Stat 104: Quantitative Methods

Class 27: Chi-Square Tests

Review

Paul's housemate Miranda is trying to convince Paul to get a piercing. Paul is very sensitive to popular opinion, however, and will only get one if "everyone else is doing it". Of course, everyone is unrealistic, so Paul will settle on getting the piercing if Miranda can show that more than 60% of people at Harvard have one. Miranda wastes no time in taking a sample of 60 people from Adam's House and performing the appropriate hypothesis test. Her sample revealed 44 people who admitted to having a piercing.

(a) Test the hypothesis $H_0 : p = 0.6$ versus $H_a : p > 0.6$ Will Paul end up getting a piercing ?

(b) Do you have any criticisms of Miranda's sample ?

Review

The federal government would like to test the hypothesis that the average age of men filing for Social Security is higher than the average age of women set using $\alpha = 0.05$ with the following data:

	Men	Women
Sample mean	64.5 years	63.6 years
Sample size	35	39
Population standard deviation	3.0 years	3.5 years

Review

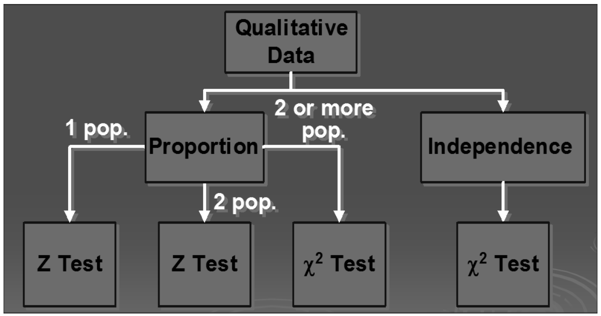
Obtain output from R and decide which one of the following statements is true? Because the *p*-value is

a) greater than .05, we fail to reject the null hypothesis and can conclude that the average age of men filling for social security is more than the average age of women.
b) greater than .05, we fail to reject the null hypothesis and cannot conclude that the average age of men filling for social security is more than the average age of women.
c) less than .05, we fail to reject the null hypothesis and conclude that the average age of men filling for social security is more than the average age of women.
d) less than .05, we reject the null hypothesis and cannot conclude that the average age of men filling for social security is less than the average age of women.

Onto Chi Square..Two Techniques…

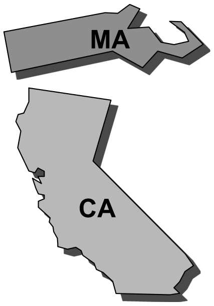
- The first is a ***goodness-of-fit test*** applied to data produced by a ***multinomial experiment***, a generalization of a binomial experiment and is used to describe one population of data.
- The second uses data arranged in a ***contingency table*** to determine whether two classifications of a population of nominal data are ***statistically independent***; this test can also be interpreted as a comparison of two or more populations.

Qualitative Data



Test for Two Proportions

- You’re an epidemiologist for the US Department of Health and Human Services. You’re studying the prevalence of disease X in two states (MA and CA). In MA, 74 of 1500 people surveyed were diseased and in CA, 129 of 1500 were diseased. At .05 level, does MA have a lower prevalence rate?



Example (cont)

- Want to test

H_0: $p_{MA} - p_{CA} = 0$

H_a: $p_{MA} - p_{CA} < 0$

$\alpha = .05$

Interpret R Output

```
> prop.test(c(74,129),c(1500,1500),alt="less")

        2-sample test for equality of proportions with continuity
correction

data:  c(74, 129) out of c(1500, 1500)
X-squared = 15.407, df = 1, p-value = 4.333e-05
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000 -0.0209544
sample estimates:
    prop 1     prop 2
0.04933333 0.08600000
```

Goodness of Fit

- ❑ A goodness-of-fit test is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.
- ❑ We use it to see if observed data follow some specified distribution.
- ❑ We will use it simply in the multinomial setting.

The Multinomial Experiment…

Unlike a binomial experiment which only has two possible outcomes (e.g. heads or tails), a ***multinomial experiment***:

- Consists of a fixed number, **n**, of trials.
- Each trial can have one of **k** outcomes, called cells.
- Each probability $\mathbf{p_i}$ remains constant.
- Our usual notion of probabilities holds, namely:

  $p_1 + p_2 + \ldots + p_k = 1$, and
- Each trial is ***independent*** of the other trials.

Chi-squared Goodness-of-Fit Test

We test whether there is sufficient evidence to reject a ***specified set*** of values for p_i.

To illustrate, our null hypothesis is:

H_0: $p_1 = a_1$, $p_2 = a_2$, …, $p_k = a_k$

where a_1, a_2, …, a_k are the values we want to test.

Our research hypothesis is:

H_a: At least one p_i is not equal to its specified value

Goodness-of-Fit Notation

O **represents the observed frequency of an outcome.**

E **represents the expected frequency of an outcome.**

k **represents the number of different categories or outcomes.**

n **represents the total number of trials.**

Example

- Two companies, A and B, have recently conducted aggressive advertising campaigns to maintain and possibly increase their respective shares of the market for **fabric softener**.
- These two companies enjoy a dominant position in the market.
- Before the advertising campaigns began, the market share of company A was 45%, whereas company B had 40% of the market. Other competitors accounted for the remaining 15%.
- Has their market share changed after the campaign?

Example

- A marketing analyst solicited the preferences of a random sample of 200 customers of fabric softener.
- Of the 200 customers, 102 indicated a preference for company A's product, 82 preferred company B's fabric softener, and the remaining 16 preferred the products of one of the competitors.
- Can the analyst infer at the 5% significance level that customer preferences have changed from their levels before the advertising campaigns were launched?

Example

We compare market share ***before*** and ***after*** an advertising campaign to see if there is a ***difference*** (i.e. if the advertising was effective in improving market share). We hypothesize values for the parameters equal to the before-market share. That is,

H_0: $p_1 = .45$, $p_2 = .40$, $p_3 = .15$

The alternative hypothesis is a denial of the null. That is,

H_a: At least one p_i is not equal to its specified value

Example

If the null hypothesis is true, we would **expect** the number of customers selecting brand A, brand B, and other to be 200 times the proportions specified under the null hypothesis. That is,

$e_1 = 200(.45) = 90$

$e_2 = 200(.40) = 80$

$e_3 = 200(.15) = 30$

In general, the expected frequency for each cell is given by

$e_i = np_i$

This expression is derived from the formula for the expected value of a binomial random variable, which we covered a few weeks ago.

Example

- If the expected frequencies and the observed frequencies are quite different, we would conclude that the null hypothesis is false, and we would reject it.
- However, if the expected and observed frequencies are similar, we would not reject the null hypothesis.
- The test statistic measures the **similarity of the expected and observed frequencies.**

Chi-squared Goodness-of-Fit Test

The Chi-squared goodness of fit test statistic is given by:

observed frequency

expected frequency

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

What is the smallest value possible for this statistic?

What do small values imply?

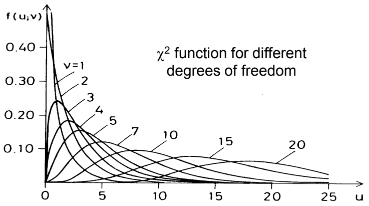
What do large values imply?

Note: this statistic is *approximately* Chi-squared with k–1 degrees of freedom provided the sample size is large. The decision rule is reject if:

$$\chi^2 > \chi^2_{\alpha,k-1}$$

The Chi-Square Distribution

- The Chi-Square distribution is a probability distribution like the Normal, Binomial, or t.
- It is a continuous probability distribution, whose shape depends on its “degrees of freedom”.



χ^2 function for different degrees of freedom

Critical Chi-Square

- ❑ Critical values for chi-square are found in tables, or using the display `qchisq(0.95,df` command in R.
- ❑ If your calculated chi-square value is greater than the critical value, you “reject the null hypothesis”.
- ❑ If your chi-square value is less than the critical value, you “fail to reject” the null hypothesis.

Chi-Square Table

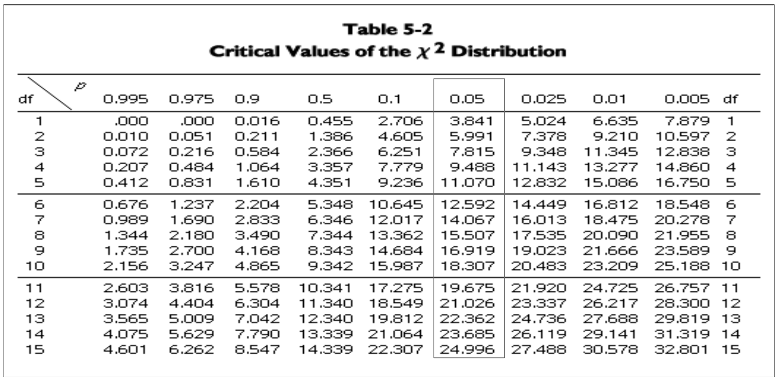


Table 5-2
Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15

Calculating Cut-Off Values in R

```
> qchisq(.95,1)
[1] 3.841459
> qchisq(.95,2)
[1] 5.991465
> qchisq(.95,3)
[1] 7.814728
> qchisq(.95,4)
[1] 9.487729
```

Example

In order to calculate our test statistic, we lay-out the data in a tabular fashion for easier calculation by hand:

Company	Observed Frequency	Expected Frequency	Delta	Summation Component
	o_i	**e_i**	**$(o_i - ei)$**	**$(o_i - e_i)^2/e_i$**
A	102	90	12	1.60
B	82	80	2	0.05
Others	16	30	-14	6.53
Total	200	200		**8.18**

Check that these are equal

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

Example

Our rejection region is:

$$\chi^2 > \chi^2_{\alpha,k-1} = \chi^2_{.05,3-1} = 5.99147$$

Since our test statistic is 8.18 which is greater than our critical value for Chi-squared, we reject H_0 in favor of H_a, that is,

"There is sufficient evidence to infer that the proportions <u>have changed</u> since the advertising campaigns were implemented"

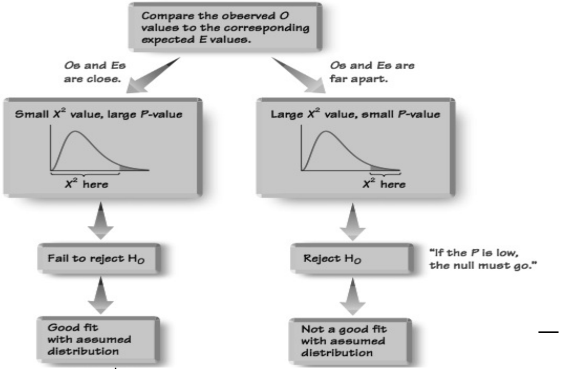
Using the Computer

```
> chisq.test(x=c(102,82,16),p=c(90,80,30)/200)

        Chi-squared test for given probabilities

data:  c(102, 82, 16)
X-squared = 8.1833, df = 2, p-value = 0.01671
```

Goodness-of-Fit Requirements

1. The data have been randomly selected.
2. The sample data consist of frequency counts for each of the different categories.
3. For each category, the expected frequency is at least 5. (The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. There is no requirement that the *observed* frequency for each category must be at least 5.)

Relationships Among the χ^2 Test Statistic, P-Value, and Goodness-of-Fit



Compare the observed *O* values to the corresponding expected *E* values.

Os and Es are close.

Small X^2 value, large *P*-value

X^2 here

Fail to reject H_0

Good fit with assumed distribution

Os and Es are far apart.

Large X^2 value, small *P*-value

X^2 here

Reject H_0

"If the *P* is low, the null must go."

Not a good fit with assumed distribution

Example

- A market analyst wished to see whether consumers have any preference among five flavors of a new fruit soda.
- A sample of 100 people provided the following data.
- Is there enough evidence to reject the claim that there is no preference in the selection of fruit soda flavors?

	Cherry	Strawberry	Orange	Lime	Grape
Observed	32	28	16	14	10
Expected	20	20	20	20	20

Example

- **State the hypotheses and identify the claim.**
  - H_0: Consumers show no preference (claim).
  - H_a: Consumers show a preference.
- Compute the test statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

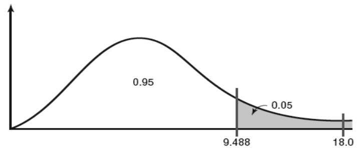
$$= \frac{(32-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(14-20)^2}{20}$$

$$+ \frac{(10-20)^2}{20} = 18$$

Example

Make the decision.

The decision is to reject the null hypothesis, since 18.0 > 9.488.



Summarize the results.

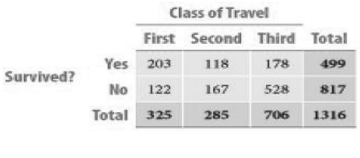
There is enough evidence to reject the claim that consumers show no preference for the flavors.

Computer Output

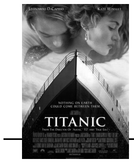
```
chisq.test(x=c(32,28,16,14,10),p=c(1/5,1/5,1/5,1/5,1/5))

        Chi-squared test for given probabilities

data:  c(32, 28, 16, 14, 10)
X-squared = 18, df = 4, p-value = 0.001234
```

The Other Chi-Square Test

- The Chi-Square test is a bit confusing since there are two of them; the one-way GOF and the two-way.
- The two-way Chi-Square test looks at contingency tables to see if the rows and columns are independent or dependent:



		Class of Travel			
		First	Second	Third	Total
Survived?	Yes	203	118	178	499
	No	122	167	528	817
	Total	325	285	706	1316



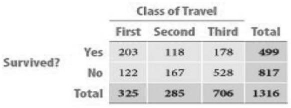
Chi-squared Test of a Contingency Table

- For a contingency table that has *r* rows and *c* columns, the chi square test can be thought of as a test of independence
- In a test of independence the null and alternative hypotheses are:

Ho: The two categorical variables are independent.

Ha: The two categorical variables are related.

Example-The Titanic



		Class of Travel			
		First	Second	Third	Total
Survived?	Yes	203	118	178	499
	No	122	167	528	817
	Total	325	285	706	1316

```
> mydata=matrix(nrow=2,ncol=3,c(203,122,118,167,178,528))
> rownames(mydata)=c("Survived","Didn't Survive")
> colnames(mydata)=c("Fist","Second","Third")
> mydata
               Fist Second Third
Survived        203    118   178
Didn't Survive  122    167   528
> chisq.test(mydata)

        Pearson's Chi-squared test

data:  mydata
X-squared = 133.05, df = 2, p-value < 2.2e-16
```

Example

- ❑ Is there a relationship between undergraduate major and MBA major?
- ❑ Suppose the undergraduate degrees are BA, BEng, BBA, and several others.
- ❑ There are three possible majors for the MBA students, accounting, finance, and marketing.
- ❑ Can the statistician conclude that the undergraduate degree affects the choice of major?

Example

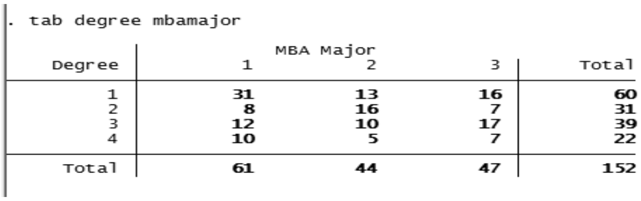
The data are stored in two columns. The first column consist of integers 1, 2, 3, and 4 representing the undergraduate degree where

1 = BA
2 = BEng
3 = BBA
4 = other

The second column lists the MBA major where

1= Accounting
2 = Finance
3 = Marketing

Tabulated Data



```
. tab degree mbamajor

                       MBA Major
    Degree |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        31         13         16 |        60
         2 |         8         16          7 |        31
         3 |        12         10         17 |        39
         4 |        10          5          7 |        22
-----------+---------------------------------+----------
     Total |        61         44         47 |       152
```

Example

The problem objective is to determine whether two variables (undergraduate degree and MBA major) are related. Both variables are categorical. Thus, the technique to use is the chi-squared test of a contingency table. The alternative hypotheses specifies what we test. That is,

H_a: The two variables are **dependent**

The null hypothesis is a denial of the alternative hypothesis.

H_0: The two variables are **independent**.

Test Statistic

The test statistic is the same as the one used to test proportions in the goodness-of-fit-test. That is, the test statistic is

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

Note however, that there is a major difference between the two applications. In this one the null does not specify the proportions p_i, from which we compute the expected values e_i, which we need to calculate the χ^2 test statistic. That is, we **cannot use**

$e = np_i$

because we don't know the p_i (they are not specified by the null hypothesis). It is necessary to **estimate** the p_i from the data.

Example

We start with what is called the contingency table of counts.

	MBA Major			
Undergrad Degree	Accounting	Finance	Marketing	Total
BA	31	13	16	60
BEng	8	16	7	31
BBA	12	10	17	39
Other	10	5	7	22
Total	61	44	47	**152**

Example

If the null hypothesis is true (Remember we always start with this assumption.) and the two categorical variables are independent, then, for example

P(BA and Accounting) = [P(BA)] [P(Accounting)]

Since we don't know the values of P(BA) or P(Accounting)

We need to use the data to estimate the probabilities.

Estimating the Probabilities

There are 152 students of which 61 who have chosen accounting as their MBA major. Thus, we estimate the probability of accounting as

P(Accounting) $\approx \frac{61}{152} = .401$

Similarly P(BA) $\approx \frac{60}{152} = .395$

Example

If the null hypothesis is true

P(BA and Accounting) = (60/152)(61/152)

Now that we have the probability we can calculate the expected value. That is,

E(BA and Accounting) = 152(60/152)(61/152)
= (60)(61)/152 = 24.08

We can do the same for the other 11 cells.

Example

We can now compare ***observed*** with ***expected*** frequencies…

	MBA Major					
Undergrad Degree	Accounting		Finance		Marketing	
BA	31	24.08	13	17.37	16	18.55
BEng	8	12.44	16	8.97	7	9.59
BBA	12	15.65	10	11.29	17	12.06
Other	10	8.83	5	6.37	7	6.80

and calculate our test statistic:

$$\chi^2 = \frac{(31-24.08)^2}{24.08} + \frac{(13-17.37)^2}{17.37} + \ldots + \frac{(7-6.80)^2}{6.80} = 14.70$$

Using R

```
> mydata=matrix(nrow=4,ncol=3,c(31,8,12,10,13,16,10,5,16,7,17,7))
> rownames(mydata)=c("BA","BEng","BBA","Other")
> colnames(mydata)=c("Acctg","Finance","Marketing")
> mydata
      Acctg Finance Marketing
BA       31      13        16
BEng      8      16         7
BBA      12      10        17
Other    10       5         7
> chisq.test(mydata)

        Pearson's Chi-squared test

data:  mydata
X-squared = 14.702, df = 6, p-value = 0.02271
```

Example

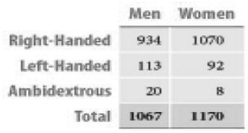
The p-value is .023. There is enough evidence to infer that the MBA major and the undergraduate degree are related.

We can also interpret the results of this test in two other ways.

1. There is enough evidence to infer that there are differences in MBA major between the four undergraduate categories.
2. There is enough evidence to infer that there are differences in undergraduate degree between the majors.

Example

- ❑ Is there a difference between handedness patterns in men and women?
- ❑ A good set of data to help you answer this question comes from the government’s 5-year Health and Nutrition Survey (HANES) of 1976–80, which recorded the gender and handedness of a random sample of 2237 individuals from across the country.



	Men	Women
Right-Handed	934	1070
Left-Handed	113	92
Ambidextrous	20	8
Total	1067	1170

R Output

```
> mydata=matrix(nrow=3,ncol=2,c(934,113,20,1070,92,8))
> rownames(mydata)=c("Right","Left","Ambi")
> colnames(mydata)=c("Men","Women")
> chisq.test(mydata)

        Pearson's Chi-squared test

data:  mydata
X-squared = 11.806, df = 2, p-value = 0.002731
```

Example

- ❑The p-value is 0.003 so we Reject Ho.
- ❑It appears that women are more likely to be right-handed and men are more likely to be left-handed or ambidextrous.

Detailed Example

- Derek wants to know if the geographical area that a student grew up in is associated with whether or not that the student drinks alcohol. Below are the results he obtained from a random sample of Harvard students

	No	Yes	Total
Big City	21	65	86
Rural	11	130	141
Small Town	18	198	216
Suburban	37	345	382
Total	87	738	825

Detailed Example

1. H_o: There is no relationship between the geographical area that a student grew up and whether or not that the student drinks alcohol.

   H_a: There is relationship between the geographical area that a student grew up and whether or not that the student drinks alcohol.

2. To check the conditions we need to calculate the expected counts for each cell.

   $E_{11} = (R_1xC_1)/n = (86x87)/825 = 9.07$,
   $E_{12} = (R_1xC_2)/n = (86x738)/825 = 76.93$, …

   $E_{32} = (R_3xC_2)/n =$ ____________________, …

Detailed Example

Here is the output with the Observed and Expected counts for each cell. We can see that the conditions are satisfied!

```
             No     Yes     All

Big_City     21      65      86
           9.07   76.93   86.00

Rural        11     130     141
          14.87  126.13  141.00

SmallTow     18     198     216
          22.78  193.22  216.00

Suburban     37     345     382
          40.28  341.72  382.00

All          87     738     825
          87.00  738.00  825.00
```

Detailed Example

3. Chi- Square statistic and P-value:

$$\chi^2 = \text{sum } \{(\text{Observed} - \text{Expected})^2/\text{Expected}\}$$
$$= (21-9.07)^2/9.07 + (65-76.93)^2/76.93$$
$$+ (11-14.87)^2/14.87 + (130-126.13)^2/126.13$$
$$+ (18-22.78)^2/22.78 + (198-193.22)^2/193.22$$
$$+ (37-40.28)^2/40.28 + (345-341.72)^2/341.72$$
$$= 20.091$$

df = (4-1)x(2-1) =3

4. Cutoff = 7.81 so the test is significant, and we can reject the null.

5. We can conclude that there is a relationship between the geographical area that a student grew up and whether or not that the student drinks alcohol.

R Output

```
> mydata=matrix(nrow=4,ncol=2,c(21,11,18,37,65,130,198,345))
> rownames(mydata)=c("Big","Rural","Small","Suburban")
> colnames(mydata)=c("No","Yes")
> mydata
         No Yes
Big      21  65
Rural    11 130
Small    18 198
Suburban 37 345
> chisq.test(mydata)

        Pearson's Chi-squared test

data:  mydata
X-squared = 20.091, df = 3, p-value = 0.0001625
```

Things you should know

❑Chi-Square Goodness of Fit Test

❑Chi-Square Test for Independence