EEOS 601
UMASS/Online
Intro. Prob & Applied Statistics
Handout 6, Week 3
6/14 Tu - 6/20 M
Revised: 2/10/11

Week 3: Ch 3 Random Variables

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 3.1-3.6[part] 3.9 (p. 128-183, 193-199), 3.9 (p 226-236, has important implications for propagation of error)
    - Skim or skip entirely 3.7-3.8, 3.10-3.13

Understanding by Design Templates

Understanding By Design Stage 1 — Desired Results Week 3

LM Chapter 3 Random Variables

G Established Goals

- Apply the binomial and hypergeometric probability distribution functions (pdf's) and their cumulative distribution functions (cdf's) to environmental problems
- (MCAS 10th Grade standard 10.D.1) Select, create, and interpret an appropriate graphical representation (e.g., scatterplot, table, stem-and-leaf plots, box-and-whisker plots, circle graph, line graph, and line plot) for a set of data and use appropriate statistics (e.g., mean, median, range, and mode) to communicate information about the data. Use these notions to compare different sets of data.

U Understand

- Integral calculus is the language used to describe continuous probability functions (e.g., the cdf is the integral of the pdf)
- Casino games, based on well-defined pdfs, are predictable, but many real-life events like Wall Street crashes and climate change disasters follow poorly described or unknown pdfs. Modeling some real life processes like casino games is an example of Taleb's ludic fallacy **http://en.wikipedia.org/wiki/Ludic_fallacy**
- Models of population genetics, natural selection, and biodiversity are based on pdfs.
- Monte Carlo simulations are used not only to simulate analytical probabilities (LM Chapter 2) but also to generate pdfs for Bayesian inference.

Q Essential Questions

- Can you do statistics without understanding probability?
- What is the best bet on the Casino craps table?
- How can you estimate the number of species in a random draw of 100 individuals from a community?
- How can the US average income increase while the median income stays the same?

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

- **Bernoulli trial**, **binomial** & hypergeometric distributions, boxplot [not in text, but will be covered below], cumulative distribution function, discrete random variable, expected value, mean, median, probability density function, probability function (discrete [**Def. 3.3.1**] & continuous), standard deviation, variance

S *Students will be able to*

- Write Matlab programs to solve applied problems involving the binomial & hypergeometric distributions
- Compute the odds for craps
- Interpret and apply the Sanders-Hurlbert rarefaction equation for estimating species diversity
- Explain why Sewall Wright's genetic drift and the potential extinction of the Norther Right Whale is a consequence of the binomial probability distribution function.

Understanding by Design Stage 2 — Assessment Evidence Week 3 6/14-6/20

LM Chapter 3 Random Variables. Read 3.1-3.6 (Pp 128-198), 3.9 (Pp 226-236). Skip or skim 3.7-3.8 3.10-3.1

- **Post in the discussion section by 6/22/11 W**
    - What is the relationship between Hurlbert's $E(S_n)$ and the hypergeometric probability distribution function (what are the white balls, red balls, etc.)?
- **Problems due Wednesday 6/22/11 W 10 PM**
    - Each problem must be solved using Matlab
    - **Basic problems (4 problems 10 points)**
        - **Problem 3.2.2 P. 136 Nuclear power rods**
            - Use Example 3.2.2 as a model
        - **Problem 3.2.26 P. 147 Sock matching**
        - **Problem 3.5.2 P 184** Expected value of a Cracker Jack Prize
        - **Problem 3.6.14 P 199 Variance of a linear function**
            - Not much of a Matlab problem The answer to 3.6.5 is in the back of the book,
    - **Advanced problems (2.5 points each)**
        - Problem 3.3.6 p. 160 pdf for non-standard dice
            - Use Example 3.3.1 as a model
        - **Problem 3.4.14 page 173**
    - **Master problems (2 only, 5 points each, Choose only 1)**
        - With fair dice, the probability of the shooter winning in craps is (244/495≈0.493; the house or fader has a 1.42% advantage). What is the probability of the shooter winning with the crooked dice called ace-six flats described in Example 3.3.1 (p. 150) and sold at **http://www.gamblingcollectibles.com/equipment.html**?
        - Write a Matlab program that calculates the probability of the shooter winning with double-6 dice (the 1 on both dice is replaced with a 6)
            - **http://www.gamblingcollectibles.com/equipment2.html**

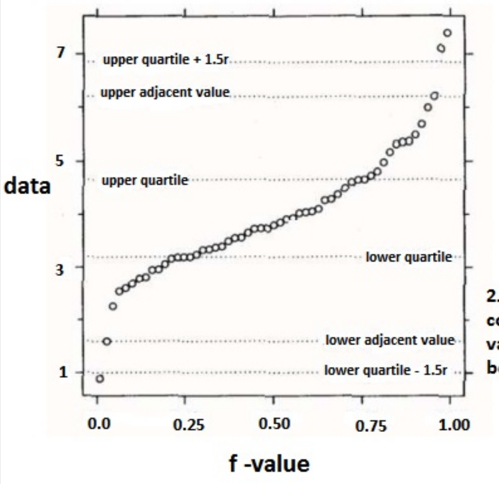
Introduction

Chapter 3 — Random Variables — is the longest (146 p) and most difficult chapter in **Larsen & Marx (2006)**. The first six sections introduce concepts that are the foundation for the rest of the course and that are essential for virtually all statistical analysis. The final 7 sections of chapter 3 are less important for the course and for future statistical analyses, so I'm recommending that we skip all but section 3.9. Section 3.9 deals with concepts fundamental to propagation of error, which is one of the important concepts that students earning graduate degrees in the sciences must learn. Feel free to read and study this material that we are skipping We simply don't have the time to cover everything, and we won't use these concepts in the latter part of the course. I've programmed many of the case studies for sections 3.7-3.13, so you can work through these sections on your own.

I'm an ecologist and this is the key chapter needed for understanding a variety of biodiversity indices, especially the Sanders-Hurlbert rarefaction method.

Boxplots

Boxplot Invented by Tukey and displaying an approximate interquartile range, **median**, range and extreme data points. A box marks the the **interquartile range (IQR)** with lower and upper limits approximately equal to the 1st and 3rd quartiles. Tukey didn't define the boxes in terms of quartiles, but used the term **hinges**, to eliminate ambiguity. Tukey wanted a quick method to graph data and calculating quartiles using cumulative distribution functions isn't quick. Tukey found his hinges by first finding the median. He then found the hinges as the median of the points larger and smaller than the median. This is NOT how quartiles are defined. There are a number of different ways of defining the 1st and 3rd quartiles, which mark the 25th and 75th % of the cumulative frequency distributior



2.7 The diagram illustrates the computation of the adjacent values, which are used in the box plot display method.

Figure 1. SPSS boxplots

Whiskers extend to the adjacent values, which are actual data outside the IQR but within 1.5 IQR's from the median. Points more than 1.5 IQR's from the IQR are outliers. Points more than 3 IQR's from the box are extreme outliers. See also **http://mathworld.wolfram.com/Box-and-WhiskerPlot.html**

Annotated outline (with Matlab programs) for Larsen & Marx Chapter 3

Jakob (Jacques) Bernoulli (1654-1705)

3 **Random Variables**

  3.1 **Introduction**

    3.1.1 Random variables

      3.1.1.1 discrete & continuous

  3.2 Binomial and Hypergeometric probabilities

    3.2.1 The binomial probability distribution

Theorem 3.2.1 *Consider a series of n independent trials, each resulting in one of two possible outcomes, "success" or "failure." Let p=P(success occurs at any given trial) and assume that p remains constant from trial to trial. Then*

$$P(k\ \textit{successes}) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k=0,1,\ldots,n$$

Comment The probability assignment given by the Equation in Theorem 3.2.1 is known as the binomial distribution.

Example 3.2.1

```
% LMex030201_4th.m
% Written by E. Gallagher, Eugene.Gallagher@umb.edu
% Dept. of Environmental, Earth & Ocean Sciences
% Revised 10/19/2010
% Reference: Larsen & Marx (2006) Introduction to Mathematical
% Statistics, 4th edition, page 131.
% Your client is guilty of a crime and a blood sample test indicates
% that the defendant's blood matches that at the scene of the crime.
% In one tenth of 1% of cases these tests are in error. If the
% test is sent out to 6 independent labs, what is the probability that
% at least one lab will make a mistake and conclude that your client is
% innocent?
P=1-binopdf(0,6,0.001)
```

```
% LMex030202_4th.m
% Larsen & Marx Example 3.2.2 Page 132 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th Edition
% Based on LMEx030306_3rd.m
% see LMEx030202_4th.m, p. 137, 3rd edition
% Written by Eugene.Gallagher@umb.edu
% revised 1/11/2011.
% A drug company is testing drugs on 30 patients. It has a rule rejecting a
% drug if fewer than 16 of 30 patients show improvement and accepting the
% drug if 16 or more patients show improvement
n=30;
po=0.5;
pa=0.6;
% What is the probability that a drug that is 60% effective will be
% rejected using the decision rule of 16 patients or more have to improve?
PFR=sum(binopdf(0:15,n,pa))
% or equivalently, use the binomial cumulative distribution function
PFR2=binocdf(15,n,pa)
fprintf(...
'Probability of a false rejection=%6.4f | effectiveness = %3.1f.\n',...
PFR,pa);
% What is the probability that a drug that is 50% effective will be
% accepted for further development using the decision rule of 16 patients
% or more have to improve?
PFA=sum(binopdf(16:30,n,po))
% or equivalently, use the binomial cumulative distribution function
PFA2=1-binocdf(15,n,po)
fprintf(...
'Probability of a false acceptance=%6.4f | effectiveness = %3.1f.\n',...
PFA,po);
```

Example 3.2.3

```
% LMex030203_4th
% Probability of winning a series in 4, 5, 6 or 7 games
% if the probability of winning each game is 0.5
% Nahin's Dueling idiots also solves the problem
% Written by E. Gallagher, revised 10/19/10
% Eugene.Gallagher@umb.edu
% Dept. of Environmental, Earth & Ocean Sciences
% Reference: Larsen & Marx (2006) Introduction to Mathematical
% Statistics, 4th edition, page 133-134.
% Call Matlab's stats toolbox binomial pdf function:
% p4 is the probability of 1 team winning 3 of 3 games * p winning a
% 4th game; the total probability of a series ending is the probabilty
% of the 2nd team winning too, so since p=0.5, just multiply by 2
% p4=binopdf(4,4,0.5)*2 or
p4=binopdf(3,3,0.5)*0.5; p4=2*p4
```

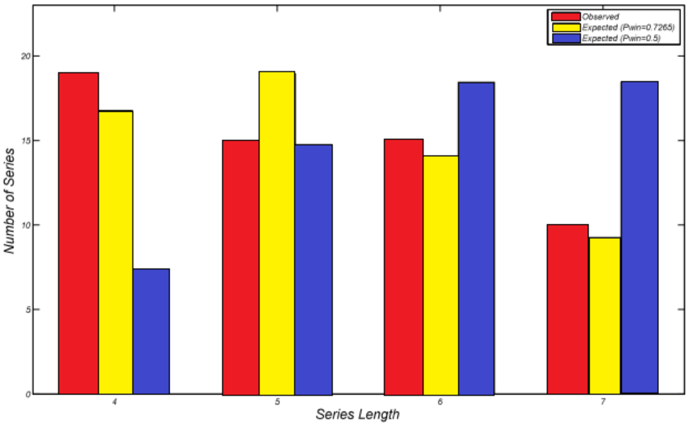


Figure 2. Observed Stanley Cup Series lengths (red , left) with expectations if probability of better team winning is 0.7265 (yellow, middle) or 0.5 (blue , right).

```
p5=binopdf(3,4,0.5)*0.5; p5=2*p5
p6=binopdf(3,5,0.5)*0.5; p6=2*p6
p7=binopdf(3,6,0.5)*0.5; p7=2*p7
% These 4 calculations can be handled in 1 Matlab line
% pdfseries=2*binopdf(3,3:6,0.5)*0.5
pdfseries=binopdf(3,3:6,0.5)
 % The observed and theoretical probabilities are shown in Table 3.2.3
 SL=[4:7]';
 ObservedP=[19/59 15/59 15/59 10/59];
 Table030203=[SL ObservedP' pdfseries'];
 fprintf('Table 3.2.3\n\n')
 fprintf('Series  Observed    Theoretical\n')
 fprintf('Length  Proportion  Probability\n')
 % Note that a loop structure can be avoided by using fprintf with
 % a matrix. The fprintf comman reads entries row-wise, hence the transpose
 fprintf('  %1.0f       %5.3f        %5.3f\n',Table030203')
 fprintf('Series  Observed    Expected\n')
% Optional produce a bar plot, with each bar of width=1;
Bardata=[ObservedP' pdfseries'];
bar(4:7,Bardata,1,'grouped')
% ylim([0 23])
legend('Observed P','Theoretical P (Pwin=0.5)',...
    'Location','NorthEast')
xlabel('Series Length')
ylabel('Proportion of Series')
title('Example 3.2.3');figure(gcf);prnmenu('LM323')
  % This next section uses some advance fitting algorithms to fit the model.
 % Optional fitting algorithms [Not required at all for EEOS601]
 % The observed proportions had more short games, what is the
 % series length if the odds of winning differ from 0.5? More short
 % games would result if one team had a higher probability of winning.
 % This can be solved iteratively by plugging different values into
 % equations and visually examining the output. This was done and 0.75
 % seemed to provide a good fit.
 % Matlab has a suberb set of fitting tools, and fminbnd can be used
 % to find the value of pwin that provides the best fit. There are
 % several different ways of examining the best fit, one is minimizing
 % the sum of squared residuals sum((Observed-Expected).^2). The other,
 % more classic approach, is to minimize the chi-square statistic
 % Chi square = sum((Observed-Expected).^2./Expected). Both approaches
 % will be used with these Stanley cup data.
  pwin=0.75;
 % Can no longer multiply the prob of one team winning by 2
 pdfseries2=binopdf(3,3:6,pwin)*pwin + binopdf(3,3:6,1-pwin)*(1-pwin);
 FittedTable030203=[SL ObservedP' pdfseries2'];
```

```
fprintf('\n\n\nFitted Table if pwin=%5.3f\n\n',pwin)
fprintf('Series  Observed    Theoretical\n')
fprintf('Length  Proportion  Probability\n')
% Note that a loop structure can be avoided by using fprintf with
% a matrix. The fprintf comman reads entries row-wise, hence the transpose
fprintf('  %1.0f       %5.3f        %5.3f\n',FittedTable030203')
% find the optimum value of pwin using Matlab's bounded search algorithm,
% fminbnd
% First, the minimization of squaredvvresiduals will be used, then
% using chi square stat: sum((O-E).^2./E)
% I'm following the help file here and using an implicit function that
% I'll call f that will take as input the observed proportions and pwin
% and return the sum of the squared residuals
f=@(pwin,ObservedP) ...
sum(  ( (binopdf(3,3:6,pwin)*   pwin   + ...
        binopdf(3,3:6,1-pwin)*(1-pwin))-ObservedP).^2);
        % Call the function f to find pwin that minimizes the function
        % with bounds 0.49 <= p <= 1
        pwino = fminbnd(@(pwin) f(pwin,ObservedP),0.49,1);
        % Note that a search, with 0,1 bounds returned 0.272
        % So the bounds should be set to find the p for the best team
        % with an allowance for a tie, with 0.49 as lower bound,
        % fminbind returns 0.728
% Use this 'best' estimate to find the pdf
pdfseries3=binopdf(3,3:6,pwino)*pwino + binopdf(3,3:6,1-pwino)*(1-pwino);
FittedTable=[SL ObservedP' pdfseries3'];
fprintf('\n\n\nFitted Table if pwin=%5.3f\n\n',pwino)
fprintf('Series  Observed    Theoretical\n')
fprintf('Length  Proportion  Probability\n')
fprintf('  %1.0f       %5.3f        %5.3f\n',FittedTable')
 % Now fit the problem minimizing the chi square statistic. Divide
% each square residual by the expected value. Observed & Expected values
% should be in the form of counts, so although unimportant, I'll redefine
% the function to use the observed & expected games (multiply by 59)
f_cs=@(pwin,ObservedP) ...
sum(  ((ObservedP-(binopdf(3,3:6,pwin)*   pwin   + ...
        binopdf(3,3:6,1-pwin)*(1-pwin)))*59).^2./ ...
       (binopdf(3,3:6,pwin)*   pwin   +     ...
        binopdf(3,3:6,1-pwin)*(1-pwin))*59);
        %
        pwino_cs = fminbnd(@(pwin) f_cs(pwin,ObservedP),0.49,1);
pdfseries3=binopdf(3,3:6,pwino_cs)*pwino_cs + ...
           binopdf(3,3:6,1-pwino_cs)*(1-pwino_cs);
       Games=59;
       CS=sum((Games*(ObservedP-pdfseries3)).^2./(Games*pdfseries3));
 % What is the probability of observing such a Chi Square statistic by
 % chance? This is covered in Chapter 10 (page 615 in the text).
 % The degrees of freedom is equal to 1-the number of cells in
 % the data, since if the total number of games played is known
 % and the number that ended in less than 7 games is known,
 % then the number of series taking 7 games can
 % be calculated by difference. Only seriesnumber-1 degrees of freedom
 % are present. Also, an addtional degree of freedom is lost for every
 % parameter fit
 % with the data. In the chi-square fitting routine, I fit the best
 % estimate of the winning percentage, or P=0.7265, so another degree
 % of freedom is lost
 df=4-1-1;
 P = 1- chi2cdf(CS,df);
 FittedTable=[SL Games*ObservedP' Games*pdfseries3'];
 fprintf(...
     '\n\n\n Pwin=%6.4f, Chi Square=%5.2f, df=%1.0f, P=%4.2f\n\n',...
      pwino_cs, CS,df, P)
 fprintf('Series  Observed    Expected\n')
 fprintf('Length  Games       Games\n')
 % Note that a loop structure can be avoided by using fprintf with
 % a matrix. The fprintf comman reads entries row-wise, hence the transpose
 fprintf('  %1.0f    %5.0f         %4.1f\n',FittedTable')
  % What was the chi-square distribution and p value for the initial
 % fit, which assumed that each team was equally likely to win? Note
 % that this fit requires only the loss of a single degree of freedom
 % since the 0.5 did not have to be estimated from the data.
CS2=sum((Games*(ObservedP-pdfseries)).^2./(Games*pdfseries));
df2=4-1;
 P2 = 1- chi2cdf(CS2,df2);
 FittedTable2=[SL Games*ObservedP' Games*pdfseries'];
 fprintf(...
     '\n\n\n Pwin=%6.4f, Chi Square=%5.2f, df=%1.0f, P=%6.2g\n\n',...
      0.5, CS2,df2,P2)
 fprintf('Series  Observed    Expected\n')
 fprintf('Length  Games       Games\n')
 fprintf('  %1.0f    %5.0f        %4.1f\n',FittedTable2')
Bardata=[FittedTable(:,2:3) FittedTable2(:,3)];
bar(4:7,Bardata,1,'grouped')
ylim([0 23])
legend('Observed','Expected (Pwin=0.7265)','Expected (Pwin=0.5)',...
    'Location','BestOutside')
xlabel('Series Length')
ylabel('Number of Series')
title('Example 3.2.3');figure(gcf);prnmenu('LM323')
```

Example 3.2.4

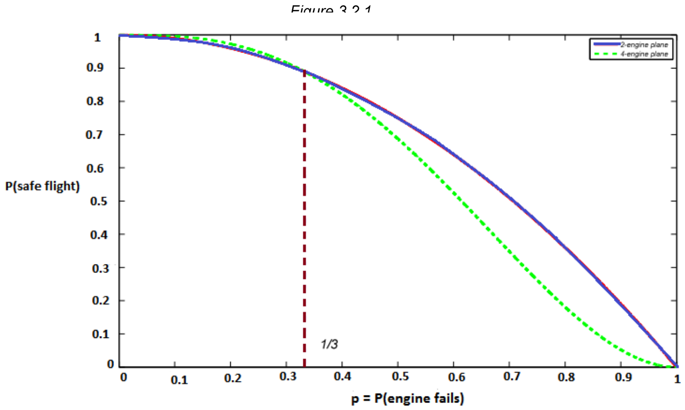


Figure 3. Probability of landing safely in a 2- or 4-engine plane. Choose the 2-engine plane if the probability of engine failure is greater then 1/3.

```
% LMex030204_4th.m or
LMex030308_3rd.m
% Exam 3.2.4, page 135-136 in
% Larsen & Marx (2006)
Introduction to Mathematical
Statistics, 4th edition
% and Larsen & Marx (2001) Third
Edition Example 3.3.8 p. 139
% Written by E. Gallagher
Eugene.Gallagher@umb.edu
% 10/6/2010 revised 1/11/2010
% A major goal is to produce Figure
3.2.1, which will require a large
% number of points
% initialize arrays to store results
% Doomsday airlines has 2 aircraft
--- a dilapidated 2-engine prob plane
% and an equally dilapidated under-maintained 4-enghine prop ploe. Each
% plane will land safely if at least half its engines are working properly.
% Given that you want to remain alive, under what conditions would you opt
% to fly in the two engine plane. Assume that the engines on each plane
% have the same probability of failing and that engine failures are
% independent events.
penginefails=zeros(999,1);
p2enginesafe=zeros(999,1);
p4enginesafe=zeros(999,1);
for i=1:9999
   penginefails(i)=i/10000;
   p2enginesafe(i)=sum(binopdf(1:2,2,1-penginefails(i)));
   p4enginesafe(i)=sum(binopdf(2:4,4,1-penginefails(i)));
end
plot(penginefails,p2enginesafe,'-r',...
   penginefails,p4enginesafe,'--g');figure(gcf)
xlabel('p=P(engine fails)','FontSize',16)
ylabel('P(Safe flight)','FontSize',16)
i=find(p2enginesafe>=p4enginesafe);
mi=min(i);
fprintf('Pick a 2 engine plane if P(enginefails) >= %8.6f \n',...
   penginefails(mi));
% A more precise estimate can be had using equation 3.2.3 in text, p 136
% and Matlab's built-in fzero function
Penginefails = fzero(@(p) (3*p-1)*(p-1), 0.001, 0.999);
fprintf('Now solve the problem more precisely using Matlab''s fzero.m\n')
fprintf('Pick a 2 engine plane if P(enginefails) >= %8.6f \n',...
   Penginefails);
P2enginesafe=sum(binopdf(1:2,2,1-Penginefails));
hold on
% This plot simply puts a vertical line in the graph
plot([Penginefails Penginefails],[0 P2enginesafe],'-.b');
legend('2-engine plane','4-engine plane','Location','NorthEast')
text(0.36,0.07,'1/3','FontSize',18)
title('Figure 3.2.1','FontSize',20)
figure(gcf)
hold off
```

Questions page 136-138

3.2.2 Nuclear power rods

3.2.2 The hypergeometric Distribution

Theorem 3.2.2 *Suppose an urn contains r red chips and w white chips, where r + w = N. If n chips are drawn out at random, without replacement, and if k denotes the number of red chips selected, then*

$$P(k \ red \ chips \ are \ chosen) = \frac{\binom{r}{k}\binom{w}{n-k}}{\binom{N}{n}}. \tag{3.2.4}$$

where k varies over all integers for which $\binom{\mathrm{r}}{\mathrm{k}}$ *and* $\binom{\mathrm{w}}{\mathrm{n}-k}$ *are defined. The probabilities appearing on the right-hand size of Equation 3.2.4 are known as* the hypergeometric distribution.

Example 3.2.5 Keno

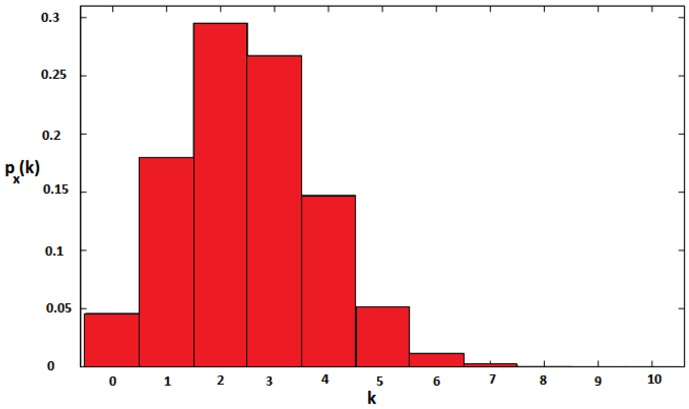


Figure 4. Hypergeometric probability distribution for a pick-10 ticket.

```
% LMex030205_4th
% Example 3.2.5 Winning at KENO with a ten-spot ticket
% Larsen & Marx (2001) 3rd edition page 130
% Larsen & Marx (2006) 4th edition page 141
% if the statitics toolbox isn't loaded with hygepdf
% calls Gallagher's m.file hypergeop.m
% N=80 numbers
% n=10 numbers selected
% r=20 number of winning numbers out of 80
% k=5  pk5 is the probability of getting
%             5 of the winning numbers
% written by E Gallagher, Eugene.Gallagher@umb.edu
% written 6/20/03; last revised 10/7/10
N=80;n=10;r=20;k=5;
if exist('hygepdf')==2
   pk5=hygepdf(k,N,r,n)
elseif  exist('hypergeop')==2
   % Gallagher's function, which calls Matlab's nchoosek.m for small N
   pk5=hypergeop(N,n,r,k)
end
fprintf('Probability of picking 5 correct = %6.4f\n',pk5);

% Optional, find the probability of winning with a 5, 6, 7 ... 10
% and plot the pdf
pkge5=sum(hygepdf(5:(min([r n])),N,n,r));
fprintf(...
   'Probability of picking 5 correct = %6.4f and 5 or more = %6.4f\n',...
   pk5,pkge5);

% Optional: plot the the full hypergeometric probability density function
% for the ten-spot ticket
p=hygepdf(0:(min([r n])),N,n,r);
bar(0:10,p,1)
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
axis([-.6 10.6 0 .31])
ylabel('p_x(k)','FontSize',16),xlabel('k','FontSize',16),
title('Example 3.2.5','FontSize',20)
figure(gcf)
```

```
pause
```

Example 3.2.6, p. 142 Hung Jury

```
% LMex030206_4th.m
% Larsen & Marx (2006) 4th edition Example 3.2.6, page 142 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% P = probability of a hung jury
% Given that 2 of the 25 jurors in the jury pool would vote 'not guilty'
% and 23 of 25 would vote guilty
% What is the probability of observing a hung jury?
% Written by Eugene.Gallagher@umb.edu, written in 2001, revised 1/12/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Can be solved in a number of mathematically identical one-liners.
P=hygepdf(1,25,2,12)+hygepdf(2,25,2,12)
% or combining in one matlab statement.
P=sum(hygepdf(1:2,25,2,12))
% or using the complement of the probability that no jurors would vote
% guilty:
P=1-hygepdf(0,25,2,12);
```

Example 3.2.7 p. 142-143 Bullet Striations

```
% LMex030207_4th.m
% Example 3.2.7, Bullet striations on page 142
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu, written in 2001, revised 1/12/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Solution to bullet striation pattern problem
% will call Matlab statistics toolbox hypergeometric probability
% distribution function or Gallagher's hypergeop.m
% The problem will also be solved using a Monte Carlo simulation.
% There were 4 striations in 25 locations in the evidence bullet
% A test bullet from the suspect's gun was examined and 1 striation out of
% 3 on the test bullet matched the striations on the evidence bullet.
% What is the probability that 1 or or more of the test bullet striations
% would match the test bullet
if exist('hygepdf','file')
   P=sum(hygepdf(1:3,25,3,4));
elseif exist('hypergeop','file')
   P=sum(hypergeop(25,3,4,1:3));
else
   disp('No hypergeometric function on path')
end
fprintf(...
'Probability of 1, 2 or 3 matches from hypergeometric distribution =%7.5f.\n',...
  P)
% According to the book, small p values indicate that the evidence
% gun and the test gun were the same.
% What if 2 of the 3 patterns from the test bullet matched the
% evidence bullet?
if exist('hygepdf','file')
  NewP=sum(hygepdf(2:3,25,3,4));
elseif exist('hypergeop','file')
  NewP=sum(hypergeop(25,3,4,2:3));
else
   disp('No hypergeometric function on path')
end
fprintf(...
'Probability of  2 or 3 matches from hypergeometric distribution =%7.5f.\n',...
  NewP)
% What if all 3 of the 3 patterns from the test bullet matched the
% evidence bullet?
if exist('hygepdf','file')
  NewP=sum(hygepdf(3,25,3,4));
elseif exist('hypergeop','file')
  NewP=sum(hypergeop(25,3,4,3));
else
   disp('No hypergeometric function on path')
end
fprintf(...
'Probability of 3 matches from hypergeometric distribution =%7.5f.\n',...
  NewP)
% Advanced problem:
% This problem is amenable to a Monte Carlo simulation
EB=[ones(1,4) zeros(1,21)]; % evidence bullet with 4 striations
TB=[1 0 0 0 1 1 zeros(1,19)];% test bullet with 3 striations, 1 overlapping
trials=9999;
matches=sum((EB+TB)==2);
teststat=sum(hygepdf(matches:3,25,3,4));
results=zeros(trials,1);
for i=1:trials
   % just need to shuffle one of the bullets, just the evidence bullet
   EBshuffled=EB(randperm(length(EB)));
   %    results(i)=sum((EBshuffled+TB)==2); % Probability of 1 match
   results(i)=sum((EBshuffled+TB)>=2);  % What is the probability of 1 OR
                                   % MORE matches
end
i=find(results>=matches);MCP=(length(i)+1)/trials;
fprintf(...
 'P from hypergeometric =%7.5f & from Monte Carlo = %6.4f +/- %6.4f\n',...
  P, MCP,norminv(0.975)*sqrt(P*(1-P))/sqrt(trials))
% Note that if there HAD been more than 1 match, the spacing of the
% striations would play a key role. The random permutations would have
% to be performed taking into account the spatial autocorrelation of
% striations on the bullets. ter Braak discusses how to shuffle in the
% presence of spatial autocorrelation.
```

Example 3.2.8, p. 144

```
% LMex030208_4th.m
% Larsen & Marx (2006) 4th Edition, p 144
% Wipe your feet carpet cleaning problem
% Written by Eugene Gallagher, Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Written, 10/7/10, revised 1/12/11
% What is the probability of contacting 100 or more customers if
% 1000 phone calls are made in a community of 60000 which has
% 5000 potential customers
if exist('hygepdf','file')
   P=1-sum(hygepdf(0:99,60000,1000,5000))
elseif exist('LMTheorem030202_4th','file')
   P=1-sum(LMTheorem030202_4th(1000,0:99,5000,60000))
else
   disp('No hypergeometric function on path')
end
% can be solved using the cumulative hypergeometric probability
% distribution function as well; The cumulative distribution function
% is introduced in Definition 3.4.3 and Theorem 3.4.1
p = 1-hygecdf(99,60000,5000,1000)
```

Case Study 3.2.1 Apple hardness

```
% LMcs030201_4th.m
% Larsen & Marx (2006) 4th edition p. 145
% Written by Eugene.Gallagher@umb.edu, written 2001, revised 1/12/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% uses the statistics toolbox m.file hygepdf, the hypergeometric pdf
% If there were 10/144 (6.9%) defective apples in the shipment,
% A shipment is accepted if 2 or fewer apples in 15 are
% found to have a firmness less than 12 lbs. What is the probability
% that a 6.9% defective rate box would be deemed acceptable?
if exist('hygepdf','file')
   P=sum(hygepdf(0:2,144,15,10));
```

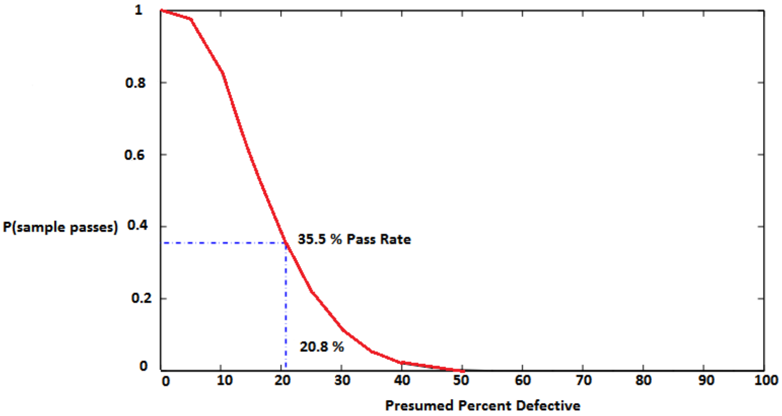


Figure 5. If 30 out of 144 apples (20.8%) are defective and the shipper ships if 2 or fewer out of 15 tested are defective, then the probability that the sample passes is 35.5%. Note that Figure 3.2.5 in **Larsen & Marx (2006, p. 146)** is wrong.

```
elseif exist ('LMtheorem030202_4th','file')
   P=sum(LMtheorem030202_4th(15,0:2,10,144));
else
   disp('Hypergeometric functions not on Matlab path.')
end
fprintf('If defective rate is %4.1f%%, acceptance rate is %4.1f%%.\n', ...
   10/144*100,P*100);
% If the defective rate was 30 of 144 or 21%.
if exist('hygepdf','file')
   P=sum(hygepdf(0:2,144,15,30));
elseif exist ('hypergeop','file')
   P=sum(hypergeop(144,15,30,0:2));
else
   disp('Hypergeometric functions not on Matlab path.')
end
fprintf('If defective rate is %4.1f%%, acceptance rate is %4.1f%%.\n', ...
   30/144*100,P*100);
% Produce the operating characteristic curve, Figure 3.2.5
PPD=0:.05:1;
DefectiveApples=round(PPD*144);
lengthApples=length(DefectiveApples);
OCP=zeros(lengthApples,1);
for i=1:lengthApples
   SoftApples=DefectiveApples(i);
   if exist('hygepdf','file')
      P=sum(hygepdf(0:2,144,15,SoftApples));
   elseif exist ('LMtheorem030202_4th','file')
      P=sum(LMtheorem030202_4th(15,0:2,SoftApples,144));
   else
      disp('Hypergeometric functions not on Matlab path.')
   end
   OCP(i)=P;
end
plot(PPD*100,OCP,'.k',PPD*100,OCP,'-k','LineWidth',1.5);
ax1=gca;
ax1=gca; % save the handle for the axes of the graph
ylabel('P(sample passes)','Color','k','FontSize',16)
set(ax1,'Ytick',[0:0.2:1],'YColor','k','FontSize',12)
title('Figure 3.2.5','FontSize',20)
set(ax1,'Xtick',[0:10:100],'XColor','k','FontSize',12)
xlabel('Presumed percent defective','FontSize',16);
hold on
% This plot simply puts a vertical line on the graph
plot([20.8 20.8],[0 .355],'-.b','LineWidth',2);
% This plot simply puts a horizontal line on the graph
```

```
plot([0 20.8],[.355 .355],'-.b','LineWidth',2);
text(22,0.08,'20.8%','FontSize',17)
text(22,0.355,'35.5% Pass Rate','FontSize',17)
title('Figure 3.2.5','FontSize',20)
figure(gcf)
hold off
```

Questions 3.2.18

3.2.18 Corporate board membership

3.2.22 Anne's history exam

3.3 **DISCRETE RANDOM VARIABLES** (p 148)

3.3.1 Assigning probabilities: the discrete case

3.3.1.1 discrete probability function

Definition 3.3.1 (p. 149) Suppose that *S* is a finite or countably infinite sample space. Let p be a real-valued function defined for each element of *S* such that

a. $0 \leq p(s)$ for each $s \in S$

b. $\sum_{for\ all\ s\ \in\ S} p(s) = 1$

Then p is said to be a **discrete probability function**.

Example 3.3.1 Ace-six loaded dice

```
% LMex030301_4th.m
% Example 3.3.1 Loaded Ace-six flats dice
% Ace-six flats dice are shortened so that
% p(1)=p(6)=1/4 and p(2)=p(3)=p(4)=p(5)=1/8;
% What is the probability of getting a 7 rolling fair dice?
% What is the probability of rolling a 7 using ace-6 dice?
% Written by Eugene.Gallagher@umb.edu
% Calls crapsmof65_ace6.m, which solves craps probabilities for fair and
% ace-six flats dice
% Written 1/12/11 for EEOS601 by Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
number=7;
[Pexact,Pa6exact,Pnumber,Pa6number]=crapsmof65_ace6(7);
fprintf('The probability of rolling a %2.0f with fair dice=%6.4f or\n',number,Pnumber)
format rat
disp(Pnumber)
format
fprintf('The probability of rolling a %2.0f with loaded ace-6 dice=%6.4f
or\n',number,Pa6number)
format rat
disp(Pa6number)
format
```

```
function [Pexact,Pa6exact,Pnumber,Pa6number]=crapsmof65_ace6(number, H)
% format [Pexact,Pa6exact,Pnumber,Pa6number]=crapsmof65_ace6(number, H)
% input: number=value of 1st roll
%        H    ='H' for Pnumber & odds of making 4, 6, 8, or 10 the hard way
% output: Pexact= Exact probability of winning, given 1st roll = number
%         Pnumber=Probability of rolling that number by rolling 2 fair dice.
%         Pa6=Probability of rolling that number with loaded ace 6 flats.
% Using the 'method of reduced sample space' from Mosteller, F. 1965. Fifty
%   challenging problems in probability with solutions, p. 9
% see craps.m for a more laborious solution using absorbing Markov chains
% written 6/2/03 by E. Gallagher for EEOS601, revised 1/12/11;
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% See Larsen & Marx Example 3.3.6 Ace-six flats are foreshortened so that
% the p(1)=p(6)=1/4 & p(2)=p(3)=p(4)=p(5)=1/8;
if nargin > 1
   if H=='H'
      if mod(number,2) | number ==2 | number==12
         error(...
           'Can only make even numbers between 4 & 10 the "hard" way')
         return
      end
   end
end
dief=1:6; % create a sample space of two dice rolls; while simultaneously
       % creating a matrix of probabilities corresponding to the faces
diep=[1/4 repmat(1/8,1,4) 1/4];
die1f=repmat(dief,6,1); % create a matrix, with each row equal to 1 to 6;
die1p=repmat(diep,6,1); % create a matrix, with each row equal to 1 to 6;
die2f=repmat(dief',1,6); % create a matrix, with each column equal to 1 to 6;
die2p=repmat(diep',1,6); % create a matrix, with each column equal to 1 to 6;
sumdicef=die1f+die2f;  % This gives the sample space, the 36 sums of two die.
sumdicep=die1p.*die2p;  % This gives the sample space,the 36 elementwise
                  % products two die.
i=find(sumdicef==number);  % find the indices for each value equal to number;
lengthi=length(i);     % find how many sums of 2 die equal the number in input;
if lengthi==0        % only true if the number isn't in the sample space.
   disp('Your number isn"t possible, not in the sample space')
   number
   % The following odd syntax will print the number
   fprintf('Your number, %6.0f, isn"t possible, not in the sample space\n',number)
   return % This stops the program after displaying the message;
end
[r,c]=size(sumdicef); % find size of sample space
Pnumber=lengthi/(r*c); % What is the probability of rolling your number with 1 roll
                 % of two dice
Pa6number=sum(sumdicep(i)); % sum the probabilities corresponding to the summed faces
% Is the craps shooter an immediate winner?
```

```
if (number==7 | number==11) % 2 logical statements using the | OR command.
   Pexact=1;            % Matlab performs statements in this if section only
                      % if the logical statement number equals 7 or equals 11
                      % is true.
   Pa6exact=1;
   return
end
% Is the craps shooter an immediate loser?
if (number==2 | number==3 | number==12)
   Pexact=0;
   Pa6exact=0;
   return  % exit the function and return to the command window.
end
% Exact probability of making a point - no need for sum of geometric series
% or Markov chains
% Use Mosteller's 'Method of reduced sample space:
% all we need is Probability of rolling a given
% number - which wins - and the probability of rolling a 7, which loses.
%  We don't need any of the other probabilities
% (e.g., probability of rolling 2, 3 ...)
%  First find probability of rolling a 7, which results in a loss.
j=find(sumdicef==7);P7=length(j)/(r*c); % P7 is the probability of rolling 7
Pa6_7=sum(sumdicep(j));
if nargin>1 & H=='H'
   ih=find(sumdicef==number & die1f==number/2);
   lengthih=length(ih);
   Pnumberh=lengthih/(r*c); % Another way to lose is to roll a soft number
   Pnumbera6h=sum(sumdicep(ih))
   Pexact=Pnumberh/(P7+Pnumber); % You win or lose if you roll your hard
                           % number (4,6,8)
   Pa6exact=Pnumbera6h/(Pa6_7+Pa6number);
else
   Pexact=Pnumber/(Pnumber+P7);
   Pa6exact=Pa6number/(Pa6_7+Pa6number);
end
```

Example 3.3.2

```
% LMex030302_4th.m
% Exmaple 3.3.2 p 150-151 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Example of Matlab's symbolic summation
% Suppose a fair coin it tossed until a head comes up for the first time.
% What are the chances of that happeing on an odd numbered toss?
% Written 1/12/11 for EEOS601 by Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
syms s
```

```
symsum(.5^s,1,inf)
% symsum(p^(2*s+1),0,inf)  % Typo in book, can't be evaluated without
% constraining p to values between 0 and 1
symsum(.5^(2*s+1),0,inf)
.5*symsum(.25^s,0,inf)
```

Case Study 3.3.1 London deaths

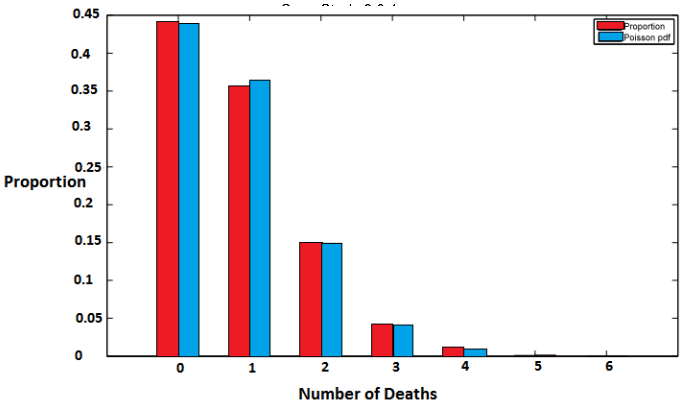


Figure 6. Observed proportions of number of deaths vs. Poisson model predictions. The fit of the Poisson model is excellent.

```
% LMcs030301_4th.m
% Case Study 3.3.1 Page 151-152 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics 4th Edition
% Fit of the Poisson distribution to Deaths.
% Written by Eugene.Gallagher@umb.edu

NumberDead=[0:6]';
NumberDays=[484 391 164 45 11 1 0]';
Proportion=NumberDays/sum(NumberDays)';
lambda=sum(NumberDead.*Proportion);
% See Larsen & Marx (2006) p. 281 on the Poisson distribution
fprintf('The Poisson parameter is %5.3f\n',lambda)
ps = poisspdf(0:1000,lambda)';
plus6=sum(ps(7:end));
ps(7)=plus6;ps(8:end)=[];
NumberDeaths=['0 ';'1 ';'2 ';'3 ';'4 ';'5 ';'6+'];
disp('Deaths Days   P     p(s)')
for i=1:7
fprintf('%3s    %3.0f  %5.3f  %5.3f\n',NumberDeaths(i,:),NumberDays(i), ...
   Proportion(i), ps(i))
end
bar(0:6,[Proportion ps],1,'grouped')
legend('Proportion','Poisson pdf','Location','Best')
xlabel('Number of Deaths','FontSize',16)
ylabel('Proportion','FontSize',16)
title('Case Study 3.3.1','FontSize',20);figure(gcf);
```

Example 3.3.3 Benford’s law

```
% LMcs030303_4th.m
% Case Study 3.3.3 Benford's Law; Page 152-153 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics 4th Edition
% Fit of the Poisson distribution to deaths.
% Written 1/12/11, Revised 1/12/11 for EEOS601 by Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
```

```
s=1:9;
BenfordP=log10(1+1./s);
P=repmat(1/9,1,9);
disp('           "Uniform" Benford"s')
disp('     s      Law      Law ')
disp([s;P;BenfordP]');
```

Example 3.3.4

3.3.2 Defining "New" Sample Spaces

Definition 3.3.2 A function whose domain is a sample space S and whose values form a finite or countably infinite set of real numbers is called a **discrete random variable**. We denote random variables by upper case letters, often X or Y.

Example 3.3.5 - Not programmed.

3.3.3 **The Probability Density Function p. 155**

Definition 3.3.3 (p 155) Associated with every discrete random variable X is a **probability density function (or pdf)** denoted $p_x(k)$, where

$$p_x(k) = P(\{ s \in S| X(s) = k \})$$

Note that $p_x(k) = 0$ for any k not in the range of Xx. For notational simplicity, ... delete all references to s and S and write $p_x(k) = P(\{X = k \})$.

Example 3.3.6 Rolling two dice

```
% LMex030306_4th.m
% Creating a discrete pdf from the sum of two dice
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Example 3.3.6, 4th edition page 155
% Written by Eugene.Gallagher@umb.edu, 10/7/2010, revised 1/12/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Find pdf for two dice by full enumeration
% See also crapsmof65.m, crapstable.m
die=1:6;
twodice=repmat(die,6,1)+repmat(die',1,6);
twodice=twodice(:); % convert to a column vector.
mink=min(twodice);maxk=max(twodice);
k=[mink:maxk]';lenk=length(k);ltwod=length(twodice);
% Simple problem, find P(X=5)
fprintf('The probability of rolling a 5 with 2 dice is:\n')
format rat
disp(sum(twodice==5)/ltwod);
fprintf('The entire discrete pdf in rational format:\n\n')
fprintf('     k          p_x(k)\n')
pxk=sum(repmat(twodice,1,lenk)==repmat(k',ltwod,1))'/ltwod;
disp([k pxk])
format
```

```
bar(k,pxk,1);xlabel('Sum of 2 dice'); ylabel('p_x(k)');
title('Example 3.3.6'),figure(gcf)
fprintf('The entire discrete pdf in numeric format:\n\n')
fprintf(' k  p_x(k)\n');
for i=1:lenk
   fprintf('%2.0f  %5.4f\n',k(i),pxk(i));
end
```

Example 3.3.7 % Profit and loss probability density function.

```
% LMex030307_4th.m
% Example 3.3.7
% Acme Industries builds 3 generators daily. Probability of a generator
% failing and needing to be retooled is 0.05. If a generator is shipped,
% the company earns $10,000, if it is retooled, the cost is $2000
% p_x(k) = the company's daily profit;
% Written 1/12/11 for EEOS601 by Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
n=3;
pfail=0.05;ppass=1-pfail;
profit=1e4;
retool=-2e3;
k=0:3;
p_xk=binopdf(k,3,ppass);
Profit=zeros(1,4); % initialize the results vector
for defects=0:3
   Profit(defects+1)=defects*retool+(3-defects)*profit;
end
disp('       Table 3.3.4')
disp('  No.')
disp('Defects  k=Profit     p_x(k)');
Out=[0:3;Profit;fliplr(p_xk)];
fprintf('%1.0f        $%5.0f
%8.6f\n',Out)
```

Example 3.3.8 Rhonda's basketball shots

```
% LMex030308_4th.m
% LMex030202 in the 3rd edition
% Example 3.2.2 Basketball
problem.
% Page 157 in
% Larsen & Marx (2006)
Introduction to Mathematical
Statistics, 4th edition
% Rhonda has a 65% probability of
making a foul shot and two baskets
are
```

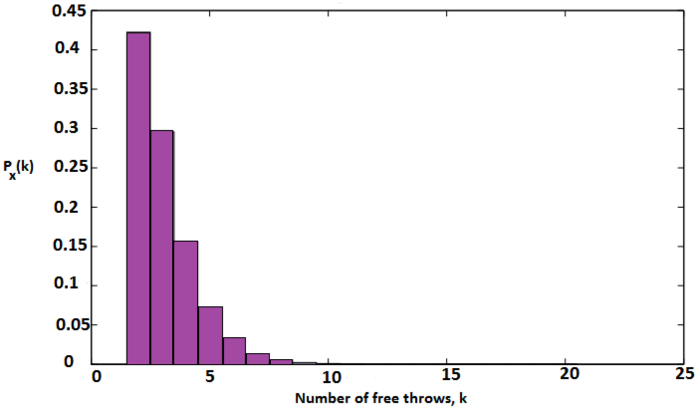


Figure 7. Discrete pdf for Rhonda hitting two free throws. The expected number of free throws is 3.077 if p=0.65 for an individual free throw.

```
% required to win
% Solve using absorbing Markov chain & by the method
% in Larsen & Marx, p. 157
% The 3 states for the Markov chain are 0 baskets made,
% 1 basket, 2 baskets.
% Solution using Larsen & Marx's methods followed by solution
% using an absorbing Markov Chain.
% Solve for the discrete pdf for up to 20 free throws.
% Need to find p_xk = P (drill ends on the kth throw)
k=2:20;
lk=length(k);
% Initialize two vectors to store results
p_xk=zeros(lk,1);  % Probability that the drill ends on the kth throw
Ek=zeros(lk,1);    % Expected values
for i=1:lk
   p_xk(i)=(k(i)-1)*(0.35).^(k(i)-2)*(0.65).^2; % Equation 3.3.4 in text
   Ek(i)=p_xk(i)*k(i);  % Expected value introduced on p 173-175
end
Ek=sum(Ek);  % Expected value, L&M 3rd edition p. 192, 4th ed p 157-158
            % this will equal tau(1) from
            % the absorbing Markov chain described in the Advanced section
disp('The probability density function:')
disp('       k      p_x(k)')
disp([k(1:12)' p_xk(1:12)])
bar(k,p_xk,1);xlabel('Number of free throws,k','FontSize',16);
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
ylabel('p_x(k)','FontSize',16),title('Example 3.3.8','FontSize',20)
figure(gcf)
disp('The expected number of free throws for Rhonda to win:')
disp(Ek)
disp('Sum the final states');
i=find(k>=8);
j=find(k==8);
p_xk(j)=sum(p_xk(i));
k=find(k>8);p_xk(k)=[];
kstring=['2  ';'3  ';'4  ';'5  ';'6  ';'7  ';'8+ '];
disp('Table 3.3.5')
disp([kstring num2str(p_xk)])
% Solution by absorbing Markov chains
% A very advanced topic: not needed for EEOS601
% Solution using absorbing Markov Chain theory, with absorb.m
% based on Kemeny & Snell (1976) see documentation in absorb.m
% The transition matrix P, with the 3 states, Made 2 is
% an absorbing state.
%         Made 0   Made 1  Made 2
% Made 0    0.35    0.65      0
% Made 1      0     0.35    0.65
% Made 2      0       0       1
P=[.35 .65 0
   0 .35 .65
   0   0   1]
tau=absorb(P);  % Tau is  the expected no. of throws to win
             % In Markov chain terminology, it is the amount of time
             % spent in transient states before being absorbed. Winning
             % is the sole absorbing state of this model.
fprintf('The number of free throws required to get 2 is %6.4f \n',tau(1))
% New game: how long until Rhonda gets 2 in a row?
% Let's use the absorbing Markov chain to find the expected number of free
% throws if winning requires 2 made throws in a row:
% Solution using absorbing Markov Chain theory, with absorb.m
% based on Kemeny & Snell (1976) see documentation in absorb.m
% The transition matrix P, with the 3 states, Made 2 is
% an absorbing state.
% P=
%        Made 0   Made 1  Made 2
% Made 0    0.35    0.65      0
% Made 1    0.35      0     0.65
% Made 2      0       0       1
P=[.35 .65 0
  .35  0 .65
   0   0   1]
tau=absorb(P);  % Tau is  the expected no. of throws to win
fprintf(...
   'The number of free throws required to get 2 in a row is %6.4f \n',...
   tau(1))
```

```
function  [tau,N,B,CP,Q,R,tau2,N2,H,MR,VR]=absorb(P);
% format: [tau,N,B,CP,Q,R,tau2,N2,H,MR,VR]=absorb(P);
% input: P an absorbing Markov transition matrix in 'from rows
%        to column' form.  Rows are probability vectors and
%        must sum to 1.0.  At least one main diagonal element=1
% output: tau: Time before absorption from a transient state.
%        B: P(ending up in absorbing state from a transient state)
%        tau2: variance matrix for tau;N: fundamental matrix
%        N2: covariance matrix for N;
%        CP: transition matrix in canonical form.
%        Q: Submatrix of transient states, R: from transient to
%        absorbing state submatrix,
%        H: hij the probability process will ever go to transient
%          state j starting in transient state i (not counting the
%          initial state (K & S, p. 61))
%       MR: expected number of times that a Markov process remains
%           in a transient state once entered (including entering step)
%       VR: Variance for MR (K & S, Theorem 3.5.6, p. 61)
% written by E. Gallagher, ECOS
% UMASS/Boston email: Eugene.Gallagher@umb.edu
% written 9/26/93, revised: 3/26/09
% refs:
% Kemeny, J. G. and J. L. Snell.  1976.  Finite Markov chains.
%       Springer-Verlag, New York, New York, U.S.A.
% Roberts, F. 1976.  Discrete Mathematical Models with applications
%       to social, biological, and environmental problems. Prentice-Hall
%
[pdim,pdim]=size(P);
if sum([sum(P')-ones(1,pdim)].^2)>0.001
     error('Rows of P must sum to 1.0')
end
dP=diag(P);
ri=find(dP==1.0);
if isempty(ri)
      error('No absorbing states (reenter P or use ergodic.m)')
end
rdim=length(ri);
qi=find(dP~=1.0);qdim=length(qi);
I=eye(qdim,qdim);
Q=P(qi,qi);
N=inv(I-Q);   % the fundamental matrix
tau=sum(N')';
CP=P([ri' qi'],[ri' qi']);  % the canonical form of P
R=CP(rdim+1:pdim,1:rdim);
B=N*R;
if nargout>6         % only if variances requested
    Ndg=diag(diag(N));
    N2=N*(2*Ndg-I)-N.^2;
    tau2=(2*N-I)*tau-tau.^2;
    H=(N-I)/Ndg;
    dQ=diag(Q);oneq=ones(qdim,1);
    MR=oneq./(oneq-dQ);
    VR=dQ./(oneq-dQ).^2;
end
```

3.3.4 Transformations

Theorem 3.3.1 See p 158

Example 3.3.9 ∅

3.3.5 The cumulative distribution function

Definition 3.3.4 Let *X* be a discrete random variable. For any real number *t*, the probability that *X* takes on a value $\leq t$ is the cumulative distribution function (cdf) of *X* (written $F_X(t)$). In formal notation, $F_X(t) = P(\{ s \in S \mid X(s) \leq t \})$. As was the case with pdfs, references to *s* and *S* are typically deleted and the cdf is written $F_X(t) = P(X \leq t)$.

Example 3.3.10 A comparison of pdfs and cumulative distribution functions

```
% LMex030310_4th.m
% Example 3.3.10 Page 159 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th Edition
% Application of binomial cdf
% Written by E. Gallagher, Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Written 10/7/2010, revised 1/12/10
% This problem is a simple demonstration that the sum of the elements of
% a discrete binomial pdf can also be computed as the difference of two
% cumulative distribution functions. This applies to any of the
% discrete pdf's.
P1=sum(binopdf(21:40,50,0.6));
fprintf('        P using sum of binomial pdfs is %6.5f\n',P1)
P2=binocdf(40,50,0.6)-binocdf(20,50,0.6);
fprintf('P using difference of cumulative pdfs is %6.5f\n',P2)
% Note numerical calculations using different algorithms
% aren't exactly equivalent, P1-P2 is equal to
% -5.55e-016
% With numerical processors, non-integer numbers can not be represented
% exactly. The precision, called eps by Matlab, is the numeric procession
% with the computer and processor on which Matlab being run. If the
% difference is less than eps, two variables should be regarded as
% identical. On my processor, eps is 2.2204e-016, so the difference above
% while trivial is not less than the numeric precision of the processor
% It is unusual for Matlab to not produce exactly equivalent results, or
% results within eps, but anything this close should be regarded as
% identical
fprintf('The difference should be 0, but is %5.3g, which is > eps=%5.3g.\n',...
   P2-P1,eps)
```

Example 3.3.11

```
% LMex030311_4th.m
% Larsen & Marx (2006) p 160
% Find F_xt for random variable X, defined as the larger of
% two dice.
% Written by Eugene Gallagher, Eugene.Gallagher@umb.edu
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Written 10/7/10, revised 1/12/10
% Brute force enumeration, but note that the book provides
% an intuitive solution;
% Nothing fancy: just creating a 36-outcome sample sapce
die=1:6;
SampleSpace=max(repmat(die,6,1),repmat(die',1,6));
SampleSpace=SampleSpace(:); % creates a 36-outcome vector
lSS=length(SampleSpace);
% creates the cumulative distribution function F_x(t)
Fxt=sum(repmat(SampleSpace,1,6)<=repmat(die,lSS,1))/lSS
% Graph
t=0:0.01:6;
 % need to use equation for Fxt from page 160 in text
 % rounding down to nearest integer with floor will create
 % stepped cdf.
Fxt=floor(t).^2./36;
plot(t,Fxt);ylabel('F_X(t)'),xlabel('t');title('Example 3.3.11');
figure(gcf);
```

Questions p 160

3.3.4 A fair die is tossed 3 times, X is the number of different faces. Find px(k). [Assigned Fall 2010]

3.4 CONTINUOUS RANDOM VARIABLES

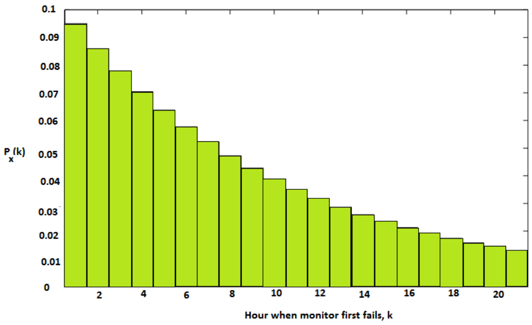


Figure 8.

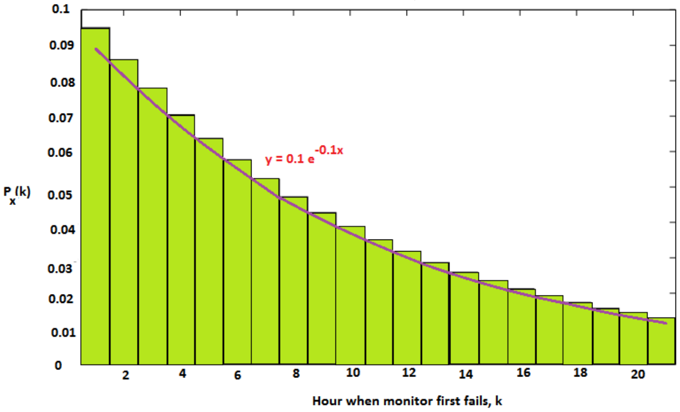


Figure 9.

Definition 3.4.1 A **probability function** *P* on a set of real numbers *S* is called **continuous** if there exists a function *f(t)* such that for any closed interval $[a,b] \subset S$, $P([a, b]) = \int_a^b f(t)\ dt$.

Comment If a probability function *P* satisfies **Definition 3.4.1**, then $P(A) = \int_a f(t)\ dt$ for any set *A* where the integral is defined.

Conversely, suppose a function *f(t)* has the two properties

1. $f(t) \geq 0$ for all *t*.
2. $\int_{-\infty}^{\infty} f(t)\ dt = 1.$

If $P(A) = \int_a f(t)\ dt$ for all A, then P will satisfy the probability axioms.

3.4.1 Choosing the function *f(t)*

Example 3.4.1 A simple application of symbolic math integration

```
% LMEx030401_4th
% Larsen & Marx (2006), p. 163-164
% Written by Eugene.Gallagher@umb.edu for EEOS601
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Revised 1/12/2011
syms t
% These two symbolic math integrations solve for two definite integrals
PA=int (1/10,t, 1, 3)
PB=int (1/10,t, 6, 8)
% Are these two definite integrals identical? Simplify will take the
% difference of these two definite integrals and the logical operator ==
% will determine if that symbolic difference is zero.
if simplify(PA-PB)==0
   disp('The definite integrals are equal.');
end
```

Example 3.4.2 p 164

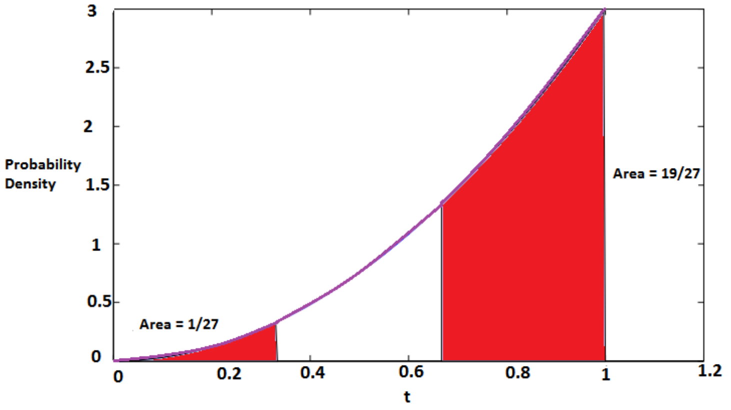


Figure 10.

```
% LMEx030402_4th
% Example 3.4.2 p 164 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu for EEOS601
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Revised 1/12/2011
syms t
% Could f(t)=3t^2 for 0 <= t <= 1 be used to define a continuous pdf?
% Does the definite integral sum to 1?
P=int(3*t^2,0,1)
fprintf('The definite integral of 3*t^2 from 0 to 1 is %2.1f.\',eval(P))
P1=int(3*t^2,0,1/3)
P2=int(3*t^2,2/3,1)
T=0:0.01:1;
Prob=3*T.^2;
plot(T,Prob);xlabel('t','FontSize',16);
ylabel('Probability density','FontSize',16)
axis([0 1.2 0 3]);
% Now fill in the areas
hold on;
T1=0:.01:1/3;yf=3*T1.^2;
fill([0 T1 1/3]',[0 yf 0]',[.8 .8 1])
text(.04,.3,'Area = 1/27','FontSize',18)
T2=2/3:.01:1;yf2=3*T2.^2;
fill([2/3 T2 1]',[0 yf2 0]',[.8 .8 1])
text(1.02,1.5,'Area = 19/27','FontSize',18)
title('Figure 3.4.4','FontSize',20)
figure(gcf);pause
hold off;
```

Example 3.4.3

```
% LMEx030403_4th
% Example 3.4.3 Plotting 3 normal curves. p 164 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
```

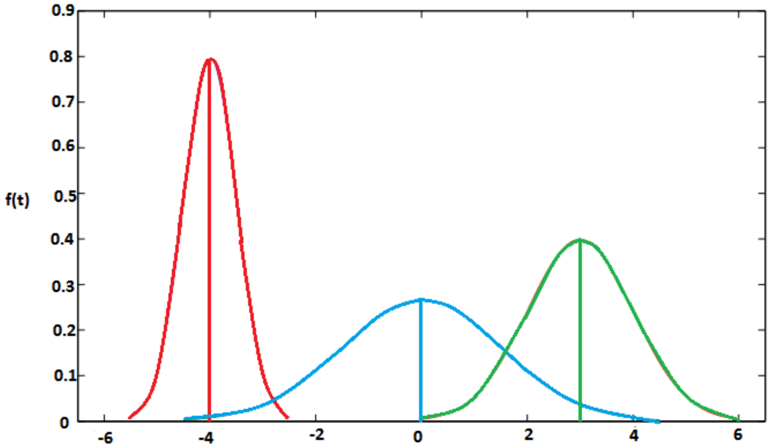


Figure 11.

```
% Written by Eugene.Gallagher@umb.edu for EEOS601
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Revised 1/12/2011
mu1=-4;
mu2=0;
mu3=3;
sigma1=0.5;
sigma2=1.5;
sigma3=1;
x1=mu1-3*sigma1:0.01:mu1+3*sigma1;
x2=mu2-3*sigma2:0.01:mu2+3*sigma2;
x3=mu3-3*sigma3:0.01:mu3+3*sigma3;
y1=normpdf(x1,mu1,sigma1);
y2=normpdf(x2,mu2,sigma2);
y3=normpdf(x3,mu3,sigma3);
plot(x1,y1,x2,y2,x3,y3);
axis([-6.5 6.5 0 0.9])
title('Figure 3.4.5','FontSize',20)
ylabel('f(t)','FontSize',16)
hold on
% This section plots 3 vertical lines
plot([mu1 mu1],[0 normpdf(mu1,mu1,sigma1)],'-b','LineWidth',1);
plot([mu2 mu2],[0 normpdf(mu2,mu2,sigma2)],'-g','LineWidth',1);
plot([mu3 mu3],[0 normpdf(mu3,mu3,sigma3)],'-r','LineWidth',1);
figure(gcf)
hold off
pause
```

3.4.2 Fitting *f(t)* to Data: the Density-scaled histogram

Case Study 3.4.1

```
% LMcs030401_4th.m
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Case Study 3.4.1 p. 167-168
% Written by Eugene.Gallagher@umb.edu
% Other m.files using exponential distribuion
%    LMEx050302.m
Tubes=[317 230 118 93 49 33 17 26 20]
% Data not sufficient to reproduce graph
syms y;
s=int(0.0056*exp(-0.0056*y),500, inf)
P=eval(s)
```

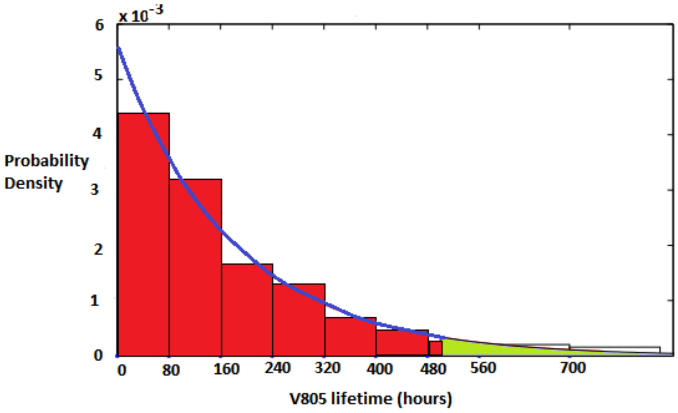


Figure 12.

```
fprintf(...
'Based on the exponential model, the P(tube life >500) is %5.4f\n',P)
% Since the exponential model describes a pdf
s2=int(0.0056*exp(-0.0056*y),0, 500)
P2=eval(s2)
fprintf(...
'Based on 2nd exponential model, the P(tube life >500) is %5.4f\n',1-P2)
x=0:750;ft=0.0056*exp(-0.0056*x);
edges=[0 80 160 240 320 400 480 560 700];
Tubes=[317 230 118 93 49 33 17 26 20];
Density=Tubes./[diff(edges) 140]./sum(Tubes);
bar(edges,Density,'histc')
hold on
plot(x,ft,'r','LineWidth',2)
axis([0 860 0 0.006])
xlabel('V805 lifetimes (hrs)','FontSize',16)
ylabel('Probability density','FontSize',16)
xf=500:860;yf=0.0056*exp(-0.0056*xf);
fill([500 xf 860]',[0 yf 0]',[.8 .8 1])
title('Figure 3.4.8','FontSize',20)
figure(gcf);pause
hold off
```

3.4.3 Continuous probability density functions

Definition 3.4.2 A function *Y* that maps a subset of the real numbers into the real numbers is called a **continuous random variable**. The pdf of *Y* is the function $f_Y(y)$ having the property that for any numbers a and b,

$$P(a \le Y \le b) = \int_a^b f_y(y)\,dy.$$

Example 3.4.4 Not a Matlab problem

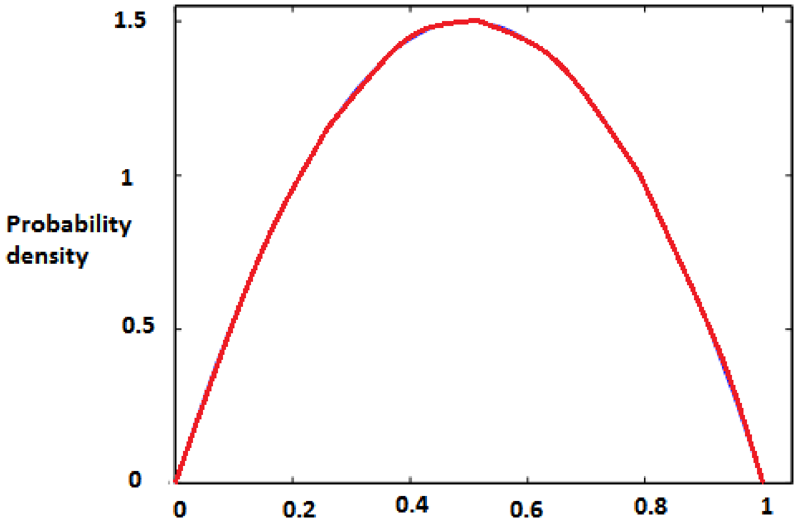


Figure 13.

Example 3.4.5

```
% LMex030405_4th
% Example 3.4.5  p 169 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu for EEOS601
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Revised 1/12/2011
% check whether fyy=6y(1-y), 0 <= y <=1 is a pdf
syms y
int(6*y*(1-y),0,1)
y=0:0.01:1;
fyy=6*y.*(1-y);
plot(y,fyy);ylabel('Probability density','FontSize',16)
axis([0 1.05 0 1.55]);
title('Figure 3.4.9','FontSize',20)
figure(gcf);pause
```

3.4.4 Continuous cumulative distribution functions

Definition 3.4.3 The **cdf** for a continuous random variable Y is an indefinite integral of its pdf:

$$F_Y(y) = \int_{-\infty}^{\infty} f_Y(r)\, dr = P(S \in S \,|\, Y(s) \le y) = P(Y \le y)$$

Theorem 3.4.1 Let $F_Y(y)$ be the **cdf** of a continuous random variable *Y*. Then

$$\frac{d}{dy} F_Y(y) = f_Y(y)$$

Theorem 3.4.2 Let *Y* be a continuous random variable with **cdf** $F_Y(y)$. Then

a. $P(Y > s) = 1 - F_Y(s)$

b. $P(r < Y < s) = F_Y(s) - F_Y(r)$

c. $\lim_{y \to \infty} F_Y(y) = 1$

d. $\lim_{y \to -\infty} F_Y(y) = 0$

3.4.5 Transformations

Theorem 3.4.3

Questions page 172-173

3.5 EXPECTED VALUES

Definition 3.5.1 Let *X* be a discrete random variable with probability function $p_X(k)$. The **expected value** of *X* is denoted *E(X)* (or sometimes μ or μ_X) and is given by

$$E(X) = \mu = \mu_X = \sum_{all\ k} k\, p_X(k)$$

Similarly, if *Y* is a continuous random variable with pdf $f_Y(Y)$,

$$E(Y) = \mu = \mu_X = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy$$

Example 3.5.1

```
% LMEx030501_4th
% Example 3.5.1 Page 175 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu for EEOS601
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Revised 1/12/2011
n=3; p=5/9;
format rat;k=0:n;EX1=k*binopdf(k,n,p)'
% Using Theorem 3.5.1, p. 176
EX2=n*p
format
```

Theorem 3.5.1 Suppose *X* is a binomial random variable with parameters *n* and *p*. Then *E(X)=n p.*

Example 3.5.2

```
% LMEx030502_4th
% Case Study 3.5.2 Page 177 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th Edition
% Written by Eugene.Gallagher@umb.edu
% Fall 2010
% An urn contains 9 chips, N=9, five red (r=5) and 4 white. Three are drawn
% out at random (n=3) without replacement, making this a problem
% involving hypergeometric probabilities. Let X denote the number of red
% chips in the sample. Find E(X), called EX1 here.
format rat
k=0:3;r=5;N=9, n=3;
EX1 = k*hygepdf(k,N,n,r)'
% Using Theorem 3.5.2, p. 177 (after the case study)
w=N-r;
EX2=r*n/(r+w)
format
```

Theorem 3.5.2 Suppose *X* is a hypergeometric random variable with parameters, *r, w,* and *n*. That is, suppose an urn contains *r* red balls and *w* white balls. A sample size *n* is drawn simultaneously from the urn. Let *X* be the number of red balls in the sample. Then

$$E(X) = \frac{r\,n}{r+w}.$$

Example 3.5.3

```
% LMEx030503_4th
% Example 3.5.3 Page 178 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu for EEOS601
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Revised 1/12/2011
% DJ bet is $5 with p=1/1000;payoff
700:1. What is the expected value?
p=1e-3;
Win=3500;Loss=-5;
Ex=Win*p+Loss*(1-p);
fprintf('The Expected Value is
$%6.2f\n',Ex)
```

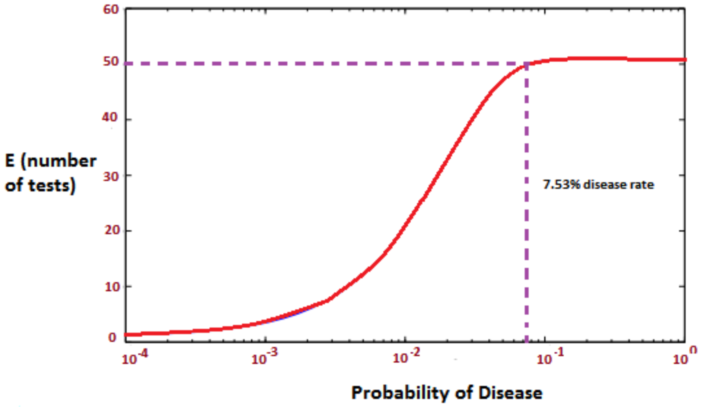


Figure 14.

Example 3.5.4 Blood testing cocktail

```
% LMex030504_4th
% Example 3.5.4 Page 178 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu for EEOS601
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Revised 1/12/2011
% 50 patients are to be tested for a disease. Their blood samples are split
% into A and B pools. All of the A's are mixed together and tested. If the
% pooled A sample is negative, none of the patients have the disease. If
% the pooled A sample comes back positive, then all of the B samples must
% be tested. Under what conditions would this pooling strategy be more
% effective than testing all 50 samples?
% Use Matlab's minimization routine to find the exact number of tests
p=[logspace(-4,0,40)]';
EX=(1-p).^50+51*(1-(1-p).^50);
disp('     Table 3.5.1')
disp('     p        E(X)')
disp([p EX])
plot(p,EX);xlabel('Probability of Disease','FontSize',16)
ylabel('E(Number of Tests)','FontSize',16)
title('Example 3.5.4','FontSize',20)
figure(gcf);pause
% find the exact p value requiring 50 tests
PEx50 = fzero(@(p)(1-p).^50+51*(1-(1-p).^50)-50, 0.001, 0.999)
semilogx(p,EX);xlabel('Probability of Disease','FontSize',16)
ylabel('E(Number of Tests)','FontSize',16)
title('Example 3.5.4','FontSize',20)
hold on
% This plot simply puts a vertical line on the graph
semilogx([PEx50 PEx50],[0 50],'-.b','LineWidth',2);
% This plot simply puts a horizontal line on the graph
semilogx([1e-4 PEx50],[50 50],'-.b','LineWidth',2);
s=sprintf('%4.2f%% disease rate',PEx50*100)
text(PEx50+0.01,25,s,'FontSize',18)
hold off
```

Example 3.5.5 St. Petersburg paradox

```
% LMex030505_4th
% Example 3.5.5 Page 180 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu for EEOS601 1/13/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% St Petersburg paradox
syms k
symsum(2^k*1/2^k,1,inf)
```

Example 3.5.6 Not working

Example 3.5.7 Rayley distribution, not working

3.5.1 A second measure of central tendency: the median

Definition 3.5.2 If X is a discrete random variable, the **median**, m, is that point for which P($X < m$)=P($X > m$). In the event that P($X \leq m$)=0.5 and P($X \geq m'$)=0.5, the median is defined to be the arithmetic average ($m+m'$)/2.

If Y is a continuous random variable, its median is the solution to the integral equation

$$\int_{-\infty}^{m} f_Y(y)\,dy = 0.5.$$

Example 3.5.8

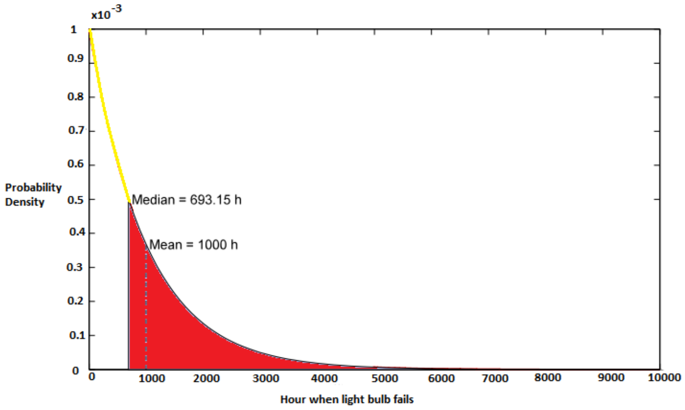


Figure 15.

```
% LMEx030508_4th.m
% Larsen & Marx 2006, p. 183 Introduction to Mathematical Statistics,
% 4th edition
% Written by E. Gallagher, EEOS, UMASS/Boston
% What is the median lifespan for a lightbulb with an average lifespan of
% 1000 hours (with lifetime exponentially distributed)
% Use Matlab's symbolic math to solve the definite integral symbolically
syms y m; int(0.001*exp(-0.001*y),y,0,m)
% use Matlab's minimization routine to find the x that solves the
% equation.
medx=fzero('1-exp(-0.001*x)-0.5',0)
y=0:1e4;
lambda=1000;
fyy=exppdf(y,lambda);  % exponential pdf
plot(y,fyy,'r','LineWidth',2)
ylabel('Probability density','FontSize',16)
xlabel('Hour when light bulb fails','FontSize',16);
title('Example 3.5.8','FontSize',20);figure(gcf)
pause
ax1=gca; % get the handle for the bar chart's axes;
hold on
axis([0 1e4 0 1e-3])
xf=medx:100:1e4;yf=exppdf(xf,lambda); ;
fill([medx xf 1e4]',[0 yf 0]',[.8 .8 1])
s=sprintf('Median = %6.2f h',medx)
text(medx+60,exppdf(medx,lambda),s, 'FontSize',18)
text(1000+60,exppdf(1000,lambda),'Mean = 1000 h', 'FontSize',18)
```

```
% This plot simply puts a vertical line on the graph
plot([1000 1000],[0 exppdf(1000,lambda)],'--k','LineWidth',1);
figure(gcf);pause
hold off
```

Questions p. 184-186

3.5.2 Cracker jack

3.5.2 The expected value of a function of a random variable

Theorem 3.5.3

Example 3.5.9

Example 3.5.10

Example 3.5.11

Example 3.5.12

Example 3.5.13

Example 3.5.14

Questions 192-193 Not too relevant for EEOS601 Summer 2011

The Variance

Definition 3.6.1 The **variance** of a random variable is the expected value of its squared deviations from μ. If *X* is discrete with pdf $p_X(k)$,

$$Var(Y) = \sigma^2 = E[(Y-\mu)^2] = \sum_{all\ k} (k - \mu)^2 \cdot p_X(k)$$

If *Y* is continuous with pdf $f_Y(y)$,

$$Var(Y) = \sigma^2 = E[(Y-\mu)^2] = \int_{-\infty}^{\infty} (y - \mu)^2 \cdot f_Y(y)\, dy$$

The standard deviation is defined to be the square root of the variance. That is,

$$\sigma = standard\ deviation = \sqrt{\sum_{all\ k} (k-\mu)^2 \cdot p_X(k)} \quad if\ X\ is\ discrete$$
$$= \sqrt{\int_{-\infty}^{\infty} (y-\mu)^2 \cdot f_Y(y)} \quad if\ Y\ is\ continuous$$

Theorem 3.6.1 Let W be any random variable, discrete or continuous, having mean μ and for which $E(W^2)$ is finite. Then

$$Var(W) = \sigma^2 = E(W^2) - \mu^2$$

Theorem 3.6.2 Let *W* be any random variable having mean μ and where $E(W^2)$ is finite. Then $Var(aW+b)=a^2 Var(W)$.

Example 3.6.1

Example 3.6.2

```
% LMex030602_4th
% An exercise in finding definite integrals and propagation of error
% Page 197-198 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu, written 2001, revised 1/12/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% A random variable is described by the pdf fY(y)=2y 0<y<1. What is the
% standard deviation of 3Y+2?
% This problem will calculate the variance of y in a symbolic expression
% and then calculate the variance of a linear equation involving y. The
% symbolic variance of the equation will then be converted to its numeric
% value using eval.
syms y
EY=int(y*2*y,0,1)
EY2=int(y^2*2*y,0,1)
VARY=EY2-EY^2
VAR3Yplus2=3^2*VARY
STD=(VAR3Yplus2)^(1/2)
% STD is a symbolic variable. It can be converted to a numeric variable
% using Matlab's eval function.
eval(STD)
```

Questions 3.6.1-3.6.17 p 198-199 All involve calculus

Higher moments (**Skip**) 199-

3.6 Joint Densities (**Skip**)

3.7 Combining Random variables (**Skip**) p. 220

3.8 Further properties of the mean and variance (**Skip**) p 226-

Theorem 3.9.1

Example 3.9.1

Example 3.9.2

Theorem 3.9.2. Let X and Y be any two random variables (discrete or continuous dependent or independent), and let *a* and *b* be any two constants. then

$$E(aX + bY) = aE(X) + bE(Y)$$

Corollary Let W_1, W_2, ..., W_n be any random variables for which $E(W_1) < \infty$, i=1, 2, ..., *n*, and let a_1, a_2, ..., a_n be any set of constants. Then

$$E(a_1 W_1 + a_2 W_2 + \ldots + a_n W_n) = a_1 E(W_1) + a_2 E(W_2) + \ldots + a_n E(W_n)$$

Example 3.9.3

Example 3.9.4

Example 3.9.5

Example 3.9.6

Example 3.9.7

3.8.1 **Expected Values of Products: a special case** p. 232

Theorem 3.9.3 If X and Y are independent random variables,

$$E(XY) = E(X) \cdot E(Y)$$

provided E(X) and E(Y) both exist.

Questions p. 233-234

3.8.2 Calculating the variance of a sum of random variables (p. 234 -236)

Theorem 3.9.4. Let $W_1, W_2, \ldots, W_n$ be a set of independent random variables for which $E(W^2)$ is finite for all i. then

$$Var(W_1 + W_2 + \ldots + W_n) = Var(W_1) + Var(W_2) + \ldots + Var(W_n)$$

Corollary Let $W_1, W_2, \ldots, W_n$ be any set of independent random variables for which $E(W^2) < \infty$ for all I. Let $a_1, a_2, \ldots, a_n$ be any set of constants. Then

$$Var(a_1 W_1 + a_2 W_2 + \ldots + a_n W_n) = a_1^2 Var(W_1) + a_2^2 Var(W_2) + \ldots + a_n^2 Var(W_n)$$

Example 3.9.8 The binomial random variable, being a sum of *n* independent Bernoulli's[1], is an obvious candidate for Theorem 3.9.4. Let X_i denote the number of successes occurring on the ith trial. Then

$$X_i = \begin{cases} 1 \ \textit{with probability } p \\ 0 \ \textit{with probability } 1 - p \end{cases}$$

and

$$X = X_1 + X_2 + \ldots + X_n = \text{total number of successes in n trials.}$$

Find Var (X).
Note that

$$E(X_i) = 1 \cdot p + 0 \cdot (1 - p)$$

and

$$E(X_i^2) = (1)^2 \cdot p + (0)^2 \cdot (1 - p) = p$$

so

$$\begin{aligned} Var(X_i) &= E(X_i^2) - [E(X_i)]^2 = p - p^2 \\ &= p \cdot (1 - p) \end{aligned}$$

It follows, then, that the *variance of a binomial random variable* is *n p (1 - p):*

$$Var(X) = \sum_{i=1}^{n} Var(X_i) = n p (1 - p)$$

[1] **Bernoulli trial** **Hogg & Tanis (1977, p. 66)** A **Bernoulli experiment** is a random experiment, the outcome of which can be classified in but one of two mutually exclusive and exhaustive ways, say success or failure … A sequence of **Bernoulli trials** occurs when a Bernoulli experiment is performed several independent times so that the probability of success remains the same from trial to trial.

Example 3.9.9 In statistics, it is often necessary to draw inferences based on $\overline{W}$, the average computed from a random sample of n observations. Two properties of $\overline{W}$ are especially important. First if the W_is come from a population where the mean is μ, the corollary to Theorem 3.9.2 implies that $E(\overline{W}) = \mu$. Second if the W_is come from a population whose variance is σ^2, then $Var(\overline{W}) = \sigma^2/n$.

Questions p. 236-237

3.8.3 Approximating the variance of a function of random variables

Case Study 3.9.1

Questions

3.9 **Order statistics**

Example 3.10.1

3.9.1 The distribution of extreme order statistics

Example 3.10.2

Example 3.10.3

Example 3.10.4

3.9.2 **A general formula for $f_{y_i'}(y)$** [not covered in EEOS601 - Summer 2011]

Theorem 3.10.1 pdf of the i'th order statisitic [not covered in EEOS601]

Example 3.10.5 Flood tide and levee height

3.9.3 **Joint pdfs of order statistics**

Example 3.10.6

Questoins p. 248-249

3.10 **Conditional Densities** [Relevant to Bayesian inference]

3.10.1 **Finding conditional pdfs for discrete random variables**

Definition 3.11.1 Let X and Y be discrete random variables. The conditional probability density function of Y given x — that is, the probability that Y takes on the value y given that X is equal to x — is denoted $p_{Y|x}$ (y) and given by

$$p_{Y|x}(y) = P(Y=y\,|\,X=x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

for $p_X(x) \neq 0$.

Example 3.11.1 Interesting example. {Will program eventually}

Example 3.11.2

Example 3.11.3 Sum of two binomials with same p is a hypergeometric

Example 3.11.4

Questions p 253-254

3.10.2 **Finding conditional pdfs for continuous random variables**

L'Hôpital's rule & The fundamental rule of calculus used to derive the formula for a continuous pdf

Example 3.11.5

Questions p. 256-257

3.11 Moment-generating functions (not covered in EEOS601 Summer 2011)

3.11.1 Calculating a radom variable's moment-generating function

Definition 3.12.1

Example 3.12.1.

Example 3.12.2

Example 3.12.3

Example 3.12.4 Moment generating function for a normally distributed random variable

Questions p. 260-261

3.11.2 Using Moment-Generating Functions to Find Moments

Theorem 3.12.1

Example 3.12.5

Example 3.12.6

Example 3.12.7

3.11.3 Using Moment-Generating Functions to Find Variances

Example 3.12.8 Variance of a binomial random variable

Example 3.12.9 Variance of a Poisson random variable. Explains why dimensionally the variance of a Poisson variable equals the dimensions of the mean of a Poisson variable. The quadratic term cancels.

Questions p. 266

3.11.4 Using Moment-Generating Functions to identify pdf's

Theorem 3.12.2

Theorem 3.12.3

Example 3.12.10 The sum of inndependent Poissons is also a Poisson. A similar property holds for independent normal random variables ... and under certain conditions, for independent binomial random variables (recall **Example 3.8.1**)

Example 3.12.11

If *Y* is a standard normal variable with mean μ and variance σ^2 the ratio $\frac{Y-\mu}{\sigma}$ is a standard normal variable, *Z*, and we call $\frac{Y-\mu}{\sigma}$ a ***Z transformation***.

3.12 Taking a second look at statistics (interpreting means)

Appendix 3.A.1 Minitab applications

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. {5, 17}

Index