Chi-square test

8.1 Goodness of Fit Test

Q: What are statisticians?
A: Mathematicians broken down by age and sex.

Deathbed Scenes. Deathbed scenes in which a dying mother or father holds on life until after the long-absent son returns home and dies immediately after are to familiar in movies. Do such things happen in real life? Can some people postpone their death until after a special event takes place? It is believed that famous people do so with respect to their birthdays to which they attach some importance. A study by David P. Philips (*Statistics: A Guide to the Unknown*, pp 52-65), seems to be consistent with the notion. Philips obtained data[1] on months of birth and death of 1251 famous Americans; the deaths were classified by the time period between the birth dates and death dates as shown in the table:

b	e	f	o	r	e	birth	a	f	t	e	r
6	5	4	3	2	1	month	1	2	3	4	5
90	100	87	96	101	86	119	118	121	114	113	106

Did Mendel cheated? In a remarkable study, the famous statistician R. A. Fisher (*Annals of Science,* **1** *1936, pp 115-137)* examined the experimental data on which Mendel advanced the basic laws of inheritance. In most of Mendel's experiments, it was found that observed ratios of phenotypes were very close to Mendel's predictions. For example, if you toss a coin 100 times, there is some chance of getting exactly 50 heads. But if you claim such a good result on most of the occasions, it arouses suspicion. Fisher showed that the precision of Mendel's data (in showing agreement with his theory) was a of a strikingly high order, which can be expected only 7 times in 100,000 trials. He commented n this rare chance: "Although no explanation can be expected to be satisfactory, it remains the possibility among others that Mendel was deceived by some assistant who knew too well what was expected".

8.2 Contingency Tables: Testing for homogeneity and independence

8.2.1 Theory

	C_1	C_2	$\cdots$	C_n	Total
R_1	n_{11}	n_{12}		n_{1n}	$n_{1\cdot}$
R_2	n_{21}	n_{22}		n_{2n}	$n_{2\cdot}$
R_m	n_{m1}	n_{m2}		n_{mn}	$n_{m\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot n}$	$n_{\cdot\cdot}$

$\chi^2 = \Sigma_{i=1}^m \Sigma_{j=1}^n \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$
where $\hat{n}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n_{\cdot\cdot}}$.
The statistics χ^2 has $(m-1)(n-1)$ degrees of freedom.

[1] 348 were people listed in *Four Hundred Notable Americans* and 903 are listed as foremost families in three volumes of *Who Was Who* for the years 1951-60, 1943-50 and 1897-1942.

School spirit at Duke. Duke has always been known for its great school spirit and support of its athletic teams, as evident by the famous Cameron Crazies. One way that school enthusiasm is shown is by donning Duke paraphernalia including shirts, hats, shorts and sweatshirts. The project of a group of STAT110 students was to explore possible links of the school spirit measured by the number of students wearing paraphernalia and some other attributes. It was hypothesized that men would wear Duke clothes more frequently than women. The data were collected on the Bryan Center walkway starting at 12:00 pm on ten different days. Each day 50 men and 50 women were tallied.

	Duke Paraphenalia	No Duke Paraphenalia	Total
Male	131	369	500
Female	52	448	500
Total	183	817	1000

Solution:

```
> duke_cbind(c(131,52),c(369,448))
> duke
     [,1] [,2]
[1,]  131  369
[2,]   52  448
> chisq.test(duke)

        Pearson's chi-square test with Yates' continuity correction

data:  duke
X-squared = 40.6927, df = 1, p-value = 0
```

Crossing Mushrooms. In an experiment in botany the results of crossing two hybrids of a species of mushrooms *Agaricus bisporus* gave observed frequencies of 120, 53, 36, and 15. Do these results disagree with theoretical frequencies which specify 9:3:3:1 ratio. Use $\alpha = 0.05$.

Solution: 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 $np_4 = 224 \cdot \frac{1}{16} = 14$.

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

Since $\chi^2_{4-1,0.95} = 7.81$, the results do not disagree with the theory.

8.2.2 Splus session

```
> cont_rbind(c(131, 369), c(52, 448))
> cont
     [,1] [,2]
[1,]  131  369
[2,]   52  448

>  chisq.test(cont)

        Pearson's chi-square test with Yates' continuity correction

data:  cont
X-squared = 40.6927, df = 1, p-value = 0
```

8.2.3 Simple formulas for 2 × 2 tables

8.2.4 Paired Contingency Tables and McNemar's Test

If the same sample is studied twice-separated by a certain interval of time or under different conditions, we are no longer dealing with independent samples but rather with dependent or paired samples. Thus the Mantel-Haen test may not be appropriate.

Example: A study[2] involved 85 patients with Hodgkin's disease. Each of these had a normal sibling (one who did not have the disease). In 26 of these pairs, both individuals had had tonsillectonies (T); in 37 pairs, both individuals had not had tonsillectonies (N); in 15 pairs, only the normal individual had had a tonsillectomy; in 7 pairs, only the one with Hodgkin's disease had had a tonsillectomy.

	Normal/T	Normal/N	Total
Patient/T	26	15	41
Patient/N	7	37	44
Total	33	52	85

A goal of the study was to determine whether there was a link between the disease and having had a tonsillectomy: Is the proportion of those who had tonsillectomies the same among those with Hodgkin's disease as among those who do not have it?

8.2.5 McNemar Test: The theory and Splus session

If the contingency table consists of N observations cross-classified on the row and column variables, which would typically have the same levels, then McNemar's statistic could be used to test the null hypothesis of symmetry, namely that the probability of an observation being classified into cell [i,j] is the same as the probability of being classified into cell [j,i].

```
>  mcnemar.test(rbind(c(26,15), c(7,37)) )

    McNemar's chi-square test with continuity correction

data:  rbind(c(26, 15), c(7, 37))
McNemar's chi-square = 2.2273, df = 1, p-value = 0.1356
```

8.3 Additional Problems

Amoebas and intestinal disease. [3] When an epidemic of severe intestinal disease occurred among workers in a plant in South Bend, Indiana, doctors said that the illness resulted from infection with the amoeba Entamoeba histolytica. There are actually two races of these amoebas, large and small, and the large ones were believed to be causing the disease. Doctors suspected that the presence of the small ones might help people resist infection by the large ones. To check on this, public health officials chose a random sample of 138 apparently healthy workers and determined if they were infected with either the large or small amoebas. Table below give the resulting data. is the presence of the large race independent of the presence of the small one?

	Large race		
Small race	Present	Absent	Total
Present	12	23	35
Absent	35	68	103
Total	47	91	138

[2]S. Johnson and R. Johnson, "Tonsillectomy history in Hodgkin's disease," *New Engl.J.Med. 287 (1972) 1122-1125.*

[3]Source: J. E. Cohen (1973). Independence of Amoebas. In *Statistics by Example: Weighing Chances*, edited by F. Mosteller, R. S. Pieters, W. H. Kruskal, G. R. Rising, and R. F. Link, with the assistance of R. Carlson and M. Zelinka, p. 72. Addison- Wesley: Reading, Mass.

Frequency of Occurrences of 'kai' in 10 Pauline Works.

Number of sentences with	Romans (1 - 15)	1st Corinth.	2nd Corinth.	Galat.	Philip.
0 kai	386	424	192	128	42
1 kai	141	152	86	48	29
2 kai's	34	35	28	5	19
3 or more kai's	17	16	13	6	12
No. of sentences	578	627	319	187	102
Total number of kai's	282	281	185	82	107

Number of sentences with	Colos.	1st Thessal.	1st Timothy	2nd Timothy	Hebrews
0 kai	23	34	49	45	155
1 kai	32	23	38	28	94
2 kai's	17	8	9	11	37
3 or more kai's	9	16	10	4	24
No. of sentences	81	81	106	88	310
Total number of kai's	99	99	91	68	253

The Importance of Being Left-Handed. It is not generally known that a coconut tree can be classified as left-handed or right-handed, depending on the direction of its foliar spiral. Some years ago, an investigation of this aspect was undertaken by T. A. Davis at the Indian Statistical Institute (ISI). the study offers a good example of a statistical approach in eliciting information from an understanding of nature.

Why are some trees left-handed and others right-handed? Is this character genetically determined? The question can be answered by considering parents plants of different combinations of foliar spirality and scoring the progeny for the same characteristic. The data collected for this purpose are shown in table below. The ratios of left to right are nearly the same for all combinations of parents indicating that there is no genetic basis for left- or right-handedness.

Pollen	Seed	Progeny		
		left	:	right
Right	Right	44	:	56
Right	Left	47	:	53
Left	Right	45	:	55
Left	Left	47	:	53

So the ratio appears to be entirely determined by external factors which act in a random way. But why is there a slight preponderance of right-handed offsprings (about 55 per cent) in the observed data. There must be something in the environment which tends to give a greater chance for a tree to twist in the right direction. And if so, does this chance depend on the geographical location of trees? This could not be determined until data from various parts of the world could be collected. if was then found that the proportion of left- handers is 0.515, in samples from the Northern Hemisphere, and 0.473 in the Southern Hemisphere. The difference may be due to the influence of the one-way rotation of the Earth, which also explains the phenomenon of the bathtub vortex which, under well-controlled conditions, is shown to be counter- clockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere.

Horse-Kick Fatalities. During the latter part of the nineteenth century, Prussian officials gathered information on the hazards that horses posed to cavalry soldiers. A total of 10 cavalry corps were monitored over a period of 20 years. Recorded for each year and each corps was X, the number of fatalities due to kicks. Table below shows the distribution of X for these 200 "corps-years".

x = Number of Deaths	Observed Number of Corps-Years in Which x Fatalities Occurred
0	109
1	65
2	22
3	3
4	1
	200

Altogether there were 122 fatalities [109(0) + 22(2) + 3(3) + 1(4)], meaning that the observed fatality *rate* was $\frac{122}{200}$, or 0.61 fatalities per corps-year. Bortkiewicz proposed a Poisson model for X with a mean of $c = .61$. Table below shows the expected frequency corresponding to $x = 0, 1, \cdots$, etc. if in fact the Poisson model for X was correct. Clearly the agreement between the observed and the expected frequencies is remarkable.

x	Observed Number of Corps-Years	Expected Number of Corps-Years
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6
	200	199.9

Drinking & Smoking. Alcohol and nicotine consumption during pregnancy are believed to be associated with certain characteristics of children. Since drinking and smoking behaviors may be related, it is important to understand the nature of this relationship when assessing the possible effects of these variables on children. In one study, 452 mothers were classified according to their alcohol intake prior to pregnancy recognition and their nicotine intake during pregnancy. The data are summarized in the following table. [4]

	Nicotine (mg/day)		
Alcohol (ounces/day)	None	1 - 15	16 or more
None	105	7	11
.01-.10	58	5	13
.11-.99	84	37	42
1.00 or more	57	16	17

(a) Calculate the column percents. In what way does the pattern of alcohol consumption vary with nicotine consumption?

(b) Calculate the row percents. In what way does the pattern of nicotine consumption vary with alcohol consumption?

(c) Write H_0 and H_a for assessing whether or not alcohol consumption and nicotine consumption are independent.

(d) Compute the table of expected counts.

(e) Find the X^2 statistic. Report the df and the P-value.

(f) What do you conclude from your analysis of these data?

Alcohol and Marriage. A national survey was conducted to obtain info on the alcohol consumption patterns of American adults by marital status. A random sample of 1772 residents 18 years old and over, yielded the data below. Do the data suggest at 5% significance level that marital status and alcohol consumption patterns are statistically dependent?

[4]Data taken from A. P. Streissguth et al., "Intrauterine alcohol and nicotine exposure: attention and reaction time in 4-year-old children," *Developmental Psychology*, 20 (1984), pp. 533-541.

	Abstain	1 - 60	over 60
Single	67	213	74
Widowed	85	633	129
Divorced	27	60	15

Cell Counts. A student takes the blood cell count of five random blood samples from a larger volume of solution to determine if it is well mixed. She expects the cell counts to be distributed uniformly. The data is given below:

sample blood cell count	expected blood cell count
35	27.2
20	27.2
25	27.2
25	27.2
31	27.2
136	136

Can she depend on the results to be uniformly distributed? Use the chi-square test to determine the answer. Take $\alpha = 0.05$.

Family Sizes. A demographer surveys 1000 randomly chosen American families and records their family sizes and family incomes:

	Family Size					
Family Income	2	3	4	5	6	7
Low	145	81	57	22	9	8
Middle	151	73	71	33	13	10
High	124	60	80	42	13	8

Do the data provide sufficient evidence to conclude that family size and family income are statistically dependent?
(a) State H_0 and H_1 hypotheses.
(b) Perform the test. Use $\alpha = 0.05$; Comment.

Aggression on Halloween. A study was designed to test whether or not aggression is a function of anonymity. The study was conducted as a field experiment on Halloween; 300 children were observed unobtrusively as they made their rounds. Of these 300children, 173 wore masks that completely covered their faces, while 127 wore no masks. It was found that 101 children in the masked group displayed aggressive or antisocial behavior versus 36 children in unmasked group.
What conclusion can be drawn. State it in terminology of the problem. Use $\alpha = 0.01$.
[$\chi^2 = 26.63, \chi^2_1 = 6.635$, Reject H_0]

Importance of bystanders. Darley and Latané (1968) asked subjects to participate in a discussion carried over an intercom. Aside from the experimenter to whom they were speaking, subjects thought that there were zero, one, or four other people (bystanders) also listening over intercom. Part way through the discussion, the experimenter feigned serious illness and ask for help. Darley and Latané noted how often the subject sought help for the experimenter as a function of the number of supposed bystanders. The data are give in the table:

	Sought Assistance	No Assistance
No bystanders	11	2
One bystander	16	10
Four bystanders	4	9

What could Darley and Latané conclude from the results?

- State H_0 and H_1.
- Perform the test. Use MINITAB.

[Sol. H_0: Assistance and the number of bystanders are independent.

```
           C1       C2    Total
    1      11        2       13
         7.75     5.25

    2      16       10       26
        15.50    10.50

    3       4        9       13
         7.75     5.25

Total      31       21       52

ChiSq =  1.363 +  2.012 +
         0.016 +  0.024 +
         1.815 +  2.679 = 7.908
df = 2
```

p-value $= 0.0192$.]

Not at all like me. You have a theory that if you ask subjects to sort one-sentence characteristics of people (e.g. "I eat to fast") into five piles ranging from *not at all like me* to *very much like me,* the percentage placed in each of five piles will be approximately 10, 20, 40, 20, and 10. You have one of your friends sort 50 statements, and you obtain the following data: 8, 9, 21, 8, and 4.

Do these data support your hypothesis?

[Sol. $\chi^2 = (5-8)^2/8 + (10-9)^2/9 + (20-21)^2/21 + (10-8)^2/8 + (5-4)^2/4 = 2.03373$. p-value $= 0.7296891$.]

Cell Counts. A student takes the blood cell count of five random blood samples from a larger volume of solution to determine if it is well mixed. She expects the cell counts to be distributed uniformly. The data is given below:

sample blood cell count	expected blood cell count
35	27.2
20	27.2
25	27.2
25	27.2
31	27.2
136	136

1. Can she depend on the results to be uniformly distributed? Do the chi-square test with $\alpha = 0.05$.

Nightmares. Over the years numerous studies have sought to characterize the nightmare sufferer. Out of these has emerged the stereotype of someone with high anxiety, low ego strength, feeling of inadequacy, and poorer-than-average physical health. What is not so well known, though, is whether the sex is independent of having frequent nightmares. Using Hersen's[5] data test this independence at level $\alpha = 0.05$..

[5] Hersen, M. (1971). Personality characteristics of nightmare sufferers. *Journal of Nervous and Mental Diseases, 153, 29-31.* Hersen looked at nightmare frequencies for a sample of 160 men and 192 women. Each subject was asked whether he or she experienced nightmares *often* (at least once a month) or *seldom* (less than once in month)

	Men	Women
Nightmares often	55	60
Nightmares seldom	105	132

SGS. Below is a part of data from 1984 and 1990 General Social Survey which is conducted annually by the National Opinion Research Center. Random samples of 1473 persons in 1984 and 1372 persons in 1990 were taken using multistage cluster sampling. One of the questions in a 67-questions-long questionary was: **Do you think most people would try to take advantage of you if they got a chance, or they try to be fair?**

	1984	**1990**
1. Would take advantage of you	507	325
2. Would try to be fair	913	515
3. Depends	47	53
4. No answer	6	6

1. Assuming that **1984** frequencies are theoretical, explore how the **1990** frequencies agree. Use $\alpha = 0.05$. State your findings clearly.

Stock Market

There are many "indicators" that investors use to predict the behavior of the stock market. One of these is the "January Indicator." Some investors believe that if the market is up in January, then it will be up for the rest of the year. On the other hand if it is down in January, then it will be down for the rest of the year. The following table gives data for 72 years from 1916 to 1987:

	Up Jan.	Down Jan.
Up Feb-Dec.	33	13
Down Feb-Dec.	13	13

1. Give the description of an appropriate null and alternative hypothesis.
2. Do the test of hypotheses from 1 at $\alpha = 0.01$ and state your decision.

Drinking trends on Duke campus. Stereotypically, drinking and partying have always been synonymous with college life. Recently, this has been a controversial issue at Duke because of the

implementation of the *alcohol policy*. As the result of the alcohol policy, fraternities are no longer permitted to openly distribute alcohol to party goers. Brandie, Randy, and Stephanie[6] decided to test how the alcohol policy, and the atmosphere it generates, affects drinking trends on the Duke campus.

Part of their data includes the following table:

	hard alcohol	beer	total
male	20	27	47
female	27	7	34
total	47	34	81

- (i) Test for independence of factors (gender, type of drink). Use $\alpha = 0.05$.
- (ii) Explain in words what constitutes the error of II kind in the above testing.

Strokes. In a long study of heart disease the day of the week on which 63 seemingly healthy men died was recorded. These men had no history of disease and died suddenly.

Day of Week	Mon.	Tues.	Weds.	Thurs.	Fri.	Sat.	Sun.
No. of Deaths	22	7	6	13	5	4	6

- (i) Test the hypothesis that these men were just as likely to die on one day as on any other. Use $\alpha = 0.05$.
- (ii) Explain in words what constitutes the error of II kind in the above testing.

[6]Brandie Littlefield, Randy Savino, and Stephanie Weiss: An analysis of the drinking trends on Duke campus, STA110E Project, Fall 1995.