

Brani Vidakovic

# Statistics for Bioengineering Sciences: Results, Hints, and Solutions to the Exercises

Springer



## Preface

This Manual provides solutions and hints to some of the exercises and constitutes a “living” document. Over time more hints and solutions will be added – these additions are always welcome by the students.

If you find an error or have a suggestion for improvement please do not hesitate to send an email.

BRANI VIDAKOVIC  
brani@bme.gatech.edu



# Contents

<b>Preface .....</b>	v
<b>2 Sample and Its Properties .....</b>	1
2.1 Additional Problems .....	7
<b>3 Probability, Conditional Probability, and Bayes Formula.....</b>	9
3.1 Additional Problems .....	17
<b>4 Sensitivity, Specificity, and Relatives .....</b>	21
4.1 Additional Problems .....	23
<b>5 Random Variables.....</b>	25
5.1 Additional Problems .....	34
<b>6 Normal Distribution .....</b>	37
6.1 Additional Problems .....	42
<b>7 Point and Interval Estimators .....</b>	43
7.1 Additional Problems .....	50
<b>8 Bayesian Approach to Inference .....</b>	53
8.1 Additional Problems .....	58
<b>9 Testing Statistical Hypotheses .....</b>	61
<b>10 Two Samples .....</b>	71
10.1 Additional Problems .....	82
<b>11 ANOVA and Elements of Statistical Design .....</b>	85
11.1 Additional Problems .....	90
<b>12 Distribution Free Tests .....</b>	95

<b>13 Goodness of Fit Tests</b> .....	99
<b>14 Models for Tables</b> .....	105
14.1 Additional Problems .....	108
<b>15 Correlation</b> .....	111
15.1 Additional Problems .....	113
<b>16 Regression</b> .....	115
16.1 Additional Problems .....	119
<b>17 Regression for Binary and Count Data</b> .....	125
17.1 Additional Problems .....	128
<b>18 Inference for Censored Data and Survival Analysis</b> .....	131
<b>19 BUGS</b> .....	133

## Chapter 2

# Sample and Its Properties

### 2.1 Auditory Cortex Spikes.

 See file  spikes.m.

### 2.2 On Average.

 The averages are: mean=85.5K, geometric mean=41.2K, median = 30K, harmonic mean=29.3K, and mode=20K. The advertising strategy in which the average salary of 85.5K is quoted will be misleading since a newly hired worker is likely to have a salary less than or equal to the median, most likely the mode.

### 2.3 Contraharmonic mean and $f$ -mean.

$$\text{(a)} 2 \frac{X_1 + X_2}{2} - \frac{2}{\frac{1}{X_1} + \frac{1}{X_2}} = X_1 + X_2 - \frac{2X_1 X_2}{X_1 + X_2} = \frac{X_1^2 + X_2^2}{X_1 + X_2}.$$

$$\text{(b)} C(x, x, x, \dots, x) = \frac{nx^2}{nx} = x.$$

(c) For functions  $f(x) = x$ ,  $f(x) = 1/x$ ,  $f(x) = x^k$ , and  $f(x) = \log(x)$ , the inverse functions are  $f^{-1}(x) = x$ ,  $f^{-1}(x) = 1/x$ ,  $f^{-1}(x) = x^{1/k}$ , and  $f(x) = \exp(x)$ . Substitution and algebra verify (c). For example, if  $f(x) = \log(x)$ ,

$$X_f = \exp \left\{ \frac{1}{n} \sum_{i=1}^n \log X_i \right\} = \exp \left\{ \log \left( \prod_{i=1}^n X_i^{1/n} \right) \right\} = \prod_{i=1}^n X_i^{1/n}.$$

### 2.4 Mushrooms.

 The following MATLAB file provides the solution:

```
 amanita = [9.2, 8.8, 9.1, 10.1, ...
            8.5, 8.4, 9.3, 8.7, ...
            9.7, 9.9, 8.4, 8.6, ...
            8.0, 9.5, 8.8, 8.1, ...
            8.3, 9.0, 8.2, 8.6, ...]
```

```

9.0, 8.7, 9.1, 9.2, ...
7.9, 8.6, 9.0, 9.1, ...
9.2, 8.8, 9.1, 10.1];

%(a)
fivens = [min(amanita) prctile(amanita,25) ...
median(amanita) prctile(amanita,75) max(amanita)]
% fivens =
%    7.9000    8.5500    8.9000    9.2000   10.1000

%(b)
[mean(amanita) mode(amanita)]
%    8.9063    9.1000

%(c)
zis = zscore(amanita);
hist(zis,15)

```

## 2.5 Manipulations with sums.



## 2.6 Emergency Calculation.

Since  $n = 12$  and  $\bar{X} = 15$ ,  $\sum_{i=1}^{12} X_i = 180$ . After the correction, the sum is 192. Thus,  $(\bar{X})_{new} = 192/12 = 16$ .

From  $s^2 = \frac{1}{n-1} (\sum_{i=1}^n X_i^2 - n(\bar{X})^2)$  it follows that

$$\sum_{i=1}^n X_i^2 = (n-1)s^2 + n(\bar{X})^2.$$

This gives  $\sum_{i=1}^{12} X_i^2 = 11 \cdot 34 + 12 \cdot 15^2 = 3074$ . After adjusting for the error,  $(\sum_{i=1}^{12} X_i^2)_{new} = 3074 - 4^2 + 16^2 = 3314$ , and  $(s^2)_{new} = \frac{1}{11} (3314 - 12 \cdot 16^2) = 22$ . Thus, the corrected values are  $(\bar{X})_{new} = 16$  and  $(s^2)_{new} = 22$ .

## 2.7 Sample Mean and Standard Deviation After a Change.

The following MATLAB file provides the solution

```

%sumy -> sum(y_i):old
%sumynew -> sum(y_i): new
%sumy2 -> sum(y_i^2): old
%sumy2new -> sum(y_i^2): new
%NEED: ybarnew and synew

sumy = 15 * 11.6;
sumynew = sumy - 7 + 10;
ybarnew = sumynew/14;

%recall sy = sqrt(1/14 (sumy2 - 15*11.6^2) )
sumy2 = 14*(4.4045)^2 + 15 * 11.6^2;
% now n=15 drops to n=14...

```

```

sumynew2 = sumy2 - 49 + 300; %300=20^2 - 10^2
synew = sqrt( 1/13 * (sumynew2 - 14 * ybarnew^2 ) )
disp(' New ybar   New sy ')
disp([ybarnew synew] )
% New ybar   New sy
% 12.6429   4.8295

```

## 2.8 Surveys on Different Scales.



```

%Surveys on Different Scales
surUK =[6, 7, 5, 10, 3, 9, 9, 6, 8, 2, 7, 5];
surUS =[67, 65, 95, 86, 44, 100, 85, 92, 91, 65];

CVUK = std(surUK)/mean(surUK);
CVUS = std(surUS)/mean(surUS);
disp(' CVUK      CVUS')
disp([CVUK CVUS])
% CVUK      CVUS
% 0.3786    0.2255

```

The UK survey is substantially more variable than the US survey.

## 2.9 Merging Two samples.



Let  $Z_i$  be the values of the merged sample,

$$(Z_1, Z_2, \dots, Z_m, Z_{m+1}, \dots, Z_{m+n}) = (X_1, \dots, X_m, Y_1, \dots, Y_n).$$

Then,

$$\bar{Z} = \frac{1}{m+n} \sum_{i=1}^{m+n} Z_i = \frac{1}{m+n} \left( \sum_{i=1}^m X_i + \sum_{i=1}^n Y_i \right) = \frac{1}{m+n} (m\bar{X} + n\bar{Y}).$$

$$\begin{aligned} s_Z^2 &= \frac{1}{m+n-1} \sum_{i=1}^{m+n} (Z_i - \bar{Z})^2 \\ &= \frac{1}{m+n-1} \left( \sum_{i=1}^m (X_i - \bar{X} + \bar{X} - \bar{Z})^2 + \sum_{i=1}^n (Y_i - \bar{Y} + \bar{Y} - \bar{Z})^2 \right). \end{aligned}$$

Since

$$\sum_{i=1}^m (X_i - \bar{X})(\bar{X} - \bar{Z}) = (\bar{X} - \bar{Z}) \sum_{i=1}^m (X_i - \bar{X}) = 0 \quad \text{and} \quad \sum_{i=1}^n (Y_i - \bar{Y})(\bar{Y} - \bar{Z}) = 0,$$

then

$$(m+n-1)s_Z^2 = (m-1)s_X^2 + m(\bar{X} - \bar{Z})^2 + (n-1)s_Y^2 + n(\bar{Y} - \bar{Z})^2.$$

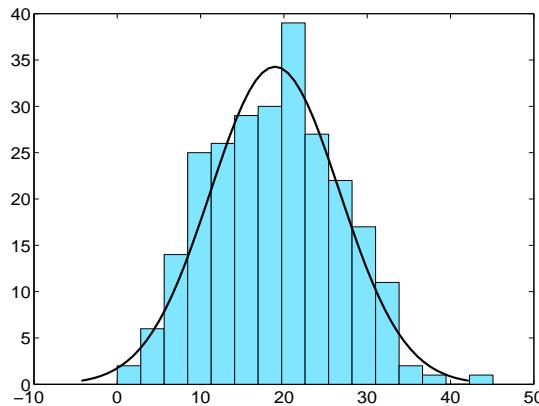
The relation for  $s_Z^2$  follows, since

$$m(\bar{X} - \bar{Z})^2 + n(\bar{Y} - \bar{Z})^2 = m \frac{n^2(\bar{X} - \bar{Y})^2}{(m+n)^2} + n \frac{m^2(\bar{X} - \bar{Y})^2}{(m+n)^2} = \frac{mn}{m+n} (\bar{X} - \bar{Y})^2.$$

### 2.10 Fitting the Histogram.



```
 load('fat.dat')
broz = fat(:,2);
histfit(broz)
h = get(gca,'Children');
set(h(2),'FaceColor',[.5 .9 1])
```



**Fig. 2.1** Histogram of Brozek index `broz` overlaid by best fitting Gaussian curve.

### 2.11 QT Syndrome.



*Hint:* QT is considered *prolonged* if it exceeds 440 ms.

### 2.12 Blowfly Count Time Series.



TBA

### 2.13 Simpson's Diversity Index.



Pure function `Eh` in Example 2.3 should be replaced by

```
Ed = @(f) (sum(f))2/(sum(f.^2)* length(f)).
```

The result is  $Ed(br) = 0.4565$ ,  $Ed(in) = 0.4602$ ,  $Ed(no) = 0.4078$ , and  $Ed(us) = 0.4429$ , and the sample from India is the most homogeneous according to Simpson's homogeneity index.

## 2.14 Speed of Light.



```
%Clean the outliers if any%
irange = iqr(light);
q1 = prctile(light, 25);
q3 = prctile(light, 75);
out1 = find(light < q1 - 2.5*irange)
    %indices for outliers smaller than q1-2.5*iqr
out3 = find(light > q3 + 2.5*irange)
    %indices for outliers larger than q3+2.5*iqr
lightc = light(setdiff((1:length(light)), union(out1,out3)))
%take indices (1:length(light)) minus (out1 union out3),
% so the outlier indices are excluded

%mean, 20% trimmed mean, Real MAD, std, variance
meli = mean(lightc)
tm20 = trimmean(lightc,20)
realmad = 1/0.6745 * mad(lightc,1)
std(lightc)
var(lightc)

figure(1)
% histogram with 30 bins
hist(lightc, 30)

figure(2)
histn(lightc,15,3,42);
hold on
[f,x,u] = ksdensity(lightc);
plot(x,f,'r-','linewidth', lw)
title('Density estimate for the cleaned light data')
```

## 2.15 Limestone Formations in Jamaica.



After loading data limestone.dat the command

`glyphplot(limestone, 'glyph', 'face', 'grid', [3, 6])` produces figure 2.3.

## 2.16 Duchenne Muscular Dystrophy.



## 2.17 Ashton's Dental Data.

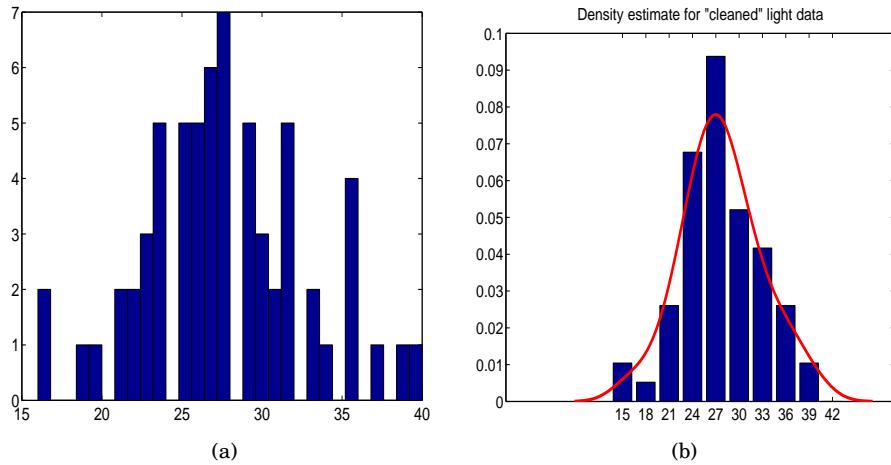


## 2.18 Andrews Plots of Iris Data.

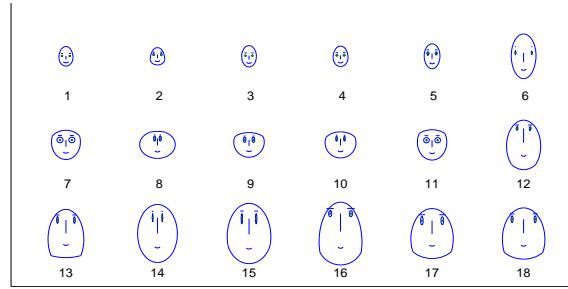


TBA

## 2.19 Cork Boring Data.



**Fig. 2.2** (a) `hist(lightc, 30)`; (b) `histn(lightc,15,3,42); hold on; [f,x,u] = ksdensity(lightc); plot(x,f,'r-','linewidth', 3)`



**Fig. 2.3** Chernoff faces from limestone data.

See [corkrao.m](#).

### 2.20 Balance.

See [balances.m](#).

### 2.21 Cats.

TBA

### 2.22 BUPA Liver Data.

TBA

### 2.23 Cell Circularity Data.

TBA

## 2.1 Additional Problems

**2.a1 Aspirin Weights.** Stoodley (1984) provides 100 weights of aspirin tablets determined using laboratory balance and rounded to the nearest mg. The data in `aspirin.dat` are given as a simple sample.

- (a) Simplify this sample using frequencies of the measurements.
- (b) Find location and spread measures of the sample.
- (c) Plot the histogram of the z-scores.

[Stoodley, K. (1984). *Applied and Computational Statistics, A First Course*. Ellis Horwood LTD, Chichester, England, 229pp.]



# Chapter 3

## Probability, Conditional Probability, and Bayes Formula

### 3.1 Event Differences.

 TBA

### 3.2. Inclusion-Exclusion principle in MATLAB.

 Hint. For example, MATLAB commands

```
numbers = 1:N; A = sum(mod(numbers, 3) == 0);
```

will count how many numbers in  $\{1, \dots, N\}$  are divisible by 3. Find appropriate counts and apply the inclusion-exclusion principle to find the number of favorable outcomes.

### 3.3 A Complex Circuit.

 TBA

### 3.4 De Mere Paradoxes.

 TBA

### 3.5 Probabilities of Some Composite Events.

 TBA

### 3.6 Deighton's Novel.

 (ii)Ans. 63.6%

### 3.7 Reliable System from Unreliable Components.

 (a) The components should be connected in parallel fashion since this increases the reliability.

(b) If single element works or fails with probabilities  $p$  or  $q$ , and if the system  $S$  has  $n$  parallel components, then the probability of  $S$  failing is  $s_S =$

$q^n$ . At least nine components are needed, since  $q_S = 0.2^9 = 5.12 \times 10^{-7} < 10^{-6}$ . Eight components will not be sufficient since  $0.2^8 = 2.56 \times 10^{-6} > 10^{-6}$ .

### 3.8 $k$ -out-of- $n$ Systems.

 A 2-out-of-4 system fails if no or only one component work. If  $p_i$  are probabilities of work and  $q_i$  are complementary probabilities, then probability of system not working is

```


p1=0.1; p2=0.8; p3=0.5; p4=0.4;
q1=0.9; q2=0.2; q3=0.5; q4=0.6;
q=q1*q2*q3*q4 + p1*q2*q3*q4+...
q1*p2*q3*q4 + q1*q2*p3*q4 + q1*q2*q3*p4;
%0.3660

```

Thus, the system works with the probability of 0.3660.

### 3.9 Number of Dominos.

 Solution for (a) is 10 by counting  $\{(0,0), (0,1), (0,2), (0,3), (1,1), (1,2), (1,3), (2,2), (2,3), (3,3)\}$  or by using combinations with repetition  $\binom{4+2-1}{2} = 10$ .

### 3.10 Counting Protocols. TBA

### 3.11 Correlation Between Events.

 The denominators are identical.  $\mathbb{P}(A^c \cap B^c) - \mathbb{P}(A^c)\mathbb{P}(B^c) = 1 - \mathbb{P}(A \cup B) - (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$ .

### 3.12 A Fair Gamble with a Possibly Loaded Coin.

 (a) Yes, one can simulate perfectly fair game with a biased coin. Flip the coin twice, ignore TT, HH outcomes and declare “heads” if you see HT and “tails” if you see TH. The probabilities of these two outcomes are identical,  $p(1-p)$  each.

If one conditions on the event that the outcomes on the two coins are different, these equal probabilities become  $1/2$  each. For any coin, possibly biased, with  $\mathbb{P}(H) = p \neq 1/2$ ,  $\mathbb{P}(HT|HT\text{ or }TH) = \mathbb{P}(TH|HT\text{ or }TH) = \frac{p(1-p)}{2p(1-p)} = 1/2$ .

Therefore a biased coin can emulate a fair coin, but at least two flips are needed to produce a “fair flip.”

(b) To emulate fair die flip the coin 4 times and consider only the cases when you observe 2 H and 2 T. If different number of H and T are obtained, ignore that case and flip the coin again.

There are 6 possibilities: HHTT, HTHT, HTTH, THHT, THTH, and TTHH. Each has the probability  $p^2(1-p)^2$ , but conditioned on the event “two heads and two tails,” each outcome has the probability

$$\frac{\frac{p^2q^2}{(4)}p^2q^2}{(2)p^2q^2} = 1/6.$$

Now assign  $\square$  to HHTT,  $\blacksquare$  to HTHT,  $\dots$ ,  $\blacksquare\blacksquare$  to TTHH.

### 3.13 Neural Signal.

Denote by  $A$  the event that neuron fires, and with  $B$  that it fires in the

$$\text{time interval } [0, t], t < T. \text{ Then } \mathbb{P}(A|B^c) = \frac{\mathbb{P}(B^c|A)\mathbb{P}(A)}{\mathbb{P}(B^c)} = \frac{\frac{T-t}{T}p}{\frac{T-t}{T}p + 1(1-p)} = \frac{(1-t/T)p}{1-t/Tp}.$$

### 3.14 Guessing.

Let  $SR$  and  $SG$  be the events that the subject guesses Red and Green, and let  $LR$  and  $LG$  be the events that the light flashes red and green, respectively. The subject's guess and the light color are independent and  $\mathbb{P}(SR|LR) = \mathbb{P}(SR|LG) = \mathbb{P}(SR) = 0.7$  and  $\mathbb{P}(SG|LR) = \mathbb{P}(SG|LG) = \mathbb{P}(SG) = 0.3$ . Also,  $\mathbb{P}(LR) = 0.7$  and  $\mathbb{P}(LG) = 0.3$ .

(i) Let  $C$  be the event that the subject guesses correctly. By the rule of total probability,

$$\begin{aligned}\mathbb{P}(C) &= \mathbb{P}(C|LR)\mathbb{P}(LR) + \mathbb{P}(C|LG)\mathbb{P}(LG) \\ &= \mathbb{P}(SR|LR)\mathbb{P}(LR) + \mathbb{P}(SG|LG)\mathbb{P}(LG) \\ &= \mathbb{P}(SR)\mathbb{P}(LR) + \mathbb{P}(SG)\mathbb{P}(LG) = 0.3^2 + 0.7^2 = 0.58.\end{aligned}$$

(ii)

$$\mathbb{P}(LR|C) = \frac{\mathbb{P}(C|LR)\mathbb{P}(LR)}{\mathbb{P}(C)} = \frac{\mathbb{P}(SR)\mathbb{P}(LR)}{\mathbb{P}(C)} = \frac{0.3^2}{0.3^2 + 0.7^2} = 0.04655.$$

### 3.15 Propagation of Genes. TBA

### 3.16 Easy Conditioning.

TBA

### 3.17 Eye Color.

Since Megan has blue eyes and both parents are brown-eyed, then the parents are both **Bb**. Without any information on Megan sister's phenotype, the distribution of her allele pairs would be

	BB	Bb	bb
probs	1/4	1/2	1/4

However, since we know that Megan's sister has brown eyes, then the conditional probabilities are calculated as

$$\mathbb{P}(\{\text{BB}\}|\{\text{BB}, \text{Bb}\}) = \frac{\mathbb{P}(\{\text{BB}\} \cap \{\text{BB}, \text{Bb}\})}{\mathbb{P}(\{\text{BB}, \text{Bb}\})} = \frac{\mathbb{P}(\{\text{BB}\})}{\mathbb{P}(\{\text{BB}, \text{Bb}\})} = \frac{1/4}{3/4} = 1/3.$$

Similarly,  $\mathbb{P}(\{\mathbf{Bb}\}|\{\mathbf{BB}, \mathbf{Bb}\}) = 2/3$  and  $\mathbb{P}(\{\mathbf{bb}\}|\{\mathbf{BB}, \mathbf{Bb}\}) = 0$ .

Thus, after information about Megan sister's phenotype her genotype distribution is

	$\mathbf{BB}$	$\mathbf{Bb}$	$\mathbf{bb}$
probs	1/3	2/3	0

Megan sister's husband allays passes **b** allele, and the child will be blue-eyed only if Megan's sister passes allele **b**. This happens with probability

$$\mathbb{P}(\{\text{Megan's sister is } \mathbf{Bb}\}) \times \mathbb{P}(\{\mathbf{b} \text{ is passed from } \mathbf{Bb}\}) = 2/3 \times 1/2 = \boxed{1/3}$$

### 3.18 Dice.

 Denote with  $A$  the event that in 10 rolls there is at least one  $\square$  and with  $B$  that there are at least two  $\square$ . Then,

$$\mathbb{P}(B|A) = 1 - \mathbb{P}(B^c|A) = 1 - \mathbb{P}(AB^c)/\mathbb{P}(A) = 1 - \frac{10 \times 1/6 \times (5/6)^9}{1 - (5/6)^{10}} = 0.6148.$$

### 3.19 Inflation and Unemployment.



		$U$		Marg
		Hi	Low	
$I$	Hi	0.16	0.24	0.4
	Low	0.36	0.24	0.6
Marg		0.52	0.48	1

(a)  $\mathbb{P}(IH) = \mathbb{P}((IH \cap UH) \cup (IH \cap UL)) = \mathbb{P}(IH \cap UH) + \mathbb{P}(IH \cap UL) = 0.16 + 0.24 = 0.40$ .

(b)  $\mathbb{P}(IH|UH) = \frac{\mathbb{P}(IH \cap UH)}{\mathbb{P}(UH)} = 0.16/0.52 = 0.30769$ .

(c) Dependent. For example  $0.16 = \mathbb{P}(IH \cap UH) \neq \mathbb{P}(IH) \times \mathbb{P}(UH) = 0.4 \times 0.52 = 0.208$ .

### 3.20 Multiple Choice.

 Let  $H_1$  be the hypothesis that the student knows the question and  $H_2 = H_1^c$ . It is given that  $\mathbb{P}(H_1) = 0.8$  and  $\mathbb{P}(H_2) = 0.2$ . Denote by  $A$  the event that the student answers the question correctly. Then,  $\mathbb{P}(A|H_1) = 1$  and  $\mathbb{P}(A|H_2) = 0.25$ . Using the rule of total probability, the required probability in (i) is

$$\mathbb{P}(A) = \mathbb{P}(A|H_1)\mathbb{P}(H_1) + \mathbb{P}(A|H_2)\mathbb{P}(H_2) = 1 \cdot 0.8 + 0.2 \cdot 0.25 = 0.85.$$

In (ii) we are interested in  $\mathbb{P}(H_1|A)$  and this can be found using Bayes' rule.

$$\mathbb{P}(H_1|A) = \frac{\mathbb{P}(A|H_1)\mathbb{P}(H_1)}{\mathbb{P}(A)} = \frac{0.8}{0.85} = 0.9412.$$

### 3.21 Manufacturing Bayes.

 Let  $A$  be the event that the randomly selected item is conforming and let  $H_1, H_2$ , and  $H_3$  be the events (hypotheses) that the item is produced on the machine-type 1, 2, or 3. >From the description of the problem,  $\mathbb{P}(A|H_1) = 0.94$ ,  $\mathbb{P}(A|H_2) = 0.95$ , and  $\mathbb{P}(A|H_3) = 0.97$ . From the distribution of total production among the machine types, it follows that  $\mathbb{P}(H_1) = 0.3$ ,  $\mathbb{P}(H_2) = 0.5$ , and  $\mathbb{P}(H_3) = 0.2$ . Note that  $\mathbb{P}(H_1) + \mathbb{P}(H_2) + \mathbb{P}(H_3) = 1$ .

The probability in (i) is found by the Rule of Total Probability:

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A|H_1)\mathbb{P}(H_1) + \mathbb{P}(A|H_2)\mathbb{P}(H_2) + \mathbb{P}(A|H_3)\mathbb{P}(H_3) \\ &= 0.94 \times 0.3 + 0.95 \times 0.5 + 0.97 \times 0.2 = 0.9510.\end{aligned}$$

The probability in (ii) in terms of  $A$  and  $H_1$  is  $\mathbb{P}(H_1|A)$ . Its calculation requires Bayes' rule,

$$\mathbb{P}(H_1|A) = \frac{\mathbb{P}(A|H_1)\mathbb{P}(H_1)}{\mathbb{P}(A)} = \frac{0.94 \times 0.3}{0.9510} = 0.2965.$$

Note that, if the item turned to be conforming, the posterior probability that it was produced on type 1 machine is slightly less than the corresponding prior probability.

### 3.22 Stanley.

 Denote with  $A$  the event that Stanley draws a favorable card (and consequently passes the exam with an  $A$ ).

- (i) If he draws the card first, then clearly  $\mathbb{P}(A) = 8/20 = 2/5$ .
- (ii) If Stanley is second in line, then one card was taken by the student before him. That first card taken might have been favorable (hypothesis  $H_1$ ) or unfavorable (hypothesis  $H_2$ ). Obviously, the hypotheses  $H_1$  and  $H_2$  partition the sample space since no other type of cards is possible in this context. Also, the probabilities of  $H_1$  and  $H_2$  are  $8/20$  and  $12/20$ , respectively. Now, after this first card has been taken, Stanley draws the second. If  $H_1$  had happened, the probability of  $A$  is  $7/19$ , and if  $H_2$  had happened, the probability of  $A$  is  $8/19$ . Thus,  $\mathbb{P}(A|H_1) = 7/19$  and  $\mathbb{P}(A|H_2) = 8/19$ . By the rule of total probability,  $\mathbb{P}(A) = 7/19 \cdot 8/20 + 8/19 \cdot 12/20 = 8/20 = 2/5$ .
- (iii) Stanley has the same probability of getting an  $A$  after two cards have already been taken. The hypotheses are  $H_1 = \{ \text{both cards taken favorable} \}$ ,  $H_2 = \{ \text{exactly one card favorable} \}$ , and  $H_3 = \{ \text{none of the cards taken favorable} \}$ .  $\mathbb{P}(H_1) = 8/20 \cdot 7/19$ ,  $\mathbb{P}(H_3) = 12/20 \cdot 11/19$ , and  $\mathbb{P}(H_2) = 1 - \mathbb{P}(H_1) - \mathbb{P}(H_3)$ . Next,  $\mathbb{P}(A|H_1) = 6/18$ ,  $\mathbb{P}(A|H_2) = 7/18$ , and  $\mathbb{P}(A|H_3) = 8/18$ . Therefore,  $\mathbb{P}(A) = 6/18 \cdot 7/19 \cdot 8/20 + 7/18 \cdot (12 \cdot 16)/(19 \cdot 20) + 8/18 \cdot 11/19 \cdot 12/20 = 8/20 = 2/5$ .

### 3.23 Kokomo, Indiana.

 By Bayes' rule,

$$\begin{aligned}\mathbb{P}(\text{liberal}|\text{did not vote}) &= \frac{(1 - 0.65) \times 0.20}{(1 - 0.82) \times 0.65 + (1 - 0.65) \times 0.20 + (1 - 0.50) \times 0.15} \\ &= 0.07/0.262 = 0.26718.\end{aligned}$$

### 3.24 Mysterious Transfer.

 The solution requires using the rule of total probability, where the event of interest is  $A$ -a ball drawn from the second box is black, and the hypotheses are  $H_1$ -transferred ball is white and  $H_2$ -transferred ball is black. By accounting for the content of the first box, we find  $\mathbb{P}(H_1) = 4/7$  and  $\mathbb{P}(H_2) = 3/7$ . The probability  $\mathbb{P}(A|H_1) = 5/9$  since after the transfer there are 4 white and 5 black balls in the second box. Similarly,  $\mathbb{P}(A|H_2) = 6/9$ . (i) The probability of selecting a black ball from the second box is

$$\mathbb{P}(A) = \mathbb{P}(A|H_1)\mathbb{P}(H_1) + \mathbb{P}(A|H_2)\mathbb{P}(H_2) = 5/9 \times 4/7 + 6/9 \times 3/7 = 38/63.$$

(ii) By Bayes' rule,

$$\mathbb{P}(H_2|A) = \frac{\mathbb{P}(A|H_2)\mathbb{P}(H_2)}{\mathbb{P}(A)} = \frac{18/63}{38/63} = 9/19.$$

### 3.25 Two Masked Robbers.

 Let  $R$  be the event that Mr. Smith is a robber and  $R^c$  its complement, that is, Mr Smith is innocent. Let  $T$  be the event that the lie detector says Mr. Smith is a robber, and  $T^c$  its complement. It is given that  $\mathbb{P}(T|R) = 0.85$  and  $\mathbb{P}(T|R^c) = 0.08$ . We are interested in  $\mathbb{P}(R|T)$ . The events  $R$  and  $R^c$  are hypotheses and  $\mathbb{P}(R|T)$  can be found using Bayes' rule. First, by the rule of total probability in which, for a randomly selected person among 40 people, the detector indicates the person is a robber is

$$\mathbb{P}(T) = \mathbb{P}(T|R)\mathbb{P}(R) + \mathbb{P}(T|R^c)\mathbb{P}(R^c) = 0.85 \times 2/40 + 0.08 \times 38/40 = 0.1185.$$

By Bayes' rule,

$$\mathbb{P}(R|T) = \frac{\mathbb{P}(T|R)\mathbb{P}(R)}{\mathbb{P}(T)} = (0.85 \times 2/40)/0.1185 = 0.38865.$$

The probability that Mr. Smith is a robber if the lie-detector said he was, is less than 39%.

### 3.26 Information Channel.

 Hint:  $\mathbb{P}(ABCA) = \mathbb{P}(ABCA|AAAA) \times 0.3 + \mathbb{P}(ABCA|BBBB) \times 0.5 + \mathbb{P}(ABCA|CCCC) \times 0.2$ . For example  $\mathbb{P}(ABCA|BBBB) = 0.2 \times 0.6 \times 0.2 \times 0.2$ . Apply Bayes' rule.  
**Sol.**

$$\begin{aligned}\mathbb{P}(ABCA) &= \mathbb{P}(ABCA|AAAAA) \times 0.3 + \mathbb{P}(ABCA|BBBBB) \times 0.5 + \mathbb{P}(ABCA|CCCCC) \times 0.2 \\ &= 0.6^2 \cdot 0.2^2 \cdot 0.3 + 0.6 \cdot 0.2^3 \cdot 0.5 + 0.6 \cdot 0.2^3 \cdot 0.2 = 0.00768.\end{aligned}$$

By Bayes' rule,  $\mathbb{P}(AAAAA|ABCA) = 0.6^2 \cdot 0.2^2 \cdot 0.3 / 0.00768 = 0.5625$ .

As an easy side result one can find  $\mathbb{P}(BBBBB|ABCA) = 0.3125$  and  $\mathbb{P}(CCCCC|ABCA) = 0.125$  and check that  $0.5625 + 0.3125 + 0.125 = 1$ .

### 3.27 Quality Control.



### 3.28 Let's Make a Deal.



### 3.29 Ternary channel.



```
 % prsissj means probability of received si if sent sj (given)
% prsi    means probability of received si (question in (a))
% pssi    means probability sent si (given 1/3 each)
% pssirsj means probability sent si if received sj (question in (b))
%
prs1ss1 = 0.75;   prs2ss1 = 0.1;   prs3ss1 = 0.15;
prs1ss2 = 0.098;  prs2ss2 = 0.9;   prs3ss2 = 0.002;
prs1ss3 = 0.02;   prs2ss3 = 0.08;  prs3ss3 = 0.9;

pss1 = 1/3; pss2 = 1/3; pss3 = 1/3;

%(a) total probabaility formula
prs1 = prs1ss1 * pss1 + prs1ss2 * pss2 + prs1ss3 * pss3  %0.2893
prs2 = prs2ss1 * pss1 + prs2ss2 * pss2 + prs2ss3 * pss3  %0.3600
prs3 = prs3ss1 * pss1 + prs3ss2 * pss2 + prs3ss3 * pss3  %0.3507

%(b) Bayes' formula
pss1rs1 = prs1ss1 * pss1/prs1 %0.8641
pss2rs1 = prs1ss2 * pss2/prs1 %0.1129
pss3rs1 = prs1ss3 * pss3/prs1 %0.0230

pss1rs2 = prs2ss1 * pss1/prs2 %0.0926
pss2rs2 = prs2ss2 * pss2/prs2 %0.8333
pss3rs2 = prs2ss3 * pss3/prs2 %0.0741

pss1rs3 = prs3ss1 * pss1/prs3 %0.1426
pss2rs3 = prs3ss2 * pss2/prs3 %0.0019
pss3rs3 = prs3ss3 * pss3/prs3 %0.8555

%0.8641  0.1129  0.0230
%0.0926  0.8333  0.0741
%0.1426  0.0019  0.8555
```

### 3.30 Sprinkler Bayes Net.



```

model
  cloudy ~ dcat(p.cloudy[]);
  sprinkler ~ dcat(p.sprinkler[cloudy,]);
  rain ~ dcat(p.rain[cloudy,]);
  wetgrass ~ dcat(p.wetgrass[sprinkler,rain,])

list(
  #hard evidence , uncomment and instantiate...
  #   sprinkler = 1,
  #   cloudy = 1,
  #   rain = 1,
  #   wetgrass = 2,
  #initial distributions
  p.cloudy = c(0.5,0.5),
  # conditionals
  p.sprinkler = structure(.Data = c(0.50, 0.50,
                                    0.90, 0.10), .Dim = c(2,2)),
  p.rain = structure(.Data = c(0.80, 0.20,
                               0.20, 0.80), .Dim = c(2,2)),
  p.wetgrass = structure(.Data = c(1.,    0.0,
                                  0.1,    0.9,
                                  0.1,    0.9,
                                  0.01,   0.99), .Dim = c(2,2,2))
) #end list

```

### 3.31 Diabetes in Pima Indians Bayes Net.



```

model
  pregnancies ~ dcat(p.pregnancies[]);  #Multiple pregnancies?
  age ~ dcat(p.age[]);      #Older than 50%
  overweight ~ dcat(p.overweight[]);  #Heavier than 50%
  diabetes ~ dcat(p.diabetes[pregnancies,age,overweight,]);
                                         #Diagnosed with diabetes?
  glycose ~ dcat(p.glycose[diabetes,]);  #Elevated glycose?
  insulin ~ dcat(p.insulin[diabetes,])  #Elevated insulin?
  bloodpressure ~ dcat(p.bloodpressure[overweight,diabetes,]);
                                         #High blood pressure

```

DATA

```

list(#put hard evidence as 1 or 2, un-comment as needed
  #pregnancies=2,
  #age = 1,
  #overweight=1,
  diabetes=2,
  #glycose = 1,
  #insulin =2,
  #bloodpressure=1,
  #next are distributions of initial nodes:
  p.pregnancies= c(0.45,0.55),
  p.age = c(0.5, 0.5),
  p.overweight = c(0.5, 0.5),
  #the rest are conditional probability distributions:
  p.diabetes = structure(.Data =
  c(0.95,0.05,          0.67, 0.33,
    0.59,0.41,          0.40, 0.60,
    0.73,0.27,          0.66, 0.34,
    0.63,0.37,          0.41, 0.60 ), .Dim = c(2,2,2,2)),
  p.glycose = structure(.Data = c(0.64,0.36,0.21,0.79),
                        .Dim = c(2,2)),
  p.insulin = structure(.Data = c(0.49,0.51,0.52,0.48),
                        .Dim = c(2,2)),
  p.bloodpressure = structure(.Data = c(0.55,0.45,
                                         0.58,0.42,
                                         0.40,0.60,
                                         0.49,0.51), .Dim = c(2,2,2))

```

Just Generate Initials by "gen inits"

### 3.32 A Simplified Probabilistic Model for Visual Pathway.



## 3.1 Additional Problems

**3.a1 Twins.** Dizygotic (fraternal) twins have the same probability of each gender as in overall births, which is approximately 51% male, 49% female. Monozygotic (identical) twins must be of the same gender. Among all twin pregnancies, about 1/3 are monozygotic.

Find the probability of two girls in  
 (a) monozygotic pregnancy,  
 (b) dizygotic pregnancy, and  
 (c) dizygotic pregnancy given that we know that the gender of the babies is the same.

If Mary is expecting twins, but no information about the type of pregnancy is available, what is the probability that the babies are

- (d) two girls;  
 (e) of the same gender;  
 (f) Find the probability that Mary's pregnancy is dizygotic if it is only known that the babies are two girls.

Retain four decimal places in your calculations.

*Hint:* (a) given; (b) genders are independent; (c) conditional probability:  $P(A|B) = P(A \cap B)/P(B)$ . Since  $A$  is subset of  $B$ ,  $A \cap B = A$  and  $P(A|B) = P(A)/P(B)$ ; (d) total probability formula/rule. What are the hypotheses? (e) similar to (d); (f) Bayes' rule.

 (a)  $P(GG|MZ) = 0.49$ . This is because a single egg is fertilized to form one zygote, which subsequently divides into two separate embryos.

(b) Because of independence, this probability is  $P(GG|DZ) = 0.49 \times 0.49 = 0.2401$ .

(c) The same gender  $S = BB \cup GG$  in dizygotic pregnancy happens with probability of  $P(S|DZ) = 0.49^2 + 0.51^2 = 0.5002$ . Then the probability is  $P(GG|DZ \cap S) = P(GG \cap S|DZ)/P(S|DZ) = P(GG|DZ)/P(S|DZ) = 0.2401/0.5002 = 0.4800$ .

(d)  $P(GG) = P(GG|MZ)P(MZ) + P(GG|DZ)P(DZ) = 0.49 \cdot 1/3 + 0.2401 \cdot 2/3 = 0.3234$ .

(e)  $S = BB \cup GG$ ;  $P(S) = P(S|MZ)P(MZ) + P(S|DZ)P(DZ) = 1 \cdot 1/3 + 0.5002 \cdot 2/3 = 0.6668$ .

(f)  $P(DZ|GG) = P(GG|DZ)P(DZ)/P(GG) = \frac{0.2401 \cdot 2/3}{0.3234} = 0.4949$ .

**3.a2 Greta.** There is a 10% chance that pure breed German shepherd Greta is a carrier of canine hemophilia A. If she is a carrier, there is a 50-50 chance that she will pass the hemophiliac gene to a puppy.

Greta has two male puppies and they are tested free of hemophilia. What is the probability that Greta is a carrier, given this information about her puppies?

*Hint:* Passing the hemophiliac gene is independent between the puppies. If the puppies are male than the only way they will get the hemophilia is from the mother carrier since hemophilia is X-chromosome-bound disorder.

 Let  $H_1$  denote the event that Greta is a carrier, then  $P(H_1) = 0.1$ . Let  $A$  be the event that the two male puppies are disease free. Then  $P(A|H_1^c) = 1$ , that is, if Greta is not a carrier, the puppies are disease free with probability 1. Because of independence  $P(A|H_1) = 0.5 * 0.5$  and according to the total probability formula

$$P(A) = P(A|H_1)P(H_1) + P(A|H_1^c)P(H_1^c) = 0.5 \cdot 0.5 \cdot 0.1 + 1 \cdot 0.9 = 0.925.$$

By Bayes formula,

$$P(H_1|A) = \frac{P(A|H_1)P(H_1)}{P(A)} = 0.025/0.925 = 0.027.$$

**3.a3 Gambling Fallacy.** An event that happened on August 18, 1913 in Le Grand Casino de Monte Carlo made headlines. The ball of a roulette wheel landed on “black” 26 times in a row. Out of 37 slots denoted by 0-36 (French roulettes have no a 00-slot), 18 slots (2, 4, 6, 8, 10, 11, 13, 15, 17, 20, 22, 24, 26, 28, 29, 31, 33, 35) are black, so the probability of a ball landing in black is 18/37.

- (a) What is the probability that in the next 26 spins of a similar roulette wheel the ball lands on “black” every single time.
- (b) After the ball landed in black slot 15 times in a row, the players in Le Grand Casino frantically started to bet on red, and that evening Casino amassed a profit in millions of Francs. If one started to bet on black with \$1, what capital he/she will have after 26th consecutive black, if Casino doubles the bet placed on winning color.



# Chapter 4

## Sensitivity, Specificity, and Relatives

### 4.1 Stacked Auditory Brainstem Response.

 TBA

### 4.2 Hypothyroidism.

 TBA

### 4.3 Alzheimer's.

  $\mathbb{P}(T|D) = 436/450 = 0.9689$  and  $\mathbb{P}(T^c|D^c) = 495/500 = 0.99$ . The first is the probability that a patient who shows symptoms of Alzheimer's disease would test positive (sensitivity) and the second is the probability that a subject who does not have symptoms of Alzheimer would test negative (specificity). Note that  $\mathbb{P}(T^c|D) = 1 - \mathbb{P}(T|D) = 0.0311$  and  $\mathbb{P}(T|D^c) = 1 - \mathbb{P}(T^c|D^c) = 0.01$ .

By Bayes' formula

$$\begin{aligned}\mathbb{P}(D|T) &= \mathbb{P}(T|D)\mathbb{P}(D)/\mathbb{P}(T) \\ &= \frac{\mathbb{P}(T|D)\mathbb{P}(D)}{\mathbb{P}(T|D)\mathbb{P}(D) + \mathbb{P}(T|D^c)\mathbb{P}(D^c)} \\ &= \frac{0.9689 \times 0.113}{0.9689 \times 0.113 + 0.01 \times 0.887} = 0.9251.\end{aligned}$$

### 4.4 Test for Being a Duchenne Muscular Dystrophy Carrier.

 TBA

### 4.5 Parkinson's Disease Statistical Excursions.

 TBA

### 4.6 Blood Tests in Diagnosis of Inflammatory Bowel Disease.

 **Sol.** The number of TP is (sensitivity  $\times$  number of people with the disease),  $0.903 \times 103 = 93.009 \approx 93$ . To find TN, we multiply specificity with the number of controls,  $0.8 \times 50 = 40$ . The table is

	disease present (D)	no disease present (C)	total
test positive (P)	93	10	103
test negative (N)	10	40	50
total	103	50	153

Prevalence can be evaluated from the table. It is the proportion of people with the disease among all 153 subjects in the experiment,  $103/153 = 0.6732 \approx 67.3\%$ .

Positive predicted value is the proportion of people who have the disease among the subjects who tested positive. In this case it happened to coincide with sensitivity,  $93/103 = 90.3\%$ .

#### 4.7 Carpal Tunnel Syndrome Tests.

 TBA

#### 4.8 Hepatic Scintigraphy.

 TBA

#### 4.9 Apparent Prevalence.

 TBA

#### 4.10 HAAH Improves the Test for Prostate Cancer.



(a) Recall that sensitivity is the ratio of true positives and total number of subjects with the disease. Since 233 subjects are with the disease, the sensitivity of 95% means that there are  $233 \cdot 0.95 = 221.35 \approx 221$  true positives. Thus  $tp = 221$ . This gives  $233 - 221 = 12$  false negatives, thus  $fn = 12$ .

Similarly, 43 subjects do not have disease. Since specificity is 0.93, the true negatives are  $43 \cdot 0.93 = 39.99 \approx 40$ . This means  $tn=40$  and  $fp = 3$ . The table is

	disease	no disease	total
test positive	tp=221	fp=3	tot.pos = 224
test negative	fn=12	tn=40	tot.neg = 52
total	tot.dis=233	tot.ndis=43	total=276

(b)

$$\begin{aligned}
 & P(\text{disease} | \text{test positive}) \\
 &= \frac{P(\text{test positive} | \text{disease})P(\text{disease})}{P(\text{test positive} | \text{disease})P(\text{disease}) + P(\text{test positive} | \text{no disease})P(\text{no disease})} \\
 &= \frac{\text{sensitivity} \cdot \text{prevalence}}{\text{sensitivity} \cdot \text{prevalence} + (1-\text{specificity}) \cdot (1-\text{prevalence})} \\
 &= \frac{221/233 \times 7/100}{221/233 \times 7/100 + 3/43 \times 93/100} \\
 &= \boxed{0.5058}
 \end{aligned}$$

$$(c) PPV = \frac{tp}{tp+fp} = 221/224 = \boxed{0.9866}.$$

In both (b) and (c) we have found positive predicted value, that is  $P(\text{disease} | \text{test positive})$ . However, (b) and (c) differ in the information where the subject comes from, which is critical for the prevalence. If the subject comes from the general population then the prevalence is 0.07 and that is used in place of  $P(\text{disease})$  in the Bayes formula.

If we selected the subject from the group involved in this study (that is, selected person is one of 276 subjects), then the “prevalence” refers to this particular group and is  $\frac{tp+fn}{\text{total } n} = 233/276$ .

#### 4.11 Creatinine Kinase and Acute Myocardial Infraction.



#### 4.12 Asthma.

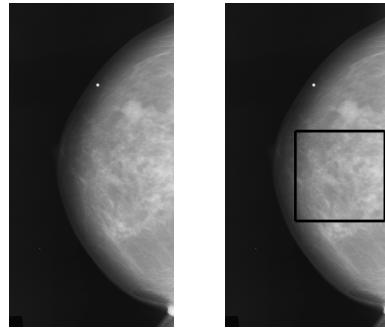


## 4.1 Additional Problems

**4.a1 Spectral Indices of Mammogram Images Predictive of BC.** Can the properties of mammogram backgrounds be indicative of breast cancer? The collection of digitized mammograms was obtained from the University of South Florida’s Digital Database for Screening Mammography (DDSM). Images from this database are coupled with cancer status verified through biopsy. For every image a slope of wavelet spectra was calculated (Hamilton et al., 2011<sup>1</sup>), and corresponding cancer status recorded. Only the craniocaudal projection images were used: the right breast image for all normal cases, and the cancerous breast (right or left) image for cancer cases. There were 105 normal (benign) cases, and 72 cancer cases considered. A malignant mammogram and subimage used to find the spectral slope are presented in Fig. 4.1.

---

<sup>1</sup> Erin K. Hamilton, Seonghye Jeon, Pepa Ramírez Cobo, Kichun Sky Lee, and Brani Vidakovic (2011). Diagnostic Classification of Digital Mammograms by Wavelet-Based Spectral Tools: A Comparative Study. Proceedings of BIBM 2011, Atlanta GA



**Fig. 4.1** *Left panel:* right CC mammogram corresponding to a malignant case. *Right panel:* subimage of size  $1024 \times 1024$  considered for the analysis

The data set `sslopesstatus.dat` contains two columns: slope of the spectra and breast cancer status. The goal is to propose and evaluate a test for BC based only on the slope of mammogram wavelet spectra. A MATLAB file `cancerslope.m` reads in the data and calculates and plots the ROC.

- (a) Find AUC. How would you grade this test?
- (b) Find Youden Index (YI - maximal distance of ROC from the  $45^\circ$  line).
- (c) What threshold for the slope would you suggest so that mammograms with slopes exceeding this threshold are considered positive for BC. Assume that the errors of misclassification are equally bad.
- (d) What are the sensitivity/specificity of the test at the threshold suggested in (c)?

*Hint.* Calculations similar to (a-d) can be found in `rocada.m`.

# Chapter 5

## Random Variables

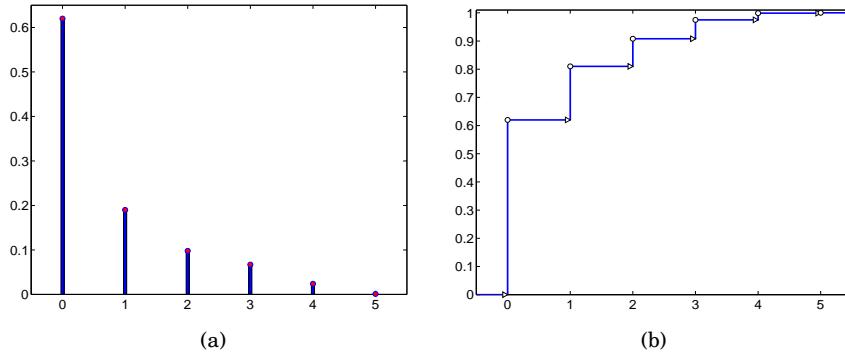
### 5.1 Phase I Clinical Trials and CTCAE Terminology.



```
X = [0 1 2 3 4 5];
px = [0.620 0.190 0.098 0.067 0.024 0.001];
sum(px) %check that the probabilities sum up to 1
w=0.02; ms=6; %plotting parameters: width of bar and marker size
xx = X; yyp= px;yyc= cumsum(px);
figure(1)
subplot(211)
bar(xx, yyp, w,'b')
hold on
plot(xx,yyp, 'bo','MarkerSize',ms, 'MarkerFaceColor','b')
hold off
axis tight
subplot(212)
stairs(xx,yyc)
hold on
plot(xx(2:end),yyc(1:end-1),'b>')
plot(xx, yyc, 'bo', 'MarkerFaceColor','b' )
hold off
% Expectation
EX=X*px' %or EX = sum(X .* px)
% k-th moment EXk = (X.^k)*px'
EX2 = (X.^2) * px' %second moment
% Variance (second central moment)
VarX = EX2 - EX^2 %or VarX=sum( (X-EX).^2 .* px )
```

### 5.2 Mendel and Dominance.

A child from hybrid parents will be *DD*, *Dd*, or *dd* with probabilities of 1/4, 1/2, and 1/4, respectively. One offspring will give outward dominant



**Fig. 5.1** CTCAE discrete random variable (a) probability mass function and (b) cdf.

appearance with probability  $1/4 + 1/2 = 3/4$ . Now we have Binomial sampling with parameters  $n = 4$  and  $p = 3/4$  and the required probability is  $27/64$ .

### 5.3 Chronic Kidney Disease.



- (a) `binopdf(3, 10, 0.17)` = 0.1600,
- (b) `1-binopdf(0, 5, 0.4)`=0.9222,
- (c) (i) `binopdf(3, 5, 6/16)`=0.2060, (ii) `1-binopdf(0, 5, 6/16)`=0.9046, and
- (d) `geopdf(3-1, 0.4)` = 0.1440.

### 5.4 Ternary channel.



### 5.5 Conditioning a Poisson.



$$\begin{aligned}
 \mathbb{P}(X_1 = k | X_1 + X_2 = n) &= \frac{\mathbb{P}(X_1 = k, X_1 + X_2 = n)}{\mathbb{P}(X_1 + X_2 = n)} \\
 &= \frac{\mathbb{P}(X_1 = k, X_2 = n - k)}{\mathbb{P}(X_1 + X_2 = n)} \\
 &= \frac{\frac{\lambda_1^k}{k!} e^{-\lambda_1} \times \frac{\lambda_2^{n-k}}{(n-k)!} e^{-\lambda_2}}{\frac{(\lambda_1 + \lambda_2)^n}{n!} e^{-\lambda_1 - \lambda_2}} \\
 &= \binom{n}{k} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \times \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n-k},
 \end{aligned}$$

which is  $\mathcal{B}in\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$  ..

### 5.6 Rh+ Plates.



**5.7 Your Colleague's Misconceptions About Density and CDF.** TBA**5.8 Falls among elderly.** TBA**5.9 Cell clusters in 3-D Petri dishes.****5.35 Heat Production by a Resistor.** TBA  TBA**5.10 Left-handed Twins.** TBA**5.11 Pot Smoking is Not Cool!**

```
 % Solution
disp('(a) Bin(12, 0.7): P(7 <= X <= 9)');
%(a) using binopdf(x,n,p)
disp('(%a)-using pdf'); binopdf(7, 12, 0.70) ...
    + binopdf(8, 12, 0.70) + binopdf(9, 12, 0.70)
% ans = 0.6293
% using binocdf(x, n, p)
disp('(%a)-using cdf'); binocdf(9, 12, 0.70) - binocdf(6, 12, 0.70)
% ans = 0.6293
%(b) at most five i.e., X <= 5
disp('(%b) Bin(12, 0.7): P(X <= 5)'); binocdf(5, 12, 0.70)
% ans = 0.0386
%(c) not less than 8 is 8,9,10,11,12 or complement of <=7
disp('(%c) Bin(12, 0.7): P(X >= 8)'); 1-binocdf(7, 12, 0.70)
% ans = 0.7237
%-----
```

**5.12 Emergency Help by Phone.**

(a)  $Y$  = number of calls until first call answered late (including the late one),  
 $Y \sim \text{Geom}(0.1)$ ,  $\mathbb{E}Y = 10$ , ( `[m var] = geostat(0.1); mean = m + 1` ).

(b)  $X$  = number of calls of the next 10 that are answered late,  $X \sim \text{Bin}(10, 0.1)$   
 $\mathbb{P}(X = 1) = \binom{10}{1} \cdot 0.1 \cdot (0.9)^9 = 0.3874$ . ( `binopdf(1, 10, 0.1)` )

**5.13 Min of Three.**

 The sample space looks like the following table and each entry has the probability  $(1/3)^3 = 1/27$ , because of independence.

$X_1 X_2 X_3$	$M$	$Mx$	$R$	$X_1 X_2 X_3$	$M$	$Mx$	$R$	$X_1 X_2 X_3$	$M$	$Mx$	$R$
111	1	1	0	211	1	2	1	311	1	3	2
112	1	2	1	212	1	2	1	312	1	3	2
113	1	3	2	213	1	3	2	313	1	3	2
121	1	2	1	221	1	2	1	321	1	3	2
122	1	2	1	222	2	2	0	322	2	3	1
123	1	3	2	223	2	3	1	323	2	3	1
131	1	3	2	231	1	3	2	331	1	3	2
132	1	3	2	232	2	3	1	332	2	3	1
133	1	3	2	233	2	3	1	333	3	3	0

By counting the equally-likely outcomes from the table, we find the probability distribution functions and cumulative distribution functions for  $M$  and  $R$ .

$$\begin{array}{c|ccc} M & 1 & 2 & 3 \\ \hline p & 19/27 & 7/27 & 1/27 \end{array} \quad F_M(m) = \begin{cases} 0, & m < 1 \\ 19/27, & 1 \leq m < 2 \\ 26/27, & 2 \leq m < 3 \\ 1, & m \geq 3 \end{cases}$$

and

$$\begin{array}{c|ccc} R & 0 & 1 & 2 \\ \hline \text{prob} & 1/9 & 4/9 & 4/9 \end{array} \quad F_R(r) = \begin{cases} 0, & r < 0 \\ 1/9, & 0 \leq r < 1 \\ 5/9, & 1 \leq r < 2 \\ 1, & r \geq 2 \end{cases}$$

This is not an elegant solution. A somewhat more elegant solution is the following: Since  $M$  is the minimum of  $X_1, X_2$  and  $X_3$ :

$$P(M > m) = P(X_1 > m, X_2 > m, X_3 > m) = [P(X_1 > m)]^3 = \begin{cases} 1, & m < 1 \\ (2/3)^3, & 1 \leq m < 2 \\ (1/3)^3, & 2 \leq m < 3 \\ 0, & m \geq 3. \end{cases}$$

Now,  $P(M = 2) = P(M > 1) - P(M > 2) = (2/3)^3 - (1/3)^3 = 7/21$ , and  $P(M = 3) = P(M > 2) - P(M > 3) = (1/3)^3 - 0 = 1/27$ . Since,  $P(M = 1) = 1 - P(M = 2) - P(M = 3)$ , the distribution for  $M$  follows.

For distribution of  $R$  the following consideration is useful. There are  $3^3$  possible realizations (number of words of length  $k$  in alphabet consisting of  $n$  symbols is  $n^k$ ).  $R$  can be 0, 1, or 2. It is easy to see that event  $\{R = 0\}$  corresponds to 3 realizations: 000, 111, and 222. Thus,  $P(R = 0) = 3/27 = 1/9$ .

The difference  $\{R = 1\}$  happens when the “words” of length 3 come from alphabets  $\{1, 2\}$  or  $\{2, 3\}$  and not all “letters” are the same. There are  $2^3 + 2^3 - 4$  such words. We subtract 4 since words 111 and 222 can be formed in alphabet

$\{1,2\}$  and 222 and 333 in alphabet  $\{2,3\}$ . Thus  $P(R = 1) = 12/27$ . Finally,  $P(X = 2) = 1 - 3/27 - 12/27 = 12/27$ .

### 5.14 Cystic Fibrosis in Japan.



### 5.15 Random Variables as Models.



(a) The rate of gastrointestinal reactions per prescription is  $538/9160000$ , and per 10000 prescriptions is  $538/9160000 \times 10000 = 0.5873$ .

(b) If  $\lambda = 0.5873$ , and  $X$  is the number of gastrointestinal reactions per 10000 prescriptions, then the suggested model is  $X \sim \text{Poi}(0.5873)$ . Furthermore,

$$P(X = 2) = \frac{0.5873^2}{2!} e^{-0.5873} = 0.0959,$$

i.e., about 9.6%.

(c) The probability is  $P(X \geq 2)$ , which is equal to  $1 - P(X < 2) = 1 - P(X \leq 1)$ , since the Poisson model is discrete and  $P(X < 2) = P(X \leq 1)$ . Then,

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 1) \\ &= 1 - P(X = 0) - P(X = 1) = 1 - \frac{0.5873^0}{0!} e^{-0.5873} - \frac{0.5873^1}{1!} e^{-0.5873} = 0.1177. \end{aligned}$$

i.e., about 11.8%. Recall,  $0! = 1$ , by definition.

If MATLAB is used, then

```
 lambda = 538/9160000 * 10000 %answer for (a)
%lambda = 0.5873

p2 = lambda^2/2 * exp(-lambda) %answer for (b)
%p2 = 0.0959

p2plus = 1 - lambda^0/0! * exp(-lambda)...
- lambda^1/1! * exp(-lambda) %answer for (c)
%p2plus = 0.1177
```

Parts (b) and (c) can be found via `poisspdf` and `poisscdf` as

```
poisspdf(2, 0.5873)
%ans = 0.0959

1-poisscdf(1, 0.5873)
%ans = 0.1177
```

### 5.16 Additivity of Gammas.

 The moment generating function for gamma  $\mathcal{G}a(r, \lambda)$  distribution is  $\frac{\lambda^r}{(\lambda-t)^r}$ .

By convolution property of moment generating functions,

$$m_Y(t) = \prod_{i=1}^n m_{X_i}(t) = \frac{\lambda^{r_1}}{(\lambda-t)^{r_1}} \frac{\lambda^{r_2}}{(\lambda-t)^{r_2}} \cdots \frac{\lambda^{r_n}}{(\lambda-t)^{r_n}} = \frac{\lambda^r}{(\lambda-t)^r},$$

which is the moment generating function of  $\mathcal{G}a(r, \lambda)$ . Since the moment generating functions uniquely determine distributions when they exist, the additivity property of Gammas with respect to the shape parameter is proved.

### 5.17 Memoryless property.

 Note that in general  $\{X \geq u+v\} \cap \{X \geq u\}$  is equivalent to  $\{X \geq u+v\}$ .

Proof for exponentials. Since, for the exponential distribution  $\mathbb{P}(X \leq x) = 1 - e^{-x/\beta}$ ,  $x \geq 0$  is the cdf, then the *residual life* is  $\mathbb{P}(X \geq x) = e^{-x/\beta}$ .

Then,

$$\mathbb{P}(X \geq u+v | X \geq u) = \frac{\mathbb{P}(\{X \geq u+v\} \cap \{X \geq u\})}{\mathbb{P}(X \geq u)} = \frac{\mathbb{P}(X \geq u+v)}{\mathbb{P}(X \geq u)} = \frac{e^{-(u+v)/\beta}}{e^{-u/\beta}} = e^{-v/\beta} = \mathbb{P}(X \geq v).$$

Proof for geometric. Denote  $q = 1-p$ . Then  $\mathbb{P}(X \geq x) = 1 - \mathbb{P}(X < x) = 1 - \mathbb{P}(X \leq x-1) = 1 - \sum_{k=0}^{x-1} q^k p$ . Since,  $\sum_{k=0}^{x-1} q^k = \frac{1-q^x}{1-q}$  and  $1-q = p$ , the residual life is  $\mathbb{P}(X \geq x) = 1 - p \frac{1-q^x}{1-q} = 1 - (1 - q^x) = q^x$ .

Now, as for the exponentials,

$$\mathbb{P}(X \geq u+v | X \geq u) = \frac{\mathbb{P}(X \geq u+v)}{\mathbb{P}(X \geq u)} = \frac{q^{u+v}}{q^u} = q^v = \mathbb{P}(X \geq v).$$

### 5.18 Rh System.

 TBA

### 5.19 Blood Types.

 (a)  $X \sim \mathcal{B}in(24, 0.374)$   $\mathbb{P}(X = 8) = \binom{24}{8} 0.374^8 (1 - 0.374)^{24-8} = 0.1566$ .

$\mathbb{E}X = np = 8.976$ ,  $\mathbb{V}ar X = npq = 5.619$ .

(b)  $\text{hygecdf}(2, 16, 8, 5) = 0.5$ .

(c)  $X \sim \mathcal{P}oi(500 \times 0.006)$ .  $\mathbb{P}(X \geq 1) = 0.9502$ .

(d)  $X \sim \mathcal{G}eo(0.085)$ .  $\mathbb{E}X = 1/p = 11.7647$ .

### 5.20 Variance of the Exponential.

  $\mathbb{E}X^2 = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = [[u = x^2; dv = \lambda e^{-\lambda x}; du = 2x dx; v = -e^{-\lambda x}]] - x^2 e^{-\lambda x} \Big|_0^\infty +$

$$2 \int_0^\infty x e^{-\lambda x} dx = 0 + 2 \int_0^\infty x e^{-\lambda x} dx = [[u = x; dv = e^{-\lambda x} dx; du = dx; v = -\frac{1}{\lambda} e^{-\lambda x}]] - \frac{2x}{\lambda} e^{-\lambda x} \Big|_0^\infty + \frac{2}{\lambda} \int_0^\infty e^{-\lambda x} dx = 0 + \frac{2}{\lambda} \int_0^\infty e^{-\lambda x} dx = \frac{2}{\lambda} \times \frac{1}{\lambda} = 2/\lambda^2.$$

Since  $\text{Var } X = \mathbb{E}X^2 - (\mathbb{E}X)^2$  and  $\mathbb{E}X = 1/\lambda$ , it follows that  $\text{Var } X = 2/\lambda^2 - (1/\lambda)^2 = 1/\lambda^2$ .

### 5.21 Equipment Aging.

 (a) Note that for exponential distribution,  $\mathbb{P}(T > x) = 1 - \mathbb{P}(T \leq t) = 1 - F(t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}$ . Then,  $0.8 = \mathbb{P}(T > 10) = e^{-10\lambda}$ . That means,  $\lambda = -\frac{1}{10} \ln(0.8) = 0.0223$ .

(b)  $ET = 1/\lambda = 44.843$  and  $Var T = 2010.9$

(c) Let  $t_p$  be  $100p$  percentile. Then  $F(t_p) = p$ . Solving  $1 - e^{-\lambda t_p} = p$  we obtain an exact formula for  $100p$  percentile,  $t_p = -\frac{1}{\lambda} \ln(1-p)$ .

- Median  $t_{0.5} = -\frac{1}{0.0223} \ln(0.5) = 31.0828$
- $Q_1: t_{0.25} = -\frac{1}{0.0223} \ln(0.75) = 12.9005$
- $Q_3: t_{0.75} = -\frac{1}{0.0223} \ln(0.25) = 62.1657$
- $IQR = Q_3 - Q_1 = 62.1657 - 12.9005 = 49.2652$

In MATLAB

```
 expinv([0.25 0.5 0.75], 1/0.0223)
%ans = 12.9005    31.0828    62.1657
```

### 5.22 A Simple Continuous Random Variable.

 TBA

### 5.23 2-D Continuous Random Variable Question.



- (a)  $C = e$   
(b)  $f_X(x) = \frac{e^{x+1}-1}{e^x}$ .  $f_Y(y) = \frac{e(1-(1+y)e^{-y})}{y^2}$ .

### 5.24 Insulin Sensitivity.

 Hint: Here MATLAB's parametrization of gamma density is used,  $\alpha = r$  and  $\beta = 1/\lambda$ . In terms of  $\alpha$  and  $\beta$ ,  $\mathbb{E}X = \alpha\beta$  and  $\text{Var } X = \alpha\beta^2$ .

### 5.25 Correlation Between a Uniform and its Power.

 TBA

### 5.26 Precision of Lab Measurements.

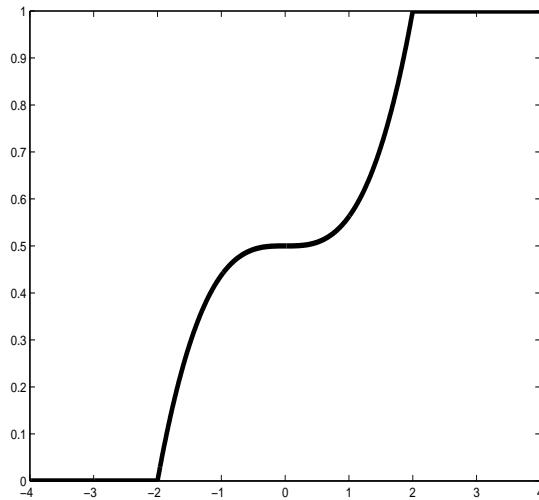
 (a)  $\mathbb{P}(\text{measurement } X \text{ accurate}) = \mathbb{P}(|X| < 0.5) = \mathbb{P}(-0.5 < X < 0.5) = \int_{-0.5}^{0.5} 3x^2/16 dx = \frac{3}{16} \frac{x^3}{3} \Big|_{-0.5}^{0.5} = 1/16(1/8 - (-1/8)) = 1/64 = 0.0156$ .  
(b) For  $x < -2$ ,  $F(x) = 0$  and for  $x > 2$ ,  $F(x) = 1$ . For  $-2 \leq x \leq 2$ ,

$$F(x) = \int_{-2}^x \frac{3}{16} t^2 dt = \frac{t^3}{16} \Big|_{-2}^x = x^3/16 - (-2)^3/16 = x^3/16 + 1/2.$$

In MATLAB,

```

```



**Fig. 5.2**  $F(x) = x^3/16 + 1/2$ ,  $-2 \leq x \leq 2$ ;  $F(x) = 0$ ,  $x < -2$ ;  $F(x) = 1$ ,  $x > 2$ .

```

x = -2:0.01:2;
y = x.^3/16 + 1/2;
plot(x, y, 'linewidth',4)
xn = -4:0.01:-2;
yn = 0 .* xn;
xp = 2:0.01:4;
yp = ones(size(xp));
hold on
plot(xp, yp, 'linewidth',4)
plot(xn, yn, 'linewidth',4)

```

$$(c) \mathbb{E}Y = \mathbb{E}X^2 = \int_{-2}^2 x^2 \frac{3}{16} x^2 dx = \frac{3}{16} \frac{x^5}{5} \Big|_{-2}^2 = \frac{3(32 - (-32))}{16 \cdot 5} = 192/80 = 2.4$$

### 5.27 Lifetime of Cells.



```

%expected life: beta = 4 mo
% 150 days = 5 mo
1-expcdf(5,4)           %(Cells(a))
%ans = 0.2865
% 1 y= 12 mo; Poisson(12/4)
% Observed on average expectation of Poisson(3) = 3.
% Or rationalize:
% 12/average life time = 3, but this is informal
1-poisscdf(5, 3)        %(Cells(b))
%ans = 0.0839
%
1 - gamcdf(12, 3, 4)    %(Cells(c))
%ans = 0.4232

```

```
%  
%(Cells(d)): By memoryless property  
% P(X>=7.2|X>=2.2)=P(X>=7.2-2.2)=P(X>=5)  
% which is equal to (Cells(a)), 0.2865
```

### 5.28 Silver-coated Nylon Fiber.

```
%Silver Coated Nylon Fibers  
% (a)  
1 - expcdf(10, 10) % 0.3679  
% (b)  
expcdf(15,10) %0.7769  
% (c) is the same as (a) because of memoryless property.
```

### 5.29 Xeroderma pigmentosum.

 TBA

### 5.30 Failure Time. TBA

### 5.31 Beta Fit.

 TBA

### 5.32 Uncorrelated but Possibly Dependent.

 Enough to show that  $\text{Cov}(Z, W) = 0$ . This follows from  $\text{Cov}(Z, W) = \mathbb{E}((X + Y)(X - Y)) = \mathbb{E}(X^2 - Y^2) = \mathbb{E}X^2 - \mathbb{E}Y^2$ , and  $\mathbb{E}X^2 = \mathbb{E}Y^2$ .

### 5.33 Nights of Mr Jones.



	Monday	Tuesday	Wednesday	Thursday	Friday
Prob(Insomnia)	1	0.4000	0.4600	0.4540	0.4546
Prob(Sleep Well)	0	0.6000	0.5400	0.5460	0.5454

### 5.34 Stationary Distribution of MC.

 TBA

### 5.35 Heat Production by a Resistor.

 TBA

## 5.1 Additional Problems

**5.a1 Africanized Honey Bee.** Matis et al. (1992), also Pal et al. (2005), modeled the 'transit-time' distribution of Africanized honey bee spread through northern Guatemala and Mexico. Data were collected on the first capture times of the bee at various monitoring transects in northern Guatemala and on the eastern and western costs of Mexico. The time intervals between consecutive sightings (in months) are reported. The transit time (months/100 km) data set consists of 45 observations.

5.3	1.8	4.2	5.7	3.8	0.8	1.4	3.5	17.5
4.6	0.8	6.3	2.9	0.6	1.9	2.0	6.7	5.5
2.5	2.2	6.7	5.7	10.0	3.3	3.5	20.0	1.6
8.3	4.8	20.0	3.6	8.2	1.3	4.0	5.0	1.7
2.0	2.9	19.2	1.1	1.4	1.5	3.2	8.6	2.2

It was suggested that gamma  $\text{Ga}(r, \lambda)$  distribution is an appropriate model for the transition time,  $T$ . Here,  $r$  is the shape parameter and  $\lambda$  is the rate parameter.

- (a) It is known that  $ET = r/\lambda$  and  $\text{Var}T = r/\lambda^2$ . Find moment matching estimators for  $r$  and  $\lambda$  by replacing  $ET$  and  $\text{Var}T$  with  $\bar{T}$  and  $s^2$ .
- (b) For  $r$  and  $\lambda$  as in (a), find the probability that transit time  $T$  exceeds 15 (month/100 km).
- (c) What is the 0.8 quantile of  $T$ , that is, find  $t^*$  for which  $P(T \leq t^*) = 0.8$ .
  - Matis, J. H., Rubink, W. L., Makela, M. (1992). Use of the gamma distribution for predicting arrival times of invading insect populations. *Environmental Entomology*, **21**, 431–440.
  - Pal, N., Jin , C., Lim W.-K. (2005). *Handbook of Exponential and Related Distributions for Engineers and Scientists*, Chapman and Hall/CRC.

*Hint: In MATLAB be careful about the parametrization of gamma distribution. MATLAB uses scale parameter  $\beta = 1/\lambda$  instead of rate parameter  $\lambda$ .*



```
% Africanized Honey Bee transit times
t =[...
5.3   1.8   4.2   5.7   3.8   0.8   1.4   3.5   17.5 ...
4.6   0.8   6.3   2.9   0.6   1.9   2.0   6.7   5.5 ...
2.5   2.2   6.7   5.7   10.0  3.3   3.5   20.0  1.6 ...
8.3   4.8   20.0  3.6   8.2   1.3   4.0   5.0   1.7 ...
2.0   2.9   19.2  1.1   1.4   1.5   3.2   8.6   2.2];

tbar = mean(t)          %5.1067
s2 = var(t)             %25.0884

lamhat = tbar/s2        %0.2035
```

```
rhat = (tbar)^2 /s2 %1.0394
%(b) 1/lamhat = 4.9129
1 - gamcdf(15, 1.0394, 4.9129) %0.0509
%(c)
gaminv(0.8, 1.0394, 4.9129) %8.1922
```

**5.a2 Imperfectly Observed Poisson.** Suppose that the number of particular experimental events in time interval  $[0, T]$  has a Poisson distribution  $\mathcal{P}oi(\lambda T)$ . A student who is observing the experiment may fail to count any of the events. An event is counted with probability equal to  $p$  and missing one event is independent of missing or counting the others. What is the distribution of events in  $[0, T]$  that are counted?



By total probability formula,

$$\begin{aligned} P(n \text{ events counted}) &= \sum_{k=n}^{\infty} (P(n \text{ events counted}|k \text{ events happened})P(k \text{ events happened})) \\ &= \sum_{k=n}^{\infty} \binom{k}{n} p^n (1-p)^{k-n} (\lambda T)^k \exp\{-\lambda T\}/k! \\ &= \exp\{-\lambda T\}(p\lambda T)^n/n! \sum_{k=n}^{\infty} \frac{(1-p)\lambda T)^{k-n}}{(k-n)!} \\ &= (p\lambda T)^n \exp\{-p\lambda T\}/n! \end{aligned}$$

after representing  $\binom{k}{n}$  by factorials and observing that  $\sum_{k=n}^{\infty} \frac{(1-p)\lambda T)^{k-n}}{(k-n)!} = \sum_{v=0}^{\infty} \frac{(1-p)\lambda T)^v}{v!} = \exp\{(1-p)\lambda T\}$ .

Thus, the number of counted events is again Poisson but with the rate  $p\lambda T$ .

**5.a3 The Smallest of  $k$  Exponentials.** Inter-event times in a particular experimental process are distributed as exponential  $\mathcal{E}(\lambda)$  where  $\lambda$  is the rate parameter. Suppose that  $k$  inter-event times are recorded. What is the distribution of the minimal inter-event time?



Let  $Y = \min_{1 \leq i \leq k} X_i$ . Then,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = 1 - P(Y > y) = 1 - P(X_1 > y, X_2 > y, \dots, X_k > y) \\ &= 1 - \prod_{i=1}^k P(X_i > y) = 1 - \left(e^{-\lambda y}\right)^k = 1 - e^{\lambda k y}. \end{aligned}$$

Thus,  $F_Y(y)$  is the cdf of exponential distribution with parameter  $\lambda k$ .



# Chapter 6

## Normal Distribution

### 6.1 Standard Normal Calculations.

 TBA

### 6.2 Nonnegative Definiteness of $\Sigma$ Constrains $\rho$ .

 TBA

### 6.3 Herrings.



 normcdf(13,10.5,1.6) - 1/2 %% normcdf(a,a,b)=1/2, why?  
% ans = 0.4409 %%about 44%  
1-chi2cdf(10, 8)  
% ans = 0.2650 %%about 26.5%  
norminv(0.9, 10.5, 1.6)  
% ans = 12.5505

### 6.4 Sea Urchins.

 Here,  $X \sim \mathcal{N}(2.83, 0.79^2)$ .

(a)  $P(2.3 \leq X \leq 4) = P\left(\frac{2.3-2.83}{0.79} \leq Z \leq \frac{4-2.83}{0.79}\right) = P(-0.67 \leq Z \leq 1.48) = \Phi(1.48) - \Phi(-0.67) = \Phi(1.48) - (1 - \Phi(0.67)) = 0.9306 - (1 - 0.7486) = 0.6792$ .

In MATLAB

 normcdf(1.48)-normcdf(-0.67)  
%ans = 0.6791  
% or more precisely  
normcdf(4, 2.83, 0.79) - normcdf(2.3, 2.83, 0.79)  
%ans = 0.6796

(b)  $P(X > t^*) = 0.95$ , is the same as  $P(X \leq t^*) = 0.05$ . The 5th percentile of standard normal is  $-1.64$ , and  $\frac{t^*-2.83}{0.79} = -1.64$ . The solution is  $t^* = 2.83 - 1.64 \times 0.79 = 1.53$ .

**In MATLAB**


```
norminv(0.05, 2.83, 0.79)
%ans = 1.5306
```

**6.5 Pyruvate Kinase for Controls is Normal.**

**TBA**
**6.6 Leptin.**

**TBA**
**6.7 Pulse Rate.**

**TBA**
**6.8 Side Effects.**

**Correct answer: (c).**


```
normcdf(0.30, 0.25, 0.08)
%ans = 0.7340
```

**6.9 Macrolepiota Procera.**



```
% (a)
normcdf(250, 230, 25)-normcdf(200, 230, 25)
% ans = 0.6731
% OR
normcdf((250 - 230)/25)-normcdf((200 - 230)/25)
%ans = 0.6731
%(b)
norminv(0.95, 230, 25)
% ans = 271.1213
```

**6.10 Duration of Gestation in Humans.**


Under the assumed model, the probability that randomly selected pregnancy case has a duration period equal or larger than 349 is  $2.6 \times 10^{-12}$ , less than 3 out of a trillion. The evidence of adultery is overwhelming. Apparently, the judges have not taken a course in statistics.

**6.11 Tolerance Design.**


Denote by  $D_{pin}$  and  $D_{hole}$  the random dimensions of interest. According to the conditions,  $\sigma_{pin} = t_{pin}/3 = 0.001$ .

$$D_{pin} \sim \mathcal{N}(5, 0.001^2),$$

$$D_{hole} \sim \mathcal{N}(5.005, \sigma_{hole}^2),$$

and

$$D_{gap} = D_{hole} - D_{pin} \sim \mathcal{N}(5.005 - 5, 0.001^2 + \sigma_{hole}^2).$$

It is given that  $\mathbb{P}(D_{gap} \geq 0.001) = 0.999$ , which, after standardizing, becomes

$$\begin{aligned} \mathbb{P}\left(Z \geq \frac{0.001 - 0.005}{\sqrt{0.001^2 + \sigma_{hole}^2}}\right) &= 0.999. \\ \frac{0.001 - 0.005}{\sqrt{0.001^2 + \sigma_{hole}^2}} &= z_{0.001} = -3.0902 \quad (\text{norminv}(0.001) = -3.0902). \end{aligned}$$

From the above,

$$0.004 = 3.0902 \sqrt{0.001^2 + \sigma_{hole}^2} \Rightarrow \sigma_{hole}^2 = \left(\frac{0.004}{3.0902}\right)^2 - 0.001^2 = 6.7551 \times 10^{-7}.$$

Finally, the tolerance  $t_{hole}$  is  $3 \cdot \sqrt{6.7551 \times 10^{-7}} = 0.0025$ .

### 6.12 Ulnar Variance.

 The ulnar variance  $X$  has normal  $\mathcal{N}(0.74, 1.46^2)$  distribution.

(a) Need  $P(X < 0) = P(Z < \frac{0-0.74}{1.46}) = \Phi(-0.74/1.46)$ . In MATLAB, there are two equivalent ways of getting the solution, using standard normal cdf from standardized argument or using general normal cdf directly.

 % (a)  
normcdf(0, 0.74, 1.46) % 0.3061  
% or  
normcdf( (0 - 0.74)/1.46 ) % 0.3061

(b)  $C$  is the difference between two normal random variables. Recall, if  $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$  and  $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ , then  $X_2 - X_1 \sim \mathcal{N}(\mu_2 - \mu_1, \sigma_1^2 + \sigma_2^2)$ . Similar problem was discussed in class – “piston problem” in combining normal random variables.

 mu1 = 0.19; mu2 = 1.52;  
sigma1 = 1.43; sigma2 = 1.56;  
  
muC = mu2 - mu1 %muC = 1.3300  
  
sigmaC = sqrt( sigma1^2 + sigma2^2 ) %sigmaC = 2.1162  
  
1 - normcdf( 1, muC, sigmaC ) %ans = 0.5620

### 6.13 Independence of Sample Mean and Standard Deviation in Normal Samples.

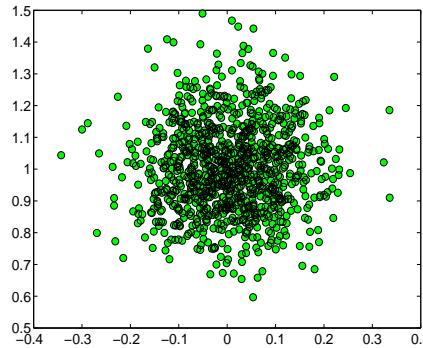


```


randn('state',1)           %fix random seed
x=normrnd(0, 1,[100, 1000]); %matrix 100 x 1000
xx = mean(x); yy=var(x);   %done columnwise
corr(xx', yy')            %0.0294
%
plot(xx, yy, 'o', 'MarkerFaceColor','g',...
'MarkerEdgeColor','k','MarkerSize',8)

```

The scatterplot in Figure 6.1 shows no dependence patterns.



**Fig. 6.1** Scatterplot of 1000 points with first coordinate being the mean of 100 standard normals and second coordinate being their variance. The scatterplot suggest no relationship between the coordinates. For this case, the coefficient of correlation is 0.0294.

#### 6.14 Sonny and Multiple Choice Exam.



```


1- normcdf(34.5, 100*0.25, sqrt(100*0.25*0.75))
%ans = 0.0141

```

#### 6.15 Amount of Liquid in a Bottle.



```


normcdf(0.48, 0.5, 0.01)
%ans = 0.0228

norminv(0.95, 0.5, 0.01)
%ans = 0.5164

```

#### 6.16 Meristem Cells in 3D.



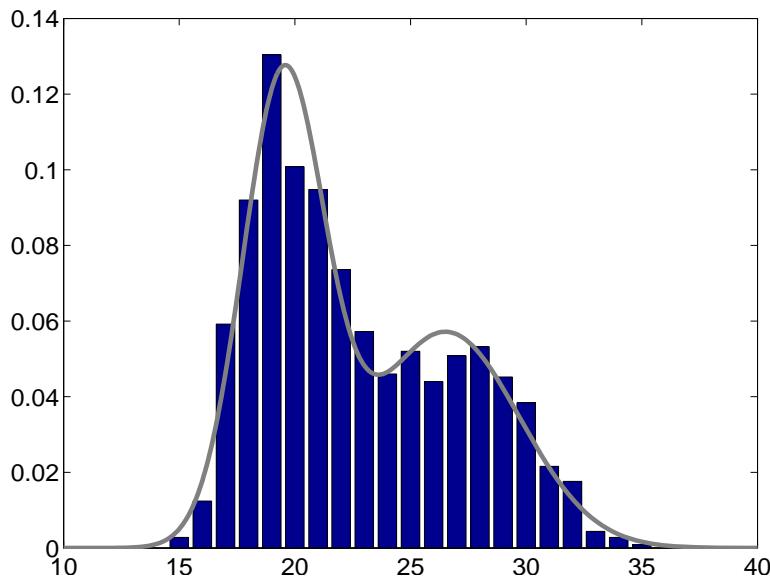
Answer: 0.0002.  $\mathbb{P}\left(\frac{X^2}{\sigma^2} + \frac{Y^2}{\sigma^2} + \frac{Z^2}{\sigma^2} \geq \frac{70^2}{250}\right) = \mathbb{P}(\chi_3^2 \geq 10.6) = 0.0002.$

#### 6.17 Glossina morsitans.

```


%Tsetse Fly
clear all
microns = 15:35;
freq = [7 31 148 230 326 252 237 184 143 ...
    115 130 110 127 133 113 96 54 44 11 7 2];
sample = [];
for i = 1:21
    sample =[sample; repmat(microns(i), freq(i),1)];
end
bar(microns, freq)
mix = gmdistribution.fit(sample,2);
mix.mu
%ans = 26.523 19.493
mix.Sigma
%ans(:,:,1) = 10.113 ans(:,:,2) = 3.141
mix.PComponents
%ans = 0.45551 0.54449

```



### 6.18 Stabilizing Variance.

 For the exponential  $\mathcal{E}(\lambda)$ ,  $\mathbb{E}(X) = \lambda$  and  $\text{Var } X = \lambda^2$ , so  $\sigma^2 = \mu^2$ . Thus, the integral in (??) is

$$g = c \int \frac{dx}{|x|} = c \log|x| + d.$$

For the binomial case,  $\sigma^2 = np(1-p) = np - np^2 = np - \frac{(np)^2}{n} = \mu - \frac{\mu^2}{n}$ . the integral in (??) is

$$g = c \int \frac{dx}{\sqrt{x-x^2/n}}.$$

### 6.19 From Normal to Lognormal.



### 6.20 The Square of a Standard Normal.



The transformation  $y = g(x) = x^2$  has two inverse branches,  $h_1(y) = \sqrt{y}$  and  $h_2(y) = -\sqrt{y}$ . Also,  $f_X(x) = \frac{1}{2\pi} e^{-x^2/2}$ . Then by the equation in (5.10) on p. 174,

$$\begin{aligned} f_Y(y) &= f_X(h_1(y))|h'_1(y)| + f_X(h_2(y))|h'_2(y)| \\ &= \frac{1}{\sqrt{2\pi}} \exp\{-(\sqrt{y})^2/2\} \left| \frac{1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} \exp\{-(-\sqrt{y})^2/2\} \left| -\frac{1}{2\sqrt{y}} \right| \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} = \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})} y^{1/2-1} e^{-y/2}, \quad y \geq 0. \end{aligned}$$

since  $\sqrt{\pi} = \Gamma(\frac{1}{2})$ .

## 6.1 Additional Problems

**6.a1 Area Spanned by Whiskers.** In MATLAB's boxplot the maximum whisker length is by default  $1.5 IQR$ , where  $IQR$  is the interquartile range  $Q_3 - Q_1$ . For a standard normal distribution, what area under the density is spanned by a box-plot with two maximal whiskers (i.e., with range  $[Q_1 - 1.5 IQR, Q_3 + 1.5 IQR]$ ).

Sol.

```
>> norminv(0.25)
ans = -0.6745 %Q_1, Q_3 = - Q_1
>> normcdf(4* 0.6745) - normcdf(-4* 0.6745) % 2*Q_3=IQR, 1+3=4
ans = 0.9930
```

# Chapter 7

## Point and Interval Estimators

### 7.1 Tricky Estimation.

 Suppose that the total number of misprints is  $T$ . Let  $p_1$  be the accuracy of first reader (probability that he/she will find a particular misprint), and let  $p_2$  be the accuracy of the second reader. Because of independence, their joint accuracy is  $p_1 p_2$  ("joint" in sense that they both find a particular misprint). None of the  $T$ ,  $p_1$  or  $p_2$  are known, but an estimate of  $p_1 T$  is 60, of  $p_2 T$  is 70, and  $p_1 p_2 T$  is 50.

Thus,

$$T = \frac{(p_1 T) \cdot (p_2 T)}{p_1 p_2 T}$$

can be estimated by  $\frac{60 \cdot 70}{50} = 84$ .

On the other hand, the total number of misprints spotted by both is  $60+70-50 = 80$ . Thus, it follows that the estimated number of remaining misprints is  $84-80=4$ .

### 7.2 Laplace's Rule of Succession.

 TBA

### 7.3 Neurons Fire in Potter's Lab.

 TBA

### 7.4 The MLE in a Discrete Case.

 TBA

### 7.5 MLE for Two Continuous Distributions.

 TBA

### 7.6 Match the Moment.

*Hint:*  $L(p) = (1-p)^{\sum_i X_i} \cdot p^n$ .  
 TBA

### 7.7 Weibull Distribution.

HINT: Recall that  $\Gamma(n) = (n - 1)!$

 TBA

### 7.8 Rate Parameter of Gamma.

 TBA

### 7.9 Estimating Parameter of Rayleigh Distribution.

 (a)  $\hat{\sigma}_{mm1}^2 = \frac{2(\bar{X})^2}{\pi}$  and  $\hat{\sigma}_{mm2}^2 = \frac{\sum_{i=1}^n X_i^2}{2n}$ .    (b)  $\hat{\sigma}_{mle}^2 = \hat{\sigma}_{mm2}^2$ .    (c)  $\sigma_{mm1}^2 = 7.7986$ ,  $\sigma_{mm2}^2 = 6.75$ .    (d) Yes. Since  $\lambda = \frac{1}{2\sigma^2}$  by the invariance of MLE,  $\hat{\lambda}_{mle} = \frac{1}{2(\hat{\sigma}_{mle})^2}$ .

### 7.10 Monocytes Among Blood Cells.

 TBA

### 7.11 Estimation of $\theta$ in $\mathcal{U}(0, \theta)$ .

 TBA

### 7.12 Estimating the Rate Parameter in a Double Exponential Distribution.

 TBA

### 7.13 Reaction Times I.



```
n=20; xb=0.9; s=0.12;
[xb-tinv(0.975,19)*s/sqrt(n), xb+tinv(0.975,19)*s/sqrt(n)]
%ans = 0.8438 0.9562
[xb-tinv(0.9925,19)*s/sqrt(n), xb+tinv(0.9925,19)*s/sqrt(n) ]
%ans = 0.8282 0.9718
[(n-1)*s^2/chi2inv(0.975, n-1), (n-1)*s^2/chi2inv(0.025, n-1)]
%ans = 0.0083 0.0307
```

### 7.14 Reaction Times II.



```
[xb-norminv(0.9925)*s/sqrt(n), xb+norminv(0.9925)*s/sqrt(n)]
%ans = 0.8347 0.9653
1.96^2*s^2*4/0.07^2
%ans = 45.1584
```

**7.15 Toxins.**

```

A X=[3 2 5 3 2 6 5 4.5 3 3 4];
Xbar = mean(X)
%Xbar = 3.6818
n=length(X)
%n = 11
s = std(X)
%s = 1.3091

[Xbar - tinv(0.995, n-1) * s/sqrt(n), ...
Xbar + tinv(0.995, n-1) * s/sqrt(n)]
%ans = 2.4309 4.9327

```

**7.16 Bias of  $s^*$ .****7.17 COPD Patients.**

```

A n=157; X=87;
phat = X/n
%phat = 0.5541
qhat = 1 - phat
%qhat = 0.4459
[phat - norminv(0.95)*sqrt(phat*qhat/n), ...
phat+norminv(0.95)*sqrt(phat*qhat/n)]
%ans = 0.4889 0.6194
n=(2*norminv(0.95)*sqrt(0.5 * 0.5)/0.03)^2
% n = 3.0062e+003

```

(i) An estimator for  $p$  is  $\hat{p} = X/n$ . Exact distribution for  $X$  is binomial  $\mathcal{B}(n, p)$ . Since  $n = 157$  is large, normal approximation to binomials hold and  $X$  has approximately normal distribution with mean  $np$  and variance  $npq$ . Thus,  $\hat{p} = X/n$  has approximately normal distribution with expectation  $np/n = p$  and variance  $npq/n^2 = pq/n$ .

(ii) From approximation in 4.1 the  $(1 - \alpha) \times 100\%$  confidence interval is:

$$\left[ \hat{p} - z_{1-\alpha/2} \sqrt{\hat{p} \hat{q}/n}, \hat{p} + z_{1-\alpha/2} \sqrt{\hat{p} \hat{q}/n} \right].$$

(iii) The sample size needed is  $n = 3007$ . Note that we conservatively took  $\hat{p} = \hat{q} = 0.5$  since sample size is prospective in nature and  $\hat{p}$  is not observed. A good estimate  $\hat{p}$  is justified especially if the population proportion  $p$  is close to either 0 or 1.

(iv)



```

phat = 0.5541; p0 = 0.5;
qhat = 1- phat; q0 = 1 - p0;
n = 157;
z = (phat - p0)/sqrt( p0 * q0/n )
%z = 1.3557
pvalue = 1-normcdf(1.3557)
% pvalue = 0.0876

```

At significance level  $\alpha = 5\%$  the hospital cannot support their claim –  $H_0$  is not rejected. If  $\alpha = 0.10$ ,  $H_0$  is rejected, the hospital's claim is statistically supported.

### 7.18 Right to Die.



```

n = 1528; X = 1238; phat=X/n
%phat = 0.8102
[phat - norminv(0.995)*sqrt(phat*(1-phat)/n),...
 phat + norminv(0.995)*sqrt(phat*(1-phat)/n)]
%sans = 0.7844 0.8360

```



```

L=2*0.01;
n = 4*norminv(0.975)^2*0.8*0.2 /L^2
%n = 6.1463e+003
% Take sample of size 6147

```

### 7.19 Exponentials Parameterized by the Scale.



(i) Since  $\mathbb{E}X = \lambda$ , the simplest moment matching estimator is  $\hat{\lambda} = \bar{X}$ . Since the variance is  $\lambda^2$ , another moment matching estimator would be  $\hat{\lambda} = \sqrt{s^2} = s$ , where  $s$  is the sample standard deviation.

The MLE is  $\bar{X}$ . Indeed,

$$L(\lambda) = \prod_{i=1}^n \frac{1}{\lambda} e^{-\frac{X_i}{\lambda}} = \frac{1}{\lambda^n} e^{-\frac{\sum_{i=1}^n X_i}{\lambda}}.$$

By taking natural logs we obtain,

$$\log L(\lambda) = \ell(\lambda) = 0 - n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^n X_i.$$

In order to find the  $\lambda$  at which the likelihood  $L(\lambda)$ , or equivalently, the log-likelihood  $\ell(\lambda)$ , is maximized, we take the derivative of  $\ell(\lambda)$  with respect to  $\lambda$  and set it equal to 0,

$$-\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^n X_i = 0.$$

By solving this equation with respect to  $\lambda$  we ultimately obtain that MLE is  $\hat{\lambda} = \hat{X}$ .

Note that MLE coincides with the moment matching estimator (corresponding to the first moment).

(ii) Since  $E(Y_1) = \frac{\lambda}{2}$  and  $E(Y_2) = 2\lambda$ , many selections of constants  $w_1$  and  $w_2$  will make  $\hat{\lambda} = w_1 Y_1 + w_2 Y_2$  an unbiased estimator, and the solution is not unique. Indeed, the equation  $\mathbb{E}\hat{\lambda} = w_1 \cdot \lambda/2 + w_2 \cdot 2\lambda = \lambda$  has a continuum of many solutions with respect to  $w_1$  and  $w_2$ .

The problem asks for a specific linear combination and a possible choice could be  $\hat{\lambda} = Y_1 + \frac{Y_2}{4}$ , ( $w_1 = 1, w_2 = 1/4$ ). For such a choice of  $w$ 's, the variance of  $\hat{\lambda}$  is:  $Var(\hat{\lambda}) = 1^2 \cdot (\frac{\lambda}{2})^2 + \frac{1}{16} \cdot (2\lambda)^2 = \frac{\lambda^2}{2}$ .

More generally, if we consider only non-negative weights  $w_1$  and  $w_2$ , any point from the line segment  $\frac{w_1}{2} + \frac{w_2}{1/2} = 1$  in the first quadrant  $w_1 O w_2$  will make the estimator  $\hat{\lambda}$  unbiased. The variance of such a general estimator is  $w_1^2(\frac{\lambda}{2})^2 + w_2^2(2\lambda)^2$ . Replacing  $w_1 = 2 - 4w_2$  (recall  $w_1$  and  $w_2$  are on the line segment) and taking the first derivative with respect to  $w_2$ , we obtain that minimum variance achieved at  $w_2 = 1/4$ , which corresponds to our original choice:  $w_1 = 1$  and  $w_2 = 1/4$ .

Legitimate choices of weights are also  $w_1 = 0, w_2 = 1/2$  leading to  $\hat{\lambda} = \frac{Y_2}{2}$ , as well as,  $w_1 = 2, w_2 = 0$  leading to  $\hat{\lambda} = 2Y_1$ . What are variances of these two estimators?

(iii) By simple inspection only  $p = 1/2$  will make the estimator  $\hat{\lambda} = pZ_1 + (1-p)Z_2$  unbiased and such an estimator trivially minimizes the (absolute value of) magnitude of bias. For  $p = 1/2$  the bias of  $\lambda$  is 0.

The variance of  $\hat{\lambda}$  is

$$\text{Var}(\hat{\lambda}) = p^2(1.1\lambda)^2 + (1-p)^2(0.9\lambda)^2.$$

Taking the derivative with respect to  $p$  we conclude that the minimum of variance is achieved as the solution of equation  $2.42p - 2(1-p)0.81 = 0$ . The solution is:  $p = 1.62/4.04 = 0.401$ .

## 7.20 Bias in Estimator for Exponential $\lambda$ .



## 7.21 Yucatan Miniature Pigs.

The solution is not unique. One can match variance of Beta,  $ab/((a+b+1)(a+b)^2)$  to  $s^2$  and solve for  $a$  assuming that  $b = a$ . Result.  $\hat{a} = 1/(8s^2) - 1/2 = 2.8711$ , where  $s^2$  is the sample variance of  $x$ .

## 7.22 Computer Games.



## 7.23 Effectiveness in Treating Cerebral Vasospasm.

 TBA

**7.24 Alcoholism and Blyth-Still Confidence Interval.**
 TBA

**7.25 Spores of *Amanita Phalloides*.**



```

amanita =[ 9.2, 8.8, 9.1, 10.1, ...
           8.5, 8.4, 9.3, 8.7, ...
           9.7, 9.9, 8.4, 8.6, ...
           8.0, 9.5, 8.8, 8.1, ...
           8.3, 9.0, 8.2, 8.6, ...
           9.0, 8.7, 9.1, 9.2, ...
           7.9, 8.6, 9.0, 9.1];
s2 = var(amanita)
% s2 = 0.3033

n=length(amanita)
% n = 28

[(n-1) * s2/chi2inv(0.95, n-1), (n-1) * s2/chi2inv(0.05, n-1)]
% ans = 0.2042 0.5071

ratint = @(n) chi2inv(0.95, n-1)./chi2inv(0.05, n-1);

ratint(315:320)
%ans = 1.3007 1.3002 1.2996 1.2991 1.2985 1.2980

ratint(317)
% ans = 1.2996

xbar = mean(amanita); s = sqrt(s2); zquant = norminv(0.975);

s/xbar
% ans = 0.0622

Lb = s/xbar - zquant * s/xbar * sqrt( (1/2 + (s/xbar)^2)/(n-1))
% Lb = 0.0456

Ub = s/xbar + zquant * s/xbar * sqrt( (1/2 + (s/xbar)^2)/(n-1))
% Ub = 0.0789

[Lb, Ub]
% ans = 0.0456 0.0789

```

**7.26 CLT-Based Confidence Interval for Normal Variance.**
 TBA

### 7.27 Stent Quality Control.

 (a) The distribution of  $X$  is Binomial with  $n = 50$  and  $p = 0.01$ , i.e.,  $\mathcal{B}(50, 0.01)$ , with the expectation  $\mathbb{E}X = 50 \cdot 0.01 = 0.5$ , and the variance is  $\text{Var } X = 50 \cdot 0.01 \cdot 0.99 = 0.495$ . The standard deviation of  $X$  is  $\sqrt{\text{Var}(X)} = \sqrt{0.495} = 0.70356$ .

Exceeding the mean plus three standard deviations is critical in the context of this example. The critical point is:  $np + 3 \cdot \sqrt{npq} = 0.5 + 3 \cdot 0.70356 \approx 2.61$ . Since  $X$  is integer-valued, i.e., takes values 0,1,2,3,4, etc., the process might be problematic when  $X \geq 3$ .

(b) If one uses exact Binomial distribution, the desired probability is  $\mathbb{P}(X \geq 3) = 1 - P(X \leq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) = 1 - \binom{50}{0}0.01^0 \cdot 0.99^{50} - \binom{50}{1}0.01^1 \cdot 0.99^{49} - \binom{50}{2}0.01^2 \cdot 0.99^{48} = 1 - 0.60500 - 0.30556 - 0.07562 = 0.01382 \approx 0.014$ .

If one uses Poisson Approximation to the binomial (recall  $n$  is large and  $p$  is small), then  $\lambda = np = 0.5$  and,  $\mathbb{P}(X \geq 3) = 1 - P(X \leq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) = 1 - \frac{0.5^0}{0!}e^{-0.5} - \frac{0.5^1}{1!}e^{-0.5} - \frac{0.5^2}{2!}e^{-0.5} = 1 - 0.60653 - 0.30326 - 0.07582 = 0.01439 \approx 0.014$ .

The normal approximation is possible, as well, but is not very precise because in this case,  $X \sim \mathcal{N}(0.5, 0.70356^2)$ .  $\mathbb{P}(X \geq 3) = \mathbb{P}(Z \geq \frac{3-0.5-1/2}{0.70356}) = \mathbb{P}(Z \geq 2.84269) = 1 - \Phi(2.84) = 0.0023$ . Such a poor approximation (compared to the exact value of 0.01381727083060) is expected since normal asymptotics require the condition:  $\min\{np/q, nq/p\} > 5$ .

(c) Since the observation  $X$  comes from binomial  $\mathcal{B}(50, p)$  distribution with  $p$  unknown, a good estimator is the sample proportion  $\hat{p} = X/50$ . This estimator is unbiased since,  $\mathbb{E}\hat{p} = \mathbb{E}X/50 = 50 \cdot p/50 = p$ .

The  $\hat{p}$  is moment matching (first moment) and MLE estimator. In MATLAB

```
 data = binornd(20,0.75,[100,1]); % Simulated data, p = 0.75
[phat,pci] = mle(data,'distribution','binomial',...
    'alpha',.05,'ntrials',20)
```

### 7.28 Right to Die.

 The margin of error is standardly assumed to be  $L/2$  in the 95% confidence interval. Since  $n \geq \frac{4 \cdot 1.96^2 \cdot 0.8 \cdot 0.2}{0.04^2} = 1536.6$ , one should sample  $n = 1537$  students.

### 7.29 Clopper-Pearson and 3/n-Rule Confidence Intervals.



```
 % Clopper-Pearson and 3/n-rule CI's when X=0
%Hint
```

```

threenrule = @(n,alpha) -log(alpha)./n;
clopper = @(n,alpha) 1-(alpha/2).^(1./n);
n = (10:10:200)'; alpha = 0.05;
[clopper(n,alpha) threenrule(n, alpha)]

```

### 7.30 Seventeen Pairs of Rats, Carbon Tetrachloride, and Vitamin B.



```

% Seventeen Pairs of Rats, Carbon Tetrachloride, and Vitamin B12
%(b)
binopdf(7, 17, 7/17)

%(c)
lb = 7/17 - norminv(0.975) * sqrt( (7/17 * 10/17)/17 ) %lb=0.1778
ub = 7/17 + norminv(0.975) * sqrt( (7/17 * 10/17)/17 ) %ub=0.6457

lb1 = 140/340 - norminv(0.975) * sqrt( (140/340 * 200/340)/340 )
%lb1 = 0.3595
ub1 = 140/340 + norminv(0.975) * sqrt( (140/340 * 200/340)/340 )
%ub1 = 0.4641

%(d)
ub - lb %ans =0.4679 too large!
%
n = norminv(0.975)^2/ 0.2^2 %n=96.0365
ssize = ceil(n) %ssize=97

```

### 7.31 Hemocytometer Counts.



### 7.32 Predicting Alkaline Phosphatase.



## 7.1 Additional Problems

**7.a1 Bernoulli's  $p^2$ .** Let  $X_1, X_2, \dots, X_n$  be a sample from Bernoulli  $\mathcal{B}er(p)$  distribution, where parameter  $p^2$  is to be estimated. The MLE is  $\delta = (\bar{X})^2$ .

- (a) What is the bias of  $\delta$ ?
- (b) What is the variance of  $\delta$ ?



It is known that  $\bar{X}$  is the MLE for  $p$ , where  $E\bar{X} = p$  and  $Var\bar{X} = p(1-p)/n$ . Thus,  $(\bar{X})^2$  is the MLE for  $p^2$  according to the invariance property of MLEs (page 235). According to (5.11), with  $\mu = p$ ,  $\sigma^2 = p(1-p)/n$  and  $g(x) = x^2$ ,

$$E\delta = p^2 + 1/2 \cdot 2 \cdot p(1-p)/n = p^2 + p(1-p)/n,$$

so the bias is  $p(1-p)/n$ .

(b) The variance is  $(2p)^2 \cdot p(1-p)/n = 4p^3(1-p)/n$ .

**7.a2 Shrinking  $\bar{X}$  Lowers MSE.** Let  $X_1, X_2, \dots, X_n$  be a sample from a distribution with mean  $\theta$  and known variance  $\sigma^2$ . The standard estimator of  $\theta$ ,  $\bar{X}$ , has MSE (and variance because  $\bar{X}$  is unbiased) equal to  $\sigma^2/n$ .

Show that MSE of  $\lambda\bar{X}$  is minimized by  $\lambda^* = \frac{\theta^2}{\theta^2 + \sigma^2/n} < 1$ . Thus, the shrinkage estimator  $\lambda^*\bar{X}$  lowers MSE which is  $\lambda^*\sigma^2/n$ .



$$\begin{aligned} MSE &= E(\lambda\bar{X} - \theta)^2 \\ &= \lambda^2 E\bar{X}^2 - 2\lambda\theta E\bar{X} + \theta^2 \\ &= \lambda^2(\sigma^2/n + \theta^2) - 2\lambda\theta^2 + \theta^2 \end{aligned}$$

as quadratic in  $\lambda$  is minimized for  $\lambda^* = \frac{\theta^2}{\sigma^2/n + \theta^2}$ .

For the value  $\lambda = \lambda^*$  the MSE becomes  $\lambda^*\sigma^2/n$ . Of course, in practice  $\theta$  is not known (it is estimated), and “plug-in” shrinker  $\hat{\lambda}^* = \frac{\bar{X}^2}{\sigma^2/n + \bar{X}^2}$  is used.



# Chapter 8

## Bayesian Approach to Inference

### 8.1 Exponential Lifetimes.

 TBA

### 8.2 Uniform - Pareto.

 TBA

### 8.3 Nylon Fibers.



(a) If  $T$  is exponential  $\mathcal{E}(\lambda)$  where  $\lambda$  is the rate parameter, then  $ET = 1/\lambda$ .  
The moment matching estimator is  $\hat{\lambda}_{mm} = 1/\bar{T}$ .

Here  $\bar{T} = \frac{3+13+8}{3} = 8$ , so  $\hat{\lambda}_{mm} = 1/8 = 0.125$ .

The likelihood is:

$$\lambda e^{-3\lambda} \times \lambda e^{-13\lambda} \times \lambda e^{-8\lambda} = \lambda^3 e^{-24\lambda}.$$

(b) The posterior for  $\lambda$  is proportional to the likelihood  $\times$  prior,

$$\lambda^3 e^{-24\lambda} \times \lambda^{-1/2} = \lambda^{5/2} e^{-24\lambda} = \lambda^{7/2-1} e^{-24\lambda},$$

which can be recognized as Gamma  $\mathcal{G}a(7/2, 24)$  distribution. Since the mean of  $\mathcal{G}a(\alpha, \beta)$  is  $\alpha/\beta$ , the mean of the posterior is

$$\hat{\lambda}_B = \frac{7/2}{24} = 0.1458,$$

which is the Bayes estimator, in this case.

(c) TBA

### 8.4 Gamma - Inverse Gamma.



The likelihood for  $X \sim \mathcal{G}a\left(\frac{n}{2}, \frac{1}{2\theta}\right)$  and the prior  $\theta \sim \mathcal{IG}(\alpha, \beta)$  are proportional to

$$\frac{1}{(2\theta)^{n/2}} \exp\left\{-\frac{x}{2\theta}\right\}, \text{ and } \frac{1}{\theta^{\alpha+1}} \exp\left\{-\frac{\beta}{\theta}\right\},$$

respectively, if all constant terms are ignored. The product is proportional to

$$\frac{1}{\theta^{n/2+\alpha+1}} \exp\left\{-\frac{x/2 + \beta}{\theta}\right\},$$

which can be recognized as the inverse gamma  $\mathcal{IG}\left(\frac{n}{2} + \alpha, \frac{x}{2} + \beta\right)$  distribution.

### 8.5 Negative Binomial - Beta.



### 8.6 Poisson - Gamma Marginal.



### 8.7 Exponential - Improper.



### 8.8 Normal Precision - Gamma.



(a) The likelihood is proportional to

$$L(\theta) \propto \sqrt{\theta} \exp\left\{-\frac{\theta x^2}{2}\right\}.$$

The log-likelihood is

$$\ell(\theta) = \frac{1}{2} \log \theta - \frac{x^2 \theta}{2} + \text{constant},$$

with first derivative

$$\ell'(\theta) = \frac{1}{2\theta} - \frac{x^2}{2}.$$

Solution of  $\ell'(\theta) = 0$  is  $\hat{\theta} = \frac{1}{x^2}$ , which represents the candidate for MLE estimator of  $\theta$ . Since  $\ell''(\theta) = -\frac{1}{3\theta^2} < 0$ , the likelihood is maximized at  $\frac{1}{x^2}$  and  $\hat{\theta}$  represents the MLE.

(b) If the prior is  $\theta \sim \mathcal{G}a(r, \lambda)$ , then the posterior is proportional to

$$\pi(\theta|x) \propto \theta^{r+1/2-1} \exp\left\{\lambda + \frac{x^2}{2}\right\},$$

which can be recognized as gamma  $\mathcal{G}a(r+1/2, \lambda+x^2/2)$  distribution. The Bayes estimator for  $\theta$  is the mean of posterior,

$$\hat{\theta}_B = \frac{r + 1/2}{\lambda + x^2/2} = \frac{2r + 1}{2\lambda + x^2}.$$

This Bayes estimator could be represented as a compromise between MLE and prior mean but with the weights depending on the observation:

$$\hat{\theta} = \frac{x^2}{2\lambda + x^2} \times \frac{1}{x^2} + \frac{2\lambda}{2\lambda + x^2} \times \frac{r}{\lambda}.$$

Note that in the case when  $X = 0$  the MLE is not defined, but its weight is 0, and the precision is estimated by the prior mean. The representation as a linear combination of MLE and prior mean with weights free of  $X$  is not possible, although one can represent the Bayes estimator as

$$\frac{1}{\omega \times \frac{1}{\hat{\theta}_{mle}} + (1 - \omega) \times \frac{1}{\hat{\theta}_{prior}}}, \quad \omega = \frac{1}{2r + 1}.$$

(c) When  $X = -2$  is observed, and  $r = 1/2$  and  $\lambda = 1$ , the posterior becomes gamma with shape parameter 1 and rate parameter 3, which is in fact the exponential distribution  $\mathcal{E}(3)$ . Indeed,

$$\pi(\theta|x = -2) \propto \theta^{1/2+1/2-1} \exp \left\{ 1 + \frac{(-2)^2}{2} \right\} = e^{-3\theta}.$$

The Bayes estimator is the mean of the posterior, in this case  $1/\lambda = 1/3$ . The equal-tail credible set is found by evaluation quantiles of the posterior. The  $p$ -quantile of exponential distribution,  $q_p$ , is easy to find by directly solving an equation involving the cdf:  $F(q_p) = p$  i.e.,  $1 - e^{-\lambda p_q} = p$ . Thus, the 0.025- and 0.975-quantiles when  $\lambda = 3$  are

$$q_{0.025} = -\frac{\log(1 - 0.025)}{\lambda} = 0.0084, \quad q_{0.975} = -\frac{\log(1 - 0.975)}{\lambda} = 1.2296,$$

which are lower and upper bounds of the equal-tail 95% credible set for  $\theta$ .

(d) The WinBUGS program approximating estimators from (c) is simple,



```
model
  x ~ dnorm(0, theta)
  theta ~ dgamma(0.5, 1)
```

```
data
  list(x=-2)

inits
  list(theta = 1)
```

The output is

```
mean sd MC_error val2.5pc median val97.5pc start sample
theta 0.3326 0.3312 0.001101 0.008439 0.2315 1.217 1001 100000
```

The MCMC approximation of Bayes' estimator for  $\theta$  is 0.3326, quite close to the exact value of 1/3. Also, the 95% credible set is [0.00844, 1.217], which is close to the exact set [0.0084, 1.2296].

### 8.9 Bayes Estimate in a Discrete Case.



### 8.10 Histocompatibility.



Gamma  $\mathcal{G}a(r, \mu)$  distribution for  $\lambda$  has a density

$$\pi(\lambda) = \frac{\mu^r \lambda^{r-1} \exp\{-\mu\lambda\}}{\Gamma(r)}, \lambda > 0.$$

Here  $r = 2$  and  $\mu = 1$ , so  $\pi(\lambda) = \lambda e^{-\lambda}$ , since  $\Gamma(2) = 1$ .

The likelihood is Poisson,  $f(x|\lambda) = \frac{\lambda^x}{x!} \exp\{-\lambda\}$ , and since  $X = 1$  is observed, the likelihood is  $\lambda e^{-\lambda}$ .

The posterior is proportional to the product of the likelihood and prior,

$$\lambda e^{-\lambda} \times \lambda e^{-\lambda} = \lambda^2 e^{-2\lambda}.$$

From this expression we conclude that the posterior is Gamma  $\mathcal{G}a(3, 2)$ . For any  $Y \sim \mathcal{G}a(r, \mu)$ , the mean  $EY$  is  $r/\mu$ . Thus, the posterior mean is  $3/2=1.5$ , and this is a Bayes estimator of  $\lambda$ . The posterior variance is  $3/2^2$  and posterior standard deviation is  $\sqrt{3}/2 = 0.8660$ .

The supplied WinBUGS program gives the following MCMC approximation to the solution:

```
mean sd      MC_error val2.5pc median val97.5pc start sample
lambda 1.495  0.863  0.002706  0.3107      1.332      3.609  10001 100000
```

The median is 1.332 and the 95% credible set for  $\lambda$  is  $[-0.3107, 3.609]$ .

### 8.11 Neurons Fire in Potter's Lab 2.



(a) The likelihood is proportional to  $\lambda^{\sum_{i=1}^{50} X_i} \exp\{-50\lambda\}$ , where  $\sum X_i = 989$  is the sum of all counts (total number of firings).

(b) A gamma prior with mean 15 is not unique, for any  $x$ ,  $\mathcal{G}a(15x, x)$  is such a prior. However, the variances depend on  $x$ , For example for priors  $\mathcal{G}a(150, 10)$ ,  $\mathcal{G}a(15, 1)$ ,  $\mathcal{G}a(1.5, 0.1)$ ,  $\mathcal{G}a(0.15, 0.01)$ , etc. have variances 1.5, 15, 150, 1500, etc. The variances indicate the degree of certainty of expert that the prior mean is 15. Large variances correspond to non-informative choices.

Since the sample variance of 50 observations is about 15, it is reasonable to take prior with larger variance, say  $\mathcal{G}a(3, 0.2)$ .

(c) Bayes estimator for  $\lambda$  is  $w \times \bar{X} + (1-w) \times 15$  = where  $w = xxxx$ . The MLE is  $\bar{X}$  and Bayes estimator slightly shrinks toward 0.

(d) The expectation of the lognormal is  $\exp\{\mu + \sigma^2\}$ . If  $\sigma^2 = 1$  then  $\mu = \log(15) - 1/2 = 2.2081$  gives the expectation 15.

```


model
for (i in 1:50)
    X[i] ~ dpois(lambda)

lambda ~ dlnorm(2.2081, 1)
#mu = 2.2081, tau =1 => mean 15

DATA
list(X=c(
20, 19, 26, 20, 24, 21, 24, 29, 21, 17,
23, 21, 19, 23, 17, 30, 20, 20, 18, 16,
14, 17, 15, 25, 21, 16, 14, 18, 22, 25,
17, 25, 24, 18, 13, 12, 19, 17, 19, 19,
19, 23, 17, 17, 21, 15, 19, 15, 23, 22))

INITS
list( lambda = 5)

```

**8.12 Eliciting a Beta Prior I.** TBA**8.13 Eliciting a Beta Prior II.** TBA**8.14 Eliciting a Weibull Prior.** TBA**8.15 Bayesian Yucatan Pigs.**

```


model
for (i in 1:nc)
    x[i] ~ dbeta(a, a)

a ~ dgamma(0.001, 0.001)

DATA
list(nc=120, x = c(
0.6121, 0.5789, 0.6053, 0.6168, 0.6237, 0.5837, 0.6500, 0.6274,
0.6726, 0.5163, 0.5374, 0.5258, 0.5374, 0.5405, 0.5184, 0.7179,
0.7332, 0.5716, 0.7521, 0.7232, 0.6884, 0.5532, 0.5268, 0.5211,
0.5484, 0.5821, 0.6205, 0.7742, 0.6421, 0.6842, 0.7405, 0.6879,
0.6532, 0.8768, 0.8221, 0.8421, 0.7853, 0.8758, 0.7853, 0.6726,
0.6411, 0.7216, 0.7416, 0.6837, 0.6879, 0.3979, 0.5789, 0.2547,
0.2758, 0.2800, 0.2495, 0.4968, 0.5679, 0.2953, 0.5679, 0.5111,
0.6884, 0.4253, 0.4095, 0.7279, 0.6789, 0.4884, 0.6858, 0.2500,
0.3405, 0.2211, 0.3547, 0.3863, 0.2674, 0.3974, 0.4921, 0.3047,

```

```
0.8495, 0.4668, 0.3832, 0.1616, 0.1137, 0.0479, 0.3347, 0.6700,
0.3168, 0.3232, 0.3711, 0.3437, 0.3021, 0.7853, 0.5000, 0.5711,
0.3984, 0.2211, 0.1479, 0.1026, 0.2737, 0.3632, 0.2279, 0.3184,
0.4321, 0.3805, 0.7626, 0.5111, 0.4316, 0.4705, 0.7974, 0.6926,
0.8200, 0.9589, 0.6000, 0.5684, 0.6489, 0.6311, 0.5032, 0.5032,
0.5779, 0.4774, 0.3074, 0.4995, 0.4384, 0.2942, 0.3132, 0.4205))
```

```
INITS
list(a =1)
```

### 8.16 Eliciting a Normal Prior.

  $\mu = 3.99382 \approx 4$ ,  $\sigma = 1.53734$ .

### 8.17 Is the Cloning of Humans Moral?



```
model {    anticlons ~ dbin(prob,npolled) ;
  anticlons.missing ~ dbin(prob,nmissing)
  prob ~ dbeta(1,1)}

Data
list(anticlons=882,npolled= 1000, nmissing=62)
```

### 8.18 Poisson Observations with Truncated Normal Rate.

 TBA

### 8.19 Counts of Alpha Particles.

 Consult the file rutherford.odc.

### 8.20 Rayleigh Estimation by Zero-Trick.

 TBA

### 8.21 Predictions in Poisson/Gamma Model.

 TBA

### 8.22 Estimating Chemotherapy Response Rate.

 TBA

## 8.1 Additional Problems

**8.a1 Fibrinogen.** Fibrinogen is a soluble plasma glycoprotein, synthesized by the liver, that is converted by thrombin into fibrin during blood coagulation. Marnie takes blood test and finds that her level of fibrinogen is 207 mg/dL. The

test results are accurate up to a random error which is normal with mean 0 and standard deviation of 12 mg/dL.

The normal range of fibrinogen is 150-400 mg/dL and Marnie puts a uniform prior over this range,  $dunif(150, 400)$ .

(a) What is the Bayes estimator of the true level of fibrinogen given this uniform prior?

(b) Report the Inference>Samples>stats output from WinBUGS. What is the 95% Credible Set for the parameter?

(c) What is the classical 95% CI (Hint: Sample Size = 1,  $\sigma$  known). Compare Bayesian answers with classical (Compare the parameter estimates and 95% CI with Bayesian counterparts).

 TBA

**8.a2 Elicitation of Gamma Prior.** You are eliciting Gamma prior on  $\theta$ ,

$$\pi(\theta) \propto \theta^{\alpha} \exp\left\{-\frac{\theta}{\beta}\right\}, \quad \theta \geq 0, \beta > 0.$$

An expert tells you that the “most probable” value for  $\theta$  is 2. If you interpret the “most probable” as the mode of this prior, fully specify the prior.



# Chapter 9

## Testing Statistical Hypotheses

### 9.1 Public Health.

 TBA

### 9.2 Testing IQ.

 TBA

### 9.3 Bricks.



```
 %(a)
n = 100; alpha = 0.05; Xbar = 395; s=20; mu0 = 400;
t = (Xbar - mu0)/(s/sqrt(n))
%t = - 2.5
% RR
% H_1: mu < mu0, RR = (-infinity, tinv(alpha, n-1))
tinv(alpha, n-1)
% ans = - 1.6604
% RR = (- infy, -1.6604), statistics t in RR, reject H0
%
% pvalue
p = tcdf(t, n-1)
% p = 0.0070

%(b)
% normal approx
n = 4 * norminv(0.975)^2 * 20^2/4^2
% n = 384.1459 approx 385.
% exact
f = @(n) n - 4 * tinv(0.975, n-1).^2 * 20^2/4^2
fzero(f, 500)
% ans = 386.5689
% n approx 387
```

**9.4 Soybeans.**

```
p-value= 0.295 is larger than alpha= 0.05 .
t-statistic= -1.058 .
The rejection region cut-point is (+/-) 1.677 .
```

**9.5 Great white shark.****9.6 Serum Sodium Levels.**

$$[t = \frac{145.55 - 140}{9.455/\sqrt{20}} = 2.625104.]$$

**9.7 Weight of Quarters.****9.8 Dwarf Plants.**


```
p0=0.75; q0=1-p0; n=200;
z = (phat - p0)/sqrt( p0*q0/n)
%z = -2.2862
normcdf(z)
% ans = 0.0111 (p-value against one sided hypothesis)
norminv(0.05)
%ans = -1.6449 (RR=(-infinity, -1.65])
%
lb = phat - norminv(1-0.05/2)*sqrt(phat * (1-phat)/n )
%lb = 0.6154
rb = phat + norminv(1-0.05/2)*sqrt(phat * (1-phat)/n )
%rb = 0.7446
```



```
model{
  X ~ dbin(p, n)
  p ~ dbeta(1,1)
  probH1 <- step(0.75-p)
  probH0 <- 1-probH1
}

DATA

list( n= 200, X = 136)
```

```
INIT
```

```
list(p=0)
```

### 9.9 Eggs in Nest.



$[\bar{X} = 4.53, (4.528571), s^2 = 1.093, t = -3.745, t_{69,0.925} \approx z_{0.975} = 1.96.]$

### 9.10 Penguins.



$\bar{X} = 44, s = 2.1122, t = -1.7714, p\text{-value}=0.0500, t_{0.05,13} = -1.7709$ . No decision at significance level  $\alpha = 0.05$ .

### 9.11 Hypersplenism and White Blood Cell Count.



The solution in MATLAB



```
n=16; xbar=5213; mu0 = 7200; s=1682; alpha = 0.05;
t = (xbar - mu0)/(s/sqrt(n))
%t = -4.7253
tcdf(t,n-1)      %p-value
%ans = 1.3543e-004
tinv(alpha, n-1) %RR bound
%ans = -1.7531

% Find the power against alternative H_1: mu=5800

esizet = abs(7200-5800)/s;
powert = nctcdf( -tinv(1-alpha, n-1), n-1,-esizet*sqrt(n))
% powert = 0.9369
ttgrc%
%power 90%, alpha 5% one sided,
%for the effect size 600/1682 = 0.3567
beta = 0.1;

%Approx sample size
ss = (norminv(1-beta) + norminv(1-alpha))^2/0.3567^2
%sss = 67.3074
%Exact sample size
f = @(n) nctcdf(-tinv(1-alpha, n-1),n-1,-sqrt(n)*0.3567)-(1-beta);
sss = fzero(f,ss)
%%sss = 68.6830
```

### 9.12 Jigsaw.



*Sol. (a)  $[5.783 \pm 2.1098 \frac{2.784}{\sqrt{18}}] = [4.3986, 7.1674]$ . (b)  $t = 1.1932$ ,  $t_{17,0.95} = 1.7396$ ,  $t < t_{17,0.95}$  - Do not Reject.*

### 9.13 Anxiety.

 TBA

### 9.14 Aptitude Test.

 TBA

### 9.15 Rats and Mazes.



```
 xbar = 15.4; mu0 = 15; s=2; mu1 = 15.5;
t = (xbar - mu0)/(s/sqrt(80))
%t = 1.7889
crit = tinv(0.99, 79)
%crit = 2.3745
pval = 1-tcdf(2.3745,79)
%pval = 0.0100
pval = 1-tcdf(1.7889,79)
%pval = 0.0387
pow = normcdf( norminv(0.01) + 0.6*sqrt(80)/s )
%pow = 0.6394
ss = 2^2 * (norminv(0.99) + norminv(0.90))^2/(0.6^2)
%ss = 144.6326
```

### 9.16 Hemopexin in DMD Cases I.

 TBA

### 9.17 Retinol and Cooper-deficient Diet.



(a) Since population variance  $\sigma^2$  is not known, we use  $t$ -quantiles in the confidence interval.

```
 xbar = 3.3; s=1.4; n=9; conf=0.95;
alpha = 1 - conf;
int =[xbar - tinv(1-alpha/2,n-1) * s/sqrt(n), ...
       xbar + tinv(1-alpha/2,n-1)*s/sqrt(n)]
%int = 2.2239 4.3761
```

The 95% CI for the unknown mean is  $[2.2239, 4.3761]$ .

(b) We test the hypothesis  $H_0 : \mu = 1.6$  versus the alternative  $H_1 : \mu > 1.6$ .



```
xbar = 3.3; s=1.4; n=9; mu0 = 1.6;
t = (xbar - mu0)/(s/sqrt(n))
%t = 3.6429
%RR approach
tinv(1-alpha, n-1)
```

```
%ans = 1.8595
%t=3.6429 > 1.8595, reject H_0
%p-value approach
p = 1-tcdf(t, n-1)
%p = 0.0033
%0.0033 < 0.05 reject H_0
```

(c) The power of the test is  $1 - \beta = \Phi\left(z_\alpha + \frac{|\mu_1 - \mu_0| \sqrt{n}}{\sigma}\right)$ .



```
alpha=0.05; mu0=1.6; mu1=2.4; n=9; sigma=1.4;
power = normcdf( norminv(alpha) + abs(mu0 - mu1)*sqrt(n)/sigma)
% or: power = 1-normcdf( norminv(1-alpha) - abs(mu0 - mu1)*sqrt(n)/sigma)
%power = 0.5277
```

Not much power is achieved with a sample of size  $n = 9$ ,  $1 - \beta \approx 53\%$ . Even this 53% is an optimistic assessment of the power.

More precise determination of power is done using t-distribution instead of normal.



```
alpha=0.05; mu0=1.6; mu1=2.4; n=9; sigma=1.4;
power = 1-nctcdf( tinv(1-alpha, n-1),n-1,abs(mu0-mu1)*sqrt(n)/sigma)
% or power = nctcdf( tinv(alpha, n-1),n-1,-abs(mu0-mu1)*sqrt(n)/sigma)
%power = 0.4693
```

(d)

```
n = ( (norminv(0.95) + norminv(0.80))*1.4/(1.6 - 2.1))^2
%n = 48.4712
```

Sample size necessary for power of 80% is  $n = 49$ . If one wants to be precise:



```
alpha=0.05; mu0=1.6; mu1=2.1; sigma=1.4;
for n=10:100
[n 1-nctcdf( tinv(1-alpha, n-1),n-1,(mu1-mu0)*sqrt(n)/sigma)]
end

% ans = 10.0000    0.2743
%
% ans = 11.0000    0.2946
% ...
%
% ans = 49.0000    0.7938
%
% ans = 50.0000    0.8011
%
% ans = 51.0000    0.8081
%
% ans = 52.0000    0.8149
% ...
```

Thus, sample size needed for power of 80% is  $n = 50$  rather than  $n = 49$  if one uses exact calculations. More elegant solution is

```
f = @(n) 1-nctcdf( tinv(1-alpha, n-1),n-1,(mu1-mu0)*sqrt(n)/sigma) - 0.8
fzero(f, 100) %49.8523
```

If the two-sided alternative is selected,  $H_1: \mu \neq 1.6$ , then the p-value in 1.2 is  $p = 0.0066$ . The power in 1.3 is only about 40%.



```
alpha=0.05; mu0=1.6; mu1=2.4; n=9; sigma=1.4;
power = normcdf( norminv(alpha/2) + abs(mu0 - mu1)*sqrt(n)/sigma)
% power =0.4030
```

Also, for the two sided alternative the sample size is approximately  $n = 62$ .

```
n = ( (norminv(0.975) + norminv(0.80))*1.4/(1.6 - 2.1))^2 %
%n = 61.5352
```

(e)



```
model{
  precxbar <- n * precx
  xbar ~ dnorm( mu, precxbar )
  mu ~ dnorm(0, 0.0001)
  #s = 1.4, s^2 = 1.96, prec = 1/1.96 =0.51
  #X gamma(a,b) -> EX=a/b, Var X = a/b^2
  precx ~ dgamma( 0.00051, 0.001 )
  indh1 <- step(mu - 1.6)
  sigx <- 1/sqrt(precx)
}

list( xbar = 3.3, n=9 )

list( mu = 1, precx = 1 )
```

### 9.18 Aniline.



TBA

### 9.19 DNA Random Walks.



(a) The sample size is calculated as  $n = \left(\frac{z_{1-\alpha/2}+z_{1-\beta}}{\epsilon}\right)^2$ , or in MATLAB for the specified  $\alpha, 1 - \beta, \epsilon$ , and two-sided alternative, as

```
n = (0.03/0.02)^2 * (norminv(0.975) + norminv(0.9))^2 %n=23.6417
```

This  $n$  is to be rounded to larger integer, here  $n = 24$ , and sampling is to follow. The exact prospective power for  $n = 24$  observations is



```
pow = normcdf( -norminv(0.975) + 0.02/(0.03/sqrt(24)) )
%pow = 0.9042
%
tstat = (mean(H) - 0.6)/(std(H)/sqrt(24))
% tstat = -1.3938
```

```

    %pvalue
2*tcdf(tstat,n-1)
% 0.1767
% p-value > 5% -- do not reject H_0
%Rejection region
%(-infinity, tinv(0.025, 24-1)] U [tinv(0.975, 24-1),infinity)
tinv(0.025,n-1) % -2.0687
tinv(0.975,n-1) % 2.0687
%(-infinity, -2.0687] U [2.0687, infinity)

%(b) Because of symmetry of t-distribution, there are several
% equivalent ways of getting the retrospective power.
% It is found to be 89.28%

pow = nctcdf( -tinv(1-alpha/2, n-1), n-1,(mu1-mu0)*sqrt(n)/s ) + ...
      1 - nctcdf( tinv(1-alpha/2, n-1), n-1,(mu1-mu0)*sqrt(n)/s )
% 0.8928
pow2 = nctcdf( tinv(alpha/2, n-1), n-1,-abs(mu1-mu0)*sqrt(n)/s ) + ...
      1- nctcdf(-tinv(alpha/2, n-1), n-1, -abs(mu0-mu1)*sqrt(n)/s )
% 0.8928
pow = nctcdf( -tinv(1-alpha/2, n-1), n-1,(mu1-mu0)*sqrt(n)/s ) + ...
      +nctcdf(-tinv(1-alpha/2, n-1), n-1,(mu0-mu1)*sqrt(n)/s )
% 0.8928
pow = nctcdf( tinv(alpha/2, n-1), n-1,(mu1-mu0)*sqrt(n)/s ) + ...
      +nctcdf(tinv(alpha/2, n-1), n-1,(mu0-mu1)*sqrt(n)/s )
% 0.8928
pow = nctcdf( -tinv(1-alpha/2, n-1), n-1,(mu1-mu0)*sqrt(n)/s ) + ...
      +nctcdf(-tinv(1-alpha/2, n-1), n-1,(mu0-mu1)*sqrt(n)/s )
% 0.8928
pow = nctcdf( tinv(alpha/2, n-1), n-1,-abs(mu1-mu0)*sqrt(n)/s ) + ...
      +nctcdf(tinv(alpha/2, n-1), n-1, abs(mu0-mu1)*sqrt(n)/s )
% 0.8928

```

## 9.20 Binding of Propofol.



(a)

```

pbar = 0.93; s=0.12; n=87; p0 = 0.96;
t = (pbar - p0)/(s/sqrt(n))
%t = -2.3318

tinv(0.05, n-1)
%ans = -1.6628

tinv(0.01, n-1)
%ans = -2.3705

pval= tcdf(-2.3318, n-1)
%pval = 0.0110

```

From the above calculations we see that  $H_0$  is rejected at 5% significance level since  $t \in (-\infty, -1.6628]$ . The  $H_0$  is not rejected at  $\alpha = 1\%$  since  $t = -2.3318$  does not fall in rejection region  $(-\infty, -2.3705]$ .

This is confirmed by looking at  $p$ -value.  $P$ -value of 0.0110 is smaller than 5% but larger than 1%.

(b) Normal approximation can be used because of Central Limit Theorem. In fact  $n = 87$  proportions are averaged. In this case,  $H_0$  is rejected even at 1% level since  $p$ -value is 0.0099.

```
 norminv(0.05)
%ans = -1.6449

norminv(0.01)
%ans = -2.3263

pval = normcdf(-2.3318)
%pval = 0.0099
```

### 9.21 Improvement of Surgical Procedure.



### 9.22 Cancer Therapy.



```
 %(a) H0 p = 0.4 vs H1: p > 0.4
%(b)
norminv(1-0.05) %1.6449; since z is in [1.6449, infinity)
% H0 rejected
%(c)
pval = 1 - normcdf(1.7321)  %0.0416 < 0.05, H0  rejected
%(d) see text page 339
p1 = 0.475;
n = p0*(1-p0)*(norminv(1-0.05) + norminv(0.85)*sqrt(p1*(1-p1)/(p0*(1-p0)))) )^2/(p0-p1)^2
%n=311.3479  approx 312
```

### 3.23 Is the Cloning of Humans Moral?



```
 clear all
n=1000; phat = 0.88;
p0=0.9; q0=1-p0;
Z = (phat - p0)/sqrt(p0 * q0/n)
%Z = -2.1082

crit = norminv(0.975)
```

```
%crit = 1.9600
% The rejection region is (-infinity, -1.96) U (1.96, infinity)
% and Z=-2.1082 falls in. Reject H_0.

pval = 2*normcdf(-2.1028)
% pval = 0.0355
% pval < 0.05 = alpha => Reject H_0.

n=1000; phat = 0.88; qhat=1-phat;
[phat - norminv(0.975)*sqrt(phat*qhat/n) ...
    phat + norminv(0.975)*sqrt(phat*qhat/n)]

% ans = 0.8599 0.9001
% 0.9 belongs to CI (tight!)

Recall that the power is  $1 - \beta = \Phi \left[ \sqrt{\frac{p_0 q_0}{p_1 q_1}} \left( z_{\alpha/2} + \frac{|p_0 - p_1| \sqrt{n}}{\sqrt{p_0 q_0}} \right) \right]$ .
normcdf( sqrt( 0.9*0.1/(0.85*0.15) ) * (norminv(0.025) +...
abs(0.9-0.85)*sqrt(1000)/sqrt(0.9*0.1)) )

%ans = 0.9973
```

**9.24 Smoking Illegal?**
 TBA
**9.25 DNA of Spider Monkey.**
 TBA



## Chapter 10

### Two Samples

#### 10.1 Testing Piaget.

 TBA

#### 10.2 Smoking and COPD.



We test hypotheses

$H_0: \mu_1 = \mu_2$  versus  $H_1: \mu_1 < \mu_2$ , that is  $H_1: \mu_1 - \mu_2 < 0$ .

Since the population variances are assumed equal we first find pooled standard deviation,

$$s_p = \sqrt{\frac{(9-1) \cdot 7,029^2 + (11-1) \cdot 7,534^2}{9+11-2}} = \sqrt{53,492,572} \approx 7,313.86.$$

Then,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{1/n_1 + 1/n_2}} = \frac{16,156 - 24,672}{7,313.86 \sqrt{1/9 + 1/11}} = -2.59.$$

The proper Rejection Region cut-point is  $tinv(0.05, 18)$ , since the statistic  $t$  has  $11 + 9 - 2 = 18$  degrees of freedom, and the rejection region is  $RR = (-\infty, -1.7341]$ . The statistic  $t$  falls in the RR and  $H_0$  is rejected at the level  $\alpha = 0.05$ . This agrees with the  $p$ -value approach since the  $p$ -value is  $tcdf(-2.59, 18) = 0.0092 < 0.05$ .

#### 10.3 Noradrenergic Activity.



(a)

```
x1bar = 279; x2bar = 198;
s1 = 122; s2 = 89; n1=17; n2 = 29;
```

```
t1 = (x1bar - 170)/(s1/sqrt(n1))
% t1 = 3.6838

pval1 = 1 - tcdf( t1, n1-1)
% pval1 = 0.0010

t2 = (x2bar - 170)/(s2/sqrt(n2))
% t2 = 1.6942

pval2 = 1 - tcdf( t2, n2-1)
% pval2 = 0.0507
```

Thus,  $H'_0$  is rejected (p-value 0.001) while  $H''_0$  is not rejected at 5% level (p-value = 0.0507).

(b)



```
s1 = 122; s2 = 89; n1=17; n2 = 29;
f = (s1^2)/(s2^2)
% f = 1.8791

pval = 2 * (1 - fcdf( 1.8791, n1-1, n2-1))
% pval = 0.1397
```

Hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2$  is not rejected, p-value is 0.1397. Thus, in testing equality of the means one should use pooled sample standard deviation.

(c)



```
x1bar = 279; x2bar = 198;
s1 = 122; s2 = 89; n1=17; n2 = 29;
%pooled
sp = sqrt( ((n1-1)*s1^2 + (n2-1)*s2^2 )/(n1 + n2 - 2 ))
% sp = 102.2399; should be between s1 and s2.

t = (x1bar - x2bar)/(sp * sqrt(1/n1 + 1/n2) )
% t = 2.5936

pval = 1-tcdf( t, n1+n2 -2) %alternative mu-nu>0
% pval = 0.0064

pval = 2*tcdf( -abs(t), n1+n2 -2) %alternative mu-nu diff 0
% pval = 0.0128
```

#### 10.4 Testing Variances.



(a)

```
2 * min( fcdf(f,n1-1, n2-1), 1- fcdf(f, n1-1, n2-1) )
%ans = 0.9727
```

(b) The problem is in the condition  $F > 1$ . The universally correct  $p$  value is obtained if the condition  $F > 1$  is replaced by  $F > \text{median}(F_{n_1-1, n_2-1})$ . The medians of  $F$ -distributions are generally close to 1, but range between 0.4549 and 2.1981, and all  $F$  statistics observed in this range may potentially lead to a wrong two-sided  $p$ -value.

```
%Observed value of F is
f = var(x)/var(y)
% f = 1.0429
%The criteria F > 1 suggest p-value
2 * (1 - fcdf(f, n1-1, n2-1))
% 1.0273 > 1
%The problem is that F < median(F(n1-1, n2-1))
med = finv(0.5, n1-1, n2 -1)
%med = 1.0687 > F > 1,
%and 2 * fcdf(f, n2-1, n1-1) should be used
```

### 10.5 Mating Calls.

 TBA

### 10.6 Fatigue.

 TBA

### 10.7 Mosaic Virus.

 TBA

### 10.8 Dopamine $\beta$ -hydroxylase Activity.



(a1) Solution when  $\sigma_1 = \sigma_2$  is assumed. Polled sample variance is  $s_p^2 = ((n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2) / (n_1 + n_2 - 2) = ((9 - 1) \times s_1^2 + (12 - 1) \times s_2^2) / (9 + 12 - 2) = 55.8399$ .

The polled sample standard deviation is  $s_p = \sqrt{s_p^2} = 7.4726$ .

Statistic:  $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{1/n_1 + 1/n_2}} = 4.3 / (7.4726 \times \sqrt{1/9 + 1/12}) = 1.3050$ .

The critical value is:  $t_{n_1+n_2-2,\alpha} = t_{19,0.05} = 1.7291$ .

Rejection Region is  $[1.7291, \infty)$ .

MATLAB code for  $p$ -value is

```
 ssiz = (s1^2 + s2^2)*(norminv(0.95)+norminv(0.9))^2/(0.005^2)
      % ssiz = 17.8128 approx 18 each
1 - tcdf(1.3050, 19)
      % ans = 0.1037
```

(b1) 99% CI for  $\mu_1 - \mu_2$  is:  $[4.3 - 7.4726 \times \sqrt{1/9 + 1/12} \times 2.8609, 4.3 + 7.4726 \times \sqrt{1/9 + 1/12} \times 2.8609] = [-5.127013, 7.7270]$ . Here  $t_{19,0.005} = 2.8609$ .

```
 ssiz = (s1^2 + s2^2)*(norminv(0.95)+norminv(0.9))^2/(0.005^2)
      % ssiz = 17.8128 approx 18 each
tinv(0.995, 19)
      % ans = 2.8609
```

(a2) Solution when no assumption about  $\sigma$ 's is made. Statistic is  $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = 4.3 / \sqrt{8.16^2/9 + 6.93^2/12} = 1.2735$ .

This statistic has approximately  $\Delta$  degrees of freedom,

$$\Delta = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \text{ nasty! } \ominus\ominus$$

$\Delta = (8.16^2/9 + 6.93^2/12)^2 / ((8.16^2/9)^2/8 + (6.93^2/12)^2/11) = 15.6627$  By rounding to the smaller integer, we get  $df = 15$ . (integer taken because of tables)  
 $t_{15,0.05} = 1.7531$ , although in principle it is possible to find  $t_{15.6627,0.05}$ .



```
ssize = (s1^2 + s2^2)*(norminv(0.95)+norminv(0.9))^2/(0.005^2)
% ssize = 17.8128 approx 18 each
tinv(0.95, 15.6627)
%ans = 1.7482
```

Rejection region is  $[1.7487, \infty)$ . Since  $t = -1.2050 > -1.7487$ , do not reject  $H_0$ .

(b2) 99% CI for  $\mu_1 - \mu_2$  is:  $[-4.3 - 2.9309 \times \sqrt{8.16^2/9 + 6.93^2/9}, -4.3 + 2.9309 \times \sqrt{8.16^2/9 + 6.93^2/9}] = [-10.23, 2.9309]$ . Here  $t_{15.5911,0.995} = 2.9309$ .

```
model{
  for(i in 1:2) {
    xbar[i] ~ dnorm(mu[i], precxbar[i])
    mu[i] ~ dnorm(0, 0.00001)
    n1[i] <- n[i]-1
    ch[i] ~ dchisqr(n1[i])
    precx[i] <- ch[i]/(n1[i] * s[i] * s[i])
    precxbar[i] <- n1[i] * precx[i]
    sigma[i] <- 1/sqrt(precx[i]) }
    teststat <- mu[1]-mu[2]
    test <- step(teststat)
  }

DATA
list( n = c(9, 12), xbar=c(39.8, 35.5), s=c(8.16, 6.93) )

INITS
list(mu=c(0,0), ch=c(1,1))


```

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
mu[1]	39.79	3.339	0.009841	33.1	39.79	46.43	1001	100000
mu[2]	35.48	2.323	0.007239	30.86	35.48	40.11	1001	100000
sigma[1]	9.036	2.646	0.009805	5.508	8.525	15.58	1001	100000
sigma[2]	7.463	1.795	0.006236	4.901	7.16	11.81	1001	100000
test	0.8653	0.3414	0.001074	0.0	1.0	1.0	1001	100000
teststat	4.309	4.061	0.012	-3.723	4.304	12.37	1001	100000

### 10.9 5-HIAA Levels.



```

Patients =[263 288 432 890 ...
           450 1270 220 350 ...
           283 274 580 285 ...
           524 135 500 120];

Controls =[60 119 153 588 ...
            124 196 14 23 ...
            43 854 400 73];

xbar1 = mean(Patients) %429
xbar2 = mean(Controls) %220.5833
s1 = std(Patients) %294.6718
s2 = std(Controls) %261.8190
n1= length(Patients) %16
n2 = length(Controls) %12
sp = sqrt( ((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2) ) %281.2413
%=====
t = (xbar1 - xbar2)/(sp * sqrt(1/n1 + 1/n2)) %1.9406

%(a) H0: mu1 = mu2 vs H1: mu1 ~ mu2 (two-sided alternative)
%RejRegion
tinv(1-0.05/2, n1 + n2 -2) %2.0555, 1.9406 is not in RR=[2.0555, inf)
                           %H0 not rejected
%p-value
pval = 2 * tcdf(-abs(t), n1 + n2 - 2) %0.0632 > 0.05, H0 not rejected
%Note: If the alternative were onesided H1: mu1 > mu2, then
%pval = 1-tcdf(t, n1 + n2 - 2) = 0.0316 < 0.05, and one would reject H0

%(b)
[xbar1 - xbar2 - sp*sqrt(1/n1 + 1/n2) * tinv(1-0.05/2, n1 + n2 -2) ...
 xbar1 - xbar2 + sp*sqrt(1/n1 + 1/n2) * tinv(1-0.05/2, n1 + n2 -2)]
% -12.3488 429.1821

```

## 10.10 Stress, Diet and Acids.



The WinBUGS solution is given below

```

model{
  for (i in 1:n){
    plasma[i] ~ dnorm(mu[smo[i]], prec[smo[i]] )
  }
  for ( j in 1:2) {
    mu[j] ~ dnorm(0, 0.0001)
    prec[j] ~ dgamma(0.0001, 0.0001)
  }
  difmu <- mu[1] - mu[2]
  testmu <- step( mu[1] - mu[2] ) #1 if mu[1]>mu[2]
  r <- prec[2]/prec[1]           #var1/var2
}

```

```

DATA
list(n=32, plasma = c(0.97, 0.72, 1.00, 0.81, 0.62, 1.32, 1.24, 0.99,
                     0.90, 0.74, 0.88, 0.94, 1.06, 0.86, 0.85, 0.58, 0.57,
                     0.64, 0.98, 1.09, 0.92, 0.78, 1.24, 1.18, 0.48, 0.71,
                     0.98, 0.68, 1.18, 1.36, 0.78, 1.64),
smo=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1))

INITS
list( mu = c(0,0), prec=c(1,1) )

```

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
difmu	-0.06288	0.1703	5.17E-4	-0.4106	-0.06287	0.2773	1001	100000
r	0.3162	0.2047	7.466E-4	0.06502	0.2699	0.8396	1001	100000
testmu	0.3393	0.4735	0.001388	0.0	0.0	1.0	1001	100000

Notice that testmu will be posterior proportion of how many times  $\mu_1 - \mu_2$  is positive. Thus, the MCMC estimate of posterior probability of hypothesis  $H_1$  that states  $\mu_1 < \mu_2$  is  $1 - 0.3393 = 0.6607$ . Note that the ratio of variances has a 95% credible set fully below 1. This is a Bayesian two-sided test for equality of variances and the conclusion is that the variances are not equal.

For comparison, a MATLAB session conducting a classical  $t$  test is provided



```

nonsmo = [0.97 0.72 1.00 0.81 0.62 1.32 1.24 0.99 ...
          0.90 0.74 0.88 0.94 1.06 0.86 0.85 0.58 0.57...
          0.64 0.98 1.09 0.92 0.78 1.14 1.18];
smo = [ 0.48 0.81 0.98 0.68 1.18 1.36 0.78 1.64];
%test hypothesis that the plasma ascorbic acid level for
%non-smokers is smaller than that of smokers. Use alpha=0.05.
X1bar = mean(nonsmo); s1 = std(nonsmo); n1 = length(nonsmo);
X2bar = mean(smo); s2 = std(smo); n2= length(smo);
%s1 = 0.2104, s2 = 0.3915; we check for equality of variances
F = s1^2/s2^2 %0.28888 is smaller than 1
pval1 = 2*fcdf(F, n1-1, n2 -1)
% pval1 =0.0223 < 5% and we will not assume equality
% of variances in comparing the two means.
% The Welch-Satterwhite df for the t test is:
ndf = (s1^2/n1 + s2^2/n2 )^2 /(( s1^2 /n1)^2/(n1-1) +
                                    (s2^2/n2)^2 /(n2-1) )
t = (X1bar - X2bar)/sqrt( s1^2/n1 + s2^2/n2 )
pval = tcdf(t,ndf)
%    ndf = 8.3892;    t =-0.4457;    pval = 0.3336
%    the mean mu1 is not significantly smaller
%    than the mean mu2 at the significance level 5%

```

### 10.11 A. pantherina and A. rubescens.



(a)  $s_p^2 = (s_1^2(m-1) + s_2^2(n-1))/(m+n-2) = (2.12^2 * 11 + 1.94^2 * 14)/(11+14) = 4.0852$  is pooled sample variance, and  $s_p = 2.0212$  is pooled sample standard deviation.

Also,  $\sqrt{1/12 + 1/15} = 0.3873$ .

The test statistic  $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{1/n+1/m}} = -1.2/2.02/0.3873 = -1.5338$ . In the two sided test, the critical value at  $\alpha = 0.05$  is  $t_{m+n-2, 1-\alpha/2} = t_{25, 0.975} = 2.060$ , and the hypothesis  $H_0$  is not rejected, since  $|t| < 2.060$ .

(b)  $z_{0.90} = 1.2816$  and  $z_{0.975} = 1.96$  and the group sample size should be  $2/0.5^2 \cdot (1.96 + 1.2816)^2 = 84.0638 \leq 85$ . To achieve desired power and detect the deviation of  $d = 0.5$ , independent samples of  $m = 85$  and  $n = 85$  spores of *A. pantherina* ("Panther") and *A. rubescens* ("Blusher") should be taken.

## 10.12 Blood Volume in Infants.



```
%Blood Volume in Infants
%X1 = early clamping measurements
X1 =[13.8 8.0 8.4 8.8 9.6 9.8 8.2 8.0 ...
       10.3 8.5 11.5 8.2 8.9 9.4 10.3 12.6];
%X2 = late clamping measurements
X2=[10.4 13.1 11.4 9.0 11.9 16.2 14.0 8.2 ...
       13.0 8.8 14.9 12.2 11.2 13.9 13.4 11.9];

X1bar = mean(X1) %9.6438
X2bar = mean(X2) %12.0938
s1 = std(X1) %1.7146
s2 = std(X2) %2.2359
n1 = 16; n2 = 16;
sp = sqrt( ((n1-1)*s1^2 + (n2-1)*s2^2 )/(n1 + n2 - 2) ) %1.9924
t = (X1bar - X2bar)/(sp * sqrt(1/n1 + 1/n2)) %-3.4781

p = 2 * tcdf(-abs(t), n1 + n2 - 2) %0.0016
```

The mean volumes of blood in infants are significantly different for early (population 1) and late (population 2) clamping of the umbilical cord.

## 10.13 Biofeedback.



(a)  $H_0 : \mu_1 = \mu_2$  versus  $H_1 : \mu_1 > \mu_2$  or in terms of differences,  $H_0 : \mu_1 - \mu_2 = 0$  versus  $H_1 : \mu_1 - \mu_2 > 0$ .

(b) To follow the alternative  $H_1$  the differences  $d_i$  should be taken as  $X_{1i} - X_{2i}$ . Here,  $d_i = \{7, 21, 17, -3, 11\}$ ,  $\bar{d} = 10.6$ ,  $s_d = 9.32$ ,  $t = 10.6/(9.32/\sqrt{5}) = 2.54$ ,  $t_{4,0.95} = 2.131847$ .

(c) Variances are the same, normal distributions.

## 10.14 Hypertension.



(a)



```
% first group
n1 = 15; X1bar = 16.16; s1 = 4.29;
% second group
n2 = 16; X2bar = 10.53; s2 = 3.33;
% population variances assumed the same, need s_p
sp2 = ((n1 - 1)* s1^2 + (n2-1) * s2^2)/(n1 + n2 - 2);
sp = sqrt(sp2) % sp = 3.8237

tstat = (X2bar - X1bar)/(sp * sqrt(1/n1 + 1/n2))
% tstat = -4.0969

%TEST two sided H_1 rejection region method
alpha = 0.05;
tcrit = tinv(1-alpha/2, n1 + n2 - 2)%tcrit = 2.0452
%and the rejection region RR is
%RR =(-inf, -tcrit)U(tcrit, inf)=
% (-inf, -2.0452)U(2.0452, inf).
%Reject H_0 since tstat falls in the RR.

%TEST using p-values
pval = 2 * tcdf(-abs(tstat), n1 + n2 - 2)
% pval = 3.0733e-004
% which is the same as 2*tcdf(tstat, n1+n2-2)
% since tstat < 0.
% Reject H_0 since pval = 0.0003 < 0.05 = alpha.
```

(b) The power is

$$\Phi\left(-z_{1-\alpha/2} + \frac{\Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right),$$

with  $\Delta = 3$ ,  $\alpha = 0.05$ , and  $\sigma_1^2, \sigma_2^2$  replaced by  $s_1^2 = 4.29^2$  and  $s_2^2 = 3.33^2$ .



```
Delta = 3;
pow = normcdf( -norminv(1 - alpha/2) + ...
    Delta/( sqrt(s1^2/n1 + s2^2/n2) ) )
% pow = 0.5813
```

(c) The sample size is

$$n = \frac{(\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2},$$

with  $\Delta = 3$ ,  $\alpha = 0.05$ ,  $\beta = 0.01$  and  $\sigma_1^2, \sigma_2^2$  replaced by  $s_1^2 = 4.29^2$  and  $s_2^2 = 3.33^2$ .



```
beta = 0.01;
n = (s1^2 + s2^2)*(norminv(1-alpha/2) + ...
    norminv(1-beta))^2 /Delta^2
% n = 60.2066, thus, take n1=n2=61.
```

**10.15 Hemopexin in DMD Cases II.**
 TBA
**10.16 Risk of Stroke.**
 TBA
**10.17 Cell Counts.**



```
%Exercise Cell Counts
A20 =[34 44 40 62 53 51 30 33 38 51 26 48];
M20 =[30 43 34 53 49 39 37 42 30 50 35 54];
A70 =[72 82 100 94 83 94 73 87 107 102];
M70 =[76 51 92 77 74 81 72 87 100 104];
n1 = length(A20); n2 = length(A70);
md20 = mean(A20-M20)
sd20=std(A20-M20)
md70 = mean(A70-M70)
sd70=std(A70-M70)

%(a)
t1 = md20/(sd20/sqrt(n1)) %0.5520
t2 = md70/(sd70/sqrt(n2)) %2.4072
pval1=2 * tcdf(-abs(t1), n1-1) %0.5920
pval2=2 * tcdf(-abs(t2), n2-1) %0.0394

%(b)
[md20 - tinv(0.975, n1-1)*sd20/sqrt(n1) ...
md20 + tinv(0.975, n1-1)*sd20/sqrt(n1)]
%-3.4853    5.8186

[md70 - tinv(0.975, n2-1)*sd70/sqrt(n2) ...
md70 + tinv(0.975, n2-1)*sd70/sqrt(n2)]
%%0.4821    15.5179

%(c)
var2p = ((n1-1)*sd20^2 + (n2-1)*sd70^2)/(n1 + n2 -2)
t0 = (md20 - md70)/sqrt(var2p * (1/n1 + 1/n2))
pval= 2 * tcdf(-abs(t0), n1 + n2 - 2 )
% -1.7935    pval = 0.0880 not significant
```

Thanks to Professor Carlos E. Fernández-Ossa from School of Engineering of Antioquia, Columbia, for pointing out typos in the original MATLAB code for (b) and (c), leading to wrong results.

**10.18 Impulses from Crayfish.**
 TBA

### 10.19 Aerobic Capacity.



```
x1bar = 46.3; x2bar = 38.0; s1 = 5; s2 = 5.2; n1 = 20; n2 = 10; sp = sqrt( ((n1 - 1)*s1^2 + (n2-1)*s2^2 )/(n1 + n2 - 2) ) t = (x1bar - x2bar)/(sp * sqrt(1/n1 + 1/n2)) tinv(0.95, n1 + n2 - 2) pval = 1 - tcdf(t, n1 + n2 - 2) sigma = 5; alpha = 0.05; beta = 0.1; delta = 4; n = 2 * sigma^2/delta^2 * (norminv(1-alpha) + norminv(1-beta))2
```

### 10.20 Cataract and Diabetes.



 [rd rdl rdu rr rrl rru or orl oru] = risk(56, 84, 552, 1927)

```
% rd = 0.1773 [ 0.0945, 0.2601]
% rr = 1.7964 [1.4477, 2.2290]
% or = 2.3273 [1.6382, 3.3063]
```

### 10.21 Beginnings of Antiseptic Surgeries.

 TBA

### 10.22 Reaction Times.



 %%

```
rg = [...
18 22;...
16 20;...
23 29;...
30 35;...
32 27;...
30 29;...
31 33;...
25 29;...
27 31;...
21 24];
```

```
d = rg(:,1) - rg(:,2);
sd = std(d);
n = length(d);
t = mean(d)/( sd/sqrt(n))
p = 2 * tcdf( - abs(t), n-1)
```

```
tcrit = tinv(0.975, n-1)
%(-inf, -tcrit) U (tcrit, inf)
% t = -2.5122
% p = 0.0332
```

```
% tcrit =2.2622
```

### 10.23 Gamma Globulin and Aspirin.



### 10.24 High/Low Protein Diet in Rats.



```
 % High/Low Protein in Rats
X1=[134 146 104 119 124 161 107 83 113 129 97 123];
X2=[70 118 101 85 107 132 94];
X1bar = mean(X1); %X1bar = 120
s1 = std(X1); %s1=21.3882
n1 = length(X1); %n1 = 12
X2bar = mean(X2); %X2bar = 101
s2 = std(X2); %s2 = 20.6236
n2= length(X2); %n2=7
%=====
F = s2^2/s1^2 %F = 0.9298
pval1 = 2*fcdf(F, n2-1, n1 -1) %pval1 =0.9788
% decide sigma_1^2 = sigma_2^2 test by
% pooled standard deviation
sp = sqrt( ( (n1-1)*s1^2 + (n2-1)*s2^2)/(n1 + n2 - 2) )
%sp =21.1215
t = (X1bar - X2bar)/(sp * sqrt(1/n1 + 1/n2)) %t=1.8914
%
pval = 1-tcdf(t, n1 + n2 -2) %pval=0.0379
%
tcrit = tinv(0.95, n1 + n2 - 2) %tcrit=1.7396
n = (450 + 450)*(norminv(0.95) + norminv(0.95))^2 / 20^2
%n =24.3499
ssize = ceil(n) %ssize=25
```

### 10.25 Spider Monkey DNA.



### 10.26 PBSC versus BM for Unrelated Donor Allogeneic Transplants.



### 10.27 Hydrogels.



```
 % Hydrogels
data =[...
20250 44250;...
51000 126000;...
77250 100500;...
39000 58500;...
```

```

40500 69750;...
42750 76500;...
78750 155250;...
42750 67500];

minutes30 = data(1:4,1)./data(1:4,2)
minutes60 = data(5:8,1)./data(5:8,2)

%minutes30 = 0.4576 0.4048 0.7687 0.6667
%minutes60 = 0.5806 0.5588 0.5072 0.6333

n1 = 4; n2 = 4;
xbar1 = mean(minutes30) %0.5744
xbar2 = mean(minutes60) %0.5700
alpha = 0.05;
s1 = std(minutes30) %0.1719
s2 = std(minutes60) %0.0522

sp = sqrt( ((n1-1)*s1^2 + (n2-1)*s2^2)/(n1 + n2 - 2)) %0.1270
[xbar1-xbar2 - tinv(1-alpha, n1+n2-2)*sp*sqrt(1/n1+1/n2),...
xbar1-xbar2 + tinv(1-alpha, n1+n2-2)*sp*sqrt(1/n1+1/n2)]
% -0.1702    0.1790

```

Since Westlake's interval  $[-0.1702, 0.1790]$  is not contained in the interval  $[-0.1, 0.1]$  the hypothesis of equivalence cannot be established. It is interesting that  $H_0 : \mu_1 = \mu_2$  is not rejected ( $p$ -value against one sided alternative is 0.48), yet the equivalence cannot be established with specified equivalence margins and significance level  $\alpha$ . The sample sizes  $n_1 = n_2 = 4$  are quite small to establish equivalence with the equivalence margins  $\theta_U = -\theta_L = 0.1$ . If the equivalence margins were  $\theta_U = -\theta_L = 0.2$ , the equivalence would be established.

## 10.1 Additional Problems

**10.a1 Tactile Sensation in Rats.** Researchers in Garrett Stanley's Lab are interested in understanding how the brain processes the sense of touch, and use the rat whisker system as a model for tactile sensation. In this particular experiment, the researchers were testing the ability of subject rats to detect very weak deflections of their whiskers resulting from a short (150ms) puff of air. Much as a person might remain very still when trying to listen for a faint sound, it was hypothesized that the animals would be more likely to succeed in the task when they held their whiskers still in anticipation of the arrival of the stimulus. To test this, the researchers recorded high speed video of the whiskers for a short interval prior to the stimulus. After recording a total of 57 trials, the researchers examined the video and separated the trials into two categories: those in which the whiskers were still prior to the arrival of the stimulus ( $n_1 = 43$ ), and those on which the whiskers were moving

( $n_2 = 14$ ). The animals correctly detected the stimulus 23 times under the first condition ( $\hat{p}_1 = 53.49\%$  correct), and only 3 times under the second condition ( $\hat{p}_2 = 21.43\%$ ).

(a) Test hypothesis for equality of proportions using normal approximation to the binomial. Argue that the sample size is small for the central limit theorem to hold (*Hint:* For applicability of normal approximation usual requirement is  $\min\{n_1, n_2\}p(1-p) > 5$  for  $p = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2)$ .) (b) Using WinBUGS, test the hypothesis in (a) using beta  $\text{Be}(1/2, 1/2)$  prior on the unknown proportions  $p_1$  and  $p_2$ . Note that the choice (1/2, 1/2) for the hyperparameters of Beta is Jeffreys noninformative prior.



```

BUGS
model{
  X1 ~ dbin( p1, n1 )
  X2 ~ dbin( p2, n2 )
  p1 ~ dbeta(0.5, 0.5)
  p2 ~ dbeta(0.5, 0.5)
  diff <- p1 - p2
  pH1 <- step(diff)
}

DATA

list( X1 = 23, X2 = 3, n1 = 43, n2 = 14)

INITS

list(p1=0.5, p2 = 0.5)

```

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample	
diff	0.301	0.1293	4.182E-4	0.02776	0.3087	0.5317	1001	100000	
pH1	0.9836	0.127	4.333E-4		1.0		1.0	1001	100000

**10.a2 Clinical Trial of Abatacept.** Abatacept is a drug proposed to treat and prevent active lupus flares in at least one of three organ systems: the skin, the heart, the lung, or four joints. If, in a double blind trial 33 out of 115 people treated with Abatacept showed cumulative damage due to Systemic Lupus Erythematosus (SLE Score  $\geq 1$ ), and 17 out of 55 people in the placebo arm also showed cumulative damage due to SLE, is Abatacept more effective than the placebo? Use a significance level of 0.05.

- (a) Answer the above question using risk differences. What is the 99% CI for the risk difference?
- (b) Answer the above question using the odds ratio. What is the 95% CI for the odds ratio?



# **Chapter 11**

## **ANOVA and Elements of Statistical Design**

### **11. 1 Nematodes.**

 TBA

### **11.2 Cell Folate Levels in Cardiac Bypass Surgery.**

 TBA

### **11.3 Computer Games.**

 TBA

### **11.4 MTHFR C677T Genotype and Levels of Homocysteine and Folate.**

 TBA

### **11.5 Beetles.**

 TBA

### **11.6 ANOVA Table from Summary Statistics.**

 TBA

### **11.7 Protein Content in Milk for Three Diets.**

 (a)

'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
'Groups'	[0.7470]	[ 2]	[0.3735]	[5.6118]	[0.0053]
'Error'	[5.0585]	[76]	[0.0666]		
'Total'	[5.8056]	[78]			

(b) The hypothesis here is that the diets don not differ in protein yield, that is

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad vs. H_1: \text{not } H_0.$$

$H_0$  is rejected since the  $p$ -value is less than 5%.

(c)  $\mu_1 - \mu_2 \in (-0.0579, 0.2844) \Rightarrow \mu_1 = \mu_2$ .  
 $\mu_1 - \mu_3 \in (0.0684, 0.4107) \Rightarrow \mu_1 > \mu_3$ .  
 $\mu_2 - \mu_3 \in (-0.0416, 0.2941) \Rightarrow \mu_2 = \mu_3$ .

### 11.8 Tasmanian Clouds.



### 11.9 Clover Varieties.



### 11.10 Cochlear Implants.



### 11.11 Bees.



### 11.12 SiRstv: NIST's Silicon Resistivity Data.



### 11.13 Dorsal Spines of *Gasterosteus aculeatus*.



```
%BENKA GARDENBAY BIG
stickleback =[...
4.2 4.4 4.9; 4.1 4.6 4.6; 4.2 4.5 4.3; 4.3 4.2 4.9; ...
4.5 4.4 4.7; 4.4 4.2 4.4; 4.5 4.5 4.5; 4.3 4.7 4.4 ];
lakes = [1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3];
[p table stats] = anova1(stickleback(:, lakes(:))

% p = 0.0267
%
% table =
%
% 'Source' 'SS' 'df' 'MS' 'F' 'Prob>F'
% 'Groups' [0.3033] [ 2] [0.1517] [4.3260] [0.0267]
% 'Error' [0.7363] [21] [0.0351] [] []
% 'Total' [1.0396] [23] [] [] []
%
% stats =
%
% gnames: 3x1 cell
% n: [8 8 8]
% source: 'anova1'
% means: [4.3125 4.4375 4.5875]
% df: 21
```

```
%           s: 0.1872
```

(a) Hypothesis  $H_0$  stating that the mean lengths of dorsal spines are the same for the three lakes is rejected at the level  $\alpha = 0.05$  since the  $p$ -value is 0.0267.

(b) If the significance level was  $\alpha = 0.01$  the hypothesis  $H_0$  would not be rejected.

#### 11.14 Incomplete ANOVA Table.

 TBA

#### 11.15 Maternal Behavior in Rats.

 TBA

#### 11.16 Comparing Dialysis Treatments.



```
 %Comparing Dialysis Treatments.
wchange = [...
2.90 2.97 2.67; 2.56 2.45 2.62;...
2.88 2.76 1.84; 1.73 1.20 1.33;...
2.50 2.16 1.27; 3.18 2.89 2.39;...
2.83 2.87 2.39; 1.92 2.01 1.66];
subject = [1 2 3 4 5 6 7 8 ...
1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8]';
treatment = [1 1 1 1 1 1 1 1 ...
2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3]';

%(a) incorrect design // independent treatments
[p1 table1 stats1] = anova1(wchange(:), treatment)
% H0 not rejected!
% p1 = 0.1616
%
% table1 =
% 'Source'    'SS'      'df'      'MS'      'F'       'Prob>F'
% 'Groups'    [1.2510]  [ 2]  [0.6255]  [1.9901]  [0.1616]
% 'Error'     [6.6004]  [21]  [0.3143]   []        []
% 'Total'     [7.8515]  [23]   []        []        []
%
% stats1 =
%     gnames: 3x1 cell
%         n: [8 8 8]
%     source: 'anova1'
%     means: [2.5625 2.4137 2.0213]
%     df: 21
```

```

%           s: 0.5606

% (b) correct solution block design
names = 'subject','treatment';
[p2 table2 stats2] = anovan(wchange(:),subject, treatment, ...
    'varnames',names)

% H0 rejected, the treatments differ.
% p2 =
% 0.0001
% 0.0028
%
% table2 =
% 'Source'    'Sum Sq.'   'd.f.'   'Mean Sq.'   'F'       'Prob>F'
% 'subject'   [ 5.6531]  [ 7]   [0.8076]   [11.9341]  [6.0748e-005]
% 'treatment' [ 1.2510]  [ 2]   [0.6255]   [ 9.2436]  [     0.0028]
% 'Error'      [ 0.9474]  [14]   [0.0677]    []        []
% 'Total'      [ 7.8515]  [23]    []        []        []

```

### 11.17 Material Scientist and Assessing Tensile Strength.



```

tensile = [ 73   68   74   71   67 ...
             73   67   75   72   70 ...
             75   68   78   73   68 ...
             73   71   75   75   69 ];
chemical = [1 1 1 1 1   2 2 2 2 2   3 3 3 3 3   4 4 4 4 4];
bolt     = [1 2 3 4 5   1 2 3 4 5   1 2 3 4 5   1 2 3 4 5];
[P,T,STATS,TERMS] = anovan( tensile, {chemical, bolt}, 'model','linear', ...
    'varnames', strvcat('chemical', 'bolt'))

```

### 11.18 Oscilloscope.



### 11.19 Magnesium Ammonium Phosphate and Chrysanthemums.



```

% Response: Height of Chrysanthemum
hchr = [...
13.2 16.0 7.8 21.0;12.4 12.6 14.4 14.8;...
12.8 14.8 20.0 19.1;17.2 13.0 15.8 15.8;...
13.0 14.0 17.0 18.0;14.0 23.6 27.0 26.0;...

```

```

14.2 14.0 19.6 21.1;21.6 17.0 18.0 22.0;...
15.0 22.2 20.2 25.0;20.0 24.4 23.2 18.2];

%50 gm/bu 100 gm/bu 200 gm/bu 400 gm/bu
treatment = [...;
1 2 3 4; 1 2 3 4; 1 2 3 4; 1 2 3 4; 1 2 3 4;...
1 2 3 4; 1 2 3 4; 1 2 3 4; 1 2 3 4];

[pval anovatab stats] = anova1(hchr(:, treatment(:))

% pval = 0.0989
% anovatab =
%    'Source'      'SS'        'df'       'MS'        'F'        'Prob>F'
%    'Groups'      [119.7870]   [ 3]      [39.9290]   [2.2522]   [0.0989]
%    'Error'       [638.2480]   [36]      [17.7291]   []         []
%    'Total'       [758.0350]   [39]      []         []         []
% stats =
%    gnames: {4x1 cell}
%          n: [10 10 10 10]
%    source: 'anova1'
%    means: [15.3400 17.1600 18.3000 20.1000]
%    df: 36
%    s: 4.2106

multcompare(stats, 'alpha',0.1,'display','off')

%    1.0000    2.0000   -6.2949   -1.8200    2.6549
%    1.0000    3.0000   -7.4349   -2.9600    1.5149
%    1.0000    4.0000   -9.2349   -4.7600   -0.2851
%    2.0000    3.0000   -5.6149   -1.1400    3.3349
%    2.0000    4.0000   -7.4149   -2.9400    1.5349
%    3.0000    4.0000   -6.2749   -1.8000    2.6749

m = stats.means
c = [1 -1 -1 1]; %mu1 + mu4 = mu2 + mu3
L = c(1)*m(1) + c(2)*m(2)+c(3)*m(3) + c(4)*m(4) %L=-0.02
LL= m * c' %LL= -0.02
stdL = stats.s * sqrt(c(1)^2/4+c(2)^2/6+c(3)^2/6+c(4)^2/8)
%stdL = 3.5437
t = LL/stdL %t = -0.0056

%test H_0: mu * c' = 0 H_1: mu * c' ~= 0
% p-value
2 * tcdf(-abs(t), 36) %pval= 0.9955; 36=40-4

```

```
%or 90% confidence interval for population contrast
[LL - tinv(0.95, 36)*stdL, LL + tinv(0.95, 36)*stdL]
%-6.0029    5.9629
```

**11.20 Color Attraction for *Oulema melanopus*.**

 TBA

**11.21 Raynaud's Phenomenon.**

 TBA

**11.22 Simvastatin.**

 TBA

**11.23 Antitobacco Media Campaigns.**

 TBA

**11.24 Orthosis.**

 TBA

**11.25 Bone Screws.**

 TBA

**11.26 R&R Study.**

 TBA

**11.27 Additive R&R ANOVA for Measuring Impedance.**

 TBA

## 11.1 Additional Problems

**11.a1 Nulatron Tumb Screws.** A manufacturer of flow chambers uses Nylon 6 (Nulatron) for production of tumb screws. The manufacturer orders Nulatron from two suppliers. The material is tested for shear strength (in PSI at  $73^{\circ}F$ ). Four batches from each supplier are selected at random and three samples from each batch used for testing. The shear strength varies from batch to batch.

(a) You are interested in testing for the difference between the two suppliers, but want to account for the differences between batches. Do your testing at  $\alpha = 0.05$  significance level.

(b) Compare the suppliers by ignoring batches, using the two-sample  $t$ -test.

Supplier Batch	1				2			
	1	2	3	4	1	2	3	4
9620	9590	9715	9690		9700	9710	9670	9695
9670	9610	9675	9665		9680	9745	9720	9680
9675	9685	9645	9710		9675	9665	9680	9730



(a) Here the suppliers are fixed effects but batches are random. The hypothesis  $H_0 : \alpha_i = 0$  is tested by  $F = MSA/MSB(A)$  which under  $H_0$  has  $F$  distribution with  $2 - 1$  and  $2(4 - 1)$  degrees of freedom.  $MSA = 6666.7, F_A = 5.4545, p_a = 0.0582$ . Thus  $H_0$  not rejected.

The hypothesis of homogeneity of batches  $H_0 : \beta_j(i) = 0$  is tested by  $F = MSB(A)/MSE$  which under  $H_0$  has  $F$  distribution with  $2(4 - 1)$  and  $2 \cdot 3(4 - 1)$  degrees of freedom.  $MSA(B) = 1222.2, F_{B(A)} = 1.1757, p_{b(a)} = 0.3670, MSE = 1039.6$ . Thus,  $H_0$  not rejected.

(b) Two-sample  $t$  statistic (pooled standard deviations) is  $t = -2.4738$ , which leads to a two-sided  $p$  value of 0.0216, suggesting that the suppliers are significantly different.

**11.a2 Chair Yoga.** The chair yoga pose (Utkatasana, Fig. 11.1) is known for improving posture and balance because of the way how it distributes the body weight over the foot (it also helps in strengthening the muscles). A group of students conducted a study with  $n = 19$  subjects to determine whether the depth (angle between the calves and thigh) of the yoga chair pose affects the uniformity of force distributions. Subjects were requested to stand on the balance board with a specific posture, they were then asked to assume the yoga chair pose until they reached 60, then 90, then 120, then 150 degrees for 10 seconds each. The balance board recorded 4 forces, right bottom, left bottom, right top, and left top, from which only their coefficient of variation, cvforce, is of interest in this problem. The data structure `chairyoga.mat` contains fields `subject`, `angle`, and `cvforce`, each as a vector  $654 \times 1$ . The original high-frequency data were subsampled to decrease time-dependence of force measurements.

The smaller the coefficient of variation of the four forces `cvforce` is, the better/safer the pose.

(a) Test whether the population means for `cvforce` are the same for the four levels of `angle`, that is, test the hypothesis  $H_0 : \mu_{60} = \mu_{90} = \mu_{120} = \mu_{150}$ . Use a block design where `angle` is the factor of interest and `subject` is a blocking variable. Copy the ANOVA-table from the output.

*Hint. Use an additive anovan with angles and subjects as the factors.*

(b) If the hypothesis of equality of means from (a) is rejected, which means differ. Which mean is the smallest (“safest” in the sense of minimum relative variability).



CHRIS ANDRE

**Fig. 11.1** The 90° chair yoga pose (Utkatasana).

(c) The procedure `multcompare` will give you simultaneous confidence intervals for all differences between the means. What is the 95% CI for  $\mu_{90} - \mu_{120}$ . Are the two means significantly different and why yes or not?

(d) Run `multcompare` for the subjects. Which subject (out of 19) was the most disbalanced (had maximum CV)? Which subject was the most stable?



```
%Chair Yoga
load 'chairyoga.mat'
figure(1)
cvforce = chairyoga.cvforce;
angle = chairyoga.angle;
subject = chairyoga.subject;
varnames ='angle','subject';
[p,table,stats] = anovan(cvforce,angle,subject, ...
    'model','linear','varnames',varnames);

% table =
%   'Source'      'Sum Sq.'     'd.f.'      'Singular?'     'Mean Sq.'     'F'       'Prob>F'
%   'angle'        [ 0.1489]    [  3]        [        0]    [  0.0496]    [3.5386]    [0.0145]
%   'subject'      [19.2803]    [ 18]        [        0]    [  1.0711]    [76.3601]   [0]
%   'Error'        [ 8.8653]    [ 632]       [        0]    [  0.0140]
%   'Total'        [28.3819]    [ 653]       [        0]           []

figure(2)
multcompare(stats,'dimension',1)

%   1.0000    2.0000    -0.0456    -0.0119    0.0218
%   1.0000    3.0000    -0.0053    0.0288    0.0630
%   1.0000    4.0000    -0.0151    0.0187    0.0525
%   2.0000    3.0000    [0.0050]   0.0408    0.0765]  <- Interval
%   2.0000    4.0000    -0.0047    0.0307    0.0660
```

```
%      3.0000    4.0000   -0.0459   -0.0101    0.0257
```

```
figure(3)
multcompare(stats,'dimension',2) ;
```



## Chapter 12

# Distribution Free Tests

**Friday the 13th.**



```
 fri6 = [9 6 11 11 3 5];  
fri13 =[13 12 14 10 4 12];  
[pvae, pvaa, n, plusses, ties] =signtst(fri6, fri13)  
  
%pvae = 0.1094  
%pvaa = 0.1103  
%n = 6  
%plusses = 1  
%ties = 0
```

The output [pvae, pvaa, n, plusses, ties] consists of the exact one-sided  $p$ -value (pvae), normal approximation to one sided  $p$ -value (pvaa), sample size  $n$  adjusted for the ties (depending on policy of tie-treatment), number of plusses (or minuses, whatever is more extreme for  $H_0$ ), and number of ties.

The exact  $p$ -value is  $\binom{6}{0} + \binom{6}{0} \frac{1}{2^6} = 7/64 = 0.1094$ .  $H_0$  is not rejected at level  $\alpha = 0.10$ .

The built-in MATLAB function `signtest` provides an alternative way to solve this problem, but is somewhat lean in reporting and options, compared to `signtst`.

**Reaction Times.**



**Simulation.**



**12.4 Grippers.**

```


(a)
%Left hand (X)
X = [ 140 90 125 130 95 121 85 97 131 110]
%Right hand (Y)
Y = [ 138 87 110 132 96 120 86 90 129 100]
[W, Zstat, pval] = wsirt(X, Y, 1)

% W = 37 [Sum of ranks pos - sum of ranks negative, should be 0 under H_0]
% Zstat = 1.8956
% pval = 0.0326
% (b)
d = X - Y %mean(d) = 3.6
n= length(X);
sd = std(d) %5.4610
tstat = mean(d)/(sd/sqrt(n)) %2.0846
pval = 1-tcdf(tstat, n-1) %0.0334 No opinion change.

```

### Iodide and Serum Concentration of Thyroxine.

 TBA

### Weightlifters.

 TBA

### Cartilage Thickness in Two Osteoarthritis Models.

 TBA

### A Claim.

 TBA

### Claustrophobia.



```


% Claustrophobia
A =[...
4.6 4.7 4.9 5.1 7.0 4.9 5.1 5.2 5.5 4.8 ...
5.7 5.0 5.8 6.1 6.5 7.0 6.4 5.2 4.6 4.7 ...
4.9 6.4 5.9 4.7 5.8 5.2 5.4 6.1 7.7 6.2 ...
5.8 5.1 6.5 2.2 6.9 5.0 6.5 7.2 8.2 6.7];

B =[...
5.2 5.3 5.4 7.7 8.1 4.9 5.6 6.2 6.3 7.0 ...
7.0 7.8 6.8 7.7 8.0 6.6 5.5 8.2 8.1 5.0];

[sumranks1, tstat, pval] = wmw(A, B, -1)
% also [sumranks1, tstat, pval] = wsurt(A, B, -1)

% sumranks1 = 1041

```

```
% tstat = -2.8094
% pval = 0.0025
%
% H0 rejected, scores for A lower than for B
```

**Nonparametric Stats with Raynaud's Phenomenon.** TBA**Cotinine and Nicotine Oxide.** TBA**Coagulation Times.** TBA**Blocking by Rats.** TBA



# Chapter 13

## Goodness of Fit Tests

### 13.1 Q-Q Plot for $\sqrt{2\chi^2}$ .



13.2 Not at all like me.



Results:  $\chi^2 = 2.55$ ,  $\chi^2_{4,1-0.05} = 9.4877$ , Do not reject  $H_0$ .

```
%Not at All Like Me
ni=[8 9 21 8 4];
n= sum(ni)
%n = 50
theopi = [10 20 40 20 10]/100
%theopi = 0.1000      0.2000      0.4000      0.2000      0.1000

npi=50*theopi
%npi = 5      10      20      10      5

ch2 = sum((ni - npi).^2 ./npi)
%ch2 = 2.5500

pval = 1 - chi2cdf(2.55, 5-1)
%pval = 0.6357

crit= chi2inv(1-0.05, 5-1)
%crit = 9.4877
```

### 13.3 Cell Counts.

Sequence  $\{(n_i - np_i)^2 / (np_i)\} : [2.2368 \ 1.9059 \ 0.1779 \ 0.1779 \ 0.5309]$ .  $\chi^2 = 5.0294$   $\chi^2_{4,1-0.05} = 9.4877$ .

### 13.4 GSS Data.

$\chi^2 = 25.0860$ ,  $\chi^2_{3,1-0.05} = 7.8147$ ,  $np_i : [309.4318 \ 557.2213 \ 28.685 \ 3.6619]$ .

### 13.5 Strokes on “Black Monday”.

$np_1 = 18.7778 \ 0.44444 \ 1 \ 1.7778 \ 1.7778 \ 2.7778 \ 1. \ \chi^2 = 27.5556 \ \chi^2_{6,1-0.05} = 12.5916$ .

### 13.6 Benford’s Law.

TBA

### 13.7 Simulational Exercise.

TBA

### 13.8 Deathbed Scenes.

TBA

### 13.9 Grouping in a Vervet Monkey Troop.

TBA

### 13.10 Crossing Mushrooms.

Total number of observations is  $n = 224$ . Theoretical frequencies are  $np_1 = 224 \cdot \frac{9}{16} = 126$ ,  $np_2 = np_3 = 224 \cdot \frac{3}{16} = 42$ , and  $np_4 = 224 \cdot \frac{1}{16} = 14$ .

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - np_i)^2}{np_i} = \frac{(-6)^2}{126} + \frac{11^2}{42} + \frac{(-6)^2}{42} + \frac{1^2}{14} = 4.095.$$

Since  $\chi^2_{4-1,0.95} = 7.81$ , the results do not disagree with the theory. In other words,  $H_0$  is not rejected at 5% significance level.

### 3.11 Renner Honey Data Revisited.

TBA

### 13.12 PCB in Yolks of Pelican Eggs.

```
 anacapa =[452    184    115    315    139    177    214    356    166    246    ...
177    289    175    296    205    324    260    188    208    109    204    ...
89     320    256    138    198    191    193    316    122    305    203    ...
396    250    230    214    46     256    204    150    218    261    143    ...
229    173    132    175    236    220    212    119    144    147    171    ...
216    232    216    164    185    216    199    236    237    206    87];

hist((anacapa), 20)
```

```

prctile(anacapa, [10 20 30 40 50 60 70 80 90])

%122.0000 148.5000 175.0000 192.0000 205.0000 216.0000
% 232.0000 256.0000 315.0000

prcs = prctile(anacapa, [12.5 25 37.5 50 62.5 75 87.5 ])

%135.7500 169.7500 187.6250 205.0000 216.2500 239.2500 299.3750

%[-inf] 46    87    89    109   115   119   122   132   138   [138.5]
% 139    143   144   147   150   164   166   171   173   [174]
% 175    175   177   177   184   185   188   191   193   [195.5]
% 198    199   203   204   204   205   206   208   212   [213]
% 214    214   216   216   216   218   220   229   230   [231]
% 232    236   236   237   246   250   256   256   260   [260.5]
% 261    289   296   305   315   316   320   324   356
% 396    452   [inf]

ni = [9 9 9 9 9 11]
ei = 65 * diff(normcdf([-1000 138.5 174 195.5 213 231 260.5 2000], ...
    mean(anacapa), std(anacapa) ))

chi2 = sum( (ni - ei).^2 ./ ei )
1-chi2cdf(chi2, 7-1-2)
% ei = 10.6016 9.5833 7.1860 6.1970 6.3072 9.2636 15.8613
% chi2 = 4.6503
% pval = 0.3251
% p = 0.3251

[h,p,stats] = chi2gof(anacapa,'cdf',...
    @(z)normcdf(z,mean(anacapa),std(anacapa)),...
    'edges',[0 138.5 174 195.5 213 231 260.5 452 1000],'nparams',2)

stats.E

% stats =
%     chi2stat: 4.6503
%         df: 4
%     edges: [0 138.5000 174 195.5000 213 231 260.5000 1000]
%         0: [9 9 9 9 9 11]
%         E: [1x7 double]
%
% ans = 10.6016 9.5833 7.1860 6.1970 6.3072 9.2636 15.8613
[h,p,stats] = chi2gof(anacapa)

```

### 13.13 Number of Leaves per Whorl in *Ceratophyllum demersum*.

 TBA

### 13.14 From 1998-2002 U.S. National Health Interview Survey (NHIS).



(a) Probabilities for Binomial  $\text{Bin}(2, 0.515)$  distribution are:  $p_0 = \binom{2}{0} 0.515^0 (1 - 0.515)^2 = 0.2352$ ,  $p_1 = \binom{2}{1} 0.515^1 (1 - 0.515)^1 = 0.4995$ ,  $p_2 = \binom{2}{2} 0.515^2 (1 - 0.515)^0 = 0.2652$ . Theoretically expected sibship counts, if the distribution for number of boys is  $\text{Bin}(2, 0.515)$ , are  $7541 \cdot 0.2352 = 1773.6$ ,  $7541 \cdot 0.4995 = 3766.7$ , and  $7541 \cdot 0.2652 = 1999.9$ . Thus,

Number of Boys	0	1	2
Observed number of sibships	1,941	3,393	2,207
Theoretical number of sibships	1,773.6	3,766.7	1,999.9
Difference	167.4	-373.7	207.1

We see that there is a difference between Observed and Theoretical numbers of sibships, especially for the case of sibships with one boy where the difference is -373.7. [Later in the course we will learn to test if this difference is significant]

(b) First note that  $n = 50936$ , i.e.,  $n$  is the total number of children. The number of boys is  $13079 + 2 \times 6545 = 26169$ . Thus,  $\hat{p} = 26169/50936$ .

The following MATLAB code calculates  $Z$  statistic and associated  $p$  value for the one-sided alternative.

```

z = (26169/50936 - 1/2)/sqrt(0.5*0.5/50936)
%z = 6.2121
1-normcdf(6.2121)
%ans = 2.6141e-010

```

Since  $Z = 6.2121$  falls in the rejection region  $RR = [1.645, \infty)$ , the hypothesis  $H_0 : p = 1/2$  is rejected. The proportion of boys in this population is significantly higher than  $1/2$ .

Note that  $p$ -value is smaller than  $\alpha = 0.05$  leading to the same decision to reject  $H_0$ .

### 13.15 Neuron Fires Revisited.



```

%neuronfires.mat
load neuronfires
[f] = hist(Y, 2.5:5:997.5)
[ni x]=hist(f,unique(f))
%ni = 6   18   21   45   39   25   25   11   6   3   1
%x = 1   2   3   4   5   6   7   8   9   10  12
ni=[ni(1:9) ni(10)+ni(11)] %join the last two cells
%ni = 6   18   21   45   39   25   25   11   6   4
npi= 200 * [poisscdf(1,mean(f))...
             poisspdf(2:9,mean(f)) 1-poisscdf(9,mean(f))]
% note that the first probability includes 0 and 1.
%npi = 8.4644 17.4079 28.6940 35.4730 35.0828
%      28.9141 20.4257 12.6256 6.9371 5.9754

```

```
ch2 = sum((ni-npi).^2./npi)
%ch2 = 8.3400

pval = 1-chi2cdf(ch2, 10-1-1) %estimated mean, -1 df
%pval = 0.4010
```

The counts in the consecutive intervals are consistent with the Poisson distribution ( $\hat{\lambda} = 4.945$ ,  $p$ -value 0.4010).

### 13.16 Cloudiness in Greenwich.

 TBA

### 13.17 Distance between Spiral Reversals in Cotton Fibers.

 TBA



## Chapter 14

### Models for Tables

#### 14.1 Amoebas and intestinal disease.

 TBA

#### 14.2 Drinking & Smoking.

 TBA

#### 14.3 Alcohol and Marriage.



	Abstain	1 - 60	over 60	Rows
Single	67:49	213:246	74:59	354
Widowed	85:116	633:589	129:142	847
Divorced	27:14	60:71	15:17	102
Columns	179	906	218	1303

$\chi^2 = 6.6 + 4.4 + 3.8 + 3.3 + 1.2 + 12.1 + 1.7 + 0.24 = 41.64$ ,  $df = (3-1) \cdot (3-1) = 4$ ,  
Critical value  $\chi^2_{4,1-0.05} = 9.448$ , Decision: Dependent.

#### 14.4 Family Sizes.



```
 [chi2,pvalue,exp]=tablerxc([145 81 57 22 9 8; ...
151 73 71 33 13 10; 124 60 80 42 13 8])

%chi2 = 16.2783
%
%pvalue = 0.0919
%
%exp = 135.2400 68.9080 66.9760 31.2340 11.2700 8.3720
% 147.4200 75.1140 73.0080 34.0470 12.2850 9.1260
% 137.3400 69.9780 68.0160 31.7190 11.4450 8.5020
```

### 14.5 Nightmares.



```
 [chi2, pvalue, exp]=tablerxc([55 60; 105 132])
%exp = 52.2727 62.7273
%      107.7273 129.2727
%chi2 =    0.3875
%pvalue =   0.5336
```

### 14.6 Independence of Segregation.



### 14.7 Site of Corpus Luteum in Caesarean Births.



### 14.8 An Easy Grade?



expected	prof A	prof B	prof C	total
grades A	15	18	17	50
grades B	24	28.8	27.2	80
grades C	21	25.2	23.8	70
total	60	72	68	200

$\chi^2 = 24.037$  exceeds critical value  $\chi^2_{4,0.99} = 13.277$ . Reject  $H_0$ .

### 14.9 Importance of Bystanders.



$H_0$ : Assistance and the number of bystanders are independent.  
MATLAB output

```
 [chisq, p, expected]=tablerxc([11 2; 16 10; 4 9])

% chisq =    7.9078

% p =     0.0192

% expected =
%    7.7500    5.2500
%   15.5000   10.5000
%    7.7500    5.2500
```

yields the  $p$ -value smaller than 2%. The statistic  $\chi^2 = 7.9078$  is significant, that is,  $H_0$  is rejected.

#### 14.10 Baseball in 2003.



#### 14.11 Psychosis in Adopted Children.

Hint. For (a) and (b) use theory from Chapter (Two Samples, page 111) since the tables are not paired.

#### 14.12 The Midtown Manhattan Study.



#### 14.13 Tonsillectomy and Hodgkin's Disease.



We used m-file unmatchch.m

```
%or = 2.1429
%chi2 = 2.2273
%lor = 0.7621
%varlor = 0.2608
%stdlor = 0.5107
%cill = -0.2388
%cilu = 1.7631
%cil = 0.7876
%ciu = 5.8303
```

#### 14.14 School Spirit at Duke.



#### 14.15 Two Halloween Questions with Easy Answers.



#### 14.16 Runners and Heart Attack.



(a)  $\chi^2 = 2.639$ , critical value is  $\chi_{1,0.95} = 3.841$ , do not reject  $H_0$ . (b) The error of second kind is to accept the hypothesis that running and heart attacks are independent, when in fact, they are dependent.]

#### 14.17 Perceptions of Dangers of Smoking.



#### 14.18 Red Dye No 2.

 TBA

#### 14.19 Cooper Hawks.

 TBA

#### 14.20 Hepatic Arterial Infusion.

 TBA

#### 14.21 Vaccine Efficacy Study.

 TBA

### 14.1 Additional Problems

**14.a1 Paired Odds Ratio.** Show, by considering Mantel-Haenszel methodology (page 555), that the odds ratio for paired table [13;42] is equal to 3/4.

 By parallelizing the paired table, there are one table [11;00], three [10;01], four [01;10], and two [00;11].

Since for multiple tables  $OR = \frac{\sum_i a_i d_i / n_i}{\sum_i b_i c_i / n_i}$  (page 551), it follows  $\sum_i a_i d_i / n_i = B/2, \sum_i b_i c_i / n_i = C/2$ , where [AB;CD] is the original paired table.

**14.a2 Chlordiazepoxide Use and Congenital Heart Defects.** Medication chlordiazepoxide (Librium) is indicated for the relief of acute agitation and hyperactivity (e.g., alcoholism, anxiety, hysterical and panic states, drug withdrawal) via its sedative, appetite-stimulating and weak analgesic actions. Rothman et al. (1979) explored the link between chlordiazepoxide use in early pregnancy and incidence of congenital heart defects in babies. The retrospective analysis is summarized in the following table:

	Chlordiazepoxide Use		
	Yes	No	Total
Case Mothers	4	386	390
Control Mothers	4	1250	1254
Total	8	1636	1644

Let  $p_1$  and  $p_2$  be the probabilities of a birth with congenital heart defect for exposed and control mothers, respectively.

(a) By using MATLAB and Fisher's exact test, test the hypothesis  $H_0 : p_1 = p_2$  versus the one sided alternative  $H_1 : p_1 > p_2$ .

(b) Compare this test with the test for two proportions (normal approximation Z, page 378).

 TBA

(a) The p-value is  $\sum_{k=4}^8 \binom{8}{k} \binom{1636}{390-k} / \binom{1644}{390} = 1 - \text{hygecdf}(3, 1644, 8, 390) = 0.0964 > 5\%$ .

Sometimes, the *mid-p value* is reported. The mid-*p* value is defined as a tail probability where the observed value is taken with weight 1/2,  $1/2 \times \binom{8}{4} \binom{1636}{386} / \binom{1644}{390} + \sum_{k=5}^8 \binom{8}{k} \binom{1636}{390-k} / \binom{1644}{390} = 0.5 * \text{hygepdf}(4, 1644, 8, 390) + (1 - \text{hygecdf}(4, 1644, 8, 390)) = 0.0590$ .

(b) The approximation is closer to mid-*p* value, and not very accurate given the fact that there are only 8 births with congenital heart defects. This could be misleading since at 5% level, the nonsignificant finding would be declared significant.

```
p1=4/390; p2 = 4/1254; n1=390; n2=1254;
pbar= n1/(n1+n2) * p1 + n2/(n1+n2) * p2 %pbar = 0.0049
z = (p1 - p2)/sqrt( pbar * (1-pbar) * (1/n1+1/n2)) %1.7515
1-normcdf(1.7515) %0.0399 %p-value
```

Rothman, K. J., Fyler, D. C., Goldblatt, A., and Kreidberg, M. B. (1979). Exogenous hormones and other drug exposures of children with congenital heart disease. *Am. J. Epidemiol.*, **109**, 433–439.



# Chapter 15

## Correlation

### 15.1 Correlation Between Uniforms and Their Squares.



```
a = 2 * rand(10000,1) - 1;  
b = a.^2;  
corr(a,b)
```

### 15.2 Muscle Strength of “Ehtanol Abusers”.



**Hints:** (a) Statistic  $t = r_{HS} \sqrt{\frac{n-2}{1-r_{HS}^2}}$  has Student  $t$  distribution with  $n-2$  degrees of freedom. The alternative is one sided (upper tail critical), p-value is  $1-tcdf(t, n-2)$ .

(b) Recall,  $r_{HS.A} = \frac{r_{HS}-r_{HA}r_{SA}}{\sqrt{(1-r_{HA}^2)(1-r_{SA}^2)}}$ . Statistic  $t = r_{HS.A} \sqrt{\frac{n-1-2}{1-r_{HS.A}^2}}$  has Student  $t$  distribution with  $n-3$  degrees of freedom.

(c) Find 95% CI for  $\omega = \frac{1}{2} \log \frac{1+\rho_{HS}}{1-\rho_{HS}}$  which is population counterpart of  $w = \frac{1}{2} \log \frac{1+r_{HS}}{1-r_{HS}}$ . The latter has normal distribution,

$$w \sim \mathcal{N}\left(\omega, \frac{1}{n-3}\right),$$

which is useful to find CI for  $\omega$ . Back transform lower and upper bounds of CI for  $\omega$  by  $r = \frac{e^{2w}-1}{e^{2w}+1}$ .

### 15.3 Vending Machine and Pharmacy Errors.



### 15.4 Vending Machine and Pharmacy Errors Revisited.



```

errors= [ 2,  3, 10,  9,  5,  7,  8,  4]';
coke =[112, 100, 220, 250, 100, 200, 160, 100]';
people = [10000, 6000, 17000, 20000, 9000, ...
           15000, 14000, 8000]';
corr(errors, coke)
% 0.8785
corr(errors, people)
% 0.8821
corr(coke, people)
% 0.9735
(0.8785-0.8821*0.9735)/(sqrt(1-0.8821^2)*sqrt(1-0.9735^2))
% 0.1836

```

### 15.5 Corn Yields and Rainfall.



### 15.6 Drosophilæ.



Grand Canyon:  $w_1 = 0.5763$  Flagstaff:  $w_2 = 0.8107$   
The test statistic for  $H_0 : \rho_1 = \rho_2$  is:  $z = \frac{0.5763 - 0.8107}{\sqrt{1/36 + 1/17}} = -0.7965$ .  
2\Phi(-0.7965) = 0.4257. Conclusion:  
Do not reject null hypothesis.

### 15.7 Confidence Interval for the Difference of Two Correlation Coefficients.



### Oxygen Intake.



### 15.9 Obesity and Pain.



(a)  $(4461.5 - 10 * 62.7 * 7.7) / (\sqrt{45141 - 10 * 62.7^2} * \sqrt{799.5 - 10 * 7.7^2}) = -0.3339$ .  
(b)  $t = \sqrt{(n-2)} * r / \sqrt{1-r^2} = -1.0019$ .  $pvaluecdf_t(-1.0019, 10-2) = 0.1729$ ,  
(c)  $(-0.3339 + 0.2089 * 0.8627) / \sqrt{1 - 0.2089^2} / \sqrt{1 - 0.8627^2} = -0.3107$ .  
(d)

```

omint=[om-1.96/sqrt(10-3) om+1.96/sqrt(10-3)]
      % omint = -1.0880    0.3936
(exp(2*omint)-1)./(exp(2*omint) + 1)
      % ans = -0.7962    0.3745

```

## 15.1 Additional Problems

**15.a1 Correlation between  $X_i$  and  $\bar{X}$ .** Let  $X_1, X_2, \dots, X_n$  be independent with common variance  $\sigma^2$ . Show that

$$\text{Corr}(X_i, \bar{X}) = 1/\sqrt{n}, \quad 1 \leq i \leq n.$$



Without loss of generality assume that  $EX_i = 0$ . Then

$$\text{Cov}(X_i, \bar{X}) = E(X_i \bar{X}) = X_i^2/n + \sum_{j \neq i} EX_i X_j/n = X_i^2/n + EX_i \sum_{j \neq i} EX_j/n = \sigma^2/n.$$

The correlation is

$$\text{Corr}(X_i, \bar{X}) = \text{Cov}(X_i, \bar{X}) / [\text{Var}(X_i) \text{Var}(\bar{X})]^{1/2} = \frac{\sigma^2/n}{\sqrt{\sigma^2 \cdot \sigma^2/n}} = 1/\sqrt{n}.$$



# Chapter 16

## Regression

### 16.1 Regression with Three Points.

 TBA

### 16.2 Age and IVF Success Rate.

 TBA

### 16.3 Sharp Dissection and Severity of Postoperative Adhesions.

 TBA

```
 lasd = [ 2.4849    3.2581    3.3322    3.5835 ...
          3.6109    3.6889    3.8918    4.4188 ...
          4.5433    4.5643    4.5951    4.5951 ...
          4.6540    4.7875    4.8752    4.8978 ...
          4.9053    5.0499    5.5255    5.8051 ...
          6.0186    6.0210]; ...

sesco = [6   7   7   7   9   9   8   14  13  10  10 ...
          10  11  12  12  12  12  15  16  18  17  18];

x = lasd';
y = sesco';
n = length(x) %n=22
p = 2; %number of parameters (beta0, beta1)
% Sums of Squares
SXX = sum( (x - mean(x)).^2 ) %17.9017
SYY = sum( (y - mean(y)).^2 ) %279.5000
SXY = sum( (x - mean(x)).* (y - mean(y)) ) %65.8276

% estimators of coefficients beta1 and beta0
b1 = SXY/SXX %3.6772
b0 = mean(y) - b1 * mean(x) %-5.0651
% predictions
yhat = b0 + b1 * x;
```

```
%residuals
res = y - yhat;
% ANOVA Identity
SST = sum( (y - mean(y)).^2 )      % 279.5000
SSR = sum( (yhat - mean(y)).^2 )    % 242.0592
SSE = sum( (y - yhat).^2 )         % 37.4408
% forming F, test of adequacy of linear regression
MSR = SSR/(p - 1) % 242.0592
MSE = SSE/(n - p) %should be sigma2hat,     1.8720
F = MSR/MSE %129.3023
pvalue = 1-fcdf(F, p-1, n-p)
%H_0: regression has beta1=0, no need for
% linear fit   pval= 3.4983e-010
% Other measures of goodness of fit
R2 = SSR/SST %0.8660
R2adj = 1 - (n-1)/(n-p)* SSE/SST %0.8593
s = sqrt(MSE) % 1.3682
% Standard error of coefficient estimators
sb1 = s/sqrt(SXX) % 0.3234
sb0 = s * sqrt(1/n + (mean(y))^2/SXX ) %3.7303

% are the coefficients equal to 0?
t1 = b1/sb1 %11.3711
pb1 = 2 * (1-tcdf(abs(t1),n-p) ) % 3.4983e-010
t0 = b0/sb0 % -1.3578
pb1 = 2 * (1-tcdf(abs(t0),n-p) ) % 0.1896
% predicting y for the new observation x, CI and PI
newx = 4;
ypred = b0 + b1 * newx % 9.6436
sym = s * sqrt(1/n + (mean(x) - newx)^2/SXX )
% for y mean 0.3343
syp = s * sqrt(1 + 1/n + (mean(x) - newx)^2/SXX )
% for y prediction 1.4085

%intervals CI and PI
alpha = 0.05;
%mean response interval
lbym = ypred - tinv(1-alpha/2, n-p) * sym;
rbym = ypred + tinv(1-alpha/2, n-p) * sym;
% prediction interval
lbyp = ypred - tinv(1-alpha/2, n-p) * syp;
rbyp = ypred + tinv(1-alpha/2, n-p) * syp;
%print the intervals
[lbym rbym] % 8.9463 10.3409
[lbyp rbyp] % 6.7055 12.5816
```

#### 16.4 Kanamycin Levels in Premature Babies.



#### 16.5 Degradation of Scaffolds.



```
% (a) [56.2193 - 6.6109 day], 0.6818.
% (b) t = -1.3713, pval = 0.0921
% (c)[-8.5882, -4.6336]
% (d) 19.8593, 11.1969
```

## 16.6 Glucosidase in Lactococcus Lactis.



**16.7 Weight and Latency in Rats.** Data consisting of rat body weight (grams) and latency to seizure (minutes)



```
p = 2; %number of parameters (beta0, beta1)
%"wei" measurement is "x", "latency" is "y".
x = wei ; %column vector
mean(x) % xbar = 411
y = latency ; %column vector
n = length(x);
% Sums of Squares
SXX = sum( (x - mean(x)).^2 ) %SXX=17754
SYY = sum( (y - mean(y)).^2 ) %SYY=8.4168
SXY = sum( (x - mean(x)).*(y - mean(y)) ) %SXY=258.53
% estimators of coefficients beta1 and beta0
b1 = SXY/SXX %0.0146
b0 = mean(y) - b1 * mean(x) %-3.6436
% predictions
y_hat = b0 + b1 * x;
%residuals
res = y - y_hat;
% ANOVA Identity
SST = sum( (y - mean(y)).^2 ) %which is SYY=8.4168
SSR = sum( (y_hat - mean(y)).^2 ) %3.7647
SSE = sum( (y - y_hat).^2 ) %sum(res.^2), 4.6521
% forming F and test of adequacy of regression
MSR = SSR/(p - 1) %3.7647
MSE = SSE/(n - p) %estimator of variance, 0.3579
s = sqrt(MSE) %0.5982
F = MSR/MSE %10.5201
pvalue = 1-fcdf(F, p-1, n-p)
%testing H_0: regression has beta1=0,
%that is no need for linear fit, p-val = 0.0064
% Other measures of goodness of fit
R2 = SSR/SST %0.4473
R2adj = 1 - (n-1)/(n-p)* SSE/SST %0.4048
% Standard deviations of coefficient estimators
sb1 = s/sqrt(SXX) %0.0045
sb0 = s * sqrt(1/n + (mean(x))^2/SXX) %1.8517
% are the coefficients equal to 0?
t1 = b1/sb1 %3.2435
pb1 = 2 * (1-tcdf(abs(t1),n-p)) %0.0064
t0 = b0/sb0 %-1.9677
pb0 = 2 * (1-tcdf(abs(t0),n-p)) %0.0708
% predicting y for the new observation x, CI and PI
newx = 410; %wei = 410
y_newx = b0 + b1 * newx % 2.3268
sym = s * sqrt(1/n + (mean(x) - newx)^2/SXX)
%st.dev. for mean response, sym = 0.1545
syp = s * sqrt(1 + 1/n + (mean(x) - newx)^2/SXX)
%st.dev. for the prediction syp = 0.6178
alpha = 0.05;
```

```
%mean response interval
lbym = y_newx - tinv(1-alpha/2, n-p) * sym;
rbym = y_newx + tinv(1-alpha/2, n-p) * sym;
% prediction interval
lbyp = y_newx - tinv(1-alpha/2, n-p) * syp;
rbyp = y_newx + tinv(1-alpha/2, n-p) * syp;
%print the intervals
[lbym rbym] % 1.9929 2.6606
[lbyp rbyp] % 0.9920 3.6615
```

### 16.8 Rinderpest virus in Rabbits.



### 16.9 Hemodilution.



### 16.10 Anscombe's Data Sets.



### 16.11 Potato Leafhopper.



### 16.12 Crossvalidating Bayesian Regression.



```
model{
  for( i in 1 : m ) {
    mu[i] <- beta0 + beta1 * x1[i] + beta2 * x2[i]
    y[i] ~ dnorm(mu[i],tau)
  }

  for( i in m+1 : n ) {
    ypred[i] <- beta0 + beta1 * x1[i] + beta2 * x2[i]
    error[i] <- ypred[i] - y[i]
    se[i] <- error[i] * error[i]
  }
  mse <- mean(se[m+1:n])

  beta0 ~ dnorm( 0.0,1.0E-5)
  beta1 ~ dnorm( 0.0,1.0E-5)
  beta2 ~ dnorm( 0.0,1.0E-5)
  tau ~ dgamma(0.001,0.001)
  sigma <- 1/sqrt(tau)
}
```

### DATA

```

list(n=40, m=20,
  x1=c(0.17, 0.39, 0.83, 0.80, 0.06, 0.39, 0.52, 0.41,
       0.65, 0.62, 0.29, 0.43, 0.01, 0.98, 0.16, 0.10,
       0.37, 0.19, 0.48, 0.33, 0.95, 0.92, 0.05, 0.73,
       0.26, 0.42, 0.54, 0.94, 0.41, 0.98, 0.30, 0.70,
       0.66, 0.53, 0.69, 0.66, 0.17, 0.12, 0.99, 0.17),
  x2=c(1, 6, 9, 7, 2, 4, 5, 10, 2, 9, 7, 4, 2, 5,
       5, 2, 6, 3, 4, 6, 3, 3, 7, 3, 9, 10, 8, 4,
       6, 2, 10, 9, 9, 3, 6, 1, 5, 4, 2, 2),
  y=c(3.038, 1.984, 3.241, 2.526, 1.532, 2.585, 1.855, -1.092,
      5.807, 1.162, 0.563, 2.660, 0.584, 4.956, 0.857, 0.877,
      1.859, 2.143, 2.280, 0.825, 5.259, 4.260, -0.394, 4.512,
      -0.623, -0.275, 1.304, 4.853, 0.748, 6.598, -2.140, 0.861,
      2.676, 3.779, 2.214, 5.466, -0.333, -0.311, 6.785, 2.789)
)

INITS
list( beta0=0, beta1=0, beta2=0, tau=1)

```

### 16.13 Taste of Cheese.

 Use the code  tastecheese.m

### 16.14 Slowing the Progression of Arthritis.

 TBA

### 16.15 Insulin on Opossum Liver.

 TBA

**16.16 Slope in EIV regression.** Show that the EIV regression slope in (??) tends to  $S_{xy}/S_{xx}$  when  $\eta \rightarrow 0$ .

 TBA

### 16.17 Interparticular Spacing and Wavelength in Nanoprisms 2.

 TBA

## 16.1 Additional Problems

**16.a1 Failures of Silver-zinc Batteries.** Silver-zinc batteries feature a water-based chemistry and contain no lithium or flammable liquids. Developed originally for satellite applications, these batteries are beginning to replace lithium-ion batteries in mobile phones, laptop computers, and battery-

powered medical devices. For example, some modern implantable hearing aids are powered by silver-zinc rechargeable batteries.

The data provided in `silverzinc.dat|mat` are collected in the 1980's when silver-zinc battery technologies have been analyzed by NASA (Sidek et al, 1980; also Johnson and Wichern, 2007). The response variable is `ctf` - the number of cycles-to-failure, while the covariates are `chr` - charge rate (in Amp), `dchr` - discharge rate (in Amp), `ddch` - depth of discharge (in % of rated Amp/hours), `temp` - temperature (in degC), and `ecv` - end of charge voltage (in Volts).

(a) Find a 95% CI for the coefficient of correlation between `temp` and `log(ctf)`.

(b) Propose a linear regression model to predict logarithm of cycles-to-failure, `log(ctf)`, that uses a subset of predictors from `chr`, `dchr.^2`, `ddch`, `sqrt(temp)`, and `ecv`. Defend the choice of your model (one paragraph).

- Sidek, S., Leibacki, H., and Bozek, J. (1980). Failure of silver-zinc cells with competing failure modes: preliminary data analysis. *NASA Technical Memorandum 81556*, Lewis Research Center, Cleveland OH.

- Johnson, R. and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*, 6th edition. Prentice Hall, Upper Saddle River, NJ.

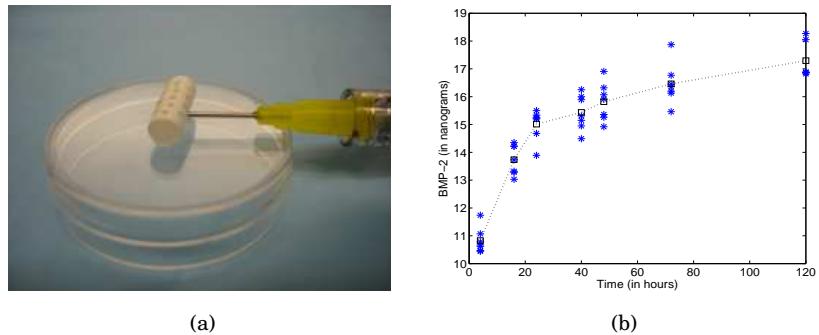
**16.a2 ANOVA Table from  $r$  and SST.** Fully recover ANOVA table in regression with  $n = 26$  pairs of observations  $(x, y)$ , for which  $r = 0.88$  and  $SST = 134.75$ .



Source	SS	DF	MS	F	p-value
Regression	$r^2SST$	1	$r^2SST$	$\frac{(n-2)r^2}{1-r^2}$	1-fcdf(F, 1, n-2)
Error	$(1-r^2)SST$	$n-2$	$\frac{(1-r^2)SST}{n-2}$		
Total	$SST$	$n-1$			

**16.a3 Release kinetics of BMP-2 from alginate hydrogels.** (Courtesy of Lauren Priddy) Polymeric biomaterials such as alginate are promising cell and protein delivery vehicles for bone tissue engineering due to their biocompatibility, moldability, and tunable degradation rates. Alginate hydrogels have been used to deliver bone morphogenetic protein-2 (BMP-2) in critically-sized rat bone defect models. Partial oxidation, whereby a small percentage of the uronate residues are oxidized, allows the polymer chains to be more susceptible to hydrolysis and increases the degradation rate in vitro. The goal of this experiment was to determine the release kinetics of BMP-2 from oxidized alginate hydrogels.

In this study, oxidized alginate hydrogels were loaded with BMP-2 and incubated in media, Figure 16.1(a). The media were collected and replaced with fresh media at the following time points: 4, 16, 24, 40, 48, 72, and 120



**Fig. 16.1** (a) Injection of alginate hydrogel into nanofiber mesh prior to incubation; (b) Scatterplot of cumulative BMP-2 over time.

hours. The amount of BMP-2 (nanograms) at each time point was quantified from the media collections using an enzyme-linked immunosorbent assay (ELISA). The cumulative amount of BMP-2 released at the preselected times is shown in the table below. The data in the table below are also given in `bmp2.mat|dat|xlsx`.

Time (hours)	BMP-2							
4	11.07	11.74	10.44	10.78	10.62	10.67	10.47	
16	13.28	14.35	13.32	14.22	13.73	14.22	13.03	
24	13.89	14.68	15.26	15.23	15.50	15.20	15.34	
40	16.00	15.31	14.95	15.15	14.49	16.25	15.89	
48	15.26	15.91	16.07	16.32	16.91	15.36	14.92	
72	15.46	16.77	17.87	16.45	16.36	16.20	16.13	
120	16.87	16.89	16.87	16.82	18.27	18.05		

From the scatterplot in Figure 16.1(b) it is evident that a linear fit, with time as a covariate and BMP2 as a response, is inadequate.

(a) Transform time to `lt=log(time)`, and inspect the scatterplot of `BMP2` against `lt`. Find the 95% confidence interval for the correlation between `lt` and `BMP2`, and comment on the adequacy of linear regression now.

(b) Find the linear relationship

$$\boxed{\text{BMP2} = b_0 + b_1 * \text{lt},}$$

where  $b_0$  and  $b_1$  are estimators of the population intercept and slope,  $\beta_0$  and  $\beta_1$ . The error  $\epsilon$  in the population equation  $\text{BMP2} = \beta_0 + \beta_1 \text{lt} + \epsilon$ , is assumed normal with mean 0 and variance  $\sigma^2$ . What is the estimator of this variance?

(c) If you are to predict the `BMP2` at `time = 100`, what are the 95% CIs for (1) a response in a single future experiment, and for (2) an average response. Comment why the intervals are not identical? [Find the intervals. Explain their difference in one or two sentences].



```

% y=[...
%   11.07    11.74    10.44    10.78    10.62    10.67    10.47 ...
%   13.28    14.35    13.32    14.22    13.73    14.22    13.03 ...
%   13.89    14.68    15.26    15.23    15.50    15.20    15.34 ...
%   16.00    15.31    14.95    15.15    14.49    16.25    15.89 ...
%   15.26    15.91    16.07    16.32    16.91    15.36    14.92 ...
%   15.46    16.77    17.87    16.45    16.36    16.20    16.13 ...
%   16.87    16.89    16.87    16.82    18.27    18.05 ];
%
% x=[ 4 4 4 4 4 4 4      16 16 16 16 16 16 16 ...
%     24 24 24 24 24 24 24      40 40 40 40 40 40 40 ...
%     48 48 48 48 48 48 48      72 72 72 72 72 72 72 ...
%     120 120 120 120 120 120 120];
%
close all
load 'bmp2.mat'
x = bmp2(:,1);
y=bmp2(:,2);
%(a)
lt = log(x);
[r pval lb ub] = corrcoef(lt, y)

% r =1.0000    0.9528
% 0.9528    1.0000
%
% lb = 1.0000    0.9169
% 0.9169    1.0000
%
% ub = 1.0000    0.9734
% 0.9734    1.0000
%
%(a) by hand
r=corr(lt, y)    % r =0.9528
%fisherz = @(x) 1/2*log( (1+x)/(1-x) );
w = 1/2 * log( (1+r)/(1-r))    % w =1.8616
n=length(y)    % n =48
LB = w - norminv(1-0.05/2) / sqrt(n-3) % LB = 1.5694
UB = w + norminv(1-0.05/2) / sqrt(n-3) % UB = 2.1537
%invfisherz = @(x) (exp(2 * x) - 1)/(exp(2 * x) + 1)
L=(exp(2*LB)- 1)/(exp(2*LB)+ 1)    % L = 0.9169
U=(exp(2*UB)- 1)/(exp(2*UB)+ 1)    % U = 0.9734

%(b)
[b] = regress(y,[ones(size(y)) lt])
stats= regstats(y, [lt]);  % 0.9079  453.3008    0.0000    0.4021

newx = log(100);
n = length(lt);
% Sums of Squares
SXX = sum( (lt - mean(lt)).^2 ) %SXX=50.5530
y_newx = b(1) + b(2) * newx %17.1912
sym = stats.mse * sqrt(1/n + (mean(lt) - newx)^2/SXX )
%st.dev. for mean response, sym = 0.0898
syp = stats.mse * sqrt(1 + 1/n + (mean(lt) - newx)^2/SXX )
%st.dev. for the prediction syp = 0.4120
alpha = 0.05;
%mean response interval
lbym = y_newx - tinv(1-alpha/2, n-2) * sym;
rbym = y_newx + tinv(1-alpha/2, n-2) * sym;
[lbym rbym]    % 17.0105    17.3719

% prediction interval

```

```
lbyp = y_newx - tinv(1-alpha/2, n-2) * syp;
rbyp = y_newx + tinv(1-alpha/2, n-2) * syp;
[lbyp rbyp] % 16.3619 18.0205
```



# Chapter 17

## Regression for Binary and Count Data

### 17.1 Blood Pressure and Heart Disease.

 TBA

### 17.2 Blood Pressure and Heart Disease in WinBUGS.

*Hint:* Beetles Example may help in setting up BUGS code.

 TBA

### 17.3 Sex of Turtles and Incubation Temperature.

 TBA

### 17.4 Health Promotion.

 TBA

### 17.5 PONV.

 TBA

### 17.6 Mannose-6-phosphate Isomerase.

 TBA

### 17.7 Arthritis Treatment Data.



```
 %arthritis2.m
load 'arthritis2.dat'
caseid = arthritis2(:,1);
treatment = arthritis2(:,2);
gender = arthritis2(:,3);
age      = arthritis2(:,4);
improve = arthritis2(:,5);
```

```

improve01 = arthritis2(:,5)>0 ;
X = [treatment gender age];
[betas, deviance, stats]=glmfit(X,improve01,'binomial','link','comploglog')

figure(1)
score = betas(1) + betas(2)*treatment + betas(3)*gender + betas(4)* age;
plot(score, improve01,'o',...
'MarkerSize',msize, 'MarkerEdgeColor','k', 'MarkerFaceColor','g')

xx = -2.7:0.01:1.4;
imp = 1 - exp(- exp(xx) );
hold on
plot(xx, imp,'r-','LineWidth',lw)
xlabel('Score')
ylabel('Probability of Improving')

[betas2, deviance2, stats2]=glmfit(X,improve01,'binomial','link','logit')
[betas3, deviance3, stats3]=glmfit(X,improve01,'binomial','link','probit')
score2 = betas2(1) + betas2(2)*treatment + betas2(3)*gender + betas2(4)* age;
score3 = betas3(1) + betas3(2)*treatment + betas3(3)*gender + betas3(4)* age;

imp = 1 - exp(- exp(score) );
imp2 = exp(score2)./(1 + exp(score2));
imp3 = normcdf(score3);
plot(score, imp,'r*')
hold on
plot(score2, imp2,'ko')
plot(score3, imp3,'bd')
xlabel('scores')
ylabel('fits')
legend('cloglog','logit','probit',2)

deviance %92.0751
deviance2 %92.0628
deviance3 %91.9286

```

### 17.8 Third-degree Burns.



### 17.9 Diabetes Data.

 TBA**17.10 Remission Ratios over Time.** TBA**17.11 Death of Sprayed Flour Beetles.** TBA**17.12 Mortality in Swiss White Mice.** TBA**17.13 Kyphosis Data.** TBA**17.14 Prostate Cancer.** TBA**17.15 Pediculosis Capitis.** TBA**17.16 Finney Data.** TBA**17.17 Shocks.** TBA**17.18 Ants.** TBA**17.19 Sharp Dissections and Postoperative Adhesions Revisited.** TBA**17.20 Airfreight breakage.** TBA**17.21 Body Fat Affecting Accuracy of Heart Rate Monitors .** TBA**17.22 Miller Lumber Company Customer Survey.** TBA**17.23  $SO_2, NO_2$ , and Hospital Admissions .**



### 17.24 Kidney Stones.



## 17.1 Additional Problems

**17.a1 Bumpus' Sparrows Data.** After an unusually severe storm in February of 1898, a number of house sparrows, *Passer domesticus*, were brought to the Anatomical Laboratory of Brown University, Providence, Rhode Island. Seventy-two of these birds revived; sixty-four perished. This event is described by Hermon Carey Bumpus, the first PhD graduate of Clark University, whose paper (Bumpus, 1898) has served as an example of natural selection in action. The data set provided by Bumpus included several anatomic measurements on 136 birds (as data structure `bumpus.mat`) and had been analyzed since by many diverse researches.

---

<code>sex</code>	1 = male; 2 = female
<code>surv</code>	1 if survived, 0 if perished
<code>lbt</code>	Length (mm) from tip of the beak to the tip of the tail
<code>ae</code>	Alar extent (mm) from tip to tip of the extended wings
<code>wei</code>	Weight (g)
<code>lbh</code>	Length of beak and head (mm), from tip of the beak to the occiput
<code>hum</code>	Length of Humerus [arm/wing bone] (in)
<code>fem</code>	Length of Femur [thigh bone] (in)
<code>tib</code>	Length of Tibiotarsus [leg bone linked to femur] (in)
<code>wos</code>	Width of Skull (in), from the postorbital bone of one side to the postorbital bone of the other
<code>kos</code>	Length of Keel of Sternum [an extension of breastbone] (in)

---

By using logistic regression, model the probability of survival for male sparrows (`sex = 1`) using the covariates `lbt`, `ae`, `wei`, `lbh`, `hum`, `fem`, `tib`, `wos`, and `kos`.

There is an agreement that lighter and shorter birds have a higher chance of survival. How is this reflected in your model?

- Bumpus, H. C. (1898) The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*. Biological Lectures at Woods Hole Marine Biological Laboratory, 11th Lecture, 209–225.

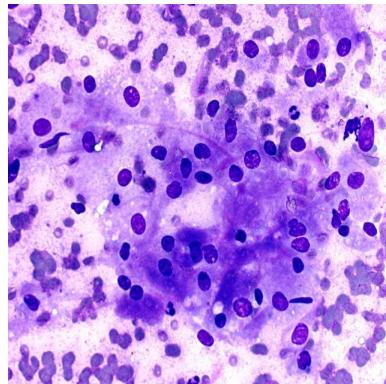


**3. Wisconsin Diagnostic Breast Cancer (WDBC).** Wolberg, Street, and Mangasarian, from the University of Wisconsin,<sup>1</sup> were interested in machine learning in diagnosing breast cancer from fine-needle aspirates.

The data set `wdbc.mat` constitutes a matrix `wdbc` with 569 rows (subjects) of which 357 correspond to controls and 212 to cancer. The matrix has 31 columns: column 1 is diagnosis (0 = control, 1 = cancer), while the columns 2-31 contain 30 features. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, see Figure 17.1. They describe characteristics of the cell nuclei present in the image.

Variable	Mean	S.Error	Extreme
Radius (average distance from the center)	Col 2	Col 12	Col 22
Texture (standard deviation of gray-scale values)	Col 3	Col 13	Col 23
Perimeter	Col 4	Col 14	Col 24
Area	Col 5	Col 15	Col 25
Smoothness (local variation in radius lengths)	Col 6	Col 16	Col 26
Compactness (perimeter <sup>2</sup> / area - 1.0)	Col 7	Col 17	Col 27
Concavity (severity of concave portions of the contour)	Col 8	Col 18	Col 28
Concave points (number of concave portions of the contour)	Col 9	Col 19	Col 29
Symmetry	Col 10	Col 20	Col 30
Fractal dimension (“coastline approximation” - 1)	Col 11	Col 21	Col 31

The mean, standard error, and extreme (largest) of nuclei measures were computed for each image, resulting in 30 features. For instance, column 2 is Mean Radius, column 12 is Radius Standard Error, column 22 is Extreme Radius.



**Fig. 17.1** FNA: A digitized image of a fine needle aspirate of a breast mass.

<sup>1</sup> Wolberg, W. H., Street, W. N., and O.L. Mangasarian, O. L., (1994). Machine learning techniques to diagnose breast cancer from fine-needle aspirates. *Cancer Letters*, **77** (1994) 163-171.

Wolberg, W. H., Street, W. N., and O.L. Mangasarian, O. L., (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative Cytology and Histology*, **17**, 2, 77-87.

(a) Propose a logistic regression model in which the incidence of malignancy is regressed on Mean Texture (3rd column), Extreme Area (25th column), and Extreme Smoothness (26th column). These three covariates are recommended by the authors as good separating variables.

(b) Find the probability of malignancy suggested by the model in (a) for a new case where Mean Texture, Extreme Area, and Extreme Smoothness, are 21.423, 654.787, and 0.118, respectively.



```
%WDBC
load 'wdbc.mat'
%
Y = wdbc(:,1);
X = wdbc(:,[3 25 26]);    %Design matrix n x (p-1) without
                           % vector of 1's (intercept)
Xdes =[ones(size(Y)) X]; %with the intercept: n x p
[n p] = size(Xdes);
alpha = 0.05;  %alpha for CIs

[b, dev, stats]=glmfit(X,Y, 'binomial','link','logit');

lin = Xdes * b;  %linear predictor, n x 1 vector

newperson= [1 21.423  654.787   0.118];
newlin = newperson * b      % -0.2020
prob = exp(newlin)/( 1 + exp(newlin) )  %0.4497

figure(1)
plot(lin, Y,'o','MarkerSize',msize,...
      'MarkerEdgeColor','k','MarkerFaceColor','g')
xx = -2:0.01:1;
mp = exp(xx)./(1 + exp(xx));
hold on
plot(xx, mp,'r-','LineWidth',lw)
plot( [newlin newlin],[0 prob],'r:')
plot([-2 newlin],[prob prob],'r:')
axis([-2 1 0 1])
xlabel('Linear Predictor','Interpreter','LaTeX')
ylabel('Probability of Cancer','Interpreter','LaTeX')
legend('Observations','Logistic Fit',2)
```

# Chapter 18

## Inference for Censored Data and Survival Analysis

### 18.1 Simulation of Censoring.



```
%survival1.m
y = exprnd(10,50,1); % Random failure times exponential(10)
d = exprnd(20,50,1); % Drop-out times exponential(20)
t = min(y,d); % Observe the minimum of these times
censored = (y>d); % Observe whether the subject failed

% Calculate and plot empirical cdf and confidence bounds
[f,x,flo,fup] = ecdf(t,'censoring',censored);
stairs(x,f,'LineWidth',2)
hold on
stairs(x,flo,'r:','LineWidth',2)
stairs(x,fup,'r:','LineWidth',2)
% Superimpose a plot of the known population cdf
xx = 0:.1:max(t);
yy = 1-exp(-xx/10);
plot(xx,yy,'g-','LineWidth',2)
legend('Empirical','LCB','UCB','Population',...
    'Location','SE')
hold off
```

### 18.2 Immunoperoxidase.



The 95% intervals are [0.0021,0.0063] for the first approximation, and [0.0026,0.0069] for the second.

The following is added to the MATLAB script in Example ??:

```
z0975 = norminv(0.975);
[hatlam1 - z0975* hatlam1/sqrt(k1), hatlam1 + z0975* hatlam1/sqrt(k1)]
%0.0021    0.0063
exp([log(lambdahat1) - z0975*sqrt(1/k1) ...
```

```
log(lambdahat1) + z0975*sqrt(1/k1])
%0.0026    0.0069
```

Note that the confidence interval found by MATLAB on the scale parameter was [153.6769, 415.7289]. By taking the reciprocals, an alternative confidence interval is [0.0024, 0.0065].

### 18.3 Massachusetts Data.



TBA

### 18.4 Expected Life-time.



$T$  is non-negative. Start with  $\mathbb{E}T = \int_0^\infty tf(t)dt$  and take  $u = t$  and  $dv = f(t)dt$ . But

$$dv = f(t)dt = d(F(t)) = d(1 - S(t)) = d(-S(t)) \rightarrow v = -S(t).$$

Now,  $\mathbb{E}T = uv|_0^\infty - \int_0^\infty (-S(t))dt = \int_0^\infty S(t)dt$ .

### 18.5 Censored Rayleigh.



TBA

### 18.6 MLE for Equally Censored Data.



TBA

### 18.7 Malignant Melanoma.



TBA

### 18.8 Rayleigh Survival Times.



(a) The cdf for Rayleigh distribution is  $F(t) = 1 - e^{-\lambda t^2}$  and  $S(t) = e^{-\lambda t^2}$  so that

$$h(t) = \frac{f(x)}{S(t)} = 2\lambda t.$$

(b) The mean survival time is  $\mu = \frac{1}{2} \sqrt{\frac{\pi}{\lambda}}$ . (c) As Example ?? . (d) The hazard is linear function of the parameter, thus the parameter is substituted by its Bayes estimator. For the survival function one can use the fact that the moment generating function for  $T \sim \mathcal{G}a(\alpha, \beta)$  is

$$\mathbb{E}e^t T = (1 - t/\beta)^{-\alpha}.$$

### 18.9 Western White Clematis.



TBA

# Chapter 19

## BUGS

### 19.1 A Coin and a Die.



```
#coin.bug:
model coin;
{
  flip12 ~ dcat(p.coin[])
  coin <- flip12 - 1
}
#coin.dat:
list(p.coin=c(0.5, 0.5))
# just generate initials
```

### 19.2 Paradox DeMere in WinBUGS.



The solution to the “paradox” deMere is simple. By taking into account all possible permutations of the above triples the sum 11 has 27 favorable permutations while the sum 12 has 25 favorable permutation.

But what if 300 fair dice are rolled and we are interested if the sum 1111 is advantageous to the sum 1112? Exact solution is unappealing, but the probabilities can be well approximated by WinBUGS model demerel.

```
model demerel;
{
  for (i in 1:300) {
    dice[i] ~ dcat(p.dice[]);
  }
  is1111 <- equals(sum(dice[]),1111)
```

```

is1112 <- equals(sum(dice[]),1112)
}

DATA
list(p.dice=c(0.1666666, 0.1666666,
0.1666667, 0.1666667, 0.1666667, 0.1666667) )

```

The initial values are generated. After five million rolls, WinBUGS outputs `is1111 = 0.0016` and `is1112 = 0.0015`, so the sum of 1111 is advantageous to the sum of 1112.

### 19.3 Simulating Probability of an Interval.



(a) `1/e - 1/e^1.6 ans =0.165982923176787`

(b) Recall that MATLAB parametrizes with scale parameter  $1/\lambda = 10$ , so  
`expcdf(16, 10) - expcdf(10,10) ans = 0.165982923176787`

(c)



```

model{
theta ~ dexp(0.1)
P <- step(theta-10)*step(16-theta)
}

```

There is no data to load in, and after checking the model in the Model's Specification Tool one proceeds directly to compiling. Also, the WinBUGS will generate a starting point for the MCMC iteration. The result after total of 10,000,000 iterations is

	mean	sd	MCerror	val2.5pc	median	val97.5pc	start	sample
P	0.1659	0.372	1.169E-4	0.0	0.0	1.0	1001	9999000

### 19.4 WinBUGS as a Calculator.

The solution is given by the following code



```

model{
F(x) <- sin(x)
int <- integral(F(x), 0, pi, 1.0E-6)
pi<- 3.141592659

y0 <- solution(F(y), 1,2, 1.0E-6)
F(y) <- pow(y,5) - 2*y
zero <- pow(y0, 5)-2*y0

randint <- integral(F(z), 0, randbound, 1.0E-6)
F(z) <- pow(z,3)*(1-pow(z,4))
randbound ~ dbeta(2,2)
}

```

NO DATA

INITS

```
list(x =1, y=0, z=NA, randbound=0.5)
```

After model checking, one should go directly to compiling (no data to load in) and to initializing the model. There is NO need to update the model or to go to Inference tool, set variables for monitoring and sample. One simply goes to Info menu and checks Node Info. In the Node Info Tool one specifies `int` for the approximation of integral, `y0` for the solution of equation, `zero` for checking that `y0` satisfies the equation (approximately), and `randint` for the value of random interval.