

WEEK 1: INTRODUCTION

TABLE OF CONTENTS

	Page:
List of Figures.	1
List of Tables.	2
List of m.files.	2
Assignment.	2
Required reading.	2
Understanding by Design Templates.	3
Understanding By Design Stage 1 — Desired Results Week 1.	3
Understanding by Design Stage 2 — Assessment Evidence Week 1.	3
Introduction.	5
Examples and Case Studies.	5
Earthquakes & Matlab.	5
A note on figures.	7
The Stock Market, Runs & Matlab.	8
Case Study 1.2.2 Golden rectangles.	11
Case Study 1.2.3 Lunacy.	11
Case Study 1.2.4 Cancer Deaths in Oregon.	12
Annotated outline (with Matlab scripts) for Larsen & Marx Chapter 1.	13
References.	26
Index.	26

List of Figures

- Figure 1.** A reanalysis of Southern California earthquake data with analysis from earthquakes_LA.m. 6
- Figure 2.** A reanalysis of Southern California earthquake data with analysis from earthquakes.m Only earthquakes of magnitude 4 or greater are plotted, producing a steeper slope than the previous figure and one comparable to the text. 6
- Figure 3.** A reanalysis of Southern California earthquake data with analysis from earthquakes_LA.m Only earthquakes of magnitude 4 or greater are plotted. When plotted on a semilog plot with base 10 logs, the slope is usually about -1, but in this case the slope is about -0.79 ± 0.15 7
- Figure 4.** The Gutenberg-Richter relation for a 5 degree latituden and longude box centered on the March 11 2011 Magnitude 9 quake in Japan. I fit the line with an without that datum. The probability of a magnitude 9 quake

went from once per 133 years to about once per century. Of course in this 38-year database, the actual datum of once per 38 years is plotted for this earthquake, which is one of the largest in recorded history..... 7

Figure 5. Closing prices for the S & P 500 from February 17, 1994 to February 9, 1996 from Yahoo.. 8

Figure 6. Closing prices for the S & P 500 from February 17, 1994 to February 9, 1996 from Yahoo.. 8

Figure 7. The observed length of runs from the S & P 500 from February 17, 1994 to February 9, 1996 (data from Yahoo) and the expected pattern of runs if the probability of an upturn is equal to the probability of a downturn on each day.. 9

Figure 8. The observed length of runs from the S & P 500 from February 17, 1994 to February 9, 1996 (data from Yahoo) and the expected pattern of runs if the probability of an upturn is equal to the probability of a downturn on each day. Also shown is the expected pattern of runs if the probability of a run persisting is $p = 0.5556$, which provides a better fit to the data.. 10

Figure 9. The observed length of runs from the S & P 500 from February 17, 1994 to February 9, 1996 (data from Yahoo) and the expected pattern of runs if the probability of an upturn is equal to the probability of a downturn on each day. The categories have been pooled so that the model expected values are equal to or greater than 5. Also shown is the expected pattern of runs if the probability of a run persisting is $p = 0.52266$, which produces the minimum chi-square statistic indicating the closest fit between model predictions and observations.. 10

Figure 10. The observed width-to-length ratios of rectangles in Shoshoni beads plotted with the expected ratio from the Golden rectangle. Also plotted is the mean width to length ratio for the Shoshone beads..... 11

Figure 11. Admission rates to the emergency room of a Virginia mental health clinic by month and phase of moon.. 11

Figure 12. Cancer deaths per 100,000 versus an Index of Exposure to the Hanford nuclear reservation.. 12

List of Tables

List of m.files

earthquakes_LA..... 13

earthquakes_Japan..... 15

function LMcs010201_4th..... 18

LMcs010202_4th..... 23

LMcs010204_4th..... 25

Assignment

Required reading

- ! Larsen, R. J. and M. L. Marx. 2006. An introduction to mathematical statistics and its applications, 4th edition. Prentice Hall, Upper Saddle River, NJ. 920 pp.
- " Read Sections 2.1-2.8

Understanding by Design Templates

Understanding By Design Stage 1 — Desired Results Week 1

LM Chapter 1 Introduction

G Established Goals

- Using Matlab for scientific computing & graphics
- Graph data using bar charts and scatter plots with regression lines (will be covered in depth in later weeks)

U Understand

- Statistics, among other things, is applied probability and a method for separating chance variability from real patterns in nature

Q Essential Questions

- What is an algorithm?

K *Students will know how to define (in words or equations)*

- algorithm

S *Students will be able to*

- Get Matlab running for the 1st time.
- Write their first Matlab programs
- Plot their first graph in Matlab and create a Word, rtf or Word processing document incorporating that figure.

Understanding by Design Stage 2 — Assessment Evidence Week 1

Read LM Chapter 1 Introduction, start on chapter 2 & Handout 5

- **Post in the discussion section by 6/9/11 W**
 - Watch the 1-h video the joy of statistics & post a comment on the online discussion <http://www.gapminder.org/videos/the-joy-of-stats/>
- **Problems due Wednesday 6/9/11 W 10 PM**
 - **Each problem must be solved using Matlab**
 - **Basic problems (4 problems 10 points)**
 - **Problem 1. S&P 500 stock prices** Go the web and download 2 years months of S&P 500 closing prices and graph them
 - <http://finance.yahoo.com/q/hp?s=%5EGSPC+Historical+Prices>
 - You may have to use a spreadsheet program to get dates on the plot, but dates aren't required. You can add dates to fulfill Advanced Problem 1 below
 - **Problem 2.** Analyze the S & P 500 closing prices with runs2.m and compare with the pattern in Table 1.2.2 in the text. Report on the results. Use LMcs010201_4th.m as a model.
 - **Problem 3.** Case Study 1.2.2 a) Calculate the sample median for the data presented in Table 1.2.3 Use my case study 1.2.2 as a guide b) plot a histogram with the sample median and expected values as vertical lines, c) transfer the histogram to a Word processor document
 - **Problem 4.** Find a particularly good graphic display of quantitative information from the popular press or the scientific literature (do a google scholar search under an area of research that interests you). Cut & paste it into your homework submission. Describe why the graphic is good
 - **Advanced problems (2.5 points each)**
 - **Problem 1.** In graph 1, replace the numbers on the x axis with actual dates (use the Matlab help file for datetick)
 - **Problem 2.** Plot the data in Table 1.2.4 as histograms after converting the data to percentages so that each month's admissions sum to 100%
 - **Master problems (1 only, 5 points)**
 - Go to the USGS website (see help documentior for earthquakes_LA.m or earthquakes_Japan.m and download the earthquake data for a region of interest to you (Boston, Haiti, London, Helena Montana). Does the Gutenberg-Richter relationship work? Just plot the data. No need for a formal test of significance. Note that there is only data compiled since 1973 (38+ years). I created a 2 degree or 5 degree latitude and longitude box for my data requests depending on the locality to ensure getting sufficient numbers of earthquakes.

Introduction

Welcome to the class. During the first week of class, the goals are rather modest. Begin reading the text (Read Chapter 1 and start on the very long chapter 2) and begin familiarizing yourself with Matlab. The most important task is to run the Matlab tutorial and play with the program to start getting familiar with the program. I do mean play. Try out different commands. Use the help file to find examples of Matlab code and run those. Run examples of my program. Make some pretty graphs.

Handout 2 and 3 are reference handouts. Handout 2 includes definitions of statistical terms and concepts. Maybe half of the terms will be covered in this course, but you'll come across others in your future work. I've provided Handout 3 which lists most of the Matlab functions that will be used in the hundreds of Matlab m.files that we'll be using this semester.

Start Reading Chapter 2. Chapters 2 and 3 are the longest and most difficult in the book. Don't get discouraged with these chapters. As you read the chapters, follow along my annotated outline included in each handout. As you read the handouts, you'll come across highlighted words, when I wrote and edited these handouts, I created hypertext links for all of the references and many of the terms. Click on the term to find the definition. To return to your original position in the pdf document from within a pdf reader, use the **alt-left arrow** combination of keystrokes, that is the pdf reader equivalent of the back arrow on a web browser. **Alt-right arrow** is the equivalent of a web browser's forward arrow. I've inserted the Matlab m.file code for most of the examples and most of the case studies. You'll never have to transcribe m.files. I distribute two zipped files for the course: EEOS601 and LMfiles. The first includes over one hundred utility m.files and m.files not specific to the text, and LMfiles includes script m.files for most of the case studies, examples and even figures in the text, labeled as Lmex*.*, Lmcs*.* and LMFig*.*

I have handouts prepared for each chapter in the book, which roughly corresponds to each week of the 10-week class. At the start of each of these handouts, I provide the Understanding by Design templates for the week. Then, I discuss some of the more important concepts, case studies and examples from the text. I'll often go beyond the text and discuss the application of these concepts for my own work in the analysis of marine benthic communities.

Examples and Case Studies

Earthquakes & Matlab

Figure 1.1.1 in the text covers a very interesting example of Matlab, plotting Southern California earthquakes on the Richter scale. If the average number of earthquakes for a year are plotted vs. the Magnitude of the quake, the points fall along a straight line. In my program **earthquakes_LA.m**, I provide an updated analysis of Southern California earthquakes. I searched online and found several databases for earthquakes, with this being the best for Southern California: http://earthquake.usgs.gov/earthquakes/eqarchives/epic/epic_rect.php. I

downloaded all of the data from 1973 through 2010 and produced Figure 1. The major difference between this figure and the text's is that I've plotted all of the earthquakes of magnitude 2 or above within 1 degree latitude and longitude of LA (LA is -118.25, 34.05), producing a slope considerably more gradual than that in the text. Figure 2 shows the fit of the Gutenberg-Richter relation for earthquakes \geq Richter scale 4.

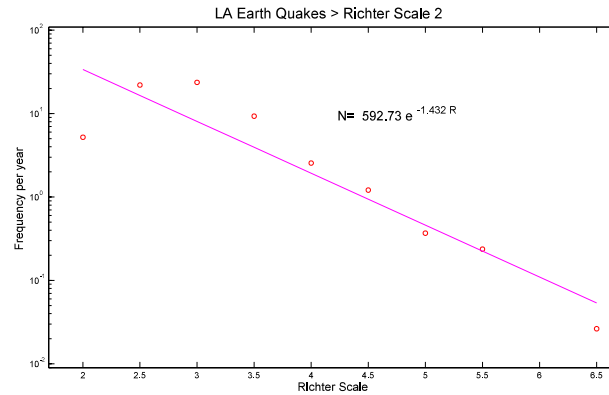


Figure 1. A reanalysis of Southern California earthquake data with analysis from earthquakes_LA.m

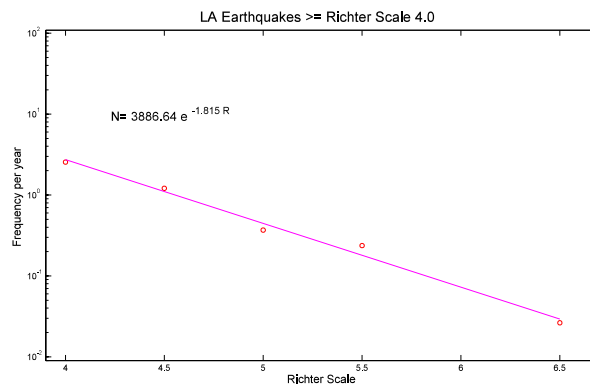


Figure 2. A reanalysis of Southern California earthquake data with analysis from earthquakes.m Only earthquakes of magnitude 4 or greater are plotted, producing a steeper slope than the previous figure and one comparable to the text.

Larsen & Marx fit the data using natural logarithms. A more interesting fit is obtained if the data are analyzed on the \log_{10} scale as shown in Figure 3. The general pattern is that no matter where in the world earthquake data are obtained, the slope of this relation is roughly -1. In the case of LA earthquake data, the slope is the slope is -0.79 ± 0.15 .

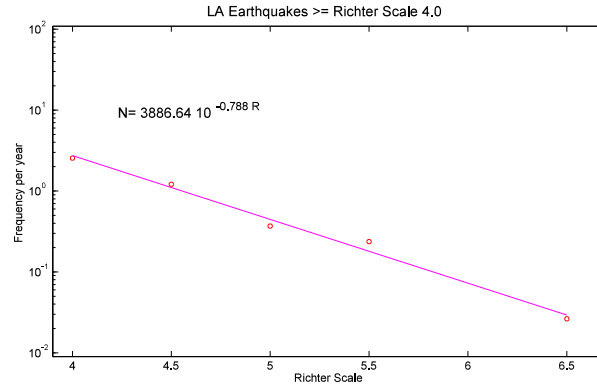


Figure 3. A reanalysis of Southern California earthquake data with analysis from earthquakes_LA.m Only earthquakes of magnitude 4 or greater are plotted. When plotted on a semilog plot with base 10 logs, the slope is usually about -1, but in this case the slope is about -0.79 ± 0.15 .

With the USGS database, I can download data from any latitude and longitude around the world. Figure 4 shows the Gutenberg-Richter relation for Japan, with and without the magnitude 9 earthquake of 11 March 2011. Without that Richter Magnitude quake, the regression model predicts a magnitude 9 quake about once every 130 years.

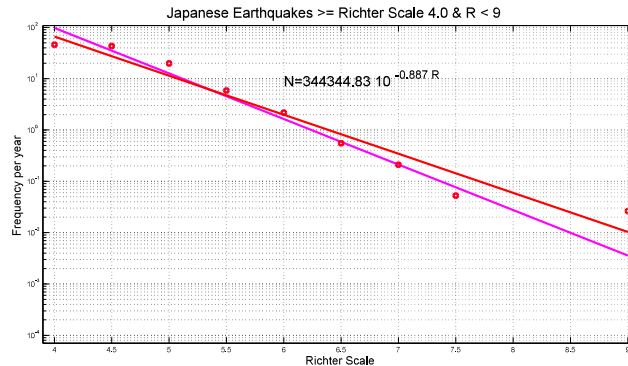


Figure 4. The Gutenberg-Richter relation for a 5 degree latitude and longitude box centered on the March 11 2011 Magnitude 9 quake in Japan. I fit the line with and without that datum. The probability of a magnitude 9 quake went from once per 133 years to about once per century. Of course in this 38-year database, the actual datum of once per 38 years is plotted for this earthquake, which is one of the largest in recorded history.

We'll be covering regression in Chapter 11 and Week 11 in the course.

A note on figures

Matlab graphics have always been optimized for display as encapsulated postscript files. When I generate a figure in Matlab, I usually export the figure as an eps file, short for encapsulated postscript. These eps files provide the commands for Adobe software to redraw the graphs using lines and fonts. Most Windows programs expect bitmapped graphics, where a picture is submitted to the program. When eps files are expanded, the fonts and lines remain jagged, but bitmapped graphics become ragged. When Matlab exports eps files it includes a bitmapped graphic version of the figure. So, when I see the figure in my word processor (WordPerfect), I see the bitmapped graphic. When I make a pdf of the file, the graphic is converted by my Adobe Acrobat Pro to a high resolution vector graphic.

The Stock Market, Runs & Matlab

Case Study 1.2.1 provides an interesting statistical analysis. In the problem, the closing price from February 17, 1994 to February 9, 1996 are analyzed for runs. I've programmed these data in **LMcs010201_4th** below. These data and most other stock price data can be downloaded from <http://finance.yahoo.com/q/hp?s=GSPC+Historical+Prices>. The key questions to be asked of these data are, "If the stock market goes up today, what is the probability that it will go up tomorrow?" and "Is the pattern of upturns and downturns consistent with a random, memory-less model?"

First, let's take a look at the data, shown in Figure 5. There is a general positive trend in the data leading one to suspect that in analyzing the day-to-day variation that there should be far more upturns than downturns. Or, so one might expect.

I kept the previous graph close to the Matlab defaults, but the default settings in Matlab often leave the axis legends too small to read. Matlab uses a programming convention common to C++ and other structured programming languages in that the graph is created by the program and then the children of the graph — the axes, legends, points, type of data plotting symbol, and line widths — are adjusted in a set of statements. In

reading my programs, bear that in mind. One of the nice features about Matlab, once you get used to it, is that virtually every component of the graph can be programmed. This can produce long programs but very nice graphs. One nice improvement in the previous graph would be to have the calendar dates on the graph. I've done that in Figure 6.

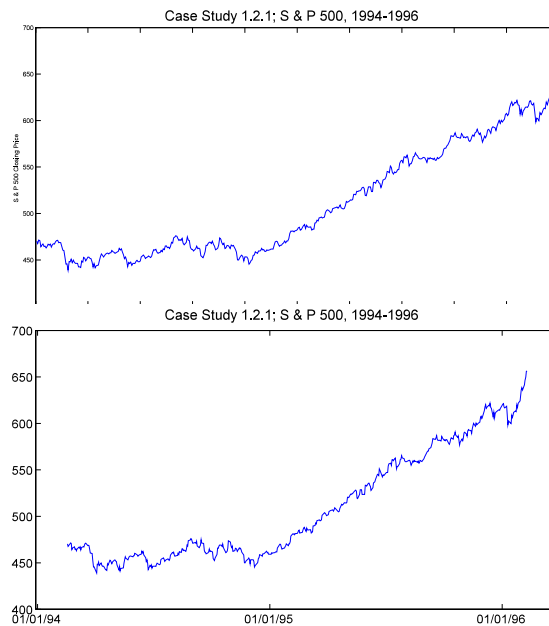


Figure 6. Closing prices for the S & P 500 from February 17, 1994 to February 9, 1996 from Yahoo.

In the text, the authors calculate just the downturns in the stock market. In my program, **runs2.m**, which calls Daniel Kaplan's **runs.m** from the resampling toolbox for Matlab (a very nice set of programs), the total number of runs, including the number of positive and negative runs (and ties) is calculated. The program **ties2.m** returns the information that there were 242 runs in the data set, including 121 positive runs and 121 negative runs. This analysis has already answered one of our questions, the daily upturn and downturn in the market for this period was as random as the tossing of a fair coin despite the nearly 200 point rise in the S & P 500 during this period. The second question deals with the run length. Are the lengths of runs consistent with a memory-less process, where the probability of an upturn or downturn is independent of the previous history of the stock price? Larsen & Marx provide the following equation to predict

the expected probabilities of runs of different lengths dependent upon the probability of an upturn or downturn on each day: $p(k) = p^{k-1} (1-p)^k$, where k is the length of run. We can apply that equation with $p=0.5$ to see how closely the observed run lengths (both up and down) compare to the expected run lengths:

No. of Runs	Observed	Expected
1	121	121
2	48	60.5000
3	43	30.2500
4	12	15.1250
5	9	7.5625
6	8	3.7813
7	0	1.8906
8	0	0.9453
9	0	0.4727
10	0	0.2363
11	0	0.1182
12	1	0.0591

As shown in Figure 7 and the above table, the fit seems quite good, matching the number of runs of length 1 exactly. There was one remarkable run of 12 straight days of increase which was not fit well by the model. Note that Larsen & Marx's Table 1.2.1 was modeling just the downturns in the stock market.

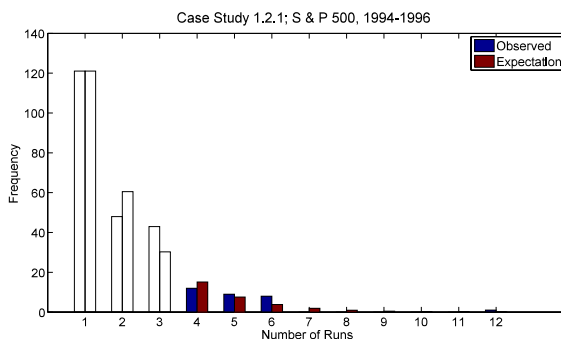


Figure 7. The observed length of runs from the S & P 500 from February 17, 1994 to February 9, 1996 (data from Yahoo) and the expected pattern of runs if the probability of an upturn is equal to the probability of a downturn on each day.

Now, as noted by Larsen and Marx, there may be a better fit to the number of runs if we $p(k) = p^{k-1} (1-p)^k$. There are a large number of possible ways to find a value of p which will provide a better fit to the data. One way is to just enter different values for p and do a 'chi-by-eye' fit to see which p provides a set of histograms closer to the observed data. The chi in that expression (from Press et al.'s Numerical Recipes) is for the chi square test, which we'll cover in chapter 10. The chi-square statistic is one of the standard ways for fitting models to data. The Chi-square statistic is the sum of the $(\text{observed}-\text{expected})^2/\text{expected}$ for every observed datum. So, we can find the value of p that provides the closest fit to the data by minimizing chi square. We could try a brute-force procedure by plugging different values of p into the equation and solving for chi square, or we could use a graphical procedure to find the p value that finds the minimum chi square on a graph, we could use a Matlab minimization routine. I've worked with all 3 approaches, and the final version of the program calls Matlab's [fminbnd.m](#), which finds the value of p that minimizes the chi-square statistic. This routine is only used a few more times in the course

(listed on Handout 3) and is not one of the methods described in the text. For standard probability models, there are more conventional way of finding estimates of model parameters. With this minimization routine, it turns out that a $p = 0.5556$ will produce an expected pattern of runs that is a closer fit to the data.

This pattern is shown in Figure 8 and the observed and expected number of runs is shown below:

Observed	Expected ($p=0.5556$)
121	115.5389
48	60.3768
43	31.5509
12	16.4874
9	8.6158
8	4.5023
0	2.3528
0	1.2295
0	0.6425
0	0.3357
0	0.1754
1	0.0917

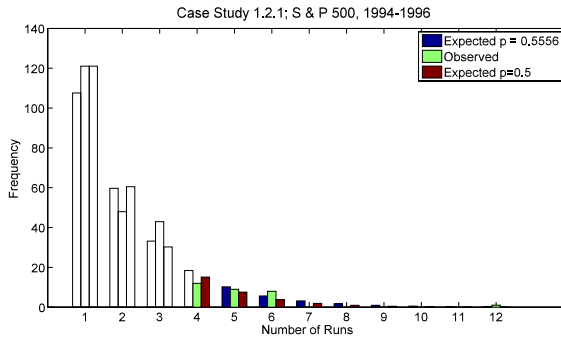


Figure 8. The observed length of runs from the S & P 500 from February 17, 1994 to February 9, 1996 (data from Yahoo) and the expected pattern of runs if the probability of an upturn is equal to the probability of a downturn on each day. Also shown is the expected pattern of runs if the probability of a run persisting is $p = 0.5556$, which provides a better fit to the data.

As we'll cover in Chapter 10 (Week 10 of the course), the chi-square test is usually modified so that the expected value in each category is greater than 5. This is done by combining adjacent categories. In the S & P 500 data, there are a number of categories with expected values less than 5. I modified the program so that it would fit the p value with this constraint. This was a little tricky, since the number of categories is a function of the p value. But, the result was that with a p value of 0.5226, the chi-square statistic was minimized. The results are shown in Figure 9 and the following table:

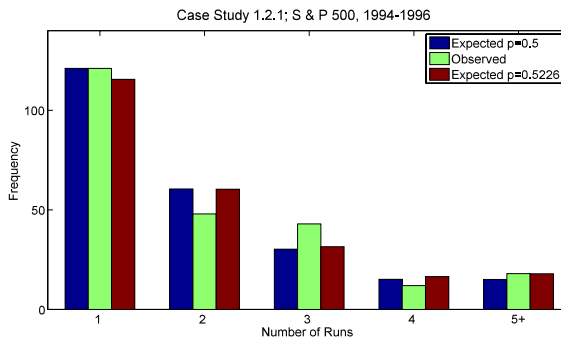


Figure 9. The observed length of runs from the S & P 500 from February 17, 1994 to February 9, 1996 (data from Yahoo) and the expected pattern of runs if the probability of an upturn is equal to the probability of a downturn on each day. The categories have been pooled so that the model expected values are equal to or greater than 5. Also shown is the expected pattern of runs if the probability of a run persisting is $p = 0.52266$, which produces the minimum chi-square statistic indicating the closest fit between model predictions and observations.

Observed	Expected
121	115.5389
48	60.3768
43	31.5509
12	16.4874
18	17.9457

Using techniques described in Chapter 10 (Week 10), the probability of observing results as different from the predictions of the null hypothesis or more extreme in their differences from the null hypothesis can be determined by examining the value of the chi-square statistic relative to cumulative Chi-square probability distribution with 3 degrees of freedom. This probability, called the p-value, is 0.0426, which indicates that the memoryless property of our model is not an adequate fit for these data. All other values of p , the probability of a run continuing for a 2nd day produce lower p values and are an even less likely choice for model parameter if the goal is to minimize the difference between model predictions and observations.

Case Study 1.2.2 Golden rectangles

Golden rectangles have the property that the width to length ratios equal the length divided by the width plus length. Is the ratio of Shoshone Indian beaded rectangles equal to 0.618, the golden ratio? Figure 10 shows the ratio and the Golden ratio expectation. One of the questions we'll be addressing at several points in the course is whether departures from expectation as severe as those shown in the figure could have been due to chance.

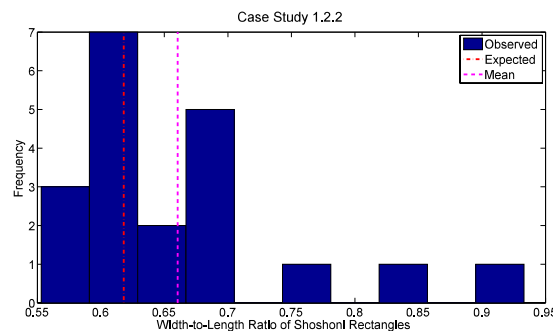


Figure 10. The observed width-to-length ratios of rectangles in Shoshoni beads plotted with the expected ratio from the Golden rectangle. Also plotted is the mean width to length ratio for the Shoshone beads.

Case Study 1.2.3 Lunacy

Do the admissions to the emergency rooms of a mental health clinic vary as a function of the phases of the moon? Figure 11 shows the admissions rate to a Virginia mental health clinic before, during, and after the twelve full moons from August 1971 to July 1972.

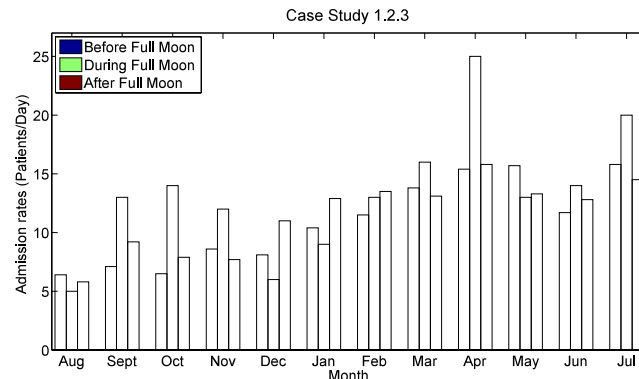


Figure 11. Admission rates to the emergency room of a Virginia mental health clinic by month and phase of moon.

We will cover methods in Week 12 of the course that will allow us to

assess whether the differences in admission rates are likely due to chance after accounting for the month-to-month differences.

Case Study 1.2.4 Cancer Deaths in Oregon

The Hanford nuclear reservation abuts the Columbia River which flows through Washington and Oregon. Over the years, it leaked radioactive material into the river. Table 1 shows the exposure index for the 9 counties of Oregon that are adjacent to the Columbia River or Pacific Ocean. The index is based partially on the distance from the Hanford nuclear reservation.

Radioactive Contamination and Cancer Mortality in Oregon		
County	Index of Exposure	Cancer Mortality per 100,000
Umatilla	2.49	147.1
Morrow	2.57	130.1
Gilliam	3.41	129.9
Sherman	1.25	113.5
Wasco	1.62	137.5
Hood River	3.83	162.3
Portland	11.64	207.5
Columbia	6.41	177.9
Clatsop	8.34	210.3

Figure 12 shows a graph of the cancer death rate versus index of exposure. In week 11 of the course, when we cover regression models, we'll learn the techniques for fitting regression lines and judging their adequacy.

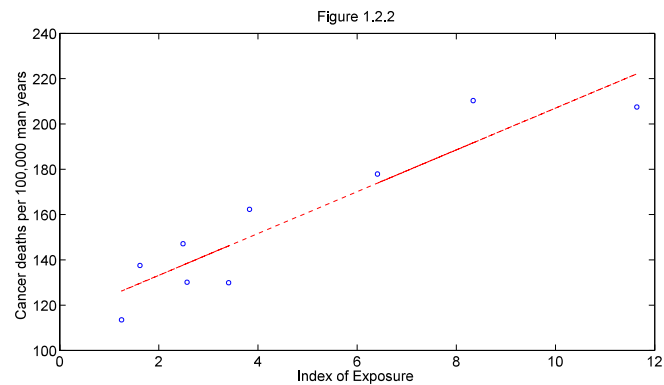


Figure 12. Cancer deaths per 100,000 versus an Index of Exposure to the Hanford nuclear reservation.

Annotated outline (with Matlab scripts) for Larsen & Marx Chapter 1

2 Chapter 1: Introduction

Francis Galton photo

2.1 A brief history

Figure 1.1.1 Average earthquake number vs. Magnitude on Richter scale

Larsen & Marx don't provide the data for their problems. This problem involved plotting the frequency of earthquakes, using techniques that we will cover near the end of the course (on regression). I downloaded all of the earthquake data (Richter magnitude ≥ 2) from California from 1973 to 2010 and fit the data in earthquakes_LA.m. This relationship is called the Gutenberg-Richter relationship. When frequency of quakes is plotted using \log_{10} , the slope of the relation should be about -1. As described in the introduction, the slope of the LA earthquake data is -0.79 ± 0.15 .

```
% earthquakes_LA.m, modeled on Bostonquakes.m
% See Also earthquakes_Japan.m
% Modeled after LMFig010101poissr.m & earthquakes_Haiti.m
% Data from USGS/NEIC (PDE) 1973 - 2010 12 31
% http://earthquake.usgs.gov/earthquakes/eqarchives/epic/epic_rect.php
% Search 1/1/73 to 12/31/10 38 years
% LA is -118.25, 34.05
% LA 33.05 to 35.05 lat Long -119.25 to -117.25
% All quakes greater than 2.0
% a) copy data from screen to Matlab window, b) save as comma delimited
% text file, c) open file in spreadsheet (excel or quatro pro) d)
% delete unnecessary columns, especially last column of characters that
% fouls up a Matlab load. e) paste Richter magnitude data into Matlab
% editor and save.
% Using example 4.3.5 as a model
% Example 4.3.5 Memphis earthquakes and the Poisson distribution
% Pp 305-306 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written Fall 2010 by Eugene Gallagher for EEOS601 Revised: 3/18/11
% Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
earthquakes_LA_38y; % calls the data file.
Richter=4; % Could be set to any value greater than 2 for LA quakes
year=38;
k=find(DATA(:,7)>=4);
edges=Richter:.5:ceil(max(DATA(k,7)));
N=histc(DATA(k,7),edges);bar(edges,N/year,'histc');
s=sprintf('LA Earthquakes >= Richter Scale %3.1f',Richter);
title(s,'FontSize',20);
```

```

set(get(gca,'Children'),'FaceColor',[.8 .8 1]);
xlabel('Richter Scale','FontSize',16)
ylabel('Frequency per year','FontSize',16)
ax1=gca;
figure(gcf);pause
i=find(~~N); % find only the positive values for N.
Y=log(N(i)/year);
X=[ones(length(i),1) edges(i)'];
B=X\Y;
Yest=X*B;
% Now use Matlab's regress to produce the same regression coefficients:
[b,bint,r,rint,stats] = regress(Y,X);
% Note for other problems, more significant figures will be needed for
% reporting the regression coefficients and 95% CI's
fprintf('\nThe Y intercept = %6.3f with 95% CI = [%6.3f %6.3f]\n',...
    b(1),bint(1,1),bint(1,2));
fprintf('The slope is %6.3f with 95% CI = [%6.3f %6.3f]\n',...
    b(2),bint(2,1),bint(2,2));
fprintf(...
    'R^2 = %4.1f%%; F = %5.1f with p-value = %7.2g; RMSE = s^2 = %6.3g\n',...
    stats(1)*100,stats(2),stats(3),stats(4))
P=2*(1-tcdf(sqrt(stats(2)),df));
fprintf('P(t >= %4.1f |Ho: slope = 0 & %3.0f df) = %7.2g\n',...
    sqrt(stats(2)),df,P)
semilogy(edges,N/year,'or',edges(i),exp(Yest),'-m');
axis([Richter-0.1 6.7 0.009 109])
xlabel('Richter Scale','FontSize',16)
ylabel('Frequency per year','FontSize',16);
s2=sprintf('N=%8.2f e^{%6.3f R}',exp(B(1)),B(2));
text(4.2296,10,s2,'FontSize',20);
title(s,'FontSize',20)
figure(gcf);pause

% Redo the analysis with for earthquakes greater than 4 with log_10 scaling to
% see whether the slope is -1 to match the figure
% in the text.
Y=log10(N(i)/year);
X=[ones(length(i),1) edges(i)'];
B=X\Y;
Yest=X*B;

% Now use Matlab's regress to produce the same regression coefficients:
[b,bint,r,rint,stats] = regress(Y,X);
% Note for other problems, more significant figures will be needed for
% reporting the regression coefficients and 95% CI's

```

```
fprintf('\nThe Y intercept = %6.3f with 95%% CI = [%6.3f %6.3f]\n',...
    b(1),bint(1,1),bint(1,2));
fprintf('The slope is %6.3f with 95%% CI = [%6.3f %6.3f]\n',...
    b(2),bint(2,1),bint(2,2));
fprintf(...
'R^2 = %4.1f%%; F = %5.1f with p-value = %7.2g; RMSE = s^2 = %6.3g\n',...
    stats(1)*100,stats(2),stats(3),stats(4))
P=2*(1-tcdf(sqrt(stats(2)),df));
fprintf('P(t >= %4.1f |Ho: slope = 0 & %3.0f df) = %7.2g\n',...
    sqrt(stats(2)),df,P)
```

```
semilogy(edges,N/year,'or',edges(i),10.^Yest,'-m');
axis([Richter-0.1 6.7 0.009 109])
xlabel('Richter Scale','FontSize',16)
ylabel('Frequency per year','FontSize',16);
s2=sprintf('N=%8.2f 10^{%6.3f R}',10.^B(1),B(2));
text(4.2296,10,s2,'FontSize',20);
title(s,'FontSize',20)
figure(gcf);pause
```

```
% earthquakes_Japan.m, modeled on earthquakes_LA.m
% Modeled after LMFig010101poissr.m & earthquakes_Haiti.m
% Data from USGS/NEIC (PDE) 1973 - 2010 12 31
% http://earthquake.usgs.gov/earthquakes/eqarchives/epic/epic_rect.php
% Search all years
% Japan 9.0 epicenter 38.22 to 142.369 (searched 2.5 n, s e & w
% All quakes greater than 2.0
% Using example 4.3.5 as a model
% Example 4.3.5 Memphis earthquakes and the Poisson distribution
% Pp 305-306 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written Fall 2010 by Eugene Gallagher for EEOS601 Revised: 1/9/11
% Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Calculate the exact probability that 2 or more Richter 5 earthquakes will
% strike next year using the Poisson distribution and compare that with the
% normal approximation based on the central limit theorem.
japaneq; % calls the data file.
Richter=4; % Could be set to any value greater than 3
year=38; % Jan 1973 to 3/16/2011
k=find(DATA>=Richter & DATA<9);
s=sprintf('Japanese Earthquakes >= Richter Scale %3.1f & R < 9',Richter);
edges=Richter:.5:ceil(max(DATA(k)));
N=histc(DATA(k),edges);bar(edges,N/year,'histc');
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
title(s,'FontSize',20);
```

```

xlabel('Richter Scale','FontSize',16)
ylabel('Frequency per year','FontSize',16)
ax1=gca;
figure(gcf);pause
i=find(~N); % find only the positive values for N.
Y=log(N(i)/year);
X=[ones(length(i),1) edges(i)'];
B=X\Y;
Yest=X*B;
% Now use Matlab's regress to produce the same regression coefficients:
[b,bint,r,rint,stats] = regress(Y,X);
% Note for other problems, more significant figures will be needed for
% reporting the regression coefficients and 95% CI's
fprintf('\nThe Y intercept = %6.3f with 95% CI = [%6.3f %6.3f]\n',...
    b(1),bint(1,1),bint(1,2));
fprintf('The slope is %6.3f with 95% CI = [%6.3f %6.3f]\n',...
    b(2),bint(2,1),bint(2,2));
fprintf(...
'R^2 = %4.1f%%; F = %5.1f with p-value = %7.2g; RMSE = s^2 = %6.3g\n',...
    stats(1)*100,stats(2),stats(3),stats(4))
P=2*(1-tcdf(sqrt(stats(2)),df));
fprintf('P(t >= %4.1f |Ho: slope = 0 & %3.0f df) = %7.2g\n',...
    sqrt(stats(2)),df,P)
semilogy(edges,N/year,'or',edges(i),exp(Yest),'-m');
axis([Richter-0.1 9.1 0.001 109])
xlabel('Richter Scale','FontSize',16)
ylabel('Frequency per year','FontSize',16);
s2=sprintf('N=%8.2f e^{ %6.3f R }',exp(B(1)),B(2));
text(6,10,s2,'FontSize',20);
title(s,'FontSize',20)
figure(gcf);pause

% Redo the analysis for earthquakes greater than 4 with log_10 scaling to
% see whether the slope is -1 to match the figure
% in the text.
Y=log10(N(i)/year);
X=[ones(length(i),1) edges(i)'];
B=X\Y;
Yest=X*B;

% Now use Matlab's regress to produce the same regression coefficients:
[b,bint,r,rint,stats] = regress(Y,X);
% Note for other problems, more significant figures will be needed for
% reporting the regression coefficients and 95% CI's
fprintf('\nThe Y intercept = %6.3f with 95% CI = [%6.3f %6.3f]\n',...

```



```

b(1),bint(1,1),bint(1,2));
fprintf('The slope is %6.3f with 95%% CI = [%6.3f %6.3f]\n',...
b(2),bint(2,1),bint(2,2));
fprintf(...
'R^2 = %4.1f%%; F = %5.1f with p-value = %7.2g; RMSE = s^2 = %6.3g\n',...
stats(1)*100,stats(2),stats(3),stats(4))
P=2*(1-tcdf(sqrt(stats(2)),df));
fprintf('P(t >= %4.1f |Ho: slope = 0 & %3.0f df) = %7.2g\n',...
sqrt(stats(2)),df,P)

```

```

semilogy(edges,N/year,'or',edges(i),10.^Yest,-'m');
axis([Richter-0.1 9.1 0.009 109])
xlabel('Richter Scale','FontSize',16)
ylabel('Frequency per year','FontSize',16);
s2=sprintf('N=%8.2f 10^{%6.3f R}',10.^B(1),B(2));
text(6,10,s2,'FontSize',20);
title(s,'FontSize',20)
figure(gcf);pause
hold off

```

```

semilogy(edges,N/year,'om',edges(i),10.^Yest,'m','LineWidth',3);
set(gca,'ytick',[0.0001 0.001 0.01 0.1 1 10 100])
axis([Richter-0.1 9.1 0.00007 109])
xlabel('Richter Scale','FontSize',16)
ylabel('Frequency per year','FontSize',16);
s2=sprintf('N=%8.2f 10^{%6.3f R}',10.^B(1),B(2));
text(6,10,s2,'FontSize',20);
title(s,'FontSize',20)
hold on
% Continue the line from R 7.5 to R 9
X=[1 7.5;1 9];
Yest=10.^(X*B);
semilogy(X(:,2),Yest,'m','LineWidth',3);
grid
figure(gcf);pause
Richter=4;
year=38; % Jan 1973 to 3/16/2011
k=find(DATA>=Richter);
s=sprintf('Japanese Earthquakes >= Richter Scale %3.1f',Richter);
edges=Richter:.5:ceil(max(DATA(k)));
N=histc(DATA(k),edges);
i=find(~N); % find only the positive values for N.
Y=log10(N(i)/year);
X=[ones(length(i),1) edges(i)'];
B=X\Y;

```

```
Yest2=X*B;  
semilogy(edges,N/year,'or',edges(i),10.^Yest2,'r','LineWidth',3);  
set(gca,'ytick',[0.0001 0.001 0.01 0.1 1 10 100])  
axis([Richter-0.1 9.1 0.00007 109])  
figure(gcf)  
pause  
hold off
```

Figure 1.1.2 Sheep astragalus, an early form of dice

2.1.1 Probability: the early years

Figure 1.13. Sample space for a fair die.

2.1.2 **Statistics from Aristotle to Quetelet**

2.1.2.1 *Staatenkunde*: The comparative description of states

2.1.2.2 Political arithmetic, “the art of reasoning by figures, upon things relating to government.”

2.1.3 *Staatenkunde*: the comparative description of states.

Figure 1.1.4

2.1.4 Quetelet the catalyst

2.2 Some Examples

Case Study 1.2.1 S& P Stock prices

function LMcs010201_4th

% LMcs010201_4th.m

% Opening and closing prices of the S&P 500 from 17 February 1994 and 9

% February 1996, the dates listed in Larsen & Marx 4th edition p. 12

% Last Revised: 3/17/2011

% Written by Eugene Gallagher, Eugene.Gallagher@umb.edu

% Data obtained from

% % <http://finance.yahoo.com/q/hp?s=GSPC+Historical+Prices>

% Note that Yahoo produces a column of closing prices with the most

% recent closing price first, so the data will have to be flipped using

% the built-in function flipud(DATA)

% An excel spreadsheet was used to get the calendar day.

```
SandP500=[656.37 723
```

```
656.07 722
```

```
{Lots of data deleted here}
```

```
467.69 2
```

```
470.34 1];
```

```
SandP500=flipud(SandP500); % day 1, 2nd column is 2/17/1994; % Calendar day
```

```
[r,c]=size(SandP500);
```

```
tradingday=[1:r]';
```

```
SandP500Close=flipud(SandP500);
```

```
SandP500=[SandP500 tradingday]; % concatenate the matrix
```

```
% Create labeling string of calendar dates using built-in functions.
```

```
% Note since 2/16/1994 is day 0, the datenum for 0 is 2/16/1994;
```

```
datenumb=SandP500(:,2)+datenum('16.02.1994','dd.mm.yyyy');
```

```
% finds the starting date, assigns it to 0 and then assigns the proper
```

```
% Matlab number indicating the calendar date for plotting with datetick
SandP500Close=SandP500(:,1);
% gets the closing prices from the 1st column of SandP500
plot(tradingday,SandP500Close)
xlabel('Trading Day');ylabel('S & P 500 Closing Price')
lt=sprintf('Case Study 1.2.1; S & P 500, 1994-1996'); % to label graphs
title(lt,'FontSize',22);figure(gcf);pause
plot(datenumb,SandP500Close);
datetick('x',2)
axis([min(datenumb)-50 max(datenumb)+50 floor(min(SandP500Close)/50)*50 ...
    ceil(max(SandP500Close)/50)*50]);
% axis([min(datenumb)-20 max(datenumb)+20 400 700]) % will scale axes
% better, but optional
title(lt,'FontSize',22)
figure(gcf); pause

% Find the number of runs up and down (L&M calculate just downturns)
% Use Gallagher's runs2 to analyze the observed number of runs
[obsruns,out]=runs2(SandP500Close)
% obsruns contains the observed number of runs, up and down
% Using the formula on page 13 (L&M 4th edition), calculate the expected
% number of runs if the probability of the S & P going up
% is 0.5:
[r,c]=size(obsruns);
k=[1:r]; % Solve for runs of length 1, 2, 3, ... max run length.
p=.5; % probability of the run going up (or down)
T=out(1); % Total number of runs
% Using the formula on page 13, calculate the expected number of runs if
% the probability of the S & P going up
% is 0.5:

% This is the formula immediately above Table
% 1.2.2 on page 13 in Larsen & Marx.
Expected=T*p.^(k-1)*(1-p);
% Comparing the Observed and Expected, they seem quite close, with maybe too many 2-day
runs:
disp('No. of Runs Observed Expected')
disp([obsruns Expected])
Observed=obsruns(:,2);
% Plot the data in a stacked histogram
bar(1:length(Observed),[Observed Expected],1,'grouped')
%set(gca,'ytick',0:100:300,'xticklabel',{'0 ','1 ','2 ','3 ','4 ','5 ','6 ','7+'}, 'FontSize',20);
legend('Observed','Expectation')
xlabel('Number of Runs','FontSize',20)
ylabel('Frequency','FontSize',20)
```

```

title(lt,'FontSize',22);figure(gcf);pause
%-----
% Advanced from here on; Uses chapter 10 concepts.
% A slightly more advanced topic. Find the p that provides the best
% chi-squared fit to all 12 categories
% f calculates the chi-square goodness-of-fit statistic, to be covered in
% chapter 10
f=@(p,Observed,T,k) sum( (Observed- T*p.^(k-1).*(1-p)) .^2 ./ (T*p.^(k-1).*(1-p)) );
pfit = fminbnd(@(p) f(p,Observed,T,k),0.1,0.9);
fprintf('The chi-square statistic for p=%6.4f is %6.3f\n',0.5,f(0.5,Observed,T,k))
fprintf('The chi-square statistic for p=%6.4f is %6.3f\n',pfit,f(pfit,Observed,T,k))
% plot the results:
Expected2=T*pfit.^(k-1)*(1-pfit);
bar(1:12,[Expected2 Observed Expected],1,'grouped')
s=sprintf('Expected p = %6.4f',pfit);
legend(s,'Observed','Expected p=0.5')
xlabel('Number of Runs','FontSize',20)
ylabel('Frequency','FontSize',20)
title(lt,'FontSize',22);figure(gcf);pause

```

```

% Following Cochran's rule, the
% expected and observed counts will be pooled for those categories in
% which the expected values are less than 5
% This minimization routine calls function pooledchi2 below
pfit3 = fminbnd(@(p) pooledchi2(p,Observed,T,k),0.1,0.9);
% finds 0.5352
fprintf('The chi-square statistic for p=%6.4f is %6.3f\n',0.5,f(0.5,Observed,T,k))
fprintf('The chi-square statistic for p=%6.4f is %6.3f\n',pfit3,f(pfit3,Observed,T,k))
% plot the results:
Expected3=T*pfit3.^(k-1)*(1-pfit3);
disp('Observed Expected')
disp([Observed Expected3])
bar(1:12,[Expected2 Expected3 Observed Expected],1,'grouped')
%set(gca,'ytick',0:100:300,'xticklabel',{'0 ':'1 ':'2 ':'3 ':'4 ':'5 ':'6 ':'7+'}, 'FontSize',20);
s2=sprintf('Expected p = %6.4f',pfit3);
legend(s,s2,'Observed','Expected p=0.5')
xlabel('Number of Runs','FontSize',20)
ylabel('Frequency','FontSize',20)
title(lt,'FontSize',22);figure(gcf);pause

```

```

% This section just prints the results
[chi2p1,O1,E1,df1,pvalue1]=pooledchi2OE(0.5,Observed,T,k);
fprintf('The Chi-square statistic for p=%6.4f, with %2.0f df is %6.2f with p-value
%6.4f\n',0.5,df1,chi2p1,pvalue1)
disp('Observed Expected')

```

```

disp([O1 E1])
[chi2p2,O2,E2,df2,pvalue2]=pooledchi2OE(pfit,Observed,T,k);
fprintf('The Chi-square statistic for p=%6.4f, with %2.0f df is %6.2f with p-value
%6.4f\n',pfit,df2,chi2p2,pvalue2)
disp('Observed Expected')
disp([O2 E2])
[chi2p3,O3,E3,df3,pvalue3]=pooledchi2OE(pfit3,Observed,T,k);
fprintf('The Chi-square statistic for p=%6.4f, with %2.0f df is %6.2f with p-value
%6.4f\n',pfit3,df3,chi2p3,pvalue3)
disp('Observed Expected')
disp([O3 E3])

% Bar chart with pooled categories
bar(1:length(E1),[E1 O1 E3],1,'grouped')
set(gca,'ytick',0:50:150,'xticklabel',{'1 ':'2 ':'3 ':'4 ':'5+'}, 'FontSize',20);
legend('Expected p=0.5','Observed','Expected p=0.5226')
xlabel('Number of Runs','FontSize',20)
ylabel('Frequency','FontSize',20)
title('t','FontSize',22);figure(gcf);pause

function chi2p=pooledchi2(p,Observed,T,k)
Expected=T*p.^(k-1).*(1-p);
i=length(Expected)+1-find(cumsum(flipud(Expected))<5); % find the indices of cells less than 5
E=Expected;
E(i)=[]; e=E(end);while(e<5);E(end)=[];e=E(end);end
E(end)=E(end)+sum(Expected)-sum(E);
O=Observed(1:length(E));
O(end)=O(end)+sum(Observed)-sum(O);
chi2p=sum((O-E).^2./E);

function [chi2p,O,E,df,pvalue]=pooledchi2OE(p,Observed,T,k)
Expected=T*p.^(k-1).*(1-p);
i=length(Expected)+1-find(cumsum(flipud(Expected))<5); % find the indices of cells less than 5
E=Expected;
E(i)=[]; e=E(end);while(e<5);E(end)=[];e=E(end);end
E(end)=E(end)+sum(Expected)-sum(E);
O=Observed(1:length(E));
O(end)=O(end)+sum(Observed)-sum(O);
chi2p=sum((O-E).^2./E);
df=length(E)-1-1;
pvalue=1-chi2cdf(chi2p,df);


---


function [out,nruns]=runs2(DATA)
% format function [out,nruns]=runs2(DATA)
% input: DATA: a row or column vector
% output:

```

```
% out matrix: column 1 length of runs (1,2,... max run length)
% column 2 number of runs of each length in col 1
% nruns: [Totalruns, posruns, negruns, tiedvalues]
% posruns: number of positive runs
% negruns: number of negative runs
% tiedvalues: elements deleted from rank calcs
% because of exact ties in DATA
% calls the resample toolbox runs.m
% written by E. Gallagher, 5/27/03, revised 9/13/2010
% Eugene.Gallagher@umb.edu for EEOS601
[r,c]=size(DATA);
if c>r
    DATA=DATA'; % change row vector to column vector
    r=c
end
if r<3; disp('Vector must have more than 2 values'); return;end
% Find the tied values in the series, and delete redundant elements
% provides the positive and negative differences:
df=diff(DATA);
% find the differences not equal to zero, make 1
df=df~=0;
j=[1; df.*[2:r]]; % change indices of redundant elements in
% tied groups to zero
k=j(find(j>0)); % find the indices,k, just for the 1st untied vals
untiedd=DATA(k);n2=length(untiedd); % create the untied vector
%
% Now, find the runs using the differencing function in Matlab, diff:
ddiff=diff(untiedd); % create the vector of differences
signd=ddiff>0; % change positive diffs to 1, negative diffs to 0
% Use the runs.m file from the resample toolbox to find runs & lengths
[len,vals]=runs(signd); % runs.m from the resample toolbox
nruns=length(len);
maxl=max(len); % find the largest run size
% create matrix of run lengths(1st column) and observed number of runs
% Using the built in Matlab function histc, designed for creating histogram
% bins, but very useful for adding categories.
% built in matlab function, histc
sizeruns=[1:maxl]';
N=histc(len,sizeruns)';
% create the output matrix out
out=[sizeruns N];

% Find the positive runs by deleting extra 1's see above;
df=diff(signd);
% find the differences not equal to zero, make 1
```

```
df=df~=0;
lsd=length(signd);
j=[1; df.*[2:lsd]]; % change index of later ties to zero
k=j(find(j>0)); % find the indices just for the 1st untied vals
posdiffs=signd(k);posruns=sum(posdiffs);
Totinruns=sum(out(:,1).*out(:,2));
% Number of elements deleted due to exact ties:
tiedd=r-Totinruns-1;
nrns=[nrns posruns nrns-posruns tiedd];
```

```
function [len, vals] = runs(data)
% RUNS --- finds runs (sequences of repeated values)
% runs(data)
% gives the lengths of runs in the data set.
%
% [len, vals] = runs(data)
% gives both the length of the runs and the value of that run.
% from the Resampling toolbox
```

```
% (c) 1998-9 by Daniel T. Kaplan, All Rights Reserved
% Version 1.0
```

```
% put the data into row format
[r,c] = size(data);
if r~=1 & c~=1
    error('runs: input must be a vector.');
```

```
end
if c==1
    data = data';
end
```

```
% find the starts of the runs
inds = [1, (1+find(0 ~= diff(data)))];
```

```
len = diff([inds,length(data)+1]);
vals = data(inds);
```

Case Study 1.2.2: Golden rectangle

```
% LMcs010202_4th.m
% p 13-15 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by E. Gallagher 9/16/2010
% Revised 3/18/2011
% Note that these data are used in problem 14.2.6 (p 810 as an application)
% of the sign test
```

```
DATA=[0.693 0.749 0.654 0.670
      0.662 0.672 0.615 0.606
      0.690 0.628 0.668 0.611
      0.606 0.609 0.601 0.553
      0.570 0.844 0.576 0.933];
DATA=DATA(:);
mn=mean(DATA)
expected=0.5*(5.^0.5-1)
hist(DATA);
xlabel('Width-to-Length Ratio of Shoshoni Rectangles')
ylabel('Frequency')
% plot a vertical line to the top of the graph
% to margins of the graph are contained in axis
v=axis;
ax1=gca; % get the identifier for the current graph
h1=line([expected expected])
set(h1
h2=line([mn mn])
set(h2
legend('Observed')
title('Case Study 1.2.2')
figure(gcf)
pause
```

Case Study 1.2.3: Lunacy

```
% LMcs010203_4th.m
% Pages 17-18 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% written by Eugene.Gallagher@umb.edu 3/18/2011
% for UMASS/Boston's EEOS601
DATA=[6.4 5.0 5.8
      7.1 13.0 9.2
      6.5 14.0 7.9
      8.6 12.0 7.7
      8.1 6.0 11.0
      10.4 9.0 12.9
      11.5 13.0 13.5
      13.8 16.0 13.1
      15.4 25.0 15.8
      15.7 13.0 13.3
      11.7 14.0 12.8
      15.8 20.0 14.5];
% plot a histogram of the data
bar(1:12,DATA,1,'grouped')
```



```
axis([0.6 12.4 0 27])
set(gca,'ytick',0:5:25,'xticklabel',{'Aug';'Sept';'Oct';'Nov';'Dec';...
    'Jan';'Feb';'Mar';'Apr';'May';'Jun';'Jul'},'FontSize',20);
legend('Before Full Moon','During Full Moon','After Full Moon','Location','NorthWest')
xlabel('Month','FontSize',20)
ylabel('Admission rates (Patients/Day)','FontSize',20)
title('Case Study 1.2.3','FontSize',22);figure(gcf);pause
```

Case Study 1.2.4 Cancer deaths

```
% LMcs010204_4th.m
% Case study 1.2.4, p. 17 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher 6/03, revised 3/17/11
IndexExp=[2.49
    2.57
    3.41
    1.25
    1.62
    3.83
    11.64
    6.41
    8.34];
Cancer=[147.1
    130.1
    129.9
    113.5
    137.5
    162.3
    207.5
    177.9
    210.3]
% Or from scratch
[r,c]=size(IndexExp); % find how many rows and columns;
% create a column vector of all 1's with the same
% number of rows as IndexExp. This is needed for the
% linear regression.
onex=ones(r,1)
% Create an X matrix with ones in the first column
% and the explanatory variable, IndexExp in the 2nd column;
X=[onex IndexExp]
% Do the leastsquares regression, solving for
% Y intercept B(1) and slope B(2))
B=X\Cancer
% Calculated the expected values
ExpY=X*B
% plot the data and expected values
```

```
plot(IndexExp,Cancer,'o',IndexExp,ExpY,'-r')
xlabel('Index of Exposure')
ylabel('Cancer deaths per 100,000 man years')
title('Figure 1.2.2')
figure(gcf)
```

2.3 Chapter Summary

References

Larsen, R. J. and M. L. Marx. 2006. An introduction to mathematical statistics and its applications, 4th edition. Prentice Hall, Upper Saddle River, NJ. 920 pp. {?}

Index

Central limit theorem.....	15
Chi-square test.....	10
degrees of freedom.....	11, 14-17, 20-23
Distributions	
Poisson.....	13, 15
Expected value.....	10
Galton.....	13
Least significant difference.....	23
Linear regression.....	25
Matlab.....	3, 5, 7-9, 13, 14, 16, 19, 22
Median.....	3
null hypothesis.....	11
P-Value.....	9-11
Parameter.....	11
Poisson.....	13, 15
Probability.....	1, 3, 7-11, 15, 18, 19
P-value.....	11
Regression.....	3, 7, 12-14, 16, 25
Runs.....	3, 8-10, 19, 21-23
sample.....	3, 18
Sign test.....	23
Statistic.....	1-3, 5, 8-11, 13, 15, 18, 20, 21, 23-26
variable.....	25