Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 13:

Two-way table for categorical data; Chi-square test for two-way table; Chi-square goodness of fit test;

Categorical response or predictor variables with more than 2 levels

- We previously covered methods and contrast measures for assessing associations between a binary response variable and binary predictor variable
  - Risk Difference
  - Risk Ratios
  - Odds Ratios
- What about when a categorical response variable and/or categorical predictor variables has more than two categories?
- How can we evaluate the evidence for an association in this setting?
  - Answer: We can assess evidence that the two categorial variables are independent.

Example: Exercise and Living Arrangement

- A survey of graduate students at a **U**niformly **W**onder institution was conducted and information was collected on how frequently they exercised and their living arrangements.
- There were 470 graduate students who responded to the survey. The following **two-way table** summarizes the data:

Living Arrangement	Exercise Level			Total
	No Regular Exercise	Sporadic Exercise	Regular Exercise	
Dormitory	32	30	28	90
On-Campus Apartment	74	64	42	180
Off-Campus Apartment	110	25	15	150
At Home	39	6	5	50
Total	255	125	90	470

- Suppose we are interested in determining whether there is an association between the row variable (living arrangement) and the column variable (exercise level) for a $r \times c$ contingency table.

Example: Exercise and Living Arrangement

Living Arrangement	Exercise Level			Total
	No Regular Exercise	Sporadic Exercise	Regular Exercise	
Dormitory	32	30	28	90
On-Campus Apartment	74	64	42	180
Off-Campus Apartment	110	25	15	150
At Home	39	6	5	50
Total	255	125	90	470

- Formally, were are interesting is assessing the following null and alternative hypotheses:
  - H_0: there is *no association* between the row variable and column variable
  - H_a: there *is* an association between the two variables

  The alternative hypothesis H_a does not specify any particular direction of the association because. It includes all of the many kinds of association that are possible, and as a result, we cannot describe H_a as either one-sided or two-sided.

Example: Exercise and Living Arrangement

- How would we test the hypotheses?
- Intuition for a test: Suppose H_0 is true. If the two variables are independent, i.e., not associated, what counts would we expect to observe?
- Recall that under the independence assumption,

$$P(A \text{ and } B) = P(A)P(B)$$

Example: Exercise and Living Arrangement

Living Arrangement	Exercise Level: No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Dormitory	32	30	28	90
On-Campus Apartment	74	64	42	180
Off-Campus Apartment	110	25	15	150
At Home	39	6	5	50
Total	255	125	90	470

- Let n be the total number graduate students in the study. Assuming independence, the expected number of graduate students who live in a Dormitory and do not get regular exercise is:

$$n \times P(\text{Lives in Dormitory and No Regular Exercise})$$

$$= n \times P(\text{Lives in Dormitory}) \times P(\text{No Regular Exercise})$$

$$= 470\left(\frac{90}{470}\right)\left(\frac{255}{470}\right) = \frac{(90)(255)}{470} = 48.83$$

Exercise and Living Arrangement: Expected Counts

- So under the null hypothesis of no association between the row variable and the column variable in the table, we have

$$\text{Expected Cell Count} = (\text{Row Total} \times \text{Column Total})/\text{Total Count}$$

- Can easily fill in the expected cell counts under the null hypothesis for the entire contingency table:

Living Arrangement	Exercise Level			Total
	No Regular Exercise	Sporadic Exercise	Regular Exercise	
Dormitory	48.8	23.9	17.2	90
On-Campus Apartment	97.7	47.9	34.5	180
Off-Campus Apartment	81.4	39.9	28.7	150
At Home	27.1	13.3	9.6	50
Total	255	125	90	470

Exercise and Living Arrangement: Observed vs. Expected Counts

- The observed counts are:

Living Arrangement	Exercise Level: No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Dormitory	32	30	28	90
On-Campus Apartment	74	64	42	180
Off-Campus Apartment	110	25	15	150
At Home	39	6	5	50
Total	255	125	90	470

- The expected counts (under the Null Hypothesis) are:

Living Arrangement	Exercise Level: No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Dormitory	48.8	23.9	17.2	90
On-Campus Apartment	97.7	47.9	34.5	180
Off-Campus Apartment	81.4	39.9	28.7	150
At Home	27.1	13.3	9.6	50
Total	255	125	90	470

- Our test for an association will be based on a measure of *how far the observed table is from the expected table.*

The Pearson χ^2-test for a $r \times c$ Table

- We will use a Pearson χ^2 test for independence of row variable and column variable. The test statistic is:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed cell} - \text{Expected cell})^2}{\text{Expected cell}}$$

- What is the the expected cell number under H_0? For each cell, we have

$$\text{Expected Cell Count} = \frac{\text{row total} \times \text{col total}}{\text{total count}}$$

- Under H_0, the X^2 test statistic has an approximate χ^2 distribution with $(r-1)(c-1) = (4-1)(3-1) = 6$ degree of freedom

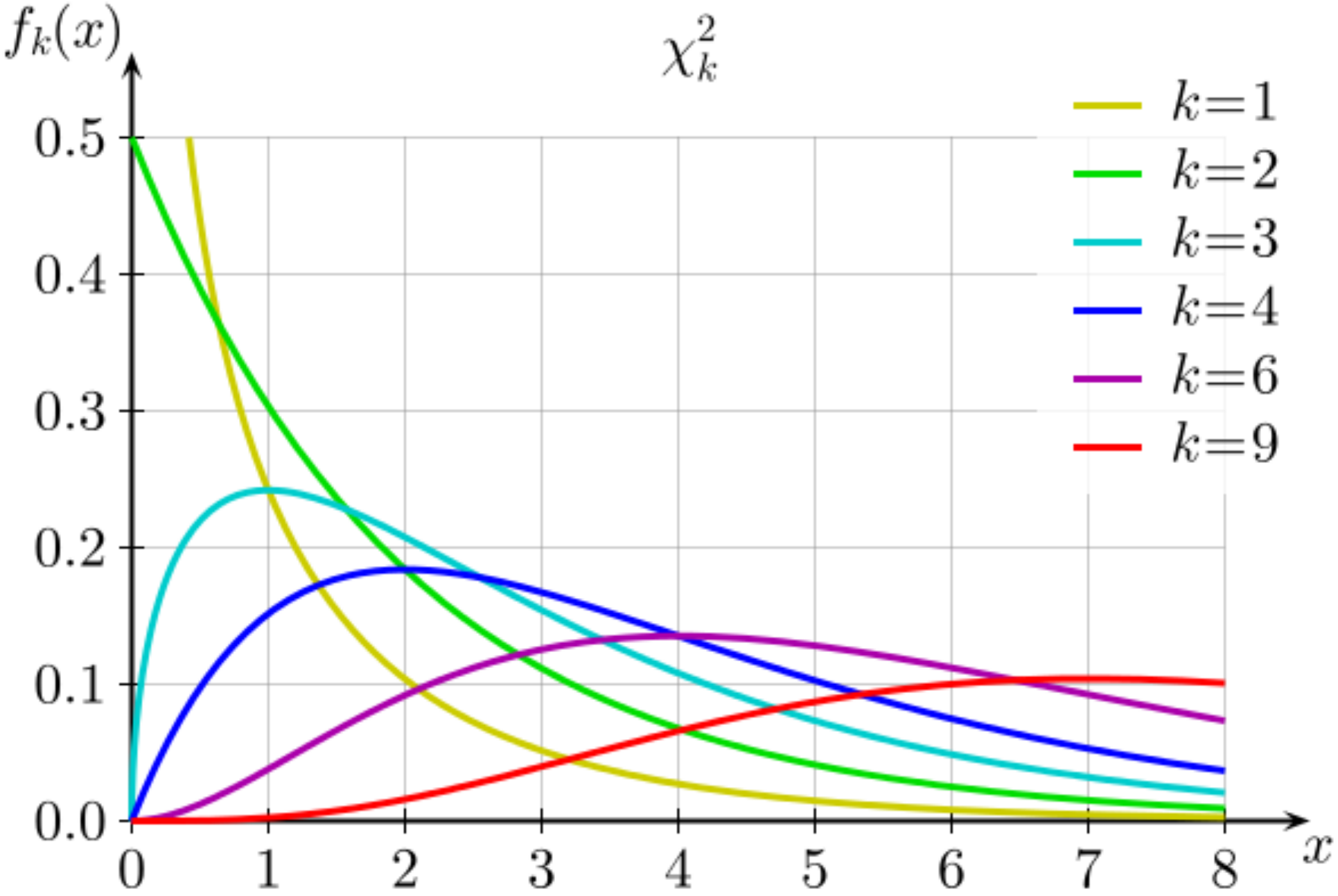
The χ^2 Distribution

- Under H_0, the X^2 test statistic has an approximate χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, denoted $\chi^2_{(r-1)(c-1)}$.
- Why $(r-1)(c-1)$?
- Recall that our "expected" table is based on some quantities estimated from the data: namely the row and column totals.
- Once these totals are known, filling in any $(r-1)(c-1)$ undetermined table entries actually gives us the whole table. Thus, there are only $(r-1)(c-1)$ freely varying quantities in the table.

The χ^2 Distribution

What does the χ^2 distribution look like?

- Unlike the Normal or t distributions, the χ^2 distribution takes values in $(0, \infty)$.
- As with the t distribution, the exact shape of the χ^2 distribution depends on its degrees of freedom k.


$f_k(x)$
χ^2_k
$k=1$
$k=2$
$k=3$
$k=4$
$k=6$
$k=9$
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
x

p-Value for the Pearson χ^2-Test

- If the observed and expected counts are very different, X^2 will be large, indicating evidence against H_0. Thus, the p-value is always based on the right-hand tail of the distribution.
- *There is no notion of a two-tailed test in this context.*
- The p-value is therefore

$$P(\chi^2_{(r-1)(c-1)} \geq X^2)$$

Sample Size Requirements for Pearson χ^2-Test

- Recall that X^2 has an *approximate* $\chi^2_{(r-1)(c-1)}$ distribution. When is the approximation valid?
- For any two-way table larger than 2×2, we require that the average expected cell count is at least 5 and each expected count is at least one (note: some statistical software might give a warning for low cell counts).
- For 2×2 tables, we require that each expected count be at least 5.

Exercise and Living Arrangement Example cont.....

- The observed counts are:

Living Arrangement	Exercise Level: No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Dormitory	32	30	28	90
On-Campus Apartment	74	64	42	180
Off-Campus Apartment	110	25	15	150
At Home	39	6	5	50
Total	255	125	90	470

- The expected counts (under the Null Hypothesis) are:

Living Arrangement	Exercise Level: No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Dormitory	48.8	23.9	17.2	90
On-Campus Apartment	97.7	47.9	34.5	180
Off-Campus Apartment	81.4	39.9	28.7	150
At Home	27.1	13.3	9.6	50
Total	255	125	90	470

The χ^2 Test Statistic and P-value

$$
\begin{aligned}
X^2 &= \frac{(32-48.8)^2}{48.8} + \frac{(30-23.9)^2}{23.9} + \frac{(28-17.2)^2}{17.2} \\
&+ \frac{(74-97.7)^2}{97.7} + \frac{(64-47.9)^2}{47.9.9} + \frac{(42-34.5)^2}{34.5} \\
&+ \frac{(110-81.4)^2}{81.4} + \frac{(25-39.9)^2}{39.9} + \frac{(15-28.7)^2}{28.7} \\
&+ \frac{(39-27.1)^2}{27.1} + \frac{(6-13.3)^2}{13.3} + \frac{(5-9.6)^2}{9.6} \\
&= 60.5
\end{aligned}
$$

- What is the p-value? Will need to use statistical software (such as R) to obtain p-value:

$$P(\chi^2_{(4-1)(3-1} \geq 60.5) = P(\chi^2_6 \geq 60.5) = 3.56e-11$$

Can also use R to do the calculations

```
> observed<-matrix(c(32,30,28, 74, 64, 42,110,25,15,39,6,5),nrow=4,byrow=TRUE)
> observed
     [,1] [,2] [,3]
[1,]   32   30   28
[2,]   74   64   42
[3,]  110   25   15
[4,]   39    6    5
> expected<-matrix(c(48.8,23.9,17.2,97.7,47.9, 34.5, 81.4, 39.9, 28.7,27.1,13.3,9.6),nrow=4,byrow=TRUE)
> expected
     [,1] [,2] [,3]
[1,] 48.8 23.9 17.2
[2,] 97.7 47.9 34.5
[3,] 81.4 39.9 28.7
[4,] 27.1 13.3  9.6
> chi2stat<- sum((observed-expected)^2/expected); chi2stat
[1] 60.50188
> pval<- 1 - pchisq(chi2stat,df=6); pval
[1] 3.558975e-11
```

Pearson χ^2 test function in R

- The **chisq.*test()*** function in R can be used to perform the Pearson χ^2 test
- It takes as input the observed data in an *r* x *c* matrix (or table)
- Performs the Pearson χ^2 test
  - Provides the test statistics
  - p-value for the null hypothesis test that the row variable and the column variable are independent

Pearson χ^2 test in R: Example

```
> observed<-matrix(c(32,30,28, 74, 64, 42,110,25,15,39,6,5),nrow=4,byrow=TRUE)
> observed
     [,1] [,2] [,3]
[1,]   32   30   28
[2,]   74   64   42
[3,]  110   25   15
[4,]   39    6    5
> chisq.test(observed)

        Pearson's Chi-squared test

data:  observed
X-squared = 60.439, df = 6, p-value = 3.664e-11
```

- Note: the test statistics when using the **chisq.test()** function is slightly different from the test statistics value we obtained "by hand" due to rounding error when we calculated our "expected counts " table under the null hypothesis

Goodness of Fit Test

- The Pearson χ^2-test for independence is a special case of a **Goodness of Fit test** for counts of data
- With a goodness of fit test we compare the observed counts to the counts we expect if the null hypothesis is true
  - the test statistic is based on comparing the squared difference of observed counts to counts that are expected under the null hypothesis:
- If the model specified by H_0 describes the data well, observed values will tend to be close to expected values
- Under the null hypothesis, the test statistic, based on the appropriate sum of $(\text{observed} - \text{expected})^2$, has a chi-squared distribution

Chi-Square Goodness of Fit Test: Categorical variable

- Supposed we have data for n observations on a categorical variable with k possible outcomes that are summarized as observed counts $n_1, n_2, \ldots, n_k$ in k cells.
- A null hypothesis specifies probabilities $p_1, p_2, \ldots, p_k$ for the possible outcomes.
- For each cell, multiply the total number of observations n by the specified probability to determine the expected counts: expected count for outcome i is np_i

Chi-Square Goodness of Fit Test: Categorical variable

- The **chi-square statistic** measures how much the observed cell counts differ from the expected cell counts. The formula for the statistic is:

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

This statistic follows a χ^2 distribution with $k - 1$ degrees of freedom.

Goodness of Fit Test: Gambler Example

- A gambler is accused of using a loaded die, but he pleads innocent. A record has been kept of the last 60 throws and below are the observed counts for the data:

Die Value	Observed Count
1	4
2	6
3	17
4	16
5	8
6	9

- We can think of the above table of counts as a one-way table with six cells.
- If the gambler is using a fair die, what would the probability be for observing a 1 on a single throw? What about a 3 on a single throw?

Gambler Example (cont.)

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$$

$$H_a : p_i \neq \frac{1}{6} \text{ for at least one value } i$$

There were 60 rolls of the die:

Die Value	Observed Count	Expected
1	4	10
2	6	10
3	17	10
4	16	10
5	8	10
6	9	10

Gambler Example (cont.)

$$
\begin{aligned}
X^2 &= \frac{(4-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} \\
&\quad + \frac{(16-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-10)^2}{10} \\
&= \frac{142}{10} = 14.2
\end{aligned}
$$

- There are 6 possible outcomes, so there there are $6 - 1 = 5$ degrees of freedom
- Finally, the p-value can be obtained using the χ^2 distribution with 5 degrees of freedom (from R):

$$P(\chi^2_5 \geq 14.2) = 0.014$$

- What is your conclusion about the gambler?