed	43	221	264
	No injury	72	44	116
	Column Summary	209	381	590

S1 = count ( )

Your table should show counts of the number of cases in each group plus row and column totals. The number in the lower right corner is the total number of (filtered) cases.

By default, a summary table computes counts for categorical attributes, but you can change that or add any additional computations you want. If the summary table isn't already selected, click on it once.

27) Choose **Add Formula** from the **Summary** menu.

28) In the resulting formula editor, type median(age) and click **OK**.

Now each cell of the table has two numbers. The top number is the count for that cell, and the bottom number is the median age for that category of people. The table also has row and column summaries.

Accidents	Summary Table			
		RestraintSystemUse		Row Summary
		belts	no restraints	
InjurySeverity	Injured	94 33.50	116 22	210 27
	Killed	43 48	221 37	264 37.50
	No injury	72 37	44 30	116 33
	Column Summary	209 37	381 29	590 32

S1 = count ( )
S2 = median (age)

Practice reading this table by answering the following questions.

- Which group has a higher median age, those who wear belts or those who don't?
- Do the people who are killed tend to be older or younger than the people who are not?
- Which group has the highest median age? (Can you rationalize why this might be?)
- What is the overall median age for this filtered portion of the collection?

29) When you are through exploring, close this file.

Questions for Further Exploration

- Are drivers more likely to be killed than passengers? Even if they're wearing seat belts?
- Do air bags appear to change the chance of being killed?
- What can you say about pedestrians?
- Come up with some of your own questions that can be answered by these data. What questions do you have that are *not* answered by these data?

Tour 4: Formulas, Functions, and Data—The Planets

This tour focuses on using formulas to define new attributes and to plot functions. We'll look at a data set with only nine cases—the planets in our solar system. We'll compute the density of the planets and investigate the relationship between a planet's distance from the sun and the length of its year.

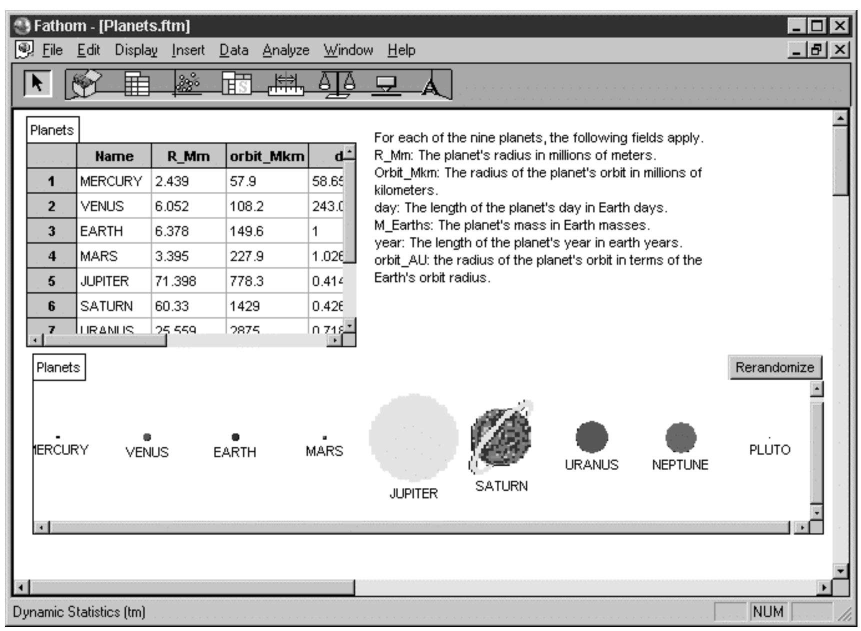
You Will Learn How to …

- define an attribute with a formula.
- plot a function in a scatter plot.
- make and use a slider.

Computing the Densities of the Planets

1) If Fathom is not already running, launch it.

2) Open the **Planets** file in the **Learning Guide Starters** folder in the **Sample Documents** folder. Your window should resemble the one below. Save, rename, move this file, and maximize the window.



First we'll compute each planet's density. For that we need to know its volume and its mass because density is the ratio of these values. The volume comes from the radius, the attribute R_Mm, measured in millions of meters. The mass, measured in Earth masses, is recorded in the attribute M_Earths. We'll compute density in two steps using two new attributes (though we could do it in one), beginning with volume. (You may want to hide the text object first, to give yourself some room. When it is selected, press **Ctrl-H** (Win) **⌘-H** (Mac).)

3) Resize the table to the right until you can see the column headed <new>.

You can also create a new attribute by selecting an existing attribute and choosing **New Attribute** *from the* ***Data*** *menu.*

4) Click once on <new>.

5) Type Volume and press **Enter**.

Now we need to tell Fathom the formula for volume.

6) From the **Display** menu, choose **Show Formulas**.

7) Double-click in the formula cell just beneath Volume.

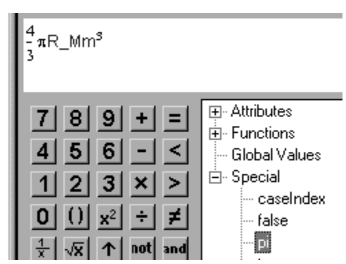
The formula editor appears. We're going to make our formula look like the one below. You can use your computer keyboard or the formula editor's keypad and list. Here are some tips.

The formula for the volume of a sphere is

$$V = \frac{4}{3}\pi r^3$$

but don't enter the left side and use R_Mm *instead of r.*

- From the keyboard, type * for multiplication, / for division, and ^ (Shift-6) for exponentiation.
- To escape the fraction, use either the right arrow button or a mouse click.
- Multiplication is sometimes indicated as a dot between terms and sometimes, as in traditional algebra, as nothing between terms.



- Enter π either by choosing it from the Special heading or by typing pi (with a space after).

Look up **Formulas** *in the help system for more information and hints.*

- Click the open/close control next to a heading in the attribute/function list to open the list of items under it.
- Double-click an item to enter it into the formula.
- Click an item to get an explanation of it in the help area at the bottom of the formula editor.
- Use the attribute R_Mm in place of *r*. You can get it from the attributes list or by typing it.

- Click the close control next to a header to close it.
- When you've typed an attribute correctly it will appear blue.

The up arrow shows you the structure of an expression.

- Use the arrow keys on the editor keypad or on your computer keyboard to move right or left. The up arrow selects the next largest portion of the expression. The down arrow selects a smaller term.

8) Close the formula window by clicking **OK** or pressing **Alt-F4** (Win) **⌘-W** (Mac).

You can resize the formula row by dragging down.

Fathom computes the volumes and displays them in the table in gray to distinguish them from ordinary data values.

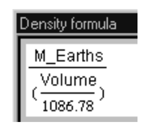
The next attribute has an easier formula:

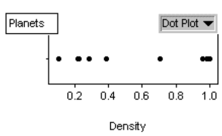
$Density = \frac{Mass}{Volume}$, but we want it relative to the earth's density.

9) Make a new attribute, **Density**.

1086.78 is the volume of the Earth in cubic megameters.

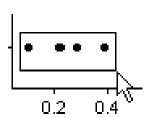
10) Give the new attribute the formula shown here. The numerator is the mass in Earth masses, and the denominator is the volume in Earth volumes, so the density comes out in Earth densities. This lets us understand the density of the planets relative to our own planet.





11) Make a graph with **Density** on the horizontal axis. It should look similar to the one at left.

When you click on a data point, the corresponding planet in the collection is selected. You can select more than one case at a time by clicking and dragging a rectangle around some points as shown below left.



Answer these questions.

- Which planet has the lowest density?
- What do the three planets on the right in the graph have in common?
- What do the four planets on the left have in common?
- What is special about Earth?
- If the Moon is a chunk of Earth, where should it lie in the graph? Look it up and see where it does lie.

Playing Kepler

Now we'll study one of the most famous relationships in physics—how the period of an orbit depends on the orbit's radius.

12) Make a scatter plot with **year** on the vertical axis and **orbit_AU** on the horizontal.

You should see points lying on a line gently curving upward. It looks as though we could predict how long a planet takes to go around the sun if we knew the radius of its orbit.

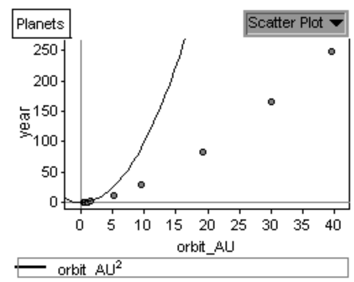
But what is the function? One way to find out is to try to plot it.

The graph has to be selected for the **Graph** *menu to appear.*

13) With the graph selected, choose **Plot Function** from the **Graph** menu.

14) Type in a plausible function, say, orbit_AU^2, and click **OK**.

It's not quite right. Let's change the exponent. We'll use a slider so we can drag continuously from one value to another.

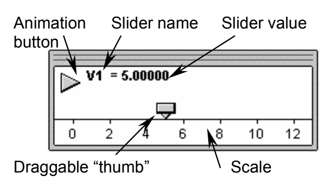


15) Choose **Slider** from the **Insert** menu.

A slider appears. The illustration below right shows its parts. By default the slider has the name V1.

16) Double-click the name and type something more reasonable, like P for power.

We think that the curve should be something between linear and quadratic, that is, between a power of 1 and 2. The slider scale is like a graph axis, so we can change it.



Zoom in with **Ctrl-** *click (Win)* **Option-** *click (Mac).*

17) Zoom in on the slider (or drag the axis) until the scale shows roughly one to two.

18) Double-click the formula below the graph, so that you can edit it. Replace the 2 in the exponent with P. Close the formula editor window.

As you drag the slider's thumb, the curve moves accordingly. Another way to change the slider value is to edit it directly.

19) Double-click the slider's value and type a new one, pressing **Tab** or **Enter** when you're done.

20) Find a value for P so the curve fits nicely through the points.

In Tour 2 you saw that you can make a residual plot for a linear fit. You can do the same with a function. Try it. Notice how small the residuals are. It's a very good fit!

You should find that $P = 1.5$ or $P = \frac{3}{2}$ fits very well. The curve's equation is year = radius$^{3/2}$, or, as Kepler would have written it, $T^2 = R^3$, where T is the length of the year and R is the planet's radius.

Kepler with Logs

Another way to figure out the relationship between period and radius of an orbit, is to take the log of both attributes.

21) Make two new attributes, named logorbit and logyear. (These are *names*, not *formulas*.)

logorbit	logyear
log (orbit_AU)	log (year)

22) Make their formulas log(orbit_AU) and log(year), respectively.

23) Make a scatter plot of logorbit versus logyear. It looks like a line!

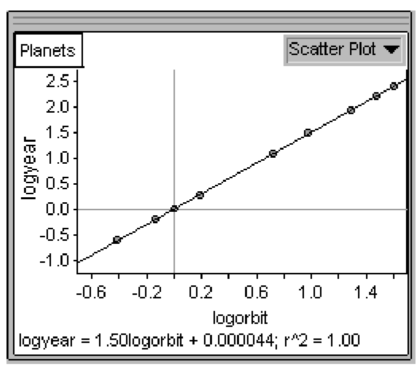
We can make some lines automatically.

24) Choose **Least-Squares Line** from the **Graph** menu.

Check out the slope. Look familiar?

If $T^2 = R^3$, then $\log(T^2) = \log(R^3)$. So $2 \log(T) = 3 \log(R)$. This means that Kepler's law implies that the slope of log(orbit) versus log(year) should be 1.5, and we see that it is.

As you can see in the graph below, there is a gap between Mars and Jupiter. What goes there?



Close the file when you're through.

Tour 5: Generating Mathematics—Triangular Numbers, Fibonacci, and Chaos

This tour also focuses on using formulas to define new attributes and to plot functions. But this time we'll work in a purely mathematical context, defining sequences recursively by writing formulas that refer to the previous case's values.

You Will Learn How to …

- use Fathom to generate a collection's data.
- define a formula that refers to other cases.
- use an if( ) construction.

Computing the Triangular Numbers

The triangular numbers are the sequence {1, 3, 6, 10…}. You can also think of the *n*th triangular number as the sum of the first *n* integers. But in this tour, we will use an equivalent, recursive definition: The *n*th triangular number is the (*n* – 1)st triangular number plus *n*.

1) If Fathom is not already running, start it by double-clicking its icon. If necessary, choose **New** from the **File** menu to get a fresh document. Maximize the window.

2) Make a new case table by choosing **Case Table** from the **Insert** menu or by dragging one off the shelf.

3) Click the <new> attribute header and type Tri (for triangular) to name our attribute. Press **Enter** to accept the name.

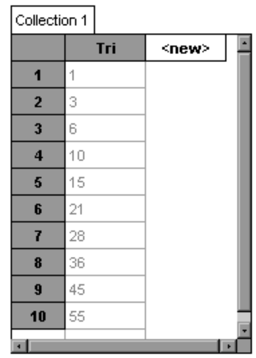
4) Select the new Tri attribute label and choose **Edit Formula** from the **Edit** menu. The formula editor opens. (In the previous tour, we chose **Show Formulas** from the **Display** menu to get access to formulas. This is another technique.)

The special constant caseIndex *is the number of the case, that is, the row number.*

The expression prev(Tri) *means "the value of* Tri *from the previous case."*

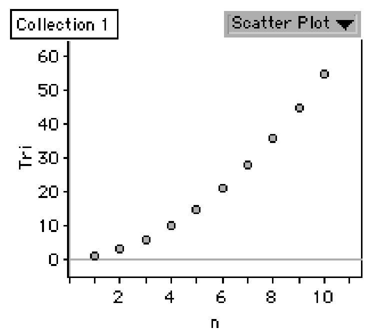
5) Enter **caseIndex + prev(Tri)**. Click **OK** to close the editor. Now we have a formula for our attribute, but no cases.

6) With the case table selected, choose **New Case(s)**... from the **Data** menu. Type 10 and click **OK**. Ten cases appear; your case table should have 10 values as shown.



7) Make a new graph and drag Tri to the *vertical* axis. You'll see a dot plot. Change the graph to **Line Plot**, and you'll see the parabolic shape of this sequence.

Notice that we cannot make a scatter plot. That's because we don't have an attribute for the horizontal axis. We'll make one now:



8) Make another new attribute; name it n.

9) Give it the formula: caseIndex. Just that one word. When you close the formula editor, you'll see that n has the row numbers in it.

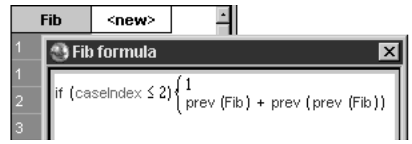
10) Now drag n to the horizontal axis.
You get the scatter plot you expect.

At this point, you could plot a function as we did in the last tour; you could try to see if you can figure out the function that goes through the points. Alternatively, you might remember that the sum of the first n integers is $n(n + 1)/2$ and see if that function fits.

Fibonacci Numbers

Now we'll enhance our collection to make the Fibonacci numbers as well as the triangular numbers. The rule for the Fibonaccis is that each number is the sum of the two previous numbers. There's an additional rule to get us started: The first two numbers are both 1.

11) Make a new attribute, Fib. Give it the formula in the illustration. Here are some hints about working with if-statements:



- The first set of parentheses encloses a *condition*—an expression that's either true or false. This one is true for the first two cases.
- When you type if(, Fathom fills in the close parenthesis with the brackets and question marks for the two clauses.
- Make "less-than-or-equal-to" by holding down **Ctrl-Shift** (Win) **Option-Shift** (Mac) and pressing **<**.
- After the brackets, the top expression is what you get if the condition is true. You get the bottom one if it's false.
- Press **Tab** to move from clause to clause.

This construction is typical for a recursive definition: Use caseIndex in an if-statement to define the initial condition. One fork of the if defines the initial values; the other defines the recursive step. Also, notice how prev(prev( )) works to give us the one-before-last that we need for Fibonacci numbers.

We didn't need this construction in the triangular numbers because, by default, Fathom defines prev(the first case) *to be 0.*

12) Make a scatter plot of Fib as a function of n. It has a characteristic exponential shape.

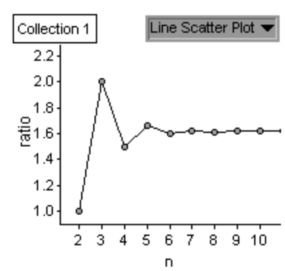
Let's figure out a formula that approximates the Fibonacci numbers. If it's exponential, the ratio of adjacent terms will be constant.

13) Make another new attribute and call it ratio.

14) Give it the formula Fib/prev(Fib).

The first value is Inf *because the* prev *of the first case is 0.*

15) Make a graph of ratio as a function of n. Make that graph a **Line Scatter Plot** so you can see the trends better.



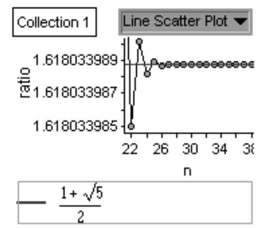
From the graph and table it looks as if ratio converges to some number between 1.6 and 1.7. Could this be the golden ratio? That ratio (called Φ, or phi) is

If you add more cases and zoom in, you can see this graph shape, alternating up and down, getting closer and closer, persists even at very small scales.

$$\Phi = \frac{1+\sqrt{5}}{2}$$

16) Click on the graph and choose **Plot Function** from the **Graph** menu. Enter the golden ratio, root and all (click the root key on the formula editor's keypad or press **Ctrl-Shift-R** (Win) or **Option-Shift-R** (Mac) for the root to make the square root). Click **OK** when you're done to make the function show.

17) The function is a constant, of course—a horizontal line. But you can see that the ratio seems to approach it, as shown in the illustration. (We added more cases to show you how precise this convergence is.)



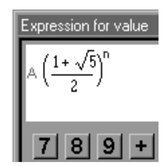
That means that we can approximate the Fibonacci numbers with a function like this:

$$Fib(n) = A\Phi^n$$

18) Make a new slider and rename it A. Set it equal to 1.0 for now. It will be our coefficient.

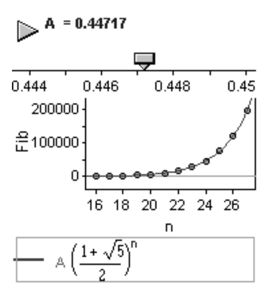
If you right-click (Win) **Ctrl-***click (Mac) on the graph, you can select* **Plot Function** *from the ensuing menu.*

19) In the scatter plot of Fib as a function of n, choose **Plot Function** from the **Graph** menu. Enter the function shown at right. Here are some tips:



- Be sure A is outside the parentheses.
- Select the 1 and the root before pressing / in order to divide them both.
- Be sure that the whole fraction is selected (or your cursor is just outside the parentheses) before pressing ^ in order to exponentiate the whole thing.

20) Play with the slider for A to get a good fit. We get A = 0.44717. You may need to zoom in to the slider axis to get fine enough control, especially if you have added more cases. A residual plot helps, too.



Alternatives

We could have taken a different path to find the function. For example, we could have found the logarithm of the Fibonaccis and fit a straight line to them. Analyzing that equation would have led us to the value of the coefficient and of Φ.

We also could have avoided re-entering **Φ** by making a measure in the collection named phi, and defining it by the formula. Then we could have used phi in the formula for the curve.

Logistic Chaos

This logistic *situation is one model for population growth where it cannot expand boundlessly, for example, if there is limited space or a limited food supply.*

It's a surprisingly short step from where we are to chaos. We'll study a famous recursive definition for population:

$$pop_n = R(pop_{n-1})(1 - pop_{n-1})$$

where *R* is a parameter of the system. Once again, we need an initial condition for *pop*. For now, we'll make its initial value 0.1.

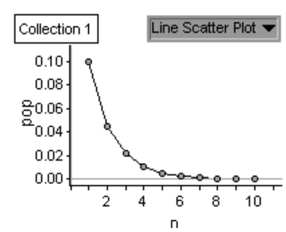
In this model, the value for *pop* is a decimal between 0 and 1. It is the size of a population compared to its maximum possible size. (You may want to free up some screen space before proceeding.)

21) Make a slider named R with a value of 0.5.

22) Make yet another new attribute in our collection, this time named pop. Give it the following formula (be sure to press * after you type R so Fathom knows you're multiplying):

$$\text{if (caseIndex} = 1) \begin{cases} 0.1 \\ \text{R prev(pop)} \cdot (1 - \text{prev(pop)}) \end{cases}$$

23) Make a **Line Scatter Plot** of pop as a function of n. The first 10 cases look something like the illustration. The population declines.



24) Play with the slider for R so you get a constant population of 0.1. This will be at about R = 1.1. Explain why 1.1 gives a constant population of 0.1.

25) Set the slider axis to extend from about 1 to 4, and set the vertical axis of the scatter plot to extend from 0 to 1.

26) Be sure you have at least 40 cases. (Use **New Cases** in the **Data** menu.)

27) Increase R slowly and observe what happens. You should get smooth behavior before R = 3, then oscillation, and, finally, stranger and stranger behavior beyond about R = 3.5.

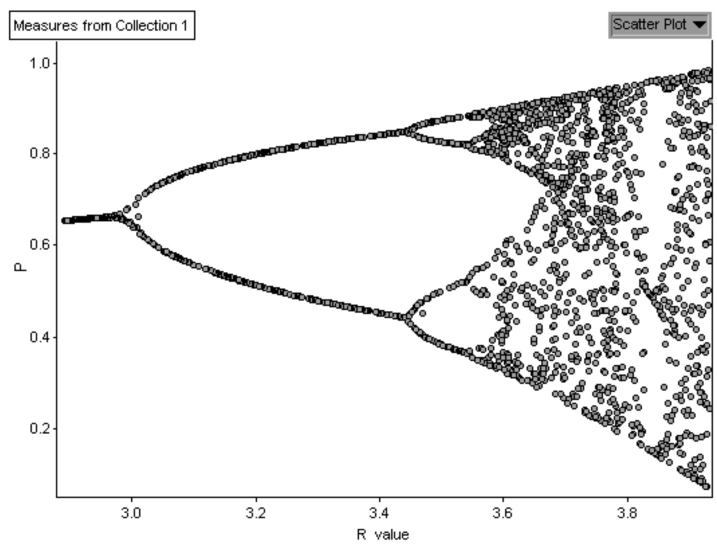
This is the chaotic regime of this equation. While before it was smooth and predictable, here it seems to jump around strangely, looking less like mathematics and more like real data.

Ideas for Further Exploration

- Create a slider to be the initial value of pop. With R < 2.5 or so, what effect does that initial value have on the *equilibrium* values after about n = 10?
- What effect does the initial value have on the eventual values when R is in the chaotic regime?
- Why does the function become less interesting with R > 4.0?
- Make a new plot, but this time put pop on the vertical axis and leave the horizontal axis empty. To get beyond the initial values, add a filter to the graph. Enter n > 20. (Now you're looking at only the last 20 cases.) For what values of R does the function converge to one value? Two values? Four? And the toughie, *three* values? (Check out R = 3.84015.)

This challenge requires collecting measures. Learn about them in the next tour.

- **Challenge**: Make this graph of the 100th iteration versus R:



Tour 6: Simulation—Polling Voters

In this tour, we simulate a population of voters, a certain proportion of whom will vote in favor of a particular proposition. We investigate the question of how accurately a random sample of voters can predict the outcome of the election.

You Will Learn How to …

- use a slider to define a population parameter.
- define an attribute based on the random function.
- define a measure.
- collect measures.

Defining the Problem

The city of Freeport has a rent control initiative, Prop A, on the ballot. The local newspaper is going to conduct a poll three weeks before the election to gauge public sentiment. Staff members need to know how big a sample to poll. Your job is to set up a simulation they can use to determine, for any given sample size, the accuracy they should expect.

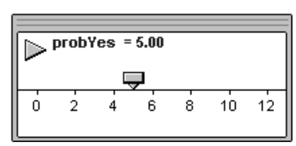
Modeling Voters

First we model the population, the people who will vote in the election either for or against the proposition. This model will consist of a single number, the proportion of voters who will vote yes.

1) Start with a new, empty Fathom document.

2) From the shelf, drag a new slider into the document.

By default a slider is named V1. We want the slider to have a more meaningful name.



3) Double-click the name portion of the slider, the V1, and type probYes and press **Enter**.

Since a proportion of voters can only lie between zero and one, we need to adjust the slider scale.

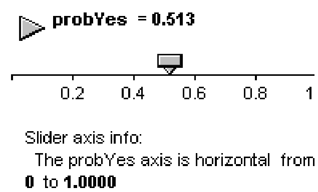
4) Drag the numbers on the slider scale until the scale goes approximately from 0 to 1.

Notice that dragging the slider scale is just like dragging the scale on a graph's numeric axis.

For the purposes of this model it isn't important that the scale go exactly from 0 to 1, but sometimes, especially when we're animating, we do care. So let's look at a more exact way to set the axis scale.

5) Double-click the numbers along the scale of the slider.

Below the slider you should see a control text object for the slider.



6) Double-click the lower bound number and change it to 0.

7) Tab to the upper bound number and change it to 1.

You should have something similar to what is shown at right.

8) Click once on Slider axis info to select it.

9) From the **Edit** menu, choose **Delete Control Text**.

10) Set the value of the slider (by dragging its thumb) to something near 0.5, simulating a close election.

The slider models the population. A collection will model our sample of voters.

11) Drag a collection from the shelf into the document.

The collection will be empty and named Collection 1.

12) Double-click the name below the box. Use the resulting dialog box to name the collection Sample of Voters.

We want this collection to contain 100 cases for starters. Later we can investigate other sample sizes.

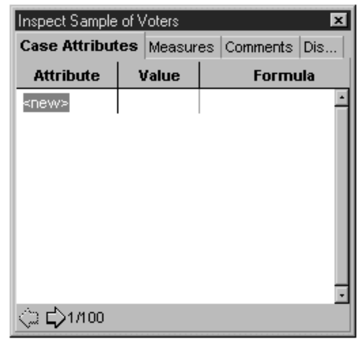
13) With the collection selected, choose **New Case(s)** from the **Data** menu.

14) Type 100 in the resulting dialog box and click **OK**.

So far, the cases in the collection have no attributes. We want one attribute, vote, whose value is either yes or no.

15) Double-click the collection.

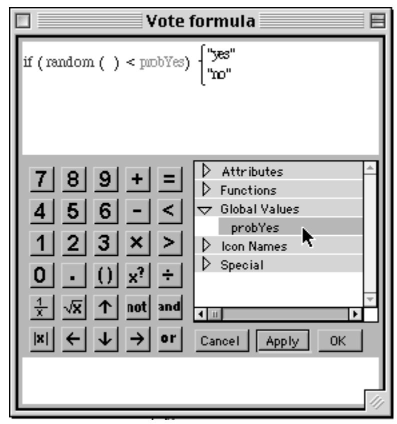
This brings up its inspector. In it you will define an attribute for a vote.



16) Click <new>, type Vote, and press **Enter**.

17) Double-click in the formula cell in the right column.

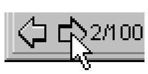
You should see a formula editor. Now you need to enter the formula that will choose "yes" the percentage of the time specified in your slider, probYes.



18) Enter the conditional formula shown in the editor at right by typing if(random()<probYes **Tab** "yes" **Tab** "no".

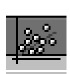
The random function, if you don't give a minimum and maximum, generates random numbers between zero and one.

19) Close the formula edit window by clicking the **OK** button.



You should see the value of the Vote attribute in the inspector become either yes or no. Scroll through a few more of the cases by clicking the right arrow in the bottom left corner of the inspector. Some of the values should be yes and some should be no.

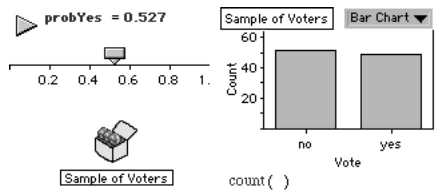
Let's make a graph of the Votes. A bar chart will show us at a glance whether the poll predicts the proposition will pass or fail.



20) Drag a graph from the tool shelf into the document.

21) Drag the vote attribute name from the inspector onto the *x*-axis of the graph.

You should now have the three objects shown at right.



The shortcut for **Rerandomize** *is* **Ctrl-Y** *(Win)* **⌘-Y** *(Mac).*

22) Choose **Rerandomize** from the **Analyze** menu. Do this several times.

Each time you rerandomize, the bars in the bar chart will change, reflecting the results of a new sample.

Another way to rerandomize is to drag the slider thumb. When the slider value is near one, most of the votes will be "yes," and when the slider value is near zero, most of the votes will be "no."

23) Move the slider back to something close to 0.5.

Simulating Repeated Surveys

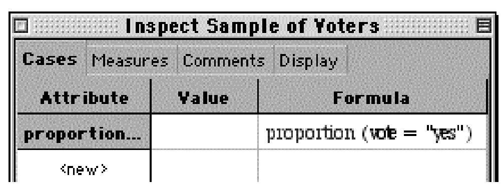
We've been hired to predict the accuracy of a poll conducted with a given sample size. To figure that out, we're going to have to record the prediction of each poll we take. We do that with a *measure*.

24) In the inspector, click on the **Measures** tab.

As yet no measures have been defined.

25) Click on <new> and type proportionOfYes for the measure's name.

26) Double-click in the formula cell in the proportionOfYes row.



Your inspector won't look exactly like this one because we've adjusted the size and column widths.

27) Enter the formula proportion(vote = "yes") as shown at right.

28) Rerandomize several times and observe the values for proportionOfYes.

Now we're ready to perform the poll of 100 voters many times, each time recording the proportion of voters that say they will vote in favor of the proposition.

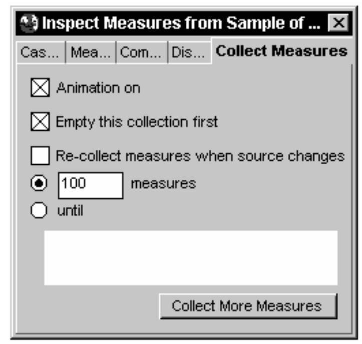
29) Close the Sample of Voters inspector.

30) With the collection selected, choose **Collect Measures** from the **Analyze** menu.

You should see a new collection, as shown above right. This collection will contain the results of repeatedly conducting the poll. Fathom has rerandomized the values in the *source collection,* each time calculating the measure, and storing that value as a case in the measures collection. By default, it collects five measures.

31) Make a case table of the measures collection.

32) Double-click the measures collection and click on the **Collect Measures** tab.



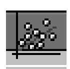
33) Change the number in the measures field from 5 to 100.

34) Click **Collect More Measures**.

If collecting the measures takes too long, press the **Esc** *key. You can speed things up by turning off animation.*

Fathom rerandomizes the Sample of Voters collection 100 times, each time computing a value for proportionOfYes and storing that value in a new case in the measures collection.

Let's take a look at the results.



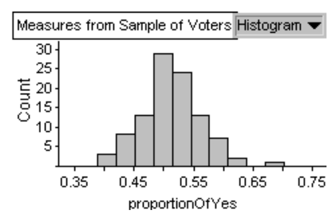
35) Drag a new graph from the shelf to the document.

36) Click on the **Cases** tab in the measures collection inspector.

37) Drag the proportionOfYes attribute from the inspector to the *x*-axis of the graph.

38) From the graph's popup menu, change the graph from a dot plot to a histogram.

An example of what you might get is shown at right.



Here are two observations you might make about the histogram of proportionOfYes.

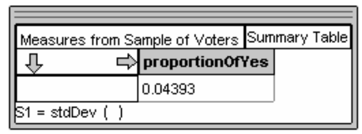
- The distribution looks plausibly normal. The central limit theorem predicts that as sample size increases, the distribution of proportions approaches normal, so we're not too surprised.
- It looks like the mean of the sample proportions is about 0.5. Since the probYes slider was set near 0.5, this bears out another prediction of the central limit theorem—that the mean of the sample proportions equals the population proportion.

The spread of proportions that we get with a sample size of only 100 is quite large. We certainly couldn't accurately predict the results of a close election! Just how large is this spread?

39) From the **Insert** menu, choose **Summary Table**.

40) Drag the proportionOfYes attribute from the measures collection into the top row of the summary table.

41) Double-click in the formula row where it says S1=mean() and replace it with stdDev(). This will compute the standard deviation of the proportions gathered in the measures collection.



Let's try a larger sample size.

42) With the Sample of Voters collection selected, choose **New Case(s)** from the **Data** menu. In the dialog box, ask for 300 new cases and click **OK**. The collection will contain 400 cases altogether.

Make sure you have unchecked the **Animation on** *option.*

43) Bring back the **Measures** pane in the measures collection inspector. Click the **Collect Measures** button to collect another 100 measures.

We expect that the larger the sample the smaller the variation in the observed population proportion in favor of Prop A. But *how much* smaller? What do you find for the standard deviation of sample proportions with a sample size of 400? What do you think would be the standard deviation if the sample size were 1000?

You now have a tool to use with the newspaper staff members. You could sit down with them and try different sample sizes until you found a sample size for which the variability is less than some desired number of percentage points.

Ideas for Further Exploration

- When we actually conduct a poll, we do not know the true proportion of the population. (After all, that's why we're doing the poll.) Probability theory predicts that we can expect to capture the true proportion in the range $\pm\sqrt{\frac{p(1-p)}{n}}$ where p is the proportion found in the sample and n is the number in the sample. Model this in Fathom and see how close your simulation comes to theory.

- We have set up this simulation to make it easy to change the population proportion. With small sample sizes, population proportions near the extremes, that is, near 0 or 1, are problematic. Investigate why this might be so.

Tour 7: Estimating Confidence Using the Bootstrap

Suppose we have a simple random sample from a population and we have measured something about the sample. What can we say about the population? In this tour we look at some measurements of body temperatures from 20 people as measured with two different kinds of thermometers. It becomes clear that the two thermometers give different measurements. With just these 20 measurements what can we say about the difference between the two kinds of thermometers?

You Will Learn How to …

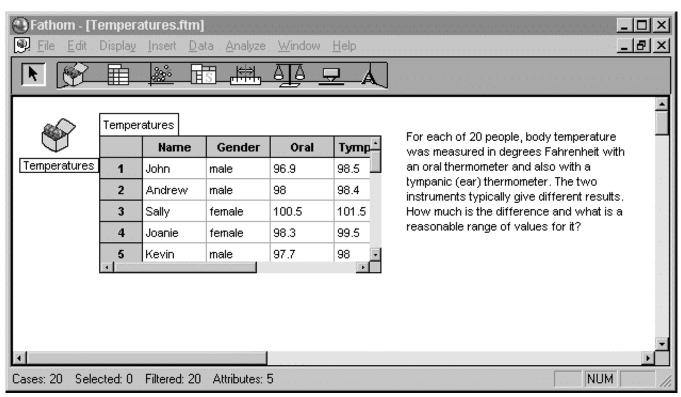
- fit a median-median line.
- do repeated sampling to estimate confidence in a measurement.

Looking at the Data

1) Open the document **Temperatures** in the **Learning Guide Starters** folder in the **Sample Documents** folder. Save, rename, and move this file.

This document contains its own documentation in a text object. It states our problem well. But, before we tackle it, let's look a bit at the data.

After reading the text, you may want to hide it.



2) Make a scatter plot of Oral versus Tympanic. (See "Graphing Two Numeric (Continuous) Attributes" on page 20, if you need instructions on how to make scatter plots.)

3) Fit a least-squares regression line through the data. (Choose **Least-Squares Line** from the **Graph** menu.)

4) Fit a median-median line through the data. (Choose **Median-Median Line** from the **Graph** menu.)

The least-squares line is "influenced" by outliers while the median-median line is "resistant" to their influence.

The two lines have somewhat different slope. The least-squares line appears to be strongly influenced by the points in the upper right.

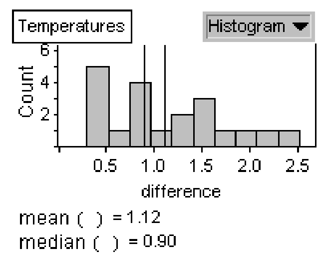
5) Drag one of the data points in the upper right and observe the difference between the response of the least-squares line and the response of the median-median line.

6) Use **Undo** to return the data point to its original values.

Measuring the Difference

The last column in the data table is difference, and it is defined formulaically as Tympanic – Oral. Scanning through this column, we see that all the differences are positive. We have no doubt that there *is* a difference, but we need to estimate a range of values within which it lies.

7) Make a histogram of the difference attribute.



This distribution is asymmetric with a tail to the right.

8) With the graph selected, choose **Plot Value** from the **Graph** menu.

9) In the resulting formula edit window, type mean() and close the window.

10) Repeat the previous two steps for the median.

As shown in the graph above, the mean and median for the difference of temperatures are separated by about two-tenths of a degree. Since the distribution of differences is skewed rather than symmetric, it makes sense to report the median rather than the mean as our estimate of the difference.

Estimating a Range for the Median

When we tell someone that the median difference in temperatures for the two thermometers is 0.90 degrees, we also need to tell them a range of values for the median that shows how certain or uncertain we are about it. There is a very general, powerful method, called the *bootstrap,* that we can use to do this.

This technique is known as the bootstrap because it feels a bit like we're pulling ourselves up by our own bootstraps.

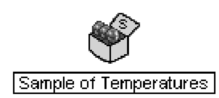
The idea is to create samples with replacement from our original sample and the same size as our original sample. Using replacement, the samples we create will, in general, be different from each other. The variability we observe among such samples is a good estimate of the variability that exists in the population.

In classical statistical methods, we typically make assumptions about the population, e.g., that values in it are normally distributed and that their standard deviation is approximately the same as the standard deviation we observe in the sample. In the bootstrap, the only assumption we make is that the sample values are representative of the population as a whole.

If you haven't selected the collection (box with gold balls), the **Sample Cases** *command won't be enabled.*

11) Select the Temperatures collection and choose **Sample Cases** from the **Analyze** menu.

You will get a new collection labeled Sample of Temperatures. It contains ten cases chosen at random from the Temperatures collection. We need to increase the sample size.

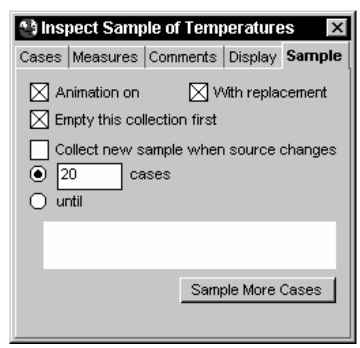


12) With the sample collection selected, choose **Inspect Collection** from the **Edit** menu.

13) Click the tab labeled **Sample** in the upper right to bring the controls for sampling to the front of the inspector.

14) Change the number of cases to sample from 10 to 20.

15) Click **Sample More Cases**.



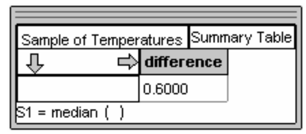
We've set things up so the sample has the same number of cases as the original collection. But because we're sampling *with replacement,* the particular cases that come into the sample will vary; thus, the median will vary as well.

16) Choose **Summary Table** from the **Insert** menu.

17) Click the leftmost tab in the inspector to bring the **Cases** pane to the front.

18) Drag the difference attribute from the inspector to the column header of the summary table.

19) Change the formula from mean() to median() (double-click the formula to edit it).

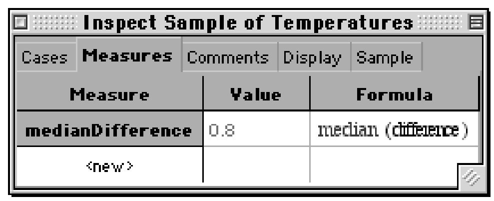


20) Press the **Sample More Cases** button in the **Sample** pane of the inspector.

Notice that each time you press the **Sample More Cases** button you get a new median for the difference attribute. It is this variability we would like to study.

21) Choose the **Measures** tab in the Sample of Temperatures inspector.

22) Click <new> in the **Measure** column. Type MedianDifference (**Enter**) to name the new measure.



23) Double-click in its formula cell to bring up the formula editor. Enter the formula median(difference).

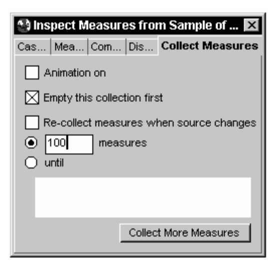
The new attribute MedianDifference is a *measure* of the sampled collection. We ask Fathom to collect this measure.

24) With the *sample collection* selected, choose **Collect Measures** from the **Analyze** menu.

A third collection, labeled Measures from Sample of Temperatures appears. It contains five cases whose values are the medians of the differences in each of the five samples.

You can speed up collection of measures by turning off animation in both the **Measures** *and the* **Sample** *collections.*

25) Double-click the measures collection to bring up its inspector.



26) Click on the **Collect Measures** tab of the inspector.

27) Type 100 in the field for number of measures.

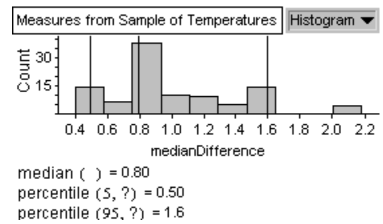
28) Click **Collect More Measures**.

After the 100 measures have been collected, you can view them in a graph.

29) Drag a new graph from the shelf into the document.

30) In the measures collection inspector, click on the **Cases** tab and drag the MedianDifference attribute from the inspector to the *x*-axis of the graph.

31) Use the popup menu in the graph to change the dot plot into a histogram.



This distribution of medians allows us to estimate the uncertainty in our measurement of the median difference. We notice that this distribution—like its parent, the distribution of differences—is not even symmetric, let alone normal, so a standard deviation will not be a particularly useful measure of spread. Instead, as shown above, we plot the median, the percentile at 5 and the percentile at 95. Ninety percent of the sample medians lie between these two percentiles, that is, between 0.5 and 1.6. We can say that this is a *plausible interval* for the median of the population.

Statistics has a precise language for describing estimation on population parameters. The bootstrapping method used here does not map exactly onto methods used to compute confidence intervals, so we are not free to report the interval we found as a 90% confidence interval.

Ideas for Further Exploration

- Use the techniques developed here to estimate the uncertainty in the mean of the difference.
- The distribution of median differences does not appear to be normal, or even symmetric. This means that our measured value of 0.8 does not lie at the center of our interval that contains 90% of the medians. Using the bootstrap technique on the median for other attributes in other collections, determine whether this asymmetry is typical or not.

Tour 8: Testing a Hypothesis

Charles Darwin believed that there were hereditary advantages in having two sexes in both the plant and animal kingdoms. Some time after he wrote *Origin of Species,* he performed an experiment in his garden at Down House in Kent. He raised two large beds of snapdragons, one from cross-pollinated seeds, the other from self-pollinated seeds. He observed, "To my surprise, the crossed plants when fully grown were plainly taller and more vigorous than the self-fertilized ones." This led him to another, more time-consuming experiment in which he raised pairs of plants, one of each type, in the same pot and measured the differences in their heights. He had a rather small sample and was not sure that he could safely conclude that the mean of the differences was greater than zero. His data for these plants were used by statistical pioneer R. A. Fisher to illustrate the use of a *t*-test.

You Will Learn How to …

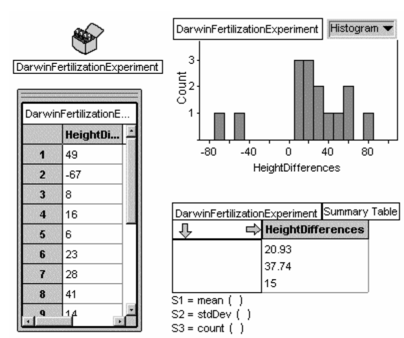
- perform a *t*-test for the mean.
- generate simulated data from a normal distribution.
- get a distribution for the *t*-statistic.
- use a text object to document your investigation.

Looking at Darwin's Data

1) Open the file **Darwin** in the **Learning Guide Starters** folder in the **Sample Documents** folder.

This document contains the data for the experiment described above: one attribute, fifteen cases.

2) Make a case table, a histogram, and a summary table similar to those shown here.

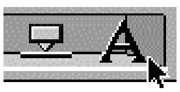


We see that most of the measurements are greater than zero, meaning that the cross-pollinated plants grew bigger. But two of the measurements are less than zero.

Darwin did not feel justified in tossing out these two values and was faced with a very real statistical question.

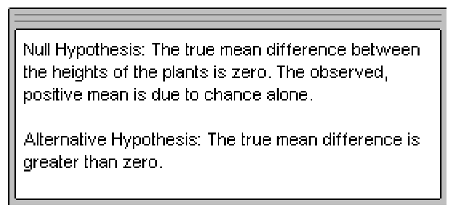
Formulating a Hypothesis

Darwin's theory—that cross-pollination produced bigger plants than self-pollination—predicts that, on the average, the difference between the two heights should be greater than zero. On the other hand, it might be that his fifteen pairs of plants have a mean difference as great as they do—21 eighths of an inch—merely by chance. You can write out these two hypotheses in Fathom, to be stored with your document.



3) From the shelf, drag a text object into the document.

4) Write the null hypothesis and the alternative hypothesis. At right you can see one way to phrase them.



Deciding on a Test Statistic

Not a statistician, Darwin decided to ask the advice of his cousin, Francis Galton, an eminent statistician, who told him that there was currently no very good theory to deal with a small sample from a population whose standard deviation is not known. In fact, it was not until some years later when William Gosset, an employee of the Guinness Brewing Company and a student of Karl Pearson, developed a statistic and its distribution. Gosset published his result under the pseudonym Student and the statistic became known as Student's *t*. When the null hypothesis is that the mean is zero, the *t*-statistic is just

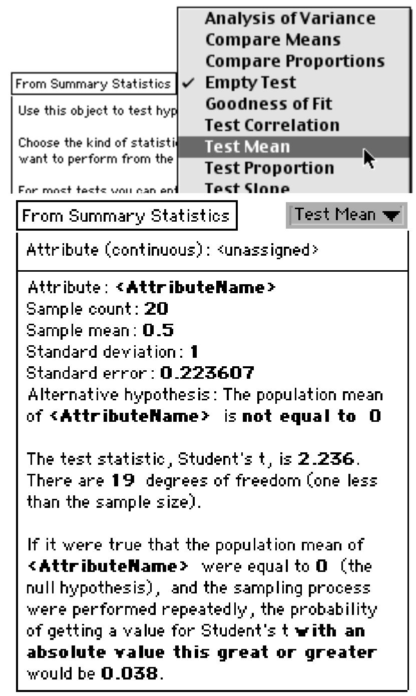
$$\frac{\bar{x}}{s/\sqrt{n}}$$

where $\bar{x}$ is the observed mean, s is the sample standard deviation, and n is the number of observations.

Let's compute this statistic for Darwin's data.

You can also drag a test object from the shelf (the balance icon).

5) From the **Analyze** menu, choose **Test Hypothesis**.



An empty test appears.

6) From the popup menu in the upper right, choose **Test Mean**.

As shown at right, the Test Mean test assumes that you are going to type in summary statistics. The blue text is editable. This is very useful when you don't have raw data.

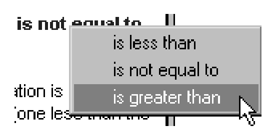
7) Try editing the blue text. You can, for example, enter the summary statistics for Darwin's data.

Here are some things to notice.

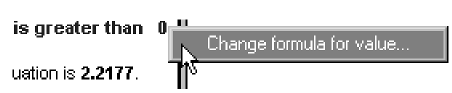
- When you change something in one part of the test, it may affect other parts. For example, editing the <AttributeName> field in the first line also changes it in the hypothesis line and the last paragraph.

In statistics terminology, we want this to be a one-tailed rather than a two-tailed test.

- In the hypothesis line, clicking on the is not equal to phrase brings up a popup menu from which you can choose one of three options. For Darwin's experiment, we want the third option because his hypothesis is that the true mean difference is greater than zero. Notice that making this change alters the phrasing of the last line of the test as well.



- The editable numbers are determined by formula. When you click them, you get a popup menu as shown at right. Choose the one available option and you get a formula editor. You can type a number or a formula here. This is particularly useful when you want to be able to change the test mean with a slider.



Checking Assumptions

Gosset's work with the *t*-statistic relied on an assumption about the population from which measurements would be drawn, namely, that the values in the population are normally distributed. Are we comfortable with this assumption for Darwin's data?

In the assumption's favor is experience with height measurements of other living things, both plants and animals. These are usually normally distributed, and so are differences between heights. But we might worry, because the two negative values give a decidedly skewed appearance to the distribution.

Fathom can help by allowing us to determine qualitatively whether this amount of skew is unusual or not. We'll generate measurements randomly from a normal distribution and compare the results with the original data.

You can make a new attribute by clicking the <new> *column header in the case table.*

8) Make a new attribute in the collection. Call it SimHeight for simulated height.

9) Select the name of the new attribute in the case table.

10) Choose **Edit Formula** from the **Edit** menu to bring up a formula editor.

11) Enter the formula shown here.

```
randomNormal (mean (HeightDifferences), stdDev (HeightDifferences))
```

This formula tells Fathom to generate random numbers from a normal distribution whose mean and standard deviation are the same as in our original data. We want to compare the distribution of these simulated heights with the distribution of the original data.

12) Make a histogram of this attribute.

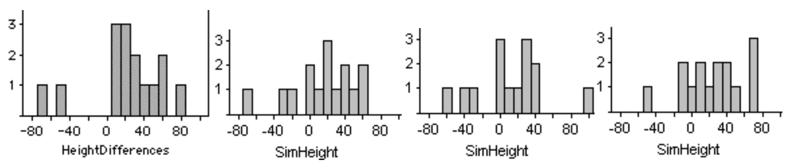
One set of simulated data doesn't tell the whole story. We need to look at a bunch.

13) With the graph selected, choose **Rerandomize** from the **Analyze** menu.

The shortcut for rerandomizing is **Ctrl-Y** *(Win)* **⌘-Y** *(Mac).*

Each time you rerandomize, you get a new set of 15 values from a population with the same mean and standard deviation as the original 15 measurements.

The leftmost histogram is of the original measurements. Next to it are three of the histograms you might get from the simulated heights.



A bit of subjectivity is called for here. Does it appear that the original distribution is very unusual or does it fit in with the simulated distributions?

Testing the Hypothesis

Once we have decided that the assumption of normality is met, we can go on to determine whether the *t*-statistic for Darwin's data is large enough to allow us to reject the null hypothesis.

In a previous section, we typed the summary values into the test as though we didn't have the raw data. But we are in the fortunate position of having the raw data, so we can ask Fathom to figure out all the statistics using that data.

14) Drag the HeightDifferences attribute from the column header of the case table to the top panel of the test where it says Attribute to test is unassigned.

15) If the hypothesis line does not already say is greater than, then select it from the popup menu.

Test of DarwinFertilizationExper Test Mean ▼

Attribute (continuous): HeightDifferences

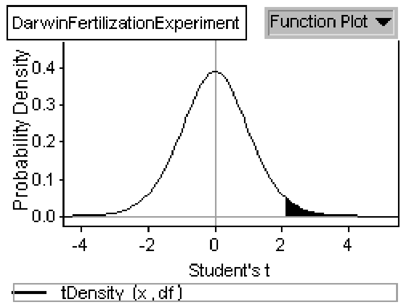
Attribute: **HeightDifferences**
Sample count: **15**
Sample mean: **20.9333**
Standard deviation: **37.7444**
Standard error: **9.74556**
Alternative hypothesis: The population mean of **HeightDifferences** is **greater than 0**

The test statistic, Student's t, is **2.148**. There are **14** degrees of freedom (one less than the sample size).

If it were true that the population mean of **HeightDifferences** were equal to **0** (the null hypothesis), and the sampling process were performed repeatedly, the probability of getting a value for Student's t **this great or greater** would be **0.025**.

The last paragraph of the test describes the results. If the null hypothesis were true and the experiment were performed repeatedly, the probability of getting a value for Student's *t* this great or greater would be 0.025. This is a pretty low *P*-value, so we can safely reject the null hypothesis and, with Darwin, pursue the theory that cross-pollination increases a plant's height compared to self-pollination.

Looking at the *t*-Distribution



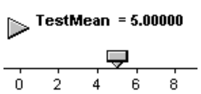
It is helpful to be able to visualize the *P*-value as an area under a distribution.

You may want to iconify the tables to free up some space.

16) With the test selected, choose **Show Test Statistic Distribution** from the **Test** menu.

The curve shows the probability density for the *t*-statistic with 14 degrees of freedom. The shaded area shows the portion of the area under the curve to the right of the test statistic for Darwin's data. We've set this up as a one-tailed test; we're only interested in the mean difference being *greater than* zero. Other times we want to know if a given mean is *different from* zero. This is a two-tailed test. You can change the test to a two-tailed test using the popup menu to change is greater than to is not equal to. If you do this, you will see a shaded region on both the positive and negative ends of the plot. The total area under the curve is one, so the area of the shaded portion corresponds to the *P*-value for a two-tailed test.

Let's investigate how the *P*-value depends on the test mean, currently set to zero.



17) Drag a slider from the shelf into the document.

18) Edit the name of the slider to TestMean.

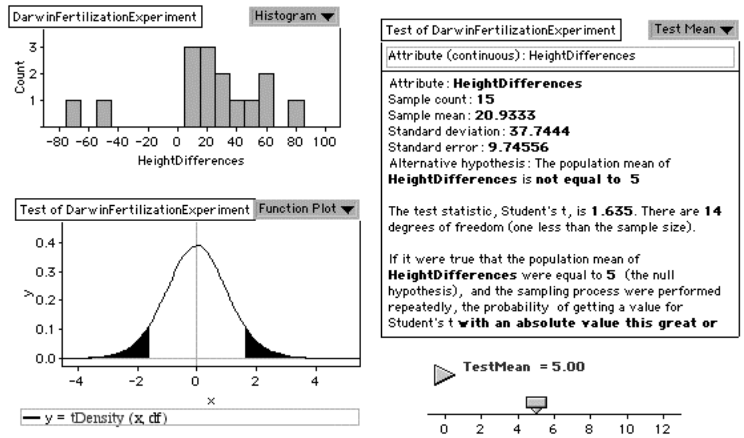
19) Select the histogram of original data and choose **Plot Value** from the **Graph** menu.

20) In the formula editor, type the name of the slider (or choose it from the **Global Values** section of the list to the right of the keypad).

You should see a vertical line in the histogram corresponding to the value of the slider.

21) Click the 0 in the statement of the hypothesis in the test. Choose **Change formula for value** from the popup menu.

22) In the formula editor enter TestMean just as you did for the plotted value.



Now the value of the null hypothesis mean in the test, and the shaded area under the *t*-distribution, change to reflect the new hypothesis.

23) Drag the slider slowly and observe the changes that take place.

For what value of the slider is half the area under the curve shaded? Why does the area under the curve decrease as the value of the slider increases?

Tour 9: Pets and Sports—Testing for Independence

A survey of 325 middle school students[1] from a city school district asks, among other things, for their gender, whether they prefer cats or dogs, and whether they prefer basketball or football. With this data we can investigate whether in this city girls are different in their preference of cats over dogs, whether gender matters in terms of favorite sport, and whether there is a relationship between favorite pet and favorite sport.

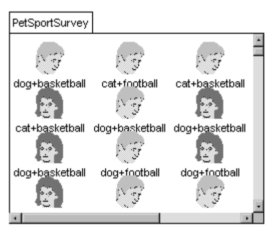
This tour is not for the faint of heart; it's a marathon. We'll be combining classical statistical inference techniques with a computer-intensive method called *scrambling*. We'll assume that you are familiar with the basic techniques of using Fathom.

You Will Learn How to …

- use a ribbon chart to judge differences between groups.
- use a summary table to compute expected values.
- perform a chi-square test for independence.
- use scrambling to simulate the null hypothesis of independence.

A Graphical Analysis Using a Ribbon Chart

1) Open the file **PetSportSurvey** in the **Guided Tour Starters** folder in the **Sample Documents** folder.



PetSportSurvey

	Sex	Pet	Sport
1	boy	dog	basketball
2	boy	cat	football
3	boy	cat	basketball
4	boy	dog	football
5	boy	dog	basketball
6	girl	cat	basketball
7	boy	dog	basketball

Look at the twelve cases that appear in the open collection. You can see that just by staring at a small number of the 325 cases in the sample, it's impossible to make any valid predictions about what trends there might be in the whole population.

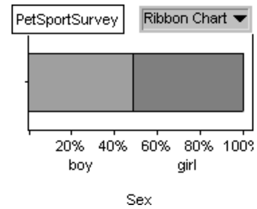
---

[1] This tour uses data collected by the Annenberg Foundation. Unfortunately, as the data were collected on a Web survey, it is far from a random selection. Please bear with the fiction.

We'll concentrate on the question of whether there is any relationship between a student's sex and his or her pet preference.

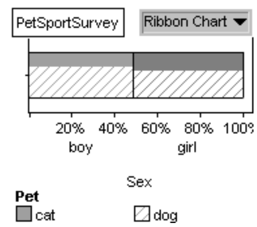
2) Make a graph and drag the Sex attribute to the *x*-axis.

3) Change the resulting bar chart to a ribbon chart by choosing **Ribbon Chart** from the graph's popup menu.



You can see from the ribbon chart that there are approximately even numbers of boys and girls in the sample.

4) Now drag the Pet attribute to the *middle* of the plot.



You should see that each rectangle of the ribbon chart is now divided into two regions, one for each kind of pet. The height of the dog portion is higher for boys than it is for girls. This translates to "A greater proportion of boys than girls prefers dogs over cats."

The Summary Table—Computing Proportions

A ribbon chart does a good job of displaying differences in proportions. But if we want to know the computed values, we need a tabular display—a summary table.

5) Drag a summary table from the tool shelf into the document.

Dropping an attribute on the arrow adds it to that portion of the table. Dropping it on top of an existing attribute replaces what is there.

6) Drag the Sex attribute to the column header of the empty summary table. (You can drop the attribute either on the gray, right-pointing arrow or in the empty column header cell.)

7) Drag the Pet attribute to the row header of the table.

PetSportSurvey | Summary Table

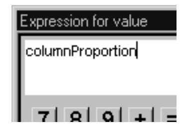
		Sex		
		boy	girl	Row Summary
Pet	cat	51	69	120
	dog	110	95	205
Column Summary		161	164	325

S1 = count ( )

A summary table into which you have dropped nominal attributes shows, by default, the number of cases in each cell defined by the categories of the attributes. We're not interested in the counts as much as we are the proportions.

columnProportion is a special keyword that applies only to formulas for a summary table. If you oriented your table the other direction, you would want to use rowProportion.

8) Double-click the formula underneath the table to bring up the formula editor.



9) Replace count( ) with columnProportion. Click **OK** in the formula editor.

You should see the computed column proportions for each cell. The number 0.3168 in the cell for boys who prefer cats means that about 32% of boys prefer cats to dogs. Similarly about 68% of boys prefer dogs to cats, whereas about 58% of girls prefer dogs to cats.

PetSportSurvey		Summary Table		
		Sex		
		boy	**girl**	Row Summary
Pet	**cat**	0.3168	0.4207	0.3692
	dog	0.6832	0.5793	0.6308
Column Summary		1	1	1

S1 = columnProportion

The Null Hypothesis and Choosing a Test Statistic

Now we're getting to the really interesting subject of statistical inference. Before moving on with Fathom, a bit of discussion about how inference works may be helpful.

The heart of statistical inference is determining whether an observed difference is due to random variation or an actual difference in the population.

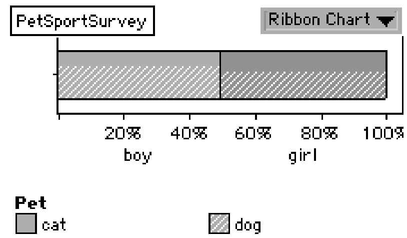
The difference of proportions, 68% – 58% or 10%, is fairly small. Perhaps it is due to chance. Just how likely is it that a random sample would have a difference of proportions this large if *there were actually no difference in pet preferences between boys and girls?*

The assumption that gender is unrelated to pet preference is the *null hypothesis in this situation.* Another way to phrase our question is, If the null hypothesis were true, what is the probability of getting a difference of proportions of 10% or greater. (We have to add the "or greater" because a greater difference is even stronger evidence in favor of there being a relationship between sex and pet preference.)

The difference of proportions is an example of a *test statistic.* Using simulation and techniques similar to those used in Tour 6, we could determine the probability of getting the observed difference of 10%. For the purposes of this tour, we're going to use a closely related test statistic that is commonly used for testing whether a relationship exists between two nominal attributes. It is called the *chi-square statistic.*

Expected Versus Observed

The chi-square statistic is based on the notion that if there were no difference between girls and boys, then the proportion of each that prefers dogs should be the same as the overall proportion of kids that prefer dogs; i.e., about 63%. For this hypothetical sample, we would get a ribbon chart like the one at right.



The fractional boys and girls may seem strange. But it's okay as long as we keep it hypothetical.

Now that we know the expected proportion of boys that prefer dogs, we can compute an expected *number*. It is just the expected proportion times the number of boys. Since there are 161 boys, that's 0.63 X 161 or 101.6 boys. Similarly for the girls, we have 63% of 164 girls or 60.6.

Let's put these numbers into the summary table.

As with rowProportion, columnTotal, rowTotal, *and* grandTotal *are all functions you can use only when writing a formula for a summary table.*

10) Double-click the S1 formula under the table.

11) In the formula editor, type the formula shown at right.

$$\text{columnTotal}\frac{\text{rowTotal}}{\text{grandTotal}}$$

12) Click **OK**.

You should see the expected values computed in each cell. Because this is such a common computation, Fathom has a shortcut for it.

13) Double-click the S1 formula again and substitute the single word expected for the more complicated expression.

You should get the same results when you click **OK**.

The term *expected*, or *expected value*, has a very general meaning in statistics. Here, in Fathom, we're applying it to the very particular situation of a chi-square test where the null hypothesis is that the row and column attributes are independent.

Now we want to compare the expected values with the observed values.

14) With the summary table selected, choose **Add Formula** from the **Summary** menu.

15) Type the formula count( ) into the formula editor and click **OK**.

Your summary table should look similar to the one shown here.

PetSportSurvey — Summary Table

		Sex: boy	Sex: girl	Row Summary
Pet	cat	59.45 51	60.55 69	120 120
	dog	101.6 110	103.4 95	205 205
Column Summary		161 161	164 164	325 325

S1 = expected
S2 = count ( )

Computing the Chi-Square Statistic

There are many different statistics we could invent using the observed and expected values. We're going to use one for which statisticians have figured out how to compute a distribution without having to resort to simulation—the chi-square statistic. It's based, as you might imagine, on the difference between the observed and the expected values.

16) Add the formula shown at right to the summary table.

$$\frac{(\text{count}(\ ) - \text{expected})^2}{\text{expected}}$$

The chi-square statistic is simply the sum of these numbers you just computed. (They correspond to the formula labeled S3 in the summary table.) In order to calculate the sum of these values in Fathom, we have to collect them.

17) With the summary table selected, choose **Collect Measures** from the **Analyze** menu.

You should see a new collection labeled Measures from PetSportSurvey Table.

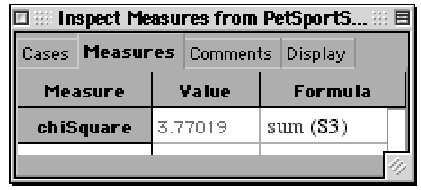
Measures from PetSportSurvey Table

	Pet	Sex	S1	S2	S3
1	cat	boy	59.4462	51	1.20004
2	cat	girl	60.5538	69	1.17808
3	dog	boy	101.554	110	0.70246
4	dog	girl	103.446	95	0.68961

18) Make a case table for the new collection.

Each case in this new collection corresponds to a single cell in the summary table—four cells and four cases. Each of the three formulas in the summary table corresponds to one attribute in the measures collection. We're interested in the sum of the values for S3.

19) Bring up an inspector for the new collection and define a measure for it as shown at right. This number is the chi-square statistic.



We now have our statistic, but we do not yet know how likely it would be to get a chi-square value this big or bigger by chance alone. But the advantage of using a well-studied statistic is that Fathom can easily compute this probability for us.

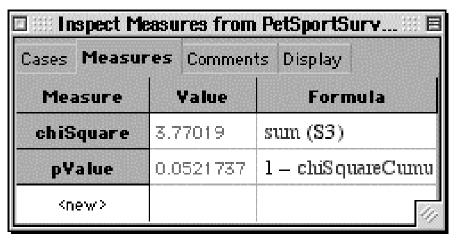
20) Define a second measure, pValue, for this collection. Make its formula as shown at right.

1 − chiSquareCumulative (chiSquare, 1)

If you know all the row totals and column totals, degrees of freedom is the number of cell counts you could fill in before all the rest were determined for you.

ChiSquareCumulative is a function built in to Fathom. It takes two arguments: the first is the value of chi-square that you have computed (the first measure you made); the second is the number of *degrees of freedom* available, in this case, one. The function computes the probability of getting that value of chi-square *or less*. Since we're interested in *or greater*, we subtract the function's value from one.

Your inspector should look similar to the one shown here. The probability of getting a chi-square statistic greater than or equal to 3.77, under the assumption of the null hypothesis, is 0.052.

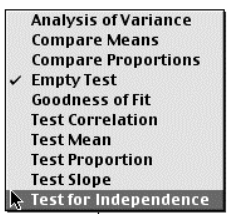


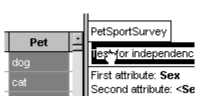
What are we to make of this result? We can say that if there were no difference between boys' and girls' pet preferences, and if we repeated the random sampling many times, we would get a result this extreme or more extreme about 1 time in 20. For many situations, especially in the social sciences, this level of probability is persuasive enough to say that we have probably found something. But in other situations, especially in medical research, we would not be able to say we had found something because the consequences of being wrong would be too great.

Test for Independence—The Simple Way

If you've followed along this far, you may be thinking that that was an awful lot of work to accomplish what is a fairly routine calculation. You're right; and Fathom has the ability to do this computation quickly and simply. Here's how.

You may want to hide or delete objects to free up some space. Keep the source collection and its case table.



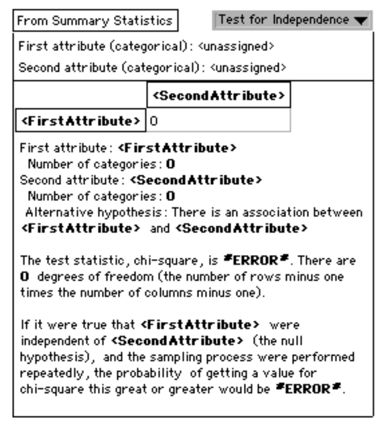


21) Choose **Test Hypothesis** from the **Analyze** menu.

You get an empty test. You choose what kind of test you want to do from the popup menu of the analysis.

22) Choose **Test for Independence** from the popup menu.

The analysis is waiting for you to assign two nominal attributes to it.



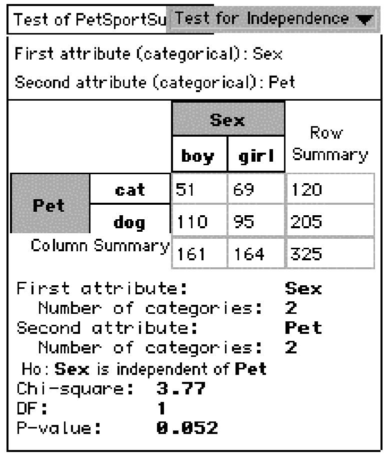
The analysis shows #ERROR# for the chi-square statistics and *P*-value because Fathom doesn't yet have information to compute them.

23) From the PetSportSurvey collection, drag Sex and then Pet to the top portion of the analysis.

Sometimes you don't need to have a full explanation of how the inference works.

24) Choose **Verbose** from the **Test** menu to turn off verbose mode and see a more compact version of the test.

You should see, as shown below, the results of a chi-square test. The value of chi-square and its *P*-value are both given. They should be the same as the values you computed in the previous section.


Test of PetSportSu Test for Independence

First attribute (categorical): Sex

Second attribute (categorical): Pet

		Sex: boy	Sex: girl	Row Summary
Pet	cat	51	69	120
	dog	110	95	205
Column Summary		161	164	325

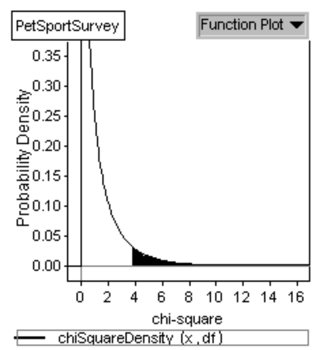
First attribute: Sex
Number of categories: 2
Second attribute: Pet
Number of categories: 2
Ho: **Sex** is independent of **Pet**
Chi-square: **3.77**
DF: **1**
P-value: **0.052**

A Graph of the Chi-Square Distribution

It's helpful to see where the computed chi-square statistic for this sample lies in the distribution of chi-square values that would result when the null hypothesis is satisfied.

25) With the test object selected, choose **Show Test Statistic Distribution** from the **Test** menu.

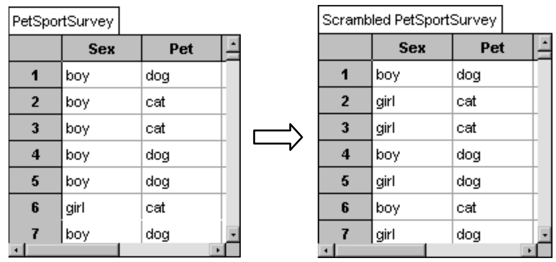
The graph you get should be similar to the one shown at right. The shaded area under the right portion of the curve corresponds to the *P*-value for the observed chi-square statistic.



Simulating the Null Hypothesis

With Fathom we can simulate conditions under which the null hypothesis *is* true and repeatedly perform the sampling and computation of a chi-square statistic. While this does not tell us anything more about the particular experiment, it does shed light on the process of statistical inference.

The null hypothesis states that there is *no* relationship between the two attributes Sex and Pet. What if we were to take all the values for the attribute Sex and scramble them so that boy and girl got reassigned randomly to each case? Any relationship that might exist between the two attributes would be wiped out by the scrambling. Any remaining relationship would have to be due to chance alone.



PetSportSurvey

	Sex	Pet
1	boy	dog
2	boy	cat
3	boy	cat
4	boy	dog
5	boy	dog
6	girl	cat
7	boy	dog

Scrambled PetSportSurvey

	Sex	Pet
1	boy	dog
2	girl	cat
3	girl	cat
4	boy	dog
5	girl	dog
6	boy	cat
7	girl	dog

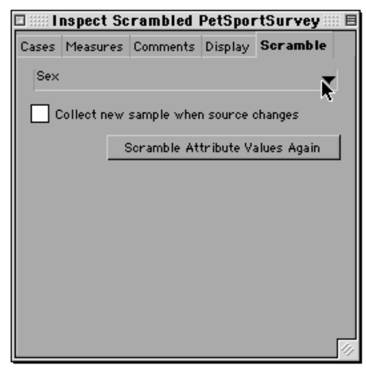
26) Select the PetSportSurvey collection.

27) Choose **Scramble Attribute Values** from the **Analyze** menu.

A new collection should appear labeled Scrambled PetSportSurvey.

28) Make a case table for the new collection.

By default, Fathom scrambles the first attribute in the collection. That works well here, but in other situations you may need to scramble a different attribute. To do so, bring up the scrambled collection's inspector and look in the **Scramble** pane for a popup menu that gives you a choice of attributes to scramble.



The keyboard shortcut for this is **Ctrl-Y** *(Win)* **⌘-Y** *(Mac).*

29) With the scrambled collection selected, choose **Scramble Attribute Values Again** from the **Analyze** menu.

You should see the values in the Sex column of the case table change each time you scramble again.

30) Make a ribbon chart, just as you did on page 76.

As you scramble, you can see the variation in the relative proportions. This variation is due solely to chance.

31) Make a new analysis for doing a test for independence. This time, drop attributes from the scrambled collection into it.

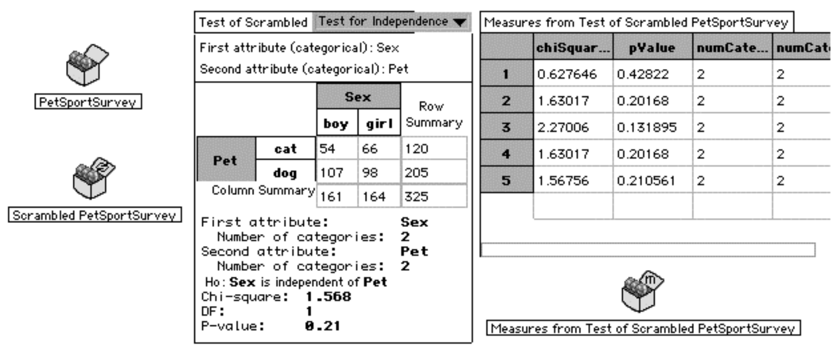
Each time you scramble again, the chi-square statistic and the *P*-value are recomputed. Because we're simulating the conditions of the null hypothesis, the chi-square values will not be very large and the *P*-values will not be very small.

Now we want to collect many chi-square values from the scrambled collection. We will build up a distribution of these values and see what shape it has.

32) Select the test object and choose **Collect Measures** from the **Analyze** menu.

Fathom will scramble the scrambled collection five times, each time collecting values computed by the test for independence and putting them in a new collection, labeled Measures from Scrambled PetSportSurvey.

33) Make a case table for the measures collection. Your screen should look similar to that shown on the next page. (We've chosen the **Verbose** command in the **Test** menu to make the output of the test shorter than it is by default.)

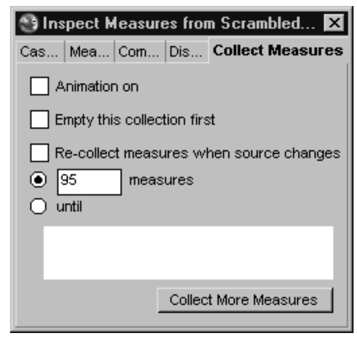


The two important columns in the table to look at are pValue and chiSquareValue.

34) Make a histogram of each of the attributes, pValue and chiSquareValue.

Your histogram won't look like much yet because you have only collected the results of five scrambles. You need some more.

35) Double-click the measures collection and bring its **Collect Measures** pane to the front as shown below.



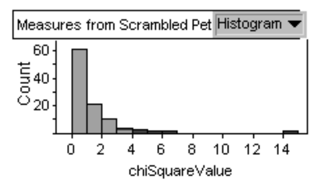
36) Uncheck both the **Animation On** option and the **Empty this collection first** option.

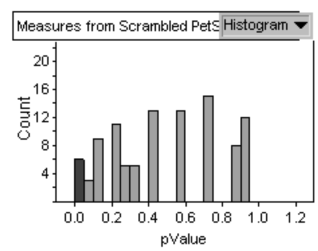
37) Specify that you want 95 measures instead of 5.

Collecting 95 measures may take a while. A progress bar will give you an idea of how long you will need to wait.

38) Click the **Collect More Measures** button.

When the collection process completes, you should have 100 *P*-values and 100 chi-square statistics.





What do these histograms tell us? We see that chi-square statistics as high as the one we got for the original sample, 3.77, don't occur very often. But they do occur; in fact, they occur about 5% of the time, corresponding to the *P*-value we computed for the original sample.

Secondly, the shape of the chi-square histogram markedly resembles the plotted chi-square distribution you made in Step 25 on page 82. That makes sense—one is from theory and the other is from simulation, but they should show the same thing.

Thirdly, the distribution of *P*-values is spread over the interval from zero to one. Select the lowest bar in the pValue histogram and notice that the highest bars in the chi-square histogram are selected. By selecting only those *P*-values less than or equal to 0.05, you can read off an approximation for the so-called *critical value* for chi-square in the chi-square plot.

Ideas for Further Exploration

- Consider the other two pairs of attributes possible: Sex versus Sport and Pet versus Sport. Would you expect them to show more or less independence than Sex versus Pet? Look at the corresponding ribbon charts. Do the observed differences in proportions look significant? Perform a chi-square test on each of these pairs. Explain why one result is so much more significant than the other.

- Invent your own test statistic that is a measure of independence between two nominal attributes. Use the scrambling technique to find a distribution of that measure under the conditions of the null hypothesis.
- Investigate the distribution of chi-square using a simulation in which the null hypothesis is *not* true. You can do this by creating a collection with randomly generated values for two attributes into which you build a dependency of one attribute's values on the other.

Index

Q

R

S