WEEK 10: CHAPTER 10

TABLE OF CONTENTS

Page:

List of Figures

List of Tables

List of m.files

Assignment

Required reading

! Larsen, R. J. and M. L. Marx. 2006. An introduction to mathematical statistics and its applications, 4th edition. Prentice Hall, Upper Saddle River, NJ. 920 pp.
" Read All of Chapter 10

Understanding by Design Templates

Understanding By Design Stage 1 — Desired Results

LM Chapter 10 Goodness-of-Fit Tests

G Established Goals

- Students will know how to perform goodness-of-fit tests to compare probability models with data
- Students will be know how to analyze contingency tables, distinguishing between tests of homogeneity and independence

U Understand

- That one degree of freedom is lost for every parameter estimated from data
- The sample-size requirements for goodness-of-fit and chi-square contingency table tests
- The difference between tests of independence and tests of homogeneity

Q Essential Questions

- What is the difference between a test of homogeneity versus a test of independence?
- What is the difference between uncorrelated and independent?

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

- **contingency table**, **goodness-of-fit tests**, **homogeneity**, **multinomial distribution**

S *Students will be able to*

- Use a chi-square goodness-of-fit tests to compare empirical observations to model predictions
- Analyze 2x2 and r x c contingency tables with chi-square and Fisher's exact tests

Understanding by Design Stage 2 — Assessment Evidence Week 10 Tu 8/2-8/8
Chapter 10 Goodness-of-Fit Tests

- **Post in the discussion section by 8/10/11**
  - In the lady tasting tea experiment, would the *p value* change if she had been allowed to put all 8 cups in the tea-first or milk-fist categories if she wished? Would it have made a difference to the *p value* if the lady wasn't told that there were exactly 4 each of the tea-first and milk-first cups?
- **HW 4 Problems due Wednesday 8/10/11 W 10 PM**
  - **Basic problems (4 problems 10 points)**
    - **Problem 10.2.2.** Mendel's peas. Hint: use LMex102010_4th.m as a model
    - Problem 10.3.6 Schizophrenia in England & Wales, hint: Use LMex100301_4th.m as a paradigm
    - Problem 10.4.4 Prussian horse kick data from question 4.2.10; Use LMcs100402_4th.m as a model; just plug in the Prussian cavalry data
    - Problem 10.5.2 Use 10.5.3 as a model
  - **Advanced problems (2.5 points each)**
    - **Problem** 10.2.4
    - **Problem 10.5.8**
  - **Master problems (1 only, 5 points)** Return to Case Study 1.2.1 Apply a goodness of fit test to the runs in the closing price of the S&P 500 for 2011.

Introduction

Chapter 10 is important. Before I started teaching EEOS601 and EEOS611 in 2000, we had 6 different professors teaching EEOS601 & EEOS611 over a 13-y period, usually using Sokal & Rohlf's Biometry as a text. Those 13 years of students had a curious blindspot. Few if any knew the difference between tests of correlation and independence. Few knew how contingency tables should be analyzed. You see, Sokal and Rohlf left tests of independence and contingency table analysis to the final portion of their text and our professors had never paced the course properly to reach and teach that material. My colleagues more versed in modern pedagogy than I tell me that in core courses, one doesn't have to cover the bases, so to speak. Just teach the students what they can absorb, and they'll be able to teach themselves the rest. Sadly, I don't think that is the case with statistics. So, with goodness of fit tests tucked in nicely between two-sample tests and regression, let's cover the bases. What is a goodness-of-fit test and how do you test contingency tables?

Chapter 10 begins by deriving the $\chi 2$ distribution from the multinomial distribution. Multinomial distributions arise in the biological and environmental sciences frequently. Figure 1 shows a ternary diagram, used commonly in geology to mark the sediment grain size samples made up of three components: sand, silt & clay. Such diagrams are also common in genetics in which they indicate the relative percentage of homozyous recessive, homozygous dominant, and heterozygous individuals in a population. The ternary diagram is the simplest graphical display of a multinomial vector. **Middleton (1999)** provides Matlab programs for programming ternary diagrams.

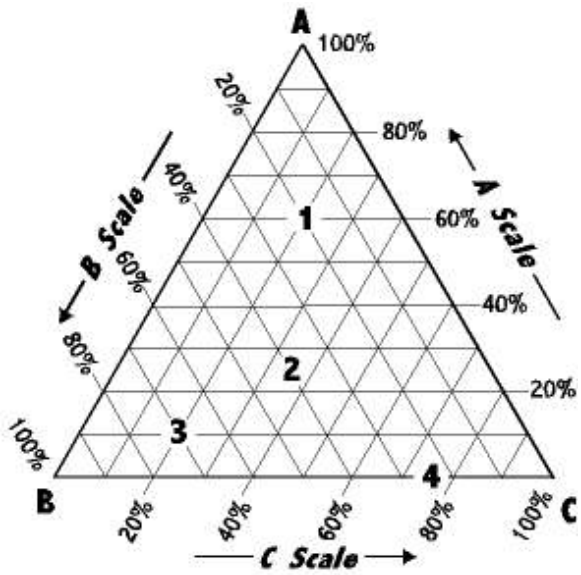

Figure 1. A ternary diagram.

Theorems and definitions

Theorem 10.2.1

Theorem 10.2.1. Let X_i denote the number of times that the outcome r_i occurs, i = 1, 2, ..., t, in a series of n independent trials, where $p_i = p(r_i)$. Then the vector (X_1, X_2, ..., X_t) has a **multinomial distribution** and

$$p_{x_1, x_2, \ldots, x_t} = P(X_1 = k_1, X_2 = k_2, \ldots, X_t = k_t)$$

$$\frac{n!}{k_1!\, k_2! \cdots k_t!}\, p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t}$$

$$k_i = 0, 1, 2, \ldots, n; \quad i = 1, 2, \ldots, t; \quad \sum_{i=1}^{t} k_i = n$$

Theorem 10.3.1

Theorem 10.3.1 Let r_1, r_2, ..., r_t be the set of possible outcomes (or ranges of outcomes) associated with each of n independent trials, where $P(r_i) = p_i$, $i = 1, 2, \ldots, t$. Let X_i = number of times r_i occurs, i = 1, 2, ..., t. Then

a. The random variable

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i}$$

has approximately a χ^2 distribution with $t-1$ degrees of freedom. For the approximation to be adequate, the t classes should be defined so that $np_i \geq 5$ for all i.

b. Let k_1, k_2, ..., k_t be the observed frequencies for the outcomes r_1, r_2, ..., r_t, respectively, and let np_{1o}, np_{2o}, ..., np_{to} be the corresponding frequencies based on the null hypothesis. At the α level of significance, H_o: $f_Y(y) = f_o(y)$ (or H_o: $p_X(k) = p_o(k)$ or H_o: $p_1 = p_{1o}$, $p_2 = p_{2o}$, ..., $p_t = p_{to}$) is rejected if

$$d = \sum_{i=1}^{t} \frac{(k_i - np_{io})^2}{np_{io}} \geq \chi^2_{1-\alpha, t-1}$$

where $np_{io} \geq 5$ for all i.

Theorem 10.4.1

Theorem 10.4.1 Suppose that a random sample of n observations is taken from $f_Y(y)$ [or $p_X(k)$], a pdf having s unknown parameters. Let r_1, r_2, ..., r_t be a set of mutually exclusive ranges (or outcomes) associated with each of the n observations. Let $\hat{p}$ i = estimated probability of ri, i = 1, 2, ... t (as calculated from $f_Y(y)$ [or $p_X(k)$] after the pdf's s unknown parameters have been

replaced by their maximum likelihood estimates). Let X_i denote the number of times r_i occurs, $i = 1, 2, ..., t$. Then

a. the random variable

$$D_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

has approximately a χ^2 distribution with t-1-s degrees of freedom. For the approximation to be fully adequate, the r_i's should be defined so that $n\hat{p}_i \geq 5$ for all i.

b. to test H_o: $f_Y(y) = f_o(y)$ [or H_o: $p_X(k) = p_o(k)$] at the α level of significance, calculate

$$d_1 = \sum_{i=1}^{t} \frac{(k_i - n\hat{p}_{io})^2}{n\hat{p}_{io}} \geq \chi^2_{1-\alpha, t-1}$$

where $k_1, k_2, ..., k_t$ are the observed frequencies of $r_1, r_2, ..., r_t$, respectively, and $n\hat{p}_{1o}, n\hat{p}_{2o}, ..., n\hat{p}_{to}$ are the corresponding expected frequencies based on the null hypothesis. If

$$d_1 \geq \chi^2_{1-\alpha, t-1-s}$$

H_o should be rejected (The r_i's should be defined so that $n\hat{p}_{io} \geq 5$ for all i)

Theorem 10.5.1

Theorem 10.5.1 Suppose that n observations are taken on a sample space partitioned by the events $A_1, A_2, ..., A_r$ and also by events $B_1, B_2, ..., B_c$. Let $p_i = P(A_i)q_j = P(B_j)$, and $P_{ij} = P(A_i \cap B_j)$, $i = 1, 2, ..., c$. Let X_{ij} denote the number of observations belonging to the intersection $A_i \cap B_j$. Then

a. the random variable

$$D_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - np_{ij})^2}{np_{ij}}$$

has approximately a χ^2 distribution with rc-1 degrees of freedom (provided $np_{ij} \geq 5$ for all i and j)

b. to test H_o: the A_is are independent of the B_js calculate the test statistic

$$d_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(k_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}$$

where k_{ij} is the number of observations in the sample that belong to $A_i \cap B_j$, $i = 1, 2, ..., r$; $j = 1, 2, ..., c$ and $\hat{p}_i$ and $\hat{q}_j$ are the maximum likelihood estimates for pi and qi respectively. The null hypothesis should be rejected at the α level of significance if

$$d_2 \geq \chi^2_{1-\alpha, (r-1)(c-1)}$$

(Analogous to the condition stipulated for all other goodness-of-fit tests, it will be assumed that $n\hat{p}_i\hat{q}_i \geq 5$ for all *i* and *j*.)

Comment The χ^2 distribution with (r-1)(c-1) degrees of freedom provides an adequate approximation to the distribution d_2 only if $n\hat{p}_i\hat{q}_i \geq 5$ for all *i* and *j*. If one or more cells in a contingency table have estimated expected frequencies that are substantially less than five, the table should be collapsed and the rows and/or columns redefined.

Gallagher comment Larsen & Marx misstate the rule here. They are being too conservative. For r x c contingency tables (i.e., *df* > 1), all cells should have expected values>1, and no more than 20% of the cells should have expected values < 5 (Dallal calls this Cochran's rule, **http://www.jerrydallal.com/LHSP/ctab.htm**). Even these more lenient restrictions may be too conservative according to Zar.

Statistical Tests

Fisher's hypergeometric test

Salsburg (2001) describes the actual story behind Fisher's exact hypergeometric test in his book, "The Lady Tasting Tea." The title is taken from Fisher's invention of this test. At a Cambridge party, a lady claimed that she could tell whether the tea or the milk was added to the cup first. So Fisher designed an experiment to test her claim and invented a statistic to test the null hypothesis that her selection was due to mere chance. He lined up 8 cups of tea and added milk first to 4 of the cups and tea first to 4 other cups and asked the lady to taste the cups and to classify them as milk first or tea first. She had to classify the 8 cups of tea into two groups of 4. He described the experiment in the second chapter of his 1935 book 'The Design of Experiments,' and produced the following table:

Table 1. Fisher's Tea-Tasting Experiment, published data

	Guess Poured First		
Poured First	**Milk**	**Tea**	**Total**
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Now using the hypergeometric probability distribution, Fisher could calculate the probability of observing that exact result if the null hypothesis was true. He calculated the probability for a single cell and since the row and marginal totals are fixed in such a design, it didn't matter which cell was chosen to calculate the probability of observing that exact result under the null hypothesis. For cell n_{11}, the upper left cell, the probability of observing a 3 could be determined

from the row and column totals as shown in the equation below. In that equation, n_{1+} denotes the marginal total for the 1st row and n_{+1} denotes the marginal total for the 1st column. The exact probability of observing a 3 in the upper left cell of the 2 x 2 contingency table is 16/70 or approximately 0.229. Now, in keeping with Fisher's own theory of hypothesis testing the *p value* for the experiment is the probability of what was observed if the null hypothesis was true plus the probability of observing more extreme observations relative to the null hypothesis. For Fisher's tea-tasting experiment, the only more extreme event for the null hypothesis would be to observe a 4 in the upper cell. The probability of observing a 4 is exactly 1/70 or approximately 0.014. The *p value* for the cell shown in Table 1 is the sum of these two probabilities or 0.243, which provides no basis for rejecting the null hypothesis at an α level of 0.05.d

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}$$

$$P(3) = \frac{\binom{4}{3}\binom{4}{4-3}}{\binom{8}{4}} = \frac{16}{70} \approx 0.229.$$

$$P(4) = \frac{\binom{4}{4}\binom{4}{4-4}}{\binom{8}{4}} = \frac{1}{70} \approx 0.014.$$

P = P(3) + P(4) ≈ 0.229 + 0.014 = 0.243

Salsburg (2001) was able to track down one of the participants in the Cambridge tea-tasting experiment and was told that the lady correctly classified every cup. So, these result would produce the results shown in Table 2. The exact *p value* is 1/70.

Table 2. Fisher's Tea-Tasting Experiment, actual results

	Guess Poured First		
Poured First	**Milk**	**Tea**	**Total**
Milk	4	0	4
Tea	0	4	4
Total	4	4	8

Fisher's test as an exact test for tests of homogeneity and independence

I've spent more than the usual amount of time describing Fisher's exact hypergeometric test (note he also introduced what is now called Fisher's exact sign test), because this test will provide an exact test for tests of independence and homogeneity for any 2 x 2 table. The test was originally defined for those rare experiments where the marginal totals and total sample size are

fixed, but the p values produced are appropriate even for 2 x 2 contingency tables where the marginal totals were not fixed in advance.

Fisher's test and Matlab

The Mathworks have not published Fisher's exact test for either 2 x 2 designs or r x c designs even though both tests exist. On the Mathworks user-contributed files is a nice program called fishertest.m, copyrighted and posted by Jos van der Geest, which performs Fisher's exact test for 2 x 2 data.

Fisher's exact test is the appropriate test for these data. For comparative purposes only, one can perform a χ^2 test on these data too, using Holsberg's contincy.m:

```
D=[3 1;1 3]
[H,P,STATS] = fishertest(D)
[p,cs,expected] = contincy(D, 'chi2')
```

These three statements produce the result that the *p value* for Fisher's exact test is 0.2429. Under the null hypothesis of independence choice is independent of classification, the χ^2 statistic with 1 *df* is 2, with a *p value* of 0.1573.

What is a test of homogeneity?

Chapter 10 in the text focuses on tests of independence, which is a hypothesis that tests for one population whether the row category is independent of the column category. Contingency tables can also be used as test for homogeneity. If each row is regarded as a separate population one can test whether the proportions in the first row π_1 equal the proportions in the second row π_2. The null hypothesis being tested is H_o: $\pi_1 = \pi_2$. If the sample sizes are large enough, the two-sample binomial test, introduced in Chapter 9 Theorem 9.4.1, is appropriate for the hypothesis test of homogeneity and reporting the effect size (the confidence limits for the difference are appropriate). A two-sample binomial test should never be used as a test of independence.

Now, confusion arises because Fisher's exact test and the chi-square tests are appropriate tests for homogeneity and independence. In the literature and in some statistics textbooks, the χ^2 test is often, and appropriately, used as a test of homogeneity. **Ramsey & Schafer (2002, p 559)** state that the chi square test for homogeneity is also a test for independence, but that confuses the matter because before the test is performed, the investigator should clearly state whether the null is for homogeneity or independence. The *p value* for the chi-square test of homogeneity will be identical to the Z statistic for the two-sample binomial test. Indeed, for a 2x2 table, the chi-square statistic will be the square of the z statistic. In reporting the results of these tests, be specific about which hypothesis is being tested even though the p values might be the same.

Sampling schemes producing contingency tables

There are 6 different sampling schemes that can produce data for a contingency table. These schemes are based on the goals of the sampling and whether the marginal totals of the contingency table are fixed in advance. The sampling schemes are:

1. **Poisson sampling**
   a. A fixed amount of time (effort money) devoted to getting a random sample from a single population
   b. None of the marginal totals is known in advance of the sampling
2. **Multinomial sampling** Similar to Poisson but grand total fixed in advance (e.g., I'll collect 100 individuals and determine the number of males)
3. **Binomial sampling**
   a. **Prospective product binomial sampling**: Two populations, size of samples fixed by the researcher
   b. **Retrospective product binomial sampling**
      i. Random samples from populations defined by the response variable (e.g., lung cancer or not)
      ii. Like prospective, but with classification based on response instead of explanatory variables
4. **Randomized binomial experiment**: Treatments assigned randomly to subjects or experimental units
5. **The hypergeometric probability distribution**
   a. Both row and columns fixed
   b. Example: Fisher's exact test & Lady drinking tea

Figure 2 from **Ramsey & Schafer (2002)** show that tests of homogeneity might be performed on data collected using all 6 sampling schemes. But, in designing a hypothesis test, it should be made clear from the outset whether a test of independence or homogeneity will be performed.

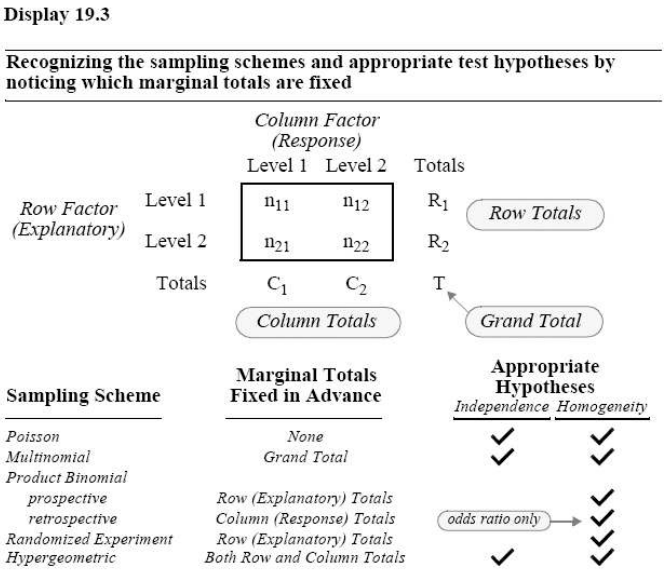

Figure 2. Six sampling schemes and whether tests of homogeneity or independence are appropriate.

In deciding whether it is a test of homogeneity or independence, think '1 population or two?,' whether the row (or column) totals are fixed, and whether the effect is to be reported as a difference in proportions. As shown in Figure 2, if the row or column totals are fixed in advance as part of the sampling scheme, the test for association between rows and columns is a test of homogeneity, not independence. The very fact that two separate populations can be identified to fix the marginal totals indicates that a test of homogeneity is involved. The converse is not true. The Poisson, multinomial and hypergeometric sampling schemes can be used for tests of homogeneity or independence.

The test used can not be used to distinguish between the homogeneity or independence hypotheses since both the Fisher's test and chi-square test can be used to test either either the homogeneity or independence hypothesis. For a test of homogeneity, the two-sample binomial test is often the appropriate test mainly because it produces a convenient estimate of the effect size and produces a 95% confidence interval for the difference in proportions. In terms of p values, the chi-square and two-sample binomial produce identical p values with the chi-square statistic being the square of the two-sample binomial's z statistic.

Case Studies and Examples

Case Study 10.3.1

Figure 3. The unhorned and horned morphs of *Ceriodaphnia cornuta*. Note that the star of Texas eyeball in Larsen & Marx (2006) was not in the original publication.

The late Tom Zaret studied zooplankton in Gatun Lake Panama, especially the role of invertebrate and vertebrate predators. He did an experiment with the horned and unhorned morphs of *Ceriodaphnia cornuta*, a small cladoceran (a shrimp-like organism), shown in Figure 3. The horned morph had evolved to be resistant to invertebrate predation, but **Zaret (1972)** wanted to know the relative vulnerabilities of the unhorned and horned morphsto fish predators. He added the prey at their natural densities. In one of the many experiments (27 April), Zaret added about 100 cladocerans in a 3:1 ratio of unhorned to horned and then added a pair of fish (*Melaniris chagresi*). After about one hour, he removed the fish, sacrificed them, and looked at the cladocera in their guts. There were 40 horned to four unhorned cladocerans in the fishes' guts. Does this constitute evidence for selectivity by the fish?

We can perform a chi-square test based on Theorem 10.3.1 to find a χ^2 statistic of 5.9394. Since there were only two categories, this would be compared to a χ^2 distribution with 1 df (the number of categories minus 1). Figure 1 shows the χ^2_1 distribution. The p value for a χ^2 statistic of 5.9394 is 0.0148. Note that this is a two-tailed *p value*.

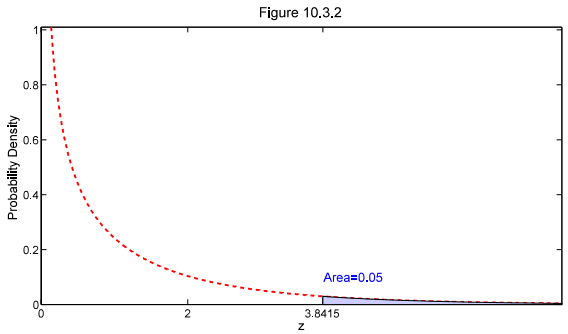

Figure 4. The χ^2_1 distribution

With these data it would have been possible to perform an exact binomial test. What is the probability of observing 40 out of 44 items in the gut being the unhorned morph if the expected proportion is 0.75? In Matlab, this is a one line problem with the solution that the 2-tailed p value (twice the 1-tailed p) is 0.0146, nearly identical to the result from the χ^2 test:

```
P=2*sum(binopdf([40:44],44,3/4));
```

Larsen & Marx comment that the fish selected the unhorned morph because of its eyespot. Another interesting possibility suggests itself based on the 3:1 ratio of unhorned to horned morphs offered to the fish. Some predators are apostatic generalists, selectively feeding on the animals that are most abundant. **Zaret (1972)** only did one experiment (10 May) in which the unhorned and horned morph were added in nearly equal proportions, and, of the 15 items eaten in that experiment, 8 were unhorned and 7 were horned. I don't need a computer to calculate the 2-tailed *p value* for that experiment. It is 1.0 (Why?)

Case Study 10.3.2

Benford's law predicts that the first nonzero digits in numbers measuring things in nature are not random but follow a law:

$pi=\log_{10}(i+1) - \log_{10}(i), \quad i=1,2, ..., 9$

The law should apply whenever the items being measured span more than one order of magnitude. Auditors can use Benford's law to find examples of fraud. Table 10.3.2 in the text provides the first 355 digits appearing in the 1997-1998 operating budget for the University of West Florida. A chi square test was used to fit the data in a very simple Matlab program:

```
O=[111 60 46 29 26 22 21 20 20]';
BP=zeros(9,1);
for i=1:9
   BP(i)=log10(i+1)-log10(i);
end
N=sum(O);
E=BP*N;
Table100302=[[1:9]' O E (O-E).^2./E]
CHI2=sum((O-E).^2./E)
df=length(E)-1;
alpha=0.05
pvalue=1-chi2cdf(CHI2,df)
criticalval=chi2inv(1-alpha,df)
if CHI2>=criticalval
   disp('Reject Ho')
else
   disp('Fail to Reject Ho')
end
```

This program returns:

```
Table100302 =
   1.0000  111.0000  106.8656    0.1599
   2.0000   60.0000   62.5124    0.1010
```

```
    3.0000   46.0000   44.3533    0.0611
    4.0000   29.0000   34.4031    0.8486
    5.0000   26.0000   28.1093    0.1583
    6.0000   22.0000   23.7661    0.1312
    7.0000   21.0000   20.5871    0.0083
    8.0000   20.0000   18.1591    0.1866
    9.0000   20.0000   16.2439    0.8685
CHI2 =
    2.5236
alpha =
    0.0500
pvalue =
    0.9606
criticalval =
   15.5073
Fail to Reject Ho
```

Case Study 10.4.1

Data were collected from all 'at bats' from opening day to mid-July 1996. Players had exactly 4 at bats 4,096 times during that period with the observed frequency of hits shown in Figure 5. Are these numbers consistent with the hypothesis that the number of hits a player gets in four at-bats is binomially distributed?

Number of Hits, i		Obs. Freq., k_i	Estimated Exp. Freq., np_i
r_is	0	1280	1289.1
	1	1717	1728.0
	2	915	868.6
	3	167	194.0
	4	17	16.3

Here the five possible outcomes associated with each four-at-bat game would be the number of hits a player makes, so $r_i = 0$, $r_1 = 1$ ... $r_5 = 4$. The presumption to be tested is that the probabilities of those r_is are given by the binomial distribution – that is:

$$P(\text{player gets I hits in 4 at-bats}) = \binom{4}{i} p^i (1-p)^{4-i}, \quad i = 0, 1, 2, 3, 4$$

Figure 5. Frequency of hits during 4096 player-games in 1996.

The Matlab program is shown below:

```
obs=[1280 1717 915 167
17];hits=[0:4]';Total=obs*hits;
estP=Total/(4096*4)
expc=binopdf(hits,4,estP)'*4096
d1=sum((obs-expc).^2./expc)
df=length(expc)-1-1;
P = 1-chi2cdf(d1,df);
fprintf(...
  'The chi-square statistic,%6.1f, with %2.0f d.f. has pvalue=%6.4f\n',...
  d1,df,P)
alpha=0.05;
criticalval=chi2inv(1-alpha,df);
fprintf('The critical value for alpha=%4.2f is %5.3f\n',alpha,criticalval)
if d1 >=      criticalval
   disp('Reject null')
else
   disp('Fail to reject null')
```

```
end
```
The program produces:

```
The chi-square statistic,   6.4, with  3 d.f. has pvalue=0.0936
The critical value for alpha=0.05 is 7.815
Fail to reject null
```

As noted by **Larsen & Marx (2006, p. 618)**, many of the assumptions of the binomial model are violated in a real baseball game and season.

Case Study 10.4.2

Do the deaths of women over the age of 80 in London follow a Poisson pdf? The following Matlab program solves this problem with the data provided in Larsen & Marx:

```
Deaths=[0:10]';
obs=[162 267 271 185 111 61 27 8 3 1 0]'; n=sum(obs);

lambda=sum(Deaths.*obs)/n
expc=poisspdf(Deaths,lambda)'*n;
% Call Matlab's goodness-of-fit test (see help file for Poisson fit)
[h,p,st] = chi2gof(Deaths,'ctrs',Deaths,'frequency',obs, ...
        'expected',expc,'nparams',1)
```

In the Matlab code below, I also provide the more explicit statements for fitting the Poisson distribution data.

Figure 6 shows the observed and expected frequencies of deaths of 80+ year old London women based on a Poisson model with an estimated Poisson λ=2.1569. The fit doesn't look unreasonable, but that just illustrates the risk of performing a 'chi-by-eye' test. If the Poisson model is true, the probability of observing results like those observed or more extreme is 0.0002. The observed chi-square statistic is 25.9 with 6 df. At an α level of 0.05 the chi-square critical value for 6 df is 12.592. As noted by the authors, there appears to more days with 0 deaths and more days with 5 or more deaths that account for the departures from Poisson expectation.

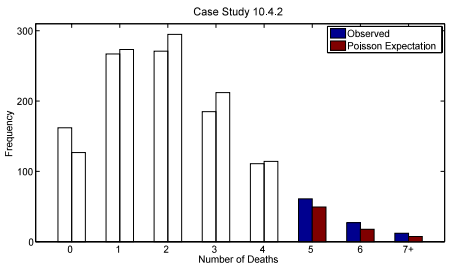

Figure 6. Observed and expected frequencies under a Poission model of deaths per day of 80+ year old women in London.

Case Study 10.4.3

Monthly percent returns for IBM stock for a 7-year period are provided. **Larsen & Marx (2006)** carry out a chi-square goodness of fit test comparing these data to a normal model. They bin the data into 5 categories and find the χ^2 statistic to be 7.52 with just 2 df, which is greater than the 5.991 critical value .

The full program is provided here:

```
DATA=[7.2 6.2 3.0 2.7 2.8 -2.7 2.4
6.8 3.6 9.2 -1.2 -0.2 -6.4 -0.8
-0.8 -14.8 -13.4 -13.6 14.1 2.6 -10.8
-2.1 15.4 -2.1 8.6 -5.4 5.3 10.2
3.1 -8.6 -0.4 2.6 1.6 9.3 -1.5
4.5 6.9 5.2 4.4 -4.0 5.5 -0.6
-3.1 -4.4 -0.9 -3.8 -1.5 -0.7 9.5
2.0 -0.3 6.8 -1.1 -4.2 4.6 4.5
2.7 4.0 -1.2 -5.0 -0.6 4.1 0.2
8.0 -2.2 -3.0 -2.8 -5.6 -0.9 4.6
13.6 -1.2 7.5 7.9 4.9 10.1 -3.5
6.7 2.1 -1.4 9.7 8.2 3.3 2.4];
DATA=DATA(:);[MUHAT,SIGMAHAT,MUCI,SIGMACI] = normfit(DATA,0.05)
[h,p,stats] = chi2gof(DATA)
df=length(stats.E)-1-2
histfit(DATA);figure(gcf)
criticalval=chi2inv(1-alpha,df);
fprintf('The critical value for alpha=%4.2f is %5.3f\n',alpha,criticalval)
```

It is more than a little disconcerting that Matlab's chi2gof finds that the normal distribution provides an adequate fit to the IBM stock data. Here is the full program, which consists of entering the data, fitting the data to the normal distribution, performing the chi-square test and plotting the data, as shown in Figure 7.

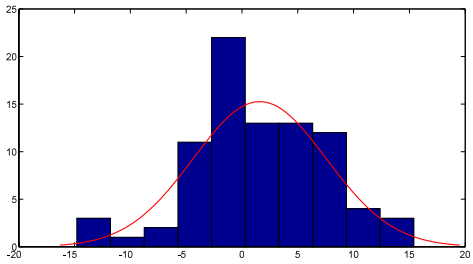

Figure 7. Histogram of 84 monthly percent returns. Matlab's χ^2 goodness-of-fit test has a p-value of 0.1996.

Matlab's χ^2 statistic is 5.99 based on 7 categories, and thus 4 df. It has a p value of 0.1996, which provides little evidence for rejecting the null hypothesis. The α=5% critical value with 4 df is 9.488. I tried a few other schemes at binning data, for example into larger groupings, only to produce even larger *p values*.

Case Study 10.5.1

Matlab doesn't have a built-in contingency table analysis program, but there is a nice program available from Anders Holsberg's stixbox toolbox called contincy.m I'll use it to solve most of the contingency analysis problems. In the first problem, the contingency table tests whether

price is independent of longevity. The data are shown in Figure 8.

The Matlab program is just two lines:

```
observed=[70 65;39 28;14 3;13 2];
% From contincy from stixbox toolbox
[p,cs,expected] = contincy(observed 'chi2')
```

which produces:

```
p =     0.0102
cs = 11.3026
expected =
   78.4615   56.5385
   38.9402   28.0598
    9.8803    7.1197
    8.7179    6.2821
criticalval =
    7.8147
Reject Ho
```

Table 10.5.6

		Still Being Published	Has Ceased Publication	Total
Cover Price	< $1.99	70 (78.5)	65 (56.5)	135
	$2.00 - $2.49	39 (38.9)	28 (28.1)	67
	$2.50 - $2.99	14 (9.9)	3 (7.1)	17
	$3.00+	13 (8.7)	2 (6.3)	15
	Total	136	98	234

Figure 8. Price versus longevity contingency table.

Case Study 10.5.2 Siskel & Ebert

Were the movie critics Siskel & Ebert rendering judgements that were independent of each other? Their reviews are shown in the following table.

Table 10.5.8

		Ebert Ratings: Down	Sideways	Up	Total
	Down	24 (11.8)	8 (8.4)	13 (24.8)	45
Siskel Ratings	Sideways	8 (8.4)	13 (6.0)	11 (17.6)	32
	Up	10 (21.8)	9 (15.6)	64 (45.6)	83
	Total	42	30	88	160

But

$$d_2 = \frac{(24-11.8)^2}{11.8} + \frac{(8-8.4)^2}{8.4} + \ldots + \frac{(64-45.6)^2}{45.6}$$

$$= 45.37$$

So the evidence is overwhelming that Siskel and Ebert's judgments are not independent.

Figure 9. Siskel & Ebert movie reviews.

The Matlab program to analyze these data is very brief:

```
observed=[24 8 13; 8 13 11; 10 9 64]
[p,cs,expected] = contincy(observed, 'chi2')
[r,c]=size(observed);df=(r-1)*(c-1);
criticalval=chi2inv(.95,df)
if chi2>criticalval
   disp('Reject Ho')
else
   disp('Fail to reject Ho')
end
```

which produces:

```
observed =
    24     8    13
```

```
    8    13    11
   10     9    64
p =
  3.3515e-009
cs =
   45.3569
expected =
   11.8125    8.4375   24.7500
    8.4000    6.0000   17.6000
   21.7875   15.5625   45.6500
criticalval =
    9.4877
Reject Ho
```

Case Study 10.5.3

Data on mobility and suicide rate for 25 American cities provided in Figure 10. The data were regrouped into above and below mean as shown in Figure 11.

Table 10.5.9

City	Suicides per 100,000, x_i	Mobility Index, y_i	City	Suicides per 100,000, x_i	Mobility Index, y_i
New York	19.3	54.3	Washington	22.5	37.1
Chicago	17.0	51.5	Minneapolis	23.8	56.3
Philadelphia	17.5	64.6	New Orleans	17.2	82.9
Detroit	16.5	42.5	Cincinnati	23.9	62.2
Los Angeles	23.8	20.3	Newark	21.4	51.9
Cleveland	20.1	52.2	Kansas City	24.5	49.4
St. Louis	24.8	62.4	Seattle	31.7	30.7
Baltimore	18.0	72.0	Indianapolis	21.0	66.1
Boston	14.8	59.4	Rochester	17.2	68.0
Pittsburg	14.9	70.0	Jersey City	10.1	56.5
San Francisco	40.0	43.8	Louisville	16.6	78.7
Milwaukee	19.3	66.2	Portland	29.3	33.2
Buffalo	13.8	67.6			

Figure 10. Mobility index & suicide

Case Study 10.5.3 is a 2 x 2 analysis and can be analyzed using a χ^2 test of independence. It can also be analyzed using Fisher's hypergeometric test. Here is the call to Matlab's contincy.m and a user-contributed m.file to perform Fisher's exact test:

```
observed=[7 4;3 11]
[p,chi2,expected] = contincy(observed, 'chi2')
[r,c]=size(observed);df=(r-1)*(c-1);
criticalval=chi2inv(.95,df)
if chi2>criticalval
   disp('Reject Ho')
else
   disp('Fail to reject Ho')
end
[H,P, STATS] = fishertest(observed, 0.05)
```

The programs produce:

```
observed =

    7     4
    3    11
p =
   0.0325
```

```
chi2 =
    4.5725
expected =
    4.4000    6.6000
    5.6000    8.4000
criticalval =
    3.8415
Reject Ho
P =
    0.0416
```

Fisher's exact hypergeometric test produces a p value of 0.0416, slightly higher but preferable to the approximate p value produced by the χ^2 test. Both tests lead to the same decision: reject H_o that suicide is independent of mobility.

Annotated outline (with Matlab scripts) for Larsen & Marx Chapter 10

TABLE 10.5.10

		Mobility Index	
		Low (<56.0)	High (>56.0)
Suicide Rate	High (>20.8)	7	4
	Low (<20.8)	3	11

10 **GOODNESS OF FIT TESTS** (**Week 10**)

Karl Pearson (1857-1936)

Figure 12. Karl Pearson

10.1 **INTRODUCTION**

10.1.1 "Any procedure that seeks to determine whether a set of data could reasonably have originated from some given probability distribution, or class of probability distribution is called a **goodness-of-fit** test.

10.2 **THE MULTINOMIAL DISTRIBUTION**

10.2.1 Goodness of fit statistics based on the chi-square statistic which is based on the multinomial which is an extension of the binomial distribution

Theorem 10.2.1. Let X_i denote the number of times that the outcome r_i occurs, $i = 1, 2, ..., t$, in a series of n independent trials, where $p_i = p(r_i)$. Then the vector $(X_1, X_2, ..., X_t)$ has a multinomial distribution and

$$p_{x_1, x_2, ..., x_t} = P(X_1 = k_1, X_2 = k_2, ..., X_t = k_t)$$

$$\frac{n!}{k_1!\, k_2! \cdots k_t!}\, p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t}$$

$$k_i = 0, 1, 2, ..., n; \quad i = 1, 2, ..., t; \quad \sum_{i=1}^{t} k_i = n$$

Example 10.2.1

```
%LMex100201_4th.m
% Application of Theorem 10.2.1, Larsen & Marx (2006) Introduction to
% Mathematical Statistics, 4th Edition. Page 601
% An application of the multinomial distribution
% Written by Eugene.Gallagher@umb.edu 11/21/10
P=exp(gammaln(13)-6*gammaln(3))...
   *(1/21)^2*(2/21)^2*(3/21)^2*(4/21)^2*(5/21)^2*(6/21)^2
```

Example 10.2.2

```
% LMex100202_4th.m
% Example 10.2.2
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu 11/21/10
syms y;
int(sym('6*y*(1-y)'),0,0.25)
p1=eval(int(sym('6*y*(1-y)'),0,0.25))
int(sym('6*y*(1-y)'),0.5,0.75)
p3=eval(int(sym('6*y*(1-y)'),0.5,0.75))
int(sym('6*y*(1-y)'),0.75,1)
p4=eval(int(sym('6*y*(1-y)'),0.75,1))
P=exp(gammaln(6)-gammaln(4))*p1*p3^3*p4
```

10.2.2 A Multinomial Binomial Relationship

Theorem 10.2.2 Suppose the vector (X_1, X_2, ..., X_t) is a multinomial random variable with parameters *n*, p_1, p_2, ..., p_t. Then the marginal distribution of X_i, *i* = 1, 2, ..., *t*, is the binomial pdf with parameters n and p*i*.

Comment Theorem 10.2.2 gives the pdf for any given Xi in a multinomial vector. Since that pdf is the binomial, we also know the mean and variance of each Xi — specifically, $E(X_i) = np_i$ and Var $(X_i) = n\ p_i(1-p_i)$. [Gallagher note: this will apply to ternary diagrams, see Jumars on feeding guilds and Matlab environmental statistics book]

Example 10.2.3

```
% LMex100203_4th.m
% Number of students and variance
% Written by Eugene.Gallagher@umb.edu 11/21/10
n=50
p1 = 1-normcdf(90,80,5)
Ex1=n*p1
Varx1=n*p1*(1-p1)
p2 = normcdf(90,80,5)-normcdf(80,80,5)
Ex2=n*p2
Varx2=n*p2*(1-p2)
p3 = normcdf(80,80,5)-normcdf(70,80,5)
Ex3=n*p3
Varx3=n*p3*(1-p3)
p4 = normcdf(70,80,5)-normcdf(60,80,5)
Ex4=n*p4
```

```
Varx4=n*p4*(1-p4)
p5 = normcdf(60,80,5)
Ex5=n*p5
Varx5=n*p5*(1-p5)
```

Questions

10.2.3 AP tests
10.2.4 Mendel
10.2.5 Hypertension
10.2.6 IQ
10.2.7 Pipeline missle
10.2.8 Baseball
10.2.9
10.2.10
10.2.11

10.3 GOODNESS OF FIT TESTS: ALL PARAMETERS KNOWN

10.3.1 First proposed by Karl Pearson in 1900.

Theorem 10.3.1 Let $r_1, r_2, ..., r_t$ be the set of possible outcomes (or ranges of outcomes) associated with each of n independent trials, where $P(r_i) = p_i$, $i = 1, 2, ..., t$. Let X_i = number of times r_i occurs, $i = 1, 2, ..., t$. Then

a. The random variable

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i}$$

has approximately a χ^2 distribution with $t-1$ degrees of freedom. For the approximation to be adequate, the t classes should be defined so that $np_i \geq 5$ for all i.

b. Let $k_1, k_2, ..., k_t$ be the observed frequencies for the outcomes $r_1, r_2, ..., r_t$, respectively, and let $np_{1o}, np_{2o}, ..., np_{to}$ be the corresponding frequencies based on the null hypothesis. At the α level of significance, $H_o: f_Y(y) = f_o(y)$ (or $H_o: p_X(k) = p_o(k)$ or $H_o: p_1 = p_{1o}, p_2 = p_{2o}, ..., p_t = p_{to}$) is rejected if

$$d = \sum_{i=1}^{t} \frac{(k_i - np_{io})^2}{np_{io}} \geq \chi^2_{1-\alpha, t-1}$$

where $np_{io} \geq 5$ for all i.

Case Study 10.3.1: Horned and unhorned *Ceriodaphnia cornuta*

```
% LMcs100301_4th.m
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu
% Predation on Ceriodaphnia cornuta by Melaniris chagresi
observed=[40 4]
expected=44*[3/4 1/4]
d=sum((observed-expected).^2./expected);
```

```
% d=(40-44*3/4)^2/(44*3/4)+(4-44*1/4)^2/(44*1/4);
fprintf('Chi-square statistic = % 6.4f\n',d);
P = 1-chi2cdf(d,1);
fprintf('Chi-squared 2-tailed p-value=%6.4f\n',P);
% the data would be better analyzed by an exact one-sample binomial test
data=40:44;
% 1 tailed exact binomial test
P=sum(binopdf(data,44,3/4));
fprintf('Exact binomial 2-tailed p value=%6.4f from the binomial pdf\n',P*2);
% or solve using the cumulative binomial pdf
P=1-binocdf(39,44,3/4);
fprintf('Exact binomial 1-tailed p value=%6.4f from the binomial cdf\n',P*2);
% Plot the figure and the critical value
% Use Figure 7.5.1 as a model
df=1;chi05=chi2inv(0.05,df);
alpha=0.05;chi95=chi2inv(1-alpha,df);
zmax=7.1; ymax=1.01;
z=0:.01:zmax;
fzz=chi2pdf(z,df);
plot(z,fzz,'linestyle','--','color','r','linewidth',3)
ylabel('Probability Density','FontSize',20)
xlabel('z','FontSize',20)
axis([0 zmax 0 ymax])
set(gca,'Ytick',[0:.2:ymax],'FontSize',18)
set(gca,'Xtick',[0:2:chi95 chi95],'FontSize',18)
ax=axis;
ax1=gca; % save the handle of the graph
title('Figure 10.3.2','FontSize',22)
hold on
fz=chi2pdf(chi95,df);
plot([chi95 chi95]',[0 fz]','-k','linewidth',1)
% Fill in the upper tail with fill
y2=chi95:.001:ax(2);
fy2=chi2pdf(y2,df);
fymax=chi2pdf(ax(2),df);
fill([chi95 y2 ax(2) ax(2)],[0 fy2 fymax 0],[.8 .8 1])
t=sprintf('Area=0.05');
text(chi95+.01,.1,t,'Color','b','FontSize',20);
figure(gcf)
hold off
```

Case Study 10.3.2

```
%LMcs100302_4th.m
% Benford's law p 609-611 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu 3/15/11
```

```
O=[111 60 46 29 26 22 21 20 20]';
BP=zeros(9,1);
for i=1:9
   BP(i)=log10(i+1)-log10(i);
end
% or the same (thanks Wikipedia for equation)
i=[1:9]';BP=log10(1+1./i)
N=sum(O);
E=BP*N;
Table100302=[[1:9]' O E (O-E).^2./E]
CHI2=sum((O-E).^2./E)
df=length(E)-1;
alpha=0.05
pvalue=1-chi2cdf(CHI2,df)
criticalval=chi2inv(1-alpha,df)
if CHI2>=criticalval
   disp('Reject Ho')
else
   disp('Fail to Reject Ho')
end
```

Example 10.3.1

A somewhat complicated problem, a nice application of the symbolic math toolbox

```
%LMex100301_4th.m
syms f y
p=zeros(5,1);
f=int(6*y*(1-y),0,0.2)
p(1)=eval(f);
f=int(6*y*(1-y),.2,0.4)
p(2)=eval(f);
f=int(6*y*(1-y),.4,0.6)
p(3)=eval(f);
f=int(6*y*(1-y),.6,0.8)
p(4)=eval(f);
f=int(6*y*(1-y),.8,1)
p(5)=eval(f)
DATA=[0.18 0.06 0.27 0.58 0.98
0.55 0.24 0.58 0.97 0.36
0.48 0.11 0.59 0.15 0.53
0.29 0.46 0.21 0.39 0.89
0.34 0.09 0.64 0.52 0.64
0.71 0.56 0.48 0.44 0.40
0.80 0.83 0.02 0.10 0.51
0.43 0.14 0.74 0.75 0.22];
DATA=DATA(:);
edges=0:0.2:1;
```

```
[N,Bin]=histc(DATA,edges);
O=[sum(N(1:2));N(3);sum(N(4:end))];
n=length(DATA);
E=p*n;
E=[sum(E(1:2));E(3);sum(E(4:end))];
chi2=sum((O-E).^2./E);
df=length(E)-1;
fprintf('The chi-square statistic with %1.0f df is %5.2f\n',df,chi2)
criticalval=chi2inv(0.95,df);
if chi2 >=      criticalval
   disp('Reject null')
else
   disp('Fail to reject null')
end
df=length(E)-1;
pvalue=1-chi2cdf(chi2,df)
```

Questions

10.4 GOODNESS OF FIT TESTS: PARAMETERS UNKNOWN

10.4.1 Many times the parameters of a probability distribution must be estimated from the data

10.4.2 One df lost for every parameter estimated, reducing the power of the test

Theorem 10.4.1 Suppose that a random sample of n observations is taken from $f_Y(y)$ [or $p_X(k)$], a pdf having *s* unknown parameters. Let $r_1, r_2, ..., r_t$ be a set of mutually exclusive ranges (or outcomes) associated with each of the *n* observations. Let $\hat{p}$ i = estimated probability of ri, i = 1, 2, ... t (as calculated from $f_Y(y)$ [or $p_X(k)$] after the pdf's s unknown parameters have been replaced by their maximum likelihood estimates). Let X_i denote the number of times r_i occurs, $i = 1, 2, ..., t$. Then

a. the random variable

$$D_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

has approximately a χ^2 distribution with $t-1-s$ degrees of freedom. For the approximation to be fully adequate, the r_i's should be defined so that $n\hat{p}_i \geq 5$ for all *i*.

b. to test H_o: $f_Y(y) = f_o(y)$ [or H_o: $p_X(k) = p_o(k)$] at the α level of significance, calculate

$$d_1 = \sum_{i=1}^{t} \frac{(k_i - n\hat{p}_{io})^2}{n\hat{p}_{io}} \geq \chi^2_{1-\alpha, t-1}$$

where $k_1, k_2, ..., k_t$ are the observed frequencies of $r_1, r_2, ..., r_t$, respectively, and $n\hat{p}_{1o}, n\hat{p}_{2o}, ..., n\hat{p}_{to}$ are the corresponding expected frequencies based on the null hypothesis. If

$$d_1 \geq \chi^2_{1-\alpha, t-1-s}$$

H_o should be rejected (The r_i's should be defined so that $n\hat{p}_{io} \geq 5$ for all i)

Case Study 10.4.1 (p. 541) Batting a binomial

```
% LMcs100401_4th.m
% p 616-617 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
obs=[1280 1717 915 167 17];hits=[0:4]';Total=obs*hits;
estP=Total/(4096*4)
expc=binopdf(hits,4,estP)'*4096
d1=sum((obs-expc).^2./expc)
df=length(expc)-1-1;
P = 1-chi2cdf(d1,df);
fprintf(...
  'The chi-square statistic,%6.1f, with %2.0f d.f. has pvalue=%6.4f\n',...
  d1,df,P)
alpha=0.05;
criticalval=chi2inv(1-alpha,df);
fprintf('The critical value for alpha=%4.2f is %5.3f\n',alpha,criticalval)
if d1 >=      criticalval
   disp('Reject null')
else
   disp('Fail to reject null')
end
```

Case Study 10.4.2 (p. 542) Fumbles in football in 3rd edition

Case Study 10.4.2 Is Death a Poisson process?

```
% LMcs100402_4th.m

% LMcs100402_4th.m
% Pp. 618-621 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu 11/21/10, revised 3/15/11
DATA=[zeros(162,1);ones(267,1);repmat(2,271,1);repmat(3,185,1)
   repmat(4,111,1);repmat(5,61,1);repmat(6,27,1);repmat(7,8,1)
   repmat(8,3,1);repmat(9,1,1)];
% These expanded data needed only for poissfit.m
Deaths=[0:10]';
obs=[162 267 271 185 111 61 27 8 3 1 0]'; n=sum(obs);
% this will fit the Poisson parameter:
lambda=poissfit(DATA)
% Here's a simpler way to calculate the Poisson parameter
lambda=sum(Deaths.*obs)/n
expc=poisspdf(Deaths,lambda)'*n;
% Call Matlab's goodness-of-fit test (see help file for Poisson fit)
[h,p,st] = chi2gof(Deaths,'ctrs',Deaths,'frequency',obs, ...
        'expected',expc,'nparams',1)
% or solve it explicitly
```

```
EXP=expc;
EXP(8)=n-sum(expc(1:7));
EXP(9:end)=[];
OBS=obs;
OBS(8)=n-sum(obs(1:7));
OBS(9:end)=[];
d1=sum((OBS'-EXP).^2./EXP);
df=length(OBS)-1-1;
P = 1-chi2cdf(d1,df);
fprintf(...
  'The chi-square statistic,%6.1f, with %2.0f d.f. has pvalue=%6.4f\n',...
  d1,df,P)
alpha=0.05;
criticalval=chi2inv(1-alpha,df);
fprintf('The critical value for alpha=%4.2f is %5.3f\n',alpha,criticalval)
if d1 >=       criticalval
   disp('Reject null')
else
   disp('Fail to reject null')
end
% Plot the results:
bar(0:7,[st.O;st.E]',1,'grouped')
axis([-.75 7.75 0 310]);
set(gca,'ytick',0:100:300,...
   'xticklabel',{'0 ';'1 ';'2 ';'3 ';'4 ';'5 ';'6';'7+'},'FontSize',20);
legend('Observed','Poisson Expectation','NorthEast','FontSize',20)
xlabel('Number of Deaths','FontSize',20)
ylabel('Frequency','FontSize',20)
title('Case Study 10.4.2','FontSize',22);figure(gcf);
```

```
Case Study 10.4.3 (p. 621)
% LMcs100403_4th.m
% Written by Eugene.Gallagher@umb.edu 11/21/10
DATA=[7.2 6.2 3.0 2.7 2.8 -2.7 2.4
6.8 3.6 9.2 -1.2 -0.2 -6.4 -0.8
-0.8 -14.8 -13.4 -13.6 14.1 2.6 -10.8
-2.1 15.4 -2.1 8.6 -5.4 5.3 10.2
3.1 -8.6 -0.4 2.6 1.6 9.3 -1.5
4.5 6.9 5.2 4.4 -4.0 5.5 -0.6
-3.1 -4.4 -0.9 -3.8 -1.5 -0.7 9.5
2.0 -0.3 6.8 -1.1 -4.2 4.6 4.5
2.7 4.0 -1.2 -5.0 -0.6 4.1 0.2
8.0 -2.2 -3.0 -2.8 -5.6 -0.9 4.6
13.6 -1.2 7.5 7.9 4.9 10.1 -3.5
6.7 2.1 -1.4 9.7 8.2 3.3 2.4];
DATA=DATA(:);[MUHAT,SIGMAHAT,MUCI,SIGMACI] = normfit(DATA,0.05)
```

```
[h,p,stats] = chi2gof(DATA)
df=length(stats.E)-1-2
histfit(DATA);figure(gcf)
criticalval=chi2inv(1-alpha,df);
fprintf('The critical value for alpha=%4.2f is %5.3f\n',alpha,criticalval)
```

Questions (Page 624-627)
4.2.10 Goodness of fit on horse kicks Assigned Carry out the details for a goodness of fit test for the horse kick data of Question 4.2.10 Use the 0.01 level of significance.

10.5 CONTINGENCY TABLES

10.5.1.1 Three types of hypothesis tests
10.5.1.1.1 Tests of parameters of pdfs
10.5.1.1.2 Goodness of fit of pdfs
10.5.1.1.3 Independence of two random variables
10.5.2 Testing for independence: a special case
10.5.2.1.1 **contingency table**

Jerry Dallal has a nice discussion of what a contingency table is:

A **contingency table** is a table of counts. A two-dimensional contingency table is formed by classifying subjects by two variables. One variable determines the row categories; the other variable defines the column categories. The combinations of row and column categories are called *cells*. Examples include classifying subjects by sex (male/female) and smoking status (current/former/never) or by "type of prenatal care" and "whether the birth required a neonatal ICU" (yes/no). For the mathematician, a two-dimensional contingency table with *r* rows and *c* columns is the set $\{x_{ij}: i=1,...,r; j=1,...,c\}$.

1. In order to use the statistical methods usually applied to such tables, subjects must fall into one and only one row and column categories. Such categories are said to be **exclusive** and **exhaustive**. **Exclusive** means the categories don't overlap, so a subject falls into only one category. **Exhaustive** means that the categories include all possibilities, so there's a category for everyone. Often, categories can be made exhaustive by creating a catch-all such as "Other" or by changing the definition of those being studied to include only the available categories.

Also, the observations must be independent. This can be a problem when, for example, families are studied, because members of the same family are more similar than individuals from different families.
Jerry Dallal, **http://www.jerrydallal.com/LHSP/ctab.htm**

Table 10.5.1

Table 10.5.1. A contingency table

		Trait B		Row Totals
		B_1	B_2	
Trait A	A_1	n_{11}	n_{12}	R_1

	A_2	n_{21}	n_{22}	R_2
Column Totals		C_1	C_2	n

Table 10.5.2. A contingency table with expected frequencies

		Trait B		Row Totals
		B_1	B_2	
Trait A	A_1	$nP(A_1)P(B_1)$	$nP(A_1)P(B_2)$	R_1
	A_2	$nP(A_2)P(B_1)$	$nP(A_2)P(B_2)$	R_2
Column Totals		C_1	C_2	n

1. Table 10.5.3 shows the estimated expected frequencies under the assumption that A and B are independent

Table 10.5.3. A contingency table with estimated expected frequencies				
		Trait B		Row Totals
		B_1	B_2	
Trait A	A_1	R_1C_1/n	R_1C_2/n	R_1
	A_2	R_2C_1/n	R_2C_2/n	R_2
Column Totals		C_1	C_2	n

2. Testing for independence, the general case

TABLE 10.5.4

	B_1	B_2	⋯	B_c	Row Totals
A_1	k_{11}	k_{12}		k_{1c}	R_1
A_2	k_{21}	k_{22}		k_{2c}	R_2
⋮		⋮	⋯	⋮	⋮
A_r	k_{r1}	k_{r2}		k_{rc}	R_r
Column totals	C_1	C_2		C_c	n

Theorem 10.5.1Theorem 10.5.1 Suppose that *n* observations are taken on a sample space partitioned by the events A_1, A_2, ..., A_r and also by events B_1, B_2, ..., B_c. Let $p_i = P(A_i)q_j = P(B_j)$, and $P_{ij} = P(A_i \cap B_j)$, *i* = 1, 2, ..., *c*. Let X_{ij} denote the number of observations belonging to the intersection $A_i \cap B_j$. Then

a. the random variable

$$D_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - np_{ij})^2}{np_{ij}}$$

has approximately a χ^2 distribution with *rc*-1 degrees of freedom (provided $np_{ij} \geq 5$ for all *i* and *j*)

b. to test H_o: the A_i*s* are independent of the B_j*s* calculate the test statistic

$$d_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(k_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}$$

where k_{ij} is the number of observations in the sample that belong to $A_i \cap B_j$, *i* = 1, 2, ..., *r*; j = 1, 2, ..., *c* and $\hat{p}_i$ and $\hat{q}_i$ are the maximum likelihood estimates for pi and qi respectively. The null hypothesis should be rejected at the α level of significance if

$$d_2 \geq \chi^2_{1-\alpha,(r-1)(c-1)}$$

(Analogous to the condition stipulated for all other goodness-of-fit tests, it will be assumed that $n\hat{p}_i\hat{q}_i \geq 5$ for all i and j.

Comment: The number of degrees of freedom associated with a goodness-of-fit statistic is given by the formula

df=number of classes - 1 - number of estimated parameters = (r-1) (c-1)

Comment: The χ^2 distribution with (r-1)(c-1) degrees of freedom provides an adequate approximation to the distribution of d_2 only if $n\hat{p}_i\hat{q}_j \geq 5$ for all i and j. If one or more cells in a contingency table have estimated expected frequencies that are substantially less than 5, the table should be "collapsed" and the rows and columns redefined.

Case Study 10.5.1

```
% LMcs100501_4th.m
observed=[70 65;39 28;14 3;13 2];
% From contincy from stixbox toolbox
[p,chi2,expected] = contincy(observed, 'chi2')
[r,c]=size(observed);df=(r-1)*(c-1);
criticalval=chi2inv(.95,df)
if chi2>criticalval
   disp('Reject Ho')
else
   disp('Fail to reject Ho')
end
```

```
function [p, cs,expected] = contincy(observed, method)
% Modified from the Stixbox toolbox
%CONTINCY  Compute the p-value for contigency table row/col independence.
%
%          p = contincy(observed, method)
%
%          The table  observed  is a count, the  method  is
%
%          'chi2': Pearson chi2-distance
%          'logL': Log likelihood quote distance
```

```
%
%          with default method 'chi2'. The p-value is computed through
%          approximation with chi-2 distribution under the null hypothesis
%          for both methods.
%
%        See also CAT2TBL

%       GPL Copyright (c) Anders Holtsberg, 1999

if nargin < 2
  method = 'chi2';
end

if any(any(observed~=round(observed) | observed<0))
  error('CONTINCY expects counts, that is nonnegative integers')
end

rowsum = sum(observed')';
colsum = sum(observed);
n = sum(rowsum);
expected = rowsum * colsum ./ n;

if strcmp(method, 'chi2')
  cs = sum(sum((expected-observed).^2 ./ expected));
elseif strcmp(method, 'logL')
  I = find(observed>0);
  cs = 2 * sum(observed(I) .* (log(observed(I)./expected(I))));
else
  error('unknown method')
end

% p = 1 - pchisq(cs, (length(rowsum)-1) * (length(colsum)-1));
p = 1-chi2cdf(cs,(length(rowsum)-1) * (length(colsum)-1));
```

Case Study 10.5.2 Siskel & Ebert

```
% LMcs100502_4th.m
observed=[24 8 13; 8 13 11; 10 9 64]
[p,cs,expected] = contincy(observed, 'chi2')
[r,c]=size(observed);df=(r-1)*(c-1);
criticalval=chi2inv(.95,df)
if chi2>criticalval
   disp('Reject Ho')
else
   disp('Fail to reject Ho')
end
```

---

3. Reducing continuous data to contingency tables

Case Study 10.5.3 Mobility and suicide

```
 % LMcs100503_4th.m
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th Edition
% Page 635-637
% An example of a 2x2 contingincy table
% Analyzed using Anders Holsberg's contincy.m & fishertest.m (Matlab
% Central)
% Written by Eugene.Gallagher@umb.edu, Written 11/21/10
% Place the data in a 2 x 2 table:
observed=[7 4;3 11]
[p,chi2,expected] = contincy(observed, 'chi2')
[r,c]=size(observed);df=(r-1)*(c-1);
criticalval=chi2inv(.95,df)
if chi2>criticalval
   disp('Reject Ho')
else
   disp('Fail to reject Ho')
end
[H,P, STATS] = fishertest(observed, 0.05)
```

```
function [H,P, STATS] = fishertest(M, alpha)
% FISHERTEST - Fisher Exact test for 2-x-2 contingency tables
```

```
%
%   H = FISHERTEST(M) performs the non-parametric Fisher exact probability
%   test on a 2-by-2 contingency table described by M, and returns the
%   result in H. It calculates the exact probability of observing the given
%   and more extreme distributions on two variables.  H==0 indicates that
%   the null hypothesis (H0: "the score on one variable is independent from
%   the score on the other variable") cannot be rejected at the 5%
%   significance level.  H==1 indicates that the null hypothesis can be
%   rejected at the 5% level. For practical convenience, the variables can
%   be considered as "0/1 questions" and each observation is casted in
%   one of the cells of the 2-by-2 contingency table [1/1, 1/0 ; 0/1, 0/0].
%
%   If M is a 2-by-2 array, it specifies this 2-by-2 contingency table
%   directly. It holds the observations for each of the four possible
%   combinations.
%   If M is a N-by-2 logical or binary array, the 2-by-2 contingency table
%   is created from it. Each row of M is a single observation that is
%   casted in the appropriate cell of M.
%
%   [H,P,STATS] = FISHERTEST(..) also returns the exact probability P of
%   observing the null-hypothesis and some statistics in the structure
%   STATS, which has the following fields:
%     .M    - the 2-by-2 contingency table
%     .P    - a list of probabilities for the original and all more extreme
%             observations
%     .phi  - the phi coefficient of association between the two attributes
%     .Chi2 - the Chi Square value for the 2-by-2 contingency table
%
%   H =FISHERTEST(M, APLHA) performs the test at the significance level
%   (100*ALPHA)%.  ALPHA must be a scalar between 0 and 1.
%
%   Example:
%     % We have measured the responses of 15 subjects on two 0-1
%     % "questions" and obtained the following results:
```

```
%    %        Q1: 1   0
%    %   Q2: 1     5   1
%    %       0     2   7
%    % (so 5 subjects answered yes on both questions, etc.)
%    M = [ 5 1 ; 2 7]
%    % Our null-hypothesis is that the answers on the two questions are
%    % independent. We apply the Fisher exact test, since the data is
%    % measured on an ordinal scale, and we have far to few observations to
%    % apply a Chi2 test. The result of ...
%    [H,P] = fishertest(M)
%    % (-> H = 1, P = 0.0350)
%    % shows that the probability of observing this distribution M or the
%    % more extreme distributions (i.e., only one in this case: [6 0 ; 1
%    8]) is 0.035. Since this is less than 0.05, we can reject our
%    null-hypothesis indicated by H being 1.
%
%   The Fisher Exact test is most suitable for small numbers of
%   observations, that have been measured on a nominal or ordinal scale.
%   Note that the values 0 and 1 have only arbitray meanings, and do
%   reflect a nominal category, such as yes/no, short/long, above/below
%   average, etc. In matlab words, So, M, M.', flipud(M), etc.  all give
%   the same results.
%
%   See also SIGNTEST, RANKSUM, KRUSKALWALLIS, TTEST, TTEST2 (Stats Toolbox)
%            PERMTEST, COCHRANQTEST (File Exchange)
%
%   This file does not require the Statistics Toolbox.

% Source: Siegel & Castellan, 1988, "Nonparametric statistics for the
%         behavioral  sciences", McGraw-Hill, New York
%
% Created for Matlab R13+
% version 1.0 (feb 2009)
% (c) Jos van der Geest
```

```
% email: jos@jasen.nl
%
% File history:
%   1.0 (feb 2009) - created

error(nargchk(1,2,nargin)) ;

if islogical(M) || all(M(:)==1 | M(:)==0)
   if ndims(M)==2 && size(M,2)== 2
      % each row now holds on observation which can be casted in a 2-by-2 contingency table
      M = logical(M) ;
      A = sum(M(:,1) & M(:,2)) ;
      B = sum(M(:,1) & ~M(:,2)) ;
      C = sum(~M(:,1) & M(:,2)) ;
      D = size(M,1) - (A+B+C) ;
      M = [A B ; C D] ;
   else
      error('For logical or (0,1) input, M should be a N-by-2 array.') ;
   end
elseif ~isnumeric(M) || ~isequal(size(M),[2 2]) || any(M(:)~=fix(M(:))) || any(M(:)<0)
   error ('For numerical input, M should be a 2-by-2 matrix with positive integers.')
end

if nargin < 2 || isempty(alpha)
   alpha = 0.05;
elseif ~isscalar(alpha) || alpha <= 0 || alpha >= 1
   error('Fishertest:BadAlpha','ALPHA must be a scalar between 0 and 1.');
end

% what is the minimum value in the input matrix
[minM, minIDX] = min(M(:)) ;

 if minM > 20,
     warning(sprintf(['Minimum number of observations is larger than 20.\n' ...
```

```
      'Other statistical tests, such as the Chi-square test may be more appropriate.'])) ;
 end

% We will move observations from this cell, and from the cell diagonal to
% it, to the other two. This leaves the sum along rows and columns intact,
% but it will make the matrix more extreme. There will be minM matrixes
% that are more extreme than the original one.
% We can do that by summing with the matrix dM (created below) until the
% cell with the least number of observations has become zero (which takes
% minM steps).
% dM will be either [-1 1 ; 1 -1] or [1 -1 ; -1 1]
dM = ones(2) ;
dM([minIDX (4-minIDX+1)]) = -1 ;

% The row and column sums are always the same
SRC = [sum(M,2).' sum(M,1)] ; % row and column sums
N = SRC(1) + SRC(2) ;          % total number of obervations

if nargout > 2,
   STATS.M = M ; % original matrix
   dt = abs(det(M)) ;
   PRC = prod(SRC) ;       % product of row and column sums
   STATS.phi = dt / sqrt(PRC) ; % Phi coefficient of association
   STATS.Chi2 = (N * (dt - (N/2)).^2) / PRC ; % Chi2 value of independence
end

% pre-allocate the P values for each matrix (the original one and the ones
% more extreme than the original one)
STATS.P = zeros(minM+1,1) ;

% now calculate the probality for observing the matrix M and all the
% matrices that have more extreme distributions. In
for i = 0:minM
   % calculate P value
```

```
    STATS.P(i+1) = local_factratio(SRC,[M(:) ; N]) ;
    % make M more extreme
    M = M + dM ;
end

P = sum(STATS.P) ; % Total probability
H = P < alpha ;    % Is it less then our significance level?,
                   % If so, we can reject our null-hypothesis

% END OF MAIN FUNCTION

function FR = local_factratio(A,B)
% See FACTORIALRATIO for detailed help and comments
% http://www.mathworks.com/matlabcentral/fileexchange/23018
A = A(A>1) - 1 ;
B = B(B>1) - 1 ;
maxE = max([A(:) ; B(:) ; 0]) ;
if maxE > 1 && ~isequal(A,B),
    R = sparse(A,1,1,maxE,2) + sparse(B,2,1,maxE,2) ;
    R = flipud(cumsum(flipud(R))) ;
    R = (R(:,1) - R(:,2)).' ;
    X = 2:(maxE+1) ;
    q = find(R) ;
    FR = prod(X(q).^R(q)) ;
else
    FR = 1 ;
end
```

Appendix 10.A.1: Minitab applications

2. **Taking a second look at statistics (outliers)**

    1. Can data fit a model too well: Gregor Mendel's data

References

Larsen, R. J. and M. L. Marx. 2006. An introduction to mathematical statistics and its applications, 4th edition. Prentice Hall, Upper Saddle River, NJ. 920 pp. {16, 17}

Middleton, G. V. 1999. Data Analysis in the Earth Sciences, using MATLAB. Prentice-Hall. {6}

Ramsey, F. L. and D. W. Schafer. 2002. The statistical sleuth: a course in methods of data analysis, Second edition Duxbury Press, Belmont CA, 742 pp & data diskette.{11, 12}

Salsburg, D. 2001. The lady tasting tea: how statistics revolutionized science in the twentieth century. W. H. Freeman & Co., New York. 340 p. {9, 10}

Zaret, T. M. 1972. Predators, invisible prey, and the nature of polymorphism in the cladocera (Class Crustacea). Limnol. Oceanogr. *17*: 171-184. {13, 14}

Index