Stat 250 Gunderson Lecture Notes
Relationships between Categorical Variables
12: Chi-Square Analysis

Inference for Categorical Variables

Having now covered a lot of inference techniques for quantitative responses, we return to analyzing categorical data, that is, analyzing count data. The three main tests described in the text that we will cover are:

1. ***Goodness of Fit Test:*** this test is for assessing if a particular discrete model is a good fitting model for a discrete characteristic, based on a random sample from the population.
   E.g. Has the model for the method of transportation (drive, bike, walk, other) used by students to get the class changed from that for 5 years ago?

2. ***Test of Homogeneity***: this test is for assessing if two or more populations are homogeneous (alike) with respect to the distribution of some discrete (categorical) variable.
   E.g. Is the distribution of opinion on legal gambling the same for adult males versus adult females?

3. ***Test of Independence***: this test helps us to assess if two discrete (categorical) variables are independent for a population, or if there is an association between the two variables.
   E.g. Is there an association between satisfaction with the quality of public schools (not satisfied, somewhat satisfied, very satisfied) and political party (Republican, Democrat, etc.)

The first test is the one-sample test for count data. The other two tests (homogeneity and independence) are actually the same test. Although the hypotheses are stated differently and the underlying assumptions about how the data is gathered are different, the steps for doing the two tests are exactly the same.

All three tests are based on an X^2 test statistic that, if the corresponding H_0 is true and the assumptions hold, follows a **chi-square distribution** with some degrees of freedom, written $\chi^2(df)$. So our first discussion is to learn about the chi-square distribution - what the distribution looks like, some facts, how to use Table A.5 to find various percentiles.

The Chi-Square Distribution

General Shape:
If we have a chi-square distribution with
df = degrees of freedom,
then the ...

Mean is equal to _____

Variance is equal to _____

Standard deviation
is equal to _____

These facts will serve as a useful frame of reference for making decision.

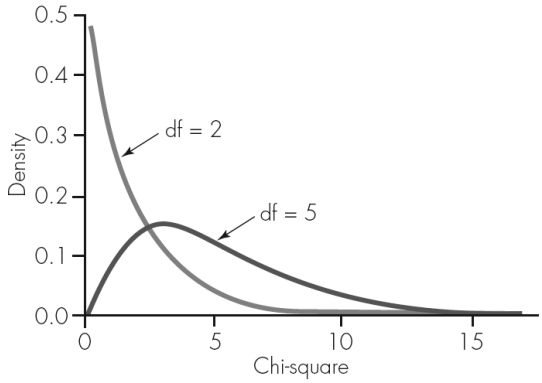

Figure 15.2 ▮ Two different chi-square distributions

All images PD-INEL

Table A.5 provides some upper-tail percentiles for chi-square distributions.

	p = Area to Right of Chi-square Value								
df	*0.50*	*0.25*	*0.10*	*0.075*	*0.05*	*0.025*	*0.01*	*0.005*	*0.001*
1	0.45	1.32	2.71	3.17	3.84	5.02	6.63	7.88	10.83
2	1.39	2.77	4.61	5.18	5.99	7.38	9.21	10.60	13.82
3	2.37	4.11	6.25	6.90	7.81	9.35	11.34	12.84	16.27
4	3.36	5.39	7.78	8.50	9.49	11.14	13.28	14.86	18.47
5	4.35	6.63	9.24	10.01	11.07	12.83	15.09	16.75	20.51
6	5.35	7.84	10.64	11.47	12.59	14.45	16.81	18.55	22.46
7	6.35	9.04	12.02	12.88	14.07	16.01	18.48	20.28	24.32
8	7.34	10.22	13.36	14.27	15.51	17.53	20.09	21.95	26.12
9	8.34	11.39	14.68	15.63	16.92	19.02	21.67	23.59	27.88
10	9.34	12.55	15.99	16.97	18.31	20.48	23.21	25.19	29.59

From Utts, Jessica M. and Robert F. Heckard. Mind on Statistics, Fourth Edition. 2012. Used with permission.

Try It! Consider the $\chi^2(4)$ distribution.

a. What is the mean for this distribution? _____

b. What is the median for this distribution? _____

c. How likely would it be to get a value of 4 or even larger?

d. How likely would it be to get a value of 10.3 or even larger?

The BIG IDEA

- The data consists of observed counts.
- We compute expected counts under the H_0 - these counts are what we would expect (on average) if the corresponding H_0 were true.
- Compare the observed and expected counts using the X^2 test statistic. The statistic will be a measure of how close the observed counts are to the expected counts under H_0. If this distance is large, we have support for the alternative H_a.

With this in mind, we turn to our first chi-square test of goodness of fit. We will derive the methodology for the test through an example. An overall summary of the test will be presented at the end.

Test of Goodness of Fit: Helps us assess if a particular discrete model is a good fitting model for a discrete characteristic, based on a random sample from the population.

Goodness of Fit Test

Scenario: We have one population of interest, say all cars exiting a toll road that has four booths at the exit.

Question: Are the four booths used equally often?

Data: One random sample of 100 cars, we record one variable *X,* which booth was used (1, 2, 3, 4). The table below summarizes the data in terms of the observed counts.

	Booth 1	**Booth 2**	**Booth 3**	**Booth 4**
Observed # cars	**26**	**20**	**28**	**26**

Note: This is only a one-way frequency table, not a two-way table as will be in the homogeneity and independence tests. We use the notation *k* = the number of categories or cells, here $k = 4$.

The null hypothesis:

Let p_i = (population) proportion of cars using booth i

H_0: p_1 = _____ , p_2 = _____ , p_3 = _____ , p_4 = _____ .

H_a: _____

The null hypothesis specifies a particular discrete model (mass function) by listing the proportions (or probabilities) for each of the k outcome categories.

The one-way table provides the OBSERVED counts. Our next step is to compute the EXPECTED counts, under the assumption that H_0 is true.

How to find the expected counts?

There were 100 cars in the sample and 4 booths.

If the booths are used equally often, H_0 is true, then we would expect

... _____ cars to use Booth #1

... _____ cars to use Booth #2

... _____ cars to use Booth #3

... _____ cars to use Booth #4

Expected Counts $= E_i = np_i$

Enter these expected counts in the parentheses in the table below.

	Observed Counts (Expected Counts)}			
	Booth 1	**Booth 2**	**Booth 3**	**Booth 4**
Number of cars	**26 ( )**	**20 ( )**	**28 ( )**	**26 ( )**

The X^2 test statistic

Next we need our test statistic, our measure of how close the observed counts are to what we expect under the null hypothesis.

X^2 =

Do you think a value of X^2 =_____ is large enough to reject H_0?

Let's find the *p*-value, the probability of getting an X^2 test statistic value as large or larger than the one we observed, assuming H_0 is true. To do this we need to know the distribution of the X^2 test statistic under the null hypothesis.

If H_0 is true, then X^2 has the χ^2 distribution with degrees of freedom = _____.

Find the p-value for our tollbooth example:

Observed X^2 test statistic value = _____ df = _____ .

Are the results statistically significant at the 5% significance level?
Conclusion at a 5% level: It appears that ....

Aside: Using our frame of reference for chi-square distributions.
Recall that if we have a chi-square distribution with df = degrees of freedom, then the mean is equal to df, and the standard deviation is equal to $\sqrt{2(df)}$

So, if H_0 were true, we would expect the X^2 test statistic to be about _____
give or take about _____ .

Since we reject H_0 for large values of X^2, and we only observed a value of _____, even less than expected under H_0, we certainly do not have enough evidence to reject H_0.

Goodness of Fit Test Summary

Assume: We have 1 random sample of size n.
We measure one discrete response *X* that has k possible outcomes

***Test*:** H_0: A specified discrete model for *x* → $p_1 = p_{10}, \quad p_2 = p_{20}, \quad \ldots, \quad p_k = p_{k0}$
H_a: The probabilities are not as specified in the null hypothesis.

Test Statistic: $X^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$
where expected $= E_i = np_{i0}$

If H_0 is true, then X^2 has a χ^2 distribution with $(k-1)$ degrees of freedom, where k is the number of categories. The necessary conditions are: at least 80% of the expected counts are greater than 5 and none are less than 1. Be aware of the sample size (pg 656).

Try It! Crossbreeding Peas

For a genetics experiment in the cross breeding of peas, Mendel obtained the following data in a sample from the second generation of seeds resulting from crossing yellow round peas and green wrinkled peas.

Yellow Round	Yellow Wrinkled	Green Round	Green Wrinkled
315	101	108	32

Do these data support the theory that these four types should occur with probabilities 9/16, 3/16, 3/16, and 1/16 respectively? Use $\alpha = 0.01$.

Try It! Desired Vacation Place

The AAA travel agency would like to assess if the distribution of *desired vacation place* has changed from the model of 3 years ago. A random sample of 928 adults were polled by the polling company *Ipsos* during this past mid-May. One question asked was "Name the one place you would want to go for vacation if you had the time and the money." The table displays the model for the distribution of desired vacation place 3 years ago and the observed results based on the recent poll.

	1 = Hawaii	2 = Europe	3 = Caribbean	4 = Other	Totals
Model 3 years ago	10%	40%	20%	30%	100%
Obs Counts from poll	124	390	125	289	928

a. Give the null hypothesis to test if there has been a significant change in the distribution of desired vacation place from 3 years ago.

b. The observed test statistic is ~ 31 and the *p*-value is less than 0.001. Interpret this *p*-value in terms of repeated random samples of 928 adults.

Test of Homogeneity: Helps us to assess if the distribution for one discrete (categorical) variable is the same for two or more populations.

Test of Homogeneity

Scenario: We have 2 populations of interest; preschool boys and preschool girls.
Question: Is Ice Cream Preference the same for boys and girls?

Data: 1 random sample of 75 preschool boys,
1 random sample of 75 preschool girls;
the two random samples are independent.

The table below summarizes the data in terms of the observed counts.
Observed Counts:

Ice Cream Preference	Boys	Girls
Vanilla (V)	**25**	**26**
Chocolate (C)	**30**	**23**
Strawberry (S)	**20**	**26**

Note: The column totals here were known in advance, even before the ice cream preferences were measured. This is a key idea for how to distinguish between the test of homogeneity and the test of independence.

The **null hypothesis**:

H_0: The distribution of ice cream preference is the **same** for the two populations, boys and girls.

A more mathematical way to write this null hypothesis is:

H_0: $P(X = i \mid population\ j) = P(X = i)$ for all i, j
where X is the categorical variable, in this case, ice cream preference.

As we can see, the null hypothesis is stating that the distribution of ice cream preference does not depend on (is independent of) the population we select from since the two distributions are the same.

The null hypothesis looks like: $P(A \mid B) = P(A)$, which is one definition of independent events, from our previous discussion of independence. This is why the test of homogeneity (comparing several populations) is really the same as the test of independence. The assumptions are different however.

For our homogeneity (comparing several populations) test, we assume we have independent random samples, one from each population, and we measure 1 discrete (categorical) response. For the independence test (discussed later) we will assume we have just 1 random sample from 1 population, but we measure 2 discrete (categorical) responses.

Getting back to **ICE CREAM** ... The table provides the OBSERVED counts. Our next step is to compute the EXPECTED counts, under the assumption that H_0 is true.

How to find the expected counts?

Let's look at those who preferred Strawberry first.

Strawberry: Since there were _____ children who preferred *Strawberry* overall, if the distributions for boys and girls are the same (H_0 is true),

then we would expect _____ of these children to be boys

and the remaining _____ of these children to be girls.

Note that our sample sizes were the same, 75 boys and 75 girls, 50% of each. If they were not 50-50, we would have to adjust the expected counts accordingly. Let's do the same for the Vanilla and Chocolate preferences.

Chocolate: Since there were _____ children who preferred Chocolate overall, if the distributions for boys and girls are the same (H_0 is true), then we would expect _____ of these children to be boys

and the remaining _____ of these children to be girls.

Vanilla: Since there were _____ children who preferred Vanilla overall, if the distributions for boys and girls are the same (H_0 is true), then we would expect _____ of these children to be boys

and the remaining _____ of these children to be girls.

Enter these expected counts in the parentheses in the table below.

Observed Counts (Expected Counts)

Ice Cream Preference	Boys	Girls	Total
Vanilla (V)	**25( )**	**26 ( )**	51
Chocolate (C)	**30( )**	**23 ( )**	53
Strawberry (S)	**20( )**	**26 ( )**	46
Total	75	75	150

A Closer Look at the Expected Counts:

Let's look at how we actually computed an expected count so we can develop a general rule: If H_0 were true (i.e., no difference in preferences for boys versus girls), then our *best* estimate of the P(a child prefers vanilla) = 51/150. Since we had 75 boys, under no difference in preference, we would expect 75 x (51/150) to prefer vanilla. That is, the expected number of boys preferring vanilla = $\frac{(75)(51)}{150} = \frac{(\text{row total})(\text{column total})}{\text{Total n}}$. This quick recipe for computing the expected counts under the null hypothesis is called the **Cross-Product Rule**.

The X^2 test statistic

Next we need to compute our test statistic, our measure of how close the observed counts are to what we expect under the null hypothesis. Below we are provided the first contribution to the test statistic value. Determine the remaining contributions which are summed to get the value.

$$X^2 = \frac{(25 - 25.5)^2}{25.5} + \ldots$$

The larger the test statistic, the "bigger" the differences between what we observed and what we would expect to see if H_0 were true. So the larger the test statistic, the more evidence we have against the null hypothesis.

Is a value of X^2 =_____ large enough to reject H_0?

We need to find the *p*-value, the probability of getting an X^2 test statistic value as large or larger than the one we observed, assuming H_0 is true. To do this we need to know the distribution of the X^2 test statistic under the null hypothesis.

If H_0 is true, then X^2 has the χ^2 distribution with degrees of freedom = _____

Brief motivation for the degrees of freedom formula:

***Find the p-value* for our ice cream example:**

Observed X^2 test statistic value = _____ df = _____

Decision at a 5% significance level: (circle one) **Reject H_0** **Fail to reject H_0**

Conclusion: It appears that ....

Test of Homogeneity Summary (Comparison of Several Populations)

Assume: We have C independent random samples of size $n_1, n_2, \ldots, n_c$ from C populations.
We measure 1 discrete response X that has r possible outcomes.

***Test*:**
H_0: The distribution for the response variable X is the same for all populations.

***Test Statistic*:** $X^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$

where expected $= \frac{(\text{row total})(\text{column total})}{\text{Total } n}$

If H_0 is true, then X^2 has a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. The necessary conditions are: at least 80% of the expected counts are greater than 5 and none are less than 1.

Try It! What is your Decision?

For a chi-square test of homogeneity, there are 3 populations and 4 possible values of the discrete characteristic.

If H_0 is true, that is, the distribution for the response is the same for all 3 populations, what is the expected value of the test statistic?

Try It! Treatment for Shingles

An article had the headline "For adults, chicken pox vaccine may stop shingles". A clinical trial was conducted in which 420 subjects were randomly assigned to receive the chicken pox vaccine or a placebo vaccine. Some side effects of interest were swelling and rash around the injection site. Consider the following results for the swelling side effect.

	Major Swelling	Minor Swelling	No Swelling
Vaccine	54	42	134
Placebo	16	32	142

```
        Pearson's Chi-squared test

data:  .Table
X-squared = 18.5707, df = 2, p-value = 9.277e-05
```

a. Give the name of the test to be used for assessing if the distribution of swelling status is the same for the two treatment populations.

b. Based on the above data, among those chicken pox vaccinated subjects, what percent had major swelling around the injection site?

c. Based on the above data, among those placebo vaccinated subjects, what percent had major swelling around the injection site?

d. Assuming the distribution of swelling status is the same for the two treatment populations, how many chicken pox vaccinated subjects would you expect to have major swelling around the injection site? ***Show your work.***

e. Compute the contribution to the Chi-square test statistic based on those vaccinated subjects who had major swelling around the injection site.

f. Use a level of 0.05 to assess if the distribution of swelling status is the same for the two treatment populations.
   Test Statistic Value: _____ *p*-value: _____
   Thus, the distribution of swelling status (circle your answer): **does** **does not**
   appears to be the same for the two treatment populations.

Test of Independence: Helps us to assess if two discrete (categorical) variables are independent for a population, or if there is an association between the two variables.

Test of Independence

Scenario: We have one population of interest - say factory workers.

Question: Is there a relationship between smoking habits and whether or not a factory worker experiences hypertension?

Data: 1 random sample of 180 factory workers, we measure the two variables:
Y = hypertension status (yes or no)
X = smoking habit (non, moderate, heavy)

The table below summarizes the data in terms of the observed counts.
Observed Counts:

		X= Smoking Habit		
Y=		Non	Mod	Heavy
Hyper	Yes	21	36	30
Status	No	48	26	19

Get the row and column totals.
Note: neither row nor column totals were known in advance before measuring hypertension and smoking habit. We only know the overall total of 180.

The **null hypothesis**:
H_0: There is no association between smoking habit and hypertension status for the population of factory workers.
(or The two factors, smoking habit and hypertension status, are independent for the population.)

One more mathematical way to write this null hypothesis is:
H_0: $P(X = i \text{ and } Y = j) = P(X = i)P(Y = j)$

The null hypothesis looks like: $P(A \text{ and } B) = P(A)P(B)$, which is one definition of independent events, from our previous discussion of independence.

Getting back to our FACTORY WORKERS …

The two-way table provides the OBSERVED counts. Our next step is to compute the EXPECTED counts, under the assumption that H_0 is true. The expected counts and the test statistic are found the same way as for the homogeneity test.

Cross-Product Rule: Expected Counts $= \dfrac{(\text{row total})(\text{column total})}{\text{Total } n}$

Compute and enter these expected counts in the parentheses in the table below.

Observed Counts (Expected Counts):

Y=		X= Non	Smoking Mod	Habit Heavy	
Hyper	Yes	21 ( )	36 ( )	30 ( )	87
Status	No	48 ( )	26 ( )	19 ( )	93
		69	62	49	180

The X^2 test statistic

Our measure of how close the observed counts are to what we expect under the null hypothesis.

$$X^2 = \frac{(21 - 33.35)^2}{33.35} + \ldots$$

Do you think a value of X^2 =_____ is large enough to reject H_0?

The next step is to find the *p*-value, the probability of getting an X^2 test statistic value as large or larger than the one we observed, assuming H_0 is true. To do this we need to know the distribution of the X^2 test statistic under the null hypothesis.

If H_0 is true, then X^2 has the χ^2 distribution with degrees of freedom =_____

Aside: Using our frame of reference for chi-square distributions.

If H_0 were true, we would expect the X^2 test statistic to be about _____

give or take about _____ .

About how many standard deviations is the observed X^2 value of 14.5 from the expected value under H_0? What do you think the decision will be?

Find the *p*-value for our factory worker example:

Observed X^2 test statistic value = _____ df = _____

Find the *p*-value and use it to determine if the results are statistically significant at the 1% significance level.

Conclusion at a 1% level: It appears that ....

Test of Independence Summary

Assume: We have 1 random sample of size n.
We measure 2 discrete responses:
X which has r possible outcomes
and *Y* which has c possible outcomes.

***Test*:** H_0: The two variables *X* and *Y* are independent for the population.

***Test Statistic*:** $X^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$

where expected $= \frac{(\text{row total})(\text{column total})}{\text{Total } n}$

If H_0 is true, then X^2 has a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. The necessary conditions are: at least 80% of the expected counts are greater than 5 and none are less than 1.

Relationship between Age Group and Appearance Satisfaction

Are you satisfied with your overall appearance? A random sample of 150 women were surveyed. Their answer to this question (very, somewhat, not) was recorded along with their age category (1 = under 30, 2 = 30 to 50, and 3 = over 50).

R was used to generate the following output from the data.

```
                        Under 30 30 to 50 Over 50
Very Satisfied                20       10      16
Somewhat Satisfied            18       20      18
Not Satisfied                 10       29       9

        Pearson's Chi-squared test

             data:  .Table
X-squared = 15.478, df = 4, p-value = 0.003805
```

a. Give the name of the test to be used for assessing if there is a relationship between age group and appearance satisfaction.

b. Assuming there is no relationship between age group and appearance satisfaction, how many old women (over 50) would you expect to be very satisfied with their appearance?

c. Compute the contribution to the Chi-square test statistic based on the older women (over 50) who were very satisfied with their appearance.

d. Assuming there is no relationship between age group and appearance satisfaction, what is the expected value of the test statistic?

e. Use a level of 0.05 to assess if there is a significant relationship between age group and appearance satisfaction.

   Test Statistic Value: _____ *p*-value: _____

   Thus, there (circle your answer): **does** **does not**

   appear to be an association between age group and appearance satisfaction.

2x2 Tables – a special case of the two proportion z test

- The z-test for comparing two population proportions is the same as the chi-square test provided the alternative is two-sided. The z-test would need to be performed for one-sided alternatives.
- When the conditions for the z-test or chi-square test are not met (sample sizes too small) there is another alternative test called the Fisher's Exact Test.

Stat 250 Formula Card:

Chi-Square Tests	
Test of Independence & Test of Homogeneity	**Test for Goodness of Fit**
Expected Count $E = \text{expected} = \frac{\text{row total} \times \text{column total}}{\text{total } n}$	**Expected Count** $E_i = \text{expected} = np_{i0}$
Test Statistic $X^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ df = $(r-1)(c-1)$	**Test Statistic** $X^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ df = $k-1$
If Y follows a $\chi^2(df)$ distribution, then E(Y) = df and Var(Y) = 2(df).	

Additional Notes

A place to … jot down questions you may have and ask during office hours, take a few extra notes, write out an extra problem or summary completed in lecture, create your own summary about these concepts.