WEEK 5: CHAPTER 5, ESTIMATION

TABLE OF CONTENTS

Page:

List of Figures

List of Tables

List of m.files

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 All of Chapter 11

Understanding by Design Templates

Understanding By Design Stage I — Desired Results Week 5

LM Chapter 5 Estimation Read 5.1-5.4, 5.8, skip 5.5-5.7

G Established Goals

- Using the method of maximum likelihood, estimate parameters for pdfs from real data
- Construct and properly interpret the meaning of **margin of error** and confidence limits
- What are the properties of good estimators?
- Distinguish between frequentist and Bayesian approaches to parameter estimation

U Understand

- That much of science is based on estimating the parameters for man's models of nature
- The meaning of confidence intervals and margin of errors, especially of polling data

Q Essential Questions

- What's the distinction between a maximum likelihood estimate and a maximum likelihood estimator (L & M p 349)?
- Why aren't we all Bayesians?
- How many random permutations are needed for a Monte Carlo simulation?

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

- **anonymous function**, **confidence interval for a binomial parameter**, confidence limits, consistency, estimator, **finite correction factor**, **geometric mean**, **maximum-likelihood estimate**, **maximum likelihood estimator**, **likelihood function**, **margin of error**, MVUE, parameter, sample standard deviation and variance, statistic, sufficiency, **unbiased**

S *Students will be able to*

- Fit the binomial, gamma, geometric, lognormal, negative binomial, normal, Pearson, and Poisson distributions to environmental data using the method of maximum likelihood
- Estimate and interpret the margin of error in a poll

Understanding by Design Stage II — Assessment Evidence Week 5 (6/28-7/4 M)

Chapter 5: 5.1-5.3, 5.9

- **Post in the discussion section by 7/6 W 10 PM W**
  - Taleb in his book 'The Black Swan' coined the term ludic fallacy to describe the mistake of thinking that probability models based on the casino can model events like climate change, the stock market, real estate and other processes that can be affected strongly be extreme events. Do you agree? Read the synopsis of the ludic fallacy on Wikipedia. I'll provide a pdf of the relevant chapter from the book.
- **HW 4 Problems due Wednesday 7/6/11 W 10 PM**
  - **Basic problems (4 problems 10 points)**
    - **Problem 5.2.24** Bird Song p 363 Don't use method of moments. Fit with Matlab's geofit as in example 5.2.1. Enter the data using X=[repmat(1,132,1); repmat(2,52,1); repmat(3,34,1); repmat(4,9,1); repmat(5,7,1); repmat(6,5,1); repmat(7,5,1); repmat(8,6,1)];
    - Problem 5.3.2 (p 376) Methylmercury in females, use case study 5.3.1 as a model. You must assume $\sigma=8$.
    - Problem 5.38 (p. 377) Tuna salads
    - Poll problem. Find a poll dealing with the upcoming 2012 elections and find out how the poll reports the margin of error and whether the difference between candidates is regarded as significant or not. Use Gallagher's pollsig.m to assess the difference in proportions of the polls.
  - **Advanced problems (2.5 points each)**
    - **Problem** 5.26 Use Example 5.2.2 as a model (1 would be a good preliminary estimate for theta)
    - Case Study 5.81. Update the posterior probability distribution incorporating the new information that there were 9 hurricanes that struck the mainland US from 2001 to 2004. Does this updated information appreciably change predictions about hurricane landfalls? http://www.nhc.noaa.gov/pastdec.shtml
  - **Master problems (1 only, 5 points)** Case Study 5.2.2 analyzes 69 years of hurricane data and fits the data to the gamma distribution. Assume that the gamma distribution is an adequate fit to these data. Calculate the expected rainfall for the storm of the century resulting from hurricanes that have moved inland.

Introduction

Chapter 5 is tough sledding. The chapter introduces a few simple concepts like the confidence limit and margin of error. But, the chapter starts off early and fast by wading into some of the most conceptually difficult areas of statistics such as deriving estimates for parameters using the method of moments and maximum likelihood and characterizing the quality of these estimates. We'll skip the methods of moments in this course, in order to focus on maximum likelihood estimates. Matlab will help in that it has functions which will solve readily for the MLE estimators. At worst, you'll have to write your own one-line **anonymous function** to find the MLE for a function. Larsen & Marx's chapter finishes with a brief and difficult section on Bayesian analysis. We'll just skim this section, noting that Bayesian inference has emerged as the state-of-the-art method for performing statistical analysis and mathematical modeling in many areas of the environmental and health sciences.

We'll use Matlab's statistics toolbox's maximum likelihood estimator programs to fit real-life data to probability models. For example, Case Study 5.2.2 fits the 2-parameter gamma distribution to hurricane rainfall data using the method of moments. We'll use a the built-in function gammafit to fit the two-parameter gamma distribution to rainfall from 36 hurricanes over 69 years, as shown in Figure 1. In modeling extreme events, it is vitally important to be able to model and predict the probability of extreme events in the right tail of the distribution. The gamma distribution is one of the more useful distributions for modeling processes with distributions that are positively skewed.

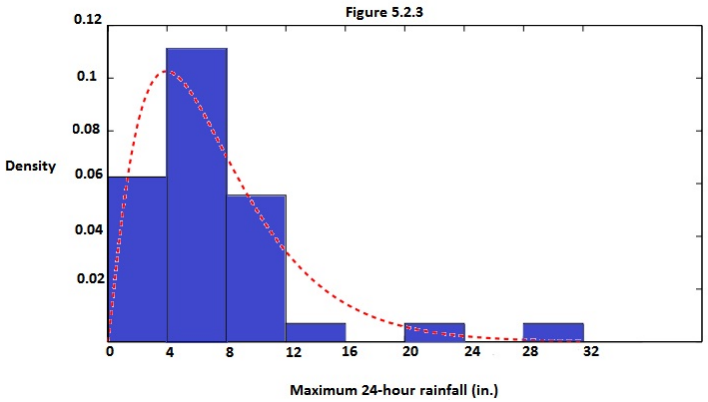

Figure 1. Modeled on Figure 5.2.2

Some definitions

Any function of a random sample whose objective is to approximate a parameter is called a statistic. If θ is the parameter being approximated, its **estimator** is called $\hat{\theta}$, and the resulting number is called an **estimate**. Most of our common statistics (mean, standard deviation) are **maximum likelihood estimators**.

Definition 5.2.1 Let k_1, k_2, ..., k_n be a random sample of size *n* from the discrete pdf $p_X(k; \theta)$ where θ is an unknown parameter. The ***likelihood function***, *L(θ)*, is the product of the pdf evaluated at the n k_i's. That is

$$L(\theta) = \prod_{i=1}^{n} p_X(k_i; \theta)$$

If $y_1, y_2, ..., y_n$ is a random sample of size *n* from a continuous pdf, fy(y;θ), where θ is an unknown parameter, the **likelihood function** is written

$$L(\theta) = \prod_{i=1}^{n} f_Y(y_i; \theta)$$

Definition 5.2.2 Let $L(\theta) = \prod_{i=1}^{n} p_X(k_i; \theta)$ and $L(\theta) = \prod_{i=1}^{n} f_Y(y_i; \theta)$ corresponding to random samples $k_1, k_2, ..., k_n$ and $y_1, y_2, ..., y_n$, drawn from the discrete pdf $p_X(k; \theta)$ and continuous pdf $f_Y(y; \theta)$, respectively, where θ is an unknown parameter. In each case let θ_e be a value of the parameter such that $L(\theta_e) \geq L(\theta)$ for all possible values of θ. Then θ_e is called a **maximum likelihood estimate** for θ.

Definition 5.3.1 The **margin of error** associated with an estimate $\frac{k}{n}$, where k is the number of successes in *n* independent trials, is 100d%, where

$$d = \frac{1.96}{2\sqrt{n}}$$

Definition 5.4.1 Suppose that $Y_1, Y_2, ..., Y_n$ is a random sample from the continuous pdf $f_Y(y; \theta)$, where θ is an unknown parameter. An estimator $\hat{\theta}$ $(=h(Y_1, Y_2, ..., Y_n))$ is said to be **unbiased** (for θ) if $E(\hat{\theta}) = \theta$ for all θ. (The same concept and terminology apply if the data consist of a random sample $X_1, X_2, ..., X_n$ drawn from a discrete pdf $p_X(k; \theta)$.

By definition, the **geometric mean** of a set of *n* numbers is the *n*th root of their product. P. 385 Note that operationally, the geometric mean is the back-transformed arithmetic mean of log transformed random variables.

Maximum likelihood estimation

Examples & Case Studies

Example 5.1.1

Three coin tosses produce the result HHT. What is the maximum likelihood estimator for the p, the probability of tossing a head? The probability that X_1 is H is p, the probability that X_2 is H is also *p* and the probability that X_3 is T is (1-*p*). If X_1, X_2, and X_3 are independent events, then the probability of X_1, X_2, X_3 is $f=p^2(1-p)$. What is the value of *p* that maximizes that probability? As shown in Figure 2, it is the maximum of the likelihood function *f*, or the point at which the 1st derivative of *f* is 0: solve(diff(p.^2*(1-p))), or 2/3. Matlab's binofit, as implemented in Gallagher's LMex050101_4th.m will solve not only for the maximum likelihood estimate for p but will calculate the 95% confidence intervals for that estimate:

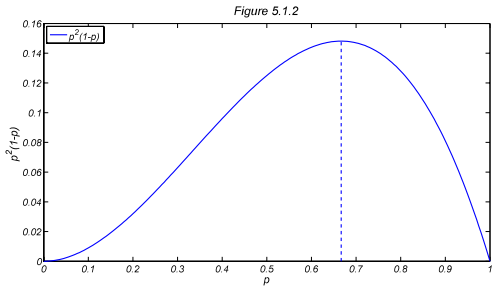

Figure 2. Probability of observing HHT for different values of p.

X=[1 1 0]; N=length(X);[Phat,PCI] = binofit(sum(X),N,0.05);

The maximum likelihood estimate $\hat{p}$ is b with a very broad 95% confidence interval from 0.094 to 0.992. Confidence intervals for binomial proportions are introduced on Larsen & Marx page 369, Theorem 5.3.1.

Example 5.2.1

This is an interesting example for me. The first time I worked with it, I thought I'd found an error in the Matlab maximum likelihood fitting function for the geometric distribution. I happily emailed the head of Matlab's statistical toolbox crew with my analysis. He reported back that there were two different ways of defining the geometric distribution. Larsen & Marx use one and the Mathworks uses the other, but the Larsen and Marx version can be quickly solved with the Matlab maximum likelihood estimator function.

The problem states that there were four independent observations representing the geometric probability model $p_X(k)=(1-p)^{k-1}p$, k = 1, 2, 3, ... Find the maximum likelihood estimate for p. Recall that the geometric distribution was introduced in section 4.4. Larsen & Marx define X as the trial on which the first success occurs if at each trial, the probability of success is p. The Mathworks follow a second widely used definition of the geometric distribution as the number of failures before the first success occurs with their geometric distribution defined as

$$p_{X\ Matlab}(k)=(1-p)^k p,\ k = 1, 2, 3, ...$$

Matlab's geometric distribution functions can be called to solve Larsen & Marx's problems with the conversion that $X_M=X-1$, where X_M is the random variable used in the Matlab functions. For

example, in L & M the four independent observations of the trial at which the first success occurred could be represented by the vector X=[3 2 1 3]. To use Matlab's maximum likelihood estimation routine, just call maximum likelihood estimate function, mle:
[Phat,PCI] =mle(X-1,'distribution','geometric')

This will produce the maximum likelihood estimate of 4/9 with the 95% confidence intervals of 0.120 and 0.769. In my m.file LMex050201_4th.m, I follow the book's approach of differentiating the likelihood function. Follow it if you wish.

Example 5.2.2

This is an interesting example, because it shows how Matlab's mle function can be used to solve for a non-standard pdf. In this case, a continuous model was specified:

$$f_Y(y;\ \theta) = \frac{1}{\theta^2}\, y\, e^{y/\theta},\ 0 < y < \infty;\ 0 < \theta < \infty$$

To solve this non-standard pdf, we'll write an **anonymous function** in Matlab. As covered in Handout 3, Matlab has script programs and named function m.files. In most of the optimization routines that are fitting parameters to data or solving equations, there is a function to be minimized, or solved, or maximized. Oftentimes, these are named. But, occasionally you'll only need a function once so why bother cluttering up your hard drive with another file. That's were anonymous files are useful. The above function can be programmed in Matlab in one line and named fyytheta. The entire program that enters the data for this problem and finds the maximum likelihood estimate for theta takes only 4 lines.

```
X=[9.2 5.6 18.4 12.1 10.7];                                  % The data
fyytheta=@(y,theta)(1./theta.^2.*y.*exp(-y./theta));         % The anonymous function
theta=10;                                                    % This is an initial guess
[theta, thetaCI]= mle(X,'pdf',fyytheta,'start',theta,'lowerbound',0)   % The mle estimation
```

These four statements find the maximum likelihood estimate for θ, 5.6 and the 95% CI for this estimate: 2.13 and 9.07.

What does it mean for 5.6 to be maximum likelihood estimate for the parameter θ? For each value of theta, the function fyytheta generates a probability density function. From this pdf, the probability of each datum can be calculated, and if we assume that the data are random variables, the probability of the event that gave rise to the sample is the likelihood. The likelihood is the product of these probabilities. The pdf for the maximum likelihood estimate of

$\hat{\theta}$ =5.6 is shown in Figure 3 with the 5 data. The likelihood is the product of these five probabilities, which is 1.84 x 10^{-7}.

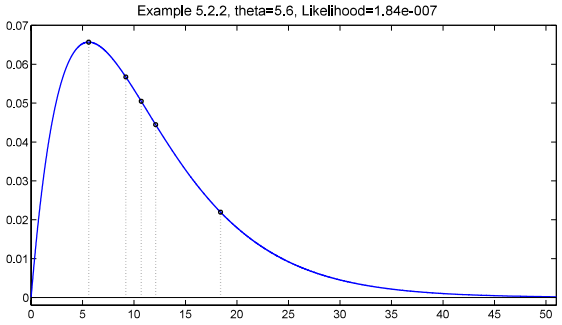

Figure 3. The pdf for the model with θ=5.6. Also shown are the probabilities for the five observations with likelihood 1.84 x 10^{-7}.

But what would the pdf look like for a different θ, say 11.2? Shown in Figure 4 is the pdf for θ=11.2 with the 5 observations. With this $\hat{\theta}$, the likelihood is 2.66 x 10^{-8}, which is only 14% of the maximum likelihood.

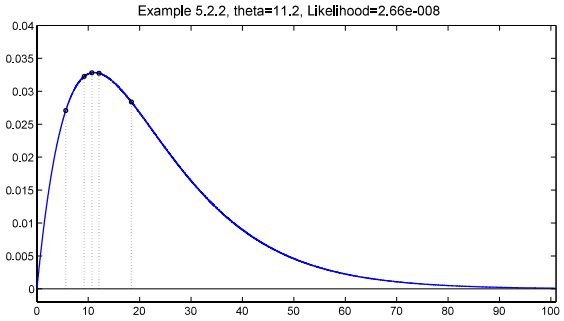

Figure 4. The pdf for the model with θ=11.2. Also shown are the probabilities for the five observations with likelihood 2.66 x 10^{-8}.

Similarly, we can plot the pdf if θ were 2.8 as shown in Figure 5 with likelihood of 8.54 x 10^{-9}, only 4.6% of the maximum likelihood with θ=5.6. So, since there is no estimate of θ that will produce a likelihood larger than $\hat{\theta}$ = 5.6, 5.6 is the maximum likelihood estimate. As noted in the book, it is a single number so it is an estimate, not an estimator. An estimator describes a random variable with a mean, variance, and pdf.

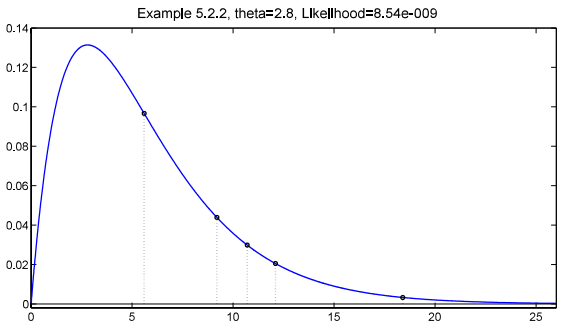

Figure 5. The pdf for the model with θ=2.8. Also shown are the probabilities for the five observations with likelihood 8.54 x 10^{-9}.

What would happen to our maximum likelihood estimate, based on 5 independent observations, if we added an additional independent observation, say 30. It would change the maximum likelihood estimate from 5.6 to 7.2 as shown in Figure 6. The likelihood is now 1.23 x 10^{-9}, but this likelihood can not be compared to the likelihoods based on only 5 samples. This likelihood must be compared with likelihoods based on the same 6 observations.

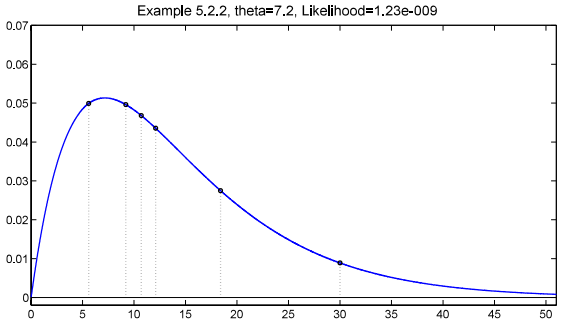

Figure 6. The pdf for the model with θ=7.2, the maximum likelihood estimate if an additional observation of 30 is added to the original 5 observations, shown above. The likelihood is 1.23 x 10^{-9}.

Figure 7 shows the pdf and likelihood if we'd used the previous maximum likelihood estimate, $\hat{\theta}$ =5.6. The likelihood is 8.29 x 10^{-10}, only 67% of the likelihood with $\hat{\theta}$ = 7.2. With that extra observation of 30, no other estimate of theta produces a likelihood as low or lower than 1.23 x 10^{-9}, so 7.2 is the maximum likelihood estimate.

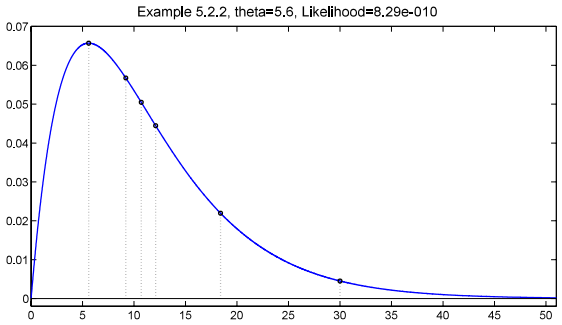

Figure 7. The pdf for the model with θ=5.6, which is no longer the maximum likelihood estimate if an additional datum of 30 is included. Also shown are the probabilities for the five observations with likelihood 8.29 x 10^{-10}.

Case Study 5.2.1

A major poll was conducted at the University of South Florida with the results shown in Table 1 below.

Table 5.2.1

Number of Major Changes	Observed Frequency	Expected Frequency
0	237	230.3356
1	90	100.2866
2	22	21.83205
3+	7	3.545751

Table 1. Results of a poll at the University of West Florida documenting the number of major changes. The final column is the result of fitting the Poisson model to these data.

It is pretty easy to fit the Poisson distribution to observations, but the Mathworks have made the task even easier by providing a maximum likelihood estimation function for the Poisson

distribution. These data were fit to the Poisson distribution using Matlab's Poisson maximum likelihood estimator function. The only non-obvious part to the fitting is that the data have to be expanded as if the poll results for each student were being entered:

```
X=[zeros(237,1);ones(90,1);2*ones(22,1);3*ones(7,1)];
[LAMBDAHAT, LAMBDACI] = poissfit(X)
```

This function produced $\hat{\lambda}$ =0.4354 with 95% confidence limits of 0.3669 and 0.5309. The fit of the data to the Poisson distribution shown in Figure 8 appears excellent. Using methods from Chapter 10, we will quantitatively analyze the fit, but a so-called chi-by-eye test indicates no problems with the use of the Poisson model to fit these data.

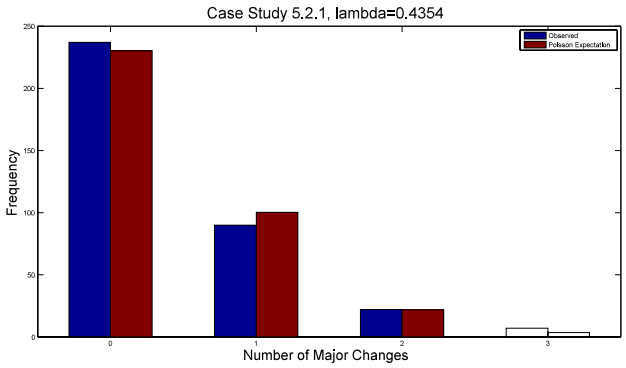

Figure 8. A grouped bar chart shows the observed number of major switches and those expected under the Poisson distribution with $\hat{\lambda}$ = 0.4354.

Case Study 5.2.2

Between 1900 and 1969, 36 hurricanes moved as far inland as the Appalachians. Listed in Table 2 are the maximum 24-h precipitation events in inches for these 36 storms.

Table 2. Maximum 24-h rainfall (inches) while 36 hurricanes were over the mountains.
RAIN=[31 2.82 3.98 4.02 9.5 4.5 11.4 10.71 6.31 4.95 5.64 5.51 13.4 9.72 6.47 10.16 4.21 11.6 4.75 6.85 6.25 3.42 11.8 0.8 3.69 3.1 22.22 7.43 5.0 4.58 4.46 8 3.73 3.5 6.2 0.67];

Figure 9 shows these 36 rainfall events with a superimposed fit to the gamma distribution. I used Matlab's gamfit to fit the two parameters of the gamma distribution. The Mathworks has a slightly different form of the gamma distribution, but Matlab's gamma b parameter is simply the inverse of Larsen & Marx's λ

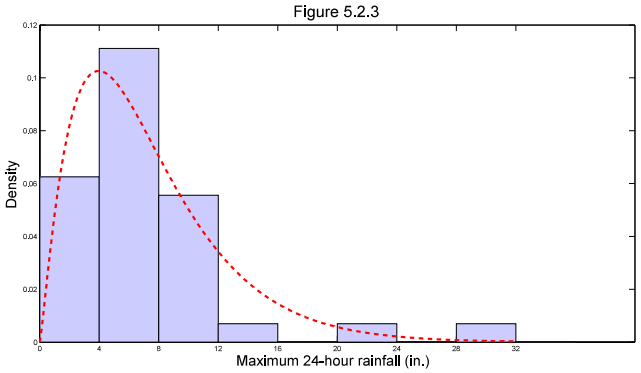

Figure 9. Maximum 24-h rainfall (inches) for 36 hurricanes from 1900 to 1969. Superimposed is the gamma distribution fit.

The gamma distribution, introduced in section 4.6, was used to fit the data using Matlab's gamfit maximum likelihood estimation routine.

The gamma distribution, introduced in Larsen & Marx section 4.6 Theorem 4.6.1, has two parameters, lambda and r.

Theorem 4.6.1 *Suppose that Poisson events are occurring at the constant rate of λ per unit time. Let the random variable Y denote the waiting time for the rth event. Then Y has pdf $f_Y(y)$ where*

$$f_Y(y) = \frac{\lambda^r}{(r-1)!} y^{r-1} e^{-\lambda y}, \quad y > 0.$$

The gamma distribution, indicated by Γ, is defined so that x! = Γ (x+1), so Theorem 4.6.1 can be rewritten as:

$$f_y(y; r, \lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}, \; y > 0$$

Matlab's gamma pdf function is described in two related parameters, *a* and *b*. Here is the equation from Mathworks' gampdf documentation:

$$y = f(x|a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-\frac{x}{b}}, \; x > 0$$

It is obvious that Matlab's a is Larsen & Marx's r and Matlab's b is the inverse of Larsen & Marx's λ.

Gamfit produces a MLE estimate of *r* of 2.187 with 95% CI: [1.421 3.367]. Larsen & Marx introduce the concept of a confidence limit later in Chapter 5 but don't provide the MLE estimator for the confidence limit for the Poisson distribution. Larsen & Marx (2006) use the methods of moments to find *r* =1.60, which is barely within the 95% CI identified by Matlab's. Matlab's *b* parameter is Larsen & Marx's 1/lambda. Matlab's MLE estimate of λ is 0.300 with 95% CI: [0.185 0.487]. Larsen & Marx find lambda=0.22, which is within Matlab's 95% CI.

The application of the gamma distribution has very little to do with the description of the gamma distribution in Theorem 4.6.1. But, the gamma distribution has some nice features in that it is a two-parameter distribution in which produces a pdf with only positive values with positive skew. This often can be a concise description of processes like rainfall events. Using the parameters from the 36 actual rainfall events, I generated random hurricane rainfall patterns with gamrand as shown in Figure 10. If

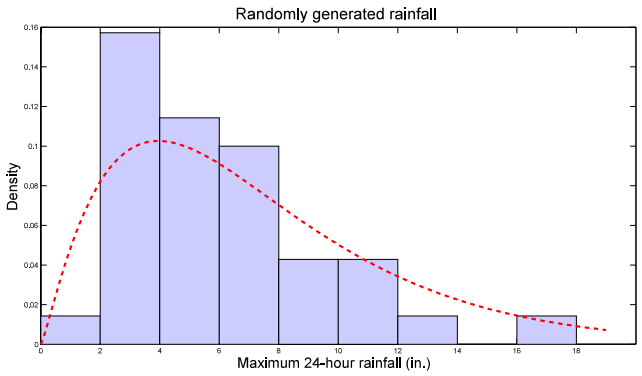

Figure 10. A random set of 36 rainfall events generated with the estimated gamma distribution parameters fit to the data in Table 2 and Figure 9. Superimposed is the gamma distribution.

one were insurance claims adjuster, it might be informative to run a century or so of storms to predict the storm of the century.

It is mildly disturbing that the Larsen & Marx estimates for the parameters found by the method of moments (r=1.6) and $\hat{\lambda}$ =0.22 differ from Matlab's maximum likelihood estimates of r=2.187 and lambda = 0.3. I calculated and plotted the two likelihood functions in Figure 11. The likelihood for the Matlab parameters is 3.5 x 10^{-45}, which is larger than the Larsen & Marx likelihood fit by the method of moments 1.4 x 10^{-45}. So, it is clear that Matlab's solution is preferred.

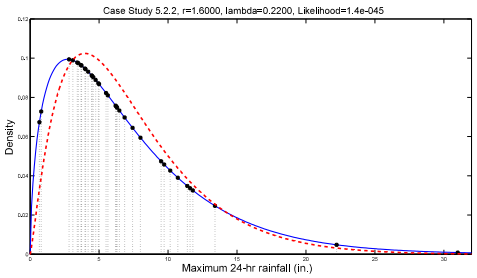

Figure 11. Larsen & Marx's fit (blue) and Matlab's MLE fit (red dashed line) are superimposed. Matlab's MLE has a larger likelihood.

Example 5.3.1

This example introduces the confidence limits for the normal distribution. If one collects a random sample from a distribution with mean μ and calculates the mean and a confidence interval for the mean, this confidence interval for μ will contain the unknown μ 95% of the time; 5% of the time, it will not contain μ. The probability that a single confidence interval contains μ is either 100% or 0%; μ is either inside or outside the CI. Figure 12 shows the 95% confidence interval for the standard normal distribution.

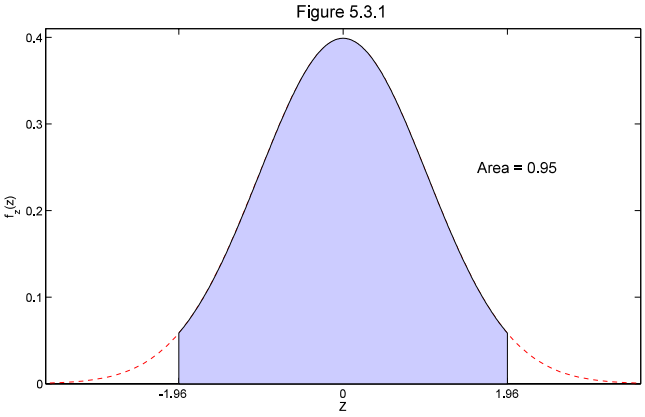

Figure 12. The 95% confidence region for the standard normal curve. Note the lower and upper cutoffs of -1.96 and 1.96.

Figures 13 and 14, based on Figure 5.3.2 in the text shows the result of 100 random samples of size 4 drawn from a normal population with μ=10 and σ = 0.8. For each one of these 100 random samples, a 95% confidence interval was created. By definition, in the long run 95% of these intervals should contain μ = 10. In the first figure, only 3 of 100 intervals did not contain μ=10, but in the second figure 5 of 100 did not contain μ=10.

Case Study 5.3.1

Larsen & Marx (p. 367-368) provide data on the maximum head breadths (mm) of 84 Etruscan males. Does the 95% CI for these skulls include the modern mean for Italian men, which is 132.4 with σ = 6?

This is mainly a computational exercise. We'll cover the appropriate 1-sample and 2-sample tests for this problem in Chapter 9. For now, we'll just calculate the 95% CI with a few Matlab statements, with the key assumption that the standard deviation is known and is 6. After we have a few more statistical tools available, we'll be able to use the sample standard deviation (s=5.9705) and a Student's *t* multiplier to calculate a more appropriate 95% confidence interval.

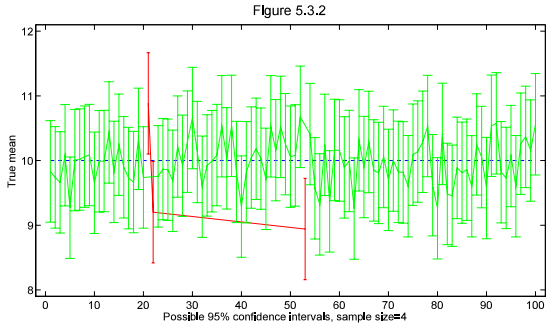

Figure 13. As described in the text, 100 confidence intervals are shown for samples of size 4 drawn from normal distribution with μ=10 and σ=0.8. While in the long run 5% of the confidence intervals wouldn't include μ, in these 100 trials, only 3 confidence intervals (marked in red and with means connected by a line) did not include μ = 10.

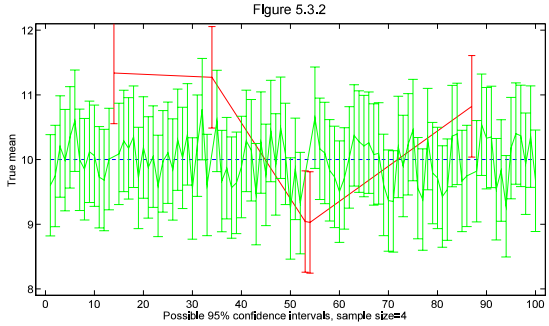

Figure 14. As described in the text, 100 confidence intervals are shown for samples of size 4 drawn from normal distribution with μ=10 and σ=0.8. While in the long run 5% of the confidence intervals wouldn't include μ, in these 100 trials, only 3 confidence intervals (marked in red and with means connected by a line) did not include μ = 10.

```
DATA=[141 148 132 138 154 142 150
   146 155 158 150 140 147 148
   144 150 149 145 149 158 143
   141 144 144 126 140 144 142
   141 140 145 135 147 146 141
   136 140 146 142 137 148 154
   137 139 143 140 131 143 141
   149 148 135 148 152 143 144
   141 143 147 146 150 132 142
   142 143 153 149 146 149 138
   142 149 142 137 134 144 146
   147 140 142 140 137 152 145];
DATA=DATA(:); % convert to a single column vector
meanD=mean(DATA);sigma=6
CI=[meanD-norminv(0.975)*sigma/sqrt(length(DATA)) ...
   meanD+norminv(0.975)*sigma/sqrt(length(DATA))];
```

```
fprintf('The mean is %4.1f with CI: [%4.1f %4.1f]\n',meanD,CI)
```

Figure 15 shows the distribution of these skull lengths, with a superimposed normal distribution (with σ=6). The formula for the 95% CI if σ is known is:

95% CI if σ known:

$$\left(\bar{y} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right).$$

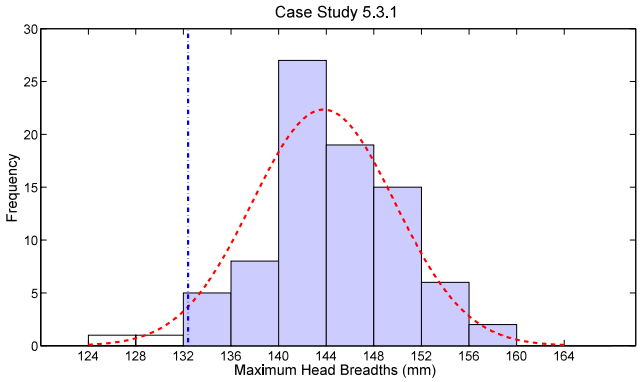

Figure 15. A histogram of 84 Etruscan maximum head breadths. I also plotted the normal pdf using the observed mean = 143.8 with σ = 6. I also plotted the maximum head breadth of modern Italian males (132.4).

The program above produces the pithy result that: The mean is 143.8 with CI: [142.5 145.1]. This result could also have been obtained using Matlab's normfit.m Since this confidence interval does not contain 132.4, one has strong evidence that the Etruscan males had head widths distinctly different from modern Italian males.

Confidence Interval for a binomial parameter *p*, Case Study 5.3.2

Theorem 5.3.1 defines the 95% **confidence interval for a binomial parameter** *p*. This is an important and useful formula, so I'll report it here:

Theorem 5.3.1 *Let k be the number of successes in n independent trials, where n is large and p = P (success) is unknown. An approximate 100(1-α)% confidence interval for p is the set of numbers*

$$\left(\frac{k}{n} - z_{\alpha/2} \sqrt{\frac{(k/n)(1-k/n)}{n}}, \frac{k}{n} + z_{\alpha/2} \sqrt{\frac{(k/n)(1-k/n)}{n}} \right)$$

I programmed this equation as LMTheorem050301_4th.m. By entering *k*, *n* and alpha, this Matlab function produces the confidence limits and observed p.

Case Study 5.3.1 applies this formula. A poll found that 713 of 517 respondents believed in the idea of intelligent life on other worlds. The observed proportion was 47. 0 (± 2.5) % (± half 95% CI). As we will see shortly, the **margin of error** for the poll is 2.5%.

Example 5.3.2

This example uses the median test to assess whether a random number generator is generating random numbers. In this example, Matlab's random number generator was used to generate 1000 samples of size 60 drawn from the exponential pdf:

```
n=60,trials=1000;Y=exprnd(1,n,trials);
```

A little calculus shows that the median of the exponential distribution with mean =1 is log(2). I calculated the proportion of samples of 60 with medians less than log(2). Under the null hypothesis that the numbers are random, the expected value is 0.5. You can run as many simulations as you want. In one simulation with 10,000 trials, 5.1% of the 95% confidence intervals did not include 0.5.

Example 5.3.3: Margin of Error

71% of 1,002 adults interviewed after Hurricane Charlie in August 2004 thought President Bush was doing a good job. What would the margin of error be associated with that poll? The two-line program finds the margin of error to be 3.1%

```
n=1002;ME=norminv(0.975)/(2*sqrt(n))*100;
fprintf('The margin of error is %3.1f%%\n',ME)
```

Example 5.4.1: Precision of an estimator

This simple example compares the precision of estimators of the binomial proportion with 10 coin tosses vs. 100 coin tosses. With 6 heads in 10 coin tosses, the probability that the true p is within 10% of the estimated value of 0.6 is 0.66647 from the binomial pdf. With 100 coin tosses, one can estimate the probability that the true p is within 10% using either the exact binomial distribution (p=0.968) or from the large-sample normal approximation to the binomial (p=0.959). Figure 16, replicating Figure 5.4.1 in the text shows these two areas.

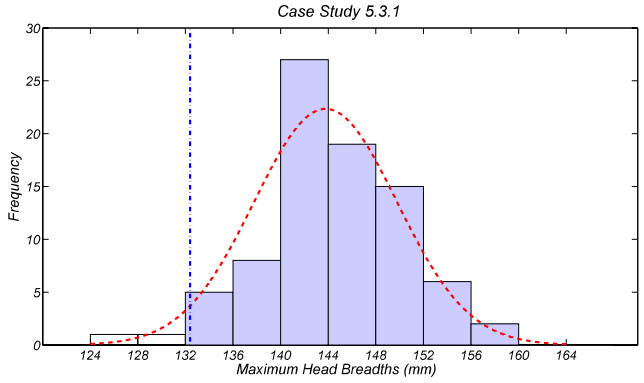

Figure 16. A histogram of 84 Estruscan maximum head breadths. I also plotted the normal pdf using the observed mean = 143.8 with σ = 6. I also plotted the maximum head breadth of modern Italian males (132.4).

Example 5.8.2 Bayesian analysis of a binomial parameter

This example uses the beta distribution as the prior and posterior distribution in a Bayesian analysis. Given prior information about the sale of videos, Larsen & Marx model the prior using the beta distribution. The store operator guessed that between 3% and 4% (mean or median 0.35) of his customers would be interested in buying his video and no more than 7%. A beta

distribution with parameters 4 and 102 produces a reasonable prior. In the text, **Larsen & Marx (2006)** provide a formula on how to update this prior distribution in light of new information. I applied this information after speculating that the a survey of customers might find that 4 out of 100 wanted to buy the video. What would the posterior distribution look like in light of the new data which indicates that far more than 3% are interested in buying the video. Figure 17 shows the prior and posterior distributions for this problem.

Obviously, the addition of the new data indicating 4 of 100 customers wanted to buy the video produces a much more precise estimate of the probability distribution for theta. The ability to update prior expectations in light of new information is one of the reasons that **Hilborn & Mangel (1997)** argued that ecologists should all become Bayesians. Bayesian inference was the best available way to update knowledge with new facts. On the flip side, ecologists would have to become more familiar with probability modeling in order to apply Bayesian methods. Most ecologists jump straight into hypothesis testing with t tests without stopping to ponder the underlying probability models that underlie statistical inference. Another argument against Bayesian methods is that it assumes that the world is parametric, that there are underlying parametric models that govern the behavior of real-life events. Bayesian analysis forces us to explicitly invoke these probabilistic models in order to generate the answers to our statistical questions.

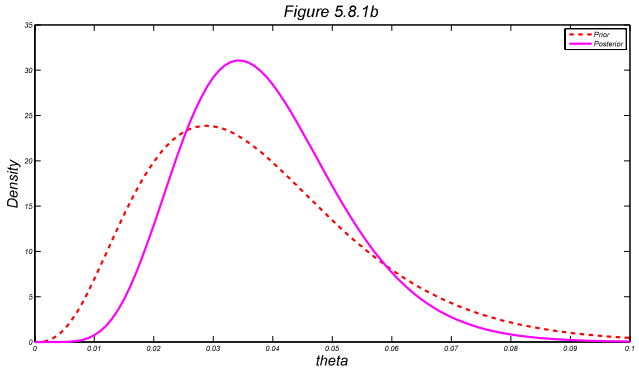

Figure 17. Prior (dashed) and posterior pdf's for theta given that a survey of users indicated that 4 of 100 wanted to buy the video.

Case Study 5.8.1: Bayesian analysis of Poisson parameter

In this case study, 150 years of hurricane landfall data were provided. Based on the first 50 years a Bayesian prior pdf using the gamma distribution was conducted. As shown in Figure 18, this Bayesian prior was updated with the next 100 years of hurricane information to produce the posterior pdf for hurricane landfall data.

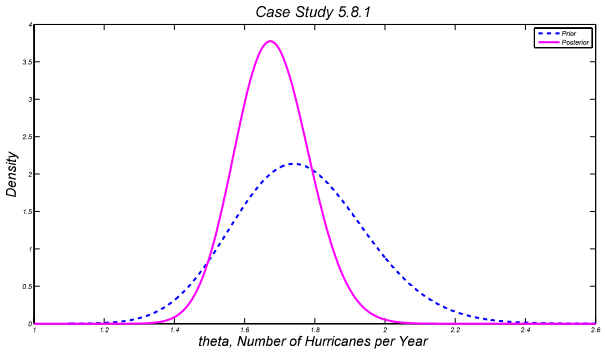

Figure 18. Prior (dashed) and posterior pdf's for theta, the number of hurricanes reaching landfall. The prior was based on the first 50 years of data and the posterior includes the subsequent 100 years of data.

Degrees of freedom **Box 1.2**

Statistical tests of significance often call upon the concept of degrees of freedom. A formal definition is the following: "The degrees of freedom of a model for expected values of random variables is the excess of the number of variables [observations] over the number of parameters in the model" (Kotz & Johnson, 1982).

In practical terms, the number of degrees of freedom associated with a statistic is equal to the number of its independent components, i.e. the total number of components used in the calculation minus the number of parameters one had to estimate from the data before computing the statistic. For example, the number of degrees of freedom associated with a variance is the number of observations minus one (noted $\nu = n - 1$): n components $(x_i - \bar{x})$ are used in the calculation, but one degree of freedom is lost because the mean of the statistical population is estimated from the sample data; this is a prerequisite before estimating the variance.

There is a different t distribution for each number of degrees of freedom. The same is true for the F and χ^2 families of distributions, for example. So, the number of degrees of freedom determines which statistical distribution, in these families (t, F, or χ^2), should be used as the reference for a given test of significance. Degrees of freedom are discussed again in Chapter 6 with respect to the analysis of contingency tables.

Figure 19. Degrees of Freedom discussed by Legendre & Legendre's (1998) Numerical Ecology.

On confidence limits & Neyman-Pearson theory

Larsen & Marx (2006) appears to be written firmly in the theory of hypothesis testing advocated by Jerzy Neyman & Egon Pearson. Neyman & Pearson, along with Fisher, were among the founders of the frequentist school of statistics. In this approach to statistical testing, an alpha level was established in advance of performing the statistical test, usually α=0.05. This is the probability of Type I error (Type I and Type II errors were Neyman-Pearson innovations), the probability of rejecting a true null hypothesis by chance. The critical value of the test statistic was identified before performing the test. After analyzing the data, the test statistic was compared to the critical value. If the test statistic, say Student's *t*, exceeded the critical value based on α and the degrees of freedom, then the test was judged as providing ***significant*** evidence that the null hypothesis was false. The results would be reported as, "We rejected the null hypothesis that $\mu_1 = \mu_2$ at α =0.05 level", or more briefly, "the test was ***significant*** at α = 0.05 {or whatever the chosen α was}." If the test statistic was less than the critical value, then the test was judged as providing ***insignificant*** evidence that the null hypothesis was false. The results would be reported as, "I failed rejected the null hypothesis that $\mu_1 = \mu_2$", or more briefly, 'the test was ***insignificant*** at α = 0.05." This approach to hypothesis testing made a great deal of sense when it was difficult to calculate the p value. A test could be performed and compared to

tabulated values of the critical values of test statistics at say the α=0.1, 0.05, and 0.001 levels of significance. Now, however, most reputable journals insist that the actual *p* values for the test be included. Matlab, for example, will calculate the *p* values for all of the major test statistics. Moreover, Matlab can perform the exact probability analysis for many of the major statistical tests and can perform randomization tests to produce *p* values to most desired levels of accuracy.

Ronald Fisher never accepted the Neyman-Pearson approach to hypothesis testing with critical values, insisting that the *p* values should be reported. He rejected the rather automatic decision rules of accept/don't accept, and significant/not significant. Indeed, **Mayo's (1996)** synthesis of frequentist hypothesis testing points out that Egon Pearson was less than enthused about the black-white, significant-nonsignificant dichotomy that the Neyman-Pearson approach entailed. Mayo argues forcefully that Neyman & Pearson's most influential contribution to scientific advance was not the significant/non significant dichotomy based on critical values but the introduction of confidence intervals. Confidence intervals, especially the 95% confidence interval allow an investigator to report the effect size and an estimate of the uncertainty on the effect size. It allows a ready assessment of not only whether the null hypothesis is false but also the relative likelihood of alternate hypotheses. In modern statistics, one of the chief allures of Bayesian inference is that it allows and assessment of the probability of a hypothesis before and after data are collected and analyzed.

Confidence limits usually have the following form: Effect size ± a multiplier based on the normal, t or F distributions * Standard Error. Surprisingly, there is a great deal of confusion among practicing scientists about the difference between the standard deviation and the standard error. Part of that confusion is that there are many standard errors, the most common being 'the standard error of the mean,' which is the conventional {standard deviation of the sample}/√n. As **Larsen & Marx (2001, Section 4.3, pp. 263-283)** discuss and demonstrate with Example 4.3.2, the underlying population distribution may take a variety of forms — including uniform, Poisson, binomial, and geometric — but as sample size increases, test statistics such as the sample mean and the difference in means between samples will tend towards the normal distribution. **Ramsey & Schafer (2002)** present a nice graphic (shown at right) showing this

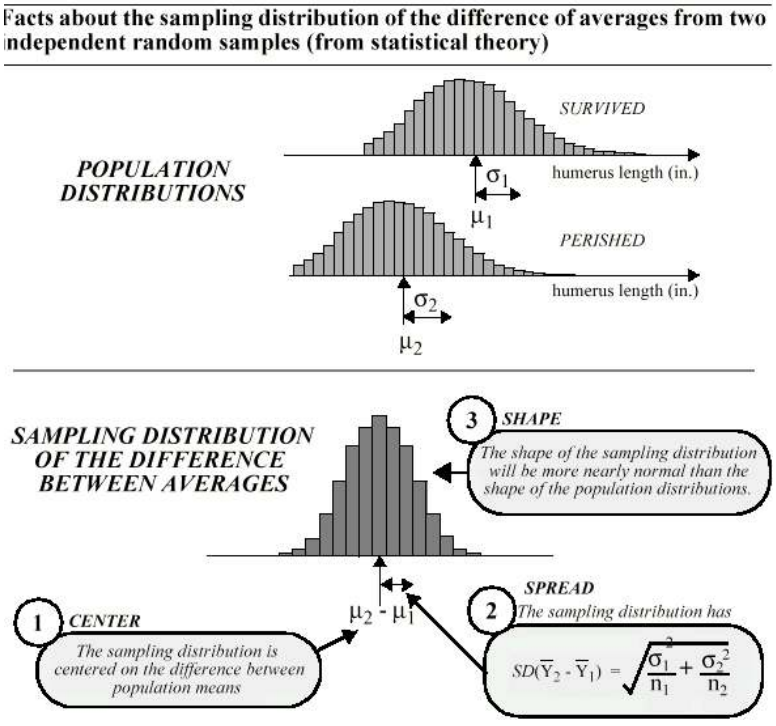

Figure 20. Demonstration of the effect of the central limit theorem on the distribution of differences in means from Ramsey & Schafer's Statistical Sleuth

effect for a 2-sample *t* test for the difference in means between a pair of populations with positive skew. The underlying distributions are non-normal, but the difference in means has a distribution that is more nearly normal. The standard deviation of the difference in means is also called the 'standard error' of the difference in means. It can be calculated from the standard deviations of the individual samples as shown in ②. Since the standard deviations are estimated from the samples, the appropriate multiplier for the 95% confidence interval around the observed difference in means would be the Student's *t* statistic with *d.f.* equal to $n+m$ - 2, where n and m are the sample sizes of the two samples and 2 df are lost since the mean and standard deviation is being estimated from the data.

The variance of the test statistic, such as the difference in means shown in Figure 20 will decrease, proportionate to $\sqrt{n}$. So, when reporting a test statistic, such as the observed difference in means, the appropriate measure of precision is either the standard error or the 95% confidence interval. In scientific meetings, it is not unusual to see the estimates of precision, or error, represented as ± 1 standard error. Indeed, that is how the MA Dept of Education reports the variability on MCAS scores. However, this is not sufficient to judge the precision of the estimate. If the variance is known from theory, one can immediately recognize that the standard error must be multiplied by the *z* statistic associated with the 97.5th percentile of the standard normal distribution, or 1.96, to obtain 95% confidence intervals [Matlab's norminv(0.975)]. However, what if the standard error is plotted, and the presenter does not present the sample size? I've often seen error bars on plots based on means with just 3 replicates. The appropriate multiplier is then not 1.96, but $t_{0.975,\ 2\ df}$ or 4.3! That could make a huge difference for a member of the audience assessing whether the results presented in a graph are different from what could be expected by chance alone. The following figure shows the effect of sample size, *d.f.*, on Student's t statistic

An investigator showing error bars for means based on 2 replicates, with no indication that n=2, may be trying to deceive his audience, since those error bars would have to use a Student's t statistic of 12.7 to convert to 95% confidence intervals.

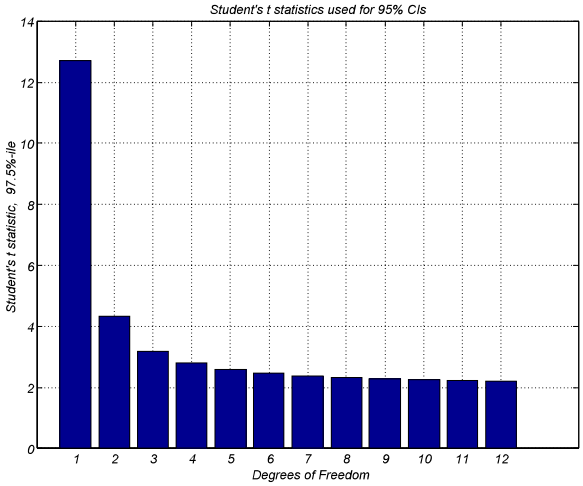

Figure 21. Effects of d.f. on the magnitude of Student's t statistic used to construct 95% confidence intervals. The Matlab program to calculate this (without labels), using tinv is: df=1:12;alpha=.05;p=1-alpha/2;fyt = tinv(p,df);bar(df,fyt);grid

Interpreting confidence intervals

Ramsey & Schafer (2002) provide a superb graphic, Figure 22 below, showing how to interpret the error bars, presented as 95% confidence intervals in presentations.

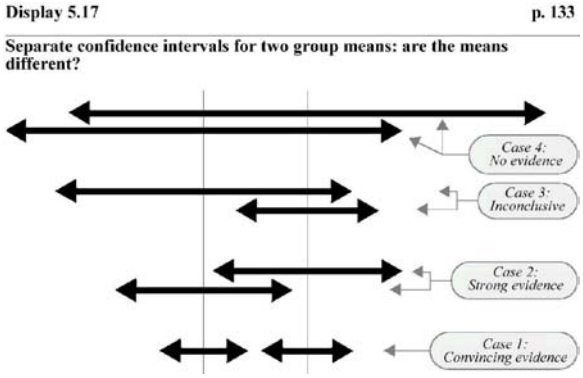

Figure 22. A rough guide to interpreting overlap in 95% confidence intervals, using the guidance that a p-value less than 0.05 is regarded as moderate to strong evidence of a difference. Note especially in Case 2 that the confidence limits can overlap and still produce a difference in sample statistics with p values less than 0.05. This interpretation of 'error bars' is not possible with standard errors if the sample sizes are not provided (and even with sample sizes provided, one would have to have a good memory of Student's t statistics for n<6-10.) [This is display 5.19 in the 2nd edition of Sleuth]

Confidence intervals, Standard Errors and significant figures for reporting results

Bevington & Robinson (1992) and **Taylor (1997)** are the two best 'how-to' guides on how to propagate errors and report errors in publications. Of the two, I prefer the more rigorous Bevington & Robinson, especially because it stresses Monte Carlo simulation, but Taylor may have the best cover photo of any recent book in statistics (see Figure 23). Neither book is in accord with current editorial practice in ecological, psychological or medication journals, because these authors recommend reporting standard errors, rather than 95% confidence intervals, in reporting results.

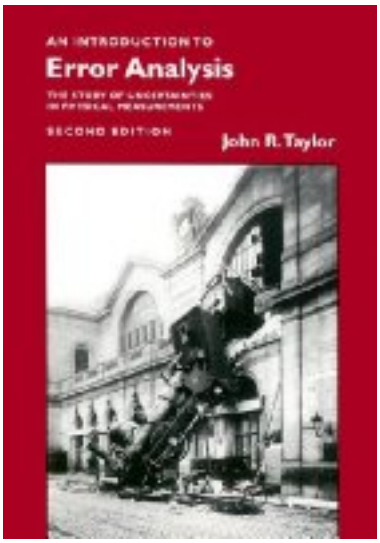

Figure 23. **Taylor (1997)** on error analysis. http://images.amazon.com/images/P/093570275X.01._SX140_SY225_SCLZZZZZZZ_.jpg

Both books argue that investigators should let the standard error (or confidence interval) dictate the level of precision reported for results. It is a sign of ignorance to report a result as 7.51478 ± 0.5672. Note, that in reporting results that you should always note whether that estimate of error to the right of the ± is a standard deviation (a poor choice), a standard error (the most common in the physical sciences) or a half-95% confidence interval. This result could be reported as 7.5 ± 0.6. Bevington & Robinson argue that standard errors should be reported to 1 significant figure only, but I'll argue shortly that there are reasons for reporting 2 significant figures. As a follow-up recommendation, there are reasons for reporting 1 additional significant figure than can be justified by the data.

One of the most persuasive arguments for retaining 1 more significant figure than warranted by the data is revealed by the answer to the following question, "What is the normal human body temperature?" Of course, the answer is 98.6ºF, but that answer is wrong. John Alan Paulos, in his book, "A mathematician reads the newspaper" recounts the story of Wunderlich who carried out the most comprehensive survey of human temperatures. They are highly variable. I wish I could report his actual results, but I have been unable to find an appropriate citation, and Paulos doesn't provide citations for his story. Paulos reports that the actual data were reported to just two significant figures, say 37ºC ± 0.8. To Wunderlich, there was no reason to add a 3rd significant figure to his data. However, when people converted his 37ºC to Fahrenheit, they produced the now highly precise 98.6º F. According to Paulos, a report that I haven't been able to confirm, the best estimate of normal human temperature is 98.2º F. The conversion to Fahrenheit has given normal human temperature a diagnosis of a mild fever. Had Wunderlich presented that single extra, non-significant at α=0.05, digit, the conversion to Fahrenheit would have been more accurate.

Theorem 5.3.2 (p. 373) solves for 'How Many Monte Carlo Simulations should you run?'

In running a Monte Carlo simulation, you face a choice about how many simulations to run. In the past, I'd used a general rule of thumb that the number of simulations should be about 10 times the inverse of the significance level of the test. As we'll see **Manly (1991, p. 80-84)** reviewed previous studies and presented a rough rule of thumb that the number of Monte Carlo simulations should be fifty times the inverse of the significance value (α) of 0.05 and five times the inverse of the significance value (α) of 0.001. Using Theorem 5.3.2 from Larsen & Marx (2006, p. 373), the number of simulations should be about four times the inverse of the significance value.

> **Theorem 5.3.2** (**Larsen & Marx, 2001, p. 331**, **2006, p. 373**) Let $\frac{X}{n}$ be the estimator for the parameter p in a binomial distribution. In order for $\frac{X}{n}$ to have at least a 100(1-α)% probability of being within a distance d of *p*, the sample size should be no smaller than
>
> $$n = \frac{z_{\alpha/2}^2}{4d^2}$$
>
> where $z_{\alpha/2}$ is the value for which P($Z \geq$) = α /2.

This theorem can be applied to solving the problem of how many Monte Carlo simulations are required to achieve a given *p* value. If you want to run a simulation that would distinguish between p=0.001 and p=0.002, you could set *d* to 0.001 in the equation and solve for n. I wrote a simple function m.file that solves this equation, called LMTheorem050302.m

```
function n=LMtheorem050302(alpha,d,p)
% format n=LMtheorem050302(alpha,d,p)
% How many samples required to achieve a
% margin of error of d for a parameter
% p in a binomial distribution
% input alpha, e.g., alpha=0.05 for
%            Margin of error=1/2 95%CI
%       d    margin of error
%       p    optional, binomial proportion
%            if not provided, p=0.5 used
%            so that d is the maximum.
% output n    number of samples required

Psn=(1-alpha/2); % Find the p value for the cumulative standard normal distribution
            % For alpha=0.05, find the z value at the 0.975 percentile of the cumulative
            % standard normal distribution, or 1.96
if nargin<3
   p=0.5;
   n=((erfinv(2*Psn-1) .* sqrt(2)).^2)./(4*d.^2);
else
   n=((erfinv(2*Psn-1) .* sqrt(2)).^2).*p.*(1-p)./d.^2; % see (5.3.4)
end
```

I wrote the m.file so that full vectors could be used for the d values. By using the following two lines, you can find the *n* for 50 values of d from 10^{-6} to 0.01

```
d=logspace(-6,-1)';% Create 50 logarithmically spaced elements from 10^-6 to 10^-1
n=LMtheorem050302(0.05,d);
```

These are plotted using loglog(d,n). I've programmed this in NforMCtrials.m The result is shown in Figure 24.

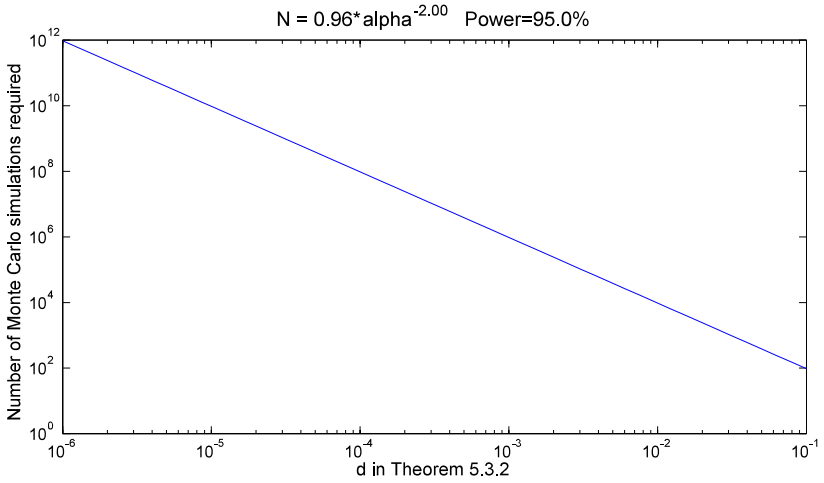

Figure 24. Graphical display of Theorem 5.3.2 for α=0.05. The number of samples required is inversely proportional to the square root of d.

Using this application of Theorem 5.3.2, if one wishes to run sufficient Monte Carlo simulations to distinguish between 0.001 and 0.002 (d=0.001), then one should use $n=0.96*0.001^{-2}$, or 960,000 Monte Carlo trials. But, this is far too conservative, because it is based on a worst-case variance when the binomial proportion is 0.5.

If you look at the proof of Theorem 5.3.2, especially equation (5.3.4), you'll see that the theorem provides a solution for equation 5.3.4 for p=0.5, the binomial proportion with the highest variance:

$$n = \frac{z_{\alpha/2}^2 \, p(1-p)}{d^2}. \qquad (5.3.4)$$

In using Monte Carlo simulations, the appropriate *p* value is the desired alpha level, not p=0.05. In running a Monte Carlo simulation, I record how many random realizations of the underlying probability model produce a result that is equal to or more extreme than that observed. It is a Bernoulli trial process, suitably modeled with the binomial probability distribution. For distinguishing between 0.001 and 0.002, the appropriate p value is 0.001, and more importantly, the variance is 0.001*0.999, 250 times smaller than 0.5*0.5. I wrote Theorem050302.m, so that

it would use equation 5.3.4, but it could also use a third input argument as the probability to insert in equation 5.3.4. To choose the appropriate n for Monte Carlo simulation to estimate p values for a null hypothesis, the call to LMTheorem050302 should be:

> n=LMtheorem050302(0.05,p,p)

where p is the desired significance value. This makes a large difference in the number of Monte Carlo simulations required, as shown in Figure 25.

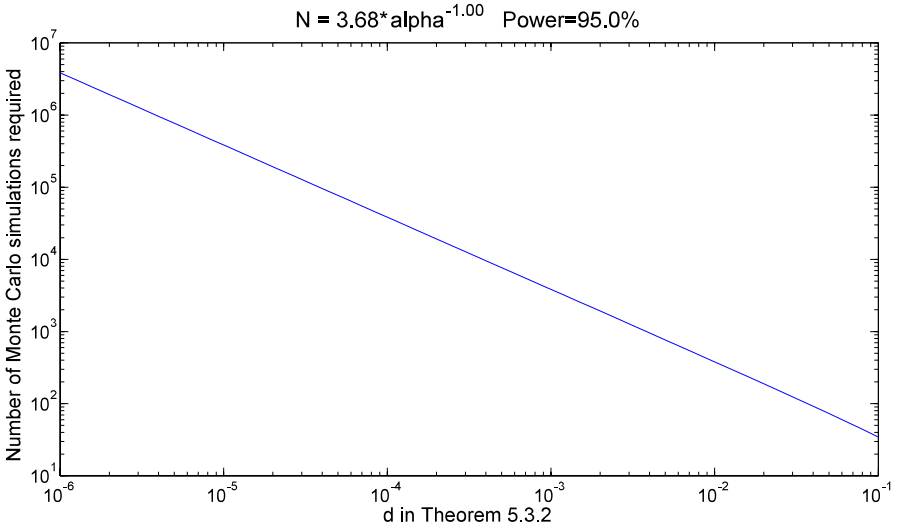

Figure 25. Graphical display of Theorem 5.3.2 for α=0.05.

Now, the number of Monte Carlo simulations is about 4*alpha$^{-1.0}$. For α=0.05 and d=0.001, instead of 960,000 Monte Carlo trials, this analysis indicates that only 3,838 Monte Carlo simulations are required.

So, I conclude that the number of Monte Carlo simulations at about 4*1/(desired precision of the p value)

What do other authors say?

Bevington & Robinson (1992, p. 92) note that the relative error in a result calculated by the Monte Carlo method is inversely proportional to the square root of the number of successful

events generated. This conclusion matches the result shown in Figure 1, but it is overly conservative. If the goal is to attain a given precision, e.g., to be able to distinguish between 0.002 and 0.003 with a fixed probability of Type I error, then the number of Monte Carlo trials increases as a linear function of 1/α, as shown in Figure 25, and not a quadratic function of 1/α as shown in Figure 24 and discussed by **Bevington & Robinson (1992, p. 92)**. The difference in expected sample sizes can be tremendous as the example discussed after Figure 26 indicates (3,838 vs. 960,000).

Manly (1991, p. 80-84) summarizes a number of studies and offers this simple recommendation, "It seems therefore that 1000 randomizations is a reasonable minimum for a test at the 5% level of significance, while 5000 is a reasonable minimum for a test at the 1% level." His rule of thumb is 5 to 50 times the inverse of the desired alpha level, which is considerably more conservative than the approach above, based on **Larsen & Marx's (2006, p. 373)** Theorem 5.3.2

Can we test our estimate of the required N?

Bevington & Robinson (1992) and **Nahin (2000)** show how Monte Carlo simulation can be used to estimate the value of pi. Figure 26 diagrams how π can be determined by placing x and y coordinates from a uniform random number distribution and calculating π = 4*successes/Trials, where success is a point 1 unit or less from the origin. This Monte Carlo simulation can be regarded as a Bernoulli trial with p= π/4. How many Monte Carlo trials would we have to run to get a simulation accurate to 10^{-5}?

```
>> n=LMtheorem050302(alpha,(1e-5)/4,pi/4)
n =
  2.9199e+011
```

If every Monte Carlo trial took 1 millisecond, this Monte Carlo simulation would take 9 years. Obviously, there are more efficient ways to estimate π.

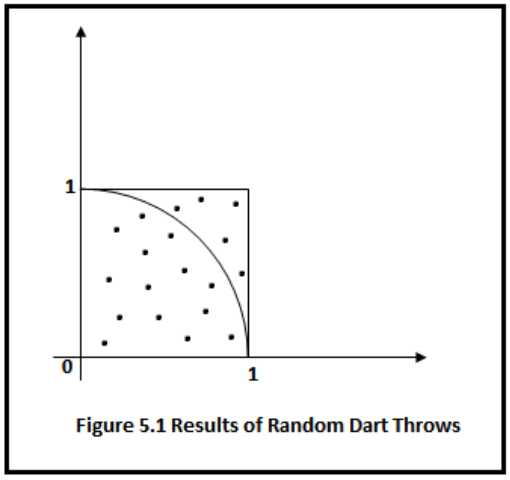

Figure 26. Figure 5.1 showing how the value of pi can be found by placing darts on a 1 x 1 grid and counting the number of darts within 1 unit of the origin. pi=4*successes/trials.

We can use this Nahin's π simulation to double check our calculations. By reducing the d value in Theorem 5.3.2, we can find a reasonable number of trials to test our equations (and to reinforce the idea of what a *p* value means)

```
>> n=LMtheorem050302(0.05,0.1/4,pi/4)
n =
  1.0360e+003
```

This indicates that if we ran Nahin's pi simulation 1036 times, we should expect to see about 5% of the Monte Carlo simulations producing estimates of π differing from the true value of π by 0.1 or more (note that due to the way Nahin set up the problem, we must divide this d=0.1 by 4).

We can apply Theorem 5.3.2 once more to find out how many times we have to run Nahin's π simulation (with a sample size of **1036**) to get a p value differing from the expected value by no more than 0.005:

```
>> n=LMtheorem050302(0.05,.005,0.05)
n =
  7.2988e+003
```

I programmed this with **7300** Monte Carlo trials and called the simulation edgpisim.m I made one or two modifications of Nahin's pisim.m

It was reassuring that edgpisim.m with **1036** and **7300** for the Monte Carlo sample sizes produces p values very close to 0.05. Each simulation takes only about 2 seconds on my Pentium 4.

```
%edgpisim.m created by PJNahin for "Duelling Idiots"(10/22/98)
%This m-file estimates pi by randomly tossing darts at the
%unit square and counting how many land inside a quarter-
%circle with unit radius contained in the square.
%
%
rand('state',100*sum(clock))  %set new seed for generator;
                        %initialize number of darts
                        %inside quarter-circle region;
Trials=7300;
kount=0;
for j=1:Trials
   x=rand(2,1036);
   x=x.^2;
   darts=sum(sum(x)<1);
   pi_estimate=4*darts/1036;
   if abs(pi_estimate-pi)>=0.1
      kount=kount+1;
   end
end
p=kount/Trials
disp('Applying Theorem 5.3.2, this estimate has a 95% probability of being within 0.005 of 0.05')
```

Some final thoughts

Monte Carlo simulations, if properly designed, produce realizations of the underlying probability model that is being tested with both nonparametric and parametric statistics. The *p*-value, which is the end result of many of the standard tests, is providing an estimate of what the results of a Monte Carlo trial would have been. If you encounter major differences between the p values from a parametric test and the results of a Monte Carlo simulation, my recommendation is to trust the Monte Carlo simulation (after you've checked your program for errors of course).

In my area of ecology, I became familiar with the utility of Monte Carlo simulations when I was analyzing a paper on spatial patterns in the deep sea benthos, the animals that live in mud and sand. Two noted benthic ecologists published a paper in one of the premier oceanography

journals describing strong spatial pattern in the deep sea. They sampled the deep sea with a coring device that took 64 samples of mud simultaneously in an 8 x 8 array. They found single individuals of a species of crustacean in 3 adjacent cores in the 64-core array and stated that the probability of finding that pattern or one more significant is **P < 0.00002** using **Cliff & Ord's [1973]** significance tests for the Moran's I spatial autocorrelation statistic, a spatially weighted version of Pearson's correlation coefficient:

$$Moran's\ I = \frac{n\sum\sum w_{ij}\,(x_i - \bar{x})\,(x_j - \bar{x})}{S_o \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

$$S_o = \sum\sum w_{ij}.$$

$$w_{ij} = weight\ for\ locality\ pair\ (i, j),\ usually\ 1/(distance)^2$$

$$\bar{x} = mean\ of\ x_i's.$$

$$\sum\sum = \sum_{i=1}^{n}\sum_{j=1}^{n}\ for\ i \neq j.$$

0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 **1**
0 0 0 0 0 0 0 **1**
0 0 0 0 0 0 0 **1**
0 0 0 0 0 0 0 0

The problem: Is the significance level of P<.00002, using **Cliff & Ord's (1973)** parametric statistical test for Moran's I correct?

There are 64 choose 3 combinations (64!/(61!3!)=41664) possible with 3 cores (See **Larsen & Marx, 2006, p. 107** for definition of combinations) , each containing a single individual of this species, placed in a 64-core array. 88 of these combinations produce patterns with Moran's I statistics greater than the one observed. The exact probability for this pattern 88/41664 = 11/5208 ≈ .00211 This is over **100** times higher the value calculated using Cliff & Ord's normal approximation.

A Monte Carlo simulation of the data, performed by randomly permuting the rows and columns of the 8x8 matrix with a calculation of the Moran's I spatial statistic after each Monte Carlo trial, produced results that rapidly converged on the p value of 0.002, not on the erroneous parametric value of 0.00002.

I went back and read Cliff & Ord's key papers describing how they had developed their significance tests for Moran's I spatial autocorrelation statistic. They'd developed their

equations using Monte Carlo simulations. But, they usually used only 300 Monte Carlo trials. Their *p* values were judged adequate only to p=0.01 (consistent with the calculations above). They were forthright that their equations were just approximations to the underlying Monte Carlo process and concluded,*"We do not recommend these approximations when...binary data are used except when the data are nearly symmetric and the lattice size is fairly large (n >50)."* The above data, consisting of 61 0's and 3 1's are certainly binary and asymmetric. The p value of 0.00002 should have never been published.

Unfortunately, Monte Carlo simulations can not be used to test many of the hypotheses that you'll encounter. Parametric statistics, especially regression & ANOVA, are the only readily available methods for analyzing many types of somewhat complex data. However, when the two approaches can be compared, parametric statistics usually produce results strikingly similar to that produced by Monte Carlo simulation. In my experience, when there is a major discrepancy between the parametric result and the Monte Carlo simulation, it can be traced to an underlying assumption of the parametric test not being met.

Example 5.3.4

How large should a sample be to have a 98% probability of being within the true proportion, p. We'll use my Matlab m.file Theorem050302_4th.m
n=LMtheorem050302_4th(0.02,.05)
This produces the answer that the sample would have to be 543.
If we knew that no more than 20% of the children are inadequately immunized, we can enter an additional proportion in the function:
n=LMtheorem050302_4th(0.02,.05,0.2)
This formula reduces the variance of the expected proportion because the variance of p=0.2 is much less than p=0.5. The answer is that only 348 samples are required.

Interpreting the significance of polls

In a March 2, 2011 NBC/Wall Street Journal poll, pollsters asked 282 likely republican primary voters who they favored for the 2012 Republican nomination for president. The results were Huckabee 25%, Romney 21%, Gingrich 13% Palin 12%, Ron Paul 6% and Tim Pawlenty 3%. I wrote a program pollsig.m to analyze polls such as this. It is called by
[ME,D,Dprob,halfCI]=pollsig(282,[.25 .21 .13 .12 .6 .3], 1e4,1)
Pollsig reports the margin of error as 5.9% and then analyzes the difference between the two leading candidates. The difference in proportions between the two leading candidates was 4%. Based on a Monte Carlo simulation, the probability of observing a 4% difference by chance is 0.31. The half 95% CI for the difference of 4% is 4 ± 7.8%

Pollsig.m takes into account that polling results are not independent. Candidate A's favorability is negatively correlated with the favorability of the other candidates. This negative covariance will produce a margin of error for the difference that is always between one and two times the margin of error for the poll.

A March 4, 2011 Gallup poll of 550 Fox News viewers had Huckabee at 18%, Romney at 17% and Palin & Gingrich tied at 13%. These data would be entered into pollsig.m
[ME,D,Dprob,halfCI]=pollsig(550,[.18 .17 .13 .13], 1e4,1)

The results are that the margin of error for a poll of size 550 is 4.2% (rounded up to 5% in the Politco news article. A Monte Carlo simulation based on 10000 trials showed that under the null hypothesis of equal proportions, and 10000 trials, the 2-sided P of observing a 1.00% difference by chance = 0.693. Lower 95% confidence limit, median, and upper 95% confidence limit based on 10000 trials. The difference between Huckabee and Romney should be reported as 1% ± 5%.

Section 5.9 in Larsen & Marx (2006) is simply wrong. They argue that in a poll with a margin of error of 5% that candidate A with a 52% to 48% lead over candidate B is a near certainty to win. This is false. First, we can use Theorem 5.3.2 to find out how many individuals would have to be polled to produce a 5% margin of error. I've programmed that theorem;
n=LMtheorem050302_4th(.05,.05)
The answer is 384.1459. We now use pollsig to calculate the probability that candidate A is even ahead of candidate B given the uncertainty in the polls.
pollsig(384,[0.52 0.48],1e5,1)
This produces that the probability of observing a difference of 4% or more by chance alone is 44.5%. I would not call a p value of 0.445 as a near certainty that the two candidates are different. This estimate was based on drawing 100,000 polls of size 384 from a population in which the candidates were equal. In 44.5% of those polls one candidate or the other had a 4% lead. One can analyze the differences and calculate the 95% confidence limit for the difference. It is 4% ± 9.9%. This result would have been different if there was another candidate in the race drawing 20% of the vote, leaving the top two candidates with 42% and 38%. The difference would be 4% ± 9% and the two-sided p value of observing a difference of 4% would be 37.5%. Imagine yet another candidate drawing an additional 20% of the vote so that the top 4 vote getters garnered [.32 .28 .2 .2]. The two-sided p value for the null hypothesis of candidate A and B being tied is now 31%, with the 4% difference having a margin of error of 7.7%. Larsen & Marx (2006) state, "Here, a 4% lead for Candidate A in a poll that has a 5% margin of error is not a "tie" — quite the contrary, it would more properly be interpreted as almost a guarantee that Candidate A will win." I wouldn't consider a p value of 0.435 as a p value upon sufficient to guarantee a winner.

```
function [ME,D,Dprob,halfCI]=pollsig(N,V,Trials,details)
% How significant are differences in poll results?
% format [ME,D,Dprob,halfCI]=pollsig(N,V,Trials,details);
% Input:
%  Required:
%    N = Number of individuals polled
%    V=  Column or row vector with proportion of responses,
%        need not sum to 1, but sum(V)<1;
%        In this implementation only the two most common items will be
%        tested.
%  Optional:
%    Trials=Number of Monte Carlo trials used to judge significance
%      if Trials not specified, 1e4 trials will be used
```

```
%       if Trials=0, then the covariance formula will be used to judge
%                   significance
%    details=0, suppress all output within the m.file.
% Output:
%   ME=Margin of Error for the poll
%   D=difference in proportions, length(V) x length(V) symmetric matrix.
%   Dprob=two-sided p value for test of equality of proportions of the 2 %
%         most common items
%         Dprob will have a minimum value of 1/Trials
%   halfCI=half the 95% CI for difference in proportions of the two most
%   common items.
%  Reference: Larsen & Marx (2006) Introduction to Mathematical Statistics
%  4th Edition p. 372

% Written in 2003 for EEOS601 by Eugene.Gallagher@umb.edu
% Dept of Environmental, Earth & Ocean Sciences.

if nargin<4
   details=1;
   MC=1;
   if nargin<3
      Trials=1e4;
   elseif Trials==0
      MC=0;
   end
end

% Calculate Margin of Error, p. 372, Larsen & Marx (2006) 4th edition
ME=norminv(0.975)/(2*sqrt(N));

if details
   fprintf('The margin of error for a poll of size %d is %3.1f%%.\n',...
      N,ME*100);
end

% Monte Carlo simulation

if details;
   fprintf('\nMonte Carlo simulation based on %d trials:\n',Trials);
end
V=V(:); % Change V to a column vector
V=flipud(sort(V));V=V(1:2);  % This m.file will only calculate significance
                        % for top two categories.
tallys=zeros(Trials,2); % A column vector with rows=Trials & 2 columns;
                    % tallys differences
```

```
tallyHo=zeros(Trials,2);% This will store the results for testing Ho:
                        % p1=p2;
ExpP=mean(V);
for i=1:Trials
   poll=rand(N,1);     % Creates a vector with uniformly distributed
                       % random numbers on the interval 0,1
   tallys(i,1)=sum(poll<=V(1));
   tallys(i,2)=sum( (poll>V(1)) & (poll <= (V(1)+ V(2))) );
   tallyHo(i,1)=sum(poll<=ExpP);
   tallyHo(i,2)=sum( (poll>ExpP) & (poll <= 2*ExpP));
end
DifferenceHo = (tallyHo(:,1) - tallyHo(:,2))/N; % Calculate the differences
% for all Trials under Ho: p1=p2
D=abs(V(1)-V(2));
Dprob=max([1 sum(abs(DifferenceHo)>=D)])/Trials;
if details & Dprob<0.001  % change the format so that it is in full form
                          % only for low p values:
   fprintf(...
  'Under the null hypothesis of equal proportions and %d trials,\n',Trials)
   fprintf(...
 'the 2-sided prob. of observing a %5.3f%% difference by chance = %d\n',...
       D*100,Dprob);
elseif details
   fprintf(...
'Under the null hypothesis of equal proportions, and %d trials,\n',...
Trials)
   fprintf(...
'the 2-sided P of observing a %4.2f%% difference by chance = %5.3f\n',...
       D*100,Dprob);
end
Diff = (tallys(:,1) - tallys(:,2))/N;
% 95% CI via Monte Carlo simulation
sortedDiff=sort(Diff);
lMC95CIpi=floor(0.025*Trials); % find the index for the lower 95% cutoff
uMC95CIpi=ceil(0.975*Trials);  % find the index for the upper 95% cutoff
medpi=round(0.5*Trials);       % find the median, should be close to or
                               % identical to the expected value.
% Save the three outputs in the row vector DLowExpUp
DLowExpUp=[sortedDiff(lMC95CIpi) sortedDiff(medpi) sortedDiff(uMC95CIpi)];
halfCI=(DLowExpUp(3)-DLowExpUp(1))/2;
if details
   fprintf(...
'\nLower 95%% confidence limit, median, and upper 95%% confidence limit based on %d
trials:\n',Trials)
   fprintf(...
```

```
'Lower 95%% CI  \tMedian \tUpper 95%% CI \n \t%4.2f%% \t\t%4.2f%% \t%4.2f%% \n',...
      DLowExpUp(1)*100,DLowExpUp(2)*100,DLowExpUp(3)*100)
   fprintf('\nDifference +/- half 95%% CI:  %4.1f%% +/- %4.1f%%\n',D*100,halfCI*100)
end
```

Annotated outline (with Matlab scripts) for Larsen & Marx Chapter 5

5 **Estimation**

Ronald Aylmer Fisher (1880-1962)

5.1 Introduction

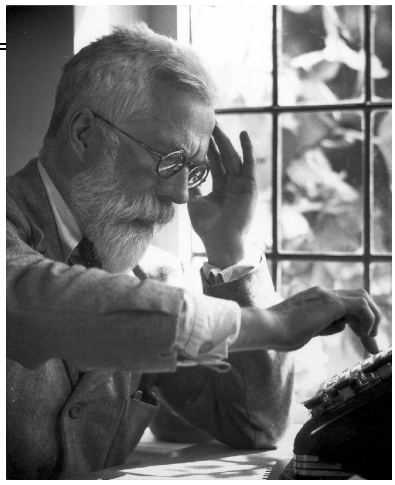

Figure 26. R. A. Fisher

Figure 5.1.1

```
% LM Fig050101_4th .m
% LM Fig050101_4th
% Plot of Poisson pdf at lambda=1 & lambda = 4
% Example of a stem plot
% Written by E. Gallagher, Eugene.Gallagher@umb.edu
% Written 10/28/10; Revised 3/3/11
%
k1=0:8;pxk1=poisspdf(k1,1);
k4=0:12;pxk4=poisspdf(k4,4);
subplot(1,2,1);
h1=stem(k1,pxk1,'Marker','none');
set(h1,'LineWidth',3);
ylabel('P_x(k)','FontSize',20)
xlabel('k','FontSize',20)
text(2.5, 0.25,'lambda=1','FontSize',18)
title('Figure 5.1.1','FontSize',22)
axis([-.5 8.5 0 0.4]);
subplot(1,2,2);
h2=stem(k4,pxk4,'Marker','none');
xlabel('k','FontSize',20)
axis([-.5 12.5 0 0.4])
text(4.5, 0.25, 'lambda = 4','FontSize',18)
set(h2,'LineWidth',3)
figure(gcf);pause
```

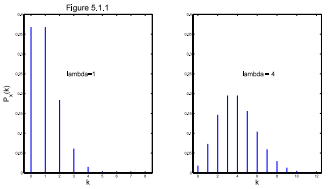

Figure 27. Figure 5.1.1 P 344

Example 5.1.1 A coin that may not necessarily be fair is handed to you and you are to determine the probability that the coin will come up heads. In three tosses, the sequence HHT is observed. As our estimate of p, choose the value that maximizes the probability of the sample. By inspection of Figure 5.1.2, we see that the value that maximizes the probability of the sample is 2/3. The value that maximizes the likelihood is the value for which the derivative is zero. In Matlab: solve(diff(p.^2*(1-p))).

5.2 ESTIMATING PARAMETERS: THE METHOD OF MAXIMUM LIKELIHOOD AND THE METHOD OF MOMENTS

5.2.1 The Method of Maximum Likelihood

Definition 5.2.1 Let $k_1, k_2, ..., k_n$ be a random sample of size *n* from the discrete pdf $p_X(k; \theta)$ where θ is an unknown parameter. The ***likelihood function***, *L(θ)*, is the product of the pdf evaluated at the n k_i's. That is

$$L(\theta) = \prod_{i=1}^{n} p_X(k_i; \theta)$$

If $y_1, y_2, ..., y_n$ is a random sample of size *n* from a continuous pdf, fy(y;θ), where θ is an unknown parameter, the **likelihood function** is written

$$L(\theta) = \prod_{i=1}^{n} f_Y(y_i; \theta)$$

Definition 5.2.2 Let $L(\theta) = \prod_{i=1}^{n} p_X(k_i; \theta)$ and $L(\theta) = \prod_{i=1}^{n} f_Y(y_i; \theta)$ corresponding to random samples $k_1, k_2, ..., k_n$ and $y_1, y_2, ..., y_n$, drawn from the discrete pdf $p_X(k; \theta)$ and continuous pdf $f_Y(y; \theta)$, respectively, where θ is an unknown parameter. In each case let θ_e be a value of the parameter such that $L(\theta_e) \geq L(\theta)$ for all possible values of θ. Then θ_e is called a **maximum likelihood estimate** for θ.

5.2.1.1 Applying the Method of Maximum Likelihood

Example 5.2.1

```
% LMEx050201_4th.m
% An example of binomial MLE fitting & a difference in the way
% Mathworks & Larsen & Marx define the geometric distribution
% From Larsen & Marx (2006). Introduction to Mathematical Statistics,
% Fourth Edition. page 348-349
% Written by Eugene.Gallagher@umb.edu 10/28/10; revised 3/3/11
% Tom Lane on 10/29/10: "unfortunately it looks like your text and MATLAB
% use different definitions for the [geometric] distribution. Our version
% has positive probability on 0,1,2,.... Yours starts at 1. The version we
% use is the one described in "Univariate Discrete Distributions" by
% Johnson, Kotz, and Kemp. Wikipedia shows both versions.
X=[3 2 1 3]; [Phat,PCI] =mle(X,'distribution','geometric')
% Matlab gives the wrong answer, because it uses a different definition
% of the geometric distribution, defined for k=0, 1, 2, 3 ... inf
% Larsen & Marx use a form defined for positive k: k=1, 2, 3, ... inf
% Larsen and Marx define the geometric distribution as the number of
% trials before a success, so k=0 is not part of the domain, but Mathworks
% defines the geometric distribution as the number of failures
% before a success, allowing k=0 to be defined.
% The correct MLE for L & M is 4/9
```

```
% 10/29 email from Tom Lane, Mathworks, says that I should call
% mle using

[Phat,PCI] =mle(X-1,'distribution','geometric');
format rat;
fprintf('The maximum likelihood estimate for p is\n')
disp(Phat)
format
fprintf('with 95%% CIs: [%5.3f %5.3f]\n',PCI);
% Just following along the text's derivation (p. 348) of the MLE:
syms p; s=diff(5*log(1-p)+4*log(p),p)
 solve(s,p)
 % find the second derivative and plot;
 s2=diff(diff(5*log(1-p)+4*log(p),p))
 % This plot shows that the second derivative is negative for all 0<p<1
 ezplot(s2,[0 1]);figure(gcf);pause
% From Larsen & Marx (2006) p. 349 provides the MLE formula
n=length(X);Phat=n/sum(X)
```

5.2.1.2 Comment (p. 349) on the distinction between maximum likelihood estimate and estimator

A maximum likelihood estimate is a number (or refers to a number), but a **maximum likelihood estimator** is a random variable... Maximum likelihood estimators, such as $\hat{p}$, have pdfs, expected values, and variances; maximum likelihood estimates, such as p_e, have none of those statistical properties.

Example 5.2.2

```
% LMEx050202_4th
% An example of customized pdf fitting.
% From Larsen & Marx (2006). Introduction to Mathematical Statistics,
% Fourth Edition. page 350
% Dept. Environmental, Earth & Ocean Sciences
% Written by Eugene.Gallagher@umb.edu 10/29/10
%
% Fit the custom pdf using a Matlab anonymous function @
%
X=[9.2 5.6 18.4 12.1 10.7];
fyytheta=@(y,theta)(1./theta.^2.*y.*exp(-y./theta));
theta=10; % This is an initial guess;
[theta, thetaCI]= mle(X,'pdf',fyytheta,'start',theta,'lowerbound',0)
% Note that must specify the lower bound for theta is 0 as specified
% in the problem or negative pdf's are produced
% Here is a plot of the pdf for the maximum likelihood estimator.
ezplot('1./5.6.^2.*y.*exp(-y./5.6)',[0.001 100])
figure(gcf);pause

y=0.001:.01:101;
```

```
fyy=1./theta.^2.*y.*exp(-y./theta);
fX=1./theta.^2.*X.*exp(-X./theta);
plot(y,fyy,'LineWidth',2);
axis([0 51 -0.002 0.07])
ax1=gca;
hold on
h1=stem(X,fX,':k','filled');
set(h1,'LineWidth',2);
s=sprintf('Example 5.2.2, theta=%3.1f, Likelihood=%7.3g',theta,prod(fX));
title(s,'FontSize',22)
figure(gcf);pause
hold off

% Plot a different value of theta
thetan=theta*2;
y=0.001:.01:101;
fyy=1./thetan.^2.*y.*exp(-y./thetan);
X=[9.2 5.6 18.4 12.1 10.7];
fX=1./thetan.^2.*X.*exp(-X./thetan);
plot(y,fyy,'LineWidth',2);
axis([0 101 -0.002 0.04])
ax1=gca;
hold on
h1=stem(X,fX,':k','filled');
set(h1,'LineWidth',2);
s=sprintf('Example 5.2.2, theta=%3.1f, Likelihood=%7.3g',thetan,prod(fX));
title(s,'FontSize',22)
figure(gcf);pause
hold off
% Plot a different value of theta
thetan=theta/2;
y=0.001:.01:26;
fyy=1./thetan.^2.*y.*exp(-y./thetan);
X=[9.2 5.6 18.4 12.1 10.7];
fX=1./thetan.^2.*X.*exp(-X./thetan);
plot(y,fyy,'LineWidth',2);
axis([0 26 -0.002 0.14])
ax1=gca;
hold on
h1=stem(X,fX,':k','filled');
set(h1,'LineWidth',2);
s=sprintf('Example 5.2.2, theta=%3.1f, Likelihood=%7.3g',thetan,prod(fX));
title(s,'FontSize',22)
figure(gcf);pause
hold off
```

```
%--------------------------
% What would happen if an additional observation of 30 were observed?
X=[X 30];
[theta, thetaCI]= mle(X,'pdf',fyytheta,'start',theta,'lowerbound',0)
% Note that must specify the lower bound for theta is 0 as specified
% in the problem or negative pdf's are produced
y=0.001:.01:101;
fyy=1./theta.^2.*y.*exp(-y./theta);
fX=1./theta.^2.*X.*exp(-X./theta);
plot(y,fyy,'LineWidth',2);
axis([0 51 -0.002 0.07])
ax1=gca;
hold on
h1=stem(X,fX,':k','filled');
set(h1,'LineWidth',2);
s=sprintf('Example 5.2.2, theta=%3.1f, Likelihood=%7.3g',theta,prod(fX));
title(s,'FontSize',22)
figure(gcf);pause
hold off
% Plot old theta of 5.6
thetan=5.6;
fyy=1./thetan.^2.*y.*exp(-y./thetan);
fX=1./thetan.^2.*X.*exp(-X./thetan);
plot(y,fyy,'LineWidth',2);
axis([0 51 -0.002 0.07])
ax1=gca;
hold on
h1=stem(X,fX,':k','filled');
set(h1,'LineWidth',2);
s=sprintf('Example 5.2.2, theta=%3.1f, Likelihood=%7.3g',thetan,prod(fX));
title(s,'FontSize',22)
figure(gcf);pause
hold off
% Plot a different value of theta
thetan=theta*2;
y=0.001:.01:101;
fyy=1./thetan.^2.*y.*exp(-y./thetan);
fX=1./thetan.^2.*X.*exp(-X./thetan);
plot(y,fyy,'LineWidth',2);
axis([0 101 -0.002 0.04])
ax1=gca;
hold on
h1=stem(X,fX,':k','filled');
set(h1,'LineWidth',2);
s=sprintf('Example 5.2.2, theta=%3.1f, Likelihood=%7.3g',thetan,prod(fX));
```

```
title(s,'FontSize',22)
figure(gcf);pause
hold off
% Plot a different value of theta
thetan=theta/2;
y=0.001:.01:51;
fyy=1./thetan.^2.*y.*exp(-y./thetan);
fX=1./thetan.^2.*X.*exp(-X./thetan);
plot(y,fyy,'LineWidth',2);
axis([0 51 -0.002 0.12])
ax1=gca;
hold on
h1=stem(X,fX,':k','filled');
set(h1,'LineWidth',2);
s=sprintf('Example 5.2.2, theta=%3.1f, Likelihood=%7.3g',thetan,prod(fX));
title(s,'FontSize',22)
figure(gcf);pause
hold off
```

5.2.2 Using Order Statistics as Maximum Likelihood Estimates

Example 5.2.3 - not easily programmed in Matlab

Case Study 5.2.1

```
% LMcs050201_4th.m
% An example of MLE Poisson fitting
% Written by Eugene.Gallagher@umb.edu 10/29/10, revised 3/3/11
% Solution of Larsen & Marx Example 5.2.1
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th Edition
% Page 352-353
% Requires the statistics toolbox
%
X=[zeros(237,1);ones(90,1);2*ones(22,1);3*ones(7,1)];
[LAMBDAHAT, LAMBDACI] = poissfit(X)
fprintf(...
'The MLE estimate of lambda is %5.3f with 95%% CI: [%5.3f %5.3f]\n',...
LAMBDAHAT,LAMBDACI)
OF=[237 90 22 7];
k=[0:3];
EF=sum(OF)*poisspdf(k,LAMBDAHAT);
% Change the 4th category pdf for k=3, to pdf for k=3:inf;
EF(4)=EF(4)+(356-sum(EF));
% Print tabular output
s=['0 ';'1 ';'2 ';'3+'];
D=num2str([OF' EF']);
fprintf('\n                  Table 5.2.1\n')
fprintf('                      Observed  Expected\n')
fprintf('Number of Major Changes    Frequency Frequency\n')
```

```
disp([repmat(' ',4,10) s repmat(' ',4,17) D])
% Plot the data using Case Study 3.3.1 as a model
h1=bar(0:3,[OF;EF]',1,'grouped')
legend('Observed','Poisson Expectation','NorthEast','FontSize',20)
xlabel('Number of Major Changes','FontSize',20)
ylabel('Frequency','FontSize',20)
s=sprintf('Case Study 5.2.1, lambda=%6.4f',LAMBDAHAT)
title(s,'FontSize',22);figure(gcf);
```

5.2.3 **Finding MLE's when more than one parameter is unknown**

Example 5.2.4

Questions p 355-357

5.2.4 **The method of moments** [Can't skip, fitting storm data]

Introduced by Pearson. More tractable than maximum likelihood when probability model has multiple parameters.

Definition 5.2.3

Example 5.2.5. Should have but couldn't get mle solution to work.

Case Study 5.2.2

```
5 % LMcs050202
% An example of pdf fitting.
% Based on Gallagher's LMEx050202.m
% From Larsen & Marx (2006). Introduction to Mathematical Statistics,
% Fourth Edition. page 359-362
% Note that LM use this example to demonstrate method of moments
% Dept. Environmental, Earth & Ocean Sciences
% Written by Eugene.Gallagher@umb.edu 1/1/2010
%
% Fit the Gamma pdf to maximum 24-h precipitation levels for 36 storms from
% 1900-1969.
%
hold off; clf
RAIN=[31 2.82 3.98 4.02 9.5 4.5 11.4 10.71 6.31 4.95 5.64 5.51 13.4 9.72 ...
   6.47 10.16 4.21 11.6 4.75 6.85 6.25 3.42 11.8 0.8 3.69 3.1 22.22 ...
   7.43 5.0 4.58 4.46 8 3.73 3.5 6.2 0.67];
[PARMHAT,PARMCI] = gamfit(RAIN)
fprintf(...
'The MLE estimate of r is %5.3f with 95%% CI: [%5.3f %5.3f]\n',...
PARMHAT(1),PARMCI(1,1),PARMCI(2,1))
fprintf('Larsen & Marx find r=1.60, so barely within the 95%% CI\n')
fprintf('Matlab"s gamma b is Larsen & Marx"s 1/lambda\n')
fprintf(...
'The MLE estimate of lambda is %5.3f with 95%% CI: [%5.3f %5.3f]\n',...
1/PARMHAT(2),1/PARMCI(2,2),1/PARMCI(1,2))
fprintf('Larsen & Marx find lambda=0.22, so within the 95%% CI\n')
```

```
Binsize=4; % maximum of 16;
X=0:32;y=gampdf(X,PARMHAT(1),PARMHAT(2));plot(X,y,'--b');figure(gcf)
% Histogram
EDGES=0:Binsize:32;N = histc(RAIN,EDGES);
bar(EDGES,N./(sum(N)*Binsize),'histc');title('Figure 5.2.3','FontSize',20)
xlabel('Maximum 24-hour rainfall (in.)','FontSize',16);
ylabel('Density','FontSize',16);
figure(gcf);
hold on
x=0:0.1:32;
y=gampdf(x,PARMHAT(1),PARMHAT(2));
% Need to multiply the gamma pdf by binsize
plot(x,y,'--r','Linewidth',2)
figure(gcf);pause
hold off
% generate a random rainfall sequence
RandRain=gamrnd(PARMHAT(1),PARMHAT(2),36,1);
Binsize=4;
EDGES=0:Binsize:ceil(max(RandRain));N = histc(RandRain,EDGES);
bar(EDGES,N./(sum(N)*Binsize),'histc');
xlabel('Maximum 24-hour rainfall (in.)','FontSize',16);
ylabel('Density','FontSize',16);
title('Randomly generated rainfall','FontSize',20); figure(gcf);
hold on
x=0:0.1:ceil(max(RandRain));
y=gampdf(x,PARMHAT(1),PARMHAT(2));
% Need to multiply the gamma pdf by binsize
plot(x,y,'--r','Linewidth',2)
figure(gcf);pause
hold off
```

Questions 362-363

5.3 **INTERVAL ESTIMATION** p. 363

Example 5.3.1

```
% LMEx050301_4th.m
% Simulation of Confidence intervals; 50 CIs each of size n = 4 were drawn
% from the normal pdf
% From Larsen & Marx  (2006) Introduction to Mathematical Statistics,
% Fourth Edition. page 364
% Dept. Environmental, Earth & Ocean Sciences
% Written by Eugene.Gallagher@umb.edu 10/29/10, revised 1/24/11, 3/4/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Call the inverse of the cumulative normal distribution
% for a 95% CI, 0.025 of the tail should be to the right, so find
% the Z value for 97.5% of the cumulative normal distribution, or 1.96
Z=norminv(0.975);
```

```
X=[6.5 9.2 9.9 12.4];meanX=mean(X);sigma=0.8;n=length(X);
CI=[meanX-Z*sigma/sqrt(n);meanX+Z*sigma/sqrt(n)]
fprintf('The mean of X is %4.2f and the 95%% CI is [%4.2f %4.2f]\n',...
   meanX,CI(1),CI(2));
 % Plot Figure 5.3.1
mu=0;
sigma=1;
X=-3.5:0.1:3.5;
Y = normpdf(X,mu,sigma);
plot(X,Y,'--r','LineWidth',2);
axis([-3.55 3.55 0 0.41]);
title('Figure 5.3.1','FontSize',20);
ax1=gca;
xlabel('Z','FontSize',16),
ylabel('f_z(z)','FontSize',16);
ax1=gca;
set(ax1,'XTick',[-1.96 0 1.96] ,'FontSize',16,...
   'XTickLabel',{'-1.96','0','1.96'},'FontSize',18)
set(ax1,'ytick',0:.1:0.4,'FontSize',16)
hold on;
xf=-1.96:.01:1.96;yf=normpdf(xf,mu,sigma);
fill([-1.96 xf 1.96]',[0 yf 0]',[.8 .8 1])
text(1.6,0.25,'Area = 0.95','FontSize',20)
title('Figure 5.3.1','FontSize',22)
figure(gcf);pause
hold off;
% Figure 5.3.2
% Generate n random samples of size 4 with mean 10 and sigma=8
% Larsen & Marx use n=50;
MU=10;SIGMA=0.8;        % Don't change unless plot axes are changed too.
fprintf('mu =%5.3f and sigma = %5.3f\n',MU,SIGMA);
n=100;                  % Larsen & Marx used 50
samsize=4;              % Larsen & Marx used 4
R = normrnd(MU,SIGMA,samsize,n);
[r,c]=size(R);
mR=mean(R)';
CIs=[mR+Z*SIGMA/sqrt(r) mR-Z*SIGMA/sqrt(r)];
% first row of CIs is the upper and 2nd row is the lower CI for 50
% random samples of size 4;
i=find(~(CIs(:,1)>=10 & CIs(:,2)<=10));
j=find(CIs(:,1)>=10 & CIs(:,2)<=10);
fprintf('With sample size %d, of %2.0f CI"s, \n',samsize,n)
fprintf('%2.0f (%5.3f%%) didn"t contain mu=10.\n',length(i),...
   100*length(i)/n)
if ~isempty(i) & n<1000
```

```
   disp(CIs(i,:))
end
% display 5.3.2
X=[1:n]';Y=repmat(MU,n,1);
plot(X,Y,'--b');hold on
errorbar(X(i),mR(i),CIs(i,1)-mR(i),CIs(i,2)-mR(i),'-r');
errorbar(X(j),mR(j),CIs(j,1)-mR(j),CIs(j,2)-mR(j),'-g');
s=sprintf('Possible 95%% confidence intervals, sample size=%d',samsize);
xlabel(s,'FontSize',18),
ylabel('True mean','FontSize',18);title('Figure 5.3.2','FontSize',22)
if samsize < 4
   axis([0-0.02*n 1.02*n 6.9 13.1])
   set(gca,'Ytick',7:13,'FontSize',18)
elseif samsize<10
   axis([0-0.02*n 1.02*n 7.9 12.1])
   set(gca,'Ytick',8:12,'FontSize',18)
else
   axis([0-0.02*n 1.02*n 8.4 11.6])
   set(gca,'Ytick',8.5:.5:11.5,'FontSize',18)
end
figure(gcf);pause;hold off
Comment
```

Case Study 5.3.1

```
% LMcs050301_4th.m
% Confidence intervals
% From Larsen & Marx (2006). Introduction to Mathematical Statistics,
% Fourth Edition. page 367
% Dept. Environmental, Earth & Ocean Sciences
% Written by Eugene.Gallagher@umb.edu 10/29/10, revised 3/4/11
DATA=[141 148 132 138 154 142 150
   146 155 158 150 140 147 148
   144 150 149 145 149 158 143
   141 144 144 126 140 144 142
   141 140 145 135 147 146 141
   136 140 146 142 137 148 154
   137 139 143 140 131 143 141
   149 148 135 148 152 143 144
   141 143 147 146 150 132 142
   142 143 153 149 146 149 138
   142 149 142 137 134 144 146
   147 140 142 140 137 152 145];
DATA=DATA(:); % convert to a single column vector
meanD=mean(DATA);sigma=6;
CI=[meanD-norminv(0.975)*sigma/sqrt(length(DATA)) ...
   meanD+norminv(0.975)*sigma/sqrt(length(DATA))];
```

```
fprintf('The mean is %4.1f with CI: [%4.1f %4.1f]\n',meanD,CI)
% Could have been obtained using normfit.
[MUHAT,SIGMAHAT,MUCI,SIGMACI]=normfit(DATA);
fprintf('Using normfit.m:\n')
fprintf('The mean is %4.1f with CI: [%4.1f %4.1f]\n',MUHAT,MUCI)

% Plot the histogram with superimposed pdf, from LMcs050202_4th.m
Binsize=4;
EDGES=124:Binsize:164;N = histc(DATA,EDGES);
bar(EDGES,N,'histc')
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
title('Case Study 5.3.1','FontSize',22)
xlabel('Maximum Head Breadths (mm)','FontSize',20);
ylabel('Frequency','FontSize',20);
figure(gcf);
hold on
x=124:0.1:164;
y=normpdf(x,MUHAT,6)*length(DATA)*Binsize;
plot(x,y,'--r','Linewidth',3)
v=axis;
plot([132.4 132.4]',[0 v(4)]','-.b','LineWidth',3)
figure(gcf);pause
hold off
```

95% CI if σ known:

$$\left(\bar{y} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right).$$

5.3.1 Confidence interval for the binomial parameter, *p*

Theorem 5.3.1

Case Study 5.3.2

```
% LMcs050302_4th
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Case Study 5.3.2 p.370
% Confidence limits for binomial proportion, k/n
% input alpha, e.g., alpha=0.05 for 95% CI's
% output CI    [Lower CI p Upper CI]
% Written by Eugene.Gallagher@umb.edu
%
alpha=0.05;
CI=LMTheorem050301_4th(713,1517,alpha)
```

LMTheorem050301_4th

```
function CI=LMTheorem050301_4th(k,n,alpha)
% format CI=LMtheorem050301_4th(k,n,alpha)
```

```
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Theorem 5.3.1 p.369
% Confidence limits for binomial proportion, k/n
% input alpha, e.g., alpha=0.05 for 95% CI's
% output CI    [Lower CI p Upper CI]
z=norminv(1-alpha/2);
CI=zeros(1,3);
p=k/n;
CI(2)=p;
halfCI=z*sqrt(p*(1-p)/n);
CI(1)=p-halfCI;
CI(3)=p+halfCI;
```

Example 5.3.2 Median test

```
% LMEx050302_4th
% LMEx050302_4th
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Case Study 5.3.2 p.370-371
% Confidence limits for binomial proportion, k/n
% input alpha, e.g., alpha=0.05 for 95% CI's
% output CI    [Lower CI p Upper CI]
% Median test for an exponential pdf
% Other m.files using the exponential distribution
%    LMcs030401_4th
% Requires the statistics and symbolic math toolboxes
% The exponential parameter, not provided, is 1.0, based on the equation
% Written October 2010, revised 3/5/11
syms y m;
fprintf('s is the integral of the exponential function')
s=int(exp(-y),y, 0, m) % integral of exponential pdf
fprintf(...
   'Solve for the value of m such that the integral is 0.5, the median\n')
solve(s-0.5,m)
median=eval(solve(s-0.5,m));
fprintf(...
'The median of the exponential distribution with mean 1 is %7.5f\n',median)
% Generate 60 random numbers from the exponential distribution
% Larsen & Marx did 1 trial of size 60; this will do 100,000 trials
trials=1e5; % 1e6 trials produced an out of memory error
n=60;
Y=exprnd(1,n,trials);
Y=Y<median;
z=norminv(0.975);
p=sum(Y)./n; % p is a vector of length trials
CI=[sum(Y)./n-z*sqrt(p.*(1-p)./n);sum(Y)./n+z.*sqrt(p.*(1-p)./n)];
Results=~[CI(1,:)<0.5 & CI(2,:)>0.5]; % a 1 only if CI doesn't include 0.5
```

```
fprintf('Median Test: In %5.0f trials, %3.1f%% outside 95%% CI.\n',...
   trials,sum(Results)/trials*100);
```

5.3.2 Margin of Error

The margin of error is half the maximum width of a 95% confidence interval.

Definition 5.3.1 The **margin of error** associated with an estimate $\frac{k}{n}$, where k is the number of successes in *n* independent trials, is 100d%, where

$$d = \frac{1.96}{2\sqrt{n}}$$

Example 5.3.3

```
% LMex050303_4th
% Written by Eugene.Gallagher@umb.edu
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th ed.
% page 372
% Written 3/4/11
n=1002;ME=norminv(0.975)/(2*sqrt(n))*100;
fprintf('The margin of error is %3.1f%%\n',ME)
```

5.3.3 Choosing sample sizes

Theorem 5.3.2

```
function n= LMtheorem050302_4th(alpha,d,p)
% format n=LMtheorem050302_4th(alpha,d,p)
% How many samples required to achieve a
% margin of error, d, for a binomial parameter, p
% p from the binomial distribution
% input alpha, e.g., alpha=0.05 for confidence limits
%       d     Margin of error=1/2 95%CI
%       p     optional, binomial proportion
%             if not provided, p=0.5 used
%             so that d is the maximum.
% output n     number of samples required
% Based on Larsen & Marx Theorem 5.3.2
% Reference Larsen & Marx Introduction to Mathematical Statistics, 3rd ed
% Same Theorem 5.3.2 in Larsen & Marx (2006) 4th edition, p. 373
% Written by E Gallagher, 2003, Eugene.Gallagher@umb.edu
% http://www.es.umb.edu/edgwebp.htm
% revised 10/17/2010, 10/29/10
%
% Find the p value for the cumulative standard normal distribution
Psn=(1-alpha/2);
% For alpha=0.05, find the z value at the 0.975 percentile of the
% cumulative standard normal distribution, or 1.96
if nargin<3
   p=0.5;
```

```
   n=norminv(Psn)^2./(4*d.^2);
else
   n=norminv(Psn)^2.*p.*(1-p)./d.^2;  % see Equation 5.3.4., p 373, 4th ed
end
```

```
% nformctrials.m
% How many Monte Carlo samples required to achieve a 95% confidence interval
% equal to alpha?  For example, how many Monte Carlo simulations would be
% required to distinguish a p value of 0.001 from 0.002. An application of
% Theorem 5.3.2 on page 331 of
% Larsen & Marx (2001). The application of the decision rule in a Monte Carlo
% trial can be viewed as the outcome of a Bernoulli trial. For most applications
% of the Monte Carlo method, there is an observed pattern. Each Monte Carlo
% simulation is evaluated to determine whether it produces a pattern equal
% to or more extreme than the observed pattern. That probability, which is
% a Bernoulli trial is a binomial process. Theorem 5.3.2 calculates the n
% required to have at least a 100(1-alpha) probability of obtaining an
% observed p value within a distance of d of p.
% Last revised 3/4/11

d=logspace(-6,-1)';% Create 50 logarithmically spaced elements from 10^-6 to 10^-1

% Power=95%, i.e., 95% chance of distinguishing a difference of size d.
alpha=0.05;
n=LMtheorem050302_4th(alpha,d);   % Theorem 5.3.2 from Larsen & Marx, p. 331 in 3rd
edition and 373 in 4th
%                        % A result in a Monte Carlo
                        % trial can be regarded as an outcome of a Bernoulli trial
X=[ones(size(d)) log(d)];      % set up the explanatory variable ln(d) for a log-log regression;
                         % A column of all 1's first to fit the Y intercept
Y=log(n);                     % ln(number of trials) from Theorem 5.3.2
B=X\Y                         % Matlab least squares regression

loglog(d,n);xlabel('d in Theorem 5.3.2','FontSize',20);
ylabel('Number of Monte Carlo simulations required','FontSize',20);
title(sprintf('N = %4.2f*{alpha}^{%4.2f}
Power=%4.1f%%',exp(B(1)),B(2),100*(1-alpha)),'FontSize',22);
figure(gcf)
pause

% The previous analysis was for a worst-case scenario, where the p values was 0.5. It would be
the
% appropriate equation if you wanted to distinguish a p value of 0.500 from 0.501. But, in
testing
% the null hypothesis that 0.005 is different from 0.006, the above
% approach is too conservative
```

```
alph=[0.05 0.01 0.001];
for i=1:length(alph)
   alpha=alph(i);
   n=LMtheorem050302_4th(alpha,d,d);   % Theorem 5.3.2 from Larsen & Marx, p. 331. A
result in a Monte Carlo
                        % trial can be regarded as an outcome of a Bernoulli trial
   X=[ones(size(d)) log(d)];      % set up the explanatory variable ln(d) for a log-log regression;
                         % A column of all 1's first to fit the Y intercept
   Y=log(n);                  % ln(number of trials) from Theorem 5.3.2
   B=X\Y;                     % Matlab least squares regression

   loglog(d,n);xlabel('d in Theorem 5.3.2','FontSize',20);
   ylabel('Number of Monte Carlo simulations required','FontSize',20);
   title(sprintf('N = %4.2f*{alpha}^{%4.2f}
Power=%4.1f%%',exp(B(1)),B(2),100*(1-alpha)),'FontSize',22);
   figure(gcf)
   pause
end
```

Example 5.3.4

```
% LMex050304_4th
% Written by Eugene.Gallagher@umb.edu
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th ed.
% page 374
% Written 10/29/10
n=LMtheorem050302_4th(0.02,.05)
fprintf('The smallest acceptable sample size is %3.0f.\n',ceil(n))
% If p is 0.2
p=.2;n=LMtheorem050302_4th(0.02,.05,0.2)
fprintf('If p=%3.1f, the smallest acceptable sample size is %3.0f.\n',...
   p,ceil(n));
```

---

finite correction factor (p. 375) If samples are drawn without replacement and that proportion is relatively large, the variance of the proportion sampled should be reduced by the **finite correction factor** $\frac{N-n}{N-1}$

Questions p 375-379

5.4 PROPERTIES OF ESTIMATORS

Example 5.4.1 coin tossing, used as an example of precision of estimator

```
% LMex050401_4th.m
% Larsen & Marx Example 5.4.1
% Analysis of precision
% Eugene.Gallagher@umb.edu, written 10/2010, revised 3/5/11.
clear all
```

```
hold off; clf
p=0.6;n=10
P1=sum(binopdf(5:7,n,p));
n=100;
z1=(0.5-0.6)/sqrt(p*(1-p)/n);
z2=(0.7-0.6)/sqrt(p*(1-p)/n);
P2=normcdf(z2)-normcdf(z1);
fprintf('P, based on large-sample normal approximation=%5.3f.\n',P2)
disp('Exact P:')
P3=sum(binopdf(50:70,n,p));
fprintf('Exact P, based on binomial distribution=%5.3f.\n',P3)
n=10;
x=0:10;
Y=binopdf(0:10,10,p);
bar(x/10,Y,1);
axis([-.05 1.05 0 .85]);
set(get(gca,'Children'),'FaceColor',[1 1 1])
hold on
bar([.5 .6 .7],binopdf(5:7,10,p),1,'FaceColor',[0.8 1 0.8]);
figure(gcf); pause;
title('Figure 5.4.1','FontSize',22)
xlabel('Value of X/n','FontSize',20)
x=-0.05:0.01:1.05;
%y2=normpdf(x,0.6,sqrt(p*(1-p)/n));
% plot(x,y2/10,'--r','LineWidth',2)
n=100;
y3=normpdf(x,0.6,sqrt(p*(1-p)/n));
plot(x,y3/10,'--r','LineWidth',3)
xf=0.5:.001:0.7;yf=normpdf(xf,0.6,sqrt(p*(1-p)/n));
fill([0.5 xf 0.7]',[0 yf/10 0]',[0.8 0.8 1])
alpha(.5); % sets the transparency to see through.
%set(get(gca,'Children'),alpha,0.5)
figure(gcf);pause
hold off
```

5.4.1 **Unbiasedness**

Definition 5.4.1 Suppose that $Y_1, Y_2, ..., Y_n$ is a random sample from the continuous pdf $f_Y(y;\theta)$, where θ is an unknown parameter. An estimator $\hat{\theta}$ $(=h(Y_1, Y_2, ..., Y_n))$ is said to be **unbiased** (for θ) if $E(\hat{\theta}) = \theta$ for all θ. (The same concept and terminology apply if the data consist of a random sample $X_1, X_2, ..., X_n$ drawn from a discrete pdf $p_X(k; \theta)$.

Example 5.4.2 Not programmed

Example 5.4.3 Not programmed

Example 5.4.4 Geometric mean

By definition, the **geometric mean** of a set of n numbers is the nth root of their product. P. 385

Note that operationally, the geometric mean is the back-transformed arithmetic mean of log transformed random variables.

Example 5.4.5

Questions p 387-388

5.4.2 **Efficiency**

Definition 5.4.2. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators for a parameter θ. If $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$ we say that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$. Aslo the relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is $\text{Var}(\hat{\theta}_2)/\text{Var}(\hat{\theta}_1)$.

Example 5.4.6

Example 5.4.7

Case Study 5.4.1 German tanks

5.5 MINIMUM VARIANCE ESTIMATORS: THE CRAMÉR-RAO LOWER BOUND *[Summer 2011, you can skip this section]*

Theorem 5.5.1

Example 5.5.1

Definition 5.5.1 minimum-variance estimator

Definition 5.5.2

Example 5.5.2

Questions 397-398

5.6 SUFFICIENT ESTIMATORS *[Summer 2011, you can skip this section]*

5.6.1 An estimator that is sufficient

5.6.2 An estimator that is not sufficient

5.6.3 A formal definition

Definition 5.6.1

Example 5.6.1

Example 5.6.2

5.6.4 A second factorization criterion

5.6.5 Sufficiency as it relates to other properties of estimators

Questions p 405

5.7 CONSISTENCY *[Summer 2011, you can skip this section]*

Definition 5.7.1

Theorem 5.7.1 Chebyshev's inequality.

Example 5.7.2

Questons p. 409

5.8 BAYESIAN ESTIMATION (p. 410-422)

Search for the U. S. Scorpion, an example of the use of Bayes' theorem

5.8.1 **Prior Distributions and Posterior Distributions**

"In a non-Bayesian analysis (which would include all the statistical methodology in this book except the present section), unknown parameters are viewed as constants; in a Bayesian analysis, parameters are treated as random variables, meaning they have a pdf." p. 411

Example 5.8.1 Phone calls in a phone bank

Definition 5.8.1

Noninformative prior, no prior information, all values of θ equally probable

---

Example 5.8.2

An application of the beta distribution to generate a posterior distribution

```
%LMex050802_4th.m
hold off; cla;clf
theta=0:0.001:1;
Density=betapdf(theta,2,4)
plot(theta,Density,'r','Linewidth',3);
ylabel('Density','FontSize',20)
title('Figure 5.8.1a','FontSize',22)
xlabel('theta','FontSize',20)
figure(gcf);pause
Prior=betapdf(theta,4,102)
plot(theta,Prior,'r','Linewidth',3);
ylabel('Density','FontSize',20)
title('Figure 5.8.1b','FontSize',22)
xlabel('theta','FontSize',20)
axis([0 0.1 0 35])
figure(gcf)
hold on;
% A work in progress to plot the posterior pdf as a function of k
% Let's assume k=4 and n=100, what is the posterior.
k=4;n=100;
%
Posterior=factorial(n+105)./(factorial(k+3).*factorial(n-k+101)).*theta.^(k+3).*(1-theta).^(n-k+
101);
Posterior=exp(gammaln(n+106)-gammaln(k+4)-gammaln(n-k+102))*theta.^(k+3).*(1-theta).^(n
-k+101); % The factorials blow up Matlab; must solve with ln(gamma)
plot(theta,Posterior,'m','Linewidth',3);figure(gcf)
legend('Prior','Posterior')
hold off
```

Example 5.8.3
If a beta posterior is used for a binomial pdf, it can be updated readily with new binomial information to produce another beta distribution. Similarly, the gamma distribution provides a suitable prior for Poisson data.

Case Study 5.8.1 Modeling Hurricanes

```
% LMcs050801_4th.m
% Bayesian analysis of the expected number of hurricanes:
% Written by Eugene.Gallagher@umb.edu, 3/5/11
% for EEOS601, UMASS/Boston
theta=1:0.001:2.6;
% Use the gamma distribution to model the prior;
% Note that Matlab's b is 1 / {L & M's mu}
mu=50;s=88;
Prior=gampdf(theta,s,1/mu);
plot(theta,Prior,'--','LineWidth',3)
xlabel('theta, Number of Hurricanes per Year','FontSize',20)
ylabel('Density','FontSize',20);
```

```
title('Case Study 5.8.1','FontSize',22)
hold on
postmu=50+100;posts=88+164;
Posterior=gampdf(theta,posts,1/postmu);
plot(theta,Posterior,'m','LineWidth',3)
legend('Prior','Posterior')
figure(gcf);pause
hold off
```

Example 5.8.4. The marginal pdf reduces to a negative binomial distribution

---

```
Case Study 5.8.2; An example of fitting the negative binomial to data.
% LMcs050802_4th.m
DATA=[repmat(0,82,1);repmat(1,57,1);repmat(2,46,1)
    repmat(3,39,1);repmat(4,33,1);repmat(5,28,1)
    repmat(6,25,1);repmat(7,22,1);repmat(8,19,1)
    repmat(9,17,1);repmat(10,15,1);repmat(11,13,1)
    repmat(12,12,1);repmat(13,10,1);repmat(14,9,1)
    repmat(15,8,1);repmat(16,7,1);repmat(17,6,1)
    repmat(18,6,1);repmat(19,5,1);repmat(20,5,1)
    repmat(21,4,1);repmat(22,4,1);repmat(23,3,1)
    repmat(24,3,1);repmat(25,3,1);repmat(26,2,1)
    repmat(27,2,1);repmat(28,2,1);repmat(29,2,1)
    repmat(30,2,1);repmat(31,13,1)];
 [parmhat,parmci] = nbinfit(DATA,0.05)
 X=0:50;
 Total=504
 Errors = Total*[nbinpdf(X,parmhat(1),parmhat(2))]'
```

5.8.2 Bayesian estimation

How do you find $\hat{\theta}$ from the posterior distribution. Differentiate and find the mode? No.

Definition 5.8.2 Let $\hat{\theta}$ be an estimator for θ based on a statistic W. The *loss function* associated with $\hat{\theta}$ is denoted $L(\hat{\theta}, \theta)$, where $L(\hat{\theta}, \theta) \geq 0$ and $L(\theta, \theta) = 0$.

Example 5.8.5

Definition 5.8.3

5.8.3 Using the risk function to find $\hat{\theta}$

Theorem 5.8.1

Example 5.8.6

5.9 TAKING A SECOND LOOK AT STATISTICS (REVISITING THE MARGIN OF ERROR)

Imagine that there is a poll with a 5% margin of error. How many individuals would have had to have been polled?

```
n=LMtheorem050302_4th(.05,.05)
384.1459
```

Imagine that the 52% support candidate A and 48% support candidate B. Is the race a statistical tie? Use pollsig.m to check.

```
[ME,D,Dprob,halfCI]=pollsig(N,V,Trials,details);
```

The margin of error for a poll of size 384 is 5.0%. Monte Carlo simulation based on 100000 trials: Under the null hypothesis of equal proportions, and 100000 trials, the 2-sided P of observing a 4.00% difference by chance = 0.445. The lower 95% confidence limit, median, and upper 95% confidence limit based on 100000 trials: Lower 95% CI Median Upper 95% CI is [-5.73% 4.17% 14.06%. The difference is 4% ± 9.9%.

APPENDIX 5.A.1 MINITAB APPLICATIONS

References

Bevington, P. R. and D. K. Robinson. 1992. Data reduction and error analysis for the physical sciences. McGraw-Hill, Boston. 328 pp. [**23**, **27**, **28**]

Cliff, A. D. and J. K. Ord. 1973. Spatial autocorrelation. Pion, London, England. [**30**]

Hilborn, Mangel. 1997. The ecological detective. [**18**]

Larsen, R. J. and M. L. Marx. 2001. An introduction to mathematical statistics and its applications, 3rd edition. Prentice Hall, Upper Saddle River, NJ 07458. [**20**, **24**]

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. [**18**, **24**, **28**]

Manly, B. F. J. 1991. Randomization, bootstrap, and Monte Carlo methods in biology. Chapman and Hall, London. [**24**, **28**]

Nahin, P. J. 2000. Duelling idiots and other probability puzzlers. Princeton University Press, Princeton & Oxford. 269 pp. [**28**]

Taylor, J. R. 1997. An introduction to error analysis: the study of uncertainties in physical measurements, 2nd edition. University Science Books, Sausalito CA 327 pp. [**23**]

Index