WEEK 4: CHAPTER 4, SPECIAL DISTRIBUTIONS

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 Sections 4.1-4.3
  - We won't cover 4.4-4.7 on the geometric, negative binomial, and gamma distributions in this course. There are some applications of these distributions in environmental science, particularly the use of the negative binomial to model aggregated plant distributions, but there is insufficient time to cover everything. I provide the Matlab code for the examples and case studies.

Understanding by Design Templates

Understanding By Design Stage 1 — Desired Results Week 4
LM Chapter 4 Special Distributions

G Established Goals
- Learn how to solve applied problems with Matlab's pdf, cdf's and inverse functions for the Poisson, exponential and normal distributions
- Learn how to model individual behavior using the standard normal (Gaussian) curve.
- Describe the effects of sampling error and rounding on the expected values and variances of computed values from observations (An MCAS 10th grade standard)

U Understand
- Most statistical inference — frequentist & Bayesian — is based on the uniform, binomial, Poisson, & Normal distributions
- The Poisson and Normal distributions were developed as approximations for the binomial distribution but are useful themselves in modeling nature
- The Poisson model is the standard model for 'random' distributions of animals and plants in nature.
- Increased precision of measurement is one of the benchmarks of scientific advance and propagation of error is the ruler for the benchmark.

Q Essential Questions
- Why is the Poisson distribution so useful in modeling random events in nature?
- Is the normal distribution the foundation of all quantitative science?
- Is estimating the variance of a climate change prediction more important than estimating its expected value?

K *Students will know how to define (in words or equations)*
- Poisson, exponential, Monte Carlo simulation, normal and standard normal distributions, propagation of error, standard error, **z-score** and z-transform

S *Students will be able to*
- Write Matlab programs using the pdf, cdf and inverse distribution functions for the binomial, Poisson, exponential and normal distributions
- Provide examples from their own fields of the Poisson, exponential & normal distributions
- Be able to state to a layman the meaning of p values based on the distribution functions for the binomial, Poisson & normal distributions

Understanding by Design Stage II — Assessment Evidence Week 4 6/21-6/27

LM Chapter 4.1-4.3 (pages 274-316) Special Distributions

- **Post in the discussion section** by 6/29/11 W
    - Find an application from the primary literature in your own field or your own work using the Poisson or normal distributions. Hint: use Google scholar: e.g., search under 'Coral abundances Poisson'). Describe the application and include a citation or link if it is from the scientific literature
- **HW 4 Problems due Wednesday 6/29/11 W 10 PM**
    - Each problem must be solved using Matlab
        - Submit your Matlab code as part a Word, Wordperfect or rtf document summarizing the answer. I must be able to run your Matlab code after a simple cut-and-paste from your document to the Matlab editor
        - You must reach a 1-2 sentence conclusion about your result
        - Submit to the course Vista 8 website
    - **Basic problems (4 problems 10 points)**
        - **Problem 4.2.2 P. 280 Prescription errors**
            - Compare the Poisson limit approximation with the exact probability from the binomial pdf
            - Use Case Study 4.2.1 as a model
        - **Problem 4.2.12 P. 287 Midwestern Skies Lost Bags**
            - Use Case Study 4.2.2 as a model
            - Matlab hints
                - Enter the data as BagsLost
                - lambda=mean(BagsLost)
                - k=[0:max(BagsLost)+4]';
                - ObservedF=hist(BagsLost,k);
                - bar(k,ObservedF);figure(gcf)
        - **Problem 4.3.20 Nevada nuclear weapons test p 306-307**
            - Use Example 4.3.5 (Memphis earthquakes as a model)
        - **Problem 4.3.22 P 314 Children's IQ's and school cost**
            - Use Example 4.3.6 (Drunk Driving, p 308-309) as a paradigm
    - **Advanced problems (2.5 points each)**
        - **Problem 4.3.20 Nevada nuclear weapons:** Graph both the Poisson pdf and superimposed normal pdf
            - Use Example the graph in Gallagher Example 4.3.5 as a model
        - **Problem 4.3.22 p 314 Children's IQ's and school cost**
            - Plot the pdf as in Figure 4.3.6 using Matlab's fill function to fill in the areas corresponding to special needs.
            - Use Example 4.3.7 as a model
    - **Master problems (1 only, 5 points)**
        - What is the probability of observing 2 or more magnitude 5 earthquakes within 1 degree latitude & longitude of Los Angeles next year
        - **http://earthquake.usgs.gov/earthquakes/eqarchives/epic/epic_rect.php**
        - Search database 1/1/73 to 12/31/10 (38 years); 33.05 to 35.05 lat; -119.25 to -117.25 longitude

Introduction

Chapter 4 is my favorite chapter in Larsen & Marx (2006). It also covers the Poisson distribution which is the null model for random distributions in nature. The case studies and examples are interesting, including the analysis of leukemia clusters in Niles Illinois (**Case Study 4.2.1**) and the probability that you'll be eating 5 or more insect parts with your next peanut butter and jelly sandwich (**Example 4.2.4**).

Annotated outline (with Matlab scripts) for Larsen & Marx Chapter 4

4 **Special distributions**

Lambert Adolphe Jacques Quetelet (1796-1874)

4.1 **Introduction**

4.1.1 To qualify as a probability model over a sample space S, a function

4.1.1.1 must be nonnegative for all outcomes in S an

4.1.1.2 must integrate to 1.

4.1.2 Not all functions qualify as a probability model.

4.2 **The Poisson distribution**

4.2.1.1 binomial distribution

$$p_x(k) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

4.2.1.2 The Poisson limit

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{e^{-np}(np)^k}{k!}.$$

Poisson limit theorem (**Larsen & Marx 2001 Theorem 4.2.1, p. 251**) If n $\to\infty$ and p $\to 0$ in such a way that λ = np remains constant, then for any nonnegative integer k,

$$\underset{n\to\infty}{LIM} \binom{n}{k} p^k (1-p)^{n-k} = \frac{e^{-np}(np)^k}{k!}.$$

Example 4.2.1. Convergence of the Poisson, very good for n$\approx$100

```
% LMex040201_4th.m
% Example 4.2.1 in
% Larsen & Marx (2006) Introduction to
Mathematical Statistics, 4th edition
% Table 4.21 & Table 4.2.2 Larsen & Marx (2006,
p. 277)
% Written by Eugene Gallagher for EEOS601 in
2001, revised 1/9/11
% Eugene.Gallagher@umb.edu
% Comparing Poisson pdf and binomial pdf
p=1/5;
n=5;
k = [0:n]';
lambda=1;
p_pdf = poisspdf(k,lambda);
p_pdf=[p_pdf;1-sum(p_pdf)];
b_pdf= binopdf(k,n,p);
b_pdf=[b_pdf;1-sum(b_pdf)];
kdisp=[0:6]';
fprintf('Table 4.2.1\n')
fprintf('%f1.0  %5.3f   %5.3f\n',[kdisp b_pdf p_pdf]');

% Plot as a grouped histogram
   bar(kdisp,[b_pdf p_pdf],'grouped');
   axis([-.6 6.6 0 0.5])
   legend('Binomial pdf','Poisson pdf')
   xlabel('k','Fontsize',14);
   ylabel('Frequency','Fontsize',14)
   title('Table 4.2.1','Fontsize',16);
   figure(gcf);pause
p=1/100;
n=100;
ndisp=[0:9]';
k = [0:n]';
lambda=1;
p_pdf = poisspdf(ndisp,lambda);
p_pdf=[p_pdf;1-sum(p_pdf)];
b_pdf= binopdf(ndisp,n,p);
b_pdf=[b_pdf;1-sum(b_pdf)];
kdisp=[0:10]';
```

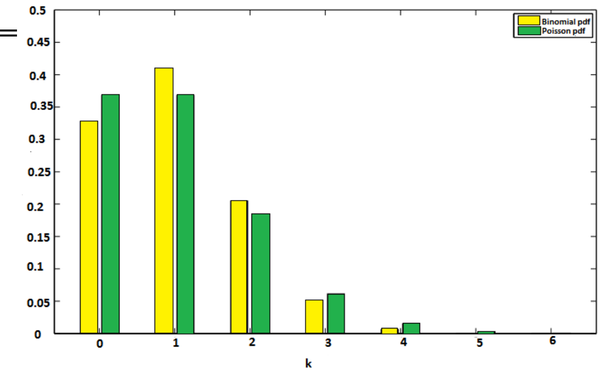


Figure 1. With n=5 (λ=1), the Poisson approximation is just adequate.

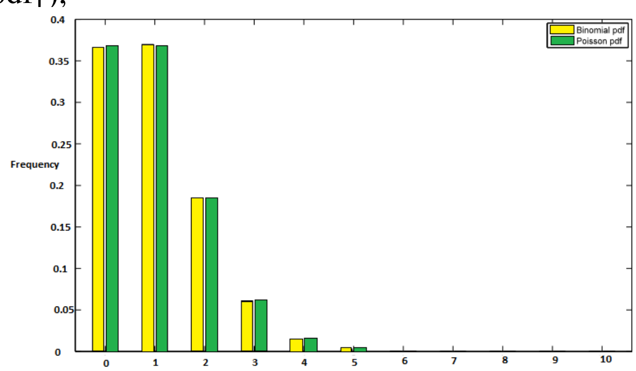


Figure 1. With n=100 (λ=1), the Poisson approximation is very good.

```
fprintf('Table 4.2.1\n')
fprintf('%f1.0  %5.3   %5.3f\n',[kdisp b_pdf p_pdf])
% Plot as a grouped histogram
   bar(kdisp,[b_pdf p_pdf],'grouped');
   legend('Binomial pdf','Poisson pdf','FontSize',16)
   axis([-.6 10.6 0 0.4])
   xlabel('k','FontSize',14);
   ylabel('Frequency','FontSize',14)
   title('Table 4.2.2','FontSize',16);
   figure(gcf);pause
```

Table 4.2.1

Table 4.2.1 Binomial Probabilities and Poisson limits; n = 5 and p = $\frac{1}{5}$ (λ=1)

k	$\binom{5}{k}(0.2)^k(0.8)^{5-k}$	$e^{-1}(1)^k/k!$
0	0.328	0.368
1	0.410	0.368
2	0.205	0.184
3	0.051	0.061
4	0.006	0.015
5	0.000	0.003
6+	0	0.001
	1.000	1.000

Table 4.2.2 Binomial Probabilities and Poisson Limits; n = 100 and p = $\frac{1}{100}$ (λ=1)

k	$\binom{100}{k}(0.01)^k(0.99)^{100-k}$	$e^{-1}(1)^k/k!$
0	0.366032	0.367879
1	0.369730	0.367879
2	0.184865	0.183940
3	0.060999	0.061313
4	0.014942	0.015328
5	0.002898	0.003066
6	0.000463	0.000511
7	0.000063	0.000073
8	0.000007	0.000009
9	0.000001	0.000001
10	0.000000	0.000000
	1.000000	0.999999

Figure 1. Table 4.2.1, Table 4.2.2

Example 4.2.2. Shadyrest hospital

```
% LMex040202_4th.m
% Example 4.2.2 Shady Rest Hospital cardiac monitoring machines, p. 277-288
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% There are 12,000 residents in the area served by Shady Rest Hospital and
% the probability that a resident will have a heart attack and need a
% monitor is 1/8000. There are 3 monitors available. What is the
% probability that the 3 monitors will be inadequate to meet the
% community's needs? That is, what is the probability of 4 or more heart
% attacks?
% Written by Eugene.Gallagher@umb.edu, Fall 2010
fprintf('Solved as the sum of binomial probabilities: %8.6f\n',...
1-sum(binopdf(0:3,12000,1/8000)))
fprintf('Or solved using the binomial cumulative distribution function: %8.6f\n',...
1-sum(binocdf(3,12000,1/8000)))
fprintf('Or solved using the Poisson distribution: %8.6f\n',...
1-sum(poisspdf(0:3,12000/8000)))
fprintf('Or solved using the cumulative Poisson distribution: %8.6f\n',...
```

```
1-sum(poisscdf(3,12000/8000)))
```

Case Study 4.2.1 Leukemia in Niles Illinois

```
% LMcs040201_4th.m
% Case Study 4.2.1 Leukemia cancer cluster. Pp 278-279 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics 4th Edition
% Written by Eugene D. Gallagher for UMASS/Boston's EEOS601
% Eugene.Gallagher@umb.edu, http://www.es.umb.edu/edgwebp.htm
% Written 10/1/2010; Revised 1/7/11
% p=incidence of leukemia in 5 1/3 years in towns around Niles IL
% What is the probability of observing 8 or more childhood leukemia cases
% by chance alone. Solve using the binomial pdf and the Poisson limit.
% N=number of people in Niles IL
N=7076;
p=24.8/1e5;
X=8:7076;X=X';
bp=binopdf(X,N,p);
BP=sum(bp); % or it could be solved using the complement
X2=0:7;
BP2=1-sum(binopdf(X2,N,p));
fprintf(...
'\nFrom the binomial pdf, the exact probability of observing \n8 or more ')
fprintf('deaths among 7076 children in 5 1/3 years is %6.3g.\n',BP);
fprintf(...
'From the complement of the binomial pdf, the exact p of observing \n8 or')
fprintf(' more deaths among 7076 children in 5 1/3 years is %6.3g.\n',BP2);
% Now solve using the Poisson pdf;
% lambda is the expected number of cases among 7076 in Niles Illionois
lambda=7076*24.8e-5;
PP= 1-sum(poisspdf(X2,lambda));
fprintf(...
'\nFrom the Poisson pdf, the approximate probability of observing \n8 or ')
fprintf('more deaths among 7076 children in 5 1/3 years with a Poisson \n')
fprintf('expected value of %6.3g is %6.3g.\n\n',lambda,PP);

% Or, one could use the Poisson cumulative distribution function:
PCP = 1-poisscdf(7,lambda);
fprintf(...
'From the cumulative Poisson pdf, the approximate probability of \n')
fprintf('observing 8 or more deaths among 7076 children in 5 1/3 years ')
fprintf('with \na Poisson expected value of %6.3g is %6.3g.\n\n',...
   lambda,PCP);
RE=abs(BP-PP)/BP * 100;
fprintf(...
'The relative error in the Poisson approximation is %4.1f%%\n',RE)
```

Questions. p 279

4.2.1.3 **The Poisson Distribution**

4.2.1.3.1 First described by Poisson as a limit theorem and then used in 1898 by Professor Ladislaus von Bortkiewicz to model the number the Prussian cavalry officers kicked to death by their horses.

Larsen & Marx 2001 Theorem 4.2.2, p. 251 The random variable X is said to have a **Poisson distribution** if

$$p_x(k) = P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}, \; k = 0,1,2,\ldots$$

4.2.1.4 **Fitting the Poisson Distribution to Data**

Case Study 4.2.2

```
% LMcs040202_4th.m
% Case Study 4.2.2 alpha decay & Geiger counter Page 282-284 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics 4th Edition
% Written by Eugene Gallagher for EEOS601
% Eugene.Gallagher@umb.edu http://www.es.umb.edu/edgwebp.htm
% Written 11/4/2010, Revised 1/7/11
% Fitting the Poisson pdf to radioactive decay: alpha-particle
k=[0:11]';
Frequency=[57 203 383 525 532 408 273 139 45 27 10 6]';
sumFrequency=sum(Frequency);
lambda=sum(k.*Frequency)./sumFrequency;
fprintf('The Poisson parameter is %5.3f\n',lambda)
ObservedP=Frequency./sumFrequency;
ExpectedP=poisspdf(k,lambda);
TableData=[k Frequency ObservedP ExpectedP];
fprintf('\n\t\t\t\t\tTable 4.2.3\n')
fprintf('No. Detected, k \tFrequency \tProportion \tp_x(k)\n')
fprintf('   %2.0f\t\t\t\t%2.0f\t\t\t%5.3f\t\t%5.3f\n',TableData');

% Using concepts from Chapter 5, a MLE fit could be made of the data:
% The only new feature added is a 95% CI for the binomial parameter.
% See LMcs050201_4th.m for another example of using Poisspdf to fit
% the Poisson distribution to data
%
DATA=[zeros(57,1);ones(203,1);2*ones(383,1);3*ones(525,1);4*ones(532,1);...
   5*ones(408,1);6*ones(273,1);7*ones(139,1);8*ones(45,1);...
   9*ones(27,1);10*ones(10,1);11*ones(6,1)];
[LAMBDAHAT, LAMBDACI] = poissfit(DATA,0.05);
fprintf(...
   '\nThe Poisson parameter =%5.3f, with 95%% CI: [%5.3f %5.3f]\n\n',...
   LAMBDAHAT, LAMBDACI)
```

Case Study 4.2.3

```
% LMcs040203_4th.m
% Fitting football fumbles to a Poisson pdf
% From Larsen & Marx (2006) Introduction to Mathematical Statistics
% 4th edition.
% Written by Eugene.Gallagher@umb.edu
% Revised 11/4/2010
% See LMcs040202_4th.m (alpha decay) for a very similar program
k=[0:7]';
Frequency=[8 24 27 20 17 10 3 1]';
sumFrequency=sum(Frequency);
meank=sum(k.*Frequency)./sumFrequency;
fprintf('The Poisson parameter is %5.3f\n',meank)
ObservedP=Frequency./sumFrequency;
ExpectedP=Poisspdf(k,meank);
TableData=[k Frequency ObservedP ExpectedP];
fprintf('\n\t\t\t\t\tTable 4.2.5\n')
fprintf('No. Detected, k \tFrequency \tProportion \tp_x(k)\n')
fprintf('   %2.0f\t\t\t\t%2.0f\t\t\t%5.3f\t\t%5.3f\n',TableData');
% Using concepts from Chapter 5, a MLE fit could be made of the data:
% The only new feature added is a 95% CI for the binomial parameter.
% See LMcs050201_4th.m for another example of using Poisspdf to fit
% the Poisson distribution to data
%
DATA=[zeros(8,1);ones(24,1);2*ones(27,1);3*ones(20,1);4*ones(17,1);...
   5*ones(10,1);6*ones(3,1);7];
[LAMBDAHAT, LAMBDACI] = poissfit(DATA,0.05);
fprintf(...
   '\nThe Poisson parameter =%5.3f, with 95%% CI: [%5.3f %5.3f]\n\n',...
   LAMBDAHAT, LAMBDACI)
```

4.2.1.5 **The Poisson Model: the law of small numbers**

4.2.1.5.1.1 Assumptions of the Poisson model

4.2.1.5.1.1.1 The probability that 2 or more events occur in any given <small> subinterval is essentially 0.

4.2.1.5.1.1.2 Events are independent

4.2.1.5.1.1.3 The probability that an event occurs during a given subinterval is constant over the entire interval from 0 to T.

4.2.1.5.2 **Calculating Poisson Probabilities**

4.2.1.5.2.1 "Calculating Poisson probabilities is an exercise in choosing T so that λT represents the expected number of occurrences in whatever "unit" is associated with the random variable X.

Example 4.2.3

```
% LMex040203_4th.m
% Example 4.2.3 Typographical Errors Page 285 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics 4th Edition
% Written by Eugene Gallagher for EEOS601
% Eugene.Gallagher@umb.edu http://www.es.umb.edu/edgwebp.htm
% Written 1/7/2011, Revised 1/7/11
% Using the Poisson pdf
% ErrorRate=typographical Errors per page
% What is the probability that fewer than three typos will appear in a 16-
% page edition.
% X=Number of errors
ErrorRate=0.4;
lambda=ErrorRate*16;
fprintf('Using the Poisson cumulative distribution function:\n')
P=poisscdf(2,lambda);
fprintf('The probability that fewer than 3 typos will appear in a \n')
fprintf(...
'16-p document with an error rate of %3.1f per page is %5.3f.\n',...
lambda,P)
fprintf('\nUsing the Poisson probability distribution function:\n')
X=0:2;
P=sum(poisspdf(X,lambda));
fprintf('If the Poisson parameter = %3.1f, the probability of \n',lambda)
fprintf('fewer than 3 typos = %5.3f.\n',P)
```

Example 4.2.4 Insect parts in peanut butter

$$P(Eating\ 5\ or\ more\ bug\ parts|\lambda = 6) = P(X \geq 5) = 1 - P(X \leq 4) = 1 - \sum_{k=0}^{4} \frac{e^{-6.0}(6.0)^k}{k!}.$$

```
% LMex040204_4th.m
% Example 4.2.4 Eating bug parts
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu, 12/12/10, revised 1/12/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% This is a simple application of the Poisson cumulative distribution
% function and Poisson pdf
% The FDA Food Detection Action Limit for peanut butter is 30 insect
% fragments per 100 grams. If your peanut butter snack weighs 20 grams,
% what is the probability that you'll ingest 5 or more insect parts.
```

```
% Calculate lambda, the expected number of insect parts per 20 grams.
lambda=30/100*20;
% solve using the Poisson cumulative distribution function.
P=1-poisscdf(4,lambda);
fprintf(...
'The probability that a 20-g snack contains 5 or more insect parts is %5.3f.\n',P)
% The problem could also be solved exactly using the Poisson pdf.
k=0:4;
P2=1-sum(poisspdf(k,lambda));
fprintf(...
'The probability that a 20-g snack contains 5 or more insect parts is %5.3f.\n',P2)
```

S
Sum of Poisson processes is a Poisson process:

$$P(x=k) = \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k}{k!}, \quad k = 0,1,2,\ldots$$

Questions Page 287
4.2.10 Prussian horse kicks.
4.2.12 Midwestern Skies Airlines bags lost

Sum of two Poisson variables is Poisson

$$P(x=k) = \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k}{k!}, \quad k = 0,1,2,\ldots$$

4.2.2 Interval between events: the Poisson/Exponential Relationship

Theorem 4.2.3 P. 290
Suppose a series of events satisfying the Poisson model are occurring at a rate of λ per unit time. Let the random variable Y denote the inteval between consecutive events. Then Y has the exponential distribution: $f_y(y) = \lambda\, e^{-\lambda y}$

126	73	3	6	37	23
73	23	2	65	94	51
26	21	6	68	16	20
6	18	6	41	40	18
41	11	12	38	77	61
26	3	38	50	91	12

To answer that question requires that the data be reduced to a density-scaled histogram and superimposed on a graph of the predicted exponential pdf. Table below details the construction of the histogram. Notice in Figure below that the shape of that histogram is entirely consistent with the theoretical model – $f_y(y) = 0.027e^{-0.027y}$ – stated in Theorem 4.2.3.

Interval (mos), y	Frequency	Density
$0 \le y \le 20$	13	0.0181
$20 \le y \le 40$	9	0.0125
$40 \le y \le 60$	5	0.0069
$60 \le y \le 80$	6	0.0083
$80 \le y \le 100$	2	0.0028
$100 \le y \le 120$	0	0.0000
$120 \le y \le 140$	1 36	0.0014

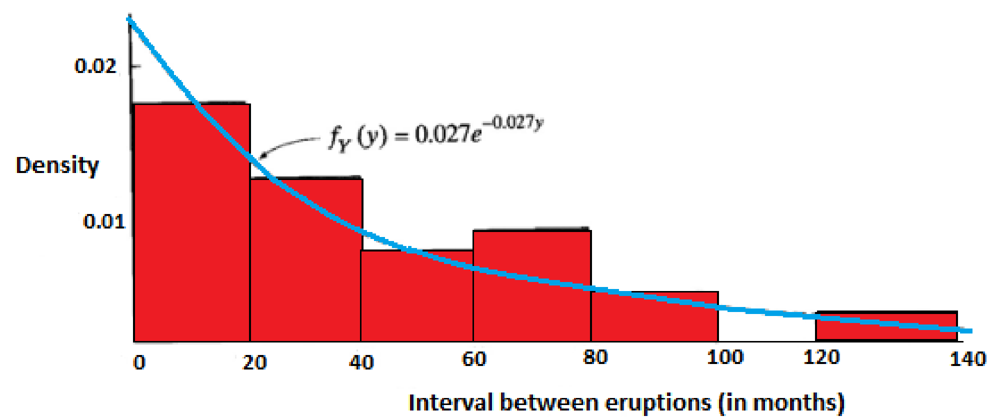


Figure 2.

Case Study 4.2.4
% LMcs040204_4th.m
% Case Study 4.2.4 Mauna Loa eruptions & exponential distribution P 290-291
% in Larsen & Marx (2006) Introduction to Mathematical Statistics 4th Ed.
% Written by Eugene Gallagher for EEOS601

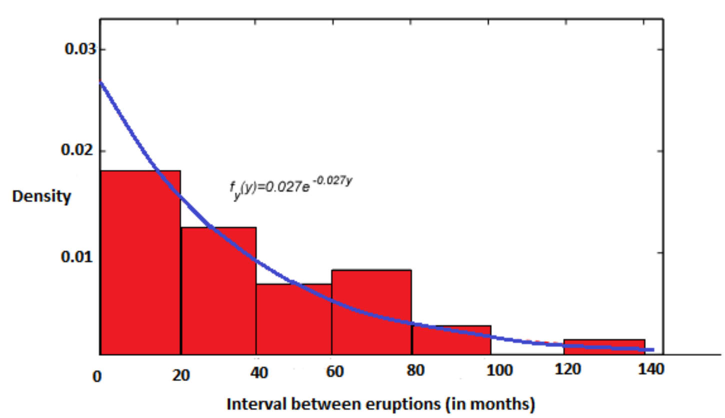


Figure 3.

```
% Eugene.Gallagher@umb.edu http://www.es.umb.edu/edgwebp.htm
% Written 1/7/11, Revised 1/7/11
% Fitting the exponential pdf to intervals between eruptons
% Data = interval between 37 consecutive eruptions of Mauna Loa
DATA=[126 73 3 6 37 23
73 23 2 65 94 51
26 21 6 68 16 20
6 18 6 41 40 18
41 11 12 38 77 61
26 3 38 50 91 12];
DATA=DATA(:);
lambda=0.027;
fprintf('Eruptions occuring at %5.3f per month or once every %3.2f ',...
   lambda,1/(12*lambda))
fprintf('years.\n')
% histc is the Matlab counting function
% Plot the data using LMFig030402_4th.m as a model
ObservedI=histc(DATA,0:20:140);
bar(0:20:140,ObservedI/(sum(ObservedI)*20),'histc');
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
axis([0 145 0 0.033]);
figure(gcf)
xlabel('Interval between eruptions (in months)')
ylabel('Density')
ax1=gca; % get the handle for the bar chart's axes;
pause
y=0:140;
fy=lambda*exp(-lambda*y);
h2=line(y,fy,'Color','r','LineStyle','--','Linewidth',2);
set(ax1,'YTick',[0:0.01:0.03])
text(34,0.015,'f_y(y)=0.027e^{-0.027y}')
title('Figure 4.2.3')
figure(gcf)
pause
```

Example 4.2.5:

```
% LMex040205_4th.m
% Example 4.2.5 page 292 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu for EEOS601
% If the Perseid meteor shower has a rate of 40 per hour, Poisson
% distributed, calculate the probability that someone would have to wait 5
% minutes before seeing another.
% y= interval in minutes between consecutive sightings
% Using the symbolic math toolbox
```

```
fprintf('Exact P=\n')
syms y; disp(int(40/60*exp(-40/60*y),5,inf))
P=eval(int(40/60*exp(-40/60*y),5,inf));
fprintf('If the Perseids have a Poisson parameter = 40/60 per min., \n')
fprintf('an observer would have p=%6.4f of waiting 5 or more \n',P)
fprintf('minutes between consecutive meteors.\n')
```

4.3 THE NORMAL DISTRIBUTION

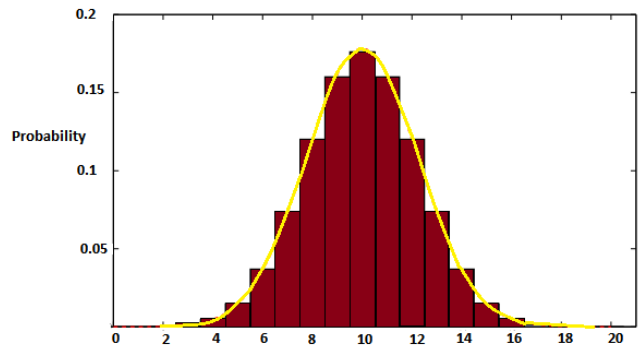


Figure 4.3.1

Figure 4.3.1

```
% LMFig040301_4th.m
% Page 293 in Larsen & Marx (2006) Introduction to Mathematical Statistics
% 4th edition
% Generate 20 binomial random numbers, p(success)=0.5
% Written in 2001 by Eugene.Gallagher@umb.edu, revised 1/7/11
% Use the statistical toolbox's binornd.m
X=0:20;n=20;p=0.5;
Y = binopdf(X,n,p);
bar(X,Y,1);axis([0 21 0 0.2]);title('Figure 4.3.1','FontSize',14);
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
ylabel('Probability','FontSize',14);
ax1=gca;set(ax1,'xtick',0:2:20,'FontSize',12);
set(ax1,'ytick',0:0.05:0.2,'FontSize',12)
figure(gcf);pause
% Superimpose the normal probability pdf:
% The normal probability equation is provided on p. 293
% This is for mean 0, and unit standard
% deviation. The more general equation (Legendre & Legendre, 1998 p. 147) is:
% f(y_j)=1/(sqrt(2*pi)*sigma_j)*exp(-1/2*((y_j-mu_j)/sigma_j)^2)
mu_j=n*p;
sigma_j=sqrt(n*p*(1-p)); % sigmaj is the standard deviation
y_j=0:0.2:20;
fy_j=1/(sqrt(2*pi)*sigma_j)*exp(-1/2*((y_j-mu_j)./sigma_j).^2)
% fyj=1/(sqrt(2*pi)*sigmaj)*exp(-1/2*((y-muj)/sigmaj).^2);
% Plot using ax1 handle, saved above,to save this graph
% on top of the previous graph.
h1=line(y_j,fy_j,'Color','r','Parent',ax1,'Linewidth',2);
set(h1,'linestyle','--','color','r','linewidth',2)
xlabel(''),ylabel('Probability')
title('Figure 4.3.1')
figure(gcf);pause
```

4.3.1 **Theorem 4.3.1** *Let X be a binomial random variable defined on n independent trials for which p=P(success). For any numbers a and b*

$$\lim_{n\to\infty} P\left(a \le \frac{X - np}{\sqrt{np(1-p)}} \le b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-z^2/2}dz.$$

Comment: The function $f_z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. is referred to as the **standard normal** (or **Gaussian**) curve [or distribution]. By convention, any random variable whose probabilistic behavior is described by a standard normal curve is denoted by *Z* (rather than by X, Y, or W)... E(Z)=0 and Var(Z)=1. Page 294 in Larsen & Marx (2006)

4.3.2 Finding areas under the standard normal curve

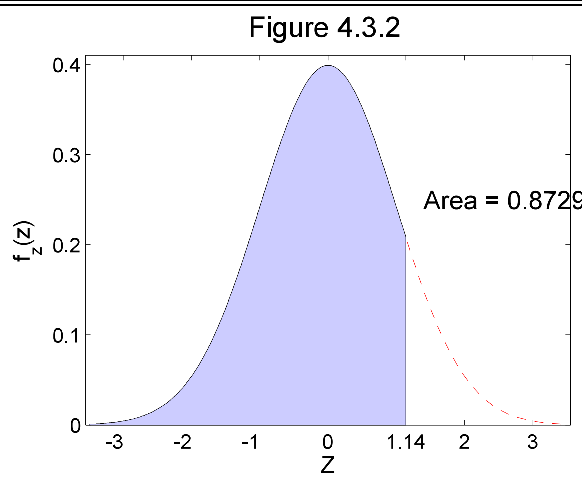


Figure 4. Figure 4.3.2

```
Figure 4.3.2
% LMFig040302_4th.m
% Larsen & Marx Figure 4.3.2 plot, page 295
% Use LMex040308_4th.m as a model
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Page 267
% Standard normal distribution using ezplot
% Integration using the symbolic math toolbox
% Written by Eugene.Gallagher@umb.edu
% October 1, 2010
clear all
hold off; clf
syms z
% Just for fun, use symbolic math integration to find the integral
INTP=eval(int(1/sqrt(2*pi)*exp(-(z^2/2)),-inf,1.14))
% or use the standard normal cumulative distribution function:
intp=normcdf(1.14)

% Plot Figure 4.3.2
mu=0;
sigma=1;
X=-3.5:0.1:3.5;
```

```
Y = normpdf(X,mu,sigma);
plot(X,Y,'--r');
axis([-3.55 3.55 0 0.41]);
title('Figure 4.3.2','FontSize',16);
ax1=gca;
xlabel(' '),
ylabel('f_z(z)','FontSize',14);
ax1=gca;
set(ax1,'XTick',[-3 -2 -1 0 1.14 2 3] ,'FontSize',12,...
   'XTickLabel',{'-3','-2','-1','0','1.14','2','3'},'FontSize',12)
set(ax1,'ytick',0:.1:0.4,'FontSize',12)
hold on;
xf=-3.5:.01:1.14;yf=normpdf(xf,mu,sigma);
fill([-3.5 xf 1.14]',[0 yf 0]',[.8 .8 1])
hold off;
```

4.3.3 The continuity correction

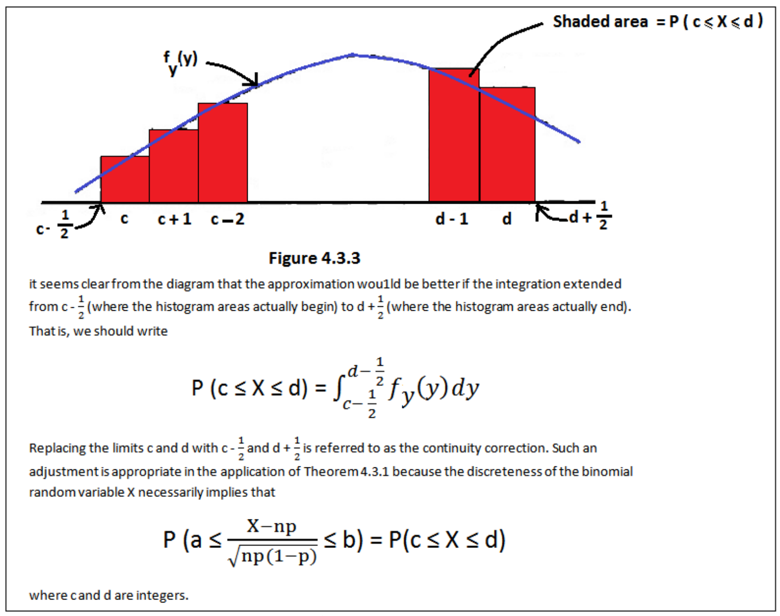


Figure 4.3.3

it seems clear from the diagram that the approximation wou1ld be better if the integration extended from c - $\frac{1}{2}$ (where the histogram areas actually begin) to d + $\frac{1}{2}$ (where the histogram areas actually end). That is, we should write

$$P\ (c \le X \le d) = \int_{c-\frac{1}{2}}^{d-\frac{1}{2}} f_y(y)\,dy$$

Replacing the limits c and d with c - $\frac{1}{2}$ and d + $\frac{1}{2}$ is referred to as the continuity correction. Such an adjustment is appropriate in the application of Theorem 4.3.1 because the discreteness of the binomial random variable X necessarily implies that

$$P\ (a \le \frac{X-np}{\sqrt{np(1-p)}} \le b) = P(c \le X \le d)$$

where c and d are integers.

Figure 5. Figure 4.3.3, the continuity correction

Example 4.3.1

```
% LMex040301_4th.m
% Example 4.3.1, Overbooking airline seats; Application of the normal
% distribution and the continuity correction. Page 297 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher, http://www.es.umb.edu/edgwebp.htm
% Written Fall 2010, Revised 1/7/11
% 178 seats sold for 168 economy class seats available, p=0.9 passenger
% will show up. What is the probability that not everyone who shows up at
% the gate on time can be accomodated?
n=178;p=0.9;n2=169;
```

```
P=normcdf((n+0.5-n*p)/sqrt(n*p*(1-p)))-...
   normcdf((n2-0.5-n*p)/sqrt(n*p*(1-p)));
fprintf(...
 '\nP not all can be accomodated, with continuity correction = %6.4f.\n',P)
% Without the continuity correction:
P2=normcdf((n-n*p)/sqrt(n*p*(1-p)))-...
   normcdf((n2-n*p)/sqrt(n*p*(1-p)));
fprintf(...
'P not all can be accomodated, without continuity correction = %6.4f.\n',P2)
k=169:178;
% Solve as a binomial pdf
 P3 = sum(binopdf(k,n,p));
 fprintf(...
'P not all can be accomodated, exact binomial = %6.4f.\n',P3)
 P4 = binocdf(n,n,p)-binocdf(n2-1,n,p);
 fprintf(...
'P not all can be accomodated, exact binomial cdf = %6.4f.\n',P4)
```

Case Study 4.3.1 ESP

```
% LMcs040301_4th.m
% Case Study 4.3.1 ESP P 298 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene Gallagher for EEOS601
% Eugene.Gallagher@umb.edu, http://www.es.umb.edu/edgwebp.htm
% Thirty two students correctly guessed 12,489 times out of 60,000
% when the expected number of successes would be just 12,000. What is
% the probability of this occuring by chance.
k=12489;
n=6e4;
p=1/5;
P1=1-binocdf(k,n,p);
fprintf('\nUsing the binomial cdf, the exact P = %9.7g\n',P1)
X=(12488.5-n*p)/sqrt(n*p*(1-p));
P2=1-normcdf(X);
fprintf('Using the Normal approximation with the continuity \n')
fprintf('correction, the exact P = %9.7g\n',P2)
X=(12489-n*p)/sqrt(n*p*(1-p));
P3=1-normcdf(X);
fprintf('Using the Normal approximation without the continuity \n')
fprintf('correction, the exact P = %9.7g\n',P3)
```

Questions Page 299

4.3.4 Central Limit Theorem (p 301)

Central limit theorem

Every binomial random variable X can be written as the sum of n independent Bernoulli random variables $X_1, X_2, ..., X_n$ where $X_i=1$ with probability P and 0 with probability 1-p, but if $X=X_1+X_2+...+X_n$, Theorem 4.3.1 can be reexpressed as

$$\lim_{n\to\infty} P\left(a \le \frac{X_1+X_2+...+X_n-np}{\sqrt{np(1-p)}} \le b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-z^2/2}dz.$$

The Russian Lyapunov made many of the key advances, and, in 1920, George Polya called it the **central limit theorem**.

Theorem 4.3.2, p 273 Let W_1,W_2, ... be an infinite sequence of of independent random variables with the same distribution. Suppose that the mean μ and the variance σ^2 of $f_w(w)$ are both finite. For any numbers a and b.

$$\lim_{n\to\infty} P\left(a \le \frac{W_1+...+W_n-n\mu}{\sqrt{n}\,\sigma} \le b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-z^2/2}dz.$$

Comment The central limit theorem is often stated in terms of the average of W_1, W_2, ..., and W_n, rather than the sum. Theorem 4.3.2 can be stated in equivalent form

$$\lim_{n\to\infty} P\left(a \le \frac{\bar{W}-\mu}{\sigma/\sqrt{n}} \le b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-z^2/2}dz.$$

Example 4.3.2

% LMEx040302_4th.m
% Larsen & Marx Example 4.3.2; Example of the central limit theorem
% Page 302-303
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Integration to find the variance of a uniform pdf using the symbolic math
% toolbox
% Written by Eugene.Gallagher@umb.edu as LMex040302_3rd.m in 2001
% revised 1/8/11
% See also LMEx040304_4th.m on rounding for variance of uniform pdf

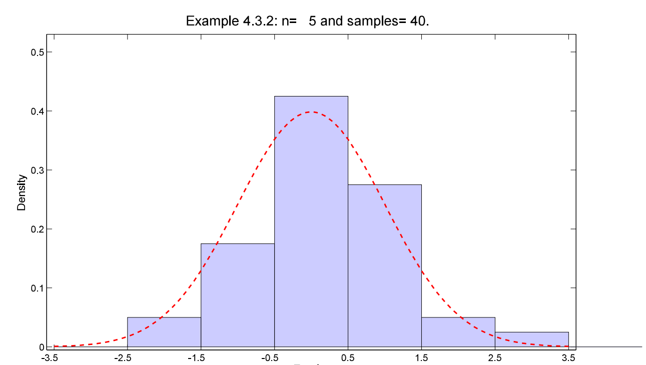


Figure 5. Example 4.3.2 Page 302 in Larsen & Marx (2006). 40 samples of 5 uniformly distributed random numbers after conversion to Z scores. The fit to the normal distribution is quite good even though each of the 40 samples is based on only 5 uniformly distributed numbers.

```
% See also LMEx040304_3rd.m  on rounding for variance of uniform pdf

% Generate 40 sets of 5 uniform random numbers as in Table 4.3.2, p 302
% Figure in Table 4.3.2 uses 7 bins and 40 samples;
% Try 1000 samples and numbins=20 for a prettier plot
 n=5;samples=40; numbins=7; % Numbins for the histogram, multiple of 7
% n=5;samples=1000; numbins=14;
% Numbins for the histogram,
multiple of 7
% n=100;samples=1000;
numbins=14; % Numbins for the
histogram, multiple of 7
% n=1000;samples=500;
numbins=14; % For the histogram
% n=1000;samples=40; numbins=7;
% For the histogram
C=rand(samples,n);
C6=sum(C')'; % sums of n replicates
% Use symbolic math toolbox to
calculate the variance of a uniform
% random distribution
mu=0.5;
syms f z y
disp('Definite integral (0,1) of y^2')
int(y^2,0,1)
EYsq=int(sym(y^2),0,1);
VARY=EYsq-mu^2;
disp('Var of a uniformly distributed random number on the interval (0,1):')
disp(VARY)
vary=eval(VARY); % convert from symbolic to numeric
% The variance is 1/12
% The Z ratios of the sums are in column C7 of Table 4.3.2
C7=(C6-n*mu)./(sqrt(n)*sqrt(vary));
Table040302=[[1:samples]' C C6 C7];
if n==5   % Reproduce Table 4.3.2
 fprintf('                     Table 4.3.2\n')
 fprintf('     y1      y2      y3      y4      y5      y    Z-ratio\n');
 fprintf('%2.0f %7.4f %7.4f %7.4f %7.4f %7.4f %7.4f %7.4f\n',Table040302');
end
% Use the code on page 5-115 of Statistics toolbox to draw histogram:
binwidthscaling=7/numbins; % This will produce 7 bins (3.5-(-3.5))
edges=-3.5:binwidthscaling:3.5;N=histc(C7,edges);
bar(edges,N/sum(N)/binwidthscaling,1);
ax1=gca;
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
```

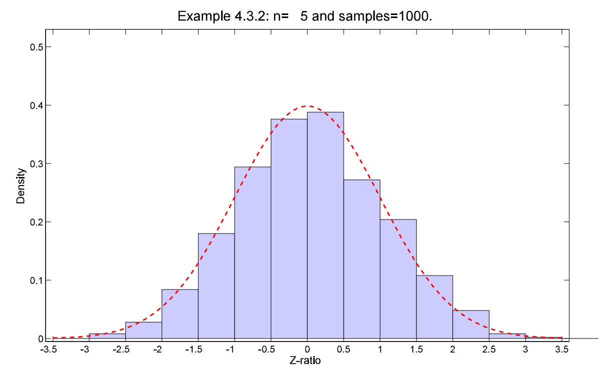


Figure 6. Example 4.3.2 Page 302 in Larsen & Marx (2006). 1000 samples of 5 uniformly distributed random numbers after conversion to Z scores. The number of bins increased from 7 to 14 for this bar chart. The normal distribution provides an excellent fit even though each of 1000 means is based on only 5 cases drawn from a uniform distribution. The uniform distribution is symmetric, allowing a rapid convergence to the normal distribution.

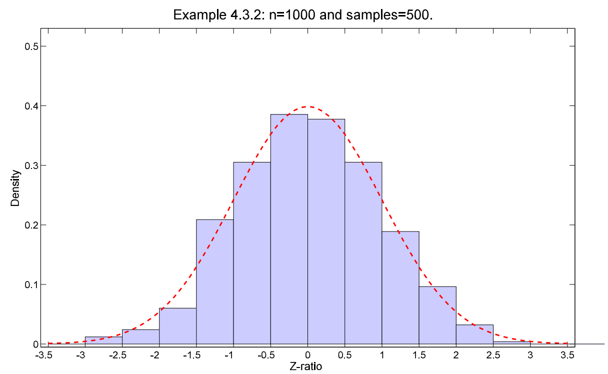


Figure 7. Example 4.3.2 Page 302 in Larsen & Marx (2006). 1000 samples of 500 uniformly distributed random numbers after conversion to Z scores. The number of bins increased from 7 to 14 for this bar chart. The fit of the normal distribution is excellent as would be expected from the large number of cases (500) on which each mean is based.

```
axis([-3.6 3.6 -0.005 0.53])
ylabel('Density','FontSize',14);xlabel('Z-ratio')
set(ax1,'xtick',edges,'FontSize',12);
set(ax1,'ytick',0:0.1:0.5,'FontSize',12)
figure(gcf);pause
% Superimpose the normal probability pdf:
% The normal probability equation is provided on p. 293
% This is for mean 0, and unit standard
% deviation. The more general equation (Legendre & Legendre, 1998 p. 147) is:
% f(y_j)=1/(sqrt(2*pi)*sigma_j)*exp(-1/2*((y_j-mu_j)/sigma_j)^2)
mu_j=0;
sigma_j=1; % sigmaj is the standard deviation; = 1 after Z transform
y_j=-3.5:0.1:3.5;
fy_j=1/(sqrt(2*pi)*sigma_j)*exp(-1/2*((y_j-mu_j)./sigma_j).^2);
% fyj=1/(sqrt(2*pi)*sigmaj)*exp(-1/2*((y-muj)/sigmaj).^2);
% Plot using ax1 handle, saved above,to save this graph
% on top of the previous graph.
h1=line(y_j,fy_j ,'Color','r','Parent',ax1,'Linewidth',2);
set(h1,'linestyle','--','color','r','linewidth',2)
s=sprintf('Example 4.3.2: n=%4.0f and samples=%3.0f.',n,samples);
title(s)
figure(gcf);pause
```

```
% LMex040302_logrnd_4th.m
% Larsen & Marx Example 4.3.2; Example of the central limit theorem using
% lognormally distributed variables, instead of uniform
% Page 302-303
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Integration to find the variance of a uniform pdf using the symbolic math
% toolbox
```

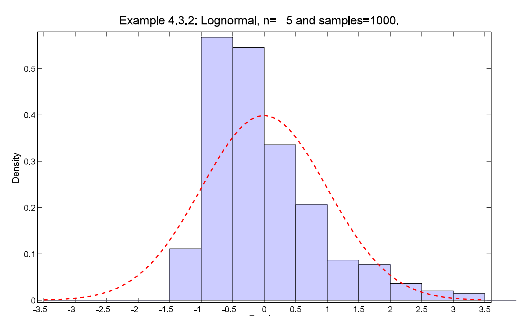


Figure 8. Example 4.3.2 Page 302 in Larsen & Marx (2006). 40 samples of 5 lognormally distributed (μ=1, σ=1) random numbers after conversion to **Z-scores**. The normal approximation is quite poor. Because each of the 40 samples is based on a mean of only 5 cases.

```
% Written by Eugene.Gallagher@umb.edu as LMex040302_3rd.m in 2001
% revised 1/8/11
% See also LMEx040304_4th.m  on rounding for variance of uniform pdf
% See also LMEx040304_3rd.m  on rounding for variance of uniform pdf

% Generate 40 sets of 5 uniform random numbers as in Table 4.3.2, p 302
% Figure in Table 4.3.2 uses 7 bins
and 40 samples;
% Try 1000 samples and
numbins=20 for a prettier plot
% n=5;samples=40; numbins=7; %
Numbins for the histogram, multiple
of 7
 n=5;samples=1000; numbins=14; %
Numbins for the histogram, multiple
of 7
% n=100;samples=1000;
numbins=14; % Numbins for the
histogram, multiple of 7
% n=1000;samples=1000;
numbins=14; % For the histogram
% n=1000;samples=40; numbins=7;
% For the histogram
```

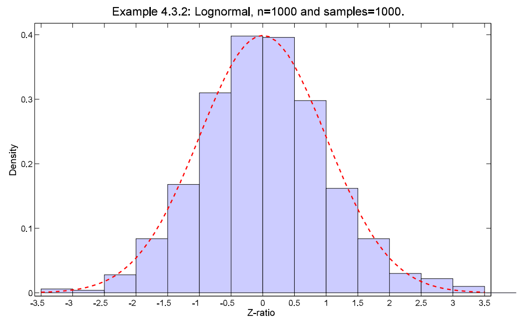


Figure 9. Example 4.3.2 Page 302 in Larsen & Marx (2006). 1000 samples of 1000 lognormally distributed (μ=1, σ=1) random numbers after conversion to Z scores. I used 14 bins for this bar chart. Even for the highly asymmetrically lognormally distributed samples, the normal approximation is excellent because of the large number of cases (1000) on which each mean was based.

```
lognormal=0;
% This example calculates the z scores based on the parametric mean and
% variance for the underlying distributions. For the uniform distribution,
% the variance is calculated from the variance of the distribution to be
% 1/12. For the lognormal distribution, lognstat is the built-in Matlab
% function that will covert the mean and variance from the log scale to the
% standard scale.
if lognormal
   mu=1;var=1;
   C=lognrnd(1,1,samples,n);
   [mu,vary]=lognstat(1,1);
s=sprintf('Example 4.3.2: Lognormal, n=%4.0f and samples=%3.0f.',n,samples);
s2=sprintf('Example 4.3.2: Lognormal distribution, 1st sample, n=%4.0f',n);
else
   C=rand(samples,n);
   mu=0.5;
   % Use symbolic math toolbox to calculate the variance of a uniform
   % random distribution
   syms f z y
   disp('Definite integral (0,1) of y^2')
   int(y^2,0,1)
   EYsq=int(sym(y^2),0,1);
```

```
    VARY=EYsq-mu^2;
    disp(...
     'Var of a uniformly distributed random number on the interval (0,1):')
    disp(VARY)
    vary=eval(VARY); % convert from symbolic to numeric
    % The variance is 1/12
 s=sprintf('Example 4.3.2: Uniform, n=%4.0f and samples=%4.0f.',n,samples);
 s2=sprintf('Example 4.3.2: Uniform distribution, 1st sample, n=%4.0f',n);
end
C6=sum(C')';  % sums of n replicates
% The Z ratios of the sums are in column C7 of Table 4.3.2
C7=(C6-n*mu)./(sqrt(n)*sqrt(vary));
Table040302=[[1:samples]' C C6 C7];
if n==5   % Reproduce Table 4.3.2
 fprintf('                    Table 4.3.2\n')
 fprintf('      y1       y2       y3       y4       y5       y    Z-ratio\n');
 fprintf('%2.0f %7.4f %7.4f %7.4f %7.4f %7.4f %7.4f %7.4f\n',Table040302');
end

% Use the code on page 5-115 of Statistics toolbox to draw histogram:
% Draw a histogram of one row of data;
hist(C(1,:));
ax1=gca;
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
ylabel('Frequency','FontSize',16);xlabel('X','FontSize',16)
title(s2,'FontSize',20)
figure(gcf);pause

% histogram of the z scores
binwidthscaling=7/numbins; % This will produce 7 bins (3.5-(-3.5))
edges=-3.5:binwidthscaling:3.5;N=histc(C7,edges);
bar(edges,N/sum(N)/binwidthscaling,1,'histc');
ax1=gca;
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
axis([-3.6 3.6 -0.005 1.05*max(N/sum(N)/binwidthscaling)])
ylabel('Density','FontSize',16);xlabel('Z-ratio','FontSize',16)
set(ax1,'xtick',edges,'FontSize',14);
set(ax1,'ytick',0:0.1:0.7,'FontSize',14)
figure(gcf);pause
% Superimpose the normal probability pdf:
% The normal probability equation is provided on p. 293
% This is for mean 0, and unit standard
% deviation. The more general equation (Legendre & Legendre, 1998 p. 147) is:
% f(y_j)=1/(sqrt(2*pi)*sigma_j)*exp(-1/2*((y_j-mu_j)/sigma_j)^2)
mu_j=0;
```

```
sigma_j=1; % sigmaj is the standard deviation; = 1 after Z transform
y_j=-3.5:0.1:3.5;
fy_j=1/(sqrt(2*pi)*sigma_j)*exp(-1/2*((y_j-mu_j)./sigma_j).^2);
% fyj=1/(sqrt(2*pi)*sigmaj)*exp(-1/2*((y-muj)/sigmaj).^2);
% Plot using ax1 handle, saved above,to save this graph
% on top of the previous graph.
h1=line(y_j,fy_j ,'Color','r','Parent',ax1,'Linewidth',2);
set(h1,'linestyle','--','color','r','linewidth',2)

title(s,'FontSize',20)
figure(gcf);pause
```

Example 4.3.3

```
% LMex040303_4th.m
% Example 4.3.3 Pp 304 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% An application of the symbolic math integration functions
% Typo in book:
% f_y(y)=y* 3*(1-y)^2 not fy(y)=3*(1-y)^2
syms y
% f_y(y)=y* 3*(1-y)^2, 0<=y,=1; Let mean y = (1/15)*Sum from 1 to 15 Yi.
int(y*3*(1-y)^2,0,1)
muy=eval(int(y*3*(1-y)^2,0,1));
% by definition of variance
int(y^2*3*(1-y)^2,0,1)-1/4^2
vary=eval(int(y^2*3*(1-y)^2,0,1)-1/4^2);
P=normcdf((3/8-muy) / (sqrt(vary)/sqrt(15)))-...
  normcdf((1/8-muy) / (sqrt(vary)/sqrt(15)))
```

Example 4.3.4

```
% LMEx040304_4th.m
% Example 4.3.4 Effects of rounding analyzed with normal approximation
% Effect of rounding on the expected value and variance of rounded
% estimates. P. 304 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Using the central limit theorem, what is the probability that the error
% on 100 transactions, each rounded to the nearest $100, would exceed $500?
% What is the variance of a uniformly distributed random number
% over the interval -50 to 50?
% From definition of Variance:
% Var(Y)=E[(Y-mu)^2]= integral from -50 to 50 {(y-mu)^2 f_y(y)} dy.
% mu=0;
% f_y(y)=1; This is the height of the uniform pdf at each value of y.
% for max(y)-min(y)=1, the value of f_y(y) is 1.
% for uniform random numbers spread over different ranges,
% for max(y)-min(y)*f_y(y) must equal 1.
% thus, for a uniform random number distributed on the interval -100 to 100,
```

```
%                       f_y(y)=1/200;
% Written by Eugene.Gallagher@umb.edu for EEOS601
% http://www.es.umb.edu/edgwebp.htm
% Written in 2001, revised 1/9/11
% The m.file goes through some applications of the symbolic math toolbox at
% the beginning to verify that the variance of 100 rounded transactions is
% 2500/3
format rat
syms y mu
int(y^2/100,-50,50)%Use the symbolic math toolbox to find the
                    % variance
               % of a uniformly distributed variable on the interval
               % -50 to 50; it is 2500/3
VarY=eval(int(sym('y^2/100'),-50,50))
% Variance of a uniform randomly distributed number
% on the interval -50 to 50 is 2500/3
% Note how the variance remains 2500/3 with a shift in mean:
% Var(Y)=E(Y^2)-mu^2;
mu=50;
int(y^2/100,0,100)
EYsq=eval(int(sym('y^2/100'),0,100))
% The previous statement produces 10,000/3 which produces an identical
% variance after subtracting the square of the expected value, mu^2, as is
% appropriate from the definition of variance.
VARY=EYsq-mu^2
% Note that Var(Y) is also defined as
% integral -inf to inf (y-mu).^2 f_y(y) dy.
% We can create the definite integral of this equation and evaluate it
% to produce the variance
syms y, mu; mu=50;
VARY2=eval(int(sym('(y-mu)^2/100'),y,0,100))
% VARY2 should also be 2500/3
format
% using Theorem 3.13.12, in Larsen & Marx (2001, p. 223)
% Calculate the variance of all 100 transactions (multiply the variance by
% 100). Note, this is for 100 indepedently distributed random variates,
% where the positive and negative errors among transactions are not
% correlated.
Varall=100*VARY;
% If you wanted to convert the variance, calculated here in dollars to the
% variance in pennies, you'd have to use Theorem 3.13.1, p. 222.  It would
% the variance of Varall in pennies would be a=100;Var(a*Varall)=a^2 * Varall
%                                             =10000 * Varall
% calculate the standard deviation;
sigmaall=sqrt(Varall);
```

```
% Calculate the Z statistic
Z=zeros(2,1);
Z(1)=(-500-0)/sigmaall;
Z(2)=(500-0)/sigmaall
% Use the cumulative normal probability function to find the probability of
% being more than $5 dollars off after 100 transactions:
P=1-(normcdf(Z(2))-normcdf(Z(1)))
fprintf('The probability of being more than $500 off after 100\n')
fprintf('transactions in which each transaction is rounded to the \n')
fprintf('nearest $100 is %6.4f\n',P)
```

Example 4.3.5

```
% LMex040305_4th.m
% Example 4.3.5 Memphis earthquakes and the Poisson distribution
% Pp 305-306 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written Fall 2010 by Eugene Gallagher for EEOS601 Revised: 1/8/11
% Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Annual earthquakes registering 2.5 on the Richter scale and within 40
% miles of downtown Memphis have a Poisson distribution with lambda=6.5
% Calculate the exact probability that nine or more such earthquakes will
% strike next year using the Poisson distribution and compare that with the
% normal approximation based on the central limit theorem.
lambda=6.5;
N=9;
P=1-poisscdf(N-1,lambda);
fprintf('\nThe exact p value using the Poisson cumulative distribution ')
fprintf('function with lambda=%5.3f is %7.4f.\n',lambda,P)
P2=1-normcdf(((N-1)-lambda)/sqrt(lambda));
fprintf('\nThe approximate p value using the central limit theorem is ')
fprintf('%7.4f.\n',P2)
continuity=0.5;
% "The continuity correction is appropriate whenever a discrete probability
% model is being approximated by the area under a curve.)
P3=1-normcdf(((N+continuity-1)-lambda)/sqrt(lambda));
fprintf('\nThe approximate p value using the central limit theorem with ');
fprintf('the continuity correction is ')
fprintf('%7.4f.\n',P3)

% Optional: plot the Poisson pdf and the normal approximation
% Use LMFig040301_4th.m as a model
X=0:18;
Y = poisspdf(X,lambda);
bar(X,Y,1);axis([-0.6 19 0 0.2]);title('Example 4.3.5','FontSize',14);
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
```

```
ylabel('Probability','FontSize',14);
ax1=gca;set(ax1,'xtick',0:2:18,'FontSize',12);
set(ax1,'ytick',0:0.05:0.2,'FontSize',12)
figure(gcf);pause
% Superimpose the normal probability pdf:
% The normal probability equation is provided on p. 293
% This is for mean 0, and unit standard
% deviation. The more general equation (Legendre & Legendre, 1998 p. 147):
% f(y_j)=1/(sqrt(2*pi)*sigma_j)*exp(-1/2*((y_j-mu_j)/sigma_j)^2)
mu_j=lambda;
sigma_j=sqrt(lambda); % sigmaj is the standard deviation
y_j=0:0.2:18;
fy_j=1/(sqrt(2*pi)*sigma_j)*exp(-1/2*((y_j-mu_j)./sigma_j).^2);
% fyj=1/(sqrt(2*pi)*sigmaj)*exp(-1/2*((y-muj)/sigmaj).^2);
% Plot using ax1 handle, saved above,to save this graph
% on top of the previous graph.
h1=line(y_j,fy_j,'Color','r','Parent',ax1,'Linewidth',2);
set(h1,'linestyle','--','color','r','linewidth',2)
xlabel('Number of Memphis Earthquakes > 2.5'),
figure(gcf);pause
```

Questions p 306

4.3.5 The Normal Curve as a Model for Individual Measurements

Definition 4.3.1 A random variable Y is said to be normally distributed with mean μ and variance σ² if

$$f_y(y) = -\sqrt{}$$

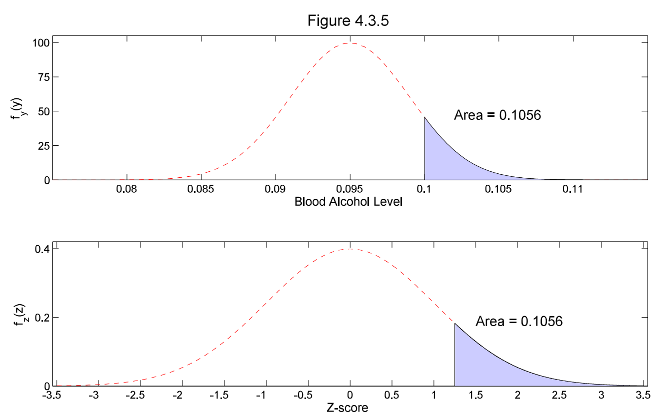


Figure 10. Figure 4.3.5 from Example 4.3.6. If a driver's true blood alcohol level is 0.095 and the breathalyzer has an instrumental error σ= 0.004% (pdf shown in upper panel), What is the probability that a driver will be erroneously charged with DUI. The problem can be analyzed using the pdf for the unscaled data or the data can be converted to z scores (lower panel) by subtracting the mean and dividing by σ. The probability of an erroneous charge is 10.56%, the area to the right of the cutpoints shown in the normal and standard normal curves.

Comment. Areas under and "arbitrary" normal distribution, $f_y(y)$ are calculated by finding the equivalent area under the standard normal distribution, $f_z(z)$:

$$P(a \le Y \le b) = P\left(\frac{a-\mu}{\sigma} \le \frac{Y-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$

The ratio $\frac{Y-\mu}{\sigma}$ is often referred to as either a Z transformation or a Z score.

Example 4.3.6
% LMex040306_4th.m

```
% Example 4.3.6 Drunk driving pages 308-309 in Larsen & Marx (2006)
% Based on LMex040305_4th.m
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written Fall 2010 by Eugene Gallagher for EEOS601 Revised: 1/9/11
% Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% A legally drunk driver in many states has a blood alcohol level of 0.10%,
% but the breathalyzer has measurement error. Repeated measurements
% indicate that the the estimate of the mean is unbiased and sigma is
% 0.004%
% If a driver's true blood alcohol content is 0.095%, what is the
% probability that he will be incorrectly booked on a DUI charge?
mu=0.095;
sigma=0.004;
Observed=0.1;
P=1-normcdf((Observed-mu)/sigma);
fprintf('\nThe approximate p value using the central limit theorem is ')
fprintf('%7.4f.\n',P)

% Plot Figure 4.3.5
subplot(2,1,1)
X=0.075:0.0005:0.115;
Y = normpdf(X,mu,sigma);
plot(X,Y,'--r');
axis([0.075 0.115 0 105]);title('Figure 4.3.5','FontSize',14);
ax1=gca;
xlabel('Blood Alcohol Level'),
ylabel('f_y(y)','FontSize',14);
ax1=gca;
set(ax1,'xtick',0.08:0.005:0.11,'FontSize',12);
set(ax1,'ytick',0:25:100,'FontSize',12)
hold on;
xf=0.1:0.0001:.115;yf=normpdf(xf,mu,sigma);
fill([0.1 xf 0.115]',[0 yf 0]',[.8 .8 1])
hold off;

subplot(2,1,2)
X=-3.5:0.01:3.5;
Y = normpdf(X,0,1);
plot(X,Y,'--r');
axis([-3.55 3.55 0 0.42]);
ax1=gca;
xlabel('Z-score'),
ylabel('f_z(z)','FontSize',14);
ax1=gca;
```

```
set(ax1,'xtick',-3.5:0.5:3.5,'FontSize',12);
set(ax1,'ytick',0:0.2:0.4,'FontSize',12)
hold on;
xf=1.25:0.001:3.5;yf=normpdf(xf,0,1);
fill([1.25 xf 3.5]',[0 yf 0]',[.8 .8 1])
figure(gcf);pause
hold off;
subplot(1,1,1)
```

Example 4.3.7 Mensa p. 309

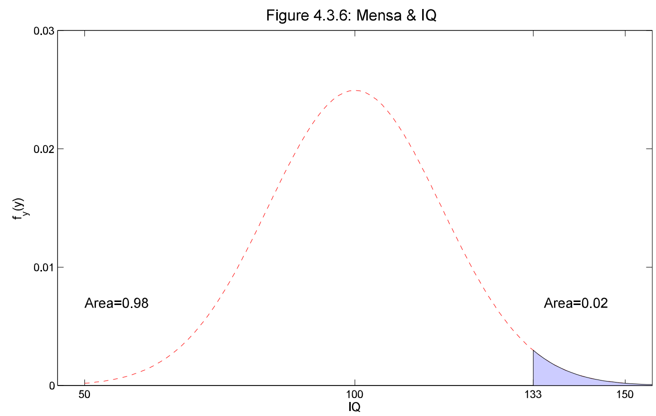


Figure 11. Figure 4.3.6 from Example 4.3.7. A Mensa score of 133 marks the upper 2% of IQ's if μ=100 and σ=16.

```
% LMex040307_4th.m
% Example 4.3.7 Mensa & IQ Pages 309-310 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Based on LMex040306_4th.m
% Written Fall 2010 by Eugene Gallagher for EEOS601 Revised: 1/9/11
% Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% A person must be in the upper 2% of the population in IQ to join. What is
% the minimum IQ that will qualify a person for membership. Assume that the
% mean IQ is 100 and sigma is 16.
mu=100;
sigma=16;
% Use the inverse of the normal pdf to find the IQ cutoff for the upper 2%
% of the population 0.98=(1-0.02)
IQmensa=norminv(0.98,100,16);
fprintf('\nThe lowest acceptable IQ for membership in Mensa is ')
fprintf('%3.0f.\n',ceil(IQmensa))

% Plot Figure 4.3.6
X=50:150;
Y = normpdf(X,mu,sigma);
plot(X,Y,'--r');
axis([45 155 0 0.03]);title('Figure 4.3.6: Mensa & IQ','FontSize',20);
ax1=gca;
xlabel('IQ','FontSize',16),
ylabel('f_y(y)','FontSize',16);
ax1=gca;
set(ax1,'xtick',[50 100 133 150],'FontSize',14);
```

```
set(ax1,'ytick',0:.01:0.03,'FontSize',14)
hold on;
xf=ceil(IQmensa):155;yf=normpdf(xf,mu,sigma);
fill([ceil(IQmensa) xf 155]',[0 yf 0]',[.8 .8 1])
text(50,0.007,'Area=0.98','FontSize',18);
text(135,0.007,'Area=0.02','FontSize',18)
figure(gcf);pause
hold off;
```

Example 4.3.8 Army Anti-Tank Missile

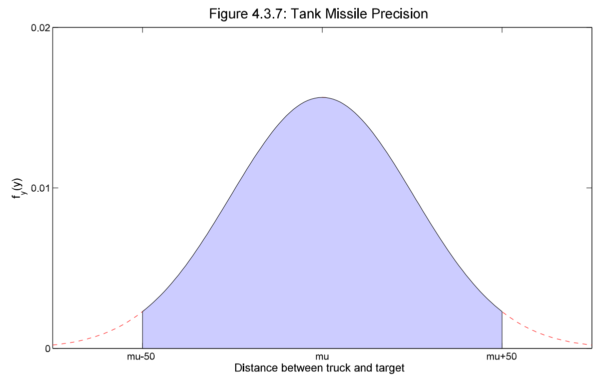


Figure 12. Figure 4.3.7 from Example 4.3.8. A Mensa score of 133 marks the upper 2% of IQ's if μ=100 and σ=16.

```
% LMex040308_4th.m
% Example 4.3.8 Truck-launced antitank missile Pages 310-11 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Based on LMex040306_4th.m
% Written Fall 2010 by Eugene Gallagher for EEOS601 Revised: 1/8/11
% Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% The army is soliciting proposlas requiring that the automatic sighting
% mechanism be sufficiently reliable to guarantee that 95% of the missiles
% fall no more than 50 feet short or 50 feet beyond the target. What is the
% largest sigma compatible with that degree of precision. Assume that Y,
% the distance a missile travels is normally distributed with its mean (mu)
% equal to the the length of separation between the truck and the target.
Zvalue=norminv(0.975);
%Zvalue=50/sigma so,
sigma=50/Zvalue;
fprintf('\nThe largest acceptable sigma is %4.2f feet.\n',sigma)

% Plot Figure 4.3.7
X=-75:75;
mu=0;
Y = normpdf(X,mu,sigma);
plot(X,Y,'--r');
axis([-75 75 0 0.02]);
title('Figure 4.3.7: Tank Missile Precision','FontSize',20);
ax1=gca;
xlabel('Distance between truck and target','FontSize',16),
ylabel('f_y(y)','FontSize',16);
ax1=gca;
```

```
set(ax1,'xtick',-50:50:50,'FontSize',14,...
   'XTickLabel',{'mu-50','mu','mu+50'},'FontSize',14)
set(ax1,'ytick',0:.01:0.02,'FontSize',14)
hold on;
xf=-50:50;yf=normpdf(xf,mu,sigma);
fill([-50 xf 50]',[0 yf 0]',[.8 .8 1])
figure(gcf);pause
hold off;
```

Example 4.3.9 Deals with Moment generating functions, not covered in EEOS601

Theorem 4.3.3 *Let Y_1 be a normally distributed random variable with mean μ_1 and variance σ^2_1 and let Y_2 be a normally distributed random variable with mean μ_2 and variance σ^2_2. Define $Y=Y_1+Y_2$. If Y_1 and Y_2 are independent, Y is normally distributed with mean $\mu_1 + \mu_2$ and variance $\sigma^2_1 + \sigma^2_2$.*

Corollary *Let Y_1, Y2, ..., Y_n be a random sample of size n from a normal distribution with mean μ and variance σ². Then the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is also normally distributed with mean μ but with variance equal to σ²/n (which implies that $\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$ is a standard normal random variable, Z).*

Corollary *Let Y_1, Y2, ..., Y_n be any set of independent normal random variables with means $\mu_1, \mu_2, ..., \mu_n$ and variances $\sigma^2_1, \sigma^2_2, ... \sigma^2_n$, respectively. Let $a_1, a_2, ..., a_n$ be any set of constants. Then $Y = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$ is normally distributed with mean $\mu = \sum_{i=1}^{n} a_i \mu_i$ and variance $\sigma^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$.*

Example 4.3.10

```
% LMex040310_4th.m
% Example 4.3.10 Swampwater Tech Elevator Pages 312-313 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Based on LMex040306_4th.m
% Written Fall 2010 by Eugene Gallagher for EEOS601 Revised: 1/8/11
% Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Swampwater Tech elevators have a maximum capacity of 2400 lb. Suppose
% that ten Swampwater Tech football players enter the elevator and that the
% football player weights are normally distributed with mean = 220 lb and
```

```
% standard deviation = 20 lb. What is the probability that there will be 10
% fewer football players at tomorrow's practice.
mu=220;
sigma=20;
n=10;
P=1-normcdf(240,mu,20/sqrt(n));
fprintf('\nThe probability of an elevator collapse is %4.2g.\n',P)
P2=1-normcdf(2400,mu*n,sqrt(n)*20);
fprintf('\nThe probability of an elevator collapse is %4.2g.\n',P2)
% If 11 players baoarded the elevator
n=11;
P3=1-normcdf(2400,mu*n,sqrt(n)*20);
fprintf('\nIf 11 players entererd the elevator, the expected weight \n')
fprintf('is %4.0f lb, and the probability of collapse is %5.3f.\n',...
    mu*11, P3);
```

Example 4.3.11

```
% LMex040311_4th.m
% Example 4.3.11 Aptitude tests for a corporation Pages 313-314 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Based on LMex040306_4th.m
% Written Fall 2010 by Eugene Gallagher for EEOS601 Revised: 1/8/11
% Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% The personnel department gives a verbal (Y1) and quantitative aptitude
% (Y2)test to all applicants. Y1 is normally distributed with mu1=50 and
% sigma1=20 and Y2 has mu2=100 and sigma2=20. Scores are independent. A
% composit score is assigned to each applicant: Y=3Y1+2Y2. The company
% rejects any applicant whose score is less than 375. If six individuals
% apply, what are the chances that fewer than half will fail the screening
% test?
mu1=50;
sigma1=10;
mu2=100
sigma2=20
cutoff=375;
n=6;
EY=3*mu1+2*mu2;
VarY=3^2*sigma1^2+2^2*sigma2^2;
fprintf('For a single applicant, the expected score is %3.0f with ',EY)
fprintf('variance = %4.0f\n',VarY);
Zscore=((375-EY)/sqrt(VarY))
P=normcdf(Zscore,0,1);
fprintf('The probability of a single applicant being rejected is %6.4f.\n',P)
fprintf('What is the probability that at most 2 of 6 will fail the test?\n');
P2=binocdf(2,6,P);
```

```
fprintf('The probability that fewer than half of 6 applicants will fail\n')
fprintf('the test is %6.4f.\n',P2)
```

Example 4.3.12

```
% LMex040312_4th.m
% Example 4.3.12 Pages 314 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Based on LMex040306_4th.m
% Written Fall 2010 by Eugene Gallagher for EEOS601 Revised: 1/8/11
% Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% An application of propagation of error
% Y=random sample of size 9 from a normal distribution with:
EY=2;
SigmaY=2;
nY=9;
% Ystar=independent random sample of size 4 from a normal distribution with
EYstar=1;
SigmaYstar=1;
nYstar=4;
% Find Probability mean(Y)>=mean(Ystar)
E_YminusYstar=EY-EYstar;
% Var_YminusYstar= VarY+VarYstar
format rat
Var_YminusYstar=SigmaY^2/nY+SigmaYstar^2/nYstar
format
P=1-normcdf((0-E_YminusYstar)/sqrt(Var_YminusYstar));
fprintf('The probability that Y >= Ystar = %6.4f\n',P)
```

4.4 **The geometric distribution** Not covered in Summer 2011

Example 4.4.1 Available on course website

Theorem 4.4.1

Example 4.4.2 Available on course website

Example 4.4.3 Not programmed. (Not really a Matlab-type problem)

Example 4.4.4 Not programmed.

Questions p. 321

4.5 **THE NEGATIVE BINOMIAL DISTRIBUTION**

Theorem 4.5.1

Example 4.5.1 Programmed in Matlab, but difficult

Questions p. 326

4.6 **THE GAMMA DISTRIBUTION, p. 296. Waiting time for the rth event to occur**

Theorem 4.6.1

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

Example 4.6.1 Programmed as an example of the symbolic math toolbox

4.6.1 **Generalizing the Waiting Time Distribution**

Theorem 4.6.2

Definition 4.6.2

Theorem 4.6.3

4.6.2 **Sums of Gamma Random Variables**

Theorem 4.6.4

Example 4.6.2

Theorem 4.6.5

4.7 TAKING A SECOND LOOK AT STATISTICS (MONTE CARLO SIMULATIONS)

4.7.1 Monte Carlo simulations are often used when the underlying probability model is complex

4.7.2 Section 4.7 analyzes a compound probability problem involving a Poisson/exponential process and a normal pdf

4.7.3 A 2-year warranty is offered for a plasma TV

4.7.4 Lambda = 0.75 service calls per year at an average cost of $100 per repair ($\sigma$=20)

4.7.5 Is the $200 2-year warranty a good deal?

4.7.6 Not mentioned by Larsen & Marx (2006), but the expected repair cost = 2*.75*$100=$150, but what is the variance?

4.7.6.1 "For any particular customer, the value of W [the warranty] will depend on (1) the number of repairs needed in the 1st two years and (2) the cost of each repair. Although we have reliability and cost assumptions that address(1) and (2), the 2-yr limit on the warranty introduces a complexity that goes beyond what we have learned in Chapters 3 and 4. What remains is the option of using random samples to simulate the repair costs that might accrue during those first two years." [p. 333]

```
% LM0407_4th.m
% A Monte Carlo simulation with graphics of the cost of a warranty
% A simulation combining the exponential pdf and Poisson pdf
% For a plasma TV screen, the sreen is expected to require 0.75 service
% calls per year. The cost of normally distributed and independently
% distributed service calls have mean=$100 and sigma=20. Should the buyer
% buy a warranty costing $200 for two years.
% Written by Eugene.Gallagher@umb.edu for EEOS601 1/10/11
% Revised
% Theorem 4.2.3 (Larsen & Marx 2006, P. 290) implies that the interval
% between repairs will follow the exponential distribution
```

```
% Figure 4.7.1
lambda=0.75;
mu=100;
sigma=20;
y=0:.1:4.5;
fyy=exppdf(y,1/lambda);  % Note
that Matlab's exp pdf is defined
                        % differently from
Larsen & Marx. Must enter
                        % 1/lambda to
observe same result.
plot(y,fyy,'--r','LineWidth',3);
axis([0. 4.4 0 0.8])
ax1=gca;
set(ax1,'xtick',[0:4],'FontSize',14,'Fon
tSize',14)
ax1=gca;
xlabel('Interval between repair calls (yr)','FontSize',16);
ylabel('f_y(y)','FontSize',16),
title('Figure 4.7.1','FontSize',20);
figure(gcf)
pause
```

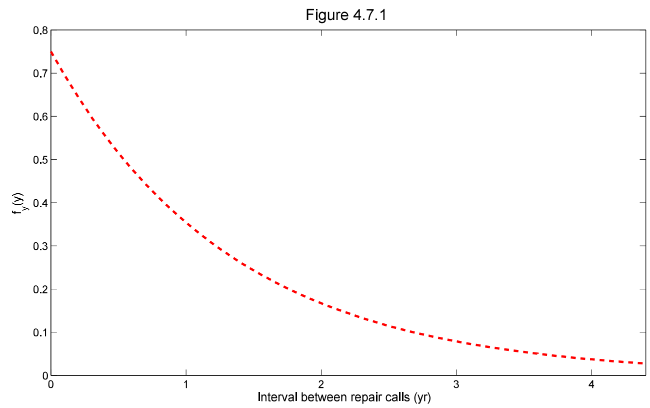


Figure 13. Figure 4.7.1 An application of Matlab's exponential distribution, which is defined differently from Larson & Marx (2006). y=0:.1:4.5;fyy=exppdf(y,1/lambda)

```
% Plot 4.7.2 using
LMex040308_4th.m as model
X=38:162;
mu=100;
sigma=20;
Y = normpdf(X,mu,sigma);
plot(X,Y,'--r','LineWidth',3);
axis([35 165 0 0.021]);
title('Figure 4.7.2: Cost of Screen
Repair','FontSize',20);
ax1=gca;
xlabel('Cost','FontSize',16),
ylabel('f_c(c)','FontSize',16);
ax1=gca;
set(ax1,'xtick',[40 100 160],'FontSize',14,...
   'XTickLabel',{'$40(mu-3 sigma)','$100','$160 (mu+3 sigma)'},...
   'FontSize',14)
set(ax1,'ytick',0:.01:0.02,'FontSize',14)
figure(gcf);pause
```

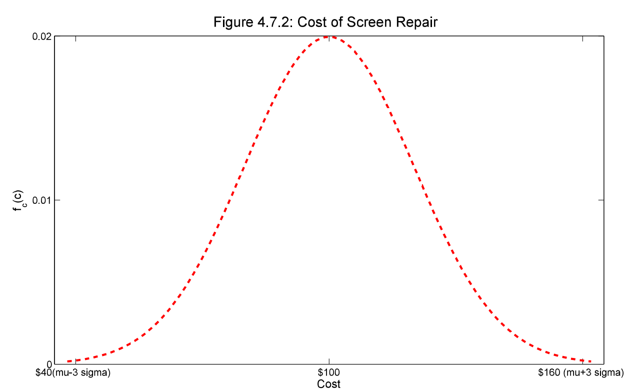


Figure 14. Figure 4.7.2 An application of Matlab's normal pdf. Page 334 in Larsen & Marx (2006), page 334.

```
% Program with nested for loop
Trials=100;
```

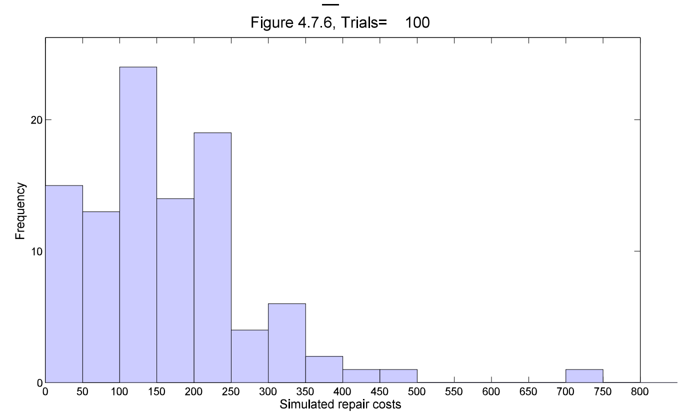
```
time=0;
COSTS=zeros(Trials,1);
for i=1:Trials
   cost=0;
   time=0;
   while time<=2
      time=exprnd(1/lambda)+time;
      if time<=2
         cost=cost+normrnd(mu,sigma);
      end
   end
   COSTS(i)=cost;
end
fprintf(...
'The max, median, and mean repair bills are $%4.0f, $%3.0f, and $%3.0f.\n',...
max(COSTS),median(COSTS),mean(COSTS));
fprintf('%4.1f%% of customers had no repair costs.\',sum(~COSTS)/Trials*100)

% Use the code from page 5-115 of Statistics toolbox & Ex040302_4th.m
% to draw histogram, Figure 4.7.6
if max(COSTS)>400;

maxbin=ceil(max(COSTS)/200)*200
;
else

maxbin=ceil(max(COSTS)/100)*100
;
end
edges=0:50:maxbin;
N=histc(COSTS,edges);
bar(edges,N,1,'histc');
maxf=ceil(max(N)/5)*5;
ax1=gca;
set(get(gca,'Children'),'FaceColor',[.8
.8 1])
axis([0 maxbin 0 1.05*maxf])
ylabel('Frequency','FontSize',16);
xlabel('Simulated repair
costs','FontSize',16)
set(ax1,'xtick',edges,'FontSize',14);
if maxf<20

set(ax1,'ytick',0:5:maxf,'FontSize',14)
elseif maxf<50
```



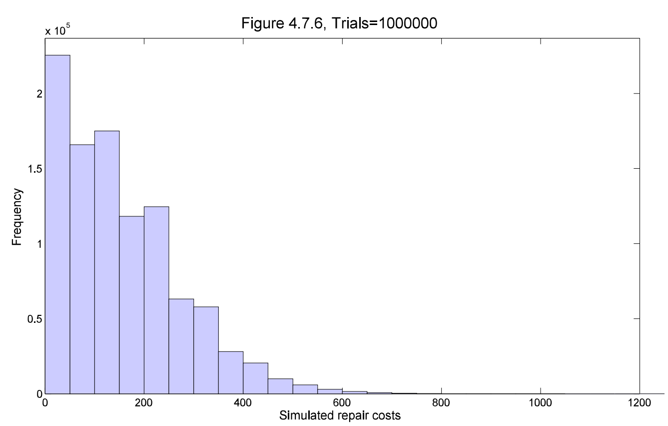


Figure 16. Figure 4.7.6 The result of 1 million Monte Carlo simulations of repair costs. Page 336 in Larsen & Marx (2006), page 334. The max, median, and mean repair bills are $1037, $119, and $150. 22.3% had no repair costs

```
    set(ax1,'ytick',0:10:maxf,'FontSize',14)
elseif maxf<100
    set(ax1,'ytick',0:20:maxf,'FontSize',14)
elseif maxf<500
    set(ax1,'ytick',0:100:maxf,'FontSize',14)
elseif maxf<1000
    set(ax1,'ytick',0:200:maxf,'FontSize',14)
elseif maxf<10000
    set(ax1,'ytick',0:1000:maxf,'FontSize',14)
elseif maxf<100000
    set(ax1,'ytick',0:10000:maxf,'FontSize',14)
end
s=sprintf('Figure 4.7.6, Trials=%7.0f',Trials);
title(s,'FontSize',20);
figure(gcf);pause
```

Appendix 4.A.1 Mintab applications
Appendix 4.A.2. A proof of the central limit theorem

References

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

Index