Chap 10 Notes: Chi-Square Testing

J. Harner A. Billings

Department of Statistics
West Virginia University

Stat 211 Fall 2007

Outline

χ^2 Tests:

- A **binomial experiment** allows only two possible outcomes on each trial.
- A **multinomial experiment** allows a countable number of possible outcomes ($k > 2$) on each trial.

Properties of a multinomial experiment:

1. A fixed number of trials, n.
2. Each trial results in exactly one of k possible outcomes.
3. p_i is the probability of getting outcome i on a single trial, and:
$$p_1 + p_2 + \ldots + p_k = 1.$$
4. Trials are independent.

A multinomial distribution can be used to analyze a multinomial experiment.

Multinomial Example (Equal Probabilities):

Theoretical probability distribution for a fair 6–sided die:

Outcome	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Theoretical probability distribution for a fair 6–sided die tossed 120 times:

Outcome	1	2	3	4	5	6
Frequency	20	20	20	20	20	20

These are called "Expected Frequencies" and are computed as:

$$E_i = np_i$$

for outcome i.

Multinomial Example (cont.):

The observed (or experimental) cell frequencies for a fair 6–sided die might look like (n = 120):

Outcome	1	2	3	4	5	6
Obs. frequency	14	18	28	17	23	20

The observed frequencies sum to n.

Note:

- A **cell** is one of the k possible outcomes (categories)
- The observed cell frequencies will be denoted by O_i.
- The expected cell frequencies will be denoted by E_i.

Goodness of Fit Test

We can determine how well the observed multinomial distribution “fits” the theoretical expected multinomial distribution by calculating the difference $O_i - E_i$, for each cell (category).

The further the observed frequencies are from their respective expected frequencies, the bigger the magnitude of the differences.

This idea will lead to a test statistic for a “goodness of fit” hypothesis test.

Multinomial Example (Cont.)

We can organize our information about a multinomial experiment into a single table:

Outcome	O	E	$O - E$	$\|O - E\|$
1	14	20	-6	6
2	18	20	-2	2
3	28	20	8	8
4	17	20	-3	3
5	23	20	3	3
6	20	20	0	0
total	120	120	0	22

Are the observed O_i reasonable if the die is fair?

We can use $D = \sum |O - E|$ as a test statistic. Use the five-step to estimate $P(D \geq 22|H_o)$.

Test Statistic

Because each multinomial experiment can have a different number of cells (categories) k and a different number of trials n, we must do something to "standardize" the differences.

In stead of computing $|O - E|$ for each cell, we compute

$$\frac{(O - E)^2}{E}.$$

Adding these values yields

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

Large values of χ^2 indicate a departure from what was expected under the assumed multinomial.

Note: χ^2 is called Chi - square.

The Chi-square Test Statistic

The statistic

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

can be used to perform a “goodness of fit” test.

χ^2 has an approximate Chi-square distribution (Sect. 10.5).

χ^2 Goodness-of-Fit Example

For the Fall '02 term, Stat 211 students were given a choice of 4 lecture sections which met at different times with distinct instructors. The data that follows shows the number of students who selected each section.

Do the data indicate that the students exhibit a preference for selections, or do the data indicate that all sections are equally likely to be chosen?
(Use $\alpha = 0.05$.)

1. H_0 : no preference for certain sections,
   i.e., all sections are equally likely to be chosen.
   We can write H_0 as:

$$H_0 : p_1 = p_2 = p_3 = p_4 = 0.25.$$

Goodness-of-Fit Example (cont.)

2. H_A: There is preference for some sections(s), i.e., at least one section has a different probability of being chosen than the other section(s).

   We can write H_A as:

   $$H_A : p_i \neq 0.25$$

   for at least one i.

3. $\alpha = 0.05$

Goodness-of-Fit Example (cont.)

We will display the data and computations of expected cell frequencies and χ^2 test statistics

Recall: $E_i = np_i$ for cell i.

Lecture	O	E	$O - E$	$\frac{(O-E)^2}{E}$
1	69	99.25	-30.25	9.2198
2	116	99.25	16.75	2.8268
3	116	99.25	16.75	2.8298
4	96	99.25	-3.25	0.1064
total	397	397	0	14.9798

4. The value of our test statistic is:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 14.9798$$

Goodness-of-Fit Example (cont.)

5. Find the P-value of test statistic.

$$\begin{aligned} P\text{-value} &= P(\chi^2 \geq \text{test statistic} | H_0) \\ &= P(\chi^2 \geq 14.9798) \end{aligned}$$

To find this we can use Table C of the Text Book. First, the degrees of freedom:

$$\mathrm{df} = k - 1 = 4 - 1 = 3.$$

Look at the row for $\mathrm{df} = 3$ and try to find the value of the test statistic. From Table C:

$$P(\chi^2 \geq 14.32) = 0.0025$$

and

$$P(\chi^2 \geq 16.27) = 0.001.$$

So our P-value is between 0.001 and 0.0025, i.e., $0.001 \leq P(\chi^2 \geq 14.98) \leq 0.0025$.

Goodness-of-Fit Example (cont.)

6. Make a decision

   Reject H_0 if P-value $\leq \alpha$.

   Do not reject H_0 if P-value $> \alpha$.

   Since
   $$0.001 < P\text{-value} < 0.0025,$$
   it is clear that
   $$P\text{-value} \leq \alpha$$
   for $\alpha = 0.05$.

Goodness-of-Fit Example (cont.)

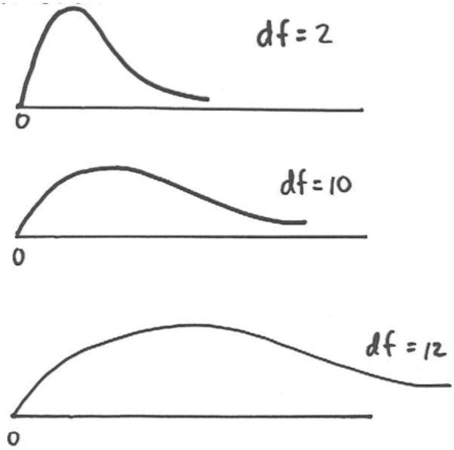
7. Conclusion:
   At the 5% significance level, the data indicates that at least one section has a different probability of being selected by Stat 211 students than the other sections.

Conditions:

1. Independent observations.
2. $E > 5$ for each cell.
3. Counts not percents or proportions.

χ^2 distributions:

The χ^2 distribution is similar to the t-distributions in that it is indexed by a single parameter, the degrees of freedom.



Properties of the Chi-square Distribution

The χ^2 density:

1. is asymmetric (positively skewed);
2. is continuous;
3. has non-negative values.

The test statistic

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

is approximately distributed as a Chi-square with $k - 1$ degrees of freedom.

χ^2 Testing

We have already addressed the concepts in **Sect. 10.6** when we did the example in **Sect. 10.4**.

χ^2 Test for Normality:

We can use the “goodness of fit” test to determine if the observed data follows (an approximate) normal distribution.

Recall the “empirical rule” for normal distributions (Ch. 6)

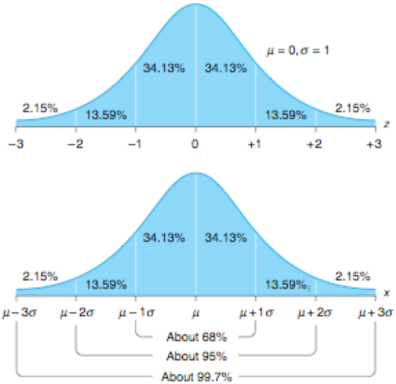


Figure 6.12 Areas under the standard normal curve derived from the normal table and the corresponding areas under the normal curve with mean μ and standard deviation σ.

Normality Test Example

The empirical rule allows us to partition the observed data values into intervals, each with width equal to σ. The percentages in the intervals allow us to compute the expected cell frequencies.

Determine if the average daily temperatures in Morgantown during the month of July follow a Normal distribution. Data was collected for the past 10 years (i.e., 310 July days), which yielded a sample mean of 75°F with a sample standard deviation of 6°F.

Use a significance level $\alpha = 0.05$.

Normality Test Example (cont.)

First note n = 310. So $\bar{X} \simeq \mu$ and $s \simeq \sigma$ by the CLT. Use $\sigma \simeq s = 6$ and $\mu \simeq \bar{X} = 75$ to construct our intervals.

	Region	Interval	Expected Proportion
1.	$\mu - 3\sigma$ to $\mu - 2\sigma$	57 to 63	0.0215
2.	$\mu - 2\sigma$ to $\mu - \sigma$	63 to 69	0.1359
3.	$\mu - \sigma$ to μ	69 to 75	0.3413
4.	μ to $\mu + \sigma$	75 to 81	0.3413
5.	$\mu + \sigma$ to $\mu + 2\sigma$	81 to 87	0.1359
6.	$\mu + 2\sigma$ to $\mu + 3\sigma$	87 to 93	0.0215

We can use these proportions to compute the E's (expected cell frequencies).

Normality Test Example (cont.)

Recall $E_i = np_i$ for cell i in our table.

$$
\begin{aligned}
E_1 = 310 \times 0.0215 &= 6.66 \\
E_2 = 310 \times 0.1359 &= 42.13 \\
E_3 = 310 \times 0.3413 &= 105.80 \\
E_4 = 310 \times 0.3413 &= 105.80 \\
E_5 = 310 \times 0.1359 &= 42.13 \\
E_6 = 310 \times 0.0215 &= 6.66
\end{aligned}
$$

Normality Test Example (cont.)

Construct the hypothesis test:

1. H_0: Data comes from a normal distribution, which we could also write as:

$$\begin{aligned} H_0 \quad : \quad & p_1 = 0.0215, p_2 = 0.1359, p_3 = 0.3413, \\ & p_4 = 0.3413, p_5 = 0.1359, p_6 = 0.0215. \end{aligned}$$

2. H_A: Data does not follow a normal distribution.
3. $\alpha = 0.05$
4. To compute the test statistic, construct a new table with E's, O's, and $\frac{(O-E)^2}{E}$'s.

Normality Test Example (cont.)

Outcome (interval)	O # days	E # days	$\frac{(O-E)^2}{E}$
57 to 63	8	6.66	0.2969
63 to 69	39	42.13	0.2324
69 to 75	106	105.80	0.0004
75 to 81	110	105.80	0.1.667
81 to 87	41	42.13	0.0303
87 to 93	6	6.66	0.0654
Totals	310	309.18	0.7649

The value of our χ^2 test statistics is

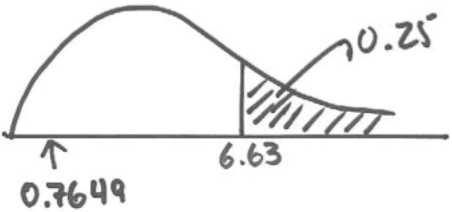
$$\chi^2 = \sum \frac{(O-E)^2}{E} = 0.7649$$

Normality Test Example (cont.)

5. Find the P-value of test statistic

$$\begin{aligned}
\text{P-value} &= P(\chi^2 \geq \text{Test statistic}|H_0) \\
&= P(\chi^2 \geq 0.7649|H_0) \\
\text{d.f.} &= k - 1 \\
&= 6 - 1 = 5.
\end{aligned}$$

- Look at the Table C row for d.f. = 5.
  Try to find test statistic value = 0.7649.
- We see that $P(\chi^2 \geq 6.63) = 0.25$



Normality Test Exampe (cont.)

Thus,

$$P(\chi^2 \geq 0.7649) > 0.25,$$

i.e., the

$$\text{P-value} > 0.25.$$

6. Make a decision

   Reject H_0 if $\text{P-value} \leq \alpha$.
   Do not reject H_0 if $\text{P-value} > \alpha$.

   In our example, $\alpha = 0.05$ and the $\text{P-value} > 0.25$.
   Thus, the $\text{P-value} > 0.05$.
   Do not reject H_0.

Normality Test Example (cont.)

7. Conclusion:
   At the 5% significance level, the observed average daily July temperatures during the past 10 years appears to follow a Normal distribution.

Conditions:
Independent observations
E > 5 for each cell.

Association Of Categorical Variables

Introduction

Recall: Categorical variable have non-numeric values which describe attributes, classes, or categories.

Example: Class Rank has values:

FR. SO. JR. SR. OTHER

The inferential methods discussed in Chap. 8 and Chap. 9 are appropriate for numeric variables.

What if our variables are categorical?

We can use contingency tables and χ^2 (Chi-square) tests to determine if two categorical variables are associated.

Contingency Tables

We can organize categorical data in a contingency table, with r rows and c columns, called an $r \times c$ contingency table.

2×2 Contingency Table Example

		Political Preference	
		Democrat	Republican
Gender	Male	60	40
	Female	90	30

More General Contingency Tables

We can also have 2×3, 4×5, or any $r \times c$ contingency tables, where $r \geq 2$ and $c \geq 2$.

2×3 Contingency Table Example

		Political Preference		
		Democrat	Republican	Other
Gender	Male	40	45	15
	Female	60	30	10

χ^2 Test for Independence

2×2 Contingency Table Example

	Democratic	Republican	Row Totals
Male	60 (68.2)	40 (31.8)	100
Female	90 (81.8)	30 (38.2)	120
Column Totals	150	70	220

H_0 : The variables “gender” and “political preference” are independent.

H_1 : The variables “gender” and “political preference” are dependent.

(Use $\alpha = 0.05$.)

2×2 Contingency Table Example (cont.)

We must calculate the expected cell frequencies, i.e., E. For each of the 4 cells,

$$E = \frac{\text{row total} \times \text{column total}}{n}.$$

- For the upper left cell: $E = \frac{100 \times 150}{220} = 68.2$
- For the lower left cell: $E = \frac{120 \times 150}{220} = 81.2$
- For the upper right cell: $E = \frac{100 \times 70}{220} = 31.8$
- For the lower right cell: $E = \frac{120 \times 70}{220} = 38.2$

2×2 Contingency Table Example (cont.)

For a 2×2 contingency table, the "test statistics" is

$$\chi^{2^*} = \sum \frac{(|O - E| - \frac{1}{2})^2}{E},$$

where O is the observed cell frequency.

$$\begin{aligned}
\chi^{2^*} &= \frac{(|60 - 68.2| - 0.5)^2}{68.2} + \frac{(|40 - 31.8| - 0.5)^2}{31.8} \\
&\quad + \frac{(|90 - 81.8| - 0.5)^2}{81.8} + \frac{(|30 - 38.2| - 0.5)^2}{38.2} \\
&= \frac{(8.2 - 0.5)^2}{68.2} + \frac{(8.2 - 0.5)^2}{31.8} + \frac{(8.2 - 0.5)^2}{81.8} \\
&\quad + \frac{(8.2 - 0.5)^2}{38.2} \\
&= 0.87 + 1.86 + 0.72 + 1.55 \\
&= 5.01
\end{aligned}$$

2×2 Contingency Table Example (cont.)

Find the "critical value" from Table C.

$$\begin{aligned} \text{d.f.} &= (r-1)(c-1) \\ &= (2-1)(2-1) = 1 \\ \alpha &= 0.05 \\ \chi^2 &= 3.84 \end{aligned}$$

We will reject H_0 if

$$\chi^{2^*} > \chi^2.$$

Since

$$5.01 > 3.84,$$

we reject H_0.

2×2 Contingency Table Example (cont.)

We conclude that there does seem to be an association between “gender” and “political preference” using $\alpha = 0.05$

Assumptions:

1. The observations are independent
   (Use random sample to ensure this)
2. The values in the contingency table are frequencies, not percentages.
3. E > 5 for each cell
   (otherwise, perform a different test)

χ^2 Test For Independence

3×3 Contingency Table Example

Does "test failure" reduce academic aspirations and thereby contribute to the decision to drop out of school?

A survey of 283 students randomly selected from schools with low graduation rates. The contingency table below reports results to the question: Do tests required for graduation discourage students from staying in school?

Does there appear to be a relation between school locations and the student responses?
(Use $\alpha = 0.05$.)

3×3 Contingency Table Example (Cont.)

		Location			Row
		Urban	Suburb.	Rural	Totals
	Yes	57 (57.86)	27 (31.48)	47 (41.66)	131
Res-ponse	No	23 (22.53)	16 (12.25)	12 (16.22)	51
	Unsure	45 (44.61)	25 (24.27)	31 (32.12)	101
	Column Totals	125	68	90	283

H_0 : School location and student response are independent.

H_1 : School location and student response are dependent.

3×3 Contingency Table Example (Cont.)

For $r \times c$ tables, the test statistics is not the same as for 2×2 tables.

Test Statistics (for $r \times c$ tables):

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

where O represents the observed frequencies. The expected cell frequencies (E) are computed as usual:

$$\mathrm{E} = \frac{\text{row total} \times \text{column total}}{n}$$

For our example,

$$\begin{aligned} \chi^{2*} &= \frac{(57-57.86)^2}{57.86} + \ldots + \frac{(31-32.12)^2}{32.12} \\ &= 0.013 + 0.638 + \ldots + 1.148 + \ldots + 0.039 \\ &= 3.655 \end{aligned}$$

3×3 Contingency Table Example (Cont.)

Critical Value: Table C

$$\begin{aligned} \text{d.f.} &= (r-1)(c-1) \\ &= (3-1)(3-1) = 4 \\ \alpha &= 0.05 \\ \chi^2 &= 9.49 \end{aligned}$$

We will reject H_0 if

$$\chi^{2^*} > \chi^2.$$

Since,

$$3.655 \not> 9.49,$$

we do not reject H_0.
We conclude that the evidence is not sufficient to show a significant association between school location and response.
Assumptions: Same as for 2×2 table.